Bayesian Data Analysis

Passenger satisfaction with Logistic Regression

Mohamed Afif Chifaoui

May. 21, 2024

Our dataset contains information about the passengers satisfaction during a trip of an airlines Company. This target variable is accompanied by other variables that describe the status of the trip. Our goal in this case study is to predict the satisfaction of a customer depending on predictor variables by means of a classical and Bayesian logistic regression.

SATISFACTION: It has 1 if satissfied. Otherwise, it is assigned to 0.

GOAL: study how trip status variables affect the probability of being satissfied, which is the dependent variable.

DATA EXPLORATION & PREPROCESSING

Let us first comprehend our data through a brief exploratory data analysis.

```
# Read data
data = read.csv('passengers.csv', header = TRUE)
head(data)
```

```
##   satisfaction Gender Age    Class Flight.Distance Seat.comfort
## 1    satisfied Female  53      Eco             519            4
## 2 dissatisfied Female  25      Eco            1145            3
## 3    satisfied Female  47      Eco             421            0
## 4    satisfied   Male  46 Business            3305            5
## 5    satisfied Female  53 Business            2097            3
## 6    satisfied   Male  59      Eco            1826            3
##   Departure.Arrival.time.convenient Food.and.drink Gate.location
## 1                                 4              4             4
## 2                                 2              3             4
## 3                                 1              1             3
## 4                                 5              5             5
## 5                                 3              3             3
## 6                                 4              4             4
##   Inflight.wifi.service Inflight.entertainment Online.support
## 1                     1                      5              2
## 2                     4                      3              4
## 3                     1                      1              1
## 4                     4                      4              5
## 5                     2                      5              5
## 6                     3                      3              3
##   Ease.of.Online.booking On.board.service Leg.room.service Baggage.handling
## 1                      3                3                4                3
## 2                      4                1                4                4
## 3                      1                3                5                3
## 4                      2                2                2                2
## 5                      5                5                5                5
## 6                      3                1                5                5
##   Checkin.service Cleanliness Online.boarding Departure.Delay.in.Minutes
## 1               4           3               4                         58
## 2               1           3               4                         24
## 3               1           5               1                          0
## 4               4           2               3                          0
## 5               3           5               4                          0
## 6               5           3               3                          0
##   Arrival.Delay.in.Minutes
## 1                       53
## 2                        9
## 3                        0
## 4                        0
## 5                        7
## 6                        0
```

```
summary(data)
```

```
##  satisfaction          Gender               Age            Class          
##  Length:2000        Length:2000        Min.   : 7.00   Length:2000       
##  Class :character   Class :character   1st Qu.:28.00   Class :character  
##  Mode  :character   Mode  :character   Median :40.00   Mode  :character  
##                                        Mean   :39.52                     
##                                        3rd Qu.:51.00                     
##                                        Max.   :80.00                     
##  Flight.Distance  Seat.comfort   Departure.Arrival.time.convenient
##  Min.   :  51    Min.   :0.000   Min.   :0.000                    
##  1st Qu.:1293    1st Qu.:2.000   1st Qu.:2.000                    
##  Median :1895    Median :3.000   Median :3.000                    
##  Mean   :1937    Mean   :2.837   Mean   :2.994                    
##  3rd Qu.:2537    3rd Qu.:4.000   3rd Qu.:4.000                    
##  Max.   :6099    Max.   :5.000   Max.   :5.000                    
##  Food.and.drink  Gate.location   Inflight.wifi.service Inflight.entertainment
##  Min.   :0.000   Min.   :1.000   Min.   :0.000         Min.   :0.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000         1st Qu.:2.000
##  Median :3.000   Median :3.000   Median :3.000         Median :4.000
##  Mean   :2.837   Mean   :3.033   Mean   :3.212         Mean   :3.373
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000         3rd Qu.:4.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000         Max.   :5.000
##  Online.support  Ease.of.Online.booking On.board.service Leg.room.service
##  Min.   :1.000   Min.   :1.000          Min.   :1.00     Min.   :0.000
##  1st Qu.:3.000   1st Qu.:2.000          1st Qu.:3.00     1st Qu.:2.000
##  Median :4.000   Median :4.000          Median :4.00     Median :4.000
##  Mean   :3.513   Mean   :3.494          Mean   :3.49     Mean   :3.496
##  3rd Qu.:5.000   3rd Qu.:5.000          3rd Qu.:5.00     3rd Qu.:5.000
##  Max.   :5.000   Max.   :5.000          Max.   :5.00     Max.   :5.000
##  Baggage.handling Checkin.service  Cleanliness    Online.boarding
##  Min.   :1.000    Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:3.000    1st Qu.:3.000   1st Qu.:3.000   1st Qu.:2.000
##  Median :4.000    Median :3.000   Median :4.000   Median :4.000
##  Mean   :3.717    Mean   :3.341   Mean   :3.721   Mean   :3.361
##  3rd Qu.:5.000    3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:4.000
##  Max.   :5.000    Max.   :5.000   Max.   :5.000   Max.   :5.000
##  Departure.Delay.in.Minutes Arrival.Delay.in.Minutes
##  Min.   :  0.00             Min.   :  0.00
##  1st Qu.:  0.00             1st Qu.:  0.00
##  Median :  0.00             Median :  0.00
##  Mean   : 15.83             Mean   : 16.05
##  3rd Qu.: 12.25             3rd Qu.: 13.00
##  Max.   :435.00             Max.   :470.00
```

```
all_columns <- names(data);all_columns
```

```
##  [1] "satisfaction"                      "Gender"
##  [3] "Age"                               "Class"
##  [5] "Flight.Distance"                   "Seat.comfort"
##  [7] "Departure.Arrival.time.convenient" "Food.and.drink"
##  [9] "Gate.location"                     "Inflight.wifi.service"
## [11] "Inflight.entertainment"            "Online.support"
## [13] "Ease.of.Online.booking"            "On.board.service"
## [15] "Leg.room.service"                  "Baggage.handling"
## [17] "Checkin.service"                   "Cleanliness"
## [19] "Online.boarding"                   "Departure.Delay.in.Minutes"
## [21] "Arrival.Delay.in.Minutes"
```

```
dim(data)
```

```
## [1] 2000   21
```

It seems that our data has initially 2000 observations and 21 variables. Let us explain our variables:

- **satisfaction**: Factor variable indicating whether the customer was satisfied or unsatisfied.
- **Gender**: Gender of the customer.
- **Age**: Age of the customer.

- **Class**: Factor variable indicating the type of class the customer was traveling in. (Eco,Eco Plus,Business)
- **Flight.distance**: Distance of the flight (in kms)
- **Seat.comfort**: Satisfaction level of seat comfort
- **Departure.Arrival.time.convenient**: Satisfaction level of departure/arrival time convenience.
- **Food.and.drink**: Satisfaction level of food and drink.
- **Gate.location**: Satisfaction level of gate location.
- **Inflight.wifi.service**: Satisfaction level of the inflight WiFi service.
- **Inflight.entertainment**: Satisfaction level of inflight entertainment.
- **Online.support**: Satisfaction level of online support.
- **Ease.of.Online.booking**: Satisfaction level of online booking.
- **On.board.service**: Satisfaction level of on-board service.
- **Leg.room.service**: Satisfaction level of leg room service.
- **Baggage.handling**: Satisfaction level of baggage handling.
- **Checkin.service**: Satisfaction level of the checkin service.
- **Cleanliness**: Satisfaction level of cleanliness.
- **Online.boarding**: Satisfaction level of online boarding.
- **Departure.Delay.in.Minutes**: Departure delay in minutes.
- **Arrival.Delay.in.Minutes**: Arrival delay in minutes.

Features corresponding to satisfaction level are in range (0-5).

```
# Check for NA's in the entire dataset
null= any(is.na(data)); null
```

```
## [1] FALSE
```

There are no missing values so we can proceed with our analysis.

```
#Let us plot the correlation

# Load necessary libraries
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.0.5
```

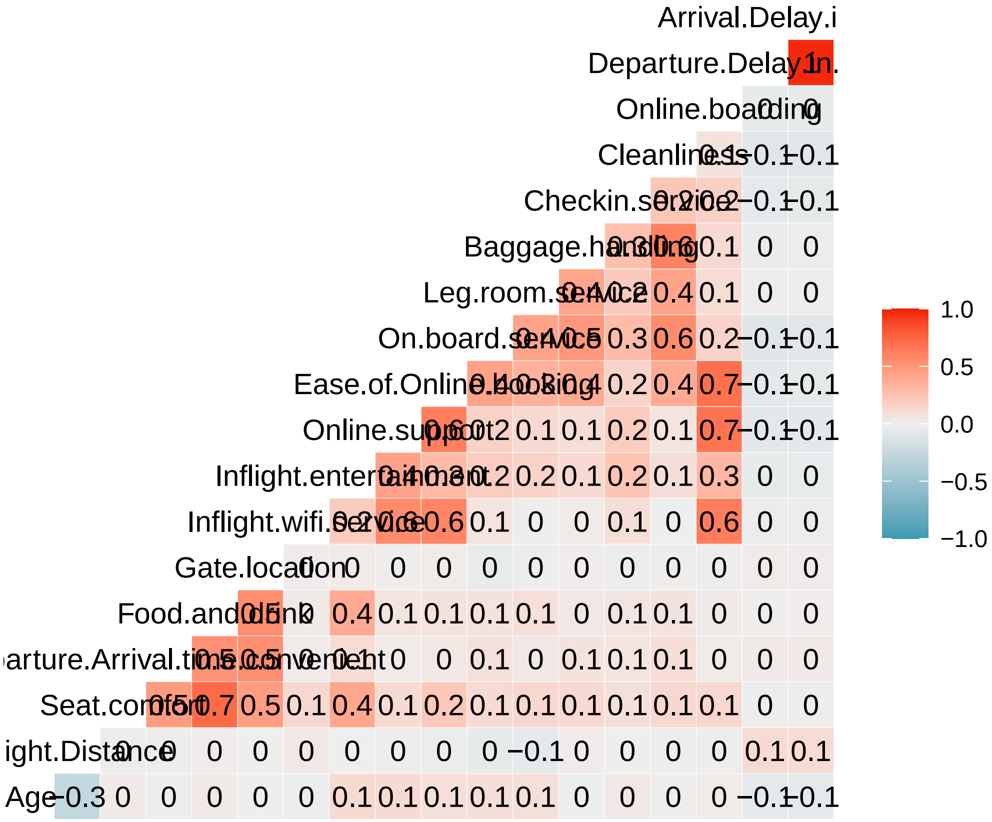
```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(ggplot2)
ggcorr(data, label = T)
```

```
## Warning in ggcorr(data, label = T): data in column(s) 'satisfaction', 'Gender',
## 'Class' are not numeric and were ignored
```

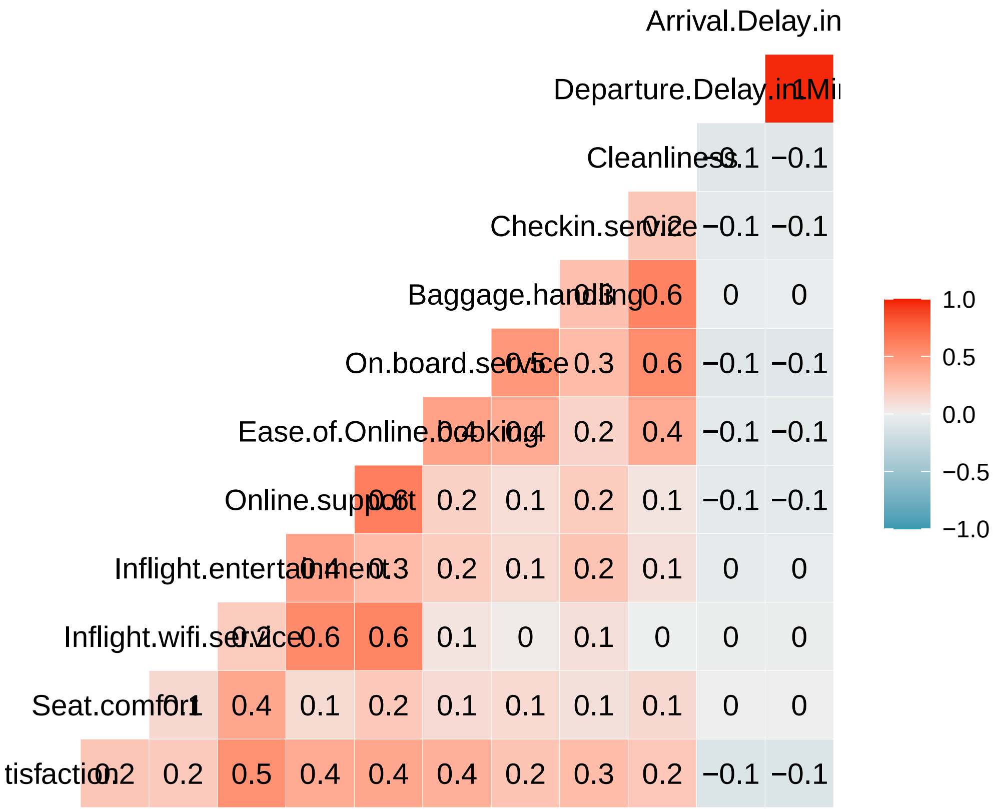


As we can see there are several variables that may not add value when predicting the binary variable satisfaction, so we will remove them.

```
# Remove variables that do not add much value
data <- subset(data, select = -c(Age, Gender, Food.and.drink, Leg.room.service, Gate.location,Class, Fl


# Encode 'satisfaction' column as 0 and 1
data$satisfaction <- ifelse(data$satisfaction == "satisfied", 1, 0)
```

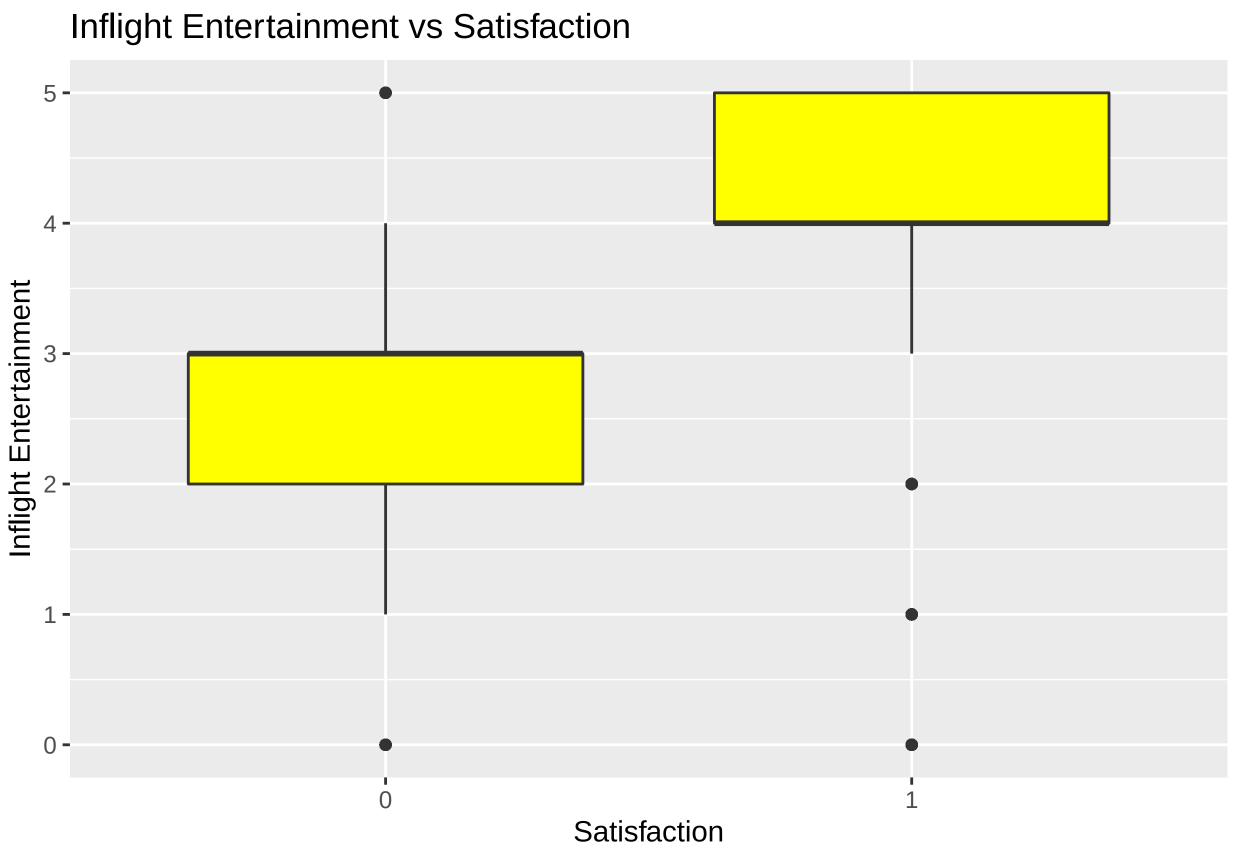
This is the new correlation with the remaining variables we will use in our task.

```
# Load necessary libraries
library(GGally)
library(ggplot2)
ggcorr(data, label = T)
```



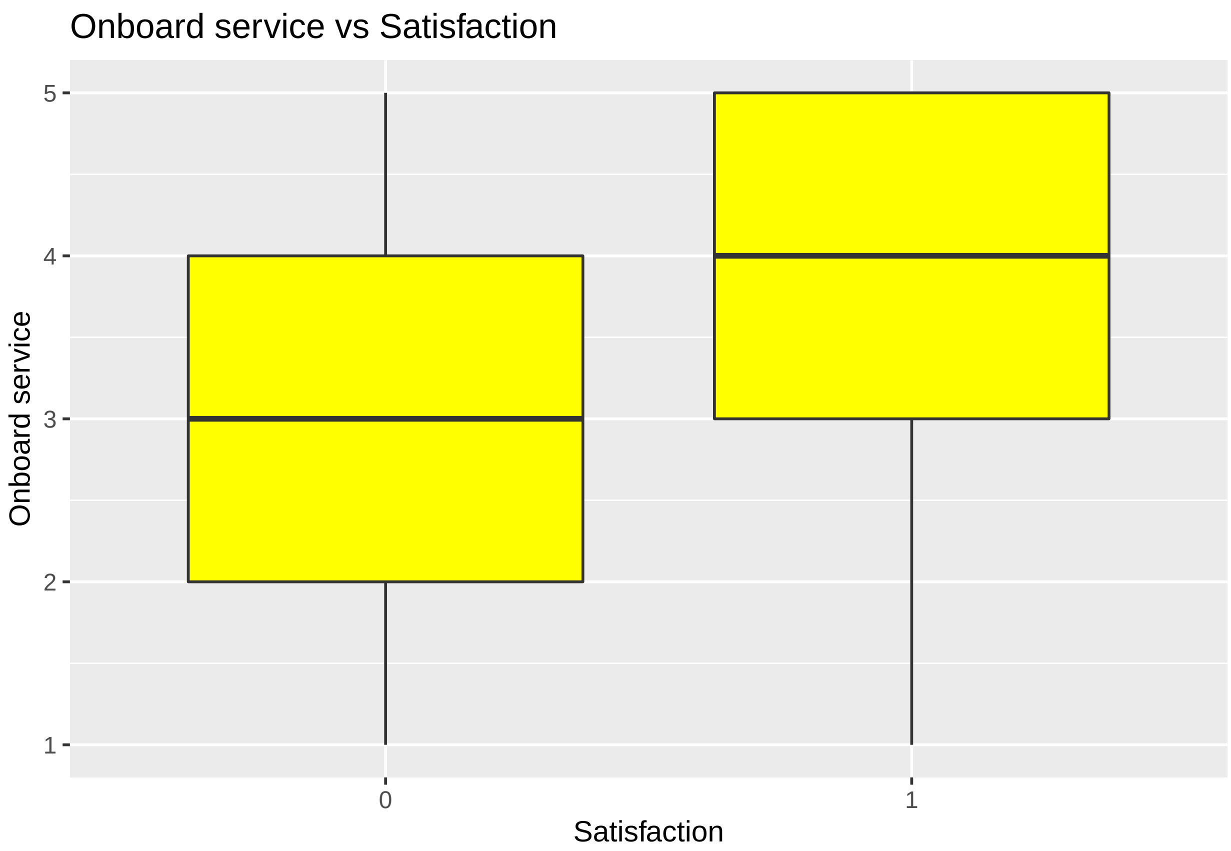
Now, let us visualize some boxplots to check how the different predictors behave with the satisfaction class. For instance, satisfied customers tend to give a higher score for the entertainment provided during the flight.

```
# Boxplot for Inflight.entertainment vs Satisfaction
ggplot(data, aes(x = factor(satisfaction), y = Inflight.entertainment)) +
  geom_boxplot(fill="yellow") +
  labs(title = "Inflight Entertainment vs Satisfaction", x = "Satisfaction", y = "Inflight Entertainmen
```



Regarding the onboard service, the median for satisfied customers is 4 while for unsatisfied ones, is 3.

```
# Boxplot for On.board.service vs Satisfaction
ggplot(data, aes(x = factor(satisfaction), y = On.board.service)) +
  geom_boxplot(fill="yellow") +
  labs(title = "Onboard service vs Satisfaction", x = "Satisfaction", y = "Onboard service")
```



CLASSICAL LOGISTIC REGRESSOR

First of all, we will try a classical logistic regression which shows the relationship between some trip status variables and the dependent one, satisfaction.

For that, we begin firstly with the complete model to see which are the variables that are significant when predicting the satisfaction.

```
classic <- glm (satisfaction ~ .,data = data, family = binomial)

summary(classic)
```

```
##
## Call:
## glm(formula = satisfaction ~ ., family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7281  -0.7295   0.3173   0.6271   3.2558
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -6.402865   0.354879 -18.042  < 2e-16 ***
## Seat.comfort                  0.009668   0.049908   0.194   0.8464
## Inflight.wifi.service      -0.073970   0.061749  -1.198   0.2309
## Inflight.entertainment      0.844564   0.057717  14.633  < 2e-16 ***
## Online.support              0.165342   0.064553   2.561   0.0104 *
## Ease.of.Online.booking      0.407317   0.075245   5.413 6.19e-08 ***
## On.board.service            0.367125   0.058751   6.249 4.14e-10 ***
## Baggage.handling            0.056835   0.065530   0.867   0.3858
## Checkin.service             0.247932   0.049501   5.009 5.48e-07 ***
## Cleanliness                -0.044159   0.069477  -0.636   0.5250
## Departure.Delay.in.Minutes -0.004323   0.005881  -0.735   0.4623
## Arrival.Delay.in.Minutes   -0.002035   0.005747  -0.354   0.7232
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2747.4  on 1999  degrees of freedom
## Residual deviance: 1807.5  on 1988  degrees of freedom
## AIC: 1831.5
##
## Number of Fisher Scoring iterations: 5
```

As we can see from the summary, the variables that have lowest p-values are those that add info to the model:

- **Inflight.entertainment**
- **Ease.of.Online.booking**
- **On.board.service**
- **Checkin.service**

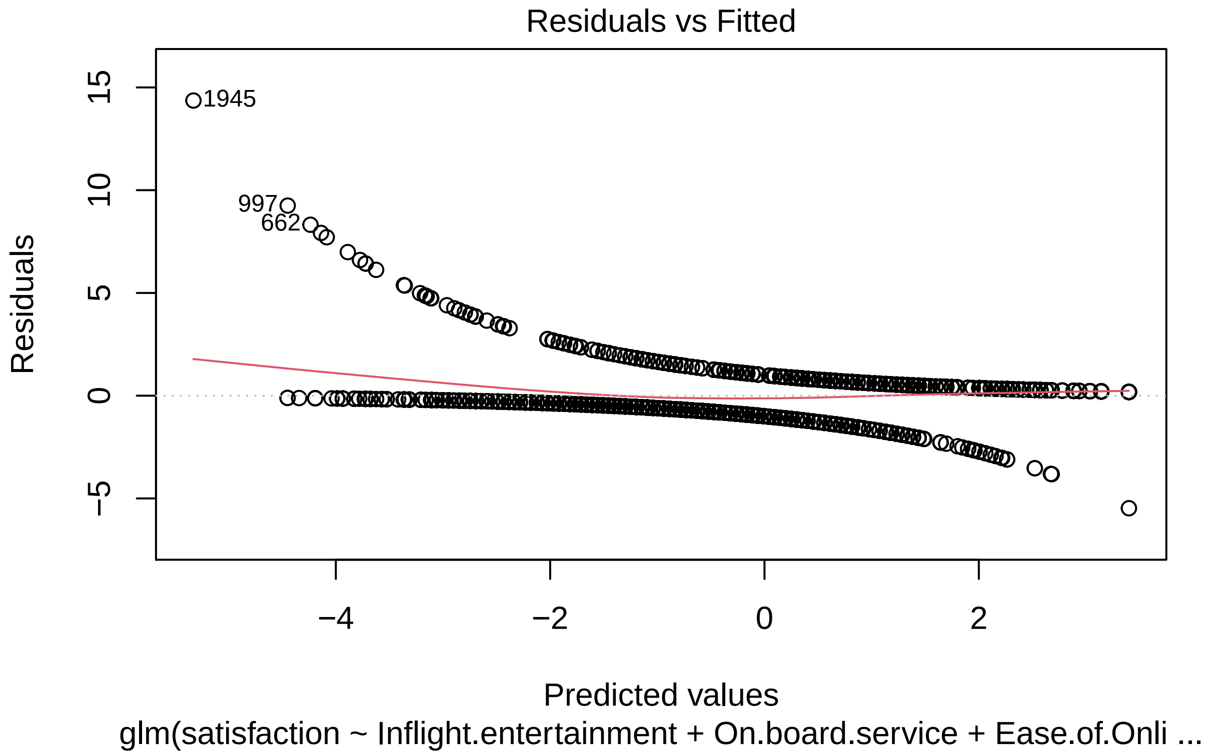
Therefore, these are the variables that we will use as predictors in our logistic regression case study.

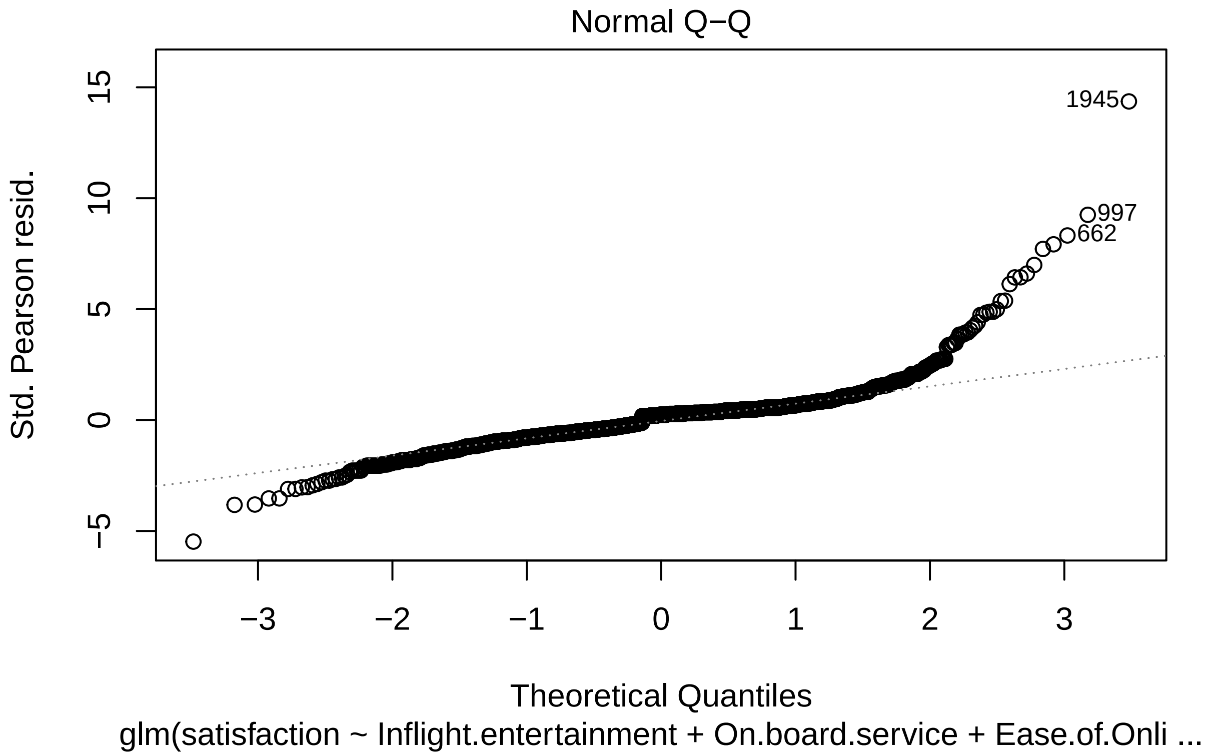
```r
# Fit logistic regression model
classic <- glm(satisfaction ~ Inflight.entertainment + On.board.service +  Ease.of.Online.booking
               + Checkin.service, data = data, family = binomial)

# Summary of the model
summary(classic)
```

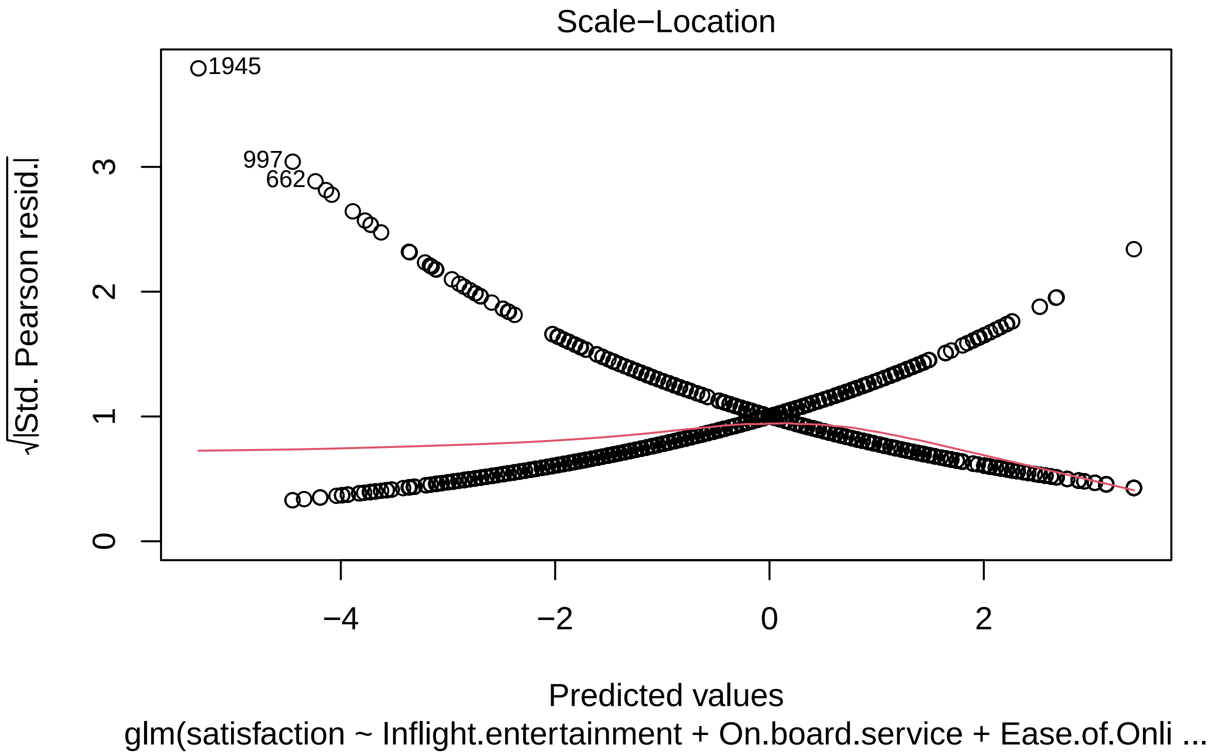
```
##
## Call:
## glm(formula = satisfaction ~ Inflight.entertainment + On.board.service +
##     Ease.of.Online.booking + Checkin.service, family = binomial,
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6205  -0.7457   0.3297   0.6528   3.2662
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -6.41345    0.30214 -21.226  < 2e-16 ***
## Inflight.entertainment  0.87843    0.05085  17.276  < 2e-16 ***
## On.board.service        0.36405    0.05025   7.245 4.33e-13 ***
## Ease.of.Online.booking  0.46237    0.04977   9.290  < 2e-16 ***
## Checkin.service         0.25797    0.04800   5.374 7.69e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2747.4  on 1999  degrees of freedom
## Residual deviance: 1832.9  on 1995  degrees of freedom
## AIC: 1842.9
##
## Number of Fisher Scoring iterations: 5
```

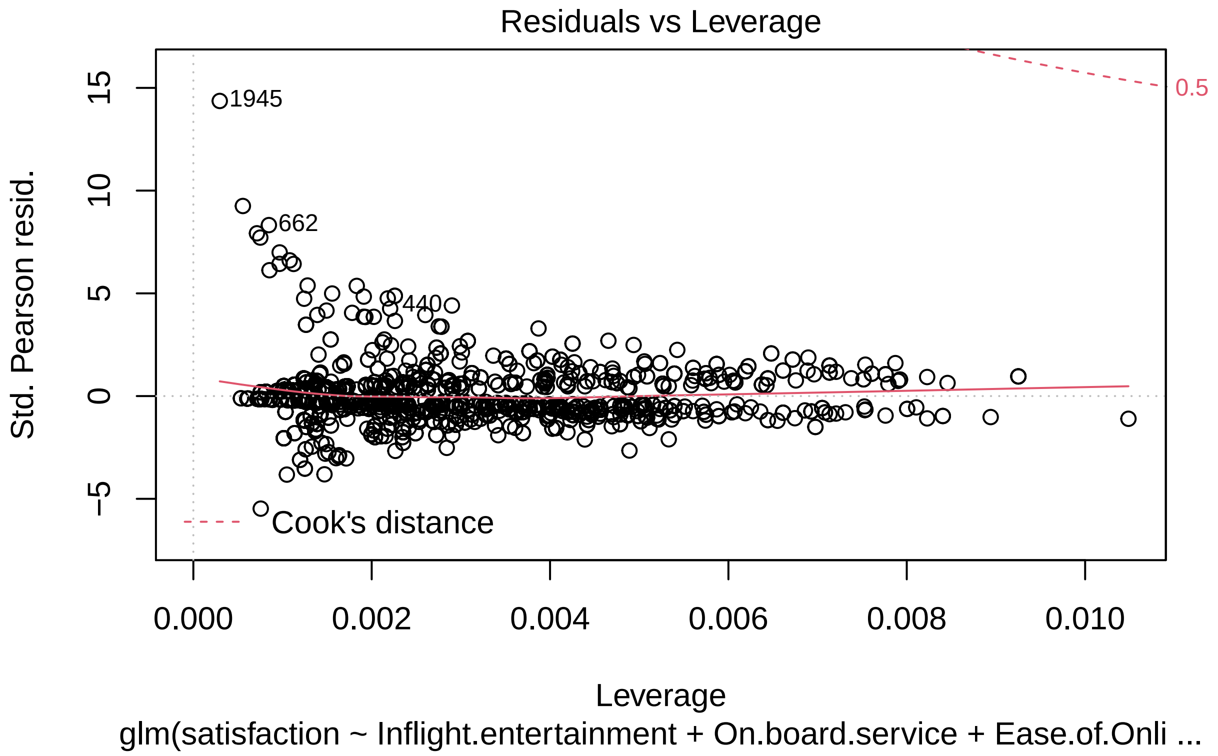
From the summary, Beta1 is 0.87 which menas that for every unit the Inflight.entertainment increases, the estimated probability of being satisfied increases 0.87. Also, Beta3 is 0.46237 and that means that when the Ease of online booking increases 1 unit, it makes probability of being satisfied increase by 0.46237.

```r
plot(classic)
```




Normal Q−Q
Std. Pearson resid.
15
10
5
0
−5
1945
997
662
−3
−2
−1
0
1
2
3
Theoretical Quantiles
glm(satisfaction ~ Inflight.entertainment + On.board.service + Ease.of.Onli ...


Scale−Location
1945
997
662
√|Std. Pearson resid.|
0
1
2
3
−4
−2
0
2
Predicted values
glm(satisfaction ~ Inflight.entertainment + On.board.service + Ease.of.Onli ...



Some remarks:

- In the Residuals vs Fitted plot, as the residuals are centered around the horizontal dashed line at 0, it indicates that the model does not systematically overestimate or underestimate the response.
- The normal Q-Q plot helps to assess the normality of residuals, however as we are in a logistic regression problem we should focus more on the overall fit and other diagnostics like residuals vs fitted plots.

Let us now introduce some new data to the model and predict the probability of being satisfied given the following scores:

```r
# New data for prediction
new_data <- data.frame(
  Inflight.entertainment = 5,
  On.board.service = 4,
  Ease.of.Online.booking = 5,
  Checkin.service = 3
)

# Predict using the model
predicted_values <- predict(classic, newdata = new_data, type = "response")

# Output the predicted values
print(predicted_values)
```

```
##         1
## 0.9255775
```

A probability of 0.9255774 of being satisfied has been obtained. Below, we will test also this new data with bayesian approaches such as MCMC and R2OPENBUGS.

BAYESIAN APPROACH

MONTE CARLO MARKOV CHAIN

Now, let us proceed but instead using a bayesian approach and the MCMC (Monte Carlo Markov Chain) to sample from the posteriors as we do not have a way to compute directly the posterior.

First we begin also with the full model to search for the important predictors.

```r
library(MCMCpack)
```

```
## Warning: package 'MCMCpack' was built under R version 4.0.5
```

```
## Loading required package: coda
```

```
## Loading required package: MASS
```

```
## ##
## ## Markov Chain Monte Carlo Package (MCMCpack)
```

```
## ## Copyright (C) 2003-2024 Andrew D. Martin, Kevin M. Quinn, and Jong Hee Park
```

```
## ##
## ## Support provided by the U.S. National Science Foundation
```

```
## ## (Grants SES-0350646 and SES-0350613)
## ##
```

```r
logit = MCMClogit(satisfaction ~ ., data = data, thin=10, burnin=1000, mcmc=21000)

summary(logit)
```

```
##
## Iterations = 1001:21991
## Thinning interval = 10
## Number of chains = 1
## Sample size per chain = 2100
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                                  Mean       SD  Naive SE Time-series SE
## (Intercept)                 -6.455596 0.376268 0.0082108      0.0199088
## Seat.comfort                 0.009002 0.050864 0.0011099      0.0024956
## Inflight.wifi.service       -0.075140 0.059555 0.0012996      0.0029505
## Inflight.entertainment       0.856372 0.057966 0.0012649      0.0028328
## Online.support               0.162570 0.064967 0.0014177      0.0034620
## Ease.of.Online.booking      0.411779 0.077069 0.0016818      0.0040361
## On.board.service            0.370942 0.059816 0.0013053      0.0031389
## Baggage.handling            0.055443 0.069938 0.0015262      0.0037839
## Checkin.service             0.249091 0.048673 0.0010621      0.0025293
## Cleanliness                -0.043639 0.071394 0.0015579      0.0037847
## Departure.Delay.in.Minutes -0.004090 0.005853 0.0001277      0.0002876
## Arrival.Delay.in.Minutes   -0.002281 0.005701 0.0001244      0.0002808
##
## 2. Quantiles for each variable:
##
##                                2.5%       25%       50%        75%     97.5%
## (Intercept)                -7.15864 -6.709179 -6.460007 -6.1972916 -5.686710
## Seat.comfort               -0.08635 -0.029819  0.008797  0.0453673  0.109170
## Inflight.wifi.service      -0.20854 -0.114324 -0.074672 -0.0339491  0.037455
## Inflight.entertainment      0.74372  0.815504  0.856495  0.8952359  0.965617
## Online.support              0.03481  0.118892  0.163165  0.2043972  0.292724
## Ease.of.Online.booking      0.25895  0.356606  0.410609  0.4656417  0.560862
## On.board.service            0.25384  0.329749  0.373114  0.4146641  0.481782
## Baggage.handling           -0.07650  0.005668  0.057177  0.1007723  0.195555
## Checkin.service             0.15822  0.215198  0.250204  0.2824293  0.342922
## Cleanliness                -0.17305 -0.091713 -0.046598  0.0046239  0.101704
## Departure.Delay.in.Minutes -0.01571 -0.008015 -0.004107 -0.0005644  0.008248
## Arrival.Delay.in.Minutes   -0.01390 -0.005856 -0.002098  0.0014895  0.008441
```

From this sumamry we can verify that the variables used before in the classical model are the significant ones as they do not contain 0 in their quantiles. The rest of variables do contain 0 in ther ci's and therefore they are not statistically significant when predicting the satisfaction variable.

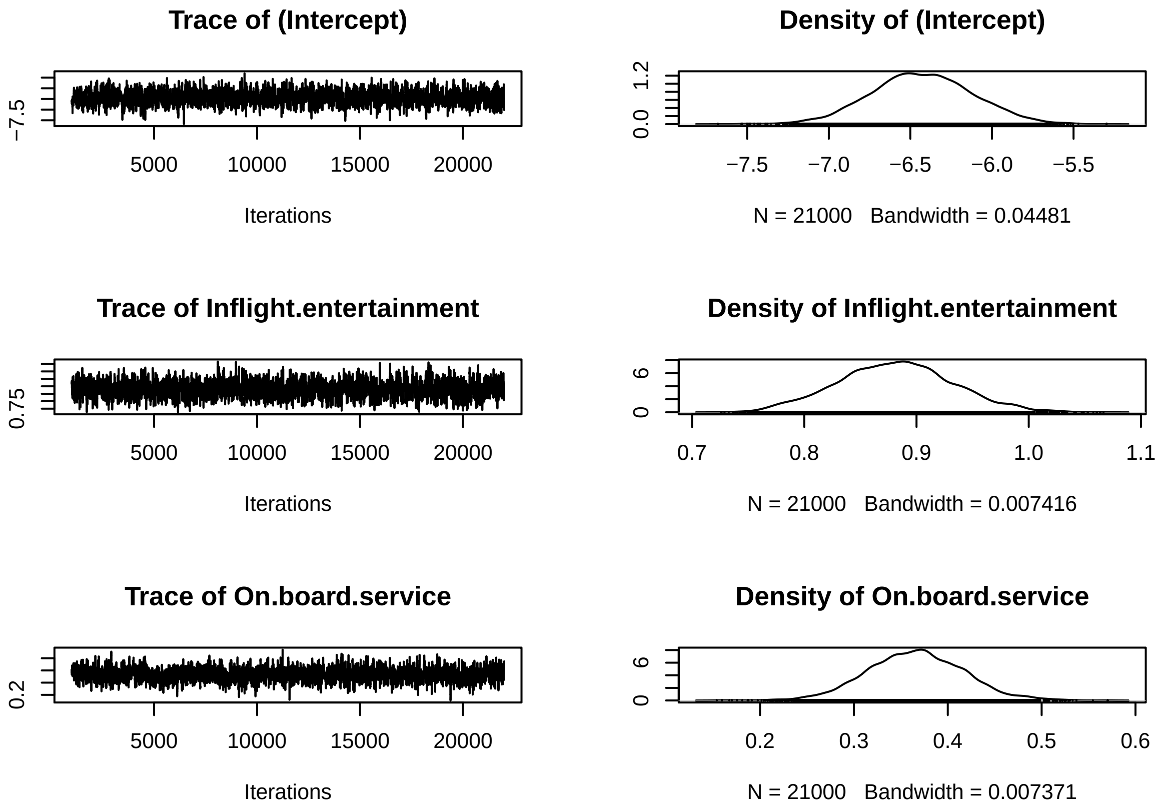
Let us implement now the model for the chosen predictors and see its convergence:

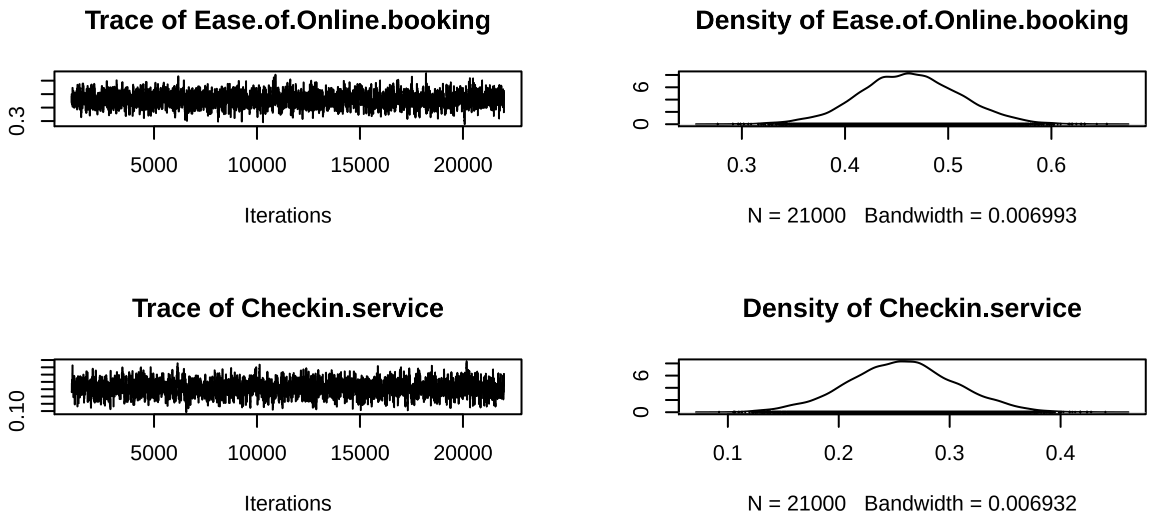
```r
logit = MCMClogit(satisfaction ~ Inflight.entertainment + On.board.service +  Ease.of.Online.booking
               + Checkin.service, burnin=1000, mcmc=21000, data = data)

summary(logit)
```

```
##
## Iterations = 1001:22000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 21000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                           Mean      SD  Naive SE Time-series SE
## (Intercept)            -6.4277 0.30939 0.0021350       0.008961
## Inflight.entertainment  0.8814 0.05121 0.0003534       0.001455
## On.board.service        0.3657 0.05132 0.0003542       0.001438
## Ease.of.Online.booking  0.4627 0.04861 0.0003354       0.001337
## Checkin.service         0.2569 0.04846 0.0003344       0.001344
##
## 2. Quantiles for each variable:
##
##                           2.5%     25%     50%     75%   97.5%
## (Intercept)            -7.0315 -6.6357 -6.4310 -6.2198 -5.8136
## Inflight.entertainment  0.7824  0.8458  0.8814  0.9150  0.9842
## On.board.service        0.2664  0.3315  0.3654  0.3997  0.4693
## Ease.of.Online.booking  0.3669  0.4305  0.4619  0.4952  0.5599
## Checkin.service         0.1603  0.2245  0.2569  0.2886  0.3524
```

Let us check model convergence and see the trace plots and posterior densities for the intercept and betas:

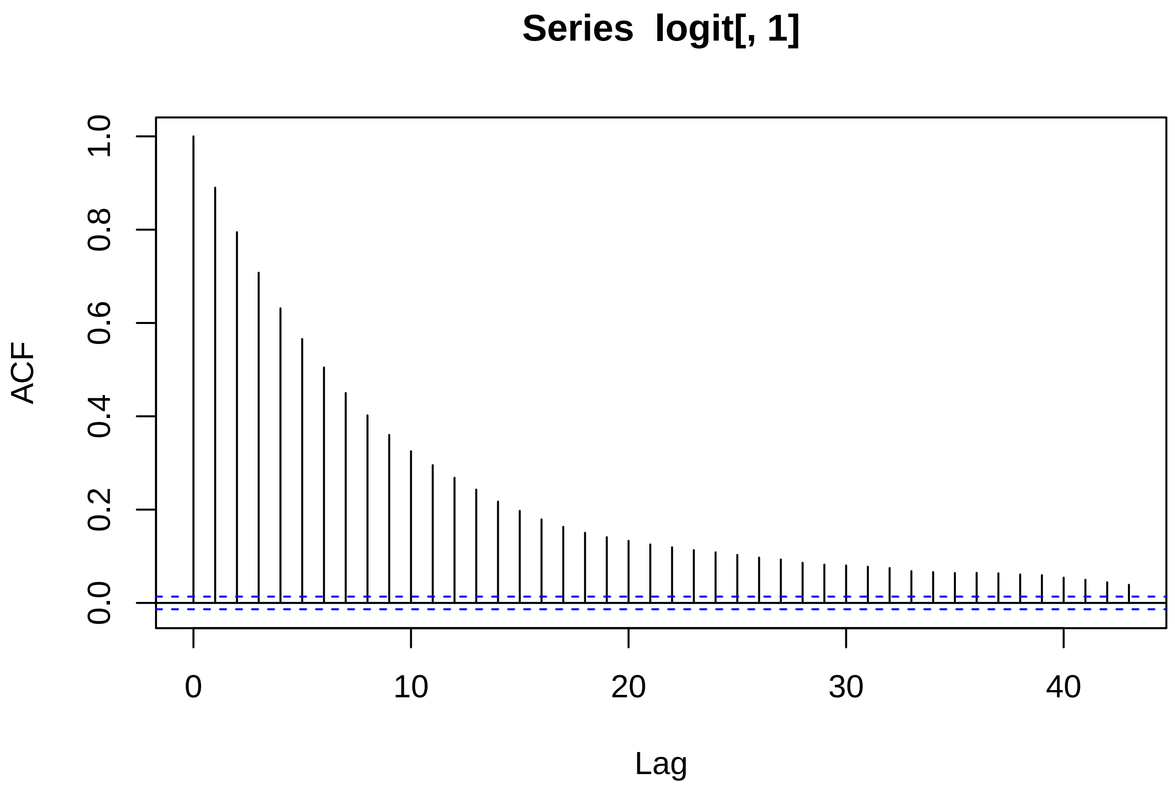
```
plot(logit)
```



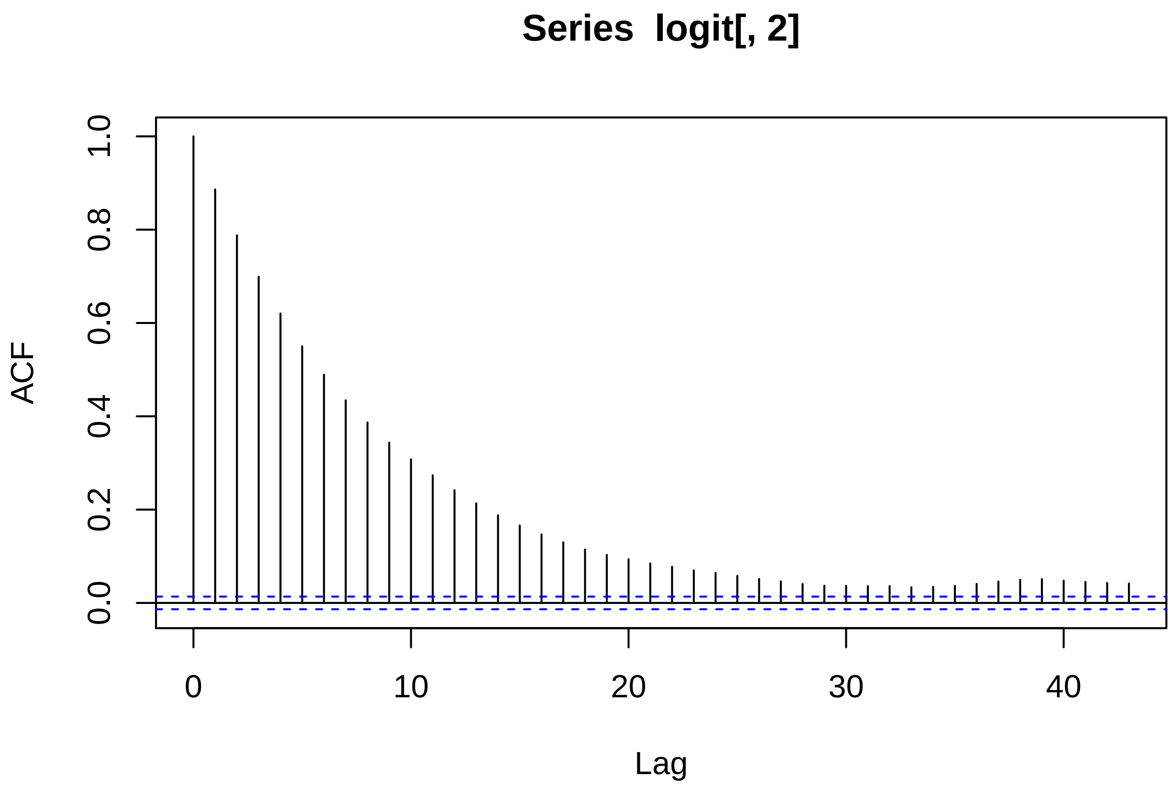


Let us also look for the autocorrelation:

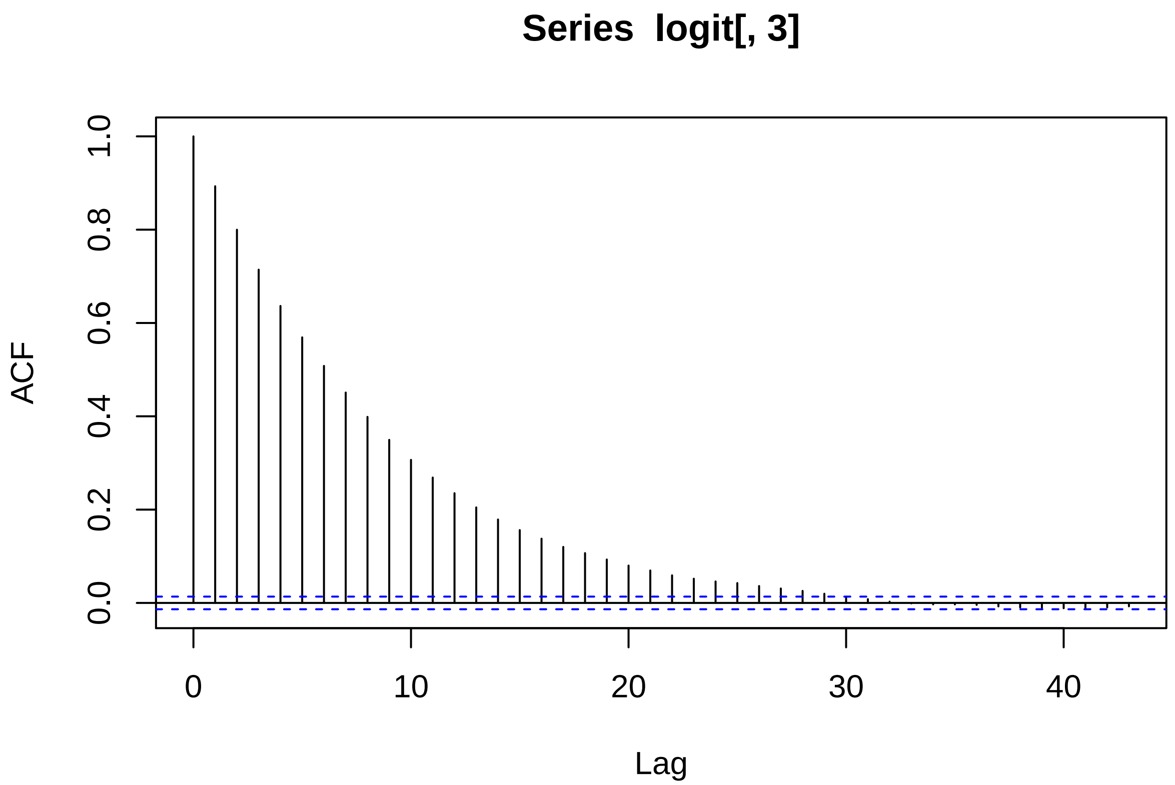
```r
acf(logit[,1])
```



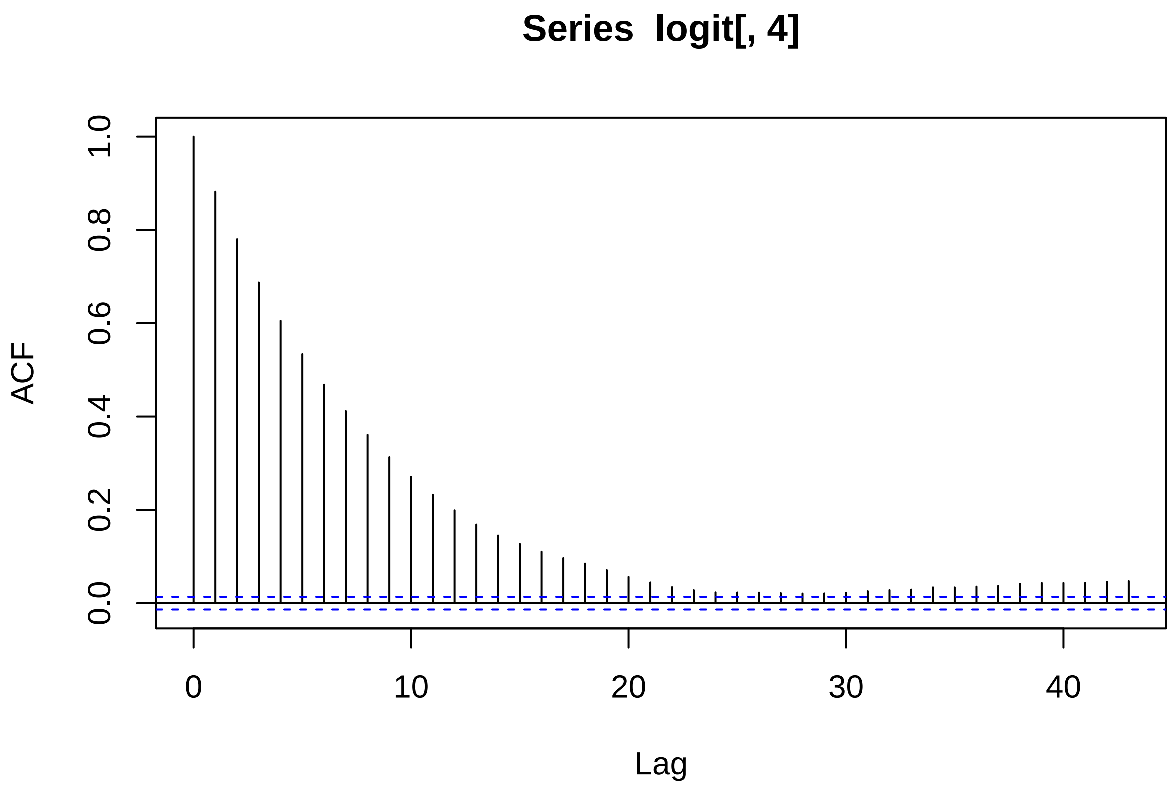
```r
acf(logit[,2])
```



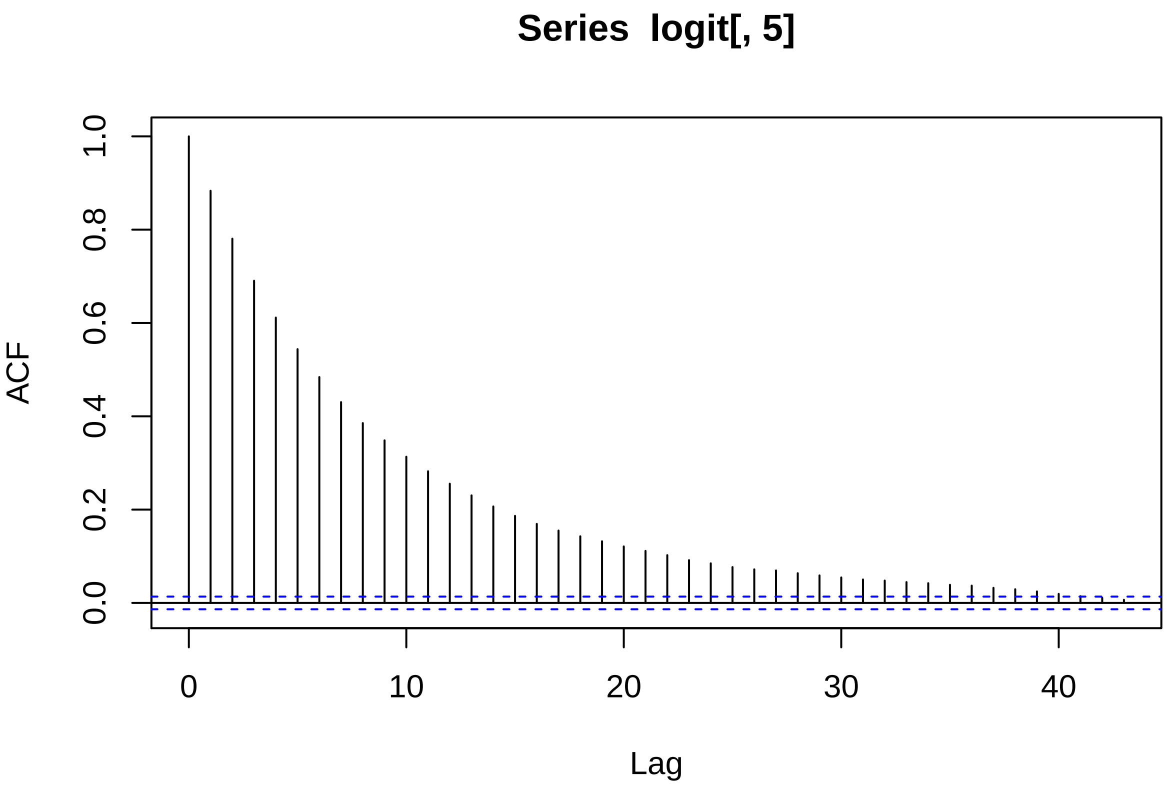
```
acf(logit[,3])
```



```r
acf(logit[,4])
```

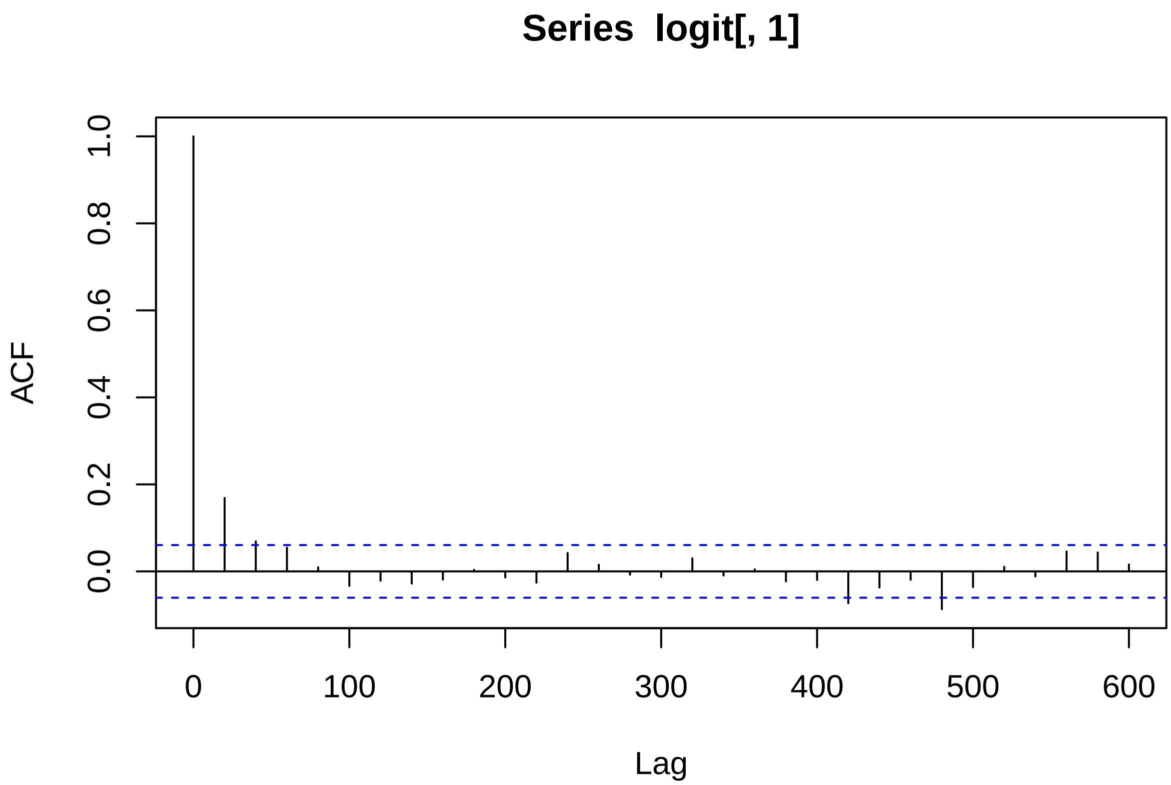


```
acf(logit[,5])
```

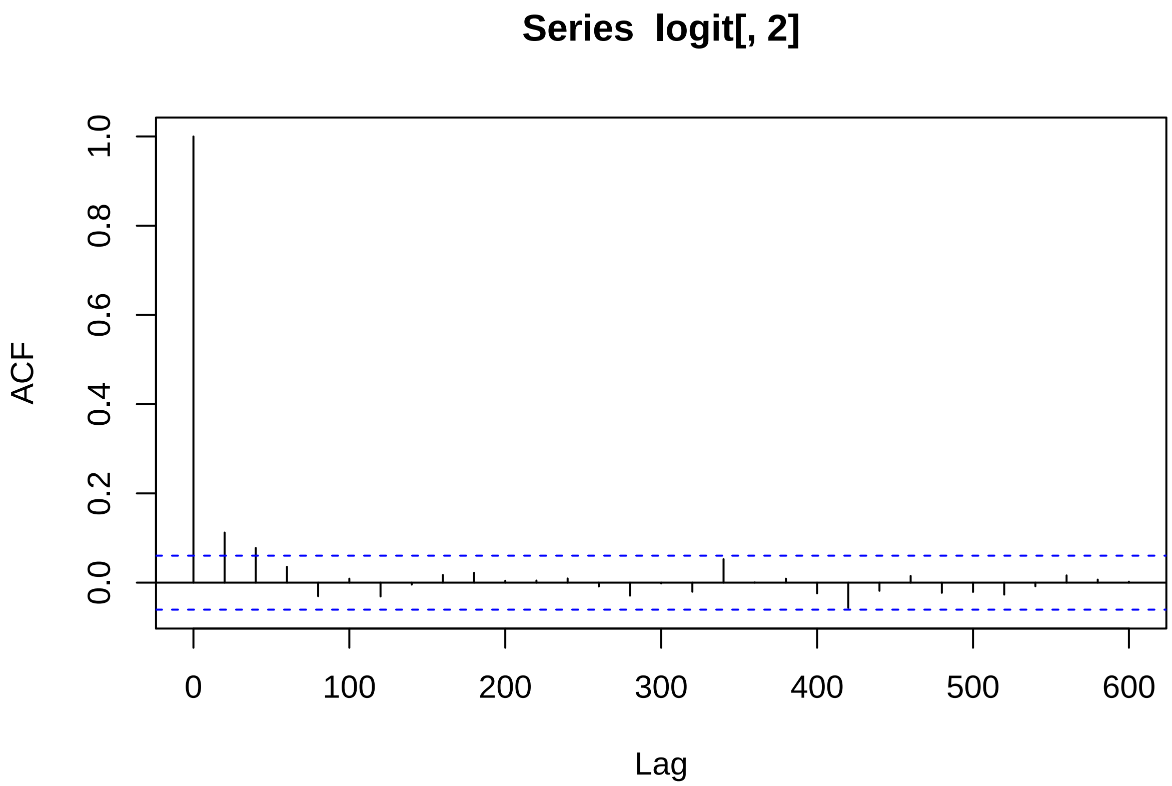


As it does not converge perfectly, we use thining of 20 iterations to achieve a better convergence and remove the autocorrelation

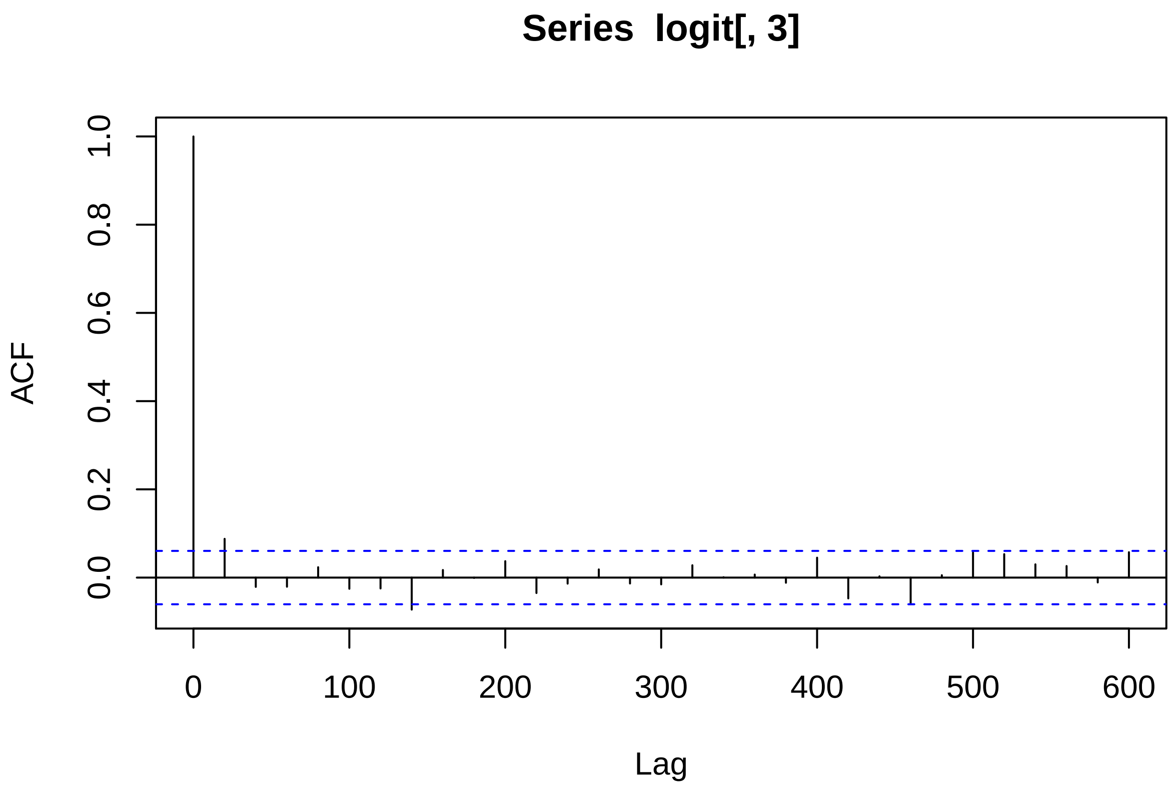
```
logit = MCMClogit(satisfaction ~ Inflight.entertainment + On.board.service +  Ease.of.Online.booking
                  + Checkin.service, burnin=1000, mcmc=21000, data = data, thin=20)
acf(logit[,1])
```



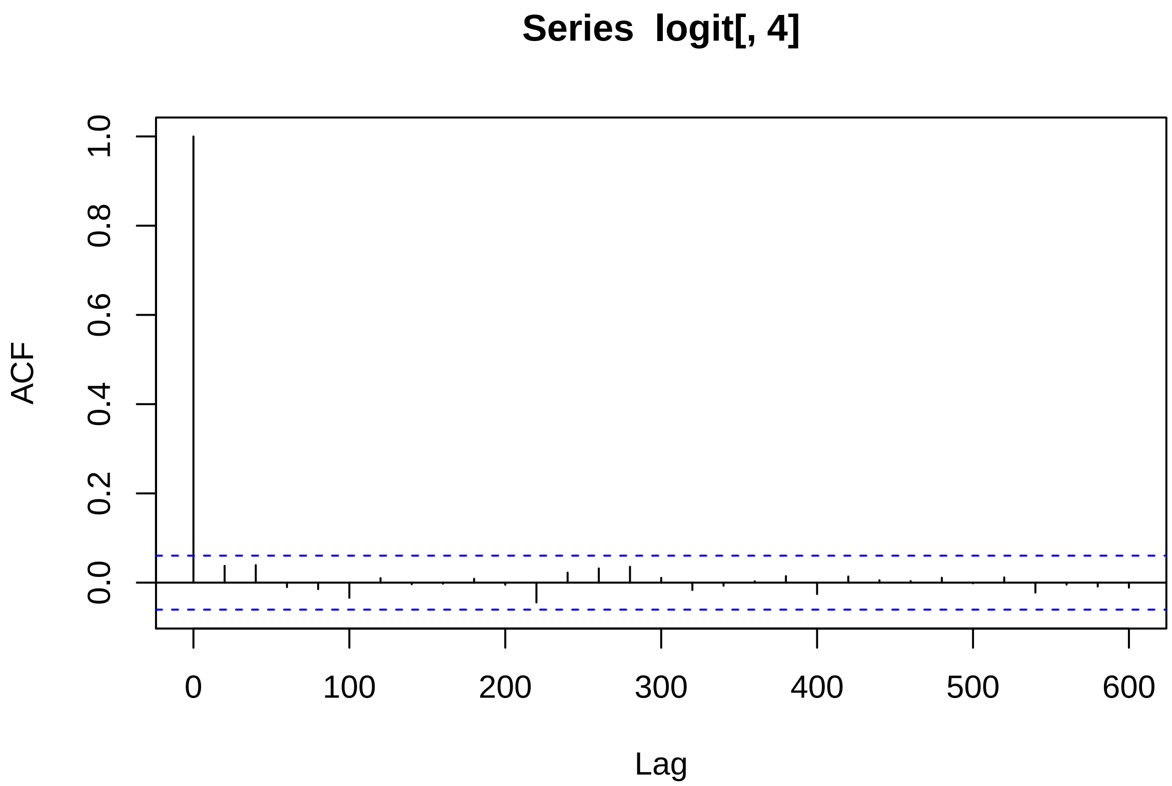
```
acf(logit[,2])
```



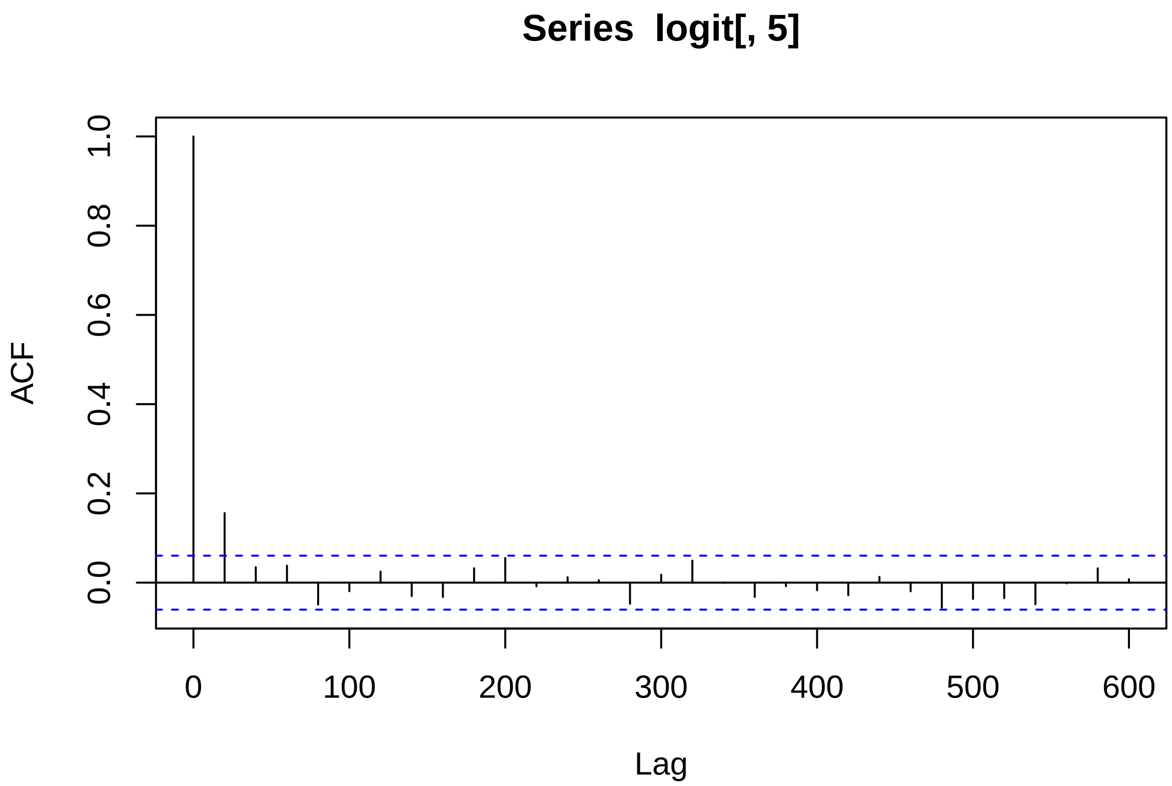
```r
acf(logit[,3])
```



```
acf(logit[,4])
```

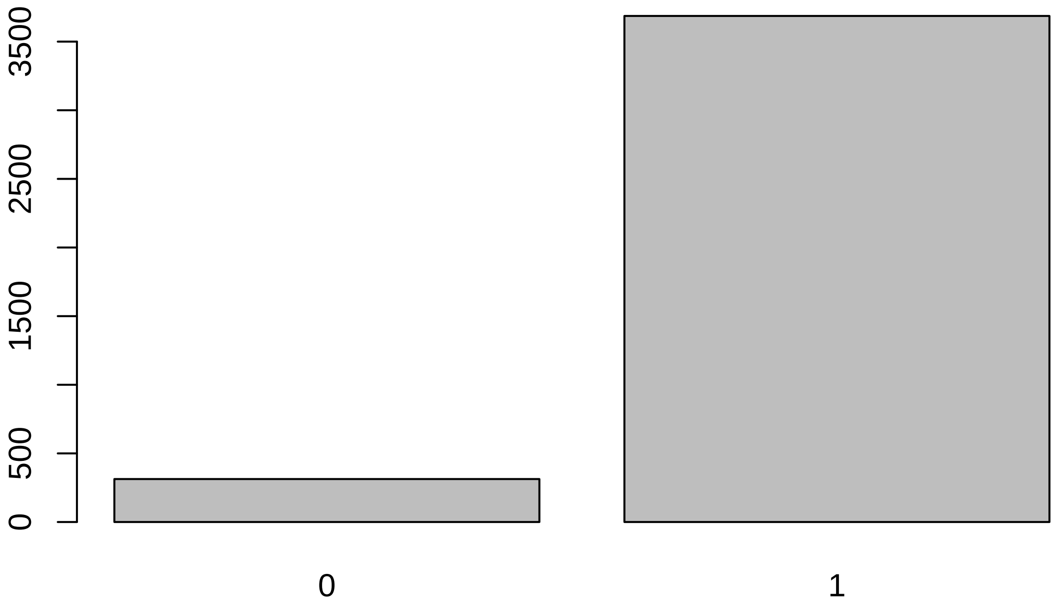


```
acf(logit[,5])
```



Now, we can play with the posterior distribution for some unseen data and check how the predictive distribution would be:

```
library(boot)

set.seed(1)

Inflight.entertainment = 5
On.board.service=4
Ease.of.Online.booking=5
Checkin.service=3

satisfaction.prob=inv.logit(logit[,1]+Inflight.entertainment*logit[,2] + On.board.service*logit[,3] + Ea
pred=rbinom(4000,size=1,prob=satisfaction.prob)
barplot(table(pred))
```



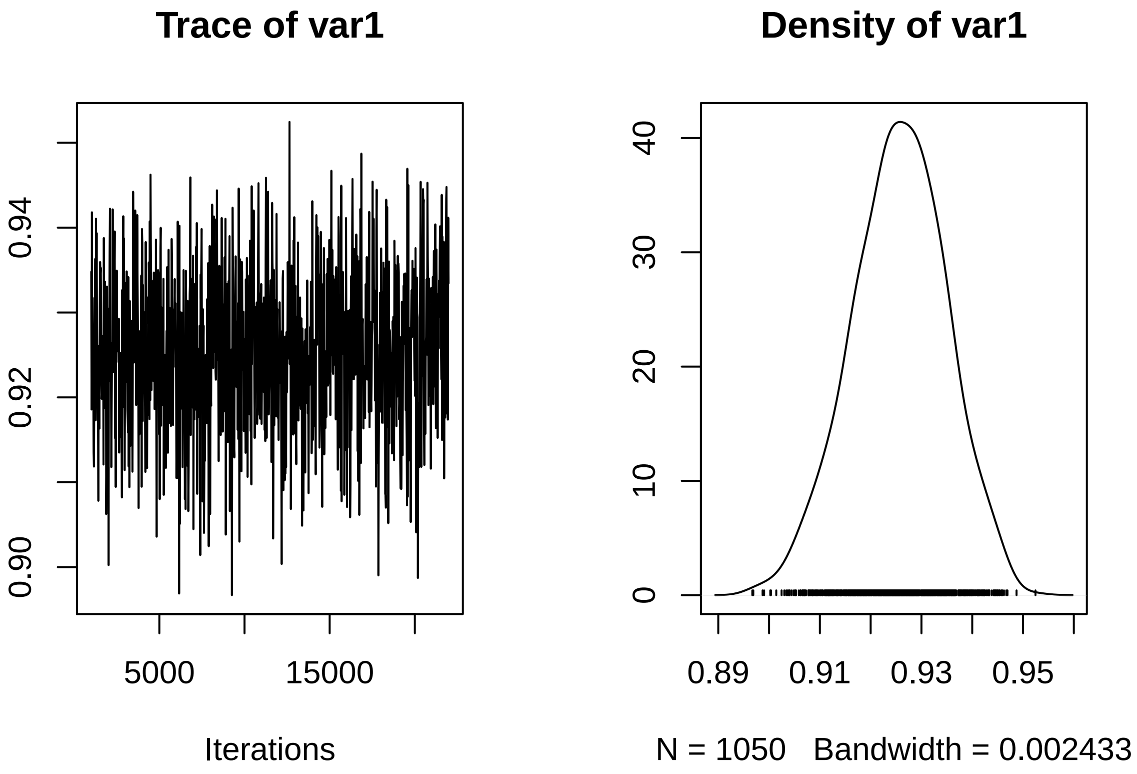
```
table(pred)
```

```
## pred
##    0    1
##  313 3687
```

We can see that for the new unseen data we gave to the model, the predictive sample composed by 4000 passengers states that 3687 passengers will be satisfied and 313 ones, disatissfied.

Trace plots and posterior density can be obtained with:

```
plot(satisfaction.prob)
```



From this plotted posterior we can see that the probability of being satisfied is high, as it is centered around 0.93, which makes sense when cheking the data sequence we inputed to the model (5,4,5,3) which has generally high rates for these trip status variables.

R2OPENBUGS

In this part of the case study, we will implement the bayesian approach but with the OPENBUGS package:

```
library(R2OpenBUGS)
```

```
## Warning: package 'R2OpenBUGS' was built under R version 4.0.5
```

```
# Define function
logit.bayes <- function() {
  for (i in 1:N) {
    y[i] ~ dbern(p[i])
    logit(p[i]) <- beta0 + beta1 * Inflight.entertainment[i] + beta2 * On.board.service[i] + beta3 * Ea
  }
  beta0 ~ dnorm(0.0, 1.0E-6)
  beta1 ~ dnorm(0.0, 1.0E-6)
  beta2 ~ dnorm(0.0, 1.0E-6)
  beta3 ~ dnorm(0.0, 1.0E-6)
  beta4 ~ dnorm(0.0, 1.0E-6)
}
```

```
# Prepare data for observed data and prior parameters
data_list <- list(
    N = nrow(data),
    y = data$satisfaction, #variable to predict
    Inflight.entertainment = data$Inflight.entertainment,
    On.board.service = data$On.board.service,
    Ease.of.Online.booking = data$Ease.of.Online.booking,
    Checkin.service = data$Checkin.service
)

# Specify initial values for parameters
inits <- function() {
    list(beta0 = 1, beta1 = 0, beta2 = 0, beta3 = 0, beta4 = 0)
}

# Run OpenBUGS
output <- bugs(data = data_list,
                   inits = inits,
                   parameters.to.save = c("beta0", "beta1", "beta2", "beta3", "beta4"),
                   model.file = logit.bayes,
                   n.chains = 1,
                   n.iter = 10000,
                   debug = F)
```

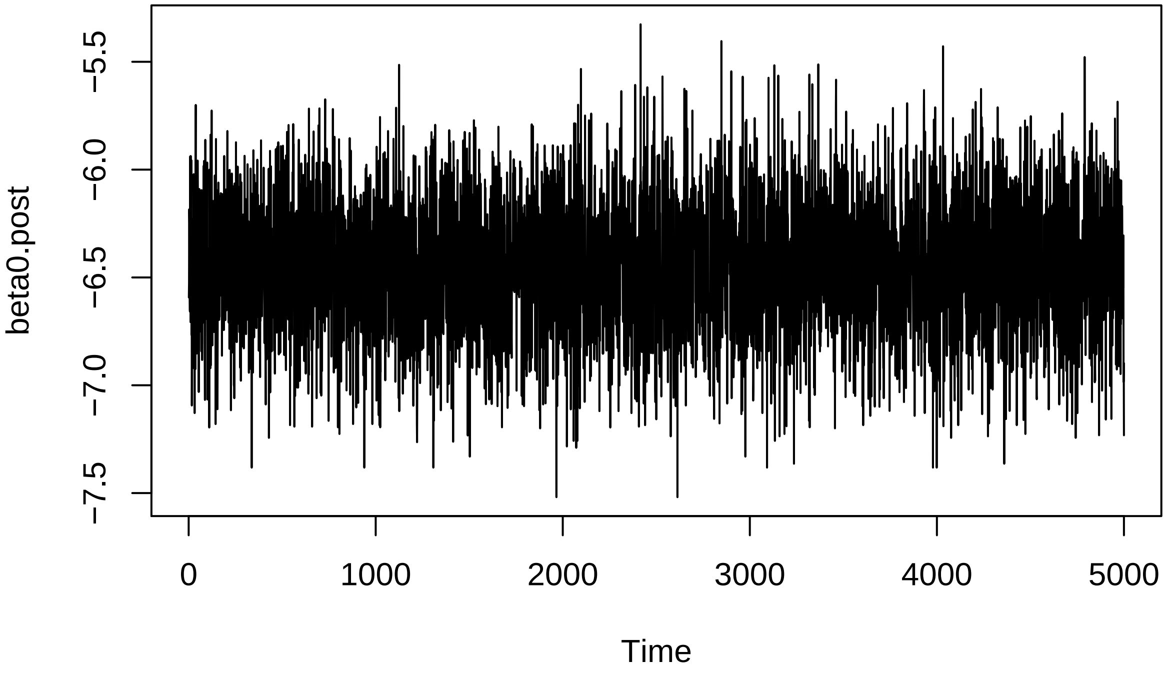
Let us inspect a summary of the output

```
summary(output)
```

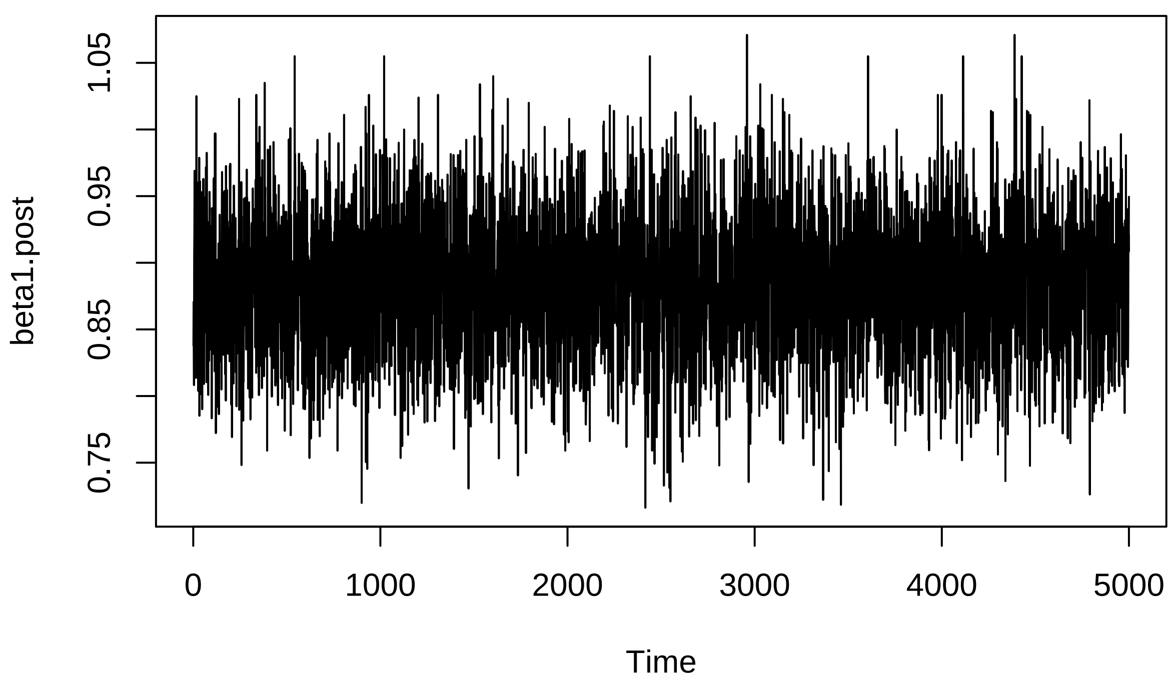
```
##                 Length Class  Mode
## n.chains            1  -none- numeric
## n.iter              1  -none- numeric
## n.burnin            1  -none- numeric
## n.thin              1  -none- numeric
## n.keep              1  -none- numeric
## n.sims              1  -none- numeric
## sims.array      30000  -none- numeric
## sims.list           6  -none- list
## sims.matrix     30000  -none- numeric
## summary            42  -none- numeric
## mean                6  -none- list
## sd                  6  -none- list
## median              6  -none- list
## root.short          6  -none- character
## long.short          6  -none- list
## dimension.short     6  -none- numeric
## indexes.short       6  -none- list
## last.values         1  -none- list
## isDIC               1  -none- logical
## DICbyR              1  -none- logical
## pD                  1  -none- numeric
## DIC                 1  -none- numeric
## model.file          1  -none- character
```

Let us now check the traces of the posterior betas:

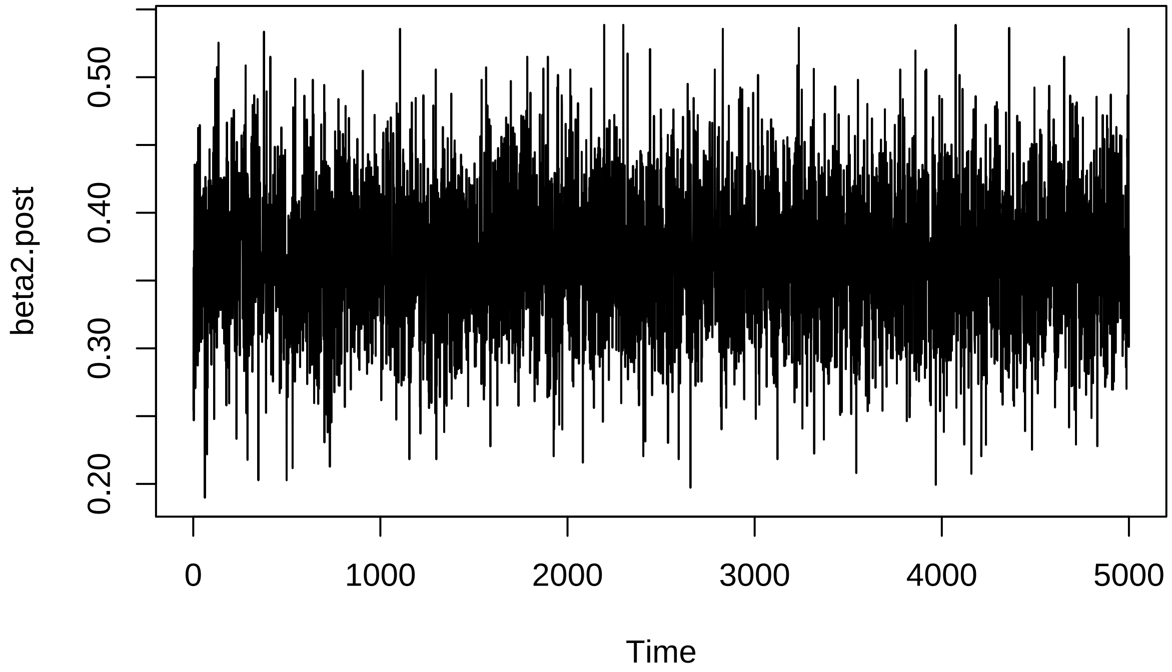
```
beta0.post <-output$sims.list$beta0
beta1.post <-output$sims.list$beta1
beta2.post <-output$sims.list$beta2
beta3.post <-output$sims.list$beta3
beta4.post <-output$sims.list$beta4

ts.plot(beta0.post)
```



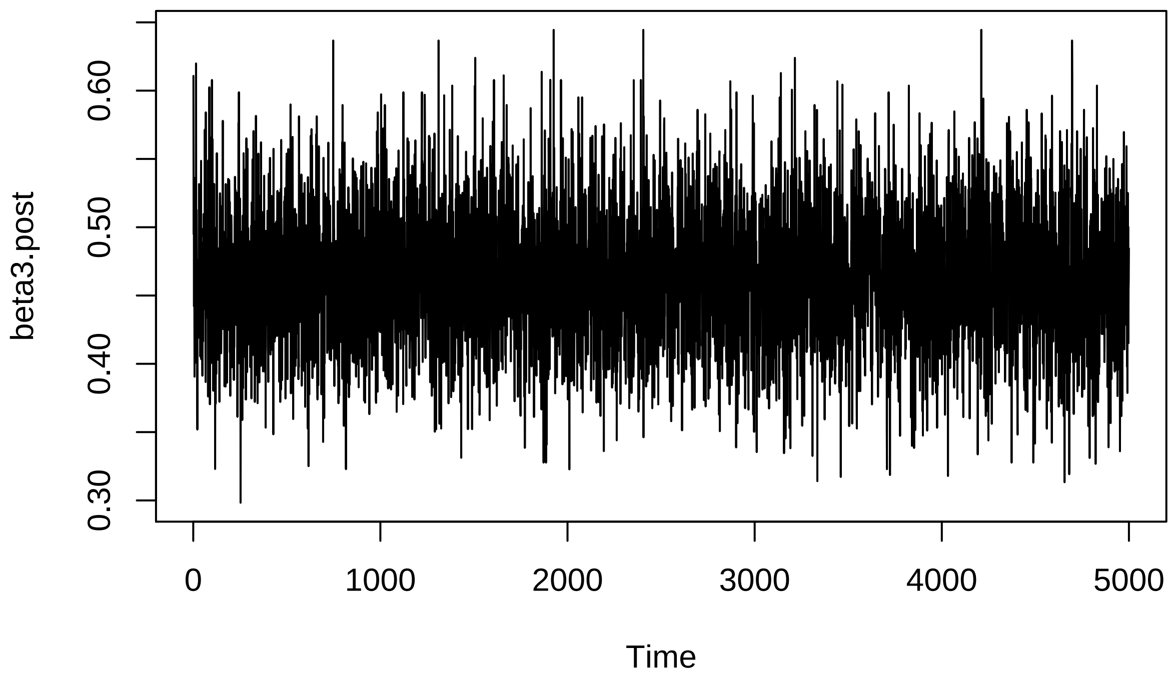
```
ts.plot(beta1.post)
```



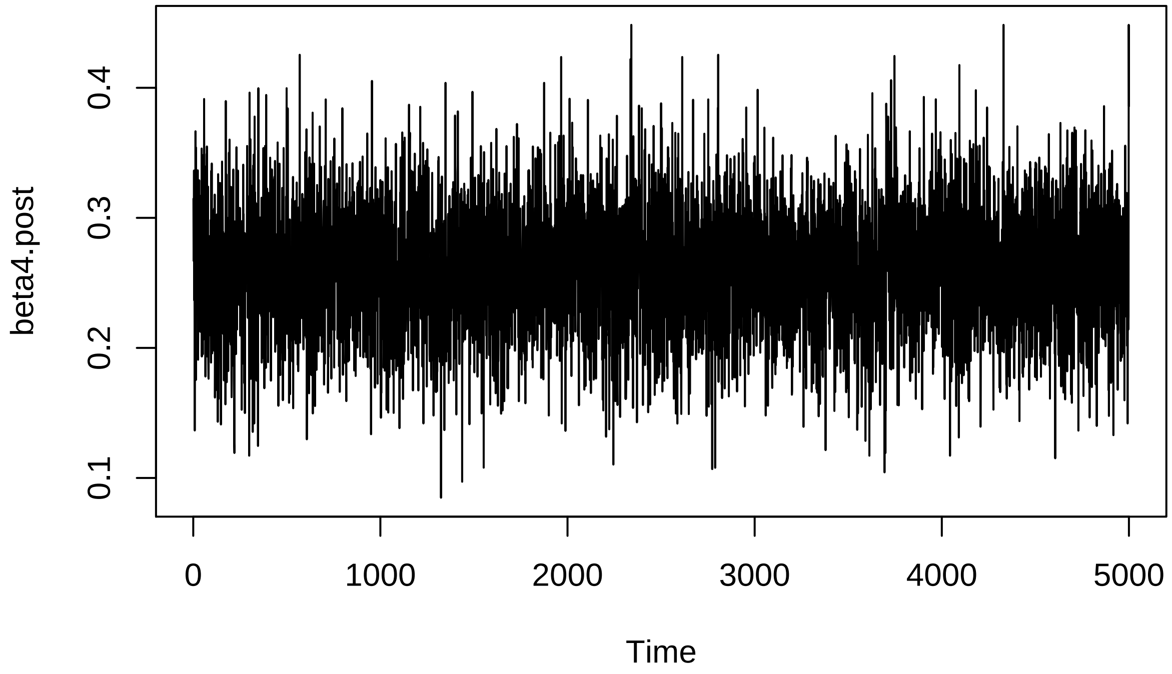
```
ts.plot(beta2.post)
```



```r
ts.plot(beta3.post)
```



```r
ts.plot(beta4.post)
```



We can see that the traces usually do converge

Finally, we can obtain the means and the 95% CI's of the betas by:

```
mean(beta0.post); quantile(beta0.post,c(0.025,0.975))
```

```
## [1] -6.449846
```

```
##      2.5%     97.5%
## -7.072075 -5.864000
```

```
mean(beta1.post); quantile(beta1.post,c(0.025,0.975))
```

```
## [1] 0.8837521
```

```
##      2.5%     97.5%
## 0.7864975 0.9824000
```

```
mean(beta2.post); quantile(beta2.post,c(0.025,0.975))
```

```
## [1] 0.3654114
```

```
##     2.5%    97.5%
## 0.270295 0.470300
```

```
mean(beta3.post); quantile(beta3.post,c(0.025,0.975))
```

```
## [1] 0.4643171
```

```
##   2.5%  97.5%
## 0.3704 0.5656
```

```
mean(beta4.post); quantile(beta4.post,c(0.025,0.975))
```

```
## [1] 0.2599757
```

```
##      2.5%     97.5%
## 0.1643975 0.3546075
```

Now, we look for the predictive distribution given the provided data using OpenBugs:

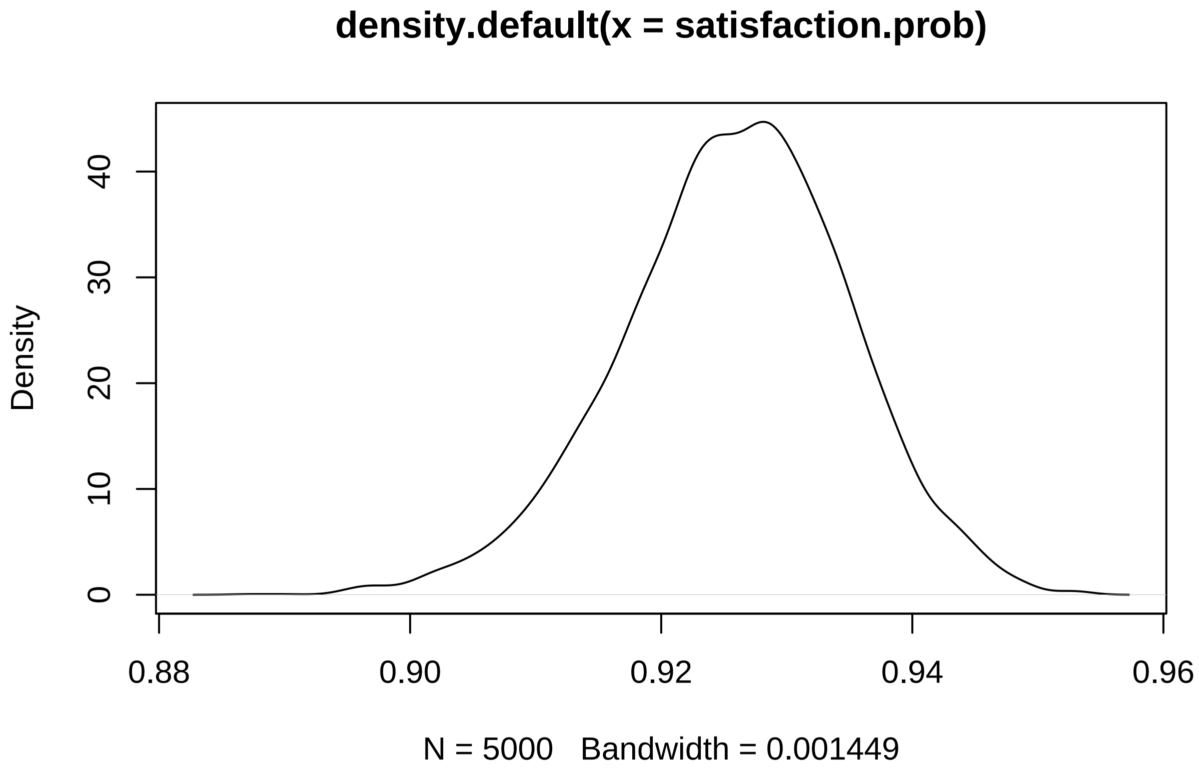
```
#New set of data
Inflight.entertainment = 5
On.board.service=4
Ease.of.Online.booking=5
Checkin.service=3

satisfaction.prob=inv.logit(beta0.post+Inflight.entertainment*beta1.post + On.board.service*beta2.post +
summary(satisfaction.prob)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.8871  0.9201  0.9262  0.9259  0.9320  0.9529
```

These values represent the posterior mean, standard deviation and 95% CI for the satisfaction probability.

And, this is the plot of this probability density:

```
plot(density(satisfaction.prob))
```



Again, we can see that our posterior probability of being satisfied is centered around 0.93 similarly with the MCMC seen previously. It is remarkably high, as it is correctly predicting the probability of being satisfied minding that the values inputted to the model are high ratings (5,4,5,3)

CONCLUSIONS

In this case study, we analyzed the factors influencing customer satisfaction using logistic regression through three distinct **approaches**: - the classical frequentist approach, - the Bayesian approach using Markov Chain Monte Carlo (MCMC), - and the Bayesian approach using R2OpenBUGS.

Our aim was to predict customer satisfaction based on the predictors: Inflight entertainment, On-board service, Ease of online booking, and Check-in service.

Classical Frequentist Logistic Regression

Using the classical frequentist approach, we fitted a logistic regression model and obtained the maximum likelihood estimates of the regression coefficients. The main findings are:

- Model Fit: The model provided a good fit to the data, with the coefficients indicating the direction and magnitude of the influence of each predictor on customer satisfaction.
- Significance of Predictors: Inflight entertainment, On-board service, Ease of online booking, and Check-in service were all significant predictors of customer satisfaction.

Bayesian Logistic Regression using MCMC

In the Bayesian approach using MCMC, we incorporated prior beliefs about the parameters and used MCMC to estimate the posterior distributions of the regression coefficients. The main findings are:

- Posterior Distributions: This approach provided a full posterior distribution for each parameter, giving a richer understanding of the uncertainty around the estimates.
- Priors and Posteriors: The use of non-informative priors ensured that the results were driven mainly by the data. The posterior distributions of the parameters were consistent with the frequentist estimates but provided additional information about the uncertainty and potential variability of the estimates.
- Credible Intervals: The 95% credible intervals for the regression coefficients indicated that all predictors had a positive and significant impact on customer satisfaction, aligning with the frequentist approach results.

Bayesian Logistic Regression using R2OpenBUGS

Using R2OpenBUGS, we implemented the Bayesian logistic regression model, combining the flexibility of BUGS language with the MCMC approach. The main findings are:

- Posterior Distributions: Similar to the MCMC approach, R2OpenBUGS provided posterior distributions for each parameter, reinforcing the robustness of our findings.
- Results Consistency: The results from R2OpenBUGS were consistent with both the frequentist approach and the Bayesian MCMC approach, validating the reliability of the insights.

Comparative Insights

- Parameter Estimates: All three methods yielded similar point estimates for the regression coefficients, indicating consistent results across different approaches.
- Uncertainty: The Bayesian approaches (both MCMC and R2OpenBUGS) offered a more comprehensive view of uncertainty through posterior distributions and credible intervals, whereas the frequentist approach provided standard errors and confidence intervals.

Final Conclusion

Both the classical frequentist and Bayesian approaches (MCMC and R2OpenBUGS) confirmed that *Inflight entertainment*, *On-board service*, *Ease of online booking*, and *Check-in service* are significant predictors of customer satisfaction. The Bayesian approaches, while computationally more demanding, provided deeper insights into the parameter uncertainties and were consistent with the results obtained from the frequentist approach.