

Stat 811: Generalized Linear Models

Assignment 1, 2018

Name: Suphdeep Singh
Student ID: 44442467

Q1 The logarithmic distribution with parameter p ($0 < p < 1$) is a discrete distribution defined on the positive integers. It has to

$$f(y; p) = \begin{cases} \frac{-1}{\ln(1-p)} \frac{p^y}{y} & y = 1, 2, 3, 4, \dots \\ 0 & \text{otherwise.} \end{cases}$$

(a) General form of the function, which belong to family of exponents.

$$f(y; \theta, \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

$$f(y; p) = \frac{-p^y}{\ln(1-p) y}$$

$$= \exp \left(\frac{p^y}{y \ln(1-p)^{-1}} \right)$$

$$= \exp \left[y \log p - \log p - \log(\log(1-p)^{-1}) \right]$$

Comparing this to the general formula we have

$$\theta = \log P$$

$$b(\theta) = \log(\log(1-P)^{-1})$$

$$c(y, \phi) = -\log y$$

hence we can conclude, log series pdf belongs to the exponential family.

$$E(Y) = \sum_{y=1}^{\infty} f(y) \times y \quad \left\{ \begin{array}{l} \text{For a discrete series} \\ E(X) = \sum f(x) \times x \end{array} \right.$$

$$= \sum_{y=1}^{\infty} \frac{-1}{\ln(1-P)} \frac{P^y}{y} \times y$$

$$= \frac{-1}{\ln(1-P)} \sum_{y=1}^{\infty} P^y$$

$$= \frac{-1}{\ln(1-P)} \times \frac{P}{1-P}$$

[As $P \in (0, 1)$ it is sum of an infinite geometric progression with common ratio P]

$$= \frac{-P}{\ln(1-P)(1-P)}$$

$$\text{Var}(Y) = b''(\theta) \quad [\text{From general formula}]$$

$$b(\theta) = \log(\log(1-p)^{-1})$$

$$\text{As } E(Y) = b'(\theta) \quad \text{and} \quad \theta = \log p$$

Substituting $p = e^\theta$ in the expression for $E(Y)$ we have

$$E(Y) = \frac{-e^\theta}{(1-e^\theta) \ln(1-e^\theta)}$$

$$b'(\theta) = \frac{-e^\theta}{(1-e^\theta) \ln(1-e^\theta)}$$

$$b''(\theta) = \text{Applying the quotient rule}$$

$$\left[\frac{f}{g} \right]' = \frac{f'g - g'f}{g^2}$$

$$= \frac{-\frac{d}{d\theta}(e^\theta)(1-e^\theta) \ln(1-e^\theta) - \frac{d}{d\theta}((1-e^\theta) \ln(1-e^\theta))e^\theta}{((1-e^\theta) \ln(1-e^\theta))^2}$$

$$\Rightarrow \frac{-e^\theta(1-e^\theta)\ln(1-e^\theta) - (-e^\theta\ln(1-e^\theta) - e^\theta)e^\theta}{((1-e^\theta)\ln(1-e^\theta))^2}$$

$$\Rightarrow \frac{e^\theta(\ln(1-e^\theta) + e^{2\theta})}{\ln^2(-e^\theta + 1)(-e^\theta + 1)^2}$$

$$\Rightarrow \frac{-e^\theta\ln(1-e^\theta) + e^{2\theta}}{\ln^2(1-e^\theta)(1-e^\theta)^2}$$

$$b''(\theta) = \frac{-e^\theta\ln(1-e^\theta) + e^{2\theta}}{\ln^2(1-e^\theta)(1-e^\theta)^2}$$

$$\text{Var}(Y) = - \frac{p \ln(1-p) + p^2}{\ln^2(1-p)(1-p)^2}$$

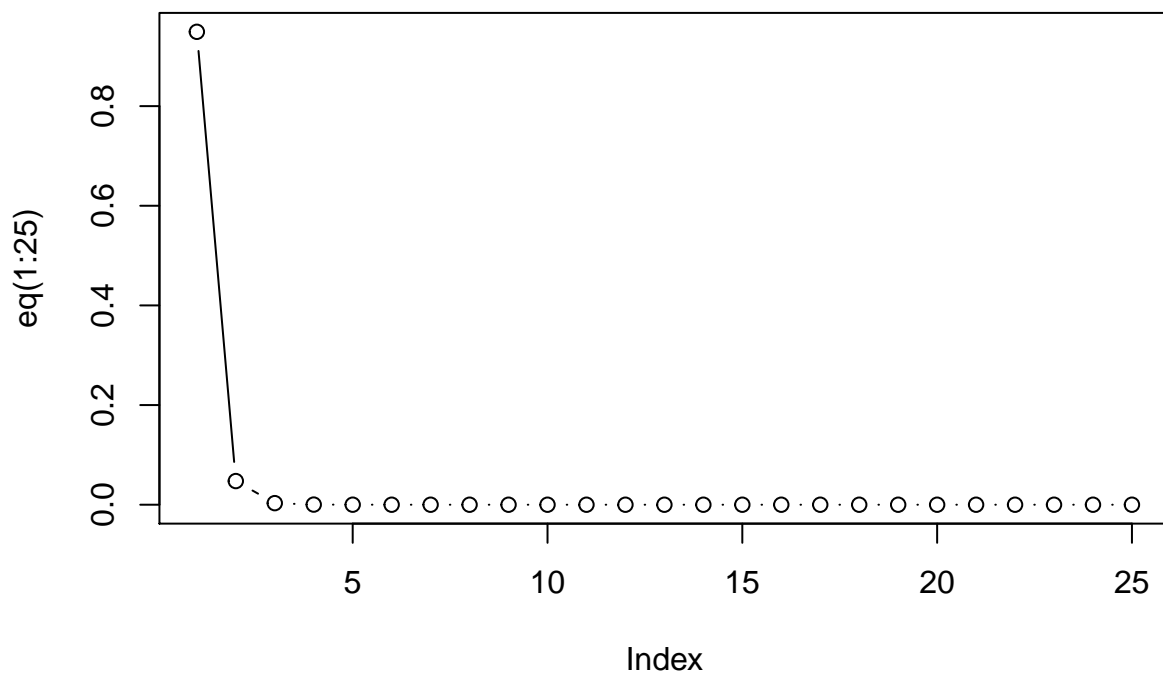
GLM

Sukhdeep Singh

August 13, 2018

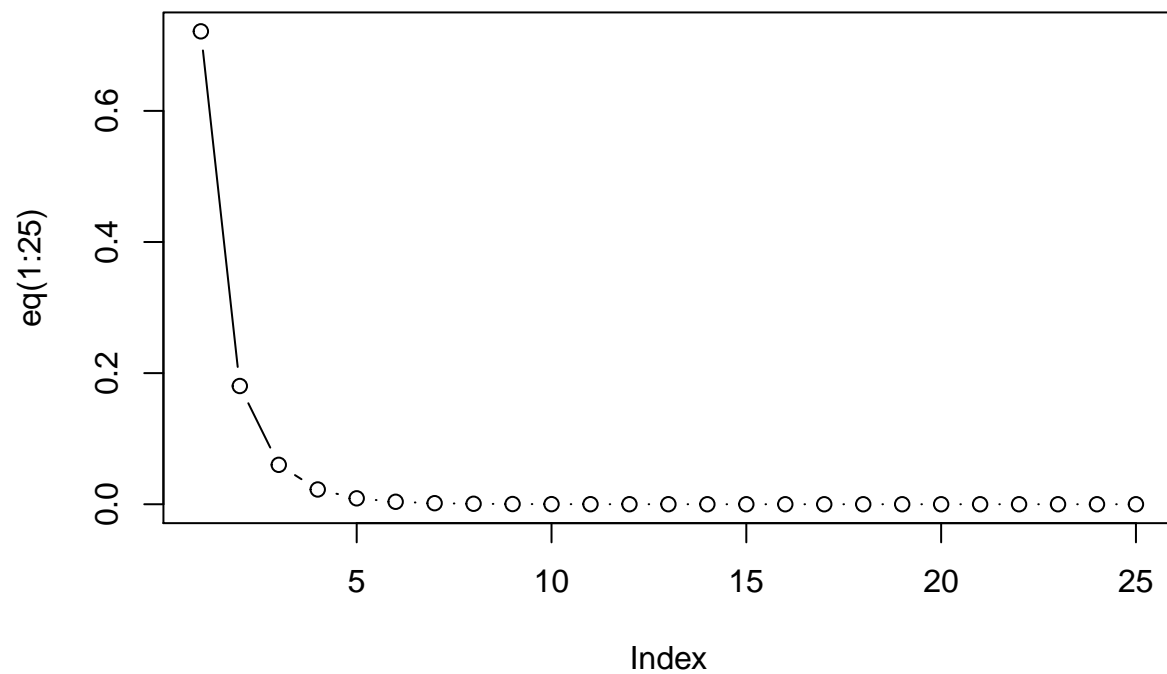
Output when $p=0.1$

```
eq=function(x){(-1*0.1^x)/(log(1-0.1)*x)}  
plot(eq(1:25), type='b')
```



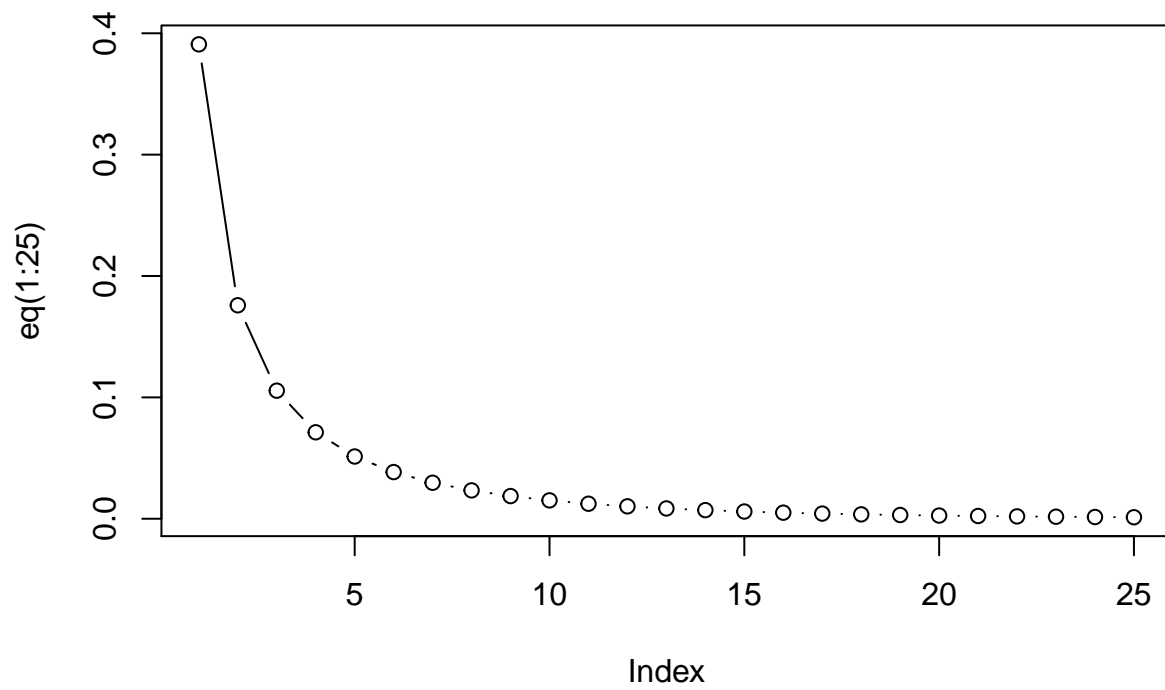
Output when $p=0.5$

```
eq=function(x){(-1*0.5^x)/(log(1-0.5)*x)}  
plot(eq(1:25), type='b')
```



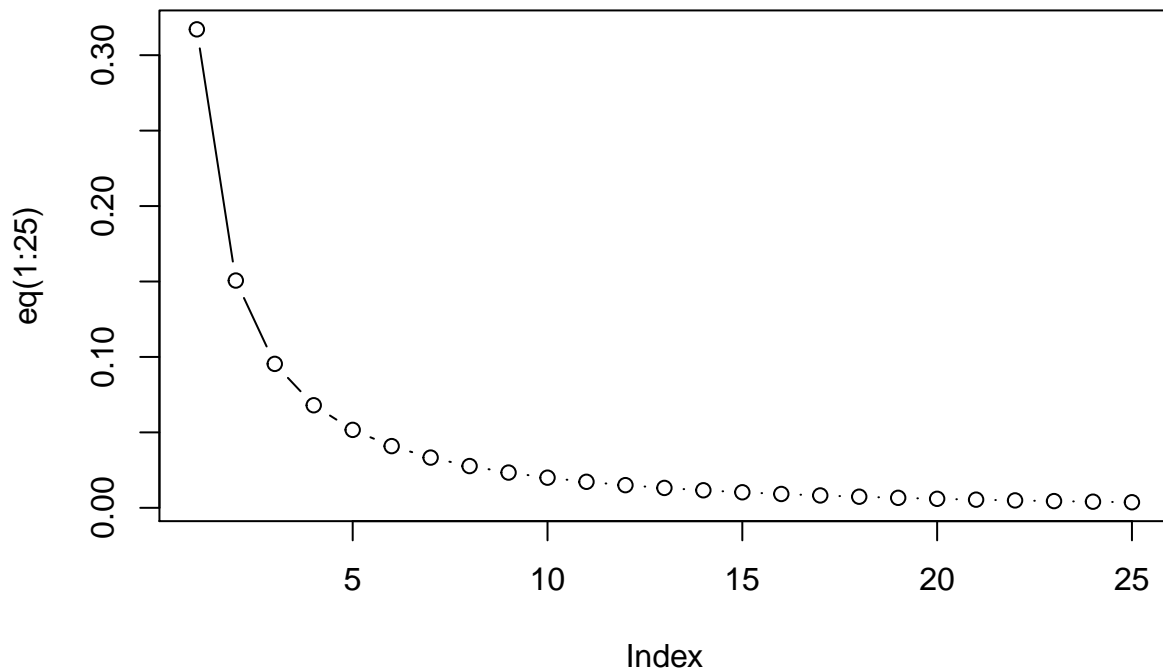
Ouput when p=0.9

```
eq=function(x){(-1*0.9^x)/(log(1-0.9)*x)}  
plot(eq(1:25), type='b')
```



Ouput when p=0.95

```
eq=function(x){(-1*0.95^x)/(log(1-0.95)*x)}  
plot(eq(1:25), type='b')
```



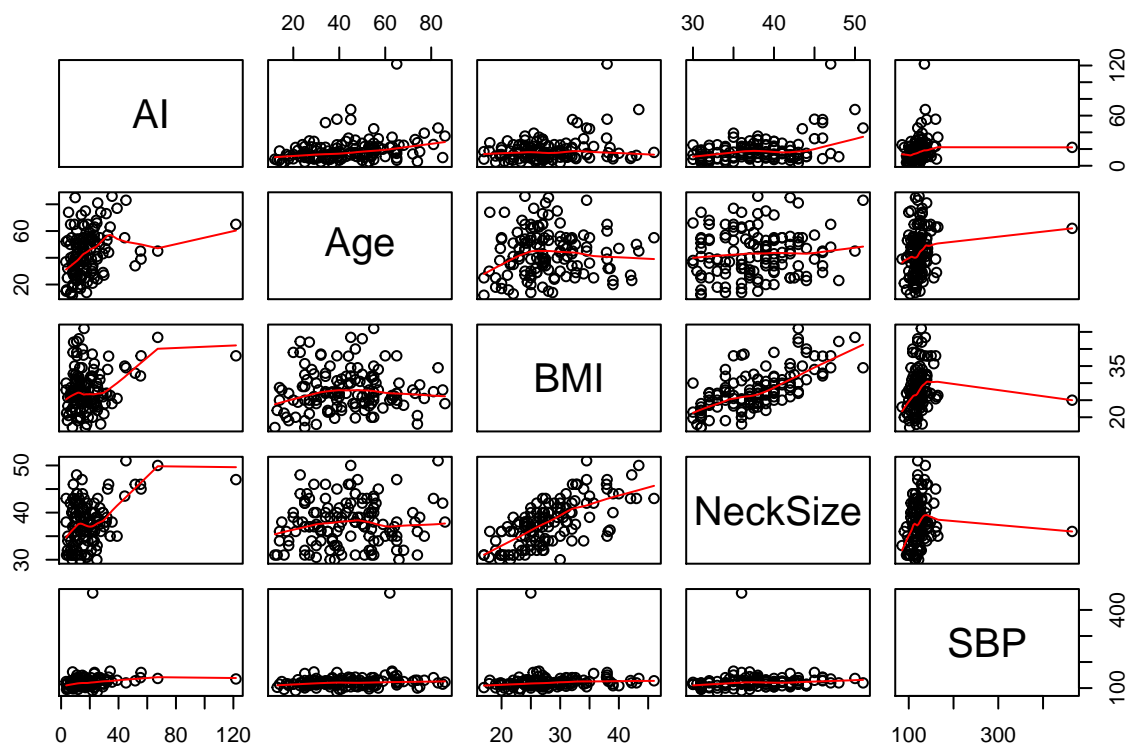
Question 2: Analysing the Sleep data

We are looking at sleep data for $n=133$ patients who have undergone a sleep study at canadian sleep clinics who are believed to have OSA: Obstructive sleep apnea.

- AI: is the response variable, which identifies the number of arousals per hour during sleep.
- Age: age of the patient in years
- Gender: Factor with 2 levels
- Alcohol: Factor with 2 levels (Yes/No)
- Caffeine: Factor with 2 levels (Yes/No)
- Neck size: Circumference of patient's neck in cm
- SBP: Systolic blood pressure (mm Hg)

Lets look at the scatterplot of the variables, and some summary statistics.

```
setwd("C:/Users/Lameware/Desktop/680/811/")
data = read.csv('sleepdata_assn1_2018.csv', header = T)
attach(data)
pairs(select(data,c('AI', 'Age', 'BMI', 'NeckSize', 'SBP')), panel = panel.smooth)
```

There appears to strong relationship b/w

1. AI and Age
2. AI and BMI
3. AI and NeckSize
4. Collinearity b/w BMI & Necksize

```
summary(data)
```

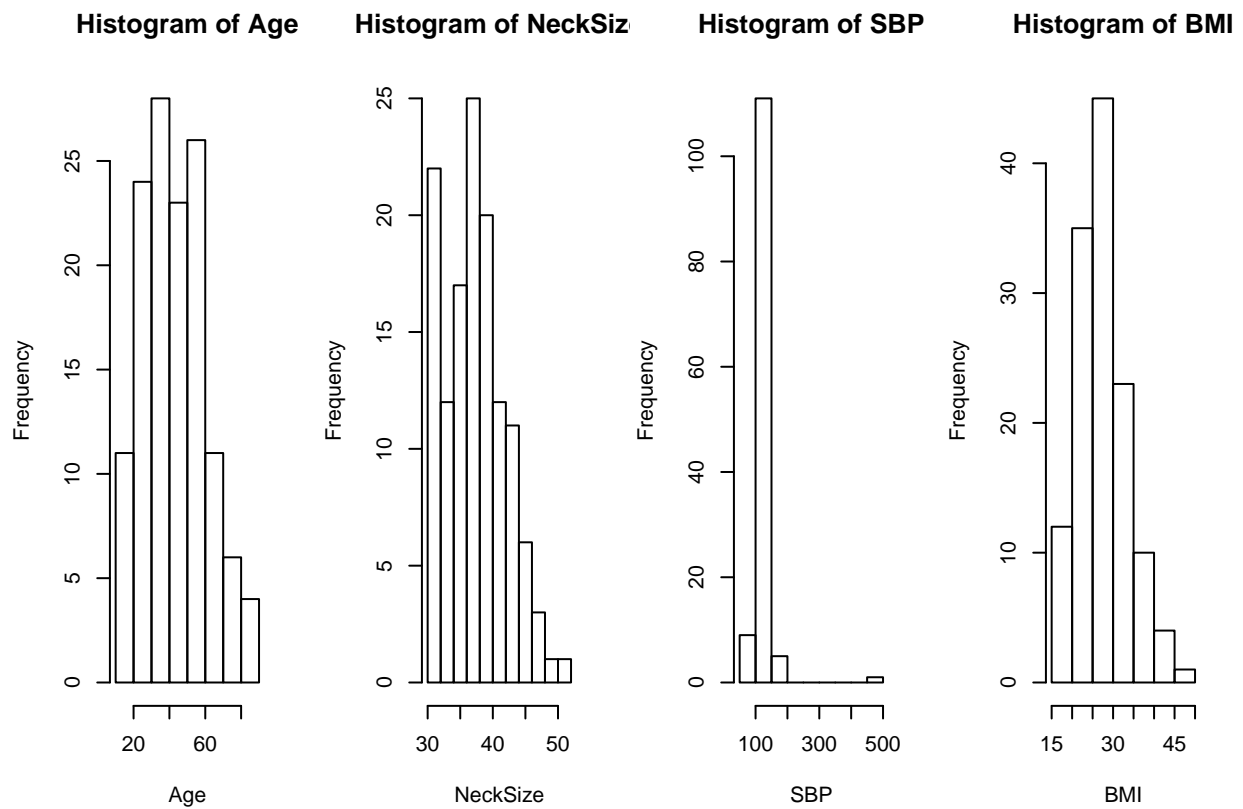
```
##      Age      Gender      Alcohol      Caffeine
##  Min.   :12.0   Min.    :1.000   Min.    :0.0000   Min.    :0.000
## 1st Qu.:30.0   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:1.000
## Median :43.0   Median :1.000   Median :0.0000   Median :1.000
## Mean   :43.5   Mean    :1.459   Mean    :0.1552   Mean    :0.771
## 3rd Qu.:55.0   3rd Qu.:2.000   3rd Qu.:0.0000   3rd Qu.:1.000
## Max.   :86.0   Max.    :2.000   Max.    :1.0000   Max.    :1.000
##
##      BMI      NeckSize      SBP      AI
##  Min.   :17.00   Min.    :30.00   Min.    : 85.0   Min.    : 3.30
## 1st Qu.:24.00   1st Qu.:34.00   1st Qu.:111.2   1st Qu.: 10.00
## Median :27.00   Median :37.75   Median :120.0   Median : 15.10
## Mean   :28.00   Mean    :37.79   Mean    :123.8   Mean    : 18.48
## 3rd Qu.:31.15   3rd Qu.:41.00   3rd Qu.:129.0   3rd Qu.: 22.70
## Max.   :46.00   Max.    :51.00   Max.    :465.0   Max.    :121.80
## NA's    :3      NA's    :3      NA's    :7      NA's    :1
```

Dataset appears to have some missing data, as denoted by the number of NA's in the corresponding variable. For the assessment, we are not dropping or substituting any missing value.

- SBP: We have a max value of 465, which is highly unlikely to be true. This could be a case of bad data and its a potential outlier for this dataset.
- AI: We have a max value of 121.80 which indicates person is waking up twice every minute, this again is a potential outlier.

Looking at this visual representaion of the continuos variables we can get a better understanding:

```
par(mfrow=c(1,4))
hist(Age, breaks=10)
hist(NeckSize,breaks=10)
hist(SBP,breaks = 10)
hist(BMI,breaks=10)
```



Frequency tables for the categorical variables

```
library(plyr)

## Warning: package 'plyr' was built under R version 3.4.4
## -----
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
## -----
##
## Attaching package: 'plyr'
```

```

## The following objects are masked from 'package:dplyr':
##
##   arrange, count, desc, failwith, id, mutate, rename, summarise,
##   summarize

## The following object is masked from 'package:purrr':
##
##   compact

count(data, 'Gender')

##   Gender freq
## 1      1   72
## 2      2   61

count(data, 'Caffeine')

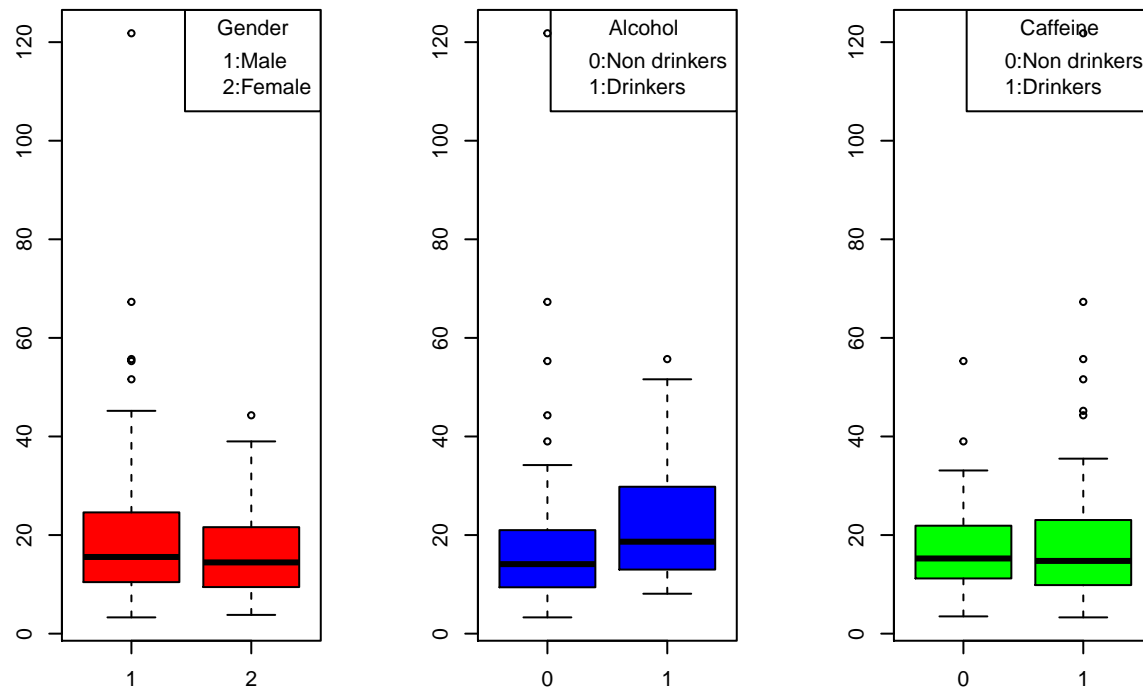
##   Caffeine freq
## 1         0   30
## 2         1  101
## 3        NA    2

count(data, 'Alcohol')

##   Alcohol freq
## 1         0   98
## 2         1   18
## 3        NA   17

par(mfrow=c(1,3))
boxplot(AI~Gender, col='red')
legend("topright", title = "Gender", c("1:Male","2:Female"))
boxplot(AI~Alcohol, col='Blue')
legend("topright", title = "Alcohol", c("0:Non drinkers","1:Drinkers"))
boxplot(AI~Caffeine, col='Green')
legend("topright", title = "Caffeine", c("0:Non drinkers","1:Drinkers"))

```

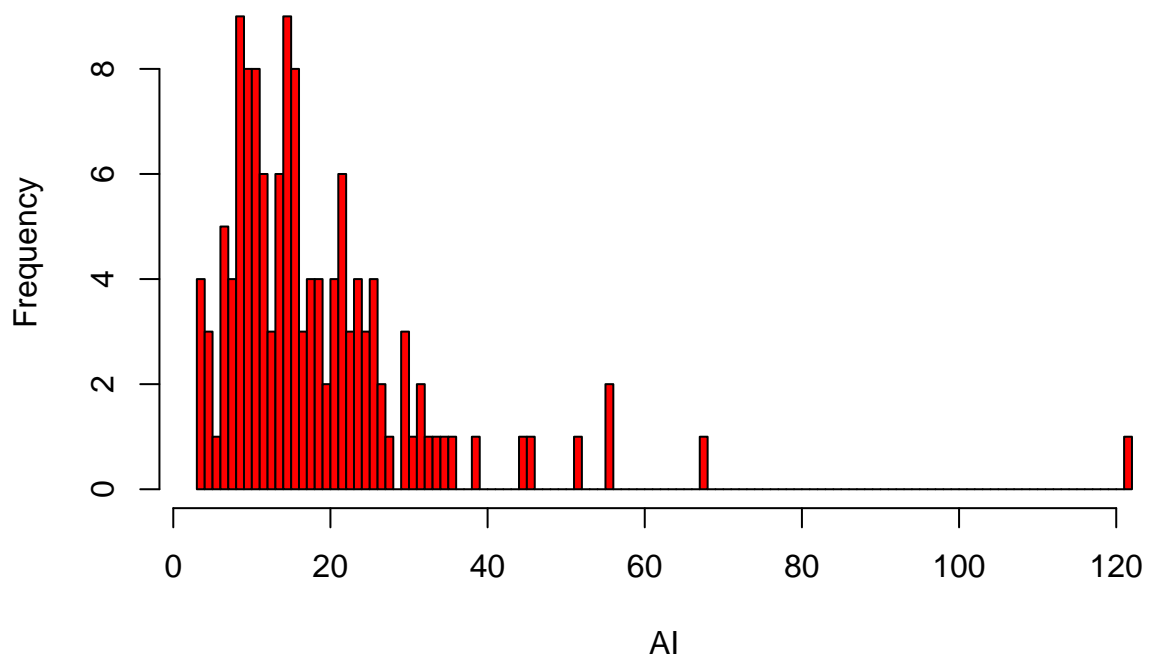


This shows that the variation in the 3 categorical factors is comparable for explaining AI.

Let's look at the distribution on the response variable

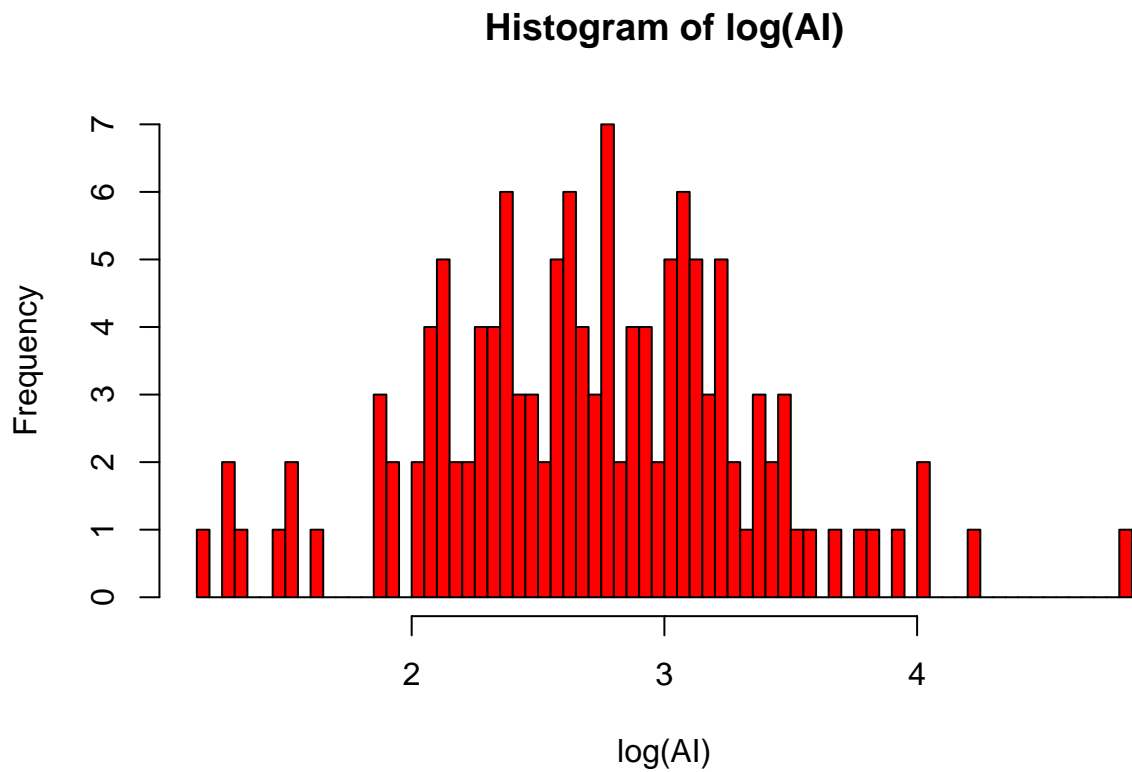
```
hist(AI,breaks = 100, col='red')
```

Histogram of AI



From the above histogram it appears that AI is heavily skewed, to meet the model assumption for normal model we apply the log transformation on the response.

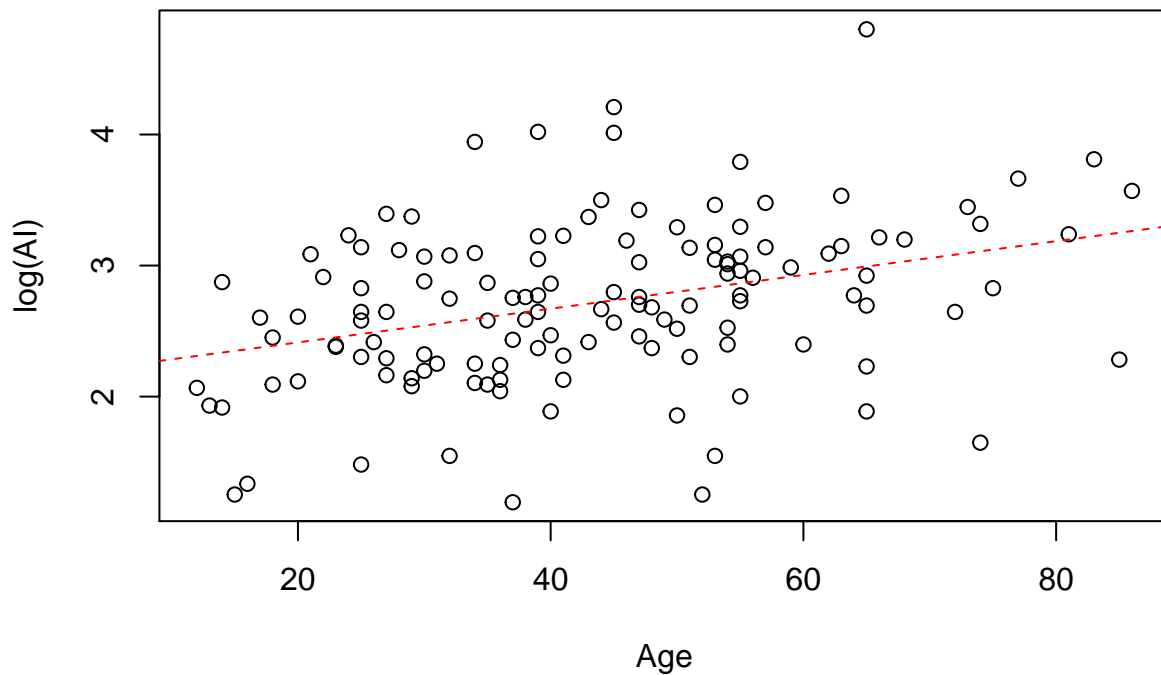
```
hist(log(AI),breaks=100,col='red')
```



Now, we can see that it is much more normally distributed. Let's proceed to observe relationship of Arousal index with continuous variables one by one.

Fitting a model: $\log(\text{AI}) \sim \text{Age}$

```
model1 = lm(log(AI) ~ Age)
plot( log(AI) ~ Age)
abline(model1, col='red', lty=2)
```



```
summary(model1)
```

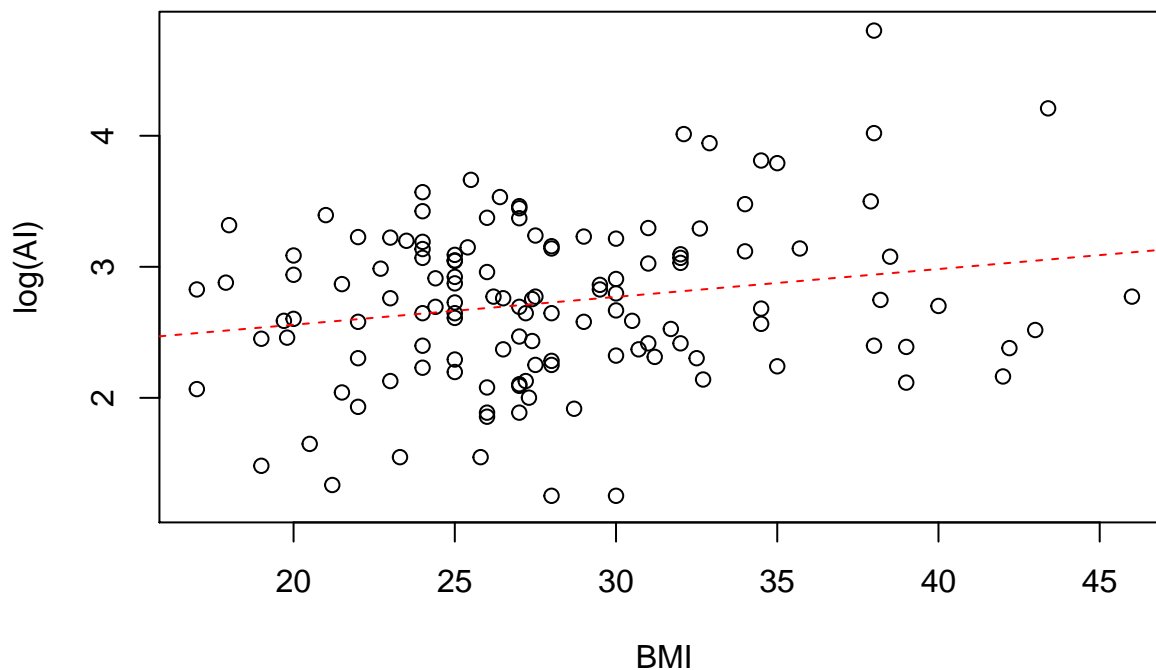
```
##
## Call:
## lm(formula = log(AI) ~ Age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.57260 -0.32906  0.01453  0.34031  1.80959
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.15563    0.14116  15.271  < 2e-16 ***
## Age          0.01288    0.00303   4.251 4.03e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5884 on 130 degrees of freedom
## (1 observation deleted due to missingness)
## Multiple R-squared:  0.122, Adjusted R-squared:  0.1153
## F-statistic: 18.07 on 1 and 130 DF, p-value: 4.033e-05
```

As, we can see from the model summary, Age is useful predictor for determining the AI in patients. Model has a F statistic of 18.07 and p-value $\rightarrow 0$ indicating a highly significant result.

Fitted model $\text{Log}(\text{AI}) = 2.15563 + 0.01288 * \text{Age}$

Fitting a model: $\log(\text{AI}) \sim \text{BMI}$

```
model1 = lm(log(AI) ~ BMI)
plot(log(AI) ~ BMI)
abline(model1, col='red', lty=2)
```



```
summary(model1)
```

```
##
## Call:
## lm(formula = log(AI) ~ BMI)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.5176 -0.4480  0.0363  0.4269  1.8620
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.132770   0.256605   8.311 1.23e-13 ***
## BMI          0.021254   0.008946   2.376  0.019 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6044 on 127 degrees of freedom
## (4 observations deleted due to missingness)
## Multiple R-squared:  0.04255,    Adjusted R-squared:  0.03501
```



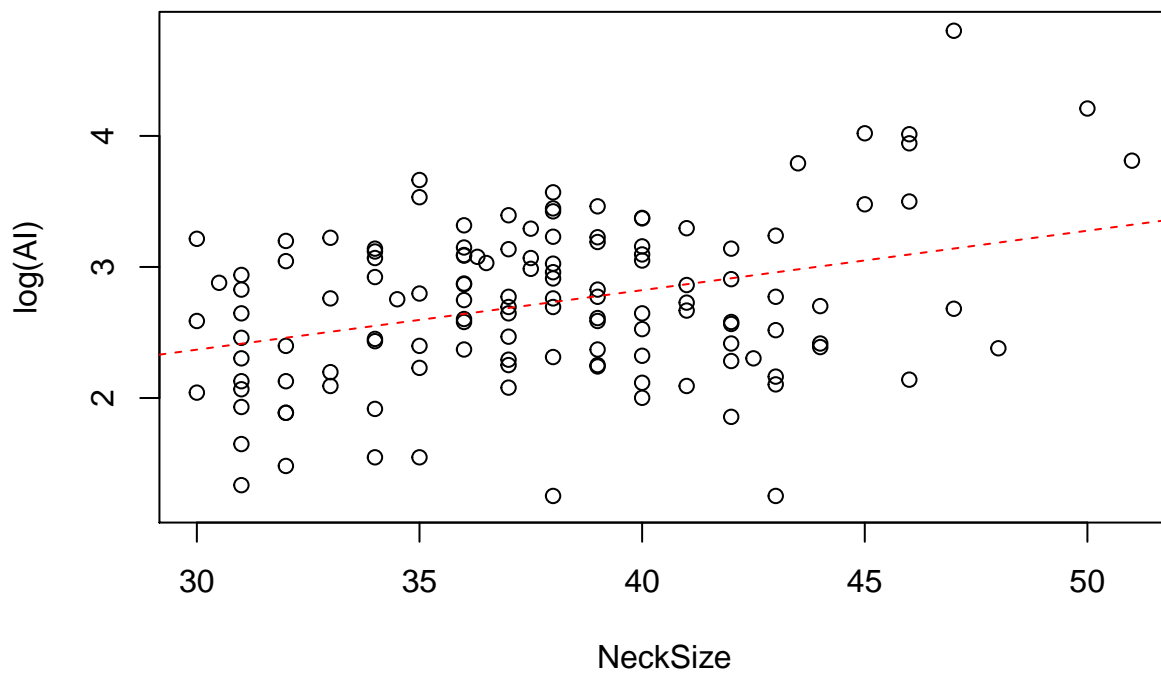
```
## F-statistic: 5.644 on 1 and 127 DF, p-value: 0.01901
```

BMI too looks like an important factor for explaining the AI, although this predictor is less significant as compared to Age.

Fitted model $\text{Log}(\text{AI}) = 2.132770 + 0.021254 * \text{BMI}$

Fitting a model: $\log(\text{AI}) \sim \text{NeckSize}$

```
model1 = lm(log(AI) ~ (NeckSize))
plot(log(AI) ~ NeckSize)
abline(model1, col='red', lty=2)
```



```
summary(model1)
```

```
##
## Call:
## lm(formula = log(AI) ~ (NeckSize))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.70627 -0.41335 -0.00492  0.44513  1.66181
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.00759    0.42130   2.392  0.0182 *
## NeckSize     0.04538    0.01105   4.107 7.14e-05 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5827 on 127 degrees of freedom
## (4 observations deleted due to missingness)
## Multiple R-squared:  0.1172, Adjusted R-squared:  0.1103
## F-statistic: 16.86 on 1 and 127 DF,  p-value: 7.139e-05
```

Necksize is good predictor of AI with an F-statistic of 16.86, p-value $\rightarrow 0$ also since there is a multicollinearity within NeckSize and BMI, we will include NeckSize as a predictor in our model.

Fitted model $\text{Log}(\text{AI}) = 1.00759 + 0.04538 * \text{NeckSize}$

Fitting a model: $\log(\text{AI}) \sim \text{SBP}$

```
model = lm(log(AI) ~ SBP)
summary(model)
```

```
##
## Call:
## lm(formula = log(AI) ~ SBP)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.47239 -0.41522 -0.02496  0.39568  2.04205
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.285317   0.203228  11.245  <2e-16 ***
## SBP          0.003519   0.001581   2.225  0.0279 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6005 on 123 degrees of freedom
## (8 observations deleted due to missingness)
## Multiple R-squared:  0.03869, Adjusted R-squared:  0.03088
## F-statistic: 4.951 on 1 and 123 DF,  p-value: 0.0279
```

SBP also makes significant contributions to the predictive model. We can include to the list of predictors

Fitted model $\text{Log}(\text{AI}) = 2.285317 + 0.003519 * \text{SBP}$

Fitting a model: $\log(\text{AI}) \sim \text{Age} + \text{SBP} + \text{NeckSize}$

```
model1 = lm(log(AI) ~ Age+SBP+NeckSize)
summary(model1)
```

```
##
## Call:
## lm(formula = log(AI) ~ Age + SBP + NeckSize)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.72537 -0.37600 -0.03262  0.38849  1.49704
```

```
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.570050    0.439372   1.297 0.196975
## Age         0.010081    0.002981   3.382 0.000972 ***
## SBP         0.002248    0.001480   1.519 0.131326
## NeckSize    0.037799    0.010708   3.530 0.000590 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5521 on 120 degrees of freedom
## (9 observations deleted due to missingness)
## Multiple R-squared:  0.2072, Adjusted R-squared:  0.1873
## F-statistic: 10.45 on 3 and 120 DF,  p-value: 3.686e-06
```

We can see that after fitting the model with our selected continuous predictors, SBP is no longer significant, hence dropping it from the the model

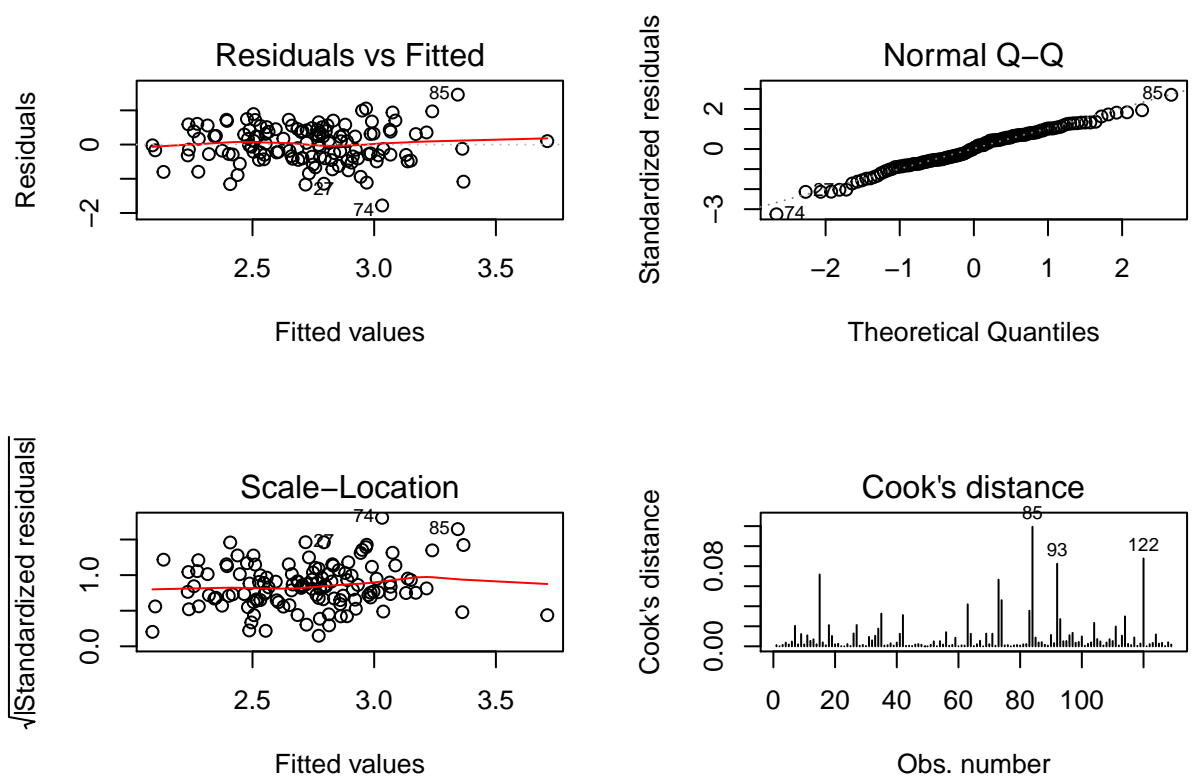
Fitted model $\text{Log}(\text{AI}) = 0.570050 + 0.010081 * \text{Age} + 0.003519 * \text{SBP} + 0.037799 * \text{NeckSize}$

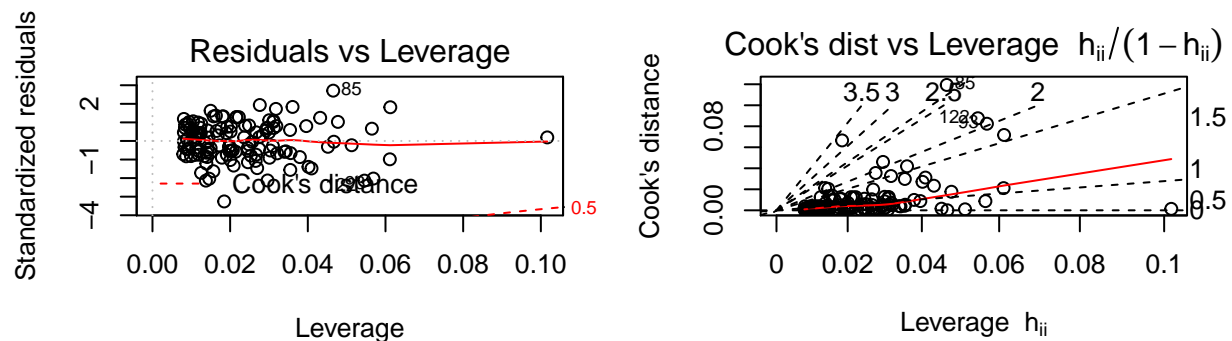
Final normal model

```
model1 = lm(log(AI) ~ Age+NeckSize)
summary(model1)

##
## Call:
## lm(formula = log(AI) ~ Age + NeckSize)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77998 -0.35526  0.02603  0.38899  1.45884
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.688739    0.406813   1.693 0.092924 .
## Age         0.011362    0.002864   3.967 0.000121 ***
## NeckSize    0.040772    0.010525   3.874 0.000171 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5516 on 126 degrees of freedom
## (4 observations deleted due to missingness)
## Multiple R-squared:  0.2152, Adjusted R-squared:  0.2028
## F-statistic: 17.28 on 2 and 126 DF,  p-value: 2.336e-07

par(mfrow=c(2,2))
plot(model1, which=1:6)
```





Fitted model $\log(AI) = 0.688739 + 0.011362 * Age + 0.040772 * NeckSize$

Diagnostics plot for normal model

- Residuals vs Fitted show a random scatter, no observable pattern (Assumptions confirmed)
- QQplot for residuals shows a straight line, therefore confirming the assumption of normal distribution of the residuals.
- Cook's distance plot for all the observations is well below the threshold for concern.
- Leverage plot and Scale plot are conforming to the model assumptions

Model parameters interpretation: We can see that with increase of 1 year in patients age, the log of AI increases by a factor of 0.011362 (considering necksize remains the same). Also for every cm increase of the necksize, log(AI) increases by a factor of 0.040772

This means a person who is older and has a larger neck circumference is more likely to suffer from OSA as compared to the others, however this model only explains a small proportion of the variation in AI (0.22), so it is quite possible that we have not yet discovered the root cause of OSA in the dataset.

Fitting categorical variables in the model,

```
model1=lm(log(AI) ~ Age + NeckSize + factor(Gender))
summary(model1)

##
## Call:
## lm(formula = log(AI) ~ Age + NeckSize + factor(Gender))
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.78257 -0.35044  0.01948  0.38584  1.45838
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.717940   0.484735   1.481 0.141097
## Age            0.011405   0.002901   3.931 0.000139 ***
## NeckSize       0.040100   0.012152   3.300 0.001261 **
## factor(Gender)2 -0.012632   0.112944  -0.112 0.911127
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5538 on 125 degrees of freedom
## (4 observations deleted due to missingness)
## Multiple R-squared:  0.2153, Adjusted R-squared:  0.1965
## F-statistic: 11.43 on 3 and 125 DF, p-value: 1.122e-06
```

We can see that the addition of categorical variable ‘Gender’ has no significant input to the model. All other combinations are checked in the appendix and none of them are significant.

Checking interactions of categorical variables with continuous variables

```
modell1 = lm(log(AI) ~ NeckSize*factor(Gender))
summary(modell1)

##
## Call:
## lm(formula = log(AI) ~ NeckSize * factor(Gender))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71202 -0.40515  0.00431  0.43506  1.62822
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.71404   0.64868   1.101 0.27312
## NeckSize       0.05234   0.01618   3.236 0.00155 **
## factor(Gender)2  0.48659   0.97613   0.498 0.61902
## NeckSize:factor(Gender)2 -0.01186   0.02613  -0.454 0.65054
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5865 on 125 degrees of freedom
## (4 observations deleted due to missingness)
## Multiple R-squared:  0.1198, Adjusted R-squared:  0.09864
## F-statistic: 5.669 on 3 and 125 DF, p-value: 0.001129
```

As we can see, interaction terms are not significant in this case and all the other possible combinations of continuous-categorical variables, for brevity we show only one of such cases. Code for rest of the cases is given in appendix.

Fitting a GLM with gamma distribution and link function as log

```
glm_model1 <- glm(AI ~ Age + NeckSize, family = Gamma(link= log))
```

```
summary(glm_model1)
```

```
##
## Call:
## glm(formula = AI ~ Age + NeckSize, family = Gamma(link = log))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.49435  -0.48887  -0.09824   0.27585   1.51674
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.410868   0.391911   1.048   0.296
## Age          0.011783   0.002759   4.270 3.81e-05 ***
## NeckSize     0.051335   0.010139   5.063 1.43e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.2823798)
##
##      Null deviance: 51.457  on 128  degrees of freedom
## Residual deviance: 36.150  on 126  degrees of freedom
## (4 observations deleted due to missingness)
## AIC: 919
##
## Number of Fisher Scoring iterations: 6
```

We can infer from the model summary that both the predictors are highly significant p-value->0

AIC for this model fit is 919

Fitted model $\text{Log}(AI) = 0.410868 + 0.011783 * \text{age} + 0.051335 * \text{necksize}$

Report

Aim: Analyse the sleep data for patients to identify the factors which could be the potential causes for OSA. This report summarizes the statistical results and procedures conducted as part of the exploratory data analysis and modelling.

Preliminary analysis (Normal model): The variable of interest is number of arousals per hour in a patient, we look at the scatterplots to determine any potential relationship b/w AI and various influence factors such as Age, BMI, NeckSize etc. We determine a positive relationship b/w 3 of these variables from observations. We also identify the some outliers in SBP and AI which report an unusually high value from rest of the observations. Dataset also contains some missing values which have been left as-is for analysis.

Statistical results and conclusion: The outcome variable AI is best explained by Age and Circumference of NeckSize. Fitted model $\text{log}(AI) = 0.688739 + 0.011362 * \text{Age} + 0.040772 * \text{NeckSize}$, this means that there's a higher chance of a older person with large neck-circumference to suffer from obstructive sleep apnea. However, these factors explain a small proportion of the variation in the data (0.21), it is possible that we have not looked at the root cause of OSA in this dataset.

Appendix

Fitting categorical variables to check for significance

```
modell1=lm(log(AI) ~ Age + NeckSize + factor(Alcohol))
summary(modell1)
```

```
##
## Call:
## lm(formula = log(AI) ~ Age + NeckSize + factor(Alcohol))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.73068 -0.35703 -0.01709  0.42122  1.51067
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.611470   0.455132   1.343 0.181926
## Age            0.010735   0.003161   3.396 0.000957 ***
## NeckSize       0.042180   0.011896   3.546 0.000580 ***
## factor(Alcohol)1 0.200852   0.147288   1.364 0.175506
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5638 on 108 degrees of freedom
## (21 observations deleted due to missingness)
## Multiple R-squared:  0.2243, Adjusted R-squared:  0.2027
## F-statistic: 10.41 on 3 and 108 DF,  p-value: 4.507e-06
```

Fitting categorical variables to check for significance

```
modell1=lm(log(AI) ~ Age + NeckSize + factor(Caffeine))
summary(modell1)
```

```
##
## Call:
## lm(formula = log(AI) ~ Age + NeckSize + factor(Caffeine))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.76229 -0.35368  0.00463  0.38599  1.47352
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.716580   0.421612   1.700 0.091730 .
## Age            0.011394   0.002911   3.914 0.000149 ***
## NeckSize       0.041422   0.010670   3.882 0.000168 ***
## factor(Caffeine)1 -0.075157   0.117883  -0.638 0.524947
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5563 on 123 degrees of freedom
## (6 observations deleted due to missingness)
## Multiple R-squared:  0.2186, Adjusted R-squared:  0.1995
## F-statistic: 11.47 on 3 and 123 DF,  p-value: 1.108e-06
```


Checking interactions of categorical variables with continuous variables

```
model1 = lm(log(AI) ~ NeckSize*factor(Alcohol))
summary(model1)
```

```
##
## Call:
## lm(formula = log(AI) ~ NeckSize * factor(Alcohol))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.66353 -0.41836  0.02802  0.43524  1.69657
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.87898    0.49667   1.770 0.079589 .
## NeckSize         0.04738    0.01326   3.574 0.000527 ***
## factor(Alcohol)1  0.80963    1.52285   0.532 0.596057
## NeckSize:factor(Alcohol)1 -0.01497    0.03884  -0.386 0.700578
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5927 on 108 degrees of freedom
## (21 observations deleted due to missingness)
## Multiple R-squared:  0.1426, Adjusted R-squared:  0.1188
## F-statistic: 5.988 on 3 and 108 DF,  p-value: 0.000817
```

Checking interactions of categorical variables with continuous variables

```
model1 = lm(log(AI) ~ NeckSize*factor(Caffeine))
summary(model1)
```

```
##
## Call:
## lm(formula = log(AI) ~ NeckSize * factor(Caffeine))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71990 -0.40987 -0.01469  0.43835  1.62544
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.98289    1.06393   1.864  0.0647 .
## NeckSize         0.02040    0.02791   0.731  0.4664
## factor(Caffeine)1 -1.20623    1.16187  -1.038  0.3012
## NeckSize:factor(Caffeine)1  0.03067    0.03047   1.007  0.3161
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5875 on 123 degrees of freedom
## (6 observations deleted due to missingness)
## Multiple R-squared:  0.1284, Adjusted R-squared:  0.1071
## F-statistic:  6.04 on 3 and 123 DF,  p-value: 0.0007169
```

Checking interactions of categorical variables with continuous variables

```
model1 = lm(log(AI) ~ Age*factor(Gender))
summary(model1)
```

```
##
## Call:
## lm(formula = log(AI) ~ Age * factor(Gender))
##
## Residuals:
```

	Min	1Q	Median	3Q	Max
	-1.65384	-0.29713	0.00628	0.35818	1.72266

```
##
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.2141512	0.1805557	12.263	< 2e-16 ***
Age	0.0133164	0.0039360	3.383	0.000951 ***
factor(Gender)2	-0.1597067	0.2883723	-0.554	0.580667
Age:factor(Gender)2	-0.0002413	0.0061610	-0.039	0.968817

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5868 on 128 degrees of freedom
## (1 observation deleted due to missingness)
## Multiple R-squared:  0.1405, Adjusted R-squared:  0.1203
## F-statistic: 6.973 on 3 and 128 DF,  p-value: 0.0002207
```

Checking interactions of categorical variables with continuous variables

```
model1 = lm(log(AI) ~ Age*factor(Caffeine))
summary(model1)
```

```
##
## Call:
## lm(formula = log(AI) ~ Age * factor(Caffeine))
##
## Residuals:
```

	Min	1Q	Median	3Q	Max
	-1.55731	-0.34416	0.01198	0.34481	1.81275

```
##
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.333764	0.287723	8.111	3.82e-13 ***
Age	0.010188	0.006481	1.572	0.118
factor(Caffeine)1	-0.241947	0.332024	-0.729	0.468
Age:factor(Caffeine)1	0.003624	0.007373	0.492	0.624

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5957 on 126 degrees of freedom
## (3 observations deleted due to missingness)
## Multiple R-squared:  0.1255, Adjusted R-squared:  0.1047
## F-statistic:  6.03 on 3 and 126 DF,  p-value: 0.0007168
```

Checking interactions of categorical variables with continuous variables

```
model1 = lm(log(AI) ~ NeckSize*factor(Alcohol))
summary(model1)
```

```
##
## Call:
## lm(formula = log(AI) ~ NeckSize * factor(Alcohol))
##
## Residuals:
```

	Min	1Q	Median	3Q	Max
	-1.66353	-0.41836	0.02802	0.43524	1.69657

```
##
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.87898	0.49667	1.770	0.079589 .
NeckSize	0.04738	0.01326	3.574	0.000527 ***
factor(Alcohol)1	0.80963	1.52285	0.532	0.596057
NeckSize:factor(Alcohol)1	-0.01497	0.03884	-0.386	0.700578

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5927 on 108 degrees of freedom
## (21 observations deleted due to missingness)
## Multiple R-squared:  0.1426, Adjusted R-squared:  0.1188
## F-statistic: 5.988 on 3 and 108 DF, p-value: 0.000817
```