## Review on ANLY545

*Introduction:* Here we are using different types of dataset, to find out the frequency distribution, relationship between two variables, contingency tables, Chi-Square test for association between the two variables, to handle zero values, find the pattern of residuals and all the two way interaction of the predictors using log linear model and logistics regression.

*Step 0- Warmup:* Using poison distribution we are generating 6 numbers. I have created dataset which has 3 columns Gender, Product and Price. Now, These 6 numbers I am assigning in all possible combination of gender and product labels. Finding the frequency distribution of females in 100 queues of length 10 in a London Underground station from WomenQueue data from VCD package.

## Code:

set.seed(234)

nums<-rpois(6,4)

gender<-sample(c("Male", "Female"), size=6, replace = TRUE)

product<-sample(c("A", "B", "C"), size=6, replace = TRUE)

data\_set<-cbind(gender,product,nums)

data<-as.data.frame(data\_set)

library(plyr)

df <- rename(data, c("nums" = "price"))

df

# Second way

dataset <- expand.grid(gender=c("female", "male"), product=c("A","B","C"))

set.seed(234)

price <- rpois(6, 4)

(ProductPrice <- cbind(dataset,price))

# Frequency Distribution

library(vcd)

data(WomenQueue)

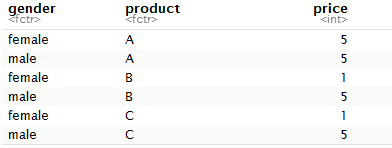
head(WomenQueue)

barplot(WomenQueue, main="Women Queue Depature Distribution",xlab="Counts",ylab="Frequency")

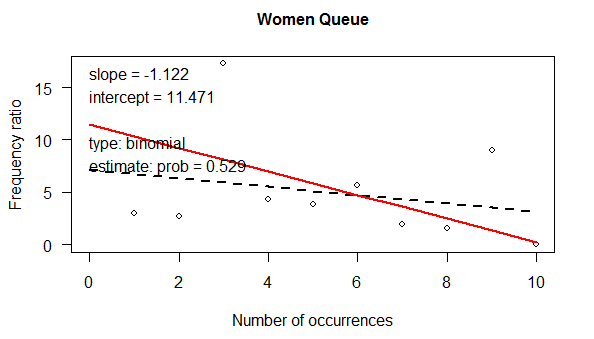
Women\_Ord=Ord\_plot(obj=WomenQueue,main="Women Queue")

Women\_Ord

## Output:







Interpretation: Since slope is Negative and Intercept is positive, it is determined that this is a Binomial Distribution.

## *Step1- Contingency tables:*

A **contingency table** is particularly useful when a large number of observations need to be condensed into a smaller format whereas a complex (flat) table is a type of contingency table that is used when creating just one single table as opposed to multiple ones. i.e. assemble it into a table that shows the layout of the original data in a manner that allows the reader to gain an overall summary of the original data.

We are using here Hospital dataset from VCD package and we are trying to find the proportions of patients with differing length of stay for each value of visit frequency.

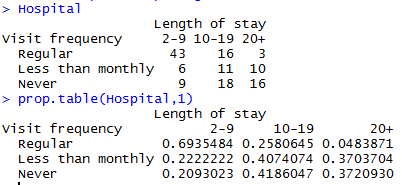
## Code:

data("Hospital", package="vcd")

Hospital

prop.table(Hospital,1)

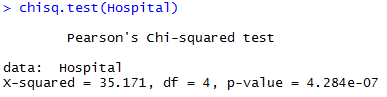
## Output:



The **chi-square** test for independence, also called **Pearson's chi-square test** or the chi-square test of association, is used to discover if there is a relationship between two categorical variables from a single population. It is used to determine whether there is a significant association between the two variables.

Here we are trying to find association of Visit Frequency like Regular, Less than monthly and Never.

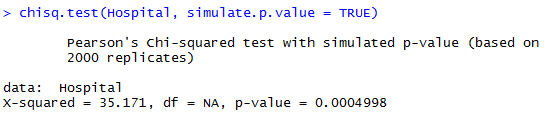
## Output:



Interpretation: As the p-value is smaller than the .05 significance level, we reject the null hypothesis. Therefore variables are associated, there is no dependency.

We don’t trust the asymptotic p-value from the chi-square test because the sample size is relatively too small. For this type scenario, we are going to use the Monte Carlo simulation technique.

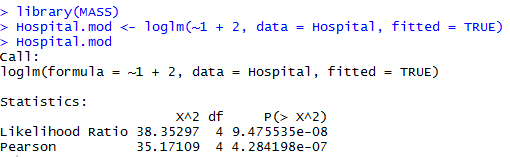
## Output:



Interpretation: Still under the Monte Carlo Simulation, we see that the p-value is smaller than alpha. It can be reassured that the variables are associated.

Under the loglm function in r we can find out the standard chi-square test value, here the log-likelohood ratio statistics for testing the current model within a saturated model. So that p-value is coming too small and we can say that the variables are associated.

## Output:



*Log Linear Model:* Log-linear models go beyond a single summary statistics and specify how the cell counts depend on the levels of categorical variables. They model the association and interaction patterns among categorical variables. The log-linear modeling is natural for Poisson, Multinomial and Product-Multinomial sampling. They are appropriate when there is no clear distinction between response and explanatory variables, or there are more than two responses. This is a major difference between logistic models and log-linear models.

Here we are considering the titanic dataset to estimate the survival on the titanic using various log linear models. We have done a 4 way cross- classification of 2201 passengers and crew, according to

- Gender (G): M vs. F  
- Age (A): Adult vs. Child  
- Class (C): 1st, 2nd, 3rd, Crew  
- Survival (S): Died vs. Survived

There are 8 cells with zero frequencies and to handle these zero values we are adding 0.5 values in the titanic dataset and used Generalized Linear models.

## Code:

data("Titanic")

Titanic

Titanic.nonzero <- Titanic + 0.5

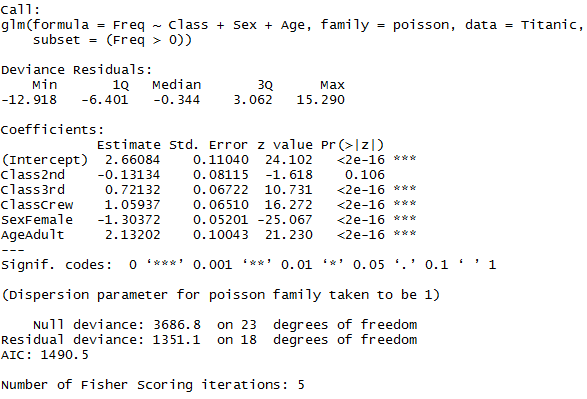
Titanic.glm1 <- glm(Freq~ Class + Sex + Age, data= Titanic, subset= (Freq> 0), family= poisson)

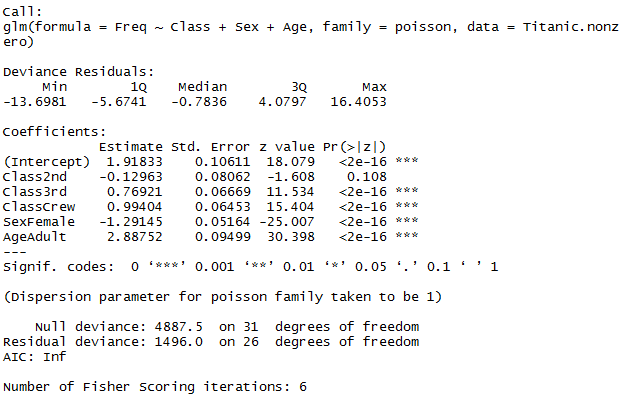
Titanic.glm2 <- glm(Freq~ Class + Sex + Age, data= Titanic.nonzero, family= poisson)

summary(Titanic.glm1)

summary(Titanic.glm2)

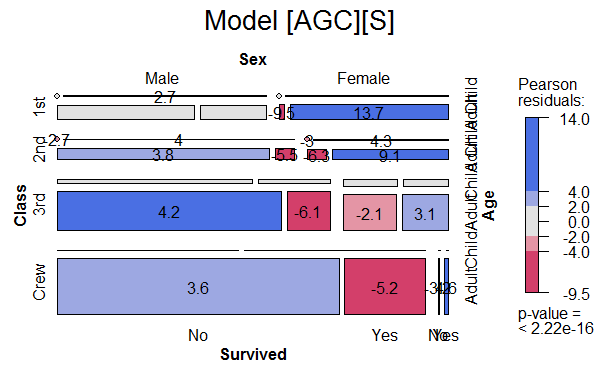
## Output:





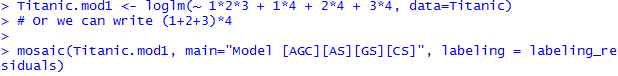
We can see that when we excluded the zero values the AIC value is small that means log linear model fitted correctly but when we are adding 0.5 values in titanic dataset the AIC value is coming Inf but p-values are coming < 0.5. So, we can conclude that GLM model can used to handle zero values.

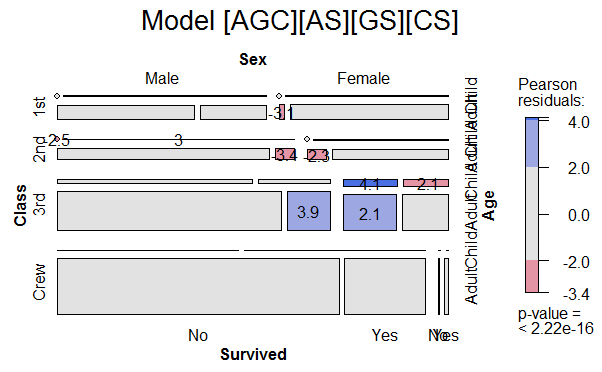
**Mosaic Plot:**

****

Interpretation:There is strong association between non-survival to crew and 3rd class. It seemed that higher class is more likely to be saved and survived at last. Regardless of the class type, children are more likely to survive. It seemed that children is first helped and saved. For first class, it seemed that female is more likely to survive.

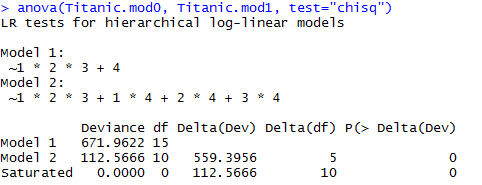
Fit a main effects model for survival, [AGC][AS][GS][CS], that includes an association of survival with each of age, gender and class.





Interpretation: After we take possible association between gender, age, and class with survival, mosaic plot showed that remaining association reduced a lot. Many associations have been captured when we have considered 1-1 association.

We want to see if we use ANOVA technique then what is association between variables.

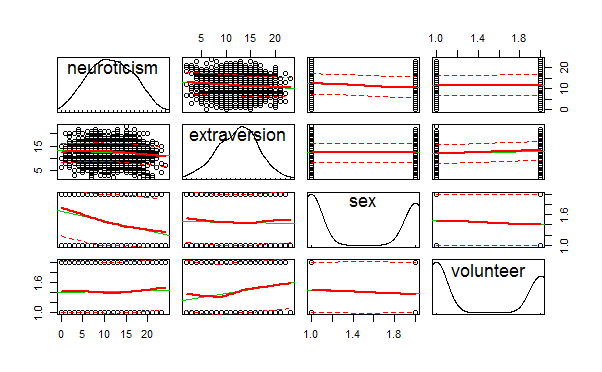


Interpretation: From ANOVA test, we find that the association decreased a lot after we add different association combination to the model.

*Logistic Regression:*  Logistic Regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary).  Like all regression analyses, the logistic regression is a predictive analysis.  Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

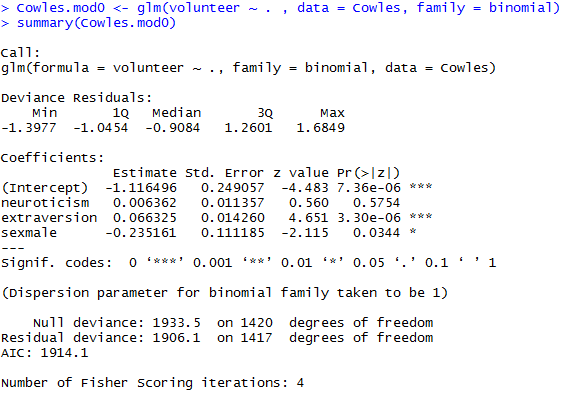
We want to see the scatter plot between male and female volunteer of Cowles dataset from the car package.





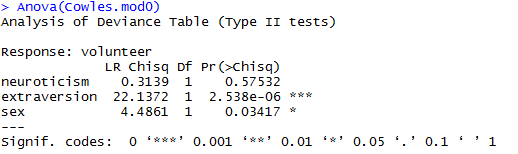
Interpretation: From scatter plot, it seemed that male and female volunteer not equally often, but the decision to volunteer seemed to be related to extraversion.

Fit a main effects model with glm(), predicting volunteer from sex, neuroticism and extraversion. And here family of distribution is binomial.



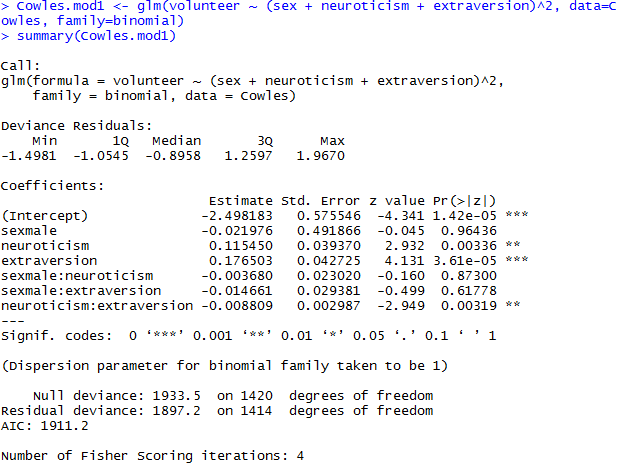
Interpretation: From the result, it seemed male is slightly unlikely to volunteer and there is positive correlation between volunteer and extraversion. While other factors are constant, if it is male, there is 20.96% decrease of the odds of volunteer. While other factors are constant, the increase of extraversion will lead to a 6.86% increase of the odds of volunteer.

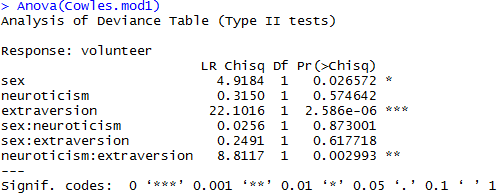
From the Anova technique-



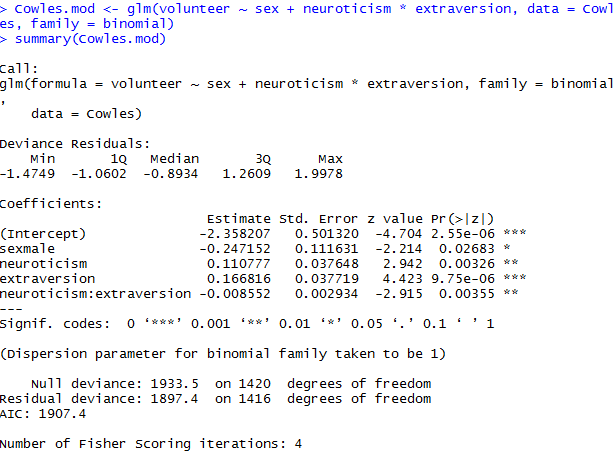
* From ANOVA result, there seemed to be association between features.

Now, we are trying to fit the logistic regression model, containing main effects and all two-way interactions of the predictors.

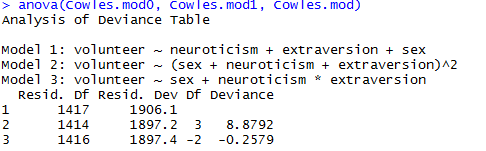




* It seemed that sex is still one important feature. The combination of neuroticisms and extraversion was significant as well.

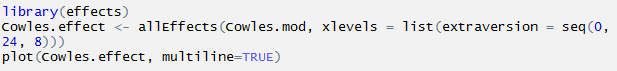


* From the above result, we verified that all factors, including the interaction factor, are statistically significant.

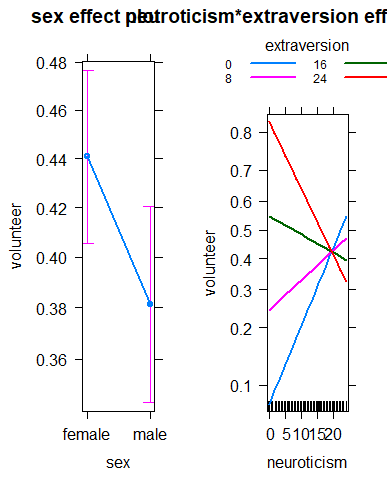


* It seemed that it is beneficial to add interaction factor to the model. However, it seemed that we needn't add the interaction of all factors. Only the interaction of neuroticism and extraversion is enough.

There is another way to check associations between male and female volunteer and the different groups of combination of extraversion and neuroticism.



## Output:



Interpretation: Effect plot showed similar conclusion, male and female are not equally often for volunteer. The different groups of combination of extraversion and neuroticism indicated different probability of volunteer.