Review on ANLY545

Rajib Kumar De

December 7, 2018

# memory clear  
rm(list = ls(all = TRUE))  
  
############ Step 0- Warmup - Making dataset #############  
  
# Question - The goal is to generate a dataset which shows the price people paid for three products  
# Set seed equal to 234  
# From a poison distrbution with ??=4 ??=4 generate 6 numbers and store it in nums  
# Make a data set which has 3 columns (gender, product and price)  
# Levels: gender (male and female), product (A, B and C)  
# Now assigne these nums to all possible combinations of gender and product levels  
  
# First way  
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

## gender product price  
## 1 Female B 5  
## 2 Female B 5  
## 3 Female A 1  
## 4 Male B 5  
## 5 Female A 1  
## 6 Female C 5

# 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))

## gender product price  
## 1 female A 5  
## 2 male A 5  
## 3 female B 1  
## 4 male B 5  
## 5 female C 1  
## 6 male C 5

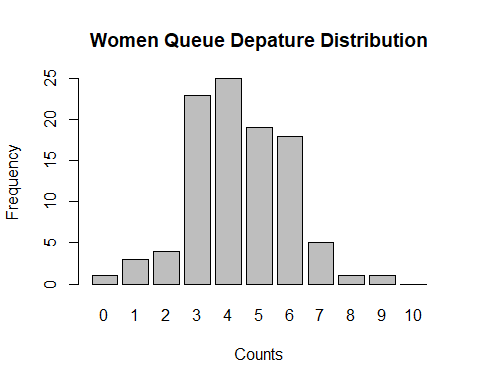
############ Distributions - WomenQueue from vcd Extra Data #############  
  
# Find the frequency distribution of females in 100 queues of length 10 in a London Underground station. - Find the best distribution for the data  
  
library(vcd)

## Loading required package: grid

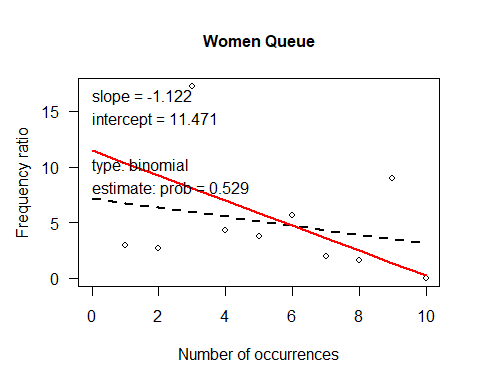
data(WomenQueue)  
head(WomenQueue)

## nWomen  
## 0 1 2 3 4 5   
## 1 3 4 23 25 19

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



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



Women\_Ord

## Intercept Slope   
## 11.471435 -1.121733

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

############ Step1- Contingency tables - Two-way tables #############  
  
# Two-way tables (Question 1)  
data("Hospital", package="vcd")  
Hospital

## Length of stay  
## Visit frequency 2-9 10-19 20+  
## Regular 43 16 3  
## Less than monthly 6 11 10  
## Never 9 18 16

# Two-way tables (Question 1) - Find the row proportions for this table, i.e., the proportions of patients with differing length of stay for each value of visit frequency.  
# What do you see there that relates to the question of independence of the table variables?  
  
prop.table(Hospital,1)

## Length of stay  
## Visit frequency 2-9 10-19 20+  
## Regular 0.6935484 0.2580645 0.0483871  
## Less than monthly 0.2222222 0.4074074 0.3703704  
## Never 0.2093023 0.4186047 0.3720930

# Two-way tables (Question 2) - Carry out a \(\chi^2\) test for association between the two variables. What do you conclude?  
  
chisq.test(Hospital)

##   
## Pearson's Chi-squared test  
##   
## data: Hospital  
## X-squared = 35.171, df = 4, p-value = 4.284e-07

# 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.  
  
# Two-way tables (Question 3) - Say you donâ t trust the asymptotic p-value from the test because the sample size is relatively small. See help(chisq.test)for how to get a Monte Carlo simulated p-value. Do it.  
  
chisq.test(Hospital, simulate.p.value = TRUE)

##   
## Pearson's Chi-squared test with simulated p-value (based on 2000  
## replicates)  
##   
## data: Hospital  
## X-squared = 35.171, df = NA, p-value = 0.0004998

# 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.  
  
# Two-way tables (Question 4) Extra Credit - Use MASS::loglm() to carry out the standard \(\chi^2\) test  
library(MASS)  
Hospital.mod <- loglm(~1 + 2, data = Hospital, fitted = TRUE)  
Hospital.mod

## Call:  
## loglm(formula = ~1 + 2, data = Hospital, fitted = TRUE)  
##   
## Statistics:  
## X^2 df P(> X^2)  
## Likelihood Ratio 38.35297 4 9.475535e-08  
## Pearson 35.17109 4 4.284198e-07

############ Loglinear Model - Survival on the Titanic #############  
  
data("Titanic")  
Titanic

## , , Age = Child, Survived = No  
##   
## Sex  
## Class Male Female  
## 1st 0 0  
## 2nd 0 0  
## 3rd 35 17  
## Crew 0 0  
##   
## , , Age = Adult, Survived = No  
##   
## Sex  
## Class Male Female  
## 1st 118 4  
## 2nd 154 13  
## 3rd 387 89  
## Crew 670 3  
##   
## , , Age = Child, Survived = Yes  
##   
## Sex  
## Class Male Female  
## 1st 5 1  
## 2nd 11 13  
## 3rd 13 14  
## Crew 0 0  
##   
## , , Age = Adult, Survived = Yes  
##   
## Sex  
## Class Male Female  
## 1st 57 140  
## 2nd 14 80  
## 3rd 75 76  
## Crew 192 20

# Loglinear Model (Question 1) - One slight complication here is that there are 8 cells with zero frequencies. Four of these (male and female children in 1st and 2nd class who died) should be considered sampling zeros, but 4 (children among the crew) should probably be considered structural zeros (cells where data could not occur. Suggest a method to take care of these zeros  
  
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)

## Warning in dpois(y, mu, log = TRUE): non-integer x = 0.500000  
  
## Warning in dpois(y, mu, log = TRUE): non-integer x = 0.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 35.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 0.500000  
  
## Warning in dpois(y, mu, log = TRUE): non-integer x = 0.500000  
  
## Warning in dpois(y, mu, log = TRUE): non-integer x = 0.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 17.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 0.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 118.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 154.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 387.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 670.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 4.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 13.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 89.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 3.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 5.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 11.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 13.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 0.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 1.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 13.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 14.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 0.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 57.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 14.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 75.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 192.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 140.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 80.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 76.500000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 20.500000

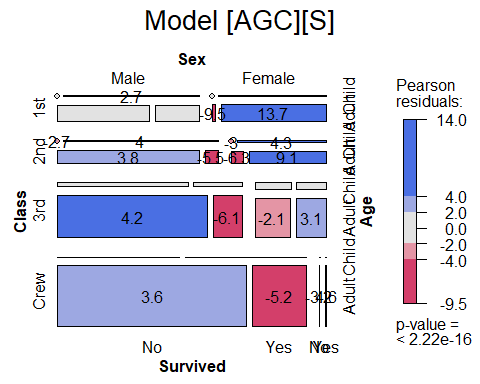
summary(Titanic.glm1)

##   
## Call:  
## glm(formula = Freq ~ Class + Sex + Age, family = poisson, data = Titanic,   
## subset = (Freq > 0))  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -12.918 -6.401 -0.344 3.062 15.290   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.66084 0.11040 24.102 <2e-16 \*\*\*  
## Class2nd -0.13134 0.08115 -1.618 0.106   
## Class3rd 0.72132 0.06722 10.731 <2e-16 \*\*\*  
## ClassCrew 1.05937 0.06510 16.272 <2e-16 \*\*\*  
## SexFemale -1.30372 0.05201 -25.067 <2e-16 \*\*\*  
## AgeAdult 2.13202 0.10043 21.230 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 3686.8 on 23 degrees of freedom  
## Residual deviance: 1351.1 on 18 degrees of freedom  
## AIC: 1490.5  
##   
## Number of Fisher Scoring iterations: 5

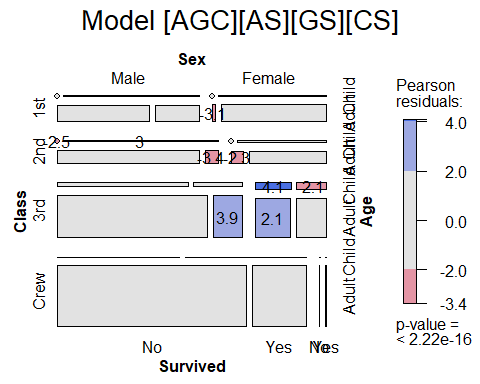
summary(Titanic.glm2)

##   
## Call:  
## glm(formula = Freq ~ Class + Sex + Age, family = poisson, data = Titanic.nonzero)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -13.6981 -5.6741 -0.7836 4.0797 16.4053   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.91833 0.10611 18.079 <2e-16 \*\*\*  
## Class2nd -0.12963 0.08062 -1.608 0.108   
## Class3rd 0.76921 0.06669 11.534 <2e-16 \*\*\*  
## ClassCrew 0.99404 0.06453 15.404 <2e-16 \*\*\*  
## SexFemale -1.29145 0.05164 -25.007 <2e-16 \*\*\*  
## AgeAdult 2.88752 0.09499 30.398 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 4887.5 on 31 degrees of freedom  
## Residual deviance: 1496.0 on 26 degrees of freedom  
## AIC: Inf  
##   
## Number of Fisher Scoring iterations: 6

# Loglinear Model (Question 2) - It is natural to consider Survival as the natural response variable, and the remaining variables as explanatory. Therefore, all models should include the high-order term among age, gender and class. Therefore, the minimal null model is [AGC][S], which asserts that survival is jointly independent of Age, Sex and Class. Fit this model, and obtain a mosaic plot. Interpret the pattern of the residuals in this mosaic plot.  
  
Titanic.mod0 <- loglm(~ 1\*2\*3 + 4, data=Titanic)  
mosaic(Titanic.mod0, main="Model [AGC][S]", labeling = labeling\_residuals)



# 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.  
# Regardiless of the class type, children is 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.  
  
  
# Loglinear Model (Question 3) Extra Credit - Fit a main effects model for survival, [AGC][AS][GS][CS], that includes an association of survival with each of age, gender and class. Is this an adequate fit? What does the pattern of residuals tell you about remaining associations?  
  
Titanic.mod1 <- loglm(~ 1\*2\*3 + 1\*4 + 2\*4 + 3\*4, data=Titanic)  
# Or we can write (1+2+3)\*4  
  
mosaic(Titanic.mod1, main="Model [AGC][AS][GS][CS]", labeling = labeling\_residuals)



# Interpretation: After we take possible association between gender, age, and class with survivual, mosaic plot showed that remaining association reduced a lot.   
## Many association has been captured when we have considered 1-1 association.  
  
  
anova(Titanic.mod0, Titanic.mod1, test="chisq")

## LR tests for hierarchical log-linear models  
##   
## Model 1:  
## ~1 \* 2 \* 3 + 4   
## Model 2:  
## ~1 \* 2 \* 3 + 1 \* 4 + 2 \* 4 + 3 \* 4   
##   
## Deviance df Delta(Dev) Delta(df) P(> Delta(Dev)  
## Model 1 671.9622 15   
## Model 2 112.5666 10 559.3956 5 0  
## Saturated 0.0000 0 112.5666 10 0

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

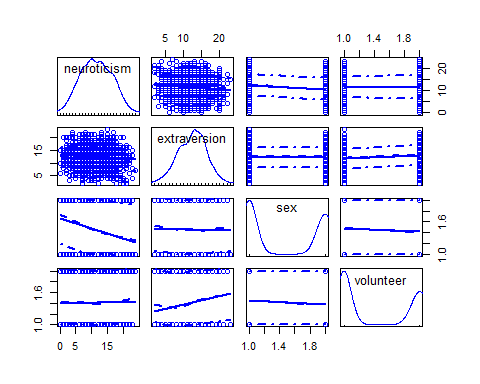
############ Logistic Regression - Volunteering for a psychology experiment #############  
  
library(car)

## Loading required package: carData

data(Cowles, package="car")

## Warning in data(Cowles, package = "car"): data set 'Cowles' not found

# Logistic regression (Question 1) - Use car::scatterplotMatrix() and describe the result  
  
scatterplotMatrix(Cowles)



# 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.  
  
# Logistic regression (Question 2) - Fit a main effects model with glm(), predicting volunteer from sex, neuroticism and extraversion. Note that it is necessary to specify family=binomial for a binary response. What do you conclude?  
  
Cowles.mod0 <- glm(volunteer ~ . , data = Cowles, family = binomial)  
summary(Cowles.mod0)

##   
## Call:  
## glm(formula = volunteer ~ ., family = binomial, data = Cowles)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.3977 -1.0454 -0.9084 1.2601 1.6849   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.116496 0.249057 -4.483 7.36e-06 \*\*\*  
## neuroticism 0.006362 0.011357 0.560 0.5754   
## extraversion 0.066325 0.014260 4.651 3.30e-06 \*\*\*  
## sexmale -0.235161 0.111185 -2.115 0.0344 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1933.5 on 1420 degrees of freedom  
## Residual deviance: 1906.1 on 1417 degrees of freedom  
## AIC: 1914.1  
##   
## Number of Fisher Scoring iterations: 4

# 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.  
  
Anova(Cowles.mod0)

## Analysis of Deviance Table (Type II tests)  
##   
## Response: volunteer  
## LR Chisq Df Pr(>Chisq)   
## neuroticism 0.3139 1 0.57532   
## extraversion 22.1372 1 2.538e-06 \*\*\*  
## sex 4.4861 1 0.03417 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# From ANOVA result, there seemed to be association between features.  
  
# Logistic regression (Question 3) - Continue the analysis, but now fit a model, mod.cowles1, containing main effects and all two-way interactions of the predictors. Use summary() and Anova() as shown above. What do you conclude about the various two-way interactions?  
  
Cowles.mod1 <- glm(volunteer ~ (sex + neuroticism + extraversion)^2, data=Cowles, family=binomial)  
summary(Cowles.mod1)

##   
## Call:  
## glm(formula = volunteer ~ (sex + neuroticism + extraversion)^2,   
## family = binomial, data = Cowles)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.4981 -1.0545 -0.8958 1.2597 1.9670   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.498183 0.575546 -4.341 1.42e-05 \*\*\*  
## sexmale -0.021976 0.491866 -0.045 0.96436   
## neuroticism 0.115450 0.039370 2.932 0.00336 \*\*   
## extraversion 0.176503 0.042725 4.131 3.61e-05 \*\*\*  
## sexmale:neuroticism -0.003680 0.023020 -0.160 0.87300   
## sexmale:extraversion -0.014661 0.029381 -0.499 0.61778   
## neuroticism:extraversion -0.008809 0.002987 -2.949 0.00319 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1933.5 on 1420 degrees of freedom  
## Residual deviance: 1897.2 on 1414 degrees of freedom  
## AIC: 1911.2  
##   
## Number of Fisher Scoring iterations: 4

Anova(Cowles.mod1)

## Analysis of Deviance Table (Type II tests)  
##   
## Response: volunteer  
## LR Chisq Df Pr(>Chisq)   
## sex 4.9184 1 0.026572 \*   
## neuroticism 0.3150 1 0.574642   
## extraversion 22.1016 1 2.586e-06 \*\*\*  
## sex:neuroticism 0.0256 1 0.873001   
## sex:extraversion 0.2491 1 0.617718   
## neuroticism:extraversion 8.8117 1 0.002993 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# It seemed that sex is still one important feature. The combination of neuroticisim and extraversion was significant as well.  
  
Cowles.mod <- glm(volunteer ~ sex + neuroticism \* extraversion, data = Cowles, family = binomial)  
summary(Cowles.mod)

##   
## Call:  
## glm(formula = volunteer ~ sex + neuroticism \* extraversion, family = binomial,   
## data = Cowles)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.4749 -1.0602 -0.8934 1.2609 1.9978   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.358207 0.501320 -4.704 2.55e-06 \*\*\*  
## sexmale -0.247152 0.111631 -2.214 0.02683 \*   
## neuroticism 0.110777 0.037648 2.942 0.00326 \*\*   
## extraversion 0.166816 0.037719 4.423 9.75e-06 \*\*\*  
## neuroticism:extraversion -0.008552 0.002934 -2.915 0.00355 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1933.5 on 1420 degrees of freedom  
## Residual deviance: 1897.4 on 1416 degrees of freedom  
## AIC: 1907.4  
##   
## Number of Fisher Scoring iterations: 4

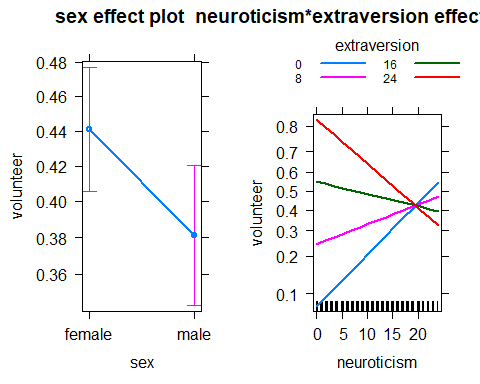
# From the above result, we verified that all factors, including the interaction factor, are statistically significant.  
  
anova(Cowles.mod0, Cowles.mod1, Cowles.mod)

## Analysis of Deviance Table  
##   
## Model 1: volunteer ~ neuroticism + extraversion + sex  
## Model 2: volunteer ~ (sex + neuroticism + extraversion)^2  
## Model 3: volunteer ~ sex + neuroticism \* extraversion  
## Resid. Df Resid. Dev Df Deviance  
## 1 1417 1906.1   
## 2 1414 1897.2 3 8.8792  
## 3 1416 1897.4 -2 -0.2579

# It seemed that it is benifical 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.  
  
library(effects)

## lattice theme set by effectsTheme()  
## See ?effectsTheme for details.

Cowles.effect <- allEffects(Cowles.mod, xlevels = list(extraversion = seq(0, 24, 8)))  
plot(Cowles.effect, multiline=TRUE)



## 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.