DASI\_Project\_proposal

myself\_vinsent

Saturday, September 20, 2014

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

### Introduction:

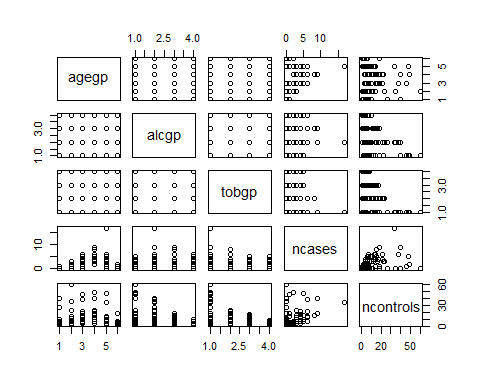
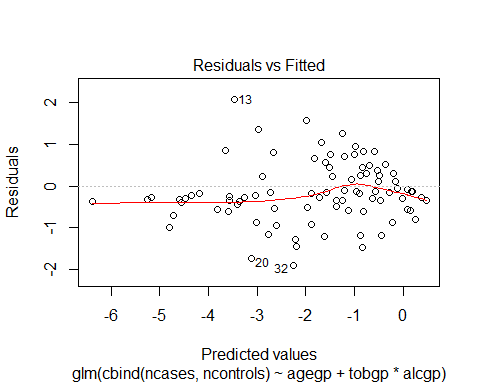
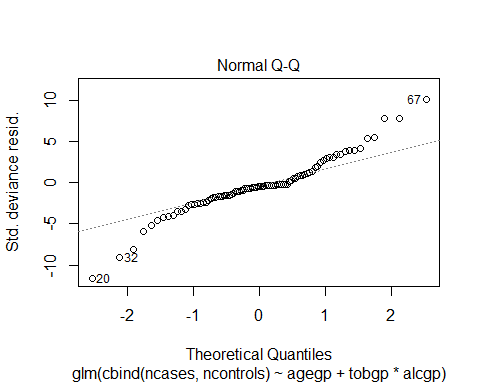
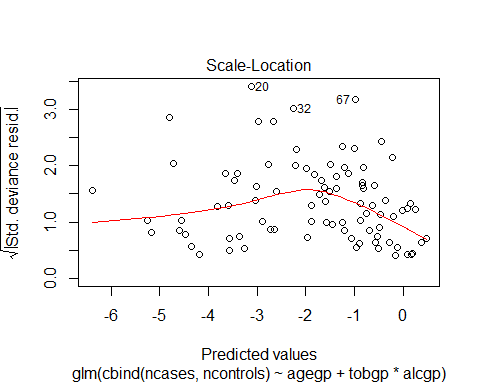
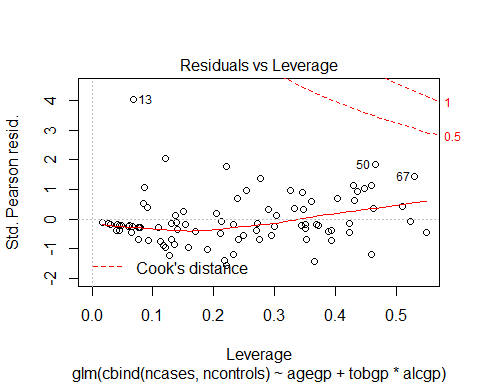
This study on dataset esoph data set to find the relation between age group, type of cancer causing agent(Alcohol and Tobacco). from the R- dataset. this dataset contains 88 observation of control and ncases.

### Data:

summary(esoph)

## agegp alcgp tobgp ncases ncontrols   
## 25-34:15 0-39g/day:23 0-9g/day:24 Min. : 0.00 Min. : 1.0   
## 35-44:15 40-79 :23 10-19 :24 1st Qu.: 0.00 1st Qu.: 3.0   
## 45-54:16 80-119 :21 20-29 :20 Median : 1.00 Median : 6.0   
## 55-64:16 120+ :21 30+ :20 Mean : 2.27 Mean :11.1   
## 65-74:15 3rd Qu.: 4.00 3rd Qu.:14.0   
## 75+ :11 Max. :17.00 Max. :60.0

### Exploratory data analysis:

## Analysis of Deviance Table  
##   
## Model 1: cbind(ncases, ncontrols) ~ agegp + tobgp \* alcgp  
## Model 2: cbind(ncases, ncontrols) ~ agegp + unclass(tobgp) + unclass(alcgp)  
## Resid. Df Resid. Dev Df Deviance  
## 1 67 47.5   
## 2 80 59.3 -13 -11.8

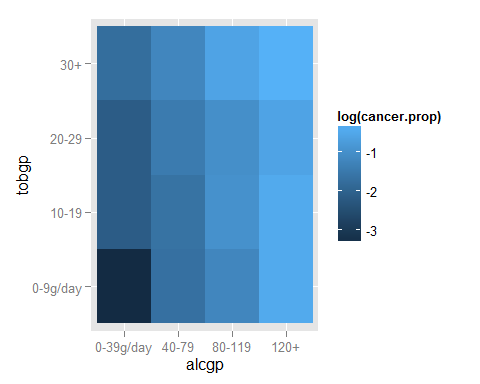
### Inference:

#### Calculate the cancer proportion by alcohol and tobacco usage

## Warning: package 'reshape2' was built under R version 3.1.1

## alcgp tobgp cancer.prop  
## 1 0-39g/day 0-9g/day 0.03448  
## 2 0-39g/day 10-19 0.11905  
## 3 0-39g/day 20-29 0.11905  
## 4 0-39g/day 30+ 0.17857  
## 5 40-79 0-9g/day 0.18994  
## 6 40-79 10-19 0.20000

## Warning: package 'ggplot2' was built under R version 3.1.1



#### The cancer caused by both tobaco and Alcohol are plotted in a log ratio,

#### the inverse log , the color are the fill is consider as cancer.prop <- sum(dfncontrols)( this states a ratio between affected persons to healthy living)

### Conclusion:

#### A high number of case in age group 30+ can be seen in log plot. for both Alcohol and Tobaco consumption.

### Reference

##### <http://davetsao.com/blog/2013-07-28-useful-r-packages.html>