BSAN 450 Assignment 14

1) In 1846 the Donner and Reed families left Springfield Illinois for California by covered wagon. In July the Donner Party, as it became known, reached Fort Bridger, Wyoming. There its leaders decided to attempt a new and untested route to the Sacramento Valley. Having reached its full size of 87 people and 20 wagons, the party was delayed by a difficult crossing of the Wasatch Range and again in the crossing of the desert west of the Great Salt Lake. The group became stranded in the eastern Sierra Nevada mountains when the region was hit by heavy snows in late October. By the time the last survivor was rescued on April 21, 1847, 40 of the 87 members had died from famine and exposure to extreme cold. This data is reported in Ramsey and Shafer in the Statistical Sleuth. The original data consisted of three variables: Surv = 1 if the party member survived and 0 if the party member died; Gender = 1 if female and 0 if male; and Age = the age of the party member. Abraham and Ledolter constructed a fourth variable MultFam = 1 if other party members with the same family name were present and 0 otherwise. The data used in this analysis consist of the Donner party members that were 15 years or older.

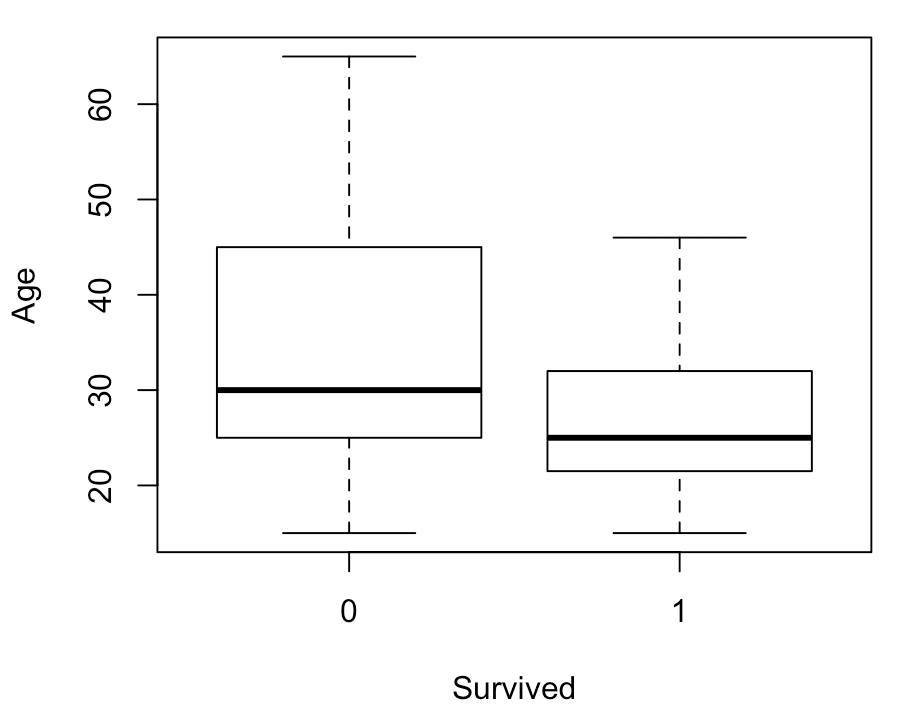
Read the data into R Studio using the following command.

Donner = read.csv("Donner.csv")

a) The set of predictor variables are of two types, a continuous variable: Age, and categorical variables: Gender and MultFam. It is a good idea to do some preliminary analysis to get an idea about whether or not the predictor variables seem to be related to the response variable.

For a continuous predictor a useful preliminary plot is a boxplot with groups formed by the response variable.

boxplot(Donner$Age~Donner$Surv,ylab="Age",xlab ="Survived")



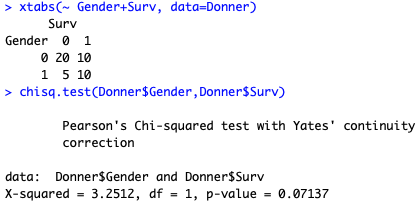
Based on the boxplot, does it appear that the variable age will be a good dependent variable in a logistic regression model with Surv as the independent variable?

**Age is an ok dependent variable. The mean of suv based on the age is somewhat different. However, the variance for surv = 0 is quite large than the variance for if surv = 1.**

b) For categorical predictors, a preliminary analysis is construction of a two way table with the predictor variable in the rows and the response variable in the columns. In this example, if the proportion who survived changes with the different values of the predictor variable, then that predictor variable is a good input variable in a logistic regression model.

A formal way to evaluate whether or not a predictor variable might be helpful is to test the following hypotheses - Null Hypothesis: the two variables are independent of each other versus Alternative Hypothesis: the two variables are not independent of each other.

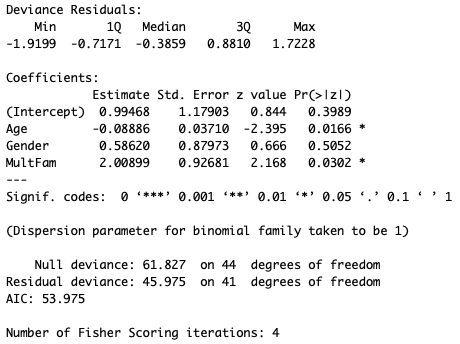
The following commands produce a cross tabulation and perform the a chisquare test of the null hypothesis of independence for the variables Gender and Surv. Produce the cross tabulation and perform the hypothesis test for the set of variables Gender and Surv and for the set of variables MultIFam and Surv.



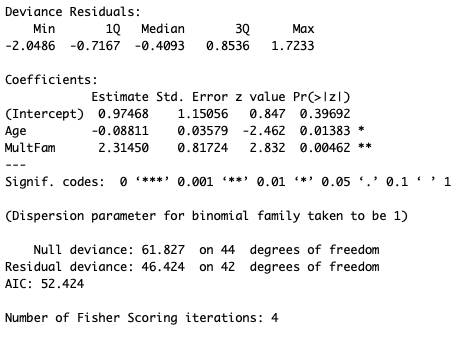
Based on this what can be said about using the variables Gender and Multifam as predictors in a logistic regression model with independent variable Surv?

**Since the p-value is > 0.05 the null hypothesis can be rejected, meaning that at least one of the two variables are helping explain the surv.**

c) Fit a logistic regression model with the variables Age, Gender, and MultFam as the input variables and Surv as the response. Interpret the output. Are both of these variables statistically significant? Make any changes to this model you believe are appropriate.

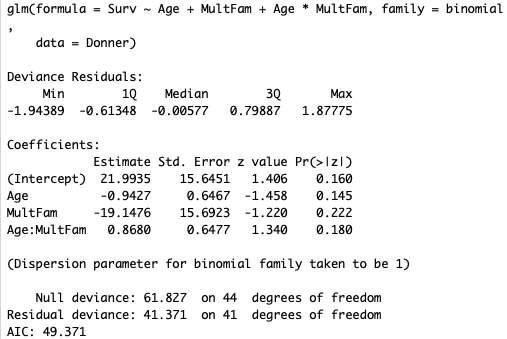


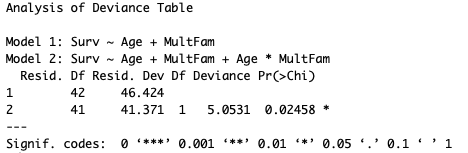
**Remove Gender**

****

**Both MultFam and Age are significant.**

d) As a diagnostic check to the model you fit in part c, fit a more complicated model and evaluate whether or not this more complicated model is necessary for this data.





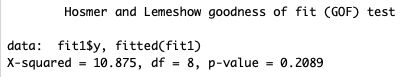
**This indicates that the addition of the variable MultFam\*Age is impactful to the model.**

e) Another way to check the model in part c is to perform the Hosmer-Lemeshow goodness of fit test. The Null Hypothesis for this test is that the model is correct and the Alternative Hypothesis is that the model is not correct. This test must be performed when there are not any replications in the predictor variables. The following R commands perform this test. Execute these commands and comment on the results. (Note it you have not installed the ResourceSelection library into R Studio you will need to do that before invoking the library command.)

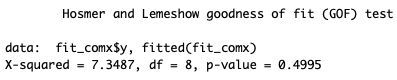
library(ResourceSelection)

hoslem.test(fit$y, fitted(fit), g=10)

**Goodness of fit for noncomplex model**



**Goodness of fit for complex model**

****

**Since the p-value on the goodness of fit is relatively high. This suggest no issues with the model.**

2) This data is taken from Shmueli et at (2010). The data set contains information of 5000 loan applications. The response is whether or not an offered loan had been accepted on an earlier occasion. The explanatory variables are:

Age = age of the customer

Exp = professional experience in years

Inc = income of the customer

Fam = family size of the customer

CCAve = average monthly credit card spending

Educ = three categories of education level: 1 = undergraduate, 2 = graduate, 3 = professional

Mort = size of mortgage

SecAcc = 1 if the customer has a securities account and otherwise = 0

CD = 1 if the customer has a CD account and otherwise = 0

Online = 1 if the customer has an online account and otherwise = 0

CreditCard = 1 if the customer has a credit card and otherwise = 0

The name of the response variable is Response.

We will use a training set of 4000 of the 5000 cases to develop a model and save the remaining 1000 as a test set. The following commands read the data into R Studio and create a training set which is stored in the dataframe loan.train. Execute these commands.

loan=read.csv("LoanAccept.csv")

set.seed(1)

train=sample(5000,4000)

test=(c(1:5000)[-train])

loan.train=loan[train,]

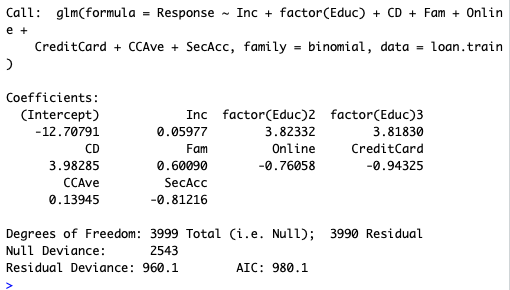
a) Since there are 11 potential independent variables we will use the stepwise procedure to identify a subset of these variables that can be used in a logistic regression model. Execute the following commands. Note that the final model is indicated as the last part of the output.

null=glm(Response~1,family=binomial,data=loan.train)

full=glm(Response~.-Educ+factor(Educ), family=binomial,data=loan.train)

step(null,scope=list(lower=null,upper=full),direction="both")

Which of the variables does this indicate should be included in the model?

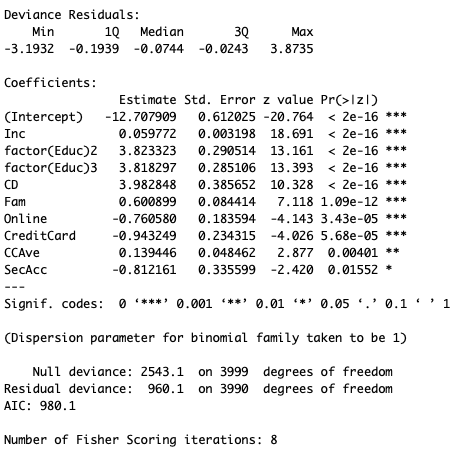


**The stepwise procedure indicates that Inc, Educ, CD, Fam, Online, CreditCard, CCAve, SecAcc.**

b) Using the training data, estimate the model that is identified in part a and print out the summary. Note that to only use the training data, you will need to include “subset = train” in the R command. An example is given below, you will need to modify this command.

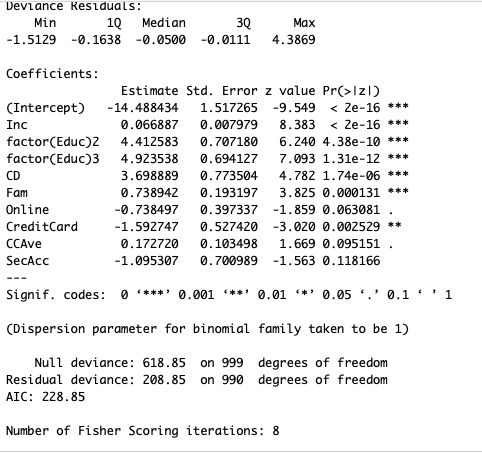
fit=glm(formula = Response ~ Inc + factor(Educ), family = binomial, data = loan, subset=train)

Are all the variables statistically significant? Make any changes that you believe are needed.



**Since all the p-values are smaller than 0.05, yes, all the variables are significant.**

c) Use the test data to estimate the model you determined was best in part b). Are all the variables statistically significant?



**Using the testing subset SecAcc, CCAve, and Online are now not significant.**