Final Project

STAT 701: Statistical Programming 701

December 15, 2015

By

Dheeraj Gadhiraju

Course Instructor: Dr. Michael Semhar

1: (Ex 14.3 R Handbook)-The data shown in Table 14.3 were collected in a follow-up study of women patients with schizophrenia (Davis, 2002). The binary response recorded at 0, 2, 6, 8 and 10 months after hospitalization was thought disorder (absent or present). The single covariate is the factor indicating whether a patient had suffered early or late onset of her condition (age of onset less than 20 years or age of onset 20 years or above). The question of interest is whether the course of the illness differs between patients with early and late onset? Investigate this question using the GEE approach.

1. Provide a two-page write-up (including graphs) explaining your analysis of the experiment and the conclusions you can draw from it.

**Output results:**

|  |
| --- |
| > summary(s2)  subject onset disorder month  5 : 5 < 20 yrs:55 absent : 0 Min. : 0.000  33 : 5 > 20 yrs:17 present:72 1st Qu.: 0.000  11 : 4 Median : 2.000  16 : 3 Mean : 2.722  18 : 3 3rd Qu.: 6.000  19 : 3 Max. :10.000  (Other):49 |
| > table(schizophrenia2$onset,schizophrenia2$disorder)    absent present  < 20 yrs 91 55  > 20 yrs 41 17  > summary(sc2\_gee3)  GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA  gee S-function, version 4.13 modified 98/01/27 (1998)  Model:  Link: Logit  Variance to Mean Relation: Binomial  Correlation Structure: Unstructured  Call:  gee(formula = disorder ~ month + onset, id = subject, data = sc2,  family = "binomial", corstr = "unstructured", scale.fix = TRUE,  scale.value = 1)  Summary of Residuals:  Min 1Q Median 3Q Max  -0.6814924 -0.2848327 -0.1149192 0.4319059 0.9260822  Coefficients:  Estimate Naive S.E. Naive z Robust S.E. Robust z  (Intercept) 0.7606389 0.31858221 2.3875750 0.31698566 2.399600  month -0.2802089 0.05462343 -5.1298298 0.05569149 -5.031449  onset> 20 yrs -0.4865597 0.49092175 -0.9911146 0.48160771 -1.010282  Estimated Scale Parameter: 1  Number of Iterations: 3  Working Correlation  [,1] [,2] [,3] [,4] [,5]  [1,] 1.00000000 0.47101780 0.09473767 0.08322270 0.08179451  [2,] 0.47101780 1.00000000 0.19946212 0.09306456 0.10156075  [3,] 0.09473767 0.19946212 1.00000000 0.19185335 0.51138945  [4,] 0.08322270 0.09306456 0.19185335 1.00000000 0.45678055  [5,] 0.08179451 0.10156075 0.51138945 0.45678055 1.00000000 |

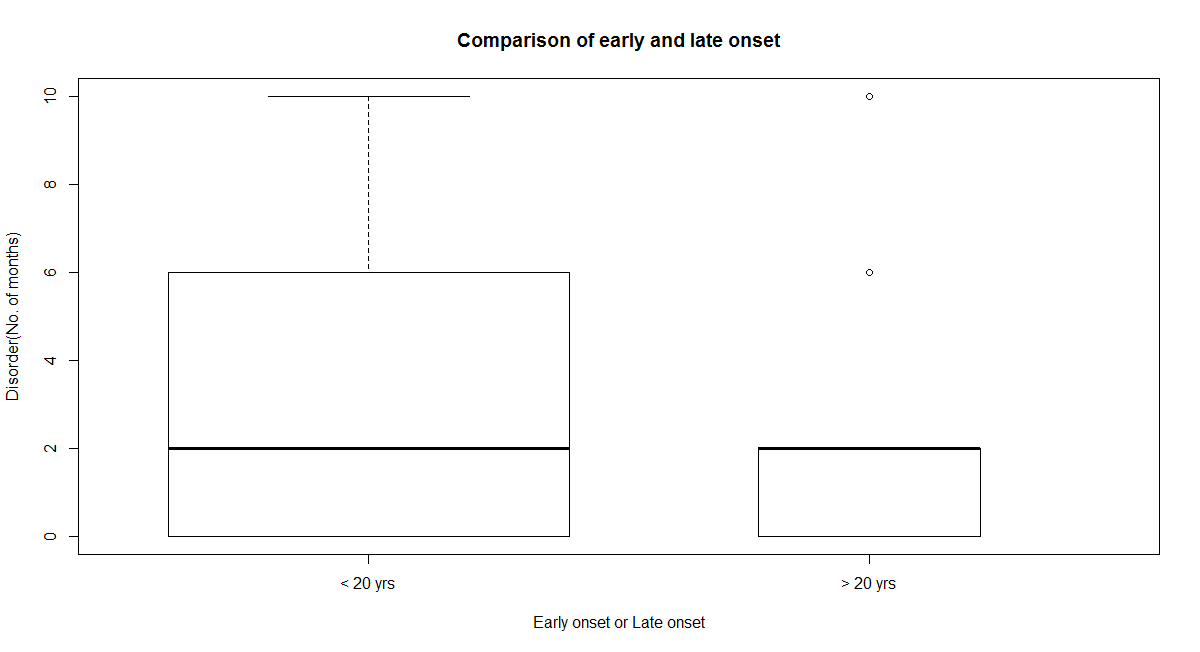


Fig. 1.1 Comparison of early and late onset for patients with disorder

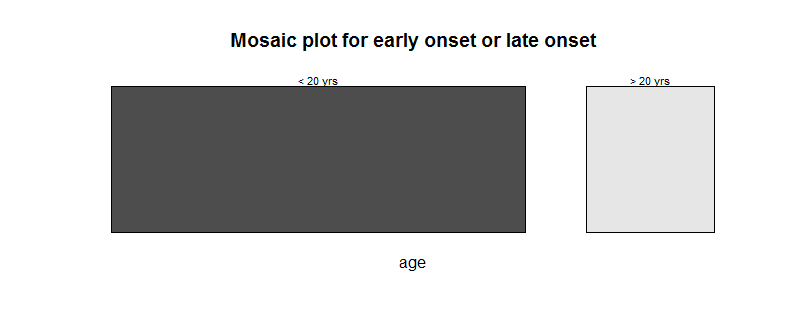


Fig. 1.2 Comparison of early and late onset for patients with < 20 yrs and > 20yrs.

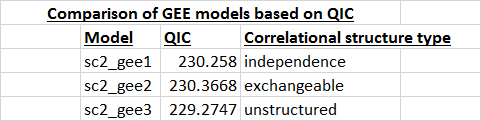


TABLE 2 – Comparison of GEE models based on QIC

**Discussion:**

The question of interest here is to find out if the course of illness i.e for patients with disorder = “present”. Clearly we have to find out if the illness differs between patients for those whose disorder = “present”. By looking at the Fig. 1.1 we can clearly tell that the illness differs between early and late onset. The course of illness by months has a wide range from months 0 to 10 with a median of 2 for early onset. Whereas, for late onset there are two outliers. By looking at the summary of the “s2” data there are about 55 patients with early onset and 17 patients with late onset. Fig. 2 clearly supports the statement that there are more number of people suffering from the disorder in early onset and the course of illness differs from early and late onset.

From my analysis, from the Mosaic plots and from the Gee approaches with different correlational structures, the course of illness does differ between patients with early and late onset, but the variable onset is insignificant. As seen in the table 2 GEE is used to analyze the longitudinal count data. GEE was performed on correlational structures of independence, exchangeable and unstructured. And based on QIC, sc2\_gee3 model with correlational structure “unstructured” is better with 229.2747.

1. **As a secondary component provide annotated code that replicates your analysis.**

**Discussion:**

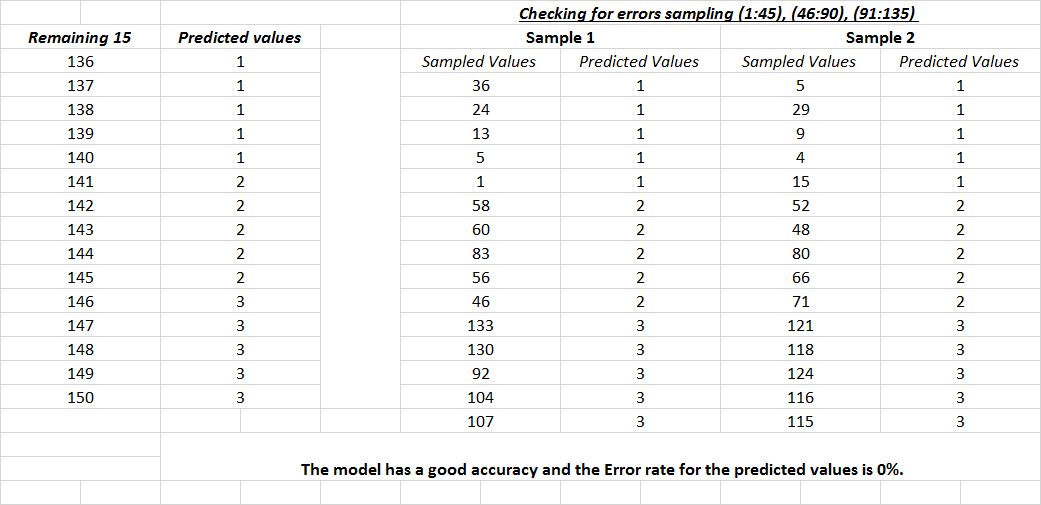
Please find the rcode in the uploaded r file.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

2: The file ”y.dat” contains a dataset consisting of 150 4-variate observations. True group labels are provided for the first 135 data points in the file ”idy.dat”. Classify the remaining 15 observations. There are three groups in this problem.

1. Provide a page write-up (including graphs) explaining what methods you used to model the groups and how you predicted the identity of the remaining 15 observations.

**Graphs and Analysis:**



Pic 2.1 The predicted values for the remaining 15 observations and error rate with two samples.

**Discussion:**

The file y.dat contains 150 observations with 4 variables and the file idy.dat has 135 observations with group labels as “1”, “2”, “3”. Idy.dat file has 15 remaining values that are to be classified with group labels based on the first 135 observations. The remaining 15 observations in idy.dat data frame were made 0 and the data was merged along with y.dat into one data frame “ydata”. The data frame ydata contains 150 observations with 5 variables and the group labels are named as id. Recursive partitioning can be used to predict the remaining 15 observations considering the first 135 observations for the model (id~V1+V2+V3+V4,data=dat1). The function rpart() from the package rpart was used on the model for the first 135 observations of the combined data frame. After using rpart, the predict function was used to predict the values and the results from 136 to 140 observations have 1.000 and the results from 141 to 145 observations have 2.02439(Close to 2) and the results from 146 to 150 observations have 2.97561 (Close to 3). We can conclude that the results for 136 to 150 observations were show cased as 1,1,1,1,1,2,2,2,2,2,3,3,3,3,3 as observed in the Fig. 2.1.

Now that I have predicted the remaining 15 values I have to check the error rate. Keeping the data set in mind, “1:45” observations have “1” as group label and “46:90” as “2” and similarly “91:135” has “3” as a group label. I have divided the data into 3 sets according to the group labels as “s1”, “s2”, “s3” and using the sample() function in r I have randomly selected 5 observations from each set. Once I had 15 observations, 5 from each set I have combined all the 15 observations into one data frame and repeated the rpart() and predict() on the data frame. The predicted values are exactly same and I had concluded that the *error rate* is 0 %. The analysis is provided in the Fig. 2.1 with two random sample values. Therefore making the model a better choice.

1. **As a secondary component provide annotated code that replicates your analysis.**

**Discussion:**

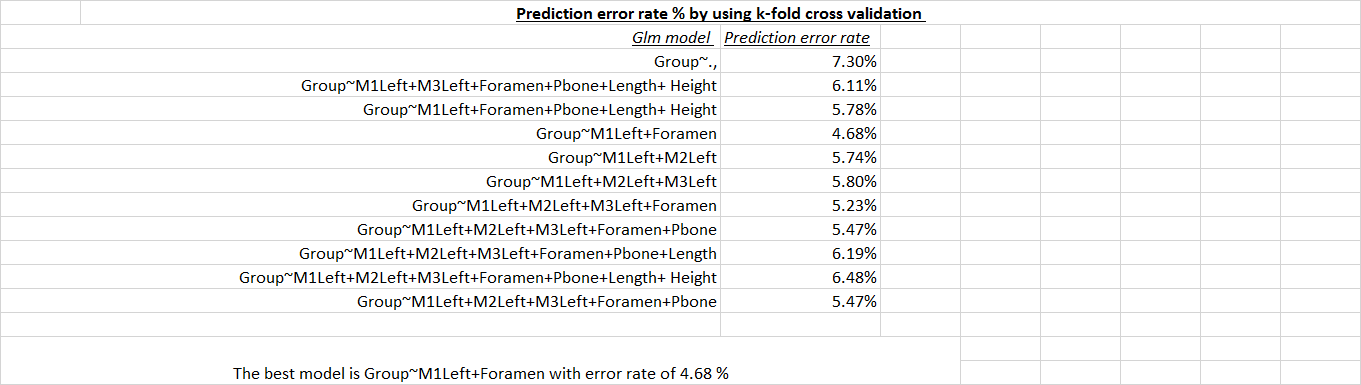
Please find the rcode in the uploaded r file.

**3. Develop a GLM model from the 89 specimens that you can use to predict the group membership of the remaining 199 specimens’.**

**i. Explain your GLM and assess the quality of the fit with the classified  
observations.  
• Use Cross Validation to predict the accuracy of your model.**

1. **Explain your GLM and assess the quality of the fit with the classified  
   observations.  
   • Use Cross Validation to predict the accuracy of your model.**

**Graphs and Analysis:**



**Fig 3.1 Prediction error rate % by using k-fold cross validation for different glm models.**

> summary(m)

Call:

glm(formula = Group ~ M1Left + Foramen, family = "binomial",

data = b)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.49687 -0.01788 0.01058 0.09923 1.28036

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 62.804452 20.661080 3.040 0.00237 \*\*

M1Left -0.047246 0.014091 -3.353 0.00080 \*\*\*

Foramen 0.006637 0.003192 2.079 0.03758 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 123.279 on 88 degrees of freedom

Residual deviance: 22.049 on 86 degrees of freedom

AIC: 28.049

Number of Fisher Scoring iterations: 8

**Conclusion:**

As seen in the Fig 3.1 the modelling was done on different models using GLM. GLM was performed on different models and as in the Fig 3.1. The 89 classified observations are taken into an object and the model Group~M1Left+Foramen is the best model. Both the variables are significant according to the summary of the model.

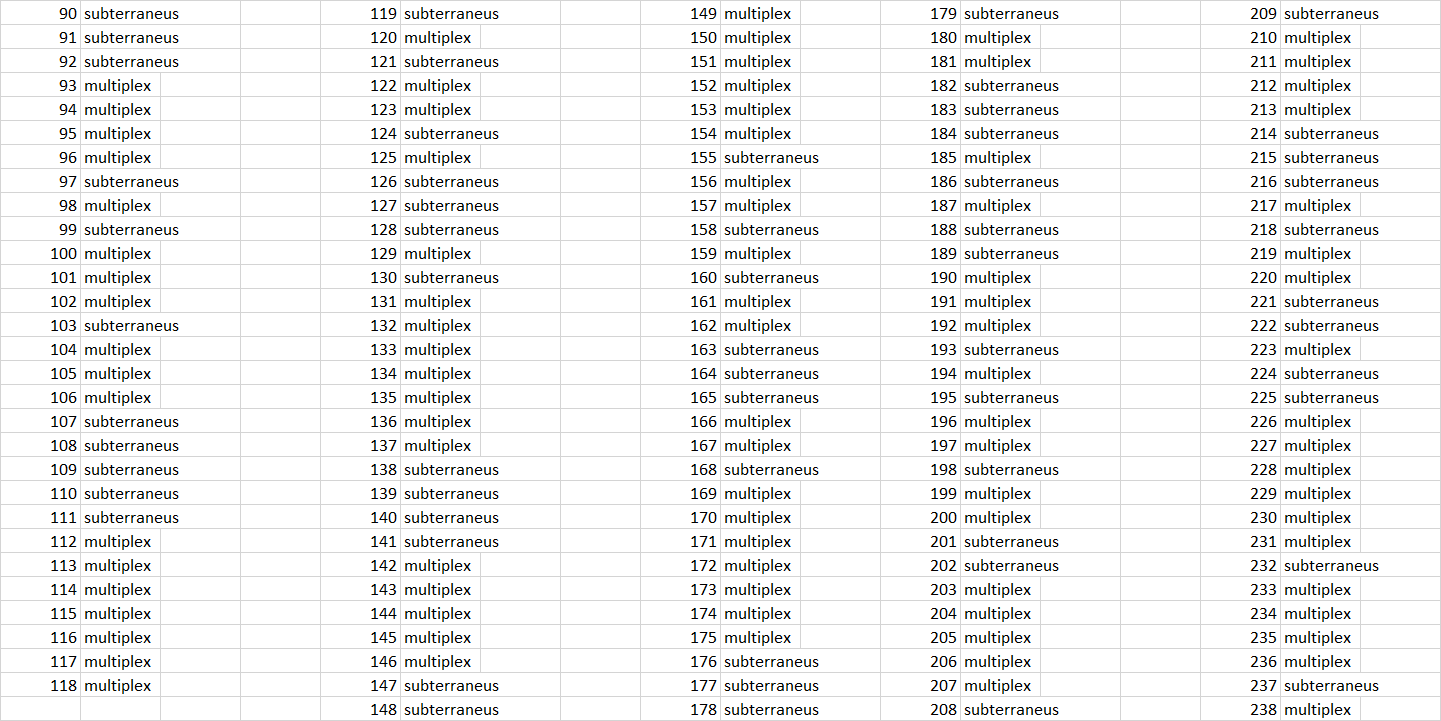
A k-fold cross validation was used to predict accuracies of various models. Cv.glm() was used on various models and after carefully analyzing the error rate Group~M1Left+Foramen has the least error rate with 4.68% and is the best model of my choice.

1. ***Provide a one-page write-up (including graphs) explaining your analysis of the*  
   *dataset and your recommendations on the usefulness of your predictions.***

**Discussion:**

The output results and my analysis along with the k-fold cross validation results can be found in Fig. 3.1 The dataset microtus was imported from the package Flury. The data set has Microtus multiplex and Microtus subterraneus with 43 and 46 species, a total of 89 observations and a further observations of 199. The first 89 classified observations are taken into one data frame and the remaining unclassified observations are taken into another data frame. From Fig 3.1 using a cross validation technique the model Group~M1Left+Foramen is considered as a best model with just 4.68% *error rate* and the same model is being used to predict the recommendations of the unclassified observations of the remaining 199 observations. The rpart() function from the package “rpart” is used on the classified observations. The results from the rpart are stored into a data frame. Similar to the Question 2, I have used predict function to predict the groups of the remaining 199 observations. After executing the predict function the results have predictions and I have exported the predictions into a file. I have included the predicted results in 3.3.

1. **Provide predictions for the unclassified observations.**



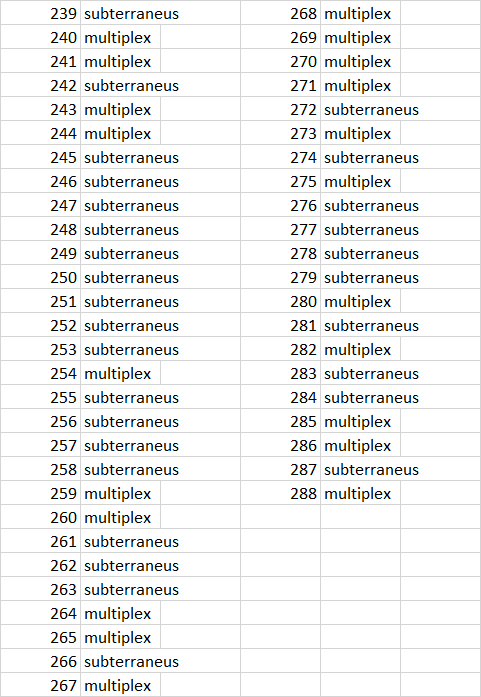


Fig 3.2 Predictions of the group memberships of the unknown 199 observations.

**Discussion:**

All the predictions for the remaining 199 unclassified observations are provided in the Fig 3.2

**iv. As a secondary component provide annotated code that replicates your analysis.**

**Discussion:**

Please find the rcode in the uploaded r file.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*