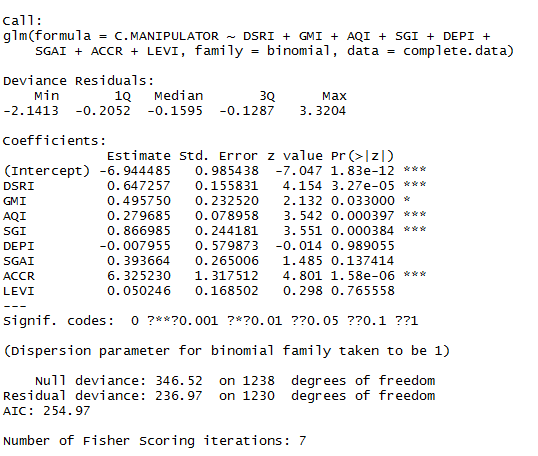
1. Beneish Model

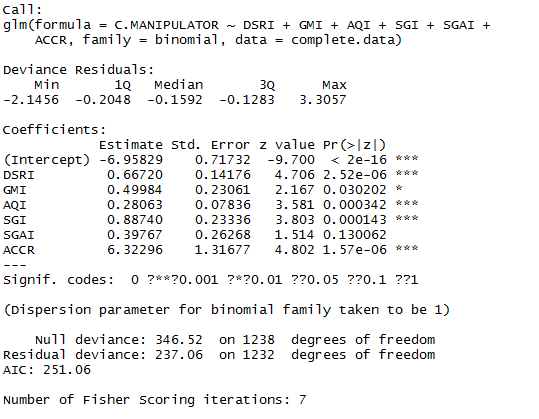
According to the Beneish Model in 1999, an indicator called M-score based on the 8 variables are used to find out the possibility of a business in public sector to be an earning manipulator. If the M-score is larger than -2.2, it implies that the company is more likely to be an earning manipulator.

From the formula below, we can first run this linear regression model to find out any statistically significance.

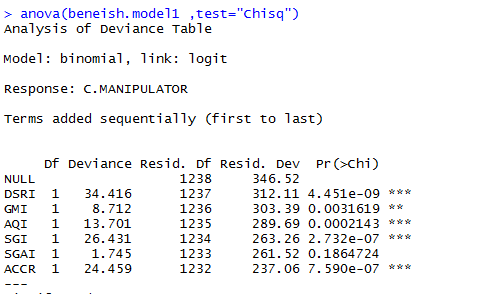
**M =β0+ β1 DSRI + β2 GMI+ β3 AQI+ β4 SGI+ β5 DEPI+ β6 SGAI+ β7 ACCR+ β8 LEVI + ε, where ε is the residual**

Since this model is probabilistic, and a classification method with binary outcome, a logistic regression model is run on the data.



Firstly, we found that the variable **“DEPI”** has a largest p-value, which cannot reject the null hypothesis, therefore, it is not statistically significant. Similarly, we found the variable**” LEVI”** is also not significant in the regression model. Therefore, we reduced the model without these two variables and run the logistic regression model again. By doing these for several times, which is the stepwise by AIC algorithm, we have a model coming up on the left side. With a lower AIC value at 251.06, we have obtained all p-values are now lower than 0.1. Therefore, we run the ANOVA test on this model compared to the original one.

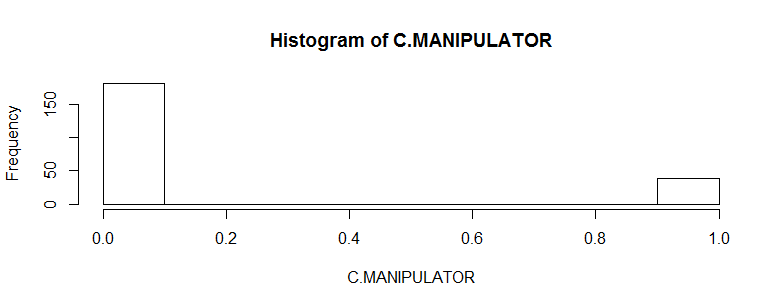
In this case, according to the second model, we found that at p<0.05, only variables **“AQI”,”SGI”,”ACCR”,”GMI”,”DSRI”** are statistically significant, while **“SGAI”** are not that significant in detecting the manipulation. In addition, variables **“DEPI”,”LEVI**” cannot be used to detect earning manipulation.



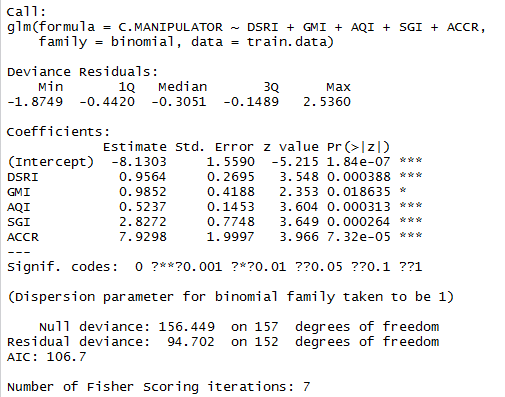


In conclusion, in the case of India, Beneish Model is only useful when the company has better prospect. AQI- Asset Quality Index, SGI-Sales and General Administration Index , ACCR-Total Accrual, Gross Profit Index, DSRI-Sales Index are the useful detector for Indian cases. The larger these indicators are, the more possible that the company might have earning manipulation activities.

**M =β0+ β1 DSRI + β2 GMI+ β3 AQI+ β4 SGI+ β5ACCR**

1. As we have an unbalanced data structure in this case, over 96% of the dataset are non-manipulator, therefore, the size would be large enough for us to build up the model to predict those companies that are non-manipulator. As a result, since the model is good at predicting the non-manipulator, the accuracy rate of the model from the confusion matrix would be large, even the “True positive “amount is lower, since the data for manipulation are not big enough. In this case, the model with high accuracy rate is not performing good on detection of the earning manipulation.
2. By running a logistic regression in the sample data, since we have concluded that only “DSRI”,” GMI”,”AQI”,”SGI”,”ACCR” are useful for the detection, we will exclude the other insignificant variables for the prediction. 

In this sample data, we found that there are 181 non-manipulators and 39 manipulators. Event rate in this dataset is 39/181=21%. Since the event rate is reasonable, the logistic regression model on the sample data from the following:



1. Alternative 1:

To measure the model accuracy, by separating the sample.data into 70% : 30% into training and testing data, we have created a confusion matrix. According to what we have constructed here, the accuracy of the model is high, which is almost 92%.

pred.1

0 1

0 53 1

1 4 4

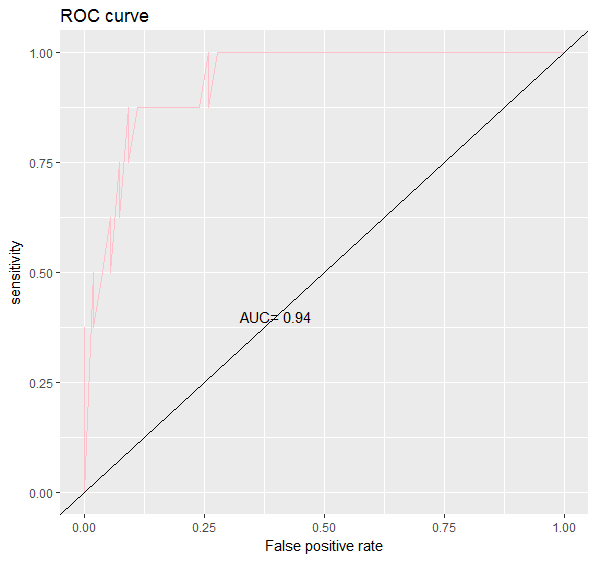
Accuracy=53+4/53+4+4+1=92%

Error rate=1-92%=8%

sensitivity=4/8=50%

However, we found that although the model has a high accuracy, it is only good at predicting the non-manipulator, while for those earning manipulators, the model is bad at detection.

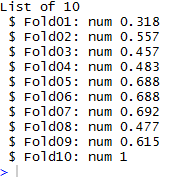
Alternative 2: use the rest of the data for testing



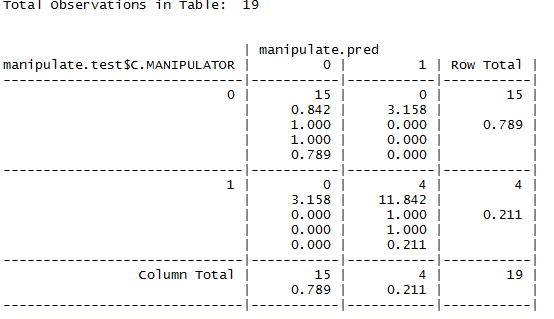
5.

Since the data structure has too many instances of non-manipulators compared to the manipulators, by keeping all the manipulators and maintaining the event rate similar to the one created in Q3(20%), we did random sampling 80% of non-manipulators to build up a new dataset of total instance of 195(39/20%). Call this new.data. Therefore, in the new data**, there are 39 manipulators and 156 non-manipulators.**

Also, we need to use cross-validation technique for better result in this new data. By using Kappa test, we have created 10 folds for the result, and as a result, by taking the mean of these kappa statistic, we have a result of **0.5975217, which is moderate agreement after cross-validation.**

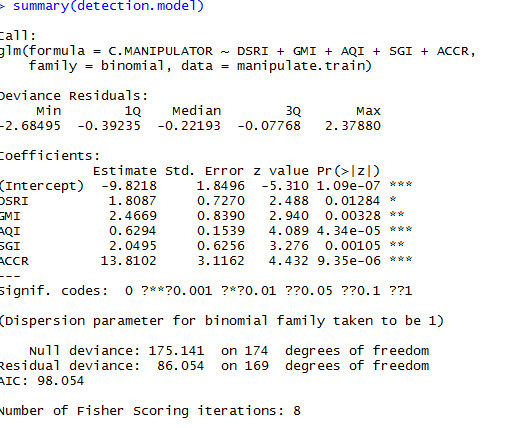


**In these cases, the confusion matrix for the 10th fold as an extreme case, accuracy= 78.9% while the sensitivity=100%, which is much better the previous one created in Q3. In general, all of these cross-validation has increase the sensitivity compared to the previous model.**

****

**#adaBoost**

**Q6. In addition, by looking on the 9th fold with kappa stat of 0.615,**

****

**The AIC result is much better now, and the final M-score could be set as the coefficient of this model.**

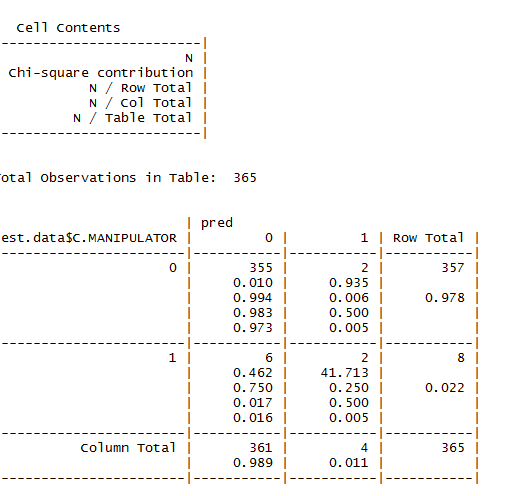
7.

From the sample.data, we found the importance of each variable



Surprisingly, the importance of each variables are different from that we have concluded from the logistic regression. In the order of its importance, we have SGI as the most important predictor while GMI is the least important predictor. Also, “LEVI” are found useless for the M-score logistic regression, while the decision tree gives importance to it in the decision tree.

8.



Prediction accuracy=355+2/355+2+2+6=97.8%

0 1 error rate=1-97.8%=2.19%

0 355 2 sensitivity=2/8=25%

1 6 2

From the complete dataset, the logistic model we have built so far has an accuracy rate of 97.8%, which is much higher than the one built from the sample dataset in Q3. However, the precision of this model for detecting manipulation is much poorer, at 25%. In this case, the reason for this poor performance of this model is due to the unbalanced data structure of the complete dataset. In the dataset, **the event rate of a manipulator is 39/1239=3.14%,with 3.57% for its training set**, while for those training data in sample dataset, the event rate **is 31/154=20%** which is relatively fair.

To visualize the performance because of unbalanced data structure, we use ROC curve

