I have a dataset containing family information of married couples, which have around 10 variables & 600+ observations.

Independent variables are ~ gender, age, years married, children, religion etc.

I have one response variable which is number of extra marital affairs.

Now, I want to know what all factor influence the chances of extra marital affair.

Since extra marital affair is a binary variable (either a person will have or not),

so we can fit logistic regression model here to predict the probability of extra marital affair.

install.packages('AER')

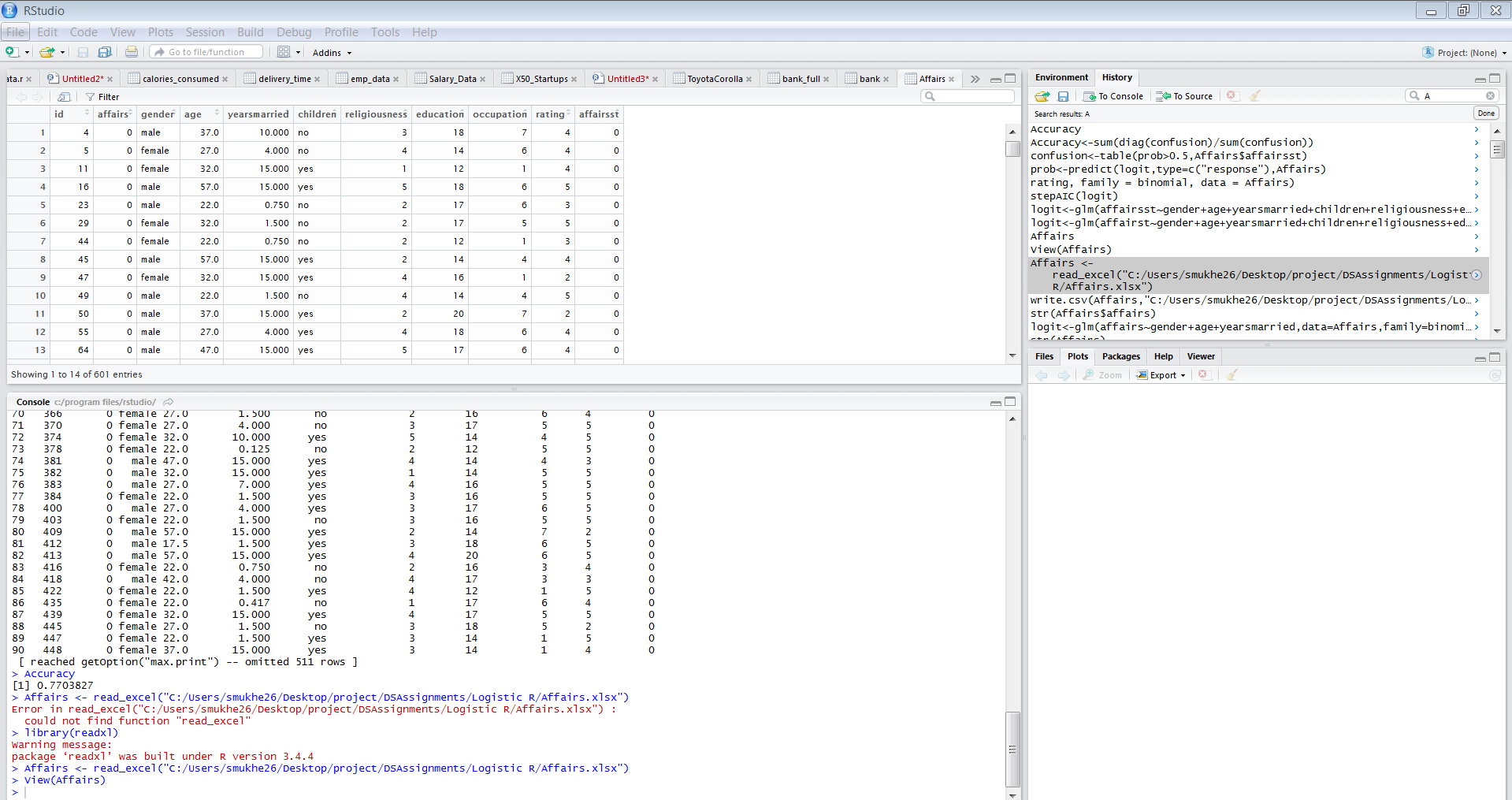
data(Affairs,package="AER")

Solution:

We convert the number of extra-marital affairs to 0 and 1 only

If (no of affairs>0 then 1 else 0)

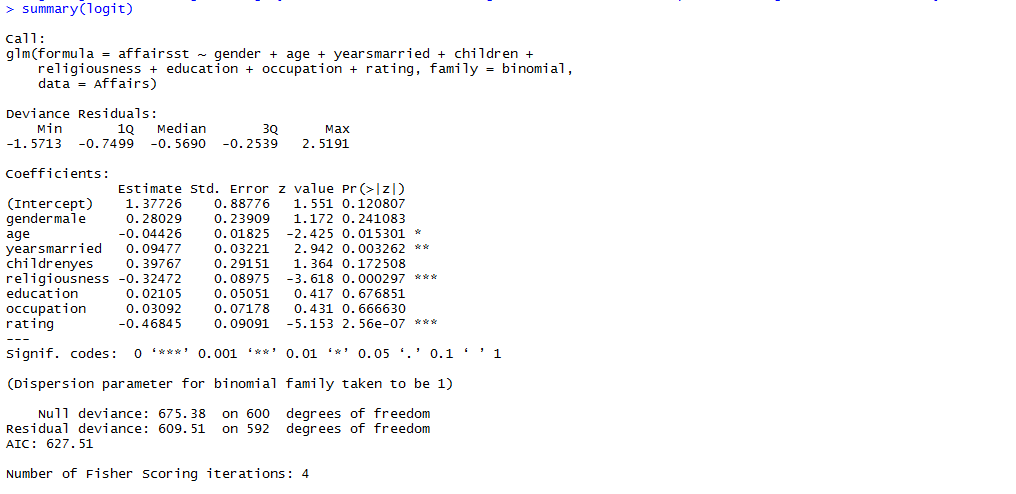
Now we read the file into R studio



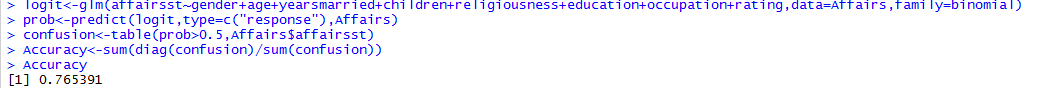
We apply the logististic regression glm on all the columns of the dataset



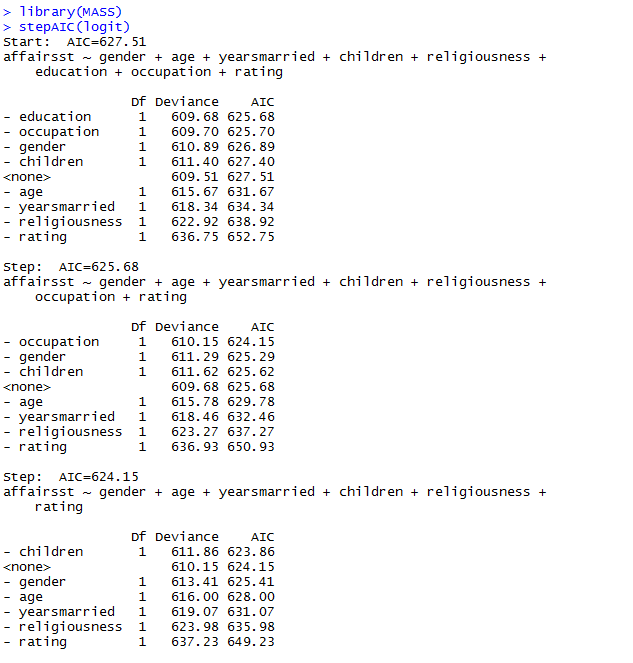
And do a summary statistics on the model

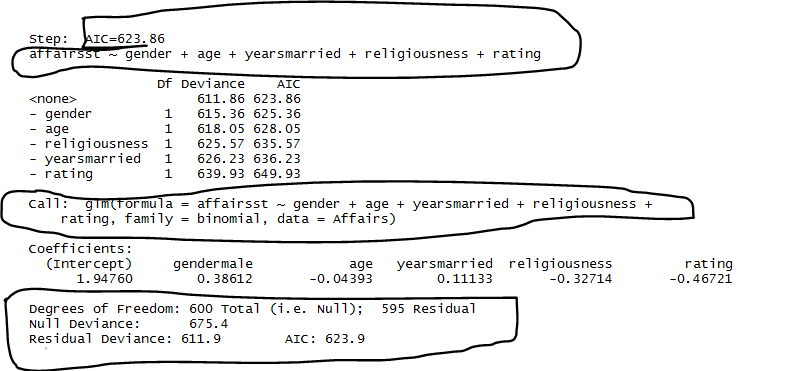


We should look into the accuracy to get



Let’s try to improve the accuracy. We can see that difference between null and residual deviance is not that noteworthy hence we can assume that model is not that much effective. Also we can see a lot many columns don’t contribute to the prediction model hence would like apply stepAIC to know the contributing variables

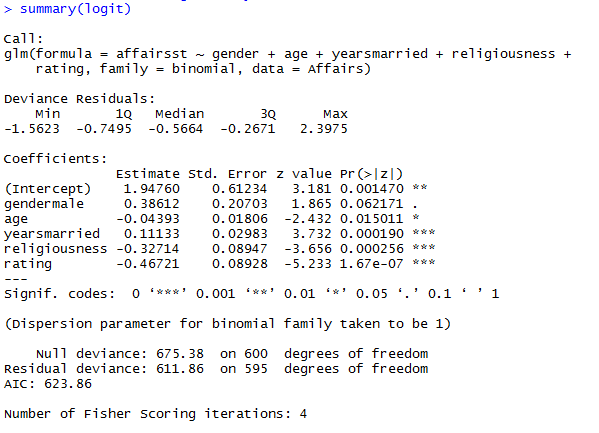




As a thumb rule, we need to select the model with the least AIC score , here we can see that AIC is 623.86 which is quite less than other combinations, hence we need to create the model again with the suggested input output combination

> logit<-glm(affairsst ~ gender + age + yearsmarried + religiousness + rating, family = binomial, data = Affairs)

We do a summary of the model as



We can see that all the column’s p value is almost near to 0 hence contributing enough to the model however the difference between null and residual deviance is still not significant.

Let’s compute the confusion matrix and accuracy

prob<-predict(logit,type=c("response"),Affairs)

confusion<-table(prob>0.5,Affairs$affairsst)

Accuracy<-sum(diag(confusion)/sum(confusion))

But even applying the transformations we are still with 77 percent accuracy

