**IMPORTING LIBRARIES:**

> library(caTools)#for splitting datasets into training and test sets; mainly for using split

> library(caret) #for evaluation metrics like RMSE, R^2 and MAE

**ASSUMPTIONS:**

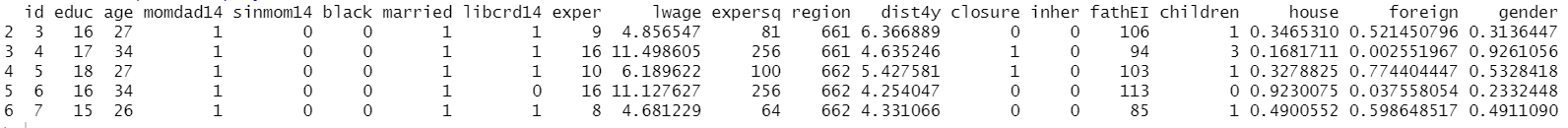
We are going to make use of the Gamma distribution to estimate the size of Wage returns, based on the covariates given.

We will be applying the GLM (Generalized Linear Model) with link function corresponding to the Gamma family.

The residuals obtained must not have any pattern with the fitted values or the explanatory variables.

**IMPORTING DATASET & PERFORMING EDA (Exploitary Data Analysis)**

> head(educCS,5) # Displays first 5 records of educCS table



> dataset <- educCS # Storing eduCS into dataset

> colnames(dataset) #Displays the column names of the dataset

[1] "id" "educ" "age" "momdad14" "sinmom14" "black" "married" "libcrd14" "exper" "lwage" "expersq" "region" "dist4y" "closure" "inher"

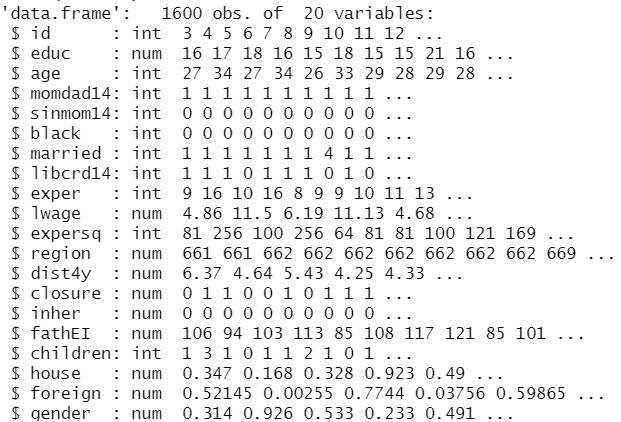
[16] "fathEI" "children" "house" "foreign" "gender"

> str(dataset)

#Command to see the structure of the dataset.

#Only population, brp, lbrp, univ\_share are numeric variables. Remaining are categorical variables.

#lwage or estimate of the wage returns of one additional year of education, is the Response or dependent or output variable



**TREATMENT OF MISSING VALUES**

> sum(is.na(dataset))

[1] 0

#Clearly, there are no missing values in the dataset. So, no imputations required.

> attach(dataset)

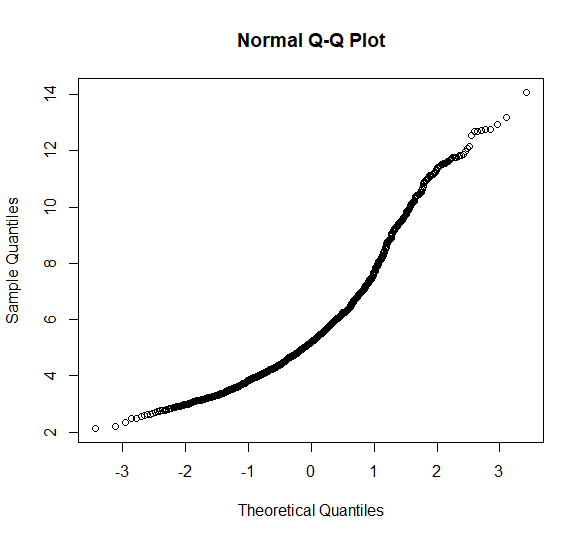
# Attaching the dataset

**ECONOMETRIC ANALYSIS:**

# Checking whether the Response variable conforms with the Normal Distribution or not

# Plotting the qqplot of the distribution, i.e. checking the Quantiles of the distribution against quantiles of the normal distribution

> qqnorm(wprem)



#This is clearly not normally distributed. From the qqplot, it deviation from the line, indicates that clearly the distribution of wprem is clearly not normal.

HISTOGRAM WITH PDF OF RESPONSE VARIABLE PLOTTED:

> hist(lwage,breaks = 20)

# To check whether the histogram corresponds to that of the Gamma distribution or not

> set.seed(2022) # Random seed set

> x <- seq(0,10,0.1)

#Estimatiion of gamma distribution paramters from Response variable ‘lwage’ values

> alpha <- ((mean(educCS$lwage))^2)/var(educCS$lwage)

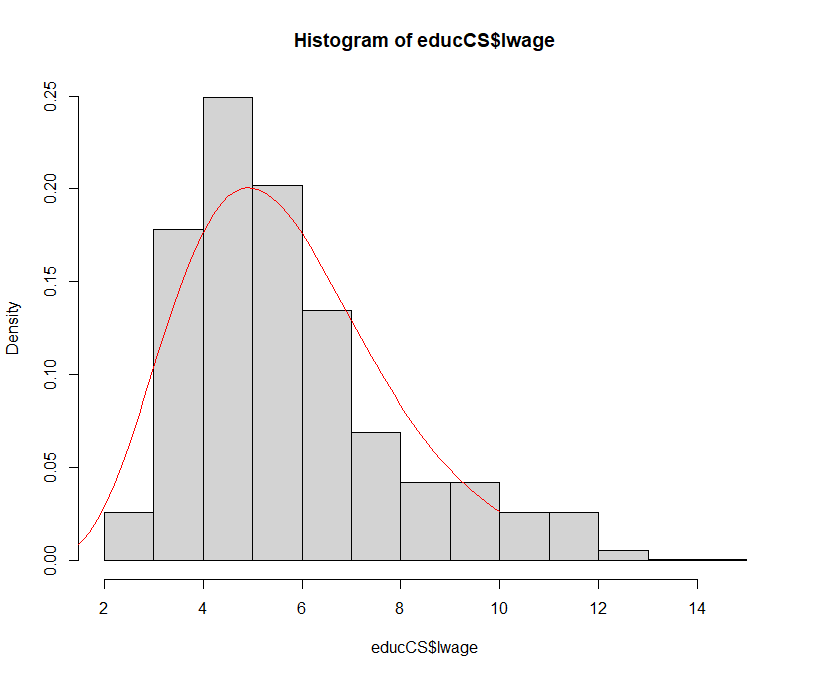
> beta <- (mean(educCS$lwage))/var(educCS$lwage)

> hist(educCS$lwage,probability = T)

> lines(x,dgamma(x,alpha,beta), type='l', col='red')

#Histogram of Dependent/Response variable resembles gamma distribution.

#So, the assumption of Gamma distribution modelling the size of wage returns holds.



#Since, all the values are positive. Also, the Histogram of Dependent/Response variable resembles gamma distribution. So, the assumption of Gamma distribution modelling the size of wage returns holds.

> unique(lwage>0)

[1] TRUE

#Since all the values of lwage are positive, gamma distribution would be apt.

**SPLITTING DATASET INTO TRAIN & TEST DATASETS:**

> set.seed(123)

> split<-sample.split(dataset, SplitRatio = 0.8)

#Split ratio splits the observations randomly into training and test data sets

> training\_set = subset(dataset, split==TRUE)

# Training\_set formed

> test\_set = subset(dataset, split==FALSE)

#Test\_set formed

> colnames(training\_set)

# Displays column names of the training dataset

[1] "id" "educ" "age" "momdad14" "sinmom14" "black" "married" "libcrd14" "exper"

[10] "lwage" "expersq" "region" "dist4y" "closure" "inher" "fathEI" "children" "house"

[19] "foreign" "gender"

> y\_train<-training\_set[c(10)]

> y\_test <- test\_set[c(10)]

# Assigning the Response variable to variable y\_train and y\_test

> detach(dataset)

> attach(training\_set)

# Detaching the dataset, and attaching the training\_set dataframe

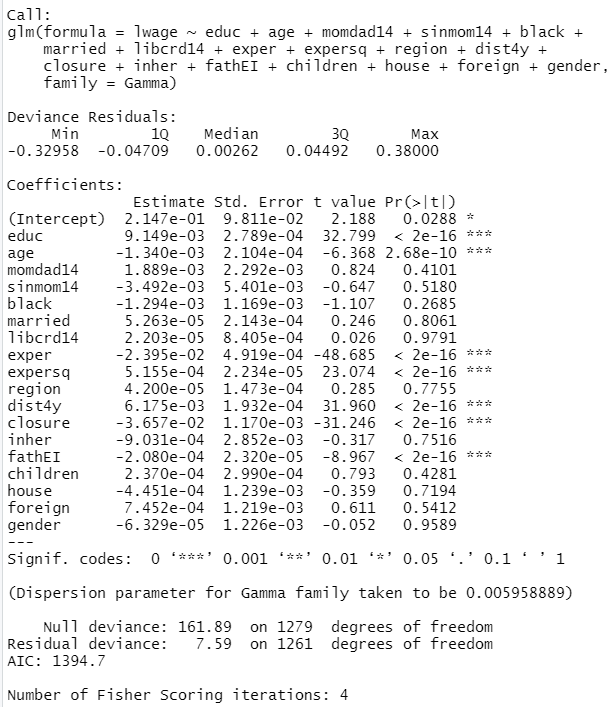
**MODEL SELECTION STAGE:**

# Generalised Linear Model (GLM) created, using inverse link function and Gamma family

FIRST MODEL:-

> model <- glm(lwage~educ+age+factor(momdad14)+factor(sinmom14)+factor(black)+factor(married)+factor(libcrd14)+exper+expersq+factor(region) +dist4y+factor(closure)+factor(inher)+fathEI+factor(children)+house+foreign+gender,family=Gamma)

> summary(model)



#AIC (Akaike's Information Criteria) for this model is 1394.7. Smaller the AIC value, better the model.

# Variables – momdad14, sinmom14, black, married, libcrd14, region, inher, children, house, foreign and gender are not significant. These variables can be subsequently removed in the next models.

> BIC(model)

[1] 1497.789.

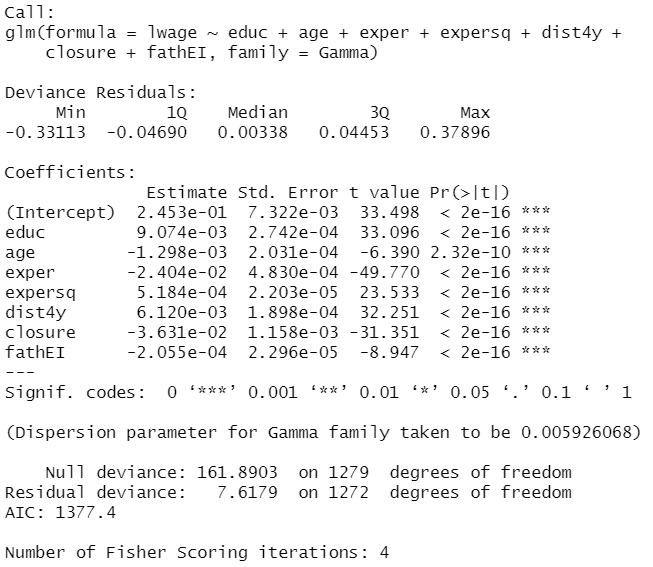
#BIC (Bayesian Information Criteria) for model is 1497.789. Smaller the BIC, better the model.

SECOND MODEL:-

#Second model created by removing insignificant coefficients. This is obtained from the last model

> model2 <- glm(lwage~educ+age+exper+expersq +dist4y+closure+fathEI,family=Gamma)

> summary(model2)

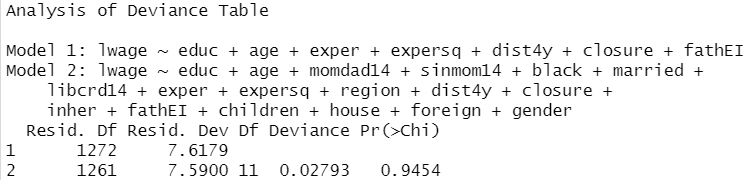


#AIC for this model is 1377.4, better than the previous model.

> BIC(model2)

[1] 1423.794

> anova(model2, model,test='Chi')



#Since, we are comparing two non-normally distributed data, we shall compare the two models by applying the Chi square test. Clearly, model2 is an improvement over model1, in terms of AIC. Also, since Pr(>Chi)= 0.9454, which is greater than 5%, indicates that the two models are not significantly different from each other. We are okay with removing the non-significant variables. Thus, model2 is better than model.

THIRD MODEL:

> model3 <- glm(lwage~educ+I(age^2)+age+exper+expersq

+dist4y+factor(closure)+fathEI,family=Gamma)

> summary(model3)

#AIC is 1377.2. Reduced as compared to model2. #Adding a higher degree variable for a significant variable (age), i.e. age^2. This has further reduced AIC of the model.

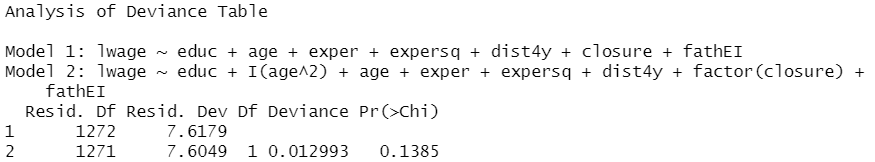
> BIC(model3)

[1] 1428.762

#BIC has increased, which is a bad sign.

> anova(model2, model3,test='Chi')

# Comparing last two models by ANOVA test Chi-square method



# Clearly, model2 is an improvement over model1, in terms of AIC. Pr(>Chi)=0.1385, which is greater than 5%, indicates that the two models are not significantly different from each other. So, model2 is better than model3. Introducing additional variables didn't add significant reduction in AIC. from each other. We can choose the model with lesser covariates, i.e. model3.

FOURTH MODEL:

> model4 <- glm(lwage~educ+age+exper+expersq+I(expersq^2)

+dist4y+factor(closure)+fathEI,family=Gamma)

> summary(model4)

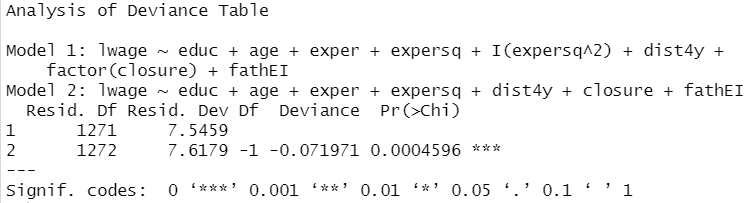
# Further reduction of AIC, with this new addition to 1367.2.

> BIC(model4)

[1] 1418.786

# BIC has also reduced.

> anova(model4, model2,test='Chi')



#Since Pr(>Chi)=0.0004596 , obtained is <5%, there is significant difference between these two models,

So, we add that new covariate expersq^2 here. Model4 is certainly better than model 2.

FIFTH MODEL:

#Adding the interaction variable policy:univ\_share

> model5 <- glm(lwage~educ+age+exper+expersq+I(expersq^2)+I(expersq^3)

+dist4y+factor(closure)+fathEI,family=Gamma)

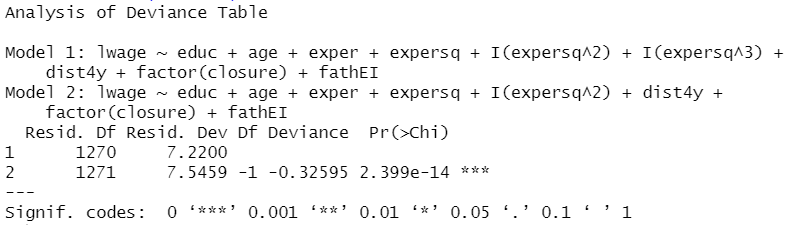
> summary(model5)

#Further reduction in AIC to 1312.7, clearly suggests that model5 is better.

> BIC(model5)

[1] 1369.367

> anova(model4, model5,test='Chi')



#Again, model5 is superior to model 4, because Pr(>Chi)=2.399e-14<5%.

SIXTH MODEL:

> model6 <- glm(lwage~educ+age+exper+expersq+I(expersq^2)+I(expersq^3)+I(expersq^4)

+dist4y+factor(closure)+fathEI,family=Gamma)

> summary(model6)

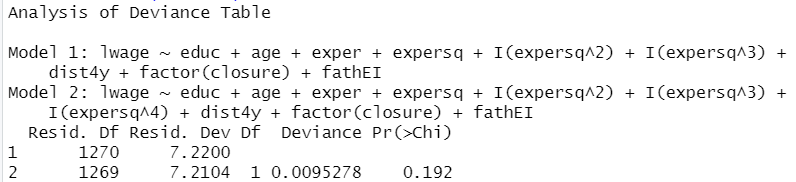
#AIC is 1313. Reduced as compared to previous model.

> BIC(model6)

[1] 1374.829

#BIC has also reduced as compared to the previous models.

> anova(model6, model5,test='Chi')



#Since, Pr(>Chi)=0.192<5%, adding additional variable doesn't significantly change the model. So, we prefer model 5 to model6.

SEVENTH MODEL:

#Adding the interaction variable expersq:edu

> model7 <- glm(lwage~educ+expersq:educ+age+exper+expersq+I(expersq^2)+I(expersq^3)

+dist4y+factor(closure)+fathEI,family=Gamma)

summary(model7)

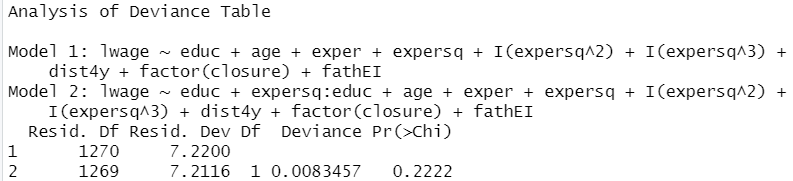
# #AIC increased, so we reject this model.

> BIC(model6)

[1] 1375.04

#BIC has also increased as compared to the previous models. So, not a good sign.

> anova(model5, model7,test='Chi')



#Since, Pr(>Chi)= 0.2222 >5%, adding additional variable doesn't significantly change the model. So, we prefer model 5 to model7.

EIGHTH MODEL:

> model8 <- glm(lwage~educ+age+exper+expersq+expersq:dist4y+I(expersq^2)+I(expersq^3)

+dist4y+factor(closure)+fathEI,family=Gamma)

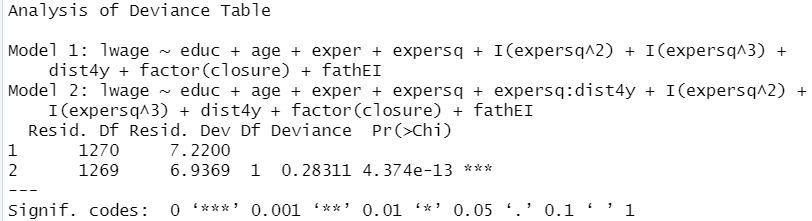
#AIC is 1263.4. Reduced as compared to previous model. So, preferable model.

> BIC(model8)

[1] 1325.272

#BIC has also reduced as compared to the previous models.

> anova(model5, model8,test='Chi')



#Model8 is certainly better than model7. Pr(>Chi)=4.374e-13<5%, suggests that model8 is significantly different from model5. Its AIC has also reduced to 1263.4. We prefer model8 to model5.

NINETH MODEL:

#Adding the interaction variable expersq:closure

> model9 <- glm(lwage~educ+age+exper+expersq+expersq:dist4y+expersq:factor(closure)+I(expersq^2)+I(expersq^3)+dist4y+factor(closure)+fathEI,family=Gamma)

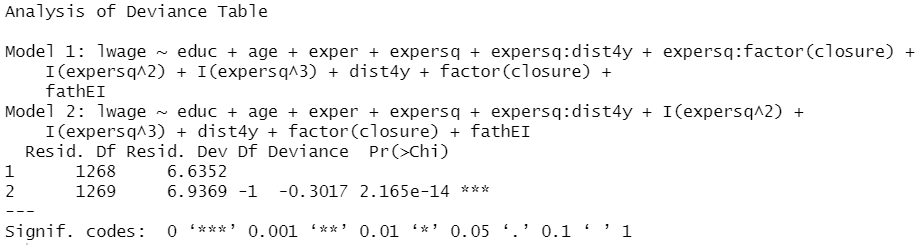
#AIC is 1208.5. Reduced as compared to previous model.

> BIC(model9)

[1] 1275.46

#BIC has also reduced as compared to the previous models.

> anova(model9, model8,test='Chi')



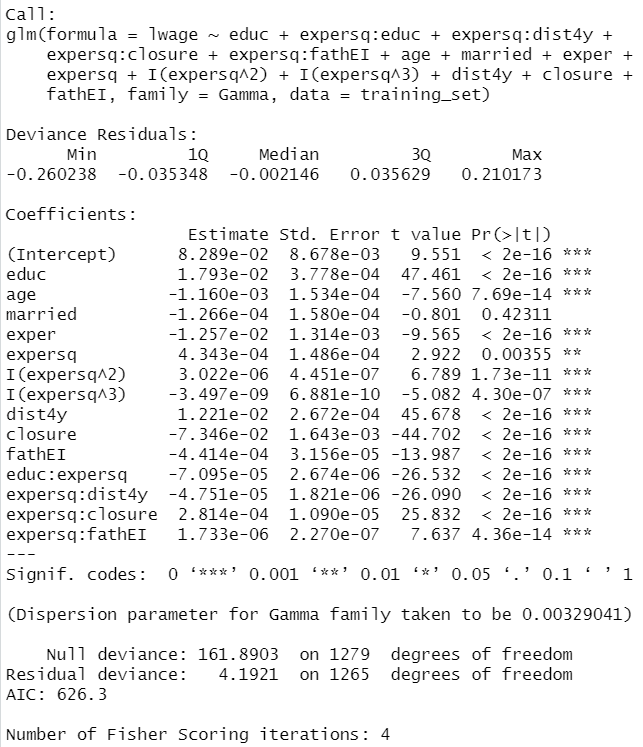
# Model9 is certainly better than model8. Pr(>Chi)=2.165e-14<5%, suggests that model9 is significantly different from model8. AIC of model9 has also reduced.

TENTH MODEL:

#Adding the interaction variable expersq:fathEI

> model10 <- glm(lwage~educ+expersq:educ+expersq:dist4y+expersq:closure+expersq:fathEI+age+married+exper+expersq+I(expersq^2)+I(expersq^3) +dist4y+closure+fathEI,training\_set,family=Gamma)

> summary(model10)



#Further reduction in AIC to 626.75, suggests that model10 is even better.

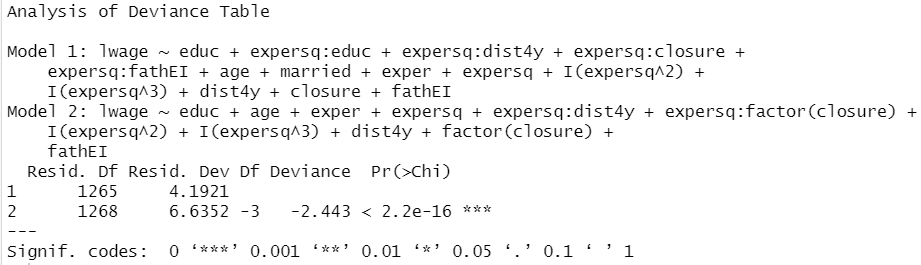
#All coefficients except that of married are significant.

> BIC(model10)

[1] 708.7788

#BIC has also reduced as compared to the previous models.

> anova(model10, model9,test='Chi')



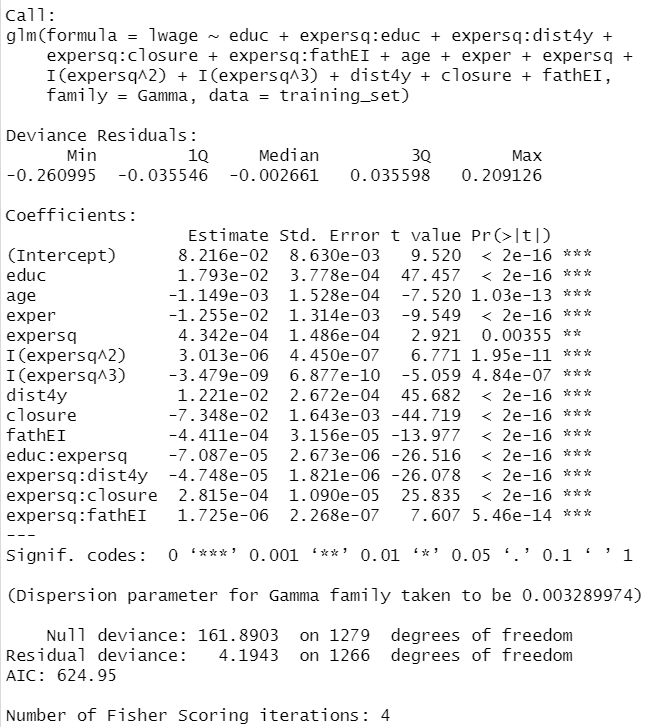
#Model10 is certainly better than model9. Pr(>Chi)=2.2e-16<5%, suggests that model10 is significantly different from model9.

ELEVENTH MODEL:

#Removing variable ‘married’

>model11 <- glm(lwage~educ+expersq:educ+expersq:dist4y+expersq:closure+expersq:fathEI+age+exper+expersq+I(expersq^2)+I(expersq^3)+dist4y+closure+fathEI,training\_set,family=Gamma)

>summary(model11)



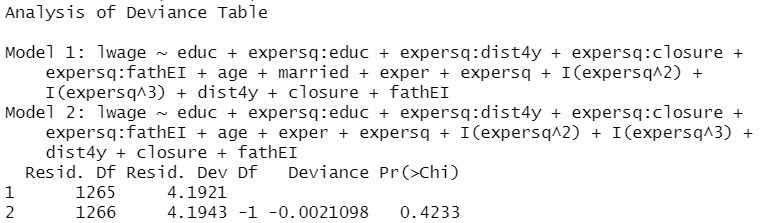
#AIC has even reduced to 624.95, lowest obtained so far!

> BIC(model11)

[1] 702.2685

#BIC has also reduced as compared to the previous models. Infact, it’s the lowest so far.

> anova(model10, model11,test='Chi')



# Removing variable ‘married’, causes both AIC and BIC, to reduce further.

#Clearly, model11 is an improvement over model10, in terms of AIC. Also, Pr(>Chi)=0.4233, which is greater than 5%, indicates that the two models are not significantly different from each other.

#So, model11 having lesser covariates, is better than model10. Infact, it’s the best model obtained so far.

**PREDICTION OF TEST DATASETS USING BEST MODEL (model11):**

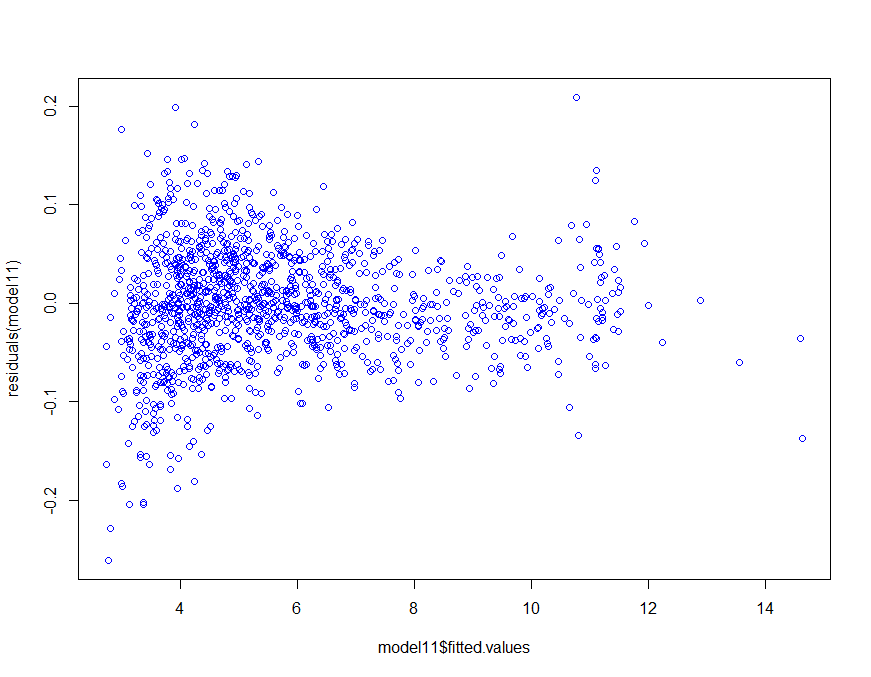
>test\_set\_covariates <- test\_set[c(-10)]

#This stores all the covariates of the test\_set in test\_set\_covariates. 10th column (response variable) has been removed.

> test\_set\_pred <- predict(model11,newdata =test\_set\_covariates,type="response")

# Storing test set predictions in test\_set\_pred

**VERIFICATION OF ERROR/RESIDUAL ASSUMPTIONS:**

plot(model11$fitted.values,residuals(model11),col='blue')

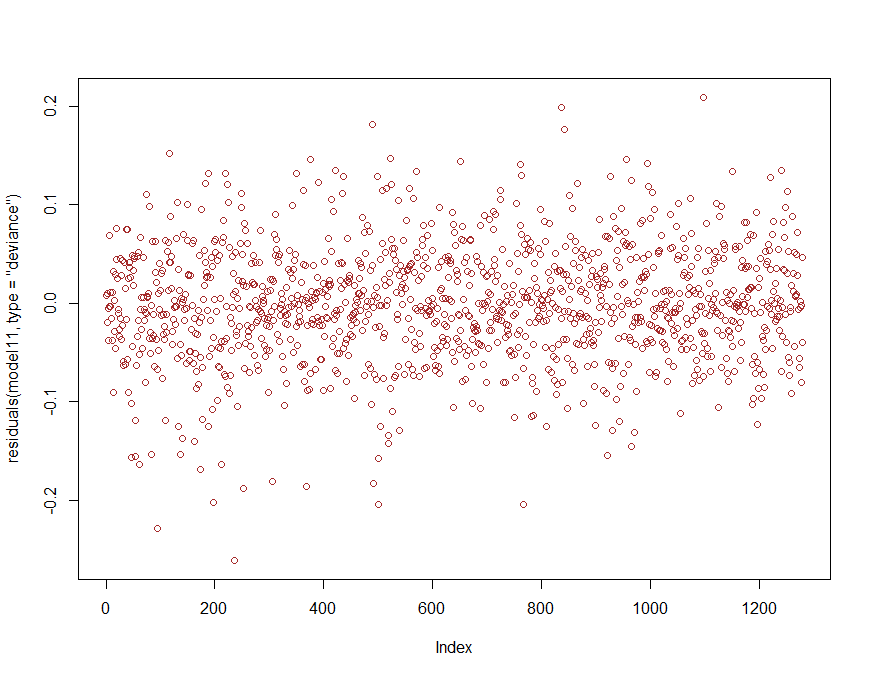
#Plot suggests that the residuals show some pattern with fitted values. For lower values of fitted values, the residuals exhibit higher variance, which decreases with increasing magnitude of the fitted values.

#This indicates that the initial selection of variables for model building can be looked into. It shows that the errors are symmetrically distributed about origin, having constant variance.

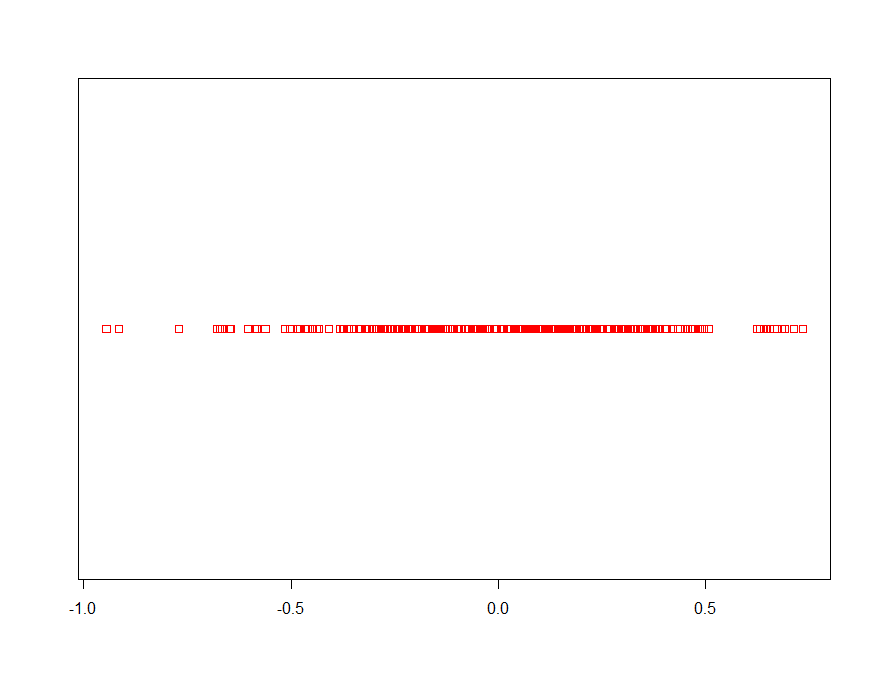
#There appears to be no pattern between the residuals and fitted values. There seems to be missing data for a certain range.

> plot(residuals(model11,type='deviance'),col='brown')

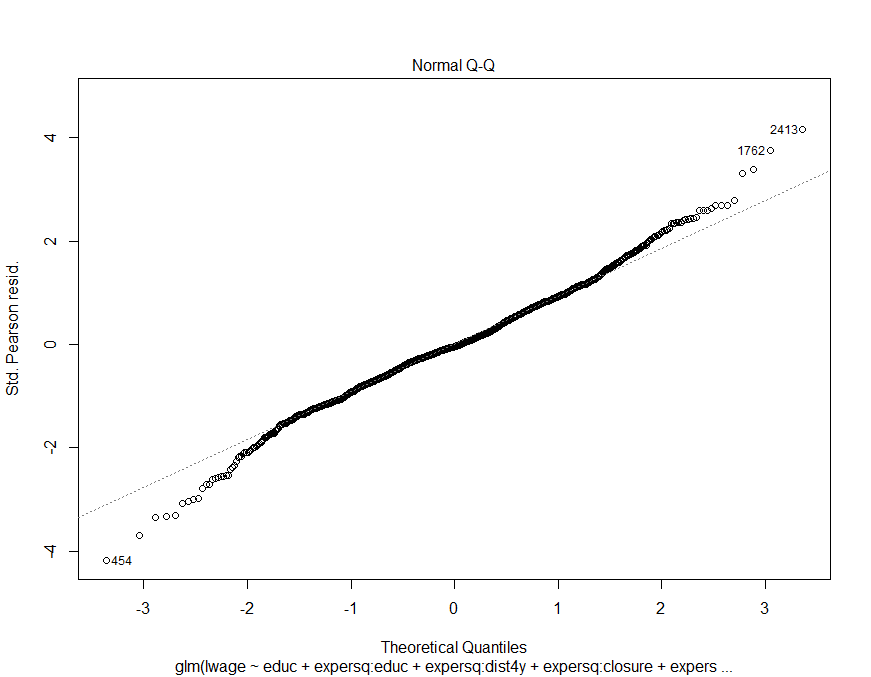
# It can be seen that the residuals of the model11 are symmetrical about origin, having a constant variance; and likely to have an approximately normal distributions. The residuals appear to be patternless.



> plot(y\_test-test\_set\_pred,col='red')



#plot suggests normally distributed errors symmetrical about origin



#Pearson residuals approximately normal, with slight deviation from normal, over the tails.

**MODEL EVALUATION METRICS:**

Evaluating Model11:

> AIC(model11)

[1] 624.9493

#AIC of the model11 is 624.9493, lowest value obtained so far.

> BIC(model11)

[1] 702.2685

# 702.2685, lowest value obtained so far. So, this is the best model, so far.

> RMSE(test\_set\_pred,y\_test$lwage)

[1] 0.001145843

#0 .2946545 is the RMSE obtained for test-dataset. Being very close to 0, it solidifies the fact that model11 is accurate in predicting the output for unknown dataset.

> MAE(test\_set\_pred,y\_test$lwage)

[1] 0.2359111

#0.0009609038 is the Mean absolute error. Being a very small value, this indicates that the model accurately predicts the output.

> R2(test\_set\_pred,y\_test$wprem)

[1] 0.9793768

# Large proportion of the variability in the model6 is explained by the covariates.