**How to do Regression Testing in R: Linear Regression**

Code:



**Step 1: Import libraries and source file**

library(dplyr) # For data manipulations

library(ggplot2) #For building graphs

library(car) #has relevant functions to perform linear regression

data<-read.csv("dm.csv")

**Step 2: Do exploratory analysis using Summary and Plot functions**.

Run summary command to check variable with ‘NA’ values.

Then, do cross tabulations between the dependent and each independent variable. This is done to get an overall understanding about variables and to see which variables have missing values.

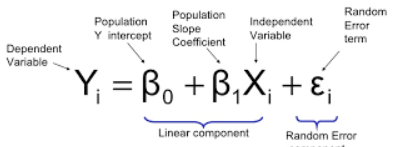
For Example: If Amount Spent is dependent variable, then plot this dependent variable against each independent variable such as Age, Salary etc.

plot(data$Age,data$AmountSpent,col="red")

plot(data$Gender,data$AmountSpent,col="red")

**These cross tabulations will later help us during model optimization.** **How?**

Dependent variable is sum of effect of all independent variables added together.



Plotting will help us to understand the contribution & direction (+ or -) of each of the independent variable to the dependent variable. When doing model summary, if we observe that any independent variable is behaving different to what we observed in the plots, then that variable can be removed from model.

Eg:- Gender male is spending more as per plotted graph (contributing positively to dependent variable). But, later when checking model summary, it is found that Gender male is having -ve beta coefficient. This is contrary to what we have seen in the plot. So, we remove Gender variable from model.

**Step 3: Handle missing values, outliers**

Substitute for missing values if possible or rename them as “Missing” category.

Handle outliers if needed. (Substitute values or remove records as suited)

Ex: data$History1<-ifelse(is.na(data$History),"Missing", as.character(data$History))

**Step 4: Rearrange categorical values into new groups (if required)**

Ex: data$Age1<-ifelse(data$Age!="Young","Middle-Old",as.character(data$Age))

data$Children1<-ifelse(data$Children==3|data$Children==2,"3-2",as.character(data$Children))

**Step 5: Create first model**

Remove unwanted columns:

data1<-data[,-c(1,7,8)] #remove columns from which other derived columns were created

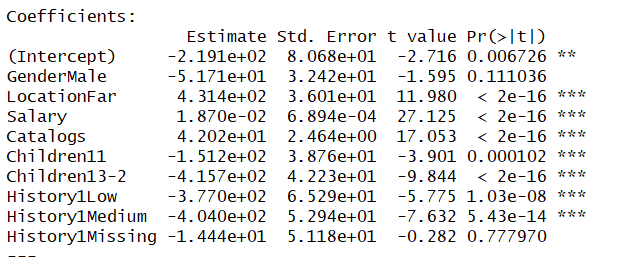
mod1<-lm(AmountSpent~.,data=data1)

summary(mod1)

**Step 6: Remove insignificant variables and create refined model**

mod2<-lm(formula = AmountSpent ~ Gender + Location + Salary + Catalogs + Children1 + History1, data = data1)

summary(mod2)



Note that Gender Male is having -ve beta coefficient, whereas, actual value plot was showing it has more prominence.

**Step 7: Create Dummy variables if required**

Example:

data1$Missing\_d<-ifelse(data$History1=="Missing",1,0)

data1$Low\_d<-ifelse(data$History1=="Low",1,0)

data1$Med\_d<-ifelse(data$History1=="Medium",1,0)

data1$High\_d<-ifelse(data$History1=="High",1,0)

**Step 8: Refine model using dummy variables and remove insignificant variables**

mod3<-lm(formula = AmountSpent ~ Male\_d + Location + Salary + Catalogs + Children1+Med\_d+Low\_d , data = data1)

mod4<-lm(formula = AmountSpent ~ Location + Salary + Catalogs + Children1+Med\_d+Low\_d, data = data1)

#Note that the adjusted R square increases when Male\_d was present in Mod3. So Male\_d can be included in the model too, but it's direction is wrong and therefore we exclude it.

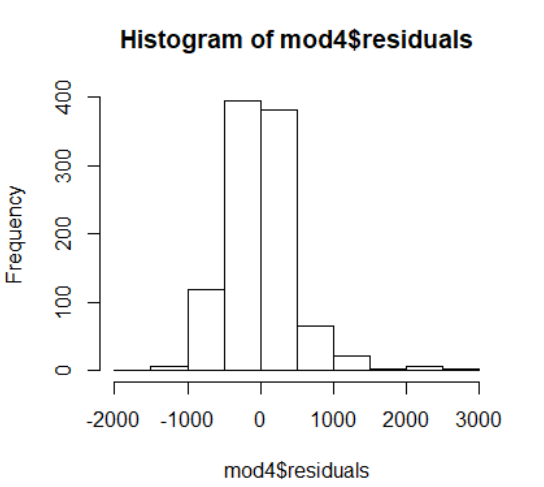
**Step 9: Assumptions check**

1.Plot histogram and check for skewness

2.Plot qqplot and check for behavior of residuals.

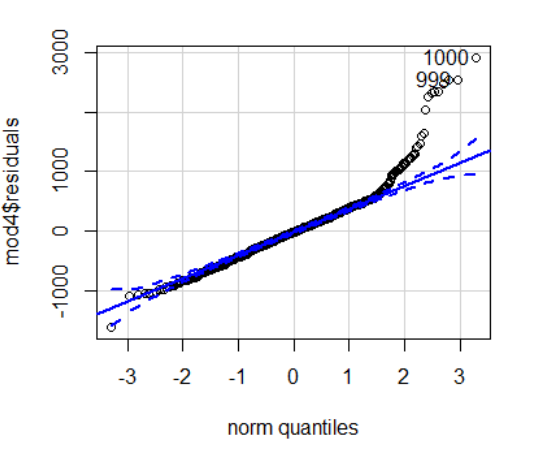
3.Multicollinearity check

4.Heteroskedasticity (constant variance check)

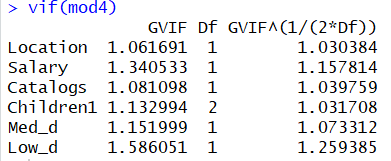
1.Histogram

hist(mod4$residuals)

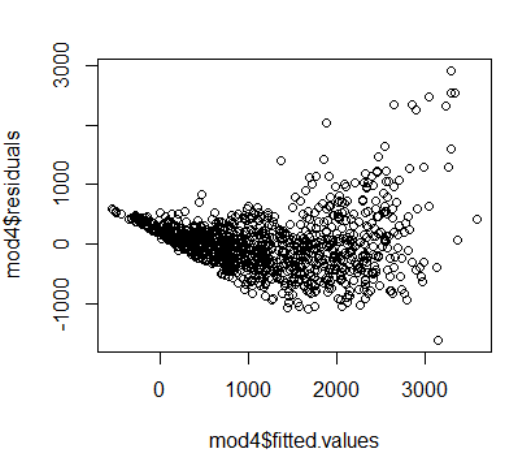
2.qqplot - qqPlot(mod4$residuals)



3.Multicollinearity check - vif(mod4) – Should be less than 5 or 10 as per business requirement



4.Heteroskedasticity - plot(mod4$fitted.values,mod4$residuals)



Above plots show skewness, non-normality of residuals in qqplot and heteroskedasticity. So, we should try remedial measures - that is - transformations.

**Step 10: Transformations**

Take log, square root etc. for dependent or independent variables. After doing transformations, check if improvement is there, in above 4 step assumptions.

In this case, after trying different combinations of transformations, the best one chosen one is

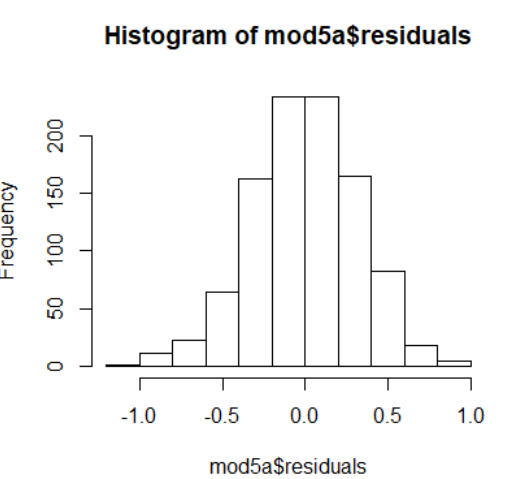
mod5a<-lm(formula = log(AmountSpent) ~ Location + log(log(Salary)) + Catalogs + Children1+Med\_d+Low\_d, data = data1)

summary(mod5a) #has highest R squared.

Now check assumptions again:

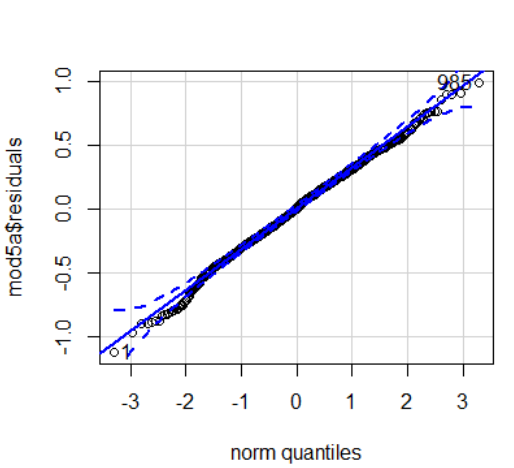
##################################

hist(mod5a$residuals) # no skewness



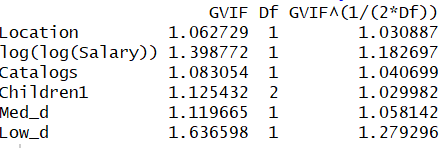
##################################

qqPlot(mod5a$residuals) # qqplot looks okay



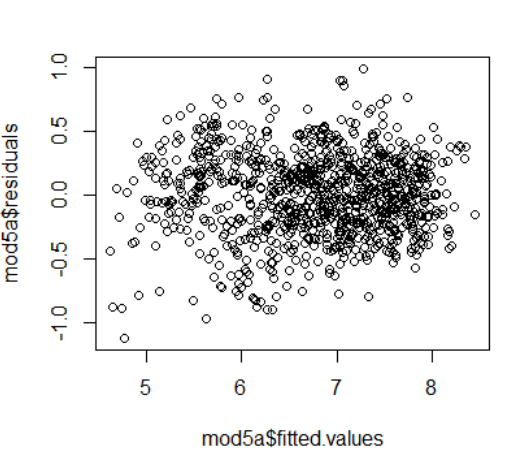
###################################

vif(mod5a) # no multicollinearity



###################################

plot(mod5a$fitted.values,mod5a$residuals) # no funneling



#For mod5a

predicted<-mod5a$fitted.values

actual<-log(data1$AmountSpent)

dat<-data.frame(predicted,actual)

**#Create run chart**

p<-ggplot(dat,aes(x=row(dat)[,2],y=predicted))

p+geom\_line(colour="blue")+geom\_line(data=dat,aes(y=actual),colour="black")

#We can also take MAPE of two models on the validation data and compare to finalize the model

#But, validation is not a necessary requirement for Linear model since meeting the assumptions means validating the model