Advanced Linear Regression

library(readxl)  
library("car")

## Loading required package: carData

library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

library("ggpubr")

## Warning: package 'ggpubr' was built under R version 4.0.3

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.0.4

library(stargazer)

## Warning: package 'stargazer' was built under R version 4.0.3

##   
## Please cite as:

## Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

library(data.table)

## Warning: package 'data.table' was built under R version 4.0.3

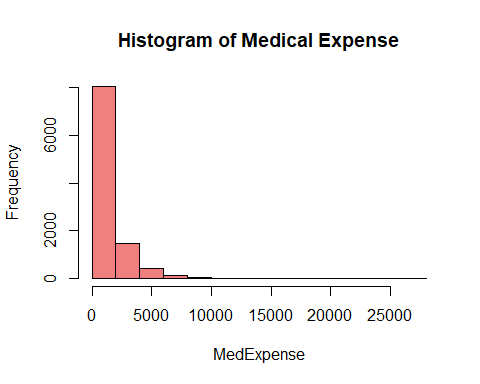
setwd("C:/Users/surya/Downloads")  
  
hi <- read\_excel("HealthInsurance.xlsx", sheet = 'Data')  
  
#NA values column wise  
sapply(hi, function(x) sum(is.na(x)))

## medexpense healthins age female blackhisp   
## 0 0 0 0 0   
## income illnesses ssiratio lowincome firmsize   
## 0 0 0 0 0   
## firmlocation educyr private hisp black   
## 0 0 0 0 0   
## married verygood good fair poor   
## 0 0 0 0 0   
## poverty midincome msa prioritylist logmedexpense   
## 0 0 0 0 0   
## agesqrd logincome vgh fph   
## 0 0 0 0

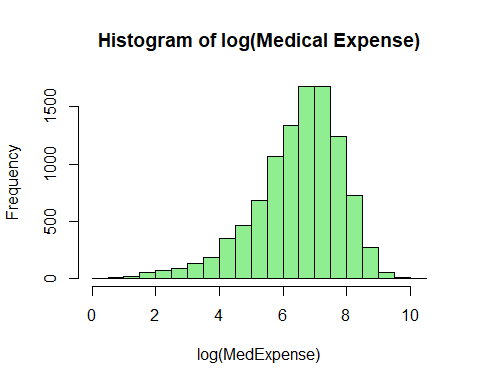
str(hi)

## tibble [10,089 x 29] (S3: tbl\_df/tbl/data.frame)  
## $ medexpense : num [1:10089] 595 1783 176 2437 330 ...  
## $ healthins : num [1:10089] 1 1 0 1 0 1 1 0 1 0 ...  
## $ age : num [1:10089] 74 73 80 70 91 82 81 90 73 79 ...  
## $ female : num [1:10089] 1 0 1 0 0 1 0 1 0 1 ...  
## $ blackhisp : num [1:10089] 0 0 0 0 0 0 0 0 0 0 ...  
## $ income : num [1:10089] 94.55 35.8 9.6 38.71 8.72 ...  
## $ illnesses : num [1:10089] 0 3 1 5 3 2 4 2 1 3 ...  
## $ ssiratio : num [1:10089] 0.15 0.396 1 0.207 0.537 ...  
## $ lowincome : num [1:10089] 0 0 0 0 0 0 0 1 0 1 ...  
## $ firmsize : num [1:10089] 0 0.1 0 0 0 0 0 0 0 0 ...  
## $ firmlocation : num [1:10089] 0 0 0 0 0 0 0 0 0 0 ...  
## $ educyr : num [1:10089] 16 8 12 17 16 15 17 12 14 12 ...  
## $ private : num [1:10089] 1 1 0 1 0 1 1 1 1 1 ...  
## $ hisp : num [1:10089] 0 0 0 0 0 0 0 0 0 0 ...  
## $ black : num [1:10089] 0 0 0 0 0 0 0 0 0 0 ...  
## $ married : num [1:10089] 1 1 0 1 1 0 0 0 0 0 ...  
## $ verygood : num [1:10089] 0 0 0 0 1 0 1 0 0 0 ...  
## $ good : num [1:10089] 1 1 0 1 0 0 0 0 1 1 ...  
## $ fair : num [1:10089] 0 0 0 0 0 1 0 0 0 0 ...  
## $ poor : num [1:10089] 0 0 0 0 0 0 0 0 0 0 ...  
## $ poverty : num [1:10089] 0 0 1 0 0 0 0 0 0 0 ...  
## $ midincome : num [1:10089] 0 0 0 0 1 0 0 0 1 0 ...  
## $ msa : num [1:10089] 0 0 1 1 1 1 0 0 1 0 ...  
## $ prioritylist : num [1:10089] 1 1 1 1 1 1 1 1 1 1 ...  
## $ logmedexpense: num [1:10089] 6.39 7.49 5.17 7.8 5.8 ...  
## $ agesqrd : num [1:10089] 5476 5329 6400 4900 8281 ...  
## $ logincome : num [1:10089] 4.55 3.58 2.26 3.66 2.17 ...  
## $ vgh : num [1:10089] 1 1 0 1 1 0 1 0 1 1 ...  
## $ fph : num [1:10089] 0 0 0 0 0 1 0 0 0 0 ...

#Checking if DV is suitable for OLS  
hist(hi$medexpense, col = 'lightcoral', main = "Histogram of Medical Expense", xlab = 'MedExpense', ylab = 'Frequency')



hist(hi$logmedexpense, col = 'lightgreen', main = "Histogram of log(Medical Expense)", xlab = 'log(MedExpense)', ylab = 'Frequency')



#Histogram shows the DV follows an exponential distribution. Log of DV makes it a ND  
#hist(hi$income)  
#hist(hi$logincome)  
#Histogram for income follows an exponential distribution. Log of income makes it a ND  
  
#Feature Engineering/Pre-processing  
#Converting categoricals into binary factor levels  
names(hi) <- tolower(colnames(hi))  
hi$health <- ifelse(hi$poor == 1, 1, ifelse((hi$verygood + hi$good + hi$fair) == 1, 2, 0))  
str(hi)

## tibble [10,089 x 30] (S3: tbl\_df/tbl/data.frame)  
## $ medexpense : num [1:10089] 595 1783 176 2437 330 ...  
## $ healthins : num [1:10089] 1 1 0 1 0 1 1 0 1 0 ...  
## $ age : num [1:10089] 74 73 80 70 91 82 81 90 73 79 ...  
## $ female : num [1:10089] 1 0 1 0 0 1 0 1 0 1 ...  
## $ blackhisp : num [1:10089] 0 0 0 0 0 0 0 0 0 0 ...  
## $ income : num [1:10089] 94.55 35.8 9.6 38.71 8.72 ...  
## $ illnesses : num [1:10089] 0 3 1 5 3 2 4 2 1 3 ...  
## $ ssiratio : num [1:10089] 0.15 0.396 1 0.207 0.537 ...  
## $ lowincome : num [1:10089] 0 0 0 0 0 0 0 1 0 1 ...  
## $ firmsize : num [1:10089] 0 0.1 0 0 0 0 0 0 0 0 ...  
## $ firmlocation : num [1:10089] 0 0 0 0 0 0 0 0 0 0 ...  
## $ educyr : num [1:10089] 16 8 12 17 16 15 17 12 14 12 ...  
## $ private : num [1:10089] 1 1 0 1 0 1 1 1 1 1 ...  
## $ hisp : num [1:10089] 0 0 0 0 0 0 0 0 0 0 ...  
## $ black : num [1:10089] 0 0 0 0 0 0 0 0 0 0 ...  
## $ married : num [1:10089] 1 1 0 1 1 0 0 0 0 0 ...  
## $ verygood : num [1:10089] 0 0 0 0 1 0 1 0 0 0 ...  
## $ good : num [1:10089] 1 1 0 1 0 0 0 0 1 1 ...  
## $ fair : num [1:10089] 0 0 0 0 0 1 0 0 0 0 ...  
## $ poor : num [1:10089] 0 0 0 0 0 0 0 0 0 0 ...  
## $ poverty : num [1:10089] 0 0 1 0 0 0 0 0 0 0 ...  
## $ midincome : num [1:10089] 0 0 0 0 1 0 0 0 1 0 ...  
## $ msa : num [1:10089] 0 0 1 1 1 1 0 0 1 0 ...  
## $ prioritylist : num [1:10089] 1 1 1 1 1 1 1 1 1 1 ...  
## $ logmedexpense: num [1:10089] 6.39 7.49 5.17 7.8 5.8 ...  
## $ agesqrd : num [1:10089] 5476 5329 6400 4900 8281 ...  
## $ logincome : num [1:10089] 4.55 3.58 2.26 3.66 2.17 ...  
## $ vgh : num [1:10089] 1 1 0 1 1 0 1 0 1 1 ...  
## $ fph : num [1:10089] 0 0 0 0 0 1 0 0 0 0 ...  
## $ health : num [1:10089] 2 2 0 2 2 2 2 0 2 2 ...

#Checking correlations  
#1,2:8,20,22:23 1,3,6,8,9,17:20,21,22,25,26,27 1,9:19,21,24:29  
hi\_corr <- hi[, c(1,2:7,23:24,30)]  
hi\_corr\_out <- hi[, c(1,8:9,12:13,16:22,26:29)]  
  
library(PerformanceAnalytics)

## Warning: package 'PerformanceAnalytics' was built under R version 4.0.3

## Loading required package: xts

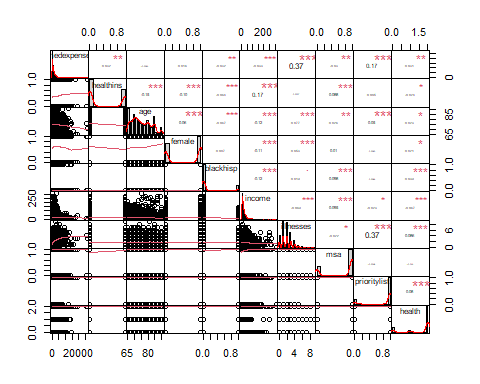
##   
## Attaching package: 'xts'

## The following objects are masked from 'package:data.table':  
##   
## first, last

##   
## Attaching package: 'PerformanceAnalytics'

## The following object is masked from 'package:graphics':  
##   
## legend

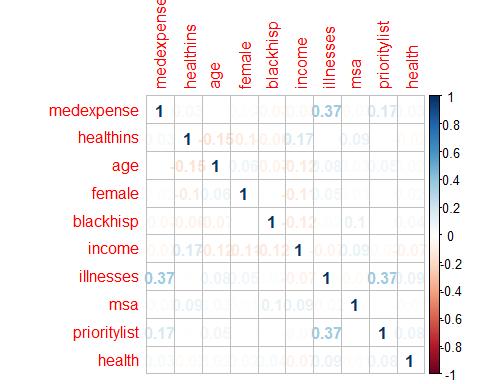
chart.Correlation(hi\_corr)



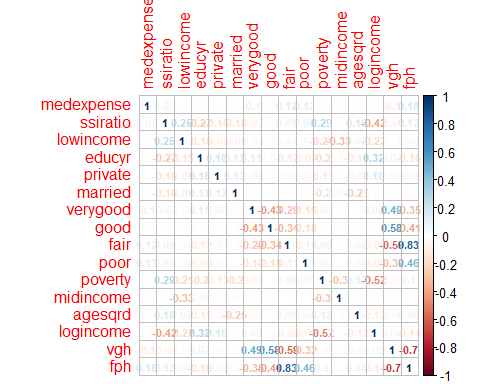
#chart.Correlation(hi\_corr\_out)  
  
library(corrplot)

## corrplot 0.84 loaded

hi\_corplot <- cor(hi\_corr)  
corrplot(hi\_corplot, method = "number")



hi\_corplot\_out <- cor(hi\_corr\_out)  
corrplot(hi\_corplot\_out, method = "number", number.cex= 0.7)



#Regression models  
names(hi)

## [1] "medexpense" "healthins" "age" "female"   
## [5] "blackhisp" "income" "illnesses" "ssiratio"   
## [9] "lowincome" "firmsize" "firmlocation" "educyr"   
## [13] "private" "hisp" "black" "married"   
## [17] "verygood" "good" "fair" "poor"   
## [21] "poverty" "midincome" "msa" "prioritylist"   
## [25] "logmedexpense" "agesqrd" "logincome" "vgh"   
## [29] "fph" "health"

m1 <- lm(log(medexpense) ~ blackhisp + illnesses + health + prioritylist, data = hi)  
summary(m1)

##   
## Call:  
## lm(formula = log(medexpense) ~ blackhisp + illnesses + health +   
## prioritylist, data = hi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.3567 -0.6737 0.1469 0.8305 4.0619   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.15562 0.04210 122.463 < 2e-16 \*\*\*  
## blackhisp -0.16381 0.03290 -4.979 6.51e-07 \*\*\*  
## illnesses 0.37969 0.01013 37.469 < 2e-16 \*\*\*  
## health 0.08816 0.01678 5.254 1.52e-07 \*\*\*  
## prioritylist 0.57884 0.03854 15.018 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.221 on 10084 degrees of freedom  
## Multiple R-squared: 0.1963, Adjusted R-squared: 0.196   
## F-statistic: 615.9 on 4 and 10084 DF, p-value: < 2.2e-16

m2 <- lm(log(medexpense) ~ healthins + age + female + blackhisp + income + illnesses + ssiratio + health + msa + prioritylist, data = hi)  
summary(m2)

##   
## Call:  
## lm(formula = log(medexpense) ~ healthins + age + female + blackhisp +   
## income + illnesses + ssiratio + health + msa + prioritylist,   
## data = hi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.4024 -0.6648 0.1393 0.8251 4.0040   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.3949334 0.1487225 36.275 < 2e-16 \*\*\*  
## healthins 0.0961037 0.0260423 3.690 0.000225 \*\*\*  
## age -0.0054223 0.0018769 -2.889 0.003874 \*\*   
## female 0.0569007 0.0248827 2.287 0.022231 \*   
## blackhisp -0.1626401 0.0334817 -4.858 1.21e-06 \*\*\*  
## income 0.0013871 0.0006445 2.152 0.031393 \*   
## illnesses 0.3773489 0.0101737 37.091 < 2e-16 \*\*\*  
## ssiratio 0.1807656 0.0383344 4.715 2.44e-06 \*\*\*  
## health 0.0895209 0.0167790 5.335 9.75e-08 \*\*\*  
## msa -0.0401874 0.0278627 -1.442 0.149239   
## prioritylist 0.5808824 0.0384830 15.095 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.219 on 10078 degrees of freedom  
## Multiple R-squared: 0.2, Adjusted R-squared: 0.1992   
## F-statistic: 251.9 on 10 and 10078 DF, p-value: < 2.2e-16

m3 <- lm(log(medexpense) ~ healthins + age + female + blackhisp + illnesses\*age + ssiratio\*age + health + msa + prioritylist, data = hi)  
summary(m3)

##   
## Call:  
## lm(formula = log(medexpense) ~ healthins + age + female + blackhisp +   
## illnesses \* age + ssiratio \* age + health + msa + prioritylist,   
## data = hi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.4200 -0.6680 0.1401 0.8243 3.9755   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.500727 0.303535 14.828 < 2e-16 \*\*\*  
## healthins 0.100468 0.026013 3.862 0.000113 \*\*\*  
## age 0.007340 0.004049 1.813 0.069873 .   
## female 0.055703 0.024823 2.244 0.024855 \*   
## blackhisp -0.169821 0.033311 -5.098 3.49e-07 \*\*\*  
## illnesses 0.703842 0.108046 6.514 7.65e-11 \*\*\*  
## ssiratio 0.809932 0.363998 2.225 0.026097 \*   
## health 0.086637 0.016756 5.171 2.38e-07 \*\*\*  
## msa -0.035281 0.027761 -1.271 0.203800   
## prioritylist 0.574217 0.038520 14.907 < 2e-16 \*\*\*  
## age:illnesses -0.004356 0.001431 -3.043 0.002348 \*\*   
## age:ssiratio -0.008838 0.004801 -1.841 0.065697 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.218 on 10077 degrees of freedom  
## Multiple R-squared: 0.2007, Adjusted R-squared: 0.1998   
## F-statistic: 230 on 11 and 10077 DF, p-value: < 2.2e-16

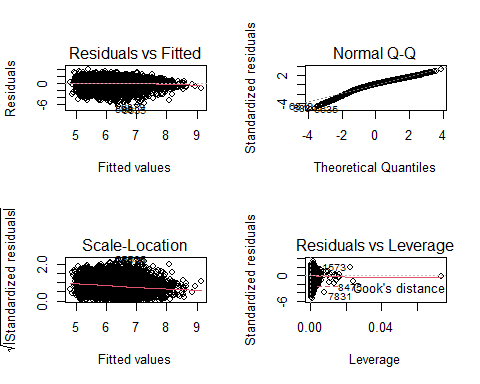
#Stargazer  
stargazer(m1, m2, m3, type='text', single.row = TRUE)

##   
## ======================================================================================================  
## Dependent variable:   
## ----------------------------------------------------------------------------------  
## log(medexpense)   
## (1) (2) (3)   
## ------------------------------------------------------------------------------------------------------  
## healthins 0.096\*\*\* (0.026) 0.100\*\*\* (0.026)   
## age -0.005\*\*\* (0.002) 0.007\* (0.004)   
## female 0.057\*\* (0.025) 0.056\*\* (0.025)   
## blackhisp -0.164\*\*\* (0.033) -0.163\*\*\* (0.033) -0.170\*\*\* (0.033)   
## income 0.001\*\* (0.001)   
## illnesses 0.380\*\*\* (0.010) 0.377\*\*\* (0.010) 0.704\*\*\* (0.108)   
## ssiratio 0.181\*\*\* (0.038) 0.810\*\* (0.364)   
## health 0.088\*\*\* (0.017) 0.090\*\*\* (0.017) 0.087\*\*\* (0.017)   
## msa -0.040 (0.028) -0.035 (0.028)   
## prioritylist 0.579\*\*\* (0.039) 0.581\*\*\* (0.038) 0.574\*\*\* (0.039)   
## age:illnesses -0.004\*\*\* (0.001)   
## age:ssiratio -0.009\* (0.005)   
## Constant 5.156\*\*\* (0.042) 5.395\*\*\* (0.149) 4.501\*\*\* (0.304)   
## ------------------------------------------------------------------------------------------------------  
## Observations 10,089 10,089 10,089   
## R2 0.196 0.200 0.201   
## Adjusted R2 0.196 0.199 0.200   
## Residual Std. Error 1.221 (df = 10084) 1.219 (df = 10078) 1.218 (df = 10077)   
## F Statistic 615.856\*\*\* (df = 4; 10084) 251.935\*\*\* (df = 10; 10078) 230.028\*\*\* (df = 11; 10077)  
## ======================================================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

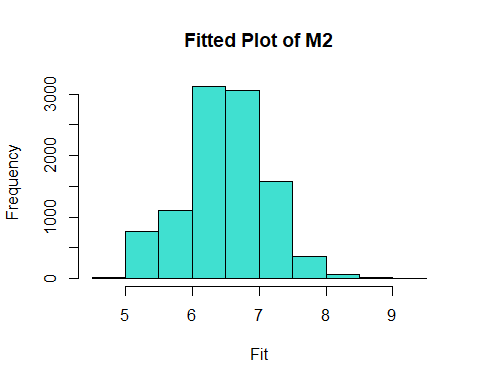
stargazer(m1, m2, m3, type='text', ci=TRUE, ci.level=0.95, single.row = TRUE)

##   
## ======================================================================================================  
## Dependent variable:   
## ----------------------------------------------------------------------------------  
## log(medexpense)   
## (1) (2) (3)   
## ------------------------------------------------------------------------------------------------------  
## healthins 0.096\*\*\* (0.045, 0.147) 0.100\*\*\* (0.049, 0.151)   
## age -0.005\*\*\* (-0.009, -0.002) 0.007\* (-0.001, 0.015)   
## female 0.057\*\* (0.008, 0.106) 0.056\*\* (0.007, 0.104)   
## blackhisp -0.164\*\*\* (-0.228, -0.099) -0.163\*\*\* (-0.228, -0.097) -0.170\*\*\* (-0.235, -0.105)   
## income 0.001\*\* (0.0001, 0.003)   
## illnesses 0.380\*\*\* (0.360, 0.400) 0.377\*\*\* (0.357, 0.397) 0.704\*\*\* (0.492, 0.916)   
## ssiratio 0.181\*\*\* (0.106, 0.256) 0.810\*\* (0.097, 1.523)   
## health 0.088\*\*\* (0.055, 0.121) 0.090\*\*\* (0.057, 0.122) 0.087\*\*\* (0.054, 0.119)   
## msa -0.040 (-0.095, 0.014) -0.035 (-0.090, 0.019)   
## prioritylist 0.579\*\*\* (0.503, 0.654) 0.581\*\*\* (0.505, 0.656) 0.574\*\*\* (0.499, 0.650)   
## age:illnesses -0.004\*\*\* (-0.007, -0.002)   
## age:ssiratio -0.009\* (-0.018, 0.001)   
## Constant 5.156\*\*\* (5.073, 5.238) 5.395\*\*\* (5.103, 5.686) 4.501\*\*\* (3.906, 5.096)   
## ------------------------------------------------------------------------------------------------------  
## Observations 10,089 10,089 10,089   
## R2 0.196 0.200 0.201   
## Adjusted R2 0.196 0.199 0.200   
## Residual Std. Error 1.221 (df = 10084) 1.219 (df = 10078) 1.218 (df = 10077)   
## F Statistic 615.856\*\*\* (df = 4; 10084) 251.935\*\*\* (df = 10; 10078) 230.028\*\*\* (df = 11; 10077)  
## ======================================================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

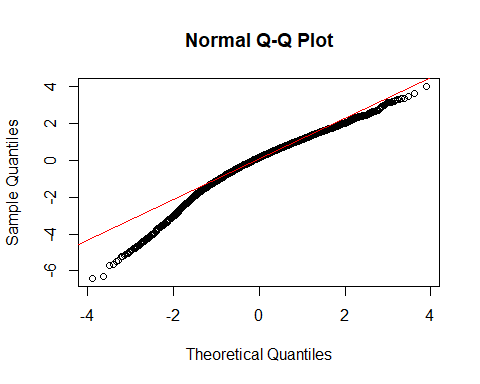
#Assumptions tests  
#Linearity  
par(mfrow = c(2, 2))  
plot(m2)



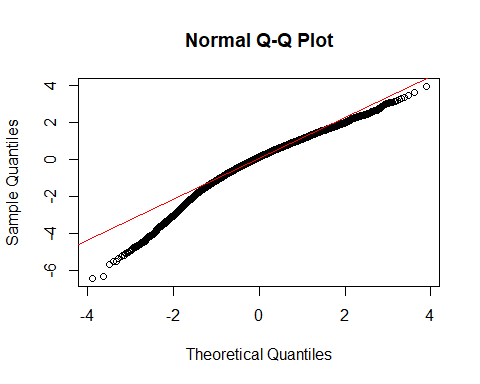
par(mfrow=c(1,1))  
  
hist(m2$fit, col = 'turquoise', main = "Fitted Plot of M2", xlab = 'Fit', ylab = 'Frequency')



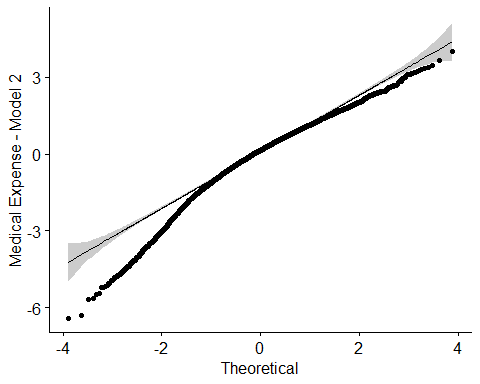
#Normality  
#Kolmogorov-Smirnov Test  
qqnorm(m2$res)  
qqline(m2$res, col = 'red')



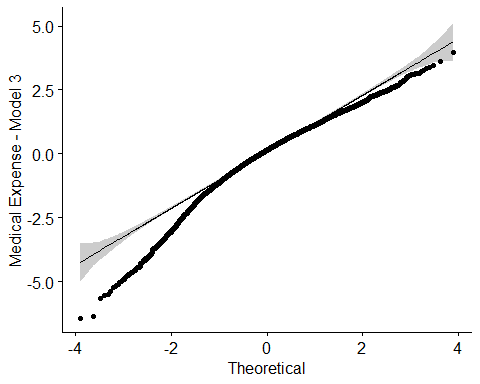
qqnorm(m3$res)  
qqline(m3$res, col = 'red')



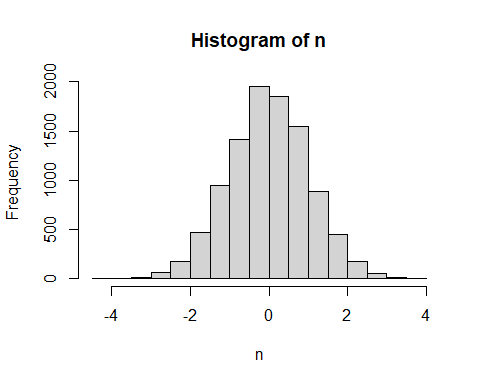
ggqqplot(m2$res, ylab = "Medical Expense - Model 2")



ggqqplot(m3$res, ylab = "Medical Expense - Model 3")



n <- rnorm(10000)  
hist(n)



ks.test(n, m2$res)

##   
## Two-sample Kolmogorov-Smirnov test  
##   
## data: n and m2$res  
## D = 0.059476, p-value = 7.772e-16  
## alternative hypothesis: two-sided

ks.test(n, m3$res)

##   
## Two-sample Kolmogorov-Smirnov test  
##   
## data: n and m3$res  
## D = 0.060772, p-value < 2.2e-16  
## alternative hypothesis: two-sided

#Homoscedasticity  
#Bartlett's Test  
bartlett.test(list(m2$res, m2$fit))

##   
## Bartlett test of homogeneity of variances  
##   
## data: list(m2$res, m2$fit)  
## Bartlett's K-squared = 4502.3, df = 1, p-value < 2.2e-16

bartlett.test(list(m3$res, m3$fit))

##   
## Bartlett test of homogeneity of variances  
##   
## data: list(m3$res, m3$fit)  
## Bartlett's K-squared = 4475.5, df = 1, p-value < 2.2e-16

#Levene's Test (System Froze)  
#leveneTest(m2$res, m2$fit, center=mean)  
#leveneTest(m3$res, m3$fit, center=mean)  
  
#Breusch-Pagan Test  
bptest(m2)

##   
## studentized Breusch-Pagan test  
##   
## data: m2  
## BP = 204.53, df = 10, p-value < 2.2e-16

bptest(m3)

##   
## studentized Breusch-Pagan test  
##   
## data: m3  
## BP = 206, df = 11, p-value < 2.2e-16

#Multicollinearity  
vif(m2)

## healthins age female blackhisp income illnesses   
## 1.087503 1.068120 1.026191 1.041426 1.333087 1.174758   
## ssiratio health msa prioritylist   
## 1.349998 1.015312 1.033319 1.160544

vif(m3)

## healthins age female blackhisp illnesses   
## 1.085958 4.973868 1.022091 1.031620 132.603830   
## ssiratio health msa prioritylist age:illnesses   
## 121.814314 1.013279 1.026594 1.163722 136.613757   
## age:ssiratio   
## 129.253930

#Autocorrelation (Independence)  
#Durbin-Watson Test  
dwtest(m2)

##   
## Durbin-Watson test  
##   
## data: m2  
## DW = 1.8036, p-value < 2.2e-16  
## alternative hypothesis: true autocorrelation is greater than 0

dwtest(m3)

##   
## Durbin-Watson test  
##   
## data: m3  
## DW = 1.802, p-value < 2.2e-16  
## alternative hypothesis: true autocorrelation is greater than 0