Team project with MLR

VishnuKant Shukla

2022-10-02

` ## Improting The Dataset and Checking Missing Values

insurance <- read.csv("C:/Users/VISHNUKANT SHUKLA/Desktop/insurance project/insurance.csv")  
miss<-read.csv("C:/Users/VISHNUKANT SHUKLA/Desktop/insurance project/insurance.csv",na.strings = c("n.a.","?","NA","n/a","na","--","<NA>"))  
sum(is.na(miss))

## [1] 0

head(insurance)

## age sex bmi children smoker region charges  
## 1 19 female 27.900 0 yes southwest 16884.924  
## 2 18 male 33.770 1 no southeast 1725.552  
## 3 28 male 33.000 3 no southeast 4449.462  
## 4 33 male 22.705 0 no northwest 21984.471  
## 5 32 male 28.880 0 no northwest 3866.855  
## 6 31 female 25.740 0 no southeast 3756.622

# This shows that there is no missing values in the data.

## Assigning Categorical Regressors Logical Values.

Data=insurance  
male=vector()   
for (i in 1:nrow(Data))   
 { if (Data$sex[i]=='male')  
 {   
 male[i]=1 }   
 else   
 { male[i]=0 }  
}  
  
Data$male = male  
  
  
smokes=vector()   
for (i in 1:nrow(Data))   
 { if (Data$smoker[i]=='yes')   
 { smokes[i]=1 }   
 else {smokes[i]=0 }  
 }  
  
Data$smokes=smokes  
  
  
  
southwest=vector()   
for (i in 1:nrow(Data))   
 { if (Data$region[i]=="southwest")  
 { southwest[i]=1 }  
 else {southwest[i]=0 }  
 }   
Data$southwest=southwest  
  
  
southeast=vector()   
for (i in 1:nrow(Data))  
 { if  
(Data$region[i]=="southeast")  
 { southeast[i]=1 }  
 else {southeast[i]=0 }  
 }   
Data$southeast=southeast  
  
  
northwest=vector()  
for (i in 1:nrow(Data))   
 { if (Data$region[i]=="northwest")  
 { northwest[i]=1 }  
 else {northwest[i]=0 }  
}  
  
Data$northwest=northwest  
  
  
head(Data)

## age sex bmi children smoker region charges male smokes southwest  
## 1 19 female 27.900 0 yes southwest 16884.924 0 1 1  
## 2 18 male 33.770 1 no southeast 1725.552 1 0 0  
## 3 28 male 33.000 3 no southeast 4449.462 1 0 0  
## 4 33 male 22.705 0 no northwest 21984.471 1 0 0  
## 5 32 male 28.880 0 no northwest 3866.855 1 0 0  
## 6 31 female 25.740 0 no southeast 3756.622 0 0 0  
## southeast northwest  
## 1 0 0  
## 2 1 0  
## 3 1 0  
## 4 0 1  
## 5 0 1  
## 6 1 0

## Deleting Columns No Longer Needed

Data1=subset(Data,select=-c(sex,smoker,region))  
#Data obtained after removing categorical columns.  
head(Data1)

## age bmi children charges male smokes southwest southeast northwest  
## 1 19 27.900 0 16884.924 0 1 1 0 0  
## 2 18 33.770 1 1725.552 1 0 0 1 0  
## 3 28 33.000 3 4449.462 1 0 0 1 0  
## 4 33 22.705 0 21984.471 1 0 0 0 1  
## 5 32 28.880 0 3866.855 1 0 0 0 1  
## 6 31 25.740 0 3756.622 0 0 0 1 0

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.1.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

Data1 <- Data1 %>% relocate(charges, .after = northwest)  
  
head(Data1)

## age bmi children male smokes southwest southeast northwest charges  
## 1 19 27.900 0 0 1 1 0 0 16884.924  
## 2 18 33.770 1 1 0 0 1 0 1725.552  
## 3 28 33.000 3 1 0 0 1 0 4449.462  
## 4 33 22.705 0 1 0 0 0 1 21984.471  
## 5 32 28.880 0 1 0 0 0 1 3866.855  
## 6 31 25.740 0 0 0 0 1 0 3756.622

## Model Fitting

library(lessR)

## Warning: package 'lessR' was built under R version 4.1.3

##   
## lessR 4.2.3 feedback: gerbing@pdx.edu   
## --------------------------------------------------------------  
## > d <- Read("") Read text, Excel, SPSS, SAS, or R data file  
## d is default data frame, data= in analysis routines optional  
##   
## Learn about reading, writing, and manipulating data, graphics,  
## testing means and proportions, regression, factor analysis,  
## customization, and descriptive statistics from pivot tables.  
## Enter: browseVignettes("lessR")  
##   
## View changes in this and recent versions of lessR.  
## Enter: news(package="lessR")  
##   
## \*\*New Feature\*\*: Interactive analysis of your data  
## Enter: interact()

##   
## Attaching package: 'lessR'

## The following objects are masked from 'package:dplyr':  
##   
## recode, rename

model\_00 <- lm(charges ~ age + male + bmi + children +  
smokes + southwest+southeast+northwest,data = Data1)  
  
model\_00

##   
## Call:  
## lm(formula = charges ~ age + male + bmi + children + smokes +   
## southwest + southeast + northwest, data = Data1)  
##   
## Coefficients:  
## (Intercept) age male bmi children smokes   
## -11938.5 256.9 -131.3 339.2 475.5 23848.5   
## southwest southeast northwest   
## -960.1 -1035.0 -353.0

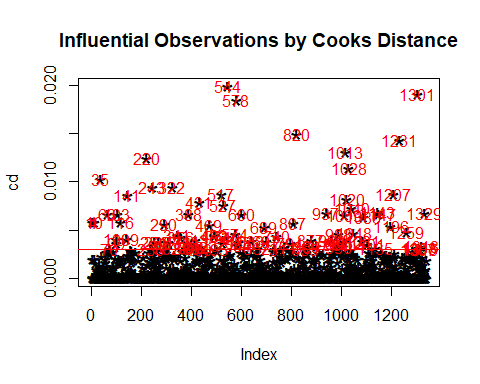
summary(model\_00)

##   
## Call:  
## lm(formula = charges ~ age + male + bmi + children + smokes +   
## southwest + southeast + northwest, data = Data1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11304.9 -2848.1 -982.1 1393.9 29992.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -11938.5 987.8 -12.086 < 0.0000000000000002 \*\*\*  
## age 256.9 11.9 21.587 < 0.0000000000000002 \*\*\*  
## male -131.3 332.9 -0.394 0.693348   
## bmi 339.2 28.6 11.860 < 0.0000000000000002 \*\*\*  
## children 475.5 137.8 3.451 0.000577 \*\*\*  
## smokes 23848.5 413.1 57.723 < 0.0000000000000002 \*\*\*  
## southwest -960.0 477.9 -2.009 0.044765 \*   
## southeast -1035.0 478.7 -2.162 0.030782 \*   
## northwest -353.0 476.3 -0.741 0.458769   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6062 on 1329 degrees of freedom  
## Multiple R-squared: 0.7509, Adjusted R-squared: 0.7494   
## F-statistic: 500.8 on 8 and 1329 DF, p-value: < 0.00000000000000022

#We found Male column and northwest column is not statistically significant.

## Detecting and Removing Influential Observations

cd=cooks.distance(model\_00)  
  
influential=as.numeric(names(cd)[(cd>4/1338)])  
  
sample\_size=1338  
plot(cd,pch='\*',cex=2,main='Influential Observations by Cooks Distance')  
abline(h=4/sample\_size,col='red')  
text(x=1:length(cd)+1,y=cd,labels=ifelse(cd>4/sample\_size,names(cd),""),col="red")



Data3 <- Data1[-influential , ]# Data obtained after removing influential data.

## Training and Testing data

set.seed(1001)  
  
n\_train <- round(0.8\* nrow(Data3))  
  
train\_indices <- sample(1:nrow(Data3), n\_train)  
  
Data2 <- Data3[train\_indices, ]   
  
Data\_test <- Data3[-train\_indices, ]  
  
Data\_train=Data2

## Fitting Model on Training Data

mq <- lm(charges ~ age + male + bmi + children + smokes + southwest+southeast+northwest, data = Data2)  
  
summary(mq)

##   
## Call:  
## lm(formula = charges ~ age + male + bmi + children + smokes +   
## southwest + southeast + northwest, data = Data2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11263.1 -1625.4 -329.5 1275.3 14737.7   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -10152.47 839.40 -12.095 < 0.0000000000000002 \*\*\*  
## age 252.55 10.04 25.143 < 0.0000000000000002 \*\*\*  
## male 107.55 282.94 0.380 0.70394   
## bmi 262.93 24.74 10.626 < 0.0000000000000002 \*\*\*  
## children 467.83 117.19 3.992 0.0000703 \*\*\*  
## smokes 24359.01 371.90 65.498 < 0.0000000000000002 \*\*\*  
## southwest -1311.01 400.75 -3.271 0.00111 \*\*   
## southeast -1054.37 407.82 -2.585 0.00987 \*\*   
## northwest -1016.04 400.21 -2.539 0.01128 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4416 on 992 degrees of freedom  
## Multiple R-squared: 0.8377, Adjusted R-squared: 0.8364   
## F-statistic: 640.1 on 8 and 992 DF, p-value: < 0.00000000000000022

# We found Male column is not statistically significant.

## Checking Assumptions on Trainig Data Model

library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

## The following objects are masked from 'package:lessR':  
##   
## bc, recode, sp

## The following object is masked from 'package:dplyr':  
##   
## recode

# Checking Multicollinearity  
  
vif(mq)

## age male bmi children smokes southwest southeast northwest   
## 1.020384 1.027167 1.125577 1.006691 1.024285 1.540657 1.657835 1.519858

## Since all VIF is less than 5 so there is no multicollinearity present in our model.  
  
# Detecting Autocorrelation  
  
durbinWatsonTest(mq,alternative = 'two.sided')

## lag Autocorrelation D-W Statistic p-value  
## 1 0.04799762 1.903661 0.124  
## Alternative hypothesis: rho != 0

## Since p-value is greater than alpha so we accept null hypothesis and this shows that there is no serial autocoreelation in error terms.  
  
# Test for normality of Residuals.  
  
library(stats)  
  
shapiro.test(rstudent(mq))

##   
## Shapiro-Wilk normality test  
##   
## data: rstudent(mq)  
## W = 0.91752, p-value < 0.00000000000000022

## Since P value< alpha so null hypothesis is rejected and assumption of normality of errors has been violated.  
  
# Check for Homoscedasticity  
  
library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

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

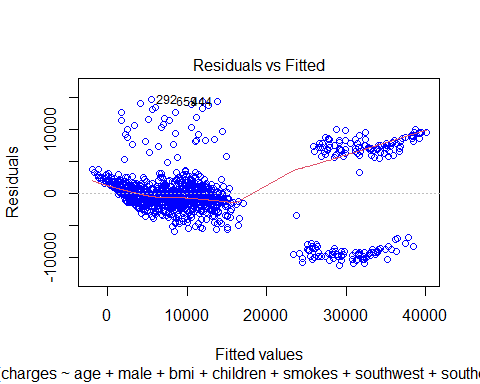
bptest(mq)

##   
## studentized Breusch-Pagan test  
##   
## data: mq  
## BP = 458.73, df = 8, p-value < 0.00000000000000022

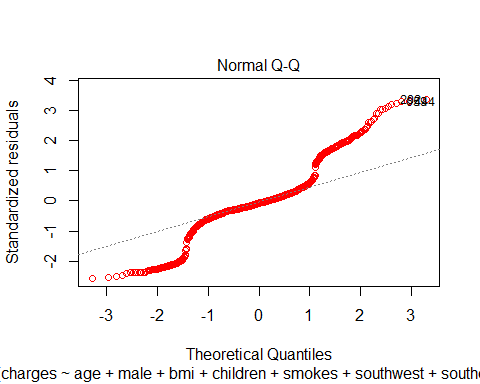
## Since P value< alpha so assumption of Homoscedasticity is violated.

## Plots on Training Data

library(ggplot2)  
 library(ggpubr)  
 library(dplyr)  
p1=plot(mq, which=1, col=c("blue")) #Plot is like outward opening funnel. This indicates variance of residual is not constant.(Why?)

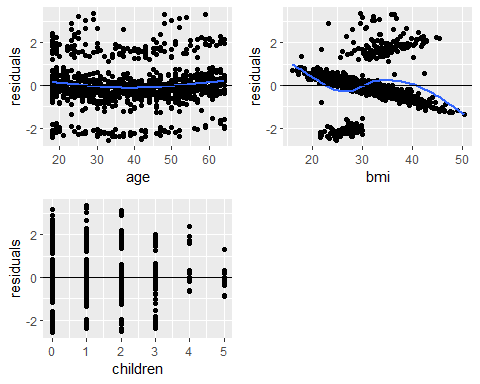


p2=plot(mq, which=2, col=c('red'))



# Checking which variable is responsible for such pattern in p1.  
## First we check for numerical variables.  
  
##Q Why we are not considering categorical variables.  
  
q1=ggplot(aes(x=Data\_train$age,y=rstudent(mq)),data=mq)+geom\_point()+geom\_smooth(se=FALSE)+geom\_hline(yintercept = 0)+labs(x='age',y='residuals')  
  
q2=ggplot(aes(x=Data\_train$bmi,y=rstudent(mq)),data=mq)+geom\_point()+geom\_smooth(se=FALSE)+geom\_hline(yintercept = 0)+labs(x='bmi',y='residuals')  
  
q3=ggplot(aes(x=Data\_train$children,y=rstudent(mq)),data=mq)+geom\_point()+geom\_smooth(se=FALSE)+geom\_hline(yintercept = 0)+labs(x='children',y='residuals')  
  
  
ggarrange(q1,q2,q3,nrow=2,ncol=2)

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'  
## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'  
## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'



# BMI vs residual plot is similar to actual fitted plot.  
  
# We plotted the graph between the charges BMI and smoker and non smoker we observe that we have two clear clusters for smokers , one cluster is for less than 30 bmi and another is for greater than 30 bmi. So we dicided to split the train data into two parts accordingly.

## Splitting of Train Dataset

Data2\_smokes=Data2[Data2$smokes==1,]  
  
nrow(Data2\_smokes)

## [1] 175

Data2\_non\_smokes=Data2[Data2$smokes==0,]  
  
nrow(Data2\_non\_smokes)

## [1] 826

Data2\_smokes\_highbmi=Data2\_smokes[Data2\_smokes$bmi>29.9,]   
  
nrow(Data2\_smokes\_highbmi)

## [1] 95

Data2\_smokes\_lowbmi=Data2\_smokes[Data2\_smokes$bmi<=29.9,]  
Data2\_smokes\_lowbmi[,9]=log((Data2\_smokes\_lowbmi$charges),base=exp(1))  
  
nrow(Data2\_smokes\_lowbmi)

## [1] 80

## Fitting models in a Splitted Data

# Smokers and High BMI  
  
m2=lm(charges~age + male + bmi + children +southwest+southeast+northwest,data=Data2\_smokes\_highbmi)  
  
summary(m2)

##   
## Call:  
## lm(formula = charges ~ age + male + bmi + children + southwest +   
## southeast + northwest, data = Data2\_smokes\_highbmi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14303.8 -295.1 290.5 1010.1 3096.0   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6429.74 2857.92 2.250 0.027 \*   
## age 246.63 19.31 12.774 < 0.0000000000000002 \*\*\*  
## male -494.57 642.30 -0.770 0.443   
## bmi 687.84 79.22 8.683 0.0000000000002 \*\*\*  
## children 287.19 252.19 1.139 0.258   
## southwest -481.14 859.87 -0.560 0.577   
## southeast -356.95 789.91 -0.452 0.652   
## northwest 1014.98 976.84 1.039 0.302   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2700 on 87 degrees of freedom  
## Multiple R-squared: 0.7514, Adjusted R-squared: 0.7314   
## F-statistic: 37.56 on 7 and 87 DF, p-value: < 0.00000000000000022

# Smokers with Low BMI  
  
m3=lm(charges~age + male + bmi + children +southwest+southeast+northwest,data=Data2\_smokes\_lowbmi)  
  
summary(m3)

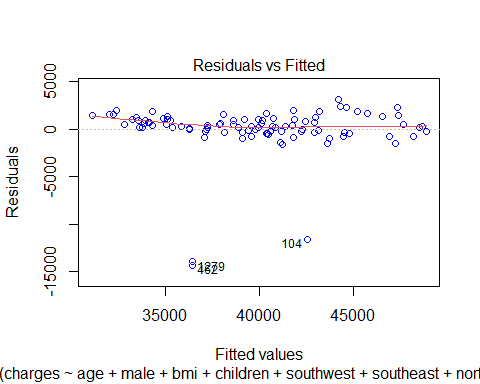
##   
## Call:  
## lm(formula = charges ~ age + male + bmi + children + southwest +   
## southeast + northwest, data = Data2\_smokes\_lowbmi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.18427 -0.06827 -0.03559 0.00567 0.63259   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.310605 0.197579 47.123 < 0.0000000000000002 \*\*\*  
## age 0.012555 0.001306 9.611 0.0000000000000154 \*\*\*  
## male 0.031801 0.037537 0.847 0.400   
## bmi 0.006701 0.007611 0.881 0.382   
## children 0.007257 0.016765 0.433 0.666   
## southwest -0.014779 0.051339 -0.288 0.774   
## southeast -0.021272 0.050826 -0.419 0.677   
## northwest -0.018756 0.046852 -0.400 0.690   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1551 on 72 degrees of freedom  
## Multiple R-squared: 0.5815, Adjusted R-squared: 0.5408   
## F-statistic: 14.29 on 7 and 72 DF, p-value: 0.00000000001711

# Non Smokers  
  
m4=lm(charges~age + male +bmi + children +southwest+southeast+northwest,data=Data2\_non\_smokes)  
  
summary(m4)

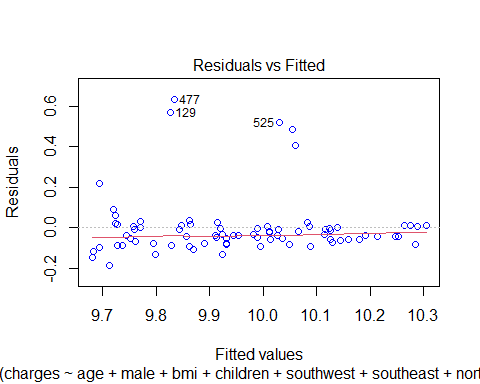
##   
## Call:  
## lm(formula = charges ~ age + male + bmi + children + southwest +   
## southeast + northwest, data = Data2\_non\_smokes)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2683.7 -987.7 -563.5 5.8 15486.0   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3740.121 532.142 -7.028 0.0000000000044 \*\*\*  
## age 258.780 6.527 39.646 < 0.0000000000000002 \*\*\*  
## male -401.517 180.891 -2.220 0.026715 \*   
## bmi 58.384 15.674 3.725 0.000209 \*\*\*  
## children 401.683 74.481 5.393 0.0000000906884 \*\*\*  
## southwest -1363.509 256.393 -5.318 0.0000001354517 \*\*\*  
## southeast -1259.862 266.156 -4.734 0.0000025993968 \*\*\*  
## northwest -901.142 255.963 -3.521 0.000454 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2591 on 818 degrees of freedom  
## Multiple R-squared: 0.6782, Adjusted R-squared: 0.6755   
## F-statistic: 246.3 on 7 and 818 DF, p-value: < 0.00000000000000022

## Plots For Checking Assumptions of AIC.

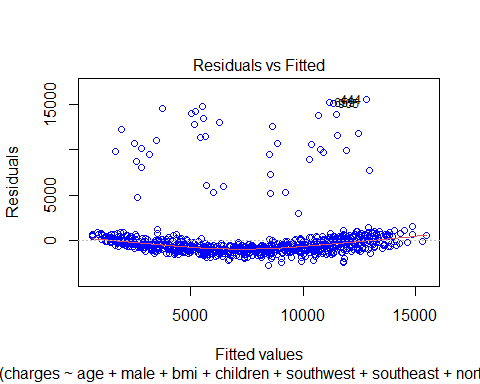
p3=plot(m2, which=1, col=c("blue"))



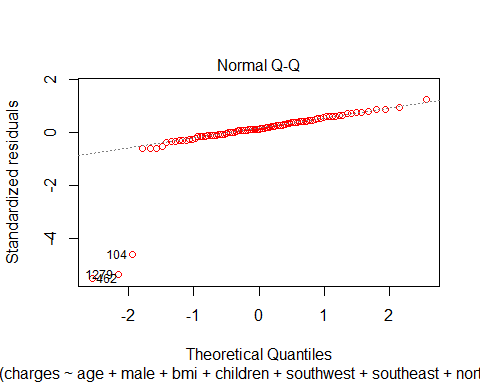
p4=plot(m3, which=1, col=c("blue"))



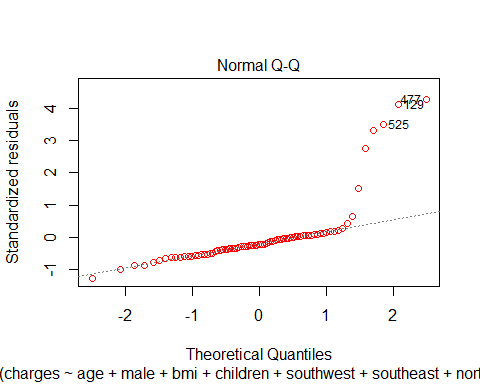
p5=plot(m4, which=1, col=c("blue"))



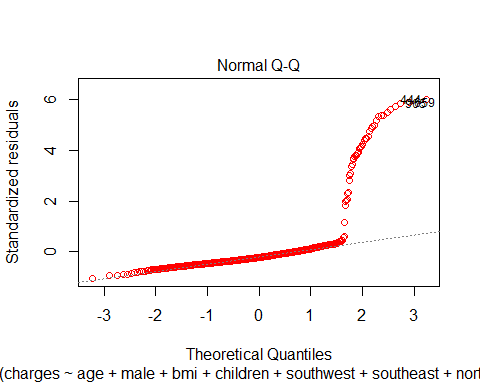
p6=plot(m2, which=2, col=c('red'))



p7=plot(m3, which=2, col=c('red'))



p8=plot(m4, which=2, col=c('red'))



## Variable Selection Using AIC Method

library(MASS)

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

stepAIC(m2,direction = 'both')

## Start: AIC=1508.87  
## charges ~ age + male + bmi + children + southwest + southeast +   
## northwest  
##   
## Df Sum of Sq RSS AIC  
## - southeast 1 1489106 635924316 1507.1  
## - southwest 1 2283164 636718374 1507.2  
## - male 1 4323591 638758801 1507.5  
## - northwest 1 7872929 642308139 1508.0  
## - children 1 9456748 643891958 1508.3  
## <none> 634435210 1508.9  
## - bmi 1 549760580 1184195790 1566.2  
## - age 1 1189979045 1824414255 1607.2  
##   
## Step: AIC=1507.09  
## charges ~ age + male + bmi + children + southwest + northwest  
##   
## Df Sum of Sq RSS AIC  
## - southwest 1 952128 636876444 1505.2  
## - male 1 3843046 639767362 1505.7  
## - children 1 9487003 645411318 1506.5  
## <none> 635924316 1507.1  
## - northwest 1 15740498 651664814 1507.4  
## + southeast 1 1489106 634435210 1508.9  
## - bmi 1 580780801 1216705117 1566.7  
## - age 1 1194646095 1830570411 1605.5  
##   
## Step: AIC=1505.23  
## charges ~ age + male + bmi + children + northwest  
##   
## Df Sum of Sq RSS AIC  
## - male 1 3844818 640721261 1503.8  
## - children 1 9559548 646435992 1504.7  
## <none> 636876444 1505.2  
## - northwest 1 18638256 655514700 1506.0  
## + southwest 1 952128 635924316 1507.1  
## + southeast 1 158070 636718374 1507.2  
## - bmi 1 595335319 1232211762 1565.9  
## - age 1 1196827541 1833703985 1603.7  
##   
## Step: AIC=1503.8  
## charges ~ age + bmi + children + northwest  
##   
## Df Sum of Sq RSS AIC  
## - children 1 9177546 649898807 1503.2  
## <none> 640721261 1503.8  
## - northwest 1 21744636 662465897 1505.0  
## + male 1 3844818 636876444 1505.2  
## + southwest 1 953899 639767362 1505.7  
## + southeast 1 51702 640669559 1505.8  
## - bmi 1 592297179 1233018441 1564.0  
## - age 1 1206477387 1847198648 1602.4  
##   
## Step: AIC=1503.15  
## charges ~ age + bmi + northwest  
##   
## Df Sum of Sq RSS AIC  
## <none> 649898807 1503.2  
## + children 1 9177546 640721261 1503.8  
## + male 1 3462815 646435992 1504.7  
## + southwest 1 1024963 648873844 1505.0  
## - northwest 1 27572771 677471578 1505.1  
## + southeast 1 50317 649848490 1505.2  
## - bmi 1 606936323 1256835129 1563.8  
## - age 1 1275151163 1925049970 1604.3

##   
## Call:  
## lm(formula = charges ~ age + bmi + northwest, data = Data2\_smokes\_highbmi)  
##   
## Coefficients:  
## (Intercept) age bmi northwest   
## 6032.5 250.1 685.0 1552.3

stepAIC(m3, direction = 'both')

## Start: AIC=-290.66  
## charges ~ age + male + bmi + children + southwest + southeast +   
## northwest  
##   
## Df Sum of Sq RSS AIC  
## - southwest 1 0.00199 1.7332 -292.57  
## - northwest 1 0.00385 1.7350 -292.48  
## - southeast 1 0.00421 1.7354 -292.46  
## - children 1 0.00451 1.7357 -292.45  
## - male 1 0.01726 1.7484 -291.87  
## - bmi 1 0.01864 1.7498 -291.80  
## <none> 1.7312 -290.66  
## - age 1 2.22107 3.9522 -226.62  
##   
## Step: AIC=-292.57  
## charges ~ age + male + bmi + children + southeast + northwest  
##   
## Df Sum of Sq RSS AIC  
## - northwest 1 0.00215 1.7353 -294.47  
## - southeast 1 0.00251 1.7357 -294.45  
## - children 1 0.00397 1.7371 -294.38  
## - bmi 1 0.01749 1.7507 -293.76  
## - male 1 0.01770 1.7509 -293.75  
## <none> 1.7332 -292.57  
## + southwest 1 0.00199 1.7312 -290.66  
## - age 1 2.25441 3.9876 -227.91  
##   
## Step: AIC=-294.47  
## charges ~ age + male + bmi + children + southeast  
##   
## Df Sum of Sq RSS AIC  
## - southeast 1 0.00126 1.7366 -296.41  
## - children 1 0.00421 1.7395 -296.27  
## - bmi 1 0.01738 1.7527 -295.67  
## - male 1 0.01968 1.7550 -295.56  
## <none> 1.7353 -294.47  
## + northwest 1 0.00215 1.7332 -292.57  
## + southwest 1 0.00029 1.7350 -292.48  
## - age 1 2.26625 4.0016 -229.63  
##   
## Step: AIC=-296.41  
## charges ~ age + male + bmi + children  
##   
## Df Sum of Sq RSS AIC  
## - children 1 0.00434 1.7409 -298.21  
## - bmi 1 0.01650 1.7531 -297.65  
## - male 1 0.02125 1.7578 -297.44  
## <none> 1.7366 -296.41  
## + southeast 1 0.00126 1.7353 -294.47  
## + northwest 1 0.00090 1.7357 -294.45  
## + southwest 1 0.00004 1.7365 -294.41  
## - age 1 2.26500 4.0016 -231.63  
##   
## Step: AIC=-298.21  
## charges ~ age + male + bmi  
##   
## Df Sum of Sq RSS AIC  
## - bmi 1 0.01659 1.7575 -299.45  
## - male 1 0.02511 1.7660 -299.06  
## <none> 1.7409 -298.21  
## + children 1 0.00434 1.7366 -296.41  
## + southeast 1 0.00138 1.7395 -296.27  
## + northwest 1 0.00100 1.7399 -296.26  
## + southwest 1 0.00000 1.7409 -296.21  
## - age 1 2.35121 4.0921 -231.84  
##   
## Step: AIC=-299.45  
## charges ~ age + male  
##   
## Df Sum of Sq RSS AIC  
## - male 1 0.04158 1.7991 -299.58  
## <none> 1.7575 -299.45  
## + bmi 1 0.01659 1.7409 -298.21  
## + children 1 0.00443 1.7531 -297.65  
## + northwest 1 0.00133 1.7562 -297.51  
## + southeast 1 0.00045 1.7571 -297.47  
## + southwest 1 0.00007 1.7574 -297.45  
## - age 1 2.37413 4.1316 -233.07  
##   
## Step: AIC=-299.58  
## charges ~ age  
##   
## Df Sum of Sq RSS AIC  
## <none> 1.7991 -299.58  
## + male 1 0.04158 1.7575 -299.45  
## + bmi 1 0.03306 1.7660 -299.06  
## + children 1 0.01003 1.7890 -298.03  
## + northwest 1 0.00313 1.7960 -297.72  
## + southeast 1 0.00169 1.7974 -297.65  
## + southwest 1 0.00089 1.7982 -297.62  
## - age 1 2.33778 4.1369 -234.97

##   
## Call:  
## lm(formula = charges ~ age, data = Data2\_smokes\_lowbmi)  
##   
## Coefficients:  
## (Intercept) age   
## 9.49768 0.01253

stepAIC(m4, direction = 'both')

## Start: AIC=12992.09  
## charges ~ age + male + bmi + children + southwest + southeast +   
## northwest  
##   
## Df Sum of Sq RSS AIC  
## <none> 5489748958 12992  
## - male 1 33065260 5522814218 12995  
## - northwest 1 83182194 5572931152 13002  
## - bmi 1 93118381 5582867339 13004  
## - southeast 1 150374160 5640123117 13012  
## - southwest 1 189803013 5679551971 13018  
## - children 1 195200188 5684949146 13019  
## - age 1 10548686924 16038435882 13876

##   
## Call:  
## lm(formula = charges ~ age + male + bmi + children + southwest +   
## southeast + northwest, data = Data2\_non\_smokes)  
##   
## Coefficients:  
## (Intercept) age male bmi children southwest   
## -3740.12 258.78 -401.52 58.38 401.68 -1363.51   
## southeast northwest   
## -1259.86 -901.14

m5= lm(formula = charges ~ age + bmi + northwest, data = Data2\_smokes\_highbmi)  
  
summary(m5)

##   
## Call:  
## lm(formula = charges ~ age + bmi + northwest, data = Data2\_smokes\_highbmi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14939.9 -313.6 302.8 1080.9 3463.1   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6032.51 2736.04 2.205 0.0300 \*   
## age 250.06 18.71 13.362 < 0.0000000000000002 \*\*\*  
## bmi 684.96 74.30 9.219 0.000000000000011 \*\*\*  
## northwest 1552.34 790.04 1.965 0.0525 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2672 on 91 degrees of freedom  
## Multiple R-squared: 0.7453, Adjusted R-squared: 0.7369   
## F-statistic: 88.78 on 3 and 91 DF, p-value: < 0.00000000000000022

m6=lm(formula = charges ~ age, data = Data2\_smokes\_lowbmi)  
  
summary(m6)

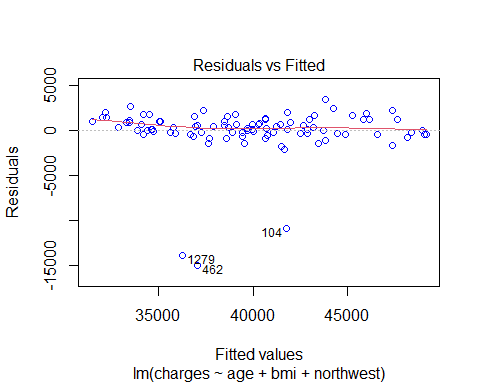
##   
## Call:  
## lm(formula = charges ~ age, data = Data2\_smokes\_lowbmi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.20013 -0.06939 -0.02597 0.01429 0.66888   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.497681 0.048720 194.94 < 0.0000000000000002 \*\*\*  
## age 0.012531 0.001245 10.07 0.000000000000000936 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1519 on 78 degrees of freedom  
## Multiple R-squared: 0.5651, Adjusted R-squared: 0.5595   
## F-statistic: 101.4 on 1 and 78 DF, p-value: 0.0000000000000009358

m7=lm(formula = charges ~ age + male + bmi + children + southwest +   
 southeast + northwest, data = Data2\_non\_smokes)  
  
summary(m7)

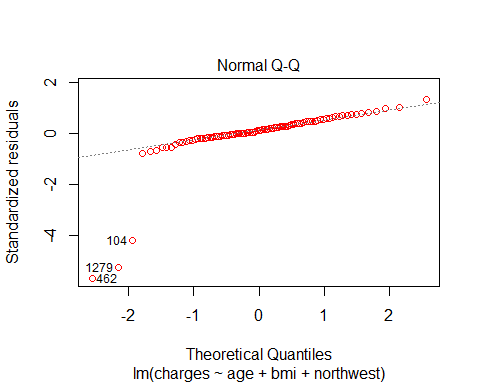
##   
## Call:  
## lm(formula = charges ~ age + male + bmi + children + southwest +   
## southeast + northwest, data = Data2\_non\_smokes)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2683.7 -987.7 -563.5 5.8 15486.0   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3740.121 532.142 -7.028 0.0000000000044 \*\*\*  
## age 258.780 6.527 39.646 < 0.0000000000000002 \*\*\*  
## male -401.517 180.891 -2.220 0.026715 \*   
## bmi 58.384 15.674 3.725 0.000209 \*\*\*  
## children 401.683 74.481 5.393 0.0000000906884 \*\*\*  
## southwest -1363.509 256.393 -5.318 0.0000001354517 \*\*\*  
## southeast -1259.862 266.156 -4.734 0.0000025993968 \*\*\*  
## northwest -901.142 255.963 -3.521 0.000454 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2591 on 818 degrees of freedom  
## Multiple R-squared: 0.6782, Adjusted R-squared: 0.6755   
## F-statistic: 246.3 on 7 and 818 DF, p-value: < 0.00000000000000022

## Plots on New Models

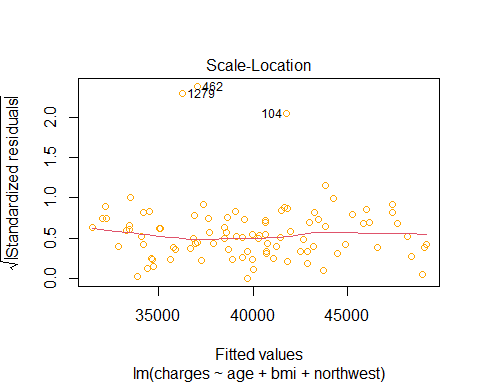
plot(m5, which=1, col=c('blue'))



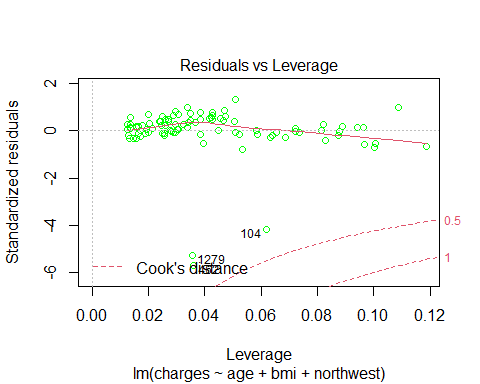
plot(m5, which=2, col=c('red'))



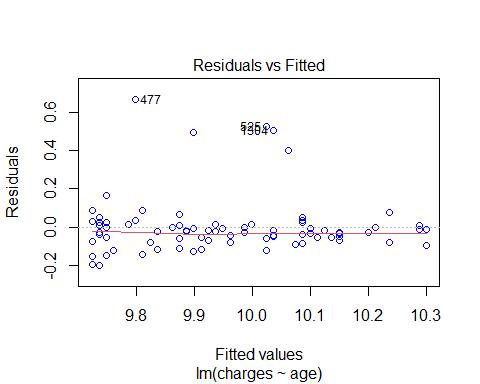
plot(m5, which=3, col=c('orange'))



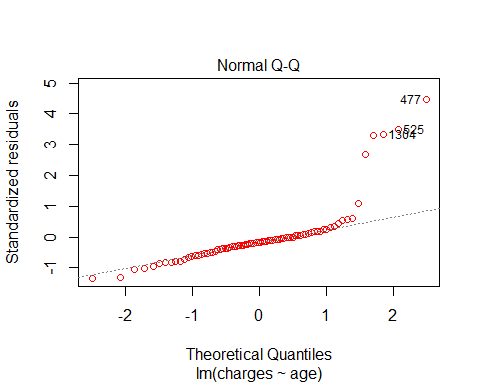
plot(m5, which=5, col=c('green'))



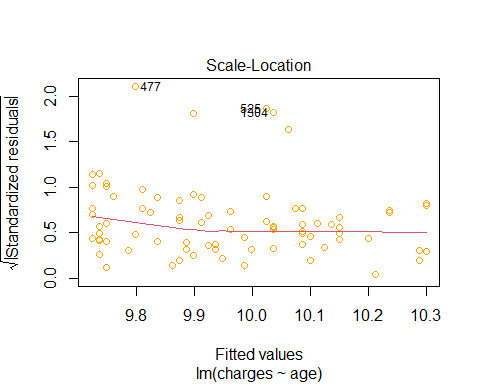
plot(m6, which=1, col=c('blue'))



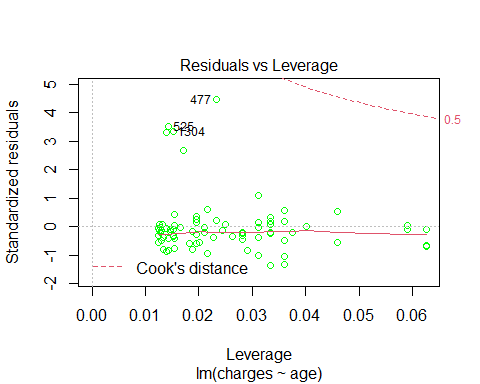
plot(m6, which=2, col=c('red'))



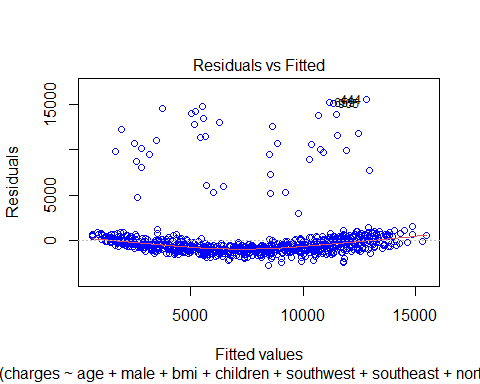
plot(m6, which=3, col=c('orange'))



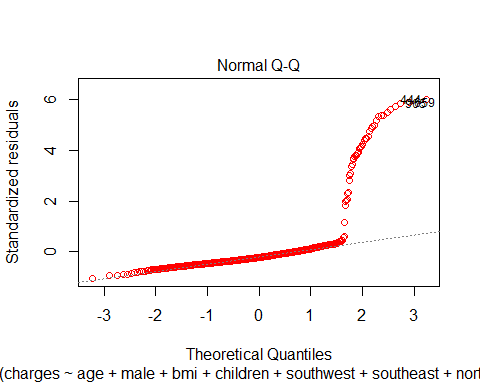
plot(m6, which=5, col=c('green'))



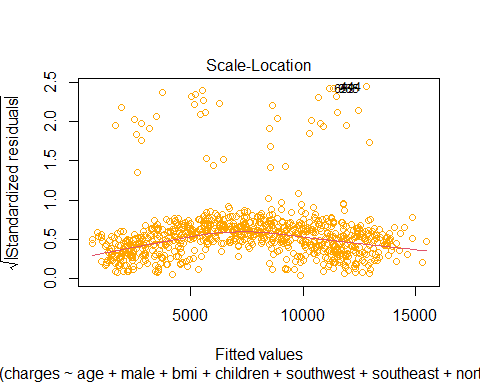
plot(m7, which=1, col=c('blue'))



plot(m7, which=2, col=c('red'))



plot(m7, which=3, col=c('orange'))



## Assumptions For Model

durbinWatsonTest(m5)

## lag Autocorrelation D-W Statistic p-value  
## 1 -0.005827226 2.007968 0.936  
## Alternative hypothesis: rho != 0

durbinWatsonTest(m6)

## lag Autocorrelation D-W Statistic p-value  
## 1 -0.04979207 2.09027 0.722  
## Alternative hypothesis: rho != 0

durbinWatsonTest(m7)

## lag Autocorrelation D-W Statistic p-value  
## 1 -0.04140807 2.082688 0.194  
## Alternative hypothesis: rho != 0

library(lmtest)  
  
bptest(m5)

##   
## studentized Breusch-Pagan test  
##   
## data: m5  
## BP = 8.2763, df = 3, p-value = 0.04063

bptest(m6)

##   
## studentized Breusch-Pagan test  
##   
## data: m6  
## BP = 0.36269, df = 1, p-value = 0.547

bptest(m7)

##   
## studentized Breusch-Pagan test  
##   
## data: m7  
## BP = 20.887, df = 7, p-value = 0.003941

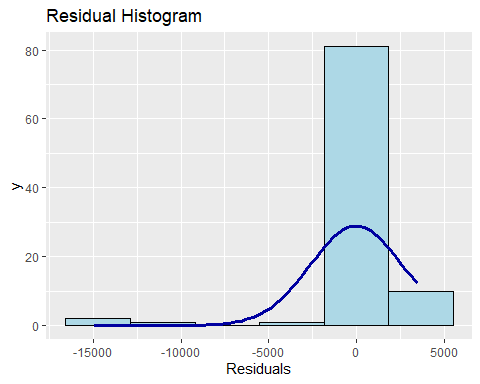
library(olsrr)

##   
## Attaching package: 'olsrr'

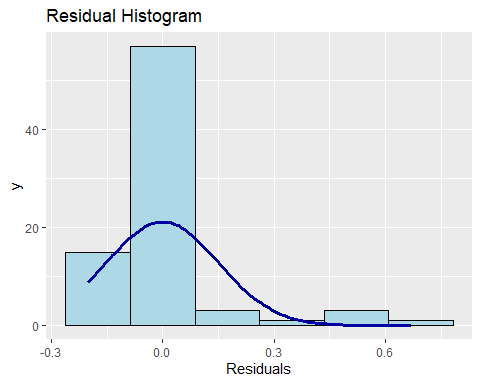
## The following object is masked from 'package:MASS':  
##   
## cement

## The following object is masked from 'package:datasets':  
##   
## rivers

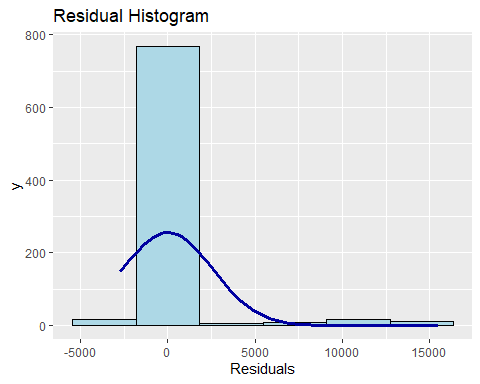
ols\_plot\_resid\_hist(m5)



ols\_plot\_resid\_hist(m6)



ols\_plot\_resid\_hist(m7)



#library(stats)   
  
#shapiro.test(rstudent(m5))  
  
#shapiro.test(rstudent(m6))  
  
#shapiro.test(rstudent(m7))

#Actual vs Fitted Plot for Test Data

Data\_smoker\_test=Data\_test[Data\_test$smokes==1,]  
  
nrow(Data\_smoker\_test)

## [1] 64

Data\_nonsmoker\_test=Data\_test[Data\_test$smokes==0,]   
  
nrow(Data\_nonsmoker\_test)

## [1] 186

Data\_smoker\_highbmi\_test=Data\_smoker\_test[Data\_smoker\_test$bmi>29.9,]   
  
nrow(Data\_smoker\_highbmi\_test)

## [1] 40

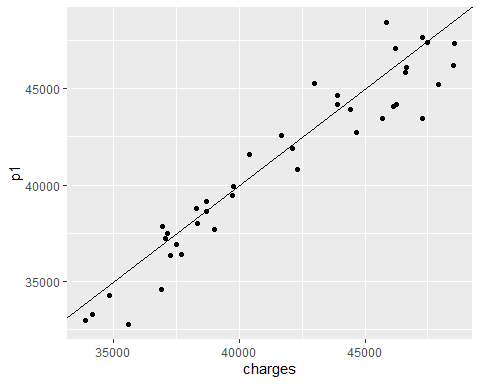
Data\_smoker\_lowbmi\_test=Data\_smoker\_test[Data\_smoker\_test$bmi<=29.9,]  
  
nrow(Data\_smoker\_lowbmi\_test)

## [1] 24

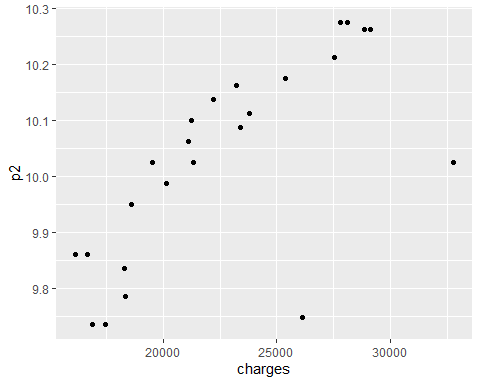
p1=predict(m5,Data\_smoker\_highbmi\_test)  
p1

## 24 30 39 50 56 95 159 204   
## 36398.55 38648.48 39902.17 39145.34 47401.16 43475.76 37871.00 37497.54   
## 257 282 315 329 331 382 422 442   
## 44623.55 47321.28 34291.93 45188.16 46208.59 40803.94 45848.98 37230.72   
## 550 630 669 675 678 737 851 853   
## 48430.34 45266.78 43465.38 47098.14 44062.09 41580.43 36367.89 41872.01   
## 884 954 1022 1032 1037 1050 1091 1097   
## 44163.46 37721.05 32781.33 43896.53 36925.34 39450.83 42574.14 42731.89   
## 1119 1121 1123 1140 1241 1289 1309 1324   
## 38771.88 33291.68 46085.78 34590.37 47667.08 38021.17 32969.86 44186.96

ggplot(data=Data\_smoker\_highbmi\_test,aes(x=charges,y=p1))+geom\_point()+geom\_abline(slope = 1, intercept = 0)



p2=predict(m6,Data\_smoker\_lowbmi\_test)  
  
ggplot(data=Data\_smoker\_lowbmi\_test,aes(x=charges,y=p2))+geom\_point()+geom\_abline(slope = 1, intercept = 0)



p3=predict(m7,Data\_nonsmoker\_test)  
  
ggplot(data=Data\_nonsmoker\_test,aes(x=charges,y=p3))+geom\_point()+geom\_abline(slope = 1, intercept = 0)

