House price prediction

Vishnukant Shukla

2022-08-06

library(MASS)   
library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

library(car)

## Loading required package: carData

library(corrplot)

## corrplot 0.92 loaded

library(olsrr)

##   
## Attaching package: 'olsrr'

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

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

library(moments)

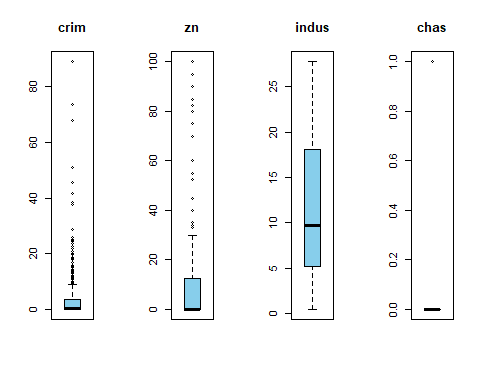
# Importing the data and checking for missing values  
  
housing = read.csv("C:/Users/VISHNUKANT SHUKLA/Desktop/Housing project/housing.csv")  
#View(housing)  
Data=data.frame(housing)  
summary(Data)

## crim zn indus chas   
## Min. : 0.00632 Min. : 0.00 Min. : 0.46 Min. :0.00000   
## 1st Qu.: 0.08205 1st Qu.: 0.00 1st Qu.: 5.19 1st Qu.:0.00000   
## Median : 0.25651 Median : 0.00 Median : 9.69 Median :0.00000   
## Mean : 3.61352 Mean : 11.36 Mean :11.14 Mean :0.06917   
## 3rd Qu.: 3.67708 3rd Qu.: 12.50 3rd Qu.:18.10 3rd Qu.:0.00000   
## Max. :88.97620 Max. :100.00 Max. :27.74 Max. :1.00000   
## nox rm age dis   
## Min. :0.3850 Min. :3.561 Min. : 2.90 Min. : 1.130   
## 1st Qu.:0.4490 1st Qu.:5.886 1st Qu.: 45.02 1st Qu.: 2.100   
## Median :0.5380 Median :6.208 Median : 77.50 Median : 3.207   
## Mean :0.5547 Mean :6.285 Mean : 68.57 Mean : 3.795   
## 3rd Qu.:0.6240 3rd Qu.:6.623 3rd Qu.: 94.08 3rd Qu.: 5.188   
## Max. :0.8710 Max. :8.780 Max. :100.00 Max. :12.127   
## rad tax ptratio b   
## Min. : 1.000 Min. :187.0 Min. :12.60 Min. : 0.32   
## 1st Qu.: 4.000 1st Qu.:279.0 1st Qu.:17.40 1st Qu.:375.38   
## Median : 5.000 Median :330.0 Median :19.05 Median :391.44   
## Mean : 9.549 Mean :408.2 Mean :18.46 Mean :356.67   
## 3rd Qu.:24.000 3rd Qu.:666.0 3rd Qu.:20.20 3rd Qu.:396.23   
## Max. :24.000 Max. :711.0 Max. :22.00 Max. :396.90   
## lstat medv   
## Min. : 1.73 Min. : 5.00   
## 1st Qu.: 6.95 1st Qu.:17.02   
## Median :11.36 Median :21.20   
## Mean :12.65 Mean :22.53   
## 3rd Qu.:16.95 3rd Qu.:25.00   
## Max. :37.97 Max. :50.00

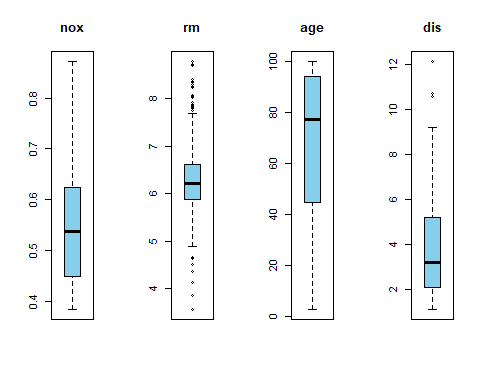
table(is.na(Data))

##   
## FALSE   
## 7084

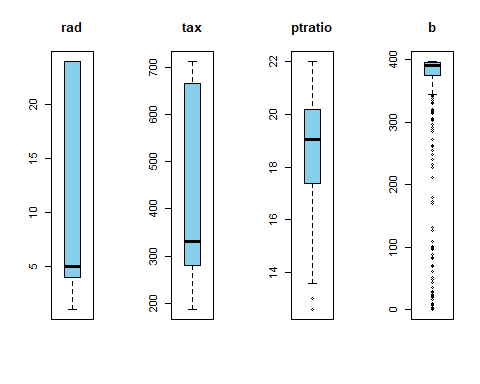
# some basic EDA  
  
par(mfrow=c(1,4))  
boxplot(Data$crim, main='crim',col='Sky Blue')  
boxplot(Data$zn, main='zn',col='Sky Blue')  
boxplot(Data$indus, main='indus',col='Sky Blue')  
boxplot(Data$chas, main='chas',col='Sky Blue')



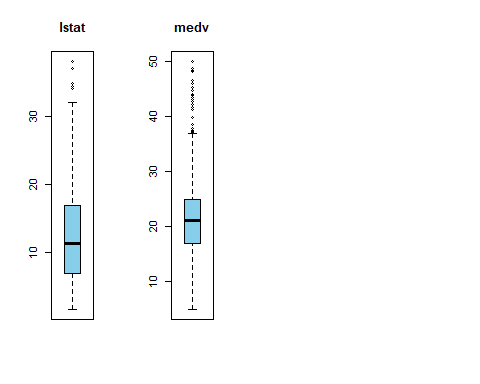
par(mfrow=c(1,4))  
boxplot(Data$nox, main='nox',col='Sky Blue')  
boxplot(Data$rm,main='rm',col='sky Blue')  
boxplot(Data$age,main='age',col='sky Blue')  
boxplot(Data$dis,main='dis',col='sky Blue')



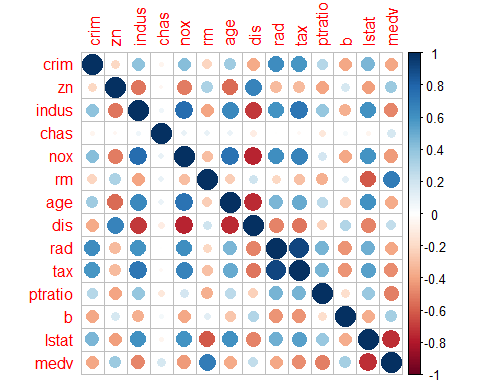
par(mfrow=c(1,4))  
boxplot(Data$rad,main='rad',col='sky Blue')  
boxplot(Data$tax,main='tax',col='sky Blue')  
boxplot(Data$ptratio,main='ptratio',col='sky Blue')  
boxplot(Data$b,main='b',col='sky Blue')



par(mfrow=c(1,4))  
boxplot(Data$lstat,main='lstat',col='sky Blue')  
boxplot(Data$medv,main='medv',col='sky Blue')  
  
par(mfrow=c(1,1))



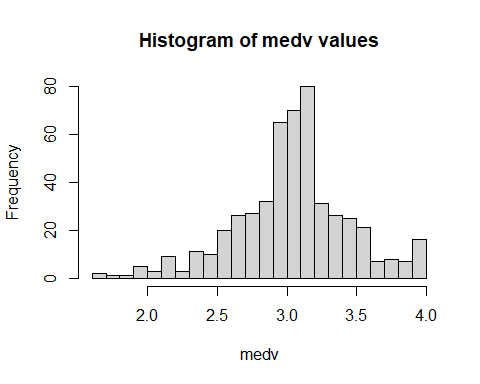
library(corrplot)  
corrplot(cor(Data))



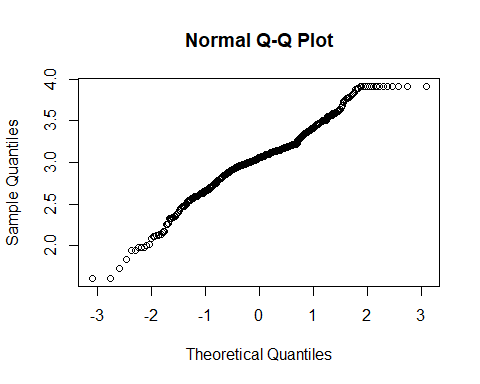
#Box-Cox Transformation for making response variable approximately normal  
attach(Data)  
y=Data$medv  
BoxCoxTrans(y)

## Box-Cox Transformation  
##   
## 506 data points used to estimate Lambda  
##   
## Input data summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 5.00 17.02 21.20 22.53 25.00 50.00   
##   
## Largest/Smallest: 10   
## Sample Skewness: 1.1   
##   
## Estimated Lambda: 0.2

#lambda=0.2 approx as 0  
y\_=log(y)  
hist(y\_,ylab = "Frequency", xlab="medv", breaks=20,main="Histogram of medv values")



qqnorm(y\_)



skewness(y\_)

## [1] -0.3293413

b1=data.frame(cbind(y\_,Data[,-14]))  
chas\_=as.factor(chas)  
model=lm(y\_~crim+zn+indus+chas\_+nox+rm+age+dis+rad+tax+ptratio+b+lstat,Data=b1)

## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :  
## extra argument 'Data' will be disregarded

summary(model)

##   
## Call:  
## lm(formula = y\_ ~ crim + zn + indus + chas\_ + nox + rm + age +   
## dis + rad + tax + ptratio + b + lstat, Data = b1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.73361 -0.09747 -0.01657 0.09629 0.86435   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.1020423 0.2042726 20.081 < 2e-16 \*\*\*  
## crim -0.0102715 0.0013155 -7.808 3.52e-14 \*\*\*  
## zn 0.0011725 0.0005495 2.134 0.033349 \*   
## indus 0.0024668 0.0024614 1.002 0.316755   
## chas\_1 0.1008876 0.0344859 2.925 0.003598 \*\*   
## nox -0.7783993 0.1528902 -5.091 5.07e-07 \*\*\*  
## rm 0.0908331 0.0167280 5.430 8.87e-08 \*\*\*  
## age 0.0002106 0.0005287 0.398 0.690567   
## dis -0.0490873 0.0079834 -6.149 1.62e-09 \*\*\*  
## rad 0.0142673 0.0026556 5.373 1.20e-07 \*\*\*  
## tax -0.0006258 0.0001505 -4.157 3.80e-05 \*\*\*  
## ptratio -0.0382715 0.0052365 -7.309 1.10e-12 \*\*\*  
## b 0.0004136 0.0001075 3.847 0.000135 \*\*\*  
## lstat -0.0290355 0.0020299 -14.304 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1899 on 492 degrees of freedom  
## Multiple R-squared: 0.7896, Adjusted R-squared: 0.7841   
## F-statistic: 142.1 on 13 and 492 DF, p-value: < 2.2e-16

#bd\_model = ols\_step\_backward\_p(model)  
#ols\_step\_backward\_p(model,0.05,details = TRUE)  
  
#plot(bd\_model) #age and indus eliminated  
model1=lm(y\_~crim+zn+chas\_+nox+rm+dis+rad+tax+ptratio+b+lstat,Data=b1)

## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :  
## extra argument 'Data' will be disregarded

summary(model1)

##   
## Call:  
## lm(formula = y\_ ~ crim + zn + chas\_ + nox + rm + dis + rad +   
## tax + ptratio + b + lstat, Data = b1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.73400 -0.09460 -0.01771 0.09782 0.86290   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.0836823 0.2030491 20.112 < 2e-16 \*\*\*  
## crim -0.0103187 0.0013134 -7.856 2.49e-14 \*\*\*  
## zn 0.0010874 0.0005418 2.007 0.045308 \*   
## chas\_1 0.1051484 0.0342285 3.072 0.002244 \*\*   
## nox -0.7217440 0.1416535 -5.095 4.97e-07 \*\*\*  
## rm 0.0906728 0.0162807 5.569 4.20e-08 \*\*\*  
## dis -0.0517059 0.0074420 -6.948 1.18e-11 \*\*\*  
## rad 0.0134457 0.0025405 5.293 1.82e-07 \*\*\*  
## tax -0.0005579 0.0001351 -4.129 4.28e-05 \*\*\*  
## ptratio -0.0374259 0.0051715 -7.237 1.77e-12 \*\*\*  
## b 0.0004127 0.0001071 3.852 0.000133 \*\*\*  
## lstat -0.0286039 0.0019002 -15.053 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1898 on 494 degrees of freedom  
## Multiple R-squared: 0.7891, Adjusted R-squared: 0.7844   
## F-statistic: 168.1 on 11 and 494 DF, p-value: < 2.2e-16

#VIF Calculation  
vif(model1) #high vif values for rad and tax

## crim zn chas\_ nox rm dis rad tax   
## 1.789704 2.239229 1.059819 3.778011 1.834806 3.443420 6.861126 7.272386   
## ptratio b lstat   
## 1.757681 1.341559 2.581984

cor(b1)

## y\_ crim zn indus chas nox  
## y\_ 1.0000000 -0.52794637 0.36334450 -0.54155616 0.158411939 -0.51060029  
## crim -0.5279464 1.00000000 -0.20046922 0.40658341 -0.055891582 0.42097171  
## zn 0.3633445 -0.20046922 1.00000000 -0.53382819 -0.042696719 -0.51660371  
## indus -0.5415562 0.40658341 -0.53382819 1.00000000 0.062938027 0.76365145  
## chas 0.1584119 -0.05589158 -0.04269672 0.06293803 1.000000000 0.09120281  
## nox -0.5106003 0.42097171 -0.51660371 0.76365145 0.091202807 1.00000000  
## rm 0.6320212 -0.21924670 0.31199059 -0.39167585 0.091251225 -0.30218819  
## age -0.4534217 0.35273425 -0.56953734 0.64477851 0.086517774 0.73147010  
## dis 0.3427803 -0.37967009 0.66440822 -0.70802699 -0.099175780 -0.76923011  
## rad -0.4819707 0.62550515 -0.31194783 0.59512927 -0.007368241 0.61144056  
## tax -0.5614657 0.58276431 -0.31456332 0.72076018 -0.035586518 0.66802320  
## ptratio -0.5017286 0.28994558 -0.39167855 0.38324756 -0.121515174 0.18893268  
## b 0.4023818 -0.38506394 0.17552032 -0.35697654 0.048788485 -0.38005064  
## lstat -0.8050341 0.45562148 -0.41299457 0.60379972 -0.053929298 0.59087892  
## rm age dis rad tax ptratio  
## y\_ 0.63202122 -0.45342171 0.34278032 -0.481970711 -0.56146566 -0.5017286  
## crim -0.21924670 0.35273425 -0.37967009 0.625505145 0.58276431 0.2899456  
## zn 0.31199059 -0.56953734 0.66440822 -0.311947826 -0.31456332 -0.3916785  
## indus -0.39167585 0.64477851 -0.70802699 0.595129275 0.72076018 0.3832476  
## chas 0.09125123 0.08651777 -0.09917578 -0.007368241 -0.03558652 -0.1215152  
## nox -0.30218819 0.73147010 -0.76923011 0.611440563 0.66802320 0.1889327  
## rm 1.00000000 -0.24026493 0.20524621 -0.209846668 -0.29204783 -0.3555015  
## age -0.24026493 1.00000000 -0.74788054 0.456022452 0.50645559 0.2615150  
## dis 0.20524621 -0.74788054 1.00000000 -0.494587930 -0.53443158 -0.2324705  
## rad -0.20984667 0.45602245 -0.49458793 1.000000000 0.91022819 0.4647412  
## tax -0.29204783 0.50645559 -0.53443158 0.910228189 1.00000000 0.4608530  
## ptratio -0.35550149 0.26151501 -0.23247054 0.464741179 0.46085304 1.0000000  
## b 0.12806864 -0.27353398 0.29151167 -0.444412816 -0.44180801 -0.1773833  
## lstat -0.61380827 0.60233853 -0.49699583 0.488676335 0.54399341 0.3740443  
## b lstat  
## y\_ 0.40238181 -0.8050341  
## crim -0.38506394 0.4556215  
## zn 0.17552032 -0.4129946  
## indus -0.35697654 0.6037997  
## chas 0.04878848 -0.0539293  
## nox -0.38005064 0.5908789  
## rm 0.12806864 -0.6138083  
## age -0.27353398 0.6023385  
## dis 0.29151167 -0.4969958  
## rad -0.44441282 0.4886763  
## tax -0.44180801 0.5439934  
## ptratio -0.17738330 0.3740443  
## b 1.00000000 -0.3660869  
## lstat -0.36608690 1.0000000

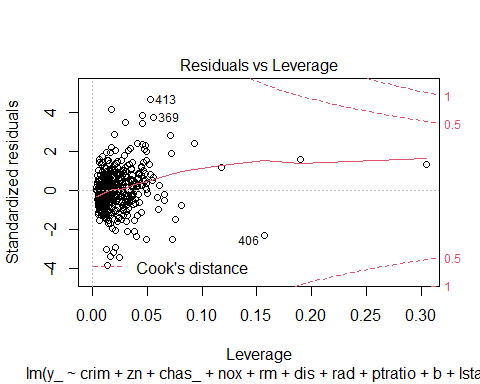
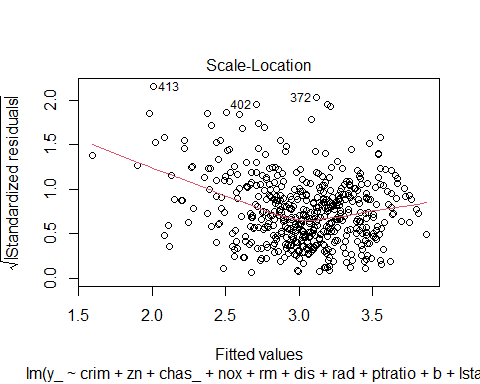
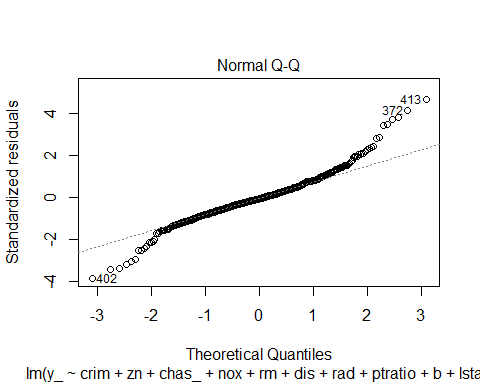
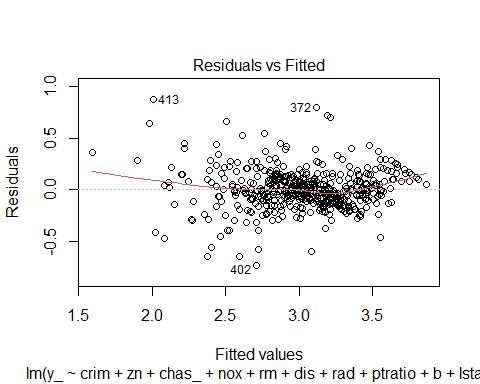
#drop tax  
  
model2=lm(y\_~crim+zn+chas\_+nox+rm+dis+rad+ptratio+b+lstat,Data=b1)

## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :  
## extra argument 'Data' will be disregarded

#summary(model2)  
vif(model2)

## crim zn chas\_ nox rm dis rad ptratio   
## 1.787963 2.154054 1.052428 3.564036 1.806735 3.410587 2.776775 1.717222   
## b lstat   
## 1.338982 2.579040

plot(model2)



summary(model2)

##   
## Call:  
## lm(formula = y\_ ~ crim + zn + chas\_ + nox + rm + dis + rad +   
## ptratio + b + lstat, Data = b1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.73447 -0.10493 -0.01084 0.09297 0.87348   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.0065321 0.2054382 19.502 < 2e-16 \*\*\*  
## crim -0.0101496 0.0013339 -7.609 1.41e-13 \*\*\*  
## zn 0.0006511 0.0005400 1.206 0.228470   
## chas\_1 0.1169498 0.0346573 3.374 0.000798 \*\*\*  
## nox -0.8609244 0.1397957 -6.158 1.52e-09 \*\*\*  
## rm 0.0989867 0.0164154 6.030 3.21e-09 \*\*\*  
## dis -0.0487057 0.0075256 -6.472 2.33e-10 \*\*\*  
## rad 0.0053533 0.0016421 3.260 0.001191 \*\*   
## ptratio -0.0406652 0.0051938 -7.830 3.00e-14 \*\*\*  
## b 0.0004321 0.0001088 3.973 8.15e-05 \*\*\*  
## lstat -0.0288688 0.0019297 -14.960 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1928 on 495 degrees of freedom  
## Multiple R-squared: 0.7819, Adjusted R-squared: 0.7775   
## F-statistic: 177.4 on 10 and 495 DF, p-value: < 2.2e-16

# To test for autocorrelation/independent among residuals.  
durbinWatsonTest(model2)

## lag Autocorrelation D-W Statistic p-value  
## 1 0.4436896 1.08952 0  
## Alternative hypothesis: rho != 0

#homoscedasticity  
#install.packages("lmtest")  
library(lmtest)

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

## Loading required package: zoo

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

##   
## Attaching package: 'zoo'

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

bptest(model2)

##   
## studentized Breusch-Pagan test  
##   
## data: model2  
## BP = 66.265, df = 10, p-value = 2.32e-10

summary(model2)

##   
## Call:  
## lm(formula = y\_ ~ crim + zn + chas\_ + nox + rm + dis + rad +   
## ptratio + b + lstat, Data = b1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.73447 -0.10493 -0.01084 0.09297 0.87348   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.0065321 0.2054382 19.502 < 2e-16 \*\*\*  
## crim -0.0101496 0.0013339 -7.609 1.41e-13 \*\*\*  
## zn 0.0006511 0.0005400 1.206 0.228470   
## chas\_1 0.1169498 0.0346573 3.374 0.000798 \*\*\*  
## nox -0.8609244 0.1397957 -6.158 1.52e-09 \*\*\*  
## rm 0.0989867 0.0164154 6.030 3.21e-09 \*\*\*  
## dis -0.0487057 0.0075256 -6.472 2.33e-10 \*\*\*  
## rad 0.0053533 0.0016421 3.260 0.001191 \*\*   
## ptratio -0.0406652 0.0051938 -7.830 3.00e-14 \*\*\*  
## b 0.0004321 0.0001088 3.973 8.15e-05 \*\*\*  
## lstat -0.0288688 0.0019297 -14.960 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1928 on 495 degrees of freedom  
## Multiple R-squared: 0.7819, Adjusted R-squared: 0.7775   
## F-statistic: 177.4 on 10 and 495 DF, p-value: < 2.2e-16

confint(model2)

## 2.5 % 97.5 %  
## (Intercept) 3.6028937297 4.4101704225  
## crim -0.0127704270 -0.0075287937  
## zn -0.0004098262 0.0017120394  
## chas\_1 0.0488561821 0.1850434309  
## nox -1.1355904980 -0.5862583348  
## rm 0.0667343215 0.1312391613  
## dis -0.0634916642 -0.0339197289  
## rad 0.0021268657 0.0085797323  
## ptratio -0.0508699414 -0.0304605482  
## b 0.0002184193 0.0006457901  
## lstat -0.0326602254 -0.0250774181