**Boston assignment**

Jitesh Rawat

20237

2/18/2021

**library**(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3 v purrr 0.3.4  
## v tibble 3.0.4 v dplyr 1.0.2  
## v tidyr 1.1.2 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

**library**(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

**library**(MVN)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

## sROC 0.1-2 loaded

**library**(MASS)

##   
## Attaching package: 'MASS'

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

bosdata <- tibble**::as\_tibble**(Boston)  
**head**(bosdata)

## # A tibble: 6 x 14  
## crim zn indus chas nox rm age dis rad tax ptratio black  
## <dbl> <dbl> <dbl> <int> <dbl> <dbl> <dbl> <dbl> <int> <dbl> <dbl> <dbl>  
## 1 0.00632 18 2.31 0 0.538 6.58 65.2 4.09 1 296 15.3 397.  
## 2 0.0273 0 7.07 0 0.469 6.42 78.9 4.97 2 242 17.8 397.  
## 3 0.0273 0 7.07 0 0.469 7.18 61.1 4.97 2 242 17.8 393.  
## 4 0.0324 0 2.18 0 0.458 7.00 45.8 6.06 3 222 18.7 395.  
## 5 0.0690 0 2.18 0 0.458 7.15 54.2 6.06 3 222 18.7 397.  
## 6 0.0298 0 2.18 0 0.458 6.43 58.7 6.06 3 222 18.7 394.  
## # ... with 2 more variables: lstat <dbl>, medv <dbl>

#summary(bosdata)

**featurePlot**(x= bosdata[],y= bosdata**$**crim)



*#from this feature plot we can determine correlation of crime with all other variables*  
*#from these plots we can see that areas with less median vakue of owner occupied homes have more crime rates*  
*#Also the areas with less weighted distances to five Boston employment centers have more number of crime rates*  
*#Areas with high index of accessibility to radial highways have high crime rates*  
*#Higher tax and rad value areas have higher crime rates whereas the lower ones have almost near to 0 crime rates*

**library**(corrplot)

## corrplot 0.84 loaded

**corrplot**(**cor**(bosdata), type = "upper", tl.col = "black")



*#From the below plot 1st row itself we can see correlation of crim with other variables*   
*#the bigger and darker circle depicts strong relattion*

**library**(funModeling)

## Loading required package: Hmisc

## Loading required package: survival

##   
## Attaching package: 'survival'

## The following object is masked from 'package:caret':  
##   
## cluster

## Loading required package: Formula

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':  
##   
## src, summarize

## The following objects are masked from 'package:base':  
##   
## format.pval, units

## funModeling v.1.9.4 :)  
## Examples and tutorials at livebook.datascienceheroes.com  
## / Now in Spanish: librovivodecienciadedatos.ai

**plot\_num**(bosdata)



*#from this plot we can see that some of the variables have many outliers and are highly undstable*

*#Lets create the first model with simply putting all variables against crime variable*  
model1= **lm**(crim**~** ., data = bosdata)  
**par**(mfrow=**c**(2,2),mar=**c**(4,4,2,0.5))  
**plot**(model1)



**summary**(model1)

##   
## Call:  
## lm(formula = crim ~ ., data = bosdata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.924 -2.120 -0.353 1.019 75.051   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 17.033228 7.234903 2.354 0.018949 \*   
## zn 0.044855 0.018734 2.394 0.017025 \*   
## indus -0.063855 0.083407 -0.766 0.444294   
## chas -0.749134 1.180147 -0.635 0.525867   
## nox -10.313535 5.275536 -1.955 0.051152 .   
## rm 0.430131 0.612830 0.702 0.483089   
## age 0.001452 0.017925 0.081 0.935488   
## dis -0.987176 0.281817 -3.503 0.000502 \*\*\*  
## rad 0.588209 0.088049 6.680 6.46e-11 \*\*\*  
## tax -0.003780 0.005156 -0.733 0.463793   
## ptratio -0.271081 0.186450 -1.454 0.146611   
## black -0.007538 0.003673 -2.052 0.040702 \*   
## lstat 0.126211 0.075725 1.667 0.096208 .   
## medv -0.198887 0.060516 -3.287 0.001087 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.439 on 492 degrees of freedom  
## Multiple R-squared: 0.454, Adjusted R-squared: 0.4396   
## F-statistic: 31.47 on 13 and 492 DF, p-value: < 2.2e-16

*#From this model it is certain that model does not follow normality*  
*#some variables like dis rad... seem to be significant*

*#Creating a model with all significant variables from last model summary*  
model2 <- **lm**(crim**~**zn**+**indus**+**dis**\***rad**+**black**+**medv**+**nox, data = bosdata)  
**par**(mfrow=**c**(2,2),mar=**c**(4,4,2,0.5))  
**plot**(model2)



**summary**(model2)

##   
## Call:  
## lm(formula = crim ~ zn + indus + dis \* rad + black + medv + nox,   
## data = bosdata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.620 -1.540 -0.392 0.807 72.065   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.44468 3.64495 2.317 0.0209 \*   
## zn 0.01477 0.01713 0.862 0.3891   
## indus -0.04271 0.06997 -0.610 0.5419   
## dis 0.66096 0.33998 1.944 0.0524 .   
## rad 1.13264 0.09174 12.346 < 2e-16 \*\*\*  
## black -0.00765 0.00342 -2.237 0.0257 \*   
## medv -0.16313 0.03750 -4.351 1.65e-05 \*\*\*  
## nox -7.10015 4.45778 -1.593 0.1119   
## dis:rad -0.27602 0.03621 -7.622 1.28e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.104 on 497 degrees of freedom  
## Multiple R-squared: 0.5043, Adjusted R-squared: 0.4964   
## F-statistic: 63.21 on 8 and 497 DF, p-value: < 2.2e-16

*#Still model is unstable i.e not holding normality*  
*#Certainly this model is not giving best Rsquared value*

*#Trial 3 trying to find good fit with strongly correlated variables*  
model3 <- **lm**(crim**~**dis**\***rad**+**indus**\***medv, data = bosdata)  
**summary**(model3)

##   
## Call:  
## lm(formula = crim ~ dis \* rad + indus \* medv, data = bosdata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.288 -1.332 -0.162 0.519 70.765   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.201933 2.325520 -1.807 0.071383 .   
## dis 1.057736 0.263394 4.016 6.84e-05 \*\*\*  
## rad 1.140970 0.085357 13.367 < 2e-16 \*\*\*  
## indus 0.326410 0.127286 2.564 0.010628 \*   
## medv 0.020866 0.059547 0.350 0.726174   
## dis:rad -0.281215 0.034403 -8.174 2.48e-15 \*\*\*  
## indus:medv -0.016116 0.004394 -3.668 0.000271 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.059 on 499 degrees of freedom  
## Multiple R-squared: 0.5097, Adjusted R-squared: 0.5038   
## F-statistic: 86.45 on 6 and 499 DF, p-value: < 2.2e-16

model4 <- **lm**(crim**~**dis**\***rad**+**medv**\***rad, data = bosdata)  
**summary**(model4)

##   
## Call:  
## lm(formula = crim ~ dis \* rad + medv \* rad, data = bosdata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.250 -1.008 -0.186 0.600 69.793   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5.088684 1.388706 -3.664 0.000274 \*\*\*  
## dis 0.939637 0.221286 4.246 2.59e-05 \*\*\*  
## rad 1.529388 0.099660 15.346 < 2e-16 \*\*\*  
## medv 0.084046 0.047016 1.788 0.074444 .   
## dis:rad -0.261031 0.033552 -7.780 4.19e-14 \*\*\*  
## rad:medv -0.023094 0.003597 -6.421 3.16e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.903 on 500 degrees of freedom  
## Multiple R-squared: 0.5337, Adjusted R-squared: 0.529   
## F-statistic: 114.5 on 5 and 500 DF, p-value: < 2.2e-16

*#This model has shown better performance than previous ones*  
*#we should check its normality for residuals*

**shapiro.test**(model4**$**residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: model4$residuals  
## W = 0.4787, p-value < 2.2e-16

*#From the test of normality we see that data is not normal as p-value is so close to 0*

*#Thats why we apply boxcox to normalise it*  
bc<-**boxcox**(model4, data=bosdata)



lambda <- bc**$**x[**which.max**(bc**$**y)]  
lambda

## [1] 0.06060606

bosdata**$**y <- ((bosdata**$**crim)**^**lambda-1**/**lambda)  
new\_model <- **lm**(y**~**dis**\***rad**+**medv**\***rad, data= bosdata)  
**summary**(new\_model)

##   
## Call:  
## lm(formula = y ~ dis \* rad + medv \* rad, data = bosdata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.137233 -0.037881 -0.004526 0.031691 0.146492   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.553e+01 1.198e-02 -1296.260 < 2e-16 \*\*\*  
## dis -1.606e-02 1.909e-03 -8.410 4.32e-16 \*\*\*  
## rad 1.145e-02 8.598e-04 13.320 < 2e-16 \*\*\*  
## medv -1.647e-03 4.056e-04 -4.059 5.71e-05 \*\*\*  
## dis:rad -6.054e-04 2.895e-04 -2.091 0.037 \*   
## rad:medv -1.176e-05 3.103e-05 -0.379 0.705   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.05093 on 500 degrees of freedom  
## Multiple R-squared: 0.8463, Adjusted R-squared: 0.8447   
## F-statistic: 550.5 on 5 and 500 DF, p-value: < 2.2e-16

pred\_data <- **sample\_n**(bosdata,10)  
**predict**(model4,pred\_data)

*#Trying to predict some random data*

## 1 2 3 4 5 6 7   
## 14.7876748 0.7070797 1.1670688 -4.5856790 0.3826902 -0.6896931 0.2865019   
## 8 9 10   
## 1.1706184 0.7871443 0.7385331

pred\_data**$**crim

## [1] 15.02340 0.25915 0.10084 0.08221 0.13587 0.31533 0.09164 0.35809  
## [9] 0.18337 0.05602