Practice assignment Solution
Linear Regression in R

```
#Boston Pricing case study

setwd("C:\\")

##Loading Data
prices<-read.csv("boston_prices.csv",header=TRUE,stringsAsFactors=FALSE)

##Checking Data Characteristics
dim(prices)
## [1] 506  14
str(prices)
## 'data.frame':    506 obs. of  14 variables:
##  $ CRIM                         : num  0.00632 0.02731 0.02729 0.03237 0.06
905 ...
##  $ ZN                           : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ INDUS                        : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7
.87 7.87 7.87 ...
##  $ Charles.River.dummy.variable: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ nitric.oxides.concentration : num  0.538 0.469 0.469 0.458 0.458 0.458
0.524 0.524 0.524 0.524 ...
##  $ X.rooms.dwelling             : num  6.58 6.42 7.18 7 7.15 ...
##  $ AGE                          : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 9
6.1 100 85.9 ...
##  $ DIS                          : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ RAD                          : int  1 2 2 3 3 3 5 5 5 5 ...
##  $ TAX                          : num  296 242 242 222 222 222 311 311 311
311 ...
##  $ PTRATIO                      : num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 1
5.2 15.2 15.2 ...
##  $ B                            : num  397 397 393 395 397 ...
##  $ LSTAT                        : num  4.98 9.14 4.03 2.94 5.33 ...
##  $ MEDV                         : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.
1 16.5 18.9 ...
```

```r
head(prices)
```

```
##      CRIM ZN INDUS Charles.River.dummy.variable
## 1 0.00632 18  2.31                            0
## 2 0.02731  0  7.07                            0
## 3 0.02729  0  7.07                            0
## 4 0.03237  0  2.18                            0
## 5 0.06905  0  2.18                            0
## 6 0.02985  0  2.18                            0
##   nitric.oxides.concentration X.rooms.dwelling  AGE    DIS RAD TAX PTRATIO
## 1                       0.538            6.575 65.2 4.0900   1 296    15.3
## 2                       0.469            6.421 78.9 4.9671   2 242    17.8
## 3                       0.469            7.185 61.1 4.9671   2 242    17.8
## 4                       0.458            6.998 45.8 6.0622   3 222    18.7
## 5                       0.458            7.147 54.2 6.0622   3 222    18.7
## 6                       0.458            6.430 58.7 6.0622   3 222    18.7
##        B LSTAT MEDV
## 1 396.90  4.98 24.0
## 2 396.90  9.14 21.6
## 3 392.83  4.03 34.7
## 4 394.63  2.94 33.4
## 5 396.90  5.33 36.2
## 6 394.12  5.21 28.7
```

```r
names(prices)
```

```
##  [1] "CRIM"                         "ZN"
##  [3] "INDUS"                        "Charles.River.dummy.variable"
##  [5] "nitric.oxides.concentration"  "X.rooms.dwelling"
##  [7] "AGE"                          "DIS"
##  [9] "RAD"                          "TAX"
## [11] "PTRATIO"                      "B"
## [13] "LSTAT"                        "MEDV"
```

```r
#summary statistics
summary(prices)
```

```
##       CRIM                ZN             INDUS
##  Min.   :0.00000   Min.   :  0.0   Min.   : 0.000
```

```
##  1st Qu.:0.04944   1st Qu.:  0.0   1st Qu.: 3.440
##  Median :0.14466   Median :  0.0   Median : 6.960
##  Mean   :1.26920   Mean   : 13.3   Mean   : 9.205
##  3rd Qu.:0.81962   3rd Qu.: 18.1   3rd Qu.:18.100
##  Max.   :9.96654   Max.   :100.0   Max.   :27.740
##
##  Charles.River.dummy.variable nitric.oxides.concentration
##  Min.   :0.0000               Min.   :0.385
##  1st Qu.:0.0000               1st Qu.:0.449
##  Median :0.0000               Median :0.538
##  Mean   :0.1408               Mean   :1.101
##  3rd Qu.:0.0000               3rd Qu.:0.647
##  Max.   :1.0000               Max.   :7.313
##
##  X.rooms.dwelling       AGE               DIS              RAD
##  Min.   :  3.561   Min.   :  1.137   Min.   : 1.130   Min.   :  1.00
##  1st Qu.:  5.962   1st Qu.: 32.000   1st Qu.: 2.431   1st Qu.:  4.00
##  Median :  6.322   Median : 65.250   Median : 3.926   Median :  5.00
##  Mean   : 15.680   Mean   : 58.745   Mean   : 6.173   Mean   : 78.06
##  3rd Qu.:  6.949   3rd Qu.: 89.975   3rd Qu.: 6.332   3rd Qu.: 24.00
##  Max.   :100.000   Max.   :100.000   Max.   :24.000   Max.   :666.00
##
##       TAX           PTRATIO              B              LSTAT
##  Min.   : 20.2   Min.   :  2.60   Min.   :  0.32   Min.   : 1.730
##  1st Qu.:254.0   1st Qu.: 17.00   1st Qu.:365.00   1st Qu.: 6.878
##  Median :307.0   Median : 18.90   Median :390.66   Median :10.380
##  Mean   :339.3   Mean   : 42.62   Mean   :332.79   Mean   :11.538
##  3rd Qu.:403.0   3rd Qu.: 20.20   3rd Qu.:395.62   3rd Qu.:15.015
##  Max.   :711.0   Max.   :396.90   Max.   :396.90   Max.   :34.410
##
##       MEDV
##  Min.   : 6.30
##  1st Qu.:18.50
##  Median :21.95
```

```
##   Mean   :23.75
##   3rd Qu.:26.60
##   Max.   :50.00
##   NA's   :54
```

```r
#Missing values treatment
colSums(is.na(prices)) #MEDV has a lot of missing values
```

```
##                           CRIM                             ZN
##                              0                              0
##                          INDUS Charles.River.dummy.variable
##                              0                              0
##   nitric.oxides.concentration              X.rooms.dwelling
##                              0                              0
##                            AGE                            DIS
##                              0                              0
##                            RAD                            TAX
##                              0                              0
##                        PTRATIO                              B
##                              0                              0
##                          LSTAT                           MEDV
##                              0                             54
```

```r
summary((prices$MEDV))
```

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     6.30   18.50   21.95   23.75   26.60   50.00      54
```

```r
prices$MEDV[is.na(prices$MEDV)]<-mean(prices$MEDV,na.rm=TRUE)


#Outlier plots
par(mfrow=c(2,7)) #This allows you to plot 14 charts on a single page; It is optional.
list<-names(prices) #Store the names of the dataset in a list format
list<-list[-4]
for(i in 1:length(list)) #Plot the boxplots of all variables and shortlist which ones need outlier treatment.
{
  boxplot(prices[,list[i]],main=list[i])
}
```

```
#Restore the par parameters to normal
dev.off()
## null device
##           1
#In this solution, We have replaced the outlier values by the median values
#You can decide to replace by max or mean values based on business objectives

#Outlier treatment
for(i in 1:length(list)) ##For loop to replace all the outlier values with th
e mean value ; if you want you can replace with median value as well.
{
    x<-boxplot(prices[,list[i]])
    out<-x$out
    index<-which(prices[,list[i]] %in% x$out)
    prices[index,list[i]]<-mean(prices[,list[i]])
    rm(x)
    rm(out)
}


#Exploratory analysis
library(ggplot2)
## Warning: package 'ggplot2' was built under R version 3.2.2
#Study the histogram of the DV and the transformed histogram
hist(prices$MEDV)
#hist(prices$log_MEDV) #Once you create the transformations;look down

#You can look at the correlation between each IDV and the DV
#An eg :
ggplot(prices,aes(x=MEDV,y=LSTAT)) +geom_point()
ggplot(prices,aes(x=MEDV,y=DIS)) +geom_point()
ggplot(prices,aes(x=MEDV,y=AGE)) +geom_point()

#Inorder to quicken the process, lets write a function :
```

```r
#Below is a function that gives you the correlation values between all IDV's and the DV
#Simply taking a look at the output of this function, you can quickly shortlist
#Which all IDV's are correlated to the DV


#Function to get the list of correlations between : DV and the IDV's
list1<-list[-13]
for(i in 1:length(list1))
{
  x<-cor(prices$MEDV,prices[list[i]])
  print(x)
}
```

```
##            CRIM
## [1,] -0.4622118
##             ZN
## [1,] 0.4172775
##           INDUS
## [1,] -0.4981729
##      nitric.oxides.concentration
## [1,]                  -0.1336743
##      X.rooms.dwelling
## [1,]        0.2442593
##             AGE
## [1,] -0.4615408
##            DIS
## [1,] 0.3322806
##             RAD
## [1,] 0.03218742
##             TAX
## [1,] -0.2989045
##         PTRATIO
## [1,] 0.01971893
##              B
## [1,] 0.1439405
```

```
##            LSTAT
## [1,] -0.6535546
```

```r
#Significant variables are : B LSTAT AGE X.rooms.dwelling nitric.oxides.conce
ntration INDUS


#You can also try to use data transformations


#Log transformations
#Create the log transformation for all variables
prices$log_CRIM<-log(prices$CRIM)
prices$log_ZN<-log(prices$ZN)
prices$log_NOX<-log(prices$nitric.oxides.concentration)
prices$log_RM<-log(prices$X.rooms.dwelling)
prices$log_AGE<-log(prices$AGE)
prices$log_DIS<-log(prices$DIS)
prices$log_RAD<-log(prices$RAD)
prices$log_TAX<-log(prices$TAX)
prices$log_PTRATIO<-log(prices$PTRATIO)
prices$log_B<-log(prices$B)
prices$log_LSTAT<-log(prices$LSTAT)
prices$log_MEDV<-log(prices$MEDV) #DV
prices$log_INDUS<-log(prices$INDUS)



#Refer to the profiling excel sheet to see all the correlations documented


#Function to get the list of correlations between : log_DV and log of IDV's


list_log<-names(prices)[c(15:25,27)]
for(i in 1:length(list_log))
{
  xlog<-cor(prices$log_MEDV,prices[list_log[i]])
  print(xlog)
}
```

```
##       log_CRIM
```

```
## [1,]      NaN
##       log_ZN
## [1,]    NaN
##         log_NOX
## [1,] -0.193495
##         log_RM
## [1,] 0.316107
##          log_AGE
## [1,] -0.3442683
##         log_DIS
## [1,] 0.3981884
##           log_RAD
## [1,] -0.08203473
##          log_TAX
## [1,] -0.2861208
##       log_PTRATIO
## [1,] -0.01558666
##         log_B
## [1,] 0.14549
##        log_LSTAT
## [1,] -0.6326763
##       log_INDUS
## [1,]       NaN
```

```r
#Function to get the list of correlations between : log_DV and IDV's

list_log_DV<-names(prices)[1:13]
list_log_DV<-list_log_DV[-4]
for(i in 1:length(list_log_DV))
{
  xlogdv<-cor(prices$log_MEDV,prices[list_log_DV[i]])
  print(xlogdv)
}
```

```
##              CRIM
## [1,] -0.4942302
```

```
##             ZN
## [1,] 0.4082063
##           INDUS
## [1,] -0.5102838
##      nitric.oxides.concentration
## [1,]                  -0.1237709
##      X.rooms.dwelling
## [1,]        0.2651325
##             AGE
## [1,] -0.4876028
##            DIS
## [1,] 0.3526956
##             RAD
## [1,] 0.04812929
##             TAX
## [1,] -0.2916202
##         PTRATIO
## [1,] 0.04396186
##              B
## [1,] 0.1444425
##           LSTAT
## [1,] -0.6669138
```

```r
sampling<-sort(sample(nrow(prices), nrow(prices)*.7))

#Select training sample
train<-prices[sampling,]
test<-prices[-sampling,]

##Building SimpLe Linear Regression Model

#Metrics :
#Rsquare
#Coefficients
#P values : Significance levels of the IDV's
```

```
#Residuals distribution


#Factor variables as IDV's

#All good modelssummm

Reg<-lm(log_MEDV~CRIM+INDUS+RAD+TAX+B+
                 Charles.River.dummy.variable+
                  DIS+ZN+PTRATIO+LSTAT+AGE+X.rooms.dwelling+nitric.oxides.con
centration,data=train)


summary(Reg)
```

```
##
## Call:
## lm(formula = log_MEDV ~ CRIM + INDUS + RAD + TAX + B + Charles.River.dummy
.variable +
##     DIS + ZN + PTRATIO + LSTAT + AGE + X.rooms.dwelling + nitric.oxides.co
ncentration,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36319 -0.09713 -0.01709  0.09100  0.46697
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   3.6817197  0.2363575  15.577  < 2e-16 ***
## CRIM                         -0.0909268  0.0280985  -3.236 0.001331 **
## INDUS                        -0.0015383  0.0020858  -0.738 0.461319
## RAD                           0.0007643  0.0014872   0.514 0.607650
## TAX                          -0.0002513  0.0001576  -1.595 0.111712
## B                            -0.0001346  0.0004983  -0.270 0.787217
## Charles.River.dummy.variable  0.0688731  0.0313088   2.200 0.028493 *
## DIS                          -0.0272362  0.0070547  -3.861 0.000135 ***
## ZN                            0.0011734  0.0009458   1.241 0.215591
## PTRATIO                      -0.0036298  0.0027451  -1.322 0.186964
## LSTAT                        -0.0180644  0.0021377  -8.451  8.6e-16 ***
```

```
## AGE                            -0.0010204  0.0004969  -2.053 0.040793 *
## X.rooms.dwelling                0.0242811  0.0085611   2.836 0.004839 **
## nitric.oxides.concentration    -0.3263059  0.1210834  -2.695 0.007391 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1503 on 340 degrees of freedom
## Multiple R-squared:  0.5855, Adjusted R-squared:  0.5696
## F-statistic: 36.94 on 13 and 340 DF,  p-value: < 2.2e-16
```

```
#Getting the formula
formula(Reg)
```

```
## log_MEDV ~ CRIM + INDUS + RAD + TAX + B + Charles.River.dummy.variable +
##     DIS + ZN + PTRATIO + LSTAT + AGE + X.rooms.dwelling + nitric.oxides.co
ncentration
```

```
#Getting the formula
formula(Reg)
```

```
## log_MEDV ~ CRIM + INDUS + RAD + TAX + B + Charles.River.dummy.variable +
##     DIS + ZN + PTRATIO + LSTAT + AGE + X.rooms.dwelling + nitric.oxides.co
ncentration
```

```
#Remove insignificant variables :


Reg1<-lm(log_MEDV~
          Charles.River.dummy.variable+
          DIS+PTRATIO+LSTAT+AGE+X.rooms.dwelling+nitric.oxides.concentration,
data=train)
summary(Reg1)
```

```
##
## Call:
## lm(formula = log_MEDV ~ Charles.River.dummy.variable + DIS +
##     PTRATIO + LSTAT + AGE + X.rooms.dwelling + nitric.oxides.concentration
,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40671 -0.09551 -0.01474  0.09706  0.48335
```

```
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   3.500469   0.065967  53.064  < 2e-16 ***
## Charles.River.dummy.variable  0.073992   0.031930   2.317 0.021069 *
## DIS                          -0.019626   0.006458  -3.039 0.002555 **
## PTRATIO                      -0.002046   0.002387  -0.857 0.392011
## LSTAT                        -0.017785   0.002095  -8.489 6.20e-16 ***
## AGE                          -0.001658   0.000468  -3.543 0.000451 ***
## X.rooms.dwelling              0.038625   0.007067   5.466 8.83e-08 ***
## nitric.oxides.concentration  -0.524614   0.096037  -5.463 8.98e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1538 on 346 degrees of freedom
## Multiple R-squared:  0.558,  Adjusted R-squared:  0.549
## F-statistic: 62.39 on 7 and 346 DF,  p-value: < 2.2e-16
```

```r
#Reg2 : remove insignificant values

Reg2 <- lm(log_MEDV ~CRIM+INDUS+RAD+TAX+B+
                 Charles.River.dummy.variable+
                 DIS+ZN+PTRATIO+LSTAT+X.rooms.dwelling+nitric.oxides.concentr
ation, data=train)
summary(Reg2)
```

```
##
## Call:
## lm(formula = log_MEDV ~ CRIM + INDUS + RAD + TAX + B + Charles.River.dummy
.variable +
##     DIS + ZN + PTRATIO + LSTAT + X.rooms.dwelling + nitric.oxides.concentr
ation,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36466 -0.10644 -0.01634  0.09679  0.46954
```

```
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 3.6716657  0.2374187  15.465  < 2e-16 ***
## CRIM                       -0.1093257  0.0267569  -4.086 5.48e-05 ***
## INDUS                      -0.0017660  0.0020926  -0.844  0.39931
## RAD                         0.0014879  0.0014517   1.025  0.30610
## TAX                        -0.0002315  0.0001580  -1.465  0.14381
## B                          -0.0002661  0.0004965  -0.536  0.59229
## Charles.River.dummy.variable  0.0669525  0.0314421   2.129  0.03394 *
## DIS                        -0.0217655  0.0065631  -3.316  0.00101 **
## ZN                          0.0012962  0.0009483   1.367  0.17259
## PTRATIO                    -0.0033593  0.0027548  -1.219  0.22353
## LSTAT                      -0.0197241  0.0019883  -9.920  < 2e-16 ***
## X.rooms.dwelling            0.0230146  0.0085791   2.683  0.00766 **
## nitric.oxides.concentration  -0.3426291  0.1213907  -2.823  0.00504 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.151 on 341 degrees of freedom
## Multiple R-squared:  0.5803, Adjusted R-squared:  0.5655
## F-statistic: 39.29 on 12 and 341 DF,  p-value: < 2.2e-16
```

```
#Reg3 _ remove insignificant values
Reg3 <- lm(log_MEDV ~CRIM+RAD+
             Charles.River.dummy.variable+
             DIS+ZN+PTRATIO+LSTAT+nitric.oxides.concentration, data=train)
summary(Reg3)
```

```
##
## Call:
## lm(formula = log_MEDV ~ CRIM + RAD + Charles.River.dummy.variable +
##     DIS + ZN + PTRATIO + LSTAT + nitric.oxides.concentration,
##     data = train)
##
## Residuals:
```

```
##       Min       1Q   Median       3Q      Max
## -0.39258 -0.09839 -0.01661  0.09708  0.48898
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   3.6261434  0.0807548  44.903  < 2e-16 ***
## CRIM                         -0.1454189  0.0234353  -6.205 1.57e-09 ***
## RAD                           0.0040390  0.0012692   3.182 0.001594 **
## Charles.River.dummy.variable  0.0836902  0.0313997   2.665 0.008053 **
## DIS                          -0.0206337  0.0061179  -3.373 0.000829 ***
## ZN                            0.0016759  0.0009496   1.765 0.078492 .
## PTRATIO                      -0.0033216  0.0027816  -1.194 0.233255
## LSTAT                        -0.0227573  0.0017663 -12.884  < 2e-16 ***
## nitric.oxides.concentration  -0.3007190  0.1132555  -2.655 0.008293 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.153 on 345 degrees of freedom
## Multiple R-squared:  0.5638, Adjusted R-squared:  0.5537
## F-statistic: 55.74 on 8 and 345 DF,  p-value: < 2.2e-16
```

```
#Some other combination
Reg4<-lm(log_MEDV~INDUS  +ZN + X.rooms.dwelling + LSTAT+CRIM + Charles.River.
dummy.variable,data=train)
summary(Reg4)
```

```
##
## Call:
## lm(formula = log_MEDV ~ INDUS + ZN + X.rooms.dwelling + LSTAT +
##     CRIM + Charles.River.dummy.variable, data = train)
##
## Residuals:
##       Min       1Q   Median       3Q      Max
## -0.45312 -0.10307 -0.02106  0.09953  0.51897
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)                    3.3044861  0.0361648  91.373  < 2e-16 ***
## INDUS                         -0.0010489  0.0018887  -0.555   0.5790
## ZN                             0.0007041  0.0009199   0.765   0.4446
## X.rooms.dwelling               0.0062007  0.0037141   1.669   0.0959 .
## LSTAT                         -0.0225876  0.0017457 -12.939  < 2e-16 ***
## CRIM                          -0.1044110  0.0230301  -4.534 7.99e-06 ***
## Charles.River.dummy.variable  0.0633570  0.0320178   1.979   0.0486 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1553 on 347 degrees of freedom
## Multiple R-squared:  0.548,  Adjusted R-squared:  0.5402
## F-statistic: 70.12 on 6 and 347 DF,  p-value: < 2.2e-16
```

```
#The best model happens to be : Reg3

##Getting predicted values
predicted<-predict(Reg3)
plot(predicted)
length(predicted)
```

```
## [1] 354
```

```
##Finding Residuals
residuals<-resid(Reg3)
plot(residuals)
length(residuals)
```

```
## [1] 354
```

```
##Plotting Residuals vs Predicted Values
##Checking Heteroskedastcity
##There should be no trend between predicted values and residual values
plot(predicted,residuals,abline(0,0))

#You can notice that there seems to be an inverse pattern for some points

#So this model may not be the preferred model.
```

```r
#atttching predicted values to test data
predicted<-predict(Reg3,newdata=test)
length(predicted)
## [1] 152
test$p<-predicted

#Calculating error in the test dataset - (Actual- predicted)/predicted values
test$error<-(test$log_MEDV-test$p)/test$log_MEDV
mean(test$error)*100 #you get to know the average error in the given dataset
## [1] 0.02728412
##Plotting actual vs predicted values
plot(test$p,col="blue",type="l")
lines(test$log_MEDV,col="red",type="l")

#checking for Correlation between variables
library(car)
vif(Reg3)
##                        CRIM                         RAD
##                    2.172322                   13.509577
## Charles.River.dummy.variable                        DIS
##                    1.515615                    2.490395
##                          ZN                     PTRATIO
##                    1.667336                    7.083645
##                       LSTAT  nitric.oxides.concentration
##                    1.412882                    8.546674
#You can drop variables if they have a vif>10 ; means high correlation betwee
n variables
```