Predictive Analytics - Assignment 4

746 - Subhodeep Bhattacharjee

2026-02-15

PS-3

Q.5

Problem to demonstrate the utility of nonlinear regression over linear regression

Get the fgl data set from “MASS” library

```
library(MASS)
data("fgl")
str(fgl)
```

```
## 'data.frame':    214 obs. of  10 variables:
##  $ RI  : num  3.01 -0.39 -1.82 -0.34 -0.58 ...
##  $ Na  : num  13.6 13.9 13.5 13.2 13.3 ...
##  $ Mg  : num  4.49 3.6 3.55 3.69 3.62 3.61 3.6 3.61 3.58 3.6 ...
##  $ Al  : num  1.1 1.36 1.54 1.29 1.24 1.62 1.14 1.05 1.37 1.36 ...
##  $ Si  : num  71.8 72.7 73 72.6 73.1 ...
##  $ K   : num  0.06 0.48 0.39 0.57 0.55 0.64 0.58 0.57 0.56 0.57 ...
##  $ Ca  : num  8.75 7.83 7.78 8.22 8.07 8.07 8.17 8.24 8.3 8.4 ...
##  $ Ba  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Fe  : num  0 0 0 0 0 0.26 0 0 0 0.11 ...
##  $ type: Factor w/ 6 levels "WinF","WinNF",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
head(fgl)
```

```
##      RI    Na   Mg   Al    Si    K   Ca Ba   Fe type
## 1  3.01 13.64 4.49 1.10 71.78 0.06 8.75  0 0.00 WinF
## 2 -0.39 13.89 3.60 1.36 72.73 0.48 7.83  0 0.00 WinF
## 3 -1.82 13.53 3.55 1.54 72.99 0.39 7.78  0 0.00 WinF
## 4 -0.34 13.21 3.69 1.29 72.61 0.57 8.22  0 0.00 WinF
## 5 -0.58 13.27 3.62 1.24 73.08 0.55 8.07  0 0.00 WinF
## 6 -2.04 12.79 3.61 1.62 72.97 0.64 8.07  0 0.26 WinF
```

The dataset contains:
RI → Refractive Index (response)
Na → Sodium
Mg → Magnesium
Al → Aluminium
Si → Silicon
K → Potassium
Ca → Calcium
Ba → Barium

Fe → Iron
type → glass type (we focus on Vehicle Window glass)

```
fgl_veh = subset(fgl,fgl$type == "Veh")
head(fgl_veh)
```

```
##        RI    Na   Mg   Al    Si    K   Ca Ba   Fe type
## 147 -0.31 13.65 3.66 1.11 72.77 0.11 8.60  0 0.00  Veh
## 148 -1.90 13.33 3.53 1.34 72.67 0.56 8.33  0 0.00  Veh
## 149 -1.30 13.24 3.57 1.38 72.70 0.56 8.44  0 0.10  Veh
## 150 -1.57 12.16 3.52 1.35 72.89 0.57 8.53  0 0.00  Veh
## 151 -1.35 13.14 3.45 1.76 72.48 0.60 8.38  0 0.17  Veh
## 152  3.27 14.32 3.90 0.83 71.50 0.00 9.49  0 0.00  Veh
```

(a) Considering the refractive index (RI) of "Vehicle Window glass" as the variable of interest and assuming linearity of regression, run multiple linear regression of RI on different metallic oxides. From the p value, report which metallic oxide best explains the refractive index.

We regress: RI= β_0+β_1Na+β_2Mg+β_3Al+β_4Si+β_5K+β_6Ca+β_7Ba+β_8Fe+ϵ

```
Full_model = lm(RI~Na+Mg+Al+Si+K+Ca+Ba+Fe, data=fgl_veh)
summary(Full_model)
```

```
##
## Call:
## lm(formula = RI ~ Na + Mg + Al + Si + K + Ca + Ba + Fe, data = fgl_veh)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29194 -0.08582  0.00072  0.10740  0.33524
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 131.4641    47.2669   2.781  0.02388 *
## Na           -0.4333     0.3509  -1.235  0.25190
## Mg           -0.2866     1.0075  -0.285  0.78325
## Al           -0.8909     0.5550  -1.605  0.14713
## Si           -1.8824     0.4993  -3.770  0.00547 **
## K            -2.4232     0.9725  -2.492  0.03743 *
## Ca            1.5326     0.5818   2.634  0.02998 *
## Ba            0.3517     2.6904   0.131  0.89922
## Fe            3.8931     0.9581   4.063  0.00362 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2621 on 8 degrees of freedom
```

```
## Multiple R-squared:  0.9906, Adjusted R-squared:  0.9813
## F-statistic: 105.9 on 8 and 8 DF,  p-value: 2.622e-07
```

Interpretation

We know that the element with lowest p-value and highest t-statistic value will be the best predictor.

Here we observe that "Fe" (Iron) is the best oxide for predicting the refractive index of vehicle glass window.

(b) Run a simple linear regression of RI on the best predictor chosen in (a).

RI= $\beta_0 + \beta_1 \text{Fe} + \epsilon$

```
Simple_model= lm(RI ~ Fe, data= fgl_veh)
summary(Simple_model)

##
## Call:
## lm(formula = RI ~ Fe, data = fgl_veh)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2324 -1.0693 -0.2715  0.2907  3.7707
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.5007     0.4861  -1.030   0.3193
## Fe            8.1362     4.0780   1.995   0.0645 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.759 on 15 degrees of freedom
## Multiple R-squared:  0.2097, Adjusted R-squared:  0.157
## F-statistic: 3.981 on 1 and 15 DF,  p-value: 0.06452
```

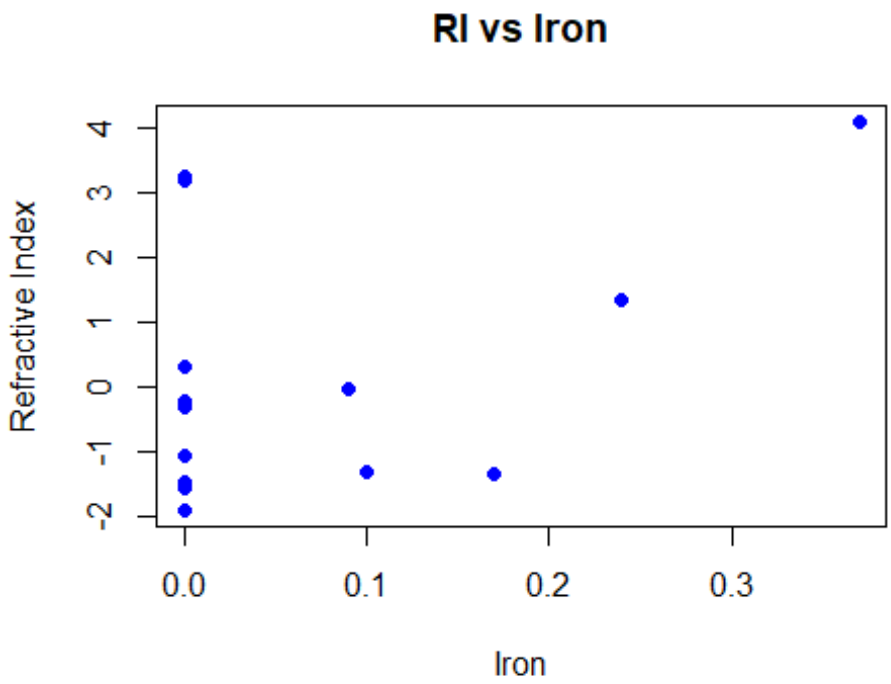
Interpretation

The R^2 value of this model is 0.2097, hence we conclude that the oxide "Fe" (Iron) explains 20.97% variability in the Refractive Index (RI) of the vehicle glass window.

(c) Can you further improve the regression of the refractive index of "Vehicle Window glass" on the predictor chosen by you in part (a)? Give the new fitted model and compare its performance with the model in (b).

```
# Plotting to check the relationship between RI and Fe
plot(fgl_veh$Fe, fgl_veh$RI,
     pch=19, col="blue",
     main="RI vs Iron",
```

```
     xlab="Iron",
     ylab="Refractive Index")
```



Interpretation

Here, we observe that there is no linear relationship between RI and Fe, rather we see a curvy relation between the two, hence we try fitting a quadratic regression to improve the fit.

RI= $\beta_0 + \beta_1 \text{Fe} + \beta_2 \text{Fe}^2 + \epsilon$

```
Quad_model = lm(RI ~ Fe + I(Fe^2), data= fgl_veh)
summary(Quad_model)

##
## Call:
## lm(formula = RI ~ Fe + I(Fe^2), data = fgl_veh)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6215 -1.1715 -0.1345  0.5985  3.5485
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.2785     0.4712  -0.591    0.564
## Fe          -12.1810    12.0408  -1.012    0.329
## I(Fe^2)      65.9600    37.0798   1.779    0.097 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.645 on 14 degrees of freedom
## Multiple R-squared:  0.3554, Adjusted R-squared:  0.2633
## F-statistic:  3.86 on 2 and 14 DF,  p-value: 0.04623
```

Interpretation
The R^2 value of this Quadratic model is 0.3554, hence we see that this model explains 35.54% variability on refractive index of vehicle glass window.

```
summary(Simple_model)$r.squared

## [1] 0.2097192

summary(Quad_model)$r.squared

## [1] 0.355413
```

Interpretation
Hence we see a 14.57% increase in the explanation of variability of Iron on Refractive Index of Vehicle Glass Window.

PS- 4

Q.1

Problem to demonstrate multicollinearity

Consider the Credit data in the ISLR library. Choose balance as the response and Age, Limit and Rating as the predictors

```
library(ISLR)

## Warning: package 'ISLR' was built under R version 4.2.3

library(stargazer)

##
## Please cite as:

##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary
Statistics Tables.

##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer

library(car)

## Loading required package: carData

data("Credit")
str(Credit)
```

```
## 'data.frame':    400 obs. of  12 variables:
##  $ ID       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Income   : num  14.9 106 104.6 148.9 55.9 ...
##  $ Limit    : int  3606 6645 7075 9504 4897 8047 3388 7114 3300 6819 ...
##  $ Rating   : int  283 483 514 681 357 569 259 512 266 491 ...
##  $ Cards    : int  2 3 4 3 2 4 2 2 5 3 ...
##  $ Age      : int  34 82 71 36 68 77 37 87 66 41 ...
##  $ Education: int  11 15 11 11 16 10 12 9 13 19 ...
##  $ Gender   : Factor w/ 2 levels " Male","Female": 1 2 1 2 1 1 2 1 2 2 ...
##  $ Student  : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1 1 1 2 ...
##  $ Married  : Factor w/ 2 levels "No","Yes": 2 2 1 1 2 1 1 1 1 2 ...
##  $ Ethnicity: Factor w/ 3 levels "African American",..: 3 2 2 2 3 3 1 2 3
1 ...
##  $ Balance  : int  333 903 580 964 331 1151 203 872 279 1350 ...
```

```
head(Credit)
```

```
##   ID  Income Limit Rating Cards Age Education Gender Student Married
Ethnicity
## 1  1  14.891  3606    283     2  34        11   Male      No     Yes
Caucasian
## 2  2 106.025  6645    483     3  82        15 Female     Yes     Yes
Asian
## 3  3 104.593  7075    514     4  71        11   Male      No      No
Asian
## 4  4 148.924  9504    681     3  36        11 Female      No      No
Asian
## 5  5  55.882  4897    357     2  68        16   Male      No     Yes
Caucasian
## 6  6  80.180  8047    569     4  77        10   Male      No      No
Caucasian
##   Balance
## 1     333
## 2     903
## 3     580
## 4     964
## 5     331
## 6    1151
```

We have been specified to keep:
Response → Balance
Predictors → Age, Limit, Rating

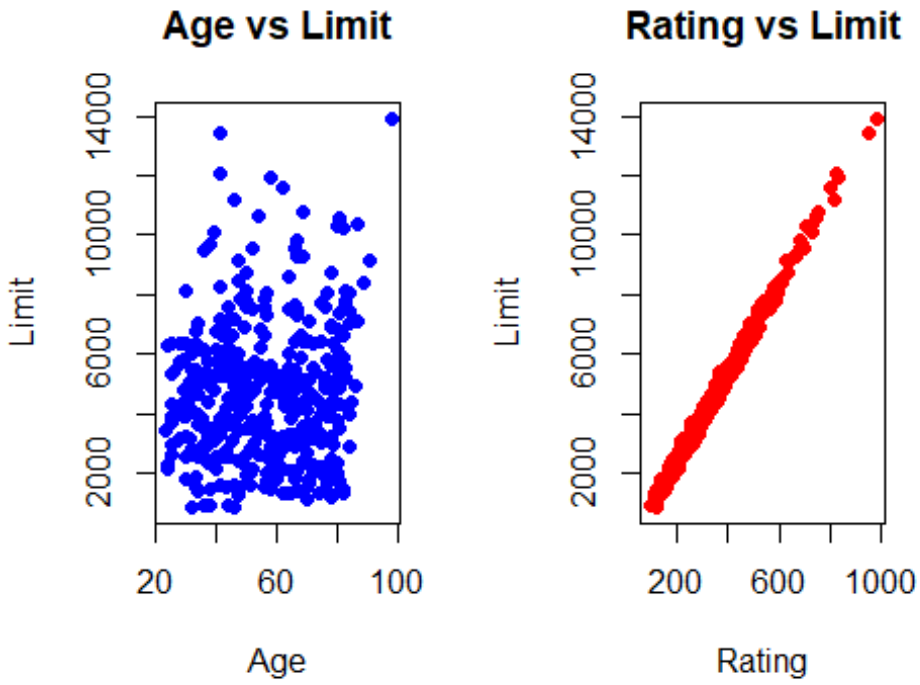
(a) Make a scatter plot of (i) Age versus Limit and (ii) Rating Versus Limit. Comment on the scatter plot.

```
par(mfrow=c(1,2))

#(i) Age vs Limit
```

```r
plot(Credit$Age, Credit$Limit,
     main="Age vs Limit",
     xlab="Age",
     ylab="Limit",
     col="blue", pch=19)


#(ii) Rating vs Limit
plot(Credit$Rating, Credit$Limit,
     main="Rating vs Limit",
     xlab="Rating",
     ylab="Limit",
     col="red", pch=19)
```



Interpretation

(i). The scatter appears random, there is proper pattern observed and there is weak collinearity between the two.

(ii). The scatter appears linear, we see a strong positive relationship between Rating and Limit.

(b) Run three separate regressions: (i) Balance on Age and Limit (ii) Balance on Age, Rating and Limit (iii) Balance on Rating and Limit. Present all the regression output in a single table using stargazer. What is the marked difference that you can observe

from the output?

Balance = β_0 + β_1Age + β_2Limit + ϵ

```
model_1 = lm(Balance ~ Age + Limit, data = Credit)
```

Balance = β_0 + β_1Age + β_2Rating + β_3Limit + ϵ

```
model_2 = lm(Balance ~ Age + Rating + Limit, data = Credit)
```

Balance = β_0 + β_1Rating + β_2Limit + ϵ

```
model_3 = lm(Balance ~ Rating + Limit, data = Credit)

stargazer(model_1,model_2,model_3, type="text",title="Regression Comparison
Table")

##
## Regression Comparison Table
##
================================================================================
=================
##                                                Dependent variable:
##                     ---------------------------------------------------------
--------------------
##                                                      Balance
##                                (1)                     (2)
(3)
## --------------------------------------------------------------------------
--------------------
## Age                          -2.291***               -2.346***
##                               (0.672)                 (0.669)
##
## Rating                                                2.310**
2.202**
##                                                       (0.940)
(0.952)
##
## Limit                         0.173***                0.019
0.025
##                               (0.005)                 (0.063)
(0.064)
##
## Constant                    -173.411***             -259.518***
-377.537***
##                               (43.828)                (55.882)
(45.254)
```

```
##
## ------------------------------------------------------------------------
-------------------
## Observations                  400                       400
400
## R2                           0.750                     0.754
0.746
## Adjusted R2                  0.749                     0.752
0.745
## Residual Std. Error    230.532 (df = 397)       229.080 (df = 396)
232.320 (df = 397)
## F Statistic         594.988*** (df = 2; 397) 403.718*** (df = 3; 396)
582.820*** (df = 2; 397)
##
========================================================================
=================
## Note:                                                         *p<0.1;
**p<0.05; ***p<0.01
```

Interpretation

In the first 2 models, Age is highly significant at 1% level implying Age has a statistically strong effect on balance. Age has negative impact on Balance since the coefficients of Age in both the models is negative.

In the first model, Limit was highly significant at 1% level but in second and third model where a new predictor Rating was introduced, Limit became statistically insignificant. This suggests, there exists multicollinearity between Rating and Limit. When both enter the model, Rating takes up most explanatory power and remains important even when Limit is included.This means Rating is a stronger determinant of balance than Limit.

(c) Calculate the variance inflation factor (VIF) and comment on multicollinearity.

The Variance Inflation Factor (VIF) reveals the presence and severity of multicollinearity across the models.

```
vif(model_1)

##      Age    Limit
## 1.010283 1.010283

vif(model_2)

##        Age     Rating      Limit
##   1.011385 160.668301 160.592880

vif(model_3)
```

```
##   Rating    Limit
## 160.4933 160.4933
```

Interpretation
In the Age and Limit model, both Age and Limit have VIF values close to 1 (≈1.01), indicating no multicollinearity meaning the predictors are almost completely independent of each other, and their coefficient estimates are stable and reliable.

However, once Rating is introduced in the Age, Limit and Rating model, the VIF values for Limit (≈160.59) and Rating (≈160.67) increase significantly, while Age remains near 1 (≈1.01).

Similarly, in the Limit and Rating model, both variables again show extremely high VIF values (≈160.49). A VIF exceeding 10 is typically considered serious; values above 100 indicate extreme multicollinearity. This means Limit and Rating are almost perfectly linearly correlated, causing inflated standard errors and making individual coefficient estimates unstable and statistically unreliable. This explains why Limit becomes insignificant when Rating is added in the model. In contrast, Age remains unaffected because it is not strongly correlated with the other predictors.

Overall, the output provides strong statistical evidence of severe multicollinearity between Limit and Rating, confirming the pattern observed in the scatter plot and regression table.

Q.2

Problem to demonstrate the detection of outlier, leverage and influential points

Attach “Boston” data from MASS library in R. Select median value of owneroccupied homes, as the response and per capita crime rate, nitrogen oxides concentration, proportion of blacks and percentage of lower status of the population as predictors.

```r
library(MASS)
data("Boston")
str(Boston)
```

```
## 'data.frame':    506 obs. of  14 variables:
##  $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ 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 ...
##  $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524
0.524 ...
##  $ rm     : 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 96.1 100 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 15.2 15.2 15.2 ...
##  $ black  : 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 ...
```

```
head(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black
lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90
4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90
9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83
4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63
2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90
5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12
5.21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

We have been specified to keep:
Response → medv (median value of owner-occupied homes in $1000s)
Predictors → crim (per capita crime rate by town), nox (nitrogen oxides concentration (parts per 10 million), black (the proportion of blacks by town), lstat (lower status of the population)

The objective is to fit a multiple linear regression model of the response on the predictors.

$\text{medv} = \beta_0 + \beta_1 \text{crim} + \beta_2 \text{nox} + \beta_3 \text{black} + \beta_4 \text{lstat} + \epsilon$

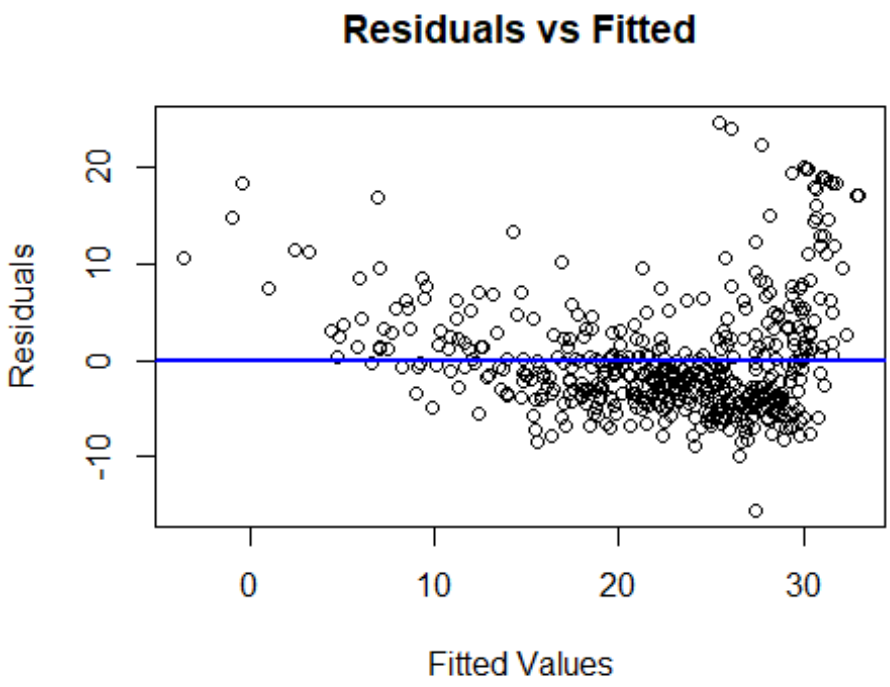
```
boston_model = lm(medv ~ crim + nox + black + lstat, data=Boston)
summary(boston_model)
```

```
##
## Call:
## lm(formula = medv ~ crim + nox + black + lstat, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.564  -4.004  -1.504   2.178  24.608
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.053584   2.170839  13.844   <2e-16 ***
## crim        -0.059424   0.037755  -1.574    0.116
## nox          3.415809   3.056602   1.118    0.264
## black        0.006785   0.003408   1.991    0.047 *
## lstat       -0.918431   0.050167 -18.307   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.183 on 501 degrees of freedom
## Multiple R-squared:  0.5517, Adjusted R-squared:  0.5481
## F-statistic: 154.1 on 4 and 501 DF,  p-value: < 2.2e-16
```

With reference to this problem, detect outliers, leverage points and influential points if any.

- Plotting a Residual Plot

```r
plot(boston_model$fitted.values, resid(boston_model),
     xlab="Fitted Values",
     ylab="Residuals",
     main="Residuals vs Fitted")
abline(h=0,col="blue",lwd=2)
```



Interpretation

After finding the residual plot we can clearly see that there are some outliers present in both positive and negative direction. Finding out leverage and influential points from this plot is not possible, i.e we cannot directly comment on the same from just this plot.

- Finding Potential Outliers

First, we find out the standardized residuals of the fitted model and then a point is declared as potential outliers if its standard residual is less than -2 and and greater than 2.

```
#Finding standardized residuals
std.res=rstandard(boston_model)

#Detecting Potential Outlier
outliers=which(abs(std.res)>2)
outliers
```

```
##  99 162 163 164 167 187 196 204 205 215 225 226 229 234 257 258 262 263
268 281
##  99 162 163 164 167 187 196 204 205 215 225 226 229 234 257 258 262 263
268 281
## 283 284 369 370 371 372 373 375 410 413 506
## 283 284 369 370 371 372 373 375 410 413 506
```

```
length(outliers)
```

```
## [1] 31
```

Interpretation
We can clearly see that there are 31 points as potential outliers in this standardized fitted model which does not fit the model well.

- Finding Leverage Points

First, we find out the diagonal elements of the hat matrix. Then we calculate a cutoff point L=3*(p+1)/n where p is the number of predictors and n is number of rows. If the hat values exceed the leverage value then we call the points as leverages points.

```
diag_val=hatvalues(boston_model)

n=nrow(Boston) #rows
p=4  #number of predictors

#Calculating the leverage values
cutoff=3*(p+1)/n
cutoff
```

```
## [1] 0.02964427
```

```
# High leverage observations
leverage_points=which(diag_val>cutoff)
leverage_points
```

```
##  49 103 142 156 157 160 375 381 399 405 406 411 413 415 416 417 419 424
425 426
##  49 103 142 156 157 160 375 381 399 405 406 411 413 415 416 417 419 424
425 426
## 427 428 438 439 451 455 457 458 467
## 427 428 438 439 451 455 457 458 467
```

```
length(leverage_points)
```

```
## [1] 29
```

Interpretation
We observe that there are 29 leverage points present in total which means these are points with high predictor values and may influence the model.

- Finding Influential Points

We find out the Cook's distance D_i which is a function of standardized residuals and elements of hat matrix.

If for a data point D_i>1, we can say that point is an influential point.

```
#Calculating the Di values
cook_distance=cooks.distance(boston_model)

influential=which(cook_distance>1)
length(influential)

## [1] 0
```

Interpretation
We see that there exist no such point where D_i>1, so we conclude that in this model there exists no such influential points.
