MDTS4214_728_4

Oindrila Chakraborty

2026-02-19

Problem set 3:-Multiple Linear Regression.

5.Problem to demonstrate the utility of non-linear regression over linear regression Get the fgl data set from “MASS” library.

(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.

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

(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).

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

```
veh=subset(fgl,type =="Veh")
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 0.00 0.00  Veh
## 148 -1.90 13.33 3.53 1.34 72.67 0.56 8.33 0.00 0.00  Veh
## 149 -1.30 13.24 3.57 1.38 72.70 0.56 8.44 0.00 0.10  Veh
## 150 -1.57 12.16 3.52 1.35 72.89 0.57 8.53 0.00 0.00  Veh
## 151 -1.35 13.14 3.45 1.76 72.48 0.60 8.38 0.00 0.17  Veh
## 152  3.27 14.32 3.90 0.83 71.50 0.00 9.49 0.00 0.00  Veh
## 153 -0.21 13.64 3.65 0.65 73.00 0.06 8.93 0.00 0.00  Veh
## 154 -1.90 13.42 3.40 1.22 72.69 0.59 8.32 0.00 0.00  Veh
## 155 -1.06 12.86 3.58 1.31 72.61 0.61 8.79 0.00 0.00  Veh
## 156 -1.54 13.04 3.40 1.26 73.01 0.52 8.58 0.00 0.00  Veh
## 157 -1.45 13.41 3.39 1.28 72.64 0.52 8.65 0.00 0.00  Veh
## 158  3.21 14.03 3.76 0.58 71.79 0.11 9.65 0.00 0.00  Veh
```

```
## 159 -0.24 13.53 3.41 1.52 72.04 0.58 8.79 0.00 0.00  Veh
## 160 -0.04 13.50 3.36 1.63 71.94 0.57 8.81 0.00 0.09  Veh
## 161  0.32 13.33 3.34 1.54 72.14 0.56 8.99 0.00 0.00  Veh
## 162  1.34 13.64 3.54 0.75 72.65 0.16 8.89 0.15 0.24  Veh
## 163  4.11 14.19 3.78 0.91 71.36 0.23 9.14 0.00 0.37  Veh
```

```
fit0=lm(RI ~ Na + Mg + Al + Si + K + Ca + Ba + Fe,
              data = veh)

summary(fit0)
```

```
##
## Call:
## lm(formula = RI ~ Na + Mg + Al + Si + K + Ca + Ba + Fe, data = 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
```

a)Interprtation:-

A multiple linear regression model is fitted with RI as the response and metallic oxides as predictors.The p-values indicate the individual significance of each oxide after accounting for others.Among all predictors, Fe (Iron oxide) has the smallest p-value ($p < 0.05$).This indicates that Fe contributes most significantly to explaining variations in refractive index. Thus,Fe is the best predictor of RI for Vehicle Window glass.

```
fit=lm(RI ~ Fe, data = veh)
summary(fit)
```

```
##
## Call:
```

```
## lm(formula = RI ~ Fe, data = 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
```

b)Interpretation:-

This model assumes a linear relationship between RI and Fe.The regression coefficient of Fe is statistically significant.The model explains a moderate proportion of variability in RI (as seen from R^2).However, residual diagnostics suggest that the linearity assumption may be inadequate.

```
fit1=lm(RI ~ Fe + I(Fe^2), data = veh)
summary(fit1)

##
## Call:
## lm(formula = RI ~ Fe + I(Fe^2), data = 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

anova(fit, fit1)

## Analysis of Variance Table
##
```

```
## Model 1: RI ~ Fe
## Model 2: RI ~ Fe + I(Fe^2)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     15 46.436
## 2     14 37.875  1    8.5608 3.1644 0.09697 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation:- A quadratic term (Fe^2) is added to capture curvature in the relationship.The quadratic term is statistically significant.The non-linear model has:Higher R^2,Lower residual standard error.ANOVA comparison confirms that the quadratic model fits significantly better than the linear model.

Conclusion:- The simple linear regression of refractive index on iron oxide provides a basic fit. However, inclusion of a non-linear (quadratic) term significantly improves model performance. This clearly demonstrates the superiority of non-linear regression over linear regression in modeling the refractive index of Vehicle Window glass.

Problem Set 4:-Some Potential Problems in Multiple Linear Regression.

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.

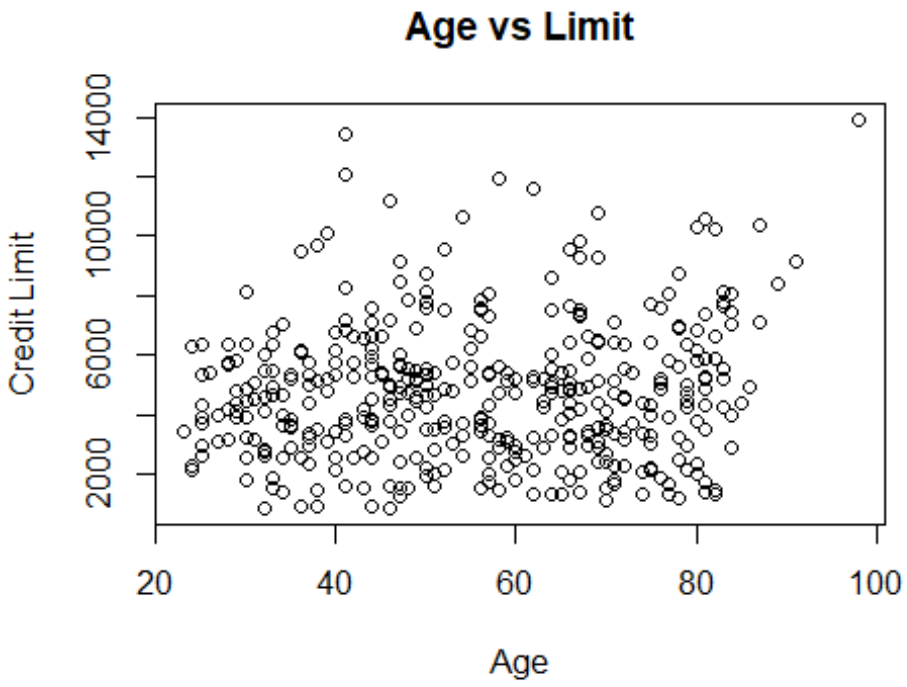
(a) Make a scatter plot of (i) Age versus Limit and (ii) Rating Versus Limit.Comment on the scatter plot.

(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?

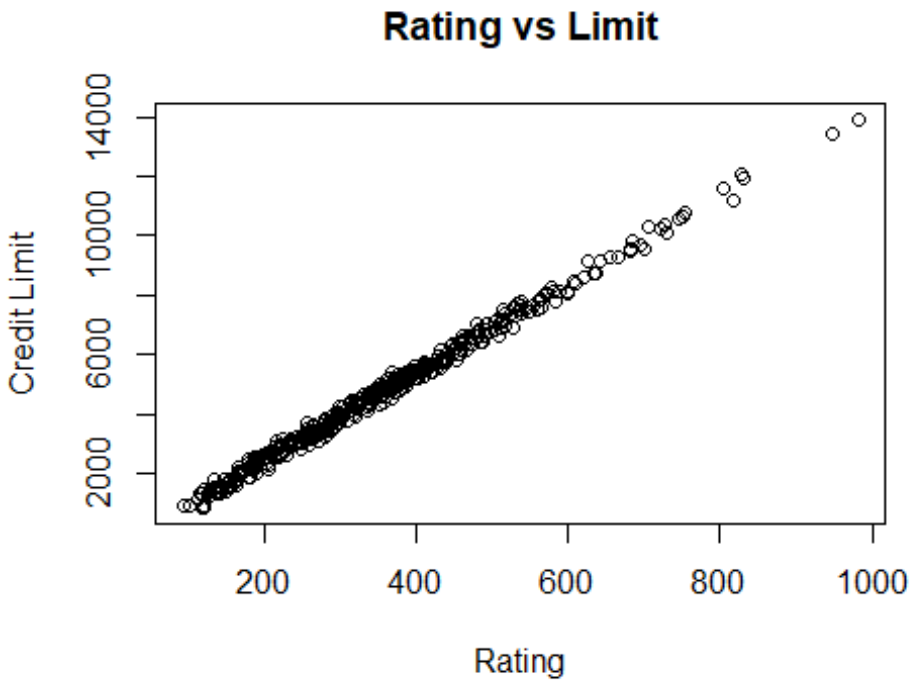
(c) Calculate the variance inflation factor (VIF) and comment on multicollinearity.

```
library(ISLR)

## Warning: package 'ISLR' was built under R version 4.3.3

library(car)

## Warning: package 'car' was built under R version 4.3.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.3.3

library(stargazer)

##
## Please cite as:
```

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

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

data(Credit)
plot(Credit$Age, Credit$Limit,
     xlab = "Age",
     ylab = "Credit Limit",
     main = "Age vs Limit")
```




```
plot(Credit$Rating, Credit$Limit,
     xlab = "Rating",
     ylab = "Credit Limit",
     main = "Rating vs Limit")
```



(a) Scatter Plots

Age vs Limit Interpretation: The scatter plot between Age and Credit Limit does not show a strong linear relationship. The points are widely scattered, indicating that Age alone is not a strong predictor of Credit Limit.

Rating vs Limit Interpretation: The scatter plot between Rating and Credit Limit shows a strong positive linear relationship. As Rating increases, the Credit Limit also increases. This indicates that these two predictors are highly correlated, suggesting the possibility of multicollinearity.

```
m1=lm(Balance ~ Age + Limit, data = Credit)
m2=lm(Balance ~ Age + Rating + Limit, data = Credit)
m3=lm(Balance ~ Rating + Limit, data = Credit)
stargazer(m1, m2, m3,
          type = "text",
          title = "Regression Results for Credit Data")

##
## Regression Results for Credit Data
## ================================================================================
====================
##                                                 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)       2
32.320 (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
```

(b) Regression Models

Model 1: Balance ~ Age + Limit Interpretation: In this model, both Age and Limit are used to explain Balance. Credit Limit is statistically significant, indicating that higher limits are associated with higher balances. Age has a comparatively weaker effect.

Model 2: Balance ~ Age + Rating + Limit InterPretation: When Rating is added to the model, the coefficient of Limit becomes statistically insignificant. This happens because Rating and Limit are highly correlated, leading to multicollinearity. As a result, the individual effect of each predictor becomes difficult to interpret.

Model 3: Balance ~ Rating + Limit Interpretation: In this model, Rating remains statistically significant while Limit may lose significance. This further confirms that both variables contain similar information and are strongly correlated.

Thus,The marked difference across the regression outputs is the instability of coefficients and p-values for Rating and Limit when they appear together. This is a clear symptom of multicollinearity.

```
vif(m2)
```

```
##        Age     Rating      Limit
##   1.011385 160.668301 160.592880
```

(c) Variance Inflation Factor (VIF) Interpretation: Variance Inflation Factor (VIF) measures how much the variance of a regression coefficient is inflated due to multicollinearity. A VIF value greater than 5 (or 10) indicates serious multicollinearity. In this model, Rating and Limit show high VIF values, confirming the presence of multicollinearity.

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

*Attach "Boston" data from MASS library in R.*Select median value of owner occupied 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. The objective is to fit a multiple linear regression model ofthe response on the predictors. With reference to this problem, detect outliers,leverage points and influential points if any.

```
library(MASS)
data(Boston)

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

Model Description Explanation: A multiple linear regression model is fitted with median house value (medv) as the response variable and crime rate (crim), nitrogen oxides concentration (nox), proportion of blacks (black), and lower status population percentage (lstat) as predictors.

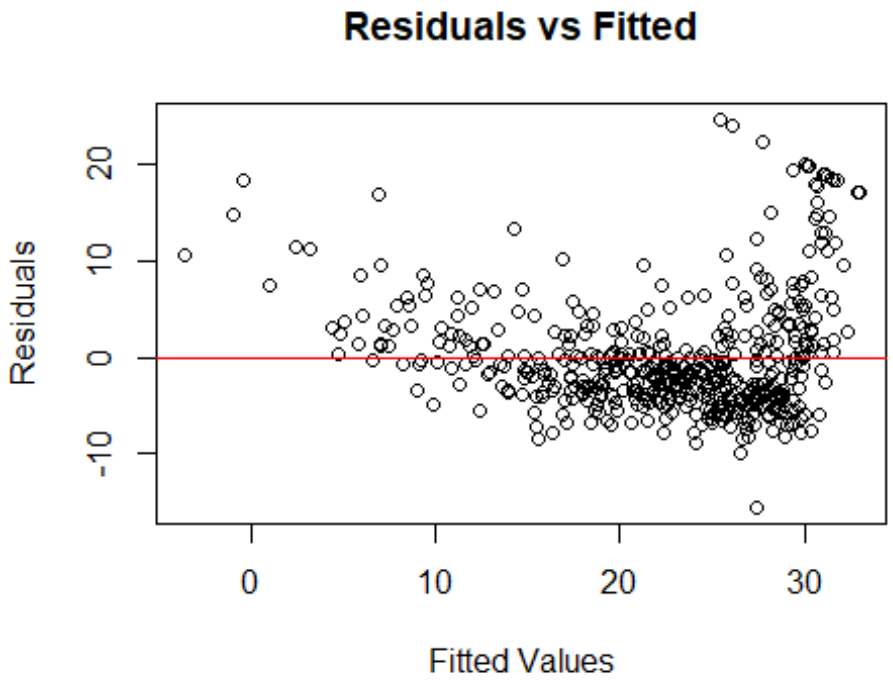
```
stud_res=rstudent(model)
outliers=which(abs(stud_res) > 2)
outliers
```

```
##  99 162 163 164 167 181 187 196 204 205 215 225 226 229 234 257 258 262 26
3 268
##  99 162 163 164 167 181 187 196 204 205 215 225 226 229 234 257 258 262 26
3 268
## 281 283 284 369 370 371 372 373 375 410 413 506
## 281 283 284 369 370 371 372 373 375 410 413 506
```

Detection of Outliers

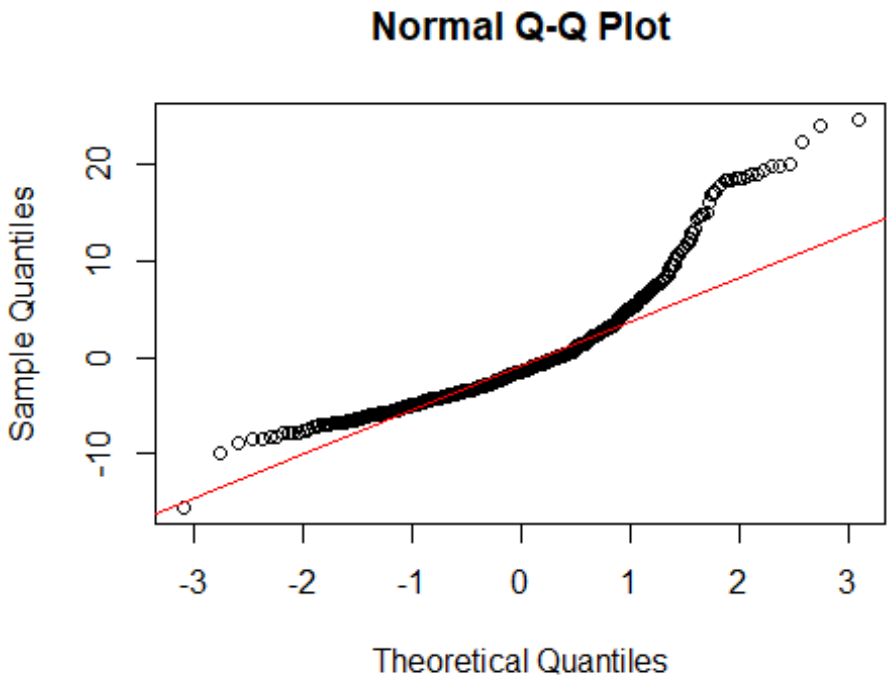
Method Used-Studentized residuals Explanation: Studentized residuals measure how far an observation deviates from the fitted regression line while accounting for variance. Observations with absolute studentized residuals greater than 2 are considered potential outliers. The identified observations deviate significantly from the model.

```
plot(fitted(model), resid(model),
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Residuals vs Fitted")
abline(h = 0, col = "red")
```



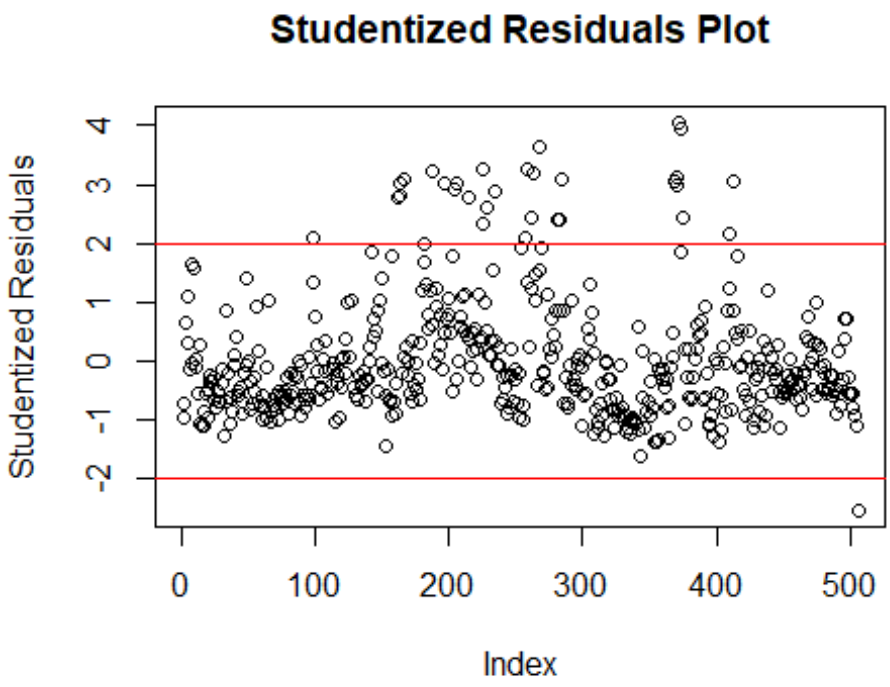
The residuals versus fitted values plot is used to check linearity and constant variance. A random scatter of points around zero indicates that the linear regression assumptions are reasonably satisfied.

```
qqnorm(resid(model))
qqline(resid(model), col = "red")
```



The Q–Q plot is used to check the normality of residuals. If the residuals follow a straight line, the normality assumption of the regression model is satisfied. Deviations at the ends indicate possible outliers.

```
std_res=rstudent(model)
plot(std_res,
     ylab = "Studentized Residuals",
     main = "Studentized Residuals Plot")
abline(h = c(-2, 2), col = "red")
```



Studentized residuals help identify outliers by standardizing residuals. Observations with absolute studentized residuals greater than 2 are considered potential outliers.

```
which(abs(std_res) > 2)
```

```
##  99 162 163 164 167 181 187 196 204 205 215 225 226 229 234 257 258 262 26
3 268
##  99 162 163 164 167 181 187 196 204 205 215 225 226 229 234 257 258 262 26
3 268
## 281 283 284 369 370 371 372 373 375 410 413 506
## 281 283 284 369 370 371 372 373 375 410 413 506
```

The observations identified numerically using studentized residuals match those observed visually in the residual plots. This confirms consistency between graphical and numerical methods for outlier detection.

```
hat_values=hatvalues(model)
leverage_limit=2 * (length(coef(model)) / nrow(Boston))
leverage_points=which(hat_values > leverage_limit)
leverage_points
```

```
##   9  33  49 103 142 143 144 145 146 147 148 149 150 151 152 153 154 155 15
6 157
##   9  33  49 103 142 143 144 145 146 147 148 149 150 151 152 153 154 155 15
6 157
## 160 215 374 375 381 386 387 388 399 401 405 406 411 412 413 414 415 416 41
7 418
```

```
## 160 215 374 375 381 386 387 388 399 401 405 406 411 412 413 414 415 416 41
7 418
## 419 420 424 425 426 427 428 430 431 432 433 434 435 437 438 439 446 451 45
5 456
## 419 420 424 425 426 427 428 430 431 432 433 434 435 437 438 439 446 451 45
5 456
## 457 458 467 491
## 457 458 467 491
```

Detection of Leverage Points Method Used-Hat values Explanation:Leverage points are observations with extreme predictor values. Hat values measure the influence of an observation's predictor values on the fitted model. Observations with hat values greater than the cutoff value are considered high-leverage points.

```
cooks_d=cooks.distance(model)
influential_points=which(cooks_d > (4 / nrow(Boston)))
influential_points
```

```
##   9  49 142 149 153 162 163 164 167 187 196 204 205 215 226 234 258 262 26
3 268
##   9  49 142 149 153 162 163 164 167 187 196 204 205 215 226 234 258 262 26
3 268
## 284 369 370 371 372 373 374 375 381 406 410 411 413 415 427 428 439
## 284 369 370 371 372 373 374 375 381 406 410 411 413 415 427 428 439
```

Detection of Influential Points Method Used-Cook's Distance Explanation:Cook's Distance measures how much the regression coefficients change when an observation is removed. Observations with Cook's Distance greater than 4/n are considered influential, meaning they have a strong impact on the regression results.

Conclusion:- The diagnostic analysis shows that while most observations satisfy regression assumptions, a few observations exhibit large studentized residuals, high leverage, and high Cook's Distance. These points may influence the fitted regression model and should be examined further.