Assignment: 1
Statistical Model and Regression analysis

Urmila Dhundhwal(M20MA207)

Q:1 Given that, we have a data set with five predictors, X1 = GPA, X2 = IQ, X3 = Level (1 for College and 0 for High School), X4 = Interaction between GPA and IQ, and X5 = Interaction between GPA and Level. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get β0 = 50, β1 = 20, β2 = 0.07, β3 = 35, β4 = 0.01, β5 = −10.
Fitted regression model eq will be-
Salary = β0 + β1*GPA + β2*IQ + β3*Level + β4*(GPA×IQ) + β5*(GPA×Level)

(a) The correct answer is (ii) For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates. Because the coefficient , β3 = 35 represents the difference in starting salary between college graduates (Level = 1) and high school graduates (Level = 0) when all other predictors are held fixed. Since β3 = 35 is positive, it implies that, on average, college graduates earn more than high school graduates, for a fixed value of IQ and GPA.

(b) We have to predict the salary of a college graduate if IQ is 110 and a GPA is 4.0.

$$\text{Salary}=50+20\times4.0+0.07\times110+35+0.01\times(4.0\times110)-10\times(4.0\times1)$$
$$\text{Salary}=50+80+7.7+35+4.4-40$$
$$\text{Salary}=136.1$$

So, the predicted salary of a college graduate with IQ of 110 and a GPA of 4.0 will be approximately $136,100.

(c) False. Even though the coefficient for the GPA/IQ interaction term ($\beta=0.01$) is small, it doesn't necessarily mean there is very little evidence of an interaction effect. The significance of an interaction effect depends on various factors, including the scale and context of the data. Therefore, the small coefficient alone does not provide conclusive

evidence for or against an interaction effect. Additional statistical tests or analysis would be needed to determine the significance of the interaction effect.

—----------

Q:2 **(a)** create a vector, x

x <- rnorm(100, mean = 0, sd = 1)

—----------

(b) create a vector, eps

eps <- rnorm(100, mean = 0, sd = sqrt(0.25))

—----------

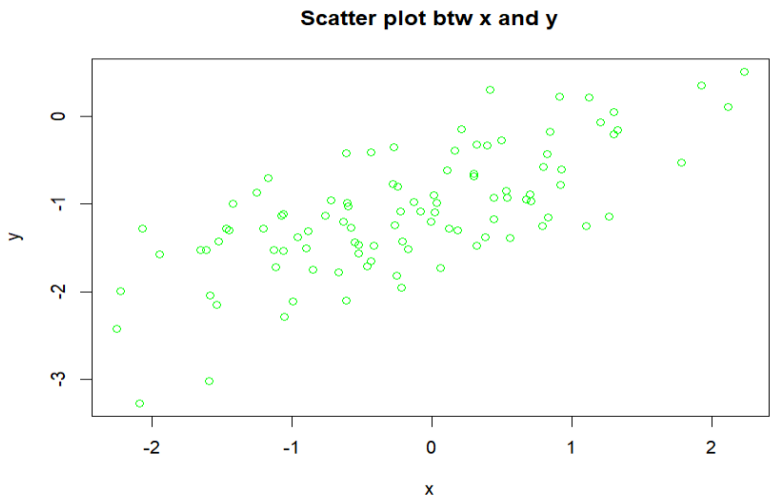
(c) Using x and eps, generate a vector y according to the model

y <- -1 + 0.5*x + eps

The length of the vector y will be 100. In this linear model, β0 = -1 and β1 = 0.5.

—----------

(d) Scatterplot displaying the relationship between x and y.



Here we can see that positive correlation between x and y, it means that as the value of one variable x increases, the value of the other variable y also tends to increase.

—----------

(e) Fit a least squares linear model to predict y using x:

Fitted least squares regression line eq will be

y = β0_cap + β1_cap*x here β0_cap = -0.99770 and β1_cap = 0.48777

```
> # (e) Fit a least squares linear model to predict y using x:
> model.fit <- lm(y ~ x)
> summary(model.fit)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.24989 -0.32768  0.02292  0.36607  1.09525

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.99770    0.04960 -20.113   <2e-16 ***
x            0.48777    0.04909   9.936   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4873 on 98 degrees of freedom
Multiple R-squared:  0.5018,    Adjusted R-squared:  0.4967
F-statistic: 98.72 on 1 and 98 DF,  p-value: < 2.2e-16
```

estimating coefficients β0_cap = -0.99770 and β1_cap = 0.48777 both are close to β0 and β1respectivey. It indicates that the linear regression model has effectively captured the underlying relationship between the predictor variable X and the target variable Y.

R2 = 0.50 (say), then it indicates the model explained 50% of the variation that is there in the response. Thus, the fitted regression model is given by y = -0.99770 + 0.48777*x

RSE = 0.4873, R2_adj = 0.4967 and R2 = 0.5018

Note that p-value< 2.2 × 10−16 is very small and hence, we should reject the overall hypothesis H0 : β0_cap = β1_cap = 0.

Check model is statisticaly significate or not: Use significance level α = 0.05

Testing Hypothisis: For β0_cap (intercept)

H01 : β0_cap = 0 vs. H11 : β0_cap != 0

p − value < 2 × 10^−16

i.e., p = value < α and hence H01 is rejected.

Testing Hypothisis: For β1_cap (slop)

H01 : β1_cap = 0 vs. H11 : β1_cap != 0

p−value < 2 × 10^−16

i.e., p-value < α and hence H01 is rejected.

Hence model a statistically significant.

—----------

(f) Display the least squares line and the population regression line on the scatterplot

—----------

Q:3 In task given that, we have to use Bostan data set. And we have to use per capita crime rate is the response and the predictors are as follows: zn, indus, nox, rm, dis, tax, medv from the Boston data set.

(a) Fit simple linear regression models for each predictor.

(i) Fit simple linear regression model for response variable crime rate and predictor zn-

```
> model1.fit <- lm(crim ~ zn , Data_set)
> summary(model1.fit)

Call:
lm(formula = crim ~ zn, data = Data_set)

Residuals:
   Min     1Q Median     3Q    Max
-4.429 -4.222 -2.620  1.250 84.523

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.45369    0.41722  10.675  < 2e-16 ***
zn          -0.07393    0.01609  -4.594 5.51e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.435 on 504 degrees of freedom
Multiple R-squared:  0.04019,   Adjusted R-squared:  0.03828
F-statistic:  21.1 on 1 and 504 DF,  p-value: 5.506e-06
```

The fitted regression model1 eq will be- crime rate = b1 + m1 * zn , here b1 = 4.45369 and m1 = -0.07393

(ii) Fit simple linear regression model for response variable crime rate and predictor indus-

```
> model2.fit <- lm(crim ~ indus , Data_set)
> summary(model2.fit)

Call:
lm(formula = crim ~ indus, data = Data_set)

Residuals:
    Min      1Q  Median      3Q     Max
-11.972  -2.698  -0.736   0.712  81.813

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.06374    0.66723  -3.093  0.00209 **
indus        0.50978    0.05102   9.991  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.866 on 504 degrees of freedom
Multiple R-squared:  0.1653,    Adjusted R-squared:  0.1637
F-statistic: 99.82 on 1 and 504 DF,  p-value: < 2.2e-16
```

The fitted regression model2 eq will be- crime rate = b2 + m2 * indus, here b2 = -2.06374 and m2 = 0.50978

(iii) Fit simple linear regression model for response variable crime rate and predictor nox-

```
> model3.fit <- lm(crim ~ zn , Data_set)
> summary(model3.fit)

Call:
lm(formula = crim ~ zn, data = Data_set)

Residuals:
   Min     1Q Median     3Q    Max
-4.429 -4.222 -2.620  1.250 84.523

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.45369    0.41722  10.675  < 2e-16 ***
zn          -0.07393    0.01609  -4.594 5.51e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.435 on 504 degrees of freedom
Multiple R-squared:  0.04019,   Adjusted R-squared:  0.03828
F-statistic:  21.1 on 1 and 504 DF,  p-value: 5.506e-06
```

The fitted regression model3 eq will be- crime rate = b3 + m3 * nox, here b3 = 4.45369 and m3 = -0.07393

(iv) Fit simple linear regression model for response variable crime rate and predictor rm-

```
> model4.fit <- lm(crim ~ rm , Data_set)
> summary(model4.fit)

Call:
lm(formula = crim ~ rm, data = Data_set)

Residuals:
   Min     1Q Median     3Q    Max
-6.604 -3.952 -2.654  0.989 87.197

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   20.482      3.365   6.088 2.27e-09 ***
rm            -2.684      0.532  -5.045 6.35e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.401 on 504 degrees of freedom
Multiple R-squared:  0.04807,   Adjusted R-squared:  0.04618
F-statistic: 25.45 on 1 and 504 DF,  p-value: 6.347e-07
```

The fitted regression model4 eq will be- crime rate = b4 + m4 * rm, here b4 = 20.482 and m4 = -2.684

(v) Fit simple linear regression model for response variable crime rate and predictor dis-

```
> model5.fit <- lm(crim ~ dis , Data_set)
> summary(model5.fit)

Call:
lm(formula = crim ~ dis, data = Data_set)

Residuals:
   Min     1Q Median     3Q    Max
-6.708 -4.134 -1.527  1.516 81.674

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.4993     0.7304  13.006   <2e-16 ***
dis          -1.5509     0.1683  -9.213   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.965 on 504 degrees of freedom
Multiple R-squared:  0.1441,    Adjusted R-squared:  0.1425
F-statistic: 84.89 on 1 and 504 DF,  p-value: < 2.2e-16
```

The fitted regression model5 eq will be- crime rate = b5 + m5 * dis, here b5 = 9.4993 and m5 = -1.5509

(vi) Fit simple linear regression model for response variable crime rate and predictor tax-

```
> model6.fit <- lm(crim ~ tax , Data_set)
> summary(model6.fit)

Call:
lm(formula = crim ~ tax, data = Data_set)

Residuals:
    Min      1Q  Median      3Q     Max
-12.513  -2.738  -0.194   1.065  77.696

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.528369   0.815809  -10.45   <2e-16 ***
tax          0.029742   0.001847   16.10   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.997 on 504 degrees of freedom
Multiple R-squared:  0.3396,    Adjusted R-squared:  0.3383
F-statistic: 259.2 on 1 and 504 DF,  p-value: < 2.2e-16
```

The fitted regression model6 eq will be- crime rate = b6 + m6 * tax, here b6 = -8.528369 and m6 = 0.029742

(vii) Fit simple linear regression model for response variable crime rate and predictor medv-

```
> model7.fit <- lm(crim ~ medv , Data_set)
> summary(model7.fit)

Call:
lm(formula = crim ~ medv, data = Data_set)

Residuals:
   Min     1Q Median     3Q    Max
-9.071 -4.022 -2.343  1.298 80.957

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.79654    0.93419   12.63   <2e-16 ***
medv        -0.36316    0.03839   -9.46   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.934 on 504 degrees of freedom
Multiple R-squared:  0.1508,    Adjusted R-squared:  0.1491
F-statistic: 89.49 on 1 and 504 DF,  p-value: < 2.2e-16
```

The fitted regression model7 eq will be- crime rate = b7 + m7 * medv, here b7 = 11.79654 and m7 = -0.36316

—----------

To determine whether there is a statistically significant association between a predictor variable and a response variable in models we will perform hypothesis testing. For testing the association between a predictor and a response, the null hypothesis usually states that there is no association (the coefficient for the predictor is zero), while the alternative hypothesis states that there is a nonzero association.

Null Hypothesis (H0): There is no association between the predictor and the response. Alternative Hypothesis (H1): There is an association between the predictor and the response.

Use significance level α = 0.05

For model1:

Testing Hypothisis: For b1 (intercept)

H01 : b1 = 0 vs. H11 : b1 != 0

p − value < 2 × 10^−16

i.e., p = value < α and hence H01 is rejected.

Testing Hypothisis: For m1 (slop)

H01 : m1 = 0 vs. H11 : m1 != 0

p−value < 5.51 × 10^−6

i.e., p-value < α and hence H01 is rejected.

For model2:

Testing Hypothisis: For b2 (intercept)

H02 : b2 = 0 vs. H12 : b2 != 0

p−value = 0.00209

i.e., p-value < α and hence H02 is rejected.

Testing Hypothisis: For m2 (slop)

H02 : m2 = 0 vs. H12 : m2 != 0

p−value < 2 × 10^−16

i.e., p-value < α and hence H02 is rejected.

For model3:

Testing Hypothisis: For b3 (intercept)

H03 : b3 = 0 vs. H12 : b3 != 0

p−value < 2 × 10^−16

i.e., p-value < α and hence H03 is rejected.

Testing Hypothisis: For m3 (slop)

H03 : m3 = 0 vs. H13 : m3 != 0

p−value = 5.51 x 10^-6

i.e., p-value < α and hence H03 is rejected.

For model4:

Testing Hypothisis: For b4 (intercept)

H04 : b4 = 0 vs. H14 : b3 != 0

p−value < 2.27 × 10^−9

i.e., p-value < α and hence H04 is rejected.

Testing Hypothisis: For m4 (slop)

H04 : m4 = 0 vs. H14 : m4 != 0

p−value = 6.35 x 10^-7

i.e., p-value < α and hence H04 is rejected.

For model5:

Testing Hypothisis: For b5 (intercept)

H05 : b5 = 0 vs. H15 : b5 != 0

p−value < 2 × 10^−16

i.e., p-value < α and hence H05 is rejected.

Testing Hypothisis: For m5 (slop)

H05 : m5 = 0 vs. H15 : m5 != 0

p−value = 2 x 10^-16

i.e., p-value < α and hence H05 is rejected.

For model6:

Testing Hypothisis: For b6 (intercept)

H06 : b6 = 0 vs. H16 : b6 != 0

p−value < 2 × 10^−16

i.e., p-value < α and hence H06 is rejected.

Testing Hypothisis: For m6 (slop)

H06 : m6 = 0 vs. H16 : m6 != 0

p−value = 2 x 10^-16

i.e., p-value < α and hence H06 is rejected.

For model7:

Testing Hypothisis: For b7 (intercept)

H07 : b7 = 0 vs. H17 : b7 != 0

p−value < 2 × 10^−16

i.e., p-value < α and hence H07 is rejected.

Testing Hypothisis: For m7 (slop)

H07 : m7 = 0 vs. H14 : m7 != 0

p−value = 2 x 10^-16

i.e., p-value < α and hence H07 is rejected.

Hence, We can see that each model there is a statistically significant association between the predictor and the response.

Check godness of Fit of models-

Model-1: RSE = 8.435, R2_adj = 0.03828 and R2 = 0.04019

Model-2: RSE = 7.866, R2_adj = 0.1637 and R2 = 0.1653,

Model-3: RSE = 8.435, R2_adj = 0.03828 and R2 = 0.04019,

Model-4: RSE = 8.401, R2_adj = 0.04618 and R2 = 0.04807

Model-5: RSE = 7.965, R2_adj = 0.1425 and R2 = 0.1441

Model-6: RSE = 6.997, R2_adj = 0.3383 and R2 = 0.3396

Model-7: RSE = 7.934, R2_adj = 0.1491 and R2 = 0.1508

So, Model-6 is the best among these seven models.

—----------

(b) Fit a multiple regression model to predict the response using all of the predictors.

The fitted regression model eq will be

crim = β0 + β1*zn + β2*indus + β3*nox + β4*rm + β5*dis + β6*tax + β7*medv

Summary of multiple regression model:

```
> multi_model <- lm(crim ~ zn+indus+nox+rm+dis+tax+medv, Data_set)
> summary(multi_model)

Call:
lm(formula = crim ~ zn + indus + nox + rm + dis + tax + medv,
    data = Data_set)

Residuals:
    Min      1Q  Median      3Q     Max
-12.573  -2.856  -0.577   1.302  74.911

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.417290   4.735683   0.510  0.60997
zn           0.039158   0.018539   2.112  0.03517 *
indus       -0.226798   0.083420  -2.719  0.00678 **
nox         -3.107230   4.934973  -0.630  0.52922
rm           0.596972   0.608758   0.981  0.32725
dis         -1.138369   0.278208  -4.092 4.99e-05 ***
tax          0.026635   0.002809   9.481  < 2e-16 ***
medv        -0.235403   0.051763  -4.548 6.82e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.795 on 498 degrees of freedom
Multiple R-squared:  0.3845,    Adjusted R-squared:  0.3759
F-statistic: 44.45 on 7 and 498 DF,  p-value: < 2.2e-16
```

R2 = 0.3845 (say), then it indicates the model explained 38% of the variation that is

there in the response. Thus, the fitted regression model is given by

crim = 2.417 + 0.039*zn - 0.226798*indus − 3.107230*nox + 0.596972*rm - 1.138369*dis + 0.026635*tax - 0.235403*medv

Note that p-value < 2.2 × 10−16 is very small and hence, we should reject the overall hypothesis H0 : β0 = β1 = β2 = β3 = β4 = β5 = β6=0

We have F0 = 44.45 i.e. model is not too good.

Test hypothesis on Individual Regression Coefficients:

Level-α-test with α = 0.05

Testing Problem 1 (For intercept): H01 : β0 = 0 vs. H11 : β0 != 0

p − value = 0.60997

i.e., p-value > α and hence H01 is accepted.

Testing Problem 2 (For coefficient of zn): H02 : β1 = 0 vs. H12 : β1 != 0

p − value = 0.03517

i.e., p-value < α and hence H02 is rejected.

Testing Problem 3 (For coefficient of indus): H03 : β2 = 0 vs. H13 : β2 != 0

p − value = 0.00678

i.e., p-value < α and hence H03 is rejected.

Testing Problem 4 (For coefficient of nox): H04 : β3 = 0 vs. H14 : β3 != 0

p − value = 0.52922

i.e., p-value > α and hence H04 is accepted.

Testing Problem 5 (For coefficient of rm): H05 : β4 = 0 vs. H15 : β4 != 0

p − value = 0.32725

i.e., p-value > α and hence H05 is accepted.

Testing Problem 6 (For coefficient of dis): H06 : β5 = 0 vs. H16 : β5 != 0

p − value = 4.99 X 10^-5

i.e., p-value < α and hence H06 is rejected.

Testing Problem 7 (For coefficient of tax): H07 : β6 = 0 vs. H17 : β6 != 0

p − value < 2e X 10^-16

i.e., p-value < α and hence H07 is rejected.

Testing Problem 8 (For coefficient of medv): H08 : β7 = 0 vs. H18 : β7 != 0

p − value = 6.82 X 10^-6

i.e., p-value < α and hence H08 is rejected.

Hence, for zn, indus, dis, tax ans medv predictors we can reject the null hypothesis.

—----------

(C) First of all we will check is there any outlier in our data or not by using Cook's Distance, DFDITS, DFBETA, COVRATIO.

Leverage points in our data set -

```
> Leverage_point_values
        58         65         67        121        122        123        124        125
0.03257538 0.03187811 0.02943673 0.04842189 0.04890818 0.04881315 0.04997225 0.04923036
       126        127        143        144        145        146        147        148
0.04878971 0.05075272 0.03496255 0.03512634 0.03927612 0.03516511 0.03465780 0.04046433
       149        150        151        152        153        154        155        156
0.03933303 0.03478728 0.03556228 0.03841873 0.04020408 0.03592916 0.03544033 0.03576184
       157        160        162        163        164        167        196        200
0.03709908 0.03763297 0.04043182 0.04112410 0.04649398 0.04178245 0.03202719 0.03113536
       201        202        204        205        226        253        254        257
0.03062041 0.02835231 0.04511526 0.04646273 0.03446388 0.02788501 0.05136878 0.03302948
       258        263        266        268        284        291        292        293
0.04268769 0.03719959 0.03008273 0.02964522 0.03789945 0.03173803 0.02933236 0.03179830
       352        353        354        355        356        365        366        368
0.03547234 0.03599164 0.05630926 0.03714091 0.03520699 0.05829266 0.08621304 0.05868991
       369        370        371        372        373        375        407
0.10628219 0.05397890 0.04904638 0.06332713 0.07293835 0.03313435 0.03134333
```

Influential Points in our data set : Detect using Cook's Distance, DFBETAS, DFFITS, COVRATIO -

```
> Influential_Points <- which(cook > cutoff)
> Influential_Points_values <- cook[Influential_Points]
> Influential_Points_values
       381        399        405        406        411        415        419        428
0.35360784 0.02011448 0.02103295 0.10510491 0.05058565 0.05590810 0.09961697 0.01638663
```

Detect using DFBETAS-

```
> DFBETAS_Points <- which(abs(def) > cutoff_)
> DFBETAS_Points_values <- abs(def)[DFBETAS_Points]
> DFBETAS_Points_values
  [1] 0.10087191 0.15899205 0.12707966 0.23223600 0.09828860 0.12279276 0.61940816 0.09549808
  [9] 0.12555279 0.09369549 0.12211434 0.15884931 0.32136545 0.13686767 0.21558832 0.37728215
 [17] 0.10692348 0.09730661 0.10779941 0.10630662 0.21078997 0.10456931 0.24761851 0.10463419
 [25] 0.13275412 0.14628877 0.13065373 0.11502323 0.11318643 0.09922125 0.11786592 0.10311781
 [33] 0.10464834 0.22307935 0.18687395 0.14395281 0.17008762 0.19186447 0.17841147 0.19157216
 [41] 0.16885314 0.12416095 0.29903560 0.39735214 0.15244756 0.13281924 0.17865030 0.16379212
 [49] 0.14643285 0.10480949 0.28645887 0.13475292 0.10759500 0.14853057 0.09659675 0.10199022
 [57] 0.15617507 0.11713195 1.47859116 0.12402442 0.16961091 0.19519190 0.18673571 0.15218081
 [65] 0.18879957 0.45392954 0.29709997 0.15899005 0.15963915 0.17074348 0.20320405 0.50404282
 [73] 0.14554614 0.12105279 0.36160941 0.34082992 0.12348900 0.15224940 0.23695866 0.11326723
 [81] 0.10238045 0.10119925 0.09125607 0.10472920 0.33074522 0.10152224 0.14325402 0.21690381
 [89] 0.34430212 0.17092424 0.20804789 0.31608836 0.12571701 0.09159583 0.13484183 0.10861910
 [97] 0.09092688 0.10578498 0.16472548 0.15468498 1.35088332 0.21534689 0.13238849 0.15182814
[105] 0.61720785 0.54718333 0.19658882
```

Detect using DFFITS-

```
> DFFITS_Points <- which(abs(dff) > cutoff)
> DFFITS_Points_values <- abs(dff)[DFFITS_Points]
> DFFITS_Points_values
      354       365       381       387       399       405       406       411       414
0.2850567 0.3119599 1.9396272 0.2384518 0.4059530 0.4171136 0.9797536 0.6574881 0.3278561
      415       419       428       489       490       491       492
0.6845059 0.9724557 0.3666508 0.2529445 0.3057590 0.2927567 0.2682802
```

A measure of model performance: COVRATIO

```
> outliers <- which(cv < l | cv > u)
> outlier_values <- cv[outliers]
> outlier_values
       57        58        67       121       122       123       124       125       126
1.0434526 1.0498043 1.0415028 1.0599996 1.0623366 1.0623284 1.0658118 1.0640576 1.0617841
      127       143       144       145       146       147       148       149       150
1.0671781 1.0524343 1.0532012 1.0564569 1.0510258 1.0514523 1.0582282 1.0575489 1.0522436
      151       152       153       154       155       156       157       160       162
1.0532227 1.0561578 1.0565047 1.0538090 1.0516166 1.0538114 1.0543331 1.0555897 1.0523521
      163       164       167       196       200       201       203       204       205
1.0522410 1.0598009 1.0525926 1.0473120 1.0465842 1.0448306 1.0440956 1.0620150 1.0630658
      226       254       255       256       257       258       262       263       266
1.0495948 1.0559746 1.0425732 1.0440888 1.0480233 1.0601112 1.0442885 1.0534956 1.0458275
      268       284       285       287       291       292       293       346       347
1.0447073 1.0499626 1.0437167 1.0419718 1.0489746 1.0467129 1.0490379 1.0417893 1.0419215
      348       352       353       354       355       356       365       366       368
1.0420992 1.0535830 1.0538272 1.0535305 1.0553819 1.0530895 1.0521909 1.1080768 1.0722724
      369       370       371       372       373       381       399       405       406
1.1364966 1.0738710 1.0678576 1.0766657 1.0901096 0.1045145 0.8368439 0.7727634 0.3512559
      411       414       415       419       428
0.5971289 0.9318098 0.7024977 0.2569325 0.8257339
```

Hence, in our data outlier present.

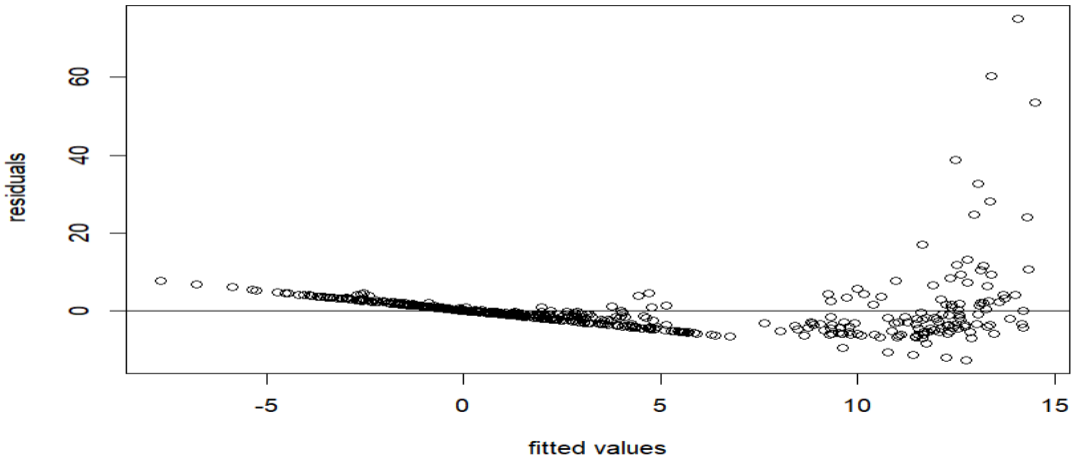
—----------

Model Assumptions:

(i) The error term has zero mean:

```
> error <- multi_model$residuals
> mean(error)
[1] 3.833105e-16
```

(ii) Check the error term has constant variance or not: Plot graph between residuals and fitted values.



By above graph we can see that error term has not constant variance.

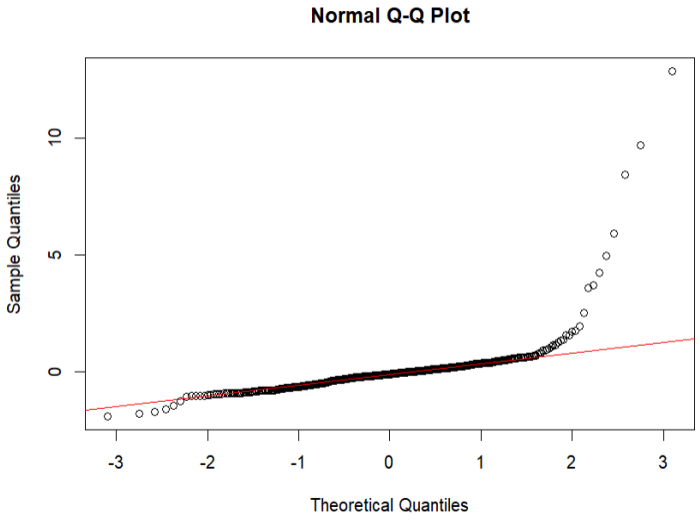
(iii) Check Uncorreleated error assumption: Use DurbinWatsonTest for this.

```
> DurbinWatsonTest(multi_model, alternative = "less")

        Durbin-Watson test

data:  multi_model
DW = 1.3535, p-value = 1
alternative hypothesis: true autocorrelation is less than 0

> DurbinWatsonTest(multi_model, alternative = "greater")

        Durbin-Watson test

data:  multi_model
DW = 1.3535, p-value = 2.531e-14
alternative hypothesis: true autocorrelation is greater than 0

> DurbinWatsonTest(multi_model, alternative = "two.sided")

        Durbin-Watson test

data:  multi_model
DW = 1.3535, p-value = 5.061e-14
alternative hypothesis: true autocorrelation is not 0
```

The Durbin-Watson test statistic (DW) is approximately 1.3535, and the p-value is very small (close to zero) for all three alternative hypotheses:Alternative = "less": The p-value is 1. This indicates that there is no evidence to suggest that the autocorrelation is less than 0 (i.e., negative autocorrelation). Alternative = "greater": The p-value is approximately 2.531 x 10^-14. This indicates strong evidence to suggest that the autocorrelation is greater than 0 (i.e., positive autocorrelation). Alternative = "two.sided": The p-value is also very small, approximately 5.061 x 10^-14. This indicates strong evidence to suggest that the autocorrelation is not equal to 0.

Based on these results, we can conclude that there is strong evidence of positive autocorrelation in the errors of the regression model. This means that consecutive residuals are correlated with each other, violating the assumption of independence of errors.

(iv) Normality assumption: Normality check by using Q-Q plot for studentized residuals-



By graph we can see that distribution of the studentized residual positive skew. It indicates that the residuals are not normally distributed.

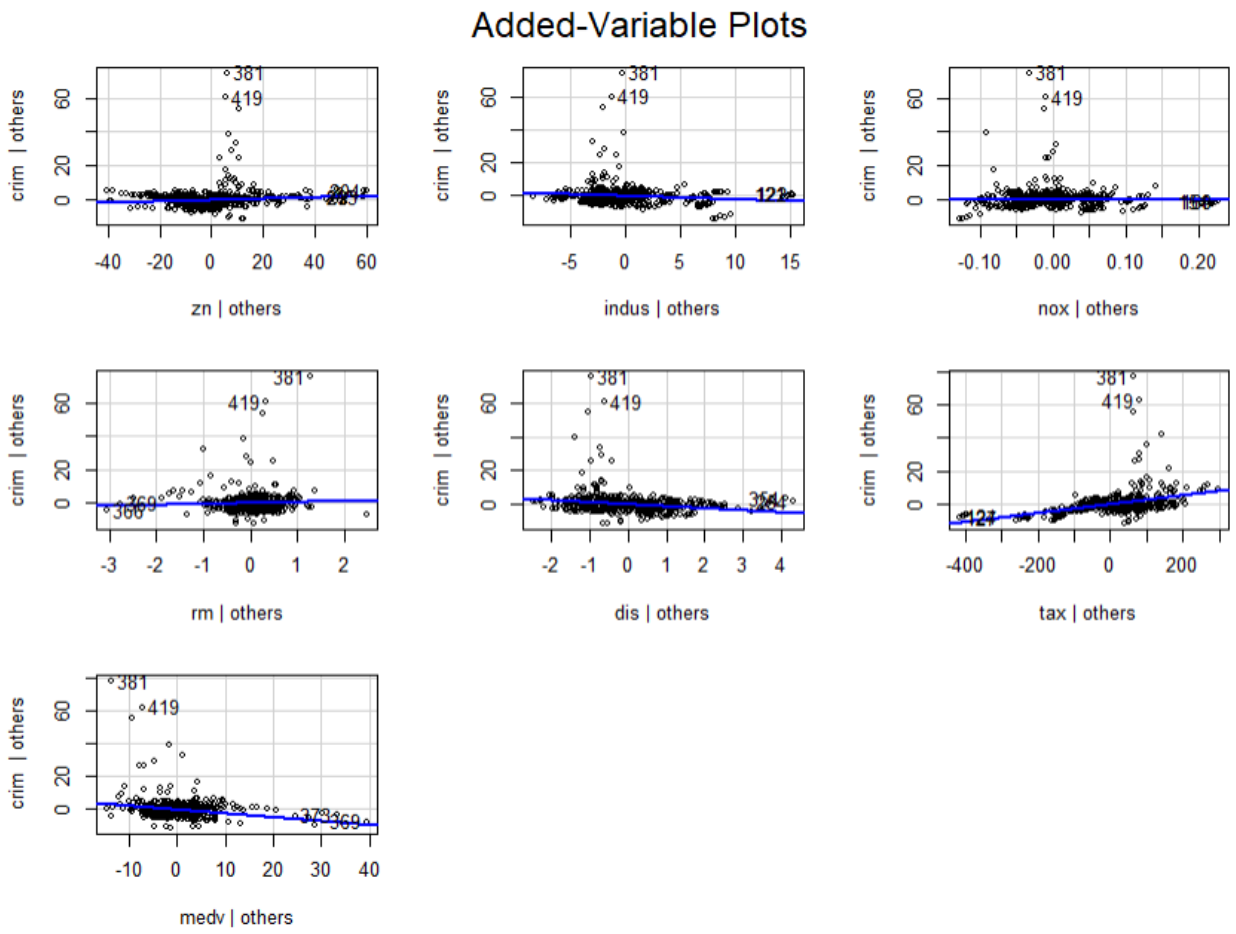
(v) Check the Multicollinearity-

The minimum eigenvalue of variance-covariance matrix is approximately 0, it is indicating that one or more variables are linear combinations of other variables, suggesting multicollinearity.

We can also check by VIF:

```
> library(mctest)
> imcdiag(multi_model, method = "VIF")

Call:
imcdiag(mod = multi_model, method = "VIF")


 VIF Multicollinearity Diagnostics

          VIF detection
zn     2.0446         0
indus  3.5818         0
nox    3.5763         0
rm     2.0008         0
dis    3.7533         0
tax    2.4517         0
medv   2.4786         0

NOTE:  VIF Method Failed to detect multicollinearity


0 --> COLLINEARITY is not detected by the test
```

(vi) Check for linearity assumption:



By above plot we can see that the relationship btw response variable and predictor variable are linear so linearity assumption hold.

Fix non-normality and non-constant variance problem be transformation of response variable: Use Box-Cox method

```
> library(car)
> p <- powerTransform(crim ~ zn+indus+nox+rm+dis+tax+medv, family = "bcPower",Data_set)
> summary(p)
bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1   -0.0405       -0.04      -0.0746      -0.0064

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
                           LRT df     pval
LR test, lambda = (0) 5.528272  1 0.018712

Likelihood ratio test that no transformation is needed
                           LRT df       pval
LR test, lambda = (1) 2786.602  1 < 2.22e-16
```

The estimated transformation parameter is $\lambda = -0.0405$

After applying the Box-Cox transformation, the updated data has approximately normal distribution with constant variance.

```
> new_MLR_model <- lm(bcPower(crim, p$roundlam, jacobian.adjusted = TRUE)~ zn+indus+nox+rm+dis+tax+
medv, Data_set)
> summary(new_MLR_model)

Call:
lm(formula = bcPower(crim, p$roundlam, jacobian.adjusted = TRUE) ~
    zn + indus + nox + rm + dis + tax + medv, data = Data_set)

Residuals:
     Min       1Q   Median       3Q      Max
-1.79179 -0.27606  0.02866  0.31543  1.04643

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.9110237  0.3016666  -9.650  < 2e-16 ***
zn          -0.0074602  0.0011810  -6.317 5.92e-10 ***
indus       -0.0063652  0.0053139  -1.198    0.232
nox          2.6853020  0.3143615   8.542  < 2e-16 ***
rm          -0.0017209  0.0387784  -0.044    0.965
dis         -0.0290243  0.0177221  -1.638    0.102
tax          0.0031805  0.0001790  17.773  < 2e-16 ***
medv         0.0002168  0.0032973   0.066    0.948
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4329 on 498 degrees of freedom
Multiple R-squared:  0.8062,    Adjusted R-squared:  0.8034
F-statistic: 295.9 on 7 and 498 DF,  p-value: < 2.2e-16
```

R2 = 0.8062 (say), then it indicates the model explained 81% of the variation that is

there in the response. Thus, the fitted regression model is given by

crim = -2.9110237 + -0.0074602*zn - 0.0063652*indus + 2.6853020*nox - 0.0017209*rm - 0.0290243*dis + 0.0031805*tax + 0.0002168*medv

Note that p-value < 2.2 × 10−16 is very small and hence, we should reject the overall hypothesis H0 : β0 = β1 = β2 = β3 = β4 = β5 = β6=0

—----------END—----------