MDTS4214_739_ Multiple Linear Regression

Debkanta Ghosh

2026-02-18

PROBLEM SET 3

PROBLEM 3:- Problem to demonstrate the role of qualitative (ordinal) predictors in addition to quantitative predictors in multiple linear regression

Consider “diamonds” data set in R. It is in the ggplot2 package. Make a list of all the ordinal categorical variables. Identify the response.

```
rm(list=ls())
library(ggplot2)
attach(diamonds)
head(diamonds)

## # A tibble: 6 × 10
##   carat cut       color clarity depth table price     x     y     z
##   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  0.23 Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
## 2  0.21 Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
## 3  0.23 Good      E     VS1      56.9    65   327  4.05  4.07  2.31
## 4  0.29 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
## 5  0.31 Good      J     SI2      63.3    58   335  4.34  4.35  2.75
## 6  0.24 Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48

dim(diamonds)

## [1] 53940    10

#HERE RESPONSE IS PRICE
```

(a) Run a linear regression of the response on the quality of cut. Write the fitted regression model.

```
fit1=lm(price~relevel(factor(as.character(cut)),ref="Ideal"),data=diamonds)
summary(fit1)

##
## Call:
## lm(formula = price ~ relevel(factor(as.character(cut)), ref = "Ideal"),
##     data = diamonds)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4258  -2741  -1494   1360  15348
```

```
##
## Coefficients:
##                                                               Estimate Std.
Error
## (Intercept)                                                    3457.54
27.00
## relevel(factor(as.character(cut)), ref = "Ideal")Fair          901.22
102.41
## relevel(factor(as.character(cut)), ref = "Ideal")Good          471.32
62.70
## relevel(factor(as.character(cut)), ref = "Ideal")Premium      1126.72
43.22
## relevel(factor(as.character(cut)), ref = "Ideal")Very Good    524.22
45.05
##                                                               t value
Pr(>|t|)
## (Intercept)                                                   128.051  < 2e-
16 ***
## relevel(factor(as.character(cut)), ref = "Ideal")Fair           8.800  < 2e-
16 ***
## relevel(factor(as.character(cut)), ref = "Ideal")Good           7.517  5.7e-
14 ***
## relevel(factor(as.character(cut)), ref = "Ideal")Premium       26.067  < 2e-
16 ***
## relevel(factor(as.character(cut)), ref = "Ideal")Very Good     11.636  < 2e-
16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3964 on 53935 degrees of freedom
## Multiple R-squared:  0.01286,    Adjusted R-squared:  0.01279
## F-statistic: 175.7 on 4 and 53935 DF,  p-value: < 2.2e-16
```

$$\widehat{\text{price}} = 3457.54 + 901.22\, I_1 + 471.32\, I_2 + 1126.72\, I_3 + 524.22\, I_4$$

$$I_1 = \begin{cases} 1, & \text{if cut = Fair} \\ 0, & \text{otherwise} \end{cases}$$

$$I_2 = \begin{cases} 1, & \text{if cut = Good} \\ 0, & \text{otherwise} \end{cases}$$

$$I_3 = \begin{cases} 1, & \text{if cut = Premium} \\ 0, & \text{otherwise} \end{cases}$$

$$I_4 = \begin{cases} 1, & \text{if cut = Very Good} \\ 0, & \text{otherwise} \end{cases}$$

(b) Test whether the expected price of diamond with premium cut is significantly different from that of the ideal cut.

ANS:-There is strong statistical evidence that the expected price of a diamond with a Premium cut is significantly different from that of an Ideal cut. Specifically, Premium cut diamonds have a significantly higher expected price than Ideal cut diamonds.

(c) What is the expected price of a diamond of ideal cut?

ANS:-The expected price of a diamond with an Ideal cut is $3457.54

(d) Modify the regression model in (a) by incorporating the predictor “table”. Write the fitted regression model.

```
fit2=lm(price~relevel(factor(as.character(cut)),ref="Ideal")+table,data=diamo
nds)
summary(fit2)

## 
## Call:
## lm(formula = price ~ relevel(factor(as.character(cut)), ref = "Ideal") + 
##     table, data = diamonds)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
##  -5630  -2694  -1458   1346  15690 
## 
## Coefficients:
##                                                               Estimate Std. 
Error
## (Intercept)                                                  -6563.672   
517.450
## relevel(factor(as.character(cut)), ref = "Ideal")Fair          345.611   
106.002
## relevel(factor(as.character(cut)), ref = "Ideal")Good          -19.957    
67.426
## relevel(factor(as.character(cut)), ref = "Ideal")Premium       626.220    
50.215
## relevel(factor(as.character(cut)), ref = "Ideal")Very Good     165.206    
48.562
## table                                                          179.105     
9.236
##                                                              t value 
Pr(>|t|)
## (Intercept)                                                  -12.685  < 2e-
16 ***
## relevel(factor(as.character(cut)), ref = "Ideal")Fair          3.260 
0.001113 ** 
## relevel(factor(as.character(cut)), ref = "Ideal")Good         -0.296 
0.767246    
## relevel(factor(as.character(cut)), ref = "Ideal")Premium      12.471  < 2e-
16 ***
## relevel(factor(as.character(cut)), ref = "Ideal")Very Good     3.402 
0.000669 ***
## table                                                         19.393  < 2e-
16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 3950 on 53934 degrees of freedom
## Multiple R-squared:  0.0197, Adjusted R-squared:  0.01961
## F-statistic: 216.7 on 5 and 53934 DF,  p-value: < 2.2e-16
```

$$\widehat{\text{price}} = -6563.672 + 345.611\, I_1 - 19.957\, I_2 + 626.220\, I_3 + 165.206\, I_4 + 179.105\, T$$

$$\begin{aligned}
I_1 &= \begin{cases} 1, & \text{if cut = Fair} \\ 0, & \text{otherwise} \end{cases} \\
I_2 &= \begin{cases} 1, & \text{if cut = Good} \\ 0, & \text{otherwise} \end{cases} \\
I_3 &= \begin{cases} 1, & \text{if cut = Premium} \\ 0, & \text{otherwise} \end{cases} \\
I_4 &= \begin{cases} 1, & \text{if cut = Very Good} \\ 0, & \text{otherwise} \end{cases}
\end{aligned}$$

$$T = \text{table (percentage of the diamond's width)}$$

(e) Test for the significance of “table” in predicting the price of diamond.

ANS:- There is strong statistical evidence that table is a significant predictor of the price of a diamond. Specifically, holding cut constant, an increase of one unit in table is associated with an average increase of approximately 179.1 units in the diamond’s price.

(f) Find the average estimated price of a diamond with an average table value and which is of fair cut.

$$\widehat{\text{price}} = -6563.672 + 345.611\, I_1 - 19.957\, I_2 + 626.220\, I_3 + 165.206\, I_4 + 179.105\, T$$

$$\begin{aligned}
\widehat{\text{price}} &= -6563.672 + 345.611 + 179.105\, \bar{T} \\
&= -6218.061 + 179.105\, \bar{T}
\end{aligned}$$

PROBLEM 5:-5 Problem to demonstrate the utility of nonlinear regression over linear regression

Get the fgl data set from “MASS” library.

```
rm(list=ls())
library(MASS)
df=fgl
attach(fgl)
head(df)

##      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
```

(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.

```
df.1=fgl[fgl$type=="Veh",]
df.1$type = NULL
fit1=lm(RI~.,data=df.1)
summary(fit1)

## 
## Call:
## lm(formula = RI ~ ., data = df.1)
## 
## 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
```

Fe best explains the refractive index.

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

```
fit2=lm(RI~Fe,data=df.1)
summary(fit2)

## 
## Call:
## lm(formula = RI ~ Fe, data = df.1)
## 
## 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
```

(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).

```
fit3=lm(RI~Fe+I(Fe^2),data=df.1)
summary(fit3)

##
## Call:
## lm(formula = RI ~ Fe + I(Fe^2), data = df.1)
##
## 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

fit4=lm(RI~Fe+I(Fe^2)+I(Fe^3),data=df.1)
summary(fit4)

##
## Call:
## lm(formula = RI ~ Fe + I(Fe^2) + I(Fe^3), data = df.1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6306 -1.1806 -0.0695  0.5621  3.5394
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.2694     0.4921  -0.548    0.593
## Fe           -16.7947    32.2946  -0.520    0.612
## I(Fe^2)      107.1214   268.4871   0.399    0.696
## I(Fe^3)      -79.0070   510.0359  -0.155    0.879
##
## Residual standard error: 1.705 on 13 degrees of freedom
## Multiple R-squared:  0.3566, Adjusted R-squared:  0.2081
## F-statistic: 2.402 on 3 and 13 DF,  p-value: 0.1146
```

Quadratic is giving substancial improvement over linear regression

Cubic is slight improvement over quadratic so we choose quadratic regression

PROBLEM SET 4

PROBLEM SET 4:- 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.

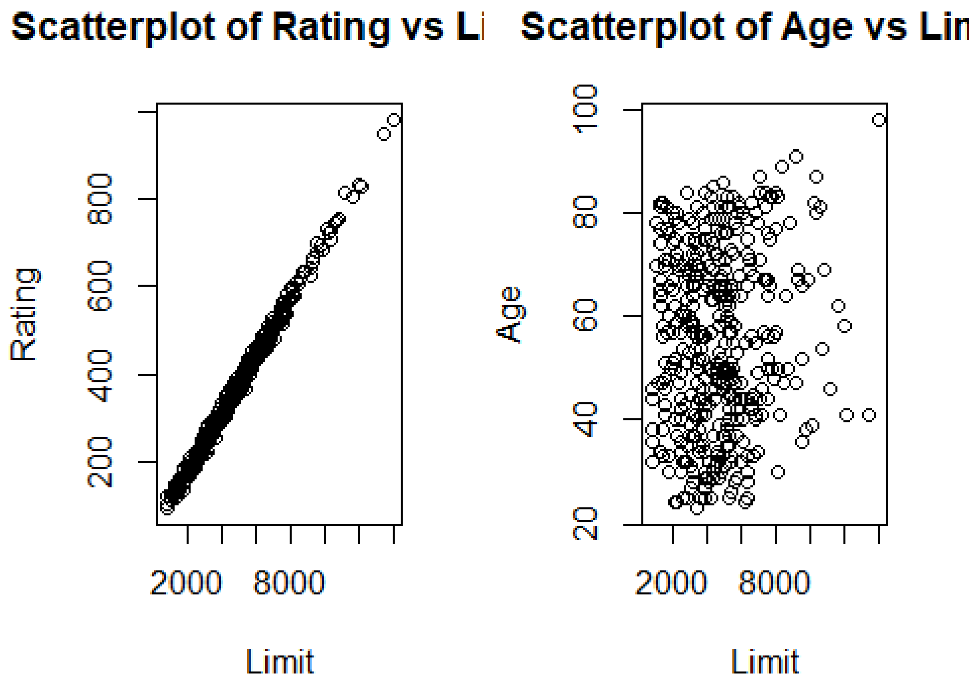
```
rm(list=ls())
library(ISLR)
attach(Credit)
head(Credit)

##   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
```

```
df=Credit[,c(3,4,6,12)]
head(df)
```

```
##   Limit Rating Age Balance
## 1  3606    283  34     333
## 2  6645    483  82     903
## 3  7075    514  71     580
## 4  9504    681  36     964
## 5  4897    357  68     331
## 6  8047    569  77    1151
```

(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))
plot(Limit,Rating,main="Scatterplot of Rating vs Limit")
plot(Limit,Age,main="Scatterplot of Age vs Limit")
```



Comment:

- **Rating vs Limit:**
  The scatterplot shows a very strong positive linear relationship between rating and credit limit. As rating increases, the credit limit increases almost proportionally, with points closely clustered around a straight line. This indicates that rating is a strong predictor of credit limit.

- **Age vs Limit:**
  The scatterplot shows no clear linear relationship between age and credit limit. The points are widely scattered, suggesting that age has little or no influence on the credit limit.

Conclusion:
Credit limit is strongly associated with rating, whereas age does not appear to be an important predictor of credit 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?

```
fit1=lm(Balance~Age+Limit)
fit2=lm(Balance~Rating+Age+Limit)
fit3=lm(Balance~Rating+Limit)
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

stargazer(fit1,fit2,fit3,type="text",out="f2.txt")

##
##
=============================================================================
=================
##                                                Dependent variable:
##                     ------------------------------------------------------
--------------------
##                                                      Balance
##                               (1)                      (2)
(3)
## --------------------------------------------------------------------------
--------------------
## Rating                                               2.310**
2.202**
##                                                      (0.940)
(0.952)
##
## Age                        -2.291***                -2.346***
##                             (0.672)                  (0.669)
##
## 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
```

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

```
library(car)

## Loading required package: carData

vif(fit1)

##      Age    Limit
## 1.010283 1.010283

vif(fit2)

##     Rating        Age      Limit
## 160.668301   1.011385 160.592880

vif(fit3)

##   Rating    Limit
## 160.4933 160.4933
```

The VIF results clearly confirm the presence of multicollinearity.

In fit1, the VIF values for Age and Limit are approximately 1, which indicates no multicollinearity. This means the predictors in that model are essentially independent of each other.

However, in fit2 and fit3, the VIF values for Rating and Limit are extremely large (around 160). A VIF above 10 is already considered problematic, so values around 160 indicate severe multicollinearity. This happens because Rating and Limit are almost perfectly linearly related.

Thus, when both Rating and Limit are included in the model, they compete to explain the same variation in Balance, leading to unstable coefficient estimates and inflated standard errors. This explains why Limit becomes insignificant once Rating is added.

Overall, the VIF results strongly support the earlier conclusion that Rating and Limit should not be included together in the same regression model.

PROBLEM 2:-Problem to demonstrate the detection of out- lier, 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 popu- lation as predictors. The objective is to fit a multiple linear regression model of the response on the predictors. With reference to this problem, detect outliers, leverage points and influential points if any.

```
#Attaching the Boston Data
rm(list=ls())
library(MASS)
attach(Boston)
df=data.frame(medv,crim,black,nox,lstat)
head(df)
```

```
##   medv    crim  black   nox lstat
## 1 24.0 0.00632 396.90 0.538  4.98
## 2 21.6 0.02731 396.90 0.469  9.14
## 3 34.7 0.02729 392.83 0.469  4.03
## 4 33.4 0.03237 394.63 0.458  2.94
## 5 36.2 0.06905 396.90 0.458  5.33
## 6 28.7 0.02985 394.12 0.458  5.21
```

```
model=lm(medv~.,data=df)
summary(model)
```

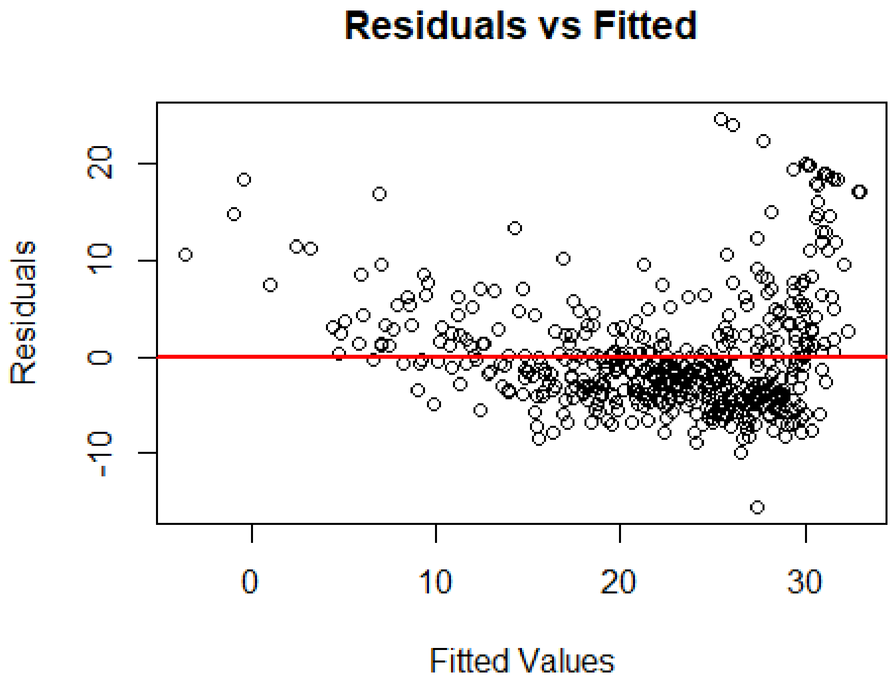
```
##
## Call:
## lm(formula = medv ~ ., data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.564  -4.004  -1.504   2.178  24.608
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.053584   2.170839  13.844   <2e-16 ***
## crim        -0.059424   0.037755  -1.574    0.116
## black        0.006785   0.003408   1.991    0.047 *
## nox          3.415809   3.056602   1.118    0.264
## 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
```

Fitted Model:

$$\widehat{medv} = 30.0536 - 0.059424\, crim + 0.006785\, black + 3.415809\, nox - 0.918431\, lstat$$

We now draw the **residual plot**

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



Comment:

From the residual plot alone we can say some outliers both in the positive and negative direction.

But from this plot we cannot comment on existence of leverage or influential points.

To find Potential Outliers:

We find out the standardized residuals from the fitted model.

A point is declared as a potential outlier if its standradized residual is greater than 2 or less than -2.

```
#Finding the standardized residuals
std.res=rstandard(model)
#Potential Outlier Detection
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
```

We can observe 31 data points which can be potentially outliers.

To find Leverge points

First, we find out the diagonal elements of the hat matrix. Now 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 hatvalues exceed the leverage value then we call the points potential leverages.

```
lev=hatvalues(model)

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

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

## [1] 0.02964427

# High leverage observations
leverage=which(lev>cutoff)
leverage

##  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
```

```r
length(leverage)
```

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

We can observe 29 potential leverage points.

To find Influential points

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

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

```r
cook=cooks.distance(model) #Calculating the Di values
influential=which(cook>1)
length(influential)
```

```
## [1] 0
```

In this model no value of Di exceeds one. So we can conclude that there exists no influential point.