ASSIGNMENT 4

SRIJITA DEY

2026-02-19

Problem Set 3

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).

```
rm(list=ls())
library(MASS)

## Warning: package 'MASS' was built under R version 4.5.2

attach(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

df3=fgl[fgl$type=="Veh",]
df3$type = NULL
head(df3)

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

```r
#a
fit1=lm(RI~.,data=df3)
summary(fit1)
```

```
##
## Call:
## lm(formula = RI ~ ., data = df3)
##
## 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
```

The Fitted Model here is:

$$\widehat{RI} = 131.4641 - 0.4333\,Na - 0.2866\,Mg - 0.8909\,Al - 1.8824\,Si - 2.4232\,K + 1.5326\,Ca + 0.3517\,Ba + 3.8931\,Fe$$

Here the R_squared = 0.9906, so the fit is very good.

The p values indicate that Fe serves as the most significant predictor in the multiple linear regression of Refractive Index against all continuous predictors, as its p-value of 0.00362 is the smallest compared to the other p-values

```r
#b
fit2=lm(RI~Fe,data=df3)
summary(fit2)
```

```
##
## Call:
## lm(formula = RI ~ Fe, data = df3)
##
## 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
```

The Fitted Model is

$$\widehat{RI} = -0.5007 + 8.1362\,Fe$$

The fit is not a very good one since the multiple R-squared is 0.2097.

```r
#c
fit3=lm(RI~Fe+I(Fe^2),data=df3)
summary(fit3)
```

```
##
## Call:
## lm(formula = RI ~ Fe + I(Fe^2), data = df3)
##
## 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
```

Here the Fitted Model is

$$\widehat{RI} = -0.2785 - 12.1810\,Fe + 65.9600\,Fe^2$$

The fit is not so good but it shows significant clear improvement from linear regression since the multipled R_square here is 0.3554.

```
fit4=lm(RI~Fe+I(Fe^2)+I(Fe^3),data=df3)
summary(fit4)
```

```
##
## Call:
## lm(formula = RI ~ Fe + I(Fe^2) + I(Fe^3), data = df3)
##
## 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
```

The Fitted Model here is

$$\widehat{RI} = -0.2694 - 16.7947\,Fe + 107.1214\,Fe^2 - 79.0070\,Fe^3$$

Here multiple R_squared is 0.3566.

Conclusion:

Quadratic regression offers significant enhancement compared to linear regression, while cubic regression provides a minor improvement over quadratic; thus, we select quadratic regression for betterment over the linear regression model. ## Problem Set 4 ##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?

```
rm(list=ls())
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 4.5.2
```

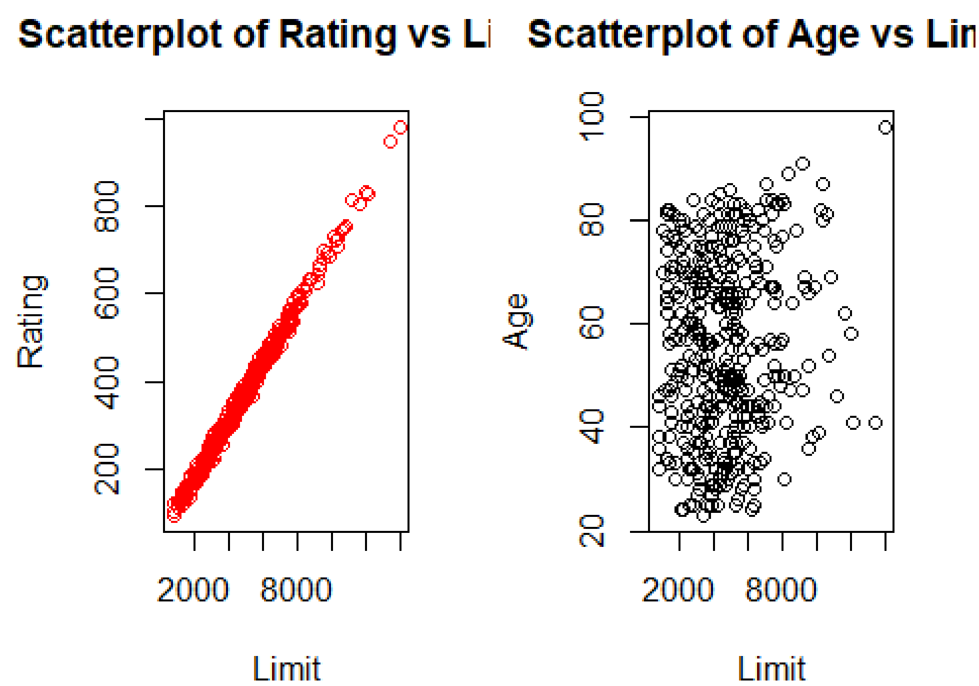
```
attach(Credit)
head(Credit)
```

```
##   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
par(mfrow=c(1,2))
plot(Limit,Rating,main="Scatterplot of Rating vs Limit",col="red")
plot(Limit,Age,main="Scatterplot of Age vs Limit")
```




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

Comment:

Rating vs Limit:

The scatterplot seems to show a very strong positive linear relationship between Rating and Limit. This suggests that when both variables are included in a regression model it may cause severe multicollinearity.

Age vs Limit:

The scatterplot seems to show a a very weak linear relationship between Age and Limit. The points are scattered without any clear trend.

```
#b
m1=lm(Balance~Age+Limit)
m2=lm(Balance~Rating+Age+Limit)
m3=lm(Balance~Rating+Limit)
library(stargazer)

## Warning: package 'stargazer' was built under R version 4.5.2

##
## 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(m1,m2,m3,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
```

Marked difference observed:

In model (1), Limit is highly significant (0.173***). In model (2), Limit becomes statistically insignificant (0.019, not significant) when Rating is added,. In model (3), Limit remains insignificant.

At the same time, Rating is significant when included (in models 2 and 3).

This indicates that Rating absorbs the explanatory power of Limit. From the earlier scatterplot, Rating and Limit seem to be almost perfectly linearly related, so this is a clear case of multicollinearity. When both are included, the model fails to separately identify their individual effects.

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

```
library(car)

## Warning: package 'car' was built under R version 4.5.2

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.5.2

vif(m1)

##      Age    Limit
## 1.010283 1.010283

vif(m2)

##     Rating        Age      Limit
## 160.668301   1.011385 160.592880

vif(m3)

##   Rating    Limit
## 160.4933 160.4933
```

The VIF outcomes distinctly indicate the existence of multicollinearity.

In m1, the VIF values for Age and Limit are around 1, suggesting no multicollinearity. This indicates that the predictors in that model are fundamentally independent from one another.

In m2 and m3, the VIF values for Rating and Limit are notably high (approximately 160). A VIF exceeding 10 is seen as problematic, thus values near 160 suggest serious multicollinearity. This occurs due to the near-perfect linear relationship between Rating and Limit.

Consequently, when both Rating and Limit are part of the model, they vie to account for the same variation in Balance, resulting in unstable coefficient estimates and increased standard errors. This clarifies why Limit becomes irrelevant once Rating is included.

In general, the VIF findings strongly affirm the previous conclusion that Rating and Limit ought not to be included together in the same regression mode

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 of the response on the predictors. With reference to this problem, detect outliers, leverage points and influential points if any.

```r
rm(list=ls())
library(MASS)
attach(Boston)
df1=data.frame(medv,crim,black,nox,lstat)
head(df1)
```

```
##   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
```

```r
fit=lm(medv~.,data=df1)
summary(fit)
```

```
## 
## Call:
## lm(formula = medv ~ ., data = df1)
## 
## 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 ***
## ---
```

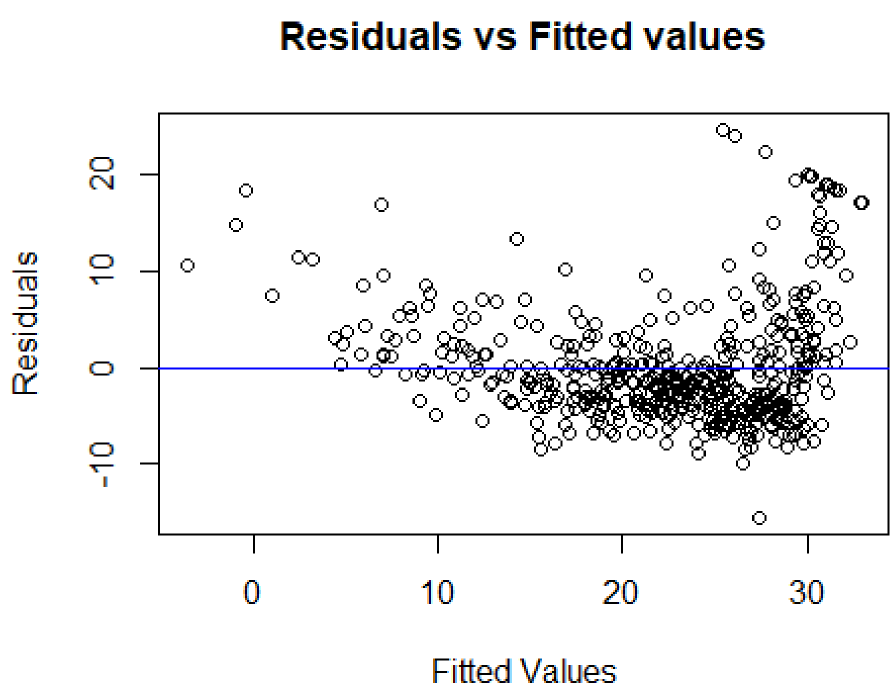
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.183 on 501 degrees of freedom
## Multiple R-squared:  0.5517, Adjusted R-squared:  0.5481
## F-statistic: 154.1 on 4 and 501 DF,  p-value: < 2.2e-16
```

Fitted Model:

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

The **residual plot** is

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



Comment:

We can comment from the residual plot that outliers are present both in the positive and negative direction but the residual plot is not sufficient to predict the presence of influential or leverage points.

To find Potential Outliers:

We find out the standardized residuals from the fitted model.

For a point to be a potential outlier its standradized residual must be either greater than 2 or less than -2.

```
std.res=rstandard(fit)
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 see 31 data points which can be potential outliers.

To find Leverge points

We obtain the diagonal elements of the Hat matrix. Then we obtain the 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 the points are called potential leverages.

```
le=hatvalues(fit)

n=nrow(df1)
p=4
cutoff=3*(p+1)/n
cutoff

## [1] 0.02964427

leverage=which(le>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

length(leverage)

## [1] 29
```

We obtain 29 potential leverage points.

To find Influential points

For this purpose we obtain 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.

```
cook=cooks.distance(fit)
influential_point=which(cook>1)
length(influential_point)

## [1] 0
```

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