MDTS4214 702 PS4

Sagnik Roy

2026-02-18

Problem Set - 3 (continued)

5.Problem to demonstrate the utility of non-linear regression over linear regression

Get the fgl data set from “MASS” library.

```
library(MASS)
attach(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 ...
```

(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.

```
summary(lm(RI~.,fgl[,-10]))
```

```
##
## Call:
## lm(formula = RI ~ ., data = fgl[, -10])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9898 -0.4273 -0.0264  0.4187  4.3833
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -64.73269   67.03935  -0.966  0.33539
## Na            1.39526    0.65505   2.130  0.03436 *
## Mg            1.84423    0.67547   2.730  0.00688 **
## Al            0.03262    0.69834   0.047  0.96278
## Si            0.16851    0.67740   0.249  0.80380
## K             1.38287    0.68998   2.004  0.04636 *
```

```
## Ca            3.11677    0.66837   4.663 5.61e-06 ***
## Ba            2.98281    0.67600   4.412 1.65e-05 ***
## Fe            0.42627    0.77869   0.547  0.58468
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.004 on 205 degrees of freedom
## Multiple R-squared:  0.8948, Adjusted R-squared:  0.8907
## F-statistic: 217.9 on 8 and 205 DF,  p-value: < 2.2e-16
```

The p-values are given by:

```
p=round(summary(lm(RI~.,fgl[,-10]))$coefficients[-1,4],5);p
```

```
##      Na      Mg      Al      Si       K      Ca      Ba      Fe
## 0.03436 0.00688 0.96278 0.80380 0.04636 0.00001 0.00002 0.58468
```

Interpretation

Based on the p-value, calcium oxide is the most statistically significant predictor in determining the refractive index of the glass.

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

```
summary(lm(RI~Ca))
```

```
##
## Call:
## lm(formula = RI ~ Ca)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1658 -0.9515 -0.1842  0.8354  6.8121
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15.12400    0.77868  -19.42   <2e-16 ***
## Ca            1.72932    0.08586   20.14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.783 on 212 degrees of freedom
## Multiple R-squared:  0.6568, Adjusted R-squared:  0.6551
## F-statistic: 405.6 on 1 and 212 DF,  p-value: < 2.2e-16
```

We obtain $R^2 = 0.6568$, indicating that, on average, calcium explains about 66% of the variability in the refractive index.

(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).

```r
summary(lm(RI~.,fgl))
```

```
##
## Call:
## lm(formula = RI ~ ., data = fgl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2131 -0.3786 -0.0186  0.3697  4.0317
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.68095   67.73022   0.172 0.863248
## Na           0.60928    0.66840   0.912 0.363100
## Mg           1.32017    0.67269   1.963 0.051086 .
## Al          -0.92790    0.70416  -1.318 0.189098
## Si          -0.61637    0.68460  -0.900 0.369022
## K            0.74007    0.68789   1.076 0.283286
## Ca           2.47150    0.66700   3.705 0.000273 ***
## Ba           1.97569    0.70198   2.814 0.005374 **
## Fe           0.36329    0.73879   0.492 0.623442
## typeWinNF    0.09996    0.17352   0.576 0.565209
## typeVeh     -0.88579    0.26111  -3.392 0.000835 ***
## typeCon      0.39910    0.39687   1.006 0.315810
## typeTabl     0.39263    0.42512   0.924 0.356822
## typeHead     1.58234    0.43890   3.605 0.000394 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9503 on 200 degrees of freedom
## Multiple R-squared:  0.9081, Adjusted R-squared:  0.9021
## F-statistic:   152 on 13 and 200 DF,  p-value: < 2.2e-16
```

From the regression results, we observe that calcium remains statistically significant at the 0.001 level. When fitting the full model, we obtain an $R^2 = 0.9081$, indicating that the model explains approximately 91% of the total variation in the refractive index of the window glass.

Comparing the adjusted values, model 2 has a higher adjusted R^2 (0.9021) than model 1 (0.6551). The adjusted R^2 is preferred for comparison here because it accounts for the number of predictors included in each model.

Problem Set 4

---

1 Problem to demonstrate multicollinearity.

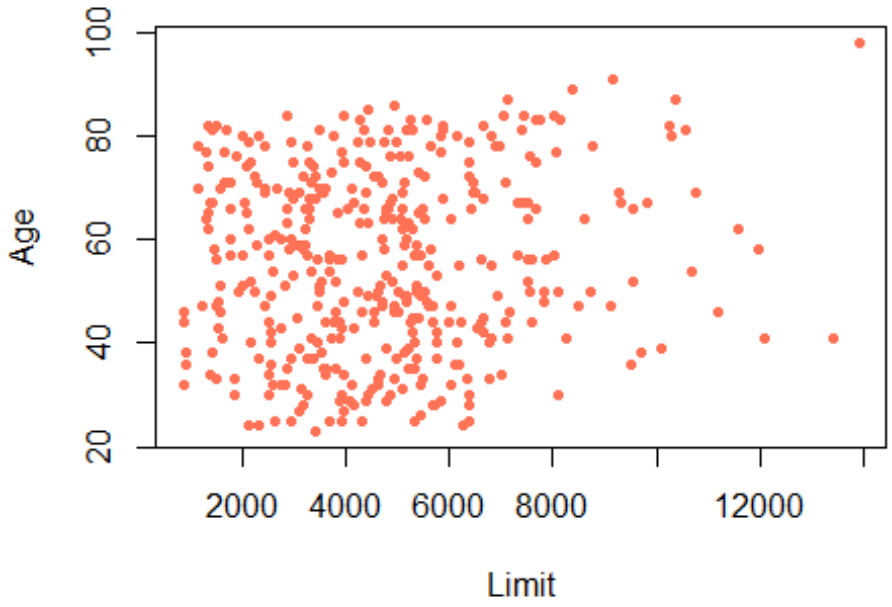
Consider the Credit data in the ISLR library. Choose Balance as the response and Age, Limit and Rating as the predictors.

```
library(ISLR)
library(stargazer)
attach(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 ...
```

(a) Make a scatter plot of (i) Age versus Limit

```
plot(Limit,Age,pch=20,col="coral1")
```

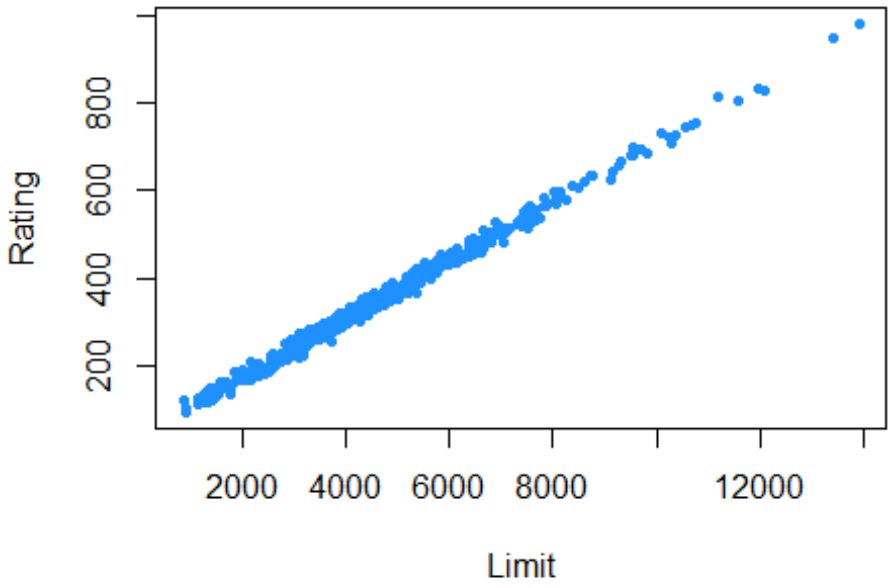


Comment

We do not observe any linear relationship

(ii) Rating Versus Limit. Comment on the scatterplot.

```
plot(Limit,Rating,pch=20,col="dodgerblue")
```



Comment

We observe a very strong positive linear relationship.

(b) Run three separate regressions:

(c) 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?

```
m1=lm(Balance~Age+Limit)
m2=lm(Balance~Age+Limit+Rating)
m3=lm(Balance~Limit+Rating)
stargazer(m1,m2,m3,type="text")
```

```
##
## ================================================================================
==============
##                                              Dependent variable:
##                     ---------------------------------------------------------
--------------
##                                                      Balance
##                                (1)                     (2)
(3)
## --------------------------------------------------------------------------------
--------------
## Age                          -2.291***               -2.346***
##                               (0.672)                 (0.669)
##
## Limit                         0.173***                 0.019                    0
.025
##                               (0.005)                 (0.063)                  (0
.064)
##
## Rating                                                2.310**                  2.
202**
##                                                       (0.940)                  (0
.952)
##
## Constant                    -173.411***             -259.518***              -377
.537***
##                               (43.828)                (55.882)                 (4
5.254)
##
## --------------------------------------------------------------------------------
--------------
## 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.

```r
r1=summary(lm(Age~Limit))$r.squared
r2=summary(lm(Rating~Age))$r.squared
r3=summary(lm(Limit~Rating))$r.squared
```

```r
R=c(r1,r2,r3)
vif=1/(1-R)
m=c("Age on Limit","Rating on Age","Limit on Rating")
data.frame("Model"=m,"R-square"=R,"VIF"=vif)
```

```
##             Model   R.square        VIF
## 1    Age on Limit 0.01017837   1.010283
## 2   Rating on Age 0.01064302   1.010758
## 3 Limit on Rating 0.99376921 160.493293
```

Comment

The Variance Inflation Factor (VIF) for the variables Limit and Rating is extremely large (far exceeding 10), indicating the presence of strong multicollinearity between these two predictors. In contrast, the pairs (Age, Limit) and (Age, Rating) show no evidence 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 of the response on the predictors. With reference to this problem, detect outliers, leverage points and influential points if any.

We take medv as the response variable and crim, nox, black, and lstat as the explanatory variables. The resulting fitted model is a multiple linear regression of the form:

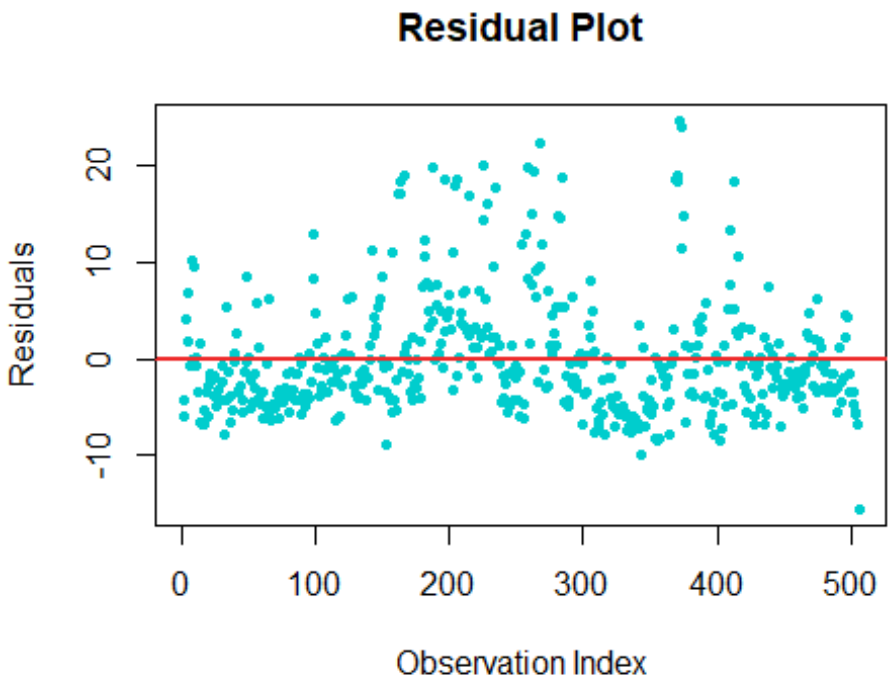
```r
library(MASS)
attach(Boston)
m4=lm(medv~crim+nox+black+lstat)
summary(m4)
```

```
##
## Call:
## lm(formula = medv ~ crim + nox + black + lstat)
##
## 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
```

The residual plot is given by:

```
plot(resid(m4),ylab="Residuals",xlab="Observation Index",main="Residual Plot",pch=20,col="cyan3")
abline(h=0,col="firebrick2",lty=1,lwd=2)
```



Comment

The residual plot does not display any clear systematic pattern; however, a few large positive residual values are observed.

We then compute the standardized residuals and identify observations with absolute values greater than 3 in order to detect potential outliers.

The following observations are outliers:

```
stdres=rstandard(m4)
outliers=which(abs(stdres)>3)
outliers
```

```
## 167 187 196 205 226 258 263 268 284 369 370 372 373 413
## 167 187 196 205 226 258 263 268 284 369 370 372 373 413
```

We compute the hat values to identify observations with high leverage.

The threshold value is:

```
t=2*5/nrow(Boston);t

## [1] 0.01976285
```

The following observations are the leverage points:

```
hat=hatvalues(m4)
lev=which(hat>t);lev

##   9  33  49 103 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157
##   9  33  49 103 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157
## 160 215 374 375 381 386 387 388 399 401 405 406 411 412 413 414 415 416 417 418
## 160 215 374 375 381 386 387 388 399 401 405 406 411 412 413 414 415 416 417 418
## 419 420 424 425 426 427 428 430 431 432 433 434 435 437 438 439 446 451 455 456
## 419 420 424 425 426 427 428 430 431 432 433 434 435 437 438 439 446 451 455 456
## 457 458 467 491
## 457 458 467 491
```

We calculate Cook's distance to identify the influential observations.

The threshold value is:

```
t2=4/nrow(Boston);t2

## [1] 0.007905138
```

The following observations are the influential points:

```
cd=cooks.distance(m4)
ip=which(cd>t2);ip

##   9  49 142 149 153 162 163 164 167 187 196 204 205 215 226 234 258 262 263 268
##   9  49 142 149 153 162 163 164 167 187 196 204 205 215 226 234 258 262 263 268
## 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
```