Predictive Analysis

Problem Set 4: Some Potential Problems in Multiple Linear Regression

Atrijo Roy

2026-02-19

1. Problem to demonstrate multicollinearity

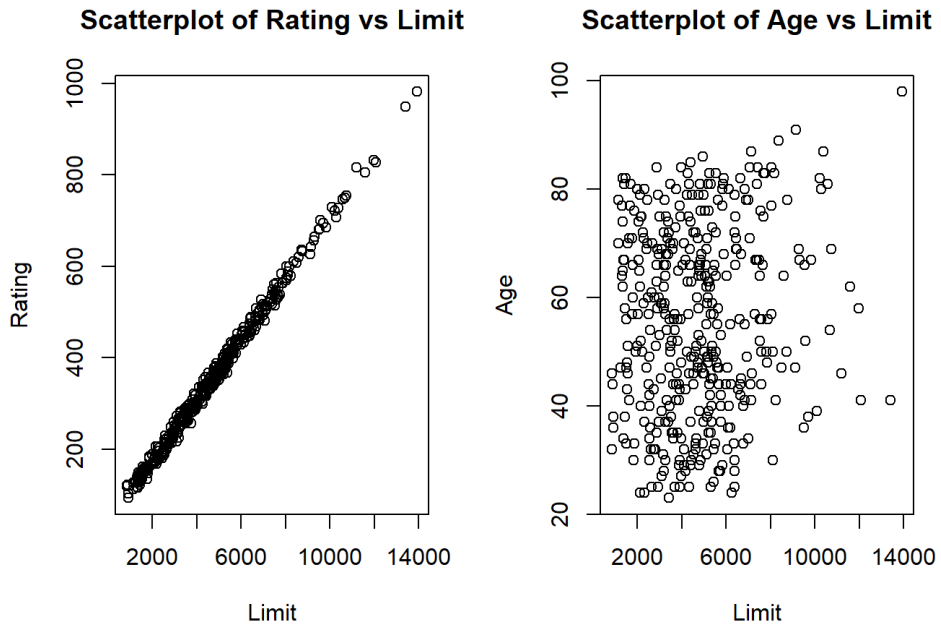
Consider the Credit data in the ISLR library. Choose balance as the response and Age, Limit and Rating as the predictors.

```
#Loading the Dataset
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) 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")
```




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

Comment:

Rating vs Limit:

The scatterplot shows an extremely strong positive linear relationship between Rating and Limit. This suggests that including both variables in a regression model may cause severe multicollinearity.

Age vs Limit:

The scatterplot shows a very weak (almost no) linear relationship between Age and Limit. The points are widely scattered without any clear trend.

(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?

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

```r
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
```

Marked difference observed:

• In model (1), Limit is highly significant (0.173***).

• In model (2), when Rating is added, Limit becomes statistically insignificant (0.019, not significant).

• 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 are almost perfectly linearly related, so this is a clear case of multicollinearity. When both are included, the model cannot 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(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.

---

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.

```
#Attaching the Boston Data
rm(list=ls())
library(MASS)

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

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)
```

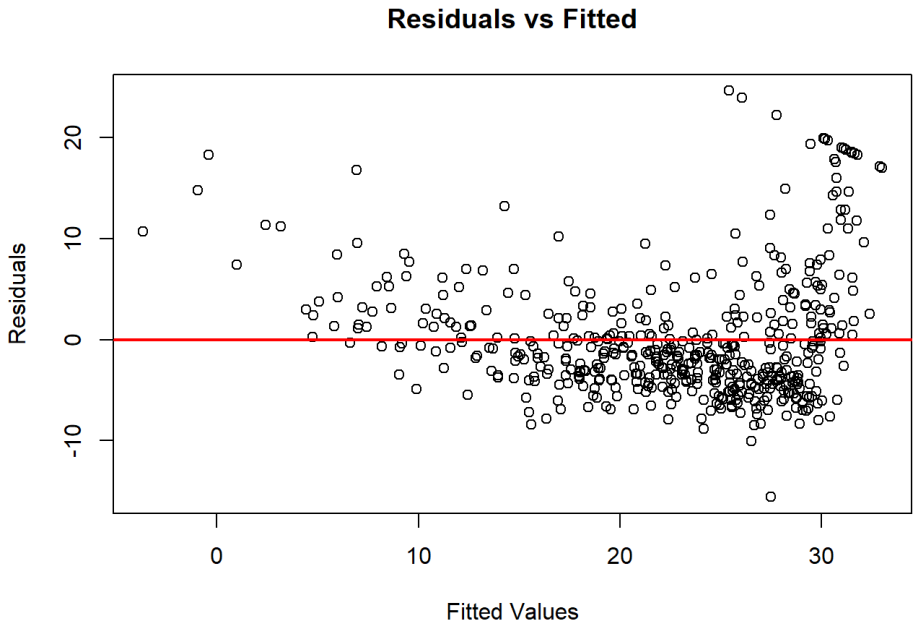
```
##
## Call:
## lm(formula = medv ~ ., data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.564  -4.004  -1.504   2.178  24.608
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.053584   2.170839  13.844   <2e-16 ***
## crim        -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
```

```
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.

```
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.