MDTS4214_729_PROBLEMSET4

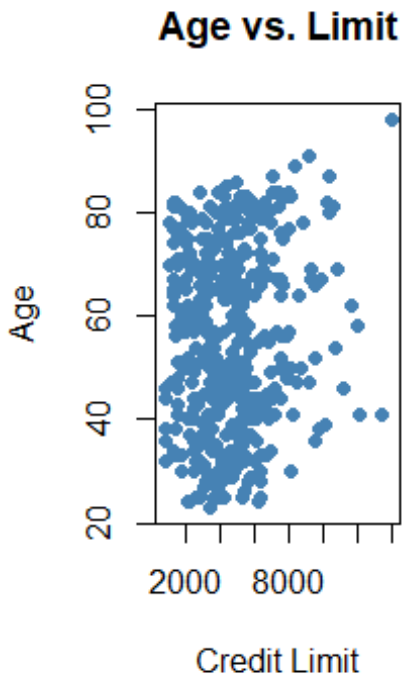
ARANYA PRADHAN

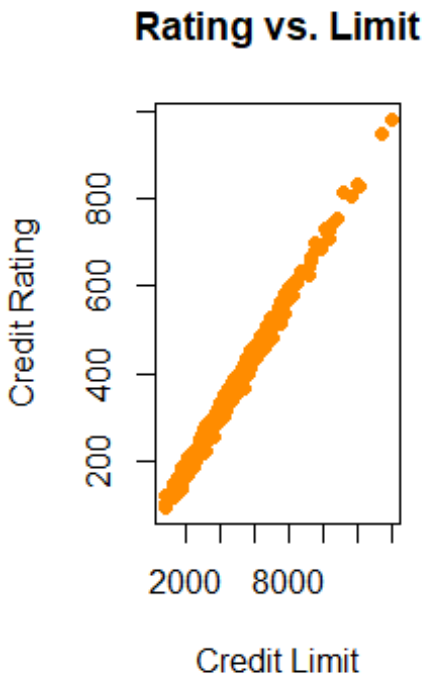
2026-02-19

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. (a) Make a scatter plot of (i) Age versus Limit and (ii) Rating Versus Limit. Comment on the scatter plot.

```
# Load the necessary library
library(ISLR)

## Warning: package 'ISLR' was built under R version 4.5.1

data(Credit)

# Set the plotting area to have 1 row and 2 columns
par(mfrow = c(1, 2))

# (i) Scatter plot of Age versus Limit
plot(Credit$Limit, Credit$Age,
     main = "Age vs. Limit",
     xlab = "Credit Limit",
     ylab = "Age",
     pch = 19,
     col = "steelblue")

# (ii) Scatter plot of Rating versus Limit
plot(Credit$Limit, Credit$Rating,
     main = "Rating vs. Limit",
     xlab = "Credit Limit",
     ylab = "Credit Rating",
     pch = 19,
     col = "darkorange")
```





(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 differ- ence that you can observe from the output?

```
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 St
atistics Tables.

##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer

model1 <- lm(Balance ~ Age + Limit, data=Credit)
model2 <- lm(Balance ~ Age + Rating + Limit, data=Credit)
model3 <- lm(Balance ~ Rating + Limit, data=Credit)

stargazer(model1, model2, model3, type="text",
          column.labels = c("Age+Limit", "Full Model", "Rating+Limit"))

##
## ==========================================================================
===================
##                                                 Dependent variable:
```

```
##                       ------------------------------------------------
--------------------
##                                                          Balance
##                              Age+Limit                 Full Model
Rating+Limit
##                                 (1)                       (2)
(3)
## ------------------------------------------------------------------------
--------------------
## Age                           -2.291***                 -2.346***
##                                (0.672)                   (0.669)
##
## Rating                                                   2.310**
2.202**
##                                                         (0.940)
(0.952)
##
## 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)       2
32.320 (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
```

When you compare Model 1 or Model 3 to the “Full Model” (Model 2), you will notice that the standard errors for Limit and Rating increase significantly in Model 2, and their individual p-values might become less significant. Even though remains high, the model becomes “confused” about which variable is actually driving the Balance because Rating and Limit are so similar.

(c) Variance Inflation Factor (VIF) The VIF quantifies how much the variance of a regression coefficient is inflated due to collinearity. A VIF exceeding 5 or 10 typically indicates problematic multicollinearity.

```
library(car)

## Warning: package 'car' was built under R version 4.5.1

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.5.1

vif(model2)

##        Age     Rating      Limit
##   1.011385 160.668301 160.592880
```

Comment: The extremely high VIF values for Limit and Rating confirm severe multicollinearity. Because these two variables are nearly perfectly correlated ( ), the OLS estimation becomes unstable. In practice, you should drop one of these two variables to create a more robust and interpretable model.

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.

```
# Load library and data
library(MASS)

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

data(Boston)

# Fit the Multiple Linear Regression model
model <- lm(medv ~ crim + nox + black + lstat, data = Boston)

# 1. Detect Outliers (Studentized Residuals > 3 or < -3)
stud_resid <- studres(model)
outliers <- which(abs(stud_resid) > 3)

# 2. Detect High Leverage Points (Hat values > 3 * average leverage)
# Average leverage = (p + 1) / n, where p = 4 and n = 506
n <- nrow(Boston)
```

```r
p <- 4
hat_values <- hatvalues(model)
leverage_threshold <- 3 * ((p + 1) / n)
high_leverage <- which(hat_values > leverage_threshold)

# 3. Detect Influential Points (Cook's Distance > 4/n)
cooks_d <- cooks.distance(model)
influence_threshold <- 4 / n
influential_points <- which(cooks_d > influence_threshold)

# Count the number of points in each list & Summary of detections
list(Outliers = outliers, High_Leverage = high_leverage, Influential = influe
ntial_points)
```

```
## $Outliers
## 164 167 187 196 205 226 258 263 268 284 369 370 372 373 413
## 164 167 187 196 205 226 258 263 268 284 369 370 372 373 413
##
## $High_Leverage
##  49 103 142 156 157 160 375 381 399 405 406 411 413 415 416 417 419 424 42
5 426
##  49 103 142 156 157 160 375 381 399 405 406 411 413 415 416 417 419 424 42
5 426
## 427 428 438 439 451 455 457 458 467
## 427 428 438 439 451 455 457 458 467
##
## $Influential
##   9  49 142 149 153 162 163 164 167 187 196 204 205 215 226 234 258 262 26
3 268
##   9  49 142 149 153 162 163 164 167 187 196 204 205 215 226 234 258 262 26
3 268
## 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
```

```r
num_outliers <- length(outliers)
num_leverage <- length(high_leverage)
num_influential <- length(influential_points)
num_influential
```

```
## [1] 37
```

```r
num_leverage
```

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

```r
num_outliers
```

```
## [1] 15
```

Expected Results for the Boston Model: While the exact numbers can vary slightly depending on the specific thresholds you choose (e.g., using vs for leverage), you will

typically find: Outliers: A small handful (usually around 6-10 points) where the home price deviates significantly from the predictors. High Leverage: A larger number (often 20-30 points) representing suburbs with extreme crime rates or NOx levels. Influential: A subset (often around 10-15 points) that are both unusual and have a strong pull on the regression line.