**STAT 40001/STAT 50001 Statistical Computing Fall 2024**

**Lab -18**

**Q.N. 1)** A hospital administrator wished to study the relationship between patients satisfaction (y)

and patient's age (x1, in years), severity of illness (x2, an index) and anxiety level (x3, an index).

46 patients are randomly selected and data is collected. The data is as below

ID y x1 x2 x3

1 48 50 51 2.3

2 70 41 44 1.8

3 46 42 50 2.2

4 77 29 50 2.1

5 47 38 55 2.2

6 66 36 49 2.0

7 60 33 49 2.1

8 52 44 58 2.9

9 43 47 53 2.5

10 72 32 46 2.6

11 59 33 42 2.0

12 47 40 48 2.2

13 82 29 48 2.5

14 42 47 50 2.6

15 37 44 51 2.6

16 92 28 46 1.8

17 57 36 46 2.3

18 89 28 43 1.8

19 54 45 48 2.4

20 89 29 48 2.4

21 51 34 51 2.3

22 79 33 56 2.5

23 49 55 51 2.4

24 60 43 50 2.3

25 34 55 54 2.5

26 57 32 52 2.4

27 83 36 49 1.8

28 36 53 57 2.8

29 64 30 51 2.4

30 66 43 53 2.3

31 68 45 51 2.2

32 66 40 48 2.2

33 36 49 54 2.9

34 26 52 62 2.9

35 67 43 53 2.4

36 57 53 54 2.2

37 88 29 46 1.9

38 77 29 52 2.3

39 86 23 41 1.8

40 63 25 49 2.0

41 55 42 51 2.7

42 76 31 47 2.0

43 80 34 49 2.2

44 37 47 60 2.4

45 83 22 51 2.0

46 59 37 53 2.1

1. Fit a multiple linear regression model to the data and state the estimated regression line.

How is b2 interpreted here?

1. Calculate the coefficient of multiple determination. What does it indicate?
2. Perform the necessary analysis to detect possible outliers and influential cases.
3. Repeat (c) using olsrr package.

> #Q1

> # file.choose()

> Q1 <- read.table("C:\\Users\\PNW\_checkout\\Downloads\\vaishak\\PNW\_COURSE-WORK\\FALL24\\STATISTICAL COMPUTING\\Assignment\\Assignment 18\\Q1\_Assignment18.txt", header = T)

> head(Q1)

ID y x1 x2 x3

1 1 48 50 51 2.3

2 2 70 41 44 1.8

3 3 46 42 50 2.2

4 4 77 29 50 2.1

5 5 47 38 55 2.2

6 6 66 36 49 2.0

> dim(Q1)

[1] 46 5

> attach(Q1)

> # note: ID is a not to be considered, its just like sl no

> model1 <- lm(y~x1+x2+x3, data = Q1)

> model1

Call:

lm(formula = y ~ x1 + x2 + x3, data = Q1)

Coefficients:

(Intercept) x1 x2 x3

158.491 -1.142 -0.442 -13.470

> cat("Fitted Model1: \n

+ y = 158.491 -1.142\*x1 - 0.442\*x2 - 13.470\*x3 ")

Fitted Model1:

y = 158.491 -1.142\*x1 - 0.442\*x2 - 13.470\*x3

> # Also

> cat("Fitted Model1: \n

+ patients satisfaction = 158.491 -1.142\*(age) - 0.442\*(severity\_Score) - 13.470\*(anxiety\_level)")

Fitted Model1:

patients satisfaction = 158.491 -1.142\*(age) - 0.442\*(severity\_Score) - 13.470\*(anxiety\_level)

> # install.packages("olsrr")

> cat("b2 interpretation: as severity\_score increase by 1 unit, the satisfcation resduces by 0.442 units")

b2 interpretation: as severity\_score increase by 1 unit, the satisfcation resduces by 0.442 units

> install.packages("olsrr")

> library(olsrr)

> model1a <- ols\_regress(y~x1+x2+x3, data = Q1)

> # model1a <- ols\_regress(model1, data = Q1)

> model1a

Model Summary

---------------------------------------------------------------

R 0.826 RMSE 9.611

R-Squared 0.682 MSE 92.366

Adj. R-Squared 0.659 Coef. Var 16.337

Pred R-Squared 0.622 AIC 348.727

MAE 8.148 SBC 357.871

---------------------------------------------------------------

RMSE: Root Mean Square Error

MSE: Mean Square Error

MAE: Mean Absolute Error

AIC: Akaike Information Criteria

SBC: Schwarz Bayesian Criteria

ANOVA

---------------------------------------------------------------------

Sum of

Squares DF Mean Square F Sig.

---------------------------------------------------------------------

Regression 9120.464 3 3040.155 30.052 0.0000

Residual 4248.841 42 101.163

Total 13369.304 45

---------------------------------------------------------------------

Parameter Estimates

-------------------------------------------------------------------------------------------

model Beta Std. Error Std. Beta t Sig lower upper

-------------------------------------------------------------------------------------------

(Intercept) 158.491 18.126 8.744 0.000 121.912 195.071

x1 -1.142 0.215 -0.591 -5.315 0.000 -1.575 -0.708

x2 -0.442 0.492 -0.111 -0.898 0.374 -1.435 0.551

x3 -13.470 7.100 -0.234 -1.897 0.065 -27.798 0.858

-------------------------------------------------------------------------------------------

> cat("R -> 0.826, Inference: ..........................")

R -> 0.826, Inference: ..........................

> # influence.measures(model1)

> # 2nd, 10th, 45th observations, seem to have outliers

> ols\_plot\_cooksd\_bar(model1)

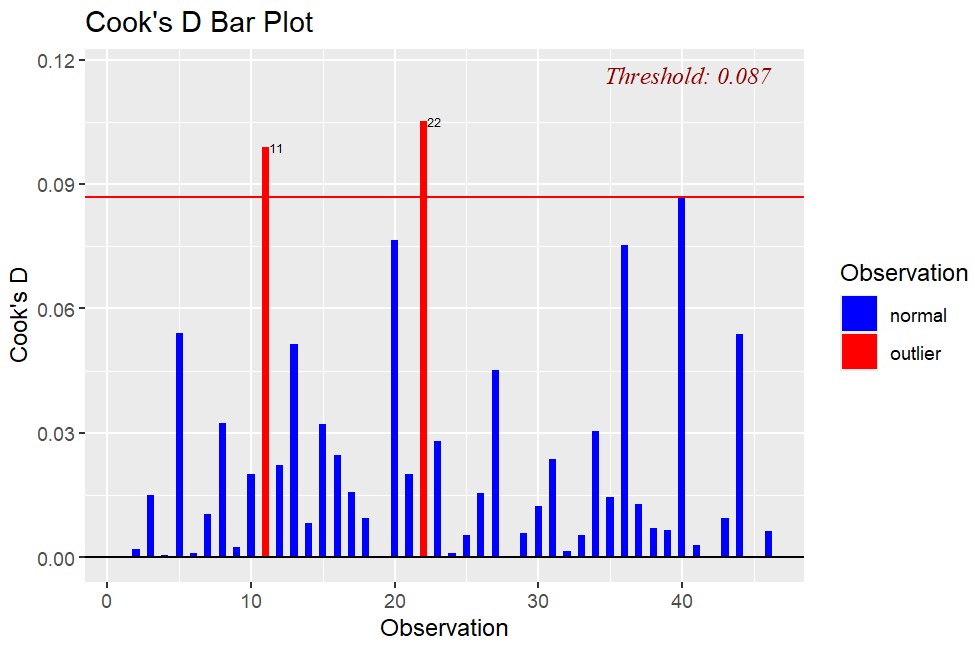
> # 11th, and 22nd observations, seem to have outliers

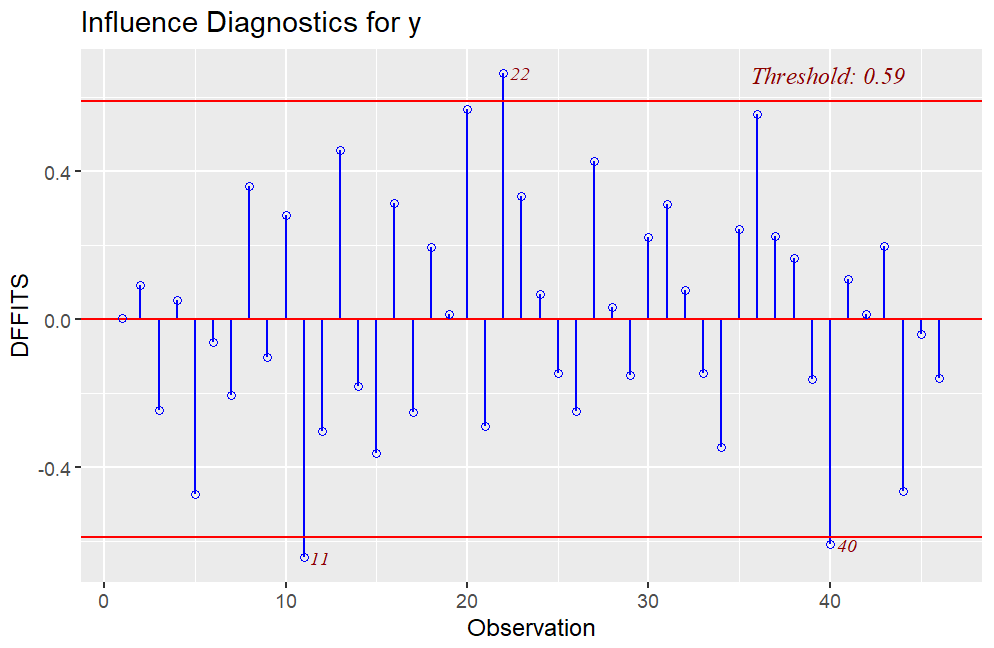
> ols\_plot\_dffits(model1)

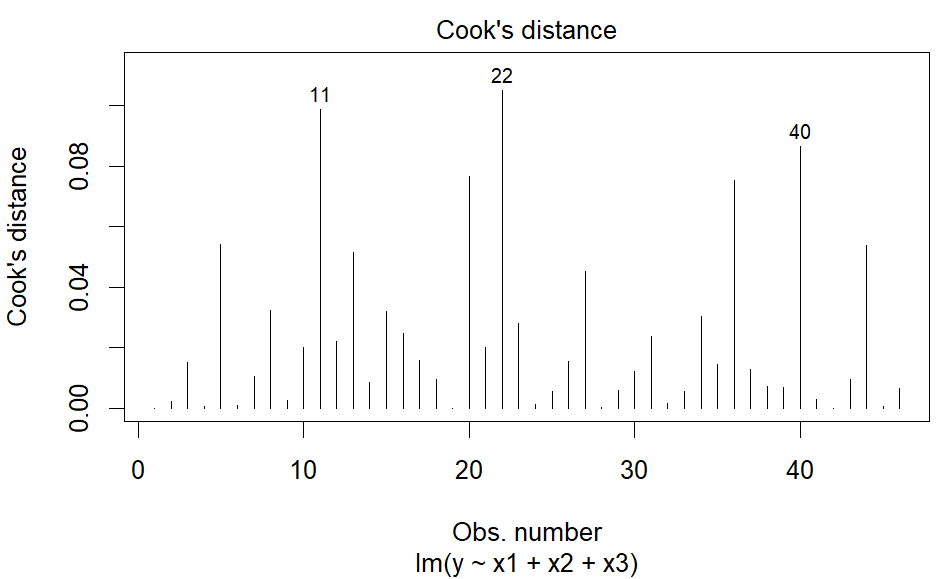
> # 11th, 22nd, and 40th observations, seem to have outliers

> # or

> plot(model1, which = 4) #plots cooks plot







**Q.N. 2)** UT Austin, like essentially every other major university in the country, asks students to evaluate the quality of instruction they have received from their professors. The rating your professors on a scale of 1 (very unsatisfactory) to 5 (excellent) on the end-of-semester course instructor survey. These ratings are significant part of what administrators use to evaluate faculty performance.

The file “profs.csv” (available with this assignment) contains data on course-instructor surveys from a sample of 463 courses at the University of Texas from 2000 - 2002. You are also given information about the individual courses and professors – including, most controversially, a rating of each professor’s physical attractiveness as judged by students. The data represent evaluations from 25,547 students and most major academic departments.

(Data originally from “Beauty in the classroom: instructors’ pulchritude and putative pedagogical productivity.” Daniel S. Hamermesh and Amy M. Parker. Economics of Education Review, August 2005, v. 24 (4) pp. 369–76.)

The variables included are:

* minority: is the professor from a non-Caucasian ethnic minority?
* age: the professor’s age.
* gender: a factor indicating the professor’s gender.
* credits: a factor indicating whether the course is a single-credit elective (e.g. scuba diving or ballroom dancing, coded “single”) or an academic course (coded “more”).
* beauty: a rating of the professor’s physical attractiveness, as judged by a panel of six students. (The score was averaged across all six panelists, and shifted to have a mean of zero)
* eval: the professor’s average teaching evaluation for courses in the sample, on a scale of 1 to 5.
* division: whether the course is an upper or lower division course.
* native: whether the professor is a native English speaker.
* tenure: whether the professor is tenured/tenure-track, or not.
* students: the number of students that participated in the evaluation.
* allstudents: the number of students enrolled in the course.
* prof: a unique numerical identifier for the professor being rated.

1. Start by examining the association between the evaluations and the beauty score graphically, and with a simple linear regression model. What do you see?
2. Perform a necessary analysis of the subject data to find other significant variables to determine the evaluation

> #Q2

> # file.choose()

> Q2 = read.csv("C:\\Users\\PNW\_checkout\\Downloads\\vaishak\\PNW\_COURSE-WORK\\FALL24\\STATISTICAL COMPUTING\\Assignment\\Assignment 18\\profs.csv")

> head(Q2,4)

minority age gender credits beauty eval division native tenure

1 yes 36 female more 0.2899157 4.3 upper yes yes

2 no 59 male more -0.7377322 4.5 upper yes yes

3 no 51 male more -0.5719836 3.7 upper yes yes

4 no 40 female more -0.6779634 4.3 upper yes yes

students allstudents prof

1 24 43 1

2 17 20 2

3 55 55 3

4 40 46 4

> dim(Q2)

[1] 463 12

> names(Q2)

[1] "minority" "age" "gender" "credits" "beauty"

[6] "eval" "division" "native" "tenure" "students"

[11] "allstudents" "prof"

> attach(Q2)

> plot(beauty, eval, pch = 17, col = "red")

> model2 <- lm(eval~beauty, data = Q2)

> model2

Call:

lm(formula = eval ~ beauty, data = Q2)

Coefficients:

(Intercept) beauty

3.998 0.133

> cat("Fitted Model: \neval = 3.998+0.133\*beauty\nComment: It's a weak model")

Fitted Model:

eval = 3.998+0.133\*beauty

Comment: It's a weak model

> summary(model2)

Call:

lm(formula = eval ~ beauty, data = Q2)

Residuals:

Min 1Q Median 3Q Max

-1.80015 -0.36304 0.07254 0.40207 1.10373

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.99827 0.02535 157.727 < 2e-16 \*\*\*

beauty 0.13300 0.03218 4.133 4.25e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5455 on 461 degrees of freedom

Multiple R-squared: 0.03574, Adjusted R-squared: 0.03364

F-statistic: 17.08 on 1 and 461 DF, p-value: 4.247e-05

> # beauty\_Score is not significant

> abline(model2, lwd = 2, col = "green")

> install.packages("MASS")

> library(MASS)

> modelsr = lm(y~Q2$minority+Q2$age+Q2$gender+Q2$credits+Q2$beauty+Q2$division+Q2$native+Q2$tenure+Q2$students+Q2$allstudents,data = Q2) #all variables considered initially

> step = stepAIC(modelsr, direction = "both")

> cat("By stepwise regression: y ~ x1 + x3")

By stepwise regression: y ~ x1 + x3

> cat("Thus x1 and x3 are significant")

Thus x1 and x3 are significant

