Regression HW 9 Solutions

Jun Lu, Mingfan Meng

Fall 2023

Contents

1. Problem 10.7. Patient satisfaction. 1
2. Problem 10.11 Patient satisfaction. 2
3. Problem 10.17 Patient satisfaction. p.417 (10 pts) 6

45 pts

Reminder

- Keep working on your project(s)
- Lab 6.

Load car package.

```
# install packages if needed.
# install.packages("car")

library(car)  # needed for avPlots(), vif()
```

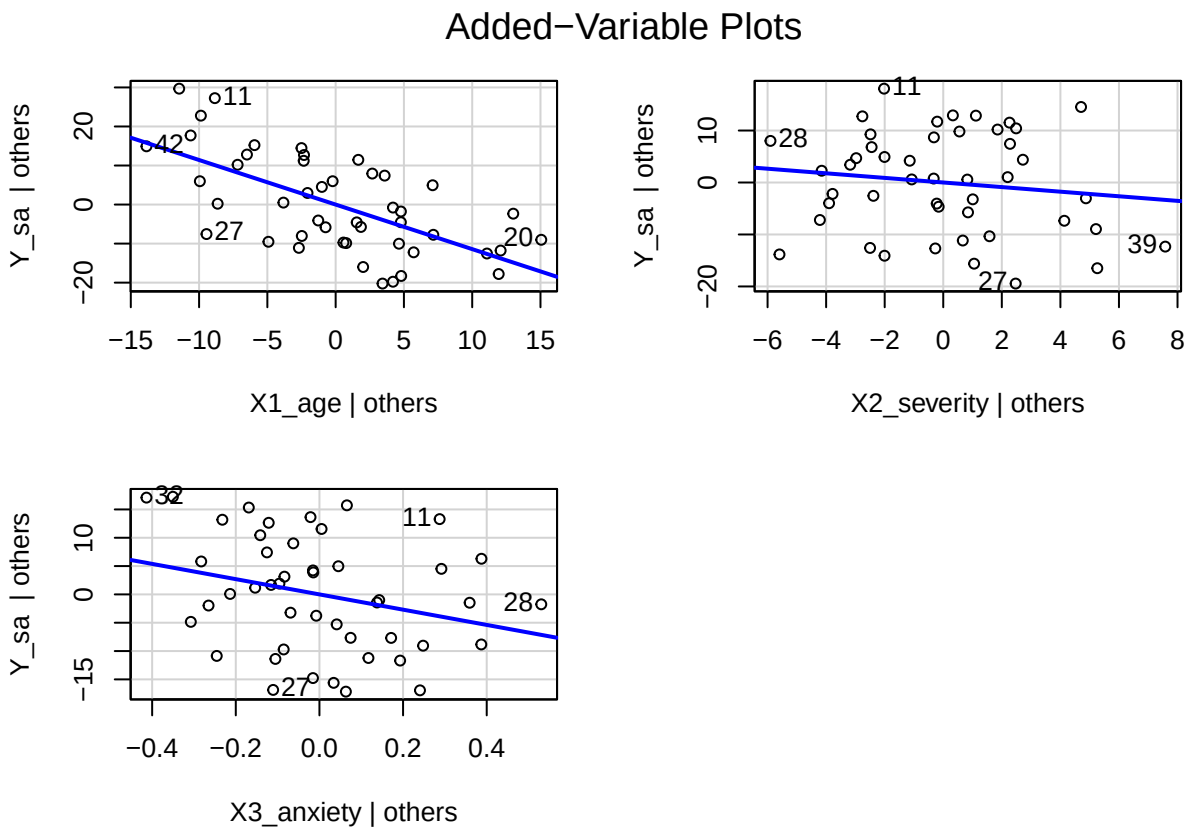
1. Problem 10.7. Patient satisfaction.

A hospital administrator wished to study the relation between patient satisfaction (Y) and patient's age (X_1, in years), severity of illness (X_2, an index), and anxiety level (X_3 , an index). The administrator randomly selected 46 patients and collected the data presented below, where larger values of Y, X_2 , and X_3 are, respectively, associated with more satisfaction, increased severity of illness, and more anxiety.

```
ps <- read.table("./CH06PR15.txt", header = F)
colnames(ps) <- c("Y_sa", "X1_age", "X2_severity", "X3_anxiety")
```

a. Prepare an added-variable plot for each of the predictor variables.

```
ps_lm <- lm(Y_sa ~ X1_age + X2_severity + X3_anxiety, x=T, data = ps)
avPlots(ps_lm)
```



b. Do your plots in part (a) suggest that the regression relationships in the fitted regression function in Problem 6.15c are inappropriate for any of the predictor variables? Explain.

- None of the three predictors' partial regression plot show any curved pattern. The predictors can be included in the model as the first-order terms. The model in problem 6.15.c seems reasonable, though we may revise it with less predictors.
- The added-variable plot (partial regression plot) for predictor AGE has a negative association with the satisfaction rate. This indicates that AGE should be kept in the model while controlling other predictors, and it has a negative slope against the response variable.
- The added-variable plot for SEVERITY and ILLNESS show no recognizable pattern, so these predictor variables do not have a strong influence on the satisfaction rate when controlling for the other predictors. At least one of them can be dropped from the model.

2. Problem 10.11 Patient satisfaction.

a. Obtain the studentized deleted residuals and identify any outlying Y observations. Use the Bonferroni outlier test procedure with $\alpha = .10$. State the decision rule and conclusion.

```
ti <- rstudent(ps_lm)
head(ti)
```

```
##           1           2           3           4           5           6
##  0.01155475 -0.93317867  0.40362503  0.21466620  0.59441714 -0.38167082
```

```
n <- 46
p <- 4
```

```
fam <- 0.1
tcrit <- qt(1-fam/(2*n), n-p-1)
tcrit
```

```
## [1] 3.271524
```

```
ti[abs(ti) > tcrit]
```

```
## named numeric(0)
```

- Decision rule:
  - if $|t_i| > t_{(1-\alpha/(2n),n-p-1)}$, conclude observation i is an outlier with respect to the response variable.
  - Otherwise the i-th observation is not an outlier on the response variable.
- Conclusion:
  - According to the studentized deleted residual, we can find the numbers are between (-2,2), which indicated that all residual value are smaller than $t = 3.27$. We can conclude that there is no outlying Y observation in the model.

b. Obtain the diagonal elements of the hat matrix. Identify any outlying X observations.

```
hii <- hatvalues(ps_lm)
head(hii)
```

```
##         1         2         3         4         5         6
## 0.07819669 0.06706793 0.03717097 0.15361084 0.09673692 0.12857668
```

```
n <- 46
p <- 4
2*p/n
```

```
## [1] 0.173913
```

```
hii[hii > 2*p/n]
```

```
##         9        28        39
## 0.1842585 0.1860192 0.1809601
```

- According to the above calculation, we can identify the case 9, 28, 39 are outliers with respect to the X-values in this model.

c. Hospital management wishes to estimate mean patient satisfaction for patients who are $X_1 = 30$ years old, whose index of illness severity is $X_2 = 58$, and whose index of anxiety level is $X_3 = 2.0$. Use (10.29) to determine whether this estimate will involve a hidden extrapolation.

```
# ps_lm <- lm(Y_sa ~ X1_age + X2_severity + X3_anxiety, x=T, data = ps)  # Already run.

designX<-ps_lm$x
newX <-c(1, 30, 58, 2)
lev<-t(newX)%*%solve(t(designX)%*%designX)%*%newX
lev
```

```
##           [,1]
## [1,] 0.3267004
```

- Because the leverage for mean patient satisfaction is 0.3267 that is greater than 0.173 ($2p/n$). We can conclude that this estimate will involve a hidden extrapolation.

d. The three largest absolute studentized deleted residuals are for cases 11, 17, and 27. Obtain the DFFITS, DFBETAS, and Cook's distance values for this case to assess its influence. What do you conclude?

```
dffits(ps_lm)[c(11, 17, 27)]
```

```
##         11         17         27
##  0.5688200  0.6657370 -0.6087397
```

```
dfbetas(ps_lm)[c(11,17, 27), ]
```

```
##    (Intercept)     X1_age X2_severity X3_anxiety
## 11  0.09910764 -0.3630892  -0.1899887 0.38998516
## 17 -0.44913479 -0.4711109   0.4432302 0.08926996
## 27 -0.01723432  0.4171827  -0.2498614 0.16136484
```

```
cooks.distance(ps_lm)[c(11, 17, 27)]
```

```
##         11         17         27
## 0.07656783 0.10513344 0.08666240
```

	DFFITS	b_0	b_1	b_2	b_3	D
case 11	0.5688	0.0991	-0.3631	-0.1900	0.3900	0.0766
case 17	0.6657	-0.4491	-0.4711	0.4432	0.0893	0.1051
case 27	-0.6087	-0.0172	0.4172	-0.2499	0.1614	0.0867

- The threshold for Cook's-D is (recommended in the text) is 0.853. According to this threshold, none of the 3 cases is highly influential.

```
p <- length(ps_lm$coefficient)
n <- nrow(ps)
qf(0.5, p, n-p)  # Median of F-distribution.
```

```
## [1] 0.8528731
```

- We can treat this data as "small" data ($n/p = 46/4 = 11.5$, small for regression analysis.) and use 1 as the threshold for DfFits and DfBetas. For larger data sets, we can use the following thresholds (recommended by the text) for DfFits and DfBetas, respectively.

```
p <- length(ps_lm$coefficient)
n <- nrow(ps)
2*sqrt(p/n)
```

```
## [1] 0.5897678
```

```
2/sqrt(n)
```

```
## [1] 0.2948839
```

f. Calculate Cook's distance D_i for each case and prepare an index plot. Are any cases influential according to this measure?

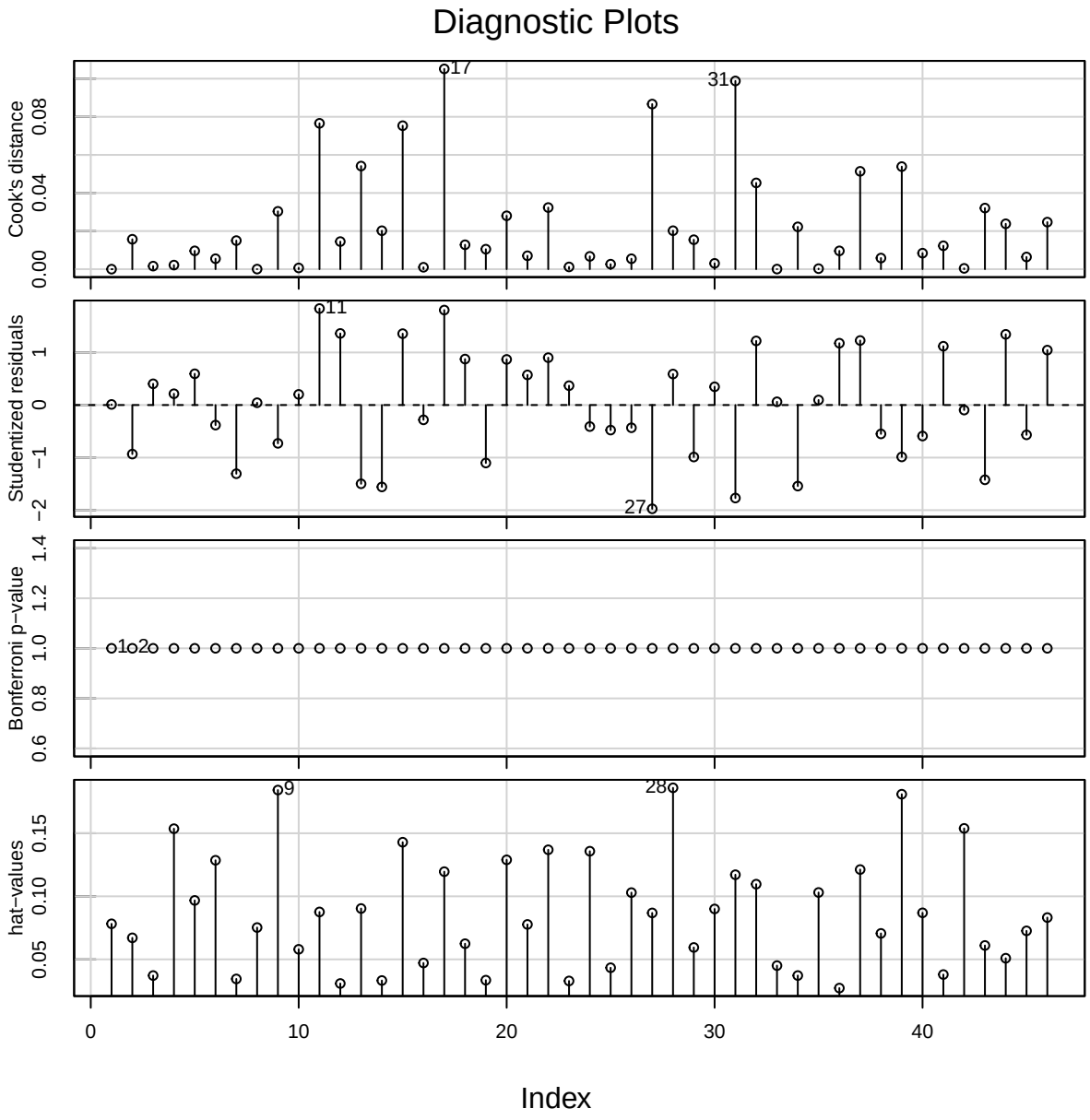
```
D <- cooks.distance(ps_lm)
head(D)
```

```
##            1            2            3            4            5            6
## 2.900516e-06 1.569901e-02 1.604332e-03 2.139423e-03 9.608143e-03 5.484993e-03
```

```
plot(D, main="Cook's Distance", ylab = "Cook's-D")
```

The style of the plots can vary.

```
influenceIndexPlot(ps_lm)
```

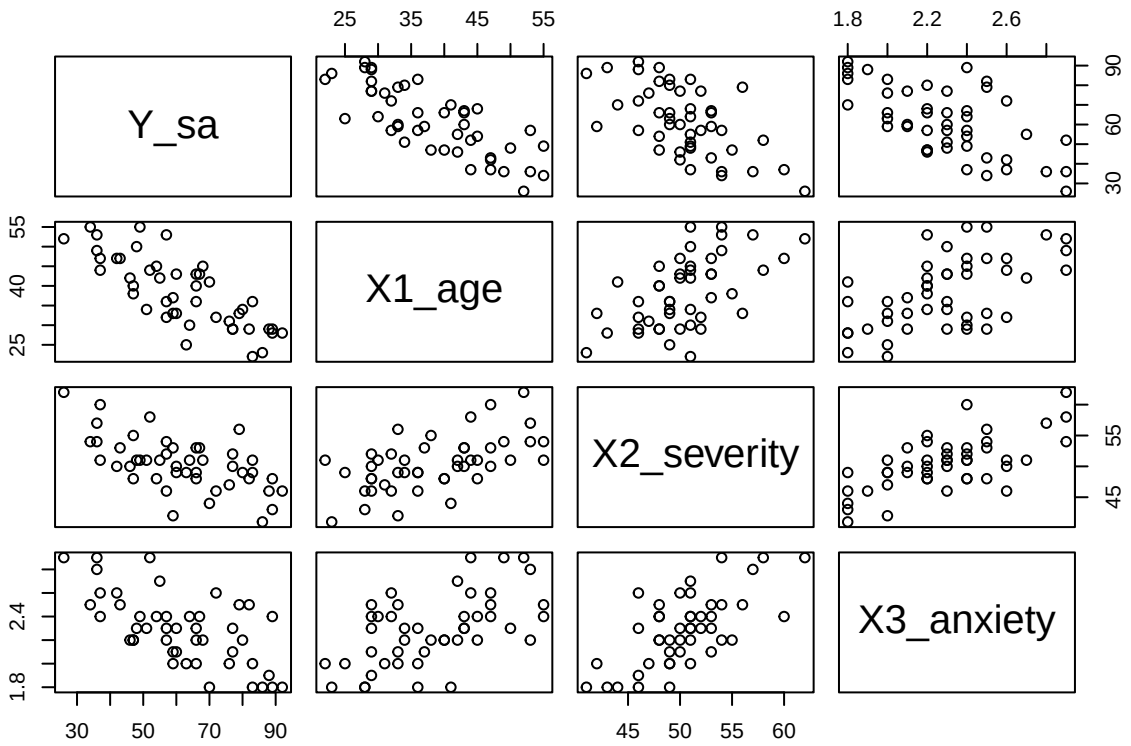


- According to the plot, we can find case 17 and 31 have large Cook's D, though they are less than $F(0.5, 4, 42) = 0.853$.

3. Problem 10.17 Patient satisfaction. p.417 (10 pts)

a. What do the scatter plot matrix and the correlation matrix show about pairwise linear associations among the predictor variables?

```
pairs(ps)
```




```
cor(ps)
```

```
##                    Y_sa     X1_age X2_severity X3_anxiety
## Y_sa          1.0000000 -0.7867555  -0.6029417 -0.6445910
## X1_age       -0.7867555  1.0000000   0.5679505  0.5696775
## X2_severity -0.6029417  0.5679505   1.0000000  0.6705287
## X3_anxiety  -0.6445910  0.5696775   0.6705287  1.0000000
```

- There is some collinearity among the predictors. In particular, the scatter plots and the correlation coefficients show association between severity and anxiety. The linear association is not very strong though.

b. Obtain the three variance inflation factors. What do these results suggest about the effects of multicollinearity here? Are these results more revealing than those in part (a)?

```
vif(ps_lm)
```

```
##      X1_age X2_severity  X3_anxiety
##    1.632296    2.003235    2.009062
```

- The VIFs are greater than 1, but they are not big. It shows that there is multicollinearity, but is not severe to cause concerns.

This is the end of HW 9.