Stat355, Assignment #2, due: Friday, October 4 in class

Zimeng Ming V00844078

2019-10-02

Multiple regression, Section 10.1 - 10.3

Instructions:

1. Complete your assignment in R Markdown using this file as a template. Insert R code in the R chunks, and type in your response after the corresponding R chunk leaving one blank line between the R chunk and your comments.
2. Execute each line of code separately to ensure that it works properly.
3. Either [knit the entire document to pdf] or [knit to HTML or Word and print to pdf].
4. Submit the pdf file to CourseSpaces in the Assignment 2 activity.

Data Description:

Question 10.3.3

In a study of factors thought to be related to patterns of admission to a large general hospital, an administrator obtained these data on 10 communities in the hospital's cachment area.

AdRate: (Y) Admission Rate, Number of hospital admissions per 1000 population.

OHealth: (X1) Index of availability of other health services. Larger values mean more health services.

SESI: (X2) Mean socio-economic status index. Larger values (here) mean less wealthy.

0. Read the data into R using the read.csv function.

```
knitr::opts_chunk$set(fig.width=8, fig.height=6) #set size of graphs
HOSP.dat<-read.csv('EXR_C10_S03_03.csv')
dim(HOSP.dat)
```

```
## [1] 10  3
```

1. Provide descriptive statistics for the data. Comment on your results and especially on any unusual features in the data. (3 marks)

```
summary(HOSP.dat)
```

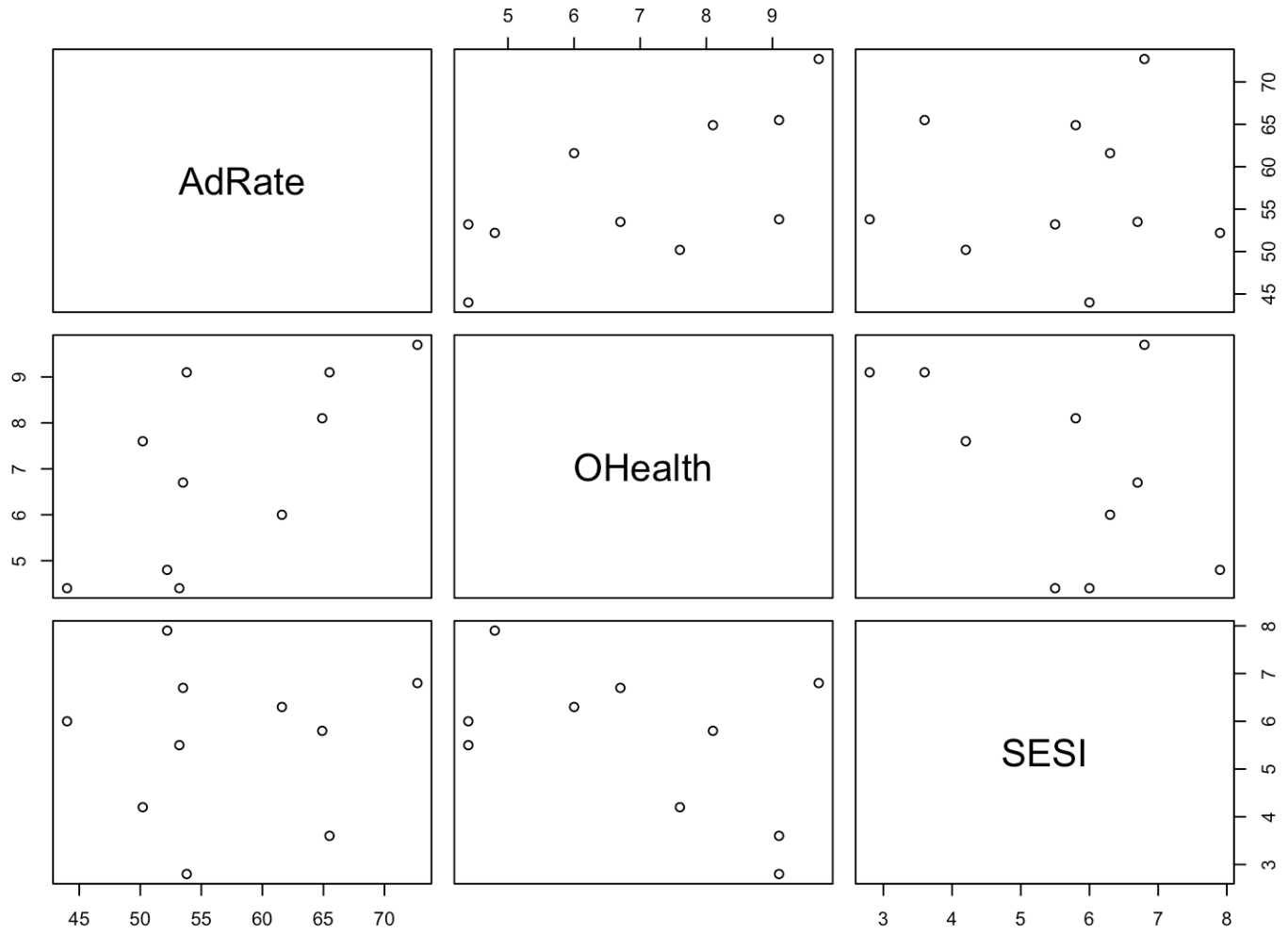
```
##      AdRate          OHealth          SESI      
##  Min.   :44.00   Min.   :4.40   Min.   :2.800  
##  1st Qu.:52.45   1st Qu.:5.10   1st Qu.:4.525  
##  Median :53.65   Median :7.15   Median :5.900  
##  Mean   :57.16   Mean   :6.99   Mean   :5.560  
##  3rd Qu.:64.08   3rd Qu.:8.85   3rd Qu.:6.600  
##  Max.   :72.70   Max.   :9.70   Max.   :7.900
```

Comment:
There are 10 data record. The Range of AdRate is from 44 to 72.7, the Range of OHealth is from 4.4 to 9.7, the Range of SESI is from 2.8 to 7.9. OHealth and SESI are looks like skewed to right.

---

2. Produce the pairs() scatterplots of all three variables. Comment. Is a linear model appropriate for this data? Why or why not? (3 marks)

```
pairs(HOSP.dat)
```



Comment:
We can see from the graph, the AdRate VS OHealth seems has an positive linear relationship,there is weakly positive relationship. the Ohealth has negatively relationship with SESI.

In general, the scattler plot shows there is a weakly positive relationship between AdRate vs Ohealth and SESI.

---

3. Fit simple linear models, Y versus X for each of the two X regressor variables. Compare the Multiple R-squared and Residual standard error for the two models. Comment. (3 marks)

```
y<-HOSP.dat$AdRate
x1<-HOSP.dat$OHealth
x2<-HOSP.dat$SESI

fit1<-lm(y~x1)
fit2<-lm(y~x2)
#fit3<-lm(y~x1+x2)
summary(fit1)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.545 -4.878  1.807  4.273  7.596
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   36.670      7.963   4.605  0.00174 **
## x1             2.931      1.098   2.669  0.02840 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.692 on 8 degrees of freedom
## Multiple R-squared:  0.4711, Adjusted R-squared:  0.4049
## F-statistic: 7.125 on 1 and 8 DF,  p-value: 0.0284
```

```
summary(fit2)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.267  -5.135  -3.312   6.826  15.237
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  55.8027    11.1771   4.993  0.00106 **
## x2            0.2441     1.9411   0.126  0.90302
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.192 on 8 degrees of freedom
## Multiple R-squared:  0.001973,   Adjusted R-squared:  -0.1228
## F-statistic: 0.01582 on 1 and 8 DF,  p-value: 0.903
```

```
#summary(fit3)
```

Comment:
First model: AdRate VS OHealth Model:y=36.670+2.931*x1

Multiple R-squared:0.4711 Residual standard error:6.692

Second model: AdRate VS SESI Model:y=55.8027+0.2441*x1

Multiple R-squared:0.001973 Residual standard error:9.192

Although for the model1, R_square is 0.4711 shows there are 47.11% of variation in y is explained by variation in x. However, it is clearly to see that the R_square for model2 is only 0.001973 shows there are only 0.1973% of variation in y is explained by variation in x which is extremly low and the residual standard error of model2 is higher then model1.Thus the second model (AdRate vs SESI) is not a good fitness model.

---

4. Fit the multiple linear regression model of Y on all of the X's.
    a. What is the estimated regression model? (2 marks)
    b. What is the hypothesis tested by the F-test. Comment on the results of the F-test. (2 marks)
    c. Compare the R-squared, adjusted R-squared and the Residual standard error with those obtained from the simple linear regression models above. (3 marks)
    d. Comment on the statistical significance of each of the X variables in the model.(3 marks)

```
fit3<-lm(y~x1+x2)
summary(fit3)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7005 -4.0547 -0.7874  4.6833  6.6122
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   13.449     13.232   1.016  0.34325
## x1             4.017      1.071   3.749  0.00718 **
## x2             2.812      1.379   2.040  0.08076 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.666 on 7 degrees of freedom
## Multiple R-squared:  0.6682, Adjusted R-squared:  0.5734
## F-statistic: 7.049 on 2 and 7 DF,  p-value: 0.02104
```

Comment:
a): Model:AdRate=13.449+4.017*OHealth*+*2.812*SESI+error

b): F-test: H0:beta_Ohealth=beta_SESI=0 As we can see that the p-value is lower then 0.05, so strong eveidence against H0, at least one of the beta_Ohealth and beta_SESI is not zero.

c): The multiple R-squared and Adjusted R-Square of multivariable regression model are 0.6682 and 0.5734 which are larger then any simple linear regression r-square. This is expected with we add both varable into the model. The resudual standard error are also smaller then both two simple variable model, Therefore, both Ohealth and SESI are needed for explaining the varibale AdRate.

d): X1(Ohealth) is very significant since the p-value is 0.00718<0.05 and the x2(SESI) is less siginificant since the p-value is 0.08076 <0.1.

e). Generate 95% confidence intervals for each of the slope parameters. Comment. (3 marks)

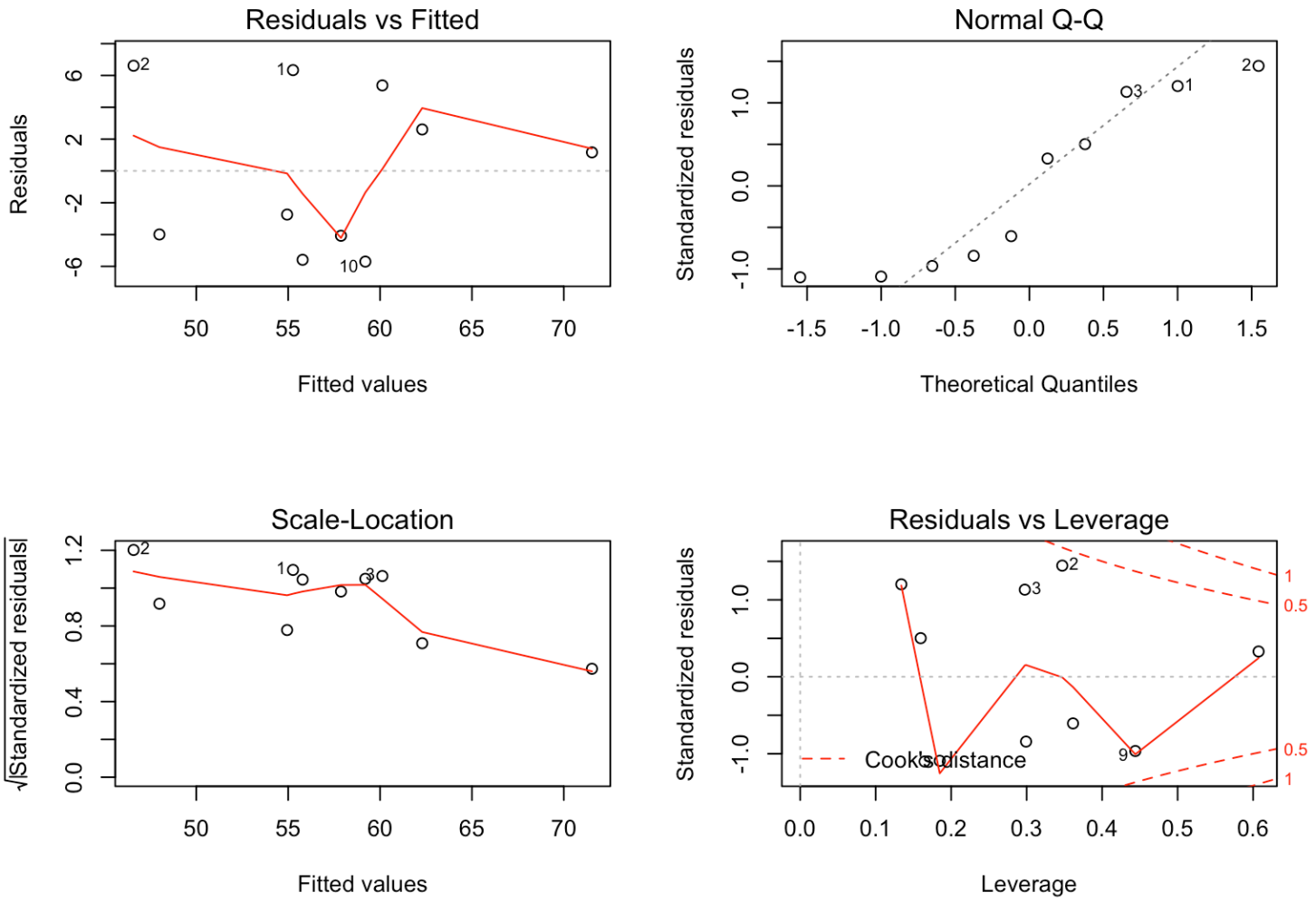
```
confint(fit3)
```

```
##                   2.5 %    97.5 %
## (Intercept) -17.8384294 44.736887
## x1            1.4834367  6.550159
## x2           -0.4480924  6.071601
```

The x1(Ohealth) do not contain the zero but x2(SESI) contain the zero, which reflect the result of the p-value. The X1 is significant and x2 is less(marginarlly) significant.

f). Check the fit of the model and comment. (3 marks)

```
par(mfrow=c(2,2))
plot(fit3)
```



Comment:
As from the Residuals vs Fitted graph we can clearly see that there are some residuals are outliers in the line, the red line are seems around the zero. The Normall Q-Q plot indicates good agreement with the assumption of normality. The scale location looks a negative smooth line and the residuals vs leverage is not shows the smooth line around zero.

Overall the plot shows there are some outliers in the model but there are only 10 data in the data set, so we need more data to fix the model.

g). Explain each of the estimated regression parameter estimates (except the intercept) in words. (3 marks)

Ohealth: The estimated expected AdRate increases by 4.017 per unit increase in Ohealth, with SESI held fixed. SESI: The estimated expected AdRate increases by 2.812 per unit increase in SESI, with Ohealth held fixed.