Class Supplement of High-Dimensional Data Analysis

Homework #3 (Revised version of 107.05.17)

Student name: Lin Ting-Yu

1. Suppose that a curve $\hat{g}$ is computed to smoothly fit a set of n points using the following formula：

$$\hat{g} = arg\ min\ \left(\Sigma_{i=1}^{n}\left(y_i - g(x_i)\right)^2 + \lambda \int \left[g^{(m)}(x)\right]^2 dx\right),$$

where $g^{(m)}$ represents the mth derivative of g (and $g^{(0)} = g$). Provide example sketches of $\hat{g}$ in each of the following scenarios.

(e) $\lambda = 0$, $m = 3$

Solution:

$$\hat{g} = arg\ min\ \left(\Sigma_{i=1}^{n}\left(y_i - g(x_i)\right)^2\right)$$

$$\Sigma_{i=1}^{n}\left(y_i - g(x_i)\right)^2 = 0, \text{when } \widehat{g(x_i)} = y_i, \forall i \in [1, n]$$

1. Suppose that $g(x_i) = \begin{cases} y_1, & x < x_2 \\ y_2, & x_2 \le x < x_3 \\ \vdots & \\ y_n, & x \ge x_n \end{cases} \Rightarrow \widehat{g(x_i)} = y_i, \forall i \in [1, n]$

$$\Rightarrow \Sigma_{i=1}^{n}\left(y_i - \widehat{g(x_i)}\right)^2 = \Sigma_{i=1}^{n}(y_i - y_i)^2 = 0$$

2. Suppose that $g(x_i) = \Sigma_{i=1}^{n} y_i \times 1 = \Sigma_{i=1}^{n} y_i \prod_{j=0, i\neq j}^{n} \frac{x_i - x_j}{x_i - x_j}$

$$\Rightarrow \Sigma_{i=1}^{n}\left(y_i - \widehat{g(x_i)}\right)^2 = \Sigma_{i=1}^{n}(y_i - y_i)^2 = 0$$

Exercise 14(Chap 3, P.125).

This problem focuses on the collinearity problem.

(a) Perform the following commands in R:

```
> set.seed(1)
> x1=runif(100)
> x2=0.5*x1+rnorm(100)/10
> y=2+2*x1+0.3*x2+rnorm(100)
```

The last line corresponds to creating a linear model in which y is a function of x1 and x2. Write out the form of the linear model. What are the regression coefficients?

(b) What is the correlation between x1 and x2? Create a scatterplot displaying the relationship between the variables.

(c) Using this data , fit a least squares regression to predict y using x1 and x2 . Describe the results obtained. What are $\widehat{\beta_0}$, $\widehat{\beta_1}$ and $\widehat{\beta_2}$? How do these relate to the true β_0 , β_1 , and β_2 ? Can you reject the null hypothesis $H_0: \beta_1 = 0$? How about the null hypothesis $H_0: \beta_2 = 0$?

(d) Now fit a least squares regression to predict y using only x1. Comment on your results. Can you reject the null hypothesis $H_0: \beta_1 = 0$?

(e) Now fit a least squares regression to predict y using only x2. Comment on your results. Can you reject the null hypothesis $H_0: \beta_1 = 0$?

(f) Do the results obtained in (c)-(e) contradict each other？ Explain yours answer.

(g) Now suppose we obtained one additional observation , which was unfortunately mismeasured.

```
> x1=c(x1, 0.1)
> x2=c(x2, 0.8)
> y=c(y,6)
```

Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on the each of the models? In each model , is this observation an outlier? A high-leverage point? Both? Explain your answers.

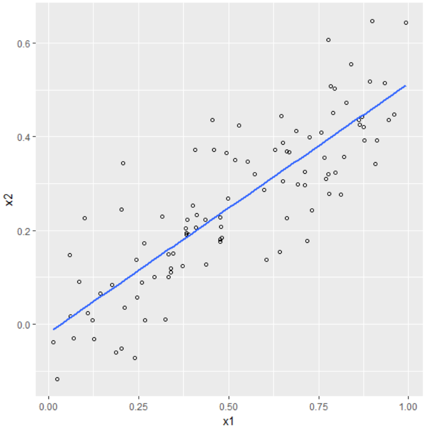
Solution:

(a)

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, where $\varepsilon \sim N(0, 1)$ and $\beta_0 = 2$, $\beta_1 = 2$, $\beta_2 = 0.3$

(b)

$$r_{12} = \frac{\Sigma_{i=1}^{100}(x_{1i}-\overline{x_1})(x_{2i}-\overline{x_2})}{\sqrt{\Sigma_{i=1}^{100}(x_{1i}-\overline{x_1})^2\Sigma_{i=1}^{100}(x_{2i}-\overline{x_2})^2}} = \frac{3.765686}{\sqrt{20.33242}} = 0.8351213$$



high positive correlation between x1 and x2.

(c)

```
> set.seed(1)
> x1=runif(100)
> x2=0.5*x1+rnorm(100)/10
> y=2+2*x1+0.3*x2+rnorm(100)
> f=lm(y~x1+x2)
> f

Call:
lm(formula = y ~ x1 + x2)

Coefficients:
(Intercept)           x1           x2
       2.13         1.44         1.01
```

```
> summary(f)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8311 -0.7273 -0.0537  0.6338  2.3359

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
x1            1.4396     0.7212   1.996   0.0487 *
x2            1.0097     1.1337   0.891   0.3754
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.056 on 97 degrees of freedom
Multiple R-squared:  0.2088,	Adjusted R-squared:  0.1925
F-statistic:  12.8 on 2 and 97 DF,  p-value: 1.164e-05
```

$$\Rightarrow y = 2.13 + 1.44x_1 + 1.01x_2\,, \widehat{\beta_0} = 2.13\,, \widehat{\beta_1} = 1.44\,, \widehat{\beta_2} = 1.01$$

$$\widehat{\beta_0}：\ 2.13$$

$$\widehat{\beta_1} = 1.44：\ \text{When } x_1 \text{ increase a unit, then y will increase 1.44}$$

$$\widehat{\beta_2} = 1.01: \text{When } x_2 \text{ increase a unit, then y will increase 1.01}$$

$$\widehat{\beta_0}、\widehat{\beta_1}、\widehat{\beta_2} \text{ are the roots of } \min\Sigma_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i})^2$$

Case 1 : $\mathrm{H_0}: \beta_1 = 0$ v.s. $\mathrm{H_1}: \beta_1 \neq 0$

$$t - statistic\ t_1 = \frac{\widehat{\beta_1}}{se(\widehat{\beta_1})} = \frac{1.44}{0.7211795} = 1.996729\,,$$

$$wit□\ p - value = Pr(|t_{df=97}| > |t|) \approx\ 0.04865697 < 0.05\ \Rightarrow reject\ \mathrm{H_0}$$

Case 2 : $\mathrm{H_0}: \beta_2 = 0$ v.s. $\mathrm{H_1}: \beta_2 \neq 0$

$$t-statistic\ t_2=\frac{\widehat{\beta_2}}{se(\widehat{\beta_2})}=\frac{1.01}{1.133723}=0.8908705\ ,$$

$$wit\square\ p-value=Pr(|t_{df=97}|>|t|)\approx\ 0.375203>0.05\ \Rightarrow not\ reject\ \mathrm{H_0}$$

(d)

```
> lm(y~x1)

Call:
lm(formula = y ~ x1)

Coefficients:
(Intercept)           x1
      2.112        1.976

> summary(lm(y~x1))

Call:
lm(formula = y ~ x1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.89495 -0.66874 -0.07785  0.59221  2.45560

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.1124     0.2307   9.155 8.27e-15 ***
x1            1.9759     0.3963   4.986 2.66e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.055 on 98 degrees of freedom
Multiple R-squared:  0.2024,	Adjusted R-squared:  0.1942
F-statistic: 24.86 on 1 and 98 DF,  p-value: 2.661e-06
```

$$\mathrm{H_0}:\ \beta_1=0\ \text{ v.s. }\ \mathrm{H_1}:\ \beta_1\neq 0$$

$$\overline{x_1}=\frac{\sum_{i=1}^{100}x_{1i}}{100}=0.5178471\ ,\bar{y}=\frac{\sum_{i=1}^{100}y_i}{100}=3.135623$$

$$S_{x_1x_1}=\sum_{i=1}^{100}(x_{1i}-\overline{x_1})^2=7.088559\ ,\ S_{x_1y}=\sum_{i=1}^{100}(x_{1i}-\overline{x_1})(y_i-\bar{y})=14.00649$$

$$\Rightarrow\widehat{\beta_1}=\frac{S_{x_1y}}{S_{x_1x_1}}=\frac{14.00649}{7.088559}=1.975929\ ,\qquad \widehat{\beta_0}=\bar{y}-\widehat{\beta_1}\times\overline{x_1}=2.112394$$

$$\Rightarrow se(\widehat{\beta_0})=\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\overline{x_1}^2}{S_{x_1x_1}}}=0.2307\ ,se(\widehat{\beta_1})=\sqrt{\frac{\hat{\sigma}^2}{S_{x_1x_1}}}=\ 0.3963$$

$$\Rightarrow t-statistic\ t_0=\frac{\widehat{\beta_0}}{se(\widehat{\beta_0})}=9.155\ ,p-value=Pr(|t_{df=198}|>|t|)\approx 8.27\times10^{-15}<0.05$$

$$t-statistic\ t_1=\frac{\widehat{\beta_1}}{se(\widehat{\beta_1})}=4.986, p-value=Pr(|t_{df=98}|>|t|)\approx 2.66\times10^{-6}<0.05$$

$$\Rightarrow \text{reject } \mathrm{H_0}\text{ , it means } \beta_1\neq 0$$

(e)

```
> lm(y~x2)

Call:
lm(formula = y ~ x2)

Coefficients:
(Intercept)           x2
       2.39         2.90

> summary(lm(y~x2))

Call:
lm(formula = y ~ x2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.62687 -0.75156 -0.03598  0.72383  2.44890

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.3899     0.1949   12.26  < 2e-16 ***
x2            2.8996     0.6330    4.58 1.37e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.072 on 98 degrees of freedom
Multiple R-squared:  0.1763,    Adjusted R-squared:  0.1679
F-statistic: 20.98 on 1 and 98 DF,  p-value: 1.366e-05
```

$$\mathrm{H_0}:\ \beta_1 = 0\ \text{ v.s. }\ \mathrm{H_1}:\ \beta_1 \neq 0$$

$$\overline{x_2} = \frac{\sum_{i=1}^{100} x_{2i}}{100} = 0.2571656\ ,\ \bar{y} = \frac{\sum_{i=1}^{100} y_i}{100} = 3.135623$$

$$S_{x_2x_2} = \sum_{i=1}^{100}(x_{2i} - \overline{x_2})^2 = 2.868343\ ,\ S_{x_2y} = \sum_{i=1}^{100}(x_{2i} - \overline{x_2})(y_i - \bar{y}) = 8.317006$$

$$\Rightarrow \widehat{\beta_1} = \frac{S_{x_2y}}{S_{x_2x_2}} = \frac{8.317006}{2.868343} = 2.90\ ,\qquad \widehat{\beta_0} = \bar{y} - \widehat{\beta_1} \times \overline{x_2} = 2.39$$

$$\Rightarrow se\left(\widehat{\beta_0}\right) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{x_1}^2}{S_{x_1x_1}}} = 0.1949\ ,se\left(\widehat{\beta_1}\right) = \sqrt{\frac{\hat{\sigma}^2}{S_{x_1x_1}}} = 0.6330$$

$$\Rightarrow t - statistic\ t_0 = \frac{\widehat{\beta_0}}{se\left(\widehat{\beta_0}\right)} = 12.26\ ,p - value = Pr\left(\left|t_{df=198}\right| > |t|\right) \approx < 2 \times 10^{-16} < 0.05$$

$$t - statistic\ t_1 = \frac{\widehat{\beta_1}}{se\left(\widehat{\beta_1}\right)} = 4.58, p - value = Pr\left(\left|t_{df=98}\right| > |t|\right) \approx 1.37 \times 10^{-5} < 0.05$$

$$\Rightarrow \text{reject } \mathrm{H_0}\text{ , it means } \beta_1 \neq 0$$

(f)

By (c) we get only x_1 have an effact on y , in (e) we get that x_2 have an effect on y , this phenomenon may be due to the collinearity of x_1 and x_2 , x_2 does not effect y directly , but it have an effect on y through x_1 .

$$VIF(\widehat{\beta_1}) = \frac{Var(\widehat{\beta_1}|multiple)}{Var(\widehat{\beta_1}|simple)} = \frac{se^2(\widehat{\beta_1}|multiple)}{se^2(\widehat{\beta_1}|simple)} = \left(\frac{0.7212}{0.3963}\right)^2 = 3.31179$$

$$VIF(\widehat{\beta_2}) = \frac{Var(\widehat{\beta_2}|multiple)}{Var(\widehat{\beta_2}|simple)} = \frac{se^2(\widehat{\beta_2}|multiple)}{se^2(\widehat{\beta_2}|simple)} = \left(\frac{1.1337}{0.633}\right)^2 = 3.207664$$

(g)

- If fit a least squares regression to predict y using x1 and x2

```
> ge=lm(y~x1+x2)
> summary(ge)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.69309 -0.68184 -0.04583  0.75224  2.29389 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.2665     0.2303   9.840 2.45e-16 ***
x1            0.1671     0.5246   0.318    0.751    
x2            3.1371     0.7703   4.073 9.37e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.079 on 99 degrees of freedom
Multiple R-squared:  0.246,	Adjusted R-squared:  0.2308 
F-statistic: 16.15 on 2 and 99 DF,  p-value: 8.501e-07
```

Case 1 : $\mathrm{H_0}: \beta_1 = 0$ v.s. $\mathrm{H_1}: \beta_1 \neq 0$

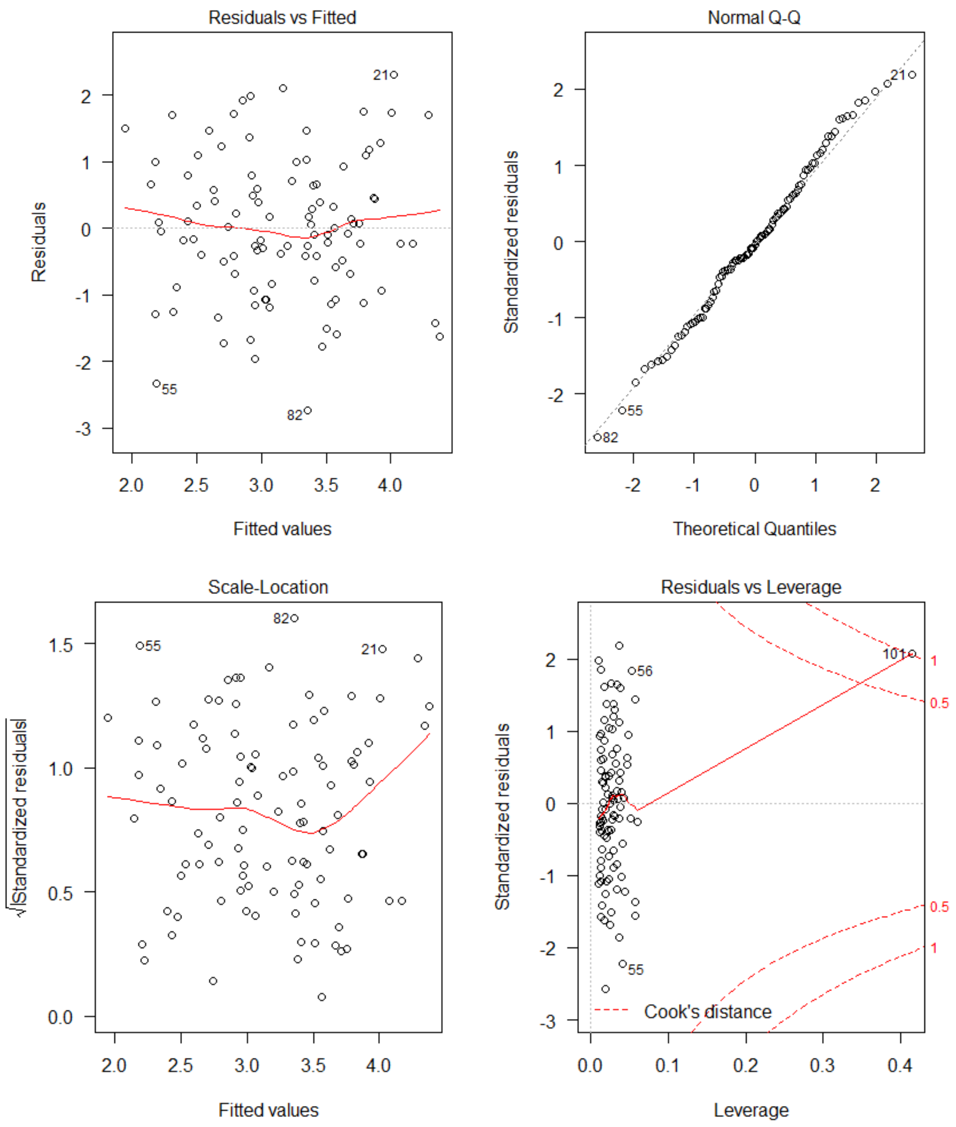
$$t-statistic\ t_1 = \frac{\widehat{\beta_1}}{se(\widehat{\beta_1})} = 0.318\,,$$

$wit\square\ p-value = Pr(|t_{df=97}| > |t|) \approx\ 0.751 > 0.05\ \Rightarrow not\ reject\ \mathrm{H_0}$, it means $\beta_1 = 0$

Case 2 : $\mathrm{H_0}: \beta_2 = 0$ v.s. $\mathrm{H_1}: \beta_2 \neq 0$

$$t-statistic\ t_2 = \frac{\widehat{\beta_2}}{se(\widehat{\beta_2})} = 4.073,$$

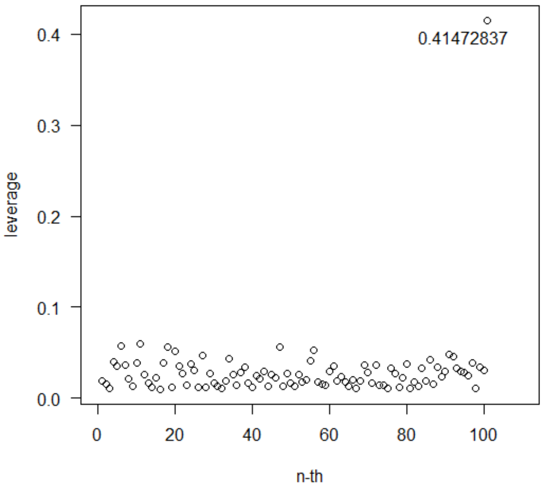
$wit\square\ p-value = Pr(|t_{df=97}| > |t|) \approx\ 0.0000937 < 0.05\ \Rightarrow\ reject\ \mathrm{H_0}$, it means $\beta_2 \neq 0$



The data 101 does not label in residual plot and qqplot may not be an outlier but it have high leverage in the plot residual.

```
> lev=hatvalues(ge)
> lev
         1          2          3          4          5          6          7          8          9         10 
0.01905793 0.01594849 0.01097584 0.04036164 0.03579376 0.05738552 0.03643856 0.02106144 0.01367365 0.03845500 
        11         12         13         14         15         16         17         18         19         20 
0.06007045 0.02568754 0.01702432 0.01222831 0.02305638 0.01005414 0.03916306 0.05642240 0.01240634 0.05168118 
        21         22         23         24         25         26         27         28         29         30 
0.03585426 0.02693541 0.01479629 0.03817939 0.03081323 0.01247725 0.04657630 0.01231205 0.02737875 0.01715106 
        31         32         33         34         35         36         37         38         39         40 
0.01280296 0.01114580 0.01868965 0.04300077 0.02576320 0.01380719 0.02829778 0.03375430 0.01677644 0.01161848 
        41         42         43         44         45         46         47         48         49         50 
0.02509987 0.02185972 0.02889894 0.01304567 0.02631106 0.02236651 0.05657476 0.01289427 0.02746237 0.01611627 
        51         52         53         54         55         56         57         58         59         60 
0.01332252 0.02653629 0.01734972 0.01989190 0.04102604 0.05274307 0.01808709 0.01499209 0.01376303 0.02960779 
        61         62         63         64         65         66         67         68         69         70 
0.03473115 0.01865022 0.02338266 0.01814573 0.01291854 0.02052027 0.01108508 0.01933631 0.03633142 0.02815957 
        71         72         73         74         75         76         77         78         79         80 
0.01648355 0.03678740 0.01433799 0.01482634 0.01027552 0.03307247 0.02688469 0.01209299 0.02294836 0.03814756 
        81         82         83         84         85         86         87         88         89         90 
0.01076185 0.01803646 0.01316211 0.03261485 0.01860412 0.04179835 0.01604374 0.03386478 0.02365038 0.02897279 
        91         92         93         94         95         96         97         98         99        100 
0.04829557 0.04627626 0.03248348 0.02958189 0.02865524 0.02514269 0.03892864 0.01139885 0.03397113 0.03088514 
       101 
0.41472837 
```

The leverage statistics for 101 points



The leverage statistic with 101_{th} =0.41472837

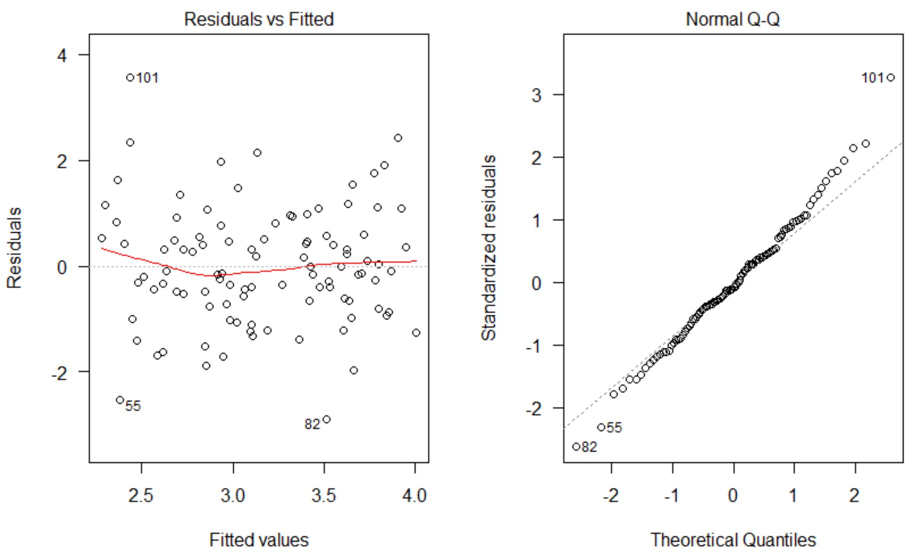
- If fit a least squares regression to predict y using only x1

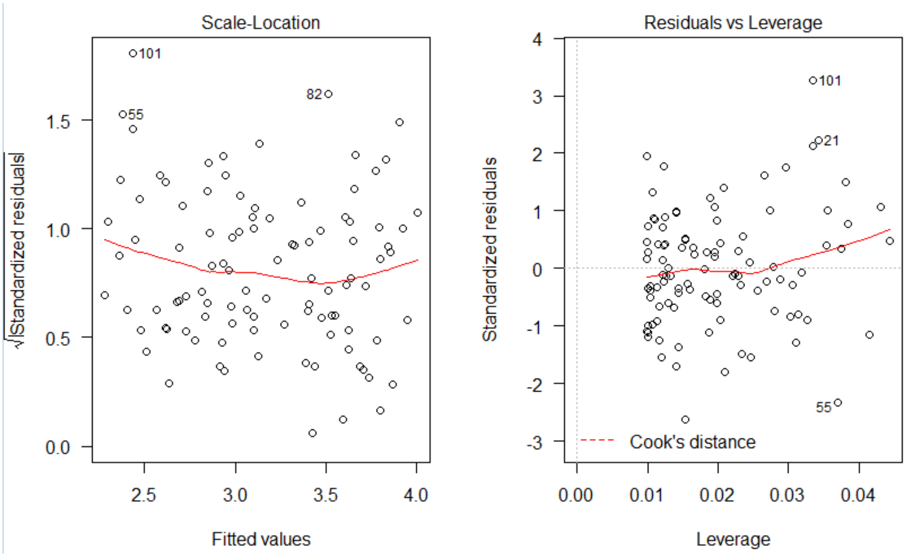
```
> x2=c(x2,0.8)
> y=c(y,6)
>
> gc=lm(y~x1)
> summary(gc)

Call:
lm(formula = y ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8848 -0.6542 -0.0769  0.6137  3.4510 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.3921     0.2454   9.747 3.55e-16 ***
x1            1.5691     0.4255   3.687 0.000369 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.16 on 100 degrees of freedom
Multiple R-squared:  0.1197,	Adjusted R-squared:  0.1109 
F-statistic:  13.6 on 1 and 100 DF,  p-value: 0.0003686
```

$$H_0 : \beta_1 = 0 \text{ v.s. } H_1 : \beta_1 \neq 0$$

$$t-statistic\ t_1 = \frac{\widehat{\beta_1}}{se(\widehat{\beta_1})} = 3.687, p-value = Pr(|t_{df=98}| > |t|) \approx 0.000369 < 0.05$$

$$\Rightarrow \text{reject } H_0 \text{, it means } \beta_1 \neq 0$$





The data 101 does label in residual plot and qqplot be an outlier but it not have high leverage in the plot residual.

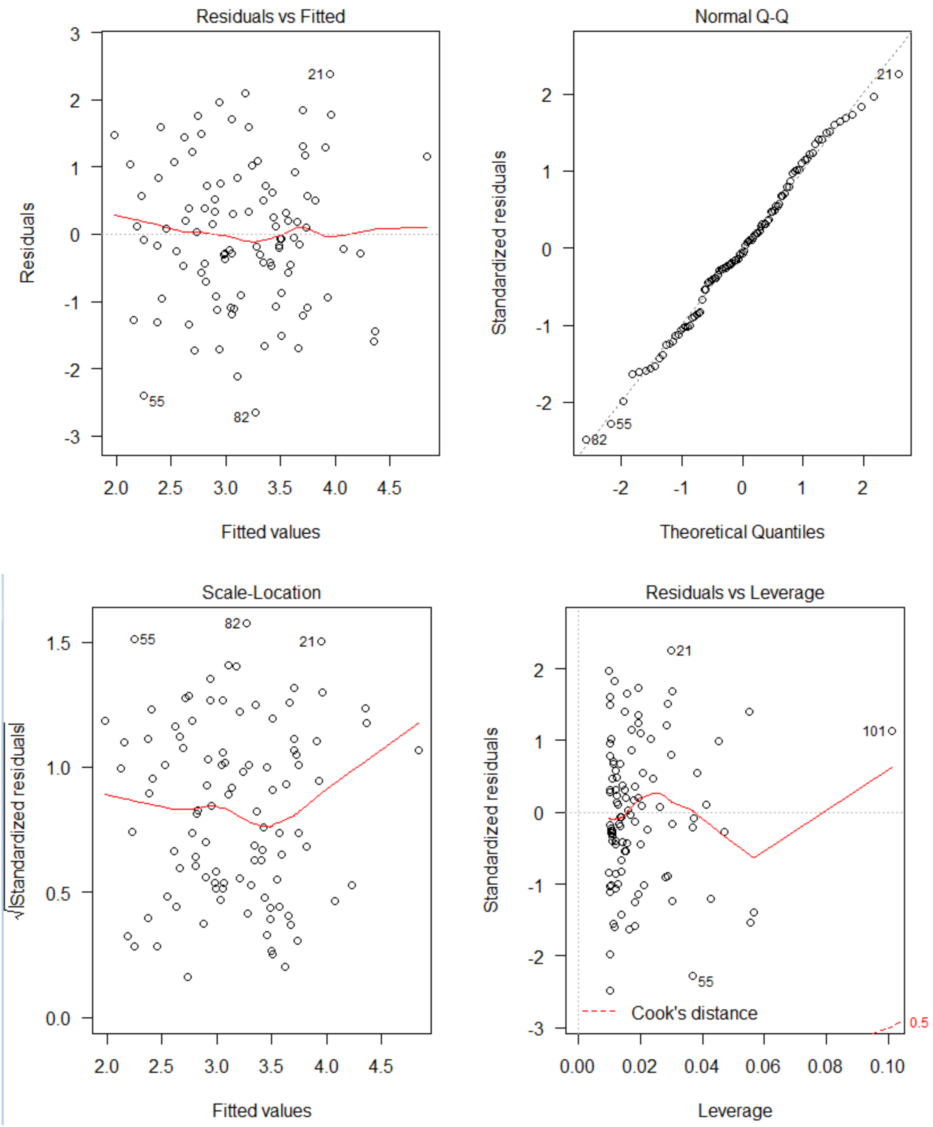
- If fit a least squares regression to predict y using only x2

```
> gd=lm(y~x2)
> summary(gd)

Call:
lm(formula = y ~ x2)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.66396 -0.67794 -0.06181  0.75541  2.32512 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.3085     0.1879   12.28  < 2e-16 ***
x2            3.2981     0.5786    5.70 1.21e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.074 on 100 degrees of freedom
Multiple R-squared:  0.2452,    Adjusted R-squared:  0.2377 
F-statistic: 32.49 on 1 and 100 DF,  p-value: 1.214e-07
```

$$\mathrm{H}_0 : \beta_1 = 0 \text{ v.s. } \mathrm{H}_1 : \beta_1 \neq 0$$

$$t - statistic\ t_1 = \frac{\widehat{\beta_1}}{se(\widehat{\beta_1})} = 5.7, p - value = Pr(|t_{df=98}| > |t|) \approx 1.21 \times 10^{-7} < 0.05$$

$$\Rightarrow \text{reject } \mathrm{H}_0 \text{, it means } \beta_1 \neq 0$$



The data 101 does not label in residual plot and qqplot may not be an outlier but it have high leverage in the plot residual.

	Outlier	leverage
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$	×	V
$y = \beta_0 + \beta_1 x_1 + \varepsilon$	V	×
$y = \beta_0 + \beta_2 x_2 + \varepsilon$	×	V

Appendix

R-code

```
########P.125########EX14########

rm(list=ls())

#####a#####

set.seed(1)

x1=runif(100)

x2=0.5*x1+rnorm(100)/10

y=2+2*x1+0.3*x2+rnorm(100)

#####b#####

r12=sum((x1-mean(x1))*(x2-mean(x2)))/(sum((x1-mean(x1))^2)*sum((x2-mean(x2))^2))^0.5

r12

library(ggplot2)

f=lm(y~x1+x2)

ggplot(f,aes(x=x1,y=x2))+geom_point(shape=1)+geom_smooth(method=lm,se=FALSE)

#####c#####

f=lm(y~x1+x2)

f

summary(f)

#####d#####

lm(y~x1)

summary(lm(y~x1))

#####e#####

lm(y~x2)

summary(lm(y~x2))

#####g#####

x1=c(x1,0.1)
```

```
x2=c(x2,0.8)

y=c(y,6)

gc=lm(y~x1)

summary(gc)

par(mfrow=c(1,2))

plot(gc,las=1)

gd=lm(y~x2)

summary(gd)

par(mfrow=c(1,2))

plot(gd,las=1)

ge=lm(y~x1+x2)

summary(ge)

par(mfrow=c(1,2))

plot(ge,las=1)
```