HW6 M072040019 梅瀚中

14. 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 $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$? 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 $H_0 : \beta_2 = 0$?

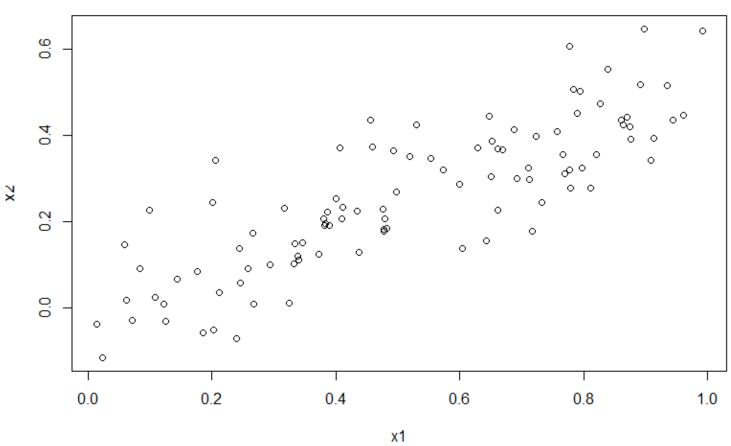
(a)

```
y2 <- 2*x1+0.3*x2+rnorm(100)+2
```

β0=2+rorm(100), β1=2, β2=0.3

(b)

```
> cor(x1,x2)
[1] 0.8351212
```



(C)

```
> fit <- lm(y2~x1+x2)
> summary(fit)

Call:
lm(formula = y2 ~ 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
```

$\hat{\beta}$0=2.1305, 跟 β0 的值很接近, $\hat{\beta}$1=1.4396 跟 β1 的值沒那麼接近, $\hat{\beta}$2=1.0097 跟 β2 的值相差更多。我們可以拒絕 H0:β1=0,當 alpha=0.05 時,因為 p-value=0.0487;我們無法拒絕 H0:β2=0,因為 P-value 高達 0.3754。

(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 your answer.

(g) Now suppose we obtain 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.

(d)

```
> fit1 <- lm(y2~x1)
> summary(fit1)

Call:
lm(formula = y2 ~ 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
```

我們可以拒絕 H0,因為 P-VALUE 很小

(E)

```
> fit2 <- lm(y2~x2)
> summary(fit2)

Call:
lm(formula = y2 ~ 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
```

我們可以拒絕 H0,因為 P-VALUE 很小

(f)

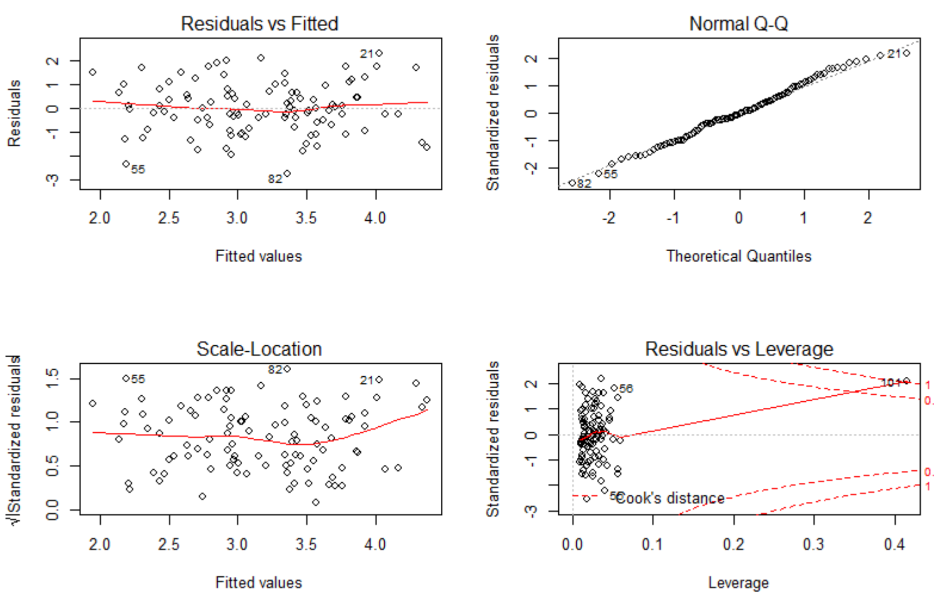
他們沒有矛盾,因為 x1 和 x2 存在高度相關性,導致共線性的問題

(g)

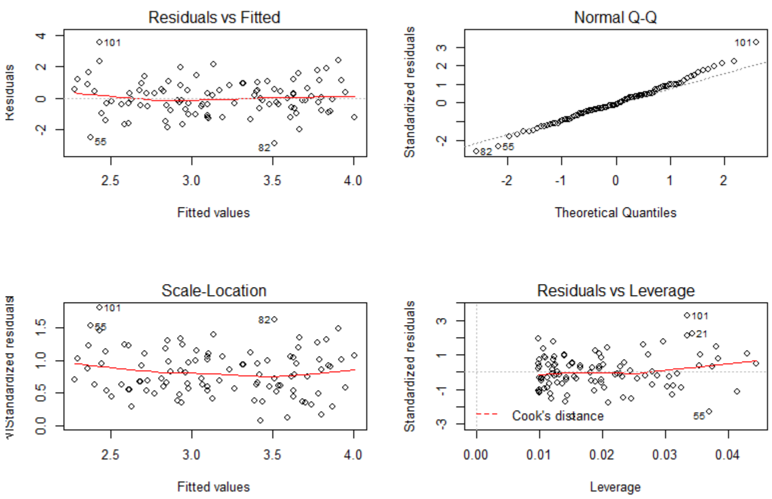
```
> x1=c(x1, 0.1)
> x2=c(x2, 0.8)
> y2=c(y2,6)
> fit <- lm(y2~x1+x2)
> summary(fit)

Call:
lm(formula = y2 ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.73348 -0.69318 -0.05263  0.66385  2.30619 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.2267     0.2314   9.624 7.91e-16 ***
x1            0.5394     0.5922   0.911  0.36458    
x2            2.5146     0.8977   2.801  0.00614 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.075 on 98 degrees of freedom
Multiple R-squared:  0.2188,	Adjusted R-squared:  0.2029 
F-statistic: 13.72 on 2 and 98 DF,  p-value: 5.564e-06

> fit1 <- lm(y2~x1)
> summary(fit1)

Call:
lm(formula = y2 ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8897 -0.6556 -0.0909  0.5682  3.5665 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.2569     0.2390   9.445 1.78e-15 ***
x1            1.7657     0.4124   4.282 4.29e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.111 on 99 degrees of freedom
Multiple R-squared:  0.1562,	Adjusted R-squared:  0.1477 
F-statistic: 18.33 on 1 and 99 DF,  p-value: 4.295e-05
```

```
> fit2 <- lm(y2~x2)
> summary(fit2)

Call:
lm(formula = y2 ~ x2)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.64729 -0.71021 -0.06899  0.72699  2.38074 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.3451     0.1912  12.264  < 2e-16 ***
x2            3.1190     0.6040   5.164 1.25e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.074 on 99 degrees of freedom
Multiple R-squared:  0.2122,    Adjusted R-squared:  0.2042 
F-statistic: 26.66 on 1 and 99 DF,  p-value: 1.253e-06
```

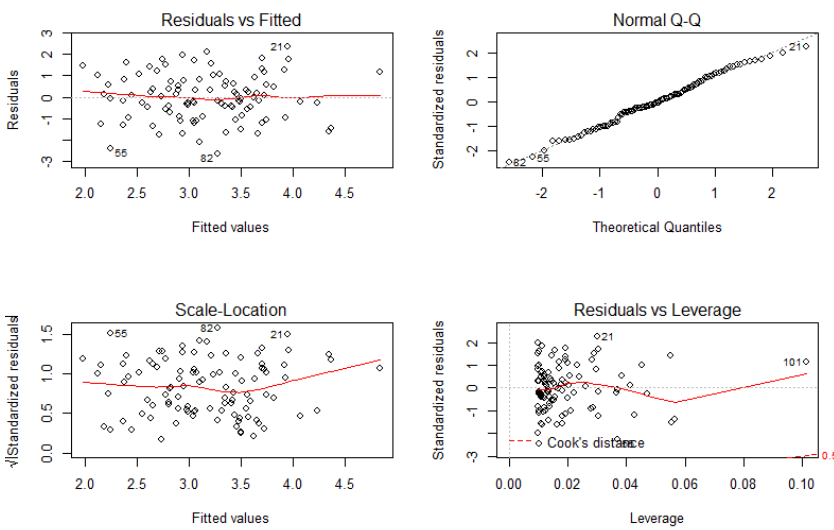
Plot(fit)



Plot(fit1)



Plot(fit2)



Plot(fit)圖中,點 101 是一個 high leverage point,plot(fit2)圖中點 101 是 high leverage point。

在 plot(fit1)圖中,點 101 是個 outlier

15. This problem involves the `Boston` data set, which we saw in the lab for this chapter. We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

(a) For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

(b) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

(c) How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.

(d) Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor X, fit a model of the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon.$$

(a)

```r
library(MASS)
colnames(Boston)
fk1 <- lm( crim~zn,data=Boston)
fk2 <- lm( crim~indus,data=Boston)
fk3 <- lm( crim~chas,data=Boston)
fk4 <- lm( crim~nox,data=Boston)
fk5 <- lm( crim~rm,data=Boston)
fk6 <- lm( crim~age,data=Boston)
fk7 <- lm( crim~dis,data=Boston)
fk8 <- lm( crim~rad,data=Boston)
fk9 <- lm( crim~tax,data=Boston)
fk10 <- lm( crim~ptratio,data=Boston)
fk11 <- lm( crim~black,data=Boston)
fk12 <- lm( crim~lstat,data=Boston)
fk13 <- lm( crim~medv,data=Boston)
fk3
```

```
> summary(fk3)

Call:
lm(formula = crim ~ chas, data = Boston)

Residuals:
   Min     1Q Median     3Q    Max 
-3.738 -3.661 -3.435  0.018 85.232 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3.7444     0.3961   9.453   <2e-16 ***
chas         -1.8928     1.5061  -1.257    0.209    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.597 on 504 degrees of freedom
Multiple R-squared:  0.003124,	Adjusted R-squared:  0.001146 
F-statistic: 1.579 on 1 and 504 DF,  p-value: 0.2094
```

只有 chas 這個自變數對 crim 在作簡單回歸時,沒有顯著,其他變數都是顯著的。

(b)

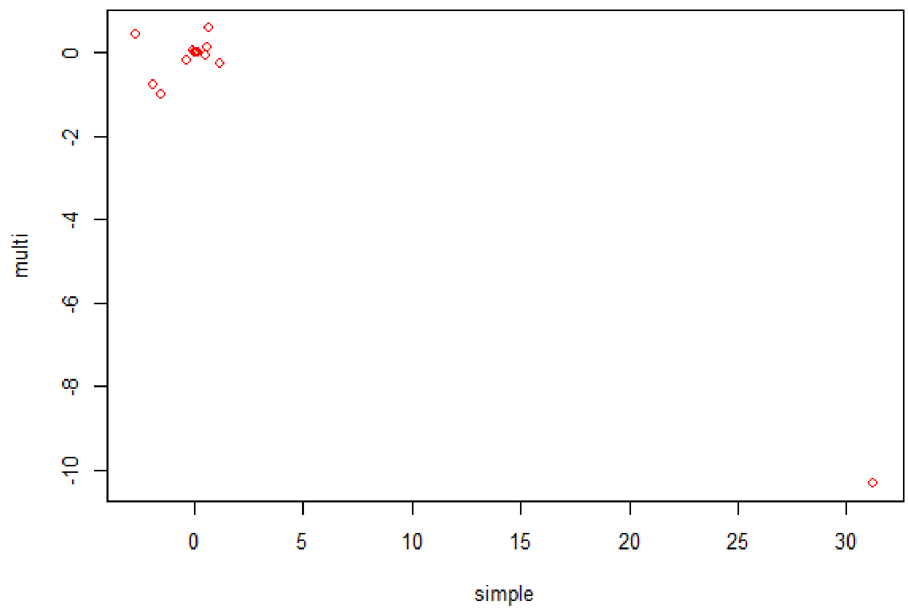
```
> fk <-lm( crim~.,data=Boston)
> summary(fk)

Call:
lm(formula = crim ~ ., data = Boston)

Residuals:
   Min     1Q Median     3Q    Max 
-9.924 -2.120 -0.353  1.019 75.051 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  17.033228   7.234903   2.354 0.018949 *  
zn            0.044855   0.018734   2.394 0.017025 *  
indus        -0.063855   0.083407  -0.766 0.444294    
chas         -0.749134   1.180147  -0.635 0.525867    
nox         -10.313535   5.275536  -1.955 0.051152 .  
rm            0.430131   0.612830   0.702 0.483089    
age           0.001452   0.017925   0.081 0.935488    
dis          -0.987176   0.281817  -3.503 0.000502 ***
rad           0.588209   0.088049   6.680 6.46e-11 ***
tax          -0.003780   0.005156  -0.733 0.463793    
ptratio      -0.271081   0.186450  -1.454 0.146611    
black        -0.007538   0.003673  -2.052 0.040702 *  
lstat         0.126211   0.075725   1.667 0.096208 .  
medv         -0.198887   0.060516  -3.287 0.001087 ** 
```

Zn,indus,black 這 3 個變數在顯著水準為 0.05 時可以拒絕 H0,

Medv 這個變數在顯著水準為 0.01 時,可以拒絕 H0

Dis,rad 這 2 個變數我們有充分證據拒絕 H0

(C)



(d)

	b_0	p-value	b_1	p-value	b_2	p-value	b_3	p-value
zn	3.613524	1.547150e-20	-38.74984	4.697806e-06	23.939832	4.420507e-03	-10.071868	2.295386e-01
indus	3.613524	3.606468e-25	78.59082	8.854243e-24	-24.394796	1.086057e-03	-54.129763	1.196405e-12
nox	3.613524	2.742908e-26	81.37202	2.457491e-26	-28.828594	7.736755e-05	-60.361894	6.961110e-16
rm	3.613524	1.026665e-20	-42.37944	5.128048e-07	26.576770	1.508545e-03	-5.510342	5.085751e-01
age	3.613524	5.918933e-23	68.18201	4.878803e-17	37.484470	2.291156e-06	21.353207	6.679915e-03
dis	3.613524	1.060226e-25	-73.38859	1.253249e-21	56.373036	7.869767e-14	-42.621877	1.088832e-08
tax	3.613524	8.955923e-29	112.64583	6.976314e-49	32.087251	3.665348e-06	-7.996811	2.438507e-01
ptratio	3.613524	1.270767e-21	56.04523	1.565484e-11	24.774824	2.405468e-03	-22.279737	6.300514e-03
black	3.613524	2.139710e-22	-74.43120	2.730082e-19	5.926419	4.566044e-01	-4.834565	5.436172e-01
lstat	3.613524	4.939398e-24	88.06967	1.678072e-27	15.888164	3.780418e-02	-11.574022	1.298906e-01
medv	3.613524	7.024110e-31	-75.05761	4.930818e-27	88.086211	2.928577e-35	-48.033435	1.046510e-12

全部變數 1 次方係數項都顯著

全部變數除了 black 變數 2 次方係數項顯著

indus、dis、nox 、age、ptratio、medv 3 次項係數顯著

1. Using a little bit of algebra, prove that (4.2) is equivalent to (4.3). In other words, the logistic function representation and logit representation for the logistic regression model are equivalent.

2. It was stated in the text that classifying an observation to the class for which (4.12) is largest is equivalent to classifying an observation to the class for which (4.13) is largest. Prove that this is the case. In other words, under the assumption that the observations in the kth class are drawn from a $N(\mu_k, \sigma^2)$ distribution, the Bayes' classifier assigns an observation to the class for which the discriminant function is maximized.

3. This problem relates to the QDA model, in which the observations within each class are drawn from a normal distribution with a class-specific mean vector and a class specific covariance matrix. We consider the simple case where $p = 1$; i.e. there is only one feature.

   Suppose that we have K classes, and that if an observation belongs to the kth class then X comes from a one-dimensional normal distribution, $X \sim N(\mu_k, \sigma_k^2)$. Recall that the density function for the one-dimensional normal distribution is given in (4.11). Prove that in this case, the Bayes' classifier is *not* linear. Argue that it is in fact quadratic.

   Hint: For this problem, you should follow the arguments laid out in Section 4.4.2, but without making the assumption that $\sigma_1^2 = \ldots = \sigma_K^2$.