2017 Big Data – R programming Homework 4

2017/11/05

范真瑋

Part I: Regression

0. 使用 package : MASS 中的 ”UScrime” dataframe 當作資料

```
> data(UScrime)
> head(UScrime)
    M So  Ed Po1 Po2  LF  M.F Pop  NW  U1 U2 GDP Ineq     Prob    Time    y
1 151  1  91  58  56 510  950  33 301 108 41 394  261 0.084602 26.2011  791
2 143  0 113 103  95 583 1012  13 102  96 36 557  194 0.029599 25.2999 1635
3 142  1  89  45  44 533  969  18 219  94 33 318  250 0.083401 24.3006  578
4 136  0 121 149 141 577  994 157  80 102 39 673  167 0.015801 29.9012 1969
5 141  0 121 109 101 591  985  18  30  91 20 578  174 0.041399 21.2998 1234
6 121  0 110 118 115 547  964  25  44  84 29 689  126 0.034201 20.9995  682
> attach(UScrime)
```

1. 將所需的 library 製成 function，並 Show 出所需的資料

```
> loadLibrary <- function() {
+   library(MASS)
+   library(car)
+   library(e1071)
+   library(class)
+   library(gmodels)
+   print("MASS")
+   print("car")
+   print("e1071")
+   print("class")
+   print("gmodels")
+   print("The library have been loaded.")
+ }
> #1
> loadLibrary()
[1] "MASS"
[1] "car"
[1] "e1071"
[1] "class"
[1] "gmodels"
[1] "The library have been loaded."
```

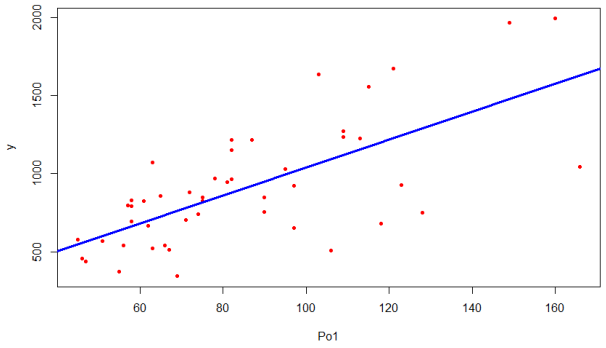
2. 使用 single regression 找出 UScrime 中的 y (犯罪率)與其他欄位的相關性(找出一條正相關以及一條負相關的回歸線)

```
> #2-1 positive correlation
> lm.fit1=lm(y~Po1)
> lm.fit1

Call:
lm(formula = y ~ Po1)

Coefficients:
(Intercept)          Po1
    144.464        8.948

> plot(Po1, y, pch = 20, col = "red")
> abline(lm.fit1, lwd = 3, col = "blue")
```

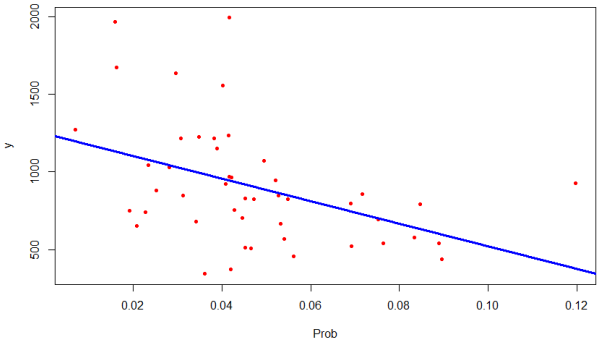



```
> #2-2 negative correlation
> lm.fit2=lm(y~Prob)
> lm.fit2

Call:
lm(formula = y ~ Prob)

Coefficients:
(Intercept)         Prob
       1247        -7271

> plot(Prob, y, pch = 20, col = "red")
> abline(lm.fit2, lwd = 3, col = "blue")
```



3. 使用 multiple regression 找出 UScrime 中的 y (犯罪率)與其他欄位的相關性並且檢測 5 個自變數共線性的問題(使用 VIF)，

```
> #3-1 multiple regression
> crime.all <- lm(y~.,UScrime)
> crime.all

Call:
lm(formula = y ~ ., data = UScrime)

Coefficients:
(Intercept)            M           So           Ed          Po1          Po2           LF          M.F          Pop
 -5984.2876       8.7830      -3.8035      18.8324      19.2804     -10.9422      -0.6638       1.7407      -0.7330
         NW           U1           U2          GDP         Ineq         Prob         Time
     0.4204      -5.8271      16.7800       0.9617       7.0672   -4855.2658      -3.4790
> #3-2
> crime5 <- lm(y~NW+Time+M.F+U2+LF,UScrime)
> crime5

Call:
lm(formula = y ~ NW + Time + M.F + U2 + LF, data = UScrime)

Coefficients:
(Intercept)           NW         Time          M.F           U2           LF
 -4109.8645       0.4476      12.2348       2.9709      11.6027       2.3692

> vif(crime5)
      NW     Time      M.F       U2       LF
1.199239 1.315131 1.851558 1.349486 1.934181
```

4. 找出 multiple regression 中最佳解 (使用逐步回歸(step)尋找)

```
> #4
> step(crime.all)
Start:  AIC=514.65
y ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 + U2 +
    GDP + Ineq + Prob + Time

       Df Sum of Sq     RSS    AIC
- So    1        29 1354974 512.65
- LF    1      8917 1363862 512.96
- Time  1     10304 1365250 513.00
- Pop   1     14122 1369068 513.14
- NW    1     18395 1373341 513.28
- M.F   1     31967 1386913 513.74
- GDP   1     37613 1392558 513.94
- Po2   1     37919 1392865 513.95
<none>              1354946 514.65
- U1    1     83722 1438668 515.47
- Po1   1    144306 1499252 517.41
- U2    1    181536 1536482 518.56
- M     1    193770 1548716 518.93
- Prob  1    199538 1554484 519.11
- Ed    1    402117 1757063 524.86
- Ineq  1    423031 1777977 525.42

Step:  AIC=512.65
y ~ M + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 + U2 + GDP +
    Ineq + Prob + Time

       Df Sum of Sq     RSS    AIC
- Time  1     10341 1365315 511.01
- LF    1     10878 1365852 511.03
- Pop   1     14127 1369101 511.14
- NW    1     21626 1376600 511.39
- M.F   1     32449 1387423 511.76
- Po2   1     37954 1392929 511.95
- GDP   1     39223 1394197 511.99
<none>              1354974 512.65
- U1    1     96420 1451395 513.88
- Po1   1    144302 1499277 515.41
- U2    1    189859 1544834 516.81
- M     1    195084 1550059 516.97
- Prob  1    204463 1559437 517.26
- Ed    1    403140 1758114 522.89
- Ineq  1    488834 1843808 525.13

Step:  AIC=511.01
y ~ M + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 + U2 + GDP +
    Ineq + Prob

       Df Sum of Sq     RSS    AIC
- LF    1     10533 1375848 509.37
- NW    1     15482 1380797 509.54
- Pop   1     21846 1387161 509.75
- Po2   1     28932 1394247 509.99
- GDP   1     36070 1401385 510.23
- M.F   1     41784 1407099 510.42
<none>              1365315 511.01
- U1    1     91420 1456735 512.05
- Po1   1    134137 1499452 513.41
Step:  AIC=506.33
y ~ M + Ed + Po1 + M.F + Pop + U1 + U2 + GDP + Ineq + Prob

       Df Sum of Sq     RSS    AIC
- Pop   1     22345 1426575 505.07
- GDP   1     32142 1436371 505.39
- M.F   1     36808 1441037 505.54
<none>              1404229 506.33
- U1    1     86373 1490602 507.13
- U2    1    205814 1610043 510.76
- Prob  1    218607 1622836 511.13
- M     1    307001 1711230 513.62
- Ed    1    389502 1793731 515.83
- Ineq  1    608627 2012856 521.25
- Po1   1   1050202 2454432 530.57

Step:  AIC=505.07
y ~ M + Ed + Po1 + M.F + U1 + U2 + GDP + Ineq + Prob

       Df Sum of Sq     RSS    AIC
- GDP   1     26493 1453068 503.93
<none>              1426575 505.07
- M.F   1     84491 1511065 505.77
- U1    1     99463 1526037 506.24
- Prob  1    198571 1625145 509.20
- U2    1    208880 1635455 509.49
- M     1    320926 1747501 512.61
- Ed    1    386773 1813348 514.35
- Ineq  1    594779 2021354 519.45
- Po1   1   1127277 2553852 530.44

Step:  AIC=503.93
y ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob

       Df Sum of Sq     RSS    AIC
<none>              1453068 503.93
- M.F   1    103159 1556227 505.16
- U1    1    127044 1580112 505.87
- Prob  1    247978 1701046 509.34
- U2    1    255443 1708511 509.55
- M     1    296790 1749858 510.67
- Ed    1    445788 1898855 514.51
- Ineq  1    738244 2191312 521.24
- Po1   1   1672038 3125105 537.93

Call:
lm(formula = y ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob,
    data = UScrime)

Coefficients:
(Intercept)            M           Ed          Po1          M.F           U1           U2         Ineq         Prob
  -6426.101        9.332       18.012       10.265        2.234       -6.087       18.735        6.133    -3796.032
```

Part II: Machine Learning

0. 使用 package : MASS 中的 ”Rabbit” dataframe 當作資料

```
> #Part II
> #0
> data(Rabbit)
> head(Rabbit)
  BPchange   Dose Run Treatment Animal
1      0.5   6.25  C1   Control     R1
2      4.5  12.50  C1   Control     R1
3     10.0  25.00  C1   Control     R1
4     26.0  50.00  C1   Control     R1
5     37.0 100.00  C1   Control     R1
6     32.0 200.00  C1   Control     R1
```

1. 使用 SVM 來分析預測 ”Animal”，並將結果以 CrossTable 呈現

```
> #1 SVM
> x <- subset(Rabbit, select = -Animal)
> y <- Rabbit$Animal
> svm_model <- svm(Animal~., data = Rabbit)
> pred_model <- predict(svm_model, x)
> CrossTable(pred_model, y)
```

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  60
```

	y					
pred_model	R1	R2	R3	R4	R5	Row Total
R1	12	0	0	0	0	12
	38.400	2.400	2.400	2.400	2.400	
	1.000	0.000	0.000	0.000	0.000	0.200
	1.000	0.000	0.000	0.000	0.000	
	0.200	0.000	0.000	0.000	0.000	
R2	0	12	0	0	0	12
	2.400	38.400	2.400	2.400	2.400	
	0.000	1.000	0.000	0.000	0.000	0.200
	0.000	1.000	0.000	0.000	0.000	
	0.000	0.200	0.000	0.000	0.000	
R3	0	0	12	0	0	12
	2.400	2.400	38.400	2.400	2.400	
	0.000	0.000	1.000	0.000	0.000	0.200
	0.000	0.000	1.000	0.000	0.000	
	0.000	0.000	0.200	0.000	0.000	
R4	0	0	0	12	0	12
	2.400	2.400	2.400	38.400	2.400	
	0.000	0.000	0.000	1.000	0.000	0.200
	0.000	0.000	0.000	1.000	0.000	
	0.000	0.000	0.000	0.200	0.000	
R5	0	0	0	0	12	12
	2.400	2.400	2.400	2.400	38.400	
	0.000	0.000	0.000	0.000	1.000	0.200
	0.000	0.000	0.000	0.000	1.000	
	0.000	0.000	0.000	0.000	0.200	
Column Total	12	12	12	12	12	60
	0.200	0.200	0.200	0.200	0.200	

2. 使用 KNN 來分析預測 ”Animal”，並將結果以 CrossTable 呈現

```
> #2 KNN
> Rabbit$Run = as.integer(as.factor(Rabbit$Run))
> Rabbit$Treatment = as.integer(as.factor(Rabbit$Treatment))
> ind <- sample(2, nrow(Rabbit), replace=TRUE, prob=c(0.67, 0.33))
> ind
 [1] 1 1 1 1 1 1 1 1 1 1 1 2 2 1 2 1 1 2 1 1 1 2 1 2 1 2 1 1 1 2 1 2 2 2 1 1 1 1 1 1 1 1 2 1 1 1 2 2 1 1 2 1 1 1 1 2 2
[60] 2
> Rabbit.training <- Rabbit[ind==1, 1:4]
> head(Rabbit.training)
  BPchange   Dose Run Treatment
1      0.5   6.25   1         1
2      4.5  12.50   1         1
3     10.0  25.00   1         1
4     26.0  50.00   1         1
5     37.0 100.00   1         1
6     32.0 200.00   1         1
> Rabbit.test <- Rabbit[ind==2, 1:4]
> head(Rabbit.test)
   BPchange   Dose Run Treatment
13     0.75   6.25   3         1
14     3.00  12.50   3         1
16    14.00  50.00   3         1
19     1.25   6.25   4         1
23    33.00 100.00   4         1
25     1.50   6.25   5         1
> Rabbit.trainLabels <- Rabbit[ind==1, 5]
> print(Rabbit.trainLabels)
 [1] R1 R1 R1 R1 R1 R1 R2 R2 R2 R2 R2 R2 R3 R3 R3 R4 R4 R4 R4 R5 R5 R5 R5 R1 R1 R2 R2 R2 R2 R2 R2 R3 R3 R3 R3 R3 R4 R4 R4
[40] R5 R5 R5
Levels: R1 R2 R3 R4 R5
> Rabbit.testLabels <- Rabbit[ind==2, 5]
> print(Rabbit.testLabels)
 [1] R3 R3 R3 R4 R4 R5 R5 R1 R1 R1 R1 R3 R4 R4 R4 R5 R5 R5
Levels: R1 R2 R3 R4 R5
> kv = round(sqrt(nrow(Rabbit)))
> Rabbit_pred <- knn(train = Rabbit.training, test = Rabbit.test, cl = Rabbit.trainLabels, k=kv)
> Rabbit_pred
 [1] R2 R2 R2 R2 R2 R2 R5 R2 R2 R4 R1 R3 R2 R5 R3 R4 R3 R4
Levels: R1 R2 R3 R4 R5
> CrossTable(x = Rabbit.testLabels, y = Rabbit_pred, prop.chisq = FALSE)
```

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  18


                 | Rabbit_pred
Rabbit.testLabels |        R1 |        R2 |        R3 |        R4 |        R5 | Row Total |
------------------|-----------|-----------|-----------|-----------|-----------|-----------|
               R1 |         1 |         2 |         0 |         1 |         0 |         4 |
                  |     0.250 |     0.500 |     0.000 |     0.250 |     0.000 |     0.222 |
                  |     1.000 |     0.222 |     0.000 |     0.333 |     0.000 |           |
                  |     0.056 |     0.111 |     0.000 |     0.056 |     0.000 |           |
------------------|-----------|-----------|-----------|-----------|-----------|-----------|
               R3 |         0 |         3 |         1 |         0 |         0 |         4 |
                  |     0.000 |     0.750 |     0.250 |     0.000 |     0.000 |     0.222 |
                  |     0.000 |     0.333 |     0.333 |     0.000 |     0.000 |           |
                  |     0.000 |     0.167 |     0.056 |     0.000 |     0.000 |           |
------------------|-----------|-----------|-----------|-----------|-----------|-----------|
               R4 |         0 |         3 |         1 |         0 |         1 |         5 |
                  |     0.000 |     0.600 |     0.200 |     0.000 |     0.200 |     0.278 |
                  |     0.000 |     0.333 |     0.333 |     0.000 |     0.500 |           |
                  |     0.000 |     0.167 |     0.056 |     0.000 |     0.056 |           |
------------------|-----------|-----------|-----------|-----------|-----------|-----------|
               R5 |         0 |         1 |         1 |         2 |         1 |         5 |
                  |     0.000 |     0.200 |     0.200 |     0.400 |     0.200 |     0.278 |
                  |     0.000 |     0.111 |     0.333 |     0.667 |     0.500 |           |
                  |     0.000 |     0.056 |     0.056 |     0.111 |     0.056 |           |
------------------|-----------|-----------|-----------|-----------|-----------|-----------|
     Column Total |         1 |         9 |         3 |         3 |         2 |        18 |
                  |     0.056 |     0.500 |     0.167 |     0.167 |     0.111 |           |
------------------|-----------|-----------|-----------|-----------|-----------|-----------|
```