APLM Final exam 610611105 黃三騰

(1)(20 pts) Please explain what the ridge and Lasso regression are, and the

difference between the least square estimator, the ridge estimator and Lasso estimator.

Sol:

Based on linear model:

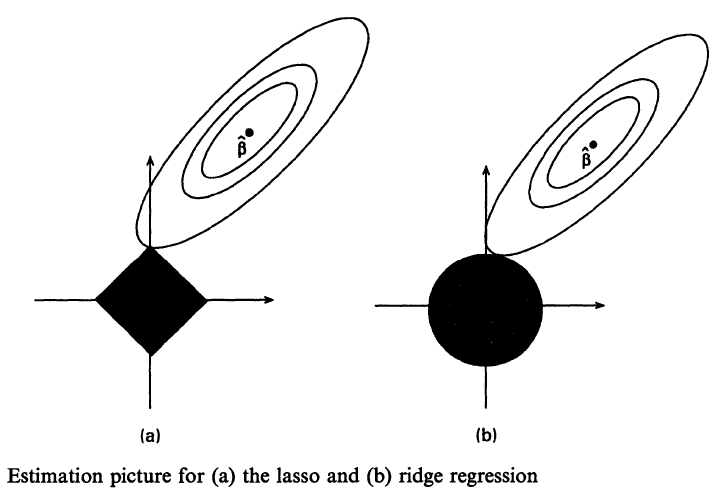
Ridge regression:

We have Ridge regression estimator is , which is

not unbiased(i.e. ).

LASSO:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Unbiased | Shrinkage coefficients | Sparsity |
| Ordinary Least Squares | ○ | ╳ | ╳ |
| Ridge | ╳ | ○ | ╳ |
| LASSO | ╳ | ○ | ○ |



Because of the L1-norm, the LASSO can set coefficients exactly equal to zero while

Ridge and Ordinary Least Squares can not.

However, the LASSO is not unbiased since the penalty function is . It

cause that the larger coefficients is, the more penalized value would get.

(3) A manager would like to increase the return on an investment, which depends

on 17 financial products. But, some of financial products can not affect the return. In

the attached data set including 2000 data, the first 17 columns show the investment

for each product, and the last one is the corresponding return.

(a) (25 pts) Please analyze the data (It is required to provide ANOVA table.)?

(b) Assume that the budget (US 100000) is limited. Based on your analysis results

in (a) and (b),

1. (15 pts) please make your comparison on two different strategies (20000,30000,0,40000,10000,0....,0) and (50000,0,20000,0,30000,0....,0).
2. (20 pts) Also, please discuss the prediction of their return throught the

theoretical way.

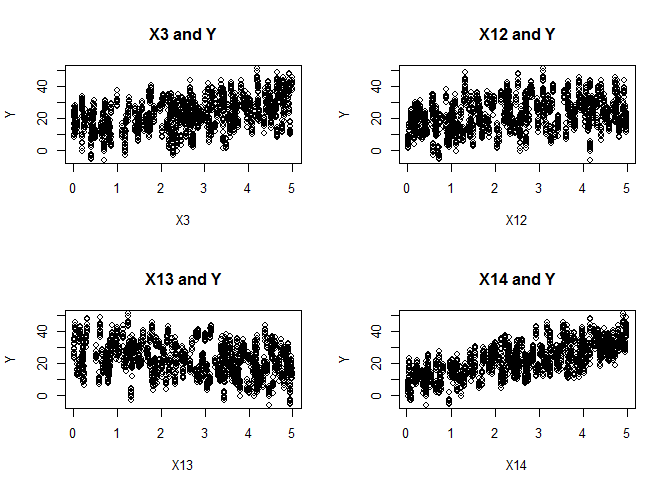
sol:

(a)

(i) Scatter plot of each Predictor() and Response().

將所有解釋變數()對反應變數()的散布圖畫出，列出有明顯線性關係的變數:

分別是 。(其它解釋變數的圖則無明顯的線性趨勢)



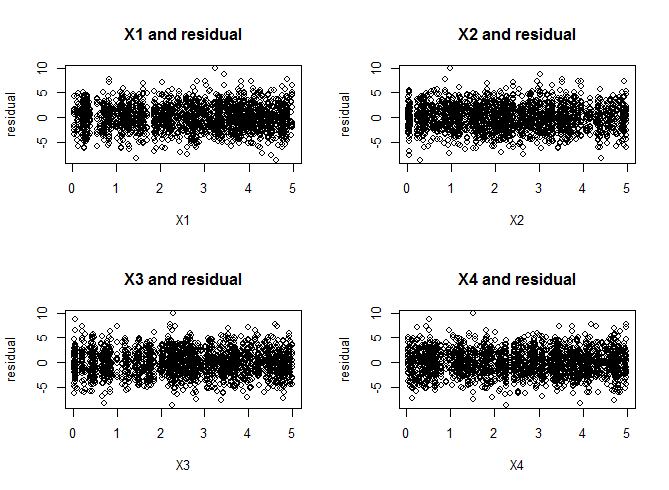
初步判斷後，決定使用線性模型分析。

備註: 這裡我們不考慮將截距項放入模型中，原因為我們的資料是有關17種金融產品投資金額()與其投資後的報酬()，而如果沒有投資金額便不會有投資後的報酬。同時也可以看到上面的散布圖是通過原點的，故截距項應為0。

(ii)

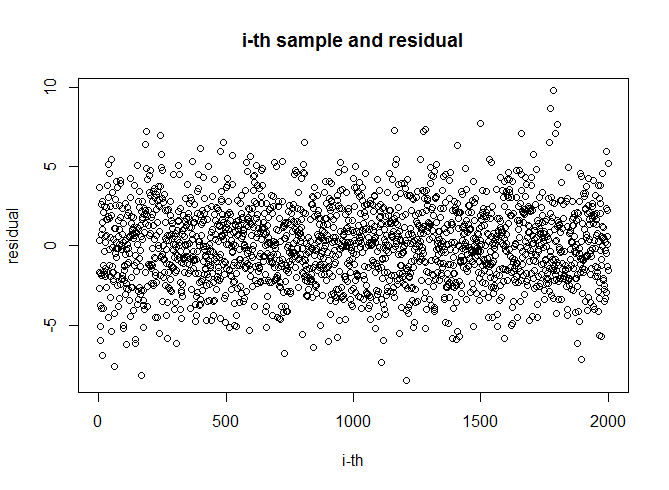
Consider the model:

1. 建立線性模型後，我們先來觀察殘差項()的畫圖:
2. 解釋變數()與殘差項()



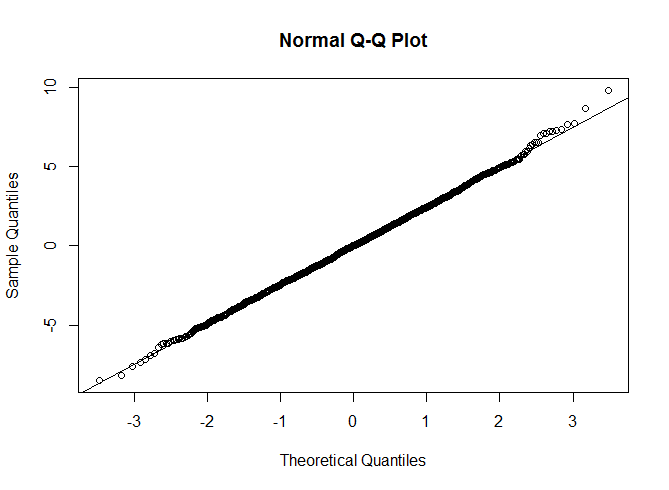
解釋變數與殘差呈現水平無特別模式的關係。(解釋變數的結果與上圖相似，故此處只列出前四個解釋變數)

1. 資料的順序(時間先後)與殘差項()



呈現水平無特別模式的關係，代表沒有違背的變異數為常數的假設，以及模型並不需考慮將資料的順序(時間先後)納入解釋變數。

1. 殘差項()的Q-Q Plot



代表沒有違背服從常態分配的假設。

1. 使用線性模型估計的結果:

|  |
| --- |
| Residuals:  Min 1Q Median 3Q Max  -8.5003 -1.6910 0.0214 1.6751 9.8326  Coefficients:  Estimate Std. Error t value Pr(>|t|)  X1 -0.4903032 0.0391813 -12.514 < 2e-16 \*\*\*  X2 -0.0216555 0.0413625 -0.524 0.6006  X3 3.0576619 0.0382537 79.931 < 2e-16 \*\*\*  X4 0.0086003 0.0378851 0.227 0.8204  X5 1.1838885 0.0399983 29.598 < 2e-16 \*\*\*  X6 0.1387319 0.0351258 3.950 8.1e-05 \*\*\*  X7 0.0812278 0.0418093 1.943 0.0522 .  X8 -0.0066523 0.0384489 -0.173 0.8627  X9 0.0006042 0.0429780 0.014 0.9888  X10 -0.0296778 0.0408369 -0.727 0.4675  X11 0.0583555 0.0374444 1.558 0.1193  X12 2.0379867 0.0363056 56.134 < 2e-16 \*\*\*  X13 -2.4571694 0.0383099 -64.139 < 2e-16 \*\*\*  X14 4.9831330 0.0399680 124.678 < 2e-16 \*\*\*  X15 0.0876630 0.0370593 2.365 0.0181 \*  X16 -0.0096647 0.0408550 -0.237 0.8130  X17 0.0285892 0.0357266 0.800 0.4237  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 2.49 on 1983 degrees of freedom  Multiple R-squared: 0.9893, Adjusted R-squared: 0.9892  F-statistic: 1.079e+04 on 17 and 1983 DF, p-value: < 2.2e-16 |

此表提供的估計值，以及

各係數是否為0之假設檢定: .

這裡和都很高，代表解釋變數()與反應變數()是有很強的線性關係。

此外，

|  |
| --- |
| ANOVA Table  Source df Sum Sq Mean Sq F P-value  Regression 16 1136975 71060.96 11462.18 < 2.2e-16  Error 1983 12294 6.1996  Total 1999 |

ANOVA表給出解釋變數()與反應變數():的假設檢定結果，解釋變數()與反應變數()是有線性關係的。

(iii)

由(ii)的分析結果，得到解釋變數()與反應變數()是有線性關係。然而，在一些解釋變數上的假設檢定顯示無法拒絕，亦即模型中可能有些是不重要的解釋變數()。

這裡我們打算利用Forward Stepwise regression的方式，進行選模:

(備註: 操作的細節是參照上課筆記，設定)

以下表格紀錄選模過程，

|  |
| --- |
| Step 0  空模型.  Step 1  加入模型中，此時: .  Step 2 and Step 3  加入模型中，此時: ；沒有變數被刪除.  加入模型中，此時: ；沒有變數被刪除.  加入模型中，此時: ；沒有變數被刪除.  加入模型中，此時: ；沒有變數被刪除.  加入模型中，此時: ；沒有變數被刪除.  加入模型中，此時: ；沒有變數被刪除.  沒有變數加入，且沒有變數被刪除，最後模型為: |

利用選模結果，建立線性模型:

|  |
| --- |
| Residuals:  Min 1Q Median 3Q Max  -8.4709 -1.6916 0.0339 1.6861 10.0568  Coefficients:  Estimate Std. Error t value Pr(>|t|)  X1 -0.44335 0.03504 -12.652 < 2e-16 \*\*\*  X3 3.08859 0.03578 86.330 < 2e-16 \*\*\*  X5 1.19762 0.03731 32.096 < 2e-16 \*\*\*  X6 0.14796 0.03306 4.476 8.04e-06 \*\*\*  X12 2.06026 0.03403 60.541 < 2e-16 \*\*\*  X13 -2.42925 0.03503 -69.348 < 2e-16 \*\*\*  X14 5.01280 0.03818 131.308 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 2.492 on 1993 degrees of freedom  Multiple R-squared: 0.9892, Adjusted R-squared: 0.9892  F-statistic: 2.615e+04 on 7 and 1993 DF, p-value: < 2.2e-16 |

得到估計值，

(b)

1.

根據兩種投資策略:

strage1

strage2

利用(a)最終模型的估計值，

求得

我們發現strage2的投資方式是可以得到比較多的報酬。

主要原因為: 同樣預算下，strage2投資在；而strage1投資在。根據我們的模型，投資在。

2.

Given a new data, , the corresponding response .

To construct the 95% confidence interval for ,

Replace by ,

Hence, confidence interval for is

Here,

confidence interval for is and

confidence interval for is .

從信賴區間我們得到兩種投資策略的報酬所落在的範圍，故判斷strage2的投資方式可以得到比strage1多的報酬。

附錄:

在(3)分析資料時，我同時也嘗試了Penalized Least Squares的方式進行選模:

我們使用LASSO、SCAD的方式估計，利用10-folds Cross Validation選出:

|  |  |
| --- | --- |
| 方法 | 選到的模型(以所對應的值大到小列出) |
| Forward Stepwise  LASSO()  SCAD() |  |

|  |  |
| --- | --- |
| 方法 | 估計值 |
| Forward Stepwise |  |
| LASSO() |  |
| SCAD() |  |

在這裡LASSO與SCAD都選到比較大的模型()，推測可能是不夠大，無法將係數估計壓到零。(註: 在LASSO與SCAD的估計結果若則將它設為零)