

國立政治大學 111 學年度第二學期
迴歸分析(一)期末 R 程式加分考題

15

Department: 統計二 ID: 110304058 Name: 郭祐丞

Subject: **Regression Analysis (I)**

Date: 2023/06/15, Time: 11:00~12:00 (60 minutes)

注意事項:

1. 本次考題以 R 程式(Rgui 或 RStudio)方式作答，其他程式不允許。
2. 考試過程中可查詢書本、教學講義或上網，禁止利用 messenger, IG, Line 等等通訊軟體。
3. 禁止疑似作弊行為。
4. 本答案卷上請務必於 **R Console** 內複製「執行後的程式碼及結果(含圖形)」，於本答案卷貼上(Courier New, 10 點字，白底黑字)，不能只有程式碼，不能只有報表。最後，將每小題之答案(不能只印出報表，要助教去找答案)，在小題最後以打字(英文)作答(Times New Roman, 12 點字，白底黑字)。
5. 請依序註明題號: (1)a, (1)b, (2)a 等等。
6. 作答完請將此 word 檔存檔，檔名為「學號-姓名-Regression-R-Midterm.docx」(更改成自己「學號、姓名」)並上傳至 <http://ftp.hmwu.idv.tw:8080/login.html?lang=tchinese> 或點選教師網站首頁【作業考試上傳區】。
7. 帳號: **reg111**，密碼: 上課教室號碼，資料夾: 「**20230615-FinalExam**」
8. 如果上傳網站出現「空白頁」，請將滑鼠移至「網址列」後，按「Enter」即可。若再不行，請換其它瀏覽器(IE/Edge/Firefox/Chrome)
9. 上傳檔案無法刪除，若要上傳更新檔，請於主檔名後加「-2」，例如:「學號-姓名-Regression-R-Midterm-2.docx」。

Notes:

1. This is an Open Book exam; you are free to use any materials including laptop, tablet and internets.
2. Smart phone and the communication software/APP (e.g., Messenger, IG, LINE, WeChat,..) are prohibited.
3. Copy the R codes and the results from **R Console** and paste it to this answer sheet.
4. Change the file name of this answer sheet according to your ID and Full Name. Upload the answer sheet to <http://ftp.hmwu.idv.tw:8080/login.html?lang=tchinese>

5. Account: **reg111** , password: classroom number.

(1)
30%

Data file: **CDI.csv**

Refer to the **CDI** data set in Appendix C.2. The number of active physicians (Y) is to be regressed against total population (X_1), total personal income (X_2), and geographic region (X_3, X_4, X_5).

- Fit a first-order regression model. Let $X_3 = 1$ if NE and 0 otherwise, $X_4 = 1$ if NC and 0 otherwise, and $X_5 = 1$ if S and 0 otherwise.
- Examine whether the effect for the northeastern region on number of active physicians differs from the effect for the north central region by constructing an appropriate 90 percent confidence interval. Interpret your interval estimate.
- Test whether any geographic effects are present; use $\alpha = .10$. State the alternatives, decision rule, and conclusion. What is the P -value of the test?

Data Set C.2 CDI

This data set provides selected county demographic information (CDI) for 440 of the most populous counties in the United States. Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county. Counties with missing data were deleted from the data set. The information generally pertains to the years 1990 and 1992. The 17 variables are:

Variable Number	Variable Name	Description
1	Identification number	1–440
2	County	County name
3	State	Two-letter state abbreviation
4	Land area	Land area (square miles)
5	Total population	Estimated 1990 population
6	Percent of population aged 18–34	Percent of 1990 CDI population aged 18–34
7	Percent of population 65 or older	Percent of 1990 CDI population aged 65 years old or older
8	Number of active physicians	Number of professionally active nonfederal physicians during 1990
9	Number of hospital beds	Total number of beds, cribs, and bassinets during 1990
10	Total serious crimes	Total number of serious crimes in 1990, including murder, rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft, as reported by law enforcement agencies
11	Percent high school graduates	Percent of adult population (persons 25 years old or older) who completed 12 or more years of school
12	Percent bachelor's degrees	Percent of adult population (persons 25 years old or older) with bachelor's degree
13	Percent below poverty level	Percent of 1990 CDI population with income below poverty level
14	Percent unemployment	Percent of 1990 CDI labor force that is unemployed
15	Per capita income	Per capita income of 1990 CDI population (dollars)
16	Total personal income	Total personal income of 1990 CDI population (in millions of dollars)
17	Geographic region	Geographic region classification is that used by the U.S. Bureau of the Census, where: 1 = NE, 2 = NC, 3 = S, 4 = W

1	2	3	4	5	6	7	8	9	10
1	Los_Angeles	CA	4060	8863164	32.1	9.7	23677	27700	688936
2	Cook	IL	946	5105067	29.2	12.4	15153	21550	436936
3	Harris	TX	1729	2818199	31.3	7.1	7553	12449	253526
...
438	Montgomery	TN	539	100498	35.7	7.9	87	188	6537
439	Maui	HI	1159	100374	26.2	11.3	192	182	7130
440	Morgan	AL	582	100043	26.3	11.7	122	464	4693

11	12	13	14	15	16	17
70.0	22.3	11.6	8.0	20786	184230	4
73.4	22.8	11.1	7.2	21729	110928	2
74.9	25.4	12.5	5.7	19517	55003	3
...
77.9	16.5	10.8	8.0	13169	1323	3
77.0	17.8	5.7	3.2	18504	1857	4
69.4	15.5	9.4	7.1	16458	1647	3

```

+8 > data <- read.csv("CDI.csv", header=F)
> data <- data[,c(8,5,16,17)]
> X3 <- ifelse(data$V17==1, 1, 0)
> X4 <- ifelse(data$V17==2, 1, 0)
> X5 <- ifelse(data$V17==3, 1, 0)
> CDI <- data.frame(Y=data[,1], X1=data[,2], X2=data[,3], X3=X3, X4=X4, X5=X5)
> # (a)
> m <- lm(Y~., data=CDI)
> summary(m)

```

Use factor

Call:

```
lm(formula = Y ~ ., data = CDI)
```

Residuals:

```

      Min       1Q   Median       3Q      Max
-1866.8 -207.7  -81.5    72.4  3721.7

```

Coefficients:

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.075e+02  7.028e+01  -2.952  0.00332 **
X1           5.515e-04  2.835e-04   1.945  0.05243 .
X2           1.070e-01  1.325e-02   8.073  6.8e-15 ***
X3           1.490e+02  8.683e+01   1.716  0.08685 .
X4           1.455e+02  8.515e+01   1.709  0.08817 .
X5           1.912e+02  8.003e+01   2.389  0.01731 *

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 566.1 on 434 degrees of freedom

Multiple R-squared: 0.9011, Adjusted R-squared: 0.8999

F-statistic: 790.7 on 5 and 434 DF, p-value: < 2.2e-16

> #(c)

> reduced <- lm(Y~X1+X2, data=CDI)

> anova(reduced,m)

Analysis of Variance Table

Model 1: Y ~ X1 + X2

Model 2: Y ~ X1 + X2 + X3 + X4 + X5

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	437	140967081				
2	434	139093455	3	1873626	1.9487	0.121

(a)

ty $\hat{Y} = -0.002075 + 0.00055X_1 + 0.1070X_2 + 0.01490X_3 + 0.01455X_4 + 0.01912X_5$

(c)

① $H_0: \beta_3 = \beta_4 = \beta_5 = 0$ versus $H_a: \text{not all } \beta_3, \beta_4, \beta_5 \text{ equal to zero}$

② Level of significance: $\alpha = 0.1$

③ Test statistic: $F^* = \frac{MSR(X_3, X_4, X_5 | X_2, X_1)}{MSE(X_1, X_2, X_3, X_4, X_5)} \sim F(3, 436)$

④ Decision rule: If $p\text{-value} < \alpha$, then we reject H_0

⑤ Decision: $p\text{-value} = 0.121 > 0.1 = \alpha \Rightarrow \text{Do not reject } H_0$

⑥ Conclusion: $\beta_3, \beta_4, \beta_5$ are not significantly different from zero.

(2)

30%

Data file: **Kidney_Function_Data.csv**

Kidney function. Creatinine clearance (Y) is an important measure of kidney function, but is difficult to obtain in a clinical office setting because it requires 24-hour urine collection. To determine whether this measure can be predicted from some data that are easily available, a kidney specialist obtained the data that follow for 33 male subjects. The predictor variables are serum creatinine concentration (X_1), age (X_2), and weight (X_3).

Subject i	X_{i1}	X_{i2}	X_{i3}	Y_i
1	.71	38	71	132
2	1.48	78	69	53
3	2.21	69	85	50
...
31	1.53	70	75	52
32	1.58	63	62	73
33	1.37	68	52	57

- Using first-order and second-order terms for each of the three predictor variables (centered around the mean) in the pool of potential X variables (including cross products of the first-order terms), find the three best hierarchical subset regression models according to the AIC_p criterion.
- Is there much difference in AIC_p for the three best subset models?

```
> data <- read.csv("Kidney_Function_Data.csv", header=F)
> colnames(data) <- c("Y", "X1", "X2", "X3")
> kidney <- data.frame(Y=data$Y,
+                      x1=data$X1-mean(data$X1),
+                      x2=data$X2-mean(data$X2),
+                      x3=data$X3-mean(data$X3),
+                      x1_2=(data$X1-mean(data$X1))^2,
+                      x2_2=(data$X2-mean(data$X2))^2,
+                      x3_2=(data$X3-mean(data$X3))^2)
> # (a)
> library(ALSM)
> BestSub(kidney[,2:7], kidney[,1], method='aic', num=3)
  p 1 2 3 4 5 6    SSEp      r2    r2.adj      Cp    AICp    SBCp    PRESSp
1 2 1 0 0 0 0 11068.478 0.6429007 0.63138133 53.382603 195.9065 198.8995 12402.603
1 2 0 1 0 0 0 17169.856 0.4460535 0.42818429 98.795117 210.3953 213.3883 19676.109
1 2 0 0 1 0 0 27286.682 0.1196571 0.09125894 174.094579 225.6823 228.6753 30198.308
2 3 1 1 0 0 0 7666.102 0.7526706 0.73618198 30.058742 185.7858 190.2754 9348.907
2 3 1 0 1 0 0 8713.158 0.7188897 0.70014903 37.851973 190.0107 194.5002 10326.342
2 3 1 0 0 1 0 9144.499 0.7049735 0.68530505 41.062439 191.6052 196.0947 11323.048
3 4 1 1 1 0 0 4499.974 0.8548186 0.83979979 8.493272 170.2055 176.1916 5963.853
```

```

3 4 1 0 1 0 1 0 5908.643 0.8093710 0.78965080 18.977987 179.1929 185.1789 7940.411
3 4 1 1 0 1 0 0 6817.335 0.7800541 0.75730112 25.741379 183.9136 189.8997 8425.011
4 5 1 1 1 1 0 0 3946.962 0.8726602 0.85446881 6.377214 167.8784 175.3609 5276.847
4 5 1 1 1 0 0 1 4122.121 0.8670091 0.84801042 7.680917 169.3113 176.7939 5795.937
4 5 1 1 1 0 1 0 4349.803 0.8596635 0.83961539 9.375554 171.0855 178.5680 6625.771
5 6 1 1 1 1 0 1 3667.570 0.8816742 0.85976199 6.297696 167.4556 176.4347 5176.363
5 6 1 1 1 1 1 0 3717.227 0.8800721 0.85786324 6.667293 167.8994 176.8785 5830.247
5 6 1 1 1 0 1 1 4017.272 0.8703918 0.84639032 8.900526 170.4611 179.4401 6515.487
6 7 1 1 1 1 1 1 3493.218 0.8872992 0.86129136 7.000000 167.8484 178.3239 5762.964

```

(a)

best subsets	aic
1,2,3,4,5,6	167.8484
1,2,3,4,6	167.4556
1,2,3,4	167.8784

(b)

There is not much difference

(3)	Data file: Performance_Ability_Data.csv
40%	

Performance ability. A psychologist conducted a study to examine the nature of the relation, if any, between an employee's emotional stability (X) and the employee's ability to perform in a task group (Y). Emotional stability was measured by a written test for which the higher the score, the greater is the emotional stability. Ability to perform in a task group ($Y = 1$ if able, $Y = 0$ if unable) was evaluated by the supervisor. The results for 27 employees were:

i :	1	2	3	...	25	26	27
X_i :	474	432	453	...	562	506	600
Y_i :	0	0	0	...	1	0	1

Logistic regression model (14.20) is assumed to be appropriate.

- Find the maximum likelihood estimates of β_0 and β_1 . State the fitted response function.
- Obtain a scatter plot of the data with both the fitted logistic response function from part (a) and a lowess smooth superimposed. Does the fitted logistic response function appear to fit well?
- Obtain $\exp(b_1)$ and interpret this number.
- What is the estimated probability that employees with an emotional stability test score of 550 will be able to perform in a task group?
- Estimate the emotional stability test score for which 70 percent of the employees with this test score are expected to be able to perform in a task group.