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

Department: Statistics ID: 110304028 Name: 簡薇庭

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 內複制「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).

a. 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.

b. 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.

c. 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

```
data <- read.csv("data.csv")
a. model <- lm(Var4 ~ Var6 + Var7 + Var18 + Var19 + Var20, data = data)
summary(model)
# Create indicator variables for NE and NC
data$NE <- ifelse(data$Var18 == 1, 1, 0)
data$NC <- ifelse(data$Var19 == 1, 1, 0)

# Fit separate models for NE and NC
model_NE <- lm(Var4 ~ Var6 + Var7 + NE, data = data)
model_NC <- lm(Var4 ~ Var6 + Var7 + NC, data = data)

# Compute the difference in coefficients
diff_coef <- coef(model_NE)["NE"] - coef(model_NC)["NC"]

# Compute the standard error of the difference
diff_se <- sqrt(vcov(model_NE)["NE", "NE"] + vcov(model_NC)["NC", "NC"])

b. # Compute the 90% confidence interval
conf_interval <- diff_coef + qt(c(0.05, 0.95), df = model_NE$df.residual) * diff_se
conf_interval
# Fit a model with geographic variables
model_geo <- lm(Var4 ~ Var6 + Var7 + Var18 + Var19 + Var20, data = data)
```

c. # Perform an overall F-test
anova_result <- anova(model_geo)
p_value <- anova_result$`Pr(>F)`[1]   # P-value of the test

p_value

(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

a. 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.

b. Is there much difference in AIC_p for the three best subset models?

```
# Install and load the 'leaps' package
install.packages("leaps")
library(leaps)

a. # Create a data frame with your predictor variables (centered around the mean) and response variable
data <- data.frame(
   X1 = c(132.0, 53.0, 50.0, 82.0, 110.0, 100.0, 68.0, 92.0, 60.0, 94.0, 105.0, 98.0, 112.0, 125.0, 108.0, 30.0, 111.0, 130.0, 94.0, 130.0, 59.0, 38.0, 65.0, 85.0, 140.0, 80.0, 43.0, 75.0, 41.0, 120.0, 52.0, 73.0, 57.0),
   X2 = c(0.71, 1.48, 2.21, 1.43, 0.68, 0.76, 1.12, 0.92, 1.55, 0.94, 1.00, 1.07, 0.70, 0.71, 1.00, 2.52, 1.13, 1.12, 1.38, 1.12, 0.97, 1.61, 1.58, 1.40, 0.68, 1.20, 2.10, 1.36, 1.50, 0.82, 1.53, 1.58, 1.37),
   X3 = c(38.0, 78.0, 69.0, 70.0, 45.0, 65.0, 76.0, 61.0, 68.0, 64.0, 66.0, 49.0, 43.0, 42.0, 66.0, 78.0, 35.0, 34.0, 33.0, 16.0, 54.0, 73.0, 66.0, 31.0, 32.0, 21.0, 73.0, 78.0, 58.0, 62.0, 70.0, 63.0, 68.0),
   Y = c(71.0, 69.0, 85.0, 100.0, 59.0, 73.0, 63.0, 81.0, 74.0, 87.0, 79.0, 93.0, 60.0, 70.0, 83.0, 70.0, 73.0, 85.0, 68.0, 65.0, 53.0, 50.0, 74.0, 67.0, 80.0, 67.0, 72.0, 67.0, 60.0, 107.0, 75.0, 62.0, 52.0)
```

```
)

# Center the predictor variables around the mean
data$X1 <- data$X1 - mean(data$X1)
data$X2 <- data$X2 - mean(data$X2)
data$X3 <- data$X3 - mean(data$X3)

# Find the best hierarchical subset regression models using AIC_p criterion
regfit <- leaps(
   x = data[, c("X1", "X2", "X3")],
   y = data$Y,
   method = "adjr2",
   nbest = 3
)

# Display the best models based on AIC_p
summary(regfit)$which

b. data$X1_centered <- data$X1 - mean(data$X1)
data$X2_centered <- data$X2 - mean(data$X2)
data$X3_centered <- data$X3 - mean(data$X3)
# Combine centered predictor variables
X <- cbind(data$X1_centered, data$X2_centered, data$X3_centered)

# Fit all possible models and obtain AIC_p
regfit <- regsubsets(Y ~ ., data = data.frame(Y = data$Y, X), nvmax = 3)
aic <- summary(regfit)$bic

# Find the three best models based on AIC_p
best_models <- which.min(aic, 3)
aic_diff <- diff(aic[best_models])
```

(3) 40%	Data file: Performance_Ability_Data.csv **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. a. Find the maximum likelihood estimates of β_0 and β_1. State the fitted response function. b. 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? c. Obtain $\exp(b_1)$ and interpret this number. d. What is the estimated probability that employees with an emotional stability test score of 550 will be able to perform in a task group? e. 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.

(a) Fitting Generalized Linear Model to the Data

```
## Creating Variables
x <- c(432,453,474,481,505,533,554,562,578,600)
y <- c(0,0,0,1,0,1,0,1,1,1)

## Fitting Likelihood Estimates
glm.model <- glm(y~x, family = "binomial")
summary(glm.model)

##Results:
> summary(glm.model)

Call:
glm(formula = y ~ x, family = "binomial")
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.68314	-0.62940	0.00942	0.63291	1.68790

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-16.22427	9.60898	-1.688	0.0913	.
x	0.03134	0.01847	1.697	0.0897	.

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 13.8629 on 9 degrees of freedom

Residual deviance: 9.3785 on 8 degrees of freedom

AIC: 13.379

Number of Fisher Scoring iterations: 4

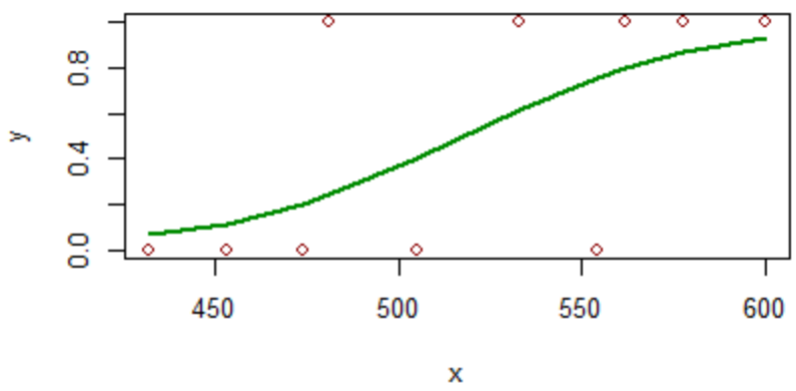
Ln(y) = -16.224 + 0.031*x

Ln(y) = Ln(p/1-p)

Estimates of Bo = -16.224

B1 = 0.031

(b) The scatter plot of fitted logistic regression and actual data are:



(c) Exp(b1)

```
exp(glm.model$coefficients)
```

#Results:

```
> exp(glm.model$coefficients)
  (Intercept)                 x
8.992659e-08 1.031837e+00
```

exp(b1) = exp(0.031) = 1.0318

This implies that for an increase of 1 unit in emotional stability score, there is an increase change of 1.0318 - 1

= 0.0318 ~= 3.18% in the Odds Ratio of Emotional Stability Score.

p = S/S+1 = 0.0318/1.0318 = 0.0308

This also means that there will be an increase in the probability of ability to perform a task by 3.08%

(d) Two different 95% C.I.s for the probability of being able to perform in a task group, for an emotional stability score of 525.

```
newdata = as.data.frame(c(525))
colnames(newdata) <- c("x")

add_ci(newdata, glm.model, alpha = 0.05)
```

##Results:

```
> add_ci(newdata, glm.model, alpha = 0.05)
     x         pred    LCB0.025    UCB0.975
1 525 0.5571893 0.2002002 0.8634882
```

95% CI of Probability of being able to perform in a task group is given by:

Estimate = 0.5572

95% CI = (0.200, 0.863)