37

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:______________ ID:__108703046_________ Name:__許祐菁

____

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

(1) a

```
(2)      > Task.data <- read.csv("data/CDI.csv", header = F)
(3)      > Task.data
(4)           V1             V2 V3    V4     V5    V6    V7    V8    V9    V10   V11   V12   V13
     V14   V15    V16 V17
(5)      1   1    Los_Angeles CA  4060 8863164 32.1  9.7 23677 27700 688936 70.0 22.
   3 11.6  8.0 20786 184230   4
(6)      2   2          Cook IL   946 5105067 29.2 12.4 15153 21550 436936 73.4 22.8
   11.1  7.2 21729 110928   2
(7)      3   3        Harris TX  1729 2818199 31.3  7.1  7553 12449 253526 74.9 25.4
   12.5  5.7 19517  55003   3
(8)      4   4     San_Diego CA  4205 2498016 33.5 10.9  5905  6179 173821 81.9 25.3
     8.1  6.1 19588  48931   4
(9)      5   5        Orange CA   790 2410556 32.6  9.2  6062  6369 144524 81.2 27.8
    5.2  4.8 24400  58818   4
(10)     6   6         Kings NY    71 2300664 28.3 12.4  4861  8942 680966 63.7 16.6
   19.5  9.5 16803  38658   1
(11)     7   7      Maricopa AZ  9204 2122101 29.2 12.5  4320  6104 177593 81.5 22.1
     8.8  4.9 18042  38287   4
(12)     8   8         Wayne MI   614 2111687 27.4 12.5  3823  9490 193978 70.0 13.7
   16.9 10.0 17461  36872   2
```

```
(13)    9   9          Dade FL  1945 1937094 27.1 13.9  6274  8840 244725 65.0 18.8
   14.2  8.7 17823  34525   3
(14)    10 10         Dallas TX   880 1852810 32.6  8.2  4718  6934 214258 77.1 26.3
   10.4  6.1 21001  38911   3
(15)    11 11    Philadelphia PA   135 1585577 29.1 15.2  6641 10494 109148 64.3 15.
   2 16.1  8.0 16721  26512   1
(16)    12 12           King WA  2126 1507319 30.1 11.1  5280  4009 124959 88.2 32.8
    5.0  4.6 23779  35843   4
(17)    13 13     Santa_Clara CA  1291 1497577 32.6  8.7  4101  3342  77009 82.0 32.6
     5.0  5.5 25193  37728   4
(18)    14 14  San_Bernardino CA 20062 1418380 30.1  8.8  2463  3349  83110 75.4 14.
   9 10.3  8.0 16399  23260   4
(19)    15 15       Cuyahoga OH   458 1412140 26.3 15.6  5620  8132  73150 74.0 20.1
    11.0  5.5 21086  29776   2
(20)    16 16      Middlesex MA   824 1398468 31.7 12.5  5158  4152  35825 84.3 35.4
     4.2  7.3 25312  35398   1
(21)    17 17      Allegheny PA   730 1336449 26.2 17.4  5281  8436  50186 79.0 22.6
     8.7  5.3 20681  27639   1
(22)    18 18        Suffolk NY   911 1321864 27.9 10.8  3021  3904  66723 82.2 23.0
     3.3  7.0 24262  32071   1
(23)    19 19         Nassau NY   287 1287348 25.7 14.2  6147  5200  43203 84.2 30.0
    2.5  5.1 31679  40782   1
(24)    20 20        Alameda CA   738 1279182 30.8 10.6  3169  3284 107338 81.4 28.8
     8.1  5.3 22148  28331   4
(25)    21 21        Broward FL  1209 1255488 25.3 20.7  2456  5543 107386 76.8 18.8
     7.1  7.4 22355  28066   3
(26)    22 22          Bexar TX  1247 1185394 29.5  9.9  3062  4086 133098 72.7 19.7
   16.2  6.7 15508  18383   3
(27)    23 23      Riverside CA  7208 1170413 27.9 13.2  1385  2435  95494 74.1 14.6
     8.4 10.7 17185  20114   4
(28)    24 24        Tarrant TX   864 1170103 32.2  8.3  1677  3672 132495 79.9 24.0
     8.2  6.6 18825  22027   3
(29)    25 25        Oakland MI   873 1083592 27.6 10.9  4020  3254  50964 84.6 30.2
     4.4  7.3 26884  29131   2
(30)    26 26     Sacramento CA   966 1041219 29.7 10.6  2464  2855  84305 82.2 23.0
     9.8  6.3 18934  19714   4
(31)    27 27       Hennepin MN   557 1032431 31.6 11.3  3706  5395  71753 88.2 31.6
     6.4  4.3 23705  24474   2
```

```
(32)    28 28      St._Louis MO   508  993529 26.1 13.1  1194  1056  42595 82.3 29.2
     4.0  5.1 24219  24062   2
(33)    29 29           Erie NY  1045  968532 27.3 15.2  2748  4632  55306 76.4 20.0
    9.4  6.8 18305  17729   1
(34)    30 30       Franklin OH   540  961437 33.5  9.6  2675  4011  82680 81.0 26.6
    9.1  4.2 19040  18306   2
(35)    31 31      Milwaukee WI   242  959275 29.3 13.6  2774  4141  73681 76.3 19.3
    12.6  4.9 18431  17680   2
(36)    32 32    Westchester NY   433  874866 26.3 14.4  4577  3540  37118 81.0 35.3
     4.7  5.4 33330  29159   1
(37)    33 33       Hamilton OH   407  866228 28.0 13.3  3164  4683  57208 75.6 23.7
    10.3  4.5 20580  17827   2
(38)    34 34     Palm_Beach FL  1974  863518 23.3 24.4  1833  3164  76142 78.8 22.1
     6.2  8.4 26798  23141   3
(39)    35 35       Hartford CT   736  851783 28.3 14.1  2851  2940  51926 77.7 25.8
     6.0  6.9 24875  21188   1
(40)    36 36       Pinellas FL   280  851659 22.4 26.0  1620  4458  62344 78.1 18.5
     6.2  6.2 21610  18404   3
(41)    37 37       Honolulu HI   600  836231 30.6 11.0  2025  2174  51032 81.2 24.6
     5.4  2.3 21307  17818   4
(42)    38 38   Hillsborough FL  1051  834054 29.4 12.2  2012  3068  89895 75.6 20.
   2  9.5  6.0 16876  14075   3
(43)    39 39      Fairfield CT   626  827645 26.7 13.3  2417  2494  44374 81.0 34.2
     4.5  5.9 32342  26768   1
(44)    40 40         Shelby TN   755  826330 29.4 10.4  2489  4918  67032 75.1 20.8
   14.7  5.4 18430  15229   3
(45)    41 41         Bergen NJ   234  825380 25.4 15.3  3226  2279  28521 81.6 31.7
    2.7  5.2 32230  26602   1
(46)    42 42  Fairfax_County VA   396  818584 29.2  6.5  1694   135  30202 91.4 49.0
     2.2  3.2 28999  23738   3
(47)    43 43      New_Haven CT   606  804219 28.7 14.7  3161  2486  52903 77.5 24.2
     6.0  7.3 22197  17851   1
(48)    44 44   Contra_Costa CA   720  803732 26.5 10.9  1761  1781  51243 86.5 31.6
     5.5  5.6 25523  20514   4
(49)    45 45         Marion IN   396  797159 30.6 11.7  2936  4654  61004 76.8 21.4
    9.3  5.0 19148  15264   2
(50)    46 46         DuPage IL   334  781666 29.0  8.7  2157  1842  29708 88.6 36.0
    1.7  4.8 26772  20927   2
```

```
47 47          Essex NJ   126  778206 28.6 12.7  2811  4841  75595 70.1 24.0 11.3  7.9 24523  19084   1
48 48      Montgomery MD   495  757027 28.6 10.2  4635  1507  34754 90.6 49.9  2.7  3.3 30081  22772   3
49 49          Clark NV  7911  741459 29.0 10.5   969  2011  52786 77.3 13.8  7.5  5.8 18625  13810   4
50 50  Baltimore_City MD    81  736014 30.0 13.7  5444  6203  87355 60.7 15.5 17.8  9.4 17263  12706   3
51 51 Prince_George's MD   486  729268 33.7  6.9  1253  1322  54469 83.2 25.5  3.7  5.0 19568  14270   3
52 52       Salt_Lake UT   737  725956 27.8  8.5  2094  2076  58610 85.3 23.8  7.7  4.5 15399  11179   4
53 53   San_Francisco CA    47  723959 32.2 14.5  4761  3640  71234 78.0 35.0  9.7  5.6 28532  20656   4
54 54         Macomb MI   480  717400 28.2 12.3   705  1202  41048 76.9 13.5  4.0  9.4 20924  15011   2
55 55         Monroe NY   659  713968 29.0 12.5  2438  3077  43780 80.1 26.3  7.7  4.4 21641  15451   1
56 56       Worcester MA  1513  709705 29.2 13.7  1902  2205   7099 77.4 22.2  6.3 10.2 19895  14120   1
57 57       Baltimore MD   599  692134 27.8 14.0  1269   641  46789 78.4 25.0  3.8  5.7 23470  16244   3
58 58      Montgomery PA   483  678111 26.1 15.0  3237  2425  20335 83.8 32.1  2.2  5.0 28462  19300   1
 [ reached 'max' / getOption("max.print") -- omitted 382 rows ]
> a <- Task.data[,c(5,8,16,17)]
> head(a)
       V5    V8    V16 V17
1 8863164 23677 184230   4
2 5105067 15153 110928   2
3 2818199  7553  55003   3
4 2498016  5905  48931   4
5 2410556  6062  58818   4
6 2300664  4861  38658   1
> colnames(a) <- c("totalpopulation", "activePhysicians", "income","geography")
> head(a)
  totalpopulation activePhysicians income geography
```

```
1         8863164          23677 184230         4
2         5105067          15153 110928         2
3         2818199           7553  55003         3
4         2498016           5905  48931         4
5         2410556           6062  58818         4
6         2300664           4861  38658         1
> #a[1,4]
> #a[2,4]
> a$x3 <- a$geography
> a$x4 <- a$geography
> a$x5 <- a$geography
> head(a)
  totalpopulation activePhysicians income geography x3 x4 x5
1         8863164          23677 184230         4  4  4  4
2         5105067          15153 110928         2  2  2  2
3         2818199           7553  55003         3  3  3  3
4         2498016           5905  48931         4  4  4  4
5         2410556           6062  58818         4  4  4  4
6         2300664           4861  38658         1  1  1  1
> for(i in 1:nrow(a)){
+   if(a[i,4] ==1){
+     a[i,5] <- 1
+     a[i,6] <- 0
+     a[i,7] <- 0
+   }
+ }
> for(i in 1:nrow(a)){
+   if(a[i,4] ==2){
+     a[i,5] <- 0
+     a[i,6] <- 1
+     a[i,7] <- 0
+   }
+ }
> for(i in 1:nrow(a)){
+   if(a[i,4] ==3){
+     a[i,5] <- 0
+     a[i,6] <- 0
+     a[i,7] <- 1
```

Use factor

```
+   }
+ }
> for(i in 1:nrow(a)){
+   if(a[i,4] !=3 && a[i,4] !=2 && a[i,4] !=1){
+     a[i,5] <- 0
+     a[i,6] <- 0
+     a[i,7] <- 0
+   }
+ }
> a <- a[,c(1,2,3,5,6,7)]
> head(a)
  totalpopulation activePhysicians income x3 x4 x5
1         8863164            23677 184230  0  0  0
2         5105067            15153 110928  0  1  0
3         2818199             7553  55003  0  0  1
4         2498016             5905  48931  0  0  0
5         2410556             6062  58818  0  0  0
6         2300664             4861  38658  1  0  0
> a$x3 <- as.factor(a$x3)
> a$x4 <- as.factor(a$x4)
> a$x5 <- as.factor(a$x5)
> a.lm <- lm(activePhysicians ~ totalpopulation + income + x3+x4+x5, data = a)
> a.lm

Call:
lm(formula = activePhysicians ~ totalpopulation + income + x3 +
    x4 + x5, data = a)

Coefficients:
    (Intercept)  totalpopulation           income              x31               x41              x51
     -2.075e+02        5.515e-04        1.070e-01        1.490e+02        1.455e+02        1.912e+02
```

+8 1a 答案

lm(formula = activePhysicians ~ totalpopulation + income + x3 +

x4 + x5, data = a)

Coefficients:

(Intercept)	totalpopulation	income	x31	x41	x5
1					
-2.075e+02	5.515e-04	1.070e-01	1.490e+02	1.455e+02	1.912e+02

~~(1)b~~

1.490e+02

(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?

(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.

+8

3.(a)

```
> Task.data <- read.csv("data/Performance_Ability_Data.csv", header = F)
> head(Task.data)
  V1  V2
1  0 474
2  0 432
3  0 453
4  1 481
5  1 619
6  0 584
> x<-Task.data$V2
> y<-Task.data$V1
> glm.model <- glm(y~x, family = "binomial")
> summary(glm.model)

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7845  -0.8350   0.5065   0.8371   1.7145
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.308925   4.376997  -2.355   0.0185 *
x             0.018920   0.007877   2.402   0.0163 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 37.393  on 26  degrees of freedom
Residual deviance: 29.242  on 25  degrees of freedom
AIC: 33.242

Number of Fisher Scoring iterations: 4

> glm.model$coefficients[1]
(Intercept)
  -10.30893
> glm.model$coefficients[2]
         x
0.01891983
```

答案:

對 beta0 的預測是 -10.30893

對 beta1 的預測是 0.01891983

Ln(y) = -10.30893+ 0.01891983*x

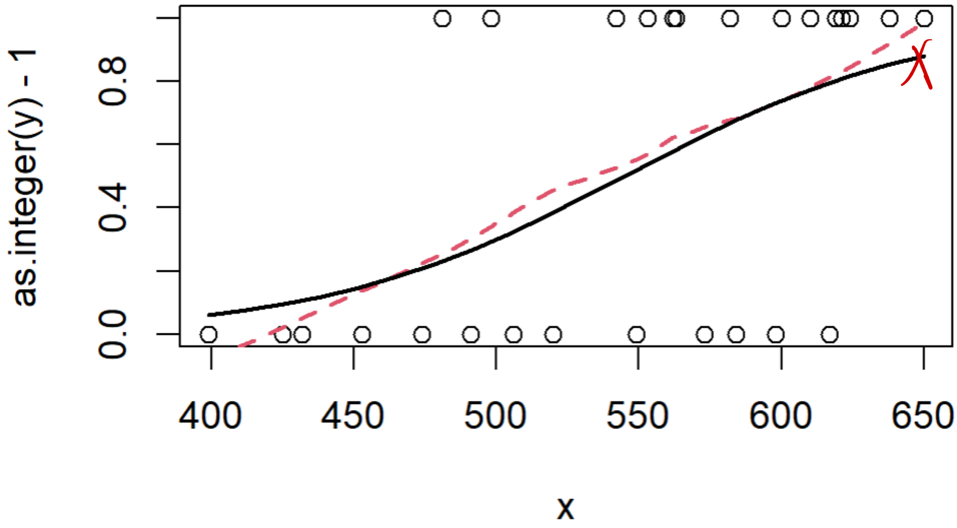
3(b)

```
> Task.data <- read.csv("data/Performance_Ability_Data.csv", header = F)
> head(Task.data)
  V1  V2
1  0 474
2  0 432
3  0 453
4  1 481
5  1 619
```

```
6  0 584
> x<-Task.data$V2
> y<-factor(Task.data$V1)
> glm.model <- glm(y~x, family = "binomial")
> summary(glm.model)

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7845  -0.8350   0.5065   0.8371   1.7145

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.308925   4.376997  -2.355   0.0185 *
x             0.018920   0.007877   2.402   0.0163 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 37.393  on 26  degrees of freedom
Residual deviance: 29.242  on 25  degrees of freedom
AIC: 33.242

Number of Fisher Scoring iterations: 4

> glm.model$coefficients[1]
(Intercept)
  -10.30893
> glm.model$coefficients[2]
        x
0.01891983
> plot(x, as.integer(y)-1)
> lines(lowess(x, as.integer(y)-1),
+      col = 2, lwd = 2, lty = 2) #color, line width
> a <- min(x)
```

```
> b <- max(x)
> new.q <- data.frame(x = seq(a, b, len = 20)) #分等
> new.q
         x
1  399.0000
2  412.2105
3  425.4211
4  438.6316
5  451.8421
6  465.0526
7  478.2632
8  491.4737
9  504.6842
10 517.8947
11 531.1053
12 544.3158
13 557.5263
14 570.7368
15 583.9474
16 597.1579
17 610.3684
18 623.5789
19 636.7895
20 650.0000
> dim(new.q)
[1] 20  1
> predicted.Y <- predict(glm.model, new.q, type="response") #response:預測是機率
> new.X <- c(new.X)
> lines(new.X, predicted.Y, lwd = 2, col = "black")
```

+5 3(B)答案 it appears to fit well



+8 3©

```
> exp(coef(glm.model))[2]
     x
1.0191
```

答案: 1.0191

+8 3(d)

```
> b = data.frame(x=550)
> a <- predict(glm.model, b, type="response")
> a
        1
0.5242263
```

預測答案: 0.5242263