國立政治大學111學年度第二學期 期中R程式考題

Department:_____應數三_____ ID:109701013____ Name:____劉恆維___

Subject: **Regression Analysis (I)**

Date: 2023/04/20

Time: 11:00~12:00 (60 minutes)

0

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　　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: 「**20230420-MidtermExam**」
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 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: **rege111**，password: classroom number.

(1) 20% Data file: Grade_Point_Average.csv

Grade point average. The director of admissions of a small college selected 120 students at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (Y) can be predicted from the ACT test score (X). The results of the study follow. Assume that first-order regression model (1.1) is appropriate.

i:	1	2	3	...	118	119	120
X_i:	21	14	28	...	28	16	28
Y_i:	3.897	3.885	3.778	...	3.914	1.860	2.948

a. Obtain the least squares estimates of β_0 and β_1, and state the estimated regression function.
b. Plot the estimated regression function and the data. Does the estimated regression function appear to fit the data well?
c. Obtain a point estimate of the mean freshman GPA for students with ACT test score $X = 30$.

a.
```
> Grade_Point_Average <- read.csv("data/Grade_Point_Average.csv ")
> Grade_Point_Average
   X3.897 X21
1   3.885  14
2   3.778  28
3   2.540  22
4   3.028  21
5   3.865  31
6   2.962  32
7   3.961  27
8   0.500  29
9   3.178  26
10  3.310  24
11  3.538  30
12  3.083  24
13  3.013  24
14  3.245  33
15  2.963  27
16  3.522  25
17  3.013  31
18  2.947  25
19  2.118  20
20  2.563  24
21  3.357  21
22  3.731  28
```

23	3.925	27
24	3.556	28
25	3.101	26
26	2.420	28
27	2.579	22
28	3.871	26
29	3.060	21
30	3.927	25
31	2.375	16
32	2.929	28
33	3.375	26
34	2.857	22
35	3.072	24
36	3.381	21
37	3.290	30
38	3.549	27
39	3.646	26
40	2.978	26
41	2.654	30
42	2.540	24
43	2.250	26
44	2.069	29
45	2.617	24
46	2.183	31
47	2.000	15
48	2.952	19
49	3.806	18
50	2.871	27
51	3.352	16
52	3.305	27
53	2.952	26
54	3.547	24
55	3.691	30
56	3.160	21
57	2.194	20
58	3.323	30
59	3.936	29
60	2.922	25
61	2.716	23
62	3.370	25
63	3.606	23
64	2.642	30
65	2.452	21
66	2.655	24
67	3.714	32
68	1.806	18
69	3.516	23
70	3.039	20
71	2.966	23
72	2.482	18
73	2.700	18
74	3.920	29
75	2.834	20
76	3.222	23
77	3.084	26
78	4.000	28
79	3.511	34
80	3.323	20

```
81   3.072  20
82   2.079  26
83   3.875  32
84   3.208  25
85   2.920  27
86   3.345  27
87   3.956  29
88   3.808  19
89   2.506  21
90   3.886  24
91   2.183  27
92   3.429  25
93   3.024  18
94   3.750  29
95   3.833  24
96   3.113  27
97   2.875  21
98   2.747  19
99   2.311  18
100  1.841  25
101  1.583  18
102  2.879  20
103  3.591  32
104  2.914  24
105  3.716  35
106  2.800  25
107  3.621  28
108  3.792  28
109  2.867  25
110  3.419  22
111  3.600  30
112  2.394  20
113  2.286  20
114  1.486  31
115  3.885  20
116  3.800  29
117  3.914  28
118  1.860  16
119  2.948  28
> colnames(Grade_Point_Average)<- c('GPA','ACT')
> Grade_Point_Average
     GPA ACT
1   3.885  14
2   3.778  28
3   2.540  22
4   3.028  21
5   3.865  31
6   2.962  32
7   3.961  27
8   0.500  29
9   3.178  26
10  3.310  24
11  3.538  30
12  3.083  24
13  3.013  24
14  3.245  33
15  2.963  27
16  3.522  25
```

17	3.013	31
18	2.947	25
19	2.118	20
20	2.563	24
21	3.357	21
22	3.731	28
23	3.925	27
24	3.556	28
25	3.101	26
26	2.420	28
27	2.579	22
28	3.871	26
29	3.060	21
30	3.927	25
31	2.375	16
32	2.929	28
33	3.375	26
34	2.857	22
35	3.072	24
36	3.381	21
37	3.290	30
38	3.549	27
39	3.646	26
40	2.978	26
41	2.654	30
42	2.540	24
43	2.250	26
44	2.069	29
45	2.617	24
46	2.183	31
47	2.000	15
48	2.952	19
49	3.806	18
50	2.871	27
51	3.352	16
52	3.305	27
53	2.952	26
54	3.547	24
55	3.691	30
56	3.160	21
57	2.194	20
58	3.323	30
59	3.936	29
60	2.922	25
61	2.716	23
62	3.370	25
63	3.606	23
64	2.642	30
65	2.452	21
66	2.655	24
67	3.714	32
68	1.806	18
69	3.516	23
70	3.039	20
71	2.966	23
72	2.482	18
73	2.700	18
74	3.920	29

```
75  2.834  20
76  3.222  23
77  3.084  26
78  4.000  28
79  3.511  34
80  3.323  20
81  3.072  20
82  2.079  26
83  3.875  32
84  3.208  25
85  2.920  27
86  3.345  27
87  3.956  29
88  3.808  19
89  2.506  21
90  3.886  24
91  2.183  27
92  3.429  25
93  3.024  18
94  3.750  29
95  3.833  24
96  3.113  27
97  2.875  21
98  2.747  19
99  2.311  18
100 1.841  25
101 1.583  18
102 2.879  20
103 3.591  32
104 2.914  24
105 3.716  35
106 2.800  25
107 3.621  28
108 3.792  28
109 2.867  25
110 3.419  22
111 3.600  30
112 2.394  20
113 2.286  20
114 1.486  31
115 3.885  20
116 3.800  29
117 3.914  28
118 1.860  16
119 2.948  28
> Grade_Point_Average_lm <- lm(GPA ~ ACT, data = Grade_Point_Average)
> summary(Grade_Point_Average_lm)


Call:
lm(formula = GPA ~ ACT, data = Grade_Point_Average)


Residuals:
     Min       1Q   Median       3Q      Max
-2.73842 -0.32556  0.04421  0.44644  1.25203


Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.06789    0.32025   6.457 2.53e-09 ***
ACT          0.04036    0.01273   3.170  0.00194 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.6193 on 117 degrees of freedom

Multiple R-squared:  0.07911,    Adjusted R-squared:  0.07124
F-statistic: 10.05 on 1 and 117 DF,  p-value: 0.001944


> Grade_Point_Average_lm_s <- summary(Grade_Point_Average_lm)
> str(Grade_Point_Average_lm_s)
List of 11
$ call        : language lm(formula = GPA ~ ACT, data = Grade_Point_Average)
$ terms        :Classes 'terms', 'formula'  language GPA ~ ACT
 .. ..- attr(*, "variables")= language list(GPA, ACT)
 .. ..- attr(*, "factors")= int [1:2, 1] 0 1
 .. .. ..- attr(*, "dimnames")=List of 2
 .. .. .. ..$ : chr [1:2] "GPA" "ACT"
 .. .. .. ..$ : chr "ACT"
 .. ..- attr(*, "term.labels")= chr "ACT"
 .. ..- attr(*, "order")= int 1
 .. ..- attr(*, "intercept")= int 1
 .. ..- attr(*, "response")= int 1
 .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
 .. ..- attr(*, "predvars")= language list(GPA, ACT)
 .. ..- attr(*, "dataClasses")= Named chr [1:2] "numeric" "numeric"
 .. .. ..- attr(*, "names")= chr [1:2] "GPA" "ACT"
$ residuals    : Named num [1:119] 1.252 0.58 -0.416 0.112 0.546 ...
 ..- attr(*, "names")= chr [1:119] "1" "2" "3" "4" ...
$ coefficients : num [1:2, 1:4] 2.0679 0.0404 0.3203 0.0127 6.4571 ...
 ..- attr(*, "dimnames")=List of 2
 .. ..$ : chr [1:2] "(Intercept)" "ACT"
 .. ..$ : chr [1:4] "Estimate" "Std. Error" "t value" "Pr(>|t|)"
$ aliased      : Named logi [1:2] FALSE FALSE
 ..- attr(*, "names")= chr [1:2] "(Intercept)" "ACT"
$ sigma        : num 0.619
$ df           : int [1:3] 2 117 2
$ r.squared    : num 0.0791
$ adj.r.squared: num 0.0712
$ fstatistic   : Named num [1:3] 10.1 1 117
 ..- attr(*, "names")= chr [1:3] "value" "numdf" "dendf"
$ cov.unscaled : num [1:2, 1:2] 0.267445 -0.010464 -0.010464 0.000423
 ..- attr(*, "dimnames")=List of 2
 .. ..$ : chr [1:2] "(Intercept)" "ACT"
 .. ..$ : chr [1:2] "(Intercept)" "ACT"
- attr(*, "class")= chr "summary.lm"
> Grade_Point_Average_lm_s$coefficients
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 2.06788792 0.32025225 6.457060 2.532802e-09
ACT         0.04036332 0.01273134 3.170392 1.943625e-03
> str(Grade_Point_Average_lm_s$coefficients)
num [1:2, 1:4] 2.0679 0.0404 0.3203 0.0127 6.4571 ...
- attr(*, "dimnames")=List of 2
 ..$ : chr [1:2] "(Intercept)" "ACT"
 ..$ : chr [1:4] "Estimate" "Std. Error" "t value" "Pr(>|t|)"
```

```
> beta0_grade <- Grade_Point_Average_lm_s$coefficients[1,1]
> beta0_grade
[1] 2.067888
> beta1_grade <- Grade_Point_Average_lm_s$coefficients[2,1]
> beta1_grade
[1] 0.04036332
```

= 2.067888

0.04036332

E(Y)= 2.067888 + 0.04036332X

b.

(2) Data file: Grade_Point_Average.csv

20%

Refer to **Grade point average**

a. Set up the ANOVA table.

b. Conduct an F test of whether or not $\beta_1 = 0$. Control the α risk at .01. State the alternatives, decision rule, and conclusion.

c. What is the absolute magnitude of the reduction in the variation of Y when X is introduced into the regression model?

(3) Data file: Grade_Point_Average_X.csv

30% Refer to **Grade point average**

a. Prepare a box plot for the ACT scores X_i. Are there any noteworthy features in this plot?

b. Prepare a dot plot of the residuals. What information does this plot provide?

c. Plot the residual e_i against the fitted values $\hat{Y}_i$. What departures from regression model (2.1) can be studied from this plot? What are your findings?

d. Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Test the reasonableness of the normality assumption here using Table B.6 and $\alpha = .05$. What do you conclude?

e. Conduct the Brown-Forsythe test to determine whether or not the error variance varies with the level of X. Divide the data into the two groups, $X < 26$, $X \geq 26$, and use $\alpha = .01$. State the decision rule and conclusion. Does your conclusion support your preliminary findings in part (c)?

f. Information is given below for each student on two variables not included in the model, namely, intelligence test score (X_2) and high school class rank percentile (X_3). (Note that larger class rank percentiles indicate higher standing in the class, e.g., 1% is near the bottom of the class and 99% is near the top of the class.) Plot the residuals against X_2 and X_3 on separate graphs to ascertain whether the model can be improved by including either of these variables. What do you conclude?

i:	1	2	3	...	118	119	120
X_2:	122	132	119	...	140	111	110
X_3:	99	71	75	...	97	65	85

(4) Data file: Solution_concentration.csv

30% **Solution concentration.** A chemist studied the concentration of a solution (Y) over time (X). Fifteen identical solutions were prepared. The 15 solutions were randomly divided into five sets of three, and the five sets were measured, respectively, after 1, 3, 5, 7, and 9 hours. The results follow.

i:	1	2	3	...	13	14	15
X_i:	9	9	9	...	1	1	1
Y_i:	.07	.09	.08	...	2.84	2.57	3.10

a. Prepare a scatter plot of the data. What transformation of Y might you try, using the prototype patterns in Figure 3.15 to achieve constant variance and linearity?

b. Use the Box-Cox procedure and standardization (3.36) to find an appropriate power transformation. Evaluate SSE for $\lambda = -.2, -.1, 0, .1, .2$. What transformation of Y is suggested?

c. Use the transformation $Y' = \log_{10} Y$ and obtain the estimated linear regression function for the transformed data.

d. Plot the estimated regression line and the transformed data. Does the regression line appear to be a good fit to the transformed data?

e. Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?

f. Express the estimated regression function in the original units.