國立政治大學111學年度第二學期 期中R程式考題

Department:_____ ID:_____ Name:_____

Subject: **Regression Analysis (I)**

Date: 2023/04/20

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

25

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.

```
> data<-read.csv(file.choose())
> data
    GPA ACT
1  3.897  21
2  3.885  14
3  3.778  28
4  2.540  22
5  3.028  21
6  3.865  31
7  2.962  32
8  3.961  27
9  0.500  29
10 3.178  26
```

11	3.310	24
12	3.538	30
13	3.083	24
14	3.013	24
15	3.245	33
16	2.963	27
17	3.522	25
18	3.013	31
19	2.947	25
20	2.118	20
21	2.563	24
22	3.357	21
23	3.731	28
24	3.925	27
25	3.556	28
26	3.101	26
27	2.420	28
28	2.579	22
29	3.871	26
30	3.060	21
31	3.927	25
32	2.375	16
33	2.929	28
34	3.375	26
35	2.857	22
36	3.072	24
37	3.381	21
38	3.290	30
39	3.549	27
40	3.646	26
41	2.978	26
42	2.654	30
43	2.540	24

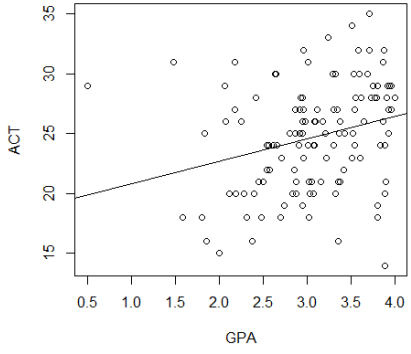
44	2.250	26
45	2.069	29
46	2.617	24
47	2.183	31
48	2.000	15
49	2.952	19
50	3.806	18
51	2.871	27
52	3.352	16
53	3.305	27
54	2.952	26
55	3.547	24
56	3.691	30
57	3.160	21
58	2.194	20
59	3.323	30
60	3.936	29
61	2.922	25
62	2.716	23
63	3.370	25
64	3.606	23
65	2.642	30
66	2.452	21
67	2.655	24
68	3.714	32
69	1.806	18
70	3.516	23
71	3.039	20
72	2.966	23
73	2.482	18
74	2.700	18
75	3.920	29
76	2.834	20

77	3.222	23
78	3.084	26
79	4.000	28
80	3.511	34
81	3.323	20
82	3.072	20
83	2.079	26
84	3.875	32
85	3.208	25
86	2.920	27
87	3.345	27
88	3.956	29
89	3.808	19
90	2.506	21
91	3.886	24
92	2.183	27
93	3.429	25
94	3.024	18
95	3.750	29
96	3.833	24
97	3.113	27
98	2.875	21
99	2.747	19
100	2.311	18
101	1.841	25
102	1.583	18
103	2.879	20
104	3.591	32
105	2.914	24
106	3.716	35
107	2.800	25
108	3.621	28
109	3.792	28

```
110 2.867  25
111 3.419  22
112 3.600  30
113 2.394  20
114 2.286  20
115 1.486  31
116 3.885  20
117 3.800  29
118 3.914  28
119 1.860  16
120 2.948  28
> n<-nrow(data)
> lsfit<-lsfit(data$GPA,data$ACT)
> lsfit$coefficients
Intercept         X
18.975441  1.870353
> b0<-lsfit$coefficients[1,1]
> b0<-lsfit$coefficients[1]
> b0
Intercept
18.97544
> b1<-lsfit$coefficients[2]
> b1
       X
1.870353
```

b.

```
> plot(data)
> abline(lsfit$coefficients)
```



no

c.

> x<-30
> y<-b0+b1*30
> y
Intercept
75.08604

(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?

a.

```
> lm<-lm(GPA~ACT,data=data)
> lm

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

Coefficients:
(Intercept)         ACT
   2.11405      0.03883

> anova(lm)
Analysis of Variance Table

Response: GPA
          Df Sum Sq Mean Sq F value   Pr(>F)
ACT        1  3.588  3.5878  9.2402 0.002917 **
Residuals 118 45.818  0.3883
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

b.

```
> alpha<-0.01
```

```
> #H0:beta1=0,Ha:beta1!=0
> ifelse(anova(lm)$'Pr(>F)'[1]<0.01,"Reject H0","fail to reject H0")
[1] "Reject H0"
```

c.

```
> summary(lm)$r.squared
[1] 0.07262044
```

(3) 30%

Data file: Grade_Point_Average_X.csv

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?

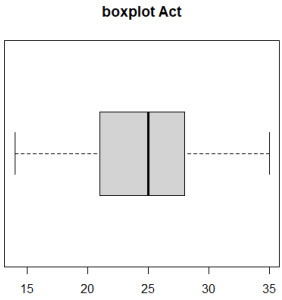
c. Plot the residual e_i against the fitted values $\hat{Y}_i$. What departures from regression model (2.1) can be studied from this plot? What are your findings?

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




```
> Grade_Point_Average_X<-read.csv(choose.files())
> Grade_Point_Average_X
    GPA ACT Intelligence RankPercentile
1   3.897  21         122            99
2   3.885  14         132            71
3   3.778  28         119            95
4   2.540  22          99            75
5   3.028  21         131            46
6   3.865  31         139            77
7   2.962  32         113            85
8   3.961  27         136            99
9   0.500  29          75            13
10  3.178  26         106            97
11  3.310  24         125            69
12  3.538  30         142            99
13  3.083  24         120            97
14  3.013  24         107            55
15  3.245  33         125            93
16  2.963  27         121            80
17  3.522  25         119            63
18  3.013  31         128            78
19  2.947  25         106            93
20  2.118  20         123            22
21  2.563  24         111            84
```

22	3.357	21	113	87
23	3.731	28	134	98
24	3.925	27	128	95
25	3.556	28	126	63
26	3.101	26	121	79
27	2.420	28	104	86
28	2.579	22	113	90
29	3.871	26	133	97
30	3.060	21	125	39
31	3.927	25	128	97
32	2.375	16	112	57
33	2.929	28	107	67
34	3.375	26	115	81
35	2.857	22	119	75
36	3.072	24	113	63
37	3.381	21	115	15
38	3.290	30	110	95
39	3.549	27	122	93
40	3.646	26	118	99
41	2.978	26	114	90
42	2.654	30	112	99
43	2.540	24	106	85
44	2.250	26	95	84
45	2.069	29	102	58
46	2.617	24	114	86
47	2.183	31	116	82
48	2.000	15	93	34
49	2.952	19	120	34
50	3.806	18	117	23
51	2.871	27	119	95
52	3.352	16	115	41
53	3.305	27	113	28
54	2.952	26	108	68

55	3.547	24	116	54
56	3.691	30	135	77
57	3.160	21	108	58
58	2.194	20	110	73
59	3.323	30	124	94
60	3.936	29	130	98
61	2.922	25	118	99
62	2.716	23	110	91
63	3.370	25	117	95
64	3.606	23	123	72
65	2.642	30	116	65
66	2.452	21	109	53
67	2.655	24	110	81
68	3.714	32	126	41
69	1.806	18	99	84
70	3.516	23	121	84
71	3.039	20	115	35
72	2.966	23	127	70
73	2.482	18	99	15
74	2.700	18	108	47
75	3.920	29	129	98
76	2.834	20	103	77
77	3.222	23	122	72
78	3.084	26	118	29
79	4.000	28	135	80
80	3.511	34	139	88
81	3.323	20	128	80
82	3.072	20	120	46
83	2.079	26	114	89
84	3.875	32	133	91
85	3.208	25	123	95
86	2.920	27	111	83
87	3.345	27	122	92

88	3.956	29	136	99
89	3.808	19	140	41
90	2.506	21	109	68
91	3.886	24	133	98
92	2.183	27	98	59
93	3.429	25	134	89
94	3.024	18	124	89
95	3.750	29	128	92
96	3.833	24	149	97
97	3.113	27	121	43
98	2.875	21	117	52
99	2.747	19	110	82
100	2.311	18	104	61
101	1.841	25	95	72
102	1.583	18	96	33
103	2.879	20	117	97
104	3.591	32	130	97
105	2.914	24	121	92
106	3.716	35	125	99
107	2.800	25	112	61
108	3.621	28	136	72
109	3.792	28	129	99
110	2.867	25	106	76
111	3.419	22	108	66
112	3.600	30	138	70
113	2.394	20	106	44
114	2.286	20	111	33
115	1.486	31	101	77
116	3.885	20	113	57
117	3.800	29	131	96
118	3.914	28	140	97
119	1.860	16	111	65
120	2.948	28	110	85

> boxplot(Grade_Point_Average_X$ACT,main="boxplot Act",horizontal = T)

b.

output ?

(4) 30% Data file: Solution_concentration.csv

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.