111 年特種考試地方政府公務人員考試試題

代號：31670
頁次：5－1

等　　別：三等考試
類　　科：統計
科　　目：迴歸分析
考試時間：2 小時　　座號：＿＿＿＿＿＿

※注意：㈠可以使用電子計算器。
㈡不必抄題，作答時請將試題題號及答案依照順序寫在試卷上，於本試題上作答者，不予計分。
㈢本科目除專門名詞或數理公式外，應使用本國文字作答。

一、一位統計分析師想瞭解身高（Y_i，以英寸為單位）是否可以用手掌張開長度（X_1，以公分為單位）和性別（X_2，男性是 1，女性是 0）來預測？他收集 66 名大學生為樣本。所配適的線性迴歸模型如下：

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i, \quad i = 1, \cdots, n.$$

請依據表 1 回答下列問題。

表1: ANOVA

Source	Sum of Squares	DF	Mean square	F test
Regression	840.8436	2		
Error	(1)	(3)	(5)	
（Lack of fit）	(2)	(4)		
（Pure error）	283.8476	45		
Total	1220.4394	65		

㈠請計算表 1 中(1)–(5)所列的線性迴歸的 ANOVA 相關訊息。（10 分）

㈡在顯著水準 5%下，請檢定身高是否與手掌張開長度(X_1)和性別(X_2)有線性關係存在。請列出虛無假設/對立假設、檢定統計量及決策法則。在無需查表之下，你的建議結論為何？（5 分）

㈢在顯著水準 5%下，請檢定線性迴歸模型是否有顯著的缺適（lack of fit）？以了解線性迴歸模型是否足以描述身高與手掌張開長度(X_1)和性別(X_2)之間的關係。請列出虛無假設/對立假設、檢定統計量及決策法則。在無需查表之下，你的建議結論為何？請說明缺適檢定所需要之假設。（10 分）

二、一位統計分析師分析奧林匹克男子田徑短跑 200 公尺數據，包含 1900 年至 2020 年間舉行的 28 次男子 200 公尺奧林匹克短跑比賽獲金牌的秒數，其中第一次和第二次世界大戰期間沒有舉辦奧運會，而 2020 年奧林匹克運動會因為 COVID-19 疫情實際是 2021 年在日本東京舉行。因此資料包含 year（以年為單位）和 Y（以秒為單位），其散布圖在圖 1。

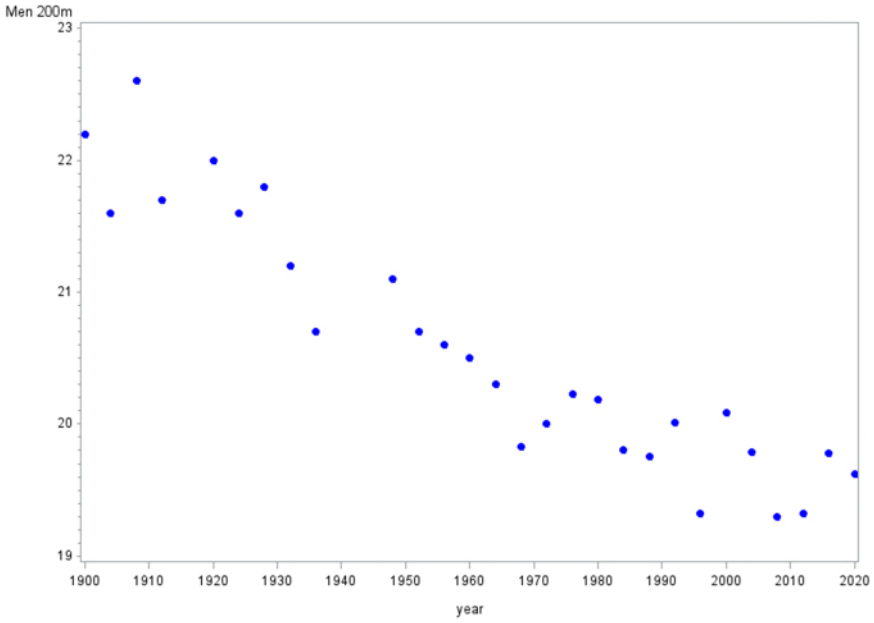

圖 1 奧林匹克年份和男子田徑短跑 200 公尺秒數散布圖

這位統計分析師重新定義變數，他把“西元年（year）”平減 1963，並定義新的解釋變數 X，也就是 X=year−1963。樣本相關資訊如下，其中 n 為樣本數，請依據這些資訊回答問題。

$$\bar{X} = -0.1429, \quad \bar{Y} = 20.5582, \quad S_{XY} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = -888.2171,$$

$$S_{XX} = \sum_{i=1}^{n}(X_i - \bar{X})^2 = 36859.4286, \quad S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = 24.3354$$

㈠請計算 (X, Y) 的皮爾森相關係數。（5 分）

㈡該統計分析師配適模型 $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$，此處 ε_i 是誤差項。請寫出以最小平方估計法所得到的估計迴歸線，並推導共變異數 $\hat{\beta}_0$ 和 $\hat{\beta}_1$，也就是 $Cov(\hat{\beta}_0, \hat{\beta}_1)$。（10 分）

㈢在顯著水準 $\alpha = 0.05$ 之下，請檢定 $H_0 : \beta_1 = 0$ 是否顯著？請詳述檢定統計量之值、決策法則和結論。請問年份和獲金牌的秒數之間是否存在線性關係？以此資料是否可以推論人類在田徑短跑越跑越快？t 分配臨界值，$t_{0.025}(26) = -2.0555, t_{0.025}(27) = -2.0518$。（10 分）

三、一位統計分析師受託分析 20 名年齡 40~60 歲高血壓患者的血壓相關數據，以評估可能影響血壓的重要因素，資料描述如下：
血壓（Y，反應變數，以 mm Hg 為單位），年齡（X_1，以年為單位），重量（X_2，公斤），體表面積（X_3，平方公尺），高血壓病史（X_4，以年為單位），基礎脈搏（X_5，以每分鐘為單位），壓力指數（X_6，0−100 為範圍）。部分統計套裝軟體輸出結果在表 2 和表 3。

表2

反應變數	5個解釋變數	判定係數 R_j^2
X_1	X_2-X_6	0.451
X_2	X_1, X_3-X_6	0.925
X_3	X_1-X_2, X_4-X_6	0.905
X_4	X_1-X_3, X_5-X_6	0.196
X_5	X_1-X_4, X_6	0.754
X_6	X_1-X_5	0.416

表3

解釋變數	Type I SS		偏判定係數	
X_1	SSR(X_1)	243.266	R^2_{Y,X_1}	0.4344
X_2	SSR(X_2|X_1)	306.886	$R^2_{Y,X_2\|X_1}$	0.96891
X_3	SSR(X_3|X_1,X_2)	0.765	$R^2_{Y,X_3\|X_1,X_2}$	0.07763
X_4	SSR(X_4|X_1,X_2,X_3)	0.250	$R^2_{Y,X_4\|X_1,X_2,X_3}$	0.02755
X_5	SSR(X_5|X_1,X_2,X_3,X_4)	0.965	$R^2_{Y,X_5\|X_1,X_2,X_3,X_4}$	0.1092
X_6	SSR(X_6|X_1,X_2,X_3,X_4,X_5)	1.023E-04	$R^2_{Y,X_6\|X_1,X_2,X_3,X_4,X_5}$	1.3E-05

㈠這位分析師一開始採用(1)式中模型 1 的複迴歸分析，他擔心有多重共線性（Multicollinearity）問題。
模型 1：
$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + \varepsilon_i, \quad i = 1,\ldots,n. \quad (1)$$
請協助這位分析師利用表 2 判斷是否有嚴重的多重共線性，並說明模型 1 是否合適？如果不合適，請詳細說明原因和判斷方法。（5 分）

㈡表 3 第二欄的定義，若 SSR（X_i|X_j）代表給定 X_j 已在模型中，X_i 加入模型中的額外平方和（extra sum of squares）。請計算 SSR（X_1,X_2,X_3,X_4,X_5,X_6）。最後一欄符號代表偏判定係數（coefficient of partial determination）。請說明偏判定係數 $R^2_{Y,X_3|X_1,X_2}$ 的計算式及其意義。請利用表 3 結果，建議分析師採用那些變數，詳細說明理由和判斷方法。（10 分）

㈢請利用表 3 結果及 SST=560，SSR（X_1,X_2,X_5）=551.568，計算 SSR（X_5|X_1,X_2）和偏判斷係數 $R^2_{Y,X_5|X_1,X_2}$。（10 分）

四、一位教師擬瞭解學生的測試表現是否受智商和教學方法所影響，以 60 名學生為實驗對象，在採用三種教學方法之下，獲得測試成績 Y，智商 X。前兩種教學方法 M_1, M_2 變數定義如下。

$$M_1=\begin{cases}1 & \text{教學法1}\\ 0 & \text{其他}\end{cases}\qquad M_2=\begin{cases}1 & \text{教學法2}\\ 0 & \text{其他}\end{cases}$$

這位教師分別考慮的模型如下：

模型 1 $Y_i=\beta_0+\beta_1X_i+\varepsilon_i,\quad i=1,\ldots,n.$

模型 2 $Y_i=\beta_0+\beta_2M_{1i}+\beta_3M_{2i}+\varepsilon_i,\quad i=1,\ldots,n.$

模型 3 $Y_i=\beta_0+\beta_1X_i+\beta_2M_{1i}+\beta_3M_{2i}+\varepsilon_i,\quad i=1,\ldots,n.$

請使用表 4 部分電腦輸出 3 個模型的變異數分析（ANOVA, Analysis of Variance）報表來回答下列問題。

㈠在考慮模型 3 之下，請檢定智商 X 該解釋變數對於解釋測試成績是否有顯著的解釋能力。請用顯著水準α=0.05檢定並詳述檢定統計量之值、決策法則、結論和所需之假設。t 分配臨界值，$t_{0.975}(56)=2.0032$。（10 分）

㈡在考慮模型 3 之下，請檢定教學方法 M_1 和 M_2 這兩個虛擬變數是否在模型 3 對預測學生測試成績有效應。請在顯著水準α=0.05，檢定 $H_0:\beta_2=\beta_3=0$，請詳述檢定統計量之值、決策法則、結論和所需之假設。F 分配左尾臨界值，$F_{0.95}(1,56)=4.0130$，$F_{0.95}(2,56)=3.1619$。（10 分）

㈢請使用表 4 說明那一種教學方法最能提升測試成績，須說明論述。（5 分）

表4

模型1 ANOVA表

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F value	P-value
Regression	1	816.928	816.928	14.72	0.0003
Error	58	3219.255	55.504		
Total	59	4036.183			

模型2 ANOVA表

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F value	P-value
Regression	2	2880.033	1440.017	71	P-value
Error	57	1156.150	20.283		
Total	59	4036.183			

模型3 ANOVA表和參數估計

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F value	P-value
Regression	3	3512.745	1170.915	125.27	<.0001
Error	56	523.438	9.347		
Total	59	4036.183			

模型3參數估計					
Variable	DF	Estimate	Standard Error	t value	P-value
Intercept	1	56.024	4.306	13.01	<.0001
X	1	0.350	0.043	8.14	<.0001
M_1	1	-15.770	0.967	-16.3	<.0001
M_2	1	-11.943	0.972	-12.28	<.0001