104年公務人員高等考試三級考試試題　　代號：22050

類　　科：統計
科　　目：迴歸分析
考試時間：2小時　　座號：_____

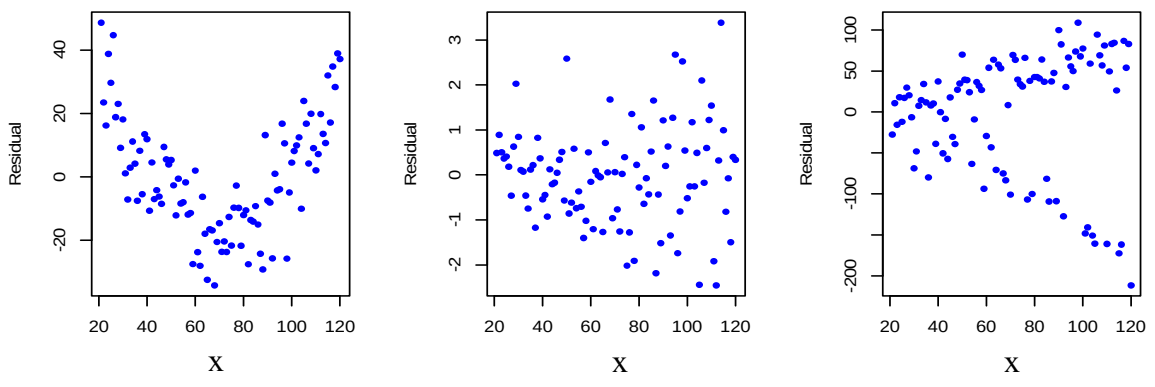
※注意：㈠可以使用電子計算器。
㈡不必抄題，作答時請將試題題號及答案依照順序寫在試卷上，於本試題上作答者，不予計分。
㈢若無特別標示，本試題採用顯著水準為 0.1 及 90%信心水準為原則。

附表：F 分佈 $\alpha=0.1$ 臨界值 $F_{df1,df2,0.1}$

df1\df2	15	16	17	18	19	20
1	3.073	3.048	3.026	3.007	2.990	2.975
2	2.695	2.668	2.645	2.624	2.606	2.589
3	2.490	2.462	2.437	2.416	2.397	2.380
4	2.361	2.333	2.308	2.286	2.266	2.249

一、若考慮配適一簡單線性迴歸模型 $y=\alpha+\beta x+\varepsilon$，其中 α、β 為參數，ε 為隨機誤差，且假設其為具均數 0，標準差 σ 之常態分配。今於配適模型後，繪出殘差對自變數 x 的分析圖。請分別針對圖(a)-(c)的結果，說明迴歸模型是否恰當？若模型不恰當時，請指出對於參數估計值是否會有偏差（bias）之影響，對於有關參數的假設檢定是否正確，另外也請提出修正的方法。（18 分）

(a)　　(b)　　(c)



二、根據下列 3 變數，6 個觀察值的資料

Y	1	0	1	1	0	0
X1	1	-2	1	0	0	0
X2	0	1	2	2	1	0

㈠令 Y、X1、X2 分表各變數觀察值所形成的向量，另定義 X0 為長度等於 6 且元素均等於 1 的向量。在以向量表示法的迴歸模型 M：$Y=\beta_0X0+\beta_1X1+\beta_2X2+\varepsilon$ 中，如何將 $\beta_0X0+\beta_1X1+\beta_2X2$ 更精簡的以矩陣與參數向量表示？另外，在一般情形下，此時 ε 之機率分佈為何？（4 分）

㈡計算迴歸模型 M 中之參數向量的最小平方估計量及估計其變異數共變異數矩陣（variance-covariance matrix）。（8 分）

㈢令 $\hat{Y}$ 為長度等於 6 的向量，其元素為迴歸模型 M 對 Y 的配適值（fitted values），則存在一矩陣 H 使得 $\hat{Y}=HY$，計算此矩陣 H。（4 分）

㈣計算迴歸模型 M 中的變異數膨脹因子（variance inflation factor, vif）vif(X1)與 vif(X2)。（4 分）

（請接第二頁）

類　　科：統計
科　　目：迴歸分析

三、三高（高血壓、高血糖、高血脂）與許多重大慢性病皆有重要關係。為了解個人體質、生活習慣等對於三高的影響因子，並對社會大眾提出建議與注意事項。因此，研究人員由臺灣數個醫學中心，採用隨機抽樣法蒐集了 10000 個就診慢性病者的資料進行調查分析。該資料測量每個人的血壓（以收縮壓為例，單位為 mmHg）及其他相關變數如下：

性別（男性為 1，女性為 0），年齡（25-85 歲），身體質量指數 BMI（定義為身高/體重^2，單位為 m/kg^2），量血壓習慣（有量血壓習慣者為 1，反之為 0），量血糖習慣（有量血糖習慣者為 1，反之為 0），量血脂習慣（有量血脂習慣者為 1，反之為 0），喝酒習慣（平均每天喝 1 瓶 600c.c.啤酒或相當之酒類以上者為 1，反之為 0），抽煙習慣（有抽煙習慣者為 1，反之為 0），外食頻率（每週外食次數），運動習慣（有運動習慣者為 1，反之為 0），睡眠品質（睡眠品質佳者為 1，反之為 0）。研究者建立血壓(y)對所有解釋變數的迴歸模型，得到如下表（LM1）之結果，其殘差分析也無明顯瑕疵。

㈠模型 LM1 之所有變數的解釋力為多少？一般來說，此解釋力算是高、中或低？並解釋表中「F-statistic：4961 on 11 and 9988 DF, p-value：<2.2e-16」之意義。（4 分）

㈡在模型 LM1 下，以兩人之不同的性別、年齡及 BMI 解釋參數估計值所代表之意義。（6 分）

㈢為了去蕪存菁，研究人員去除兩個非常不顯著的變數並得到下表模型 LM2 之結果。根據 LM1 及 LM2，請就下面 1.或 2.擇一回答（兩項均答者不予評分）。（10 分）

1. 說明 LM1 與 LM2 何者較佳或差不多，並建議大眾那些變數為三高影響因子應儘量避免或注意？
2. 此分析結果不適合用來推薦三高影響因子（說明原因及提出改進方法，此結論是否與題㈠結論矛盾？）。

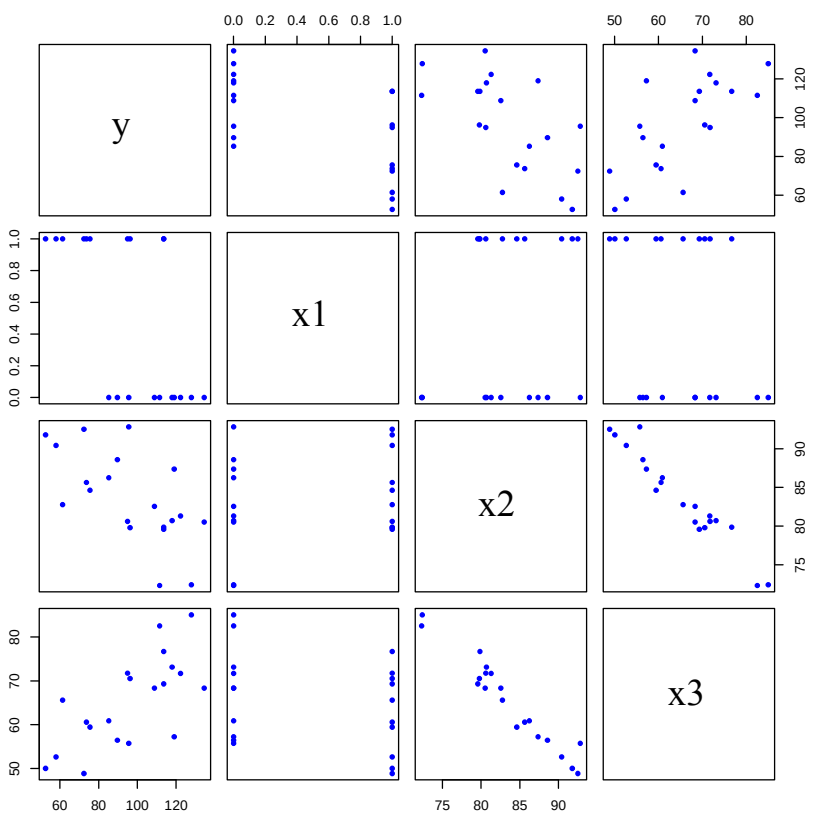
模型 LM1	Estimate	Std. Error	t value	Pr(>|t|)
（Intercept）	97.487	0.627	155.365	0.0000
性別	19.564	0.113	173.786	0.0000
年齡	0.452	0.005	86.894	0.0000
身體質量指數 BMI	1.249	0.458	2.729	0.0064
量血壓習慣	2.070	0.108	19.084	0.0000
量血糖習慣	0.557	0.100	5.545	0.0000
量血脂習慣	3.012	0.311	9.697	0.0000
喝酒習慣	-0.741	0.294	-2.522	0.0117
抽煙習慣	0.046	0.049	0.936	0.3494
外食頻率	-1.827	0.979	-1.866	0.0621
運動習慣	2.933	0.858	3.418	0.0006
睡眠品質	-0.005	0.019	-0.284	0.7764
Residual standard error：4.923 on 9988 degrees of freedom Multiple R-squared：0.8453, Adjusted R-squared：0.8451 F-statistic：4961 on 11 and 9988 DF, p-value：< 2.2e-16				

模型 LM2	Estimate	Std. Error	t value	Pr(>|t|)
（Intercept）	97.551	0.624	156.414	0.0000
性別	19.570	0.111	176.780	0.0000
年齡	0.452	0.005	86.912	0.0000
身體質量指數 BMI	1.247	0.457	2.726	0.0064
量血壓習慣	2.070	0.108	19.081	0.0000
量血糖習慣	0.556	0.100	5.532	0.0000
量血脂習慣	3.013	0.311	9.702	0.0000
喝酒習慣	-0.746	0.294	-2.539	0.0111
外食頻率	-1.836	0.979	-1.876	0.0607
運動習慣	2.934	0.858	3.420	0.0006
Residual standard error：4.923 on 9990 degrees of freedom Multiple R-squared：0.8453, Adjusted R-squared：0.8451 F-statistic：6064 on 9 and 9990 DF, p-value：< 2.2e-16				

（請接第三頁）

類　　科：統計
科　　目：迴歸分析

四、一個學習效果評量相關分析的報告裏，資料內容由 20 人（男女各半）的 4 個變數(y,x1,x2,x3)所構成。其中 y 為學習效果（其平均值 96.2 且標準差為 24.47），x1＝1 或 0 表男性及女性，x2（其平均值 83.6 且標準差為 5.9）與 x3（其平均值 65 且標準差為 10.3）分別表某性向測驗的兩種分數。下圖為資料之 4 個變數間的散佈圖；此外，下表也列出配適學習效果 y 與不同解釋變數之迴歸模型的 R^2。



Model	Variables in model	R^2
M1	x1	0.397
M2	x2	0.413
M3	x3	0.487
M4	x2, x3	0.504
M5	x1, x2	0.676
M6	x1, x3	0.697
M7	x1,x2, x3	0.697

㈠考慮模型 M1，完成下面的分析表，說明填入之 F value 及 t value 的值所代表意義。（12 分）

Analysis of Variance Table：Response：y

	Df	Sum Sq	Mean Sq	F value
x1				
Residuals				
Total				

Coefficients：

	Estimate	Std. Error	t value
Intercept			
x1			

㈡考慮模型 M1，計算 y 在 x1＝1 之信心水準為 90%的預測區間。（5 分）

㈢在 M1-M7 模式中，給定進入模式水準（entry level）α=0.1，採用 F 檢定法，列出前進選取（forward selection）程序與其最終選定之模式。（10 分）

㈣根據準則 Akaike Information Criterion（AIC），依序列出 M1-M7 模式中的最佳 3 個模型。（10 分）

㈤針對 M7 模式，在顯著水準 α=0.1 下，檢定 x2 與 x3 之係數是否同時等於0。（5 分）