109年特種考試地方政府公務人員考試試題

代號：31570
頁次：6－1

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。

參考之查表值：F 分布 α=0.05 臨界值 $F_{0.05}$(df1, df2)

	df1		
df2		1	2
	27	4.2100	3.3541
	28	4.1960	3.3404
	402	3.8647	3.0182
	403	3.8646	3.0181

$t_{0.025}(28)$=-2.0484, $t_{0.025}(30)$=-2.0422

一、一位主管欲知道碩士級分析師的月薪是否可以用年資來預測，以作為未來給薪的參考。他收集了30個樣本觀察值，資料包含年資（X，以年為單位）和月薪（Y，以千元為單位）。請依據下面數據和圖1回答問題。

$$\bar{X}=5.34,\ \bar{Y}=76,\ S_{XY}=\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})=2198,$$

$$S_{XX}=\sum_{i=1}^{n}(X_i-\bar{X})^2=232.072,\quad S_{YY}=\sum_{i=1}^{n}(Y_i-\bar{Y})^2=21890$$

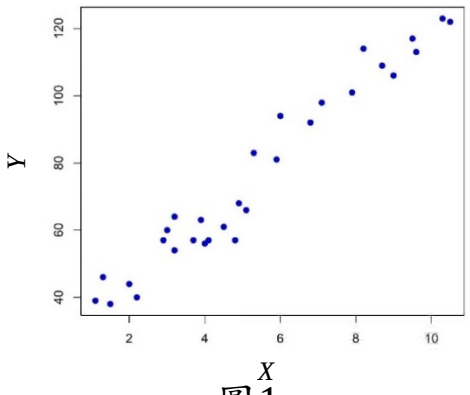

圖1

㈠在配適$Y_i=\beta_0+\beta_1X_i+\varepsilon_i$的簡單線性迴歸方程式下，請利用最小平方法計算參數$\beta_0$和$\beta_1$估計值（estimates）。如果將模型改為$Y_i=\alpha+\beta_1(X_i-\bar{X})+\varepsilon_i$，請寫出參數$\alpha$和$\beta_1$最小平方估計式（least-squares estimators）及其估計標準誤（standard errors）。（12分）

㈡假設$Y_i\sim N(\beta_0+\beta_1X_i,\sigma^2)$，請在顯著水準$\alpha=0.05$下，檢定$H_0:\beta_1=0$。請試述檢定統計量之值、決策法則和結論。請寫出在應用最大概似估計（Maximum likelihood estimation）法，σ^2的估計值。請寫出利用最小平方法，σ^2的估計值。（10分）

㈢請問年資是5年的碩士級分析師之平均薪資的95%信賴區間。（4分）

二、㈠一位分析師受託分析一組資料。資料來自於20位25歲至34歲的健康女性，其中包括反應變數 Y（身體脂肪）和三個解釋變數（X_1：皮褶厚度，X_2：大腿圓周和 X_3：中臂圓周）用作預測身體脂肪。該分析師初步配適一個迴歸模型如下：

模型1 $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i \quad i = 1, \cdots, 20.$

另外，表1計算解釋變數之間的解釋能力。

表1

反應變數	解釋變數	判定係數 R^2
X_1	X_2, X_3	99.86%
X_2	X_1, X_3	99.82%
X_3	X_1, X_2	99.04%

請由表1計算變異數膨脹因子（variance inflation factor, VIF）評論該分析師所配適的迴歸模型1是否合適？如果不合適，請詳細說明原因和解決方法。（8分）

㈡一位分析師受託分析影響縣市首長滿意度的重要因素。滿意度分數Y（以1～10為評分範圍，分數愈高代表愈滿意）作為反應變數。該分析師找到一些重要的解釋變數。依據他所配適的複迴歸模型，有些預測值有超過10的情況。請問該分析師所配適的複迴歸模型是否合適？如果不合適，請詳細說明原因和解決的方法。（6分）

㈢一位分析師分析2017年1月至2019年12月的旅遊人數月資料。該分析師配適的迴歸模型如下：

模型2

$$\ln(y_t) = \beta_0 + \beta_1 t + \beta_2 M_1 + \beta_3 M_2 + \cdots + \beta_{12} M_{11} + \varepsilon_t,\ \varepsilon_t \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$$

此處 t 是時間，ε_t 為獨立且具有共同分配其平均數為0變異數σ^2的常態分配，M_i是虛擬變數，第 i 個月為1，其他月份為0，i=1, 2,…, 11。請說明在線性迴歸模型下，如何檢查誤差項的所有假設是否有違反。圖2是模型2的標準化殘差值（studentized residual）對應時間的殘差圖。請問該分析師所配適的複迴歸模型是否合適？如果不合適，請詳細說明原因和解決的方法。（10分）

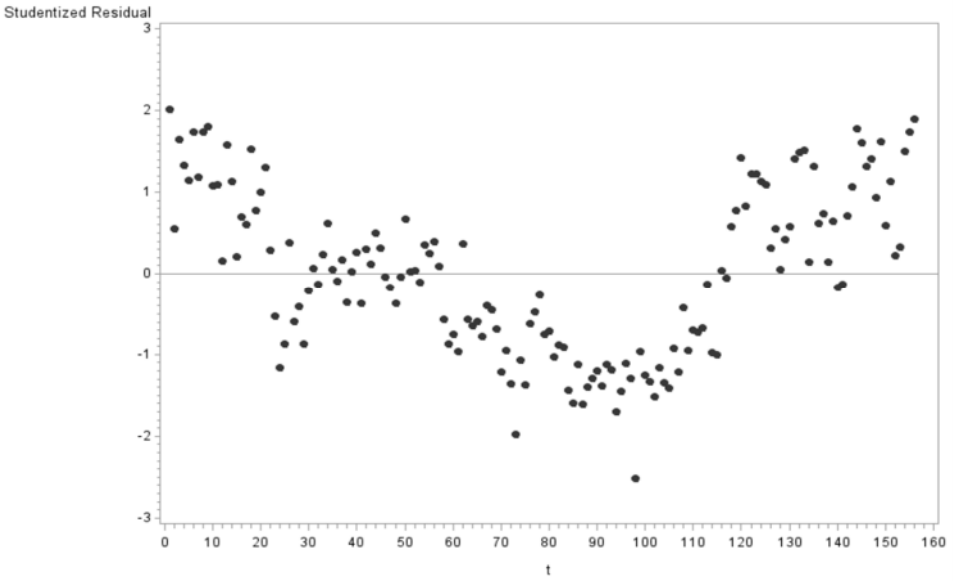

圖2

三、一位數據分析師受託分析於33（n=33）位男學生，其腳長（Y，以公分為單位）和 X 身高（以英吋為單位）的關係。所建立的簡單線性模型如下：

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

請使用表2部分電腦輸出報表來回答以下問題。表2第一欄是觀察值的順序，第二欄是殘差值。

㈠請說明何謂異常點（outlier）和高槓桿觀察值（high leverage observation），及其之間的區別。（8分）

㈡表2第三欄是標準化的殘差值（studentized residual）。請以此判斷是否有異常點存在？請說明判斷準則。

表2第五欄是 Student 化刪除殘差（Studentized deleted residuals，以 R-Student 表示）。第 i 個 R-Student 殘差是在假定將資料中的第 i 個觀察值刪除，然後以剩下的 n-1個觀察值來建立新的估計迴歸方程式而標準化獲得的 R-Student 殘差值。請以此判斷是否有異常點存在？請說明判斷準則。（8分）

㈢表2第六欄是 h_{ii}（hat value），其公式為 $h_{ii} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{n}(X_j - \bar{X})^2}$，

請問 $\sum_{i=1}^{n} h_{ii}$ 的值為何？請以此判斷是否有可能的高槓桿觀察值存在？請說明判斷準則。表2的最後一欄，第八欄是 DFFITS（Difference in Fits）值。請以此判斷是否有可能的影響點（influential observation）存在？請說明判斷準則。（8分）

表2

Obs	Residual	Student Residual	Cook's D	R-Student	Hat Diag H	Cov Ratio	DFFITS
1	0.541	0.443	0.011	0.438	0.101	1.173	0.147
2	0.906	0.718	0.009	0.712	0.035	1.070	0.136
3	-1.777	-1.410	0.041	-1.434	0.040	0.974	-0.293
4	0.390	0.308	0.002	0.304	0.033	1.097	0.056
5	-0.977	-0.772	0.010	-0.767	0.032	1.061	-0.140
6	-1.510	-1.194	0.024	-1.203	0.033	1.005	-0.222
7	1.490	1.179	0.024	1.186	0.033	1.007	0.219
8	-0.160	-0.127	0.000	-0.125	0.045	1.117	-0.027
9	1.023	0.809	0.011	0.804	0.032	1.057	0.147
10	-0.510	-0.403	0.003	-0.398	0.033	1.093	-0.073
11	1.957	1.563	0.067	1.602	0.052	0.956	0.374
12	0.157	0.125	0.000	0.123	0.052	1.125	0.029
13	1.023	0.809	0.011	0.804	0.032	1.057	0.147
14	0.556	0.444	0.005	0.438	0.050	1.110	0.101
15	-0.777	-0.614	0.006	-0.608	0.032	1.077	-0.111
16	-0.243	-0.192	0.001	-0.189	0.030	1.099	-0.034
17	-2.043	-1.632	0.073	-1.679	0.052	0.941	-0.392
18	-1.810	-1.458	0.078	-1.486	0.068	0.994	-0.402
19	0.140	0.110	0.000	0.109	0.031	1.101	0.019
20	2.356	1.944	0.236	2.041	0.111	0.926	0.721
21	0.623	0.522	0.022	0.516	0.141	1.221	0.209
22	0.490	0.388	0.003	0.382	0.033	1.093	0.071
23	0.790	0.627	0.008	0.620	0.039	1.083	0.125
24	-0.843	-0.697	0.031	-0.691	0.114	1.168	-0.248
25	-0.810	-0.641	0.007	-0.635	0.033	1.075	-0.117
26	1.490	1.179	0.024	1.186	0.033	1.007	0.219
27	0.490	0.388	0.003	0.382	0.033	1.093	0.071
28	-3.545	-3.437	3.274	-4.299	0.357	0.636	-3.200
29	0.089	0.073	0.000	0.072	0.086	1.168	0.022
30	0.257	0.203	0.001	0.200	0.030	1.098	0.035
31	-1.277	-1.013	0.021	-1.014	0.040	1.040	-0.207
32	1.323	1.065	0.040	1.067	0.066	1.061	0.283
33	0.190	0.153	0.001	0.151	0.068	1.144	0.041

四、一位統計分析師受託預測單位面積房價，欲了解房價受到那些因素所影響。收集了408筆有關於單位面積房價，屋齡（X_1，以年為單位），到最近的地鐵站的距離（X_2），便利商店數量（X_3），房屋座落的緯度（X_4）和經度（X_5）。擬考慮的模型如下：

模型1 $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i, i = 1,\ldots,n.$

模型2 $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i, \quad i = 1,\ldots,n.$

模型3 $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i, \quad i = 1,\ldots,n.$

請使用表3部分電腦輸出三個模型的變異數分析表（ANOVA, Analysis of Variance）報表來回答以下問題。

表3　模型1 ANOVA表

Response： Y	DF	Sum of squares	Mean square	F value	P-value
Model	5	44260	8852.03227	134.46	<.0001
Error	402	26465	65.83443		
Corrected Total	407	70726			

模型2 ANOVA表

Response： Y	DF	Sum of squares	Mean square	F value	P-value
Model	3	41703	13901	193.50	<.0001
Error	404	29023	71.83833		
Corrected Total	407	70726			

模型3 ANOVA表

Response： Y	DF	Sum of squares	Mean square	F value	P-value
Model	4	41879	10470	146.27	<.0001
Error	403	28847	71.57982		
Corrected Total	407	70726			

㈠在考慮模型1之下，請檢定便利商店數量（X_3）這個解釋變數是否可以從給定模型1中刪除。請用顯著水準$\alpha = 0.05$檢定並敘述對立假設、檢定統計量之值、決策法則和結論。（8分）

㈡在考慮模型1之下，請檢定房屋座落的緯度（X_4）和經度（X_5）這兩個解釋變數是否在模型1對預測單位面積房價有影響。亦即請用$\alpha = 0.05$檢定$H_0 : \beta_4 = \beta_5 = 0$，並請敘述對立假設、檢定統計量之值、決策法則和結論。（8分）

㈢請計算模型1，2和3的調整的複判定係數 R^2（the adjusted R-squared）並試述其意義。請敘述㈠㈡檢定，模型誤差項所需要的假設，並綜合㈠㈡檢定結果，請說明在模型1，2和3中，何者模式為最佳模型。（10分）