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，須詳列解答過程。
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 分佈 $\alpha=0.05$，臨界值 $F_{0.05}(df1,df2)$，$t_{0.05}(28)=1.701$，$t_{0.025}(28)=2.048$。

| | df1 | | |
|---|---|---|---|
| df2 | | 1 | 2 |
| | 28 | 4.196 | 3.340 |
| | 29 | 4.183 | 3.328 |
| | 50 | 4.034 | 3.183 |
| | 52 | 4.027 | 3.175 |

一、請回答下列問題：

㈠圖 1 是探討美國在游泳池溺斃（Swimming-pool drownings）的人數和美國核能發電廠發電（Nuclear power plants）數量數之間的關係，這兩個變數的相關係數為 90.12%。請試述以簡單線性迴歸分析是否具有因果關係或意義？請說明理由。（5 分）

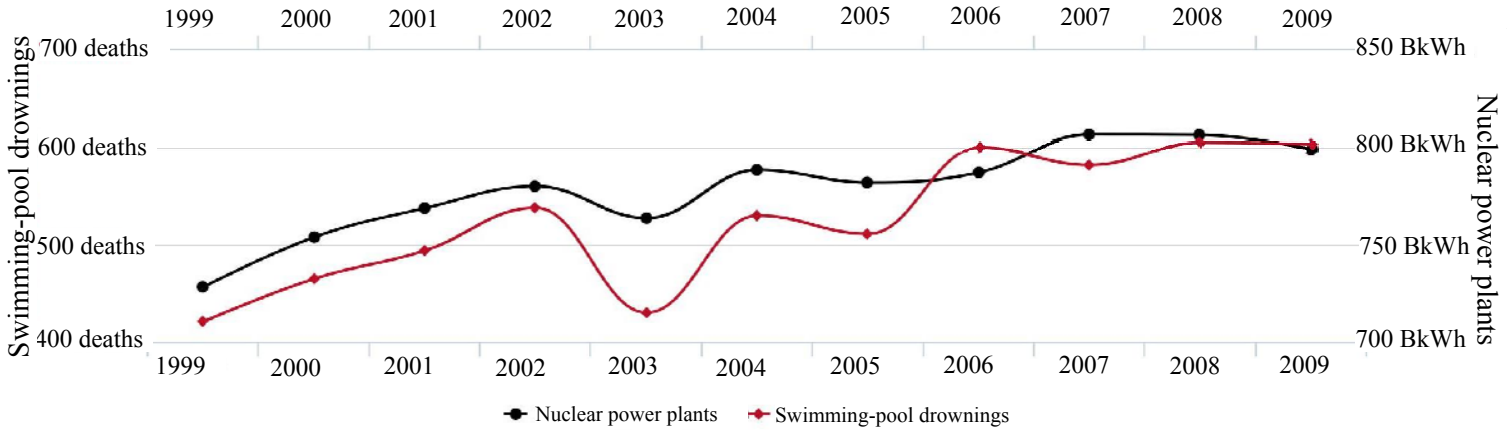


圖 1

㈡一位數據分析師擬研究滷肉飯銷售量受到那些因素所影響。所蒐集的可能解釋變數有價格、店內坪數、客流量、附近店家數、店內位置數、營業時間、店齡、配菜種類、選取肉的部位、米的種類等十個可能的解釋變數。該分析師計畫作複迴歸分析，要選擇重要解釋變數來描述反應變數（滷肉飯銷售量），請試述四種選擇重要變數的方法。又大數據的時代來臨，我們應用迴歸分析，有時會遇到高維度解釋變數的情況，解釋變數的個數（p）大到超過於樣本數（n）的情況，在高維度的解釋變數情況，請試述上述四種選擇重要變數之方法是否仍適用？如果你的答案為不適用，請說明理由。（10 分）

（請接第二頁）

類　　科：統計
科　　目：迴歸分析

二、一位分析師隨機抽取 55 位大學生並蒐集到五個變數。該分析師希望研究身高（$Y$，英吋）與受測者左前臂長度（$X_1$，公分）、左腳長度（$X_2$，公分）、頭圍（$X_3$，公分）和鼻長（$X_4$，公分）之間的關係。該分析師考慮配適下列三個迴歸模型：

模型 1：$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i$

模型 2：$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$

模型 3：$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$

請使用表 1 和表 2 中部分 R 統計軟體輸出之變異數分析表（ANOVA，Analysis of Variance）報表來回答以下問題：（每小題 10 分，共 30 分）

表 1　模型 1 ANOVA 表

| Response：Y | DF | Sum of squares | Mean square | F value |
|---|---|---|---|---|
| $X_1$ | 1 | 590.21 | 590.21 | 123.8106 |
| $X_2 \mid X_1$ | 1 | 224.35 | 224.35 | 47.0621 |
| $X_3 \mid X_1, X_2$ | 1 | 1.4 | 1.4 | 0.294 |
| $X_4 \mid X_1, X_2, X_3$ | 1 | 0.43 | 0.43 | 0.0896 |
| Error | 50 | 238.35 | 4.77 | |

表 2　模型 2 ANOVA 表

| Response：Y | DF | Sum of squares | Mean square | F value |
|---|---|---|---|---|
| $X_1$ | 1 | 590.21 | 590.21 | 127.782 |
| $X_2 \mid X_1$ | 1 | 224.35 | 224.35 | 48.572 |
| Error | 52 | 240.18 | 4.62 | |

㈠假設該分析師採用模型 1。在顯著水準 α=0.05 之下，請檢定 $X_3$ 和 $X_4$ 兩個解釋變數是否可以從給定模型 1 中刪除。也就是用 α=0.05 檢定 $H_0: \beta_3 = \beta_4 = 0$，並試述對立假設，檢定統計量之值、決策法則和結論。並請計算偏相關係數 $R^2_{Y,X_3,X_4|X_1,X_2}$（partial $R^2$）。

㈡假設該分析師採用模型 2。也就是在模型中僅考慮了兩個解釋變數，這兩個解釋變數是學生的左前臂長度（$X_1$）和左腳長度（$X_2$）。該分析師想知道這兩個解釋變數是否與身高（$Y$）有線性關係。在顯著水準 α=0.05 之下，請檢定 $H_0: \beta_1 = \beta_2 = 0$。並請試述檢定統計量之值、決策法則和結論。另請計算模型 2 的調整的複判定係數 $R^2$（adj $R^2$，the adjusted R-squared）並試述其意義。又該分析師要把身高的單位英吋轉公分（英吋乘以 2.54），試述模型 2 的 adj $R^2$ 是否改變？

㈢假設該分析師採用模型 3。只考慮模型中具有一個解釋變數，為學生的左前臂長度（$X_1$）。在顯著水準 α=0.05 下，該分析師想知道一個額外的解釋變數 $X_2$ 是否在解釋身高上具有顯著的貢獻。也就是說，該分析師想知道 $X_2$ 對模型 3 的貢獻。請協助回答此問題並說明對立假設、檢定統計量之值、決策法則和結論。在表 1 和表 2 的 F 檢定中，請試述需要做何假設，才能執行這些 F 檢定。

（請接第三頁）

類　　科：統計
科　　目：迴歸分析

三、㈠在作迴歸分析時，經常會遇到離群值和有影響力觀察值（influential data point）的問題。請試述何謂離群值和有影響力觀察值。並請分別試述兩種判斷準則偵測迴歸分析中的離群值和有影響力觀察值。（12 分）

㈡圖 2A 是一組數據的散佈圖，圖 2B 提供兩條估計線，實線估計式$\hat{Y}_i = 2.8 + 4.97X_i$包括第 51 點觀察值$((X_{51}, Y_{51}) = (4,50))$，虛線估計式$\hat{Y}_i = 3.68 + 4.98X_i$不包括第 51 點觀察值。請試述這組數據集是否包含任何離群值？並請試述這組數據是否包含任何有影響力觀察值？另請說明理由。（4 分）

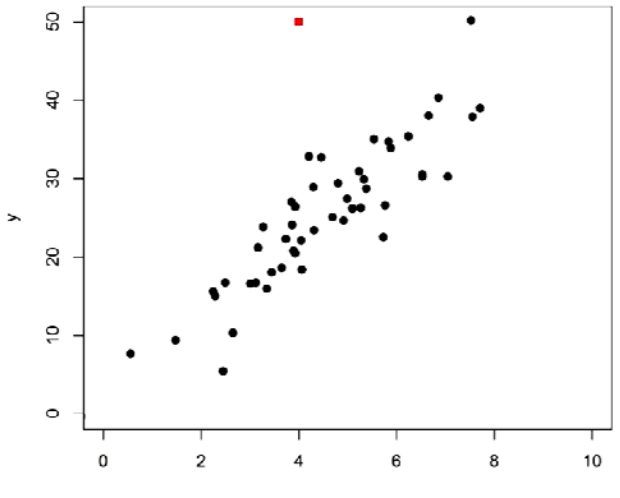


圖 2A

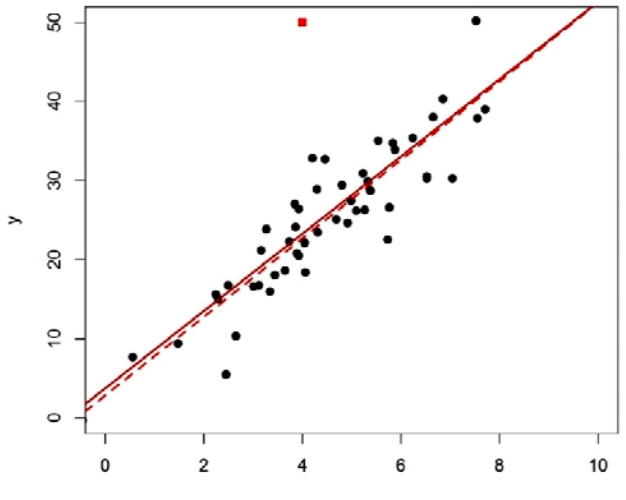


圖 2B

㈢圖 3A 是另一組數據的散佈圖，圖 3B 提供兩條估計線，實線估計式$\hat{Y}_i = 6.95 + 4.08X_i$包括第 41 點觀察值$((X_{41}, Y_{41}) = (10,16))$，虛線估計式$\hat{Y}_i = 1.93 + 5.21X_i$不包括第 41 點觀察值。請試述這組數據集是否包含任何離群值？並請試述這組數據集是否包含任何有影響力觀察值？另請說明理由。（4 分）

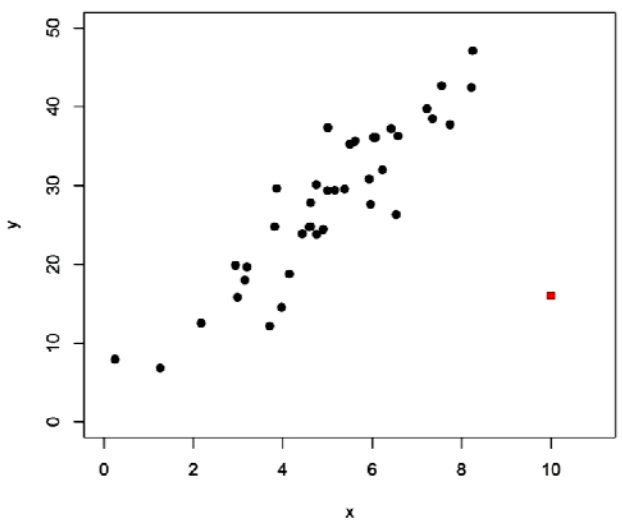


圖 3A

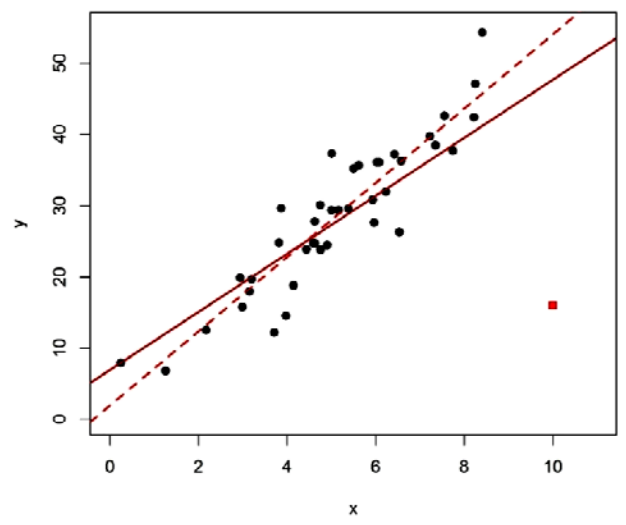


圖 3B

（請接第四頁）

類　　科：統計
科　　目：迴歸分析

四、一位數據分析師受冰飲企業老闆的委託，欲知道每日最高溫和該公司冰品銷售是否有線性關係，以作為未來商品促銷的依據。他蒐集了每日最高溫（$X$，以攝氏為單位）和冰品銷售（$Y$），共 30 個樣本點。下列是這些數據的統計量：

$n=30$，$\overline{X}=28.9892$，$\overline{Y}=34.7065$，$SXY=\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})=360.2128$

$SXX=\sum_{i=1}^{n}(X_i-\overline{X})^2=556.0186$，$SYY=\sum_{i=1}^{n}(Y_i-\overline{Y})^2=353.0085$

㈠在配適$E(Y\mid X=x)=\alpha+\beta_1(x-\overline{X})$的簡單線性迴歸方程式下，請利用最小平方法計算參數估計值（$\hat{\alpha}$和$\hat{\beta}_1$）與分別之標準誤。並請試述$\hat{\alpha}$和$\hat{\beta}_1$的共變異數，也就是 Cov($\hat{\alpha}$, $\hat{\beta}_1$)。（15 分）

㈡請在試卷上，完成下列變異數分析表。在顯著水準 α=0.05，請協助檢定$H_0$: $\beta_1=0$。並請試述檢定統計量之值、決策法則、結論和所需要之假設。（10 分）

| Source | Sum of Squares | DF | Mean square | F value |
|---|---|---|---|---|
| Regression | （1） | （4） | | |
| Error | （2） | （5） | （6） | |
| Total | （3） | | | |

五、一位分析師擬以$\tilde{\beta}_1=\frac{1}{n-1}\sum_{i=2}^{n}\left[\frac{Y_i-Y_{i-1}}{X_i-X_{i-1}}\right]$估計簡單線性迴歸模型$Y_i=\beta_0+\beta_1X_i+\varepsilon_i$，$i=1,...,n$之斜率$\beta_1$。他可以證明$\tilde{\beta}_1$是一個不偏估計式。請寫出$\beta_1$的最小平方估計式$\hat{\beta}_1$。在無須推導$\tilde{\beta}_1$的變異數下，試述相較於最小平方估計式$\hat{\beta}_1$，$\tilde{\beta}_1$和$\hat{\beta}_1$何者為最佳之估計式？請詳細敘述所依據的理由或定理。（10 分）