國立臺灣大學 114 學年度碩士班招生考試試題

科目：資訊管理導論（統計學部分）

題號：278、節次：7

考古題詳解

Question 4（多元線性迴歸）

題目 1. Amy analyzes the factors that determine how much garbage a family produces per week. Factors including the size of the house, the number of children, and the number of adults who are usually home during the day are considered. Below is the report produced by a regression routine. Answer the following questions. You should provide reasoning and justifications for your solution.

OLS Regression Results

Dep. Variable:	Garbage	R-squared:	0.170
Model:	OLS	Adj. R-squared:	[UNK1]
Method:	Least Squares	F-statistic:	29.80
No. Observations:	440	Prob (F-statistic):	1.52e-17
Df Residuals:	[UNK2]		
Df Model:	3		

	coef	std err	t	$P > \|t\|$	[0.025, 0.975]
Intercept	7.1943	1.092	[UNK3]	0.000	—
Q(“House Size”)	0.0019	0.001	—	0.001	0.001, 0.003
Children	1.1028	0.141	—	0.000	0.826, 1.379
Adults	1.0425	0.233	—	0.000	0.585, 1.500

(1) Compute the value of adjusted R-squared (UNK1) based on available information. (5%)

(2) Compute the residual degree of freedom (UNK2) and the t-value of the Intercept (UNK3) based on available information. (4%)

(3) Conduct the following hypothesis test (using 95% significance level): (6%)

H_0: Coefficients of all independent variables are jointly zero.

H_1: At least one independent variable has a non-zero coefficient.

解答. 本題為**多元線性迴歸**問題，迴歸模型為：

$$\text{Garbage}_i = \beta_0 + \beta_1 \cdot \text{HouseSize}_i + \beta_2 \cdot \text{Children}_i + \beta_3 \cdot \text{Adults}_i + \epsilon_i$$

已知資訊：

- 觀測數 $n = 440$

- 自變數個數 $k = 3$（不含截距）
- 模型自由度 Df Model $= k = 3$
- $R^2 = 0.170$
- F-statistic $= 29.80$

(1) 計算 Adjusted R-squared (UNK1)

調整後判定係數（Adjusted R^2）的公式為：

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}(1-R^2)$$

其中 n 為觀測數，k 為自變數個數（不含截距）。
代入數值：

$$\begin{aligned}\bar{R}^2 &= 1 - \frac{440-1}{440-3-1}(1-0.170)\\ &= 1 - \frac{439}{436}(0.830)\\ &= 1 - 1.00688 \times 0.830\\ &= 1 - 0.8357\\ &= 0.1643\end{aligned}$$

$$\boxed{\text{UNK1} = \bar{R}^2 \approx 0.164}$$

公式的意義：Adjusted R^2 對增加的自變數個數進行懲罰，避免僅因增加變數而使 R^2 虛高。當 k 增加而解釋力未顯著提升時，Adjusted R^2 可能下降。

(2) 計算殘差自由度 (UNK2) 與截距的 t 值 (UNK3)

殘差自由度：

$$\text{Df Residuals} = n - k - 1 = 440 - 3 - 1 = 436$$

$$\boxed{\text{UNK2} = 436}$$

截距的 t 值：
t 統計量的公式為：

$$t = \frac{\hat{\beta} - \beta_0}{\text{SE}(\hat{\beta})}$$

對於檢定 $H_0 : \beta_0 = 0$（截距為零）：

$$t = \frac{\hat{\beta}_0}{\text{SE}(\hat{\beta}_0)} = \frac{7.1943}{1.092} = 6.588$$

$$\boxed{\text{UNK3} \approx 6.59}$$

(3) 進行 F 檢定

假設：

- $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$（所有自變數係數同時為零）
- H_1: 至少有一個 $\beta_j \neq 0$（至少一個自變數有顯著影響）

檢定統計量：
F 統計量的公式為：

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}/k}{\text{SSE}/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

驗算（利用 $R^2 = 0.170$）：

$$F = \frac{0.170/3}{(1-0.170)/436} = \frac{0.0567}{0.001904} = 29.77 \approx 29.80 \quad \checkmark$$

臨界值：
自由度：$df_1 = k = 3$，$df_2 = n - k - 1 = 436$
顯著水準 $\alpha = 0.05$
由 F 表，$F_{0.05,3,436}$。由於 $df_2 = 436$ 很大，可用 $df_2 = 120$ 或 ∞ 近似：
從題目附表：$F_{0.05,3,120} = 2.6802$，$F_{0.05,3,\infty} = 2.6049$
故 $F_{0.05,3,436} \approx 2.63$
決策：
$F = 29.80 > 2.63 = F_{0.05,3,436}$，落在拒絕域內。
或者直接看 p-value $= 1.52 \times 10^{-17} < 0.05$，拒絕 H_0。
結論：

在 95% 信心水準下，拒絕 H_0。有非常強的統計證據顯示，至少有一個自變數（房屋大小、兒童數或成人數）對家庭垃圾產量有顯著影響。迴歸模型整體具有統計顯著性。

Question 5（加權估計量的性質）

題目 2. Let $Y_1, Y_2, \ldots, Y_T$ be an identical and independent sample from population Ω with mean μ and variance σ^2. T is a multiple of 5. Define estimator m as:

$$m = \frac{\sum_{i=1}^{T} g(i) Y_i}{\sum_{i=1}^{T} g(i)},$$

where $g(i) \equiv 1 + (i \mod 5)$ and mod is the modular operator.

(1) Prove that m is an unbiased estimator for μ. (10%)

(2) What is the variance of m? (10%)

解答. 前置分析：理解 $g(i)$ 函數

$g(i) = 1 + (i \mod 5)$ 的值取決於 i 除以 5 的餘數：

i	1	2	3	4	5	6	7	8	9	10	$\cdots$
$i \mod 5$	1	2	3	4	0	1	2	3	4	0	$\cdots$
$g(i)$	2	3	4	5	1	2	3	4	5	1	$\cdots$

因此 $g(i)$ 以週期 5 循環，取值為 $\{2, 3, 4, 5, 1\}$。
由於 T 是 5 的倍數，設 $T = 5N$，則 $g(i)$ 恰好完整循環 N 次。
計算 $\sum_{i=1}^{T} g(i)$：
每一個週期的總和為 $2 + 3 + 4 + 5 + 1 = 15$。
共 $N = T/5$ 個完整週期，故：

$$\sum_{i=1}^{T} g(i) = N \times 15 = \frac{T}{5} \times 15 = 3T$$

因此，估計量 m 可改寫為：

$$m = \frac{\sum_{i=1}^{T} g(i) Y_i}{3T} = \frac{1}{3T} \sum_{i=1}^{T} g(i) Y_i$$

(1) 證明 m 是 μ 的不偏估計量

證明：

取期望值：

$$\begin{aligned}
\mathbb{E}[m] &= \mathbb{E}\left[\frac{\sum_{i=1}^{T} g(i)Y_i}{\sum_{i=1}^{T} g(i)}\right] \\
&= \frac{1}{3T}\mathbb{E}\left[\sum_{i=1}^{T} g(i)Y_i\right] \\
&= \frac{1}{3T}\sum_{i=1}^{T} g(i)\mathbb{E}[Y_i] \quad (g(i) \text{ 為常數，期望值的線性性}) \\
&= \frac{1}{3T}\sum_{i=1}^{T} g(i)\cdot\mu \quad (\mathbb{E}[Y_i]=\mu \text{ 對所有 } i) \\
&= \frac{\mu}{3T}\sum_{i=1}^{T} g(i) \\
&= \frac{\mu}{3T}\cdot 3T \\
&= \mu
\end{aligned}$$

$$\boxed{\mathbb{E}[m]=\mu}$$

因此 m 是 μ 的**不偏估計量**。 □

(2) 計算 m 的變異數

計算：

由於 $Y_1, Y_2, \ldots, Y_T$ 是獨立的，且 $\mathrm{Var}(Y_i)=\sigma^2$：

$$\begin{aligned}
\mathrm{Var}(m) &= \mathrm{Var}\left(\frac{\sum_{i=1}^{T} g(i)Y_i}{3T}\right) \\
&= \frac{1}{(3T)^2}\mathrm{Var}\left(\sum_{i=1}^{T} g(i)Y_i\right) \\
&= \frac{1}{9T^2}\sum_{i=1}^{T}\mathrm{Var}(g(i)Y_i) \quad (\text{獨立性} \Rightarrow \text{變異數可加}) \\
&= \frac{1}{9T^2}\sum_{i=1}^{T}[g(i)]^2\mathrm{Var}(Y_i) \quad (\mathrm{Var}(aX)=a^2\mathrm{Var}(X)) \\
&= \frac{\sigma^2}{9T^2}\sum_{i=1}^{T}[g(i)]^2
\end{aligned}$$

計算 $\sum_{i=1}^{T}[g(i)]^2$：

每一個週期（5 個元素）的平方和為：

$$2^2+3^2+4^2+5^2+1^2 = 4+9+16+25+1 = 55$$

共 $N = T/5$ 個完整週期，故：

$$\sum_{i=1}^{T}[g(i)]^2 = N \times 55 = \frac{T}{5} \times 55 = 11T$$

代入：

$$\mathrm{Var}(m) = \frac{\sigma^2}{9T^2} \cdot 11T = \frac{11\sigma^2}{9T}$$

$$\boxed{\mathrm{Var}(m) = \frac{11\sigma^2}{9T}}$$

補充討論：與樣本平均數的比較

樣本平均數 $\bar{Y} = \frac{1}{T}\sum_{i=1}^{T} Y_i$ 的變異數為 $\mathrm{Var}(\bar{Y}) = \frac{\sigma^2}{T}$。

比較兩者：

$$\frac{\mathrm{Var}(m)}{\mathrm{Var}(\bar{Y})} = \frac{11\sigma^2/(9T)}{\sigma^2/T} = \frac{11}{9} \approx 1.22$$

因此 $\mathrm{Var}(m) > \mathrm{Var}(\bar{Y})$，表示加權估計量 m 的變異數較大，效率較低。

這是因為 m 給予不同觀測值不同的權重，而樣本平均數給予等權重。根據 Gauss-Markov 定理的推廣，在 i.i.d. 樣本下，等權重平均是最有效的線性不偏估計量。

知識點整理與歸納

Question 4 涉及的知識點

1. **多元線性迴歸模型**

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i$$

2. **自由度**

- 模型自由度：$df_{\text{model}} = k$（自變數個數）
- 殘差自由度：$df_{\text{residual}} = n - k - 1$
- 總自由度：$df_{\text{total}} = n - 1$

3. **調整後判定係數**

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}(1 - R^2) = 1 - \frac{\text{SSE}/(n-k-1)}{\text{SST}/(n-1)}$$

4. **F 檢定**（整體顯著性檢定）

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{R^2/k}{(1-R^2)/(n-k-1)} \sim F_{k,n-k-1}$$

5. **個別係數的 t 檢定**

$$t = \frac{\hat{\beta}_j}{\text{SE}(\hat{\beta}_j)} \sim t_{n-k-1}$$

Question 5 涉及的知識點

1. **加權平均估計量**

$$m = \frac{\sum_{i=1}^{n} w_i Y_i}{\sum_{i=1}^{n} w_i} = \sum_{i=1}^{n} c_i Y_i, \quad \text{其中} c_i = \frac{w_i}{\sum_j w_j}$$

2. **不偏性的證明**：若 $\sum c_i = 1$ 且 $\mathbb{E}[Y_i] = \mu$，則

$$\mathbb{E}\left[\sum c_i Y_i\right] = \sum c_i \mathbb{E}[Y_i] = \mu \sum c_i = \mu$$

3. **變異數的計算**：若 Y_i 獨立且 $\text{Var}(Y_i) = \sigma^2$，則

$$\text{Var}\left(\sum c_i Y_i\right) = \sum c_i^2 \text{Var}(Y_i) = \sigma^2 \sum c_i^2$$

4. **最佳線性不偏估計量**：在 i.i.d. 樣本下，等權重平均 $\bar{Y}$ 的變異數 σ^2/n 最小。

建議融入講義的章節

- Question 4 → **第六部分**需擴充「多元線性迴歸」內容，包括：
  - 多元迴歸模型與矩陣表示
  - 調整後判定係數 Adjusted R^2
  - 整體 F 檢定
  - 個別係數的 t 檢定
  - 多重共線性問題
- Question 5 → **第四部分**（點估計），作為估計量性質的進階例題，或作為練習題。