統計諮詢 - 作業 4

國立成功大學統計學系暨數據科學研究所

廖傑恩（RE6094028）

2021-04-02

1 Exercise 11.3

1.1 問題敘述

因為在 1970 年代晚期經歷嚴重的缺水，Concord 地區在 1980 年起開始執行節約措施。Hamilton（1983; 1992）研究了 Concord 家戶用水量的預測模型。本研究旨在以 Concord 家戶特徵對其 1981 年的用水量進行預測。

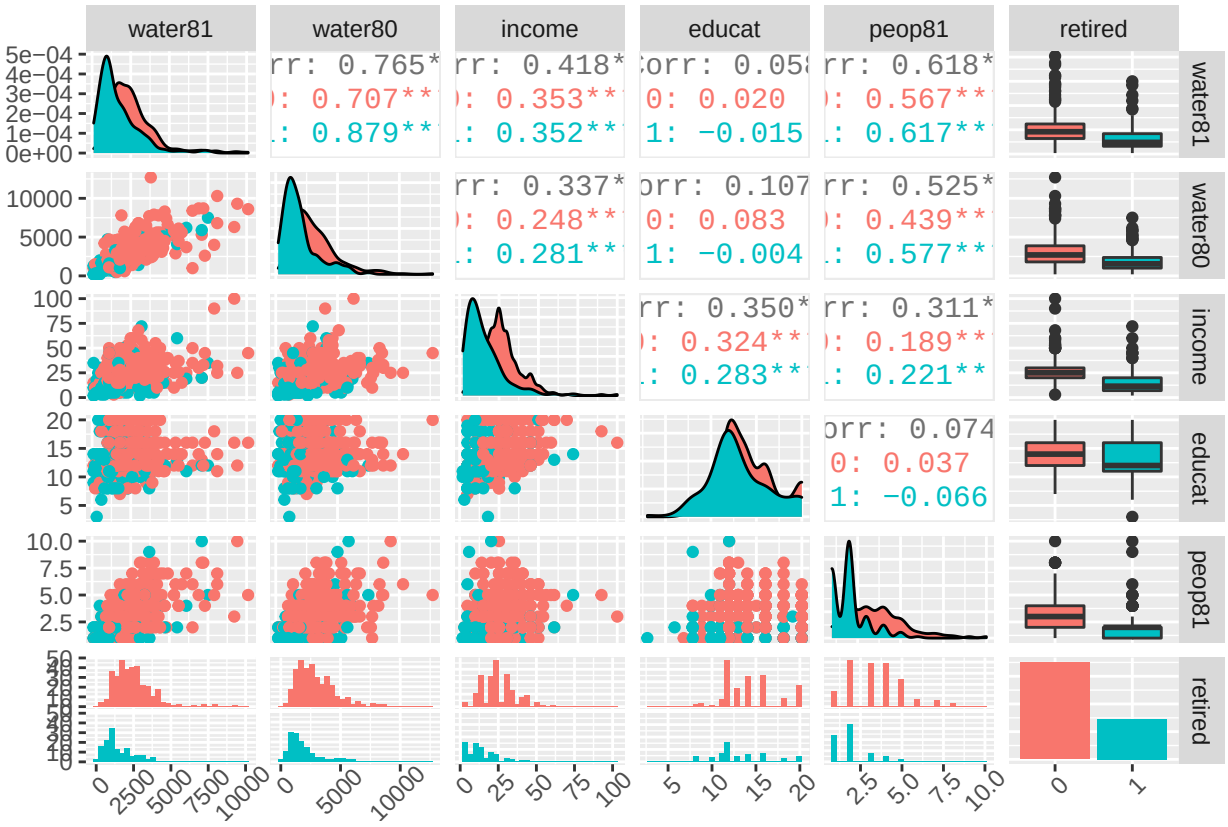
1.2 資料介紹

資料是 Concord496 個家戶的用水量以及家戶相關基本變項，共有 6 個變項，說明如下表:

變項名稱	變項意義	變項類型
water81	1981 年家戶用水量（單位：立方英尺）	連續
water80	1980 年家戶用水量（單位：立方英尺）	連續
income	1981 年家戶收入（單位：千元）	連續
educat	戶長教育程度（單位：年）	連續
peop81	1981 年夏季時家戶人數	連續
retired	戶長是否退休，是 =1，否 =0	類別

1.3 資料探索

我們繪製兩兩散佈圖、盒鬚圖或直方圖矩陣，並以二分變項分層，以快速觀察這些變項間的關聯:



由此矩陣可以發現，大部分數值型變項兩兩之間都有一定程度的線性相關，若直接進行線性迴歸的配適，可能會產生共線性（collinearity）的問題，必須要先檢查。此外，從圖中也可以發現變項多呈現右偏（right-skewed）的分布，要經過檢查才能確定是否有離群值、影響點或是槓桿點。

1.4 建立迴歸模型

1.4.1 變數與模型定義

- 定義變數

1. 令 Y 為 `water81`（1981 年家戶用水量）
2. 令 X_1 為 `water80`（1980 年家戶用水量）
3. 令 X_2 為 `income`（1981 年家戶收入）
4. 令 X_3 為 `educat`（戶長教育程度）
5. 令 X_4 為 `peop81`（1981 年夏季時家戶人數）
6. 令 X_5 為 `retired`（戶長是否退休），是 =1，否 =0。

- 定義模型

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \epsilon_i, \ \forall i = 1, ..., 496$$

其中 $\epsilon_i \overset{iid}{\sim} N(0, {\sigma_i}^2)$。

1.4.2 檢查共線性

由散佈圖可知獨變項間彼此之間有相關性，故必須先檢查獨變項間是否有共線性，以免共線性影響配適結果，檢查方式為先計算全部獨變項間的變異膨脹因子（variance inflation factor, VIF）值：

$$VIF_p = \frac{1}{1 - R_p^2}$$

其中 R_p^2 為將第 i 個獨變項視為依變項，其他變項當成獨變項建立迴歸模型，所得之判定係數。

獨變項之 VIF 值大於 5 表示其可由其他獨變項線性組合而成，表示與其他獨變項具有明顯的共線性。我們先剔除 VIF 值最大之獨變項，再進行一次 VIF 的計算，並重複以上過程直至所有獨變項之 VIF 值皆小於 5 為止。我們 5 個獨變項的 VIF 值如下表所示，均不具共線性。

變項	VIF 值
water80	1.452611
income	1.394802
educat	1.146499
peop81	1.508365
retired	1.289000

1.4.3 逐步迴歸分析（stepwise selection）

我們使用逐步選擇法中的向前選擇法（forward selection），逐步納入對模型貢獻程度最高的變項，直到模型配適度（goodness of fit）不再改善。在這裡，我們以赤池訊息量準則（Akaike information criterion, AIC）作為模型配適度指標，其公式如下：

$$AIC = -2\ln(L) + 2k = n\ln(\frac{SS_R}{n}) + 2k$$

其中 L 為概似函數（likelihood function），k 為參數數量，n 為觀察值個數，SS_R 是模型殘差平方和。

由於 SS_R 越高，表示模型不能夠解釋的變異性越大，因此 AIC 越大表示模型配適越差。

下表列出了我們進行逐步選擇的過程，其中「$+X$」意味著在上一步驟的模型中再納入變項 X。我們的模型納入了所有獨變項。

步驟編號	步驟	步驟完成後的模型 AIC
1	不考慮任何變項的原始模型	7246.494
2	+ water80	6812.380
3	+ peop81	6730.133

步驟編號	步驟	步驟完成後的模型 AIC
4	+ income	6707.949
5	+ educat	6700.821
6	+ retired	6698.726

經過變數挑選後的模型為:

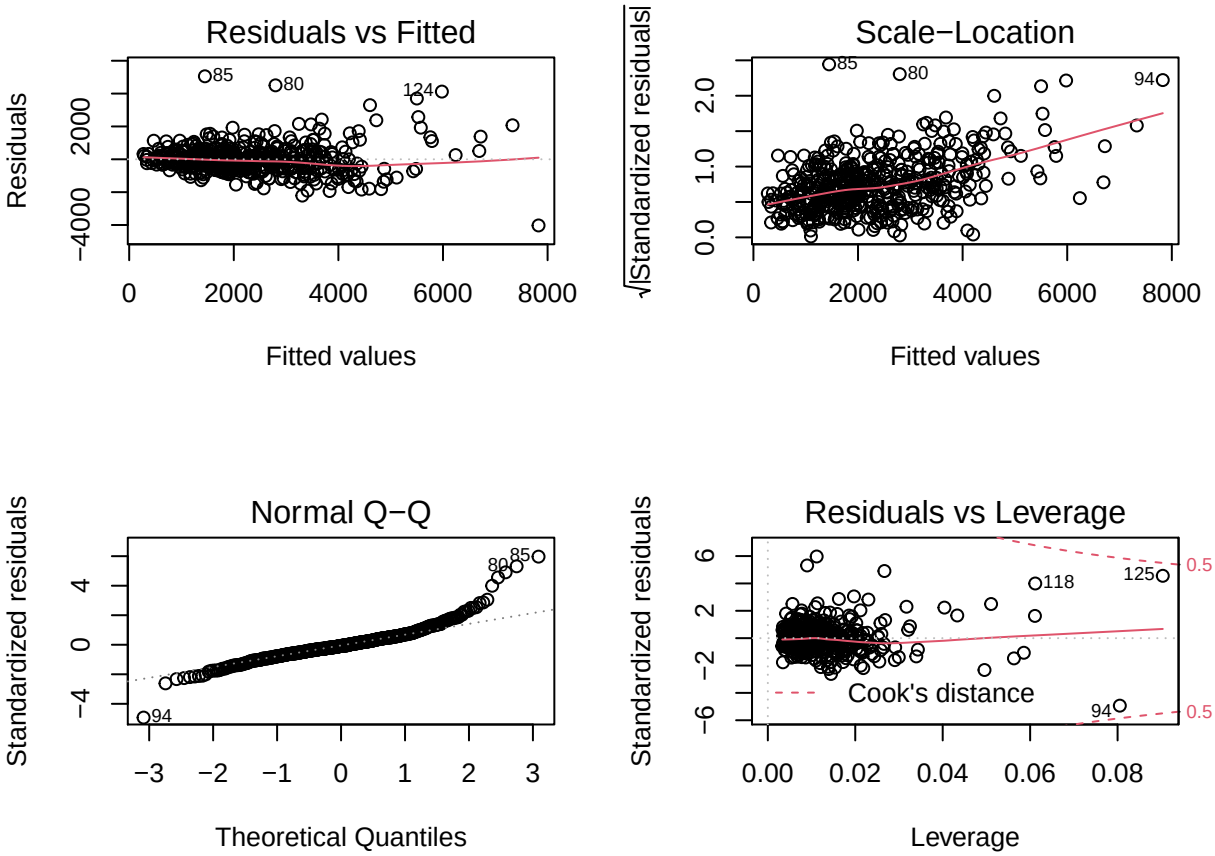
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5$$

其中 $\hat{\beta}_0 = 167.6075$，$\hat{\beta}_1 = 0.4876$，$\hat{\beta}_2 = 20.9782$，$\hat{\beta}_3 =$ -37.7761，$\hat{\beta}_4 = 255.6427$，$\hat{\beta}_5 = 191.9079$。

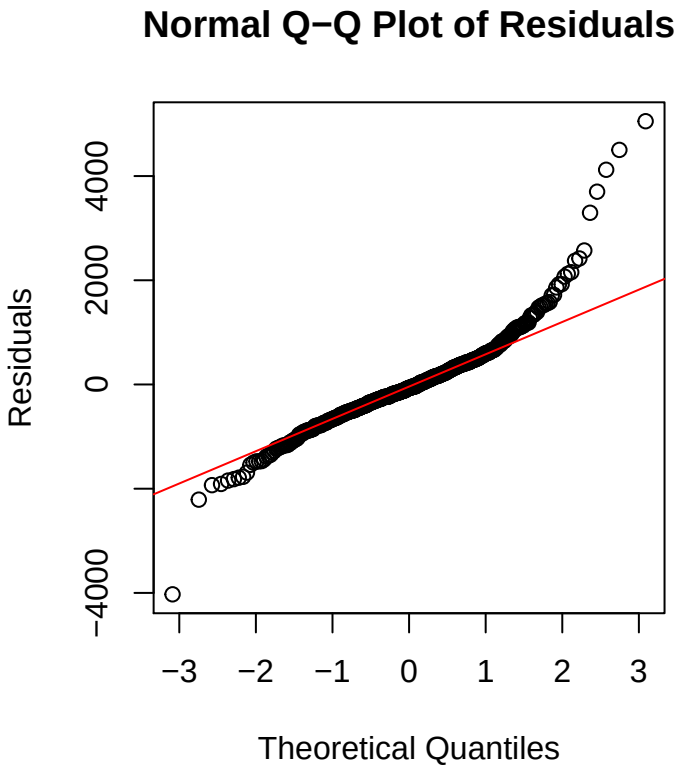
各個獨變項的係數估計值中，只有戶長教育程度 `educat`（$\hat{\beta}_4$）為負，表示隨著戶長教育程度提高，家戶在 1981 年的用水量會降低，而其他變項都與家戶在 1981 年的用水量呈現正相關，包含：家戶在 1980 年的用水量（`water80`）、家戶在 1981 年的收入（`income`）、家戶在 1981 年的人數（`peop81`）以及與戶長退休（`retired`）。

1.4.3.1 模型診斷

配適迴歸模型之後，我們應對殘差做檢定來確認線性迴歸模型中的前提假設 $\epsilon_i \overset{iid}{\sim} N(0, {\sigma_i}^2)$ 是否滿足，包含殘差常態性、殘差變異數同質與殘差獨立性。我們先繪製 4 張殘差圖以觀察殘差樣態：由左上及右上兩圖可以發現，殘差期望值相相當接近 0，不過殘差變異數有明顯變化趨勢。左下的常態 Q-Q 圖中資料點在值偏小及偏大處明顯偏離 45 度線，顯示殘差可能不服從常態分配。右下圖顯示資料中無槓桿點。



- 殘差常態性檢驗



上圖是模型殘差的 Normal Q-Q plot，圖中值較大時資料點偏離 45 度線，顯示殘差可能不服從常態分配。我們以 Shapiro-Wilk 檢定檢驗模型殘差 ϵ_i 是否為常態分配，令顯著水準為 0.05，其假設如下:

$$H_0 : \epsilon_i \sim ND \ v.s. \ H_1 : \epsilon_i \ does \ not \sim ND$$

檢定結果：檢定統計量 W 為 0.9087，p 值為 0，小於顯著水準，因此我們拒絕 H_0，表示我們有足夠的證據顯示殘差不服從常態分配，診斷未通過。

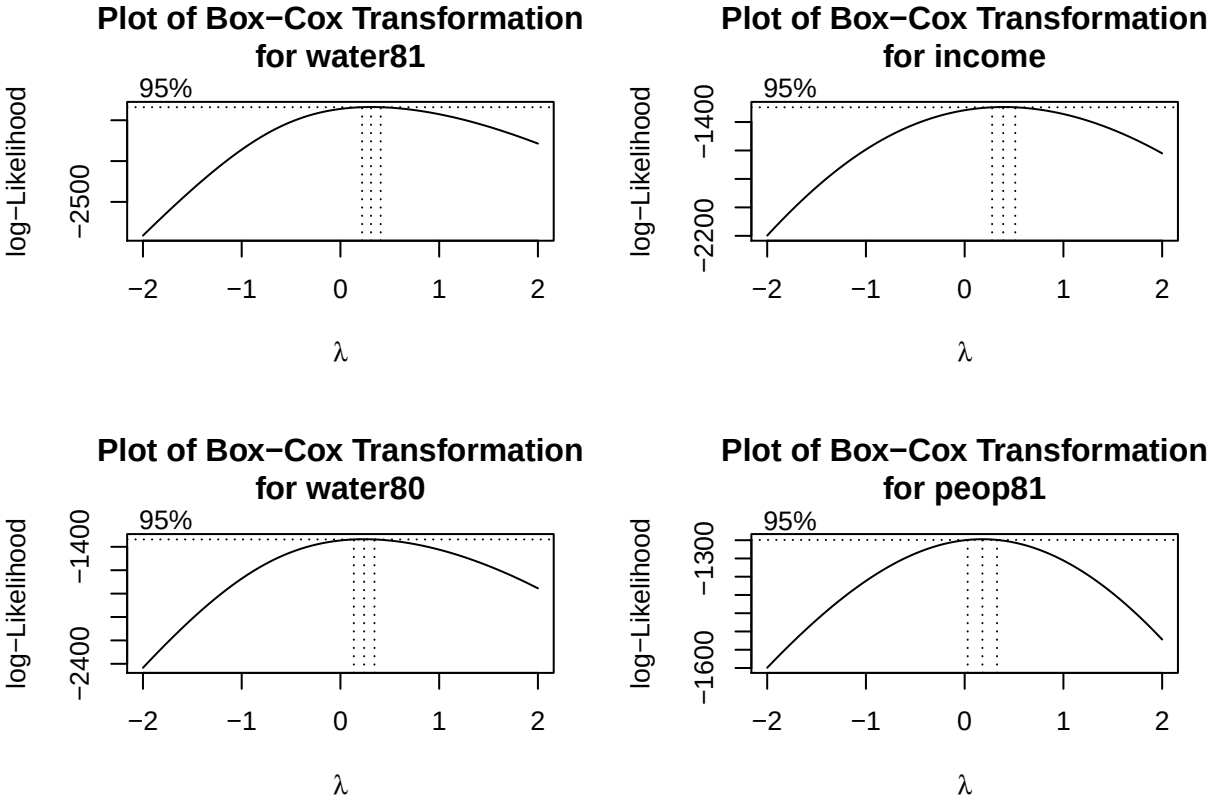
由於已知殘差常態假設未通過，便不繼續進行其他前提假設的檢驗。我們接著應對資料進行轉換，使得配適出來的迴歸模型得以符合線性迴歸的前提假設。

1.5 對資料進行轉換後再進行迴歸分析

1.5.1 Box-Cox 轉換

在資料探索時，我們已知依變項（`water81`）與獨變項 `water80`、`income` 與 `peop81` 等變項的分佈呈現明顯的右偏，有可能是因此使得模型未通過殘差常態假設檢定。我們採用 Box-Cox 轉換法對這些變項進行轉換，以壓縮其尺度。令 Y' 為轉換後的依變項、X'_j 為轉換後的獨變項 X_j，其數學式如下：

$$Y' = \begin{cases} \frac{Y^\lambda - 1}{\lambda}, & if \ \lambda \neq 0 \\ \log Y, & if \ \lambda = 0 \end{cases} \qquad X'_j = \begin{cases} \frac{X_j^{\lambda_j} - 1}{\lambda_j}, & if \ \lambda_j \neq 0 \\ \log X_j, & if \ \lambda_j = 0 \end{cases}$$



以上四張圖是分別對 `watwe81`、`water80`、`income` 與 `peop81` 進行 Box-cox 轉換時挑選 λ 的示意圖，我們在 $[-2, 2]$ 範圍中挑選使概似函數值最大的 λ，結果如下表所示：

	lambdas
water81	0.31
water80	0.24
income	0.39
peop81	0.18

由於各 95% 信賴區間都沒有包含 0，因此我們不考慮 log 轉換，而以這些非零的 λ 值對變項進行 Box-cox 轉換，轉換前與轉後的變項分佈如下 8 圖所示，可以發現轉換後的變項分佈對稱許多。

1.5.2 使用轉換後的資料配適迴歸模型

我們接著使用經 Box-Cox 轉換的資料配適線性迴歸模型。

1.5.2.1 變數與模型定義

- 定義變數

  1. 令 Y' 為經 Box-Cox 轉換的 `water81`（1981 年家戶用水量）
  2. 令 X_1' 為經 Box-Cox 轉換的 `water80`（1980 年家戶用水量）
  3. 令 X_2' 為經 Box-Cox 轉換的 `income`（1981 年家戶收入）
  4. 令 X_3 為 `educat`（戶長教育程度）
  5. 令 X_4 為 `peop81`（1981 年夏季時家戶人數）
  6. 令 X_5 為 `retired`（戶長是否退休），是 =1，否 =0。

- 定義模型

$$Y_i' = \beta_0' + \beta_1' X_{i1}' + \beta_2' X_{i2}' + \beta_3' X_{i3} + \beta_4' X_{i4} + \beta_5' X_{i5} + \epsilon_i',\ \forall i = 1, ..., 496$$

其中 $\epsilon_i' \overset{iid}{\sim} N(0, \sigma_i'^2)$。

1.5.2.2 檢查共線性

在配適模型前，我們同樣以 VIF 值來檢查是否共線性存在。獨變項之 VIF 值大於 5 表示其可由其他獨變項線性組合而成，表示與其他獨變項具有明顯的共線性。我們先剔除 VIF 值最大之獨變項，再進行一次 VIF 的計算，並重複以上過程直至所有獨變項之 VIF 值皆小於 5 為止。我們 5 個獨變項的 VIF 值如下表所示，均不具共線性。

變項	VIF 值
water80	1.599472
income	1.570261

變項	VIF 值
educat	1.159884
peop81	1.669177
retired	1.388340

1.5.2.3 逐步迴歸分析

我們同樣使用向前選擇法，逐步納入對模型貢獻程度最高的變項，直到模型配適度不再改善。我們同樣以 AIC 作為模型配適度指標，AIC 越大表示模型配適越差。下表列出了我們進行逐步選擇的過程，其中「$+X$」意味著在上一步驟的模型中再納入變項 X。

步驟編號	步驟	步驟完成後的模型 AIC
1	不考慮任何變項的原始模型	1915.723
2	+ water80	1383.638
3	+ peop81	1314.856
4	+ income	1304.614

經過變數挑選後的模型為:

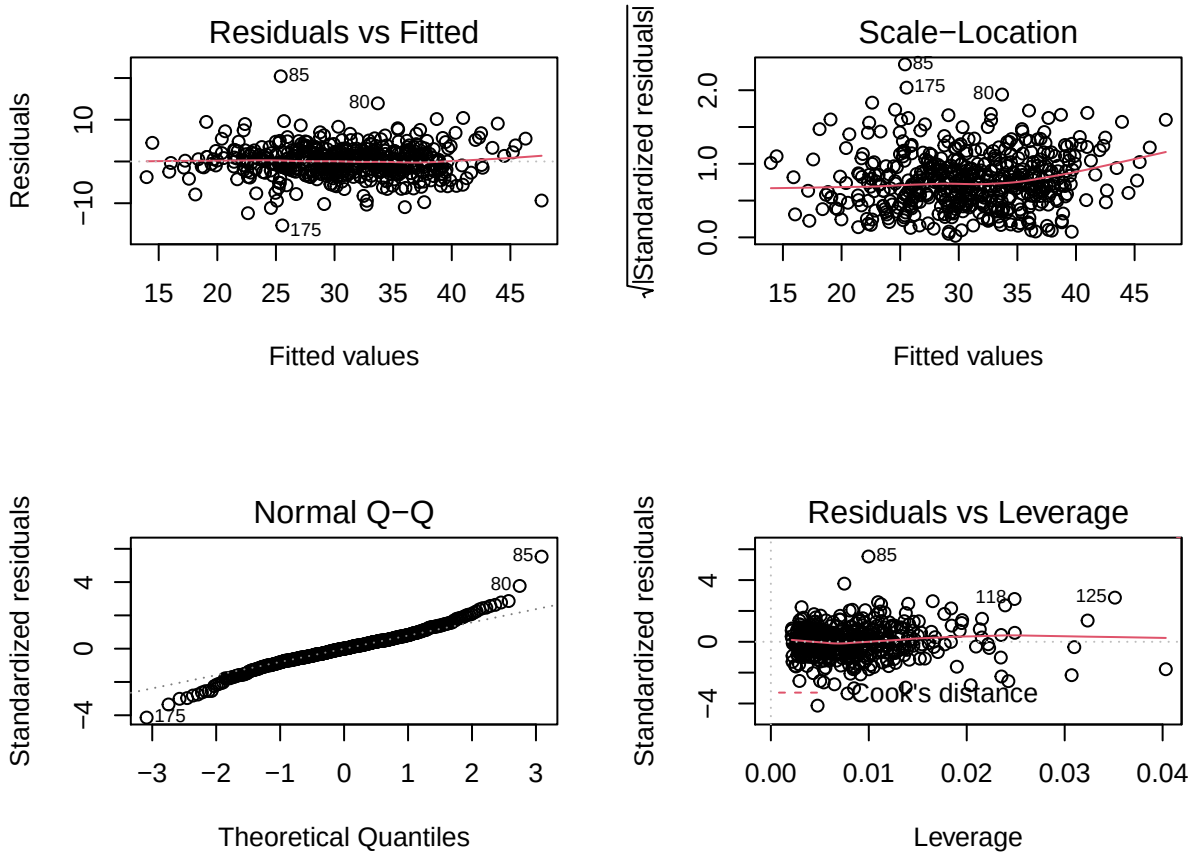
$$\hat{y}' = \hat{\beta}'_0 + \hat{\beta}'_1 x_1 + \hat{\beta}'_2 x_2 + \hat{\beta}'_4 x_4$$

其中 $\hat{\beta}'_0 = 2.5799$，$\hat{\beta}'_1 = 1.0398$，$\hat{\beta}'_2 = 0.3385$，$\hat{\beta}'_4 = 2.4669$。

各個獨變項的係數估計值都為正與家戶在 1981 年經由 Box-Cox 轉換的用水量呈現正相關，包含：經由 Box-Cox 轉換的 1980 年家戶用水量（`water80`）、經由 Box-Cox 轉換的 1981 年家戶收入（`income`）以及家戶在 1981 年的人數（`peop81`）。

1.5.2.4 模型診斷

配適迴歸模型之後，我們應對殘差做檢定來確認線性迴歸模型中的前提假設 $\epsilon'_i \overset{iid}{\sim} N(0, \sigma_i'^2)$ 是否滿足，包含殘差常態性、殘差變異數同質與殘差獨立性。我們先繪製 4 張殘差圖以觀察殘差樣態：由左上及右上兩圖可以發現，殘差期望值相相當接近 0，不過殘差變異數似乎有變化趨勢。左下的常態 Q-Q 圖中資料點在值偏小及偏大處偏離 45 度線，顯示殘差可能不服從常態分配。右下圖顯示資料中無槓桿點。



- 殘差常態性檢驗

我們以 Shapiro-Wilk 檢定與 Kolmogorov-Smirnov 檢定檢驗模型殘差 ϵ_i 是否為常態分配，令顯著水準為 0.05，其假設如下:

$$H_0 : \epsilon_i' \sim ND \ v.s. \ H_1 : \epsilon_i' \ does \ not \sim ND$$

Shapiro-Wilk 檢定結果: 檢定統計量 W 為 0.969，p 值為 0，小於顯著水準。Kolmogorov-Smirnov 檢定結果: 檢定統計量 D 為 0.2684，p 值為 0。因此我們拒絕 H_0，表示我們有足夠的證據顯示殘差不服從常態分配，診斷未通過。

- 以 Brown-Forsythe 檢定檢驗殘差變異同質性（令顯著水準為 0.05）

$$\begin{cases} H_0 : \sigma_i'^2 = \sigma_{i'}'^2 \ \forall i \neq i' \\ H_1 : Not \ H_0 \end{cases}$$

檢定結果: 檢定統計量 BF 為 71.2298，p 值為 0，小於顯著水準，因此我們拒絕 H_0，也就是說我們有充分證據支持殘差不具備變異同質性，未通過變異同質性假設。

- 以 Durbin-Waston 檢定檢驗殘差獨立性（令顯著水準為 0.05）

$$H_0 : \epsilon_i \ are \ independent. \ v.s. \ H_1 : \epsilon_i \ are \ not \ independent.$$

檢定結果如下：檢定統計量 DW 為 2.0611，p 值為 0.7495，不小於顯著水準，因此我們不拒絕 H_0，也就是說我們沒有充分證據支持殘差不具備獨立性，通過獨立假設。

1.6 依據類別變項分開建立迴歸模型

我們以經由 Box-Cox 轉換的資料建立的線性迴歸模型仍未通過殘差常態檢驗，也未通過殘差變異數同質性檢驗，有可能是極端值造成影響。此外，先前模型並未納入「戶長是否退休（`retired`）」這個變項，然而此變項與其他變項可能對依變項有交互作用，因此我們以考慮交互作用的變異數分析檢驗。結果顯示 `retired` 與 1980 年家戶用水量（`water80`）(F =,13$p=0$Box-Cox$F=, 8p=$0)。而我們使用轉換後的資料時，在逐步選擇時排除了 `retired`，顯示有可能要將 `retired` 各類別拆開，分別建立迴歸模型。

1.6.1 定義變數與模型

- 定義變數

  1. 令 Y' 為經 Box-Cox 轉換的 `water81`（1981 年家戶用水量）
  2. 令 X_1' 為經 Box-Cox 轉換的 `water80`（1980 年家戶用水量）
  3. 令 X_2' 為經 Box-Cox 轉換的 `income`（1981 年家戶收入）
  4. 令 X_3 為 `educat`（戶長教育程度）
  5. 令 X_4' 為 `peop81`（1981 年夏季時家戶人數）
  6. 令 X_5 為 `retired`（戶長是否退休），是 =1，否 =0。

- 定義模型

  - 模型 A - 針對戶長非退休之家戶建立的模型（$X_{i5}=0$）：$Y_i' = \beta_{0.0}' + \beta_{1.0}' X_{i1}' + \beta_{2.0}' X_{i2}' + \beta_{3.0}' X_{i3} + \beta_{4.0}' X_{i4} + \epsilon_i',\ \forall i = 1, ..., 350$
  - 模型 B - 針對戶長已退休之家戶建立的模型（$X_{i5}=1$）：$Y_j' = \beta_{0.1}' + \beta_{1.1}' X_{j1}' + \beta_{2.1}' X_{j2}' + \beta_{3.1}' X_{j3} + \beta_{4.1}' X_{j4} + \epsilon_j',\ \forall j = 1, ..., 146$

  其中 $\epsilon_i' \overset{iid}{\sim} N(0, \sigma_i'^2)$，$\epsilon_j' \overset{iid}{\sim} N(0, \sigma_j'^2)$。

1.6.1.1 檢查共線性

在配適模型前，我們同樣以 VIF 值來檢查是否共線性存在。我們 4 個獨變項的 VIF 值如下表所示，在兩模型中均不具共線性。

變項	模型 A 之 VIF 值	模型 B 之 VIF 值
water80	1.365276	1.477286
income	1.228893	1.301293
educat	1.120668	1.141373
peop81	1.336280	1.462180

1.6.1.2 逐步迴歸分析

我們同樣使用向前選擇法，逐步納入對模型貢獻程度最高的變項，直到模型配適度不再改善。我們同樣以 AIC 作為模型配適度指標，AIC 越大表示模型配適越差。下兩表分別列出了我們在模型 A 與模型 B 進行逐步選擇的過程，其中「$+X$」意味著在上一步驟的模型中再納入變項 X。

步驟編號	步驟	步驟完成後的模型 AIC
1	不考慮任何變項的原始模型	1269.0123
2	+ water80	985.7286
3	+ peop81	925.4181
4	+ income	914.8417
5	+ educat	913.5064

步驟編號	步驟	步驟完成後的模型 AIC
1	不考慮任何變項的原始模型	578.9767
2	+ water80	386.7738
3	+ peop81	381.7581
4	+ income	381.7283

經過變數挑選後，模型 A（針對戶長非退休之家戶建立的模型）納入的獨變項有：X_1'（經 Box-Cox 轉換的 `water80`）、X_2'（經 Box-Cox 轉換的 `income`）、X_3（`educat`）與 X_4'（經 Box-Cox 轉換的 `peop81`）；模型 B（針對戶長已退休之家戶建立的模型）納入的獨變項有：X_1'（經 Box-Cox 轉換的 `water80`）、X_2'（經 Box-Cox 轉換的 `income` 與 X_4'（經 Box-Cox 轉換的 `peop81`）。

1.6.2 迴歸係數之 t 檢定

在假設 $\epsilon_i' \overset{iid}{\sim} N(0, \sigma_i'^2)$ 成立的情況下，我們可以單樣本 t 檢定對每個迴歸係數 β 檢定其是否顯著不為零。所有檢定中我們都令顯著水準為 0.05。

- 研究假設

$$H_0 : \beta_{k.0}' = 0,\ v.s. H_1 : \beta_{k.0}' \neq 0,\ \forall k = 1, 2, 3, 4$$

$$H_0 : \beta_{k.1}' = 0,\ v.s. H_1 : \beta_{k.1}' \neq 0,\ \forall k = 1, 2, 4$$

- 檢定統計量：

$$T_{k.0} = \frac{\hat{\beta}_{k.0}'}{\sqrt{\hat{Var}(\beta_{k.0}')}} \sim t(n - K_0 - 1)$$

$$T_{k.1} = \frac{\hat{\beta}'_{k.1}}{\sqrt{\hat{Var}(\beta'_{k.1})}} \sim t(n - K_1 - 1)$$

- 模型 A（針對戶長非退休之家戶建立的模型）的檢定結果:

	估計值 9	5CI 下界.	95CI 上界 t	檢定統計量	p 值
截距 (beta_0.0)	5.4344	2.4359	8.4330	3.5646	4e-04
water80 (beta_1.0)	0.9154	0.7973	1.0335	15.2447	0e+00
income (beta_2.0)	0.5291	0.2665	0.7918	3.9624	1e-04
educat (beta_3.0)	-0.1273	-0.2651	0.0105	-1.8175	7e-02
peop81 (beta_4.0)	2.9273	2.1904	3.6641	7.8140	0e+00

針對各係數的 t 檢定結果如上表。檢定 $\beta'_{1.0}$、$\beta'_{2.0}$ 與 $\beta'_{4.0}$ 得到的 t 統計量之 p 值都小於顯著水準，因此我們在這三個檢定中可以拒絕 H_0，顯示我們有充分證據可以宣稱對於戶長非退休之家戶，「經由 Box-Cox 轉換的 1980 年用水量」、「經由 Box-Cox 轉換的 1981 年收入」與「經由 Box-Cox 轉換的 1981 年人數」對於「經由 Box-Cox 轉換的 1981 年用水量」都有顯著的預測力。

- 模型 B（針對戶長已退休之家戶建立的模型）的檢定結果:

	估計值.	95CI 下界.	95CI 上界 t	檢定統計量	p 值
截距 (beta_0)	-1.7288	-4.9093	1.4517	-1.0745	0.2844
water80 (beta_1)	1.3188	1.1472	1.4904	15.1918	0.0000
income (beta_2)	0.2322	-0.0933	0.5577	1.4100	0.1607
peop81 (beta_4)	1.4937	0.2711	2.7162	2.4152	0.0170

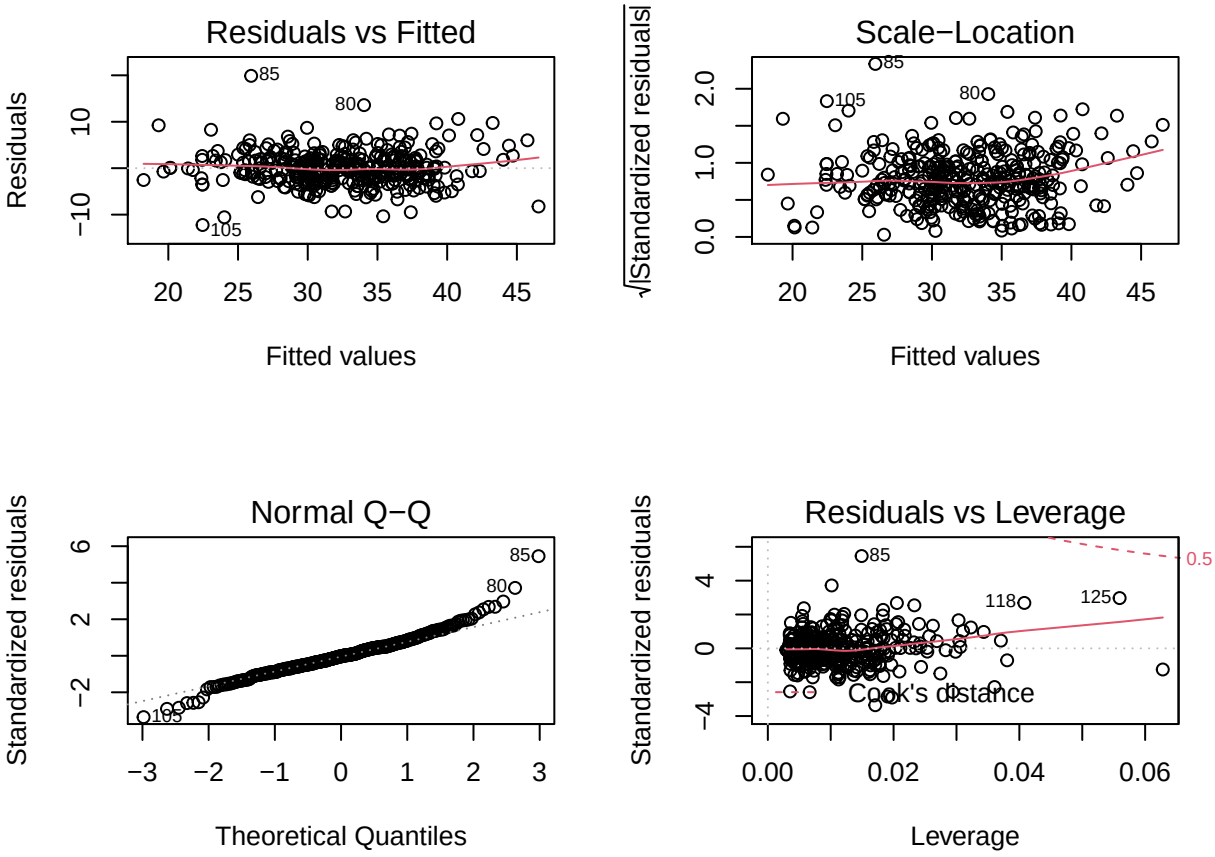
針對各係數的 t 檢定結果如上表。檢定 $\beta'_{1.1}$ 與 $\beta'_{4.1}$ 得到的 t 統計量之 p 值都小於顯著水準，因此我們在這三個檢定中可以拒絕 H_0，顯示我們有充分證據可以宣稱對於戶長已退休之家戶，「經由 Box-Cox 轉換的 1980 年用水量」與「經由 Box-Cox 轉換的 1981 年人數」對於「經由 Box-Cox 轉換的 1981 年用水量」都有顯著的預測力。

1.6.2.1 模型診斷

配適迴歸模型之後，我們應對殘差做檢定來確認線性迴歸模型中的前提假設 $\epsilon'_i \overset{iid}{\sim} N(0, \sigma'^2_i)$ 與 $\epsilon'_j \overset{iid}{\sim} N(0, \sigma'^2_j)$ 是否滿足，包含殘差常態性、殘差變異數同質與殘差獨立性。

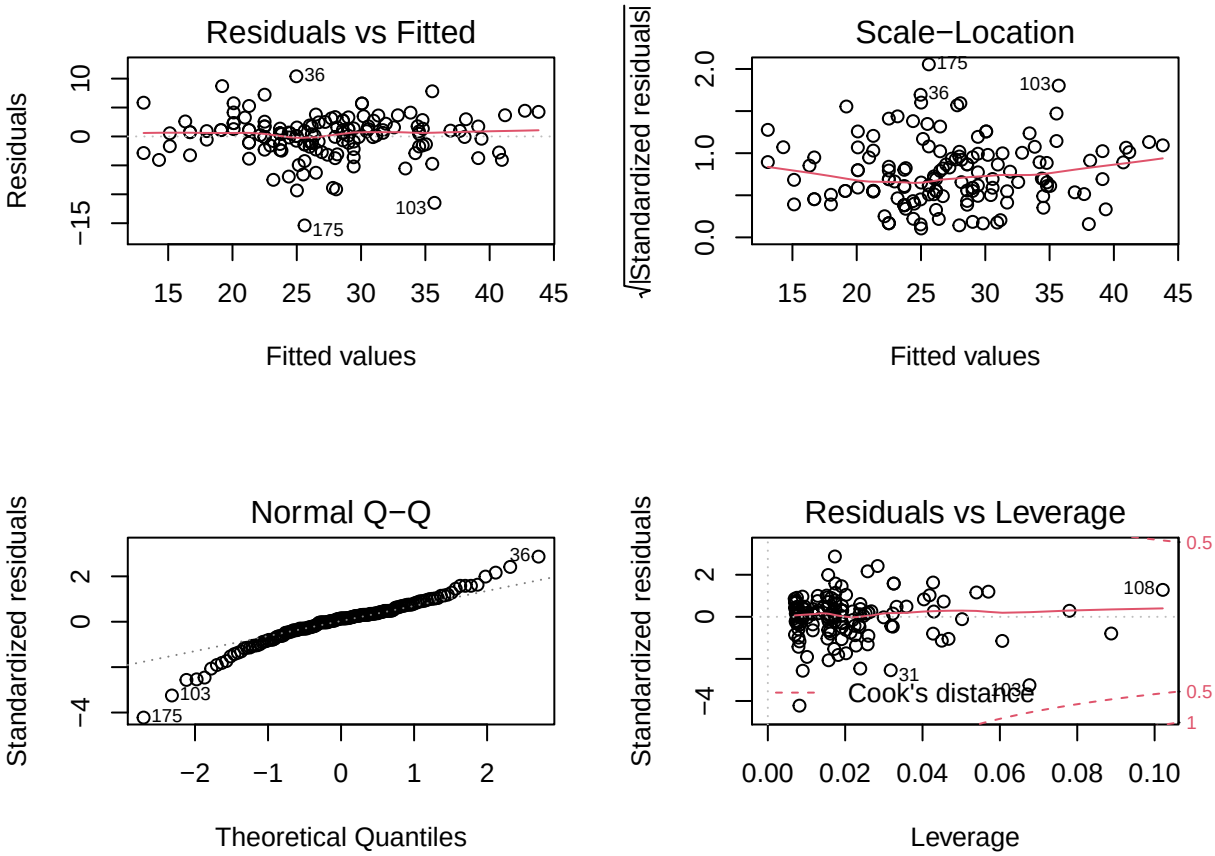
- 針對模型 A（針對戶長非退休之家戶建立的模型）繪製殘差圖以觀察殘差樣態

由左上及右上兩圖可以發現，殘差期望值相相當接近 0，不過殘差變異數似乎有變化趨勢。左下的常態 Q-Q 圖中資料點在值偏小處偏離 45 度線，顯示殘差可能不服從常態分配。右下圖顯示資料中無槓桿點。



- 針對模型 B（針對戶長已退休之家戶建立的模型）繪製殘差圖以觀察殘差樣態

由左上及右上兩圖可以發現，殘差期望值相相當接近 0，不過殘差變異數似乎有變化趨勢。左下的常態 Q-Q 圖中資料點在值偏小及偏大處非常明顯偏離 45 度線，顯示殘差可能不服從常態分配。右下圖顯示資料中無槓桿點。



- 殘差常態性檢驗

我們以 Shapiro-Wilk 檢定檢驗模型殘差是否為常態分配，令顯著水準為 0.05，其假設如下:

$$H_0 : \epsilon_i' \sim ND \ v.s. \ H_1 : \epsilon_i' \ does \ not \sim ND$$

$$H_0 : \epsilon_j' \sim ND \ v.s. \ H_1 : \epsilon_j' \ does \ not \sim ND$$

模型 A 的檢定結果：檢定統計量 W 為 0.9664，p 值為 0，小於顯著水準，因此我們拒絕 H_0，表示我們有足夠的證據顯示殘差不服從常態分配，診斷未通過。模型 B 的檢定結果：檢定統計量 W 為 0.9482，p 值為 0，小於顯著水準，因此我們拒絕 H_0，表示我們有足夠的證據顯示殘差不服從常態分配，診斷未通過。

- 以 Brown-Forsythe 檢定檢驗殘差變異同質性（令顯著水準為 0.05）

$$\begin{cases} H_0 : \sigma_i'^2 = \sigma_{i'}'^2 \ \forall i \neq i' \\ H_1 : Not \ H_0 \end{cases}$$

$$\begin{cases} H_0 : \sigma_j'^2 = \sigma_{j'}'^2 \ \forall j \neq j' \\ H_1 : Not \ H_0 \end{cases}$$

模型 A 的檢定結果：檢定統計量 BF 為 45.6035，p 值為 0，小於顯著水準，因此我們拒絕 H_0，表示我們有足夠的證據顯示殘差不具備變異數同質性，診斷未通過。模型 B 的檢定結果：檢定統計量 BF 為 29.4046，p 值為 0，小於顯著水準，因此我們拒絕 H_0，表示我們有足夠的證據顯示殘差不具備變異數同質性，診斷未通過。

- 以 Durbin-Waston 檢定檢驗殘差獨立性（令顯著水準為 0.05）

$$H_0 : \epsilon_i' \textit{ are independent. v.s. } H_1 : \epsilon_i' \textit{ are not independent.}$$

$$H_0 : \epsilon_j' \textit{ are independent. v.s. } H_1 : \epsilon_j' \textit{ are not independent.}$$

模型 A 的檢定結果：檢定統計量 DW 為 2.0484，p 值為 0.6706，不小於顯著水準，因此我們不拒絕 H_0，也就是說我們沒有充分證據支持殘差不具備獨立性，通過獨立假設。模型 B 的檢定結果：檢定統計量 DW 為 1.8611，p 值為 0.1972，不小於顯著水準，因此我們不拒絕 H_0，也就是說我們沒有充分證據支持殘差不具備獨立性，通過獨立假設。

1.7 結論

我們建立若干個迴歸模型以嘗試預測家戶用水量及解釋其變異性，模型的解釋力可由經自由度調整後的決定係數 R^2_{adj} 敘述，數學式如下：

$$R^2 = 1 - \frac{SS_E/(n-K)}{SS_T/(n-1)}$$

其中 SS_E 為模型無法解釋之變異，SS_T 為總變異，n 為樣本大小，K 為最後納入模型的獨變項個數。

我們建立的其中一個模型為以 Box-Cox 轉換的家戶相關特徵預測「經由 Box-Cox 轉換的 1981 年家戶用水量」的複迴歸模型：

$$\hat{y}' = \hat{\beta}_0' + \hat{\beta}_1' x_1 + \hat{\beta}_2' x_2 + \hat{\beta}_4' x_4$$

其中 $\hat{\beta}_0 = 2.5799$，$\hat{\beta}_1 = 1.0398$，$\hat{\beta}_2 = 0.3385$，$\hat{\beta}_4 = 2.4669$。此迴歸模型之 $R^2_{adj} = 0.7101$，顯示大部分的變異可以被此模型解釋。由於我們採用的 Box-Cox 轉換是單調函數，因此轉換後效果方向仍相同。由迴歸係數可知，在固定其餘獨變項下，家戶在 1980 年經 Box-Cox 轉換的用水量每多 1 立方英尺，其在 1981 年經由 Box-Cox 轉換的用水量便多 1.0398 立方英尺（1980 年用水量上升，1981 年用水量上升）；家戶在 1981 年的收入每多 1000 元，其在 1981 年經由 Box-Cox 轉換的用水量便多 0.3385 立方英尺（1981 年收入上升，1981 年用水量上升）；家戶在 1981 年的人數每多 1 人，其在 1981 年經由 Box-Cox 轉換的用水量便多 2.4669 立方英尺（1981 年家戶人數上升，1981 年用水量上升）。

我們也將家戶依據其戶長是否退休分層，分別針對戶長非退休以及戶長已退休的家戶各建立一個預測「經由 Box-Cox 轉換的 1981 年家戶用水量」的複迴歸模型：

- 針對戶長非退休之家戶：$\hat{y}' = \hat{\beta}'_{0.0} + \hat{\beta}'_{1.0}x_1 + \hat{\beta}'_{2.0}x_2 + \hat{\beta}'_{4.0}x_4$。

其中 $\hat{\beta}'_{0.0} = 4.0922$，$\hat{\beta}'_{1.0} = 0.9148$，$\hat{\beta}'_{2.0} = 0.4519$，$\hat{\beta}'_{4.0} = 2.9476$ 此迴歸模型之 $R^2_{adj} = 0.6396$，顯示超過一半的變異可以被此模型解釋。由於我們採用的 Box-Cox 轉換是單調函數，因此轉換後效果方向仍相同。由迴歸係數可知，在固定其餘獨變項下，戶長非退休的家戶在 1980 年經 Box-Cox 轉換的用水量每多 1 立方英尺，其在 1981 年經由 Box-Cox 轉換的用水量便多 0.9148 立方英尺（1980 年用水量上升，1981 年用水量變上升）；在 1981 年的收入每多 1000 元，其在 1981 年經由 Box-Cox 轉換的用水量便多 0.4519 立方英尺（1981 年收入上升，1981 年用水量上升）；在 1981 年的人數每多 1 人，其在 1981 年經由 Box-Cox 轉換的用水量便多 2.9476 立方英尺（1981 年家戶人數上升，1981 年用水量上升）。

- 針對戶長已退休之家戶：$\hat{y}' = \hat{\beta}'_{0.1} + \hat{\beta}'_{1.1}x_2 + \hat{\beta}'_{4.1}x_4$

其中 $\hat{\beta}'_{0.1}$ = -1.2809，$\hat{\beta}'_{1.1} = 1.3434$，$\hat{\beta}'_{4.0} = 1.6271$ 此迴歸模型之 $R^2_{adj} = 0.7444$，顯示超過一半的變異可以被此模型解釋。由於我們採用的 Box-Cox 轉換是單調函數，因此轉換後效果方向仍相同。由迴歸係數可知，在固定其餘獨變項下，戶長已退休的家戶在 1980 年經 Box-Cox 轉換的用水量每多 1 立方英尺，其在 1981 年經由 Box-Cox 轉換的用水量便多 1.3434 立方英尺（1980 年用水量上升，1981 年用水量上升）；在 1981 年的人數每多 1 人，其在 1981 年經由 Box-Cox 轉換的用水量便多 1.6271 立方英尺（1981 年家戶人數上升，1981 年用水量上升升）。

然而需要注意，這三個模型在殘差假設檢驗中，都僅通過了殘差獨立性檢定，並未通過常態性與變異數同質性的檢定。

1.8 建議

在模型篩選的過程中，我們嘗試了許多資料轉換方式與模型，不管是考慮交互作用項或考慮多項式迴歸（例如因直觀上家戶人數對用水量的效果應明顯，而新增「家戶人數的平方」，以強調人數的效果）等方法，都無法使得模型通過所有假設檢驗，也無明顯提高對資料的解釋能力。

這顯示這份資料可能不適合使用線性迴歸分析，若諮詢者主要目的為得到精準的預測模型，建議可以改採其他前提假設不同於線性迴歸模型的機器學習模型，若諮詢者主要目的為解釋各獨變項如果對依變項產生效果，則可能還有沒有收集到的其他變項與依變項有關，建議可以多增加相關的變項以做分析，例如：家戶用水習慣（習慣淋浴或泡澡、是否開伙、洗衣習慣、是否裝設私人游泳池）、家戶房屋大小、家戶浴室個數等等。此外也可以確認資料收集方式之適當性，例如家戶人數若是以戶籍資料取得，則可能與家戶實際居住人數有出入。

而就以目前的資料及分析結果而言，若要制定節水措施，建議可依據戶長是否退休否進行不同政策的擬定。

2 Exercise 12.6

2.1 研究問題

舉辦於鹽城湖市的 2002 年冬季奧運中的溜冰賽事因計分引發爭議。該溜冰賽由 9 位評審評分，主要分為「技術」與「表現」兩個面向。研究問題為分析評審在這兩個面向上的評分是否具有一致性，並做進一步的分析。

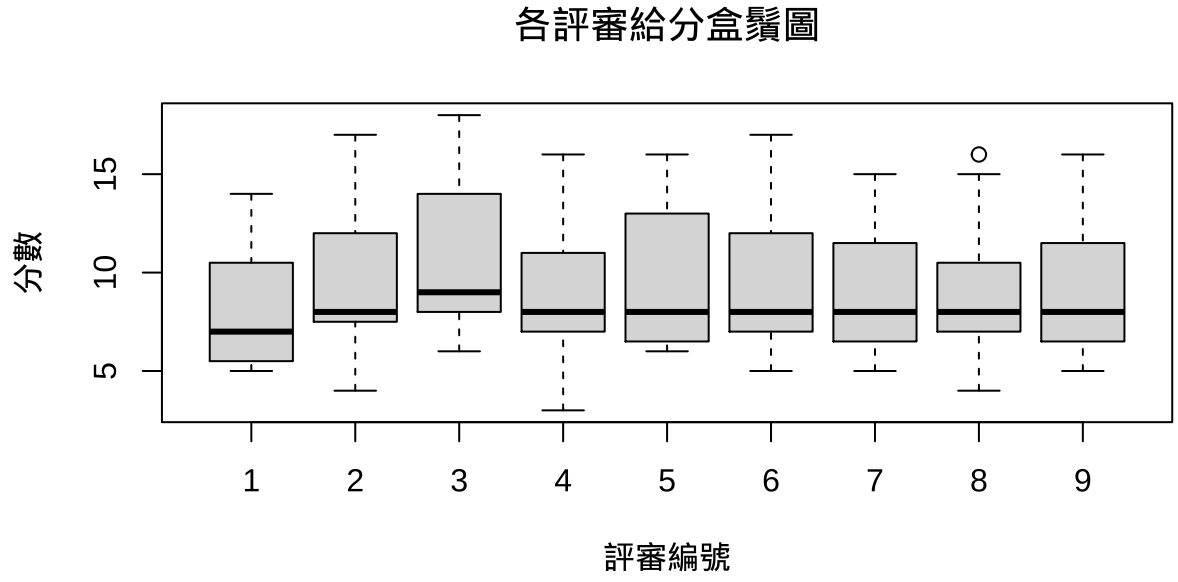
2.2 資料介紹

資料集有 45 列、4 個變項，每一列為一位溜冰選手獲得一位評審評分的紀錄，變項說明如下:

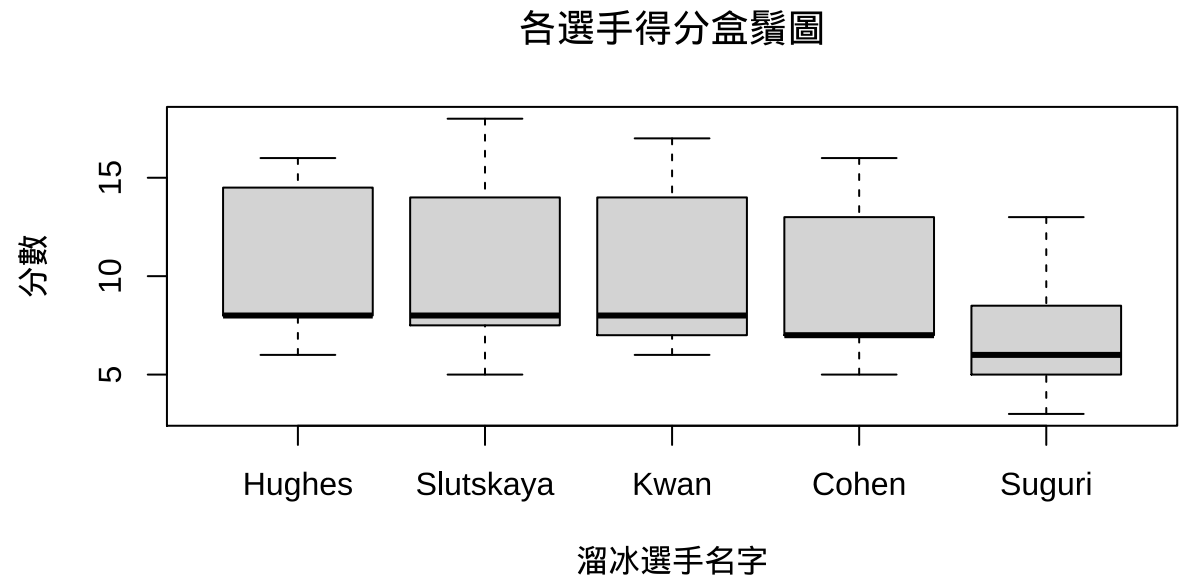
- `Technique`: 評審給予選手的技術分數
- `Presentation`: 評審給予選手的表現分數
- `Judge`: 評審編號，共有 9 位評審
- `Skater`: 溜冰選手，有 Hughes、Slutskaya、Kwan、Cohen 與 Suguri 等 5 位
- 新增變項 `Total.score`: `Technique` 與 `Presentation` 之和，為評審給予選手的總分

2.3 資料探索

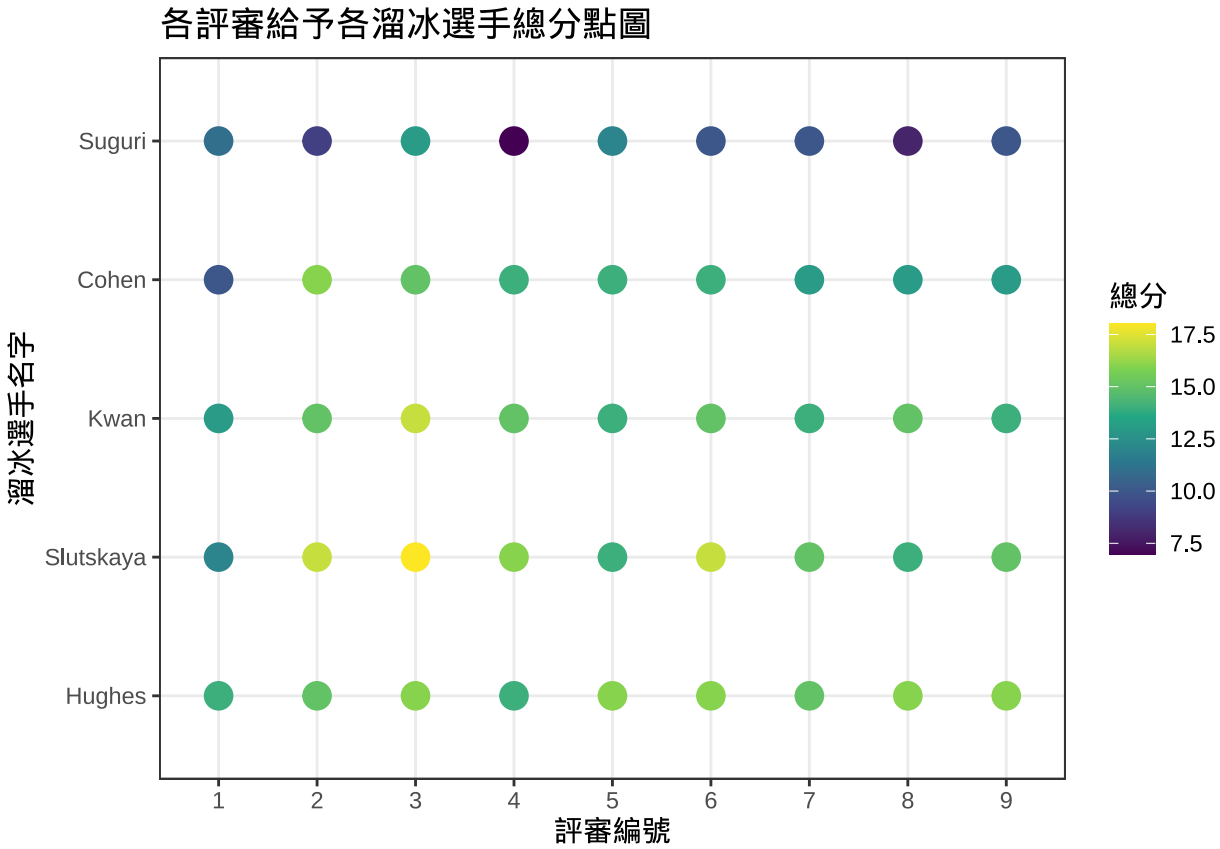
我們先以盒鬚圖來大致看各位評審給分的分佈。由圖可看出評審 1 給分偏低，而評審 3 給分偏高，評審 2 與評審 4 總分給分變異範圍較大。



我們接著以盒鬚圖來大致看各選手得分的分佈。由圖可看出 Suguri 得分偏低，而 Kwan 與 Cohen 都有獲得較極端的分數（某幾位評審給他們的評分明顯低或高於其他評審給他們的分數）。



接著，我們將評審、選手與總分三個變項都繪製在一起，來觀察交互作用是否可能存在。由於兩獨變量都是名義尺度的類別變項，我們不繪製折線圖，而繪製以顏色軸表徵依變項的點圖，圖中各點顏色表示評審給予選手的總分高低，顏色越靠近黃色者表示總分越高，而顏色越靠近紫色者則表示總分越低。可以發現，評審 2 與評審 4 給分不低，其給予 Suguri 選手的總分卻特別低，在 Suguri 選手獲得的 9 個總分中也屬於偏低者，這顯示評審與選手之間對於總分的交互作用可能存在。不過因為各交互作用組合都只有一個資料點，無法考慮交互作用



2.4 變異數分析

我們以不考慮交互作用項、隨機區集化設計的的二因子變異數分析（two-way analysis of variance with random block design, two-way ANVOA RBD）回答研究問題。

2.4.1 定義模型

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma_{ij}^2)$$

$$\sum_{i=1}^{9} \alpha_i = \sum_{j=1}^{5} \beta_i = 0; \quad \begin{cases} i = 1, ..., 9 \\ j = 1, ..., 5 \end{cases}$$

其中 Y_{ij} 為評審 i 給予選手 j 的總分，μ 為所有評審給予所有選手總分之總平均，α_i 為評審 i 對於總分的效果（處理效果，treatment effect），β_i 為選手 j 對於總分的效果（集區效果，block effect），ϵ_{ij} 為評審 i 與選手 j 之誤差項。

2.4.2 檢查分析的前提假設

1. 獨變項之間彼此獨立

評審與評審之間評分互相獨立，選手得分之間也互相獨立。符合。

2. 變異數同質性假設

我們對兩獨變項進行 Levene 檢定，顯著水準皆設定為 0.05，檢定假設如下：

$$\begin{cases} H_0 : \sigma_{1j}^2 = \sigma_{2j}^2 = ... = \sigma_{9j}^2 \ v.s. \ H_1 : Not \ H_0 \\ H_0 : \sigma_{i1}^2 = \sigma_{i2}^2 = ... = \sigma_{i5}^2 \ v.s. \ H_1 : Not \ H_0 \end{cases}$$

因子	自由度 1	自由度 2	F 檢定統計量	p 值
Judge	8	36	0.2240	0.9842
Skater	4	40	0.8857	0.4812

檢定結果如上表，兩個檢定的 p 值接不小於顯著水準，因此我們都不拒絕 H_0，表示我們無足夠證據宣稱變異數同質性不存在，通過此前提假設。

3. 殘差彼此獨立且服從相同的常態分配：待模型配適後再進行診斷。

2.4.3 研究假設

$$H_0 : \alpha_1 = \alpha_2 = ...\alpha_9 \ v.s. \ H_1 : Not \ H_0$$

$$H_0 : \beta_1 = \beta_2 = ...\beta_5 \ v.s. \ H_1 : Not \ H_0$$

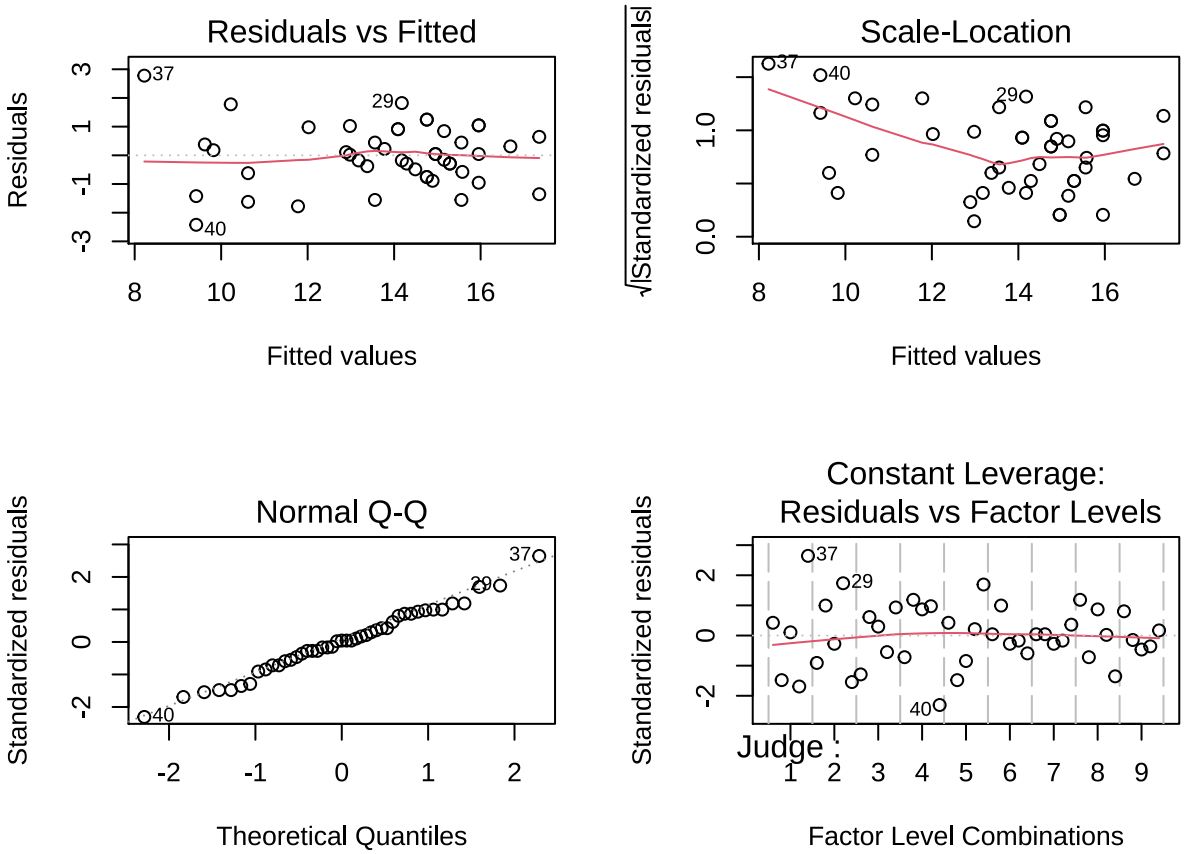
2.4.4 檢定結果

變異來源	自由度	平方和	均方	F 檢定統計量	p 值
因子（評審）	8	44.58	5.57250	3.592	0.004
集區（選手）	4	179.56	44.89000	28.935	3.18e-10
殘差	32	49.64	1.55125		

令顯著水準為 0.05。ANOVA 結果如上表。根據 ANOVA，我們在兩組研究假設中都拒絕虛無假設，表示因子效果與集區效果都顯著，也就是說，不同評審給予的總分並不完全相同，不同選手得到的分數也並不完全相同。

2.4.5 檢查殘差假設

我們先繪製殘差圖觀察殘差分佈情形。由左上圖可知殘差期望值接近 0，左下的常態 Q-Q 圖中資料點大多落在 45 度線上，顯示殘差可能服從常態分配，右下圖則顯示殘差變異性不大，變異數同質性可能存在。



我們接著進行檢定以確認這些條件成立。所有檢定的顯著水準皆設為 0.05。

1. 殘差常態檢定

我們以 Shapiro-Wilk 檢定檢驗模型殘差是否為常態分配，檢定假設如下:

$$H_0 : \epsilon_{ij} \sim ND \ v.s. \ H_1 : \epsilon_{ij} \ does \ not \sim ND$$

檢定結果：檢定統計量 W 為 0.9894，p 值為 0.9507，不小於顯著水準，因此我們 1 不拒絕 H_0，表示我們沒有足夠的證據顯示殘差不服從常態分配，診斷通過。

2. 殘差變異同質性檢驗

在常態分佈下，以 Barlett 檢定變異同數同質性的統計檢定力較高，因此我們以其檢驗殘差變異數同質性，檢定假設與檢定結果如下:

- 檢定因子間殘差變異數同質性：$H_0 : \sigma_{ij}^2 = \sigma_{i'j}^2 \ \forall i \neq i' \quad v.s. \quad H_1 : Not \ H_0$

檢定統計量 $Barlett's \ k^2$ 為 5.6072，其 p 值為 0.6911，不小於顯著水準，因此我們不拒絕 H_0，意味著沒有足夠的證據證明變異數同質性不存在。

- 檢定集區間殘差變異數同質性：$H_0 : \sigma_{ij}^2 = \sigma_{ij'}^2 \ \forall j \neq j' \quad v.s. \quad H_1 : Not \ H_0$

檢定統計量 $Barlett's \ k^2$ 為 6.2571，其 p 值為 0.1808，不小於顯著水準，因此我們不拒絕 H_0，意味著沒有足夠的證據證明變異數同質性不存在。

- 以 Brown-Forsythe Test 檢定整體殘差變異數同質性:

$H_0 : \sigma_{ij}^2 = \sigma_{i'j'}^2 \ \forall i \neq i', \ j \neq j' \quad v.s. \quad H_1 : Not \ H_0$
檢定結果：檢定統計量 BF 為 4.5058，p 值為 0.0397，小於顯著水準，因此我們拒絕 H_0，表示我們有足夠的證據顯示殘差不具備變異數同質性。殘差變異數同質性之診斷未完全通過。

3. 以 Durbin-Waston 檢定檢驗殘差獨立性

$$H_0 : \epsilon_i' \ are \ independent. \ v.s. \ H_1 : \epsilon_i' \ are \ not \ independent.$$

檢定結果：檢定統計量 DW 為 2.9218，p 值為 0.993，不小於顯著水準，因此我們不拒絕 H_0，也就是說我們沒有充分證據支持殘差不具備獨立性，通過獨立假設。

2.5 結論

藉由資料視覺化與隨機集區化設計的二因子變異數分析可以得知 9 位評審對溜冰選手的評分不具一致性。且另外也可得知 5 位選手間的表現有所差異。不過，需要注意，因為我們建立的模型無法通過診斷，而且 ANOVA RBD 無法刪除極端值，因此以該檢定推論可能會有問題。

2.6 建議

因為我們建立的模型無法通過診斷，而且 ANOVA RBD 無法刪除極端值，因此以該檢定推論可能會有問題，建議可改用前提假設較不嚴格的無母數分析，或是增加資料筆數，例如增加更多選手的紀錄，以提升分析可信度。