統計諮詢 - 作業 3

國立成功大學統計學系暨數據科學研究所

廖傑恩（RE6094028）

2021-03-26

1 Exercise 9.3

1.1 問題敘述

諮詢者想以賓州伊利在 1970 年代的房屋的特徵對其房價進行預測。

1.2 資料介紹

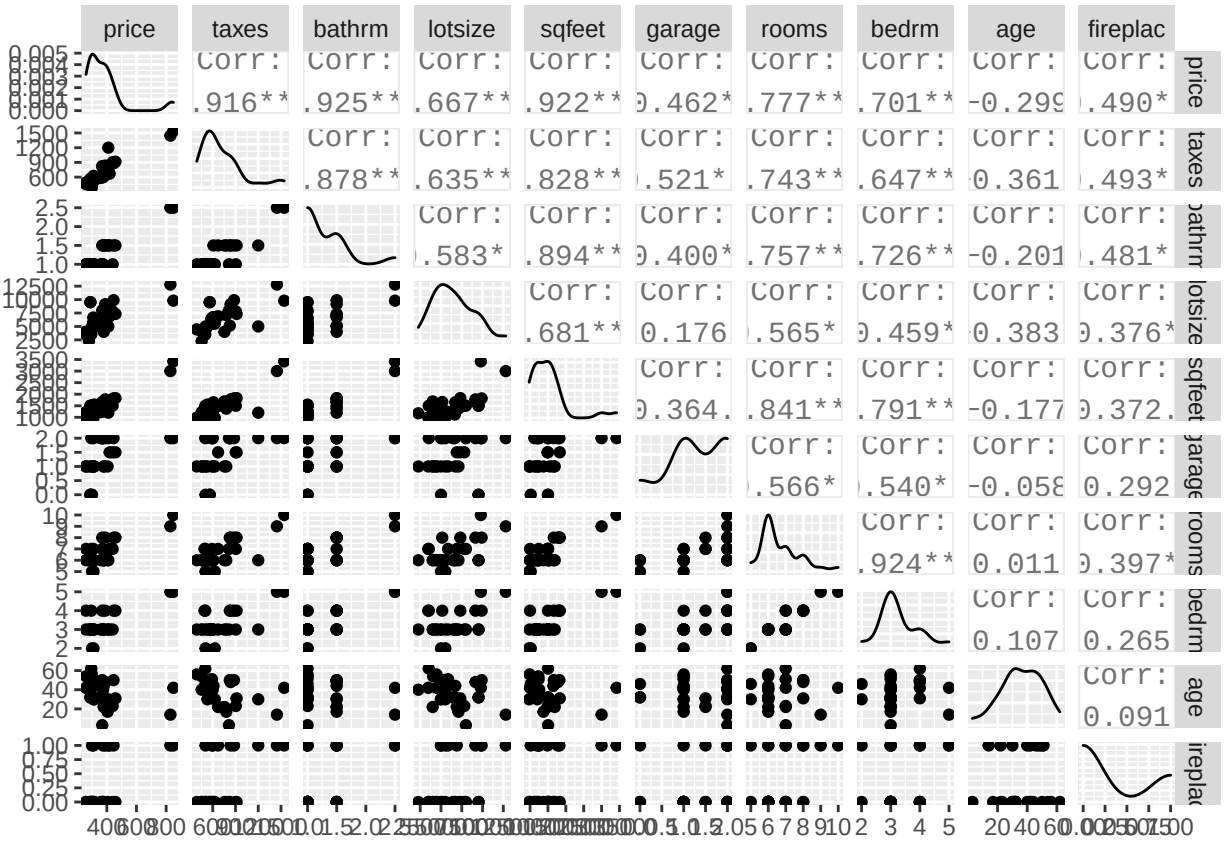
資料是由 Narula 與 Wellington（1977）提供，共有 28 列、12 個變項，每一列為一棟賓州伊利在 1970 年代房屋售價的資料。變項包含了房價與其可能的預測因子（predictor），說明如下：

1. `price`: 價格（單位：100 元）
2. `taxes`: 稅（單位：元）
3. `bathrm`: 浴室數量
4. `lotsize`: 地皮尺寸（單位：平方英呎）
5. `sqfeet`: 居住空間大小（單位：平方英呎）
6. `garage`: 車庫能停放的汽車數量
7. `rooms`: 房間數量
8. `bedrm`: 寢室數量
9. `age`: 屋齡（單位：年）
10. `type`: 房屋建材型態，有：磚（brick）、磚和木框（brick and frame）、鋁和木框（aluminum and frame）、木框（frame）4 種
11. `style`: 類型，有 2 層、1.5 層和牧場住宅（ranch）3 種
12. `fireplac`: 壁爐數量

1.2.1 資料探索

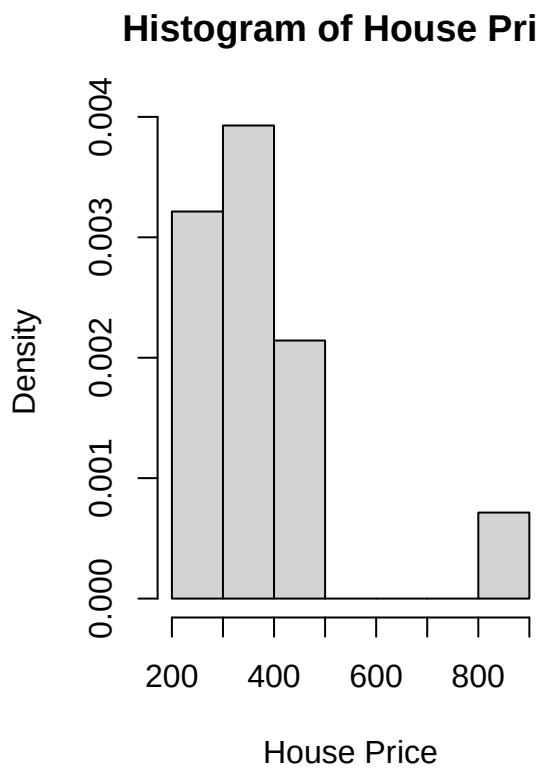
- 針對數值型變項繪製兩兩散佈圖矩陣以快速觀察這些變項間的關聯

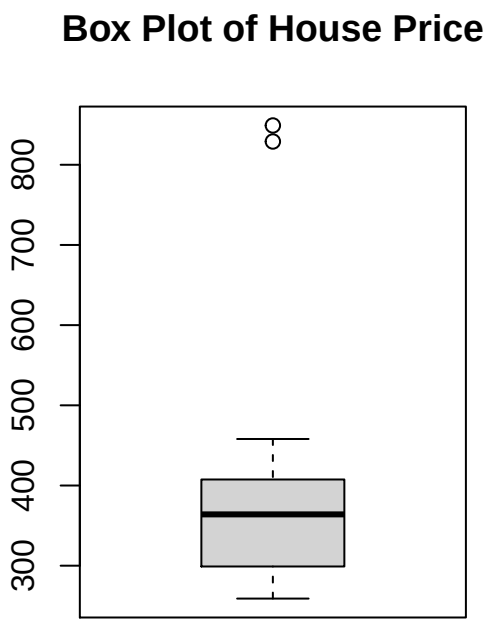
非數值型變項而沒有納入繪製的變項：`type` 與 `style`



由此矩陣可以發現，大部分數值型變項兩兩之間似乎有所關聯，若直接進行線性迴歸的配適，可能會產生共線性的問題因此必須要先檢查共線性。此外，在房價這個變項上好像有若干特別的值，我們再針對房價繪圖以了解其分布狀態。

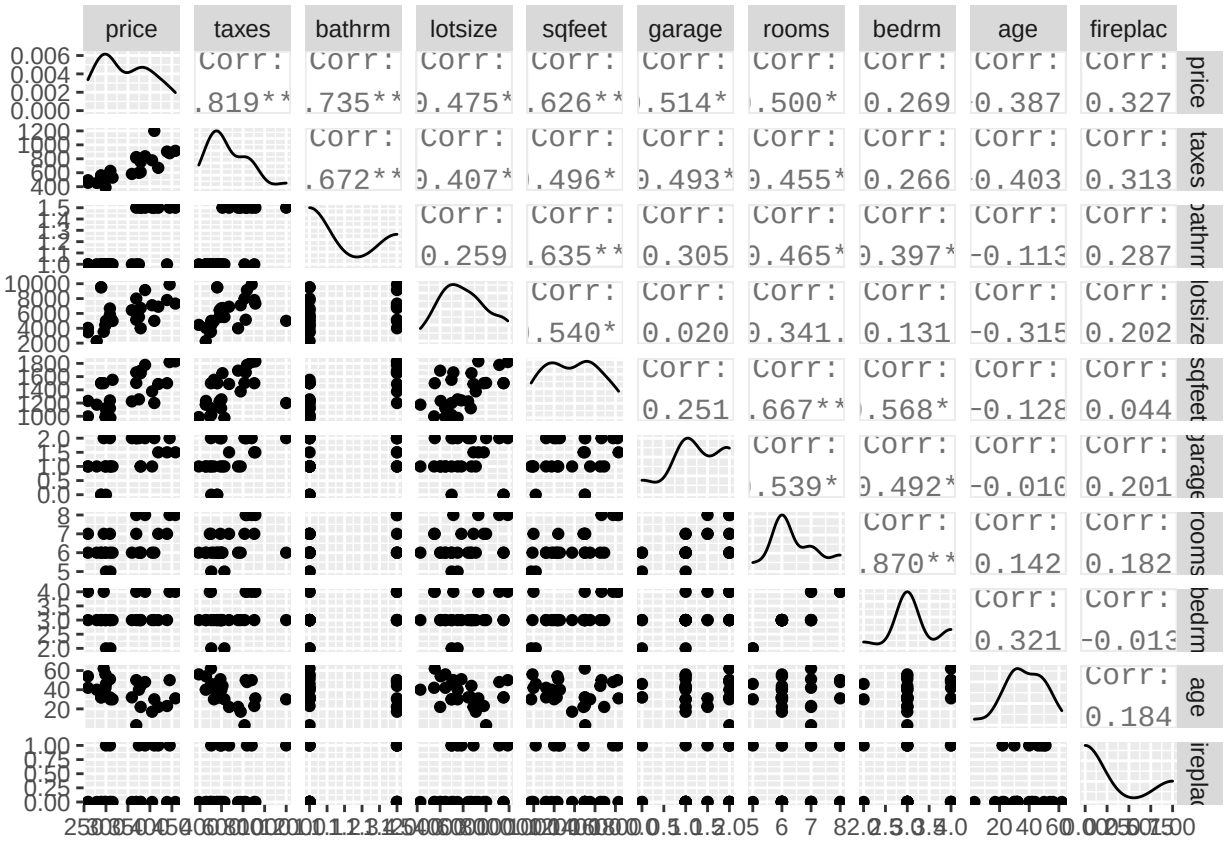
- 針對目標變項「房價」繪圖以了解其分布狀態





由直方圖（histogram）可以發現，有少部分房屋的房價特別高。盒鬚圖（box plot）則顯示兩有 2 間房屋價格的值特別大。由於樣本數不多，這樣極端的值可能就會對之後的分析造成影響，顯示我們有可能需要將資料進行轉換，或是其他處理。我們先嘗試看看直接將其移除，以剩下的資料點來繪製散步圖矩陣。

- 移除極端值，針對數值型變項繪製兩兩散佈圖矩陣以快速觀察這些變項間的關聯



由此矩陣可以發現，即使移除 2 筆房價特別大的資料，大部分數值型變項兩兩之間似乎仍有所關聯，若直接進行線性迴歸的配適，可能會產生共線性的問題因此必須要先檢查共線性。

1.2.2 建立迴歸模型

1.2.2.1 變數與模型定義

- 定義變數

1. 令 Y 為 `price`（房價）
2. 令 X_{taxes} 為 `taxes`（稅金）
3. 令 X_{bathrm} 為 `bathrm`（浴室數量）
4. 令 $X_{lotsize}$ 為 `lotsize`（地皮尺寸）
5. 令 X_{sqfeet} 為 `sqfeet`（居住空間大小）
6. 令 X_{garage} 為 `garage`（車庫能停放的汽車數量）
7. 令 X_{room} 為 `rooms`（房間數量）
8. 令 X_{bedrm} 為 `bedrm`（寢室數量）
9. 令 X_{age} 為 `age`（屋齡）
10. 令 $X_{type.bf}$ 為 `type`（房屋建材型態）是否為「磚和木框（brick and frame）」，是 =1，否 =0
11. 令 $X_{type.af}$ 為 `type`(房屋建材型態)是否為「鋁和木框(aluminum and frame)」，是 =1，否 =0

12. 令 $X_{type.f}$ 為 `type`（房屋建材型態）是否為「木框（frame）」，是 =1，否 =0
13. 令 $X_{style.1.5}$ 為 `style`（房屋類型）是否為「1.5 層」，是 =1，否 =0
14. 令 $X_{style.r}$ 為 `style`（房屋類型）是否為「牧場住宅（ranch）」，是 =1，否 =0
15. 令 $X_{fireplac}$ 為 `fireplac`（壁爐數量）

- 模型假設

$$y_i = \beta_0 + \beta_1 x_{i,taxes} + \beta_2 x_{i,bathrm} + \beta_3 x_{i,lotsize} + \beta_4 x_{i,sqfeet} + \beta_5 x_{i,garage}$$
$$+\beta_6 x_{i,room} + \beta_7 x_{i,bedrm} + \beta_8 x_{i,age} + \beta_9 x_{i,type.bf} + \beta_{10} x_{i,type.af}$$
$$+\beta_{11} x_{i,type.f} + \beta_{12} x_{i,style.1.5} + \beta_{13} x_{i,style.r} + \beta_{14} x_{i,fireplac} + \epsilon$$

其中 $\epsilon \overset{iid}{\sim} N(0, \sigma^2)$
亦可以向量與矩陣的方式表達以求簡潔:

$$\tilde{Y} = \tilde{\beta}_0 + \tilde{\beta}\mathbf{X}^T + \tilde{\epsilon},\ \tilde{\epsilon} \overset{iid}{\sim} N(\tilde{0}, \tilde{\sigma^2})$$

其中，$\tilde{Y}$ 為所有觀測值之房價形成的 28 ×1 的向量；$\tilde{\beta}_0$ 為 28; 相同的截距 β_0 形成的 28 ×1 的向量；$\tilde{\beta}$ 為除了截距之外的其他迴歸係數 $\beta_1, ..., \beta_{14}$ 形成的 1 × 13 的向量；$\mathbf{X}$ 則為 28×14 的矩陣，其列為所有觀測值、行為所有變項；$\tilde{\epsilon}$、$\tilde{0}$ 與 $\tilde{\sigma^2}$ 皆為 28×1 的向量，分別為殘差向量、零向量與變異數向量。

1.2.2.2 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膨脹係數（VIF）值:

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

其中 R_p^2 為以獨變項 p 建立的迴歸模型的決定係數。

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。最後選出之不具共線性之獨變項有: $X_{lotsize}$、X_{garage}、X_{rooms}、X_{age}、$X_{type.bf}$、$X_{type.af}$、$X_{type.f}$、$X_{style.1.5}$、$X_{style.ranch}$ 與 $X_{fireplac}$ 等 10 個

1.2.2.3 逐步迴歸分析（stepwise selection）

我們使用逐步選擇法中的向後選擇法（back selection），先將所有數值型變項都納入模型，再逐步移除對模型貢獻程度最低的變項，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：

步驟	步驟完成後的模型 AIC
納入所有變項的原始模型	247.9691
移除 garage	246.4701
移除 style.ranch	246.1375
移除 type.af	245.5462
移除 lotsize	245.0048
移除 style.1.5	244.6438

經過變數挑選後的模型為：

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_6 x_{i,rooms} + \hat{\beta}_8 x_{i,age} + \hat{\beta}_9 x_{i,type.bf} + \hat{\beta}_{11} x_{i,type.f} + \hat{\beta}_{14} x_{fireplac}$$

其中 $\hat{\beta}_0$ = -66.6624，$\hat{\beta}_6$ = 80.7649，$\hat{\beta}_8$ = -3.8603，$\hat{\beta}_9$ = 52.8837，$\hat{\beta}_{11}$ = 60.8358，$\hat{\beta}_{14}$ = 70.032。

各個獨變項的係數估計值中，只有屋齡 `age`（$\hat{\beta}_8$ =）為負，表示隨著屋齡上升，房價會較低，而其他變項都與房價呈現正相關，房間數量（`rooms`）、房屋建材型態是否為「磚和木框」（`type.bf`）、房屋建材型態是否為「木框」（`type.f`）與壁爐數量（`fireplac`）。

1.2.2.3.1 整體模型之 F 檢定

我們對以上的迴歸估計式進行顯著性為 0.05 之整體 F 檢定（Overall F test），檢定該迴歸模型是否對於房價具有解釋效力。

- 研究假設為：$H_0 : \beta_6 = \beta_8 = \beta_9 = \beta_{11} = \beta_{14} = 0$ $v.s.$ $H_1 : Not\ H_0$
- 檢定統計量：

$$F = \frac{MS_R}{MS_E} \sim F(k,\ n-k-1);\ \begin{cases} MS_R = \frac{1}{k}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ MS_E = \frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \end{cases}$$

其中 n 為樣本大小（i.e., 28），k 為最後納入模型的獨變項個數（i.e., 5）。此檢定統計量服從自由度為 k 與 $n-k-1$ 的 F 分配。

- 檢定結果：檢定統計量 F 為 16.55，其 p 值為 0，小於顯著水準，因此我們拒絕 H_0，表示此迴歸模型對於依變項房價具有解釋力。

1.2.2.3.2 迴歸係數之 t 檢定

在假設 $\epsilon \overset{iid}{\sim} N(0, \sigma^2)$ 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。

- 研究假設: $H_0 : \beta_j = 0,\ v.s. \beta_j \neq 0,\ \forall j = 6, 8, 9, 11, 14$
- 檢定統計量:

$$T_j = \frac{\hat{\beta}_j}{\sqrt{\hat{Var}(\beta_j)}} \sim t(n-k-1)$$

- 檢定結果:

	估計值 9	5%CI 下界 9	5%CI 上界 t	檢定統計量	p 值
截距 (beta_0)	-66.6624	-259.2844	125.9596	-0.7177	0.4805
rooms (beta_6)	80.7649	53.5731	107.9568	6.1598	0.0000
age (beta_8)	-3.8603	-6.0043	-1.7164	-3.7341	0.0012
type.bf (beta_9)	52.8837	-20.7961	126.5635	1.4885	0.1508
type.f (beta_11)	60.8358	-9.5870	131.2586	1.7915	0.0870
fireplac (beta_14)	70.0320	3.7398	136.3241	2.1909	0.0393

針對各係數的 t 檢定結果如上表。其中檢定 β_6、β_8 與 β_{14} 得到的 t 統計量之 p 值小於顯著水準，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 $\beta_6 = 0$、$\beta_8 = 0$ 與 $\beta_{14} = 0$ 都是錯的，也就是說，房間數量、屋齡與壁爐數量對於房價都有顯著的預測力。

1.2.2.3.3 殘差模型診斷

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 0.9823，p 值為 0.9006，不小於顯著水準，因此我們不拒絕 H_0，表示我們沒有足夠的證據顯示殘差不服從常態分配，通過診斷。

1.2.2.3.4 配適結果

我們建立一個預測房價的複迴歸模型:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_6 x_{i,rooms} + \hat{\beta}_8 x_{i,age} + \hat{\beta}_9 x_{i,type.bf} + \hat{\beta}_{11} x_{i,type.f} + \hat{\beta}_{14} x_{fireplac}$$

其中 $\hat{\beta}_0$ = -66.6624，$\hat{\beta}_6$ = 80.7649，$\hat{\beta}_8$ = -3.8603，$\hat{\beta}_9$ = 52.8837，$\hat{\beta}_{11}$ = 60.8358，$\hat{\beta}_{14}$ = 70.032。

此迴歸模型經由自由度調整後的決定係數 R^2_{adj} 為 0.7422，顯示大部分的變異可以被此模型解釋。由係數可知，在固定其餘獨變項下，房屋每多一間房間，房價便提升 8076.49 美元；屋齡每多一年，房價便下跌 386.03 美元；房屋每多一個壁爐，房價便提升 7003.2。

1.2.2.4 剔除極端值後檢驗模型解釋力

在先前資料探索的部分我們以發現原始資料中有 2 筆資料房價明顯高於其他房屋，因此嘗試將這 2 筆資料剔除，對剩餘資料進行整體 F 檢定，檢查該迴歸模型是否依舊適配。

- 研究假設為：$H_0 : \beta_6 = \beta_8 = \beta_9 = \beta_{11} = \beta_{14} = 0 \ v.s. \ H_1 : Not \ H_0$
- 檢定統計量：

$$F = \frac{MS_R}{MS_E} \sim F(k,\ n-k-1);\ \begin{cases} MS_R = \frac{1}{k}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ MS_E = \frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \end{cases}$$

其中 n 為樣本大小（i.e., 26），k 為最後納入模型的獨變項個數（i.e., 5）。此檢定統計量服從自由度為 k 與 $n-k-1$ 的 F 分配。

- 檢定結果：

檢定統計量 F 為 6.38，其 p 值為 0.0011，小於顯著水準，因此我們拒絕 H_0，表示此迴歸模型對於依變項房價仍具有解釋力。然而，R^2_{adj} 由 0.7422 大幅下滑至 0.5183，表示模型解釋力大幅下降。

1.2.2.5 創建新變項再次進行逐次選擇法

由以上分析可知，由原始資料所配適的迴歸模型可能是在有極端值的存在下才具備如此高的解釋力，一旦極端值被剔除，模型解釋力便大幅下降，因此嘗試加入改變型態的獨變項。我們以居住空間大小 `sqfeet` 這個獨變項的平方作為新變項 `sqfeetsq`，令其為 $X_{sqfeetsq}$，並再次進行共線性檢查與逐次選擇法。

1.2.2.5.1 檢查共線性

依照前方之標準與步驟檢查獨立變量中的共線性，彼此不具共線性之獨變項有：$X_{lotsize}$、X_{garage}、X_{age}、$X_{type.bf}$、$X_{type.af}$、$X_{type.f}$、$X_{style.1.5}$、$X_{style.ranch}$、$X_{fireplac}$ 與 $X_{sqfeetsq}$ 等 10 個。

1.2.2.5.2 逐次迴歸分析

我們同樣使用逐步選擇法中的向後選擇法，並以 AIC 作為模型配適度指標。下表列出了我們進行逐步選擇的過程：

最後的模型為：

經過變數挑選後的模型為：

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_5 x_{i,garage} + \hat{\beta}_8 x_{i,age} + \hat{\beta}_9 x_{i,type.bf} + \hat{\beta}_{10} x_{i,type.af} + \hat{\beta}_{11} x_{i,type.f} + \hat{\beta}_{14} x_{fireplac} + \hat{\beta}_{15} x_{sqfeetsq}$$

其中 $\hat{\beta}_0$ = 261.9893，$\hat{\beta}_5$ = 28.0564，$\hat{\beta}_8$ = -2.1739，$\hat{\beta}_9$ = 29.7857，$\hat{\beta}_{10}$ = 42.9744，$\hat{\beta}_{11}$ = 40.3324，$\hat{\beta}_{14}$ = 43.7539，$\hat{\beta}_{15}$ = 0。

各個獨變項的係數估計值中，只有屋齡 age（$\hat{\beta}_8$）為負，表示隨著屋齡上升，房價會較低，而其他變項都與房價呈現正相關。

1.2.2.5.3 整體模型之 F 檢定

我們對以上的迴歸估計式進行顯著性為 0.05 之整體 F 檢定，檢定該迴歸模型是否對於房價具有解釋效力。

- 研究假設為:

$$H_0 : \beta_5 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{14} = beta_{15} = 0 \ v.s. \ H_1 : Not \ H_0$$

- 檢定統計量:

$$F = \frac{MS_R}{MS_E} \sim F(k,\ n-k-1);\ \begin{cases} MS_R = \frac{1}{k}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ MS_E = \frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \end{cases}$$

其中 n 為樣本大小（i.e., 28），k 為最後納入模型的獨變項個數（i.e., 7）。此檢定統計量服從自由度為 k 與 $n-k-1$ 的 F 分配。

- 檢定結果：檢定統計量 F 為 81.42，其 p 值為 0，小於顯著水準，因此我們拒絕 H_0，表示此迴歸模型對於依變項房價具有解釋力。

1.2.2.5.4 迴歸係數之 t 檢定

在假設 $\epsilon \overset{iid}{\sim} N(0, \sigma^2)$ 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。

- 研究假設：$H_0 : \beta_j = 0,\ v.s. \beta_j \neq 0,\ \forall j = 5, 8, 9, 10, 11, 14, 15$

- 檢定統計量:

$$T_j = \frac{\hat{\beta}_j}{\sqrt{\hat{Var}(\beta_j)}} \sim t(n-k-1)$$

- 檢定結果:

	估計值 9	5%CI 下界 9	5%CI 上界 t	檢定統計量	p 值
截距 (beta_0)	261.9893	216.0265	307.9520	11.8901	0.0000
garage (beta_5)	28.0564	7.7940	48.3189	2.8883	0.0091
age (beta_8)	-2.1739	-3.1123	-1.2355	-4.8322	0.0001
type.bf (beta_9)	29.7857	-3.3909	62.9624	1.8728	0.0758
type.af (beta_10)	42.9744	5.7096	80.2393	2.4056	0.0259
type.f (beta_11)	40.3324	8.5541	72.1106	2.6475	0.0155
fireplac (beta_14)	43.7539	14.9962	72.5116	3.1737	0.0048
sqfeetsq (beta_15)	0.0000	0.0000	0.0001	16.1930	0.0000

針對各係數的 t 檢定結果如上表。其中檢定 β_5、β_8、β_{10}、β_{11}、β_{14} 與 β_{15} 得到的 t 統計量之 p 值小於顯著水準，因此我們在這 6 個檢定中可以拒絕 H_0，顯示我們有充分證據可以宣稱 $\beta_5 = 0$、$\beta_8 = 0$、$\beta_{10} = 0$、$\beta_{11} = 0$、$\beta_{14} = 0$ 與 $\beta_{15} = 0$ 都是錯的，也就是說，車庫能容納的汽車數量、屋齡、房屋建材型態、壁爐數量與居住空間大小（經過平方轉換）對於房價都有顯著的預測力。

1.2.2.5.5 殘差模型診斷

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 0.9666，p 值為 0.492，不小於顯著水準，因此我們不拒絕 H_0，表示我們沒有足夠的證據顯示殘差不服從常態分配，通過診斷。

1.2.2.5.6 配適結果

我們建立一個預測房價的複迴歸模型:

$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_5 x_{i,garage} + \hat{\beta}_8 x_{i,age} + \hat{\beta}_9 x_{i,type.bf} + \hat{\beta}_{10} x_{i,type.af} + \hat{\beta}_{11} x_{i,type.f} + \hat{\beta}_{14} x_{fireplac} + \hat{\beta}_{15} x_{sqfeetsq}$$

其中 $\hat{\beta}_0 = 261.9893$，$\hat{\beta}_5 = 28.0564$，$\hat{\beta}_8 = -2.1739$，$\hat{\beta}_9 = 29.7857$，$\hat{\beta}_{10} = 42.9744$，$\hat{\beta}_{11} = 40.3324$，$\hat{\beta}_{14} = 43.7539$，$\hat{\beta}_{15} = 0$。

此迴歸模型經由自由度調整後的決定係數 R^2_{adj} 為 0.9542，顯示極大部分的變異可以被此模型解釋。由係數可知，在固定其餘獨變項下，房屋車庫容納量每多一台汽車，房價便提升 2805.64 美元；屋齡每多一年，房價便下跌 217.39 美元；房屋每多一個壁爐，房價便提升 4375.39。

1.2.2.5.7 再次剔除極端值後檢驗模型解釋力

在先前資料探索的部分我們以發現原始資料中有 2 筆資料房價明顯高於其他房屋，因此嘗試將這 2 筆資料剔除，對剩餘資料進行整體 F 檢定，檢查該迴歸模型是否依舊適配。

- 研究假設:

$$H_0 : \beta_5 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{14} = \beta_{15} = 0 \; v.s. \; H_1 : Not \; H_0$$

- 檢定統計量:

$$F = \frac{MS_R}{MS_E} \sim F(k,\ n-k-1);\ \begin{cases} MS_R = \frac{1}{k}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ MS_E = \frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \end{cases}$$

其中 n 為樣本大小（i.e., 26），k 為最後納入模型的獨變項個數（i.e., 7）。此檢定統計量服從自由度為 k 與 $n-k-1$ 的 F 分配。

- 檢定結果:

檢定統計量 F 為 10.87，其 p 值為 0，小於顯著水準，因此我們拒絕 H_0，表示此迴歸模型對於依變項房價仍具有解釋力。R^2_{adj} 也仍保有 0.7342。

1.2.2.6 模型比較

新增變數之線性迴歸模型 R^2_{adj} 明顯大於原始資料的做逐步分析所配適之模型的 R^2_{adj}，表示前者可相較後者解釋較多變異。且新增變數之線性迴歸模型 R^2_{adj} 在剔除極端值後依舊具有解釋能力，亦即該迴歸模型受極端值的影響較小，說明前者擁有更高的預測能力。

2 Exercise 10.4

2.1 研究問題

Hand 等人（1994）蒐集了英格蘭和威爾斯大型市鎮中於 1958 至 1964 年的男性平均死亡率、公共飲水中的鈣濃等資料，同時也紀錄了這些市鎮是否位於 Derby 鎮的北方。研究問題是藉由資料分析解釋死亡率與鈣含量、地理位置的關係，並提出對於公衛政策的建議。

2.2 資料介紹

資料集由 Hand 等人（1994）提供，有 61 列、3 個變項，一列為一個位於英格蘭與威爾斯的大型市鎮的資料，變項說明如下:

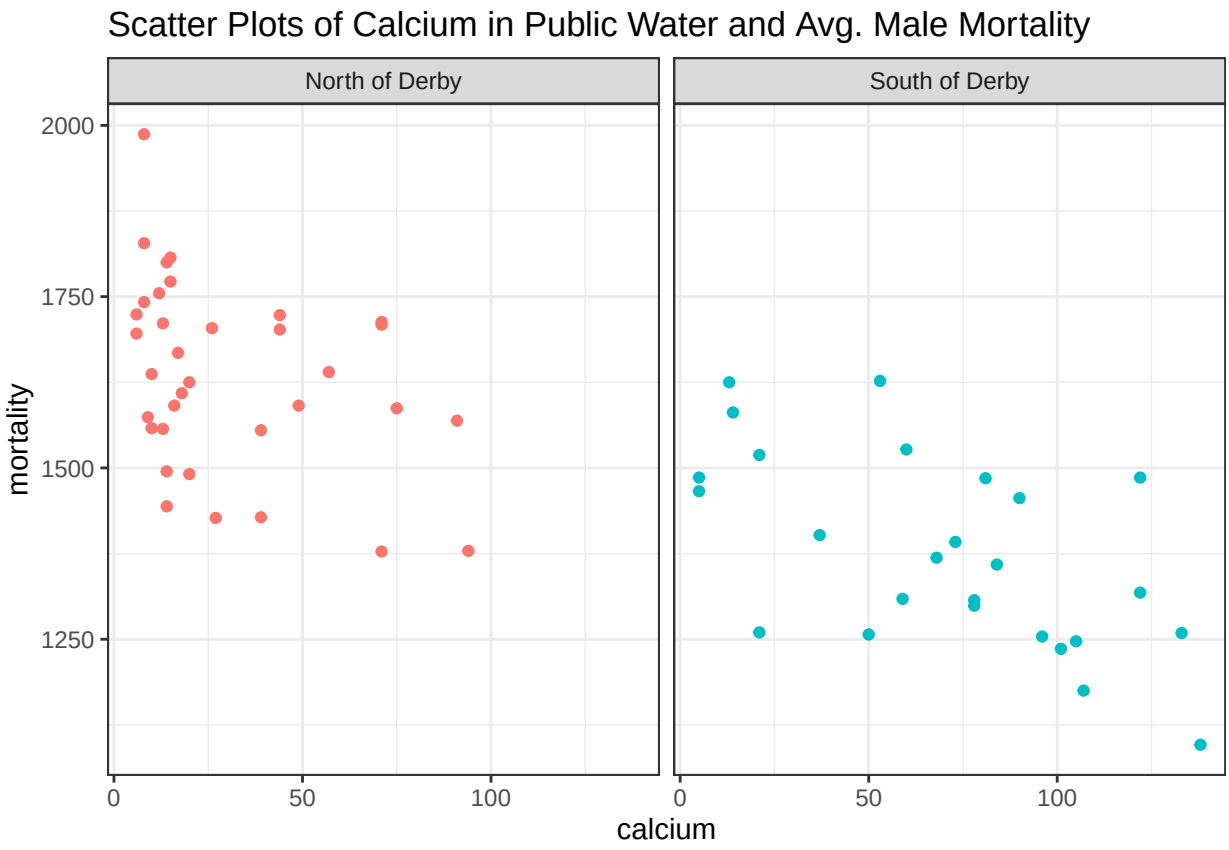
- `mortality`: 每十萬名男性的平均死亡率，取 1958-1964 年間的數值平均。

- `calcium`: 公共飲水的鈣濃度（單位：ppm），反應了水質硬的程度
- `derbynor`: 該市鎮是否比 Derby 鎮還北邊，二分變項

2.3 資料分析

2.3.1 資料探索

我們先以散佈圖看 `mortality` 與 `calcium` 這兩個變項的關聯，並以 `derbynor` 分層來看。



由上圖可發現：`mortality` 與 `calcium` 間似乎呈現負相關。而是否位於 Derby 北邊（`derbynor`）似乎對 `mortality` 變異的解釋也有貢獻：整體而言，位於 Derby 北邊的市鎮（i.e., `derbynor`=1）`mortality` 比位於 Derby 南邊的市鎮（i.e., `derbynor`=0）還高。

2.3.2 建立迴歸模型

2.3.2.1 分別針對位於 Derby 北邊與南邊的市鎮建立以鈣濃度預測死亡率的迴歸模型

- 定義變數：
  - $Y_{i,N}$ 為位於 Derby 北方之市鎮 i 的男性平均死亡率

  - $Y_{j,S}$ 為位於 Derby 南方之市鎮 j 的男性平均死亡率
  - $X_{i,ca,N}$ 為位於 Derby 北方之市鎮 i 公共飲水的鈣濃度
  - $X_{i,ca,S}$ 為位於 Derby 南方之市鎮 j 公共飲水的鈣濃度
  - 其中 $i = 1, 2, ..., 35$，$j = 1, 2, ..., 26$

- 模型:

$$\begin{cases} Y_{i,N} = \beta_{0,N} + \beta_{1,N} X_{i,ca,N} + \epsilon_{i,N} \\ Y_{j,S} = \beta_{0,S} + \beta_{1,S} X_{j,ca,S} + \epsilon_{j,S} \end{cases}$$

其中，$\beta_{0,N}$ 與 $\beta_{0,S}$ 分別為兩地區市鎮迴歸模型之截距，$\beta_{1,N}$ 與 $\beta_{1,S}$ 分別為兩地區市鎮迴歸模型之斜率，$\epsilon_{i,N}$ 與 $\epsilon_{i,S}$ 則為各自的殘差項，且 $\epsilon_{i,N} \overset{iid}{\sim} N(0, \sigma_i^2)$; $\epsilon_{j,S} \overset{iid}{\sim} N(0, \sigma_j^2)$

- 研究假設:

我們欲探討公共飲水鈣濃度對於男性平均死亡率的預測力各自在兩個區域是否顯著，也就是要分別檢定迴歸係數 $\beta_{1,N}$ 與 $\beta_{1,S}$ 是否顯著不為 0，研究假說為 $H_0 : \beta_{1,N} = 0$ $v.s.$ $H_1 : \beta_{1,N} \neq 0$ 與 $H_0 : \beta_{1,S} = 0$ $v.s.$ $H_1 : \beta_{1,S} \neq 0$。

- 檢定統計量與結果

依據資料的估計的結果為:

$$\begin{cases} \hat{Y}_{i,N} = \hat{\beta}_{0,N} + \hat{\beta}_{1,N} X_{i,ca,N} \\ \hat{Y}_{j,S} = \hat{\beta}_{0,S} + \hat{\beta}_{1,S} X_{j,ca,S} \end{cases}$$

其中 $\hat{\beta}_{0,N} = 1692.3128$，$\hat{\beta}_{1,N} =$ -1.9313，$\hat{\sigma_i}^2 = 1.6695058 \times 10^4$
$\hat{\beta}_{0,S} = 1522.815$，$\hat{\beta}_{1,S} =$ -2.0927，$\hat{\sigma_j}^2 = 1.3063908 \times 10^4$。
令 n 為樣本大小，k 為獨變項個數，檢定統計量 t_{β_j}（$j = 1, ..., k$）服從自由度 $n - k$ 的 t 分配，數學式如下:

$$t_{\beta_j} = \frac{\hat{\beta}_j - 0}{Var(\hat{\beta}_j)} \sim t(n - k)$$

我們令顯著水準 α 為 0.05。針對 Derby 北方市鎮的檢定結果如下表所示。其中，我們關心的 $\beta_{1,N}$ 之估計值 $\hat{\beta}_{1,N}$ 為，$1 - \alpha = 1 - 0.05 = 95\%$ 信賴區間（confident interval, CI）為 [-3.6564, -0.2063]，統計檢定量 t_{TS} 為-2.2778，其 p 值為 0.0293，小於顯著水準，因此我們拒絕 H_0，表示我們有充分證據支持 $\beta_{1,N} = 0$ 這個宣稱是錯的，也就是說，在位於 Derby 北方市鎮中，公共飲水鈣濃度對於男性平均死亡率的預測力顯著。

	估計值	95%CI 下界	95%CI 上界	t 檢定統計量	p 值
截距 (beta_0)	1692.3128	1623.5775	1761.0481	50.0912	0.0000
鈣濃度 (beta_1)	-1.9313	-3.6564	-0.2063	-2.2778	0.0293

針對 Derby 南方市鎮的檢定結果如下表所示。其中，我們關心的 $\beta_{1,S}$ 之估計值 $\hat{\beta}_{1,S}$ 為，95% 信賴區間為 [-3.2617, -0.9238]，統計檢定量 t_{TS} 為-3.6949，其 p 值為 0.0011，小於顯著水準，因此我們拒絕 H_0，表示我們有充分證據支持 $\beta_{1,S} = 0$ 這個宣稱是錯的，也就是說，在位於 Derby 南方市鎮中，公共飲水鈣濃度對於男性平均死亡率的預測力顯著。

	估計值	95%CI 下界	95%CI 上界	t 檢定統計量	p 值
截距 (beta_0)	1522.8150	1429.0503	1616.5798	33.5194	0.0000
鈣濃度 (beta_1)	-2.0927	-3.2617	-0.9238	-3.6949	0.0011

- 診斷前提假設是否滿足

線性迴歸分析有 4 項基本前提假設，我們分別針對兩個模型進行這些前題假設的診斷:

1. 線性關係：依變數和獨變項之間的關係必須是線性。

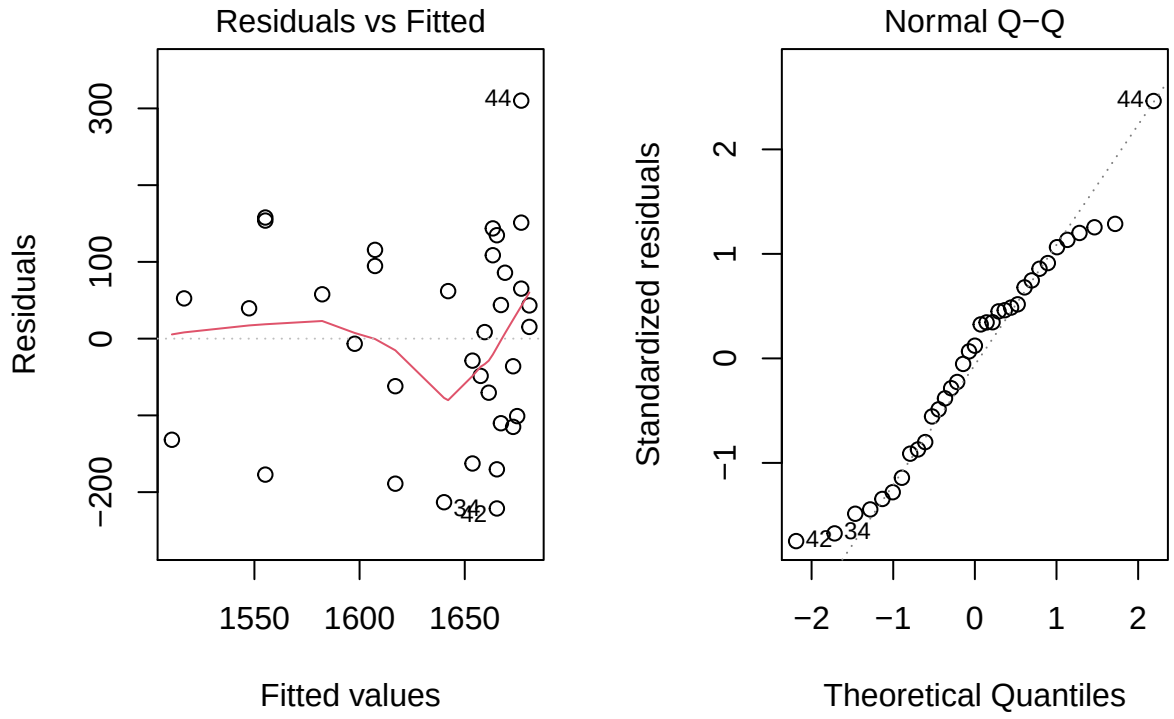
根據先前以是否位於 Derby 北方分層繪製的鈣濃度與死亡率的散佈圖，可知在兩個模型中，依變數（死亡率）和獨變項（鈣濃度）之間都呈現的線性負相關。符合。

2. 殘差（ϵ）服從常態分配。

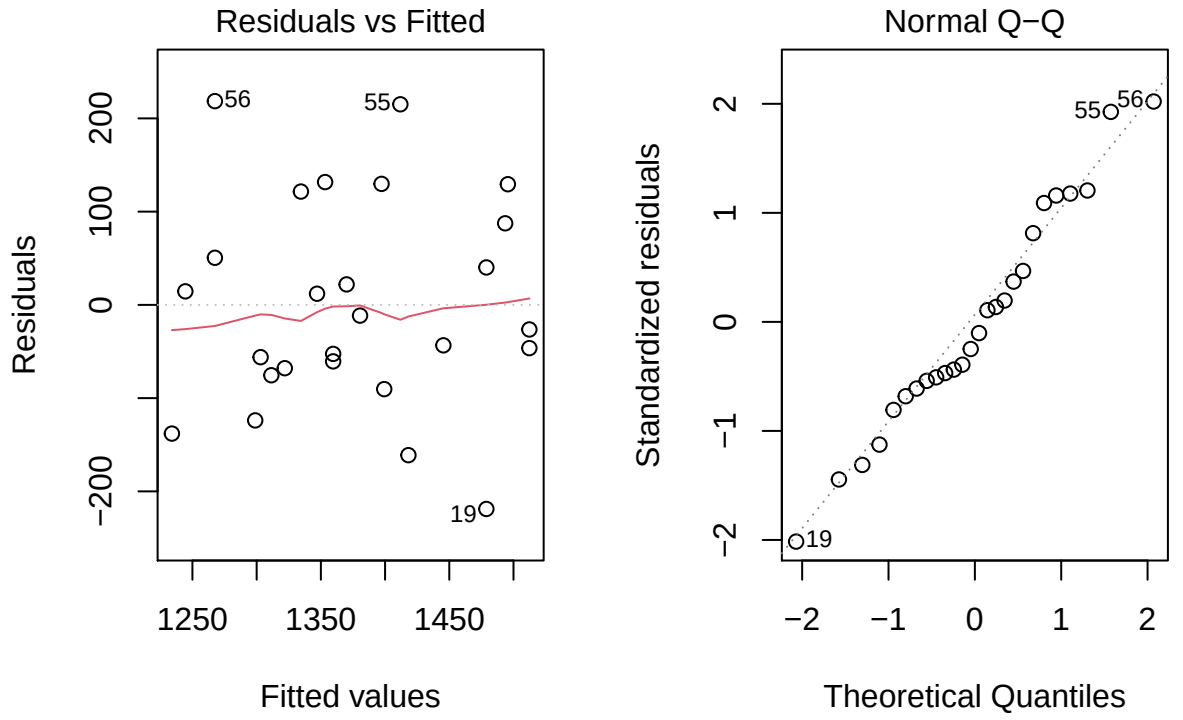
3. 殘差具備獨立性。

4. 殘差具備變異數同質性。

2 - 4 這 3 個假設可表示成：$\epsilon_{i,N} \overset{iid}{\sim} N(0, \sigma_i^2)$; $\epsilon_{j,S} \overset{iid}{\sim} N(0, \sigma_j^2)$



針對位於 Derby 北方市鎮的模型，Normal Q-Q plot（上右圖）中殘差 quantile 資料點大部分落在 45 度線上，顯示殘差可能服從常態分布。而殘差與配適值散佈圖（上左圖）則顯示在各配適值殘差之變異數差異不大。我們接著會進行檢定來確認這些假設是否成立（顯著水準均設為 0.05）。



針對位於 Derby 南方市鎮的模型，Normal Q-Q plot（上右圖）中殘差 quantile 資料點大部分落在 45 度線上，顯示殘差可能服從常態分布。而殘差與配適值散佈圖（上左圖）則顯示在各配適值殘差之變異數差異不大。我們以下面檢定來確認這些假設是否成立（顯著水準均設為 0.05）

- 以 Shapiro-Wilk 檢定檢驗殘差常態假設:

$$\begin{cases} H_0 : \epsilon_{i,N} \sim ND \ v.s. \ H_1 : \epsilon_{i,N} \ does \ not \sim ND \\ H_0 : \epsilon_{j,S} \sim ND \ v.s. \ H_1 : \epsilon_{j,S} \ does \ not \sim ND \end{cases}$$

針對位於 Derby 北方市鎮的模型的檢定結果如下：檢定統計量 W 為 0.9686，p 值為 0.406128，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並不服從常態分配，通過常態假設。

針對位於 Derby 南方市鎮的模型的檢定結果如下：檢定統計量 W 為 0.9729，p 值為 0.6990357，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並不服從常態分配，通過常態假設。

- 以 Durbin-Waston 檢定檢驗殘差獨立性:

$$\begin{cases} H_0 : \epsilon_{i,N} \text{ are independent. v.s. } H_1 : \epsilon_{i,N} \text{ are not independent.} \\ H_0 : \epsilon_{j,S} \text{ are independent. v.s. } H_1 : \epsilon_{j,S} \text{ are not independent.} \end{cases}$$

針對位於 Derby 北方市鎮的模型的檢定結果如下：檢定統計量 DW 為 2.5233，p 值為 0.9411，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。

針對位於 Derby 南方市鎮的模型的檢定結果如下：檢定統計量 DW 為 1.1304，p 值為 0.0092，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獨立性，j 未通過獨立假設。

- 以 Brown-Forsythe 檢定檢驗殘差變異同質性：

$$\begin{cases} H_0 : {\sigma_1}^2 = {\sigma_2}^2 = ... = {\sigma_i}^2 \ v.s. \ H_1 : Not \ H_0 \\ H_0 : {\sigma_1}^2 = {\sigma_2}^2 = ... = {\sigma_j}^2 \ v.s. \ H_1 : Not \ H_0 \end{cases}$$

針對位於 Derby 北方市鎮的模型的檢定結果如下：檢定統計量 BF 為 0.5731，p 值為 0.4544，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變異同質性，通過變異同質性假設。

針對位於 Derby 南方市鎮的模型的檢定結果如下：檢定統計量 BF 為 0.0079，p 值為 0.9301，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變異同質性，通過變異同質性假設。

- 結論

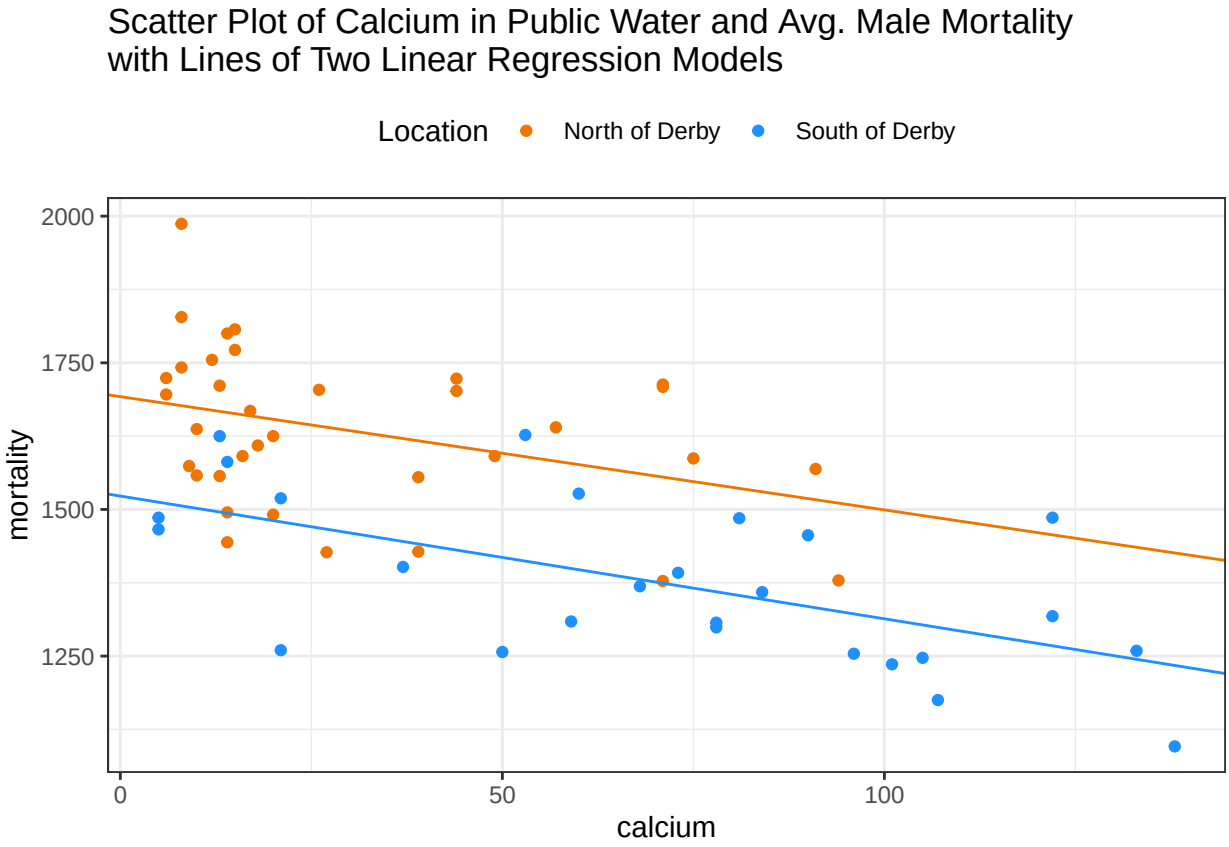
我們分別針對位於 Derby 北方與南方的市鎮建立了以公共飲水鈣濃度（$X_{ca,N}$ 與 $X_{ca,S}$）來預測湖水水位（Y_N 與 Y_S）的線性迴歸模型：

1. $\hat{Y}_{i,N} = 1692.3128 + \text{-}1.9313 X_{i,ca,N}, \forall i = 1, 2, ..., 35$
2. $\hat{Y}_{j,S} = 1522.815 + \text{-}2.0927 X_{j,ca,S}, \forall j = 1, 2, ..., 26$

模型解釋力可由經由自由度校正的決定係數 R^2_{adj} 描述，分別為 0.1097 與 0.336，顯示我們建立的兩個迴歸模型都只能解釋依變項少部分的變異，配適結果不佳。此外，亦要注意針對南方市鎮建立的模型並未通過殘差獨立假設。R^2_{adj} 公式如下：

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

將迴歸線繪製於兩變項的散佈圖上，可看出在公共飲水中擁有相同鈣含量的情況下，北方市鎮死亡率較南方市鎮高。



2.3.2.2 建立以鈣濃度與是否位於 Derby 北方預測死亡率的複迴歸模型

- 定義變數:
  - Y_i 為市鎮 i 之男性平均死亡率
  - $X_{i,ca}$ 為市鎮 i 公共飲水之鈣濃度
  - $X_{i,nor}$ 為市鎮 i 之相對地理位置，位於 Derby 北方者記為 1，位於 Derby 南方者為 0
  - $X_{i,ca\times nor}$ 為公共飲水中鈣濃度與相對地理位置之交互作用

- 模型:

$$Y_i = \beta_0 + \beta_1 X_{i,ca} + \beta_2 X_{i,nor} + \beta_3 X_{i,ca\times nor} + \epsilon_{i,N}$$

其中，β_0 為迴歸模型之截距，β_1、β_2 與 β_3 分別為 $X_{i,ca}$、$X_{i,nor}$ 與 $X_{i,ca\times nor}$ 之迴歸係數，ϵ_i 為模型殘差，且 $\epsilon_i \overset{iid}{\sim} N(0, \sigma_i^2)$

- 研究假設:

我們欲探討市鎮公共飲水鈣濃度、相對地理位置（是否為於 Derby 北方）以及這兩者的交互作用對於男性平均死亡率的預測力是否顯著，也就是要分別檢定迴歸係數 β_1、β_2 與 β_3 是否顯著不為 0，研究假說為:

$$\begin{cases} H_0 : \beta_1 = 0 \ v.s. \ H_1 : \beta_1 \neq 0 \\ H_0 : \beta_2 = 0 \ v.s. \ H_1 : \beta_2 \neq 0 \\ H_0 : \beta_3 = 0 \ v.s. \ H_1 : \beta_3 \neq 0 \end{cases}$$

。

- 檢定統計量與結果

依據資料的估計的結果為:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i,ca} + \hat{\beta}_2 X_{i,nor} + \hat{\beta}_3 X_{i,ca \times nor}$$

其中 $\hat{\beta}_0 = 1522.815$, $\hat{\beta}_1$ = -2.0927, $\hat{\beta}_2 = 169.4978$, $\hat{\beta}_3 = 0.1614 \hat{\sigma}_i^2$ =1.4652046×10^4

令 n 為樣本大小, k 為獨變項個數, 檢定統計量 t_{β_j} ($j = 1, ..., k$) 服從自由度 $n-k$ 的 t 分配, 數學式如下:

$$t_{\beta_j} = \frac{\hat{\beta}_j - 0}{Var(\hat{\beta}_j)} \sim t(n-k)$$

	估計值	95%CI 下界	95%CI 上界	t 檢定統計量	p 值
截距 (beta_0)	1522.8150	1424.7944	1620.8357	31.1096	0.0000
鈣濃度 (beta_1)	-2.0927	-3.3147	-0.8707	-3.4293	0.0011
相對地理位置 (beta_2)	169.4978	52.1702	286.8253	2.8929	0.0054
交互作用 (beta_3)	0.1614	-1.8665	2.1892	0.1594	0.8740

依據資料的估計的結果為:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i,ca} + \hat{\beta}_2 X_{i,nor}$$

其中 $\hat{\beta}_0 = 1518.7263$, $\hat{\beta}_1$ = -2.0341, $\hat{\beta}_2 = 176.7108$, $\hat{\sigma}_i^2$ =1.4911308×10^4

	估計值	95%CI 下界	95%CI 上界	t 檢定統計量	p 值
截距 (beta_0)	1518.7263	1435.9853	1601.4673	36.7419	0e+00
鈣濃度 (beta_1)	-2.0341	-3.0007	-1.0675	-4.2124	1e-04
相對地理位置 (beta_2)	176.7108	102.8649	250.5567	4.7900	0e+00

我們令顯著水準 α 為 0.05。檢定結果如上表所示。其中 β_1 與 β_2 的估計值分別為-2.0341 與 176.7108, 其 p 值為 10^{-4} 與 0, 都不小於顯著水準, 因此針對兩係數我們都拒絕 H_0, 表示我們有充分證據支持 $\beta_1 = 0$ 與 $\beta_2 = 0$ 是錯的, 也就是說, 公共飲水鈣濃度與相對地理位置對於男性平均死亡率的預測力都顯著。

- 診斷前提假設是否滿足

我們接著針對不考慮交互作用項的線性迴歸模型進行 4 項基本前提假設的診斷:

1. 線性關係:依變數和獨變項之間的關係必須是線性。

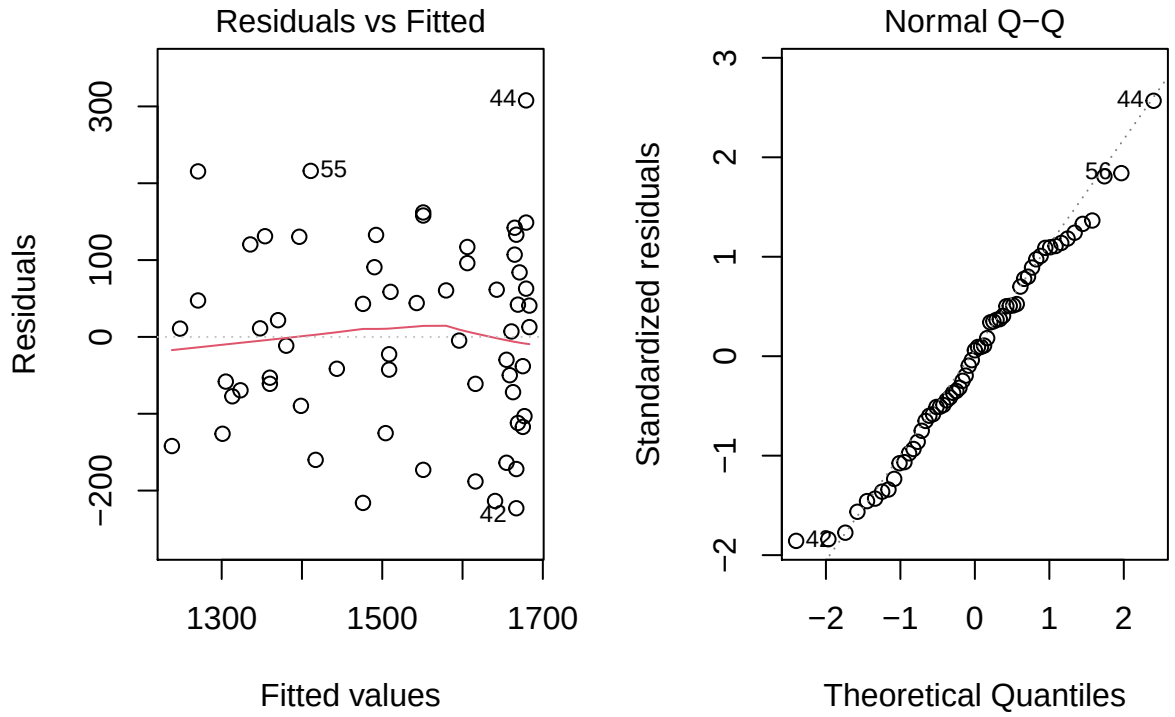
根據先前繪製的鈣濃度與死亡率的散佈圖,可知依變數(死亡率)和獨變項(鈣濃度)之間都呈現的線性負相關。符合。

2. 殘差(ϵ)服從常態分配。

3. 殘差具備獨立性。

4. 殘差具備變異數同質性。

2 - 4 這 3 個假設可表示成:$\epsilon_i \overset{iid}{\sim} N(0, \sigma_i^2)$



Normal Q-Q plot(上右圖)中殘差 quantile 資料點大部分落在 45 度線上,顯示殘差可能服從常態分布。而殘差與配適值散佈圖(上左圖)則顯示在各配適值殘差之變異數差異不大。我們接著會進行檢定來確認這些假設是否成立(顯著水準均設為 0.05)。

- 以 Shapiro-Wilk 檢定檢驗殘差常態假設:

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

檢定結果如下:檢定統計量 W 為 0.9846,p 值為 0.638777,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並不服從常態分配,通過常態假設。

- 以 Durbin-Waston 檢定檢驗殘差獨立性:

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

檢定結果如下: 檢定統計量 DW 為 2.1303，p 值為 0.6882，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。

- 以 Brown-Forsythe 檢定檢驗殘差變異同質性:

$$H_0 : {\sigma_1}^2 = {\sigma_2}^2 = ... = {\sigma_i}^2 \ v.s. \ H_1 : Not \ H_0$$

檢定結果如下: 檢定統計量 BF 為 0.9305，p 值為 0.3387，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變異同質性，通過變異同質性假設。

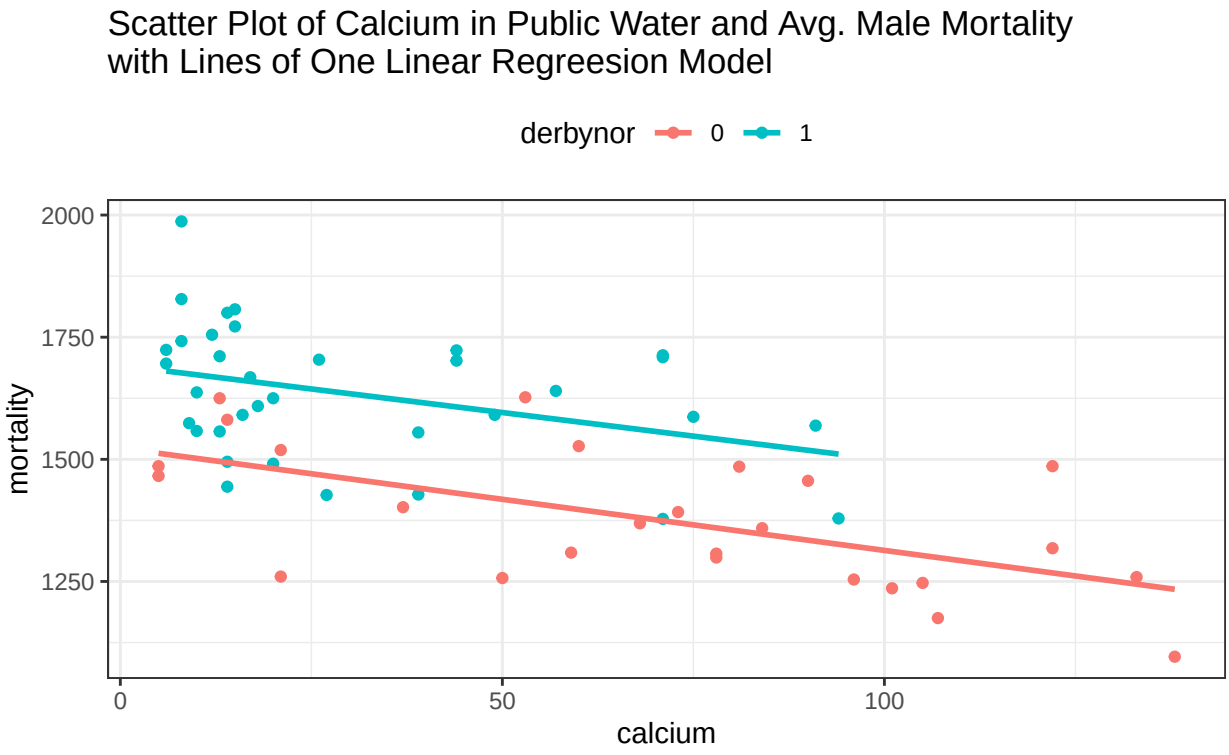
- 結論

我們建立了以公共飲水鈣濃度（X_{ca}）與是否位於相對地理位置（X_{nor}）來男性平均死亡率（Y）的線性迴歸模型:

$\hat{Y_i}$ = 1518.7263 + -2.0341$X_{i,ca}$+ 176.7108$X_{i,nor}$, $\forall i = 1, 2, ..., 61$

模型解釋力可由經由自由度校正的決定係數 R^2_{adj} 描述，為 0.5766，顯示我們建立的兩個迴歸模型都能解釋依變項超過一半的變異，配適結果尚可。

將迴歸線繪製於兩變項的散佈圖上，一樣可看出在公共飲水中擁有相同鈣含量的情況下，北方市鎮死亡率較南方市鎮高（北方線與南方線截距差為正，也就是 $\hat{\beta}_2$ 為正）。此外，由於交互作用項未納入模型中，兩線平行（斜率相同，皆為 $\hat{\beta}_1$）。



2.3.2.3 兩種模型建立方式之比較

R^2_{adj} 越大，表示線性迴歸模型可解釋越多的變異。將「是否位於 Derby 北方」以虛擬變數（dummy variable）納入複迴歸模型的 R^2_{adj} 為 0.5766，而針對位於 Derby 北方與南方市鎮分開建立的兩單迴歸模型的 R^2_{adj} 分別為 0.1097 與 0.336，複迴歸模型明顯高出許多。此外，針對位於 Derby 南方市鎮建立的單迴歸模型未通過殘差獨立假設的檢驗。因此以複迴歸模型作為最終模型。

不過兩者皆給出了相同的結論：一市鎮公共飲水中鈣濃度越高，其男性平均死亡率越低，且在相同公共飲水鈣濃度下，位於 Derby 北方的市鎮男性平均死亡率比位於 Derby 南方的市鎮高。

2.4 研究結論

1. 一市鎮公共飲水中的鈣濃度與其男性平均死亡率呈現線性負相關。當局可以嘗試增加公共飲水中的鈣濃度。不過以目前的分析不能做出因果關係的推斷。

2. 在相同公共飲水鈣濃度下，位於 Derby 北方的市鎮男性平均死亡率比位於 Derby 南方的市鎮高。這可能與氣候、生活條件或其他環境因素有關，可以蒐集相關資料進行更進一步的分析。

2.5 建議

1. 可再蒐集市鎮的氣候、降雨量、飲食習慣、生活習慣、社會經濟等資料，或許可以進而解釋為何位於 Derby 北方市鎮死亡率較高。

2. 若能取得各年或各時間點較高頻率的資料，而非一段時間平均後的單一資料點，分析時或許可以考慮時間的效果，並有可能可以做出因果推論。