**統計諮詢作業三**

**統計所 李權恩 R26094022**

**第一題(ex 9.3)**

資料簡述：

該資料由1977年Narula和Wellington所提供，共28筆，每筆資料代表1970年代初賓夕法尼亞州伊利市的房屋銷售價以及11個認為可能是影響價格的因素，各個變數名稱與其解釋如下表所示。

|  |  |
| --- | --- |
| 變數名稱 | 解釋 |
| price | 房屋價格，以百為單位 |
| taxes | 稅收，以美元計算 |
| bathrm | 浴室數量 |
| lotsize | 地皮大小，以平方英尺為單位 |
| sqfeet | 建坪大小，以平方英尺為單位 |
| garage | 車庫數量 |
| rooms | 房間數量 |
| bedrm | 臥室數量 |
| age | 屋齡 |
| type | 房屋類型 (磚、磚+框架、鋁+框架、框架) |
| style | 房屋風格 (2層、1.5層、牧場式) |
| fireplac | 壁爐數量 |

分析：

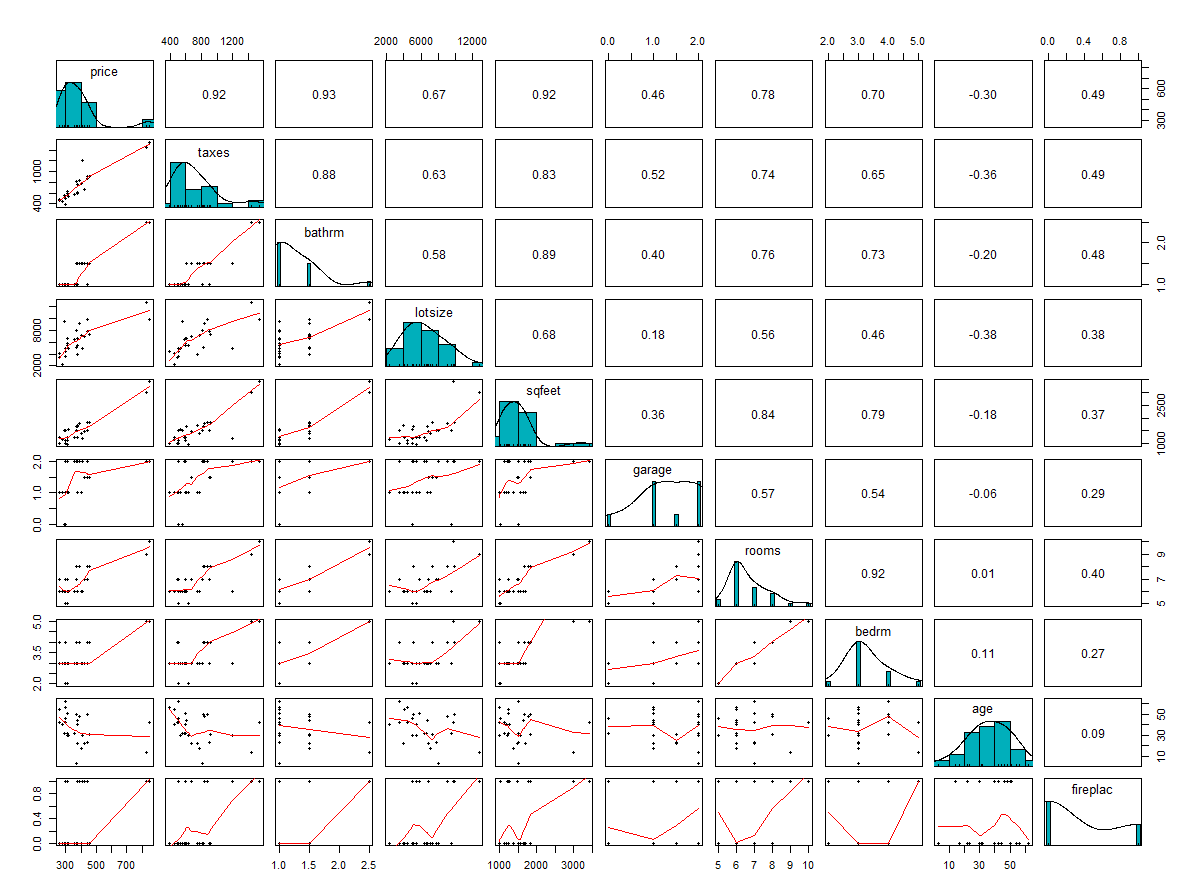
分析的主要流程如下：

1. 了解各項變數之間的關係
2. 分析11個變數當中影響價格的重要因素為何
3. 解釋會影響的因素是如何影響價格

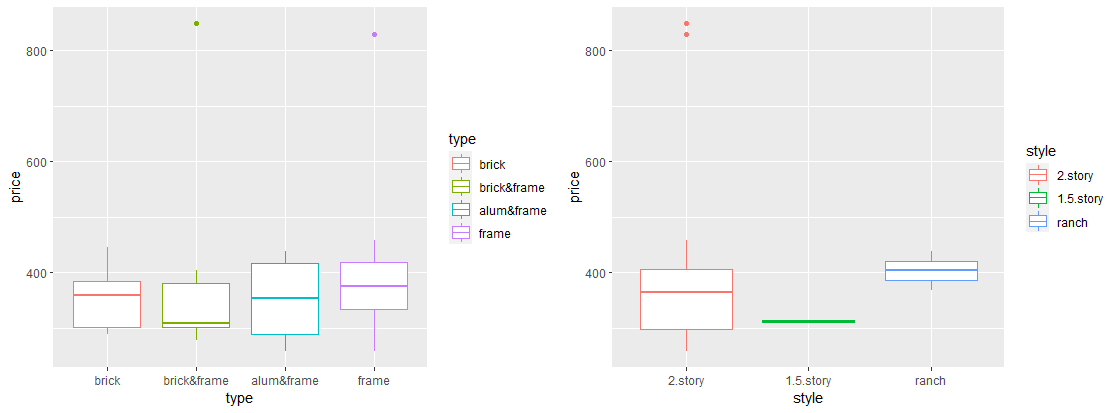
首先我們先對資料進行探索式的初步分析，在資料的變數當中有10個是連續變數，2個是類別變數，為了可以快速了解變數之間的關係為何，我們使用散佈圖來觀察10個連續變數的兩兩關係。

下圖為散佈圖矩陣，圖的上三角為兩兩變數之間的相關係數，對角線為該變數的直方圖，下三角為兩兩變數間的散佈圖，透過下圖我們可以發現：

除了屋齡以外，其他變數兩兩之間都是正相關，隨著建坪、房間數等等的設備增加，稅收、售價等等的變數也跟著上升，這很符合一般的房屋狀況。另外在中間對角線的分佈圖上，售價、稅收及坪數上有幾個特別高的數值，這也代表資料內可能有些觀測值是紀錄那些比較高級的豪宅。



不過在散佈圖矩陣裡我們只有辦法討論連續變數，所以在離散變數的部分我們觀察其與價格之間的關係。透過下圖左方，房屋的類型的中位數都在300至400之間，又以鋁+框架的類型變異最大，在磚+框架與框架兩個類別中各有一筆離群值超過800。另外在圖右可以看到兩筆離群值皆為兩層式，但在1.5層及牧場式都只有兩筆資料。



在對資料有了初步的認識之後來考慮我們分析最主要的問題：影響房價的因素。在我們的問題當中，最感興趣的變數房價為一連續型變數，所以我們考慮使用複迴歸分析，將房價外的其他變數作為解釋變數建立模型，模型如下：

代表房價，代表截距項、代表當變動一單位時，平均房價上升多少單位，最後為各個解釋變數，雖然課本當中不考慮將兩個類別變數納入模型，不過我基於以下兩點不移除，(1)迴歸模型可以分析類別變數；(2)每筆資料都額外收集了這兩項資訊，應多加利用。各變數符號如下所列，因為有兩類以上的類別變數，所以。在房屋類型有4類所以使用三個變數來各自區分，並選擇磚作為基準組記做(0,0,0)，房屋風格有三類所以使用兩個變數，選擇2層作為基準組。

在建立好全模型(full model)之後，我們想先了解殘差是否通過前提假設，主要是因為接下來不論是變數篩選、參數推論都是基於殘差符合模型一開始的假設。下面我們檢定殘差的常態性、同質性以及無關性。

在**常態性**的部分使用Shapiro-Wilk test來檢驗殘差是否服從常態分配。虛無假設與對立假設如下：

而得到的檢定統計量與p-value分別為，表示不拒絕虛無假設，我們沒有充分的證據顯示殘差不服從常態分配。

在**同質性**的部分使用Breusch–Pagan test檢驗殘差是否具有同質性。虛無假設與對立假設如下：

而得到的檢定統計量與p-value分別為，表示不拒絕虛無假設，我們沒有充分的證據顯示殘差不具同質性。

最後在**無關性**的部分，由於資料點之間是否有關連很難做檢測，但我們可以退而求其次使用Durbin-Watson test來檢驗資料點是否有一階自我相關。而其虛無假設與對立假設如下：

而得到的檢定統計量與p-value分別為，表示不拒絕虛無假設，我們沒有充分的證據說明殘差有一階自我相關的存在。

上述殘差的前提假設檢定皆是通過的，所以進一步做檢定該模型是否具有解釋能力，虛無假設與對立假設如下：

而得到的檢定統計量與p-value分別為，表示拒絕虛無假設，我們有充分的證據顯示至少有一個，至少有一個自變數對於反應變數是有線性關係的。

再進行個別參數估計量的檢定之前，我們需要考慮模型是否有共線性的狀況，主要的原因在於共線性會讓原本與反應變數無關的解釋變數在經過組合之後變成與反應變數有關，進而導致產生錯誤的結論。

在檢查共線性時，最主要的是看膨脹係數(VIF)，一般來說VIF小於5最為理想，代表模型沒有共線性的狀況，而VIF超過10代表模型內有嚴重的共線性問題。下表是關於膨脹係數(VIF)的檢驗表，可以看到除了地皮大小、車庫數量、屋齡、房屋類型、房屋風格及壁爐數量以外，變數之間都有嚴重的共線性問題，代表我們需要額外的準則來判斷哪些變數留在模型裡面是最有解釋能力。

|  |  |
| --- | --- |
| 變數名稱 | VIF |
| 稅收 | **13.96** |
| 浴室數量 | **16.88** |
| 地皮大小 | 3.11 |
| 建坪大小 | **16.79** |
| 車庫數量 | 2.98 |
| 房間數量 | **13.37** |
| 臥室數量 | **20.23** |
| 屋齡 | 2.91 |
| 房屋類型 | 1.67 |
| 3.22 |
| 1.71 |
| 房屋風格 | 1.64 |
| 5.15 |
| 壁爐數量 | 2.49 |

於是我們可以再進一步針對各個參數估計進行檢查，虛無假設與對立假設如下:

下表為各個參數估計量的檢定，透過下表可以看見的p-value<0.05，拒絕虛無假設，代表建坪大小與房屋類型是顯著影響我們的反應變數(房價)的。

|  |  |  |  |
| --- | --- | --- | --- |
| 變數名稱 | 迴歸係數 | 估計量 | p-value |
| 稅收 |  | -0.125 | 0.902 |
| 浴室數量 |  | 1.338 | 0.204 |
| 地皮大小 |  | 1.095 | 0.294 |
| 建坪大小 |  | 3.258 | **0.006** |
| 車庫數量 |  | 1.772 | 0.100 |
| 房間數量 |  | -0.523 | 0.610 |
| 臥室數量 |  | -0.261 | 0.798 |
| 屋齡 |  | -1.068 | 0.305 |
| 房屋類型 |  | 2.613 | **0.021** |
|  | 2.032 | 0.063 |
|  | 1.870 | 0.084 |
| 房屋風格 |  | 0.902 | 0.383 |
|  | 0.302 | 0.767 |
| 壁爐數量 |  | 1.424 | 0.178 |

從表2結果可看出，為顯著，拒絕虛無假設，有足夠證據顯示其不等於0，表示對價格有顯著的線性影響。然而這樣子的模型過於複雜且有些不必要的自變數，並且從前面資料初步探索有發現，某些自變數之間具有高度線性關係，對於迴歸模型會有共線性的問題發生，此問題會導致某個兩個變數可能是顯著的，但因為共線性的問題，使這兩個變數變成不顯著，因此考慮將模型進行簡化。

在簡化模型的部分，我們使用AIC準則搭配逐步選取法來篩選變數。其中*k*為參數個數，*n*為樣本數，*RSS*為殘差平方和，我們希望AIC越小越好，AIC的在考量殘差平方和越來越小的同時也不可以讓參數個數過大。而逐步選取法像是向前選取、向後選取的綜合版，最大的不同點在於逐步選取法可以允許被排除的變數又被選進模型，也允許被選進的變數最後又被模型刪除。

從下表可以發現，一開始什麼變數都沒有的模型AIC為278.34，在step-1考慮加入單個解釋變數時，發現加入浴室數量後AIC達最低為226.12。而step-2再考慮要加入新變數或是刪除浴室數量，多加入稅收時可使AIC達最低為216.96。step-3考慮刪除前面加入的兩個變數或是加入其他新變數，多加入建坪大小時可使AIC達最低為208.90。此時不論是刪除這三個變數或是加入其他新變數都會使AIC變大，因此到此停止。

|  |  |  |
| --- | --- | --- |
| 行為 | 模型內自變數 | AIC |
| 無 | 沒有自變數 | 278.34 |
| 加入:浴室數量 | 浴室數量 | 226.12 |
| 加入:稅收 | 浴室數量、稅收 | 216.96 |
| 加入:建坪大小 | 浴室數量、稅收、建坪大小 | 208.90 |

最後簡化的模型如下:

並且針對簡化模型，再做一次殘差檢定。

在**常態性**的部分一樣使用Shapiro-Wilk test來檢驗簡化模型的殘差是否服從常態分配。虛無假設與對立假設如下：

而得到的檢定統計量與p-value分別為，表示不拒絕虛無假設，我們沒有充分的證據顯示簡化模型的殘差不服從常態分配。

在**同質性**的部分一樣使用Breusch–Pagan test檢驗簡化模型的殘差是否具有同質性。虛無假設與對立假設如下：

而得到的檢定統計量與p-value分別為，表示不拒絕虛無假設，我們沒有充分的證據顯示簡化模型的殘差不具同質性。

最後在**無關性**的部分使用Durbin-Watson test來檢驗簡化模型的殘差是否有一階自我相關。而其虛無假設與對立假設如下：

而得到的檢定統計量與p-value分別為，表示不拒絕虛無假設，我們沒有充分的證據說明簡化模型的殘差有一階自我相關的存在。

到目前為止，模型 ，，已經是利用原始資料找到的一個不錯的模型，不過我們也可以將原始資料的解釋變數做轉換之後當作新的變數再試著重新篩選一次，這裡我們將建坪大小取平方之後再重新執行AIC逐步選取法。

挑選出來的模型如下：

根據前一章節，已得到簡化後的最終模型，該模型的R平方為0.932，表示房屋的售價有93.2%可以被稅收、浴室數量及建坪大小所解釋。

控制稅收以外的因素，稅收每增加一單位，房屋價格平均會增加0.19單位，亦即平均會增加1美元。

控制浴室數量以外的因素，浴室數量每增加一單位，房屋價格平均會增加87.87單位，亦即平均會增加8787美元。

控制建坪大小以外的因素，建坪大小每增加一單位，房屋價格平均會增加0.1單位，亦即平均會增加10美元。

然而其實這樣所做出的迴歸模型並無經濟上的意義，從現實層面考量，房屋售價本就與三者息息相關，因此建議資料可以多收集一些外部因素做為參考，而非單純只有房屋交易之資訊。

第二題(ex 10.4)

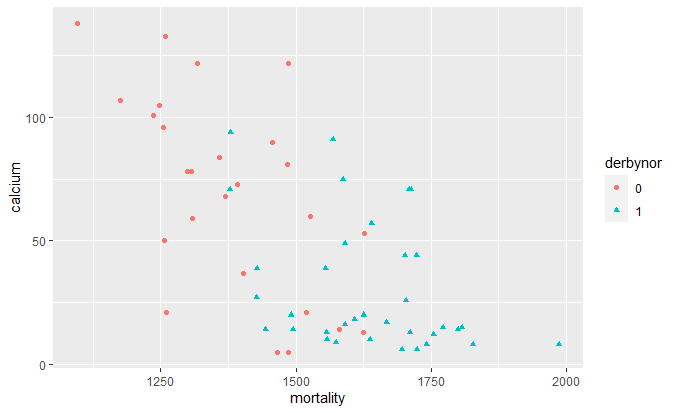
資料簡述：

資料由1994年Hand所收集，共61筆，紀錄英格蘭和威爾士的61個大城鎮中的每十萬男性中死亡人數以及公共飲用水中的鈣濃度。紀錄是從1958年至1964年的平均值，以及另外紀錄城鎮是在Derby的北方或南方。

|  |  |
| --- | --- |
| 變數名稱 | 解釋 |
| mortality | 每十萬男生死亡人數 |
| calcium | 公共飲用水中的鈣濃度(ppm) |
| derbynor | 1:derby的北方  0: derby的南方 |

分析：

首先，我們可以藉由視覺化的方式將三個變數呈現出來，下圖是死亡人數鈣濃度的散佈圖，而點的顏色及形狀代表不同的城鎮位置(北方或南方)。透過下圖，若我們僅考慮南方與北方的死亡人數，可以從圖中看出青色點的集中位置死亡人數較高，不過從這張圖中還有其他資訊是，青色的樣本點多位於死亡人數高但鈣濃度較低，紅色的點則較傾向死亡人數低但鈣濃度較高，所以或許死亡人數同時受到地區及飲用水鈣濃度的影響。



因此本次分析的目的為：

1. 了解死亡人數是否與水中的鈣濃度有關
2. 了解城鎮的位置是否會影響死亡人數
3. 了解城鎮位置與鈣濃度是否有交互作用

本次分析使用的模型如下:

其中*Y*表示為死亡人數，表示為飲用水中的鈣濃度，表示為城鎮相對位置，這裡我們先考慮有交互作用的配置。

在建立好全模型(full model)之後，我們想先了解殘差是否通過前提假設，主要是因為接下來不論是變數篩選、參數推論都是基於殘差符合模型一開始的假設。下面我們檢定殘差的常態性、同質性以及無關性。

在**常態性**的部分使用Shapiro-Wilk test來檢驗殘差是否服從常態分配。虛無假設與對立假設如下：

而得到的檢定統計量與p-value分別為，表示不拒絕虛無假設，我們沒有充分的證據顯示殘差不服從常態分配。

在**同質性**的部分使用Breusch–Pagan test檢驗殘差是否具有同質性。虛無假設與對立假設如下：

而得到的檢定統計量與p-value分別為，表示不拒絕虛無假設，我們沒有充分的證據顯示殘差不具同質性。

最後在**無關性**的部分使用Durbin-Watson test來檢驗殘差是否有一階自我相關。而其虛無假設與對立假設如下：

而得到的檢定統計量與p-value分別為，表示不拒絕虛無假設，我們沒有充分的證據說明殘差有一階自我相關的存在。

上述殘差的前提假設檢定皆是通過的，所以進一步做檢定該模型是否具有解釋能力，虛無假設與對立假設如下：

而得到的檢定統計量與p-value分別為，表示拒絕虛無假設，我們有充分的證據顯示至少有一個，至少有一個自變數對於反應變數是有線性關係的。

於是我們可以再進一步針對各個參數估計進行檢查，虛無假設與對立假設如下:

下表為各個參數估計量的檢定，透過下表可以看見的p-value<0.05，拒絕虛無假設，代表鈣濃度與城鎮相對位置是顯著影響我們的反應變數(死亡人數)的。另外，我們也發現交互作用並不顯著，這代表我們接受飲用水鈣濃度對死亡人數的影響程度在不同地區是相同的。

|  |  |  |  |
| --- | --- | --- | --- |
| 變數名稱 | 迴歸係數 | 估計量 | p-value |
| 飲用水鈣濃度 |  | -2.093 | 0.001 |
| 城鎮相對位置 |  | 169.498 | 0.005 |
| 交互作用 |  | 0.161 | 0.874 |

接著在交互作用不顯著的狀況下，我們建立另一個不含交互作用的簡化模型，並對全模型與簡化模型進行LR test來檢驗是否全模型的解釋能力較佳。虛無假設與對立假設如下：

而得到的檢定統計量與p-value分別為，因此不拒絕虛無假設，沒有充分的證據顯示全模型的解釋能力較佳。簡化模型的參數估計如下。

|  |  |  |  |
| --- | --- | --- | --- |
| 變數名稱 | 迴歸係數 | 估計量 | p-value |
| 飲用水鈣濃度 |  | -2.034 | <0.001 |
| 城鎮相對位置 |  | 176.711 | <0.001 |

結論：

最終由LR test的結果我們推薦使用簡化模型作為解釋依據，迴歸模型如下:

當固定城鎮的相對位置時，水中鈣濃度每上升一單位(ppm)時，平均死亡人數減少2.03人；當固定飲用水鈣濃度時，位在北方城鎮的平均死亡人數會比位在南方城鎮的平均死亡人數多176.71人。

由本次分析結果可以知道，飲用水中的鈣濃度對死亡人數是有負相關的，代表當濃度升高，平均死亡人數會下降，因此我們建議可以對飲用水進一步處理提高鈣濃度，不過想當然並非無上限的提高，在我們的資料當中，鈣濃度最低為5ppm最高為138ppm，所以我們只能保證在5ppm到138ppm之間，提升鈣濃度是會帶來上述的影響。而城鎮的相對位置對於死亡人數也有所影響，位於北方城鎮的死亡人數較多，認為可能是北方城鎮有著其他影響死亡的因素存在，例如氣候、飲食等等導致其死亡人數較多。