Ex9.3

1. 資料說明

該資料為1977年由Narula和Wellington所收集的資料，該資料共有28筆，為1970年代初賓夕法尼亞州伊利市的28套房屋的銷售價格數據，以及有11個認為可能是影響價格的因素，表1即給出了這11個因素分別為什麼以及其含義。以下為本次分析之目的:

1. 除價格以外之變數的11個因素是否會影響價格
2. 上述會影響的因素是如何影響價格

|  |  |
| --- | --- |
| 變數 | 說明 |
| price | 房屋價格，以百為單位 |
| taxes | 稅收，以美元計算 |
| bathrm | 房屋內浴室的數量 |
| lotsize | 地皮大小，以平方英尺為單位 |
| sqfeet | 房屋建坪大小，以平方英尺為單位 |
| garage | 附有幾間車庫 |
| rooms | 房屋內房間數量 |
| bedrm | 房屋內臥室數量 |
| 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 | 房屋內壁爐數量 |

表1 變數說明

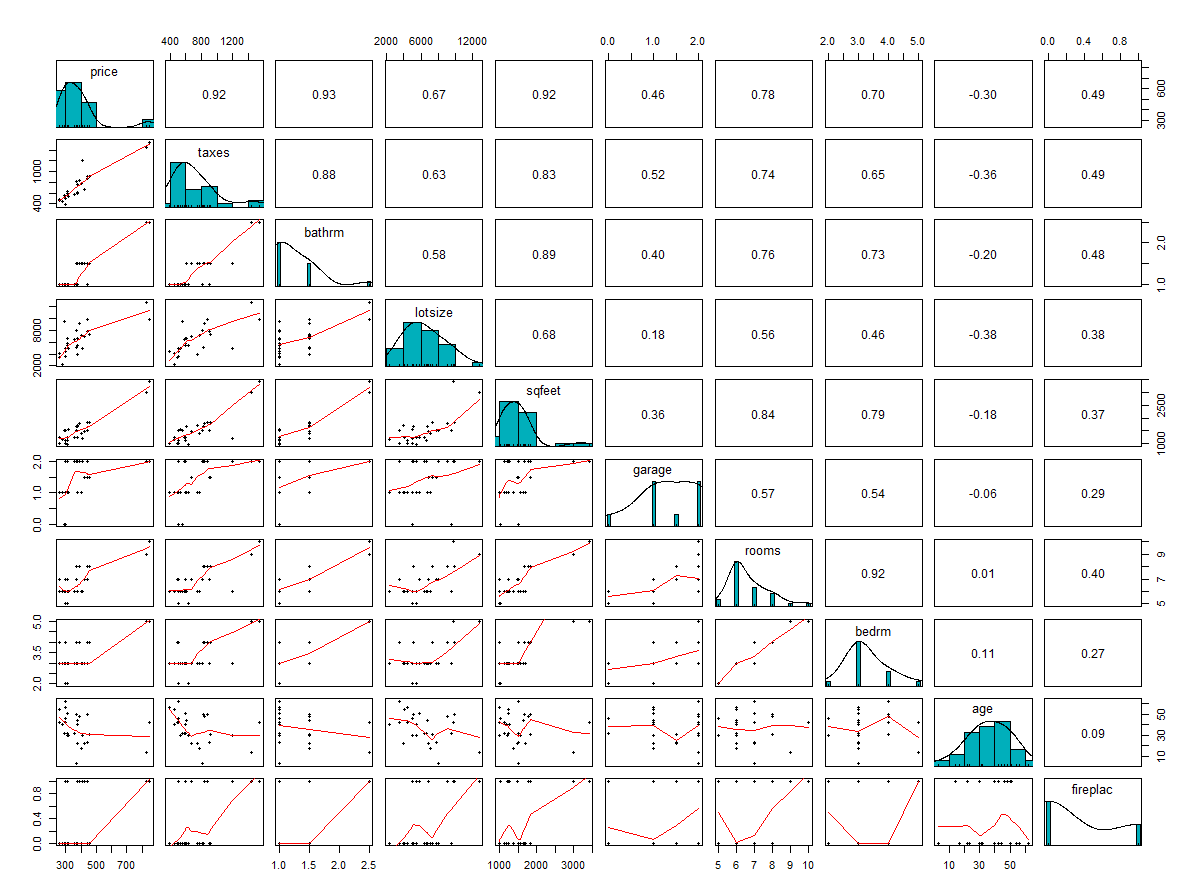
1. 資料初步探索

從資料說明已知道此資料集共有兩個類別變數，其餘包含感興趣價格的十個皆為連續型變數，因此將這些連續型變數畫散佈圖矩陣對資料進行初步探索。圖1右上角表示兩變數之間的皮爾森相關係數，對角線為該變數直方圖，左下角為兩變數之間散佈圖。

從圖1第一列以價格為主軸可以看到價格與屋齡以外之因素，皆成正相關，當價格高時，這些變數也會較高，而與屋齡成負相關，當價格低時，屋齡反而較高，也蠻符合一般房屋情況。

再以圖1對角線之直方圖，整體來看每個變數皆較集中，而在價錢、稅收、地坪、建坪、房間數以及臥室數有幾個特別高之數值，是後續分析所需要注意。

最後圖一第一行以價格為主軸與其餘變數的散佈圖，可以看出從前段直方圖看到特別高數值的變數對應到了高價格，這兩筆較高價格的交易與整體差距甚大，且某些變數之間具有高度線性關係，對後續分析可能有影響需特別注意。

圖1 散佈圖矩陣

前面我們只探討到連續型變數，接著觀察感興趣的價格和另外兩個類別變數之間的關係，從圖2來看，房屋類型整體對於價格似乎沒太大影響，而兩筆較高房屋價格的房屋類型為磚+框架與框架。而房屋風格多數集中在兩層式，而1.5層與牧場式各只有兩筆交易，因此不另外畫圖觀察，兩筆較高的房屋價格也為兩層式。

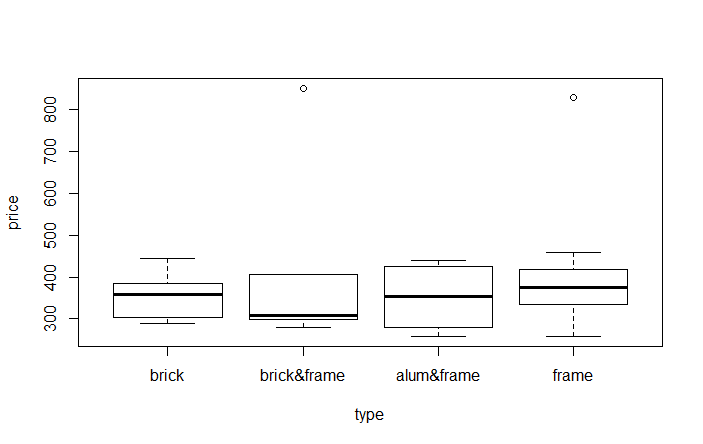


圖2 以房屋類型分組之價格盒型圖

1. 分析方法

從第一章節資料說明提到該分析是想了解哪些因素會影響房屋售價，這邊採取的分析方法為複迴歸分析，針對感興趣變數是連續型時即可使用，其模型如下:

其中*Y*表示為房屋價格，為截距項，為偏迴歸係數，當控制某個變數以外的所有變數時，該變數對應變數*Y*的線性影響，為殘差項必須彼此獨立且服從平均數為0變異數為的常態分配。而所代表的變數涵義如下:

配適好迴歸模型後需要針對各個變數進行檢定，然而檢定各個變數是否對價格有影響的t檢定是建立於殘差獨立且服從相同常態分配下所推導出的結果，因此需先檢查殘差是否通過常態性、同質性以及獨立性。

1. 常態性

利用Shapiro-Wilk檢定殘差是否服從常態分配，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，認為殘差通過常態性檢定。

1. 獨立性

利用durbinWatson檢查殘差之間是否存在自我相關，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，認為殘差不存在一階自我相關。

1. 同質性

利用Breusch–Pagan檢查殘差是否符合同質性，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，認為殘差通過同質性檢定。

總結上述殘差是通過的，因此前一章節所發現兩筆特別高的房屋交易資料不做另外的處理，進一步做檢定該模型是否有用，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

表示拒絕虛無假設，有足夠證據顯示。因此認為至少有一個自變數對於應變數是有線性影響的，進一步針對各個變數進行檢查，虛無假設與對立假設如下:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 迴歸係數 | 檢定統計量 | p-value | 迴歸係數 | 檢定統計量 | p-value |
|  | -0.125 | 0.9021 |  | -1.068 | 0.30506 |
|  | 1.338 | 0.20394 |  | 2.613 | 0.02147 |
|  | 1.095 | 0.29354 |  | 2.032 | 0.06313 |
|  | 3.258 | 0.00623 |  | 1.87 | 0.08422 |
|  | 1.772 | 0.09989 |  | 0.902 | 0.38329 |
|  | -0.523 | 0.60997 |  | 0.302 | 0.76748 |
|  | -0.261 | 0.79826 |  | 1.424 | 0.178 |

表2 迴歸係數檢定表

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。

1. 簡化模型

從前一章節已了解到目前模型雖然符合迴歸的前提假設，但過於複雜且有不必要的自變數與共線性的狀況，因此考慮簡化模型，採用以下AIC作為簡化原則，該準則為若模型計算出的AIC越小，表示此模型越好。篩選變數的步驟採去逐步選取法，逐步選取法是結合向前選取及向後選取兩種方法，最大不同處是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。

從表3可得，一開始什麼變數都沒有的模型AIC為278.34，此時只考慮加入變數，發現加入浴室數量後AIC達最低為226.12。而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。下一步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。

其中*k*為參數個數，*n*為樣本數，*RSS*為殘差平方和

|  |  |
| --- | --- |
| 模型 | AIC |
| 沒有自變數 | 278.34 |
| +浴室數量 | 226.12 |
| +浴室數量+稅收 | 216.96 |
| +浴室數量+稅收+建坪大小 | 208.90 |

表3 逐步選取法

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，確保此模型是否有被那兩筆過高價格的房屋交易所影響。

1. 常態性

利用Shapiro-Wilk檢定殘差是否服從常態分配，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，認為殘差通過常態性檢定。

1. 獨立性

利用durbinWatson檢查殘差之間是否存在自我相關，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，認為殘差不存在一階自我相關。

1. 同質性

利用Breusch–Pagan檢查殘差是否符合同質性，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，認為殘差通過同質性檢定。

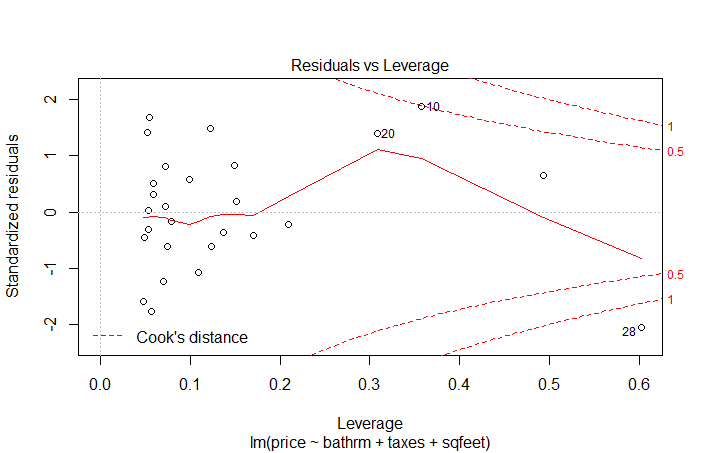
而因簡化模型通過殘差的三大前提假設的檢定，因此雖然從圖3的cook’s distance發現資料有影響點，也不考慮將該資料做刪除，因資料集中含有此筆資料，也就代表現實是有該情形，認為不可隨意做刪除資料的動作。

圖3 殘差分析圖

而進一步使用LR檢定，對原始模型與簡化模型比較優劣，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

因此不拒絕虛無假設，認為簡化後的模型較原始模型好。

1. 結論

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。

Ex10.4

1. 資料說明

該資料為1994年由Hand所收集的資料，該資料一共有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，該數據是紀錄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。表1即為這些變數之名稱以及說明。以下為本次分析之目的:

1. 死亡率與水中的鈣濃度是否有關
2. 城鎮的位置是否會影響死亡率與水中鈣濃度之間的關係

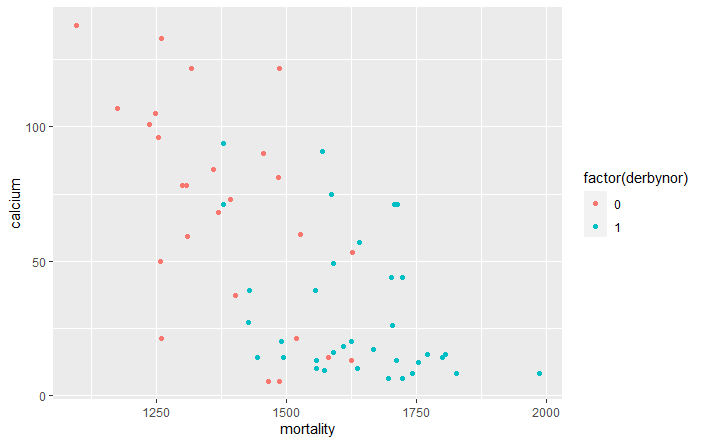
|  |  |
| --- | --- |
| 變數 | 說明 |
| mortality | 每10萬男性死亡人數 |
| calcium | 公共飲用水中的鈣濃度，以ppm為單位 |
| derbynor | 位於derby北方記作1，位於南方則記作0 |

表1 變數說明

1. 資料初步探索

從資料說明已知此資料集中有兩個連續型變數與一個類別型變數，因此我們以類別變數做分割，劃出死亡人數與水中鈣濃度的散佈圖，觀察一下兩個連續型變數之間的關係，以及城鎮位置是否會影響此兩者之間的關係。

從圖1可看出，若不考慮城鎮位置時，兩個變數之間的關係為負相關，計算出相關係數為-0.65，表示若死亡人數越高時，反而水中鈣濃度會較低。而考量城鎮位置後，發現位於南方的負相關似乎較位於北方的城鎮高，進一步計算發現位於南方城鎮的死亡人數與水中鈣濃度的相關係數為-0.6；而位於北方的兩者之間相關係數為-0.37。

圖1 死亡人數與水中鈣濃度散佈圖

1. 分析方法

從第一章節資料說明提到該分析想了解水中鈣濃度與死亡人數之間的關係，以及城鎮位置是否會影響兩者之間的關係，因此我們把死亡人數當作感興趣的反應變數，而此變數為連續型，將使用迴歸模型進行分析，其模型如下:

其中*Y*表示為死亡人數，表示為水中鈣濃度，表示為城鎮相對位置，為截距項，為偏迴歸係數，當控制某個變數以外的所有變數時，該變數對應變數*Y*的線性影響，為殘差項必須彼此獨立且服從平均數為0變異數為的常態分配。

配適好迴歸模型後需要針對各個變數進行檢定，然而檢定變數是否對死亡人數有影響的t檢定是建立於殘差獨立且服從相同常態分配下所推導出的結果，因此需先檢查殘差是否通過常態性、同質性以及獨立性。

1. 常態性

利用Shapiro-Wilk檢定殘差是否服從常態分配，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，認為殘差通過常態性檢定。

1. 獨立性

利用durbinWatson檢查殘差之間是否存在自我相關，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，認為殘差不存在一階自我相關。

1. 同質性

利用Breusch–Pagan檢查殘差是否符合同質性，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，認為殘差通過同質性檢定。

總結上述殘差是符合前提假設的，進一步做檢定該模型是否有用，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

表示拒絕虛無假設，有足夠證據顯示。因此認為兩者中至少有一個對於應變數是有線性影響的，進一步針對各個變數進行檢查，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，有足夠證據顯示，因此認為水中鈣濃度對於死亡人數是有線性影響的。而另一個變數的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，有足夠證據顯示，因此認為城鎮的相對位置對於死亡人數是有線性影響的。

1. 結論

因此我們可以得到一個迴歸模型如下:

當城鎮相對位置固定時，水中鈣濃度每上升一單位時也就是上升1ppm時，死亡人數平均會減少2人

當水中鈣濃度固定時，位在北方城鎮的死亡人數會較位在南方城鎮的死亡人數多176人。

因此本分析得知，水中鈣濃度與死亡人數是有負向的關係存在，當水中鈣濃度越高時，對於死亡人數有降低的趨勢，但這裡只能保證上限到138ppm，為資料中最高的比例。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，導致其死亡人數較多。