5. 與線性迴歸模型的比較

5.1 建立迴歸模型

爲了瞭解房產均價與四個特徵因子之間是否存在線性迴歸關係，在這邊利用了訂定關聯法則的 99 筆資料來建立線性迴歸模式。在建立線性迴歸模型部分仍然以房產均價與捷運站距離、公園距離、學校距離、醫院距離爲線性迴歸模型的因變數及自變數。應用 MINITAB 14.0 軟題。以下爲 MINITAB 14.0 軟體得出的線性迴歸模型的結果

Regression Analysis: 房產均價 versus 捷運站距離, 公園距離, 學校距離, 醫院距離

房產均價 = 69086 ($\tilde{\beta}_0$) - 0.23 ($\tilde{\beta}_1$) 捷運站距離+ 74.5 ($\tilde{\beta}_2$) 公園距離 - 7.4 ($\tilde{\beta}_3$) 學校距離-10.2 ($\tilde{\beta}_4$) 醫院距離

Analysis of Variance

Source	DF	Seq SS
捷運站距離	1	1239461806
公園距離	1	3733974670
學校距離	1	1552036157
醫院距離	1	3366710341

Source	DF	SS	MS	F	P
Regression	4	9892182974	2473045743	3.49	0.011
Residual Error	94	66567547636	708165400		
Total	98	76459730610			

S = 26611.4 R-Sq = 12.9% R-Sq (adj) = 9.2%

Unusual Observations

Obs	捷運站距離	房產均價	Fit	SE Fit	Residual	St Resid
2	1757	186267	83938	6490	102329	3.97R
11	879	142561	72187	4809	70374	2.69R
18	2490	50350	35334	16033	15016	0.71X
27	823	42825	36740	15427	6085	0.28X
49	652	180062	78203	3689	101859	3.86R
56	3121	34196	29236	17041	4959	0.24X
65	4915	56365	63141	15227	-6775	-0.31X
89	1968	141727	63980	6273	77746	3.01R
91	417	175824	83144	5601	92680	3.56R

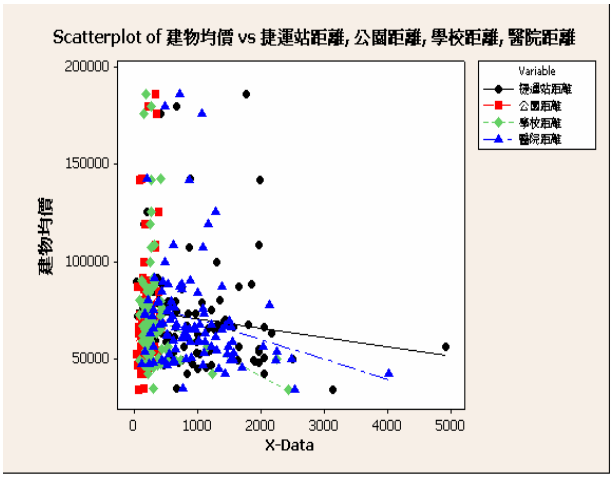
R denotes an observation with a large standardized residual.

X denotes an observation whose X value gives it large influence

Predictor	Coefficient	SE Coefficient	T	P
Constant	69086	8002	8.63	0
捷運站距離	-0.226	4.382	-0.05	0.959
公園距離	74.48	30.68	2.43	0.017
學校距離	-7.45	10.25	-0.73	0.469
醫院距離	-10.164	4.661	-2.18	0.032

歸與各點散佈圖爲圖 5.1 圖形爲下圖所

示：



由 Minitab 14.0 整理出來簡單線性迴歸的結果，可以發現所建立起的線性迴歸模型與特徵因子值相去甚遠，由 Minitab 14.0 跑出的 SSE 值太高，使得整個線性迴歸模式的準確度讓人十分懷疑其精確度。又在表中的 T 表示著各個特徵因子與房產均價建立簡單線性迴歸模型時的領導係數(即 $\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2, \tilde{\beta}_3, \tilde{\beta}_4$) 的是否可以爲 0 的 T 值，倘若以 95%信賴區間來研究，建立假設檢定， $H_0 : \tilde{\beta}_1 = 0$ versus $H_1 : \tilde{\beta}_1 \neq 0$ ，利用雙尾檢定發現， $t = -0.05 < t_{0.025}(96) = 1.96$ ，所以接受 H_0 ，這表示當房產均價與其他特徵因子所建立起的簡單線性迴歸模式 $\tilde{\beta}_1 = 0$ 是被接受的，及被 $\tilde{\beta}_1$ 所領導的捷運站距離對間物均價並無影響，可將其去除。由假設檢定的步驟，利用假設檢定個別驗證每一個領導係數是否對房產均價有影響，發現唯有公園距離對房產均價有影響，及其他特徵因子前的領導係數均可爲 0。

而在離群值部分，有 9 個影響著我們所建立的簡單線性迴歸模型，其狀況分別有 R 所代表的其觀測值與標準誤差中有極大的問題，因爲標準誤差是由平方差形成的，而所謂的標準誤差其實是代表著利用最小平方法所建構而成的迴歸線與觀測值之間的差距，在這 99 筆資料中有 5 個觀測值與迴歸線的標準誤差過高，其實就是一邊的離群值。而 X 代表的是此觀測值對於整體觀測值有著很大的影響，因爲此觀測值對於整個觀測值的平均值有很大的影響，若將此觀測值隨意刪除不加入其中探討，將會導致整個迴歸模式參數的估計上有很大的誤差。無論是帶有 R 值或帶有 X 值的觀測值均會在建構迴歸模型極大的影響，因此最好的方法應爲，先將這 9 筆觀測值刪除，再將每一種不同的特徵因子個別與房產均價做一簡單迴歸模型，將得出的結果利用假設檢定的方法，判別是否與房產均價相關，方能建立最有效且最具信度的迴歸模型。

5.2 與迴歸模型的討論

在大樣本資料中若想尋求出各個特徵因子與均價之間的關係，除了可使用上述的關聯法則外，尚有許多方法可做，如以多變量函數模擬出適當的模型或利用線性迴歸模式將特徵因子與房產均價建立一線性模式，用以預測均價與各特徵因子之關係…等。在這邊將提出線性迴歸模式，用以討論線性迴歸及關聯法則之優缺點。

線性迴歸模型的優劣

優：在上述線性迴歸模型可看出線性迴歸模型之優點在於，它不需經過特徵因子的分組來探討均價與特徵因子之間的關係，此方法較爲直接及易於瞭解，而且線性迴歸模型至古典統計發展至今，迴歸模式有著非常嚴謹的制約，所以在建立起線性迴歸模型後，此模型將有著一定的可靠性及信度。

劣：由於多類因子的線性迴歸得出的模型，如上節所得出的線性迴歸方程式，

房產均價 = 69086 - 0.23 捷運站距離 + 74.5 公園距離 - 7.4 學校距離 - 10.2 醫院距離

Predictor	Coefficient	SE Coefficient	T	P
Constant	69086	8002	8.63	0
捷運站距離	-0.226	4.382	-0.05	0.959
公園距離	74.48	30.68	2.43	0.017
學校距離	-7.45	10.25	-0.73	0.469
醫院距離	-10.164	4.661	-2.18	0.032

R-Sq = 12.9%　R-Sq (adj) = 9.2%

可看出捷運站距離及學校距離的 P-value 過高及 R-Sq = 12.9% ，R-Sq (adj) = 9.2%，代表著在 MINITAB14.0 所得的線性迴歸方程式，在捷運站距離及學校距離並未與房產均價有著密切的關係，甚至從 R-Sq = 12.9%中發現這兩個特徵因子影響著整個線性迴歸模型的不準確性。倘若想讓此線性迴歸模型的 R-Sq 降低，除了必須先將房產均價與每一個特徵因子建立線性迴歸模型，考慮每一個線性迴歸模型的 R-Sq，才能做出房產均價與哪幾個特徵因子間有著相關性。倘若仍然找不出準確的線性迴歸模型，則並需將特徵因子中的所有資料作一轉換，接著在去討論各個轉換後的特徵因子與房產均價的線性迴歸模型，是否其 R-Sq 有一理想的值。當每一個特徵因子(或轉換後的特徵因子)與房產均價的線性迴歸模型可順利建構，接下來才會建構多個特徵因子與房產均價線性迴歸模型。這個流程不但費時，而且建構出來多個特徵因子(或轉換後的特徵因子)與房產均價間的線性迴歸模型也有可能因爲離群值極具有相當影響力的觀測值將左右整個迴歸模型，使得線性迴歸模型並不一定十分準確。且由於假設檢定中可檢定其參數是否可爲 0，即此參數是否對迴歸模型有影響，由 5.1 節所討論的結果，唯有公園距離通過假設檢定

的檢驗，代表著整個均價區間只與公園距離有迴歸關係，其他特徵因子並無影響，這不但無法解釋多個特徵因子區間與均價區間的關係，反而讓整個迴歸模型更爲狹隘。

關聯法則的優劣

優：關聯法則可以將多個特徵因子與房產均價間的關係一次釐清，無論在布林值關聯法則，亦或者是複合維度的關聯法則均可建立。且在整個過程當中，不但步驟明確且相當合理，將有可能出現的關聯法則全部建立後，再做比對在，誤差方面單一不準確的特徵因子並不會否決關聯法則的正確性，更說明關聯法則的多樣化及人性化，如此說來關聯法則不但準確度較線性迴歸模式來的好，討論起來也較爲全面，線性迴歸模式因受古典統計的定義及限制較多，更顯得關聯法則使用起來較爲人性化，且結果也較爲直接。

劣：在關聯法則中，若每一個特徵因子區間的劃分過於繁多，將導致在複合維度關聯法則中找不出具有信度及效度的關聯法則，唯有減少特徵因子區間的個數或加大每一個特徵因子區間方能取得有效的關聯法則。還有在資料的離群值方面，有些離群值不但影響了關聯法則的建立，還會造成整個關聯法則的不準確，所以在資料的收集方面還需再做進一步的篩選，排除離群值，以便建立更有支持度及信度的關聯法則。

由各個層面來看，雖關聯法則的應用與迴歸模型的應用各有利弊。但在多個特徵因子討論時，迴歸模式較關聯法則來的單調，甚至無法全面的探討到多個特徵因子與均價區間的關係，而迴歸模式無法建立有效的模型，這樣子的迴歸模型並未滿足購屋消費者與房地產業者的需求，所以整體而言關聯法則真的能幫助到購屋消費者與房地產業者的需求。