112fir第7組迴歸分析報告

摘要

此研究主要想了解波士頓房價資料集經過變數選取法、離群值和影響點的移除後，嘗試進行殘差分析與滿足線性回歸的三大假設。後使用欲先分割的資料集嘗試驗證模型的有效性，最終想了解能有效影響波士頓房價的變數以及這些變數影響的強度。

目錄

第壹章 緒論

第一節 前言

當談及波士頓的房價時，我們進入了一個動態、充滿活力的話題。波士頓不僅是一個擁有豐富歷史、文化底蘊的城市，更是科技、教育的據點。在這片土地上，每一座建築都承載著過去與現在的交織，而房價則成為一個窗口，透視這座城市的經濟、社會脈動。

這份報告將探討波士頓房價的複雜面向，從市場趨勢到區域差異，以及房價背後可能隱含的故事。過去，波士頓房地產市場是否繼續保持穩健，或者經歷了何種變遷？我們將透過數據分析和專業見解，助您全面瞭解這座城市的房價格局。

波士頓的房價不僅僅是一個經濟指標，更是居民生活、城市發展的一個重要指標。透過這份報告，我們將探討波士頓房價的背後故事，揭示其與城市發展、社會變遷的密切關聯。

第二節 研究動機與目的

研究動機:

本研究的啟發來自於對波士頓房地產市場的好奇與求知慾，考慮到當前社會經濟環境的不斷變遷，以及房價波動對個人和機構的深遠影響。波士頓地區作為一個重要的經濟樞紐，其房地產市場的穩健運作對於區域經濟發展至關重要。因此，透過深入研究波士頓房價資料集，我們迫切希望解析影響房價波動的各種因素，從而獲得對市場變化的更深層次理解。

此外，隨著科技進步，我們能利用電腦程式，使原本需要龐大人力以及時間才可進行的複迴歸分析，只要運用一些電腦程式進行運算，便可透過十人不到的小組建構一個準確而可靠的迴歸模型，以預測未來波士頓房價的走勢。這對於房地產業者、政府機構、投資者以及一般市民都具有實質價值，有助於制定明智的投資和政策決策。因此，透過深入分析波士頓房價資料，我們尋求揭示潛在趨勢和關聯性，以促進更為智慧和可持續的城市發展，同時提供更精確的預測工具，以應對不確定性的挑戰。這個研究動機不僅有助於學術界對於房地產市場的理論探索，同時也對實際應用和社會發展產生積極而深遠的影響。

研究目的:

本研究旨在利用波士頓房價資料集，以房價作為反應變數，探索並確定其中的解釋變數，以構建適切的迴歸模型。透過精確挑選相關變數，本研究尋求建立一個有效的統計模型，以深入理解房價與其他相關因素之間的關聯性。最終目標在於提高預測模型的準確性，以滿足實際需求，並為房價變動提供可靠的解釋。透過此研究，期望能夠為不同需求背景下的房價預測提供實質可行的方法，進一步拓展對於波士頓房地產市場的深入了解。

第三節 研究對象

變數解釋:

CRIM(以城鎮劃分的人均犯罪率):

衡量該地區城市犯罪率的變數，犯罪率的高低可能影響房價，因為人們傾向於選擇住在相對安全的社區。

ZN(面積超過 25,000 平方英尺的住宅用地比例):

表示居住區的土地面積比例，可能影響該地區的房屋密度和居住環境。

INDUS(每個城鎮非零售商業面積的比例):

指示城市中非零售業商用土地的比例，可能反映了該地區的工業化程度。

CHAS(查爾斯河虛擬變數-是否鄰近查爾斯河(如果區域邊界為河流，則為 1;否則為 0)):

是一個虛擬變數，表示房屋是否鄰近查爾斯河，可能與景觀、環境品質有關。

NOX(一氧化氮濃度(千萬分之一)):

衡量空氣中的一氧化氮濃度，可能與環境品質和居住健康相關。

RM（每套住宅的平均房間數）：

反映了房屋的大小，對於房價有直接的影響，通常房間數越多，房價越高。

AGE（1940 年以前建成的自住房屋比例）：

表示社區中老舊房屋的比例，可能與居住狀態和建築品質相關。

DIS（到五個波士頓就業中心的加權距離）：

考慮到就業機會的接近程度，對於房價有一定影響。

RAD（高速公路的可及性指數）：

表示附近高速公路的便利性，可能影響居民的交通狀況和生活便利度。

TAX（每 10,000 美元的全額財產稅率）：

衡量地區的財政負擔，高稅率可能影響購房者的選擇。

PTRATIO（以城鎮劃分的師生比例）：

反映當地的教育資源和教育品質，對於有子女的家庭可能影響房屋選擇。

B（1000(Bk - 0.63)^2 其中 Bk 是按城鎮劃分的黑人比例）：

衡量社區的種族結構，可能影響社區的多樣性和文化氛圍。

*因SAS程式對 'B'有預設功能，後基本以BLAC或blac表示

LSTAT（社會地位較低的人口占全部的百分比）：

反映社區中低收入人群的比例，可能與社區的經濟狀況和房價相關。

MEDV（自住房屋的中位數價值（1/1000 美元））：

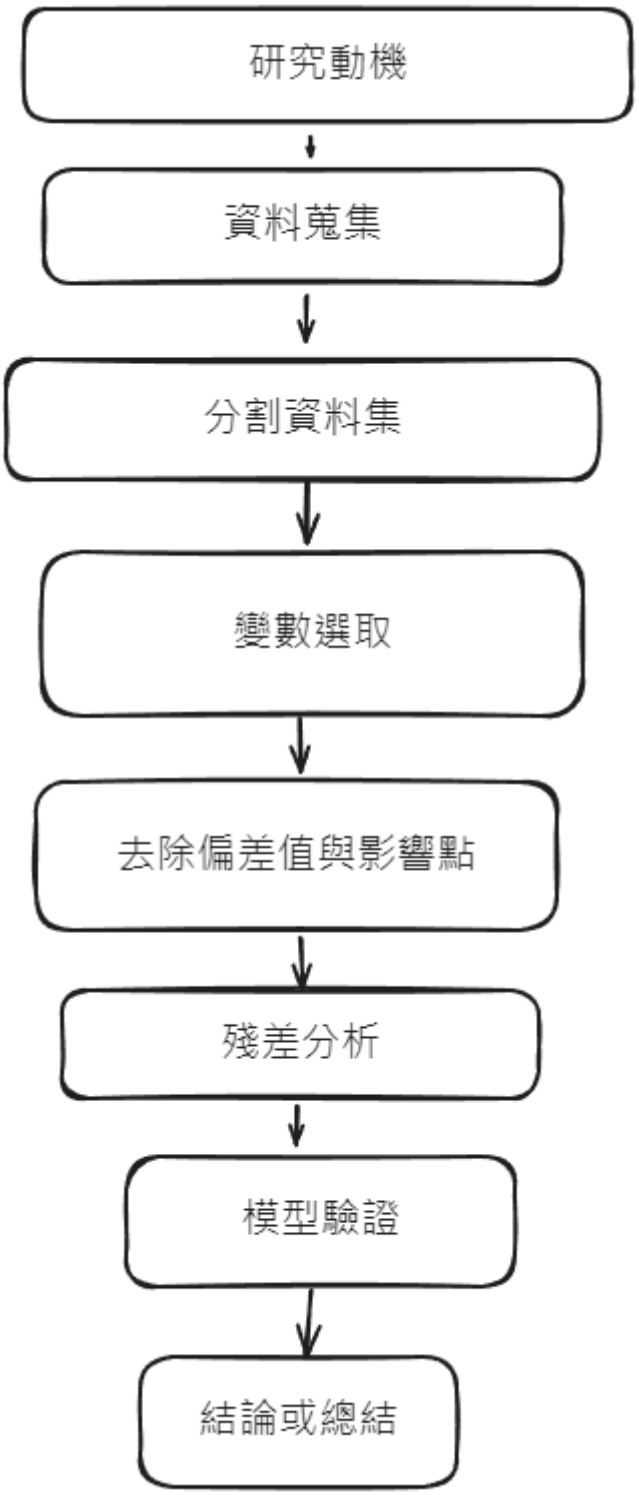
是本研究的反應變數，代表自住房屋的中位數價格，是研究房價波動的核心目標。

這些特徵提供了綜合性的信息，可用於深入分析波士頓不同區域的房地產市場特徵及其對房價的影響。

第四節 研究方法

我們根據多倫多大學提供的完整資料集為基準。首先，使用R把原始資料隨機分割成9:1的訓練集和測試集(原始資料集共506筆，訓練集456筆，測試集50筆資料/觀察值)，進行基本資料分析，。之後建立以房屋中位數價值(MEDV)為反映變數的初步的回歸模型並參考各種選取法以R^2為參考的參數用來篩選掉不合適的解釋變數，並將同時具有離群值及影響點特值的樣本刪除，得到最適迴歸模型後進行齊一性、常態性、獨立性三項檢定。用以預測訓練集資料。

第五節 研究架構



第貳章 基本資料分析

第一節 基本敘述統計量

簡單統計值							
變數	N	平均值	標準差	總和	最小值	最大值	標籤
MEDV	456	22.58618	9.14159	10299	5.00000	50.00000	MEDV
CRIM	456	3.69767	8.91517	1686	0.00630	88.97620	CRIM
ZN	456	11.42325	23.36337	5209	0	100.00000	ZN
INDUS	456	11.03110	6.91879	5030	0.74000	27.74000	INDUS
CHAS	456	0.07018	0.25572	32.00000	0	1.00000	CHAS
NOX	456	0.55451	0.11650	252.85570	0.38500	0.87100	NOX
RM	456	6.29093	0.70239	2869	3.56100	8.72500	RM
AGE	456	68.58860	28.31564	31276	2.90000	100.00000	AGE
DIS	456	3.80823	2.13934	1737	1.13700	12.12650	DIS
RAD	456	9.44737	8.67822	4308	1.00000	24.00000	RAD
TAX	456	406.42544	168.70166	185330	187.00000	711.00000	TAX
PTRATIO	456	18.43180	2.15662	8405	12.60000	22.00000	PTRATIO
BLAC	456	359.13844	88.69873	163767	0.32000	396.90000	BLAC
LSTAT	456	12.65057	7.24054	5769	1.73000	37.97000	LSTAT

一、自住房屋的中位數價值 (MEDV)($)

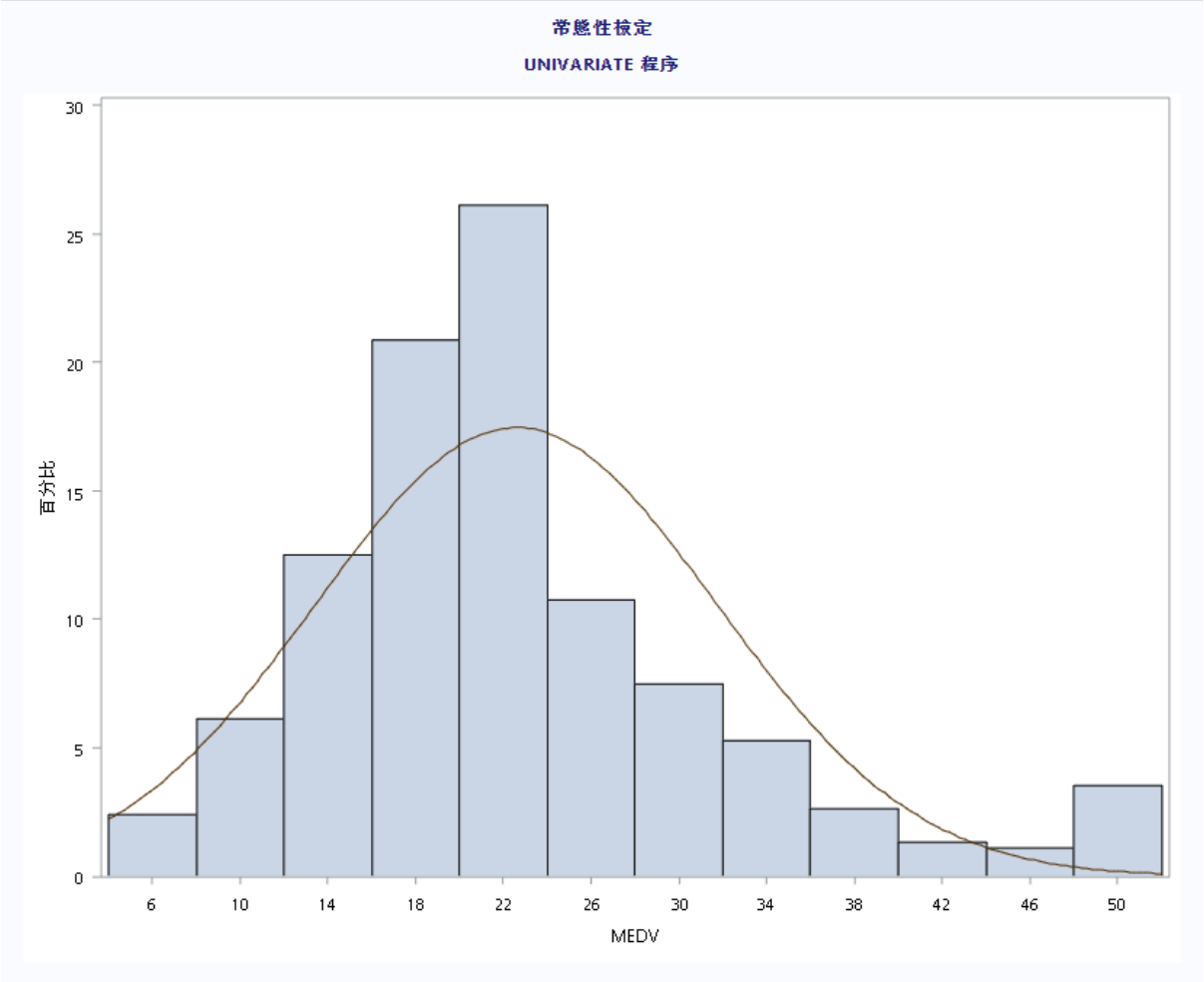


圖2-1(MEDV)相對直方圖

由上圖可知資料呈現右尾分配，資料以價值$22000的自住房屋數量占比最高，而占比最低則是價值$46000的自住房屋。

二、犯罪率(CRIM)(%)

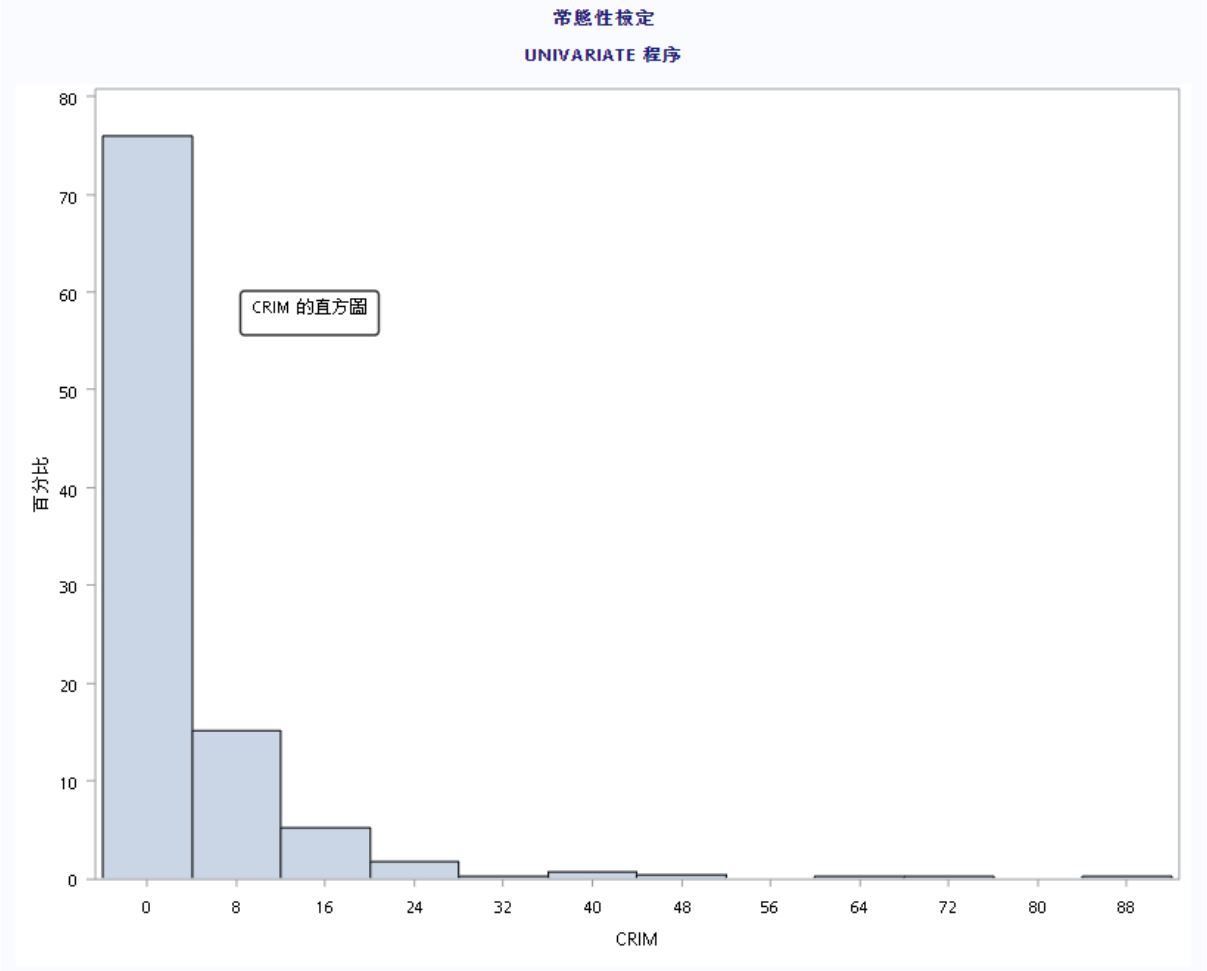


圖2-2(CRIM)相對直方圖

由上圖可知呈現右尾分配，以犯罪率0%為最大占比，可得知大部分的城鎮地區犯罪率接近0%，房價可能較高，而犯罪率88%的城鎮地區的房價可能較低。

三、住宅用地比例(ZN)

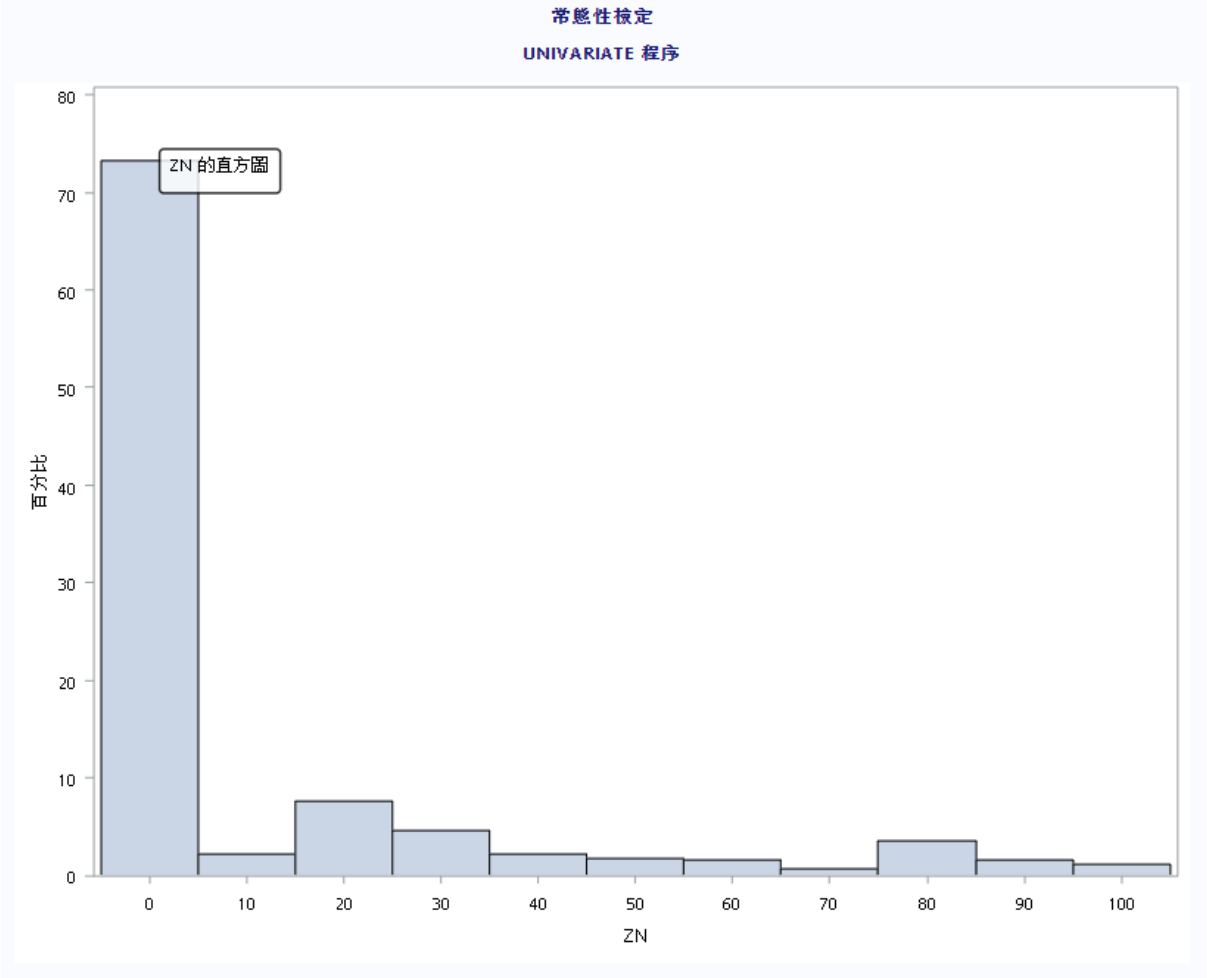


圖2-3(ZN)相對直方圖

由上圖可知呈現右尾分配，以住宅用地比例0%為最大占比，可得知大部分的住宅用地比例接近0%，可能城鎮都聚集在同一地區，其他非城鎮地區可能為荒野。

四、城鎮非零售商業面積的比例(INOUS)

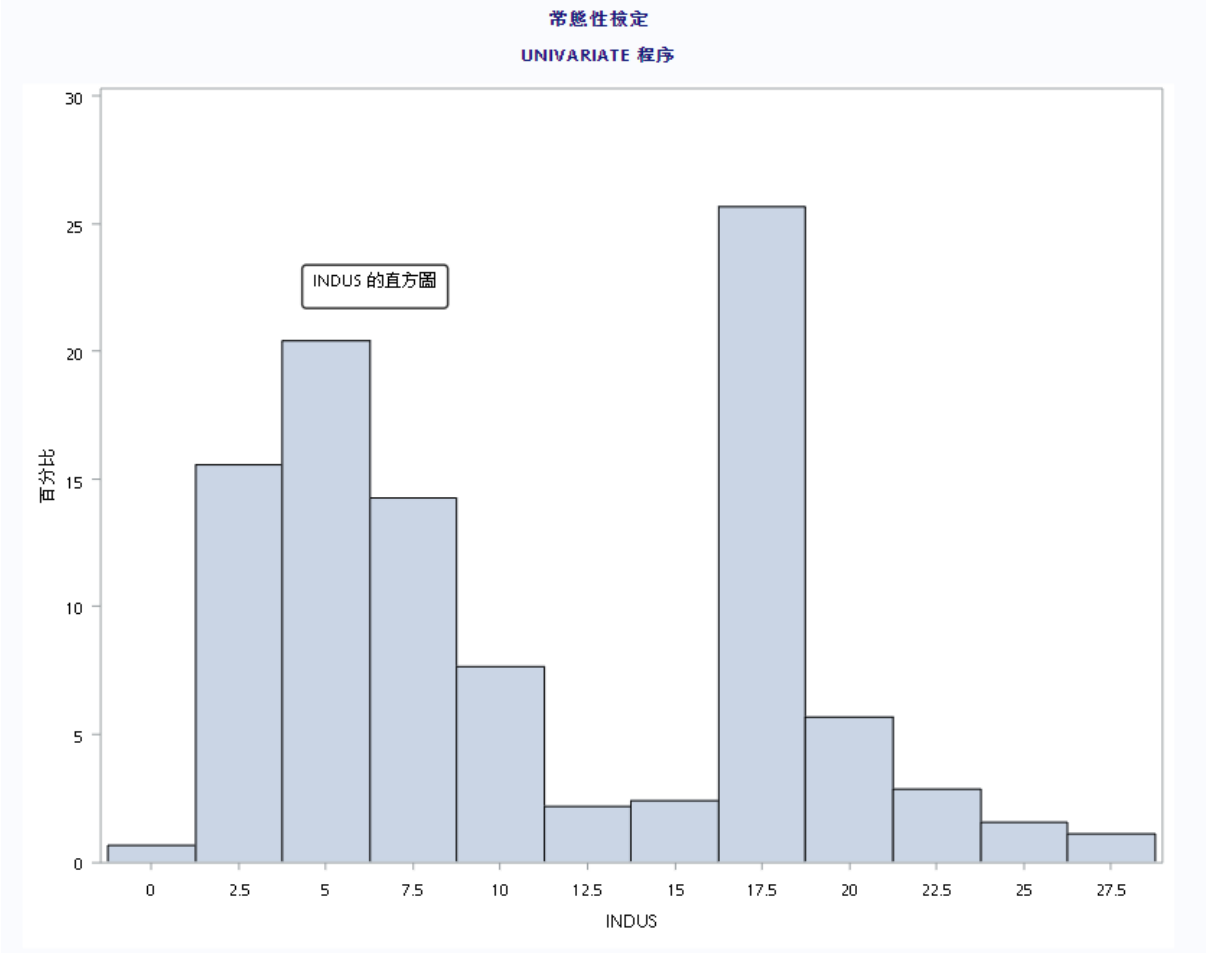


圖2-4(INOUS)相對直方圖

以比例17.5%的城鎮地區數量最多，高比例可能表示一個地區以商業活動為主，而低比例則可能表示主要為住宅區或其他用途。

五、查爾斯河虛擬變數-是否鄰近查爾斯河(CHAS)

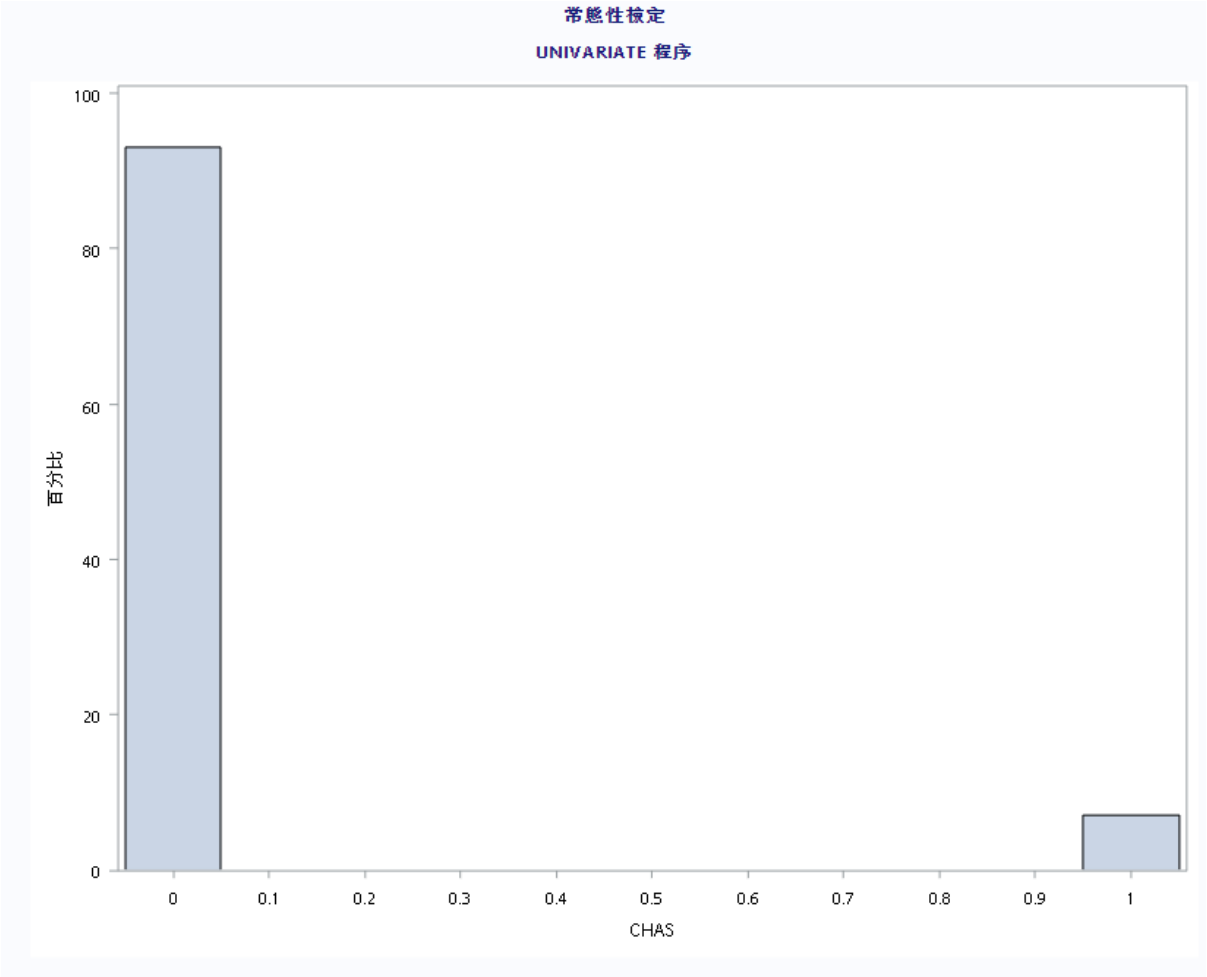


圖2-5(CHAS)相對直方圖

如果區域邊界為河流，則為 1;否則為 0。

鄰近查爾斯河，可能影響景觀、環境品質。

六、一氧化氮濃度(NOX)

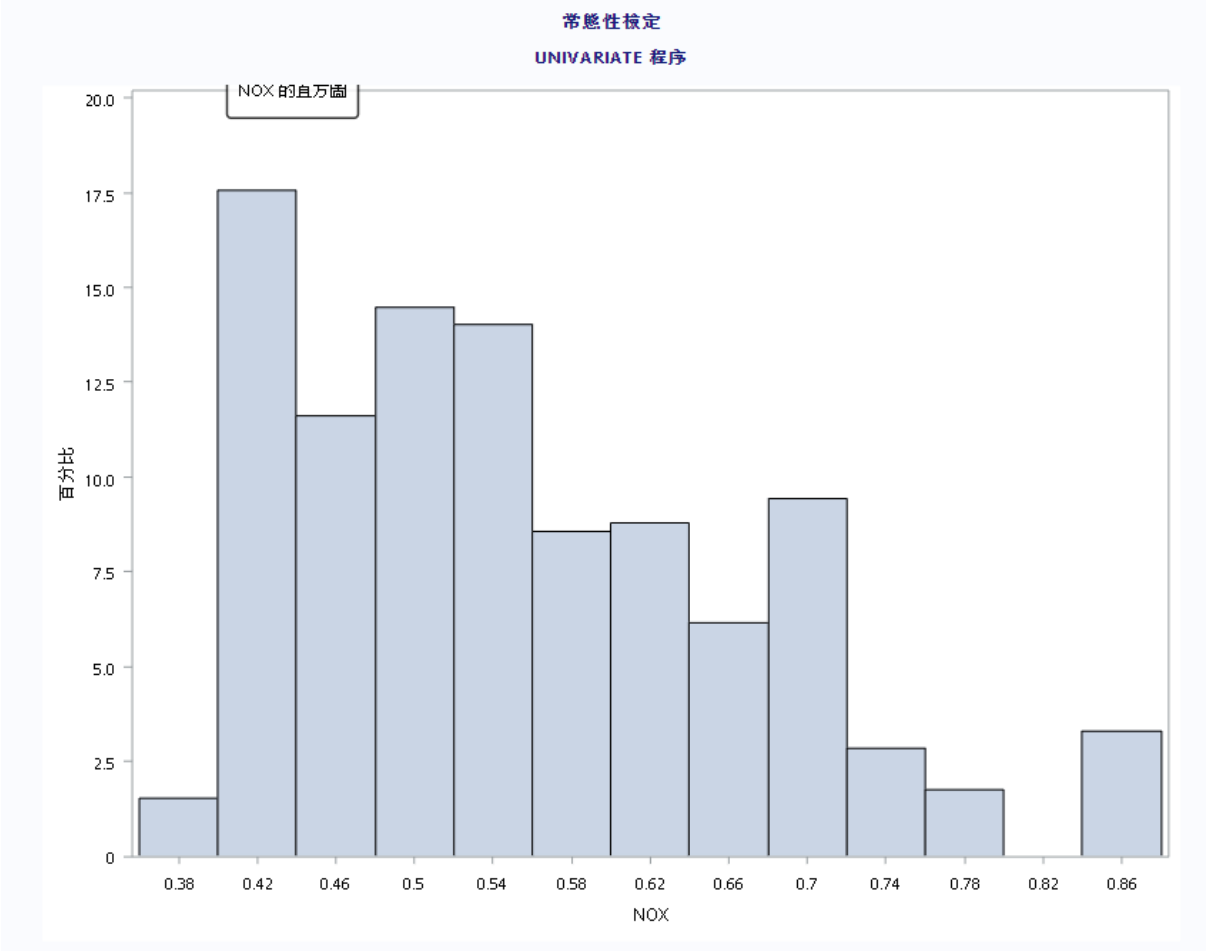


圖2-6(NOX)相對直方圖

一氧化氮的濃度可能受到以下因素的影響:

1.交通排放

2.工業排放。

3.自然源燃燒過程

4.燃燒過程

高濃度的一氧化氮可能代表都市化程度越高，反之則越低;大部分的地區一氧化氮濃度皆達於7.5%。

七、每套住宅的平均房間數(RM)

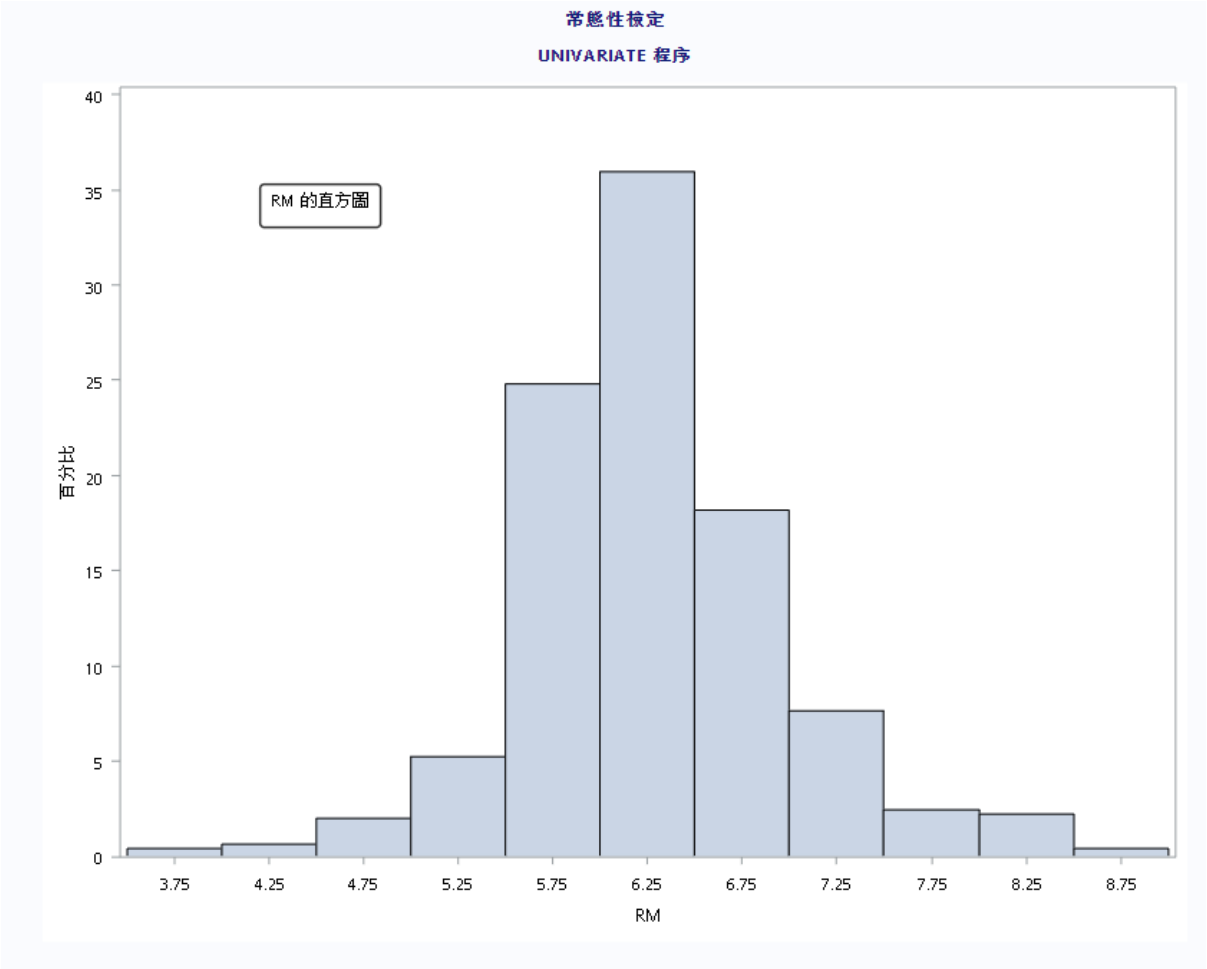


圖2-7(RM)相對直方圖

由上圖可知近似鐘形分配，集中在5.75~6.75個房間數，表示大部分的家庭皆屬於這個情況。

八、1940 年以前建成的自住房屋比例(AGE)

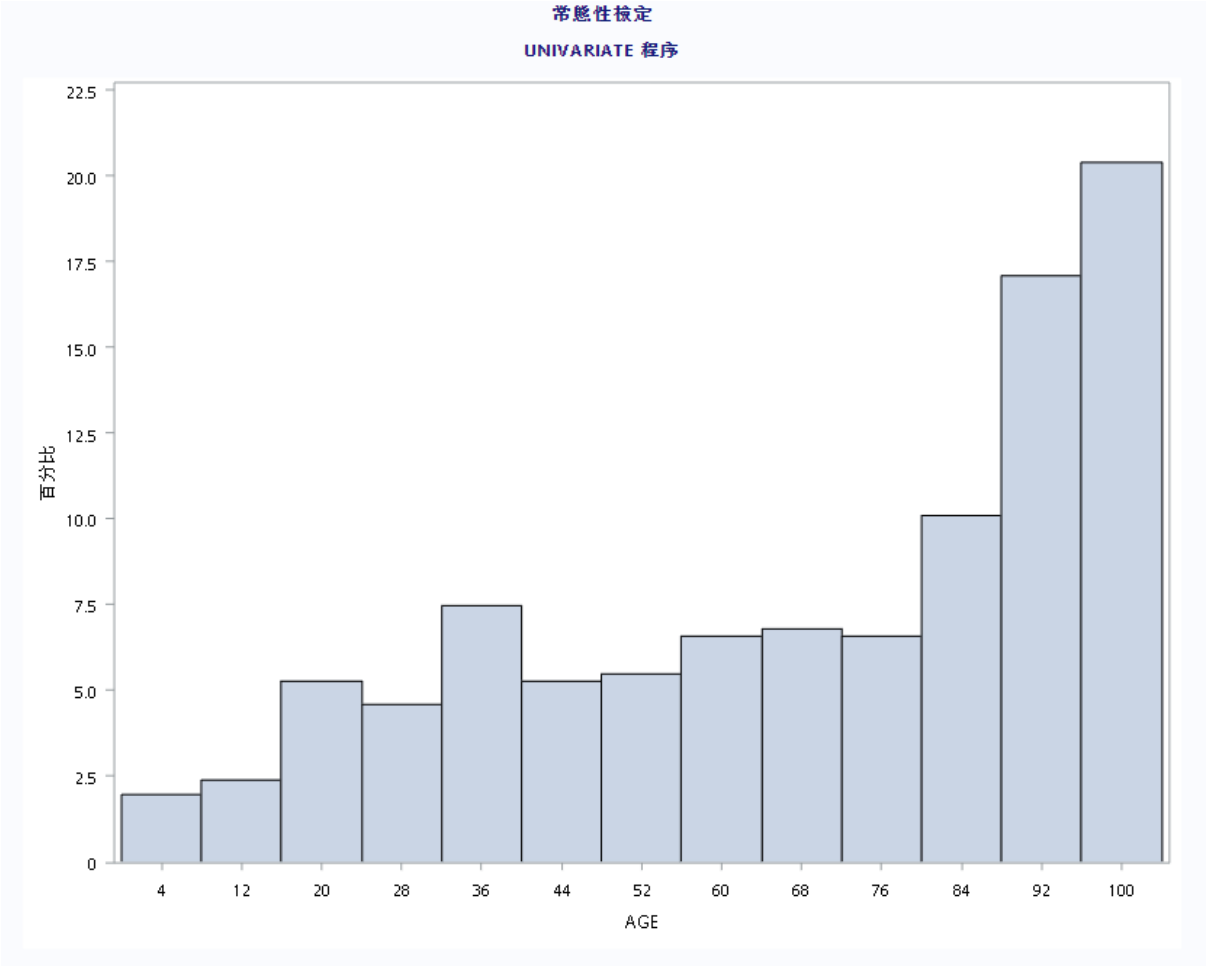


圖2-8(AGE)相對直方圖

由上圖可知呈現左尾分配，為100%皆為老舊社區的社區最多，較高比例的老舊建築可能意味著社區的整體建築品質和設施可能較為陳舊。

九、到五個波士頓就業中心的加權距離(DIS)

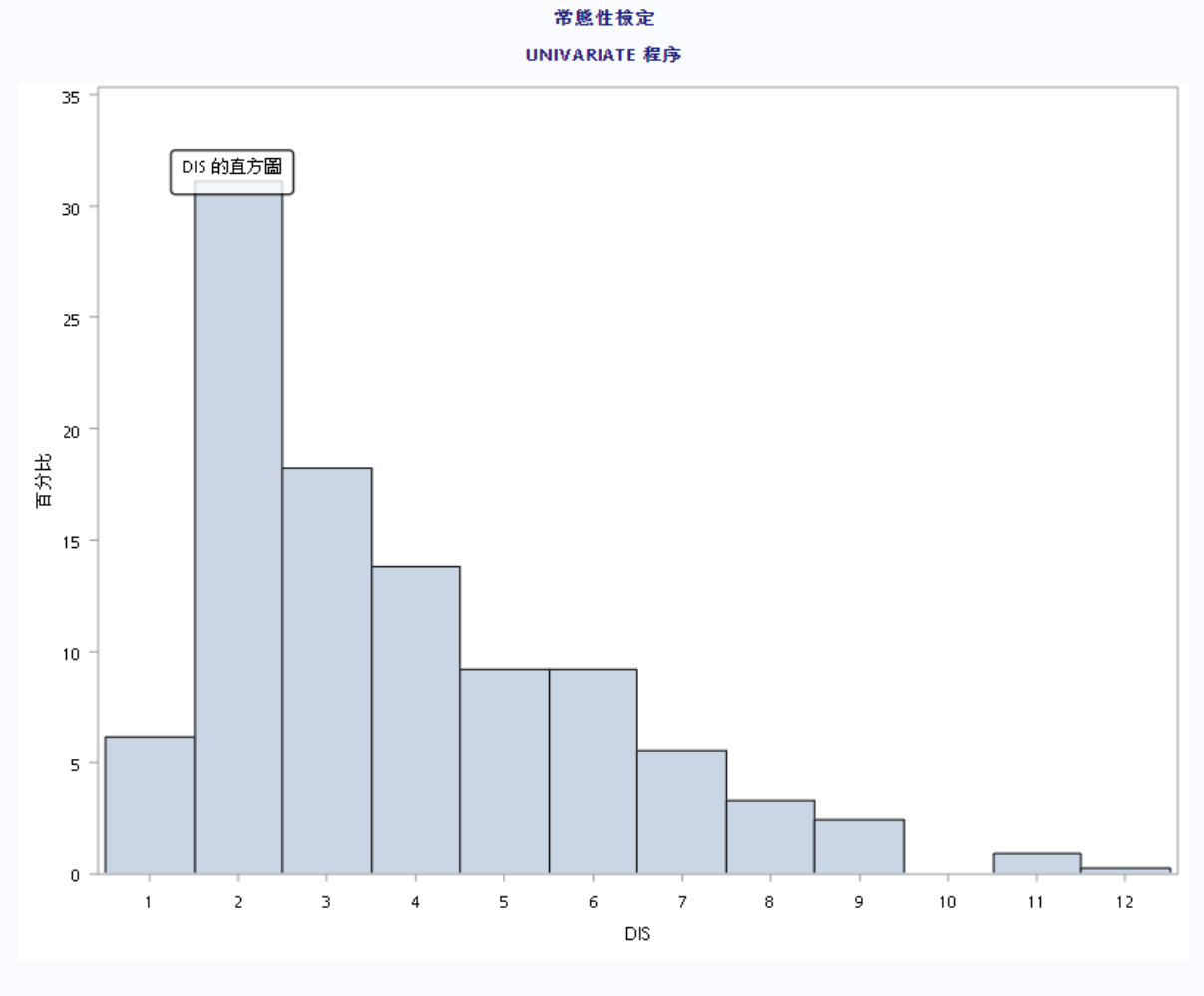


圖2-9(DIS)相對直方圖

由上圖可知呈現右尾分配，以加權距離=2　的比例最高，可能意味著就業中心的距離對於房價和居住者的便利性有一定的影響，加權距離越低房價也可能越高。

十、高速公路的可及性指數(RAD)

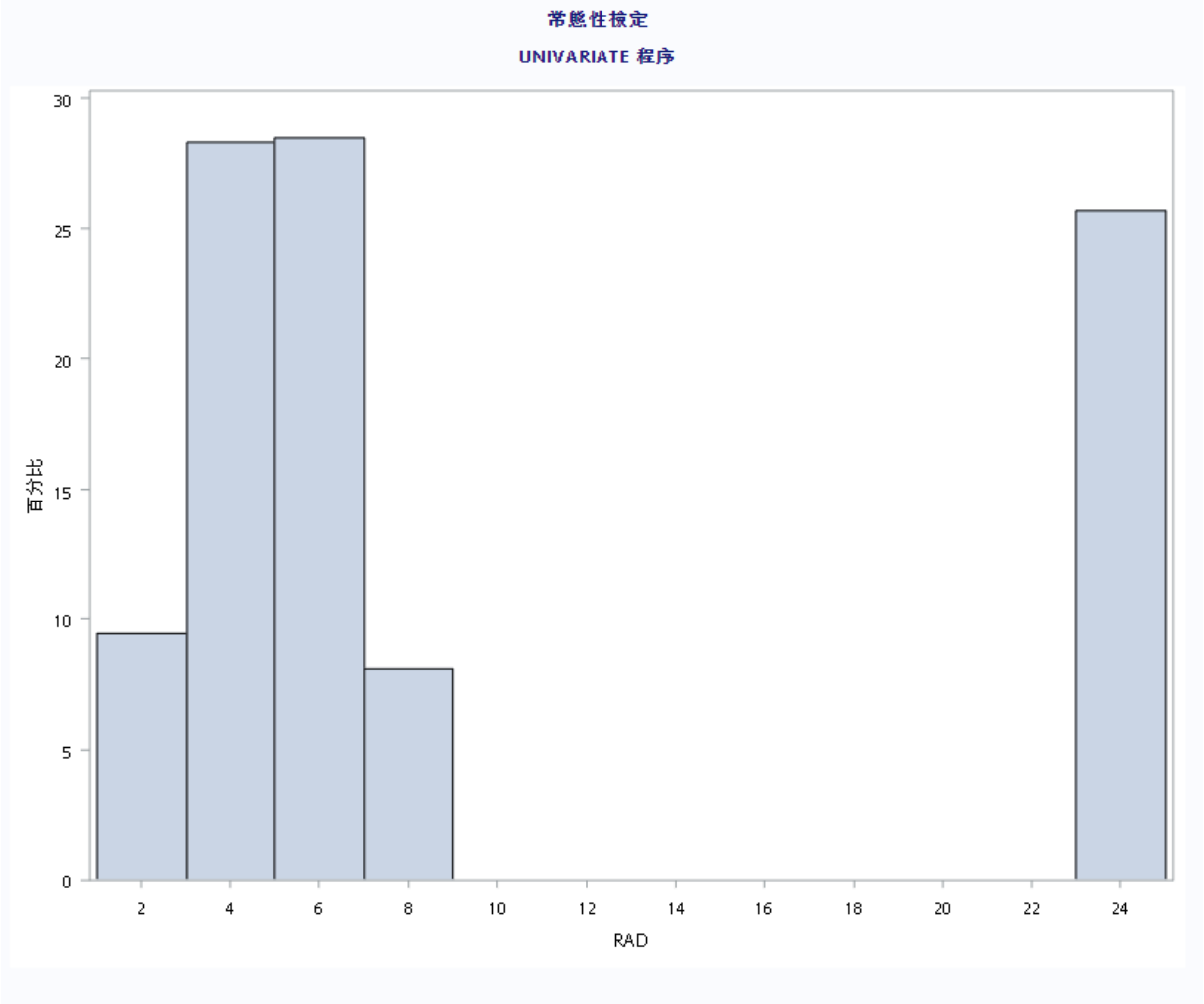


圖2-10(RAD)相對直方圖

以RAD=6的比例最高，當RAD值較高時，可能意味著該區域的居民更容易使用高速公路，通勤和生活更加便利。這對於評估房地產價值、居住品質和城市規劃都具有重要意義。相反，較低的RAD值可能表示居民在交通和出行方面可能面臨一些挑戰。

十一、全額財產稅率(TAX)

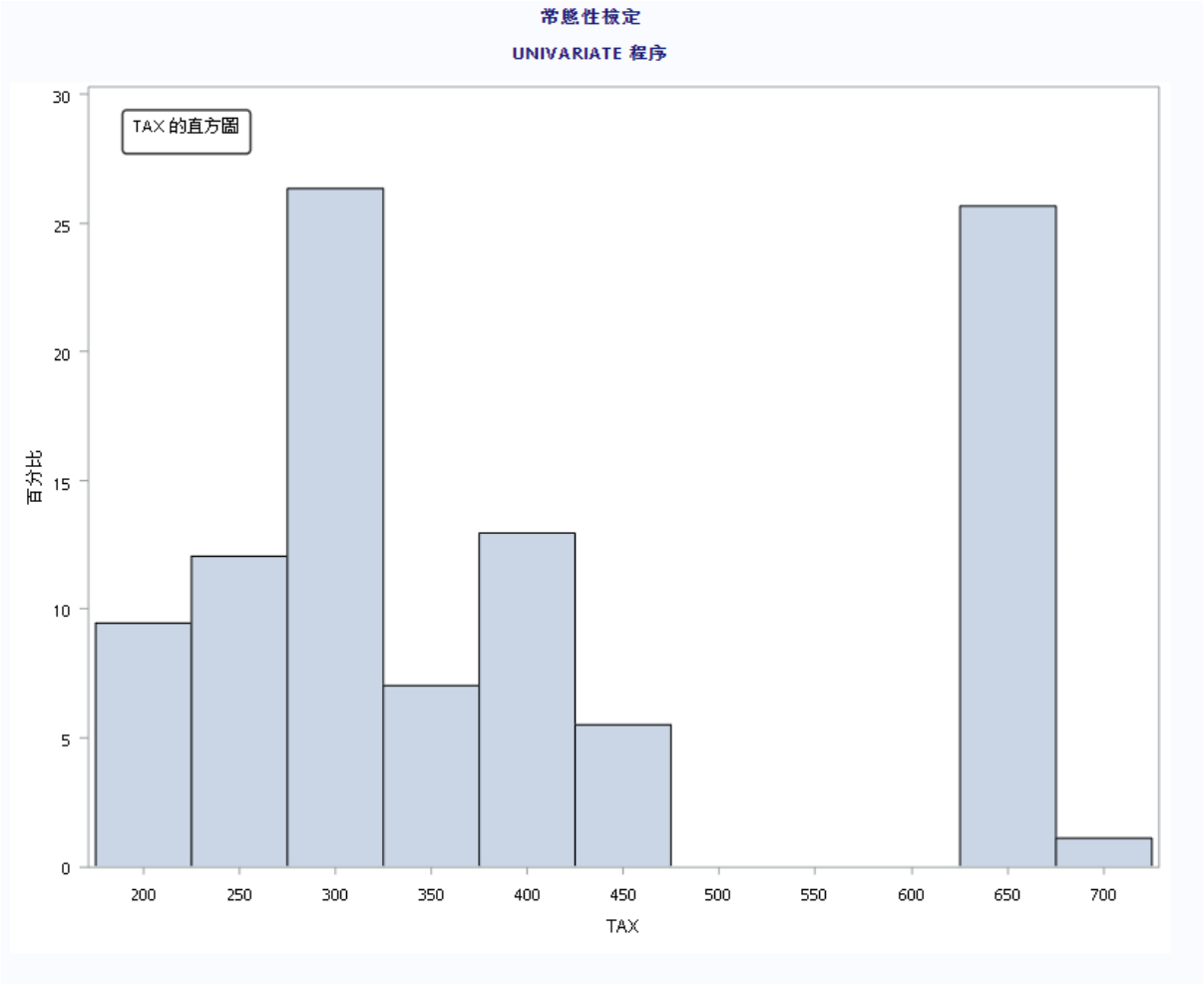


圖2-11(TAX)相對直方圖

這一指標可能對購房者的選擇產生影響，因為高稅率通常會增加擁有房產的成本，可能影響人們購房的意願。

同時又以TAX=300的比例最高，可能代表人們購房的意願此時達到最高。

十二、以城鎮劃分的師生比例(PTRAYIO)

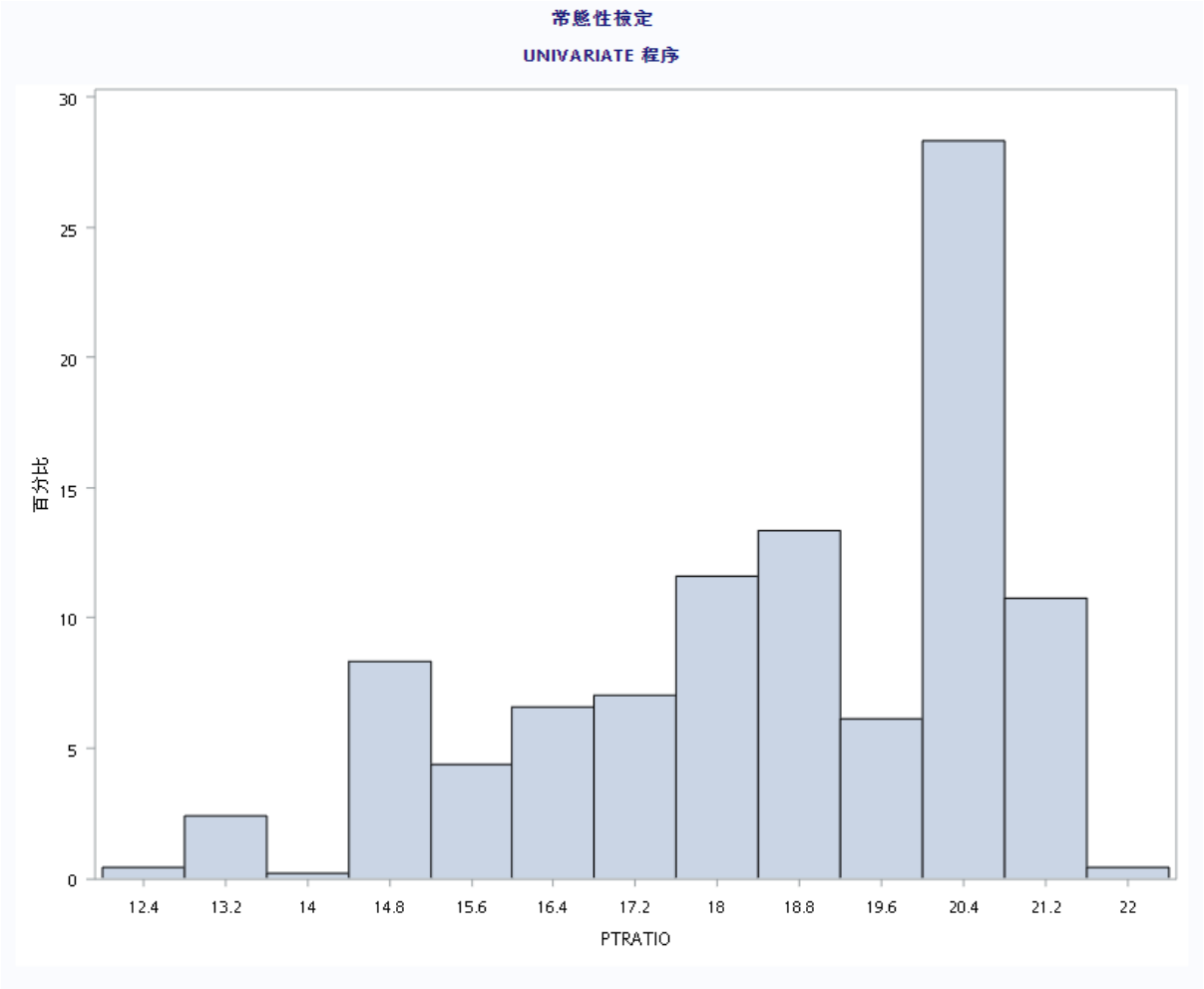


圖2-12(PTRAYIO)相對直方圖

這一比例通常用於反映當地的教育資源和教育品質，對於有子女的家庭可能影響他們的房屋選擇。同時又以比例20.4%的數量為最多，可得知大部分的家庭都選擇居住在此區域。

十三、按城鎮劃分的黑人比例(BLAC)

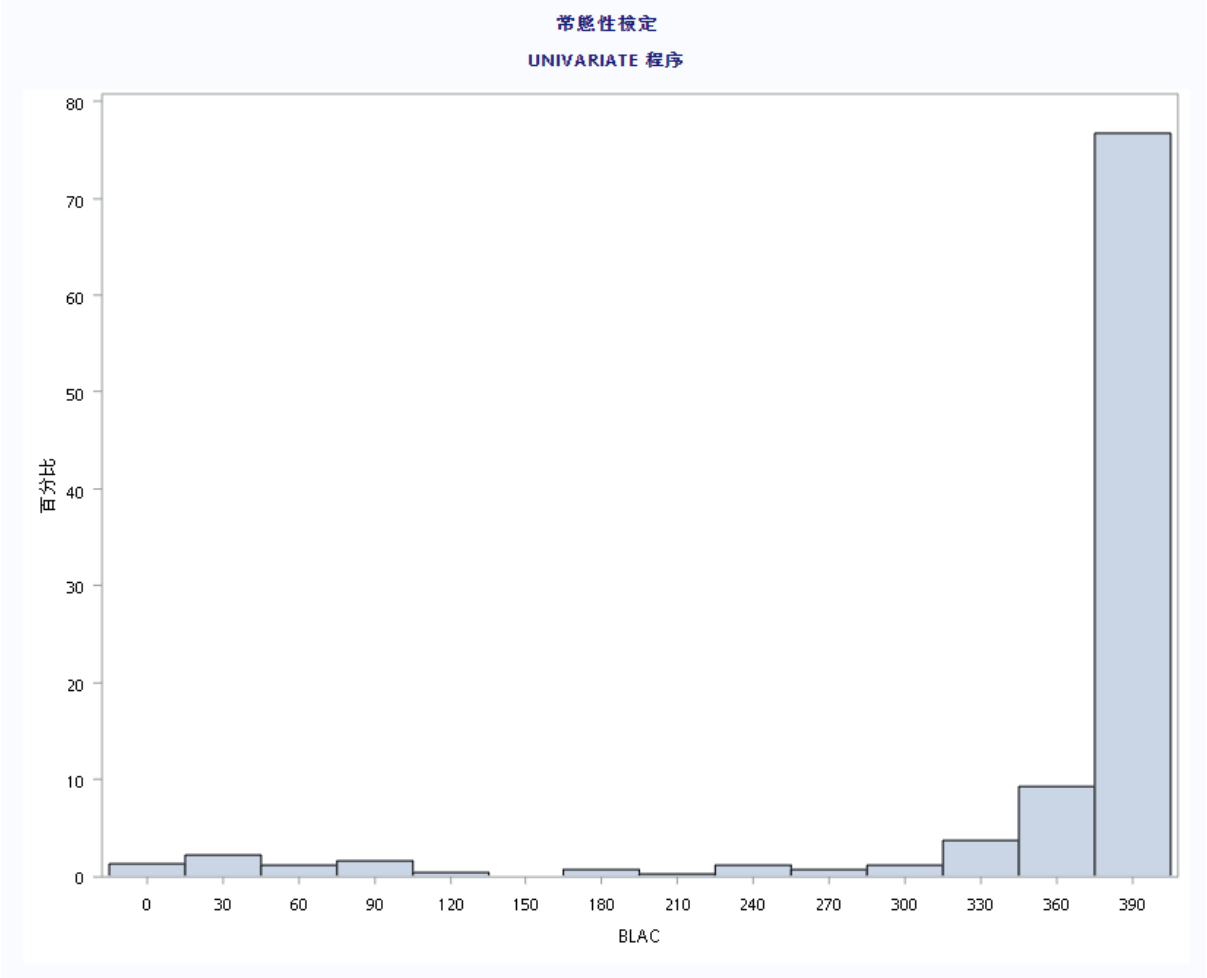


圖2-13(BLAC)相對直方圖

由上圖可知呈現左尾分配，其中以B=390的數量為最多，可能代表代表大部分的城鎮黑人比例都占多數，可能連帶影響房價高低。

十四、社會地位較低的人口占全部的百分比(LSTAT)

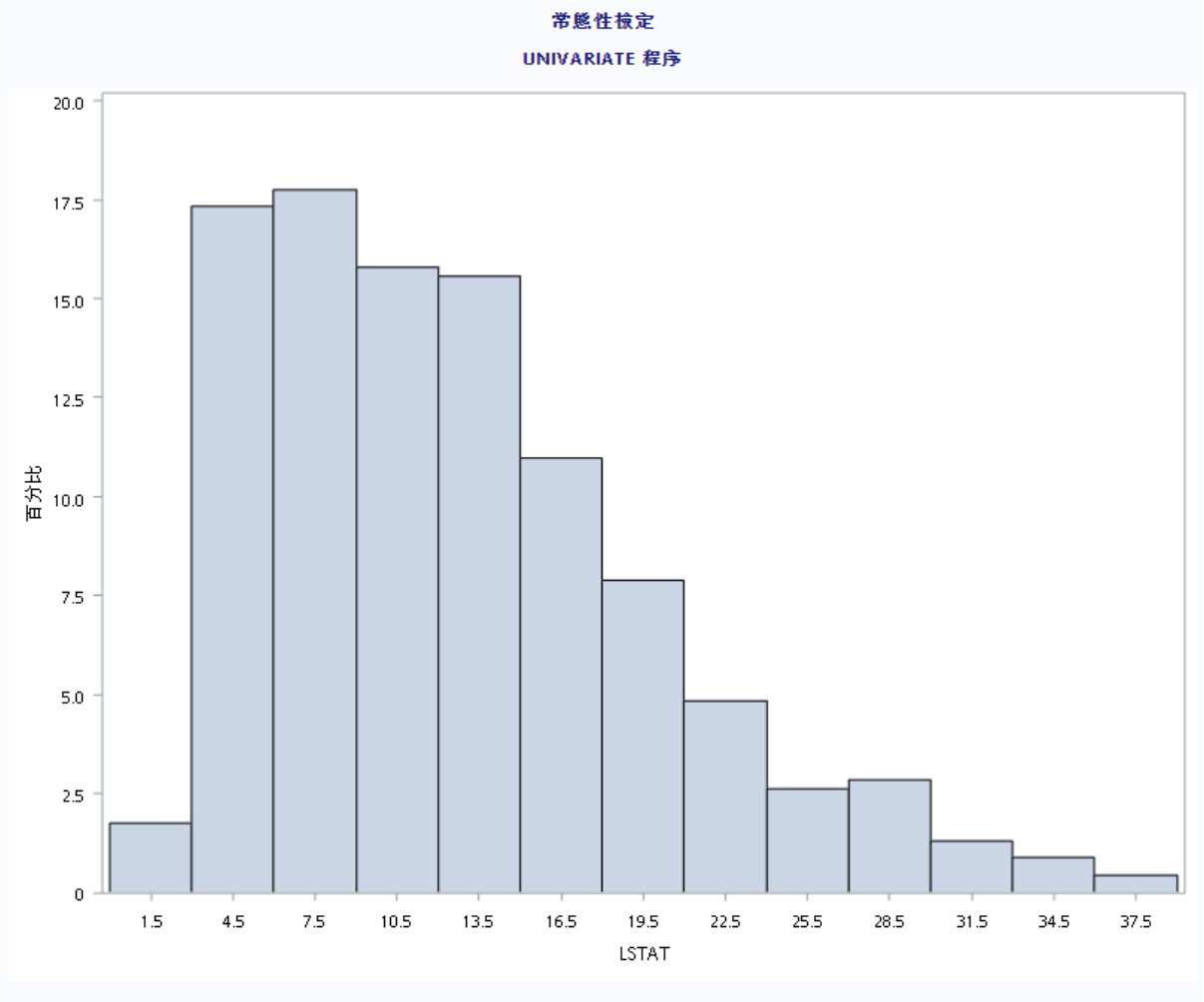


圖2-14(LSTAT)相對直方圖

由上圖可知呈現右尾分配，可得知大部分的地區，社會地位較低的人口占人口的百分比較高，房價也可能連帶降低，反之百分比越低，房價則越高。

第二節 CORRELATION

「相關係數」用於描述兩數值變數X、Y間的關係強度，一般用r來表示，範圍介於-1與1之間。相關係數的正負符號代表變數間關係的方向。正稱為「正相關」，表示兩變數為正向關係；負則稱為「負相關」，意指兩變數為反向關係。而數值大小則為相關性的強弱。當數值越接近”1”，表示相關性越高；若數值越接近”0”，則相關性越低。其判定原則如下:

|r|=1	完全相關
0.7＜|r|＜1	高度相關
0.3＜|r|≦0.7	中度相關
|r|≦0.3低度相關r=0	零相關(無線性關係)

Pearson 相關係數, N = 430
Prob > |r| (低於 H0): Rho=0

	MEDV	CRIM	ZN	CHAS	NOX	RM	DIS	RAD	TAX	PTRATIO	BLAC	LSTAT
MEDV MEDV	1.00000	-0.48150 <.0001	0.39107 <.0001	0.07115 0.1408	-0.49945 <.0001	0.75083 <.0001	0.31539 <.0001	-0.47995 <.0001	-0.58088 <.0001	-0.53947 <.0001	0.33339 <.0001	-0.76989 <.0001
CRIM CRIM	-0.48150 <.0001	1.00000	-0.21999 <.0001	-0.06623 0.1704	0.46876 <.0001	-0.25986 <.0001	-0.41066 <.0001	0.69806 <.0001	0.64443 <.0001	0.31754 <.0001	-0.37011 <.0001	0.51377 <.0001
ZN ZN	0.39107 <.0001	-0.21999 <.0001	1.00000	-0.05486 0.2563	-0.50961 <.0001	0.30351 <.0001	0.68247 <.0001	-0.30578 <.0001	-0.29340 <.0001	-0.38218 <.0001	0.16389 0.0006	-0.42312 <.0001
CHAS CHAS	0.07115 0.1408	-0.06623 0.1704	-0.05486 0.2563	1.00000	0.09142 0.0582	0.00208 0.9657	-0.08193 0.0897	-0.03760 0.4367	-0.07410 0.1250	-0.12444 0.0098	0.03149 0.5148	0.00850 0.8604
NOX NOX	-0.49945 <.0001	0.46876 <.0001	-0.50961 <.0001	0.09142 0.0582	1.00000	-0.29590 <.0001	-0.76198 <.0001	0.61116 <.0001	0.65838 <.0001	0.18467 0.0001	-0.36233 <.0001	0.62725 <.0001
RM RM	0.75083 <.0001	-0.25986 <.0001	0.30351 <.0001	0.00208 0.9657	-0.29590 <.0001	1.00000	0.18031 0.0002	-0.18795 <.0001	-0.28523 <.0001	-0.32988 <.0001	0.03565 0.4609	-0.60443 <.0001
DIS DIS	0.31539 <.0001	-0.41066 <.0001	0.68247 <.0001	-0.08193 0.0897	-0.76198 <.0001	0.18031 0.0002	1.00000	-0.48022 <.0001	-0.50761 <.0001	-0.23186 <.0001	0.27386 <.0001	-0.52688 <.0001
RAD RAD	-0.47995 <.0001	0.69806 <.0001	-0.30578 <.0001	-0.03760 0.4367	0.61116 <.0001	-0.18795 <.0001	-0.48022 <.0001	1.00000	0.90078 <.0001	0.45363 <.0001	-0.41679 <.0001	0.53331 <.0001
TAX TAX	-0.58088 <.0001	0.64443 <.0001	-0.29340 <.0001	-0.07410 0.1250	0.65838 <.0001	-0.28523 <.0001	-0.50761 <.0001	0.90078 <.0001	1.00000	0.45992 <.0001	-0.41602 <.0001	0.58537 <.0001
PTRATIO PTRATIO	-0.53947 <.0001	0.31754 <.0001	-0.38218 <.0001	-0.12444 0.0098	0.18467 0.0001	-0.32988 <.0001	-0.23186 <.0001	0.45363 <.0001	0.45992 <.0001	1.00000	-0.16258 0.0007	0.36791 <.0001
BLAC BLAC	0.33339 <.0001	-0.37011 <.0001	0.16389 0.0006	0.03149 0.5148	-0.36233 <.0001	0.03565 0.4609	0.27386 <.0001	-0.41679 <.0001	-0.41602 <.0001	-0.16258 0.0007	1.00000	-0.33220 <.0001
LSTAT LSTAT	-0.76989 <.0001	0.51377 <.0001	-0.42312 <.0001	0.00850 0.8604	0.62725 <.0001	-0.60443 <.0001	-0.52688 <.0001	0.53331 <.0001	0.58537 <.0001	0.36791 <.0001	-0.33220 <.0001	1.00000

CRIM 與 MEDV 之間的相關係數為 -0.48150,

CHAS 與 MEDV 之間的相關係數為 0.07115,

ZN 與 MEDV 之間的相關係數為 0.39107,

NOX 與 MEDV 之間的相關係數為 -0.49945,

RM 與 MEDV 之間的相關係數為 0.75083,

DIS 與 MEDV 之間的相關係數為 -0.47995,

RAD 與 MEDV 之間的相關係數為 -0.58088,

TAX 與 MEDV 之間的相關係數為 -0.53947,

PTRATIO 與 MEDV 之間的相關係數為 -0.53947,

BLAC 與 MEDV 之間的相關係數為 0.33339,

LSTAT 與 MEDV 之間的相關係數為 -0.76989,

以下是 MEDV（房屋價格中位數）與其他變數之間相關性的整理：

正相關：

RM（每房間平均房間數）：高度正相關，相關係數為 0.75083。

- 解釋：每房間平均房間數的增加與房屋價格中位數的上升呈現高度正相關。

中度正相關：

ZN（住宅用地的比例）：中度正相關，相關係數為 0.39107。

- 解釋：住宅用地比例的增加與房屋價格中位數的上升呈現中度正相關。

輕度正相關：

CHAS（查爾斯河虛擬變數）：輕度正相關，相關係數為 0.07115。

- 解釋：查爾斯河虛擬變數與房屋價格中位數呈現輕度正相關，但相關性較弱。

負相關：

LSTAT（社會地位較低的人口占全部的百分比）：高度負相關，相關係數為 -0.76989。

- 解釋：社會地位較低的人口比例的增加與房屋價格中位數的下降呈現高度負相關。

PTRATIO（以城鎮劃分的師生比例）：中度負相關，相關係數為 -0.53947。

- 解釋：師生比例的增加與房屋價格中位數的下降呈現中度負相關。

DIS（到五個波士頓就業中心的加權距離）：中度負相關，相關係數為 -0.47995。

- 解釋：到就業中心的距離的增加與房屋價格中位數的下降呈現中度負相關。

NOX（一氧化氮濃度）：中度負相關，相關係數為 -0.49945。

- 解釋：一氧化氮濃度的增加與房屋價格中位數的下降呈現中度負相關。

RAD（高速公路可及性指數）：中度負相關，相關係數為 -0.58088。

- 解釋：高速公路可及性指數的增加與房屋價格中位數的下降呈現中度負相關。

TAX（財產稅率）：中度負相關，相關係數為 -0.53947。

- 解釋：財產稅率的增加與房屋價格中位數的下降呈現中度負相關。

第三節Variance Inflation Factor(VIF)

變異數膨脹因子(VIF) 為診斷多元共線性嚴重程度的指標。

參數估計值							
變數	標籤	DF	參數估計值	標準誤差	t 值	Pr > |t|	變異數膨脹
Intercept	Intercept	1	33.77742	5.23870	6.45	<.0001	0
CRIM	CRIM	1	-0.11701	0.03269	-3.58	0.0004	1.79029
ZN	ZN	1	0.04474	0.01408	3.18	0.0016	2.28013
CHAS	CHAS	1	2.50410	0.87744	2.85	0.0045	1.06116
NOX	NOX	1	-17.30442	3.61340	-4.79	<.0001	3.73498
RM	RM	1	4.03910	0.41984	9.62	<.0001	1.83289
DIS	DIS	1	-1.44387	0.18940	-7.62	<.0001	3.46033
RAD	RAD	1	0.30166	0.06420	4.70	<.0001	6.54243
TAX	TAX	1	-0.01217	0.00342	-3.56	0.0004	7.02135
PTRATIO	PTRATIO	1	-0.89425	0.13419	-6.66	<.0001	1.76531
BLAC	BLAC	1	0.00871	0.00281	3.10	0.0021	1.31035
LSTAT	LSTAT	1	-0.49879	0.04833	-10.32	<.0001	2.58095

VIFj=(1-Rj2)-1, j=1, 2, ..., 11

其中Rj2為Xj對其他所有X變數做的迴歸模型之複判定係數，Rj2越大Xj與其他自變數的關係越密切，VIF值也越大，當VIF超過10將造成估計值的不穩定，表示Xj幾乎是其他自變數的線性組合，即有嚴重的共線性問題，我們將考慮把Xj從模型中去除。

而由上表可知VIF皆小於十，表示此筆資料沒有嚴重的共線性問題。

第參章 原始模型檢定

第一節 建立迴歸模型

為了深入分析波士頓不同區域的房地產市場特徵及其對房價的影響，我們將探討CRIM（以城鎮劃分的人均犯罪率）、ZN（面積超過 25,000 平方英尺的住宅用地比例）、INDUS（每個城鎮非零售商業面積的比例）、CHAS（查爾斯河虛擬變數-是否鄰近查爾斯河（如果區域邊界為河流，則為 1；否則為 0））、NOX（一氧化氮濃度（千萬分之一））、RM（每套住宅的平均房間數）、AGE（1940 年以前建成的自住房屋比例）、DIS（到五個波士頓就業中心的加權距離）、RAD（高速公路的可及性指數）、TAX（每 10,000 美元的全額財產稅率）、PTRATIO（以城鎮劃分的師生比例）、blac（1000(Bk - 0.63)^2 其中 Bk 是按城鎮劃分的黑人比例）、LSTAT（社會地位較低的人口占全部的百分比）、MEDV（自住房屋的中位數價值（1/1000 美元））。

$$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\beta_3X_{3i}+.....+\beta_{12}X_{12i}+\beta_{13}X_{13i}+\varepsilon_i$$

i=1,2,.....,456

$$E(Y_i)=\hat{\beta}_0+\hat{\beta}_1X_{1i}+\hat{\beta}_2X_{2i}+\hat{\beta}_3X_{3i}+.....+\hat{\beta}_{12}X_{12i}+\hat{\beta}_{13}X_{13i}+\varepsilon_i$$

i=1,2,.....,456

為了印證以上的基本假設，我們將所有的可能變數皆列入解釋變數。

首先令:

X_1:以城鎮劃分的人均犯罪率

X_2:面積超過 25,000 平方英尺的住宅用地比例

X_3:每個城鎮非零售商業面積的比例

X_4:查爾斯河虛擬變數-是否鄰近查爾斯河(如果區域邊界為河流, 則為 1;否則為 0)

X_5:一氧化氮濃度(千萬分之一)

X_6:每套住宅的平均房間數

X_7:1940 年以前建成的自住房屋比例

X_8:到五個波士頓就業中心的加權距離

X_9:高速公路的可及性指數

X_{10}:每 10,000 美元的全額財產稅率

X_{11}:以城鎮劃分的師生比例

X_{12}:1000(Bk - 0.63)^2 其中 Bk 是按城鎮劃分的黑人比例

X_{13}:社會地位較低的人口占全部的百分比

以Y為反應變數,

Y:自住房屋的中位數價值(1/1000 美元)

參數估計值							
變數	DF	參數估計值	標準誤差	t 值	Pr > |t|	允差	變異數膨脹
Intercept	1	33.90330	5.28030	6.42	<.0001	.	0
crim	1	-0.11626	0.03277	-3.55	0.0004	0.55782	1.79269
zn	1	0.04512	0.01426	3.16	0.0017	0.42891	2.33151
indus	1	0.03779	0.06253	0.60	0.5459	0.25441	3.93065
chas	1	2.45448	0.88528	2.77	0.0058	0.92906	1.07636
nox	1	-17.73700	3.91954	-4.53	<.0001	0.22836	4.37897
rm	1	4.08181	0.43160	9.46	<.0001	0.51811	1.93010
age	1	-0.00296	0.01355	-0.22	0.8270	0.32368	3.08950
dis	1	-1.43143	0.20316	-7.05	<.0001	0.25207	3.96721
rad	1	0.31154	0.06695	4.65	<.0001	0.14107	7.08868
tax	1	-0.01310	0.00378	-3.47	0.0006	0.11724	8.52940
ptratio	1	-0.90406	0.13648	-6.62	<.0001	0.54959	1.81955
blac	1	0.00881	0.00282	3.12	0.0019	0.75939	1.31685
lstat	1	-0.49781	0.05145	-9.68	<.0001	0.34311	2.91448

經上表資料, 我們可得:

β_0=33.90330, β_1=-0.11626, β_2=0.04512, β_3=0.03779

β_4=2.45448, β_5=-17.73700, β_6=4.08181, β_7=-0.00296

β_8=-1.43143, β_9=0.31154, β_{10}=-0.01310, β_{11}=-0.90406

β_{12}=0.00881, β_{13}=-0.49781

建立原始模型:

$$\hat{Y}=33.90330-0.11626X_1+0.04512X_2+0.03779X_3+2.45448X_4-17.73700X_5$$
$$+4.08181X_6-0.00296X_7-1.43143X_8+0.31154X_9-0.01310X_{10}-0.90406X_{11}$$
$$+0.00881X_{12}-0.49781X_{13}$$

第二節 單一參數t檢定

單一參數t檢定得知β值之後，接著判斷各解釋變數

X_1:(以城鎮劃分的人均犯罪率)、

X_2:(面積超過 25,000 平方英尺的住宅用地比例)、

X_3:(每個城鎮非零售商業面積的比例)、

X_4:(查爾斯河虛擬變數-是否鄰近查爾斯河（如果區域邊界為河流，則為1；否則為 0）)、

X_5:(一氧化氮濃度（千萬分之一）)、

X_6:(每套住宅的平均房間數)、

X_7:(1940 年以前建成的自住房屋比例)、

X_8:(到五個波士頓就業中心的加權距離)、

X_9:(高速公路的可及性指數)、

X_{10}:(每 10,000 美元的全額財產稅率)、

X_{11}:(以城鎮劃分的師生比例)、

X_{12}:(1000(Bk - 0.63)^2 其中 Bk 是按城鎮劃分的黑人比例)、

X_{13}:(社會地位較低的人口占全部的百分比)、

與Y(自住房屋的中位數價值（1/1000 美元）)是否存在線性關係。

一、β_1之t檢定

欲了解X_1(以城鎮劃分的人均犯罪率)與Y(自住房屋的中位數價值（1/1000 美元）)是否存在線性相關，首先我們先假定其他變數為固定的情況下統計基本假設:

H_0:$\beta_1 = 0$

H_1:$\beta_1 \neq 0$

在顯著水準α=0.05之下虛無假設拒絕域為:p−value<α=0.05

檢定如下:p−value=0.0004<α=0.05，因此拒絕H0的假設，表示有充分資料顯示β_1=0。即表示X_1(以城鎮劃分的人均犯罪率)與Y(自住房屋的中位數價值（1/1000 美元）)有存在線性相關。S

二、β_2之t檢定

欲了解X_2(面積超過 25,000 平方英尺的住宅用地比例)與Y(自住房屋的中位數價值（1/1000 美元）)是否存在線性相關，首先我們先假定其他變數為固定的情況下統計基本假設:

H_0:$\beta_2 = 0$

H_1:$\beta_2 \neq 0$

在顯著水準α=0.05之下虛無假設拒絕域為:p−value<α=0.05

檢定如下:p−value=0.0017<α=0.05，因此拒絕H0的假設，表示有充分資料顯示β_2=0。即表示X_2(面積超過 25,000 平方英尺的住宅用地比例)與Y(自住房屋的中位數價值（1/1000 美元）)有存在線性相關。

三、β_3之t檢定

欲了解X_3(每個城鎮非零售商業面積的比例)與Y(自住房屋的中位數價值(1/1000　美元))是否存在線性相關，首先我們先假定其他變數為固定的情況下統計基本假設:

H_0:$\beta_3 = 0$

H_1:$\beta_3 \neq 0$

在顯著水準α=0.05之下虛無假設拒絕域為:p−value<α=0.05

檢定如下:p−value=0.5459>α=0.05，因此不拒絕H0的假設，表示無充分資料顯示β_3≠0。即表示X_3(每個城鎮非零售商業面積的比例)與Y(自住房屋的中位數價值(1/1000 美元))沒有存在線性相關。

四、β_4之t檢定

欲了解X_4(查爾斯河虛擬變數-是否鄰近查爾斯河(如果區域邊界為河流，則為 1;否則為 0))與Y(自住房屋的中位數價值(1/1000 美元))是否存在線性相關，首先我們先假定其他變數為固定的情況下統計基本假設:

H_0:$\beta_4 = 0$

H_1:$\beta_4 \neq 0$

在顯著水準α=0.05之下虛無假設拒絕域為:p−value<α=0.05

檢定如下:p−value=0.0058<α=0.05，因此拒絕H0的假設，表示有充分資料顯示β_4=0。即表示X_4(查爾斯河虛擬變數-是否鄰近查爾斯河(如果區域邊界為河流，則為　1;否則為　0))與Y(自住房屋的中位數價值(1/1000　美元))有存在線性相關。

五、β_5之t檢定

欲了解X_5(一氧化氮濃度(千萬分之一)與Y(自住房屋的中位數價值(1/1000　美元))是否存在線性相關，首先我們先假定其他變數為固定的情況下統計基本假設:

H_0:$\beta_5 = 0$

H_1:$\beta_5 \neq 0$

在顯著水準α=0.05之下虛無假設拒絕域為:p−value<α=0.05

檢定如下:p−value<0.0001，因此拒絕H0的假設，表示有充分資料顯示β_5=0。即表示X_5(一氧化氮濃度(千萬分之一)與Y(自住房屋的中位數價值(1/1000 美元))有存在線性相關。

六、β_6之t檢定

欲了解X_6(每套住宅的平均房間數)與Y(自住房屋的中位數價值(1/1000 美元))是否存在線性相關，首先我們先假定其他變數為固定的情況下統計基本假設:

H_0:$\beta_6 = 0$

H_1:$\beta_6 \neq 0$

在顯著水準α=0.05之下虛無假設拒絕域為:p−value<α=0.05

檢定如下:p−value<0.0001，因此拒絕H0的假設，表示有充分資料顯示β_6=0。即表示X_6(每套住宅的平均房間數)與Y(自住房屋的中位數價值(1/1000 美元))有存在線性相關。

七、β_7之t檢定

欲了解X_7(1940年以前建成的自住房屋比例)與Y(自住房屋的中位數價值（1/1000　美元）)是否存在線性相關，首先我們先假定其他變數為固定的情況下統計基本假設:

H_0:$\beta_7 = 0$

H_1:$\beta_7 \neq 0$

在顯著水準α=0.05之下虛無假設拒絕域為:p−value<α=0.05

檢定如下:p−value=0.8270>α=0.05，因此不拒絕H0的假設，表示無充分資料顯示β_7≠0。即表示X_7(1940年以前建成的自住房屋比例)與Y(自住房屋的中位數價值（1/1000 美元）)沒有存在線性相關。

八、β_8之t檢定

欲了解X_8(到五個波士頓就業中心的加權距離)與Y(自住房屋的中位數價值（1/1000　美元）)是否存在線性相關，首先我們先假定其他變數為固定的情況下統計基本假設:

H_0:$\beta_8 = 0$

H_1:$\beta_8 \neq 0$

在顯著水準α=0.05之下虛無假設拒絕域為:p−value<α=0.05

檢定如下:p−value<0.0001，因此拒絕H0的假設，表示有充分資料顯示β_8=0。即表示X_8(到五個波士頓就業中心的加權距離)與Y(自住房屋的中位數價值（1/1000 美元）)有存在線性相關。

九、β_9之t檢定

欲了解X_9(高速公路的可及性指數)與Y(自住房屋的中位數價值（1/1000 美元）)是否存在線性相關，首先我們先假定其他變數為固定的情況下統計基本假設:

H_0:$\beta_9 = 0$

H_1:$\beta_9 \neq 0$

在顯著水準α=0.05之下虛無假設拒絕域為:p−value<α=0.05

檢定如下:p−value<0.0001，因此拒絕H0的假設，表示有充分資料顯示β_9=0。即表示X_9(高速公路的可及性指數)與Y(自住房屋的中位數價值（1/1000 美元）)有存在線性相關。

十、β_{10}之t檢定

欲了解X_{10}(每10,000美元的全額財產稅率)與Y(自住房屋的中位數價值（1/1000　美元）)是否存在線性相關，首先我們先假定其他變數為固定的情況下統計基本假設:

H_0:$\beta_{10} = 0$

H_1:$\beta_{10} \neq 0$

在顯著水準α=0.05之下虛無假設拒絕域為:p−value<α=0.05

檢定如下:p−value=0.0006<α=0.05，因此拒絕H0的假設，表示有充分資料顯示β_{10}=0。即表示X_{10}(每10,000美元的全額財產稅率)與Y(自住房屋的中位數價值（1/1000 美元）)有存在線性相關。

十一、β_{11}之t檢定

欲了解X_{11}(以城鎮劃分的師生比例)與Y(自住房屋的中位數價值（1/1000 美元))是否存在線性相關，首先我們先假定其他變數為固定的情況下統計基本假設:

H_0:$\beta_{11} = 0$

H_1:$\beta_{11} \neq 0$

在顯著水準α=0.05之下虛無假設拒絕域為:p−value<α=0.05

檢定如下:p−value<0.0001，因此拒絕H0的假設，表示有充分資料顯示β_{11}=0。即表示X_{11}(以城鎮劃分的師生比例)與Y(自住房屋的中位數價值（1/1000 美元))有存在線性相關。

十二、β_{12}之t檢定

欲了解X_{12}(1000(Bk - 0.63)^2 其中 Bk 是按城鎮劃分的黑人比例)與Y(自住房屋的中位數價值（1/1000 美元))是否存在線性相關，首先我們先假定其他變數為固定的情況下統計基本假設:

H_0:$\beta_{12} = 0$

H_1:$\beta_{12} \neq 0$

在顯著水準α=0.05之下虛無假設拒絕域為:p−value<α=0.05

檢定如下:p−value=0.0019<α=0.05，因此拒絕H0的假設，表示有充分資料顯示β_{12}=0。即表示X_{12}(1000(Bk - 0.63)^2 其中 Bk 是按城鎮劃分的黑人比例)與Y(自住房屋的中位數價值（1/1000 美元))有存在線性相關。

十三、β_{13}之t檢定

欲了解X_{13}(社會地位較低的人口占全部的百分比)與Y(自住房屋的中位數價值（1/1000 美元）)是否存在線性相關，首先我們先假定其他變數為固定的情況下統計基本假設：

H_0:$\beta_{13} = 0$

H_1:$\beta_{13} \neq 0$

在顯著水準α=0.05之下虛無假設拒絕域為:p−value<α=0.05

檢定如下：p−value<0.0001，因此拒絕H0的假設，表示有充分資料顯示β_{13}=0。即表示X_{13}(社會地位較低的人口占全部的百分比)與Y(自住房屋的中位數價值（1/1000 美元）)有存在線性相關。

第三節 模型適合度檢定

以下我們使用SAS做適合度檢定，建立回歸模型

變異數的分析					
來源	DF	平方和	均方	F 值	Pr > F
模型	13	28448	2188.29931	101.01	<.0001
誤差	442	9575.89197	21.66491		
已校正的總計	455	38024			

統計假設如下:

H0:$\beta_1=\beta_2=\beta_3=.....=\beta_{12}=\beta_{13}=0$

H1:β_i不全為0,i=1，2，...，13

虛無假設之拒絕域為：p−value<α=0.05，p−value<0.0001<α=0.05所以拒絕H0之假設，表示我們有充分證據顯示β_i不全為0，即表示根據資料所配適之迴歸模型是合適的。

第四節 模型解釋能力

R^2=$\frac{SSR}{SSTO}$=**0**.7482>**0.5**調整R^2=**1**−$\frac{MSE}{MSTO}$=0.7408全模型下的R^2，表示此迴歸能解釋74.82%的Y變異而經過校正的調整R^2只低0.0074，所以代表X_i，i=1，2，...，13對Y(犯罪率)具有相當程度的解釋能力，因此下一步我們將進行模型的篩選，去除掉影響力較小的變數，以精簡模型。

根 MSE	4.65456	R 平方	0.7482
應變平均值	22.58618	調整 R 平方	0.7408
變異係數	20.60799		

第四章 變數選取

x_1 : CRIM，x_2 : ZN，x_3 : INDUS，x_4 : CHAS，x_5 : NOX，x_6 : RM

x_7 : AGE，x_8 : DIS，x_9 : RAD，x_{10} : TAX，x_{11} : PTRATIO，

x_{12} : blac x_{13} : LSTAT

向前選取法 ： 一開始模型中沒有任何的變數，且須設定進入模型的條件，一般來說會假設進入模型的值以0.15為準，接著，針對各別的變數做檢定，檢定結果須滿足P-value小於0.15中最顯著的變數，則此變數將優先選入模型中，依此類推至沒有變數可選入模型為止。

已輸入變數	變數數目	偏 R 平方	模型 R 平方	C(p)	F 值	Pr > F
lstat	1	0.545	0.545	346.511	543.87	<.0001
rm	2	0.1019	0.6469	169.749	130.66	<.0001
ptratio	3	0.039	0.6859	103.262	56.16	<.0001
dis	4	0.0108	0.6967	86.2824	16.08	<.0001
nox	5	0.0186	0.7153	55.5965	29.44	<.0001
blac	6	0.007	0.7223	45.3228	11.31	0.0008
chas	7	0.006	0.7284	36.7241	9.96	0.0017
zn	8	0.0037	0.7321	32.1946	6.21	0.0131
crim	9	0.0033	0.7354	28.4124	5.55	0.0189
rad	10	0.0054	0.7407	21.0161	9.19	0.0026
tax	11	0.0072	0.7479	10.4114	12.65	0.0004

由上表可知，上述所有變數的p-value都<0.15，因此上述所有變數都應選入模型中。

$\hat{Y}$=33.77742 - 0.11701x_1 + 0.04474x_2 + 2.5041x_4 -17.30442x_5 +4.0391x_6 -1.44387x_8 +0.30166x_9 -0.01217x_{10} -0.89425x_{11} +0.00871x_{12} -0.49879$x_{13'}$

向後消去法 ： 一開始把所有的變數放在模型中，且須假設離開模型的值(通常為0.15)，接著對整個模型做檢定，選出所有p-value大於0.15中最不顯著的變數，將優先離開模型，依此類推至沒有變數離開模型。

$\hat{Y}$ = 33.77742 - 0.11701x_1 + 0.04474x_2 + 2.5041x_4 -17.30442x_5 +4.0391x_6 -1.44387x_8 +0.30166x_9 -0.01217x_{10} -0.89425x_{11} +0.00871x_{12} -0.49879x_{13}

逐步選取法：是向前選取法與向後選取法的結合。一開始模型中沒有任何的變數，且須假設進出模型的值(通常進出模型的準則預設為0.15)。首先檢定所有的變數，接著須滿足p-value小於0.15中最顯著的變數優先放入模型，接著再選出第二顯著的變數進入，此時，模型中已有兩個變數，再將這兩個變數做檢定，如果符合標準則留在模型裡，若不符合標準則離開模型，反覆執行直到沒有任何的變數進出為止。

複判定係數法：

R^2表示Y之總變異中和使用X變數有關的部份。R^2值高並不一定表示配適模型最好。且即使R^2很高，其MSE可能太大而實際上卻需要高精確度故不

符實用。

投入變數 x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 x_{10} x_{11} x_{12} x_{13} 時，R^2=0.7482，與投入

x_1 x_2 x_4 x_5 x_6 x_8 x_9 x_{10} x_{11} x_{12} x_{13} 時的R^2=0.7417 差異不大，即多了 x_3 x_7 其貢獻度不大，所以我們最後選擇 x_1 x_2 x_4 x_5 x_6 x_8 x_9 x_{10} x_{11} x_{12} x_{13} 等11個解釋變數來建立回歸模型。

C_P準則

此準則關心的是每一子集迴歸模型n個配適值的總均方誤差。當子集有較小的CP值時，總均方誤差會很小，迴歸模型的偏誤也會較小。我們選擇投入 x_1 x_2 x_4 x_5 x_6 x_8 x_9 x_{10} x_{11} x_{12} x_{13}來建立回歸模型組合。

SBC、AIC法

從AIC法中，我們選擇變數x_1 x_2 x_4 x_5 x_6 x_8 x_9 x_{10} x_{11} x_{12} x_{13}時，其AIC值=1412.7214為最小值，因此選取以上變數為最佳回歸模型。

SBC法中，選取同AIC法的變數，其SBC值=1462.19127為最小值。

結論

method	selected variable
向前選取法	x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 x_{10} x_{11} x_{12} x_{13}
向後選取法	x_1 x_2 x_4 x_5 x_6 x_8 x_9 x_{10} x_{11} x_{12} x_{13}
逐步回歸法	x_1 x_2 x_4 x_5 x_6 x_8 x_9 x_{10} x_{11} x_{12} x_{13}
複判定係數法	x_1 x_2 x_4 x_5 x_6 x_8 x_9 x_{10} x_{11} x_{12} x_{13}
Cp準則	x_1 x_2 x_4 x_5 x_6 x_8 x_9 x_{10} x_{11} x_{12} x_{13}
AIC法	x_1 x_2 x_4 x_5 x_6 x_8 x_9 x_{10} x_{11} x_{12} x_{13}
SBC法	x_1 x_2 x_4 x_5 x_6 x_8 x_9 x_{10} x_{11} x_{12} x_{13}

由以上圖表可知，我們應剔除x_3 x_7，即可求得最佳回歸模型：

$$\hat{Y} = 33.77742 - 0.11701x_1 + 0.04474x_2 + 2.5041x_4 - 17.30442x_5 + 4.0391x_6 - 1.44387x_8 + 0.30166x_9 - 0.01217x_{10} - 0.89425x_{11} + 0.00871x_{12} - 0.49879x_{13}$$

第伍章 離群值及影響點之檢定

第一節 離群值

最初模型:$\hat{Y}$ = 33.77742 - 0.11701x_1 + 0.04474x_2 + 2.5041x_4 -17.30442x_5 +4.0391x_6 -1.44387x_8 +0.30166x_9 -0.01217x_{10} -0.89425x_{11} +0.00871x_{12} -0.49879x_{13}

一、標準化殘差值

取絕對值大於兩個標準化殘差的值，即為離群值。

本資料共選出22筆資料(第60、129、148、149、170、193、203、206、211、230、331、332、333、334、335、336、338、339、365、376、378、456)為離群值。

觀測值	應變數	預測值	標準誤差平均值預測	殘差	標準誤殘差	Student 殘差
60	33.0	23.7873	0.8786	9.2127	4.562	2.019
129	14.4	4.2065	0.9995	10.1935	4.537	2.247
148	50.0	36.3966	0.7751	13.6034	4.581	2.969
149	50.0	40.0582	1.0760	9.9418	4.520	2.200
170	50.0	36.1753	0.6637	13.8247	4.599	3.006
193	23.7	11.4635	0.8649	12.2365	4.565	2.680
203	50.0	40.1909	0.8995	9.8091	4.558	2.152
206	46.7	35.6420	0.5753	11.0580	4.610	2.398
211	48.3	37.4501	0.7331	10.8499	4.588	2.365
230	42.8	30.4653	1.0662	12.3347	4.522	2.728
331	27.5	13.4829	1.4988	14.0171	4.398	3.187
332	23.1	10.2655	1.3302	12.8345	4.452	2.883
333	50.0	23.1513	1.1563	26.8487	4.500	5.966
334	50.0	32.2131	1.0978	17.7869	4.515	3.940
335	50.0	34.2089	1.1121	15.7911	4.511	3.500
336	50.0	24.7038	0.6461	25.2962	4.601	5.498
338	13.8	0.7146	1.0268	13.0854	4.531	2.888
339	15.0	25.3333	0.7522	-10.3333	4.585	-2.254
365	7.2	17.8595	0.5665	-10.6595	4.612	-2.311
376	17.9	1.8742	1.1512	16.0258	4.501	3.560
378	7.0	-4.3135	1.2943	11.3135	4.462	2.535
456	11.9	22.3213	0.6927	-10.4213	4.594	-2.268

二、Y outlier

由Student residual，當$|t_i| \geqq t_{(1-\frac{\alpha}{2n},n-p-1)}$，則第i個點即為Y outlier，p為參數β的個數。

顯著水準α=0.05之下，p=11，n=456，則$|t_i| \geqq t_{(1-\frac{0.05}{2*456},456-11-1)}$=3.902431時為離群值，找出第333、334、336筆資料$|t_i|$=>3.902431，共找到3筆離群值。

觀測值	應變數	預測值	標準誤差平均值預測	殘差	標準誤殘差	Student 殘差	-2-1 0 1 2	Cook's D	RStudent
333	50.0	23.1513	1.1563	26.8487	4.500	5.966	| |******|	0.196	6.2139
334	50.0	32.2131	1.0978	17.7869	4.515	3.940	| |******|	0.076	4.0060
336	50.0	24.7038	0.6461	25.2962	4.601	5.498	| |******|	0.050	5.6887

三、X outlier

利用帽子矩陣，當H>$\frac{2p}{n}$=$\frac{2*11}{456}$=0.04824，此時資料即為離群值。由SAS可得知，第9 46 52 130 131 132 133 134 135 136 139 140 141 142 143 146 147 149 150 184 185 189 190 199 230 259 320 321 322 324 325 331 332 333 334 335 338 344 362 368 369 374 376 378 381 384 386 440 441 442 443 444筆資料皆為離群值。

觀測值	Hat 對角 H
9	0.0507
46	0.0502
52	0.0485
130	0.0697
131	0.0487
132	0.0537
133	0.0586
134	0.0568
135	0.0540
136	0.0502
139	0.0528
140	0.0816
141	0.0487
142	0.0653
143	0.0905
146	0.0556
147	0.0497
149	0.0536
150	0.0631
184	0.0526
185	0.0537
189	0.0515
190	0.0532
199	0.0491
230	0.0527
259	0.0681
320	0.0635

觀測值	Hat 對角 H
321	0.0598
322	0.0601
324	0.0520
325	0.0544
331	0.1041
332	0.0820
333	0.0619
334	0.0558
335	0.0573
338	0.0488
344	0.3134
362	0.0510
368	0.0493
369	0.1600
374	0.1246
376	0.0614
378	0.0776
381	0.1952
384	0.0483
386	0.0642
440	0.0770
441	0.0793
442	0.0866
443	0.0790
444	0.0789

第二節 影響點

一、DFFITS

$$DFFITS_i = \frac{\hat{Y}_1 - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}}$$

這個方程式表示為個案 i 對配適值$\hat{Y}_1$之影響力，當|DFFITS|> $2\sqrt{\frac{P}{N}}$

= $2\sqrt{\frac{12}{456}}$ = 0.3244428，即為影響點。此資料中，第 60、129、135、136、148、149、150、170、184、185、193、203、211、230、238、243、259、331、332、333、334、335、336、337、338、339、344、376、378、381、456即為影響點。

觀測值	DFFITS
60	0.3902
129	0.4971
135	0.3304
136	0.4168
148	0.5069
149	0.5259
150	0.4903
170	0.4379
184	0.3387
185	0.3597
193	0.5114
203	0.4264
211	0.3798
230	0.6478
238	0.3484
243	0.3866
259	0.3336
331	1.0976
332	0.8687
333	1.5967
334	0.9741
335	0.8741
336	0.7988
337	0.3268
338	0.6598
339	-0.3714
344	-0.5962
376	0.9228
378	0.7399
381	0.3528
456	-0.3436

二、COOKS'D

此為個案 i 對所有配適值$\widehat{Y}_1$的影響力，當COOKS'D > $F_{0.5}$(p , n - p)

時，即為影響點。此資料中，當COOKS'D > $F_{0.5}$(11,445)

=0.941509，此資料並無大於0.94的值，故在此檢驗中，找不到影響點。

三、DFBETAS

此為個案 i 對第 k 個迴歸係數的影響力，當|$DFBETAS_{k(i)}$| > 1

(or n →∞ DFBETAS > $\frac{2}{\sqrt{n}}$)，表第 i 個案對迴歸係數b_k有影響力。此資料有

一筆|$DFBETAS_{k(i)}$| > 1的影響點，為第333筆資料。

		DFBETAS											
觀測值	應變數	Intercept	CRIM	ZN	CHAS	NOX	RM	DIS	RAD	TAX	PTRATIO	BLAC	LSTAT
333	50.0	0.8493	-0.2299	0.2173	-0.1220	-0.1233	-1.1706	-0.5447	0.4518	-0.0159	-0.0707	0.0184	-1.2224

第三節 結論

將離群值與影響點經過整理過後，可以得知下表:

離群值	9、46、52、60、129、130、131、132、133、134、135、136、139、140、141、142、143、146、147、148、149、150、170、184、185、189、190、193、199、203、206、211、230、259、320、321、322、324、325、331、332、333、334、335、336、338、339、344、362、365、368、369、374、376、378、381、384、386、440、441、442、443、444、456
影響點	60、129、135、136、148、149、150、170、184、185、193、203、211、230、238、243、259、331、332、333、334、335、336、337、338、339、344、376、378、381、456

當某值同時為離群值與影響點時，應要刪除。由上表可知兩者之交集為:第60、129、135、136、148、149、150、170、184、185、193、203、211、230、259、331、332、333、334、335、336、338、339、344、376、378、381、456筆資料，且將其刪除。

觀測值	應變數	Student 殘差	Hat 對角 H	RStudent	DFFITS	Cook's D
60	33.0	2.019	0.0358	2.0263	0.3902	0.013
129	14.4	2.247	0.0463	2.2569	0.4971	0.020
135	14.6	1.382	0.0540	1.3831	0.3304	0.009
136	17.8	1.808	0.0502	1.8128	0.4168	0.014
148	50.0	2.969	0.0278	2.9960	0.5069	0.021
149	50.0	2.200	0.0536	2.2091	0.5259	0.023
150	50.0	1.884	0.0631	1.8897	0.4903	0.020
170	50.0	3.006	0.0204	3.0339	0.4379	0.016
184	48.5	1.436	0.0526	1.4374	0.3387	0.010
185	50.0	1.508	0.0537	1.5100	0.3597	0.011
193	23.7	2.680	0.0347	2.6994	0.5114	0.021
203	50.0	2.152	0.0375	2.1608	0.4264	0.015
211	48.3	2.365	0.0249	2.3772	0.3798	0.012
230	42.8	2.728	0.0527	2.7476	0.6478	0.034
259	50.0	1.233	0.0681	1.2337	0.3336	0.009
331	27.5	3.187	0.1041	3.2207	1.0976	0.098
332	23.1	2.883	0.0820	2.9071	0.8687	0.062
333	50.0	5.966	0.0619	6.2139	1.5967	0.196
334	50.0	3.940	0.0558	4.0060	0.9741	0.076
335	50.0	3.500	0.0573	3.5458	0.8741	0.062
336	50.0	5.498	0.0193	5.6887	0.7988	0.050
338	13.8	2.888	0.0488	2.9120	0.6598	0.036
339	15.0	-2.254	0.0262	-2.2642	-0.3714	0.011
344	10.4	-0.883	0.3134	-0.8824	-0.5962	0.030
376	17.9	3.560	0.0614	3.6081	0.9228	0.069
378	7.0	2.535	0.0776	2.5510	0.7399	0.045
381	8.8	0.717	0.1952	0.7165	0.3528	0.010
456	11.9	-2.268	0.0222	-2.2790	-0.3436	0.010

經過校正之後，可建立最適迴歸模型為:

$$\hat{Y} = 19.40031 - 0.10844X_1 + 0.03743X_2 + 0.80540X_4 - 12.03860X_5 + 5.03998X_6 - 1.05617X_8 + 0.21599X_9 - 0.01313X_{10} - 0.78058X_{11} + 0.01101X_{12} - 0.38099X_{13}$$

第陸章 殘差分析

第一節 均質性

均質性檢定

第一個和第二個動差規格的檢定		
DF	卡方	Pr > ChiSq
76	94.16	0.0774

H0:誤差項無異質變異

H1:誤差項有異質變異

那看表可知0.0774>0.05=α,所以不拒絕H0, 故我們沒有足夠的證據顯示誤差項有異質變異, 即符合均質性假設。

由SAS分析的結果,可以發現殘差都落在正負10之間,以0為界線上下隨機分布,我們也就可以滿足變異數為常數的假設。

分位數 (定義 5)	
層級	分位數
100% Max	11.057981
99%	8.399118
95%	5.638559
90%	4.099491
75% Q3	1.143001
50% 中位數	-0.800203
25% Q1	-2.839178
10%	-4.873178
5%	-5.925292
1%	-7.183323
0% 最小值	-10.659452

第二節 常態性

未轉換常態性檢定

常態性檢定				
檢定	統計值		p 值	
Shapiro-Wilk	W	0.955929	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.11864	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.995013	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	5.354574	Pr > A-Sq	<0.0050

統計假設如下:

H0誤差項服從常態分配

H1:誤差項不服從常態分配

取顯著水準α=0.05在虛無假設下的拒絕域為：$p-value < \alpha = 0.05$。

檢定如下：因為p−value<0.0001>0.05=α，所以拒絕H0，代表有足夠的證據顯示誤差項不服從常態分配。→即不符合常態性假設。

所以我們必須接著去做轉換

box-cox 轉換

模型陳述式規格詳細資料				
類型	DF	變數	描述	值
Dep	1	BoxCox(MEDV)	使用的 Lambda	0.41
			Lambda	0.41
			對數概度	-480.9
			Conv. Lambda	0.5
			Conv. Lambda LL	-481.8
			CI 界限	-482.8
			Alpha	0.05
			標籤	MEDV
Ind	1	Identity(CRIM)	標籤	CRIM
Ind	1	Identity(ZN)	標籤	ZN
Ind	1	Identity(CHAS)	標籤	CHAS
Ind	1	Identity(NOX)	標籤	NOX
Ind	1	Identity(RM)	標籤	RM
Ind	1	Identity(DIS)	標籤	DIS
Ind	1	Identity(RAD)	標籤	RAD
Ind	1	Identity(TAX)	標籤	TAX
Ind	1	Identity(PTRATIO)	標籤	PTRATIO
Ind	1	Identity(BLAC)	標籤	BLAC
Ind	1	Identity(LSTAT)	標籤	LSTAT

box-cox 轉換常態性檢定

常態性檢定				
檢定	統計值		p 值	
Shapiro-Wilk	W	0.986023	Pr < W	0.0004
Kolmogorov-Smirnov	D	0.084735	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.474652	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2.432317	Pr > A-Sq	<0.0050

統計假設如下:

H0誤差項服從常態分配

H1:誤差項不服從常態分配

取顯著水準α=0.05在虛無假設下的拒絕域為 : $p-value< \alpha= 0.05$。

檢定如下：因為p−value=0.0004<0.05=α，所以拒絕H0，代表有足夠的證據顯示誤差項不服從常態分配。→即不符合常態性假設。

log轉換-常態性檢定

常態性檢定				
檢定	統計值		p 值	
Shapiro-Wilk	W	0.968117	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.090025	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.722941	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	3.95645	Pr > A-Sq	<0.0050

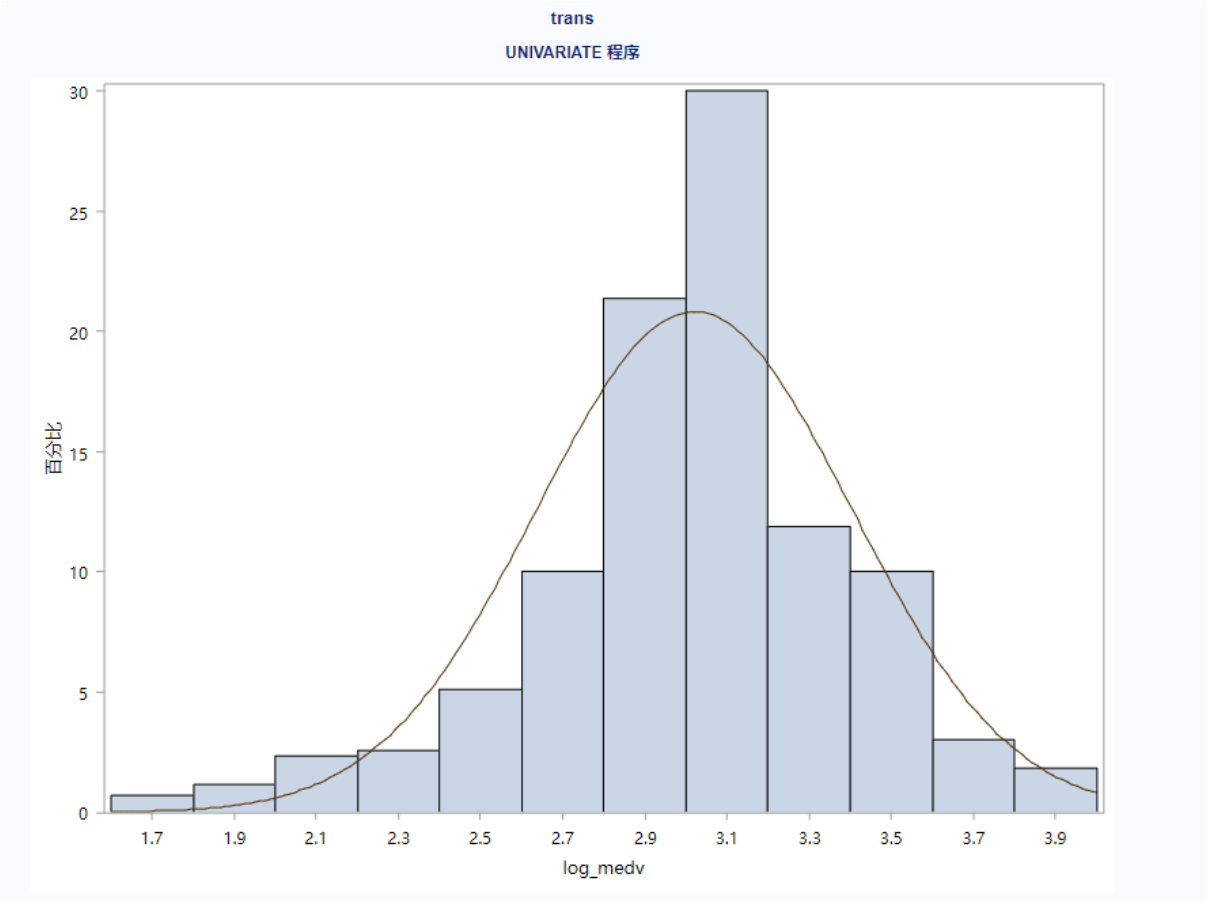
統計假設如下:

H0誤差項服從常態分配

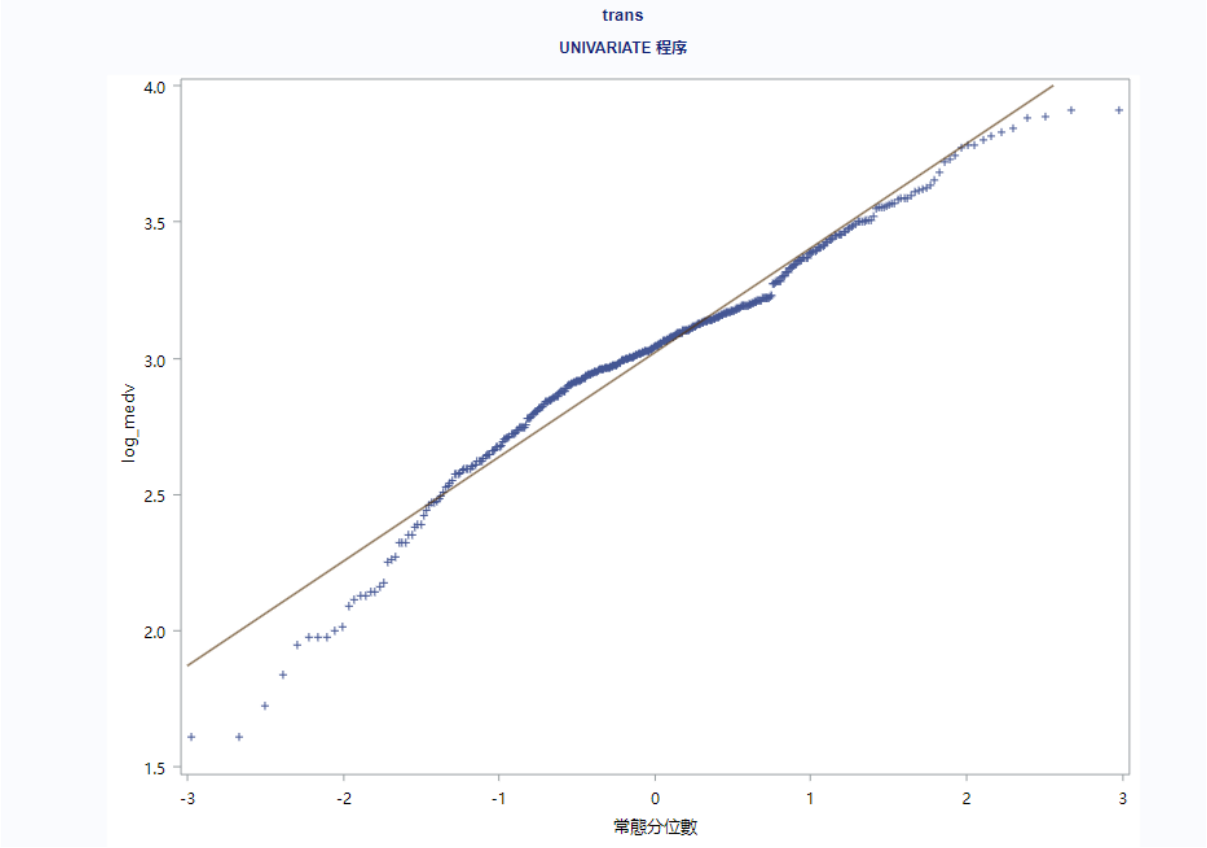
H1:誤差項不服從常態分配

取顯著水準α=0.05在虛無假設下的拒絕域為：$p-value < \alpha = 0.05$。

檢定如下：因為p−value<0.0001<0.05=α，所以不拒絕H0，代表沒有足夠的證據顯示誤差項不服從常態分配。→即不符合常態性假設。


trans
UNIVARIATE 程序
30
25
20
15
10
5
0
百分比
1.7
1.9
2.1
2.3
2.5
2.7
2.9
3.1
3.3
3.5
3.7
3.9
log_medv

常態 分布的配適度檢定				
檢定	統計值		p 值	
Kolmogorov-Smirnov	D	0.09002488	Pr > D	<0.010
Cramer-von Mises	W-Sq	0.72294072	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	3.95645031	Pr > A-Sq	<0.005



平方根轉換-常態性檢定

常態性檢定				
檢定	統計值		p 值	
Shapiro-Wilk	W	0.985344	Pr < W	0.0002
Kolmogorov-Smirnov	D	0.089098	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.490891	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2.50829	Pr > A-Sq	<0.0050

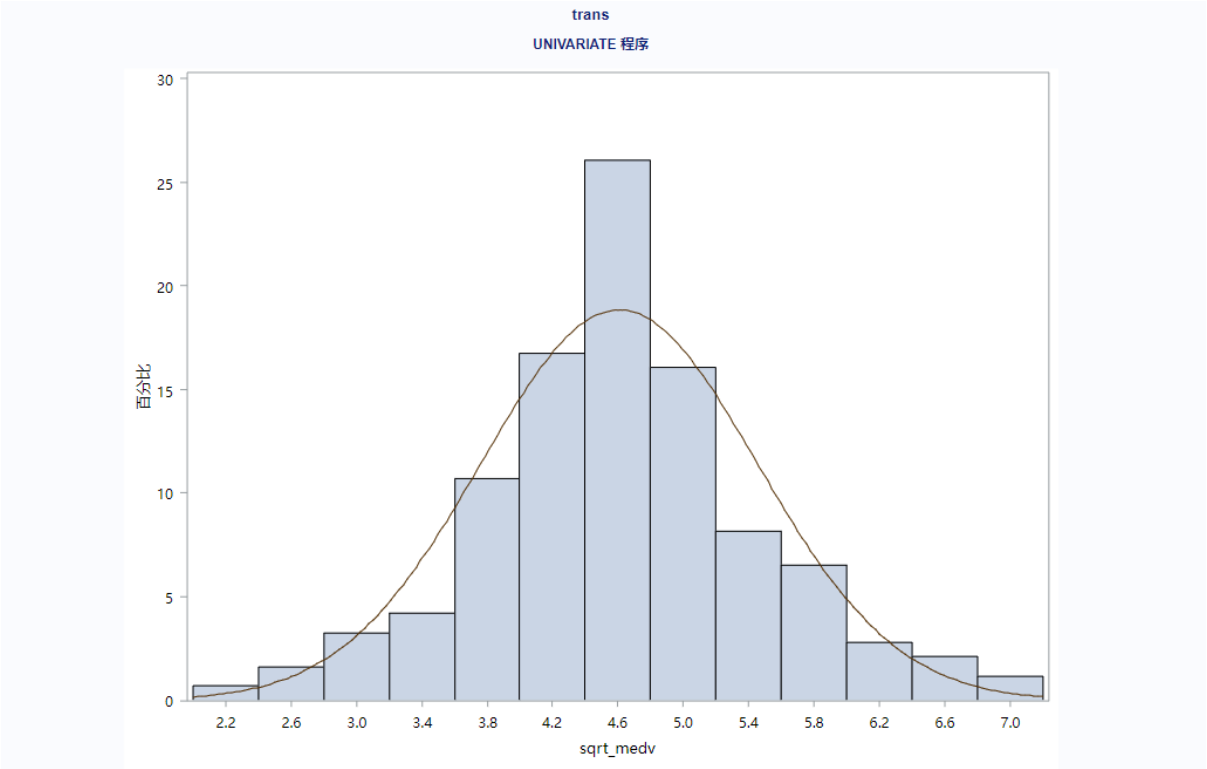
統計假設如下:

H0誤差項服從常態分配

H1:誤差項不服從常態分配

取顯著水準α=0.05在虛無假設下的拒絕域為：$p-value < \alpha = 0.05$。

檢定如下：因為p-value=0.0002<0.05=α，所以不拒絕H0，代表沒有足夠的證據顯示誤差項不服從常態分配。→即不符合常態性假設。



常態 分布的配適度檢定				
檢定	統計值		p 值	
Kolmogorov-Smirnov	D	0.08909761	Pr > D	<0.010
Cramer-von Mises	W-Sq	0.49089146	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	2.50828962	Pr > A-Sq	<0.005

第三節 獨立性

獨立性檢定

REG 程序
模型: MODEL1
應變數: MEDV MEDV

Durbin-Watson D	1.075
觀測值數目	456
一階自相關	0.455

統計假設如下:

$H0:\varrho=0$

$H1:\varrho\neq0$

Durbin-Watsontest

判斷方法:當檢定統計量D, 越趨近於2, 即代表樣本間沒有一階自我相關性存在。由表可知, D=1.075≅2, 所以不拒絕$H0$, 表示沒有足夠證據顯示, 誤差項有一階自我相關性存在。→即誤差項符合獨立性假設。

第柒章 模型確認

建立最終回歸模型時:

WARNING: The average covariance matrix for the SPEC test has been deemed singular which violates an assumption of the test. Use caution when interpreting the results of the test.

這個警告顯示在進行 SPEC 測試時，模型沒有符合，意味著模型中的某些變數之間存在線性相依性，或者說存在共線性的問題。但是其他測試沒有反應不服。因此，雖然值得注意，但不影響接下來的流程。

最終回歸模型:

$$\widehat{Y} = 19.40031 - 0.10844X_1 + 0.03743X_2 + 0.80540X_4 - 12.03860X_5 + 5.03998X_6 - 1.05617X_8 + 0.21599X_9 - 0.01313X_{10} - 0.78058X_{11} + 0.01101X_{12} - 0.38099X_{13}$$

第一節 最終模型解釋能力

表 7-1 配適模型解釋能力分析

根 **MSE**	**3.32592**	**R^2**	**0.8312**
應變平均值	**21.94674**	調整**R^2**	**0.8268**
變異係數	**15.15449**		

配適模型:

R^2=0.8312

調整R^2=0.8268

配適模型的R^2, 表示此迴歸能解釋 83.12%的 Y(波士頓房價)變異, 校正過後的調整R^2些微低了 0.0044, 表示配適模型Xi ， i = 1.2 … 13, 對Y(犯罪率)具有相當程度的解釋能力。

表 7-2 原始模型與配適模型解釋能力之比較

模型	R^2	調整R^2
原模型	**0.7482**	**0.7408**
配適模型	**0.8312**	**0.8268**

由表 7-2 可知, 配適模型比原模型能多解釋 8.3%的 Y(波士頓房價)變異, 表示經修正過後, 模型對變異量的解釋有所提升, 此模型也較精簡, 故我們選擇此配適模型作為最終模型。

第二節 最終模型預測能力

一、MAPE

為證明此模型為正確的，我們利用預先分割出的50筆資料建立一個「確認資料集」， 檢測此模型的平均絕對誤差(MAPE)，來確認此模型是否具有預測能力。

Step1:

表 7-3「最終模型資料集」ANOVA 表

來源	自由度	平方和	平均值平方	F 值	Pr > F
模型	11	22770	2070.03532	187.13	<.0001
誤差	418	4623.80192	11.06173		
已校正的總計	429	27394			

Step2:確認模型資料集

變異數的分析					
來源	DF	平方和	均方	F 值	Pr > F
模型	13	28448	2188.29931	101.01	<.0001
誤差	442	9575.89197	21.66491		
已校正的總計	455	38024			

Step3:MAPE(平均絕對預測誤差)

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\widehat{y_i}}{y_i}\right|=0.13697$$

y_i=新資料的反應變數，$\widehat{y_i}$=y_i的預測值，n=確認資料集筆數

經運算後，MAPE=0.13697，表示所選的最終迴歸模型預測偏離的程度很小，表示此模型有建模資料以外的應用能力。

二、最終模型的應用

現假設波士頓有一房子,

而該地區城市犯罪率為0.5341;

面積超過 25,000 平方英尺的住宅用地比例為20;

不再查爾斯河附近(虛擬變數0);

空氣中的一氧化氮濃度的千萬分之一為0.647;

每套住宅的平均房間數為7.52間;

到五個波士頓就業中心的加權距離為2.1398;

高速公路的可及性指數為5;

每 10,000 美元的全額財產稅率為264;

師生比例為13;

黑人比例指數為388.37;

社會地位較低的人口占全部的百分比為7.26。

有了這些訊息, 可以利用最終回歸模型:

$$\hat{Y} = 19.40031 - 0.10844X_1 + 0.03743X_2 + 0.80540X_4 - 12.03860X_5 + 5.03998X_6 - 1.05617X_8 + 0.21599X_9 - 0.01313X_{10} - 0.78058X_{11} + 0.01101X_{12} - 0.38099X_{13}$$

得出最終我們判斷自住房屋的中位數價值(1/1000美元)的結果為 18.130412, 將此數值乘以1000, 得出實際自住房屋的中位數價值為 18130.412美元。

第捌章 結論

本次的迴歸報告主要是想藉由美國波士頓房價資料集了解該地區房價組成因素。

本是想探討犯罪率、房屋密度和居住環境、工業化程度、景觀、環境品質、居住健康、房間數、居住狀態和建築品質、到五個波士頓就業中心的加權距離、交通狀況和生活便利度、地區的財政負擔、教育資源和教育品質、種族結構、社區中低收入人群的比例，這13項因素對於房價的影響。

以此建立的原始模型為:

$$\widehat{Y}=33.90330-0.11626X_1+0.04512X_2+0.03779X_3+2.45448X_4-17.73700X_5+4.08181X_6-0.00296X_7-1.43143X_8+0.31154X_9-0.01310X_{10}-0.90406X_{11}+0.00881X_{12}-0.49781X_{13}$$

後使用向前選取法、後退刪去法、逐步迴歸法、其他選取法(R^2、調整R^2、$C(p)$準則法、AIC法和SBC法)，來篩選影響力較小的變數以得到最適合的迴歸模型。

以此建立最適合的迴歸模型為:

$$\mathbf{Y} = 33.77742 - 0.11701x_1 + 0.04474x_2 + 2.5041x_4 - 17.30442x_5 + 4.0391x_6 - 1.44387x_8 + 0.30166x_9 - 0.01217x_{10} - 0.89425x_{11} + 0.00871x_{12} - 0.49879x_{13}$$

再檢測並移除同時為離群值和影響點的資料，經過程式判定，刪去了26筆觀察值，使原本456筆的訓練資料集剩下430筆。後以此資料調整回歸模型的迴歸係數。

以此建立的最終迴歸模型為:

$$\hat{Y} = 19.40031 - 0.10844X_1 + 0.03743X_2 + 0.80540X_4 - 12.03860X_5 + 5.03998X_6 - 1.05617X_8 + 0.21599X_9 - 0.01313X_{10} - 0.78058X_{11} + 0.01101X_{12} - 0.38099X_{13}$$

之後以此迴歸模型進行殘差檢定，結果不符合線性回歸三大假設:同質性、常態性、獨立性，故沒有理由選擇此迴歸模型作為最佳迴歸模型，但依舊可以嘗試以此模型進行驗證資料集的模型驗證。

附錄

回歸:資料集**https://lib.stat.cmu.edu/datasets/boston**

ppt**https://docs.google.com/presentation/d/10j1tf4tf2_m9pPgsnTUltDCxjHbcre-rrOsyPrB2ZTw/edit?usp=sharing**
excelhttps://docs.google.com/spreadsheets/d/1PdGmGvZNZo_YA2-r8FgSLWgqTqxf2dm5r68dKhW91d0/edit?usp=sharing
因子google翻譯

CRIM-- 以城鎮劃分的人均犯罪率

ZN - 面積超過 25,000 平方英尺的住宅用地比例。

INDUS - 每個城鎮非零售商業面積的比例。

CHAS - 是否鄰近查爾斯河（如果區域邊界為河流，則為 1；否則為 0）

NOX - 一氧化氮濃度（千萬分之一）

RM一每套住宅的平均房間數

AGE - 1940 年之前建造的自住單元的比例

DIS - 到五個波士頓就業中心的加權距離

RAD - 放射狀高速公路的可及性指數

TAX - 每 10,000 美元的全額財產稅率

PTRATIO - 以城鎮劃分的師生比

blac - 1000(Bk - 0.63)^2 其中 Bk 是按城鎮劃分的黑人比例

LSTAT - 社會地位較低的人]口占全部的百分比

MEDV - 自住房屋的中位數價值（1/1000 美元）

變異數的分析					
來源	DF	平方和	均方	F 值	Pr > F
模型	11	22770	2070.03532	187.13	<.0001
誤差	418	4623.80192	11.06173		
已校正的總計	429	27394			

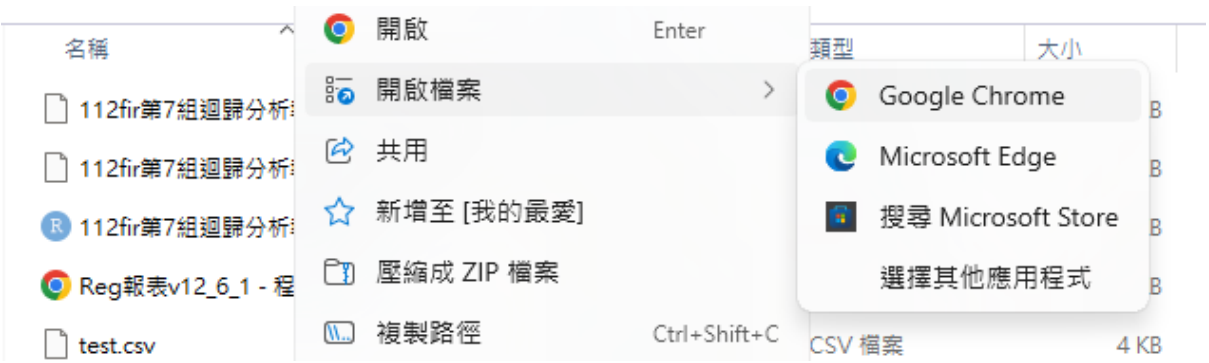
根 MSE	3.32592	R 平方	0.8312
應變平均值	21.94674	調整 R 平方	0.8268
變異係數	15.15449		

參數估計值							
變數	標籤	DF	參數估計值	標準誤差	t 值	Pr > |t|	變異數膨脹
Intercept	Intercept	1	19.40031	4.25661	4.56	<.0001	0
CRIM	CRIM	1	-0.10844	0.03371	-3.22	0.0014	2.17389
ZN	ZN	1	0.03743	0.01067	3.51	0.0005	2.39066
CHAS	CHAS	1	0.80540	0.68080	1.18	0.2375	1.06028
NOX	NOX	1	-12.03860	2.67457	-4.50	<.0001	3.77463
RM	RM	1	5.03998	0.36783	13.70	<.0001	1.97868
DIS	DIS	1	-1.05617	0.14535	-7.27	<.0001	3.68095
RAD	RAD	1	0.21599	0.04899	4.41	<.0001	6.79017
TAX	TAX	1	-0.01313	0.00251	-5.23	<.0001	6.79241
PTRATIO	PTRATIO	1	-0.78058	0.09805	-7.96	<.0001	1.71086
BLAC	BLAC	1	0.01101	0.00213	5.18	<.0001	1.31077
LSTAT	LSTAT	1	-0.38099	0.03984	-9.56	<.0001	2.85142

Durbin-Watson D	1.198
觀測值數目	430
一階自相關	0.396

下載後用瀏覽器開啟有目前所有表

https://drive.google.com/file/d/1HQTVUlIZsGOiYpIzXPhLo_g4mOa68AoP/view?usp=sharing



R程式碼

```
#使用R將資料分割:#
data <- read.csv("112fir第7組迴歸分析報告 - 完整506筆資料.csv", header = TRUE)
index <- sample(1:2, nrow(data), replace = TRUE, prob = c(0.9, 0.1))
train <- data[index == 1, ]
test <- data[index == 2, ]
write.csv(train, "train.csv", row.names = TRUE)
write.csv(test, "test.csv", row.names = TRUE)
```

```
/*SAS程式碼:訓練集分析*/
LIBNAME mydata 'C:\Users\410650229\Desktop';
PROC IMPORT DATAFILE='C:\Users\410650229\Desktop\112fir第7組迴歸分析報告.xlsx'
        OUT=mydata.train
        DBMS=XLSX REPLACE;
        /* 表示要匯入的工作表名稱 */
        SHEET="訓練集n=456";
RUN;
PROC IMPORT DATAFILE='C:\Users\410650229\Desktop\112fir第7組迴歸分析報告.xlsx'
        OUT=mydata.tests
        DBMS=XLSX REPLACE;
        /* 表示要匯入的工作表名稱 */
        SHEET="測試集n=50";
RUN;
/*常態性檢定*/
proc univariate data=mydata.train normal;
  var medv;
  histogram / normal;
TITLE "常態性檢定";
run;
/* 生成所有變數的直方圖 */
```

```
PROC UNIVARIATE DATA=mydata.train;
  VAR crim zn indus chas nox rm age dis rad tax ptratio blac lstat medv;
  HISTOGRAM;
RUN;
/* 查看匯入的資料集 */
PROC PRINT DATA=mydata.train;
RUN;
proc reg DATA=mydata.train;
  model medv = crim zn indus chas nox rm age dis rad tax ptratio blac lstat / r partial tol vif collinoint ;
TITLE "檢定是否有共線性關係";
run;
proc reg DATA=mydata.train;
  model medv = crim zn indus chas nox rm age dis rad tax ptratio blac lstat /
selection=Forward sle=0.05 ;
TITLE "前選取法";
run;
proc reg DATA=mydata.train;
  model medv = crim zn indus chas nox rm age dis rad tax ptratio blac lstat /
selection=backward sle=0.05;
TITLE "後選取法";
run;
proc reg DATA=mydata.train;
  model medv = crim zn indus chas nox rm age dis rad tax ptratio blac lstat /
selection=stepwise sle=0.05;
TITLE "逐步選取法";
run;
proc rsquare DATA =mydata.train adjrsq cp aic mse sbc;
  model medv = crim zn indus chas nox rm age dis rad tax ptratio blac lstat ;

TITLE "其他選取法";
run;
proc corr DATA=mydata.train;
var medv crim zn indus chas nox rm age dis rad tax ptratio blac lstat ;
TITLE "檢定是否有相關";
run;
/*訓練集分析_資料選取法後的分析*/
proc reg DATA=mydata.train;
  model medv = crim zn chas nox rm dis rad tax ptratio blac lstat /vif r influence dw;;
output out=outlier r=r h=h
rstudent=rs student=student
cookd=cookd dffits=dffits;
title'極端值與影響點';
run;
proc univariate data=outlier normal plot;
var r;
```

```
title'未刪除離群值的資料';
run;
/*告訴你離群值的資料*/
data train_ol;
  set outlier;
  p=12; n=456;
  hh=2*p/n;
  dif=((p/n)**0.5)*2;
  c=finv(0.5,p,n-p);
  t=tinv(1-0.05/(2*n),n-p-1);
if(abs(student)>2or h>hh or abs(rs)>t) and (cookd>c or abs(dffits)>dif )then output;
  drop p n hh dif c t;
run;
proc print data=train_ol;
title'告訴你離群值的資料';
run;
/*刪除離群值的資料*/
data train_do;
set outlier;
p=12;n=456;
hh=2*p/n;
dif=((p/n)**0.5)*2;
c=finv(0.5,p,n-p);
t=tinv(1-0.05/(2*n),n-p-1);
if(abs(student)>2or h>hh or abs(rs)>t) and (cookd>c or abs(dffits)>dif )then delete;
drop p n hh dif c t;
run;
proc print data=train_do;
title'刪除離群值的資料';
run;
proc reg data=train_do;
model medv = crim zn chas nox rm dis rad tax ptratio blac lstat / r ;
output out=train_res r=r p=pd;
title'刪除離群值的資料regl';
run;
proc autoreg data=train_res;
model medv = crim zn chas nox rm dis rad tax ptratio blac lstat;
hetero;
title'檢定';
run;
/*常態性檢定*/
proc univariate data=train_do normal;
   var MEDV;
TITLE "常態性檢定";
run;
```

```
/*boxcox*/
proc transreg details data=train_do;
model boxcox(MEDV/lambda=-2to 2 by 0.01)=identity(crim zn chas nox rm dis rad tax ptratio blac lstat);
run;
data train_doo;
set train_do;
medvnew=((MEDV**0.41)-1)/0.41;
run;
proc univariate data=train_doo normal ;
  var medvnew;
TITLE "常態性檢定";
run;
/* 對數轉換 */
data train_do_transformed;
   set train_do;
   log_medv = log(MEDV);

run;

/* 檢查轉換後的殘差是否符合正態分佈 */
proc univariate data=train_do_transformed  normal;
   var log_medv;
   qqplot log_medv / normal(mu=est sigma=est);
   histogram log_medv / normal;
            TITLE "trans";
run;

/* 方根轉換 */
data train_do_transformed2;
   set train_do;
   sqrt_medv = sqrt(MEDV);

run;

/* 檢查轉換後的殘差是否符合正態分佈 */
proc univariate data=train_do_transformed2  normal;
   var sqrt_medv;
   qqplot sqrt_medv / normal(mu=est sigma=est);
   histogram sqrt_medv / normal;
run;

/* fo轉換 */
data train_do_transfo;
   set train_do;
   sqrt_medv = sqrt(MEDV)-1;
```

```
run;

/* 檢查轉換後的殘差是否符合正態分佈 */
proc univariate data=train_do_transfo  normal;
    var sqrt_medv;
    qqplot sqrt_medv / normal(mu=est sigma=est);
    histogram sqrt_medv / normal;
run;

proc transreg details data=train_do;
model boxcox(MEDV/lambda=-2to 2 by 0.01)=identity(crim zn chas nox rm dis rad tax ptratio blac lstat);
run;
data train_doo;
set train_do;
medvnew=((MEDV**0.41)-1)/0.41;
run;
proc univariate data=train_doo normal ;
   var medvnew;
TITLE "常態性檢定";
run;

proc reg data=train_doo;
model medvnew = crim zn chas nox rm dis rad tax ptratio blac lstat /
r vif dw influence partial collinoint spec;
title'建立最終回歸模型';
run;

proc univariate data=train_res normal plot;
var r;
title'殘差檢定';
run;
proc corr data=train_do;
var medv crim zn chas nox rm dis rad tax ptratio blac lstat ;
title'檢定是否有相關性';
run;
proc reg data=train_do;
model medv = crim zn chas nox rm dis rad tax ptratio blac lstat /
r vif dw influence partial collinoint spec;
title'建立最終回歸模型';
run;
/*test to prove reg_line*/
data test_maper;
set mydata.tests;
/* 回歸估計線*/
```

```
MEDVhat=19.40031-0.10844*CRIM
+0.03743*ZN +0.80540*CHAS
-12.03860*NOX+5.03998*RM
-1.05617*DIS +0.21599*RAD
-0.01313*TAX-0.78058*PTRATIO
+0.01101*BLAC -0.38099*LSTAT;
mape=abs((MEDV-MEDVhat)/MEDV);
title 'test_prove';
run;
proc univariate data=test_maper;
var mape;
output out=test_con mean= m;
run;
data test_fin;
set test_con;
MAPE=m;
keep MAPE;
title 'ending';
RUN;
proc print data=test_fin;
run


/*變數轉換*/
/*常態性檢定*/
proc univariate data=train_do normal;
  var MEDV;
TITLE "常態性檢定";
run;
/*boxcox*/
proc transreg details data=train_do;
model boxcox(MEDV/lambda=-2to 2 by 0.01)=identity(crim zn chas nox rm dis rad tax ptratio blac lstat);
run;
data train_doo;
set train_do;
medvnew=((MEDV**0.41)-1)/0.41;
run;
proc univariate data=train_doo normal ;
  var medvnew;
TITLE "常態性檢定";
run;
/* 對數轉換 */
data train_do_transformed;
   set train_do;
   log_medv = log(MEDV);
```

```
run;

/* 檢查轉換後的殘差是否符合正態分佈 */
proc univariate data=train_do_transformed  normal;
   var log_medv;
   qqplot log_medv / normal(mu=est sigma=est);
   histogram log_medv / normal;
            TITLE "trans";
run;

/* 方根轉換 */
data train_do_transformed2;
   set train_do;
   sqrt_medv = sqrt(MEDV);

run;

/* 檢查轉換後的殘差是否符合正態分佈 */
proc univariate data=train_do_transformed2  normal;
   var sqrt_medv;
   qqplot sqrt_medv / normal(mu=est sigma=est);
   histogram sqrt_medv / normal;
run;

/* fo轉換 */
data train_do_transfo;
   set train_do;
   sqrt_medv = sqrt(MEDV)-1;

run;

/* 檢查轉換後的殘差是否符合正態分佈 */
proc univariate data=train_do_transfo  normal;
   var sqrt_medv;
   qqplot sqrt_medv / normal(mu=est sigma=est);
   histogram sqrt_medv / normal;
run;


proc reg data=train_do_transfo;
model sqrt_medv = crim zn chas nox rm dis rad tax ptratio blac lstat /
r vif dw influence partial collinoint spec;
title'建立最終回歸模型';
run;
```

```
proc univariate data=train_do_transfo normal plot;
var r;
title'残差検定';
run;
```