

# 紐約州犯罪率的真相

陳柏璋 賴奕宏 曾樟傑

June 14, 2025

## 摘要

美國長期以來被視為全球最受關注的高犯罪風險國家之一，暴力犯罪如搶劫、槍擊與謀殺等案件時常見於新聞報導中，特別是在某些大城市，犯罪情況更為嚴峻。鑑於此現象，本研究以美國紐約州 62 個郡（County）作為分析對象，進行重大犯罪率的多元線性迴歸分析，試圖找出影響紐約各郡 2023 年犯罪率的主要社會經濟因素。本研究選取八項社會經濟變數作為解釋變數，包括貧窮率、失業率等指標，並以 2023 年各郡的重大犯罪率作為反應變數。本研究使用 R 語言建構多元線性迴歸模型，採用逐步迴歸法與全子集迴歸法等方法選出最佳模型。並進行Box-Cox轉換，使最終模型符合多元迴歸分析中的假設，包括殘差的常態性、獨立性與變異數齊一性，以確保模型推論結果的正確性。最終模型選出三項顯著的解釋變數，分別為「收入中位數」、「擁有學士學位以上人口比例」與「各郡總人口的對數值」。這些變數與重大犯罪率之間呈現顯著的線性關係，並可作為區域犯罪差異的重要解釋因素。模型的解釋力為 0.6338，顯示本模型在社會科學研究中具有良好的解釋力，具備實證分析與政策參考之價值。

財精三C 11155312 陳柏璋

財精三C 11155339 賴奕宏

財精三B 11155220 曾樟傑

## 目錄

1. 緒論 .....	3
1.1. 研究背景與動機 .....	3
1.2. 文獻探討 .....	4
2. 研究架構與變數說明 .....	5
2.1. 研究方法概述 .....	5
2.2. 流程圖 .....	6
2.3. 資料來源 .....	7
2.4. 資料探索 .....	8
2.4.1. 敘述性統計量 .....	8
2.4.2. 自變數與依變數關係分析 .....	9
3. 模型建構 .....	11
3.1. 初始模型 .....	11
3.2. 模型選擇 .....	12
3.2.1. 模型指標 .....	12
3.2.2. 向前選取法 .....	13
3.2.3. 向後選取法 .....	14
3.2.4. 逐步迴歸 .....	14
3.2.5. 全子集迴歸 .....	15
3.2.6. 最佳模型 .....	16
4. 模型診斷與假設檢定 .....	17
4.1. 迴歸假設檢驗 .....	17
4.2. Box-Cox轉換 .....	19
4.3. Box-Cox轉換後的迴歸假設檢驗 .....	20
4.4. 離群值分析 .....	21
4.5. 敏感性分析 .....	22
5. 結果分析 .....	24
6. 結論 .....	25
7. 附錄 .....	25
7.1. 參考文獻 .....	25
7.2. 程式碼 .....	26

# 1. 緒論

## 1.1. 研究背景與動機

2023年4月，美國紐約市地鐵站內發生一起隨機持刀攻擊事件，引起社會大眾對於城市犯罪問題的再度關注。類似事件並非偶發個案，反映出美國部分地區暴力犯罪仍持續存在的現象。美國向來是全世界最受關注的犯罪高風險國家之一，尤其在暴力犯罪方面，像是搶劫、槍擊、謀殺等案件時常見諸新聞。根據FBI（2023）統計，美國整體暴力犯罪率相較於2020年疫情高峰期已有所下降，每十萬人下的犯罪次數從386.3次降至363.8次，但在特定城市如芝加哥、紐約、費城等地，暴力與財產犯罪仍然頻繁，呈現出地區落差明顯、集中於特定熱區的現象。Bhattac harya（2020）透過迴歸模型指出，犯罪率與失業、貧困、人口密度及槍枝持有率顯著相關，顯示社會經濟壓力越大的地區，其犯罪風險也越高。Chang等人（2021）則指出，美國城市的犯罪事件會隨著人口規模「不成比例增加」，也就是說城市越大，單位人口的犯罪率反而越高。

這些研究結果反映，犯罪問題不能僅從國家或州的層級平均來理解，而應深入探討不同地區之間的差異與分布情形。圖一為2023年紐約州各郡的犯罪率分布圖，顯示出犯罪率在各郡之間差異明顯，其中包括紐約市所屬的郡（如紐約NEW、布魯克林KIN、皇后區QUE），犯罪率相對偏高，顯示出犯罪主要集中在部分都市與人口密集地區。相較之下，中南部與部分農業為主的郡則犯罪率較低，呈現出明顯的地理落差。因此，本研究以紐約州為例，針對其62個郡（County）進行犯罪率的多元線性回歸分析，藉此探討紐約州62個郡犯罪率的地區差異以及潛在影響犯罪率的社會經濟因素。

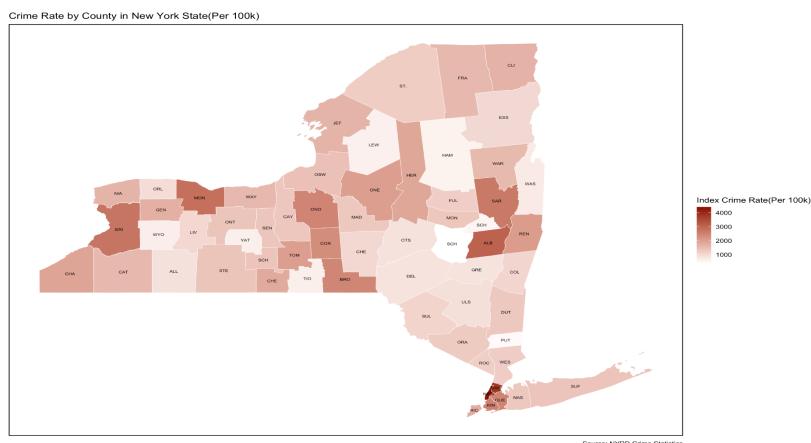


圖 1-1 紐約州各郡的犯罪率分布圖

資料來源：NYPD Crime Statistics

## 1.2. 文獻探討

Schulessler (1962) 探討了美國大城市之間犯罪率變異的社會來源，研究聚焦於1950年人口超過10萬的105座城市，試圖了解犯罪是否可由少數普遍的社會因素解釋，或是反映地區特有的情境。研究中彙整了七項主要犯罪（如謀殺、搶劫、盜竊等）的平均發生率，並搭配美國人口普查中20項社會經濟變項（如非白人比例、平均收入、擁擠程度等），進行統計分析。研究結果顯示五個潛在社會因素可用來解釋城市犯罪率的變異。其中，第一個因素與非白人比例高、收入低、居住擁擠有關，代表一種「社會資源缺乏」的情況，和謀殺與傷害等暴力犯罪有明顯關聯。第二個因素則與搶劫和竊盜等財產犯罪有關，但具體社會原因較不清楚，其他三個因素雖然也與某些社會條件有關，像是制度控制、工業化程度與人口年齡結構等面向，但對犯罪率的影響相對較弱。

Gibbs & Erickson (1976) 從生態學角度出發，探討美國都市犯罪率的統計解釋問題。本研究並非以城市總人口作為分母的「犯罪率公式」，因為許多犯罪行為的加害者或受害者可能來自鄰近非本地居民，可能低估實際潛在犯罪人口。研究提出以「社區 / 城市人口」比值作為一種更合理的分析角度，利用美國372個人口超過5萬人的城市資料，建立犯罪率與社區規模之間的相關性分析。透過相關係數計算結果顯示，在「單一都市」情況下（即所在都會區僅有該一座主要城市），社區 / 城市人口的比值與犯罪率之間存在顯著正相關，其中以搶劫與汽車竊盜的相關性最高。研究結果說明了城市的生態位置與其犯罪率密切相關，而不應僅將犯罪率解釋為城市內部社會、經濟或文化的單一反映。

Osgood (2000) 針對美國青少年的暴力犯罪率進行研究，試圖找出哪些社會因素與犯罪率之間有關聯。研究蒐集了來自美國四個州、共264個非都市縣市的資料，以每千名青少年人口中的搶劫逮捕人數作為犯罪率指標，並運用Poisson回歸模型來分析資料，這種模型適合用來處理像犯罪案件這類「發生次數」的資料，避免使用傳統OLS模型可能出現的統計偏誤。研究結果顯示，青少年比例高、家庭破碎率（單親家庭比例）高、失業率高，以及人口流動率高的地區，其青少年暴力犯罪率也顯著較高。這些結果支持社會解組理論的核心假設：當社區缺乏穩定結構與社會控制能力時，容易讓青少年更容易涉入犯罪。

Chang, Kim & Jeon (2019) 研究美國758個城市，探討城市規模是否與較低的犯罪率有關。研究不同犯罪類型（如財產犯罪與暴力犯罪）與城市人口數量之間的數學關係。分析城市內的社會現象與其人口規模之間的變化。研究結果發現，大多數犯罪類型（如搶劫、竊盜、襲擊）

隨著城市人口增加而呈現「超線性」成長，意即越大的城市，其犯罪率不但沒有降低，反而相對更高。這種現象顯示，大城市可能因人口密集、社會互動頻繁與資源競爭，反而成為犯罪較集中的熱點。此外，研究亦指出，不同犯罪類型的隨人口增加的變化趨勢也有所差異，反映出犯罪背後的社會結構與互動機制並不完全相同。

Bhattacharya (2020) 針對美國50州與哥倫比亞特區2019年的數據進行研究，目的於分析影響美國暴力犯罪率的主要社會經濟因素。研究假設槍枝擁有率上升會導致暴力犯罪增加，並以每10萬人口中的暴力犯罪發生率作為反應變數Y，透過最小平方法（OLS）建構迴歸模型。其解釋變數X包括：每平方英哩的人口密度、失業率、貧困人口比例，以及槍枝擁有率等變數。研究結果顯示，上述四項變數皆與暴力犯罪率呈現統計上顯著的正相關。具體而言，該模型的解釋力 ( $R^2$ ) 為0.532，表示這些因素能解釋美國各州暴力犯罪變異的53%以上。其中，四項解釋變數的P值皆低於顯著性水準0.05，顯示其對暴力犯罪率的解釋具有高度統計意義。研究也發現，都市地區相較於鄉村地區，有更高的暴力犯罪率，支持犯罪在空間上集中於城市的假說。綜合而言，Bhattacharya的分析突顯了美國暴力犯罪問題的多因性，並從統計模型中提供具體可行的社會政策參考方向。

Bothos & Thomopoulos (n.d.) 的研究探討了社會和經濟因素影響美國各州犯罪率的社會和經濟因素。研究使用了2004到2014年間49個州的統計資料，並透過計量模型來分析數據，試圖了解不同變數之間的關聯性。研究主要關注四個因素：失業率、貧困人口比例、退學率，以及警察人力與執法效率，並使用了ARIMA時間序列模型來分析1950年到2012年間的犯罪變化趨勢。研究結果顯示，失業、貧困、教育程度低等問題，都和犯罪率上升有關。相對地，如果一個地區的警察人力充足、執法效率高，當地犯罪率則可能下降。研究也指出，犯罪率不只是個人行為的結果，背後往往與地區的經濟環境與社會資源有密切關聯。

## 2. 研究架構與變數說明

### 2.1. 研究方法概述

本研究採用多元線性迴歸作為主要分析方法，目的是在同時控制多個社會經濟背景條件下，評估各變數對紐約州62個郡重大犯罪率的影響程度。相較於單變數分析，多元迴歸能夠在其他條件不變的前提下，釐清個別變數與犯罪率之間的獨立關係。本研究系統性地考量所有解

釋變數的可能組合，從中篩選出在統計上具顯著性、且對模型整體表現具有實質貢獻的變數組合，試圖建構精簡、同時保持著高度解釋力的模型，找出影響犯罪率最關鍵的解釋變數。

## 2.2. 流程圖

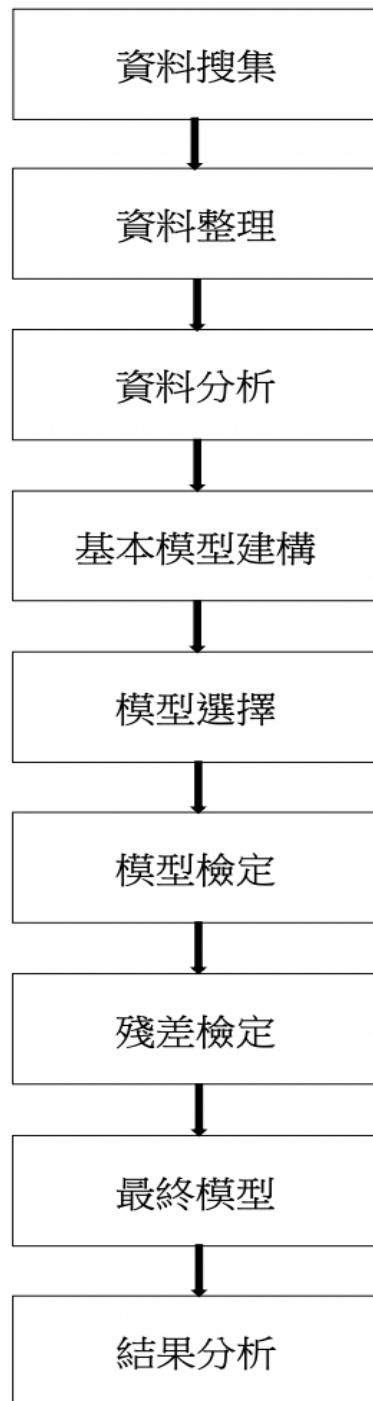


圖 2-1

## 2.3. 資料來源

紐約各州犯罪率資料來自於美國紐約市政府網站（NYC.gov）中紐約市警察局（NYPD）官方發布在2023年的犯罪統計資料。另外，所有的解釋變數資料皆來自來自美國人口普查局（U.S. Census Bureau）在2023年的統計資料。

### 變數定義

反應變數（Y）：

ICR：紐約州各郡（County）在2023年的指標性犯罪率（Index Crime Rate）

$$ICR = \frac{\text{number of index crimes occurred in the County}}{\text{Total Population of the County}}$$

其中指標性犯罪（index crime）定義為重大犯罪：謀殺、非過失殺人、強暴、搶劫、重傷害、入室竊盜、竊盜、汽車竊盜、縱火以及店內行竊

解釋變數（X）：

1) PI：貧窮率（Poverty Incidence）

$$PI = \frac{\text{number of people below the poverty thresholds in the county}}{\text{Total Population of the County}}$$

其中貧窮門檻（Poverty Thresholds）採用美國人口普查局訂定之貧窮門檻（如表2-1），依據不同家庭組成人數有不同貧窮門檻的標準。

Size of family unit	Weighted average thresholds	Related children under 18 years								
		None	One	Two	Three	Four	Five	Six	Seven	Eight or more
One person (unrelated individual):	15,480									
Under 65 years.....	15,850	15,852								
65 years and over.....	14,610	14,614								
Two people:	19,680									
Householder under 65 years.....	20,490	20,404	21,002							
Householder 65 years and over.....	18,430	18,418	20,923							
Three people.....	24,230	23,834	24,526	24,549						
Four people.....	31,200	31,428	31,942	30,900	31,008					
Five people.....	36,990	37,901	38,452	37,275	36,363	35,807				
Six people.....	41,860	43,593	43,766	42,864	41,999	40,714	39,952			
Seven people.....	47,670	50,159	50,472	49,393	48,640	47,238	45,602	43,808		
Eight people.....	52,850	56,099	56,594	55,575	54,683	53,416	51,809	50,136	49,710	
Nine people or more.....	62,900	67,483	67,810	66,908	66,151	64,908	63,198	61,651	61,268	58,907

表 2-1 美國人口普查局訂定之貧窮門檻

- 2) MI：收入中位數（Median Income）
- 3) UR：失業率（Unemployment Rate）
- 4) PD：人口密度（Population Density）單位：平方英里
- 5) TP：總人口數（Total Population）
- 6) Age25\_44：年齡為25到44歲人口比例（Age from 25 to 44）

7) BoH：18歲以上學士學位或學士以上學位人口比例（Bachelor's or higher Education Attainment Rate）

$$BoH = \frac{\text{number of people have Bachelor's degree or higher in the county}}{\text{Total Population of citizens 18 years and over in the County}}$$

8) ND：18歲以上不具有高中畢業文憑比例（No High School Diploma）

$$ND = \frac{\text{number of people without High School Diploma in the county}}{\text{Total Population of citizens 18 years and over in the County}}$$

9) 虛擬變數（Dummy Variable）：是否為紐約市五大郡（County）

在紐約62個郡中，各個郡有各自的政府，獨自管理。唯獨紐約市五大郡是由紐約市府統一管理五個郡，分別為Manhattan（曼哈頓）、Brooklyn（布魯克林）、Queens（皇后）、The Bronx（布朗克斯）以及Staten Island（史坦頓島）。其社會結構與紐約州其他郡有明顯的差異，因此設置虛擬變數，屬於紐約市五大郡給定 dummy variable 值為 1，否則為 0。希望藉此判斷紐約市五大郡的犯罪率是否跟其他郡有顯著的不同。

## 2.4. 資料探索

### 2.4.1. 敘述性統計量

表一呈現本研究中各變數的敘述性統計結果，有助於我們初步理解各變數在樣本資料中分布情況與特性。整體而言，大部分變數的平均值與中位數相差不大，如失業率、學歷比例與年齡結構等，顯示其資料分布大致對稱，無明顯偏態，資料亦相對集中。然而，值得注意的是人口密度（Population Density, PD）與總人口數（Total Population, TP），兩者的敘述性統計結果顯示出明顯的右偏及高峰的現象。人口密度（Population Density, PD）平均為 3038.92 人/ $\text{mi}^2$ ，但中位數僅為 112.20 人/ $\text{mi}^2$ ，兩者落差極大，偏度（Skewness）和峰度（Kurtosis）分別為 4.59 與 22.49。說明人口密度（Population Density, PD）分布極度不對稱，且存在極端值，顯示都市與鄉村發展落差極大。另一方面，總人口數（Total Population, TP）平均值為 320521.27 人，中位數僅為 86966.5 人，落差同樣極大，偏度和峰度分別為 2.55 與 6.11。可觀察到總人口數（Total Population, TP）分布同樣不對稱，顯現出紐約州人口分配不平均的現象。進一步觀察圖二所示的指標性犯罪率（Index Crime Rate, ICR）盒鬚圖，我們發現整體分布趨於集中，唯右側存在兩筆明顯的離群值，分別來自 Bronx 郡與 New York 郡。這兩地犯罪率顯著高於其他地區，導致整體平均值（1533.25）被明顯拉高，遠高於中位數（1353.95）。此結果反映出城市地區的高犯罪率對全州平均值造成的影響。

	size	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Index Crime Rate	62	1533.25	790.55	1353.95	1445.30	634.18	379.90	4403.00	4023.10	1.25	1.90	100.40
Median Income	62	76432.69	18436.69	70695	73195.62	9748.84	49036	143408	94372	1.74	2.79	2341.46
Unemployment Rate	62	0.05	0.01	0.05	0.05	0.01	0.03	0.11	0.08	1.32	4.74	0.00
Population Density	62	3038.92	10862.42	112.20	321.56	101.24	3.01	69871.04	69868.03	4.59	22.49	1379.53
Bachelor Or Higher (%)	62	29.22	9.42	25.65	28.20	7.78	16.60	62.00	45.40	1.09	0.94	1.20
No Diploma (%)	62	6.25	1.78	6.05	6.13	1.70	2.20	12.10	9.90	0.68	0.84	0.23
Age 25-44 (%)	62	0.24	0.03	0.24	0.24	0.02	0.16	0.36	0.20	1.20	4.42	0.00
Total Population	62	320521.27	555885.19	86966.50	175956.46	66273.70	5102.00	2646306	2641204	2.55	6.11	70597.49
Below Poverty Level (%)	62	13.09	3.46	13.05	13.05	2.89	5.3	26.90	21.60	0.67	2.78	0.44

表 2-2 敘述性統計量

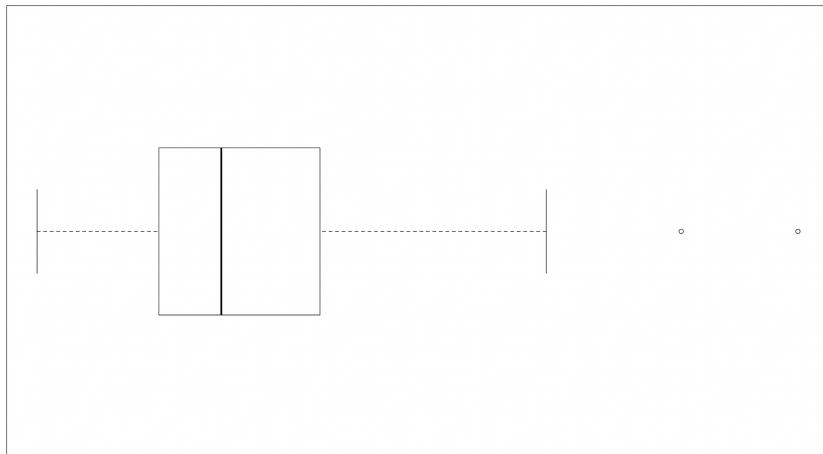


圖 2-2 紐約州各郡(County)在2023年的指標性犯罪率-盒鬚圖

#### 2.4.2. 自變數與依變數關係分析

從圖2-3可觀察到貧窮率（Poverty Incidence, PI）、失業率（Unemployment Rate, UR）、18歲以上學士學位或學士以上學位人口比例（Bachelor's or higher Education Attainment Rate, BoH）與年齡為25到44歲人口比例（Age from 25 to 44, Age25\_44）個別與反應變數大致呈現線性關係。收入中位數（Median Income）與18歲以上不具有高中畢業文憑比例（No High School Diploma, ND）則個別與反應變數之間沒有明顯的線性關係。此外，可看出人口密度（Population Density, PD）與總人口數（Total Population, TP）的嚴重右偏現象，雖然與反應變數的趨勢明顯，但是可以觀察到紐約市五大郡跟其他郡的表現明顯不同，趨勢線很可能很大程度受到紐約市五大郡的影響。因此我們決定對人口密度（Population Density, PD）與總人口數（Total Population, TP）取其對數值，希望能夠減緩這兩個變數的偏態性以及降低極端值對整體趨勢的影響。從圖2-4可看出，取完對數值後的人口密度（Population Density, PD）與總人口數（Total Population, TP）個別與反應變數的線性關係更為明顯，紐約市五大郡雖然表現還是與其他地區明顯不同，但是對於整體趨勢的影響比較沒那麼大，因此我們最終決定不直接使用人口密度（Population Density, PD）與總人口數（Total Population, TP）的數值，而是將兩個變數取對數值後放入模型

中。另外，由皮爾森相關係數表可觀察到變數之間沒有太高的相關性，可初步判斷原始資料間沒有嚴重的多重共線性。

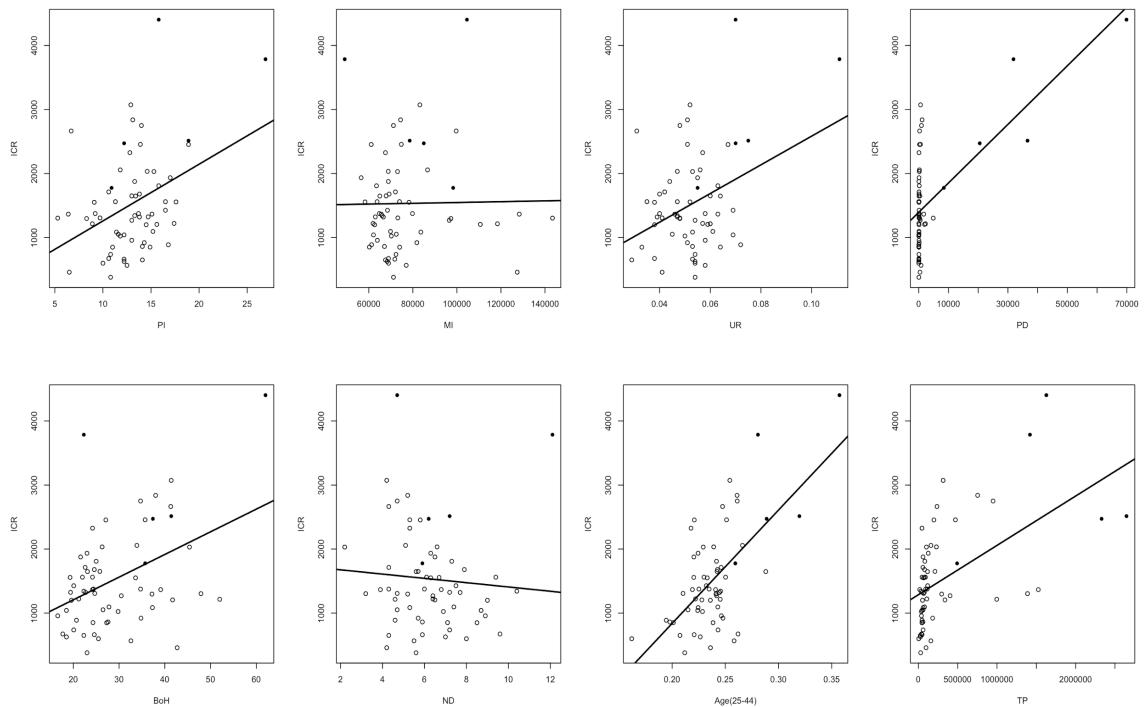


圖 2-3 各解釋變數與指標性犯罪率（Index Crime Rate, ICR）之散佈圖與趨勢線

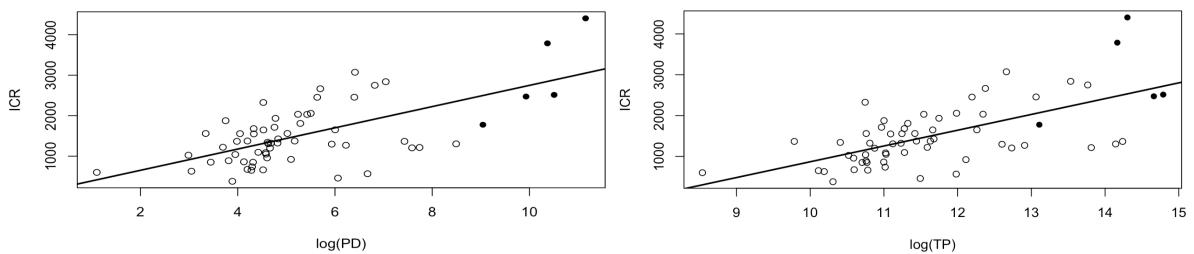


圖 2-4 PD與TP取對數值後與ICR之散佈圖與趨勢線

	Index Crime Rate	Median Income	Unemployment Rate	Population Density	Bachelor Or Higher	No Diploma	Age 25-44	Total Population	Below Poverty Level
Index Crime Rate	1.000	0.015	0.365	0.629	0.422	-0.076	0.637	0.542	0.388
Median Income	0.015	1.000	-0.190	0.173	0.774	-0.515	0.235	0.422	-0.628
Unemployment Rate	0.365	-0.190	1.000	0.488	0.028	0.274	0.289	0.420	0.643
Population Density	0.629	0.173	0.488	1.000	0.473	0.067	0.730	0.714	0.326
Bachelor Or Higher	0.422	0.774	0.028	0.473	1.000	-0.577	0.461	0.570	-0.298
No Diploma	-0.076	-0.515	0.274	0.067	-0.577	1.000	0.059	-0.097	0.463
Age 25-44	0.637	0.235	0.289	0.730	0.461	0.059	1.000	0.658	0.171
Total Population	0.542	0.422	0.420	0.714	0.570	-0.097	0.658	1.000	0.103
Below Poverty Level	0.388	-0.628	0.643	0.326	-0.298	0.463	0.171	0.103	1.000

表 2-3 皮爾森相關係數

### 3. 模型建構

#### 3.1. 初始模型

我們利用所選的八個解釋變數與一個虛擬變數來建構迴歸模型，分別為貧窮率（PI）、收入中位數（MI）、失業率（UR）、人口密度（PD）的對數值、總人口數（TP）的對數值、年齡為25到44歲人口比例（Age25\_44）、18歲以上學士學位或學士以上學位人口比例（BoH）、18歲以上不具有高中畢業文憑比例（ND）以及虛擬變數（D）：是否為紐約市五大郡。模型如下：

$$Y = \beta_0 + \beta_1 PI + \beta_2 MI + \beta_3 UR + \beta_4 \log(PD) + \beta_5 ND + \beta_6 Age25\sim44 + \beta_7 \log(TP) + \beta_8 BoH + \beta_9 D + \varepsilon$$

然而，由圖3-1的迴歸分析的結果顯示，大部分的解釋變數在所有解釋變數一起加入模型時的表現都不顯著。只有MI、BoH以及logTP有一定程度的顯著性。模型的F-Statistic為10.32、P Value小於0.01的顯著水準，可推知有足夠的信心認定模型至少存在一個解釋變數對反應變數有顯著的影響。此外，模型的調整後 $R^2$ 為0.6272，可見模型對於反應變數存在一定的解釋力。然而，我們認為導致大部分解釋變數沒有顯著性的原因除了某些解釋變數的確與犯罪率關聯薄弱之外，資料間的多重共線性也可能是其中抑制顯著性的原因。Kutner, Nachtsheim, & Neter (2004) 建議當  $VIF_j > 10$  時表示第j個解釋變數是一個多重共線性的變數。由表3-1可見， $\log(PD)$ 的VIF值高達16.3471，有明顯的多重共線性，另外， $\log(TP)$ 的VIF值為8.0655，在10附近，可能也有存在多元共線性的風險。我們認為將所有解釋變數一同放入模型時，會有不可忽視的多重共線性存在。

Residuals:					
	Min	1Q	Median	3Q	Max
	-1259.46	-249.72	-80.09	211.81	1164.84
Coefficients:					
	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	-1.947e+03	1.438e+03	-1.354	0.18167	
PI	2.758e+01	3.696e+01	0.746	0.45883	
MI	-2.749e-02	8.647e-03	-3.179	0.00249 **	
UR	-5.864e+03	7.301e+03	-0.803	0.42554	
$\log(PD)$	3.494e+01	1.299e+02	0.269	0.78896	
BoH	4.006e+01	1.447e+01	2.769	0.00777 ***	
ND	-2.133e+01	5.278e+01	-0.404	0.68780	
Age25_44	5.189e+03	3.776e+03	1.374	0.17525	
$\log(TP)$	2.606e+02	1.344e+02	1.938	0.05800 .	
D	1.400e+02	4.540e+02	0.308	0.75902	
---					
Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’	0.1 ‘ ’ 1
Residual standard error:	482.7	on 52 degrees of freedom			
Multiple R-squared:	0.6822		Adjusted R-squared:	0.6272	
F-statistic:	12.4	on 9 and 52 DF,	p-value:	3.25e-10	

圖 3-1 初始模型迴歸結果

	VIF
PI	4.2802
MI	6.6551
UR	2.3114
$\log(PD)$	16.3471
BoH	4.8609
ND	2.3210
Age25_44	3.0155
$\log(TP)$	8.0655
D	4.0669

表 3-1 VIF值

## 3.2. 模型選擇

### 3.2.1. 模型指標

在建立迴歸模型時，AIC（Akaike Information Criterion）、「 $C_p$ 」與 BIC（Bayesian Information Criterion）是常見的模型選擇準則，這些指標皆屬於以資訊量為基礎的評估指標，用來平衡模型的「解釋能力」與「複雜度」。

#### AIC（赤池資訊量準則）

AIC 的目的是尋找具有良好預測能力的模型，其計算公式為：

$$AIC = -2\log(L) + 2k$$

其中L為模型的最大概似函數值，k為模型中的解釋變數個數。AIC 越小，表示模型在擬合資料與避免過度複雜之間取得更佳平衡。AIC 通常較適合用於預測導向的研究，偏好保留更多變數以提升模型表現。

#### BIC（貝氏資訊量準則）

BIC 與 AIC 類似，但對模型複雜度的懲罰較重，其公式為：

$$BIC = -2\log(L) + k\log(n)$$

其中n為樣本數。由於懲罰項中加入 $\log(n)$ ，BIC 在樣本數較大時會更嚴格限制變數的數量，因此傾向選擇較簡潔的模型。BIC 適合用於解釋導向的研究，尤其在強調模型可解釋性時更為合適。

#### Mallow's $C_p$

Mallow's  $C_p$ 統計量也是一種評估模型的指標，它同時考慮模型的擬合程度以及所使用的變數數量。其原始定義為：

$$C_p = \frac{RSS(p)}{\sigma_{full}^2} - N + 2p$$

其中，

$$RSS(p) = \sum_{n=1}^N (y_n - x_n \widehat{\beta}_p)^2$$

$\widehat{\sigma_{full}}^2$ : 完整模型的誤差變異估計

$N$ : 觀察值數量

$p$ : 模型中解釋變數個數

若是  $C_p$  大於  $1 + p$ ，則視此模型預測誤差偏大。因此通常選擇  $C_p$  小於  $1 + p$  的模型，而理想模型的  $C_p$  值會接近  $p$ 。

### 3.2.2. 向前選取法

向前選取法是一種從「空模型」開始建構迴歸模型的方法。具體操作流程為：一開始先找出對結果（反應變數）影響最大的一個解釋變數，把它加入模型中。接著，再從剩下的解釋變數中挑一個能讓模型變得更好的（使模型評估指標（AIC 或 BIC）下降幅度最大），繼續加入，重複這個過程，直到沒有其他解釋變數能再顯著改善模型為止。

本研究以此法搭配 AIC 為準則時選入 logTP、MI、BoH 與 Age25\_44 這四項解釋變數，AIC 值為 944.7516，調整後  $R^2$  值為 0.6518。而以 BIC 為準則時，選入 logTP、MI 與 BoH 這三項解釋變數，BIC 值為 957.5144，調整後  $R^2$  值為 0.6354。

<pre> Call: lm(formula = Y ~ logTP + MI + BoH + Age25_44, data = data)  Residuals:     Min      1Q   Median      3Q     Max  -1292.62 -254.80  -55.64  201.59 1169.62   Coefficients:             Estimate Std. Error t value Pr(&gt; t )     (Intercept) -2.446e+03  6.573e+02 -3.721 0.000456 *** logTP        3.106e+02  7.686e+01  4.041 0.000161 *** MI          -3.078e-02  5.314e-03 -5.793 3.13e-07 *** BoH         4.482e+01  1.177e+01  3.807 0.000347 *** Age25_44    5.779e+03  2.995e+03  1.929 0.058656 .   --- Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  Residual standard error: 466.5 on 57 degrees of freedom Multiple R-squared:  0.6746,    Adjusted R-squared:  0.6518  F-statistic: 29.54 on 4 and 57 DF,  p-value: 2.57e-13 </pre>	<pre> Call: lm(formula = Y ~ logTP + MI + BoH, data = data)  Residuals:     Min      1Q   Median      3Q     Max  -1217.85 -278.50  11.51  197.15 1183.57   Coefficients:             Estimate Std. Error t value Pr(&gt; t )     (Intercept) -1.989e+03  6.275e+02 -3.170 0.002434 **  logTP        3.975e+02  6.371e+01  6.239 5.49e-08 ***  MI          -3.354e-02  5.238e-03 -6.403 2.92e-08 ***  BoH         4.887e+01  1.186e+01  4.122 0.000121 ***  --- Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  Residual standard error: 477.3 on 58 degrees of freedom Multiple R-squared:  0.6533,    Adjusted R-squared:  0.6354  F-statistic: 36.44 on 3 and 58 DF,  p-value: 2.278e-13 </pre>
---	--

圖 3-2 向前選取法(AIC)

圖 3-3 向前選取法(BIC)

### 3.2.3. 向後選取法

向後選取法剛好跟向前法相反，是從「全模型」開始，也就是先把所有的解釋變數都放進去，然後一步一步把影響力最小、解釋力最弱的解釋變數移除。每次剔除一個解釋變數後就重新檢查模型，看看還有沒有解釋變數是不需要的，直到剩下來的解釋變數都對模型有幫助使得模型評估指標（AIC 或 BIC）下降幅度最大。

本研究以此法搭配 AIC 為準則時選入 logTP、MI、BoH 與 Age25\_44 這四個解釋變數，AIC 值為 944.7516，調整後  $R^2$  值為 0.6518。而以 BIC 為準則時，選入 logTP、MI 與 BoH 這三項解釋變數，BIC 值為 957.5144，調整後  $R^2$  值為 0.6354。

```
Call:  
lm(formula = Y ~ MI + BoH + Age25_44 + logTP, data = data)  
  
Residuals:  
    Min      1Q   Median      3Q     Max  
-1292.62 -254.80  -55.64  201.59 1169.62  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) -2.446e-03 6.573e-02 -3.721 0.000456 ***  
MI           -3.078e-02 5.314e-03 -5.793 3.13e-07 ***  
BoH          4.482e-01 1.177e-01  3.807 0.000347 ***  
Age25_44     5.779e+03 2.995e+03  1.929 0.058656 .  
logTP        3.106e-02 7.686e-01  4.041 0.000161 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 466.5 on 57 degrees of freedom  
Multiple R-squared:  0.6746,   Adjusted R-squared:  0.6518  
F-statistic: 29.54 on 4 and 57 DF,  p-value: 2.57e-13
```

圖 3-4 向後選取法(AIC)

```
Call:  
lm(formula = Y ~ MI + BoH + logTP, data = data)  
  
Residuals:  
    Min      1Q   Median      3Q     Max  
-1217.85 -278.50  11.51  197.15 1183.57  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) -1.989e+03 6.275e+02 -3.170 0.002434 **  
MI           -3.354e-02 5.238e-03 -6.403 2.92e-08 ***  
BoH          4.887e+01 1.186e+01  4.122 0.000121 ***  
logTP        3.975e+02 6.371e+01  6.239 5.49e-08 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 477.3 on 58 degrees of freedom  
Multiple R-squared:  0.6533,   Adjusted R-squared:  0.6354  
F-statistic: 36.44 on 3 and 58 DF,  p-value: 2.278e-13
```

圖 3-5 向後選取法(BIC)

### 3.2.4. 逐步迴歸

逐步迴歸可以想像成是「向前選取」和「向後選取」的綜合版。它會在每加入一個新解釋變數的同時，也去檢查原本加入的解釋變數是不是還有存在的必要。如果發現某個解釋變數不再顯著，就會把它移除。這樣模型能在過程中不斷微調，確保保留的解釋變數都有貢獻。這個方法比單純的向前或向後來得靈活，能更有效控制解釋變數的多寡與模型的複雜程度。不過，最後得到的模型結果還是會受到起始變數、選擇順序與資料特性影響，不一定就是「唯一正確」的模型，因此常常會搭配全子集方法一起使用，來確認結果是否穩定。

本研究以此法搭配 AIC 為準則時選入 logTP、MI、BoH 與 Age25\_44 這四個解釋變數，AIC 值為 944.7516，調整後  $R^2$  值為 0.6518。而以 BIC 為準則時，選入 logTP、MI 與 BoH 這三項解釋變數，BIC 值為 957.5144，調整後  $R^2$  值為 0.6354。

```

Call:
lm(formula = Y ~ logTP + MI + BoH + Age25_44, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-1292.62 -254.80  -55.64  201.59 1169.62 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.446e+03 6.573e+02 -3.721 0.000456 ***  
logTP        3.106e+02 7.686e+01  4.041 0.000161 ***  
MI          -3.078e-02 5.314e-03 -5.793 3.13e-07 ***  
BoH         4.482e+01 1.177e+01  3.807 0.000347 ***  
Age25_44    5.779e+03 2.995e+03  1.929 0.058656 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 466.5 on 57 degrees of freedom
Multiple R-squared:  0.6746, Adjusted R-squared:  0.6518 
F-statistic: 29.54 on 4 and 57 DF,  p-value: 2.57e-13

```

圖 3-6 逐步選取法(AIC)

```

Call:
lm(formula = Y ~ logTP + MI + BoH, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-1217.85 -278.50  11.51  197.15 1183.57 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.989e+03 6.275e+02 -3.170 0.002434 **  
logTP        3.975e+02 6.371e+01  6.239 5.49e-08 ***  
MI          -3.354e-02 5.238e-03 -6.403 2.92e-08 ***  
BoH         4.887e+01 1.186e+01  4.122 0.000121 ***  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 477.3 on 58 degrees of freedom
Multiple R-squared:  0.6533, Adjusted R-squared:  0.6354 
F-statistic: 36.44 on 3 and 58 DF,  p-value: 2.278e-13

```

圖 3-7 逐步選取法(BIC)

方法	評估指標	選出模型變數	AIC 值	BIC 值	Adjusted R^2 值
向前選取法	AIC	logTP、MI、BoH、Age25_44	944.7516	957.5144	0.6518
	BIC	logTP、MI、BoH	946.6742	957.3099	0.6354
向後選取法	AIC	logTP、MI、BoH、Age25_44	944.7516	957.5144	0.6518
	BIC	logTP、MI、BoH	946.6742	957.3099	0.6354
逐步選擇法	AIC	logTP、MI、BoH、Age25_44	944.7516	957.5144	0.6518
	BIC	logTP、MI、BoH	946.6742	957.3099	0.6354

表 3-2 各種模型選擇比較結果

### 3.2.5. 全子集迴歸

全子集迴歸（All Subset Regression）會窮舉所有解釋變數組合的可能，因此在指定模型選擇指標（如AIC、BIC、 $C_p$ ）後，全子集迴歸即可從所有模型組合中找到指標表現最好的組合，作為候選的最佳模型。表3-3以及表3-4列出進行全子集迴歸後AIC值前十小的模型組合，以及BIC值前十小的模型組合。可以看到結果顯示，AIC值最低的模型選到MI、BoH、logTP與 Age25\_44，結果與使用向前選取法、向後選取法以及逐步選取法使用AIC作為選取依據的結果一致。另一方面，BIC值最低的模型選到MI、BoH與logTP，結果與使用向前選取法、向後選取法以及逐步選取法使用BIC作為選取依據的結果同樣一致。

Combination	Adj R <sup>2</sup>	AIC	BIC	Cp
MI + BoH + log_TP + Age25_44	0.6517626	944.7516	957.5144	1.250404
MI + BoH + log_TP + log_PD + Age25_44	0.6475662	946.3969	961.2868	2.946618
MI + BoH + log_TP + PI + Age25_44	0.6470629	946.4854	961.3753	3.022231
MI + BoH + log_TP + Age25_44 + D	0.6465730	946.5714	961.4613	3.095823
MI + BoH + log_TP + UR + Age25_44	0.6462073	946.6355	961.5254	3.150766
MI + BoH + log_TP	0.6354145	946.6742	957.3099	2.728335
MI + BoH + log_TP + ND + Age25_44	0.6458045	946.7060	961.5960	3.211271
MI + BoH + log_TP + log_PD	0.6401561	946.7843	959.5471	3.025203
MI + BoH + log_TP + D	0.6367169	947.3741	960.1369	3.551106
MI + BoH + log_TP + PI + UR + Age25_44	0.6438003	947.9387	964.9558	4.556788

表 3-3 AIC值最好的前十名模型組合

Combination	Adj R <sup>2</sup>	AIC	BIC	Cp
MI + BoH + log_TP	0.6354145	946.6742	957.3099	2.728335
MI + BoH + log_TP + Age25_44	0.6517626	944.7516	957.5144	1.250404
MI + BoH + log_TP + log_PD	0.6401561	946.7843	959.5471	3.025203
MI + BoH + log_TP + D	0.6367169	947.3741	960.1369	3.551106
MI + BoH + log_PD	0.6150744	950.0402	960.6758	5.893187
MI + BoH + log_TP + ND	0.6310444	948.3347	961.0975	4.418509
MI + BoH + log_TP + PI	0.6303885	948.4448	961.2076	4.518805
MI + BoH + log_TP + log_PD + Age25_44	0.6475662	946.3969	961.2868	2.946618
MI + BoH + log_TP + UR	0.6298223	948.5397	961.3025	4.605378
MI + BoH + log_TP + PI + Age25_44	0.6470629	946.4854	961.3753	3.022231

表 3-4 BIC值最好的前十名模型組合

### 3.2.6. 最佳模型

我們分別以AIC、BIC、 $C_p$ 與調整後 $R^2$ 為準則進行模型選擇的參考依據，根據向前選取法、向後選取法、逐步選取法以及全子集迴歸的窮舉考量，結果顯示 AIC值最佳的模型納入4個解釋變數分別為MI、BoH、logTP與 Age25\_44，得到0.6518的調整後 $R^2$ ， $C_p$ 值約為1.25小於模型中解釋變數數量。而 BIC值最佳的模型則納入3個變數，分別為MI、BoH與logTP，得到0.6354的調整後 $R^2$ ， $C_p$ 值約為2.73，接近模型中解釋變數數量。考量到本研究希望在不失解釋力的情況下，尋求最簡潔的模型。雖然在以 AIC為準則下比起以BIC為準則額外納入了 Age25\_44，使模型的調整後  $R^2$  從0.6354 提升至0.6518，增加了約0.02的解釋力，但相對地也提高了模型的複雜度。我們認為此提升幅度有限，相較之下模型增加的複雜度更為明顯。此外， $C_p$ 指標的模型選擇方法指出，理想模型的 $C_p$ 應近似於模型中解釋變數的數量，以AIC為準則選出的模型 $C_p$ 值明顯小於解釋變數數量，而以BIC為準則選出的模型的 $C_p$ 接近解釋變數數量。故本研究最終採用以 BIC為準則所選出之三變數MI、BoH與logTP作為最佳模型。亦即：

$$y = \beta_0 + \beta_1 MI + \beta_2 BoH + \beta_3 log(TP) + \varepsilon$$

為最佳模型，此外，模型中變數MI的VIF值為2.496、BoH為3.337、log(TP)為1.852，沒有明顯的多重共線性。

```

Call:
lm(formula = Y ~ MI + BoH + log_TP, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-1217.85 -278.50   11.51  197.15 1183.57 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.989e+03 6.275e+02 -3.170 0.002434 **  
MI          -3.354e-02 5.238e-03 -6.403 2.92e-08 ***  
BoH          4.887e+01 1.186e+01  4.122 0.000121 ***  
log_TP       3.975e+02 6.371e+01  6.239 5.49e-08 ***  
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 477.3 on 58 degrees of freedom
Multiple R-squared:  0.6533,    Adjusted R-squared:  0.6354 
F-statistic: 36.44 on 3 and 58 DF,  p-value: 2.278e-13

```

圖 3-8 最佳模型回歸結果

## 4. 模型診斷與假設檢定

### 4.1. 迴歸假設檢驗

多元線性迴歸的模型的推論應在模型符合多元線性迴歸對於模型的假設成立時才有意義，考慮多元線性回歸模型如下：

$$Y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} + \varepsilon_i, i = 1, \dots, n$$

其中多元線性迴歸對於模型的假設為：

$$\varepsilon_i \sim N(0, \sigma^2), i = 1, \dots, n$$

且殘差之間彼此獨立。因此我們經由各種模型選擇方法選出的最佳模型還需要檢定模型是否符合多元線性迴歸中的假設，才能確保對於模型的各種推論是正確的。

#### 殘差常態性檢定

本研究利用 Shapiro-Wilk 檢定（Shapiro & Wilk, 1965），同時也參考Q-Q Plot（Quantile-Quantile Plot）以及P-P Plot（Probability-probability Plot）的結果來檢測殘差的常態性。

Shapiro-Wilk檢定的假設為：

$$H_0: \text{資料為常態分佈}$$

$$H_1: \text{資料不為常態分佈}$$

由R程式計算結果得出檢定統計量 $W = 0.97295$ ，P-Value為0.187。結果顯示在0.05的顯著水準下，資料無法拒絕殘差服從常態分配的虛無假設。然而，因為P-Value接近顯著水準門檻，並且從Q-Q圖中可觀察到尾端部分偏離理論值。為了結果的穩定性，我們決定於後續對反應變數進行轉換以調整殘差的分佈結構。

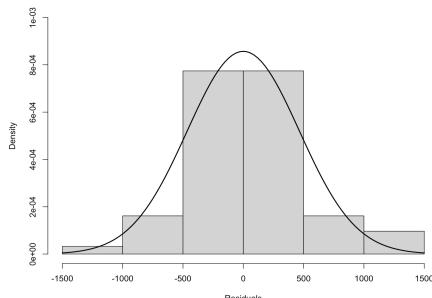


圖 4-1 直方圖

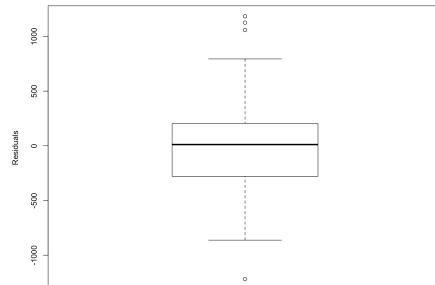


圖 4-2 盒型圖

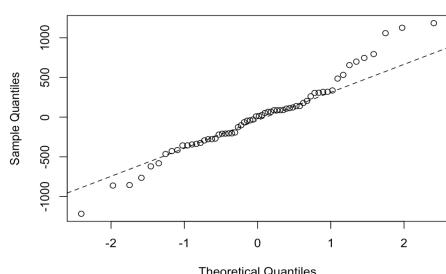


圖 4-3 Q-Q Plot

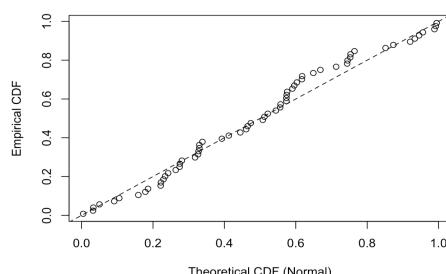


圖 4-4 P-P Plot

## 殘差變異數齊一性檢定

本研究使用Breusch - Pagan test作為殘差變異數齊一性的檢定。Breusch - Pagan test的核心理念為檢定殘差是否會隨解釋變數而改變。其檢定流程如下：

考慮多元迴歸模型

$$Y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} + \varepsilon_i, i = 1, \dots, n$$

估計迴歸式

$$\varepsilon_i^2 = \gamma_0 + \gamma_1 x_{1i} + \dots + \gamma_p x_{pi} + error_i, i = 1, \dots, n$$

另外，求出  $R_{\varepsilon^2}^2$ ，即可得到檢定統計量

$$LM = n * R_{\varepsilon^2}^2 \sim \chi_k^2$$

n為樣本數，k為模型中解釋變數數量。Breusch - Pagan test的假設為：

$$H_0: \gamma_0 = \dots = \gamma_p = 0$$

意即所有解釋變數對於殘差都沒有顯著的影響。由R程式計算得出  $LM = 17.084$ ，P-Value為0.000679。有足夠的證據拒絕  $H_0$ 。此外，由圖4-6所示的之 Spread - Level Plot可看出Studentized殘差的絕對值隨著預測值變大而增加，支持上述檢定結果。因此我們判定我們的模型的殘差變異數存在異質性，違反多元迴歸假設。此結果亦增強了我們需要對資料進行轉換之必要性，嘗試降低殘差變異數的異質性。

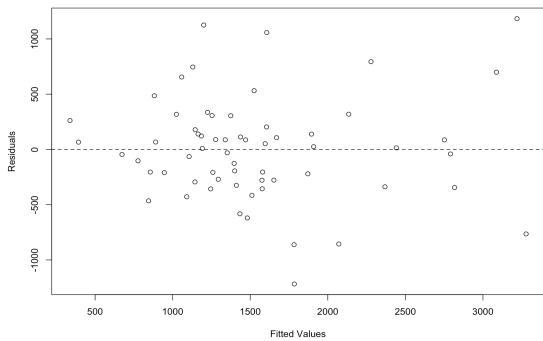


圖 4-5 殘差圖

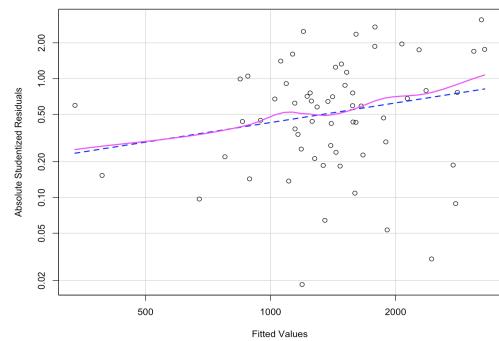


圖 4-6 Spread - Level Plot

## 殘差獨立性檢定

本研究使用Durbin-Watson Test來檢定殘差的獨立性。Durbin - Watson Test會先計算出統計量(D值)，再根據樣本數(n)、模型中自變數個數(p)以及所要求的顯著水準，查表得到臨界值下界  $d_L$  與上界  $d_U$ ，用來判斷是否存在一階自相關。其假設如下：

$$H_0: \text{殘差之間沒有一階自相關性}$$

$$H_1: \text{殘差之間存在一階自相關性}$$

檢定統計量

$$d = \frac{\sum_{i=2}^n (\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^n (\varepsilon_i)^2}$$

其中  $\varepsilon_i$  為第i個殘差值。

若是 $d < d_L$ ，則拒絕 $H_0$ ，若是 $d > d_U$ ，則不拒絕 $H_0$ ，若 $d_L < d < d_U$ ，則無法得出結論。

本研究樣本數為62，檢驗的模型中包含MI、BoH以及log(TP)三個解釋變數，此外設定顯著水準為0.05，查表得到 $d_L$ 約為1.48， $d_U$ 約為1.69。經由R程式計算得出d值為2.3495，明顯超出 $d_U$ ，此外也計算出P-Value為0.9171。因此無法拒絕 $H_0$ ，判斷殘差之間不存在一階相關性。然而，受限於Durbin-Watson Test只能檢定資料的一階自相關性，無法檢測更高階的自我相關性。本研究另外參考了自相關函數(Autocorrelation Function, ACF)以及偏自我相關函數(Partial Autocorrelation Function, PACF)來檢視殘差之間更高階數的延遲下的自我相關性。由圖4-8可觀察到在95%的信心水準下，大部分階數的延遲下都沒有明顯的自我相關性，唯獨在PACF圖中Lag 9的地方出現高點，顯示該階可能出現相關性。因為殘差整體的相關性表現不明顯，因此我們決定於後續進行對反應變數的轉換後再觀察此現象是否持續存在。

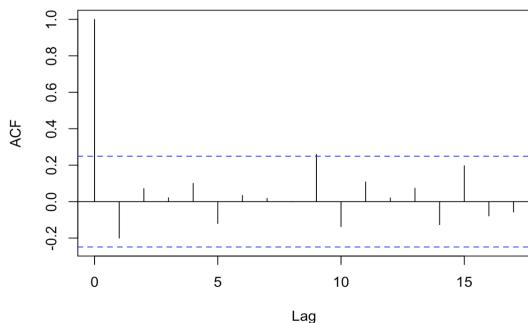


圖 4-7 ACF圖

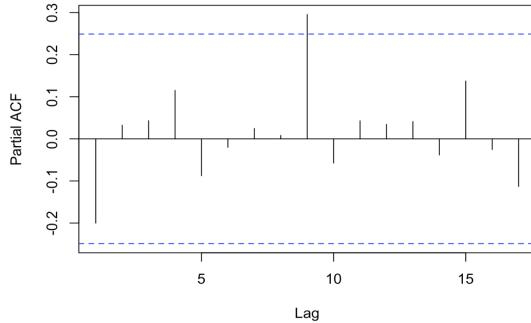


圖 4-8 PACF圖

## 4.2. Box-Cox轉換

由於殘差的Shapiro-Wilk檢定結果之P-Value距離顯著水準的臨界值不遠，為提升模型結果的穩定性，並同時因應後續變異數齊一性檢定中顯示的異質性問題，故考慮對反應變數進行Box-Cox轉換，以期改善殘差結構並提升模型對基本假設的符合程度。

### Box-Cox轉換

Box-Cox 轉換 (Box-Cox transformation) 由Box和Cox (1964) 所提出，用於將不符合常態性的資料轉換成近似常態的型態。其轉換公式如下：

假設所有資料y皆大於0，對y使用以下轉換

$$\psi(y; \lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(y) & \text{if } \lambda = 0 \end{cases}$$

而 $\lambda$ 的數值則是由轉換後迴歸模型的最大概似估計值。

### Box-Cox轉換後的迴歸模型

經由R程式的計算，我們計算出的 $\lambda$ 值為0.424。實務上，為了結果的易解讀性， $\lambda$ 值接近0.5常會直接對反應變數取開根號的轉換。因此我們選擇採用開根號的方式。轉換後模型如下：

$$\sqrt{y} = \beta_0 + \beta_1 MI + \beta_2 BoH + \beta_3 \log(TP) + \varepsilon$$

Box-Cox轉換後的模型與未做Box-Cox轉換的模型相較之下， $R^2$ 下降了0.0205。另外，截距項顯著性下降，原本在0.01顯著水準下顯著，轉換後不再顯著。然而，其餘三個解釋變數在轉換前後皆維持高度顯著。

```
Call:
lm(formula = Y_BoxCox ~ MI + BoH + log_TP, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-17.0344 -3.2460  0.2697  2.9639 14.2637 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.538e+00  7.822e+00 -0.836 0.406670    
MI          -3.913e-04  6.528e-05 -5.995 1.39e-07 ***  
BoH         5.123e-01  1.478e-01  3.467 0.000998 ***  
log_TP      5.074e+00  7.942e-01  6.390 3.08e-08 ***  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 5.95 on 58 degrees of freedom
Multiple R-squared:  0.6338,   Adjusted R-squared:  0.6149 
F-statistic: 33.47 on 3 and 58 DF,  p-value: 1.099e-12
```

圖 4-9 Box-Cox轉換後模型表現

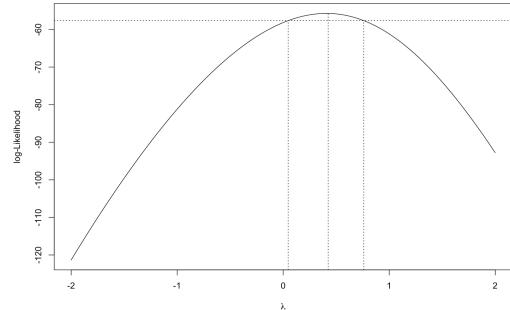


圖 4-10 Box-Cox lambda值

### 4.3. Box-Cox轉換後的迴歸假設檢驗

#### 殘差常態性檢定

轉換後的模型的殘差在進行Shapiro-Wilk 檢定後得到檢定統計量 $W = 0.9857$ ，P-Value為0.6871。明顯無法拒絕殘差為常態分佈的虛無假設。此外，Q-Q Plot以及P-P Plot上的點皆更貼近直線。因此我們判斷Box-Cox轉換後的模型之殘差比起原本的模型更接近常態分配。

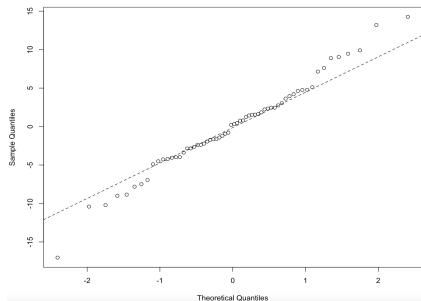


圖 4-11 Box-Cox轉換後Q-Q Plot

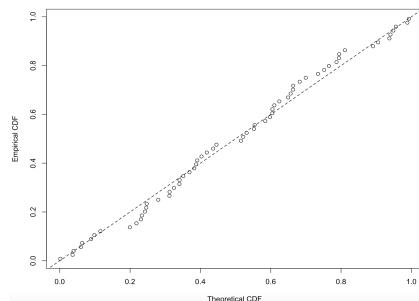


圖 4-12 Box-Cox轉換後P-P Plot

#### 殘差變異數齊一性檢定

轉換後的模型的殘差在進行Breusch - Pagan檢定後得到檢定統計量 $LM = 4.787$ ，P-Value為0.1881。在0.1或0.05的顯著水準下皆無法拒絕殘差變異數為同質變異的虛無假設。另外，由Spread - Level Plot可看出殘差之間不再有原先Studentized殘差的絕對值隨著預測值變大而增加的趨勢，而是變得平緩。殘差圖也呈現隨機分佈在0附近的狀態，沒有特別的趨勢。因此我們認為Box-Cox轉換後的模型之殘差符合同質變異性的假設。

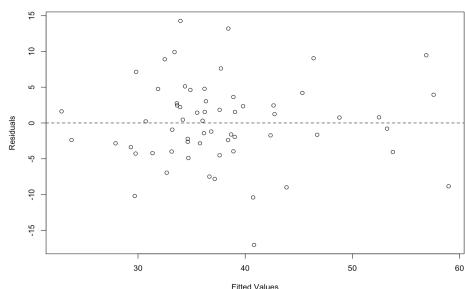


圖4-13 Box-Cox轉換後殘差圖

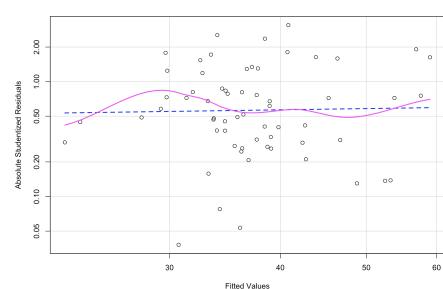


圖4-14 Box-Cox轉換後Spread-Level Plot

## 殘差獨立性檢定

轉換後的模型的殘差在進行Durbin-Watson 檢定後得到檢定統計量d=2.2659，P-Value為0.8528。轉換後的模型同樣使用62筆樣本、模型中同樣為三個解釋變數，在顯著水準設為0.05的條件下， $d_L$ 約為1.48， $d_U$ 約為1.69，檢定統計量d大於 $d_U$ ，因此不拒絕殘差為獨立的虛無假設。另一方面，在原先模型的PACF圖中有發現在Lag 9的地方發現異常高的相關性。不過，在經過轉換後的模型之ACF或PACF圖中均沒有發現過高的相關性。因此我們認為經過Box-Cox轉換的模型之殘差符合獨立性之假設。

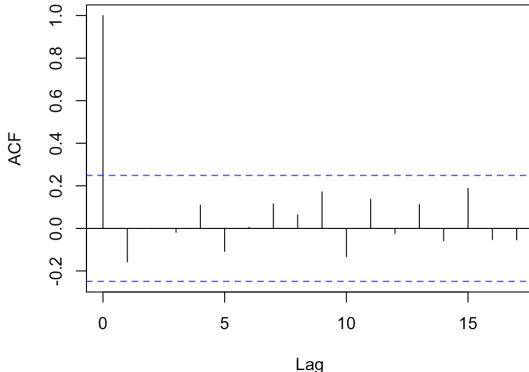


圖 4-15 Box-Cox轉換後PAC圖

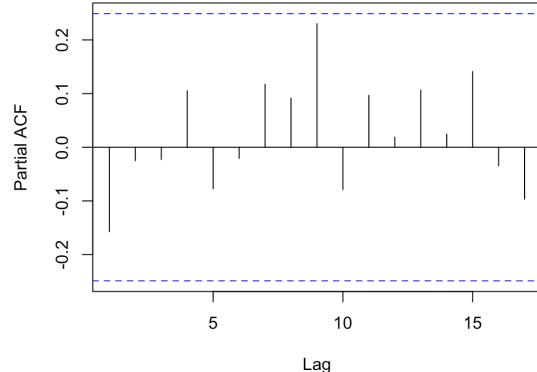


圖 4-16 Box-Cox轉換後FPAC圖

## 4.4. 離群值分析

### Cook Distance

Cook distance為Cook在1977年所提出，用以判斷資料中的影響點。其定義如下：

$$Cook_i = \frac{\sum_{j=1}^n [\widehat{y}_{j(-i)} - \widehat{y}_j]^2}{(p+I)s^2}$$

其中 $\widehat{y}_{j(-i)}$ 表示刪除第*i*個觀察值後模型對於 $y_j$ 的預測值， $s^2$ 為殘差之變異數，p為模型中解釋變數數量。

Fox (2022) 指出當 $Cook_i > \frac{4}{n-(p+I)}$ 時可視為影響點。本研究亦採用此做法。結果顯示資料中有四個樣本點可視為影響點，分別為Kings、New York、Saratoga以及Westcheste這四個郡。

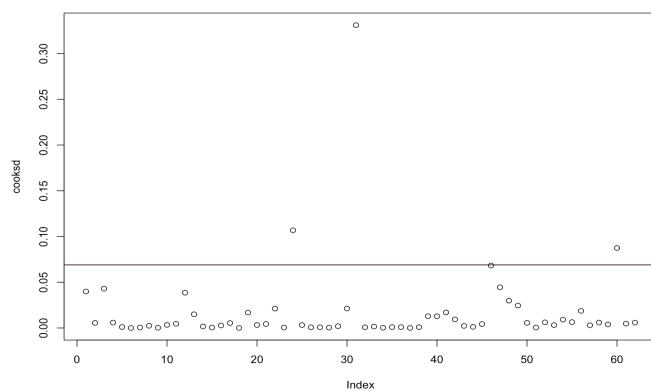


圖 4-17 Cook Distance

## DFFITS

DFFITS亦為另一個常用來判斷影響點的指標，其定義為：

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i(-i)}}{s_{i(-i)} \sqrt{h_{ii}}}$$

當  $DFFITS_i > \sqrt{\frac{p+1}{n-(p+1)}}$  時，可視該值為影響點。

結果顯示，與Cook distance之結論相同。同樣選出Kings、New York、Saratoga以及Westchester這四個郡為影響點。

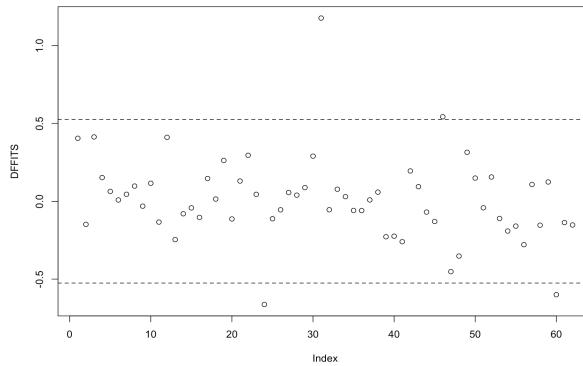


圖 4-18 DEFFITS

## 4.5. 敏感性分析

因為在資料中發現影響點的存在，為了觀察這些影響點是否明顯地改變整個模型，我們將進一步進行敏感性分析。透過分別將62筆觀察值抽離模型訓練集中，來觀察每一筆觀察值影響模型結果的程度。我們將分析三個層面的敏感性分析，第一，每筆觀察值抽離後對迴歸係數的改變。第二，每筆觀察值抽離後對解釋變數顯著性的改變。最後，每筆觀察值抽離後對調整後 $R^2$ 的改變。

首先，可以觀察到大多數情況抽離每一筆資料後，BoH的迴歸係數還是接近整體模型結果0.5123488，不過在抽離第31筆New York郡後，BoH的係數降為0.3712257。雖然BoH對犯罪率的影響方向並未改變，不過New York郡資料的加入確實很大程度的改變BoH對於犯罪率影響程度的估計。其餘兩個變數在敏感性分析下雖然還是有一定程度的高低起伏震盪，不過起伏程度不大。

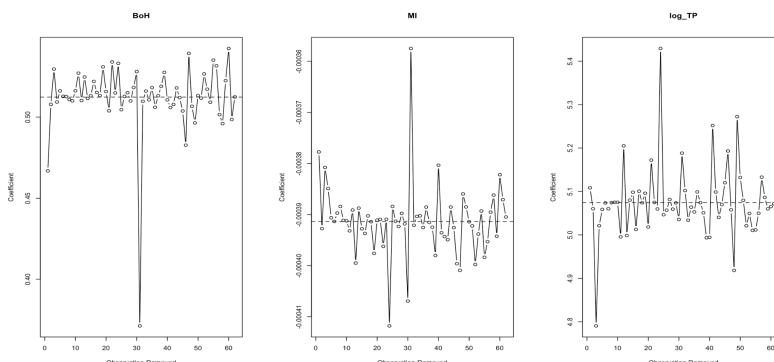


圖 4-19 每筆觀察值抽離後對迴歸係數的改變

另一方面，從敏感性分析對於解釋變數之P-Vaule的影響之結果來看，可以發現BoH在抽離New York郡後的顯著性降低，不過結果依然在0.05的顯著水準下具有顯著性。其餘變數則維持極低的P-Value。

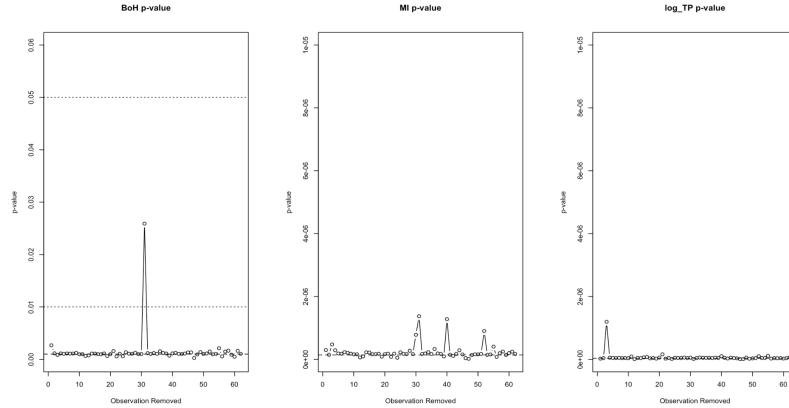


圖 4-20 每筆觀察值抽離後對解釋變數顯著性的改變

最後，觀察敏感性分析對調整後 $R^2$ 的改變，可以看出結果雖有高低起伏，最高點到0.658，最低至0.576，距離使用所有資料訓練之模型得到之調整後 $R^2=0.6149$ 距離最遠為0.0431。我們認為此變異程度不大，顯示模型在資料擬合的穩定性良好。

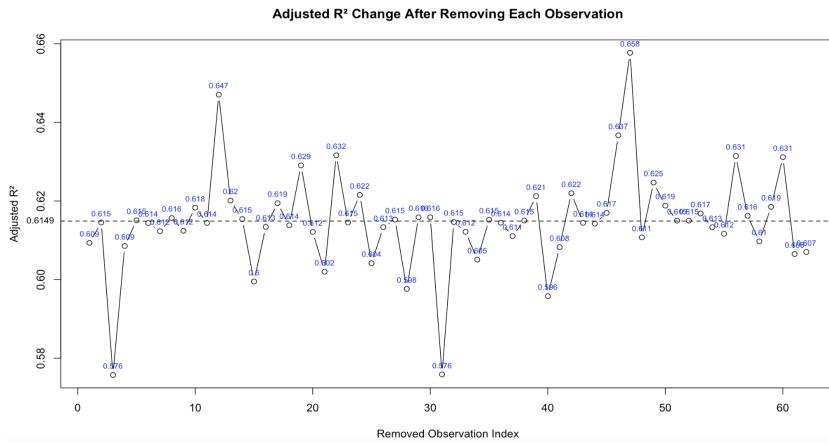


圖 4-21 每筆觀察值抽離後對調整後 $R^2$ 的改變

綜合敏感性分析結果，特別針對紐約郡（第31筆觀察值）的影響進行探討。我們發現，儘管紐約郡資料的納入確實對BoH（Bachelor degree or higher）的迴歸係數估計產生明顯影響，導致其數值由整體模型的 0.5123488 下降至抽離後的 0.3712257，且顯著性從極高水準略降至 0.05 顯著水準。然而，值得強調的是，BoH 對犯罪率的影響方向始終保持一致，未發生符號上的改變。這顯示模型中 BoH 對於犯罪率方向上的影響是穩定的。另一方面，其餘解釋變數的係數估計與顯著性在敏感性分析下均表現穩健，而調整後  $R^2$  的波動幅度亦在可接受範圍內。儘管單一觀察值對 BoH 係數的精確估計帶來了波動，促使我們對其數值穩定性抱持謹慎態度，但考量到影響方向的穩定性，以及其他變數的穩健表現，我們認為維持包含所有資料的原模型是合理的選擇。

## 5. 結果分析

在以AIC、BIC、 $C_p$ 等指標作為模型選擇依據，經過向前選取法、向後選取法、逐步迴歸以及全子集迴歸的選取後，從原先的八個解釋變數，分別為PI（貧窮率）、MI（收入中位數）、UR（失業率）、log\_PD（人口密度對數值）、log\_TP（總人口數對數值）、Age25\_44年齡為25到44歲人口比例）、BoH（18歲以上學士學位或學士以上學位人口比例）、ND（18歲以上不具有高中畢業文憑比例）以及一個虛擬變數（Dummy Variable）：是否為紐約市五大郡，選出了MI（收入中位數）、BoH（18歲以上學士學位或學士以上學位人口比例）以及log\_TP（總人口數對數值）作為最佳模型的解釋變數組合。此外，為了使殘差符合迴歸模型中的常態性、變異數齊一性及獨立性的假設，對反應變數進行了Box-Cox轉換，使模型最終符合所有假設，最終得到模型

$$\sqrt{y} = -6.538 - 0.0003913 MI + 0.5123 BoH + 5.074 \log(TP)$$

作為本研究之最終模型。此模型得到  $R^2 = 0.6338$  以及調整後  $R^2 = 0.6149$  的解釋力，並且 MI 的 P-Vaue 小於 0.01，達高度顯著，BoH 的 P-Vaue 小於 0.05，亦達顯著，log\_TP 的 P-Vaue 小於 0.01，同樣達高度顯著。就社會科學領域的實證分析結果來看，我們認為這樣的解釋力可被視為良好，模型能夠相當高程度地掌握實際犯罪率的變異。

另一方面，從迴歸係數的結果來看，可解釋為當一郡的 MI（收入中位數）增加一單位時，在其他條件不變的情況下，該郡之每十萬人口犯罪率之平方根值約減少 0.00039 單位。當一郡的 BoH（18 歲以上學士學位或學士以上學位人口比例）增加 1 單位時，在其他條件不變的情況下，該郡每十萬人口之犯罪率的平方根值約增加 0.512 單位。當一郡的 log\_TP（總人口數之對數值）增加一單位時，在其他條件不變的情況下，該郡每十萬人口之犯罪率的平方根值約增加 5.074 單位。就解釋變數對於犯罪率之影響方向來看，MI（收入中位數）與犯罪率是負相關。表示平均而言收入越高的地方犯罪率越低，符合我們的直覺，通常一個地區的發展良好，就會抑制人們想犯罪的念頭，尤其是對於重大犯罪。TP（總人口數）與犯罪率呈現正相關，這也十分符合我們的直覺，人口龐大的地方可能面臨資源分配以及管理難以全面顧及的問題。此外，人口眾多意味著人與人之間的互動更加頻繁，意味著衝突也可能增加。種種因素之下我們認為人口數與犯罪率呈現正相關是個合理的關係。然而，出乎我們意外的是 BoH（18 歲以上學士學位或學士以上學位人口比例）與犯罪率呈現正相關，我們認為可能的原因之一是因為高學歷者通常傾向到人口密集的大都市發展，而由原始資料可發現，紐約州各郡犯罪率最高的城市即為高度發展的大都市，可推測不一定是高學歷者犯罪率高，也可能是因為高學歷者通常會到高犯罪率地區發展。除此之外，高學歷人口比例地區的法律意識可能更為強烈，對於不法行為容忍度較低，因此犯罪案件更容易被揭露，由於本研究所使用的資料為官方統計的資料，意即犯罪案件需要被通報才會有記錄，因此也不排除這項可能。

## 6. 結論

本研究以紐約州62個郡為分析對象，透過多元線性迴歸模型探討影響紐約州各郡在2023年重大犯罪率的社會經濟因素。在所納入的八個潛在解釋變數中，最終模型篩選出「收入中位數」、「學士學位以上人口比例」以及「總人口數對數值」三個具有統計顯著性的重要變數，並建立其與紐約各郡重大犯罪率之間的定量關係。模型的決定係數  $R^2 = 0.6338$ ，調整後  $R^2 = 0.6149$ ，顯示本模型對於郡層級犯罪率的解釋力在社會科學領域中屬於良好水準，具有實證分析的參考價值。

從變數解釋方向來看，「收入中位數」與犯罪率呈現顯著負相關，顯示在其他條件不變下，收入越高的地區其重大犯罪率越低，這與 Schuessler (1962)、Bhattacharya (2020) 等研究指出貧困與犯罪正相關的結論一致。「總人口數」亦與犯罪率呈顯著正相關，支持 Gibbs & Erickson (1976) 與 Chang, Kim & Jeon (2019) 對於都市化與高犯罪風險的觀察結果，意即人口聚集可能加劇資源壓力以及增加人際互動，進而提升犯罪機率。然而，雖然 Bothos & Thomopoulos (n.d.) 的研究在探討社會和經濟因素影響美國各州犯罪率的社會和經濟因素後指出教育程度是犯罪率的顯著因子，指出教育程度與犯罪率的負向關係，與本研究學士學位以上人口比例與犯罪率呈現正相關的結論有落差，我們認為此差異可能是紐約州地域性的現象，由於本研究以紐約州62個郡為分析單位，高學歷人口多集中於人口密度高、社經活動頻繁的大都市區，而這些區域亦常伴隨較高的犯罪發生率。因此得出與文獻相反的結論。

綜合而言，本研究驗證了多篇文獻中收入以及人口數對於犯罪率之間相關性的結論。除此之外，揭示了在紐約州各郡之中，高學歷人口比例與犯罪率呈現正相關的關係，指出紐約州地域性的現象與全國各州之間的現象之間存在差異。然而，本研究無法釐清導致這一現象的原因，為了更全面了解這其中的關聯，未來研究可考慮納入更多教育程度相關的微觀行為解釋變數，以補足本研究僅限於宏觀變數的侷限，進一步釐清教育與犯罪率之間複雜的社會關係。

## 7. 附錄

### 7.1. 參考文獻

- Bhattacharya, A. (2020). *Analysis of the Factors Affecting Violent Crime Rates in the US*. International Journal of Engineering and Management Research, 10(5).18
- Bothos, M., & Thomopoulos, R. (n.d.). Bothos, J. M. A., & Thomopoulos, S. C. A. (n.d.). *Factors influencing crime rates: An econometric analysis approach*. Integrated Systems Laboratory, National Center for Scientific Research "Demokritos", Athens, Greece.

Chang, Y. S., Kim, H. E., & Jeon, S. (2019). *Do Larger Cities Experience Lower Crime Rates? A Scaling Analysis of 758 Cities in the U.S.* Sustainability, 11, 3111.

Gibbs, J. P., & Erickson, M. L. (1976). *Crime Rates of American Cities in an Ecological Context*. American Journal of Sociology, 82(3), 605 – 620.

Osgood, D. W. (2000). *Poisson-based regression analysis of aggregate crime rates*. Journal of Quantitative Criminology, 16(1)

Schuessler, K. (1962). *Components of Variation in City Crime Rates*. Social Problems, 9(4), 314 – 323.

## 7.2. 程式碼

```
1 # Read Data
2 data <- read.csv("/Users/chenpowei/Documents/大三下/RegressionAnalysis/紐約犯罪率/紐約州各個城市犯罪率RegressionData.csv", fileEncoding = "UTF-8")
3 colnames(data) <- c("County", "Y", "PI", "MI", "UR", "PD", "BoH", "ND", "Age25_44", "TP", "NYC")
4 Y = data$Y
5 PI = data$PI
6 MI = data$MI
7 UR = data$UR
8 log_PD = log(data$PD)
9 BoH = data$BoH
10 ND = data$ND
11 Age25_44 = data$Age25_44
12 log_TP = log(data$TP)
13 D = data$NYC
14 df <- data.frame(Y, PI, MI, UR, log_PD, BoH, ND, Age25_44, log_TP, D)
15 colnames(df) <- c("Y", "PI", "MI", "UR", "log_PD", "BoH", "ND", "Age25_44", "log_TP", "D")
16 #
17 # Initial Model
18 model = lm(Y~PI+MI+UR+log_PD+BoH+ND+Age25_44+log_TP+D)
19 summary(model)
20 library(car)
21 vif(model)
22 #
23 # null model
24 null.model <- lm(Y ~ 1, data=df)
25 summary(null.model)
26 # full model
27 full.model <- lm(Y ~ ., data=df)
28 summary(full.model)
29 #
30 # Forward Selection
31 n <- nrow(data)
32 # Forward Selection AIC
33 forward_model_aic <- step(null.model,
34 scope = list(lower = null.model, upper = full.model),
35 direction = "forward",
36 k = 2) # k = 2 表示使用 AIC
37 summary(forward_model_aic) # Y ~ log_TP + MI + BoH + Age25_44 (Adjusted R-squared: 0.6518)
38 vif(forward_model_aic) # log_TP:2.821697 MI:2.690206 BoH:3.446688 Age25_44 :2.031464
39 AIC(forward_model_aic) # 944.7516
```

```

40 BIC(forward_model_aic)      # 957.5144
41 # Forward Selection BIC
42 forward_model_bic <- step(null.model,
43                           scope = list(lower = null.model, upper = full.model),
44                           direction = "forward",
45                           k = log(n)) # k = log(n) 表示使用 BIC
46 summary(forward_model_bic) # Y ~ log_TP + MI + BoH (Adjusted R-squared: 0.6354)
47 vif(forward_model_bic)   # log_TP:1.852179 MI:2.496284 BoH:3.337434
48 AIC(forward_model_bic)  # 946.6742
49 BIC(forward_model_bic)  # 957.3099
50 # -----
51 # 向後選取
52 n <- nrow(data)
53 # Backward Selection AIC
54 backward_model_aic <- step(full.model,
55                           scope = list(lower = null.model, upper = full.model),
56                           direction = "backward",
57                           k = 2)
58 summary(backward_model_aic) # Y ~ MI + BoH + Age25_44 + log_TP (Adjusted R-squared: 0.6518)
59 vif(backward_model_aic)   # log_TP:2.821697 MI:2.690206 BoH:3.446688 Age25_44 :2.031464
60 AIC(backward_model_aic)  # 944.7516
61 BIC(backward_model_aic)  # 957.5144
62 # Backward Selection AIC
63 backward_model_bic <- step(full.model,
64                           scope = list(lower = null.model, upper = full.model),
65                           direction = "backward",
66                           k = log(n))
67 summary(backward_model_bic) # Y ~ log_TP + MI + BoH (Adjusted R-squared: 0.6354)
68 vif(backward_model_bic)   # log_TP:1.852179 MI:2.496284 BoH:3.337434
69 AIC(backward_model_bic)  # 946.6742
70 BIC(backward_model_bic)  # 957.3099
71 # -----
72 # 逐步迴歸
73 n <- nrow(data)
74 # Stepwise AIC
75 stepwise_model_aic <- step(null.model,
76                           scope = list(lower = null.model, upper = full.model),
77                           direction = "both",
78                           k = 2)
79 summary(stepwise_model_aic) # Y ~ MI + BoH + Age25_44 + log_TP (Adjusted R-squared: 0.6518)
80 vif(stepwise_model_aic)   # log_TP:2.821697 MI:2.690206 BoH:3.446688 Age25_44 :2.031464
81 AIC(stepwise_model_aic)  # 944.7516
82 BIC(stepwise_model_aic)  # 957.5144
83 # Stepwise BIC
84 stepwise_model_bic <- step(full.model,
85                           scope = list(lower = null.model, upper = full.model),
86                           direction = "both",
87                           k = log(n))
88 summary(stepwise_model_bic) # Y ~ log_TP + MI + BoH (Adjusted R-squared: 0.6354)
89 vif(stepwise_model_bic)   # log_TP:1.852179 MI:2.496284 BoH:3.337434
90 AIC(stepwise_model_bic)  # 946.6742
91 BIC(stepwise_model_bic)  # 957.3099
92
93 # -----
94 # 全子集迴歸
95 # install.packages("leaps") # 如果還沒裝的話
96 library(leaps)
97 allpossible_model_result <- regsubsets(Y ~ MI + BoH + log_TP + PI + UR + log_PD + ND + Age25_44 + D,
98                                         data = df,
99                                         nvmax = 9,
100                                        nbest = 512,
101                                        really.big = TRUE)
102 # 獲取摘要結果
103 summary_all_models <- summary(allpossible_model_result)
104
105 # 計算每個模型的 AIC (因為 summary_all_models 不直接提供)
106 num_models_considered <- nrow(summary_all_models$which)
107 all_aics <- numeric(num_models_considered)
108 all_bics <- numeric(num_models_considered)

```

```

110 ~ for (i in 1:num_models_considered) {
111   vars_in_model_logical <- summary_all_models$which[i, ]
112   vars_in_model_names <- names(vars_in_model_logical)[vars_in_model_logical == TRUE]
113   vars_in_model_names <- vars_in_model_names[vars_in_model_names != "(Intercept)"]
114
115 ~ if (length(vars_in_model_names) == 0) {
116   current_formula <- as.formula("Y ~ 1")
117 ~ } else {
118   current_formula <- as.formula(paste("Y ~", paste(vars_in_model_names, collapse = " + ")))
119 ~ }
120
121   current_lm_model <- lm(current_formula, data = df)
122
123   all_aics[i] <- AIC(current_lm_model)
124   all_bics[i] <- BIC(current_lm_model)
125 ~}
126
127 # --- 2. 創建結果數據框 ---
128
129 # 初始化列表來儲存每一列的數據
130 model_summary_list <- list()
131
132 # 儲存變數組合字符串
133 model_summary_list$Variables <- apply(summary_all_models$which, 1, function(x) {
134   vars <- names(x)[x == TRUE & names(x) != "(Intercept)"]
135   if (length(vars) == 0) {
136     return("Intercept Only")
137   } else {
138     return(paste(vars, collapse = " + "))
139   }
140 })
141
142 # 儲存 Adjusted R-squared
143 model_summary_list$Adj_R_squared <- summary_all_models$adjr2
144
145 # 儲存 AIC
146 model_summary_list$AIC <- all_aics
147 # 儲存 BIC
148 model_summary_list$BIC <- all_bics
149
150 # 儲存 Cp
151 model_summary_list$Cp <- summary_all_models$cp
152
153 # 轉換為 data.frame
154 all_models_df <- as.data.frame(model_summary_list)
155
156 # --- 3. 顯示結果數據框的頭部和尾部 ---
157 cat("\n--- All Models Summary Dataframe (First 10 rows) ---\n")
158 print(head(all_models_df, 10))
159
160 cat("\n--- All Models Summary Dataframe (Last 10 rows) ---\n")
161 print(tail(all_models_df, 10))
162
163
164 # --- 4. 找出在數據框中各指標的最佳模型 ---
165
166
167 # 最佳 Adjusted R-squared
168 best_adjr2_row <- all_models_df[which.max(all_models_df$Adj_R_squared), ]
169 cat("\n--- Best Model by Adjusted R-squared ---\n")
170 print(best_adjr2_row)
171
172 # 最佳 AIC
173 best_aic_row <- all_models_df[which.min(all_models_df$AIC), ]
174 cat("\n--- Best Model by AIC ---\n")
175 print(best_aic_row)
176
177 # 最佳 BIC
178 best_bic_row <- all_models_df[which.min(all_models_df$BIC), ]
179 cat("\n--- Best Model by BIC ---\n")
180 print(best_bic_row)

```

```

182 # 最佳 Cp (找最接近 p+1 的模型)
183 # 我們需要模型中的變數數量 (p)
184 num_predictors_in_models <- rowSums(summary_all_models$which[, -1]) # -1 排除截距
185 cp_target_values <- num_predictors_in_models + 1
186 best_cp_row_index <- which.min(abs(all_models_df$Cp - cp_target_values))
187 best_cp_row <- all_models_df[best_cp_row_index, ]
188 cat("\n--- Best Model by Cp (closest to p+1) ---\n")
189 print(best_cp_row)
190
191
192 # --- 5. 按照特定指標排序整個數據框並查看 ---
193 cat("\n--- All Models Sorted by AIC (Top 10) ---\n")
194 all_models_df_sorted_by_aic <- all_models_df[order(all_models_df$AIC), ]
195 print(head(all_models_df_sorted_by_aic, 10))
196
197 cat("\n--- All Models Sorted by BIC (Top 10) ---\n")
198 all_models_df_sorted_by_bic <- all_models_df[order(all_models_df$BIC), ]
199 print(head(all_models_df_sorted_by_bic, 10))
200
201 cat("\n--- All Models Sorted by Adjusted R-squared (Top 10) ---\n")
202 all_models_df_sorted_by_adjr2 <- all_models_df[order(-all_models_df$Adj_R_squared), ] # 負號表示降序
203 print(head(all_models_df_sorted_by_adjr2, 10))
204 #
205 # 全子集畫圖
206 summary_result = summary(allpossible_model_result) # 呈現1個解釋變數的最佳模型到11個解釋變數的最佳模型
207 best_adjr2 <- which.max(summary_result$adjr2)
208 summary_result$which[best_adjr2, ] # TRUE 表示選進模型
209 best_cp <- which.min(summary_result$cp)
210 summary_result$which[best_cp, ]
211 best_bic <- which.min(summary_result$bic)
212 summary_result$which[best_bic, ]
213
214 plot(allpossible_model_result, scale = "r2")      # R2
215 plot(allpossible_model_result, scale = "adjr2")    # Adjusted R2
216 plot(allpossible_model_result, scale = "bic")       # BIC
217 plot(allpossible_model_result, scale = "Cp")        # Cp
218 #
219 # Best Model
220 model_best <- lm(Y ~ MI + BoH + log_TP, data = df)
221 summary(model_best)
222 # install.packages("car")
223 library(car)
224 vif(model_best)
225
226 # [Normality Test]-----
227 resid_best <- residuals(model_best)
228
229 # Shapiro-Wilk test
230 shapiro.test(resid_best)
231 # W = 0.97295, p-value = 0.187
232
233 # Jarque-Bera test
234 # install.packages("tseries")
235 library(tseries)
236 jarque.bera.test(resid_best)
237 # X-squared = 1.8822, df = 2, p-value = 0.3902
238
239 # Kolmogorov-Smirnov (Lilliefors) test
240 # 標準化殘差
241 resid_std <- scale(resid_best)
242 # 與標準常態分布比較
243 # install.packages("nortest")
244 library(nortest)
245 # Lilliefors修正的K-S檢定 (適用於迴歸殘差)
246 lillie.test(resid_best)
247 # D = 0.108, p-value = 0.06956
248
249 # QQ Plot
250 qqnorm(resid(model_best),
251         main = "QQ Plot of Residuals")
252 qqline(resid(model_best), col = 'black', lty = 2)
253 # PP Plot
254 # 排序後的殘差
255 resid_sorted <- sort(resid_best)

```

```

257 # 樣本累積機率
258 n <- length(resid_sorted)
259 empirical_p <- ppoints(n) # = (1:n - 0.5)/n
260
261 # 計算常態分布下的理論累積機率
262 theoretical_p <- pnorm(resid_sorted, mean = mean(resid_sorted), sd = sd(resid_sorted))
263
264 # 繪圖
265 plot(theoretical_p, empirical_p,
266       main = "P-P Plot of Residuals",
267       xlab = "Theoretical CDF (Normal)",
268       ylab = "Empirical CDF",
269       pch = 1)
270 abline(0, 1, col = "black", lty = 2) # 對角線
271 # Histogram
272 # 計算樣本數
273 n <- length(resid_best)
274
275 # 根據 Sturges 規則計算 breaks 數量 (log10 版本)
276 k <- ceiling(1 + 3.322 * log10(n)) # 向上取整
277
278 # 畫直方圖 + 常態曲線
279 hist(resid_best,
280       breaks = k,
281       ylim = c(0, 0.001),
282       main = paste("Histogram of Residuals (k =", k, "bins"),
283       xlab = "Residuals",
284       col = "lightgray",
285       border = "black",
286       freq = FALSE)
287
288 # 加上常態分布曲線
289 x_vals <- seq(min(resid_best), max(resid_best), length = 100)
290 curve(dnorm(x, mean = mean(resid_best), sd = sd(resid_best)),
291       col = "black", lwd = 2, add = TRUE)
292
293 # Box Plot
294 boxplot(resid_best,
295       main = "Boxplot of Residuals",
296       ylab = "Residuals",
297       col = "white",
298       border = "black")
299 # [Independence Test]-----
300 library(lmtest)
301 # Durbin-Watson
302 dwtest(model_best)
303 # DW = 2.3495, p-value = 0.9171
304 # ACF plot
305 acf(residuals(model_best))
306 # PACF plot
307 pacf(residuals(model_best))
308 # Residual Plot
309 plot(model_best$fitted.values, resid(model_best),
310       xlab = "Fitted Values", ylab = "Residuals",
311       main = "Residuals vs Fitted Values")
312 abline(h = 0, col = "red", lty = 2)
313
314 plot(model_best$fitted.values, resid(model_best),
315       xlab = "Fitted Values", ylab = "Residuals",
316       main = "Residuals vs Fitted Values")
317 abline(h = 0, col = "black", lty = 2)
318 plot(cooks.distance(model_best),
319       type = "h",
320       main = "Cook's Distance",
321       ylab = "Cook's D")
322 abline(h = 4/nrow(df), col = "red", lty = 2) # 通常界線為 4/n
323 shapiro.test(resid(model_best))
324 library(lmtest)
325 dwtest(model_best)
326 # [Homoscedasticity Test]-----
327 library(lmtest)
328 bptest(model_best, studentize = FALSE)
329 library(car)
330 spreadLevelPlot(model_best)

```

```

331 # Box-Cox 轉換-----
332 library(MASS)
333 boxcox(model_best, lambda = seq(-2, 2, 0.1), data = df)
334 bc <- boxcox(model_best, lambda = seq(-2, 2, 0.1), data = df)
335 lambda_opt <- bc$x[which.max(bc$y)]
336 print(lambda_opt)
337
338 df$Y_BoxCox = sqrt(df$Y)
339 model_BoxCox <- lm(Y_BoxCox ~ MI+BoH+log_TP, data = df)
340 summary(model_BoxCox) # Adjusted R-squared: 0.6149
341 # install.packages("car")
342 library(car)
343 vif(model_BoxCox)
344 # Box-Cox 轉換後常態性分析-----
345 resid_bc <- residuals(model_BoxCox)
346 # Shapiro-Wilk test
347 shapiro.test(resid_bc)
348 # W = 0.9857, p-value = 0.6871
349 # Freedman-Diaconis Rule 決定 bin width
350 iqr <- IQR(resid_bc)
351 n <- length(resid_bc)
352 bin_width <- 2 * iqr / (n^(1/3))
353 bins <- ceiling((max(resid_bc) - min(resid_bc)) / bin_width)
354 # 畫 histogram
355 hist(resid_bc,
356       breaks = bins,
357       main = "Histogram of Residuals (Box-Cox)",
358       xlab = "Residuals",
359       col = "white",
360       border = "black")
361 # Box Plot
362 boxplot(resid_bc,
363           main = "Boxplot of Residuals (Box-Cox)",
364           ylab = "Residuals",
365           col = "white",
366           border = "black")
367 # QQ plot
368 qqnorm(resid_bc,
369           main = "QQ Plot of Residuals (Box-Cox)")
370 qqline(resid_bc, col = "black", lty = 2)
371 # PP plot
372 # 實際與理論分布的累積機率
373 resid_sorted <- sort(resid_bc)
374 n <- length(resid_bc)
375 empirical_p <- ppoints(n)
376 theoretical_p <-pnorm(resid_sorted, mean = mean(resid_bc), sd = sd(resid_bc))
377 # PP Plot
378 resid_sorted <- sort(resid_bc)
379 n <- length(resid_bc)
380 empirical_p <- ppoints(n)
381 theoretical_p <-pnorm(resid_sorted, mean = mean(resid_bc), sd = sd(resid_bc))
382 plot(theoretical_p, empirical_p,
383       main = "PP Plot of Residuals (Box-Cox)",
384       xlab = "Theoretical CDF",
385       ylab = "Empirical CDF",
386       pch = 1, col = "black")
387 abline(0, 1, col = "black", lty = 2)
388 # Box-Cox 轉換後獨立性分析-----
389 # Durbin-Watson
390 dwtest(model_BoxCox)
391 # DW = 2.2659, p-value = 0.8528
392 # ACF plot
393 acf(residuals(model_BoxCox))
394 # PACF plot
395 pacf(residuals(model_BoxCox))
396 plot(model_BoxCox$fitted.values, resid(model_BoxCox),
397       xlab = "Fitted Values",
398       ylab = "Residuals",
399       main = "Residuals vs Fitted")
400 abline(h = 0, col = "red", lty = 2)
401 # Residual Plot
402 plot(model_BoxCox$fitted.values, resid(model_BoxCox),
403       xlab = "Fitted Values", ylab = "Residuals",
404       main = "Residuals vs Fitted Values")

```

```

405 abline(h = 0, col = "black", lty = 2)
406 # Box-Cox 轉換後齊一性分析-----
407 library(lmtest)
408 bptest(model_BoxCox, studentize = FALSE)
409 library(car)
410 spreadLevelPlot(model_BoxCox)
411
412 # -----
413 # 1. 使用 Cook's Distance
414 cooksD <- cooksdistance(model_BoxCox)
415 plot(cooksD, main="Cook's Distance for Influence Points")
416 abline(h = 4/58, col="black") # 標記常用的閾值
417
418 # 找出 Cook's D 值超過閾值的點
419 influential_cooks <- which(cooksD > 4/nrow(df))
420 print("Potentially influential points based on Cook's Distance:")
421 print(influential_cooks)
422
423 # 計算 DFFITS
424 dffits_values <- dffits(model_BoxCox)
425
426 # 繪製 DFFITS 圖
427 plot(dffits_values,
428       main = "DFFITS for Influence Detection",
429       ylab = "DFFITS",
430       xlab = "Index",
431       pch = 1, col = "black")
432 abline(h = 2 * sqrt(4/58), col = "black", lty = 2)
433 abline(h = -2 * sqrt(4/58), col = "black", lty = 2)
434
435 # 找出超過標準的點
436 threshold <- 2 * sqrt(4 / 62)
437 influential_dffits <- which(abs(dffits_values) > threshold)
438
439 # 顯示影響觀測值的索引
440 print("Influential points based on DFFITS:")
441 print(influential_dffits) # 24 31 46 60
442
443 # 敏感性分析
444 influential_indices <- c(24, 31, 46, 60)
445 df_no_influential <- df[-influential_indices, ]
446 df_no_influential$Y_BoxCox = sqrt(df_no_influential$Y)
447 model_no_influential <- lm(Y_BoxCox ~ MI + BoH + log_TP, data = df_no_influential)
448 summary(model_no_influential) # Adjusted R-squared: 0.6158
449
450
451 # -----
452 # 建立結果表格
453 n <- nrow(df)
454 results <- data.frame(
455   obs = 1:n,
456   coef_BoH = numeric(n),
457   coef_MI = numeric(n),
458   coef_logTP = numeric(n)
459 )
460
461 # 逐筆移除並建模
462 for (i in 1:n) {
463   temp_data <- df[-i, ]
464   model_i <- lm(Y_BoxCox ~ MI + BoH + log_TP, data = temp_data)
465
466   results$coef_BoH[i] <- coef(model_i)[["BoH"]]
467   results$coef_MI[i] <- coef(model_i)[["MI"]]
468   results$coef_logTP[i] <- coef(model_i)[["log_TP"]]
469 }
470
471 # 設定三個圖並排
472 par(mfrow = c(1, 3))
473
474 # 畫出每個係數的變化圖
475 plot(results$obs, results$coef_BoH, type = "b",
476       main = "BoH", xlab = "Observation Removed", ylab = "Coefficient")
477 abline(h = coef(model_BoxCox)[["BoH"]], col = "black", lty = 2)

```

```

479 plot(results$obs, results$coef_MI, type = "b",
480       main = "MI", xlab = "Observation Removed", ylab = "Coefficient")
481 abline(h = coef(model_BoxCox)[["MI"]], col = "black", lty = 2)
482
483 plot(results$obs, results$coef_logTP, type = "b",
484       main = "log_TP", xlab = "Observation Removed", ylab = "Coefficient")
485 abline(h = coef(model_BoxCox)[["log_TP"]], col = "black", lty = 2)
486
487 # -----
488 # 逐筆移除觀測值並重新建模，記錄調整後 R2
489 for (i in 1:n) {
490   temp_data <- df[-i, ]
491   model_i <- lm(Y_BoxCox ~ MI + BoH + log_TP, data = temp_data)
492   results$R2[i] <- summary(model_i)$adj.r.squared
493 }
494
495 # Plot: Adjusted R2 change under the removal of each observation
496 plot(results$obs, results$R2, type = "b",
497       main = "Adjusted R2 Change After Removing Each Observation",
498       xlab = "Removed Observation Index",
499       ylab = "Adjusted R2",
500       pch = 1, col = "black")
501
502 # 加上調整後 R2 數值
503 text(results$obs, results$R2,
504       labels = round(results$R2, 3),
505       pos = 3, cex = 0.7, col = "blue")
506
507 # Draw the horizontal baseline
508 abline(h = summary(model_BoxCox)$adj.r.squared, col = "black", lty = 2)
509
510 # Add label on the y-axis
511 text(x = par("usr")[1], # far left of x-axis
512       y = summary(model_BoxCox)$adj.r.squared,
513       labels = round(summary(model_BoxCox)$adj.r.squared, 4),
514       pos = 2, # to the right of x=par("usr")[1]
515       col = "black", cex = 0.8, xpd = TRUE)
516
517 # 初始化結果表格
518 n <- nrow(df)
519 results_p <- data.frame(
520   obs = 1:n,
521   p_BoH = numeric(n),
522   p_MI = numeric(n),
523   p_logTP = numeric(n)
524 )
525
526
527 # 逐一移除觀測值並提取 p-value
528 for (i in 1:n) {
529   temp_data <- df[-i, ]
530   model_i <- lm(Y_BoxCox ~ MI + BoH + log_TP, data = temp_data)
531   p_vals <- summary(model_i)$coefficients[, "Pr(>|t|)"]
532
533   results_p$p_BoH[i] <- p_vals["BoH"]
534   results_p$p_MI[i] <- p_vals["MI"]
535   results_p$p_logTP[i] <- p_vals["log_TP"]
536 }
537
538 # 畫三張圖橫向排列
539 par(mfrow = c(1, 3))
540
541 plot(results_p$obs, results_p$p_BoH, type = "b",
542       main = "BoH p-value", xlab = "Observation Removed", ylab = "p-value", ylim = c(0, 0.06))
543 abline(h = summary(model_BoxCox)$coefficients["BoH", "Pr(>|t|)"], col = "black", lty = 2)
544 abline(h = 0.01, col = "black", lty = 3)
545 abline(h = 0.05, col = "black", lty = 3)
546
547 plot(results_p$obs, results_p$p_MI, type = "b",
548       main = "MI p-value", xlab = "Observation Removed", ylab = "p-value", ylim = c(0, 0.00001))
549 abline(h = summary(model_BoxCox)$coefficients["MI", "Pr(>|t|)"], col = "black", lty = 2)
550
551 plot(results_p$obs, results_p$p_logTP, type = "b",
552       main = "log_TP p-value", xlab = "Observation Removed", ylab = "p-value", ylim = c(0, 0.00001))
553 abline(h = summary(model_BoxCox)$coefficients["log_TP", "Pr(>|t|)"], col = "black", lty = 2)

```