作業二:應用多種模型迴歸資料點

撰寫者:數金三甲 **B1245025** 洪子貽　　日期:**2025**年**10**月**17**日

一、實驗目的

此實驗旨在評估七種迴歸模型，在一個具非線性與週期性特徵的數據集 (exercise_dataset.csv) 上的擬合性能。所有模型的超參數選擇與性能評估，均基於 5折交叉驗證 (5-fold Cross-Validation) 框架，並以均方根誤差 (RMSE) 作為衡量指標。

核心目標:

1. 模型評估:量化評估多種迴歸模型在真實數據上的性能
2. 模型複雜度:探討模型複雜度與「偏誤-變異數權衡」之間的關係
3. 過擬合與正則化:比較 L1 與 L2 正則化在抑制過度擬合上的效果與差異
4. 最佳模型選擇:根據實驗結果，為數據集找出最佳迴歸模型

二、實驗原理與設定

(一)實驗框架

- 數據集:*exercise_dataset.csv*，含200個 (x, t) 資料點

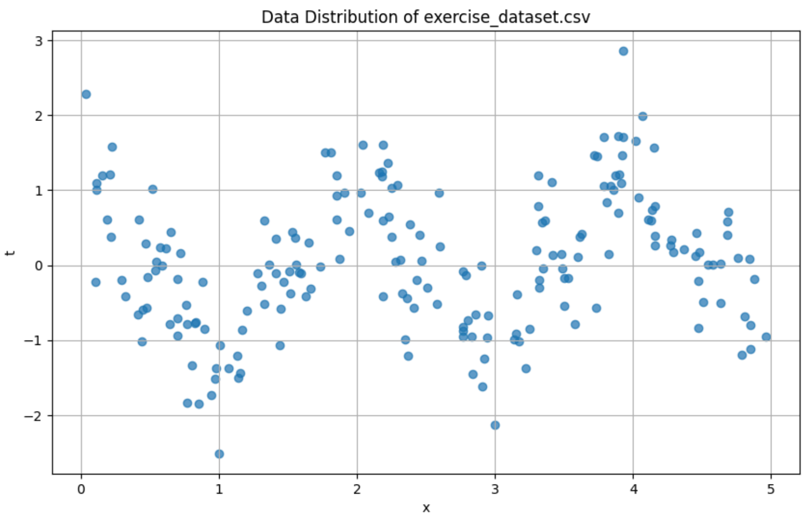

圖***1***:資料集散佈圖
數據具非線性、週期性與高噪聲，擬合難度高。

- 驗證策略:5折交叉驗證

- 將數據集分為五份，輪流使用其中四份進行訓練，一份進行驗證，重複五次後取平均RMSE ⇒ 可有效降低評估結果的隨機性，使其更能反映模型的真實泛化能力。

- 評估指標：均方根誤差 (RMSE)
  - $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(t_i - y(x_i))^2}$，反映模型預測與真實值間的平均差距。

(二)模型原理

- 正弦函數模型：基於數據週期性的參數化模型，假設數據可由正弦函數 $t = a \cdot \sin(b \cdot x + c) + d$ 擬合。模型透過數值優化找出最佳參數 [a, b, c, d] 以最小化誤差。其優點是可解釋性強，但缺點是函數形式固定，若真實數據模式與假設不符，表現將嚴重受限。
- 基底函數模型：將輸入 x 透過非線性基底函數 $\phi_j(x)$ 映射至高維空間，再進行線性迴歸，通用形式為 $y(x, w) = \sum_{j=0}^{M-1} w_j\phi_j(x)$。
  - 多項式基底 ($\phi_j(\mathbf{x}) = x^j$)：為全局基底，權重 (w_j) 調整會影響整個函數曲線，超參數 (M) 決定多項式的最高次數，直接控制模型複雜度。
  - 高斯基底 ($\phi_j(\mathbf{x}) = exp(-\frac{(x-\mu_j)^2}{2s^2})$)：為局部基底，形如鐘形曲線 (以 s 為寬度)，其影響範圍局限於中心點 μ_j 附近，適合捕捉局部特徵。
  - **S型(羅吉斯)基底** ($\phi_j(\mathbf{x}) = \sigma(\frac{x-\mu_j}{s})$, σ 是Sigmoid函數)：同為局部基底，形如平滑階梯，可用於組合擬合複雜模式。
- 正則化模型：在標準最小二乘法損失函數 $L_{MSE} = \sum_{i=1}^{N}(t_i - y(x_i, w))^2$的基礎上，增加一個由超參數 α 控制的懲罰項 $P(w)$，來約束模型權重、控制複雜度並防止過度擬合。其總損失為 $L = L_{MSE} + \alpha P(w)$。
  - 嶺迴歸 **(Ridge)**：採 L2 正則化，懲罰項為權重平方與 $P(w) = \Sigma w_j^2$，傾向於將權重縮小 (不會變為零)，使擬合曲線更平滑。
  - **Lasso**迴歸：採 L1 正則化，懲罰項為權重絕對值與$P(w) = \Sigma|w_j|$，能將不重要的特徵權重精確壓縮到零，具自動特徵選擇的能力。

○ 彈性網路 **(Elastic Net)**:結合 L1 與 L2 正則化,試圖融合兩者優點,既能處理特徵間相關性,又能進行特徵選擇。

三、實驗結果與分析

(一)非正則化模型

1. 正弦函數模型 (RMSE = 0.837)

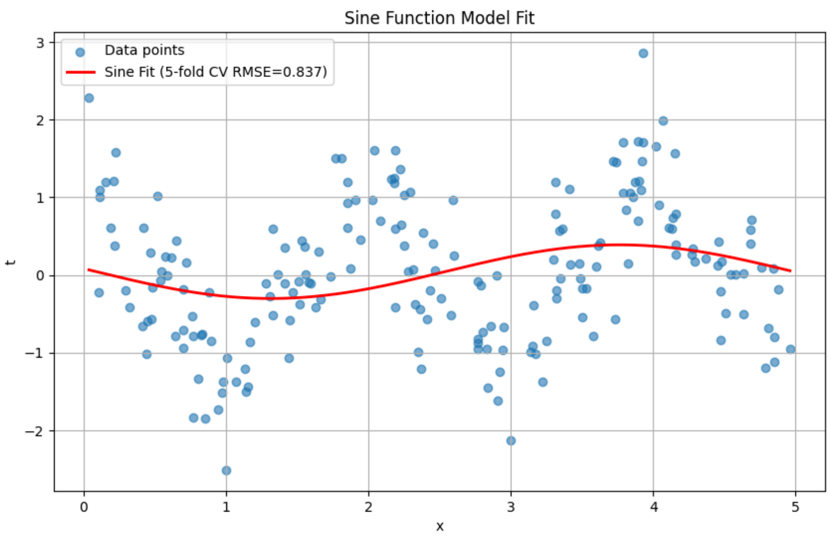

圖**2**:正弦函數模型擬合結果

分析:此模型表現最差。如圖 2 所示,擬合曲線過於平滑,完全無法捕捉數據的局部峰谷和動態變化。此為典型的高偏誤 (High Bias),即模型自身假設過於簡單,無法描述數據的複雜性,導致無論如何訓練都無法取得良好性能。

2. 多項式基底函數模型 (最佳RMSE = 0.627 at M=11)

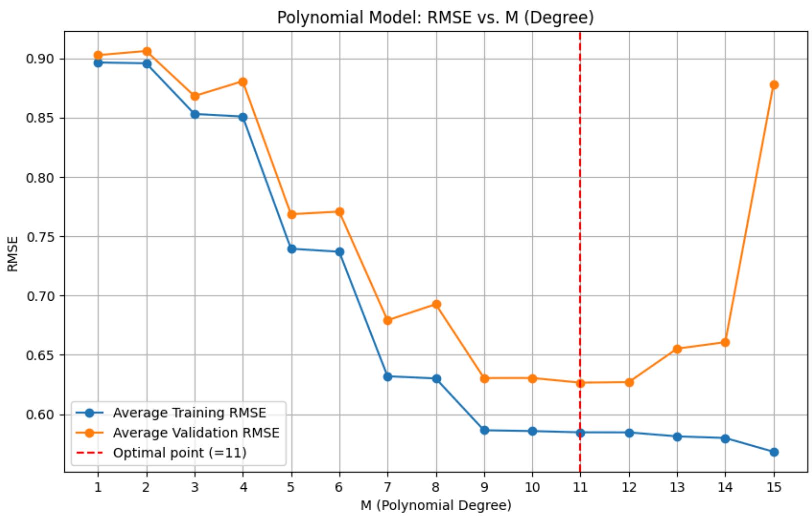

圖**3**:多項式模型交叉驗證曲線

分析:圖 3 完美展示偏誤-變異數權衡的過程。

- **M < 11 (高偏誤區域)**：訓練（藍線）和驗證誤差（橘線）都很高且同步下降，表模型複雜度不足。
- **M = 11 (最佳點)**：驗證誤差達最低點，為模型泛化能力的峰值。
- **M > 11 (高變異數/過度擬合區域)**：驗證誤差開始反彈並急劇上升，而訓練誤差仍在下降。兩條曲線的分離為過度擬合的明確信號。模型開始過度擬合訓練數據中的噪聲，導致其在未見過的驗證數據上表現變差。

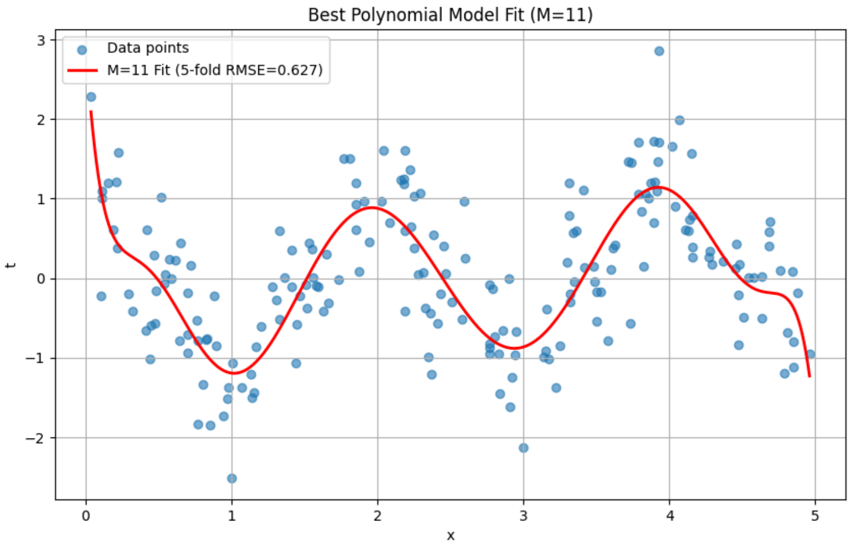

圖*4*：最佳多項式模型*(M=11)*擬合結果

分析：圖 4 中 M=11 的擬合曲線雖捕捉了主要趨勢，但其在數據兩端 (x 接近0 & 5) 出現劇烈且不自然的彎曲。此為高階多項式作為全局模型的固有問題 (龍格現象)，為遷就某些數據點，可能導致在其他區域產生劇烈振盪。

3. 高斯基底函數模型 (最佳RMSE = 0.621 at M=7)

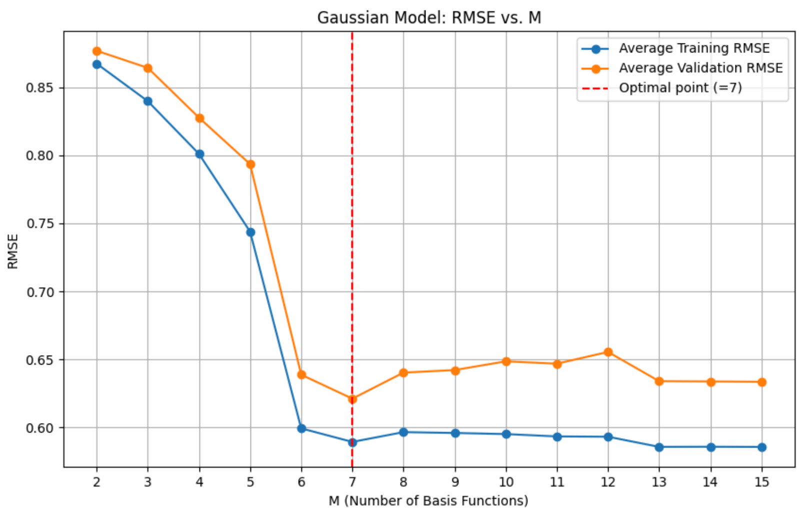

圖*5*：高斯基底模型交叉驗證曲線

分析：圖 5 顯示，其在 M=7 時就迅速達到最佳性能點，之後驗證誤差保持在一個相對穩定的低水平，顯示模型對超參數 M 的敏感度較低，魯棒性更好。

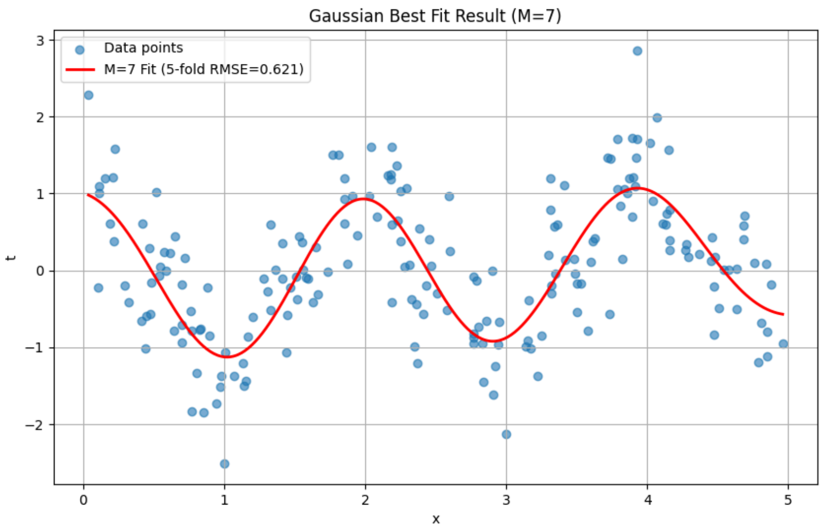

圖*6*：最佳高斯基底模型*(M=7)*擬合結果

分析：其擬合曲線（圖 6）極其平滑且精準地貼合數據的趨勢，完全沒有多項式模型在邊界的振盪問題。這得益於其局部性，模型可像搭積木一樣，用7個局部的「小山丘」組合出複雜的全局曲線，靈活性和穩定性都遠超多項式模型。

4. S型(羅吉斯)基底函數模型 (最佳RMSE = 0.630 at M=15)

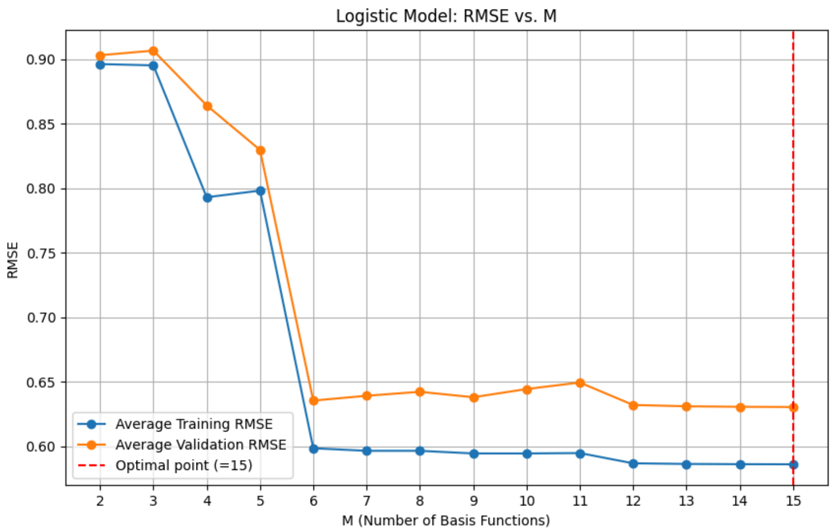

圖*7*：***S***型基底模型交叉驗證曲線

分析：圖 7 顯示，其驗證誤差在 M=6 時大幅下降後便進入一個平坦期，直到 M=15 才達最低點，表明增加更多的基底函數對性能的提升微乎其微，但也並未導致嚴重過度擬合。

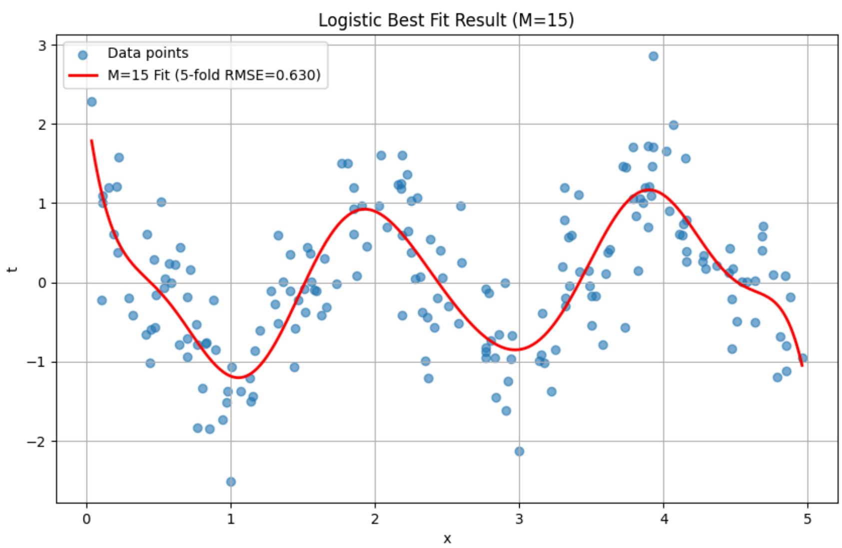

圖*8*:*S*型基底模型*(M=15)*擬合結果

分析:作為另一種局部模型,S型基底的擬合效果同樣出色,曲線平滑。但其達到最佳效果所需的模型複雜度(M=15)遠高於高斯模型(M=7),顯示出較低的模型效率。

(二)正則化模型 (基於 M=14 多項式)

實驗目的:給定一個已知會過度擬合的模型(M=14多項式),比較不同正則化策略「抑制過度擬合」的效果。

1. 嶺迴歸 (Ridge) - 成功的正則化 (最佳RMSE = 0.627)

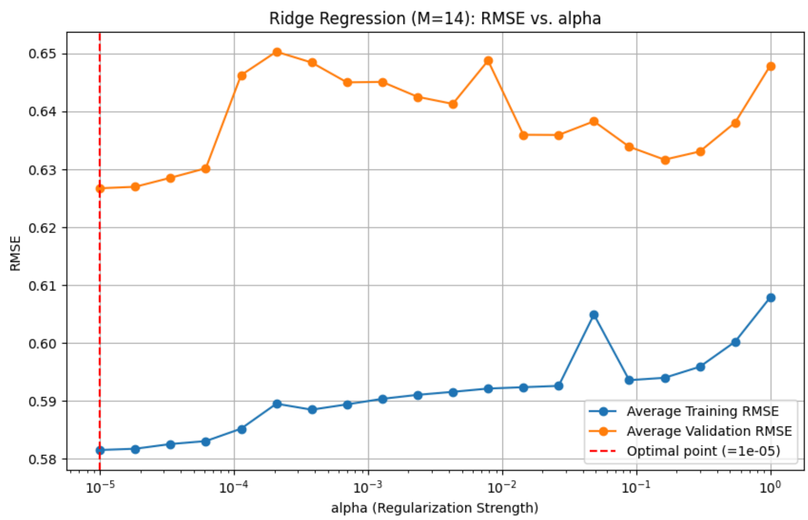

圖*9*:嶺迴歸交叉驗證曲線

分析:圖 9 展示完美的正則化效果。當正則化強度 α 極小(圖左側)時,驗證誤差很高,模型處於過擬合狀態。隨著 α 增加,驗證誤差迅速下降,在 $\alpha = 10^{-5}$ 達最低點。若 α 繼續增大,則會因懲罰過度而導致欠擬合,驗證誤差重新上升。嶺迴

歸成功找到最佳平衡點，將過擬合模型的性能恢復至與最佳多項式模型完全相同的水平。

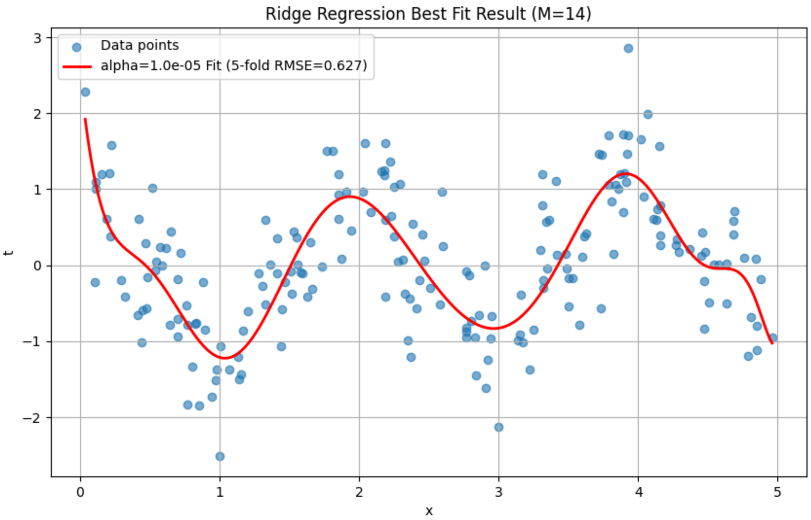

圖*10*：最佳嶺迴歸模型擬合結果

分析：圖 10 中平滑的擬合曲線顯示，一個原本會過度擬合的14次多項式模型，在L2正則化下表現得如一個更簡單的最佳模型。嶺迴歸透過收縮權重，有效降低模型的有效複雜度，成功過濾掉過擬合所導致的劇烈振盪，同時保留捕捉數據主要趨勢的能力。

2. Lasso迴歸＆彈性網路 - 失敗的正則化 (最佳RMSE = 0.777)

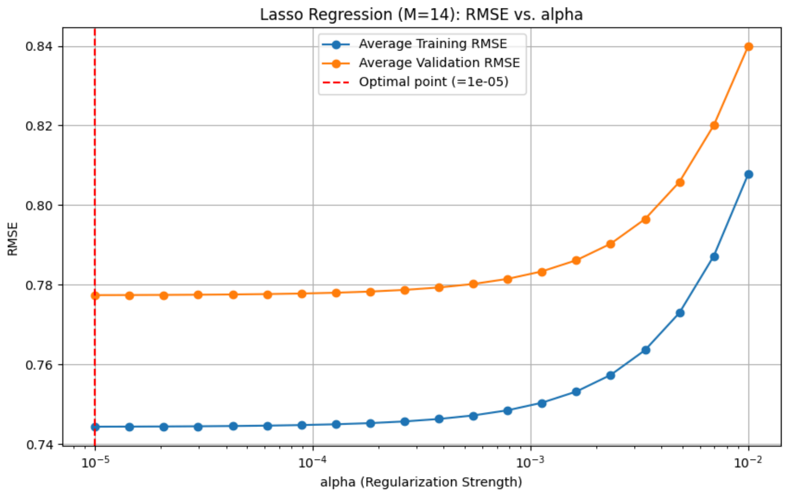

圖*11*：*Lasso*迴歸交叉驗證曲線

分析：圖 11 與嶺迴歸形成鮮明對比，揭示其根本性失敗。驗證誤差最低點（0.777）不僅遠高於嶺迴歸，且出現在正則化最弱處。隨懲罰強度 α 增加，誤差幾乎單調遞增，顯示L1從一開始便損害模型。訓練與驗證誤差間的巨大鴻溝，進一步證

實模型喪失泛化能力。高位且平坦的誤差曲線，則是正則化策略與問題本質不匹配的強烈信號。

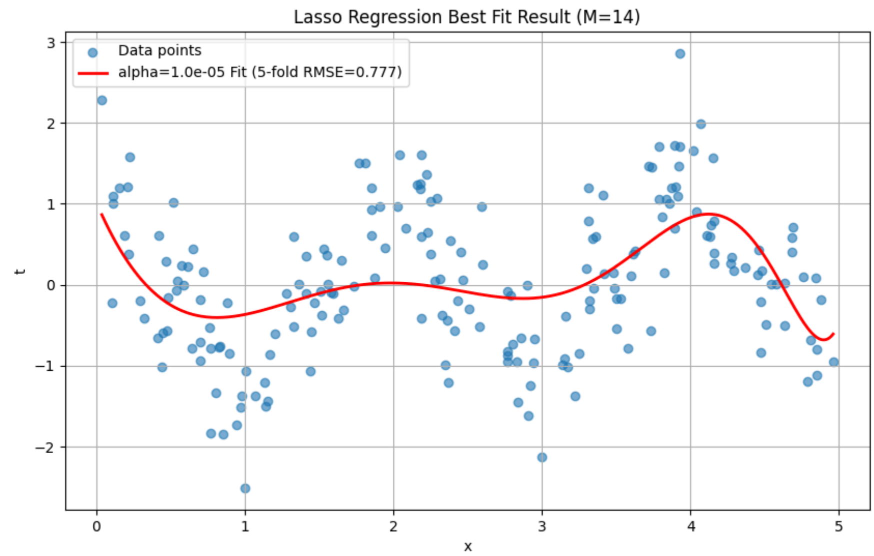

圖*12*：最佳***Lasso***迴歸模型擬合結果

分析：圖 12 展示Lasso的失敗，擬合曲線呈嚴重欠擬合，過於平滑和簡化，完全丟失數據的週期性結構。根本原因為 L1 正則化的稀疏化效應，錯誤地將構建複雜曲線所需的許多高階項權重強制歸零。Lasso的「特徵選擇」在此摧毀模型的表達能力，將錯誤的歸納偏誤應用於問題。

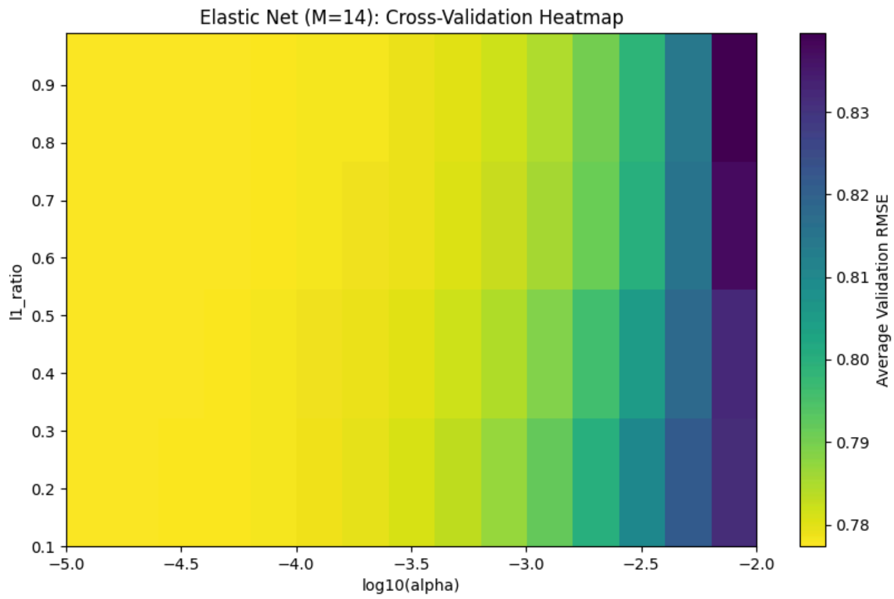

圖*13*：彈性網路交叉驗證熱圖

分析：圖 13 顯示彈性網路在所有超參數組合下的性能都普遍很差 (RMSE約 0.78-0.82)，無任何區域接近最佳值。最低誤差（最亮的黃色）集中在正則化強度

α 最小的左側區域。說明只要L1懲罰存在，無論其與L2如何混合，都無法改善模型性能。最終，最佳點出現在 l1_ratio 極高 (0.99) 的位置，再次證實模型的行為被表現不佳的Lasso所主導。

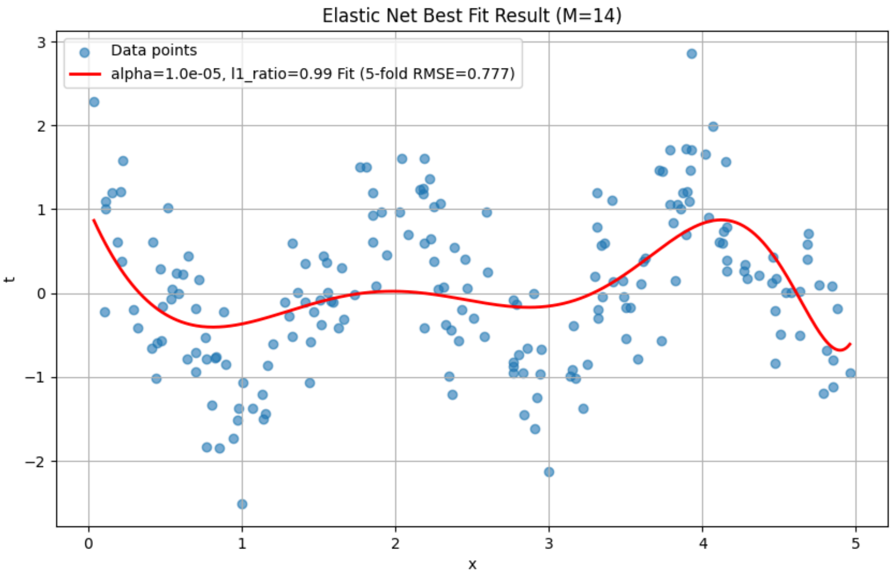

圖*14*:最佳彈性網路擬合結果

分析：因最佳超參數 (l1_ratio=0.99) 使模型行為極度接近純Lasso，圖 14 與Lasso的結果（圖12）幾乎完全相同，同樣呈現出嚴重欠擬合狀態。此圖有力地證明，對於此數據擬合問題，L1正則化是一個不恰當的策略。

四、綜合比較與分析

排名	模型名稱	最佳超參數	最小平均驗證 RMSE	核心分析
1	高斯基底函數	M = 7	0.621	性能冠軍。局部基底，效率與效果俱佳。
2	多項式基底函數	M = 11	0.627	全局基底的次優選擇，但需更高複雜度且在邊界不穩定。
2	嶺迴歸 (多項式 M=14)	$\alpha = 10^{-5}$	0.627	成功的正則化。L2懲罰有效平滑過擬合模型，使其性能恢復至最佳。
4	S型(羅吉斯)基底函數	M = 15	0.63	表現良好，但模型效率低於高斯模型。
5	Lasso迴歸 (多項式 M=14)	$\alpha = 10^{-5}$	0.777	失敗的正則化。L1懲罰過於激進，導致欠擬合。
5	彈性網路 (多項式 M=14)	$\alpha = 10^{-5}$, l1_ratio = 0.99	0.777	本質上等同於Lasso，表現同樣不佳。
7	正弦函數模型	N/A	0.837	模型假設過強，偏誤高，完全無法擬合數據。

（一）局部 vs. 全局基底函數

此實驗最顯著的結論之一為局部基底模型(高斯、S型)的整體表現優於全局基底模型(多項式)。多項式模型的一個權重調整會牽一髮而動全身，影響整個曲線。而局部模型則可獨立調整特定區域的擬合，而不過多干擾其他部分，因此在擬合具有複雜局部特徵的數據時更具優勢。

(二)過度擬合與正則化

- 觀察過擬合：多項式模型次數 M > 11 時，訓練與驗證誤差曲線的顯著分離，為模型過度擬合的可視化證據，為後續正則化提供明確目標。
- **L2 的調和之道 (權重縮減而非剔除)**：嶺迴歸的成功，證明當模型中多數特徵都對結果有潛在貢獻時，L2 的「柔性」懲罰是控制複雜度的穩健策略。透過縮減所有權重而非輕易歸零，有效平滑模型，在抑制過擬合的同時，完整保留模型的表達潛力。
- **L1 的雙面刃 (稀疏性假設的侷限)**：Lasso 的失敗，則警示 L1 並非適用於所有問題的通用解法。其「剛性」懲罰背後是一個強烈的稀疏性假設——即認定只有少數特徵是重要的。若此假設與數據真實模式不符(如本次實驗中，所有高次項都有其作用)，L1 便會錯誤地將重要特徵的權重強制歸零，從而摧毀模型的表達能力，導致災難性的欠擬合。

(三)模型效率

評估模型時，除了 RMSE 所代表的準確性，其效率(即模型複雜度)同樣至關重要。高斯基底模型僅用 M=7 便取得最佳擬合，而多項式及S型模型則分別需要 M=11 與 M=15 的更高複雜度才達到次級性能。此對比凸顯出，高斯基底是描述本數據集最有效的基礎，能以最少的參數實現最優的擬合效果。這種以簡馭繁的特性，是理想模型的標誌，它不僅意味著更高的運算效率，更代表模型具有更強的泛化能力與更低的過度擬合風險。

五、結論

此實驗通過對七種迴歸模型的評估分析，成功為給定數據集找到最優的擬合方案。主要結論如下：

- 最佳模型：高斯基底函數模型以其高效的局部建模能力，在 M=7 時達到 0.621的最低 RMSE，為本次實驗的最佳選擇。

- 模型選擇啟示：局部基底函數模型（高斯、S型）整體表現優於全局基底模型（多項式）。這顯示在面對未知數據時，應優先考慮更靈活、更能適應局部變化的模型。
- 正則化實踐：嶺迴歸 (L2) 是處理過擬合問題時一個非常穩健的工具。而 Lasso (L1) 的應用則需謹慎，須深刻理解其稀疏性假設是否與問題的內在特性相符，否則可能適得其反。