DOI: 10.6277/TER.202606_54(2).0001

因果推論下的機器學習

黃柏鈞*

傳統機器學習方法著重於提升預測準確度, 但近年研究顯示, 這些方法亦可應用於高維度資料中, 協助選取控制變數, 進而提高因果推論的可信度。本文首先說明預測與因果推論在研究目標與方法設計上的根本差異, 接著介紹兩種整合方法: Post-Double Selection LASSO 與 Double/Debiased Machine Learning。這兩種方法皆透過機器學習, 在高維度資料下輔助因果推論。本文亦簡要介紹兩個實證上的應用, 分別探討生育補助對生育率的影響, 以及制度對經濟成長的影響, 以說明機器學習在因果推論中的實際運用。

關鍵詞: 因果推論, 機器學習, Post-Double Selection LASSO, Double/Debiased Machine Learning
JEL 分類代號: C14, C21, C52, C55

1 前言

隨著大型調查資料與行政資料的可及性提升, 以及電腦運算能力的進步, 機器學習 (machine learning) 在社會科學與商業實務中的應用愈來愈廣。以勞動經濟學為例, Cengiz et al. (2022) 利用機器學習模型預測哪些勞工的時薪可能低於即將實施的最低工資水準, 藉此評估最低工資政策的影響範圍; 在需求預測的應用中, Bajari et al. (2015) 採用機器學習建構更具彈性的需求模型, 以提升對消費者購買行為的預測準確度; 在醫療與公共安全領域, Kleinberg et al. (2015) 則強調機器學習在資源分配與識別高風險

*國立政治大學經濟學系副教授, Email: huangpo5@nccu.edu.tw。作者感謝兩位匿名審查人及 2025 年台灣經濟學會年會課程參與者的寶貴建議。

臺灣大學學術期刊資料庫

對象上的應用, 例如預測病患再入院的風險或被告再次犯案的可能性。這些研究的共同特徵是關注「預測」, 即利用一組變數 (如年齡、教育、工作經驗) 來預測某個結果 (如是否低薪、是否再犯), 以輔助決策與資源配置。

相較之下, 在應用個體經濟學 (applied microeconomics) 中, 研究者更常關注的問題並非「某個結果會是什麼」, 而是「某項政策或事件是否導致結果的改變」。舉例而言, 我們可能關心提高最低工資對就業有何影響 (如 Cengiz et al. (2019))、勞動市場是否存在性別或種族的歧視 (如 Goldin and Rouse (2000); Bertrand and Mullainathan (2004)), 或生育對勞動供給的影響 (如 Angrist and Evans (1998)) 等。這類問題屬於因果推論 (causal inference) 的範疇, 重點在於如何在給定其他條件不變下, 準確地評估變數 X 對結果 Y 的影響, 而非預測 Y 的大小。2021 年諾貝爾經濟學獎頒發給 Joshua D. Angrist、David Card 與 Guido W. Imbens, 即表彰他們對因果推論方法的開創性貢獻。

為了進一步說明預測與因果推論在目標參數與方法設計上的差異, 我們可以從簡單線性迴歸模型切入思考:

$$Y_i = \alpha + \beta X_i + \epsilon_i。$$

在預測類型的問題中, 研究者的目標是根據 X_i 預測 Y_i, 因此關注的是條件期望值 $\mathbb{E}[Y|X]$ 的估計。在這個框架下, β 通常不代表兩變數間的因果關係, 其功能是協助預測 Y 的一個參數。相對地, 在因果推論中, 研究者關心的是 β 是否可以被解釋為 X 對 Y 的影響。為了讓 β 的估計具備因果意涵, 經濟學家發展出一系列實證設計 (empirical designs), 包括迴歸控制法、工具變數 (Instrumental Variables)、斷點迴歸 (Regression Discontinuity)、差異中的差異 (Difference-in-Differences) 等, 用以處理 X 的內生性 (Endogeneity) 問題, 提升推論的可信度。[1]

事實上, 內生性有多種來源, 包括遺漏變數偏誤 (Omitted Variable Bias)、反向因果 (Reverse Causality) 或衡量誤差 (Measurement Error) 等。針對不同的內生性來源, 研究者須採用對應的認定假設 (Identifying Assumptions)

[1] 見 Angrist and Pischke (2015)、Angrist and Pischke (2009) 或 Cunningham (2021) 關於上述這些因果推論方法的介紹。

臺灣大學學術期刊資料庫

以進行因果推論: 例如, 迴歸控制法仰賴條件獨立假設 (Conditional Independence, 或 Selection on Observables); 差異中的差異法 (Difference-in-Differences) 要求平行趨勢假設 (Common Trend); 而工具變數法 (Instrumental Variables) 則利用外生變異來排除不可觀測因素的干擾。本文聚焦於條件獨立架構下的因果推論。在此框架中, 若能妥善控制變數 X, 處理項即可視為外生。然而, 實務上面臨的挑戰在於控制變數可能具有高維度特性 (High-dimensional Controls), 這使得變數選擇至關重要 — 武斷地選擇某些變數可能導致遺漏變數偏誤, 若納入過多變數, 則可能有變數數量超過樣本數的的困境, 導致傳統最小平方法無唯一解。

機器學習方法主要著眼於預測, 不宜直接用於估計因果關係。然而, 正如 Belloni, Chernozhukov, and Hansen (2014a) 所指出的, 在任何具有眾多控制變數的實證研究中, 研究者應設法避免過度擬合 (Overfitting) 的問題。這個問題同樣存在於有高維度控制變數或工具變數的因果推論模型中。為了善用機器學習在處理高維度資料上的優勢, 同時滿足因果推論的需求, 近年來已有多項研究致力於整合這兩類方法。本文簡介兩種以機器學習協助因果推論的方法: Post Double Selection LASSO (PDSLASSO) (Belloni, Chernozhukov, and Hansen, 2014a) 與 Double/Debiased Machine Learning (DDML) (Chernozhukov et al., 2018)。PDSLASSO 假設資料中的變數關係可以用線性方式描述, 並透過 LASSO 方法, 分別挑選出與結果變數及處理變數關聯性較強的一小部分控制變數, 再將這兩組控制變數得聯集做為控制變數, 以降低遺漏重要變數所造成的偏誤。相較之下, DDML 不需要假設變數之間的關係是線性的, 也不需要事先假定只有少數幾個控制變數是重要的。它使用更彈性的機器學習工具, 來估計被解釋變數與處理變數的條件期望函數, 並透過正交化 (orthogonalization) 技巧, 減少遺漏變數偏誤。考量篇幅限制, 本文不涉及機器學習於工具變數選擇上的應用, 而將重點放在控制變數的處理與其應用潛力。

本文對監督式機器學習的介紹主要依據 James et al. (2013), 該書對機器學習方法提供深入淺出的說明, 是初學者入門機器學習的經典教科書。關於 PDSLASSO 與 DDML 的介紹, 則主要參考 Belloni, Chernozhukov, and Hansen (2014a) 與 Chernozhukov et al. (2018), 兩篇文章除闡述方法

臺灣大學學術期刊資料庫

論外, 亦涵蓋相關的實證應用。Mullainathan and Spiess (2017) 與 Athey and Imbens (2019) 廣泛地討論機器學習在經濟學中的應用, 包括其在因果推論上的應用。Chernozhukov et al. (2024) 一書則系統性地回顧因果推論的核心方法, 並說明機器學習技術如何輔助各種因果推論方法, 為有志於深入了解因果推論下的機器學習的研究者提供系統性的介紹。本文在上述基礎上, 嘗試以直觀方式概略地介紹 PDSLASSO 與 DDML 的概念與應用, 並說明其在高維控制變數下的優勢與限制。

以下章節安排如下: 第 2 節回顧監督式機器學習的基本概念; 第 3 節介紹 PDSLASSO 與 DDML 的理論架構與估計步驟; 第 4 節以兩個應用為例說明實作細節; 最後一節總結。

2 監督式機器學習簡介

在本章中, 我將簡介監督式學習 (Supervised Learning) 的方法。所謂監督式學習, 是指利用具有已知的輸入 (X) 與對應的輸出 (y) 資料, 來估計一個模型, 進而預測新輸入資料所對應的輸出結果, 常見的監督式學習方法包括 LASSO、Ridge Regression 與隨機森林 (Random Forest) 等。相對地, 非監督式機器學習 (Unsupervised Learning) 則是僅有輸入資料 X, 但沒有明確定義的輸出結果 y。在非監督式機器學習中, 研究者的目的是從資料中發掘其潛在結構、群體分類或變異來源, 而非預測特定變數, 常見的非監督式學習方法包括群聚分析 (clustering) 與主成分分析 (Principal Component Analysis) 等。[2]

舉例而言, 若我們擁有勞工特徵資料 (如年齡、教育程度、產業別、工作年資) 與對應的薪資資訊, 便可運用監督式學習方法訓練模型, 以預測勞工的工資水準。相對地, 若資料中僅包含勞工特徵而無薪資資訊, 則可採用非監督式學習方法進行群聚分析, 識別勞工在勞動市場中的潛在類型 (例如: 高技術專業職、基層勞工、初入職場青年等), 進一步應用於政策設計或薪資結構分析。

[2]見 James et al. (2013) 與 Hastie, Tibshirani, and Friedman (2009) 關於上述方法的介紹, 這些方法在經濟學中的應用, 可參見 Mullainathan and Spiess (2017) 與 Athey and Imbens (2019)。

臺灣大學學術期刊資料庫

本文以下的介紹將聚焦於監督式機器學習方法; 關於非監督式方法的介紹, 可參考 James et al. (2013)。我將先介紹衡量模型預測準確度的方式, 接著說明兩大類常見的監督式學習方法: 收縮方法 (shrinkage methods) 與樹基方法 (tree-based methods)。

2.1 模型準確度的衡量

一般而言, 不同的機器學習方法會產生不同的預測結果, 而究竟哪一種方法能提供較高的預測準確性, 是機器學習領域中的核心問題之一。因此, 在介紹各種機器學習方法之前, 應先說明如何衡量模型的預測能力。

首先, 我們考慮以下的迴歸模型:

$$y = f(X) + \varepsilon。$$

其中, y 為連續變數, $X = (X_1, X_2, \cdots, X_p)$, f 為未知的 X 函數, ε 為誤差項, 滿足 $E(\varepsilon) = 0$ 且與 X 獨立。在評估模型表現時, 均方誤差 (Mean Squared Error, MSE) 是最常用的衡量指標。它衡量模型預測值與實際值之間的平均平方差, 定義如下:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2。 \tag{1}$$

其中, y_i 為第 i 筆資料的實際數值, $\hat{y}_i$ 為模型對該筆資料的預測值, n 為樣本數。這 n 筆資料稱為訓練資料 (training data), 用以估計模型 $f(x)$, 因此式 (1) 稱為訓練誤差 (Training MSE)。然而, 需要強調的是, 我們關心的是模型應用於測試資料或未觀察資料 (test data) 時的預測表現, 而非在訓練資料上的擬合程度。也就是說, 在我們所考慮的模型中, 我們希望所建立的模型具有較小的測試誤差 (Test MSE), 而非在訓練資料上有較小的訓練誤差。

訓練資料指的是用來估計模型 $f(x)$ 的資料, 而測試資料則用於評估模型的預測能力。舉例來說, 假設一家銀行蒐集了數千名貸款申請人的資訊, 例如年齡、收入、信用分數、職業類別與過往還款紀錄等, 並且知道他們之中哪些人曾發生貸款違約。這些資料可以用來訓練一個模型, 預測新客

臺灣大學學術期刊資料庫

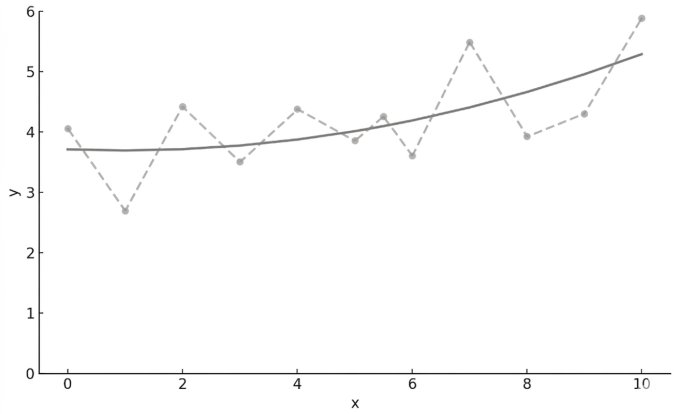


圖 1: 選擇簡單或是複雜的模型?

說明: 上圖展示過度擬合的情形。圖中複雜模型能完美擬合 12 筆訓練資料的 (x, y) 關係, 但本質上等同為每個 x 值設置一個虛擬變數, 導致對未觀察資料無法預測。相較之下, 較為簡單的模型雖在訓練資料上雖有誤差, 卻在未觀察資料上具備預測能力。

戶違約的風險。然而, 真正重要的是這個模型是否能準確預測新客戶的違約可能性, 而不是對已知結果的樣本的解釋能力。因為我們已經知道訓練樣本中哪些人違約, 模型在這些資料上的表現不是我們最關心的重點; 能否對新客戶產生準確的風險評估, 才是模型是否成功的關鍵。

在圖 1 中, 共有 12 組 (x, y) 的觀察值。圖中的複雜模型能完美連結所有訓練資料, 對可觀察到的 y 值具備完美的預測能力, 然而這樣的模型實際上並無預測能力。其本質等同於為每個 x 值設置一個虛擬變數, 因此當面對一個未曾觀察的新 x 值時, 便不具預測力。相對地, 另一個較為簡化的模型雖然在訓練資料上存在誤差, 卻在未觀測資料上具有某種程度的預測能力。也就是說, 在這個例子中, 複雜的模型有過度擬合的問題: 一個模型在訓練資料中的解釋力高, 但在未觀察資料的預測能力差。

圖 2 取自 James et al. (2013) 的圖 2.9, 模型複雜度與訓練誤差與測試誤差之間的關係呈現顯著不同的趨勢。隨著模型複雜度上升, 訓練誤差持續下降, 表示模型對訓練資料的擬合能力愈來愈強。然而, 測試誤差則呈現先下降後上升的 U 形曲線, 顯示模型的複雜度超過某個程度之後, 對於未觀察資料的預測能力反而下降。當模型的訓練誤差小, 但測試誤差大, 則

臺灣大學學術期刊資料庫

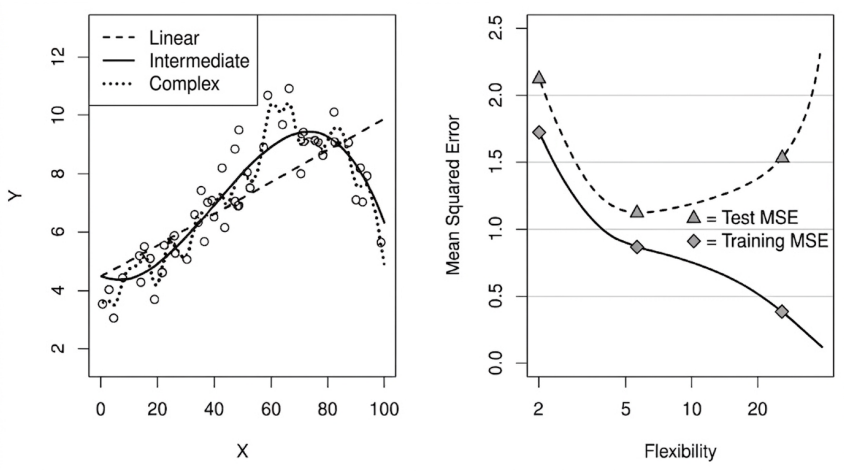


圖 2: 模型複雜度與訓練/預測誤差之間的關係

說明: 上圖源自 James et al. (2013) 圖 2.9, 說明模型複雜度與訓練/預測誤差之間的關係。左圖展示不同模型對資料的擬合情形; 右圖顯示模型複雜度與訓練/測試誤差之間的關係。

代表模型過度擬合, 一個良好的模型應避免過度擬合的現象。

為什麼會出現過度擬合的問題呢? 我們可以從模型對一個新資料點 (x_0, y_0) 的預測誤差來理解。其均方誤差可拆解為以下三項組成:

$$E\left[\left(y_0 - \hat{f}\,(x_0)\right)^2\right] = \left[\text{Bias}\left(\hat{f}(x_0)\right)\right]^2 + \text{Var}\left[\hat{f}(x_0)\right] + \text{Var}(\varepsilon)。 \quad (2)$$

其中, $\text{Bias}(\hat{f}(x_0))$ 為預測上的偏誤, 其定義為:

$$\text{Bias}\left(\hat{f}(x_0)\right) = E\left[\hat{f}\,(x_0)\right] - f\,(x_0)\,。$$

偏誤衡量的是模型預測的平均值與真實函數值之間的差距, 通常源自於模型對真實資料生成過程的簡化。例如, 若 $f(x)$ 是高度非線性的函數, 而我們使用線性模型進行預測, 則無法捕捉變數間的非線性關係, 導致預測上的偏誤。另一方面, $\text{Var}(\hat{f}(x_0))$ 為預測上的變異, 其定義為:

$$\text{Var}\left(\hat{f}\,(x_0)\right) = E\left[\left(\hat{f}\,(x_0) - E\left[\hat{f}\,(x_0)\right]\right)^2\right]。$$

臺灣大學學術期刊資料庫

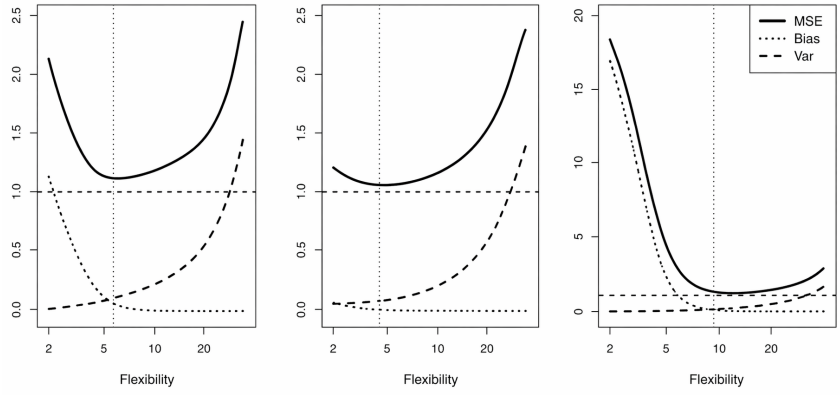


圖 3: 偏誤與變異數的權衡

說明: 上圖源自 James et al. (2013) 圖 2.12, 展示模型複雜度與偏誤平方、變異數與均方誤差之間的關係, 說明在偏誤 — 變異數權衡下, 預測誤差與模型複雜度間呈現 U 型關係。三圖分別對應不同資料產生過程下的情境。

變異反映的則是模型對於不同訓練樣本的敏感程度。若變異高, 表示在不同訓練資料下, 模型對同一點的預測結果差異大。例如, 像決策樹這類高彈性的模型, 雖然能更貼近資料的分布、降低偏誤, 但也更容易學習到資料中的雜訊, 導致預測結果在不同樣本下有較大的變化, 增加預測上的變異。

由式 (2) 可見, 均方誤差主要來自偏誤與變異這兩個來源, 兩者之間存在典型的偏誤 — 變異權衡關係 (Bias-Variance Trade-off): 當模型越複雜或靈活, 偏誤通常較低, 但變異會提高; 相反地, 較簡化的模型雖然偏誤較高, 但變異較低。因此, 我們的目標並非追求偏誤或變異中的任一極小值, 而是權衡兩者, 找出能夠最小化測試誤差的模型。從圖 3 (見 James et al. (2013) 圖 2.12) 中清楚看出: 當模型複雜度提升時, 誤差下降, 但變異上升, 測試誤差因此呈現 U 形曲線。

雖然測試誤差才是衡量模型預測能力的關鍵指標, 但在實務上, 我們往往無法取得測試資料。此時, 交叉驗證 (Cross Validation) 是一種廣泛使用且有效的方法, 用來估計模型的測試誤差。其中, k 折交叉驗證 (k-fold cross validation) 是常見的做法。其步驟如下: 首先, 隨機將資料分成 k 個大小相近的子集。接著, 輪流將第 i 個子集作為驗證資料 (validation data), 其餘 $k-1$ 個子集作為訓練資料, 使用訓練資料估計模型, 並計算模型在

臺灣大學學術期刊資料庫

第 i 折驗證資料上的均方誤差 (MSE_i)。將所有 k 次的驗證誤差取平均, 即可得到交叉驗證對測試誤差的估計值:

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\text{MSE}_i。$$

我們通常會比較不同模型下的 $CV_{(k)}$ 值, 並選擇使其最小者作為預測模型。附帶一提, 當 k 等於樣本數時, 每次只留下一筆資料作為驗證資料, 其餘作為訓練資料, 此即為「逐一驗證」(Leave-One-Out Cross Validation, LOOCV), 是交叉驗證的極端情形。交叉驗證能在沒有額外測試資料的情況下, 評估模型的預測能力, 是機器學習不可或缺的工具。

2.2 收縮方法

為了有效降低模型在未觀察資料上的預測誤差, 我們除了透過交叉驗證等方法來評估模型表現外, 有時也需要在模型建構階段採取適當的策略, 以避免過度擬合的發生。特別是在高維度資料的情境中, 當變數數量 p 超過樣本數 n, 傳統最小平方法 (Ordinary Least Squares, OLS) 並不適用。在此背景下, 機器學習與統計學發展出一類稱為「收縮方法 (shrinkage methods)」的技術, 作為傳統線性迴歸模型的延伸。這類方法的核心概念是: 在估計模型參數時, 於損失函數中加入懲罰項 (penalty term), 約束模型的複雜度, 進而避免過度擬合的問題, 提升模型在新資料上的預測能力。常見的收縮方法包括 LASSO 與 Ridge Regression, 兩者皆屬於線性模型。以下分別介紹這兩種方法。

2.2.1 LASSO

LASSO 是一種結合參數收縮與變數選擇功能的線性迴歸方法, 其目標是最小化以下損失函數:

$$\hat{\beta}^{\text{LASSO}} = \arg\min_{\beta}\sum_{i=1}^{n}\left(y_i - x_i^{\top}\beta\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|。$$

其中, $\lambda \geq 0$ 為使用者指定的懲罰強度參數。當 λ 越大時, 模型對迴歸係數絕對值的加總的懲罰越強, 部分係數可能收縮為 0, 從而剔除重要性較低的

臺灣大學學術期刊資料庫

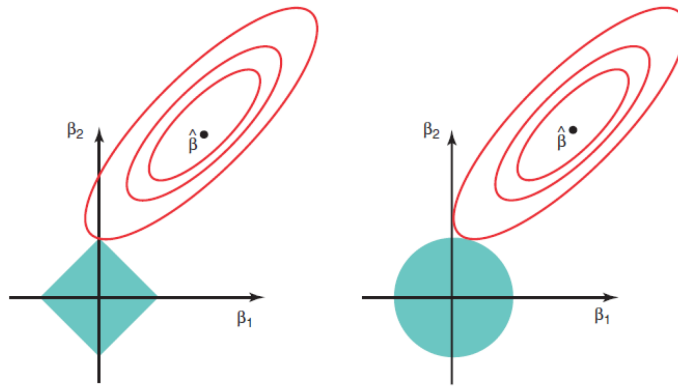


圖 4: LASSO v.s. Ridge Regression

說明: 上圖源自 James et al. (2013) 圖 6.7, 展示 Lasso (左) 與 Ridge (右) 迴歸的幾何直覺。紅色等高線為殘差平方和等高線, 其中心為無限制條件下之 OLS 解 $\hat{\beta}$。Lasso 的限制條件對應至菱形區域, 易使某些係數為 0, 具有變數選擇功能; Ridge 之限制條件為圓形區域, 會將所有係數縮小但不為 0。

變數, 達到變數篩選的效果, 讓預測模型更為精簡且易於解釋。簡而言之, 若模型中僅有兩個解釋變數 (x_1 與 x_2), 從圖 4 可見, LASSO 的懲罰項會在 β_1 與 β_2 之二維空間中形成一個菱形的限制區域, 與殘差平方和的等高線相交時容易在座標軸上產生角解, 而這個角解對應的 β_1 為 0, 也就是 LASSO 將僅將 x_2 選入模型。

與 OLS 相比, LASSO 的目標函數在最小化殘差平方和的同時, 亦懲罰迴歸係數的絕對值, 使得在 $p > n$ 的情況下仍有唯一解。因此, 實務上, LASSO 特別適用於變數數量遠多於樣本數 ($p \gg n$) 且僅有少數變數具預測力的情況 (sparsity)。例如, 研究者可能想要預測一名勞工在的離職風險, 並擁有包含教育程度、工作年資、產業別、公司規模、過往職業歷史、居住地區與家庭組成等數十甚至上百個變數。在這種情況下, LASSO 可以自動篩選出與離職風險高度相關的關鍵因素, 排除次要變數, 進而建構出更有預測能力且易於解釋的預測模型。

臺灣大學學術期刊資料庫

2.2.2 Ridge Regression

Ridge Regression 是另一種常見的收縮方法, 透過在損失函數中加入 $\sum_j \beta_j^2$ 懲罰項, 在高維度資料下, 估計迴歸係數。其目標函數為:

$$\hat{\beta}^{\text{Ridge}} = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - x_i^\top \beta\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2。$$

與 LASSO 不同, Ridge Regression 一般不會將迴歸係數收縮為 0, 而是將係數向 0 收縮 (見圖 4)。因此, 它無法進行變數選擇, 但相對於最小平方法, 在高維度資料下仍有唯一解, 且能在對偏誤影響有限的情況下, 大幅度減少預測上的變異, 提高預測能力。這使 Ridge Regression 特別適合用於眾多變數皆具預測力, 且變數間高度相關的情況。

無論是 LASSO 或 Ridge Regression, 實作上皆需選定懲罰強度參數 λ, 以控制模型對複雜度的容忍程度。λ 與模型的測試誤差間亦存在 U 型關係, 較大的 λ 會加強係數收縮, 降低變異但可能提高偏誤; 較小的 λ 則允許更大的模型彈性, 但可能導致過度擬合。實務上, λ 的選擇通常透過 k-fold 交叉驗證完成, 以找出能最小化測試誤差的懲罰強度 (圖 5)。

此外, 在應用這些收縮方法前, 通常須對所有輸入變數 X 進行標準化 (standardization), 亦即將各變數轉換為平均數為 0、標準差為一的形式。這是因為懲罰項 (例如 Ridge 中的 $\sum_j \beta_j^2$) 的存在, 可能使估計結果因為變數的單位不同而產生差異。舉例而言, 若某變數為「房屋面積」, 原本以「平方公尺」為單位, 若改以「坪」為單位 (1 坪約等於 3.3 平方公尺), 則該變數的數值會縮小約三分之一。對 OLS 而言, 這僅會使該變數的係數相應放大三倍, 預測結果維持不變, 這種性質稱為尺度等變性 (scale equivariance)。然而 LASSO 與 Ridge Regrssion 不具備此性質, 因為其懲罰項直接作用於係數的大小, 使模型結果依賴於變數的尺度, 可能因某一變數單位的改變, 間接影響其他變數的係數估計。為避免此問題, 在執行 LASSO 或 Ridge 前, 通常會先對所有變數進行標準化, 確保懲罰項對所有變數公平一致, 避免模型結果受到單位選擇的影響。

總而言之, LASSO 與 Ridge Regression 各具優勢, 適用於不同的模型建構情境。LASSO 透過變數選擇機制產生稀疏解 (Sparse Solution), 有助

臺灣大學學術期刊資料庫

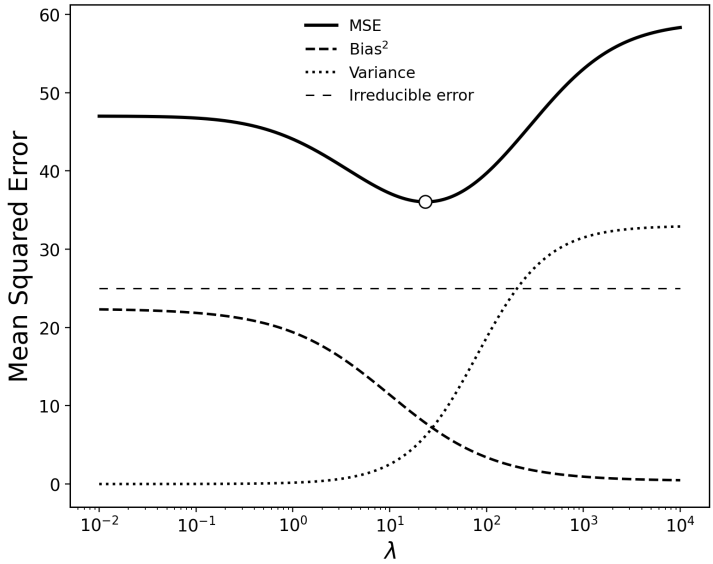


圖 5: λ的選擇—偏誤與變異數的權衡

說明: 上圖源自 James et al. (2013) 圖 6.5, 展示不同 λ 下的測試誤差 (粉紅色曲線)、變異數 (綠色曲線) 與誤差平方之 (黑色曲線)。

於提升模型的可解釋性, 特別適用於僅有少數變數與被解釋變數具關聯的情況。相對地, Ridge Regression 則在所有變數皆略具預測力、且係數大小分布較為平均的高維度資料中有更好的預測能力。然而, 在真實應用中, 資料背後的生成機制往往不可得, 因此難以事先判斷哪一種方法更合適。此時, 研究者可以利用交叉驗證挑選最適合的收縮方法。

目前介紹的 LASSO 與 Ridge Regression 雖具有其理論基礎與實務應用價值, 但皆屬於線性模型。然而, 真實世界的資料生成機制往往比線性模型複雜許多, 可能涉及非線性關係與變數間的交互作用。在此情況下, 樹基方法提供了一種更具彈性的選擇。這類方法不需事先指定函數形式, 而是藉由分割特徵空間來捕捉複雜的結構關係, 尤其適用於變數間存在高度非線性或交互效果的情境。以下將說明樹基方法的基本原理與程序。

臺灣大學學術期刊資料庫

2.3 樹基方法

2.3.1 決策樹 (Decision Trees)

決策樹是一種非參數預測方法, 其核心概念是將特徵空間 (predictor space) — 即所有解釋變數 $X_1, X_2, \cdots, X_p$ 所能取值的組合 — 遞迴地劃分為若干個不重疊的區域, 並在每個區域內對應相同的預測。若研究對象為類別型變數, 則為分類樹 (Classification Tree), 每個區域的預測值為該區域內出現最多的類別 (即眾數); 當研究對象為連續型變數時, 所建構的模型稱為迴歸樹 (Regression Tree), 每個區域的預測值為該區域內觀察值的平均; 以下說明建構分類樹的主要步驟, 並補充迴歸樹的差異:

1. 將 $(X_1, X_2, \cdots, X_p)$ 的所有可能組合, 分割為 J 個互不重疊的區域 $R_1, R_2, \cdots, R_J$。這些區域是根據遞迴式二元分割 (recursive binary splitting) 的方式逐步劃分而形成。
2. 對於落入某個區域 R_j 的每一筆觀察值, 其預測類別為訓練資料中所有位於該區域內最常見的類別: $\hat{c}_{R_j}$, 若為迴歸樹, 其預測為該區域的樣本平均數, $\hat{y}_{R_j}$。因此, 每個區域對應一個固定的類別或平均數, 只要確定新觀察值所屬區域, 即可給出對應的預測。

關於第一步的遞迴式二元分割, 其流程如下, 從所有解釋變數 X_j 中, 尋找一個變數 X_j 及其對應的切點 s, 將資料分為兩個子區域:

$$R_1(j,s) = \left\{X | X_j < s\right\}, \quad R_2(j,s) = \left\{X | X_j \geq s\right\}。$$

對於每組 (j, s), 計算分割後區域 m 的吉尼指數 (Gini index), 其定義如下:

$$G(m) = \sum_k \hat{p}_{mk}\left(1 - \hat{p}_{mk}\right)。$$

其中, $\hat{p}_{mk}$ 為該區域中屬於第 k 類的比例。吉尼指數愈小, 表示該區域的分類愈一致, 即大多數觀察值屬於同一類別。[3]。接者, 將兩個區域的吉尼指數依樣本比例加權平均:

$$G = q_1 G(1) + q_2 G(2)。$$

[3]另一常見的分類誤差指標為熵 (Entropy), 其定義為 $-\sum_k \hat{p}_{mk} \log \hat{p}_{mk}$。與吉尼指數類似, 熵也衡量區域內的不純度, 且在數值上兩者通常相近。

臺灣大學學術期刊資料庫

選擇能最小化吉尼指數 G 的變數 X_j 與切點 s 作為最佳的第一層分割方式。類似地, 若為迴歸樹, 則選擇一 X_j 與切點 s, 以最小化殘差平方和:

$$\sum_{i:x_i \in R_1(j,s)} \left(y_i - \hat{y}_{R_1}\right)^2 + \sum_{i:x_i \in R_2(j,s)} \left(y_i - \hat{y}_{R_2}\right)^2 。$$

其中, $\hat{y}_{R_1}$ 與 $\hat{y}_{R_2}$ 分別為 R_1 與 R_2 區域的樣本平均數。

完成第一層切分後, 對目前存在的區域進行下一層分割, 且每次僅針對一個區域進行一次切割, 目標是讓吉尼指數下降的程度最大, 並依據尚未分割的區域持續遞迴, 直到符合預設的停止條件 (如每個區域的最小樣本數 minimum training observations per leaf)、樹的最大深度 (maximum tree depth) 等。最終, 特徵空間將被劃分為 J 個不重疊的矩形區域 $R_1, R_2, \cdots, R_J$, 每個區域對應一個固定的預測類別。對於任一新觀察值 x, 其預測為:

$$\hat{f}(x) = \sum_{j=1}^{J} \hat{c}_{R_j} \cdot \mathbf{1}\left\{x \in R_j\right\} 。$$

其中, $\hat{c}_{R_j}$ 為訓練資料中落在區域 R_j 內的觀察值中最常見的類別。若為迴歸樹, 其預測即為 R_j 區域的樣本平均數。這種由上而下、逐步尋找局部最佳切分點的策略, 使得決策樹能有效捕捉資料中的非線性結構與變數間的交互關係, 其邏輯簡單, 且易於理解預測的結果。

舉例來說, 假設我們想要建立一個分類樹模型, 根據員工的特性來預測其薪資水準是否低於下一年度的最低工資。我們的資料中包含兩個解釋變數: 年資 (Years of Experience) 與教育程度 (Years of Education)。如圖 6 所示, 在第一步分割中, 演算法可能會發現以「教育年數是否大於 13 年」作為分割準則, 這個分割方式能最大幅度地降低吉尼指數, 於是將資料分割為兩個區域。接著, 對於教育程度較高的員工, 模型再進一步以年資是否小於 10 年進行分割。最後, 針對教育程度未滿 13 年的樣本區分為年資大於等於 5 年或小於五年的樣本, 每一步的切割方式皆以最大幅度降低吉尼指數為準則。根據以上的步驟, 圖 7 的決策樹將整體資料空間切分為四個

臺灣大學學術期刊資料庫

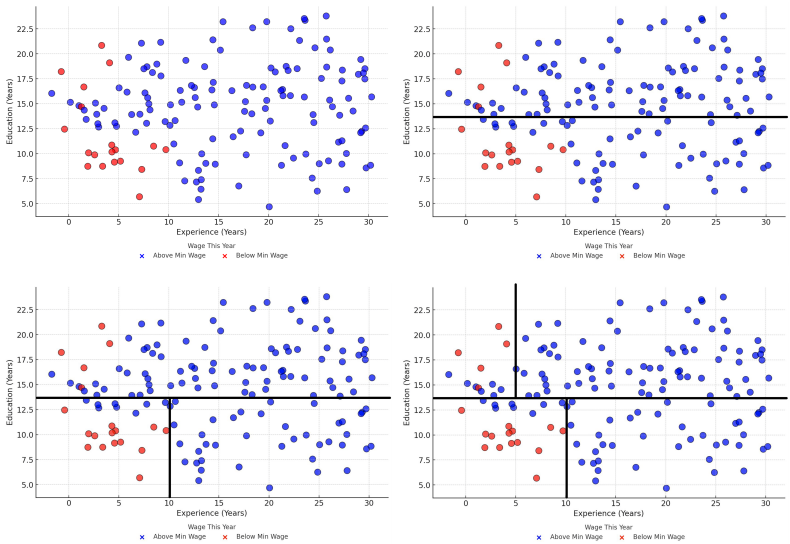


圖 6: 樹基方法

說明: 圖 6(a) 至 (d) 顯示決策樹分類過程中模型如何在特徵空間中進行逐步切割。圖中以教育年數與工作經驗為特徵, 將觀察值依是否低於最低工資 (紅點) 或高於最低工資 (藍點) 進行分類。每一步的垂直或水平線代表一次分割, 將特徵空間依據某個變數的切點區分成更小的子區域。

不重疊的區域:

$$R_1 = \{\text{Education} > 13\text{且 Experience} < 10\};$$
$$R_2 = \{\text{Education} > 13\text{且 Experience} \geq 10\};$$
$$R_3 = \{\text{Education} \leq 13\text{且 Experience} \geq 5\};$$
$$R_4 = \{\text{Education} \leq 13\text{且 Experience} < 5\}。$$

這樣一來, 對於任一新的觀察值 x, 我們可以依據其教育年數與年資, 判定其所屬的區域 R_j, 並根據該區域內訓練資料中最多數的類別作為預測結果。比方說, 如果一位新觀察值落入區域 R_4, 而該區域中有 80%的觀察值屬於「低薪」類別, 則分類樹會預測這位員工可能落入低薪行列。

在實務應用中, 單棵決策樹的表現很大程度取決於其樹的深度。若樹的深度過淺, 模型往往無法有效捕捉資料中的非線性模式與交互作用, 導

臺灣大學學術期刊資料庫

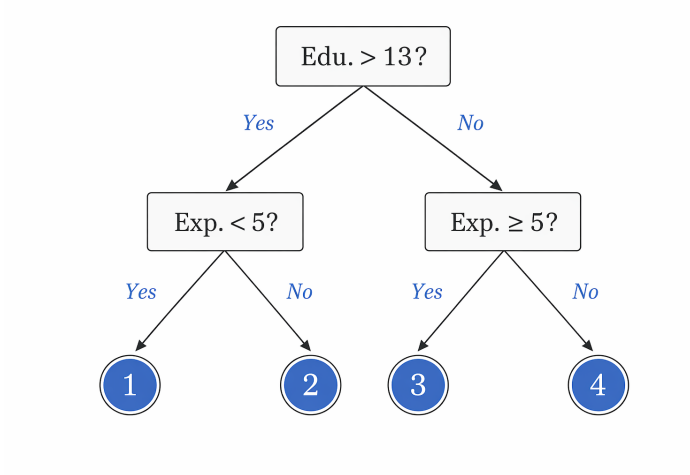


圖 7: 樹基方法

說明: 圖 7 為最終形成的決策樹結構, 其中每一個內部節點對應一個變數切點, 每個終端節點代表一個分類預測結果。呈現決策樹如何以遞迴方式在特徵空間中劃分區域並建立分類規則。

致低度擬合 (Under-fitting) 與較高的偏誤。相反地深度過深則易陷入過度擬合, 將資料中的雜訊誤認為規律, 導致模型對訓練樣本過於敏感而產生高變異。這種偏誤與變異之間的權衡, 使得單棵決策樹在面對複雜資料時, 往往難以同時兼顧預測的精準度與穩定性。為了在不增加偏誤的前提下有效降低變異, 學者進一步提出了隨機森林的概念, 試圖透過匯集多棵決策樹的力量, 來達到更為穩健的預測結果。

2.3.2 隨機森林

隨機森林立基於決策樹的方法, 是一種透過整合多棵差異化的決策樹來提升預測準確性的機器學習方法, 其核心概念結合了裝袋算法 (bagging) 和去相關化 (decorrelation) 兩項技術。

首先, 裝袋算法透過隨機抽樣與預測平均來降低預測上的不確定性。雖然單一決策樹產出的預測對資料變動極為敏感, 但若能對訓練資料多次抽樣, 建立多棵決策樹, 並將預測結果進行平均 (針對連續變數) 或多數決 (針對類別變數), 即可有效降低變異。其步驟如下:

臺灣大學學術期刊資料庫

1. Bootstrap: 從原始的 n 筆資料中進行有放回的抽樣 B 次, 每次抽出 n 筆觀察值, 產生 B 組樣本。
2. 利用這 B 組樣本, 分別建立 B 棵決策樹。
3. 整合這 B 棵決策樹的預測結果。若為迴歸問題, 則對 B 棵樹的預測值取平均; 若為分類問題, 則以多數決方式選出預測次數最多的類別作為最終預測結果。

這 B 棵決策樹可能都相當複雜, 導致較大的預測變異, 但由於平均值的變異會隨樣本數上升而下降, 裝袋算法能有效地降低變異, 提高預測準確度。

雖然裝袋算法透過整合多棵樹的預測, 提高預測準確度, 但是當部分變數特別重要時, 這些變數可能反覆出現在多棵樹的上層分割節點, 導致不同樹之間高度相似, 裝袋算法降低變異的效果因而受限。因此, 隨機森林採用隨機特徵選擇來達成去相關化 — 在每次分割時, 演算法不再從全部 p 個變數中尋找最佳分割點, 而是隨機選取其中 m 個變數 (通常 $m \approx \sqrt{p}$) 作為候選。這樣的作法避免所有樹都依賴相同的預測變數, 使得樹的結構更具多樣性、降低預測結果間的相關性, 整合 B 棵決策樹後的預測結果便較能達到降低預測變異的效果。

2.4 小結

本章節僅介紹監督式學習中最常見的收縮方法 (如 LASSO 與 Ridge Regression) 與樹基方法 (如決策樹與隨機森林), 針對其他進階方法 (如 Boosting、Support Vector Machines 等), 建議讀者參閱相關教科書 (James et al., 2013)。在下一章中, 我們將介紹如何應用機器學習的方法在因果推論的問題。

3 機器學習與因果推論的結合

在簡介了監督式機器學習後, 在這一章中, 我將介紹兩種結合監督式機器學習與因果推論的方法: PDSLASSO 與 DDML。

臺灣大學學術期刊資料庫

3.1 控制變數的選取

在高維度的經濟模型中, 當解釋變數的數量 p 超過樣本數 n 時, 最小平方法不具有唯一解。而在進行因果推論時, 若有眾多的控制變數, 也有類似的問題。首先, 考慮以下模型:

$$y_i = \alpha D_i + X_i'\beta + \epsilon_i。 \tag{3}$$

其中, y_i 為被解釋變數 (outcome); D_i 為處理變數 (treatment), 可以是連續或是虛擬變數, 例如生育補助的金額或者是否接受職業訓練的虛擬變數; X_i 是維度為 p 的控制變數向量; ϵ_i 為誤差項, 假設其滿足 $E(\epsilon_i|D_i, X_i) = 0$。此為誤差項條件外生的假設, 在這個假設下, 給定 X, 實驗組與對照組平均而言有相同的誤差項, 故又稱 Selection on Observables 的假設。當這個假設成立時, 最小平方法的估計為不偏的估計。然而, 若有為數眾多的控制變數 ($p > n$), 尤其是在考慮控制變數的高階項與彼此間的交乘項的時候, 潛在的控制變數數量往往迅速增加, 這個時候控制變數要如何選擇呢? 一方面, 若錯誤地排除控制變數可能導致遺漏變數偏誤; 另一方面, 全部納入則因 $p > n$, 而造成估計上的困難。

我們可以用遺漏變數偏誤的公式來思考該納入哪些控制變數, 假設我們沒有納入任何控制變數, 僅以簡單最小平方法估計 α, 則

$$\hat{\alpha}_{\text{bivariate}} \to \alpha + \beta \cdot \frac{\text{Cov}(D_i, X_i)}{\text{Var}(D_i)}。$$

由上式可見, 遺漏某個控制變數 X_i 所造成的遺漏變數偏誤為 $\beta \cdot (\text{Cov}(D_i, X_i)/\text{Var}(D_i))$, 其中 β 為 X_i 對 y_i 的影響, $(\text{Cov}(D_i, X_i)/\text{Var}(D_i))$ 則捕捉 X_i 與 D_i 之間的關係。因此, 控制變數應包含會影響 y_i 且與 D_i 有關的變數, 若排除影響 y_i 且與 D_i 有關的變數則會造成估計 α 時的偏誤。許多實證文章在選取控制變數時是以選擇被解釋變數的決定因子的角度納入控制變數, 但納入 D_i 的影響因素在正確估計 α 時同等重要。舉例而言, 在估計職業訓練對在就業薪資的影響時, 控制變數不僅應考量薪資的決定因素, 也應該考量影響參與職業訓練的因素。

傳統方法在選擇控制變數時, 多是人為依據經濟學的理論、實證文獻或是經濟直覺。然而, 研究者可能受主觀偏好影響, 而選擇某些控制變數, 令

臺灣大學學術期刊資料庫

人質疑估計結果的正確性。在接下來的兩個章節, 我將介紹兩個利用機器學習協助我們估計因果關係的方法。

3.2 PDSLASSO

Belloni, Chernozhukov, and Hansen (2014b) 與 Belloni, Chernozhukov, and Hansen (2014a) 提出了 PDSLASSO, 透過兩次的 LASSO 選擇控制變數, 再將選擇的控制變數納入迴歸模型以最小平方法估計, 因此稱為「雙重選擇程序」(double selection procedure)。

首先, 考慮以下模型:

$$y_i = \alpha D_i + g(X_i) + \varepsilon_i。 \tag{4}$$

相較於式 (3), X_i 同為一高維度的控制變數向量, 但上式較為一般化, 因為 $g(X_i)$ 可以為非線性的函數。將上式取給定 X下的條件期望值得:

$$E\left[y_i|X_i\right] = \alpha E\left[D_i|X_i\right] + g\left(X_i\right)。$$

PDSLASSO 假設 $E[Y_i|X_i]$ 與 $E[D_i|X_i]$ 這兩個條件期望值函數可以用 X_i 中少數變數的線性組合加以近似:

$$E[y_i|X_i] \approx X_i^{'}\gamma_y + r_i; \quad E[D_i|X_i] \approx X_i^{'}\gamma_D + s_i。$$

其中, r_i 與 s_i 為近似誤差。具體而言, PDSLASSO 假設在高維度的 X_i 中, 只有少數的變數對預測 y_i 或 D_i 來說是重要的; 亦即 γ_y 與 γ_D 的非 0 係數的數量相對於樣本數 n 而言很小, 此即 sparse linear approximation。

PDSLASSO 的執行步驟如下:

1. 使用 LASSO 預測 y_i, 選出與被解釋變數關聯較強的 X_i;
2. 使用 LASSO 預測 D_i, 選出與處理變數關聯較強的 X_i;
3. 將第一步與第二步所選出的變數聯集, 並用 OLS 估計下列模型:

$$y_i = \alpha D_i + X_i'\gamma + u_i。$$

臺灣大學學術期刊資料庫

其中 X_i 僅包含被選中的變數。得到 α 的估計後, 其標準誤即為傳統的 OLS 標準誤, 也可使用常用的標準誤如異質變異穩健標準誤 (heteroskedasticity-robust standard error) 與叢聚標準誤 (clustering-robust standard error)。

以上控制變數選擇的步驟確保納入所有與 y_i 與 D_i 有顯著相關的控制變數, 即使這些變數本身對 y_i 的解釋力不高, 如果它與 D_i 高度相關, 仍可避免被排除在控制變數的集合。

有人可能會問: 既然我們的被解釋變數是 y_i, 為何不只針對 y_i 做一次 LASSO 選則控制變數即可? 原因是這樣做會忽略控制變數與處理變數 D_i 間的關係, 可能造成遺漏變數的偏誤。舉例來說, 某些控制變數可能與 D_i 高度相關, 但與 y_i 的關係不強, 因此在第一步的 LASSO 中被排除。然而, 若這些變數實際上與 y_i 中度的關聯, 則仍會導致一定程度的遺漏變數偏誤。這樣的現象, 即使在已透過交叉擬和的方式選取 LASSO 損失函數中的 λ 時仍會存在。

我們可以借助 Frisch-Waugh-Lovell 定理 (Frisch and Waugh, 1933; Lovell, 1963, 2008) 理解雙重選擇程序的道理, 根據 Frisch-Waugh-Lovell 定理, α 的最小平方估計量可由以下方式求得:

1. 以最小平方法估計以 y_i 為被解釋變數, X_i 為解釋變數的簡單線性迴歸, 得到殘差 $\tilde{y}_i = y_i - \hat{y}_i$, 其中 $\hat{y}_i = X_i(X'X^{-1})X'y$;
2. 以最小平方法估計以 D_i 為被解釋變數, X_i 為解釋變數的簡單線性迴歸, 得到殘差 $\tilde{D}_i = D_i - \hat{D}_i$, 其中 $\hat{D}_i = X_i(X'X^{-1})X'D$;
3. 以最小平方法估計以$\tilde{y}_i$為被解釋變數, $\tilde{D}_i$解釋變數的簡單線性迴歸:

$$\hat{\alpha}_{OLS} = \frac{\sum_i \tilde{y}_i \times \tilde{D}_i}{\sum_i \tilde{D}_i^2}。$$

由上式可見, α 的估計實質上是利用在控制 X_i 之後, D_i 所剩餘的變異來解釋 y_i 的變異。換言之, 控制變數的角色在於將 y_i 與 D_i 中與 X_i 相關的部分剔除 (partial out), 使得 $\tilde{D}_i$ 可以被視為條件於 X_i 的外生變異。在高維度情境下, 關鍵問題在於如何選取一組適當的控制變數, 使上述的過

臺灣大學學術期刊資料庫

程能夠有效進行。若遺漏與 D_i 相關的重要變數, 則 $\tilde{D}_i$ 仍可能與誤差項相關, 導致估計偏誤。PDSLASSO 的雙重選擇程序正是為了解決此問題: 透過同時考量與 y_i 與 D_i 相關的變數, 以降低遺漏變數偏誤。

接者, 我以「就讀美國私立大學對未來薪資的影響」的研究文獻, 說明 PDSLASSO。美國私立大學一般較難申請, 也被認為有較佳的品質, 同時學費也較高, 究竟高昂的學費值不值得是值得探討的問題。雖然美國私立大學的畢業生較公立大學畢業生有較高的畢業後薪資, 但這兩組人的各方面也有許多差異, 包含家庭背景、個人能力與對未來的志向等, 因此不易論斷兩組人薪資的差異全然反映私立大學的報酬。Angrist and Pischke (2015) 延伸 Dale and Krueger (2002) 的研究, 以美國 College and Beyond 的調查資料估計就讀私立大學對學生畢業後的薪資的影響, College and Beyond 的調查資料的特別之處在於有個人大學申請哪些學校與被那些學校錄取的資訊, Dale and Krueger (2002) 認為申請相同學校與被相同學校錄取的學生有類似的個人能力與未來發展的抱負; 的確, 他們發現申請相同學校與被相同學校錄取的學生有類似的 SAT 成績與類似的家庭背景。Angrist and Pischke (2015) 與 Dale and Krueger (2002) 依據學生申請與被錄取的學校區分為 151 組, 他們發現在控制了個人所屬的組別後, 就讀私立大學對未來薪資即無顯著的影響, 也就是說, 相較於公立大學畢業生, 美國私立大學畢業生較高的薪資反映的可能是其較強的個人能力或對未來發展較高的期許。[4]

然而, Dale and Krueger (2002) 所採用的方法面臨一項挑戰: 在納入多達 150 個虛擬變數作為控制項後, 樣本數大幅縮減約三分之二。這是因為在某些細分的組別中, 受試者非全數就讀私立大學, 便是全數就讀公立大學, 導致該類別內無法進行有效比較。樣本的大幅減少可能削弱估計結果的代表性, 並降低統計上的顯著性。Angrist and Frandsen (2022) 以相同資料運用 PDSLASSO 重新估計就讀私立大學對薪資的影響。PDSLASSO 一方面以客觀的方式選取控制變數, 另一方面也避免樣本因分組過細而大量流失。所幸, 其估計結果與 Angrist and Pischke (2015) 及 Dale and

[4]Dale and Krueger (2002) 考慮的處理變數是學生所就讀大學的平均 SAT 成績, 以此估計菁英大學對學生畢業後的薪資的影響。

臺灣大學學術期刊資料庫

Krueger (2002) 的研究一致, 不支持私立大學教育對薪資具有正向效果的結論。

總的來說, PDSLASSO 為因果推論下控制變數的選擇提供了一個較為客觀的篩選機制, 不僅考量了與 y_i 有關的控制變數也考量了 D_i 有關的控制變數, 大幅度地限縮遺漏變數偏誤的程度。然而, PDSLASSO 有兩個方法上的限制, 一個是假設 X_i 中僅有少數重要的變數與 y_i 和 D_i 有關 (sparsity), 另一個是假設 X_i 與 y_i 和 D_i 之間為線性關係 (linearity)。下一節介紹的 DDML 就是在放寬這兩個限制下, 發展出來較為一般化的估計方法。

3.3 DDML

DDML 是另一種將機器學習工具整合至因果推論框架中的方法, 由 Chernozhukov et al. (2018) 提出。DDML 允許 X_i 中有許多重要的變數與 y_i 和 D_i 有關, 且允許 X_i 與 y_i 和 D_i 之間為非線性關係。DDML 結合正交化與交叉擬合 (Cross-fitting) 估計迴歸模型 (4) 中的 α。

直觀上, DDML 可視為 Frisch-Waugh-Lovell 定理的延伸。DDML 以機器學習的方法分別得到 y 與 D 的殘差, 再以最小平方法估計以 y 的殘差為被解釋變數, D 的殘差為解釋變數的簡單線性迴歸, 得到的斜率係數即為 α 的估計值。因此, DDML 可使用各種非參數化的機器學習方法, 允許同時有眾多重要的控制變數, 也未限制控制變數與 y_i 和 D_i 為線性關係。其執行步驟如下

1. 用 X_i 結合機器學習方法預測 Y_i, 並計算殘差 $\tilde{y}_i = y_i - \hat{y}_i$;
2. 同樣地, 也用 X_i 結合機器學習方法預測 D_i, 並計算殘差 $\tilde{D}_i = D_i - \hat{D}_i$;
3. OLS 估計以 $\tilde{y}_i$ 為被解釋變數, $\tilde{D}_i$ 為解釋變數的簡單線性迴歸, 即可得到 DML 下 α 的估計值。

$$\hat{\alpha}_{\text{DML}} = \frac{\sum_i \tilde{y}_i \times \tilde{D}_i}{\sum_i \tilde{D}_i^2}。$$

臺灣大學學術期刊資料庫

另外, 為了避免過度擬合所引起之估計上的偏誤, DDML 採用交叉擬合的技術。研究者將樣本資料隨機劃分為 K 個子樣本 (folds), 對於每一個子樣本 k, 僅使用剩下的 $K-1$ 個子樣本訓練機器學習模型, 並用這個模型來預測子樣本 k 的資料中的 y_i 和 D_i, 並得到對應的殘差。如此一來, 每一筆資料的預測值都是在排除該資料時所訓練出來的結果, 避免模型在同一筆資料上同時訓練與預測而引起的偏誤。

具體來說, 我們將樣本資料 $\{(y_i, D_i, X_i)\}_{i=1}^n$ 隨機分成 K 個大小相近的子集合, 記第 k 個子集合為 I_k, 其補集合為 I_k^c。對每一個 $k = 1, \cdots, K$, 在不包含第 k 子集合的樣本 I_k^c 上, 套用機器學習方法估計條件期望函數 $\ell(X) = E[y|X]$, 並將所得到的模型應用於第 k 子集合中的觀察值, 得到預測值 $\hat{\ell}_{I_k^c}(X_i)$。同樣地, 在 I_k^c 上以機器學習方法估計條件期望函數 $m(X) = E[D|X]$, 並對 $i \in I_k$ 的觀察值產生預測值 $\hat{m}_{I_k^c}(X_i)$。將所有子樣本的預測結果合併後, 可得到 $\hat{\alpha}_{\text{DML}}$:

$$\hat{\alpha}_{\text{DML}} = \frac{\sum_{i=1}^n \left(y_i - \hat{\ell}_{I_{k(i)}^c}(X_i)\right)\left(D_i - \hat{m}_{I_{k(i)}^c}(X_i)\right)}{\sum_{i=1}^n \left(D_i - \hat{m}_{I_{k(i)}^c}(X_i)\right)^2}。$$

其中 $k(i)$ 表示觀察值 i 所屬的子集合。這個估計量即為結合交叉擬合與正交化的雙重機器學習估計量。其標準誤可使用線性模型下常用的標準誤, 如異質變異穩健標準誤與叢聚標準誤等。

相較於 PDSLASSO, DDML 放寬了線性模型與稀疏性 (僅有少數重要控制變數) 的假設。然而, DDML 雖允許更一般的函數形式並可納入大量控制變數, 其實務應用通常需要較大的樣本數, 以確保機器學習估計的穩定性。因此, 在樣本數較小或模型結構相對簡單 (近似線性) 的情境下, PDSLASSO 仍具有優勢。

3.4 小結

我們已概略介紹了 PDSLASSO 與 DDML 這兩種結合機器學習與因果推論的工具。需特別強調的是, 這些方法並未創造新的認定模型參數的策略, 其因果推論仍然仰賴誤差項條件獨立的假設。然而, 相較於傳統人為選擇

臺灣大學學術期刊資料庫

變數的做法, 這些方法提供了在高維度控制變數下系統性的變數選擇機制, 也可作為客觀的穩健性檢驗。

值得注意的是, 無論是 PDSLASSO 中可供選擇的控制變數集合 X, 或 DDML 中納入機器學習的預測變數 X, 仍需由研究者事前設定與挑選。若 X 中缺乏良好的變數, 即便運用強大的機器學習方法, 也無法正確認定變數間的因果關係。

最後, 研究者也應避免將受到處理變數影響的變數納入 X 作為控制變數。這類變數為內生變數, 不僅會使控制變數的估計係數產生偏誤, 更會導致處理效果本身的估計偏誤, 亦即所謂的 Bad Control Problem (Angrist and Pischke, 2015)。PDSLASSO 與 DDML 的演算法不具備辨認 X 是否具有內生性的機制。

4 應用分析

介紹了 PDSLASSO 與 DDML 的方法論後, 本章將透過兩個實證研究範例 — (1) 生育津貼對生育率的影響與 (2) 制度對經濟成長的影響 — 說明 PDSLASSO 與 DDML 在實證分析中的應用。相關的 STATA 程式碼已整理於附錄中, 而完整說明可參考 Ahrens, Hansen, and Schaffer (2018) 和 Ahrens et al. (2024); 至於 R 與 Python 的實作方式, 可分別參考 Bach et al. (2021) 與 Bach et al. (2022)。

4.1 生育津貼對生育率的影響

第一個應用分析是生育補助對生育率的影響, 圖 8 呈現 1999–2019 台灣各縣市平均的生育補助與育齡婦女總生育率的時間趨勢, 縣市平均育齡婦女總生育率從 1.63 降到不到 1.04, 但在生育補助於 2010 年開始大幅提高後, 育齡婦女總生育率下降的速度似乎有趨緩的現象。兩條時間序列的趨勢雖呈現負向的關係, 但在未考慮其他影響因素下, 縣市的生育補助對提振生育率的效果仍需進一步探究。

黃柏鈞與蔡欣潔 (2023) 以縣市生育補助與育齡婦女總生育率的資料探討生育補助對生育率的效果, 並估計以下雙向固定效果的模型:

$$y_{ct} = \alpha_c + \lambda_t + \delta MB_{ct} + X_{ct}'\beta + u_{ct}。 \tag{5}$$

臺灣大學學術期刊資料庫

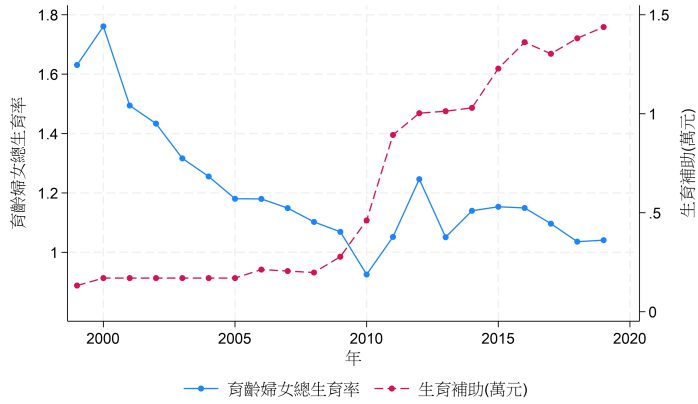


圖 8: 育齡婦女總生育率與生育補助趨勢

說明: 上圖呈現 1999–2019 年縣市平均的生育補助 (虛線) 與育齡婦女總生育率 (實線)。

其中, y_{ct} 為縣市 c 在 t 年的育齡婦女總生育率, MB_{ct} 為縣市 c 在 t 年提供第一胎生育補助的金額, $X_{ct}^{'}$為控制變數, 包括女性教育年數、男性教育年數、家戶可支配所得、每人社福支出與每人教科文支出。利用 1999–2019 縣市層級的追蹤資料, 排除縣市合併的縣市, 黃柏鈞與蔡欣潔 (2023) 的估計結果顯示生育補助每提高 1 萬元, 育齡婦女總生育率提高約 0.1 人。

這裡的應用分析延伸黃柏鈞與蔡欣潔 (2023) 的方法, 採用 PDSLASSO 檢驗黃柏鈞與蔡欣潔 (2023) 的估計結果是否受到控制變數的選取影響, 這裡的控制變數考慮女性教育年數、男性教育年數、家戶可支配所得、每人社福支出、每人教科文支出與失業率、這六個變數的平方項、這六個變數與其平方項的一期落後項與兩期落後項, 以及落後項外的所有變數的兩兩交乘項, 總共 102 個潛在的控制變數。要特別注要的是, 在執行 PDSLASSO 時, 我們要求 PDSLASSO 選入各縣市與各年度的虛擬變數 (即控制縣市與年度固定效果)。

表 1 呈現黃柏鈞與蔡欣潔 (2023) 的原始估計結果、納入 102 個控制變數的結果, 以及 PDSLASSO 的估計結果。無論是納入所有控制變數, 或透過 PDSLASSO 選擇控制變數 (PDSLASSO 選入女性教育程度平方的兩期落後項以及可支配所得平方與教科文支出平方的交乘項), 所得結果皆

臺灣大學學術期刊資料庫

表 1: 生育補助對生育率的影響

	(1)	(2)	(3)
育齡婦女總生育率			
生育津貼 (萬元)	0.114***	0.082***	0.118***
	(0.018)	(0.012)	(0.013)
黃柏鈞與蔡欣潔 (2023)	✓		
102 個控制變數		✓	
PDSLASSO			✓

說明: 各欄所使用的資料皆為 1999–2019 年全台各縣市 (不含台中市、台南市、高雄市、連江縣和金門縣) 之年資料。第 (1) 欄所使用的控制變數包含女性教育年數、男性教育年數、家戶可支配所得、每人社福支出、每人教科文支出與失業率; 第 (2) 欄所使用的控制變數包含前述這六個變數, 這六個變數的平方項、這六個變數與其平方項的一期落後項與兩期落後項, 以即落後項外的所有變數的兩兩交乘項, 總共 102 個控制變數。第 (3) 欄所使用 PDSLASSO 選擇控制變數。* 代表在 10%的顯著水準下呈現顯著, ** 代表在 5%的顯著水準下呈現顯著, *** 代表在 1%的顯著水準下呈現顯著, 括弧內為縣市層級聚集之標準誤 (clustered standard errors)。

與黃柏鈞與蔡欣潔 (2023) 原始估計相近 — 生育補助每增加 1 萬元, 育齡婦女總生育率約提升 0.1 人。

根據估計結果, 我們可推算若 2019 年各縣市平均的生育補助水準仍維持在 1999 年的水準, 則育齡婦女總生育率約為多少。由於縣市平均生育補助自 1999 年的 1,325 元成長至 2019 年的 14,380 元, 增加幅度為約 1.3 萬元, 依據估計結果, 若補助未提高, 則 2019 年總生育率約為 $1.04 - 1.3 \times 0.1 = 0.93$。然而, 這個估計無法回答若政府大幅提高生育補助 (例如每月發放 3 萬元、持續 3 年) 能多大程度提升生育率。這是因為我們的實證策略基於各縣市現有補助水準之變異, 推估其與總生育率之間的關聯。現有資料並未涵蓋如此大規模的補助政策, 因此現有結果不見得能夠外推至大幅度政策改變的情境。

臺灣大學學術期刊資料庫

4.2 制度對經濟成長的影響

4.2.1 研究簡介

Acemoglu, Johnson, and Robinson (2001) 探討國家間人均所得差異的原因, 著重在制度品質 (institutional quality) 的重要性。他們主張, 制度較健全的國家能更有效率地累積與運用物質與人力資本, 進而達到較高的經濟發展水準。他們將制度區分為兩種類型: 包容性制度 (inclusive institutions) 與掠奪性制度 (extractive institutions)。包容性制度保障財產權、提供平等機會並鼓勵創新與投資, 有助於促進長期成長; 相對地, 掠奪性制度則由少數菁英掌控政治與經濟資源, 扼殺生產誘因與社會流動性, 限制經濟發展。然而, 實證上認定制度對經濟表現的因果效果相當困難。首先, 制度與經濟成長可能存在反向因果: 較富裕的國家可能更有能力建立良好制度。其次, 也可能存在遺漏變數偏誤 — 其他未觀察到的因素 (如文化、地理條件等) 可能同時與制度的品質與收入高低有關。

為了解決制度的內生性, Acemoglu, Johnson, and Robinson (2001) 提出使用歐洲殖民者殖民時期的死亡率 (settler mortality) 作為制度品質的工具變數, 主張該變數可合理反映當初殖民策略的選擇, 進而形塑當地制度型態。殖民者在面對高死亡風險時, 通常不願定居, 而是採取掠奪性策略, 例如建立榨取資源與勞力的制度, 將當地財富輸送回母國。反之, 在死亡率較低的地區, 歐洲人較願意長期定居, 並建立保障財產權、法治與政治參與的包容性制度。這些初期制度類型具有高度的歷史持續性 (path dependence), 即使在殖民統治結束後, 這些制度架構往往延續下來, 影響後續的政府治理、投資誘因與人力資本累積, 最終反映在現代的人均所得水準。因此, Acemoglu, Johnson, and Robinson (2001) 認為殖民者死亡率與制度品質高度相關 (relevancy)。另一方面, Acemoglu, Johnson, and Robinson (2001) 認為, 在控制地理環境等變數後, 殖民者死亡率不太可能透過制度以外的其他途徑影響當前的經濟發展水平, 因此滿足工具變數所需的外生性 (exogeneity) 條件。

具體而言, Acemoglu, Johnson, and Robinson (2001)將模型設定為:

$$\log(\text{GDP per capita}_i) = \alpha \text{Protection from Expropriation}_i + x_i'\beta + \varepsilon_i。$$

臺灣大學學術期刊資料庫

GDP per capita$_i$ 為國家 i 的人均 GDP, Protection from Expropriation$_i$ 為國家 i 之人民財產免於被政府徵收的保障程度。在 Acemoglu, Johnson, and Robinson (2001) 的主要設定中, x_i 僅包括國家的緯度 (latitude), 以線性函數的方式納入作為控制變數。Acemoglu, Johnson, and Robinson (2001) 以殖民者死亡率 (Settler Mortality$_i$) 作為 Protection from Expropriation$_i$ 之工具變數估計以上模型。其估計結果顯示制度對經濟成長有顯著的影響 ($\alpha > 0$)。

Belloni, Chernozhukov, and Hansen (2014b) 運用 PDSLASSO 重新檢視 Acemoglu, Johnson, and Robinson (2001) 的研究。他們針對下列三個縮減式 (reduced form equation) 分別套用 LASSO, 並將所有被選中的變數的聯集作為控制變數集合:

$$\begin{aligned}\log(\text{GDP per capita}_i) &= x_i'\tilde{\beta} + \varepsilon_i;\\ \text{Protection from Expropriation}_i &= x_i'\tilde{\pi}_2 + v_i;\\ \text{Settler Mortality}_i &= x_i'\tilde{\gamma} + u_i。\end{aligned}$$

其中, x_i 為高維度的潛在控制變數集合。在 Belloni, Chernozhukov, and Hansen (2014b) 的實證設計中, 控制變數僅包含非洲、亞洲、北美洲與南美洲的虛擬變數, 以及 12 個緯度相關指標。而我們在實作中則納入廣泛的控制變數集合, 包括土壤、溫度、濕度、殖民母國、法律淵源、歐洲移民後代比例等地理與歷史因素等。

Chernozhukov et al. (2018) 則進一步以 DDML 檢視 Acemoglu, Johnson, and Robinson (2001) 的結果。在工具變數架構下, 他們考慮如下模型:

$$y_i = \alpha D_i + g(X_i) + \varepsilon_i。$$

其中, D_i 為處理變數, X_i 為高維度的控制變數, $g(\cdot)$ 為未知的非線性函數。Z_i 為 D_i 之工具變數。DDML 結合機器學習與交叉擬合估計以下三個條

臺灣大學學術期刊資料庫

表 2: 制度對經濟成長的影響

	(1)	(2)	(3)
Protection from Expropriation	0.94***	0.841***	0.772***
	(0.16)	(0.249)	(0.207)
Acemoglu, Johnson, and Robinson (2001)	✓		
PDSLASSO		✓	
DDML			✓

說明: 第 (1) 欄僅控制緯度變數 (Acemoglu, Johnson, and Robinson (2001) 的主要結果); 第 (2) 欄使用 PDSLASSO 選擇控制變數; 第 (3) 欄使用 DDML。三個欄位皆以殖民者殖民時期的死亡率作為制度品質的工具變數。* 代表在 10%的顯著水準下呈現顯著, ** 代表在 5%的顯著水準下呈現顯著, *** 代表在 1%的顯著水準下呈現顯著, 括弧內為標準誤。

件期望函數:

$$\ell(X_i) = \mathbb{E}[y_i|X_i];$$
$$m(X_i) = \mathbb{E}[D_i|X_i];$$
$$r(X_i) = \mathbb{E}[Z_i|X_i]。$$

在得到 $\hat{\ell}(X_i)$、$\hat{m}(X_i)$ 與 $\hat{r}(X_i)$ 後, 建構以下估計式:

$$\hat{\alpha}_{\text{DDML}} = \frac{\mathbb{E}\left[\left(y_i - \hat{\ell}(X_i)\right)\left(Z_i - \hat{r}(X_i)\right)\right]}{\mathbb{E}\left[\left(D_i - \hat{m}(X_i)\right)\left(Z_i - \hat{r}(X_i)\right)\right]}。$$

也就是說, DDML 的估計式為以殘差化的 Y 作為被解釋變數, 殘差化的 Z 作為殘差化的 D 的工具變數, 作為 α 的估計式。如表 2 所示, 其估計結果與 Belloni, Chernozhukov, and Hansen (2014b) 一致, 皆支持 Acemoglu, Johnson, and Robinson (2001) 的實證結果 — 制度對經濟成長有深遠的影響。

4.2.2 討論

不論是用 PDSLASSO 或是 DDML, 認定制度對經濟成長的影響仍仰賴一項關鍵假設: 殖民者死亡率僅透過制度影響經濟發展。若該工具變數同時影響其他未納入模型的變數, 則估計結果仍可能產生偏誤。

臺灣大學學術期刊資料庫

Glaeser et al. (2004) 指出, 歐洲人在可定居的殖民地不僅建立了制度, 也引進了自身所具備的人力資本, 例如較高的教育程度與知識體系。他們主張, 正是人力資本的提升促進了經濟成長, 而隨著人民變得更富裕, 才更有能力建立良好的制度。因此, Glaeser et al. (2004) 認為 Acemoglu, Johnson, and Robinson (2001) 的實證策略無法有效區分制度品質與人力資本對經濟成長的相對貢獻。若殖民者死亡率同時影響制度建構與當地人力資本的累積, 則違反作為制度工具變數的外生性條件。為了區分制度與人力資本的影響, Glaeser et al. (2004) 進一步引入法系是否源自法國和1500年的人口密度作為教育年數 (人力資本) 的工具變數。他們的實證結果顯示, 制度的影響在控制人力資本之後顯著下降。Chanda, Cook, and Putterman (2014) 的後續研究亦支持此一觀點: 相較於制度品質, 人力資本是影響經濟發展更根本的因素。Glaeser et al. (2004) 與後續學者的研究並不是說制度不重要, 但確實指出區分制度與人力資本對經濟成長的困難之處。[5]

雖然 PDSLASSO 與 DDML 能協助研究者系統性地選取控制變數, 或準確估計條件期望值, 但針對 Glaeser et al. (2004) 的批判, 模型內有未考慮的內生變數 — 人力資本, 此時應再尋找合適的工具變數以處理內生性問題, 而非將其當作控制變數交由機器學習選取。對此, 研究者應留意 PDSLASSO 與 DDML 方法上的侷限。

5 結語

本文簡要介紹了監督式機器學習, 以及兩種整合機器學習與因果推論的方法: PDSLASSO 與 DDML。希望藉此協助讀者理解預測與因果推論的區別、監督式機器學習的基本概念, 以及如何運用機器學習輔助因果推論。國內已有學者應用 DDML 方法於實證研究中, 如 Yang, Chuang, and Kuan

[5] Glaeser et al. (2004) 提及其觀點與 Lipset (1960) 的早期主張一致 — 他認為教育能讓公民更理性地透過選舉與協商解決衝突, 提升政府的問責與司法制度的正常運作, 識字能力也能揭露政府問題。換句話說, 制度品質反映的是社會在人力與社會資本上的累積, 而這些資本可以透過政策培養, 即使在威權體制下也有可能。這套理論似乎能夠解釋南韓、台灣與中國的經驗: 這三個國家都在威權統治下達成了高度經濟成長, 南韓與台灣後來也成功轉型為民主國家。Barro (1999) 與 Alvarez et al. (2000) 的實證研究也支持 Lipset 的看法: 經濟發展與教育進步會帶來對民主與法治更高的需求與接受度。

臺灣大學學術期刊資料庫

(2020), 未來可望有更多應用。但值得強調的是, 領域內的專業知識仍扮演關鍵角色, 特別是在選擇控制變數時, 需避免納入潛在內生變數所導致的偏誤。

本文主要聚焦於「條件獨立假設」下的因果推論, 並未涵蓋機器學習於其他實證方法中的應用, 例如工具變數法、差異中的差異法與合成控制法等。在工具變數架構中, 機器學習可用於挑選工具變數, 惟其認定假設依然仰賴工具變數的外生性。[6] 機器學習亦可應用於差異中的差異與合成控制法 (Synthetic Control Method), 如 Doudchenko and Imbens (2016) 即提出應用機器學習於合成控制法, 估計各個控制組的最適權重。

此外, 本文未涵蓋異質性處理效果的估計, 這是近期因果機器學習領域的重要發展方向之一。如 Wager and Athey (2018) 提出的因果隨機森林 (causal forest), 即結合隨機森林與因果推論估計不同群體的處理效果。Britto, Pinotti, and Sampaio (2022) 即應用該方法探討失業與失業保險對不同個體犯罪行為的影響有何差異。有興趣的讀者可參考 Athey and Imbens (2019) 或是 Chernozhukov et al. (2024) 的介紹, 進一步了解因果推論下的機器學習。

[6] Angrist and Frandsen (2022) 指出, 雖可利用 PDSLASSO 選取工具變數, 但由此建構的估計式在有限樣本下表現不如傳統的 LIML (limited information maximum likelihood) 等方法。

臺灣大學學術期刊資料庫

附錄

安裝套件

首先, 使用下列指令安裝必要的套件:

```
ssc install pdslasso // Post-Double Selection Lasso
ssc install ddml // Double Machine Learning
ssc install rforest // 隨機森林 (用於 ddml)
```

更多教學與範例可參考以下網站: https://statalasso.github.io/

STATA 實作 — 生育津貼對生育率的影響

以下程式碼展示本文所使用的資料處理與估計步驟, 包括變數標準化、多項式與落後項的產生、交乘項的建立, 以及使用 PDSLASSO ('pdslasso') 在高維度控制變數下選擇適當的控制變數。

首先, 下載黃柏鈞與蔡欣潔 (2023) 的資料與標準化變數:

```
use "https://raw.githubusercontent.com/huangpo5/github/main/reg.dta", clear
tsset v2 yr

local xx "EF EM DI_1 sc_wel edu UN"
foreach x of local xx {
    egen temp = std ('x')
    replace 'x' = temp
    drop temp
}
```

接者, 建立變數的平方項與落後項:

臺灣大學學術期刊資料庫

```
local xx2
foreach x of local xx {
    gen ‘x’2 = ‘x’2̂
    local xx2 “ ‘xx2’ ‘x’2”
}

local lagvars
foreach x in ‘xx’ ‘xx2’ {
    gen L_‘x’ = L.‘x’
    gen L2_‘x’ = L2.‘x’
    local lagvars “ ‘lagvars’ L_‘x’ L2_‘x’ ”
}
```

建立變數間的交乘項:

```
local intvars
foreach x in ‘xx’ ‘xx2’ {
    foreach y in ‘xx’ ‘xx2’ {
      if “ ‘x’ ” < ‘ “y’ ” {
        gen ‘x’_x_‘y’ = ‘x’ * ‘y’
        local intvars “ ‘intvars’ ‘x’_x_‘y’ ”
      }
    }
}
local controls “ ‘xx’ ‘xx2’ ‘lagvars’ ‘intvars’ ”
```

納入所有控制變數的估計:

```
reg TFR MP1 i.v2 i.yr ‘controls’ if v2! = 18 & v2! = 14 & v2! = 9, cl (v2)
```

建立虛擬變數:

```
tab v2, gen (county)
tab yr, gen (year)
```

最後, 以 PDSLASSO 估計:

```
pdslasso TFR MP1 (‘controls’ county* year*) if v2! = 18 & v2! = 14 & v2!
= 9, partial (county* year*) cl (v2)
```

臺灣大學學術期刊資料庫

STATA 實作 — 制度對經濟成長的影響

以下呈現以 PDSLASSO 與 DDML 實作 Acemoglu, Johnson, and Robinson (2001) 的相關程式碼。

PDSLASSO

PDSLASSO 的實作範例:

```
use https://statalasso.github.io/dta/AJR.dta, clear
pdslasso logpgp95 (lat_abst edes1975 avelf temp* humid* / / /
steplow-oilres) (avexpr=logem4), first
```

在這個範例中, `logpgp95` 是被解釋變數 (1995年人均 GDP), `avexpr` 是制度品質變數, `logem4` 則為其對應的工具變數 (殖民者死亡率)。括號中的其他變數則構成初始的控制變數集合, 包含氣候、教育與地理等因素。

DDML

至於 DDML 之 STATA 實作流程, 先設定資料與變數:

```
use https://statalasso.github.io/dta/AJR.dta, clear
global Y logpgp95
global D avexpr
global Z logem4
global X lat_abst edes1975 avelf temp* humid* steplow-oilres
set seed 42
```

其中, `Y` 為被解釋變數、`D` 為處理變數、`Z` 為工具變數, `X` 為潛在控制變數集合。

初始化 DDML 模組, 指定使用工具變數法並設定交叉擬合的折數為 30:

```
ddml init iv, kfolds (30)
```

接著, 利用隨機森林方法估計三個條件期望函數:

臺灣大學學術期刊資料庫

```
ddml E[Y|X], vtype (none): rforest $Y $X, type (reg)
ddml E[D|X], vtype (none): rforest $D $X, type (reg)
ddml E[Z|X], vtype (none): rforest $Z $X, type (reg)
```

完成模型訓練後, 執行交叉擬合以取得樣本外的預測值並殘差化各變數, 最後估計處理變數對被解釋變數的影響, 並使用異質變異數穩健的標準誤:

```
qui ddml crossfit
ddml estimate, robust
```

臺灣大學學術期刊資料庫

參考文獻

黃柏鈞與蔡欣潔 (2023), “生育補助對生育率的影響,” 《經濟論文》, 51(4), 423–459。(Huang, Po-Chun and Hsin-Chieh Tsai (2023), “The Effect of Baby Bonuses on Fertility: Evidence from Taiwan,” *Academia Economic Papers*, 51(4), 423–459.)

Acemoglu, Daron, Simon Johnson, and James A. Robinson (2001), “The Colonial Origins of Comparative Development: An Empirical Investigation,” *American Economic Review*, 91(5), 1369–1401.

Ahrens, Achim, Christian B. Hansen, and Mark E. Schaffer (2018), “PDSL ASSO: Stata Module for Post-Selection and Post-Regularization OLS or IV Estimation and Inference,” Statistical Software Components, Boston College Department of Economics.

Ahrens, Achim, Christian B. Hansen, Mark E. Schaffer, and Tobias Wiemann (2024), “ddml: Double/debiased machine learning in Stata,” *Stata Journal*, 24(1), 3–45.

Alvarez, Michael, José Antonio Cheibub, Fernando Limongi, and Adam Przeworski (2000), *Democracy and Development: Political Institutions and Well-Being in the World, 1950–1990*, New York: Cambridge University Press.

Angrist, Joshua D. and William N. Evans (1998), “Children and Their Parents’ Labor Supply: Evidence from Exogenous Variation in Family Size,” *American Economic Review*, 88(3), 450–477.

Angrist, Joshua D. and Brigham Frandsen (2022), “Machine Labor,” *Journal of Labor Economics*, 40(1), 97–140.

Angrist, Joshua D. and Jörn-Steffen Pischke (2009), *Mostly Harmless Econometrics: An Empiricist’s Companion*, Princeton, NJ: Princeton University Press.

——— (2015), *Mastering ’Metrics: The Path from Cause to Effect*, Princeton, NJ: Princeton University Press.

Athey, Susan and Guido Imbens (2019), “Machine Learning Methods Economists Should Know About,” *Annual Review of Economics*, 11(1), 685–725.

Bach, Philipp, Victor Chernozhukov, Martin Spindler Kurz, and Martin Spindler (2021), “DoubleML — An Object-Oriented Implementation of Double Machine Learning in R,” arXiv:2103.09603.

臺灣大學學術期刊資料庫

——— (2022), "DoubleML - An Object-Oriented Implementation of Double Machine Learning in Python," *Journal of Machine Learning Research*, 23(53), 1–6.

Bajari, Patrick, Denis Nekipelov, Stephen P. Ryan, and Miaoyu Yang (2015), "Machine Learning Methods for Demand Estimation," *American Economic Review*, 105(5), 481–485.

Barro, Robert J. (1999), "Determinants of Democracy," *Journal of Political Economy*, 107(S6), S158–S183.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (2014a), "High-Dimensional Methods and Inference on Structural and Treatment Effects," *Journal of Economic Perspectives*, 28(2), 29–50.

——— (2014b), "Inference on Treatment Effects After Selection Among High-Dimensional Controls," *Review of Economic Studies*, 81(2), 608–650.

Bertrand, Marianne and Sendhil Mullainathan (2004), "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review*, 94(4), 991–1013.

Britto, Diogo G. C., Paolo Pinotti, and Breno Sampaio (2022), "The Effect of Job Loss and Unemployment Insurance on Crime in Brazil," *Econometrica*, 90(4), 1393–1423.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and David Zentler-Munro (2022), "Seeing Beyond the Trees: Using Machine Learning to Estimate the Impact of Minimum Wages on Labor Market Outcomes," *Journal of Labor Economics*, 40(S1), S203–S247.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer (2019), "The Effect of Minimum Wages on Low-Wage Jobs," *Quarterly Journal of Economics*, 134(3), 1405–1454.

Chanda, Areendam, C. Justin Cook, and Louis Putterman (2014), "Persistence of Fortune: Accounting for Population Movements, There Was No Post-Columbian Reversal," *American Economic Journal: Macroeconomics*, 6(3), 1–28.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins (2018), "Double/Debiased Machine Learning for Treatment and Structural Parameters," *Econometrics Journal*, 21(1), C1–C68.

Chernozhukov, Victor, Christian Hansen, Nathan Kallus, Martin Spindler, and Vasilis Syrgkanis (2024), "Applied Causal Inference Powered by ML and AI," CausalML-book.org; arXiv:2403.02467.

臺灣大學學術期刊資料庫

Cunningham, Scott (2021), *Causal Inference: The Mixtape*, New Haven, CT: Yale University Press.

Dale, Stacy Berg and Alan B. Krueger (2002), "Estimating the Payoff to Attending A more Selective College: An Application of Selection on Observables and Unobservables," *Quarterly Journal of Economics*, 117(4), 1491–1527.

Doudchenko, Nikolay and Guido W. Imbens (2016), "Balancing, Regression, Difference-in-Differences and Synthetic Control Methods: A Synthesis," NBER Working Paper, No. 22791.

Frisch, Ragnar and Frederick V. Waugh (1933), "Partial Time Regressions As Compared with Individual Trends," *Econometrica*, 1(4), 387–401.

Glaeser, E. L., R. La Porta, F. Lopez-de-Silanes, and A. Shleifer (2004), "Do Institutions Cause Growth?" *Journal of Economic Growth*, 9(3), 271–303.

Goldin, Claudia and Cecilia Rouse (2000), "Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians," *American Economic Review*, 90(4), 715–741.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., New York: Springer.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2013), *An Introduction to Statistical Learning with Applications in R*, New York: Springer.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer (2015), "Prediction Policy Problems," *American Economic Review*, 105(5), 491–495.

Lipset, Seymour Martin (1960), *Political Man: The Social Bases of Politics*, New York: Doubleday.

Lovell, Michael C. (1963), "Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis," *Journal of the American Statistical Association*, 58(304), 993–1010.

——— (2008), "A Simple Proof of the FWL Theorem," *Journal of Economic Education*, 39(1), 88–91.

Mullainathan, Sendhil and Jann Spiess (2017), "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 31(2), 87–106.

臺灣大學學術期刊資料庫

Wager, Stefan and Susan Athey (2018), "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association*, 113(523), 1228–1242.

Yang, Jui-Chung, Hui-Ching Chuang, and Chung-Ming Kuan (2020), "Double Machine Learning with Gradient Boosting and Its Application to the Big N Audit Quality Effect," *Journal of Econometrics*, 216(1), 268–283.

投稿日期: 2025年7月18日, 接受日期: 2026年5月14日

Machine Learning and Causal Inference

Po-Chun Huang

Department of Economics, National Chengchi University

Traditional machine learning methods focus on improving predictive accuracy. Recent research, however, shows that these methods can also be applied to high-dimensional data to assist in selecting control variables, thereby improving the credibility of causal inference. This article first explains the fundamental differences between prediction and causal inference in terms of research objectives and methodology. It then introduces two integrated approaches: Post-Double Selection LASSO and Double/Debiased Machine Learning. Both methods use machine learning to support causal inference in high-dimensional settings. The article also briefly discusses two empirical applications: the effect of fertility subsidies on fertility rates and the effect of institutions on economic growth. These examples illustrate the practical use of machine learning in causal inference.

Keywords: causal inference, machine learning, post-double selection LASSO, double/debiased machine learning

JEL classification: C14, C21, C52, C55

臺灣大學學術期刊資料庫