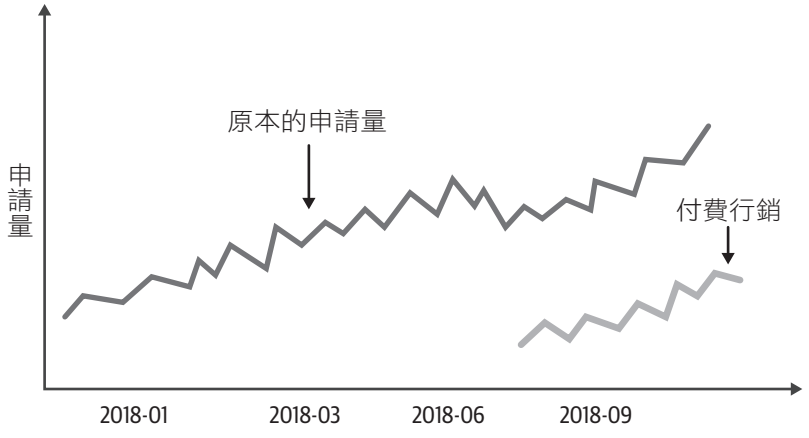
前言

想像您是一位資料科學家新手，剛進入一家正在快速成長的新創企業，雖然還沒有完全掌握機器學習的精髓，但對自己的技能充滿信心。您完成數十個線上課程，甚至在一些預測競賽中獲得不錯的名次。您現在準備將這一切知識應用於現實世界，迫不及待地想開始。前途一片光明。

然後，您的團隊領袖拿著一張圖表走過來，看起來像下圖這樣：



然後他隨口說出：「嘿，你能不能搞清楚付費行銷實際上為我們帶來多少額外客戶。一開始啟動時，的確看到一些來自付費行銷管道的客戶，但似乎原本的申請量也有所下降。我們認為**就算沒有付費行銷，還是會出現這些客戶**。」是啦！您很期待挑戰，不過……怎麼會這樣？！您哪有辦法知道沒有付費行銷會發生什麼事。我猜可以比較舉辦行銷活動前後的總申請數量，包括有或沒有付費申請。但在一家快速成長又動態變化的公司中，如何確定行銷活動辦下去後，其他一切都沒有變化（見圖 P-1）？

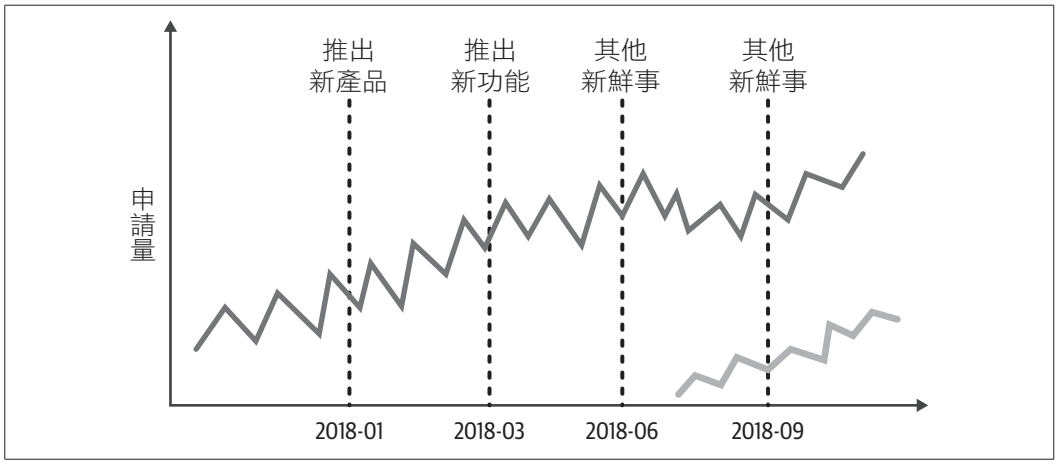
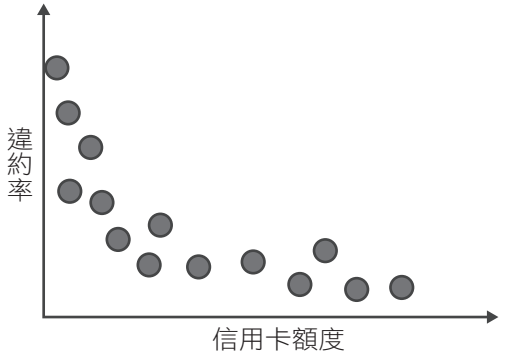


圖 P-1　產品不斷變化的快速成長公司

換個角度來說，或完全不換也行，假設您是一位出色的風險分析師，一家貸款公司剛聘用您，而您的第一個任務是完善其信用風險模型。目標是建立一個良好的自動決策系統，評估客戶的信用狀況，包括核定公司可以借款的信用額度。毫無疑問，如果信用額度很高，這個系統將為錯誤付出非常高昂的代價。

自動決策的一個關鍵組成部分，是了解較高的信用額度對客戶違約可能性的影響。他們能否管理大量的信用額度並還清，還是會陷入過度消費和無法管理的債務漩渦？為了建模這種行為，您開始繪製特定信用額度的平均違約率，讓人訝異的是，資料顯示出這種意想不到的模式：



圖表顯示，信用額度與違約率之間的關係似乎是負相關的，為什麼給予更多信用額度後，違約率反而會降低呢？您懷疑這個結果的合理性，於是去找其他分析師想試著搞懂這點。結果非常簡單：不出所料，貸款公司會給那些違約風險較低的客戶較高的信用額度。所以，並不是高額度降低違約風險，而是相反，低風險增加了信用額度。雖然這能解釋問題，但仍然沒有解決最初的問題：如何用這些資料來建模信用風險與信用額度之間的關係？當然，您不希望系統認定高信用額度等於低違約風險，也不可能天真地隨機分配額度以進行 A/B 測試，只為了看看會發生什麼事，因為錯誤的信用決策代價太高。

這兩個問題的共同點在於，您需要了解更改某些可以控制的因素，如行銷預算和信用額度，對您希望造成影響，如客戶申請和違約風險的業務後果。影響（impact）或效應（effect），可說是現代科學幾個世紀以來的支柱，但直到近期將這些工具系統化到**因果推論**（*causal inference*）領域，才取得了重大進展。此外，機器學習的進步，以及世人對於能夠透過資料來自動化和告知決策過程的普遍願望，更讓因果推論進入產業界和公共機構。然而，因果推論工具包在決策者或資料科學家界仍未廣為人知。

因為希望改變這一點，我寫了《Causal Inference for the Brave and True》這本線上書籍，涵蓋因果推論的傳統工具和最新發展，所有內容均使用開源 Python 軟體，並以嚴謹但輕鬆的方式呈現。現在，我將更進一步，從產業角度審視這一切內容，提供更新的範例，並希望提供更直觀的解釋。我的目標是讓這本書成為解答所有資料決策問題的起點。

先備知識

這本書是一本 Python 中的因果推論入門書，但不是通用的入門書。之所以稱它為入門書，是因為我將專注於因果推論的應用，而不是相關嚴格證明和定理；此外，萬不得已的情況下，我也會使用較為簡單和直觀的解釋，而非完整或複雜的說明。

它不是通用的入門書，因為我假設您已經具備一些機器學習、統計和 Python 程式設計的基礎知識；但它也不至於太進階，我仍會拋出一些您應該要事先知道的術語。

例如，可能會出現以下這段文字：

> 「首先需要解決的挑戰是連續型變數在任何地方的機率為零，即 $P(T = t) = 0$ ，這是因為機率由密度下的面積表達，而單一點的面積始終為零。一種可能的解決方案是使用條件密度函數 $f(T|X)$ 代替條件機率 $(T = t|X)$。」

我不會詳細解釋密度，以及它與機率的區別。這裡有另一個關於機器學習的例子：

> 「也可以使用機器學習模型來估計傾向分數，以作為替代方式，但這樣更需要小心。首先，必須確保機器學習模型輸出機率預測校準過；其次，需要使用摺外（out-of-fold）預測，來避免因過度擬合而產生的偏差。」

我不會在這裡解釋機器學習模型，也不會解釋校準預測的方法，或過度擬合、摺外預測的意思，這些都算是相當基本的資料科學概念，我認為您在開始前就知道了。

事實上，這裡有一份我建議閱讀本書之前，先了解的內容清單：

- 基本的 Python 知識，包括資料科學家最常用的程式庫：pandas、NumPy、Matplotlib、scikit-learn。我是念經濟學出身的，所以您不必擔心我會使用多花哨的程式碼，只要能確實掌握基礎知識就好。
- 基本的統計概念，如分布、機率、假設檢定、迴歸、雜訊、期望值、標準差和獨立性等。第 2 章會包含統計回顧，以防您需要刷新記憶。
- 基本的資料科學概念，如機器學習模型、交叉驗證、過度擬合，以及一些最常用的機器學習模型，如梯度提升、決策樹、線性迴歸、邏輯迴歸等。
- 高中數學知識，如函數、對數、根、矩陣和向量，以及一些大學等級的數學知識，如微分和積分。

本書的主要讀者是產業中的資料科學家，如果您就是，那很有可能已經涵蓋我提到的先備知識。此外，請記住這將是一個廣泛的讀者群體，擁有非常多樣的技能，因此，可能會包括一些為最進階讀者準備的筆記、段落或章節，所以如果您無法理解本書中的每一句話，也不用擔心，仍然能夠從中提取很多內容。也許您在掌握一些基礎知識後，會再次回來閱讀。

大綱

第一部分涵蓋因果推論的基本概念。第 1 章介紹因果推論的關鍵概念，並將其應用於價格調整的效應。第 2 章討論 A/B 測試（或隨機化對照試驗）的重要性，這不僅是決策工具，也是測試其他因果推論工具的黃金標準；更是回顧一些統計概念的大好機會。第 3 章偏理論性，涵蓋因果識別和圖形模型，這是一種強大的方法，就字面上的意義而言，能用於描繪對因果過程的假設，並推理您解開關聯和因果之間關係的必要步驟。完成第一部分後，您應該具備因果推論的基本思維方式。

實際範例

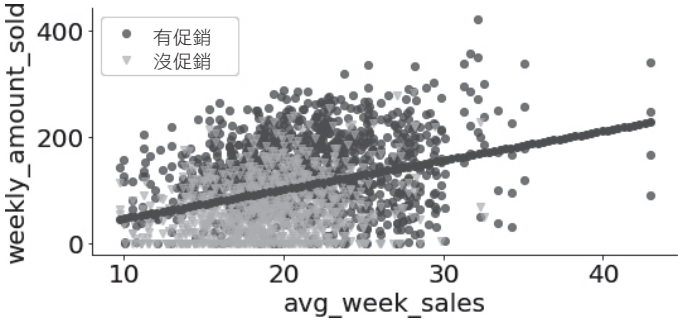
每天一杯酒，醫生遠離我

一個流行的觀念是，適量飲用紅酒對健康有益，這個觀點認為，地中海文化如義大利和西班牙以每天飲用一杯紅酒著稱，當地人的壽命同時也比較長。

但您應該對這個說法抱持懷疑態度。要將壽命延長歸因於紅酒，那些飲酒和不飲酒的人就必須是可交換的，但事實並非如此。反之，義大利和西班牙都有易取得的醫療系統和較高的人類發展指數（Human Development Index），用技術術語來說，$E[Lifespan_0 | WineDrinking = 1] > E[Lifespan_0 | WineDrinking = 0]$，因此，偏差可能會掩蓋真正的因果效應。

偏差的視覺指南

您不必只用數學和直覺來討論可交換性。在我們的例子中，甚至可以透過繪製不同處理組的變數和結果之間的關係，來檢查它們是否不可交換。如果根據企業規模（由 `avg_week_sales` 衡量）繪製結果（`weekly_amount_sold`），並用 `is_on_sale` 此處理來區分每個點的顏色，可以看到處理組（有促銷的企業）更集中在圖的右側，這表示它們通常是較大的企業；也就是說，處理組和未處理組之間是不平衡的。



這是強而有力的證據，證明您的假設 $E[Y_0 | T = 1] > E[Y_0 | T = 0]$ 是正確的。存在上升的偏差，因為降價的企業數量（$T = 1$），和這些企業在未降價時的結果（這些企業的 Y_0），會隨著企業規模的增大而增加。

如果您聽說過辛普森悖論（Simpson's Paradox），這種偏差就像是它沒那麼極端的版本。在辛普森悖論中，兩個變數之間的關係起初是正的，但一旦調整第三個變數，就變

成負的。在我們的例子中，偏差並沒有極端到改變關聯性的正負號（見圖 1-3）。這裡，起點是降價與銷售量之間的關聯性過高，控制第三個變數會減少這種關聯性的大小；如果聚焦到相同規模的企業，降價與銷售量之間的關係會減小，但仍然是正數。

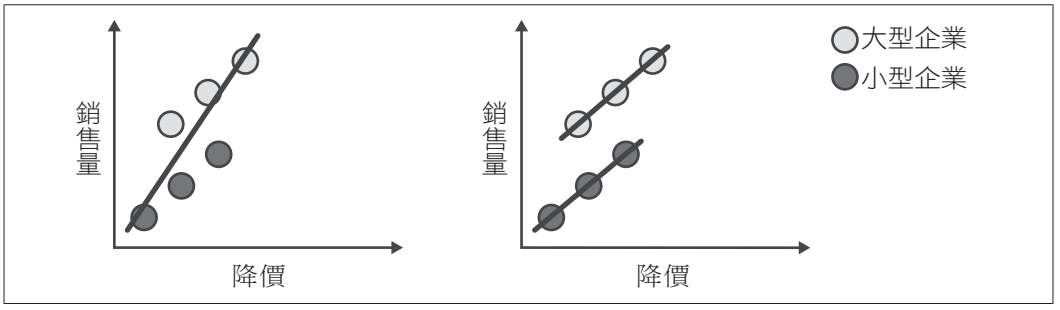
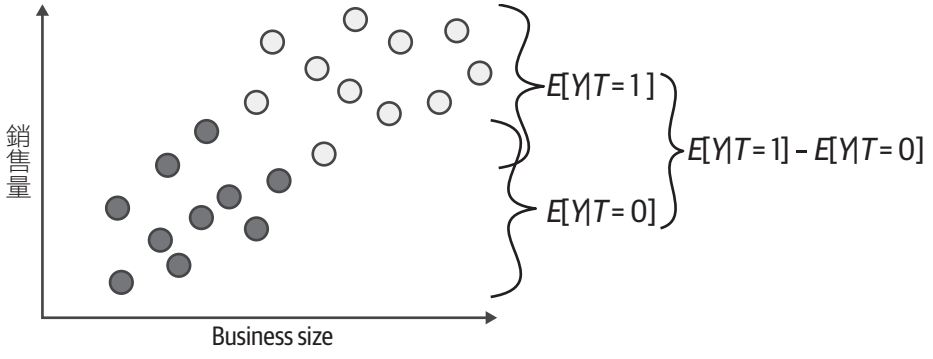


圖 1-3　偏差與辛普森悖論的關係

我必須再次強調，這一點非常重要，值得再爬梳一遍，這次用一些影像來解釋，這些影像可能不現實，但能有效解釋偏差問題。假設用一個變數表達企業規模，繪製銷售量與企業規模的關係圖後，會看到呈現一個上升趨勢，即企業規模越大，銷售量越多。接下來，根據處理的狀態來為點上色：白點表示降價的企業，黑點表示未降價的企業。如果只比較處理組和未處理組企業的平均銷售量，會得到以下結果：

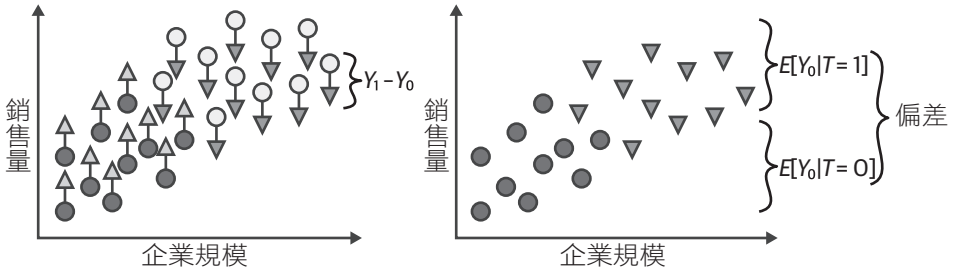


請注意兩組之間銷售量的差異，可能有兩個原因導致，且很可能就是：

1. 處理效應。由於降價導致的銷售量增加。
2. 企業規模。較大的企業能夠同時賣較多商品，也更能壓低價格。這種處理組和未處理組之間的差異並不是由於降價引起的。

因果推論的挑戰在於解開這兩個原因。

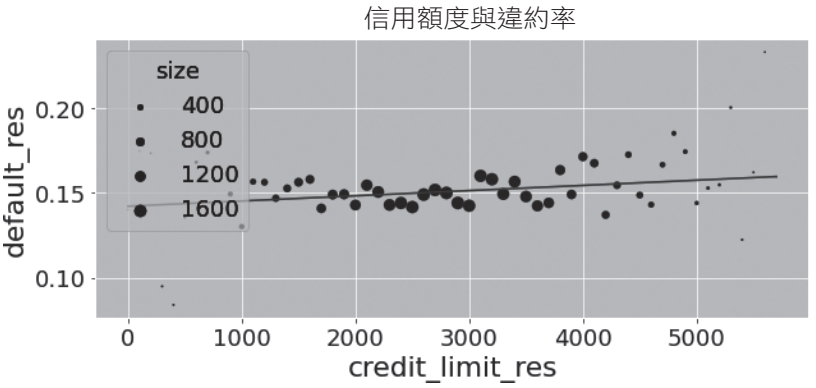
與此對比，如果您將潛在結果加入圖中（反事實結果用三角形表示）。個體處理效應是單位的結果，與相同單位在接受另一種處理時的理論結果之間的差異，您想要估計的平均處理效應，是每個個體單位潛在結果之間的平均差異 $Y_1 - Y_0$。這些個體差異遠小於您在前一個圖中所見，處理組和未處理組之間的差異。這是由於偏差，且會呈現在右圖中：



可以藉由將所有人設定為未接受處理來表達偏差，在這種情況下，您只剩下潛在結果 Y_0。然後，可以看到在無處理的情況下，處理組和未處理組在那些潛在結果上的差異。如果它們確實存在差異，也是其他因素而不是處理導致了處理組和未處理組的不同，這正是我一直在說的偏差，它是掩蓋真實處理效應的因素。

確定處理效應

現在您了解問題所在，是時候看看它的解決方案了，或者至少是某一種解決方案。識別（identification）是任何因果推論分析的第一步，第 3 章會有更多關於識別的內容，但現在先了解一下它也很值得。記住，您無法觀察因果量化指標，因為只有一個潛在結果是可觀察的；無法直接估計像 $E[Y_1 - Y_0]$ 這樣的東西，因為無法觀察到任何資料點的這種差異。但也許，您可以找到一些其他可觀察的量化指標，並可以用來找出您關心的因果量化指標，這就是識別的過程：**找出從可觀察資料中找到因果量化指標的方法**。例如，如果某種奇蹟讓 $E[Y|T=t]$ 能夠找出 $E[Y_t]$（識別 $E[Y_t]$），就可以透過簡單地估計 $E[Y|T=1] - E[Y|T=0]$ 來得到 $E[Y_1 - Y_0]$。這可以透過估計處理組和未處理組的平均結果來完成，而它們都是可觀察的量。



FWL 總結

我不知道您是否感覺得出來，但我真的很喜歡具有說明性的圖表，即使它們不反映任何真實資料，仍然可以有效幫助我們視覺化一些相當技術性的概念；這在 FWL 中也不例外。因此，總結一下，假設想估計處理變數 T 與結果變數 Y 之間的關係，但有一些混淆變數 X，您將處理變數繪製在 x 軸上，結果變數繪製在 y 軸上，混淆變數則用顏色來表示。一開始看到處理變數與結果變數之間的負相關，但有充分理由，即一些領域知識相信這種關係應該是正相關的，所以您決定對資料進行去偏差。

為此先使用線性迴歸估計 $E[T|X]$，然後建構一個去偏差版本的處理變數：$T-E[T|X]$（見圖 4-1）。使用這個去偏差處理變數，已經可以看到希望發現的正相關，但雜訊仍然很多。

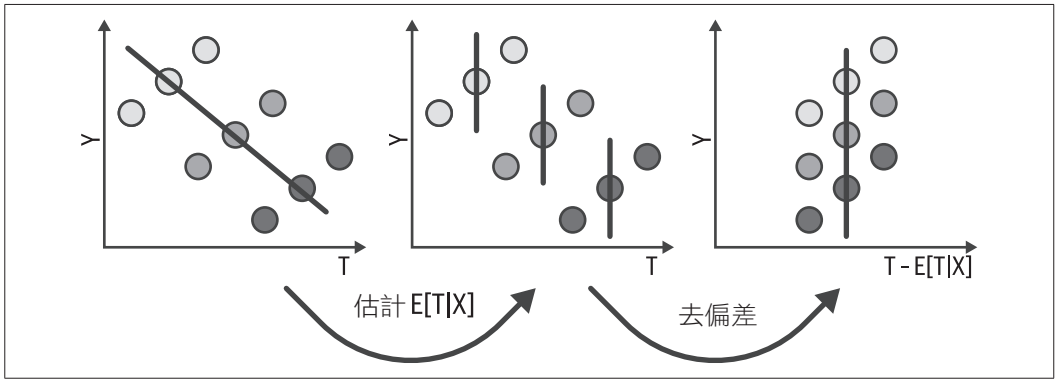


圖 4-1　正交化消除偏差的方法

為了處理雜訊，可以同樣使用迴歸模型來估計 $E[Y|X]$，然後可以建構一個去雜訊版本的結果：$Y-E[Y|X]$（見圖 4-2），將這個去雜訊結果視為在考慮所有由 X 解釋的變異數之後的結果。如果 X 解釋了 Y 中很大一部分變異數，去雜訊結果將會比較沒有雜訊，就更容易看到您真正關心的 T 和 Y 之間的關係。

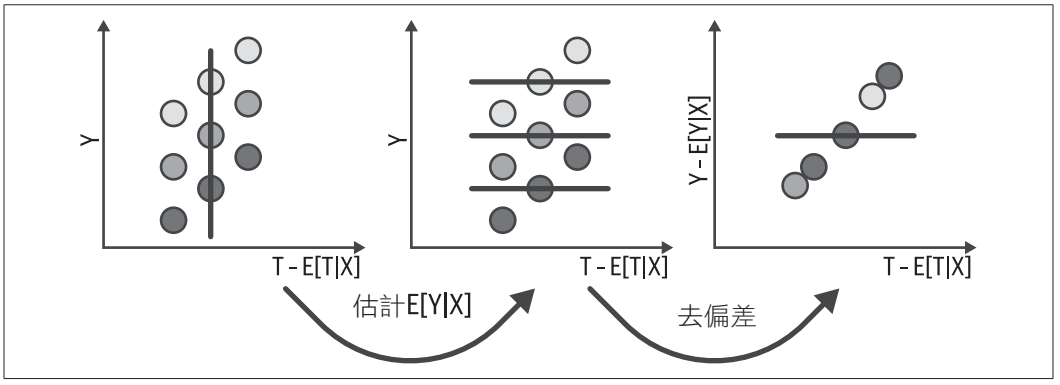
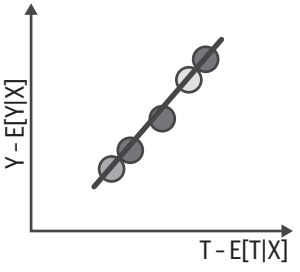


圖 4-2　正交化消除雜訊的方法

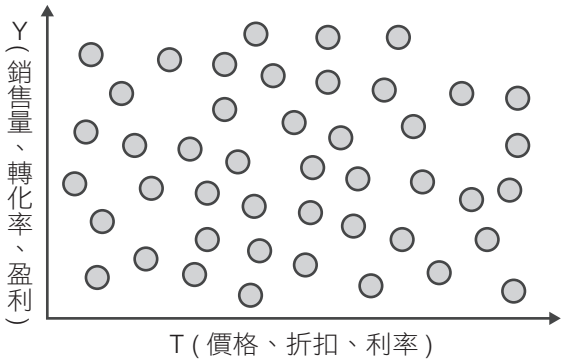
最後，在去偏差和去雜訊之後，可以清楚地看到 T 和 Y 之間的正相關，剩下的工作就是對這些資料擬合一個最終模型：



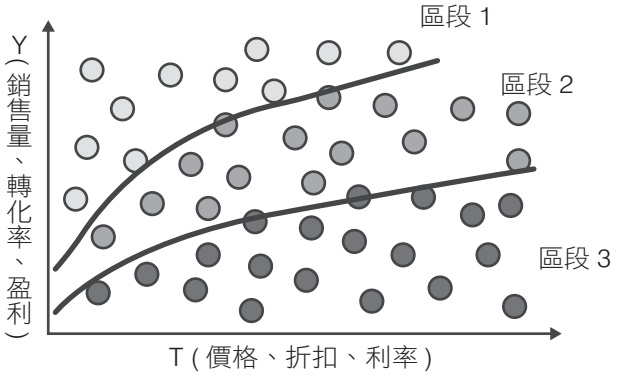
這個最終迴歸的斜率，與同時將 Y 迴歸到 T 和 X 上的結果完全一致。

去偏差和截距

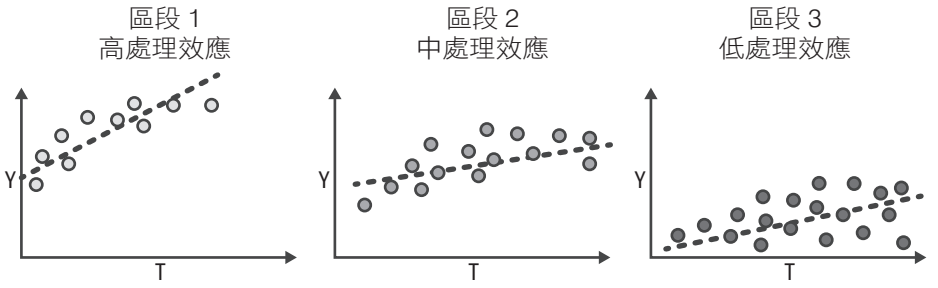
不過，有一點需要注意。在因果推論中，您大多關注這條迴歸線的斜率，因為斜率是對連續處理的效應 $\frac{\partial}{\partial t}E[Y|t]$ 線性近似。但是，如果您也關心截距，例如，如果試圖做反事實預測，應該知道去偏差和去雜訊會使截距等於零。



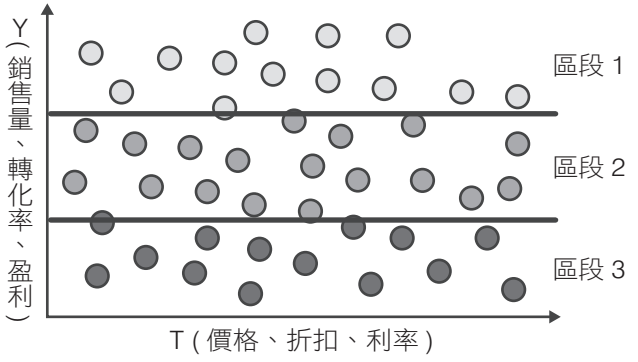
您可以將個人化任務看作是客戶細分的問題，希望根據客戶對處理的反應來創建不同的客戶群體，例如，假設想要找到對折扣反應良好的客戶，和對折扣反應不佳的客戶，則客戶對處理的回應由條件處理效應 $\delta Y/\delta T$ 提供。因此，如果能以某種方式為每個客戶估計出這個值，就可以將那些對處理回應良好的，即高處理效應，和回應不佳的客戶分組。這樣做之後，會將客戶空間分割成如下圖所示：



這樣很棒，因為現在可以估計每個群體的不同處理效應。同樣地，因為效應只是處理回應函數的斜率，如果能夠劃分出斜率不同的群體，則這些區間內的實體將會對處理有不同的反應：



現在，與傳統的機器學習方法相比，您可能會嘗試預測 Y，而不是為每個實體預測微分 $\frac{\delta Y}{\delta T}$。假設您的預測模型可以有效逼近目標，本質上是在 y 軸上分割空間，然而，這並不一定會導向擁有不同處理效應的群體，因此，只預測結果對決策來說並不總是有用：



好吧，您可能會說，我明白除了預測結果以外，還必須估計效應，但這有點棘手，如果看不到它，要怎麼預測斜率 $\frac{\delta \text{銷售量}}{\delta \text{折扣}}$ 呢？

這是個好問題。與原始結果 Y 不同，斜率（或變化率）本質上是不可觀察的單位層級資料。要觀察到個別單位的斜率，需要在不同的處理水準下觀察每個單位，並計算結果隨著每個處理變化而變化的程度：

$$\frac{\delta Y_i}{\delta T_i} \approx \frac{Y(T_i) - Y(T_i + \epsilon)}{T_i - (T_i + \epsilon)}$$

這同樣也是因果推論中的一個基本問題，永遠無法在不同處理條件下觀察到同一個體。因此，該怎麼辦？

為了簡化 CATE 模型的開發，可以創建變數來儲存處理、結果和共變數，以及訓練和測試集。一旦有了這些變數，建構幾乎所有的元學習器都會變得相當簡單：

```
In [2]: y = "next_mnth_pv"
        T = "mkt_email"
        X = list(data_rnd.drop(columns=[y, T]).columns)

        train, test = data_biased, data_rnd
```

現在已經準備好所有資料，來看看第一個元學習器。

> **因果推論程式庫**
>
> 以下所有的元學習器都已經在大多數因果推論程式庫中實作。然而，由於它們的程式非常容易編寫，我將不依賴外部程式庫，而是從零開始教您建構它們的方式。此外，在撰寫本文時，所有的因果推論套件都還處於早期階段，很難預測哪一個將在產業中占據主導地位。但這並不意味著您不應該自己去看看，當然。我特別喜歡的兩個是 Microsoft 的 `econml` 和 Uber 的 `causalml`。

T- 學習器

如果您有一個類別型處理變數，第一個應該嘗試的學習器是 T- 學習器，它非常簡單，而我猜這可能就是您已經想到的東西。它為每個處理擬合一個結果模型 $\mu_t(x)$，以估計潛在結果 Y_t，在二元情況下，只需要估計兩個模型，因此取名 T：

$$\mu_0(x) = E[Y \mid T = 0, X]$$

$$\mu_1(x) = E[Y \mid T = 1, X]$$

一旦有了這些模型，就可以為每個處理做反事實預測，並得到以下 CATE：

$$\hat{\tau}(x)_i = \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$$

圖 7-1 呈現此學習器的架構圖。

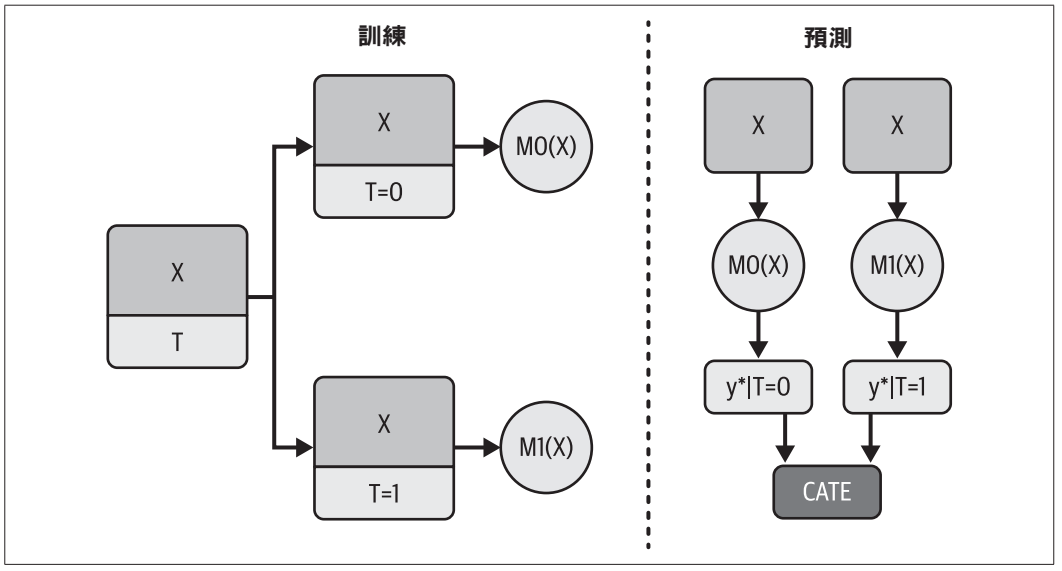


圖 7-1　T- 學習器會分別在 T = 1 和 T = 0 的資料上訓練機器學習模型，並在預測時使用這兩個模型來估計處理組與對照組之間的差異

為了寫它的程式，我將使用提升迴歸樹（boosted regression tree）作為結果模型。具體來說，我將使用 `LGBMRegressor`，這是一個非常流行的迴歸模型。我還將使用預設參數，但如果您願意，可以優化：

```
In [3]: from lightgbm import LGBMRegressor

        np.random.seed(123)

        m0 = LGBMRegressor()
        m1 = LGBMRegressor()

        m0.fit(train.query(f"{T}==0")[X], train.query(f"{T}==0")[y])
        m1.fit(train.query(f"{T}==1")[X], train.query(f"{T}==1")[y]);
```

有了這兩個模型後，在測試集上進行 CATE 預測相當容易：

```
In [4]: t_learner_cate_test = test.assign(
            cate=m1.predict(test[X]) - m0.predict(test[X])
        )
```

為了評估這個模型，我將使用第 6 章曾說的相對累積增益曲線，和該曲線下的面積。請記住，這種評估方法只關心是否正確地排序客戶，從處理效應最高的客戶到效應最低的客戶。

不連續性設計

除了傳統的工具變數和不遵從設計外，迴歸不連續設計（regression discontinuity design, RDD）也是值得一提的方法。雖然學術界廣泛使用 RDD，但它在業界的應用可能較為有限，RDD 利用處理指派中的人工不連續性來識別處理效應。例如，假設政府實施一項現金轉帳計畫，向收入低於 50 美元的低收入家庭，每月發放 200 美元的補貼，這在計畫指派中創造一個 50 美元的不連續性，允許研究人員比較收入略高於和略低於這一門檻的家庭，以衡量該計畫的有效性，前提是這兩組家庭具有相似性。

不光是這裡提到的現金轉帳計畫範例，RDD 還可以應用在許多其他情境。由於不連續性在各種場景中廣泛存在，這使得 RDD 對研究人員來說非常具有吸引力。例如，想了解念大學的影響，研究人員可以比較在入學考試中低空飛過及差一點就要通過門檻的人；要評估女性在政治中的影響，研究人員可以比較女性候選人以微小差距失敗和微小差距獲勝的城市，這樣的應用幾乎不勝枚舉。

RDD 在業界中也能發揮作用，但應用範圍可能較為有限。例如，假設一家銀行向所有客戶提供信用卡，但對那些帳戶餘額低於 5,000 美元的客戶會收取費用，這就創造了一個信用卡提供方式上的不連續性，餘額超過門檻的客戶更可能選擇高級信用卡，而餘額低於門檻的客戶則可能不會選擇。因此，RDD 可以用來比較擁有高級卡與普通卡的效應，前提是門檻上下的客戶在其他方面有其相似性。

關於不連續性設計在業界的關聯性，我認為它的應用較少，因為企業可以輕易地進行實驗來隨機化資格，就像之前討論的那樣。然而，假設在這個例子中，進行這樣的實驗會非常耗時，這可能是因為由於低遵從性造成所需的樣本量太大。

相反地，該銀行已經擁有遵循先前所描述的不連續性設計資料。因此，銀行可以利用這些資料來確定高級信用卡的效應，但要如何利用不連續性來達到這個目的呢？最簡單的方法是意識到可以將該門檻視為一個工具變數，因為越過它會增加獲得處理，即拿到高級信用卡的可能性。

下圖可以看到不連續性設計與工具變數產生關聯的方式。下半部分顯示帳戶餘額對應的反事實處理情況。由於工具變數是越過 5,000 美元的門檻，可以觀察到當**餘額** < 5,000 美元時的 T_0，以及餘額大於該門檻時的 T_1。此外，由於工具變數增加了獲得處理，也就是高級信用卡的可能性，一旦越過門檻，$P(T = 1)$ 就會上升。圖的上半部分反映這些處理機率變化影響結果的方式。

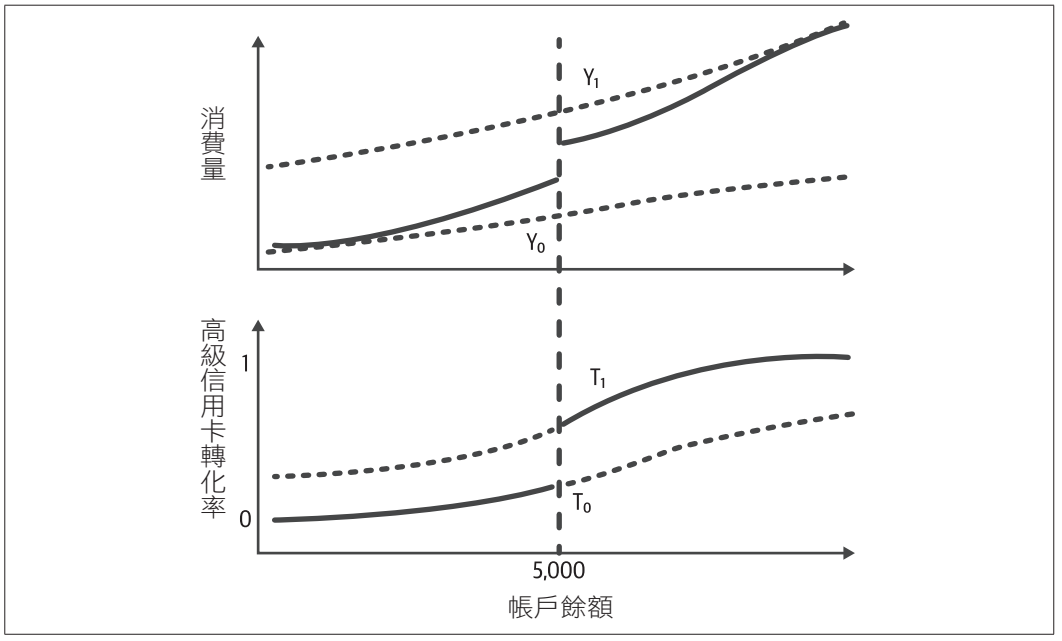


圖 11-4　不連續設計中的潛在結果與潛在處理效應

儘管即使在超過門檻後處理的機率仍然小於 1，但這使得可觀察的結果小於真正的潛在結果 Y_1。同樣地，低於門檻時，可觀察的結果高於真正的潛在結果 Y_0，這讓處理效應看起來比實際的要小，必須使用工具變數來修正這個偏差。

不連續性設計的假設

除了工具變數的假設外，不連續性設計還需要進一步的假設，即潛在結果和潛在處理函數的平滑性。以下定義一個變數 R，使得處理機率是一個在門檻 $R = c$ 處不連續的函數，在銀行例子中，R 代表帳戶餘額，而 $c = 5{,}000$。

現在，需要假設：

$$\lim_{r \to c^-} E\left[Y_t \middle| R = r\right] = \lim_{r \to c^+} E\left[Y_t \middle| R = r\right]$$
$$\lim_{r \to c^-} E\left[T_z \middle| R = r\right] = \lim_{r \to c^+} E\left[T_z \middle| R = r\right]$$