Logistic Regression Introduction

National Kaohsiung University of Sciences and Technology
Department of Finance and Information, Professor
AI Fintech Center, Director
Lin, Ping-Chen

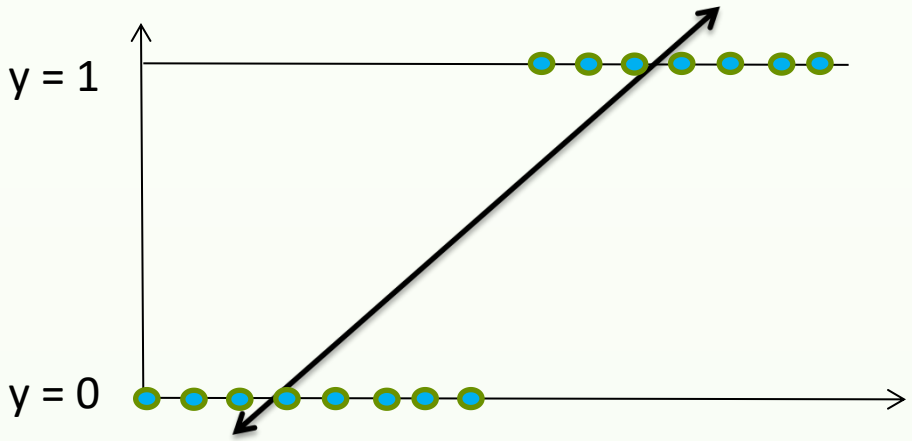
Outline

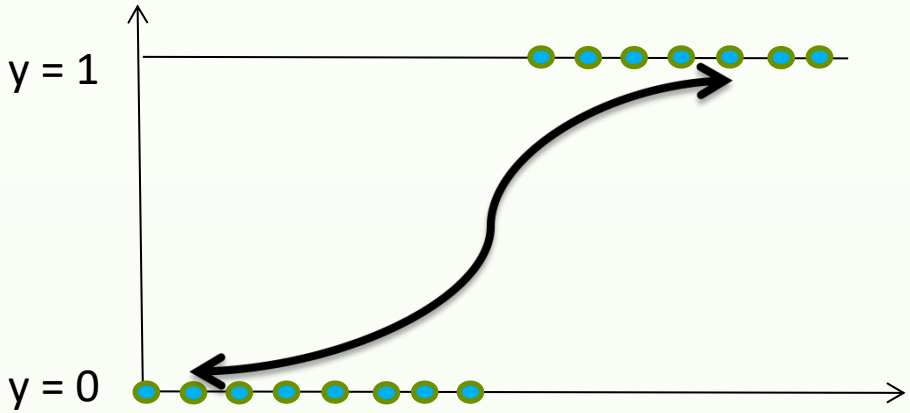
- 什麼是邏輯式迴歸 (LR)?
- 邏輯式迴歸的優點
- LR 方程式
- LR方程式的關鍵特性
- 邏輯式迴歸的假設

Logistic Regression Introduction

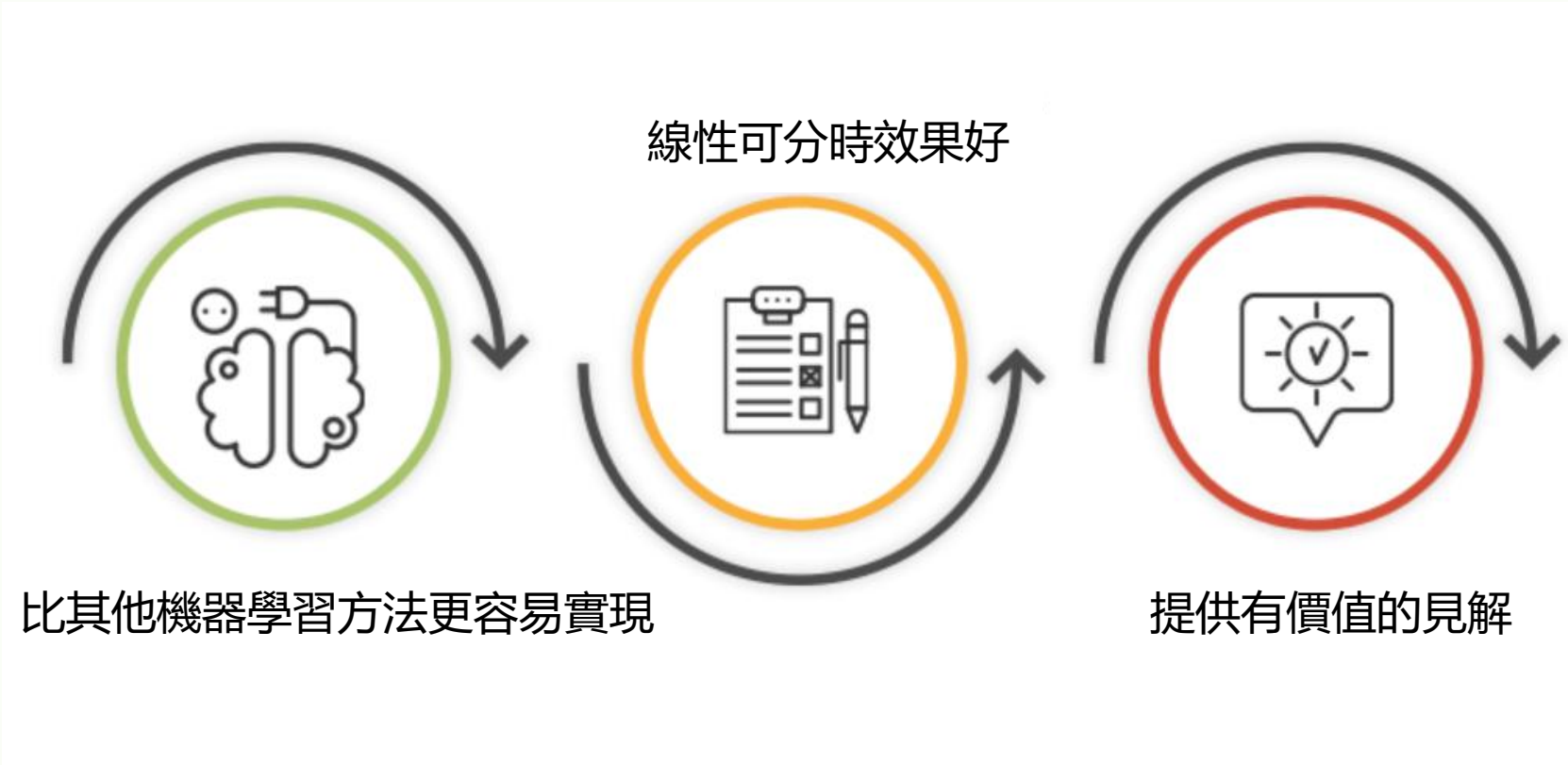
什麼是邏輯式迴歸 (LR)?

- 邏輯式迴歸是一種監督式機器學習算法，通過預測結果觀察值的機率來完成二元分類任務
- 該模型產生二元或二分結果，僅限於兩種可能的結果：是/否、0/1 或真/假
- 邏輯式迴歸分析一個或多個自變量之間的關係，並將數據分類為離散的類別
- 0 代表負類；1 代表正類。邏輯迴歸通常用於二元分類問題中，當結果變量顯示為兩個類別之一（0 和 1）





邏輯式迴歸的優點



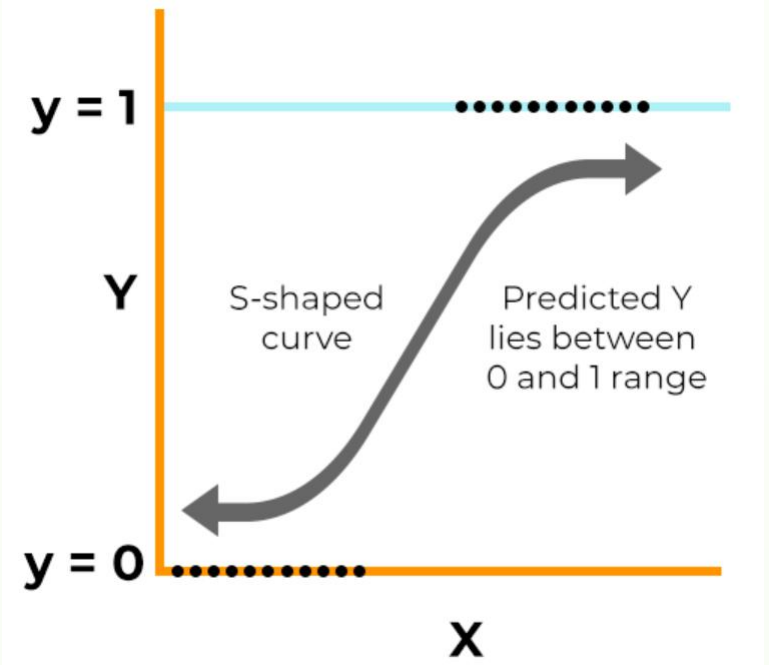
邏輯式迴歸的優點

- **1.更容易實現機器學習方法**
  - 機器學習模型可以通過訓練和測試有效設置
  - 訓練過程識別輸入數據（圖像）中的模式，並將其與某種形式的輸出（標籤）關聯起來
  - 使用迴歸算法訓練邏輯模型不需要高計算能力
  - 與其他機器學習方法相比，邏輯迴歸更容易實現、解釋和訓練
- **2.**適合線性可分的數據集
  - 線性可分的數據集指的是在圖表上可以用一條直線分隔兩個數據類別
  - 在邏輯式迴歸中，y變量只有兩個值
  - 如果使用線性可分的數據，它可以有效地將數據分類為兩個不同的類別
- **3.**提供有價值的見解
  - 邏輯式迴歸衡量獨立/預測變量的相關性或適當性（係數大小），並揭示其關係或關聯的方向（正或負）

LR 方程式

- 邏輯式迴歸使用一個稱為S型函數的邏輯函數來映射預測及其機率。S型函數指的是一條將任何實數值轉換為0到1之間範圍的S形曲線
- 如果S型函數的輸出值大於0.5，則輸出被視為1。另一方面，如果輸出值小於0.5，則輸出被分類為0
- S型函數被稱為邏輯式迴歸的激活函數，其定義為：

$$f(x) = \frac{1}{1 + e^{-x}}$$



LR 方程式

- 以下方程式表示邏輯式迴歸：
- 如果S型函數的輸出值大於0.5，則輸出被視為1。另一方面，如果輸出值小於0.5，則輸出被分類為0
- S型函數被稱為邏輯式迴歸的激活函數，其定義為：

x =輸入值
y =預測輸出
b0 =偏差或截距項
b1 =輸入(x)的係數

$$y = \frac{e^{(b_0+b_1x)}}{1+e^{(b_0+b_1x)}}$$

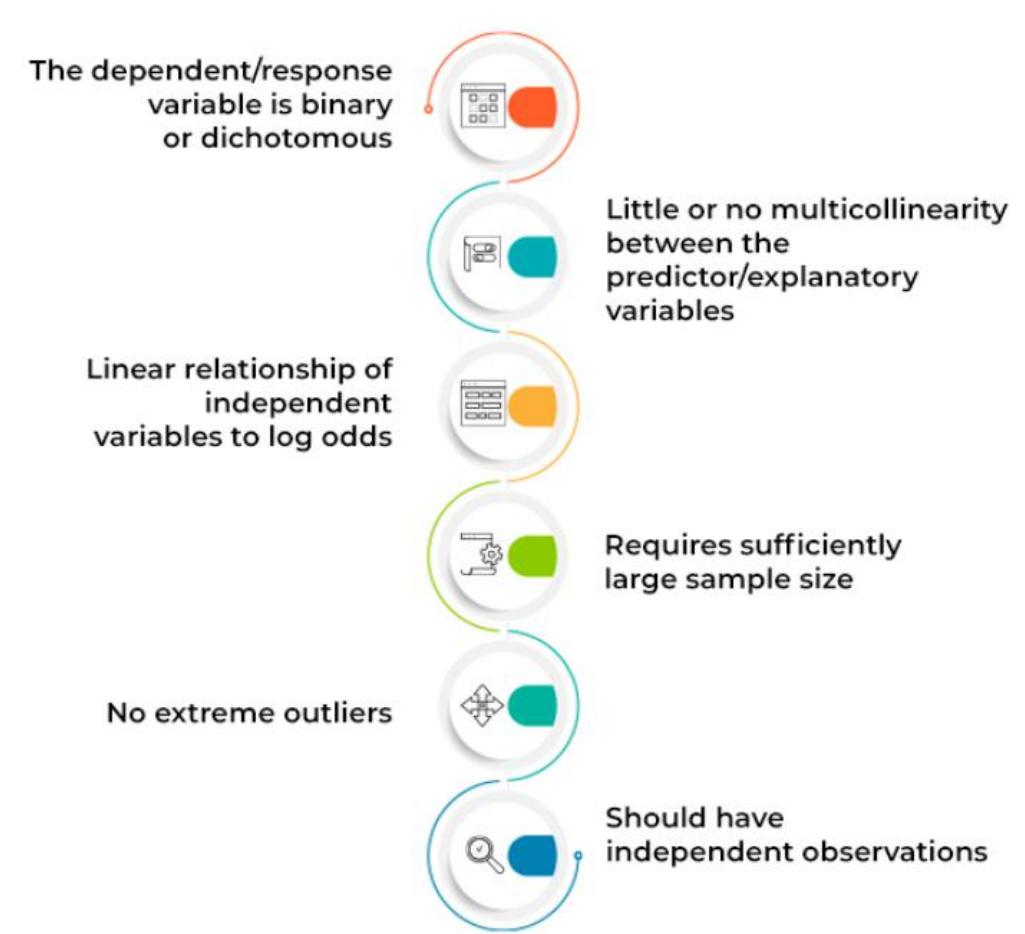
- 此方程式類似於線性迴歸，其中輸入值線性組合，使用權重或係數值來預測輸出值
- 與線性迴歸不同，此處建模的輸出值是一個二元值（0或1），而不是數值

LR方程式的關鍵特性

- 邏輯式回歸的依變量遵循伯努利分佈。
- 估計/預測基於最大概似估計
- 邏輯回歸不會像線性回歸那樣評估決定係數（或 R 平方）。相反，模型的擬合度是通過一致性來評估

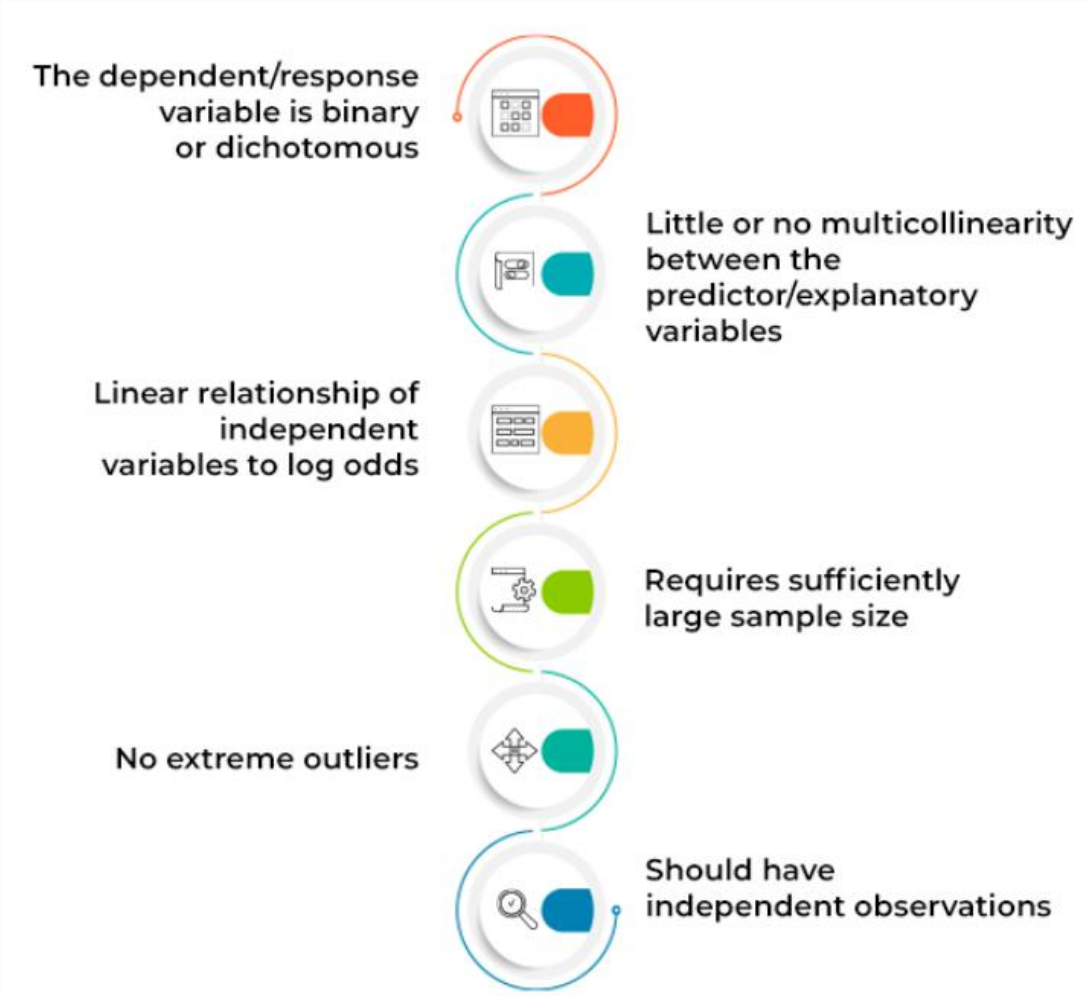
邏輯式迴歸的假設

- 1.依變量/響應變量是二元或二分的
  - 只能有兩個可能的結果——例如，通過/不通過、男性/女性、惡性/良性。
  - 可以通過簡單地計數依變量的唯一結果來檢查這一點。如果出現超過兩個可能的結果，那麼可以認為這一假設被違反了。



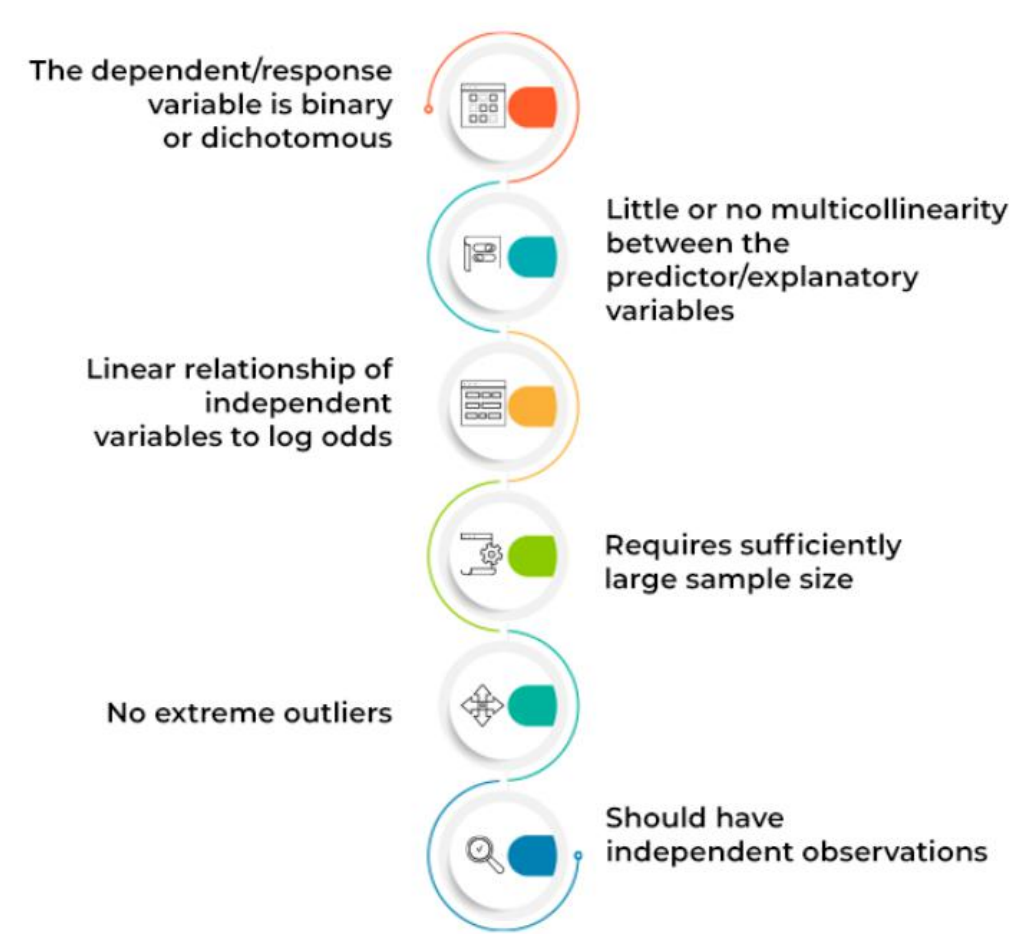
邏輯式迴歸的假設

- 2.預測變數/自變數之間幾乎沒有多重共線性
  - 預測變數（或自變數）應該彼此獨立
  - 多重共線性指的是兩個或多個高度相關的自變數
  - 這些變量在回歸模型中不提供獨特的信息，並可能導致錯誤的解釋
  - 可以通過變異數膨脹膨脹因子（VIF）來驗證這一假設，VIF 用於確定回歸模型中自變量之間的相關強度



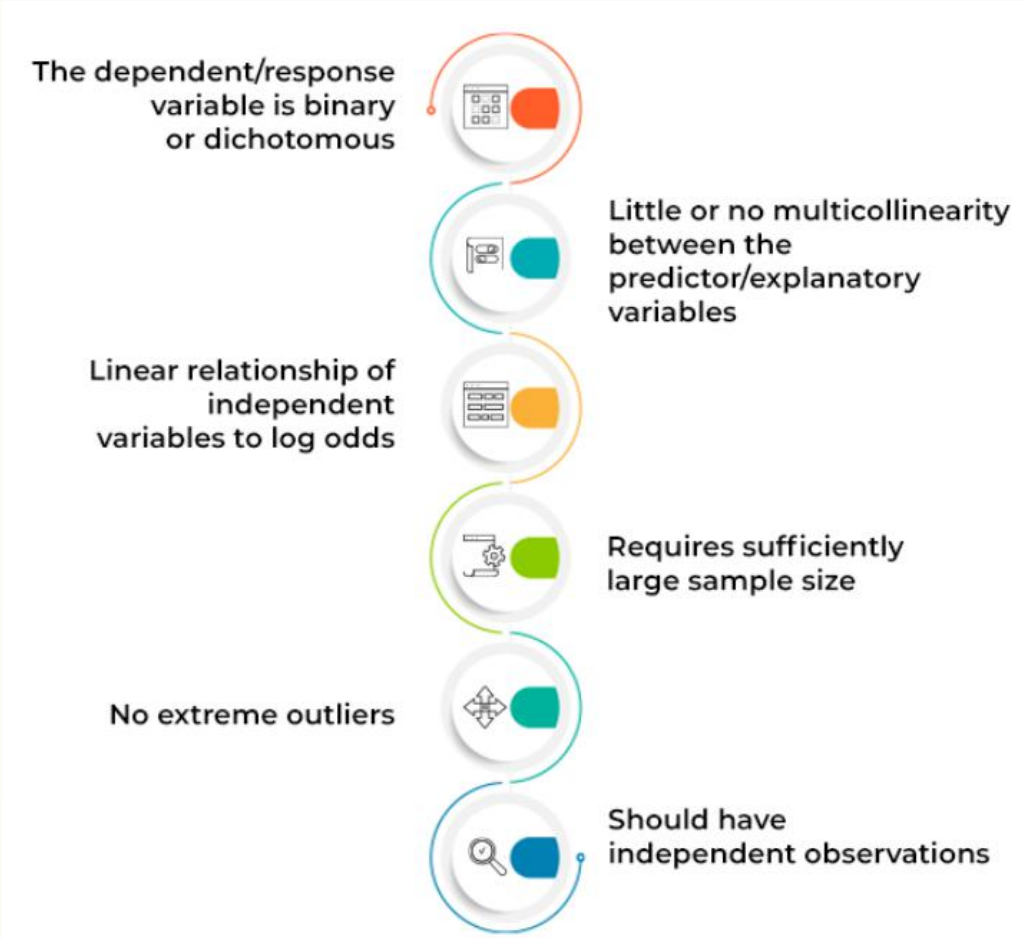
邏輯式迴歸的假設

- 3.自變數與對數發生率之間的線性關係
  - 對數發生率是表示機率的一種方式
  - 對數機率不同於概率
  - 機率指的是成功與失敗的比率，而概率指的是成功與所有可能發生事件的比率
- 4.偏好較大的樣本量
  - 當數據集的樣本量較大時，邏輯式回歸分析會產生可靠、穩健和有效的結果



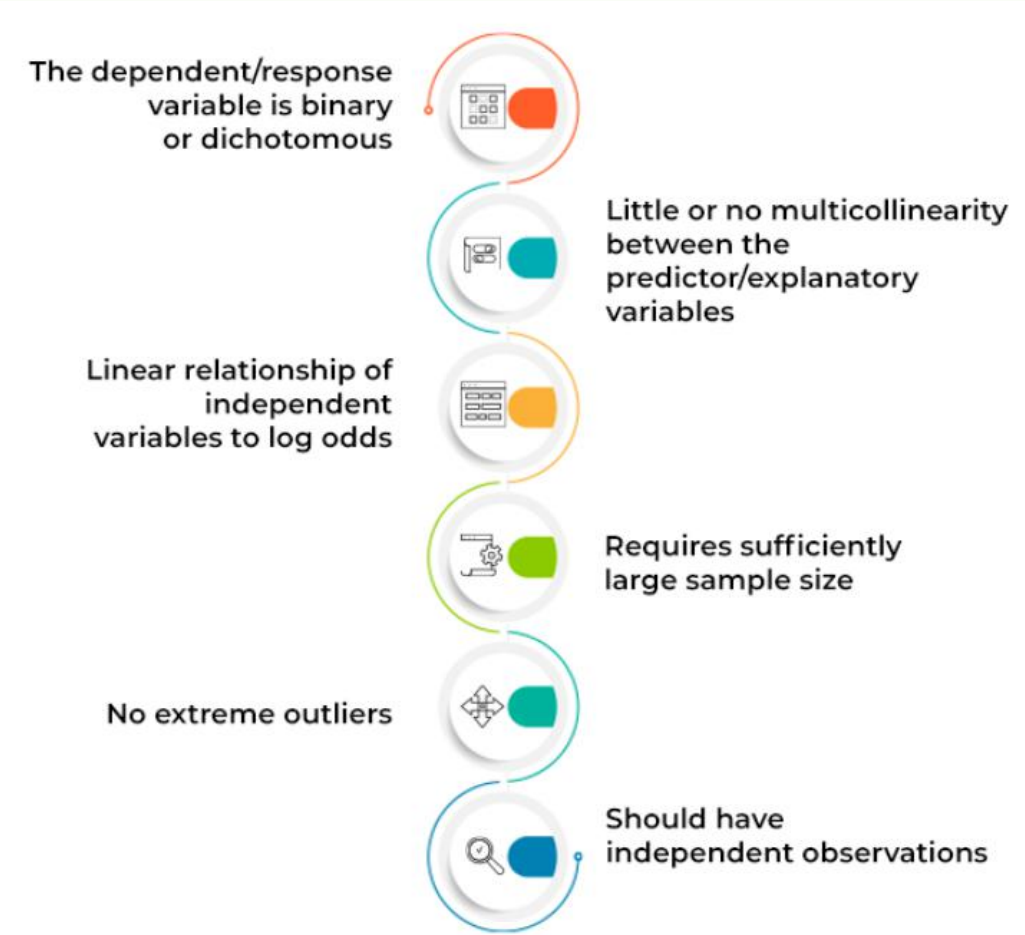
邏輯式迴歸的假設

- 5.極端異常值問題
  - 要求數據集中沒有極端異常值
  - 存在異常值的情況下，可以實施以下解決方案：
    - 消除或移除異常值
    - 考慮使用均值或中位數代替異常值，保留異常值在模型中，但在報告迴歸結果時記錄它們



邏輯式迴歸的假設

- 6.考慮獨立觀察
  - 這表示數據集中的觀察值應該彼此獨立。觀察值不應該彼此相關或來自於對相同個體類型的重複測量
  - 可以通過將殘差與時間繪圖來驗證這一假設，這顯示了觀察的順序
  - 該圖有助於確定是否存在隨機模式。如果發現或檢測到隨機模式，則可以認為這一假設被違反



Reference

- https://medium.com/hackernoon/introduction-to-machine-learning-algorithms-logistic-regression-cbdd82d81a36
- https://dzone.com/articles/machinex-simplifying-logistic-regression
- https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/
- https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/

Thank you.