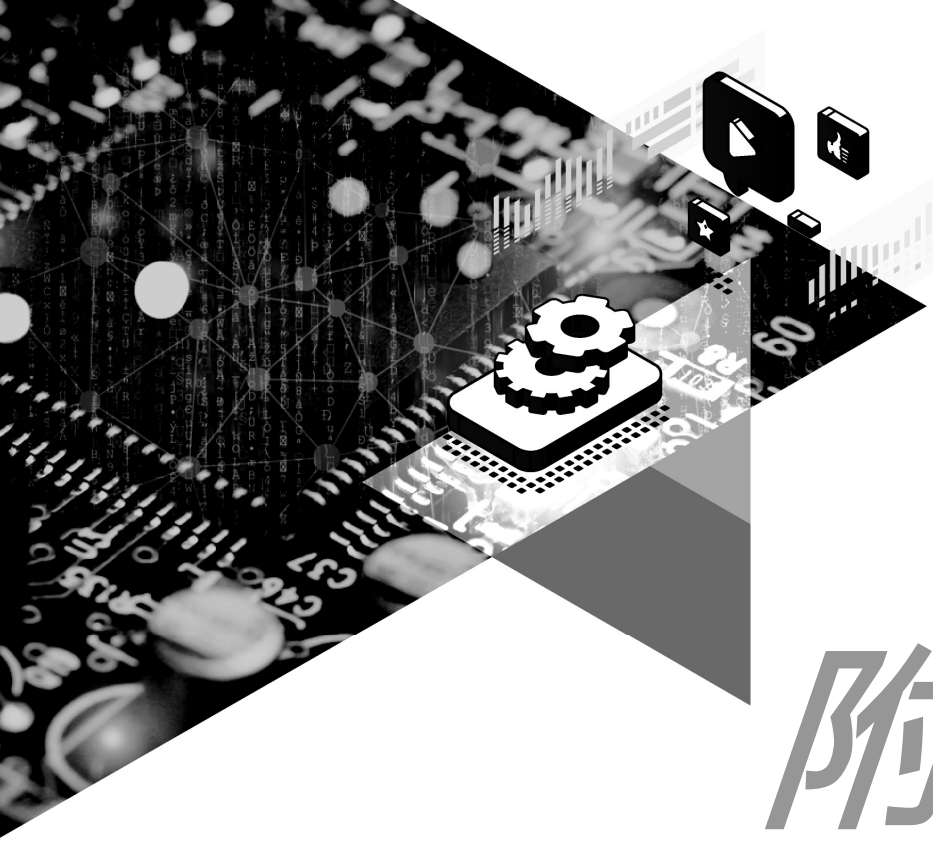

# 附錄A

# 線性迴歸

## Linear Regression

本章介紹的是「**線性迴歸**」(linear regression),此方法屬於監督式學習用於迴歸問題中的基礎模型,在數據科學中扮演著舉足輕重的角色。「線性迴歸」充分展現了統計與數據科學的精神,掌握線性迴歸的統計思維、模型解釋性以及建模步驟(包含建模前後的視覺化與評估),便能有效率地學習更進階的方法論。在本章中,首先將介紹的是「迴歸背景與觀念」,並依此精神依序說明單一特徵的「簡單線性迴歸」(simple linear regression)與多個特徵的「多元線性迴歸」(multiple linear regression)。

## A.1 迴歸背景與觀念

為解說迴歸的背景與觀念,先以一紅酒製造品質的案例出發,接著說明迴歸分析的思維。

### A.1.1 紅酒製造品質

首先,此章節以 UCI Machine Learning Repository 公開數據當中「酒品質」(wine quality dataset)的紅酒數據為例(https://archive.ics.uci.edu/ml/datasets/Wine+Quality),以紅酒品質作為目標變數,其為 0~10 分,由於品質為排序的整數,本書在目標變數加上一點擾動較符合現實品質或良率連續的特性。此數據一共包含了 11 個特徵以及 1,599 筆觀測值。作為範例僅探討「酒精度」(alcohol)、「硫酸鹽」(sulphates)、「揮發酸」(volatile acidity)三個特徵。假想若我們今天身為一個製造紅酒的工廠,我們可能會想預測紅酒品質以提升出貨品質,抑或是找出影響紅酒品質的重要特徵,在製程加工時多加留意它們的變化,以維持或提升紅酒製造的品質。於是我們該提供什麼樣的資訊滿足上述需求?以下我們先點出幾個問題供讀者思考:

- **上述三個特徵與紅酒品質是否存在相關性呢?**
  倘若我們所分析的特徵與紅酒品質沒有存在一定的關聯時,不論進行建模或是後續的分析都無法提供任何有用的資訊與效益。
- **要如何簡單快速判定相關性的存在呢?除了線性相關外,是否可能為非線性相關呢?**
  我們可用「數據視覺化」來觀察,但倘若視覺化的結果顯示為非線性相關時,是否該對數據轉換?又或是選用更合適的非線性模

型呢？

- **當某一特徵與紅酒品質存在相關性時，他們之間的關聯有多高呢？**

  假設特徵酒精度確實與紅酒品質存在相關性，他們彼此的關聯是否為給定酒精度就能直接預測紅酒品質的高相關？還是可能為比隨機亂猜來的好一些的低相關呢？

- **要如何解釋特徵與紅酒品質的關聯？又是否與領域知識相符呢？**

  在已知特徵與紅酒品質相關性的存在與高低下，這樣的線性關係要怎麼去形容與描述呢？又該如何與「領域知識」（domain knowledge）連結來說明「因果關係」呢？在紅酒的案例中，酒的發酵是由糖份在酵母菌的作用下所產生的化學變化，因此「酒精度」與葡萄的糖度高度相關，因此一般來說酒精度越高紅酒品質越好；而「硫酸鹽」的作用則是為保護紅酒不被氧化，因此一般來說適量的硫酸鹽紅酒品質越好。有趣的是，當「酒精度」與「硫酸鹽」非常高時反而會降低紅酒品質，也就是呈現兩特徵的交互作用。最後，由於高含量「揮發酸」的酒大多有強烈刺鼻的味道，會將酒的香氣蓋過，因此一般來說揮發酸越高紅酒品質越差。具備上述紅酒的領域知識後，我們便可以對數據特性與模型解釋有更高層次的理解，這也是我們一再強調的「領域知識」的重要性。

- **探討多個特徵與紅酒品質的關係，彼此間是否存在相關性或特殊的效應呢？**

  例如酒精度是否與硫酸鹽高相關呢？不論在各式各樣的模型中，特徵之間的關聯性都會連帶影響模型的建構與效果，該如何鑑別與處理特徵間關係呢？

- **數據中可能會出現極端值（異常值）的特殊案例，又該如何找出他們呢？**

  這邊提到的極端值並非特徵本身的極端值，而是特徵與紅酒品質關係的特殊情形。例如存在一瓶酒精度低但品質卻很高的紅酒，而它的出現必然會干擾我們想建立的相關性，因此需要找出及進行數據處理以提升模型與分析品質。

在上述一系列的問題中，討論了特徵與紅酒品質間的「數據視覺

化」、「相關性」、「解釋性」、「特徵間的影響」以及「極端值的存在」等問題，以期更進一步瞭解「統計學習」的本質。

### A.1.2 迴歸分析的思維

在統計學習的思維中，「**迴歸分析**」(regression analysis) 是由探討變數間的相關性所延伸發展出的分析方法，主要的目標在於找出一個合適的模型以描述特徵（自變數）與目標變數（應變數）的關係 (Montgomery and Runger, 2003; James et al., 2021)。而一般這個模型須符合嚴謹的統計假設以及具備線性可解釋的特性，而「預測」(在統計觀點常以「估計」作為用詞) 則是這個模型對新數據期望發生的結果。

迴歸分析具有一套嚴謹的推論流程，如圖 A.1 所示。首先，在收集到數據後，經由初步的「**數據視覺化**」(data visualization) 與簡單的相關性判斷，能初步瞭解數據的特性以及過濾低相關的特徵。其次，我們將對這些特徵與目標變數提出「**模型假設**」(model assumption)。接著，依照此假設模型中的迴歸係數進行「**估計與檢定**」(estimation and testing)，最後我們將分析模型的殘差與模型的配適度作為「**模型評估**」(evaluation)，若殘差符合假設以及不存在極端值的影響時，則最後即可使用這個模型對數據進行「**解釋與預測**」(interpretation and prediction)；倘若殘差不符合假設時，則回過頭進行「**模型調整**」(model adjustment)，修正模型假設或是對數據進行轉換。此流程圖與數據科學分析流程的思維與架構非常相似，而「迴歸分析」則更著重於嚴謹的模型假設與模型解釋性。

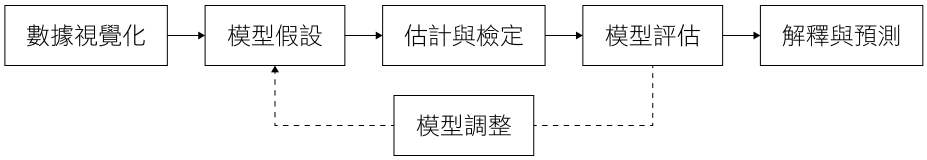


圖 A.1 迴歸分析流程

## A.2 簡單線性迴歸

### A.2.1 相關性分析

介紹簡單線性迴歸之前，我們先來探討相關性。在統計學中我們常使用相關性指標來探討兩個變量之間的關聯性，例如「**皮爾森相關係數**」

（Pearson correlation coefficient）是衡量兩變數相似性的度量，其母體與樣本的皮爾森相關係數計算公式分別為(A.1)與(A.2)。分子呈現的共變異數代表兩變數之間的關係，分母則代表兩變數各自的變異，除去各自變異的效果如同去除各自單位。

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{A.1}$$

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{A.2}$$

此方法適用於連續變數對連續變數的情形且其值介於-1 和 1 之間，其正負號則代表這兩個變數變化的方向性是否一致。若為正相關時，X 越大 Y 也會越大；若為負相關時，X 越大則 Y 會越小。如圖 A.2 第一列所示，當此係數越接近-1 或 1 表示這兩個變數具有很高的線性相關，也就是取絕對值後接近 1。此外，圖 A.2 第二列呈現當此係數為零時並非一定代表兩變數完全無關，可能是潛在的非線性相關。

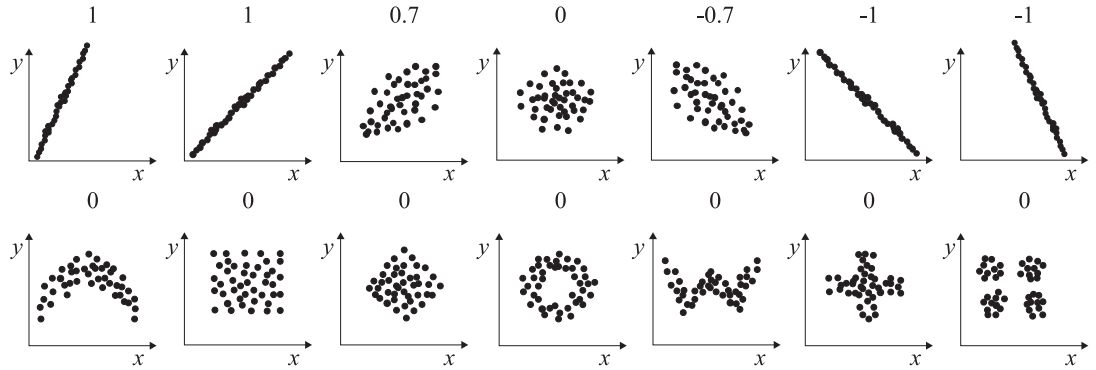


圖 A.2　相關係數的表現

## A.2.2　模型假設

「簡單線性迴歸」可就字面上來解釋。若我們定義一個特徵為$x$，而目標變數為$y$，「迴歸」是期望找出前二者的關聯，而「簡單」則是說明我們僅使用了一個特徵，而「線性」則是假設特徵$x$與目標變數$y$的關聯為線性相關，一個簡單線性迴歸的數學模型可如公式(A.3)表示。

$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{系統性的}} + \underbrace{\varepsilon}_{\text{隨機性的}} \tag{A.3}$$

上述公式將真實的目標變數$y$拆解成兩部分，前半部為「**系統性（線性）相關**」（systematic correlation），後半部則是「**隨機性誤差**」（random error）。

### A.2.2.1 系統性相關

首先，「系統性（線性）相關」是由公式(A.3)中兩個未知的迴歸係數（母體參數）$\beta_0$與$\beta_1$所描述出的，分別代表了簡單線性模型的截距項與斜率項，如圖 A.3 所示。實務上，我們不知道他們實際值為多少，因此藉由訓練集（樣本）估計出迴歸係數的估計值$\hat{\beta}_0$與$\hat{\beta}_1$。在統計學中，我們使用符號"∧"（hat）表示為某一未知參數的估計值或是目標特徵的預測值，

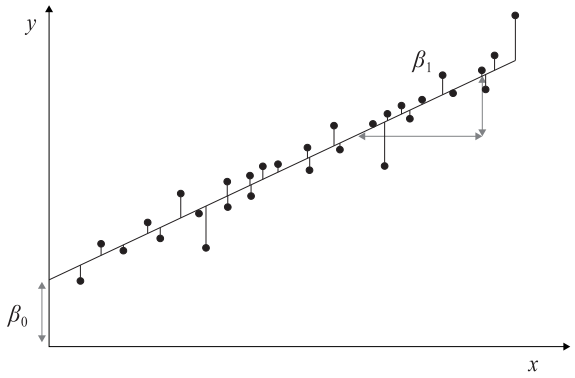


圖 A.3　線性迴歸的截距項與斜率項

### A.2.2.2 隨機性誤差

「隨機性誤差」是由公式(A.3)中迴歸模型的誤差項$\varepsilon$所描述出的，一般來說會假設服從一個平均數為 0 而變異數為常數的常態分配，並且滿足「獨立同分布」（independent and identically distributed, i.i.d.）的假設，如公式(A.4)所示。

$$\varepsilon \overset{\text{i.i.d.}}{\sim} N(0, \sigma_\varepsilon^2) \tag{A.4}$$

也就是對於真實數據的每個觀測值而言，並非完美落於線性方程式上，而

是再加上一個隨機的誤差，並且這些誤差假設是從相同的常態分配隨機產生而來的。正如同統計學在探討隨機性的精神，現實中並非完美的數學等號關係，必然會有隨機誤差的存在。舉例而言，在製造現場裡，即便我們使用的是相同製程參數進行加工，但最後品質依然會存在一定的變異。在線性迴歸中則是對這個隨機誤差做出「**同質性**」(homoscedasticity)(變異數相等)、「**獨立性**」以及「**常態分配**」的三大假設，如上述公式(A.4)所示(進階的迴歸分析方法則會打破上述假設，例如加權最小平方法用於處理異質性)。因此，如圖 A.4 所示，想像一個新的維度(機率函數)，在固定某一個$x$的數值下(與$y$軸平行的虛線)，看數據在$z$軸上的密度。基於前述假設，這個機率密度為一個常態分配，因此當我們從側面看時，可以更明顯地看出常態分配的圖形，並且不論$x$為哪一個數值，這些常態分配會是一模一樣的(i.i.d.假設)。

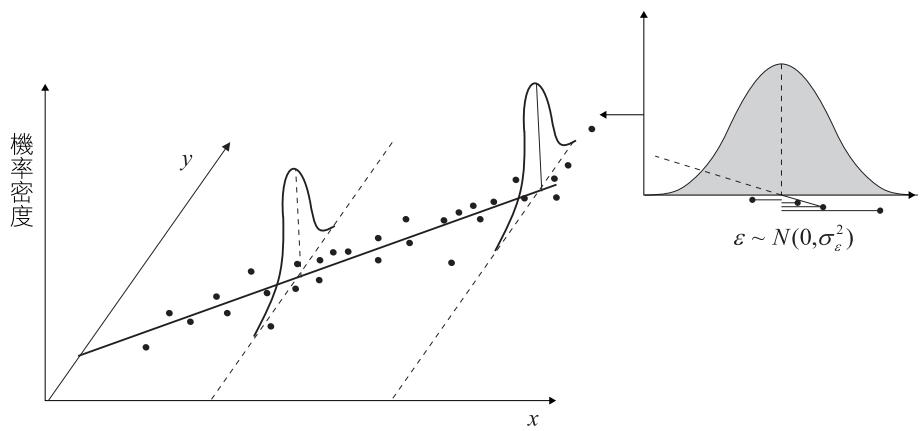


圖 A.4　殘差的常態性解釋

最後，當我們使用訓練集估計出迴歸係數的估計值$\hat{\beta}_0$與$\hat{\beta}_1$後，並可預測(估計)新的數據，而迴歸的估計式則如公式(A.5)所示，

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_{new} \tag{A.5}$$

$x_{\text{new}}$為新數據特徵的值，而$\hat{y}$代表基於某一筆新數據所產生的預測值(估計值)。若我們比較真實的迴歸模型$(\beta_0 + \beta_1 x)$與估計的迴歸模型$(\hat{\beta}_0 + \hat{\beta}_1 x)$，如圖 A.5 所示。前者真實模型是無法觀察的，除非我們能收集到所有樣本(等價於母體)，而後者估計模型則是視訓練集為母體抽樣

的樣本結果，因此當收集的數據量越多時，估計就越精確，便能更接近理想中的真實模型。

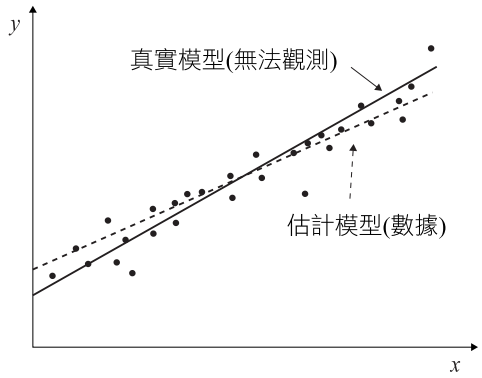


圖 A.5　真實模型與估計模型

**紅酒品質案例：相關性分析**

上述介紹了簡單線性迴歸的模型假設，而在提出模型假設之前，可藉由數據視覺化事先瞭解數據的特性。承續紅酒的數據，我們對三個不同的特徵分別與紅酒品質畫出散佈圖，如圖 A.6 所示。從圖中可看出雖所有特徵與品質的線性相關並非很明顯，但或許可看出分別的趨勢性，「酒精度」以及「硫酸鹽」與紅酒品質較為正相關，而「揮發酸」則是與紅酒品質呈現負相關。特別的是，當「硫酸鹽」過高時品質反而會下滑，而這些趨勢與數據特性與前述所提及的紅酒知識是相符的，因此我們可接續以線性模型的假設推論模型與其迴歸係數。

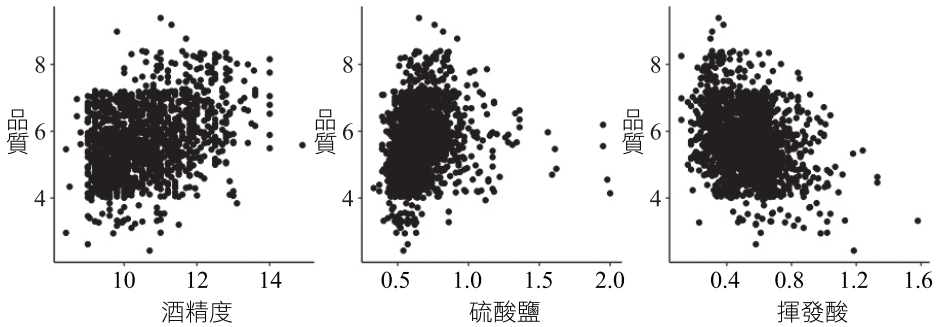


圖 A.6　特徵與目標變數間的相關性

於是我們可以對紅酒品質與酒精度提出一個線性模型的假設，特徵$x$為酒精度（$alcohol$），而目標變數$y$為紅酒品質（$quality$），如公式(A.6)所示，

$$y_{quality} = \beta_0 + \beta_1 x_{alcohol} + \varepsilon, \qquad \varepsilon \overset{\text{i.i.d.}}{\sim} N(0, \sigma_\varepsilon^2) \tag{A.6}$$

上述公式可以解釋紅酒品質與酒精度的關係為，當酒精度為0時，紅酒品質將等於截距項迴歸係數$\beta_0$。而當酒精度增加一單位時，紅酒品質將增加斜率項迴歸係數$\beta_1$，而此係數為正時，紅酒品質與酒精度便為**正相關**；反之，當$\beta_1$係數為負時，則為**負相關**。我們期望運用數據估計出此兩係數來加以預測。在真實數據中，某一特定的酒精度下紅酒品質會具有一個隨機常態（此處假設）的誤差，並非完美的線性關係。

## A.2.3 估計與檢定

在提出線性模型的假設後，如同前述提到實際$\beta_0$與$\beta_1$是未知的，因此「**估計**」此兩個迴歸係數，並接著再對估計的結果做「**統計檢定**」，以驗證估計值是具有統計顯著性。

### A.2.3.1 迴歸係數估計

迴歸係數的估計可以由兩個視角切入，分別為最小化誤差的「**最小平方法**」（ordinary least squares, OLS）與最大化可能性假設的「**最大概似估計**」（maximum likelihood estimation, MLE），有趣的是這兩種方法的估計結果是一致的。而在進入估計方法前，我們先釐清「誤差」（error）$\varepsilon$與「殘差」（residual）$\hat{\varepsilon}$的差別。如同前述統計母體與樣本思維，「誤差」是無法直接觀察得知的，需由真實模型（公式(A.3)）與估計模型（公式(A.5)）相減的「殘差」來估計，如公式(A.7)所示，

$$\begin{aligned} \hat{\varepsilon} &= y - \hat{y} \\ &= (\beta_0 + \beta_1 x + \varepsilon) - (\hat{\beta}_0 + \hat{\beta}_1 x) \\ &= y - (\hat{\beta}_0 + \hat{\beta}_1 x) \end{aligned} \tag{A.7}$$

上述公式第二列說明了，當迴歸係數的估計值越接近真實值時，殘差會越接近真實誤差；並且也同時說明了，當迴歸係數的估計值離真實值越遠

時，殘差會跟著逐漸擴大。因此，最小化殘差是估計迴歸係數的一種思維。

1. 最小平方法：「最小平方法」運用了上述最小化殘差的思維，但倘若直接將每個觀測值的殘差加總會有正負相消的情形，因此最小平方法的目標式（objective function）採用將「殘差平方和」（residual sum of squares, RSS）的方式，如公式(A.8)所示：

$$RSS = S(\beta_0, \beta_1) = \sum_{i=1}^{n} \hat{\varepsilon}^2 = \sum_{i=1}^{n} \left[ y_i - \left( \hat{\beta}_0 + \hat{\beta}_1 x_i \right) \right]^2 \tag{A.8}$$

$S$為平方加總函數，而$n$為樣本數，而殘差平方和的目標式在幾何空間上為一個凸函數（convex function）。因此，欲求出此目標式的最小值（圓點），我們可以分別對$\hat{\beta}_0$與$\hat{\beta}_1$偏微分求極小值，如公式(A.9)。

$$\begin{aligned} \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} &= -2 \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)] \overset{\text{set}}{=} 0 \\ \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} &= -2 x_i \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)] \overset{\text{set}}{=} 0 \end{aligned} \tag{A.9}$$

經求解上述的二元一次方程式，可得到估計值$\hat{\beta}_0$與$\hat{\beta}_1$，如公式(A.10)所示，其中$S_{xx}$為特徵的樣本變異數、$S_{xy}$為特徵與目標變數的樣本共變異數。

$$\begin{aligned} \hat{\beta}_0 &= \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^{n} x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \end{aligned} \tag{A.10}$$

2. 最大概似估計：接著，我們介紹迴歸係數的第二種估計方法，最大化可能性假設的「最大概似估計」。在迴歸模型中，我們假設誤差來自常態分配，因此，殘差（估計的誤差）須符合常態分配假設，如公式(A.11)所示。

$$\hat{\varepsilon} = y - \left(\hat{\beta}_0 + \hat{\beta}_1 x\right) \overset{\text{i.i.d.}}{\sim} N(0, \sigma_\varepsilon^2) \tag{A.11}$$

而最大概似估計是最大化概似函數的連乘，如公式(A.12)所示，

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \prod_{i=1}^{n} \mathcal{L}(\theta | x_i) \tag{A.12}$$

其中$\theta$為欲估計的參數，$\mathcal{L}$為概似函數，因此最大概似估計（將公式(A.11)代入公式(A.12)）求出的迴歸係數估計值如公式(A.13)所示：

$$\begin{aligned}
\hat{\beta}_{\text{MLE}} &= \arg\max_{\beta} \sum_{i=1}^{n} \log \mathcal{L}(\beta | x_i) \\
&= \arg\max_{\beta} \sum_{i=1}^{n} \log \left( \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} exp\left\{ \frac{-\left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) - 0\right)^2}{2\sigma_\varepsilon^2} \right\} \right) \\
&= \arg\max_{\beta} \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} - \frac{\left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right)^2}{2\sigma_\varepsilon^2} \\
&= \arg\max_{\beta} \sum_{i=1}^{n} -\left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right)^2 \\
&= \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right)^2
\end{aligned} \tag{A.13}$$

經公式(A.13)推導後，可發現「最大概似估計」推論的結果，即是最小化殘差的「最小平方法」。因此，經微分極值求解後的迴歸係數估計結果與公式(A.10)是相同的。

#### A.2.3.2 迴歸係數檢定

在估計完迴歸係數後，下一步則是評估這些估計值的精確度，而造成精確度有所差異的原因就來自於數據本身的變異。舉例而言，根據統計學的抽樣分配理論，當我們在估計一個隨機變數$X$的平均數$\mu$時，樣本數大小$n$將對於估計的精確度有所影響，如公式(A.14)呈現平均數估計的變異數，其中$\sigma^2$為隨機變數$X$的變異數。

$$\mathrm{Var}(\hat{\mu}) = \frac{\sigma^2}{n} \tag{A.14}$$

同理可證，迴歸係數估計的變異代表著數據線性關係的不確定性，如圖 A.7 所示。比較兩個不同特性的數據，左圖的線性關係較緊密，因此迴歸係數的變異較小，右圖則線性關係較寬鬆，係數估計變異相對較大。

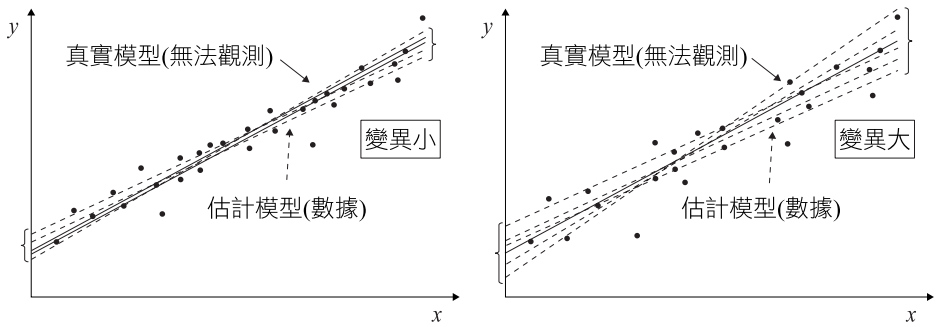


圖 A.7　迴歸係數的變異估計

接著，從公式(A.10)的迴歸係數估計式可推論估計上的變異，如公式(A.15)。

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_1) = \sigma_{\hat{\beta}_1}^2 &= \left(\frac{S_{xy}}{S_{xx}}\right)^2 = \left(\frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^n (x_i-\bar{x})^2}\right)^2 \\
&= \frac{\sum_{i=1}^n (y_i-\bar{y})^2}{\sum_{i=1}^n (x_i-\bar{x})^2} = \frac{\sigma_\varepsilon^2}{S_{xx}} \\
\mathrm{Var}(\hat{\beta}_0) = \sigma_{\hat{\beta}_0}^2 &= \mathrm{Var}(\bar{y}-\hat{\beta}_1\bar{x}) = \bar{y} + \bar{x}^2\mathrm{Var}(\hat{\beta}_1) \\
&= \frac{\sigma_\varepsilon^2}{n} + \bar{x}^2\frac{\sigma_\varepsilon^2}{S_{xx}} = \sigma_\varepsilon^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)
\end{aligned} \tag{A.15}$$

基於公式(A.4)的假設，我們也可以將目標值表示成$y \sim N(\hat{\beta}_0 + \hat{\beta}_1 x, \sigma_\varepsilon^2)$，因此迴歸係數的分配可表示如公式(A.16)所示。

$$\begin{aligned}
\beta_0 &\sim N\left(\hat{\beta}_0, \sigma_\varepsilon^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)\right) \\
\beta_1 &\sim N\left(\hat{\beta}_1, \frac{\sigma_\varepsilon^2}{S_{xx}}\right)
\end{aligned} \tag{A.16}$$

此外，由於誤差的變異$\sigma_\varepsilon^2$無法直接觀察，因此我們也可透過殘差的變異$\hat{\sigma}_\varepsilon^2$估計，如公式(A.17)。

$$\hat{\sigma}_\varepsilon^2 = \mathrm{Var}(\hat{\varepsilon}) = \frac{RSS}{n-2} \tag{A.17}$$

上式分母之所以為$n-2$，是由於殘差估計包含了兩個估計值因而減少了兩個自由度。因此，由上述迴歸係數的分配公式(A.16)，可對迴歸係數建構信賴區間，如公式(A.18)所示，

$$\begin{aligned}\hat{\beta}_0 \pm t_{\frac{\alpha}{2},n-2}\hat{\sigma}_{\beta_0} &\Rightarrow \left[\hat{\beta}_0 - t_{\frac{\alpha}{2},n-2}\hat{\sigma}_{\beta_0}, \hat{\beta}_0 + t_{\frac{\alpha}{2},n-2}\hat{\sigma}_{\beta_0}\right] \\ \hat{\beta}_1 \pm t_{\frac{\alpha}{2},n-2}\hat{\sigma}_{\beta_1} &\Rightarrow \left[\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2}\hat{\sigma}_{\beta_1}, \hat{\beta}_1 + t_{\frac{\alpha}{2},n-2}\hat{\sigma}_{\beta_1}\right]\end{aligned} \tag{A.18}$$

同理可證，基於上述信賴區間公式，可對迴歸係數做假設檢定。在給定虛無假設$H_0$為$X$與$Y$之間不存在線性關係，對立假設$H_a$則是$X$與$Y$之間存在某種線性關係，如公式(A.19)。

$$\begin{aligned}H_0&: \beta_0 = 0 \text{ v.s. } H_a: \beta_0 \neq 0 \\ H_0&: \beta_1 = 0 \text{ v.s. } H_a: \beta_1 \neq 0\end{aligned} \tag{A.19}$$

接著，並可得到迴歸係數的檢定統計量、拒絕域與顯著程度，檢定統計量如公式(A.20)所示。

$$\begin{aligned}t_{\beta_0} &= \frac{\hat{\beta}_0 - 0}{\hat{\sigma}_{\beta_0}} \\ t_{\beta_1} &= \frac{\hat{\beta}_1 - 0}{\hat{\sigma}_{\beta_1}}\end{aligned} \tag{A.20}$$

若迴歸係數是顯著的，則我們認為迴歸係數代表的線性關係是成立的。

**紅酒品質案例：估計與檢定**

承續上述紅酒數據案例，我們對紅酒品質與酒精度假設一個線性模型如公式(A.6)，接著，由上述最小平方法所得出的估計式(A.8)「估計迴歸係數」，並由公式(A.19)與公式(A.20)進行「統計檢定」，結果如表 A.1 所

示。在 95%的信心水準下，經檢定後兩迴歸係數的 p-value 均遠小於 0.05，因此統計上是顯著的，可相信他們均不為零；其次，當酒精度為0時，紅酒品質將等於截距項迴歸係數$\beta_0 = 1.868$，而斜率項迴歸係數$\beta_1 = 0.362$則代表著每當酒精度上一單位時紅酒品質上升的量。最後若將迴歸模型視覺化後，如圖 A.8 所示，藍色線為迴歸線，而灰色區域則是迴歸模型的信賴區間。

表 A.1　迴歸係數的統計檢定

| 特徵 (parameter) | 迴歸係數 (coefficient) | 標準誤 (standard error) | $t$ 檢定統計量 (t-statistic) | 顯著程度 (p-value) |
|---|---|---|---|---|
| 截距項：$\beta_0$ (intercept) | 1.868 | 0.193 | 9.643 | <0.0001*** |
| 酒精度：$\beta_1$ (alcohol) | 0.362 | 0.018 | 19.577 | <0.0001*** |

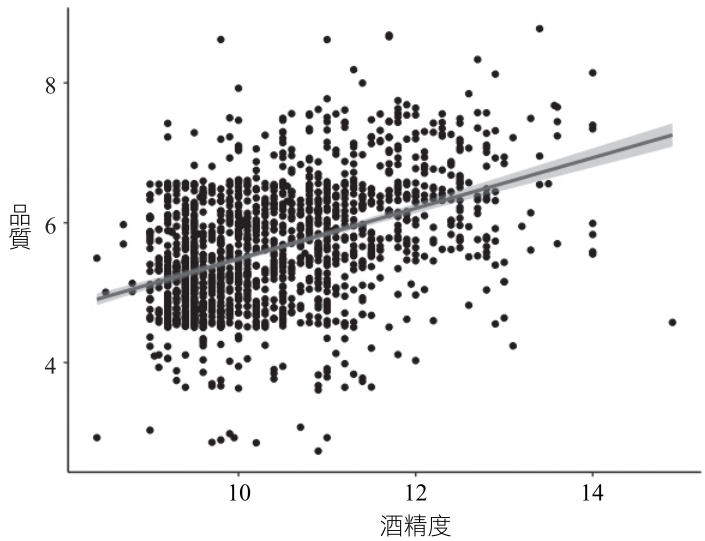


圖 A.8　紅酒品質迴歸模型與信賴區間

## A.2.4　模型評估

接著，我們進行對整體模型的評估，將其分為「**模型配適度**」與「**殘差分析**」兩部分，前者是判斷模型解釋特徵與目標變數的能力，後者則是判斷模型的假設是否滿足。

### A.2.4.1 模型配適度

模型配適度一般會使用一些指標來評估，例如「**殘差標準差**」（residual standard error, RSE）、「**判定係數**」（coefficient of determination）以及「**F 統計量**」（F-Statistic），以下我們分別介紹這三項指標。

1. 殘差標準差：「殘差標準差」呈現的是數據與模型偏離的程度，其接近「殘差均方差」（mean square error, MSE）的定義，如公式(A.21)所示。

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{A.21}$$

因此，我們期望建構的模型其殘差標準差越小越好，然而這個指標一般只能用以比較不同假設的模型（例如不同特徵子集），無法以絕對的角度直接說明該模型的表現。

2. 判定係數：「判定係數」又可被稱為「R 平方（$R^2$）」，是一項用以說明「**模型解釋變異的能力**」的指標，如圖 A.9 所示。首先我們可將「整體變異」（total variance）拆解為系統性的「解釋變異」（explained variance）與隨機性的「殘差變異」（unexplained variance）。因此，「判定係數」可解釋為「解釋變異」佔「整體變異」的比例，代表模型的解釋力，如公式(A.22)。

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{A.22}$$

$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$代表數據的「整體變異」，而分子為「解釋變異」，是由「整體變異」扣除「殘差變異」，也就是（TSS − RSS）所得到的。因此這項指標提供了一個介於 0 與 1 之間的值，這個值代表模型能解釋的變異所佔的比例。舉例而言，若一個線性迴歸模型的$R^2 = 0.7$，則我們會說此線性模型可以解釋數據 70% 的變異，當這項指標越接近 1 時，代表配適度越高。而「R 平方」的命名由來，則是由於當我們計算實際值$\boldsymbol{y}$與預測值$\hat{\boldsymbol{y}}$的相關係數$r_{y,\hat{y}}$時，如公式(A.23)，

$$r_{y,\hat{y}} = \frac{Cov(y,\hat{y})}{\sqrt{Var(y)Var(\hat{y})}} = \frac{Cov(\hat{y}+\varepsilon,\hat{y})}{\sqrt{Var(y)Var(\hat{y})}} = \frac{Var(\hat{y})}{\sqrt{Var(y)Var(\hat{y})}} = \sqrt{\frac{Var(\hat{y})}{Var(y)}} \tag{A.23}$$

可發現，若將此相關係數$r_{y,\hat{y}}$取平方後會等於「R 平方」，如公式(A.24)所示。

$$r_{y,\hat{y}}^2 = \frac{Var(\hat{y})}{Var(y)} = \frac{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\text{RSS}}{\text{TSS}} = R^2 \tag{A.24}$$

因此，「R 平方」也代表著預測模型所提供的預測值與實際值的相關性有多高，越相關自然能解釋的變異越多。

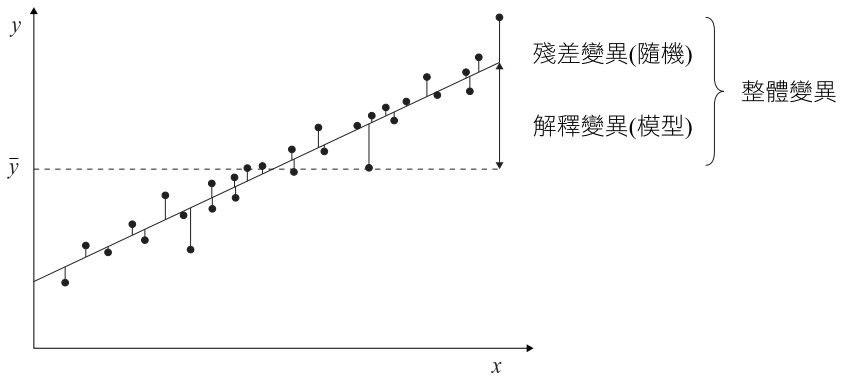


圖 A.9　整體變異、解釋變異與殘差變異

模型評估若討論「多個特徵」的情形，理論上當我們使用越多特徵建模時，殘差會越來越小，因而模型的解釋能力會越強。為了權衡「特徵個數」的影響力，放入愈多特徵在迴歸模型中，應給予$R^2$一些懲罰，因此建議「**調整後 R 平方**」（Adjusted $R^2$, $R_{adj}^2$）是相對公平的指標，如公式(A.25)所示。

$$R_{adj}^2 = 1 - \frac{\frac{\text{RSS}}{n-p-1}}{\frac{\text{TSS}}{n-1}} = 1 - (1-R^2)\frac{(n-1)}{(n-p-1)} \tag{A.25}$$

上式分別對「殘差變異」（分子）與「整體變異」（分母）除上各自的自由度，分子殘差變異的自由度為樣本數$n$減去「特徵個數與截距項」$p+1$，分母整體變異的自由度則為樣本數$n$減去一項平均數的估計1。因此，當模型使用越多特徵時，「調整後 R 平方」會對特徵個數進行懲罰。

3. F 統計量：除了調整後 R 平方外，另一項考慮特徵個數與樣本數的指標為「F 統計量」。這個指標延續了前述對單一迴歸係數檢定（$\beta_0$ or $\beta_1 = 0$）的思維，拓展至對所有迴歸係數的假設檢定。若給定$p$個特徵，其虛無假設$H_0$為所有的迴歸係數均為零，對立假設$H_a$則是至少一項迴歸係數不為零，如公式(A.26)。

$$\begin{aligned} &H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0 \\ &H_a: \text{至少有一者 } \beta_j \text{ 不為零} \end{aligned} \tag{A.26}$$

而此檢定的統計量為 F 統計量如公式(A.27)所示。

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)} \tag{A.27}$$

上式的分子為「解釋變異」，分母則是「殘差變異」。這樣的思維與調整後 R 平方相當接近，兩者除了皆考慮變異的自由度外，分子的部分皆為「解釋變異」。不同的地方在於調整後 R 平方的分母為「整體變異」而非「殘差變異」，因此 F 統計量相較調整後 R 平方對於特徵個數更為敏感。一般來說，「F 統計量」常被用以作為「特徵挑選」的指標，而「調整後 R 平方」則是常被用以說明「模型的解釋力」。

**紅酒品質案例：模型評估**

承續上述紅酒數據案例，我們對紅酒品質與酒精度假設一個線性模型，並進行迴歸係數的估計與統計檢定，最後我們將分析模型的配適度，結果如表 A.2。首先，殘差標準差的數值代表著模型對新數據的預測有 0.7865 的誤差；其次，R 平方與調整後 R 平方約莫可解釋 0.193 的數據變異，代表依舊有一大部分的變異是無法用既有特徵解釋的。最後，F 統計

量在統計上是顯著的，也代表模型至少有一迴歸係數不為零。

表 A.2　線性迴歸模型的配適結果

| 指標 | 數值 |
|---|---|
| 殘差標準差 | 0.7865 |
| R 平方 | 0.194 |
| 調整後 R 平方 | 0.193 |
| F 統計量 | 383.6 （p-value<0.0001***） |

### A.2.4.2　殘差分析

除了分析模型的配適度外，迴歸分析的最後一個步驟是檢驗隨機誤差的假設。若假設滿足了，我們才會認為這個線性模型是可用的，若不滿足，則代表著在殘差中藏著未被找出的潛在效應，需修正模型假設或模型設定（model specification）。理論上，我們期望殘差為隨機發生的誤差，這是迴歸分析中相當重要的思維。前述介紹了線性迴歸的三大假設分別為「**同質性**」（變異數相等）、「**獨立性**」以及「**常態分配**」，而檢驗的方法除了視覺化殘差外，也可對三大假設分別作出檢定，以下我們將分別介紹殘差視覺化與統計檢定的兩種方法。

1. **殘差視覺化**：殘差的視覺化是檢驗誤差三大假設最常使用的方法。首先，「同質性」假設的視覺化是將預測值放在 x 軸，殘差放在 y 軸，觀察殘差是否隨著預測值的大小而有不同的「樣型」（pattern）。如圖 A.10 為殘差常見的三種樣型，左圖殘差變異維持在一個固定大小的區間，是一致且符合殘差假設的；中圖殘差變異隨著預測值越大跟著越大，屬於殘差不同質的情形違反假設；而右圖殘差隨著預測值有非線性的樣型，同樣地違反假設。

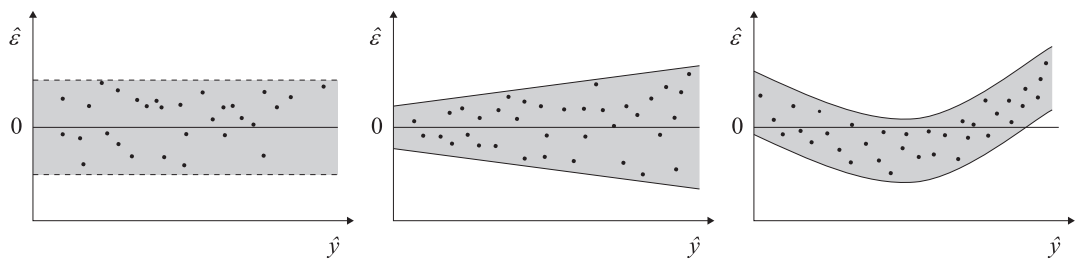


圖 A.10　殘差同質性視覺化

其次，「獨立性」假設的視覺化則是將觀測值順序（數據收集的時間）放在 x 軸，殘差放在 y 軸，觀察殘差是否在時間上存在某種關係而違反獨立性的假設。理論上，若殘差在時間上是獨立的，會呈現如圖 A.11 所示，同樣地，若有其他的樣型則代表違反了此假設。

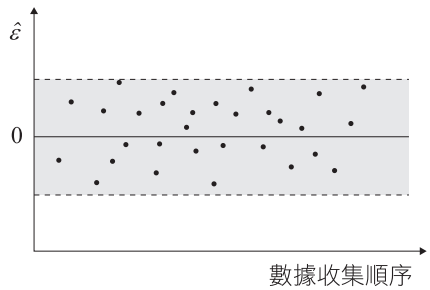


圖 A.11　殘差獨立性視覺化

最後，「常態分配」假設的視覺化則是將標準常態分配的「分位數」（quantile）放在 x 軸，數據標準化後的分位數放在 y 軸，這種圖被稱為「Q-Q 圖」（quantile-quantile plot），如圖 A.12 的右圖所示。若數據近似常態分配時，數據的分位數會與常態分配的相同，因而落於斜直線上；當數據為一個「右偏」（right-skewed）的分配時，則我們可從圖中看到其「Q-Q 圖」右上的分位數會往上偏，這是由於當我們在 x 軸上的同一點時，對到 y 軸實線代表著常態分配應處在的位置，然而數據右偏使得其對應到的分位數較大，而「左偏」（left-skewed）的分配可同理推得。因此藉由「Q-Q 圖」我們能觀察殘差與一常態分配的近似程度，判斷是否違反常態假設。

2. **殘差統計檢定**：假設檢定相較於殘差的視覺化更有實質統計上嚴謹的判定，以下我們將分別介紹三大假設所對應到的虛無對立假設與檢定方法。首先，「同質性」檢定的虛無對立假設如公式(A.28)所示，

$$\begin{aligned} &H_0\text{: 變異數為常數} \\ &H_a\text{: 變異數不為常數} \end{aligned} \tag{A.28}$$

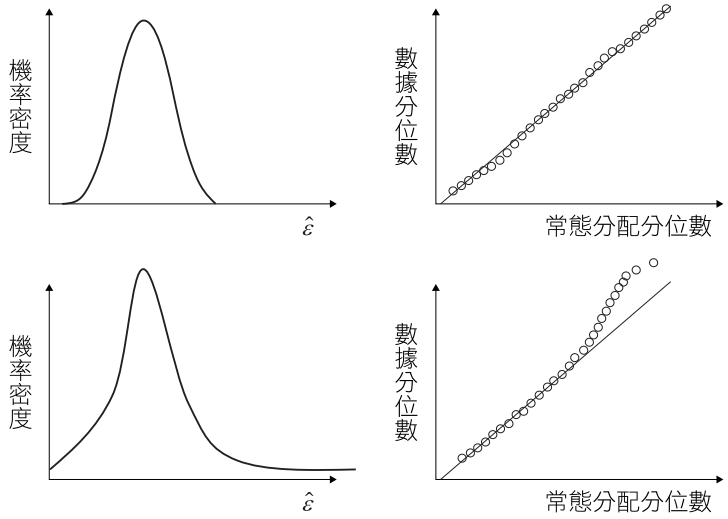


圖 A.12　殘差常態性視覺化

而檢定方法一般可使用無母數的 Levene's 檢定以及基於常態假設的 Breusch-Pagan 檢定。

其次，「獨立性」檢定的虛無對立假設如公式(A.29)所示，

$$\begin{aligned} H_0&: \rho = 0 \\ H_a&: \rho \neq 0 \end{aligned} \tag{A.29}$$

$\rho$代表殘差間的一步相關($\varepsilon_t$ and $\varepsilon_{t+1}$)，而檢定方法一般可使用 Durbin-Watson 檢定。

最後，「常態分配」檢定的虛無對立假設如公式(A.30)所示，

$$\begin{aligned} H_0&: \text{樣本服從常態分配} \\ H_a&: \text{樣本不服從常態分配} \end{aligned} \tag{A.30}$$

而檢定方法一般可使用 Shapiro-Wilk 檢定。對於統計檢定有興趣的讀者，可參閱相關書籍（陳順宇，2009；Montgomery et al., 2012）。

**紅酒品質案例：殘差分析**

承續上述紅酒數據案例，在評估模型的配適度後，我們最後進行模型的殘差分析，結果如圖 A.13 與表 A.3 所示。首先，從殘差的視覺化可判斷，殘差變異有大有小的情形，可能違反同質性假設，而依觀測值排序的殘差無特殊樣型，未違反獨立性假設，而從「Q-Q 圖」可發現殘差的分配兩邊均有厚尾的情形，可能為一個高狹峰分配違反常態分配的假設。其次，從三個假設的統計檢定結果中，僅有獨立性檢定 p-value 相當大，而未拒絕虛無假設的殘差獨立性假設，其餘的同質與常態假設在統計上則有顯著的證據說明殘差不符合此二假設。

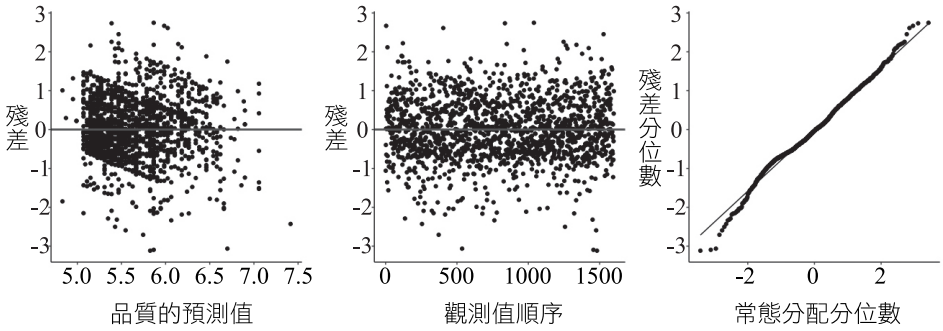


圖 A.13　殘差視覺化

表 A.3　殘差統計檢定結果

| 假設 | 統計檢定 | p-value |
|---|---|---|
| 同質性 | Breusch-Pagan 檢定 | <0.0001*** |
| 獨立性 | Durbin-Watson 檢定 | 0.492 |
| 常態分配 | Shapiro-Wilk 檢定 | <0.0001*** |

由本小節的模型評估後，可發現此紅酒品質與酒精度的線性模型還需有所調整，在模型的配適度上，調整後 R 平方$R^2_{adj} = 0.193$在解釋力上仍然有大幅的進步空間，可以加入更多特徵（多元線性迴歸）或是調整模型的複雜度（無母數迴歸或非線性迴歸）。此外，在殘差分析上，違反了同質性與常態分配假設，可由數據轉換或使用假設不同質的迴歸模型（例如加權最小平方法）。

此外，殘差分析不僅限於線性迴歸中，殘差分析的思維是將未被模型解釋或未被建構的關係進行深入的探討。不論是任何模型完成建模後，皆可透過殘差分析能輔助我們重新檢視數據的特性，不僅能改進特徵工程的處理（特徵與目標變數的轉換、極端值的判定等），更能協助領域專家從殘差的特性中找出尚未收集的特徵（鄭宇翔，2020；Lee et al., 2021）。

## A.3 多元線性迴歸

經由上述對「簡單線性迴歸」的詳細介紹後，本節將從單一特徵的「簡單線性迴歸」拓展至多個特徵的「多元線性迴歸」。我們將介紹多元線性迴歸的「模型假設與多元迴歸係數的估計」、「特徵間的多元共線性」以及「潛在的問題與討論」。

### A.3.1 模型假設與估計

#### A.3.1.1 模型假設與預測模型

多元線性迴歸同時考慮了多個特徵的線性關係，其數學模型如公式(A.31)所示。

$$y = \underbrace{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}_{\text{系統性的}} + \underbrace{\varepsilon}_{\text{隨機性的}} \tag{A.31}$$

$p$為特徵個數，因此，一共$p+1$個迴歸係數（$\beta_0, \beta_1, \ldots, \beta_p$）需要被估計。同樣地，我們可將目標值拆解成「系統性相關」與「隨機誤差」，而「系統性相關」又可拆解成多個特徵線性關係的加總，數學上為了更簡單的表示，一般會將公式以矩陣的形式表示如公式(A.32)。

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{A.32}$$

其中，

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$$

當我們使用訓練集估計出迴歸係數的估計值$\widehat{\boldsymbol{\beta}}$後，便可預測（估計）新的數據，而多元線性迴歸的估計式則如公式(A.33)所示。

$$\widehat{\boldsymbol{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} \tag{A.33}$$

### A.3.1.2 多元迴歸係數的估計

從簡單線性迴歸中，我們知道「最小平方法」與「最大概似估計」的結果是一致的，同理，多個迴歸係數的估計使用最小平方法的推導如下。首先，「殘差平方和」（residual sum of squares, RSS）如公式(A.34)所示。

$$\begin{aligned} RSS = S(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \varepsilon^2 = (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'\mathbf{X}'\boldsymbol{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \end{aligned} \tag{A.34}$$

接著，我們對所有的$\beta$進行偏微分求極值，如公式(A.35)所示。

$$\frac{\partial S(\beta)}{\partial \beta} = -2\mathbf{X}'\boldsymbol{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \overset{\text{set}}{=} \mathbf{0} \tag{A.35}$$

最後，整理上式後可得到$\hat{\beta}$的估計值與其估計的變異，如公式(A.36)與(A.37)所示。

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y} \tag{A.36}$$

$$\begin{aligned} \mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\right) &= \mathrm{Var}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}\right] = \mathrm{Var}(\boldsymbol{y})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned} \tag{A.37}$$

因此，同樣地能後續對所有迴歸係數進行各別的檢定，判斷線性關係是否成立。

## A.3.2 特徵間的多元共線性

「**多元共線性**」（multi-collinearity）一般發生於特徵彼此間有高度的線性相關，以資訊的角度則代表特徵間所貢獻的資訊量是重疊的。這情況會造成矩陣X中有兩行元素線性相依，使得最小平方法估計係數有「多重解」（multiple solution）的情況發生。換言之，估計的迴歸係數變異很大。通常可使用變異數膨脹因子（variance inflation factor, VIF）來偵測多

元共線性的狀況，判斷多元線性迴歸模型的特徵之間是否獨立。一般來說，當 VIF>5 時，迴歸模型有初步的共線性狀況；若 VIF 值>10，表示特徵間存在嚴重的共線性，會造成迴歸係數估計的變異過大。此時，可建議刪除該特徵、以脊迴歸（ridge regression）方式處理、或是進行特徵挑選或特徵轉換使彼此間獨立，詳細的觀念與解決方法請參閱章節「特徵挑選與維度縮減」。

### A.3.3 潛在的問題與討論

除了多元共線性的問題外，當我們配適線性迴歸還有可能會發生下列的潛在問題：

- **目標值與特徵間的非線性關係**（non-linearity of the response-predictor relationships）

  目標值與特徵間的非線性關係，可從殘差的視覺化中觀察到，這時若用線性迴歸配適會發生**模型設定錯誤**（model misspecification）。除了對特徵或目標值做非線性的轉換外，無母數與非線性迴歸則是配適非線性關係更進階的方法，我們將於章節「無母數迴歸與分類」中詳細介紹。

- **誤差彼此間不獨立，存在相關性**（correlation of error terms）

  誤差彼此間不獨立，一般發生於時間序列的數據中，在短時間內觀測值與觀測值間會有高度的相關。想像我們將收集數據的時間（週期）縮小至非常短，相當於在某時間點複製$k$筆相同的觀測值，對一個數據則是將樣本數乘上$k$倍，造成誤差估計的信賴區間縮小$1/\sqrt{k}$倍，高估了估計的精確度。而透過殘差分析中的視覺化與獨立性檢定可判斷出此問題，當發現誤差彼此間不獨立時，可改善數據收集的頻率與實驗設計，或使用特徵工程移除冗餘（高相關）的觀測值。

- **誤差不同質、不等變異**（non-constant variance of error terms）

  誤差不同質，發生在一些特殊的情形下，例如馬達的振動訊號將隨著老化的過程其振幅與變異都會跟著越來越大，同樣地，可透過殘差分析中的視覺化與同質性檢定可判斷出此問題。當發現誤差不同質時，可轉而使用不同質假設的迴歸模型（例如加權最小平方法），或是透過目標值的轉換（例如取對數 log），使得殘差滿

足同質性假設。

- **極端值與槓桿點的存在**（existence of outliers and leverage points）
  極端值發生於特徵固定於某個值時（$y|x$），某一樣本的真實值與預測值相距非常遠，一般發生於極少數的意外情形，例如收集機台的振動或聲音時堆高機從機台旁邊駛過，抑或是機台加工時的意外當機或感測器故障導致數據收集的異常等。另一方面，槓桿點則是發生於某一樣本的特徵數值與其餘樣本相距非常遠，因而該樣本在固定特徵數值下的樣本數會相當稀少，此情形並非異常狀況所產生的結果，而是一般發生的機率或比例很稀少，例如製造現場工程貨或實驗貨，某商品經過許久再次小批量生產、或是加工生產特別少量的特殊產品時所收集的數據等。而這些極端值與槓桿點通常是我們不納入模型考量的特殊情形，若將它們納入模型訓練時，可能造成模型的估計與配適度受到影響（特別是在線性模型中，但非線性模型則可能導致過度配適）。然而殘差的視覺化僅能協助我們判別極端值與部分槓桿點，因此我們需要更精確的指標協助我們判斷，其中包含了「庫克距離」（Cook's distance）與「標準化殘差」（standardized residual）。

## A.4 結語

線性迴歸是機器學習與數據科學方法的根本，也是許多進階模型的基礎模型（base model）。在實務上，雖然線性預測效果表現欠佳，但有相當高的解釋能力，對於變數間的相關性與因果關係能充分地說明，並提供數學方程式（closed form）來描述變數間的關係，這是相當有用的。事實上，從線性迴歸出發，也延伸許多各式各樣的模型與應用，例如在迴歸式中考慮高階次的因子，變數的二次項或兩兩交互作用；可作為時間序列分析（time series analysis）、實驗設計（design of experiments, DOE）、反應曲面法（response surface methodology, RSM）等進階方法學習的基礎；在特徵挑選上的應用，包含逐步迴歸（stepwise regression）與正則化（regularization）的特徵壓縮方法（shrinkage method）；增強迴歸預測能力的非線性與無母數方法，例如迴歸樹（regression tree）與多變量適應性迴

歸樣條（multivariate adaptive regression splines, MARS）；以及透過隨機組合數學運算子與函數，以搜尋樹的表達方式建構數學模型的符號迴歸（symbolic regression）等。在這些各式各樣的進階模型與應用下，更突顯出線性迴歸重要的數理基礎。

## 參考文獻

[1] 陳順宇（2009），迴歸分析，四版，三民書局。

[2] 鄭宇翔（2020），資料科學於 CNC 機台健康管理之摩擦力估測與補償，國立成功大學製造資訊與系統研究所碩士論文，台灣。

[3] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning*. 2nd edition, Springer.

[4] Lee, C.-Y., Hung, Y.-H., and Chen, Y.-W. (2021). Hybrid data science and reinforcement learning in data envelopment analysis. Book chapter edited in: Zhu, J. and Charles V. (Editor), *Data-Enabled Analytics: DEA for Big Data*, Springer.

[5] Montgomery, D. C., Peck, E. A., and Geoffrey Vining, G. (2012). *Introduction to Linear Regression Analysis*. 5th edition, Wiley.

[6] Montgomery, D. C., and Runger, G. C. (2003). *Applied Statistics and Probability for Engineers*. 3rd edition, John Wiley & Sons.

## 問題與討論

1. 試繪製出迴歸分析的推論流程，並簡述流程中每個步驟所進行的分析。
2. (a)試說明線性迴歸的四大假設為何？（提示：模型的假設，尤其是誤差項所服從的分配特性，以視覺化呈現尤佳）(b)如何確保建構的模型滿足這些假設呢？其殘差的視覺化與統計檢定為何？
3. (a)試說明線性迴歸為何使用最小平方法與最大概似估計法的迴歸係數估計結果是一致的；(b)其中估計出的迴歸係數的數值大小代表的含義又為何？（提示：解釋迴歸係數在特徵變化對目標值的影響）(c)而對於每個迴歸係數檢定的顯著程度又代表了什麼含義？

4. 在線性迴歸的模型評估中，作為衡量模型配適度的常見指標有調整後 R 平方 $R^2_{adj}$與 F 統計量，(a)試說明此二指標的定義為何；(b)分別的使用時機又為何？
5. 在線性迴歸中，(a)什麼是共線性？(b)試使用網路資源學習，如何衡量共線性的發生？
6. 在線性迴歸中除了檢驗四大假設外，可能存在著樣本為極端值或槓桿點：(a)試說明兩者的定義為何；(b)有哪些方法可以偵測此二者？
7. 在 UCI Machine Learning Repository 開放數據中包含了一個具有紅酒與白酒的酒品質數據（wine quality dataset，https://archive.ics.uci.edu/ml/datasets/Wine+Quality），其中紅酒數據一共包含了 1,599 個觀測值，而每個觀測值具有 11 個特徵以及作為目標值的紅酒品質。試著參考網路資源學習並撰寫程式，使用此數據回答下列問題：
   (a) 試將迴歸分析的結果呈現如下表，並試著解釋任一特徵與目標值之間的關係。

| | estimate | std. error | t value | p-value |
|---|---|---|---|---|
| intercept | | | | |
| fixed acidity | | | | |
| volatile acidity | | | | |
| … | | | | |
| alcohol | | | | |

R-squared: 0.xxxx, Adjusted R-squared: 0.xxxx

   (b) 試問配適一個線性迴歸模型是否合適？試說明原因。其中配適不佳的結果可能為何？
   (c) 基於上述結果，將上述特徵以 t value 進行排序後，哪些特徵的迴歸係數在統計上是顯著的呢（p-value<0.01）？t value 與 p-value 差異為何？而不顯著的特徵可能的原因為何？（提示：試檢驗共線性）
   (d) 試用統計檢定檢驗線性迴歸其中殘差的常態性（normality）、獨立性（independency）、同質性（homoscedasticity）的三大假設，若不符合假設該如何修正呢？
   (e) 試檢驗是否存在極端值與槓桿點？