Statistic Concept

teuton[1]

March 17, 2021

[1]Still learning

Contents

Chapter 1

基本統計介紹

1.1 什麼是統計學習?

通常使用 X 代表輸入變數，如果有多個變數則用 $X_1, X_2...$ 代表，用 Y 代表輸出變數，假設 $X = (X_1, X_2, \ldots, X_p)$ 代表一些已決定好的輸入變數，通常可以用一個未知但固定的函數 f 和隨機誤差項 ϵ 表示:

$$Y = f(X) + \epsilon \tag{1.1}$$

1.1.1 為何估計 f?

預測

假設你有一個預測函數 $\hat{Y} = \hat{f}(X)$ 而且預測函數和預測因子是固定的，那 Y 的差平方的期望值就是:

$$\begin{aligned} E(Y - \hat{Y})^2 &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= [f(X) - \hat{f}(X)]^2 + Var(\epsilon) \end{aligned} \tag{1.2}$$

其中 $[f(X) - \hat{f}(X)]^2$ 是可以調整的，$Var(\epsilon)$ 是自然誤差的變異數，是一個常數，訓練模型主要是為了降低前者以達到更好的預測結果

推理

除了預測結果，還需要對資料進行推理:

- 哪些變數會影響結果?
- 結果與各個變數的關係是什麼?
- 輸出與輸入的關係可以用線型函數描述嗎? 還是說兩者關係更複雜?

1.1.2 如何估計 f?

訓練組資料顧名思義是為了訓練模組而使用的資料，使模組產生一個接近未知函數 f 的函數 $\hat{f}$，如果用 x_{ij} 表示第 i 筆資料中變數 X_j 的值，y_i 則代表第 i 筆資料輸出結果，整個資料可以寫成一個矩陣: 訓練組資料

$$\begin{pmatrix} x_{11} & \cdots & x_{1p} & y_1 \\ x_{21} & \cdots & x_{2p} & y_2 \\ \vdots & \ddots & \vdots & \vdots \\ x_{n1} & \cdots & x_{np} & y_n \end{pmatrix}$$

模型有兩種方法取得預測函數 $\hat{f}$: 參數化方法以及非參數化方法 參數化方法 非參數化方法

參數化方法

分成二步進行:

1 假設函數的樣式或形狀，例如線性:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{1.3}$$

下一章會詳細介紹本方法，只要知道 $p+1$ 個變數就可以建構此線性模型，當然一開始的預測不見得是線性，有時候其他形狀可以的到更好的預測結果

1. 接著將訓練組資料代入等式，在這個例子裡，目標是找到適當的參數使得

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{1.4}$$

這裡最常見的作法是最小平方法

注意誤差值過低可能導致過度套入，進而影響實際表現 過度套入

非參數化方法

在圖形不要太粗糙或搖晃、但又能盡量貼和訓練組資料的前提下，估計出 f 的大致圖形，或使用數學函數逼近

1.1.3 預測精度與模型可解釋性之間的權衡

通常兩者不可兼得

1.1.4 監督式學習與非監督式學習

監督式學習就像是做選擇題練習，每筆訓練組資料的輸出都有對應值 (標準答案)，模型的目標是使預測結果盡量逼近真實輸出結果 (也就是遇上問題，回答要盡量正確)，例如氣溫預測；非監督式學習的訓練組資料沒有輸出值 (沒有正確答案)，模型要自行從訓練組資料中歸納出規則，進而輸出，例如鳶尾花分類 監督式學習 非監督式學習

1.1.5 回歸與分類問題

1.2 評估模型準確性

1.2.1 測量擬合質量

用來估計模型效能最常見的數值就是均方誤差 **(MSE)** : 均方誤差 (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \tag{1.5}$$

其中 $\hat{f}(x_i)$ 指的是第 i 筆資料的預測輸出，如果 MSE 越小，表示模型預測結果與真實輸出越接近，MSE 分成兩種: 訓練 MSE 和測試 MSE，分別是使用訓練組資料和測試組資料所得到的 MSE，降低訓練 MSE 相對容易，由於模型不知道測試組資料的結果，測試 MSE 難以降低

1.2.2 偏差-變異折衷

1.2.3 分類設定

如果預測的不是數字，而是某個標籤 (例如晴天雨天、蘋果番茄...)，則改用錯誤率來評估模型: 錯誤率

$$\frac{1}{n}\sum_{i=1}^{n}I(y_i \neq \hat{y}_i) \tag{1.6}$$

這裡 $\hat{y}_i$ 指的是第 i 筆資料的預測輸出標籤，而 $I(y_i \neq \hat{y}_i)$ 的定義是:

$$I(y_i \neq \hat{y}_i) = \begin{cases} 1, & \text{如果 } y_i \neq \hat{y}_i\text{，也就是預測結果錯誤} \\ 0, & \text{如果 } y_i = \hat{y}_i\text{，也就是預測結果正確} \end{cases}$$

貝氏分類

貝氏分類是建立在條件機率上，觀察特定原始資料 x_0 然後計算每個類別機率: 貝氏分類

$$Pr(Y = j|X = x_0) \tag{1.7}$$

選出適當的輸出類別 j 使上面等式有最大值，對每筆資料操作，就能把原始資料分類成各種區域，每一種對應到一個最有可能的輸出類別，由這方法可以得到最低的錯誤率，稱為貝氏錯誤率，因為每筆資料 $X = x_0$ 的錯誤率是 $1 - \max_j Pr(Y = j|X = x_0)$，貝氏錯誤率的公式為: 貝氏錯誤率

$$1 - E(\max_j Pr(Y = j|X = x_0)) \tag{1.8}$$

K-近鄰分類

但是在現實的資料中無從得知條件隨機分布，貝式分類法不可行，有些方法嘗試估計條件隨機分布，**K-**近鄰分類就是其中一個，給定一正整數 K 和一觀察資 K-近鄰分類

料 x_0，選擇離 x_0 最近的 K 個觀察資料，這些資料形成一個集合，以 $\mathcal{N}_0$ 表示，這樣對每種輸出結果 j 的條件機率是:

$$Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j) \tag{1.9}$$

有這個就能套用貝氏分類法，K 值太高或太低都不好，找到一個適當的 K 值可以最高程度的降低測試誤差

Chapter 2

線性迴歸

2.1 簡單線性迴歸

對於簡單線性迴歸，輸入和輸出各只有一個: 簡單線性迴歸

$$Y \approx \beta_0 + \beta_1 X \tag{2.1}$$

β_0 稱為截距，β_1 稱為斜率，這些需要決定的數 (在本例，二個) 統稱為係數或數，藉由訓練資料可以得到這些係數，進而估計資料: 截距 斜率 係數 參數

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{2.2}$$

2.1.1 估計係數

為了使模型梗準確，最常見的方法是降低誤差的最小平方，$e_i = y_i - \hat{y}_i$ 稱為第 i 筆殘差，由此可以定義殘差平方和 **(RSS)**: 最小平方 殘差 殘差平方和 (RSS)

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2$$

或，根據定義:

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2 \tag{2.3}$$

最小平方法選擇 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 使 RSS 最小，根據微積分，係數的公式為:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned} \tag{2.4}$$

其中 $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ 以及 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$，也就是取樣本平均

2.1.2 評估係數估計的準確性

如果變數間關係可用線性模型表達，可寫成總體回歸線的格式: 總體回歸線

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{2.5}$$

自行由最小平方法定義出的直線稱為最小平方線 ，重複預測多次，真實平均數以 μ 表示，不同資料的平均數集合以 $\hat{\mu}$ 表示，這樣就可以計算其標準誤差 (以 $SE(\hat{\mu})$ 表示): 最小平方線 標準誤差

$$Var(\hat{\mu}) = SE(\hat{\mu})^2 = \frac{\sigma^2}{n} \tag{2.6}$$

其中 σ 是 y_i 於隨機變數的標準偏差，同樣的操作可以對模型係數做:

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{x^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right], \quad SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.7}$$

在這裡 $\sigma^2 = Var(\epsilon)$，通常未知但可以從資料估計，σ ，別名殘差標準誤差，可以用 $RSE = \sqrt{RSS/(n-2)}$ 公式估計，後續會解說

標準誤差可用於估計信心區間 ，例如 95% 信心區間代表未知參數有 95% 機率落在此區間範圍內，以線性迴歸而言，β_1 的 95% 信心區間形式大約為: 信心區間

$$\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1) \tag{2.8}$$

β_0 的信心區間則是

$$\hat{\beta}_0 \pm 2 \cdot SE(\hat{\beta}_0) \tag{2.9}$$

也可使用標準差對係數進行假設檢驗，最常見的假設檢驗包括零假設: 假設檢驗 零假設

$$H_0 : There\ is\ no\ relationship\ between\ X\ and\ Y \tag{2.10}$$

並且與替代假設進行比較: 替代假設

$$H_a : There\ is\ some\ relationship\ between\ X\ and\ Y \tag{2.11}$$

以數學的語言描述，對應到假設:

$$H_0 : \beta_1 = 0$$

對比

$$H_a : \beta \neq 0$$

但 β_1 與 0 要" 多靠近" 才能確定 H_0 是對的? 對此可以用 **t** 統計量 來估計，如果輸入與輸出沒有關係，這個就會接近 t 分布，誤差 $n-2$ 級: t 統計量

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \tag{2.12}$$

p 值計算 $\beta_1 = 0$ 的機率，也就是輸入與輸出沒有任何關係的機率，夠小的 p 值 (5% 或 1%) 表示 H_0 是錯的，也就是輸入與輸出有關係存在 p 值

In Table 2.1, a small p-value for TV indicates that we can reject the null hypothesis that $\beta_1 = 0$, which allows us to conclude that there is a relationship between TV and sales.

Table 2.1: coefficients of the least squares model for the regression

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	<0.0001
TV	0.0475	0.0027	17.67	<0.0001

(An increase of \$1000 in the TV advertising budget is associated with an increase in sales by around 50 units)

2.1.3 評估模型準確度

殘差標準誤差

由於 ϵ 的存在，就算有真實線性模型，也無法完美預測每個結果，殘差標準誤差 **(RSE)**，也就是 ϵ 標準偏差的估計，定義為: 殘差標準誤差 (RSE)

$$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.13}$$

R^2 統計量

R^2 統計量用來測量模型與資料的擬和度，由於這個值介於 0 到 1 之間，不會受到資料值大小的影響，越接近 1 代表模型越能解釋資料: R^2 統計量

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \tag{2.14}$$

其中 $TSS = \sum(y_i - \bar{y})^2$ 代表總平方和 ，也就是輸出的總變異數 總平方和

輸入與輸出的相關係數定義為: 相關係數

$$r = Cor(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.15}$$

在簡單線性迴歸的設置下，$R^2 = r^2$

2.2 多重線性迴歸

多重線性迴歸的模型形式為:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \tag{2.16}$$

這裡 X_j 代表第 j 個輸入變數，β_j 對應其係數

2.2.1 估計迴歸係數

與簡單線性迴歸類似，多重線性迴歸的預測為:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p \tag{2.17}$$

一樣使用最小平方法預測係數，只是數量有 p 個:

$$
\begin{aligned}
RSS &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2
\end{aligned} \tag{2.18}
$$

由於一次考慮多個輸入，數值可能會與簡單線性迴歸不同

2.2.2 一些重要的問題

1. 至少有一個輸入變數對預測輸出實用嗎**?**
   使用多變數的零假設:

   $$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

   與替代假設相比:

   $$H_a : at\ least\ one\ \beta_j\ is\ not\ zero$$

   這個假設檢驗可以由計算 **F** 統計量來驗證 (在 p 值相對小時也適用): F 統計量

   $$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \tag{2.19}$$

   此處 $TSS = \sum(y_i - \bar{y})^2$、$RSS = \sum(y_i - \hat{y})^2$，這些定義和簡單線性迴歸時相同，如果線性模組假設是對的，那可以證明:

   $$E\{RSS/(n-p-1)\} = \sigma^2 \tag{2.20}$$

   另外，如果 H_0 是真的:

   $$E\{(TSS - RSS)/p\} = \sigma^2 \tag{2.21}$$

   也就是說，如果輸入變數與輸出結果關係不大，F 會接近 1，如果 H_a 是真的，F 會比 1 大，也就是至少有一個輸入與輸出有關係，但是要注意 n 值對 F 可能會造成誤判，例如 n 值很大時，H_0 為真可能只是假象
   有時候只想測試特定 q 個輸入變數，對應到的假設檢驗為:

   $$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$$

   為了方便起見，把這 q 個輸入變數移到最後方、在不使用這 q 個輸入變數的情況下做了第二個線性模型，並假設其殘差平方和為 RSS_0，此時適當的 F 統計量為:

   $$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)} \tag{2.22}$$

2. 對於解釋結果有幫助的輸入變數有多少? 全部? 一部份?
   決定哪些輸入變數用來計算輸出，以用單一模型呈現，這叫變數選擇，一般的做法是嘗試各種變數選擇，並從中選出最佳模組，在輸入變數很多的時候，有三個方法協助選擇: 變數選擇

   (a) 向前選擇，從空模組(只有截距沒有輸入變數的模組) 開始，先做出 p 個不同變數的簡單線性迴歸模型，從中選出最低 RSS 的模組加入空模組，從剩下的模組中選一個加入後的 RSS 最低的模組，重複直到中止條件 向前選擇 空模組
   (b) 向後選擇，先建立包含所有輸入的多重線性迴歸模組，移除擁有最大 p 值的輸入 (因為它與結果的關聯最差)，重新計算模組後再移除一個，直到中止條件 向後選擇
   (c) 混合選擇，先使用向前選擇法建立模型，如果途中有輸入的 p 值超過一定程度，就把此輸入移除，重複直到模組的所有變數 p 值都比較小，而且加入任意輸入會有超標 p 值 混合選擇

3. 模組對資料的擬合度如何?
   最常見的二種模型擬和度測量數為 RSE 和 R^2，當輸入變數增加時，就算其與輸出的關係不大，R^2 總是會增加，通用 RSE 的定義是:

$$RSE = \sqrt{\frac{1}{n-p-1}RSS} \tag{2.23}$$

4. 給定一些輸入變數，我們應該預測什麼輸出? 預測結果多精準?
   除了前述的信心區間可用於預測結果，預測區間 也可，它比信心區間長，預測區間是針對特定條件，信心區間則是取平均 預測區間

2.3 迴歸模型的其他考量因素

2.3.1 定性預測因子

雖然多數輸入是數值，但還是有些許輸入不是，例如性別、國籍...

只有二種可能性的定性預測因子

如果要針對性別做研究，可以先把它轉成虛擬變量，創建一個新變數來代表: 虛擬變量

$$x_i = \begin{cases} 1 & \text{如果第 } i \text{ 人是女性} \\ 0 & \text{如果第 } i \text{ 人是男性} \end{cases} \tag{2.24}$$

然後將此變數當成預測因子帶入迴歸模型:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{如果第 } i \text{ 人是女性} \\ \beta_0 + \epsilon_i & \text{如果第 } i \text{ 人是男性} \end{cases} \tag{2.25}$$

選擇的虛擬變量不同，得到的係數解釋方法，根據代入的迴歸式也不同

超過二種可能性的定性預測因子

在這種情況下，要建造更多虛擬變量:

$$x_{i1} = \begin{cases} 1 & \text{如果第 } i \text{ 人是亞洲人} \\ 0 & \text{如果第 } i \text{ 人不是亞洲人} \end{cases} \tag{2.26}$$

以及第二個:

$$x_{i2} = \begin{cases} 1 & \text{如果第 } i \text{ 人是高加索人} \\ 0 & \text{如果第 } i \text{ 人不是高加索人} \end{cases} \tag{2.27}$$

再把這些變數加入迴歸式以獲得模組:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{如果第 } i \text{ 人是亞洲人} \\ \beta_0 + \beta_2 + \epsilon_i & \text{如果第 } i \text{ 人是高加索人} \\ \beta_0 + \epsilon_i & \text{如果第 } i \text{ 人是非裔美國人} \end{cases} \tag{2.28}$$

2.3.2 線性模組延伸

線性模組建立在兩個假設上: 獨立性 與線性，前者假設變動其中一個輸入不會影響到其他輸入，而後者假設無論變數值多少，每次往上加一單位的特定變數對結果的影響量始終相同

獨立性

線性

移除獨立性假設

其中一個做法是增加互動型:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon \tag{2.29}$$

等級原則聲稱如果要增加一互動型至模組，就算那兩個變數的 p 值偏高，還是要加入

等級原則

如果互動型的其中一項是定性預測因子:

$$balance_i \approx \beta_0 + \beta_1 \times income_i + \begin{cases} \beta_2 + \beta_3 \times income_i & \text{如果第 } i \text{ 人是學生} \\ 0 & \text{如果第 } i \text{ 人不是學生} \end{cases} \tag{2.30}$$

非線性關係

這種情況使用多項式迴歸，例如二次:

多項式迴歸

$$mpg = \beta_0 + \beta_1 \times horsepower + \beta_2 \times horsepower^2 + \epsilon \tag{2.31}$$

雖然是多項式，但這種回歸還是一種線性模組，因為多出的次方項可以當成新變數使用

2.3.3 潛在問題

對一資料使用線性模組擬和可能會發生許多問題，以下這些最常見:

1. 非線性的因果關係
   殘差圖適合用來偵測這個問題，如果是簡單線性迴歸，畫出 $e_i = y_i - \hat{y}_i$ 和 x_i 的對應關係，如果是多重線性迴歸則改為 e_i 和 $\hat{y}_i$，如果圖形能看出某種規則那就表示此線性模組的某個觀點有問題，嘗試新增 $\log X$ 、$\sqrt{X}$ 或 X^2 等來調整

   殘差圖

2. 誤差項的相關係數
   一個線性迴歸模組的重要假設是誤差項彼此之間沒有關係，如果有，估計標準誤差通常會低於真正的標準誤差，進而導致信心和預測區間比真實情況來的短，p 值也會被低估

3. 誤差項的非常數變異數
   例如，如果輸出越高誤差項變異數就越高，其中一個可能的解決方法是把輸出結果用凹函數變形，像是 $\log Y$ 或 $\sqrt{Y}$，如果第 i 項輸出對應到平均 n_i 個原始觀察，且這些原始觀察的變異數為 σ^2，彼此不相關，那他們的平均變異為 $\sigma_i^2 = \sigma^2/n_i$，在這情形下使用加權最小平方法，在這例子加權數 $w_i = n_i$

   加權最小平方法

4. 離群值
   離群值是指某個點 y_i 離模組的預測值太遠，可能是數據收集過程中觀察記錄不正確，離群值可能導致 RSE 過高，進而導致信心區間和 p 值失真，R^2 也會受影響，殘差圖可用來辨識離群點，但辨識離群點需要一個標準，為此可以改用學生化殘差來處理，將每個殘差 e_i 除以其估計標準誤差即可得到，此時通常絕對值大於 3 的資料就是離群點，可以選擇移除離群點，但有時離群點可以代表模型的不足之處

   離群值

   學生化殘差

5. 高槓桿點
   高槓桿點是針對不正常的資料 x_i，對於最小平方線的影響比離群值高，計算槓桿統計值 可以協助篩選，以簡單線性迴歸為例:

   高槓桿點

   槓桿統計值

   $$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \tag{2.32}$$

   針對多變數的槓桿統計值也有公式可以計算，這值總是介於 $1/n$ 和 1 之間，而所有觀察資料的平均槓桿值一定是 $(p+1)/n$，若有資料槓桿統計值大幅超過此值，則可以懷疑其為高槓桿點

6. 共線性共線性代表至少有二個變數有很高的相關性，難以從共線性分離出個別變數對輸出的影響程度，由於共線性會降低迴歸係數的估計準確度，估計係數標準差會上升，降低 t 統計量，進而降低成功偵測到非零係數的機率，查看相關矩陣是發現共線性的一個簡單方法，有時候共線性發生在三或更多變數，但兩兩之間沒有共線性，稱之為多重共線性，對這個比較好的辨識方法是方差膨脹因子 **(VIF)**，最小值是 1，也就絕對沒有共線性，

   共線性

   多重共線性

   方差膨脹因子 (VIF)

超過 5 或 10 表示可能會導致問題的共線性:

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}} \tag{2.33}$$

其中 $R^2_{X_j|X_{-j}}$ 是使用 X_j 以外的所有變數，針對 X_j 進行迴歸所得到的 R^2，$R^2_{X_j|X_{-j}}$ 越接近 1，共線性越明顯，共線性有二個處理方法，一是移除其中一個共線性的變數，二是把這些變數集合成單一變數，例如將它們標準化後取平均

2.3.4 範例: 市場分析

1. 廣告與銷售有幫助嗎?
   將電視、廣播和報紙作為輸入，針對銷售進行多重線性迴歸模型分析，並檢驗零假設 H_0，在本例裡對應到 F 統計值的 p 值非常小，顯示廣告和銷售有關係

2. 關係有多強?
   兩種方法估計，RSE 估計總體迴歸線輸出的標準差 (本例為 1681 單位，輸出平均為 14022，得到誤差百分率約為 12%)；R^2 統計數紀錄預測變量解釋的輸出變異百分比 (本例為 90%)

3. 哪些媒體對銷售有貢獻?
   對每個預測變量的 t 統計值檢驗 p 值，夠低的 p 值對應的預測便量和輸出有關係 (本例中為電視和廣播)

4. 每種媒介對銷售的影響有多大?
   使用 $\hat{\beta}_j$ 的標準誤差可以建立 β_j 的信心區間 (本例中電視和廣播的 95% 信心區間比較窄而且不接近零，證明了這些媒體與銷售有關；但報紙的信心區間包含零，統計上比較不重要)，共線性會造成非常寬的標準誤差 (本例中，報紙的信心區間是受共線性影響的嗎? 三個輸入變數的 VIF 都只比 1 大一點點，所以無法證明這個假設是對的)，為了評估個別變數對輸出的關係，對他們使用簡單線性迴歸

5. 預測未來銷售有多準確?
   將資料代入得到的多重現行回歸模組就能預測，如果要得到單項輸出則使用預測區間；如果要得到長期平均的預測結果則使用信心區間。

6. 關係是線性的嗎?
   可使用殘差圖來辨識非線性，如果是線性，那殘差圖的理想值應該接近常數函數

7. *Is there synergy among the advertising media?*
   A small p-value associated with the interaction term indicates the presence of non-additive relationships.

Bibliography

[1] An introduction to statistical learning *with Applications in R*
Gareth Jams,
Daniela Witten,
Trevor Hastie,
Robert Tibshirani

[2] TOP 100 R TUTORIALS : STEP BY STEP GUIDE
https://www.listendata.com/p/r-programming-tutorials.html