Multiple Linear Regression

A. 複回歸模式(Multiple Regresion Model)

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

其中 y 為應變數(response variable)，$x_1 \cdots x_k$ 為 k 個自變數(Independent variable)，β_0為截距(intercept)，$\beta_1 \cdots \beta_k$為迴歸係數(coefficients of regression)，ε和為誤差變數或稱為誤差項(error term)。

B. 模式假設(Model's Assumptions)

a. 線性關係(Linearity)
   I. 依變數和自變數之間的關係必須是線性，也就是說，依變數與自變數存在著相當固定比率的關係，若是發現依變數與自變數呈現非線性關係時，可以透轉換(transform)成線性關係，再進行迴歸分析。

b. 誤差項的變異數相等(Homoscedasticity)
   I. 自變數的誤差項除了需要呈現常態性分配外，其變異數也需要相等，變異數的不相等(heteroscedasticity)會導致自變數無法有效的估計應變數，可以使用 Levene test，來測試變異數的一致性，當變異數的不相等發生時，我們可以透過轉換(transform)成變異數的相等後，再進行迴歸分析。

c. 常態性(Normality)
   I. 若是資料呈現常態分配 (normal distribution)，則誤差項也會呈現同樣的分配，當樣本數夠大時，檢查的方式是使用簡單的 Histogram (直方圖)，樣本數較小時，檢查的方式是使用 normal probability plot (常態機率圖)。

d. 誤差項的獨立性(Independence of error)
   I. 自變數的誤差項，相互之間應該是獨立的，也就是誤差項與誤差項之間沒有相互關係，否則，在估計迴歸參數時，會降低統計的檢定力，我們可以藉由殘差(Residuals)的圖形分析來檢查，尤其是與時間序列和事件相關的資料，特別需要注意去處理。

e. 沒有多元共線性(Lack of multicollinearity)
   I. 檢視自變項間是否有多元共線性（multicollinearity）的問題，也就是自變項間是否有高度相關的問題。如果自變項間高度相關的話，會影響到對迴歸係數之假設測定。最簡單的方式就是，先以簡單迴歸或 Pearson 相關，以每一個自變項個別與依變項跑相關，假設我們有五個自變項，當在跑簡單迴歸的時候，其迴歸係數都是正的，可是當我們五個自變項聯合預測依變項的時候，卻

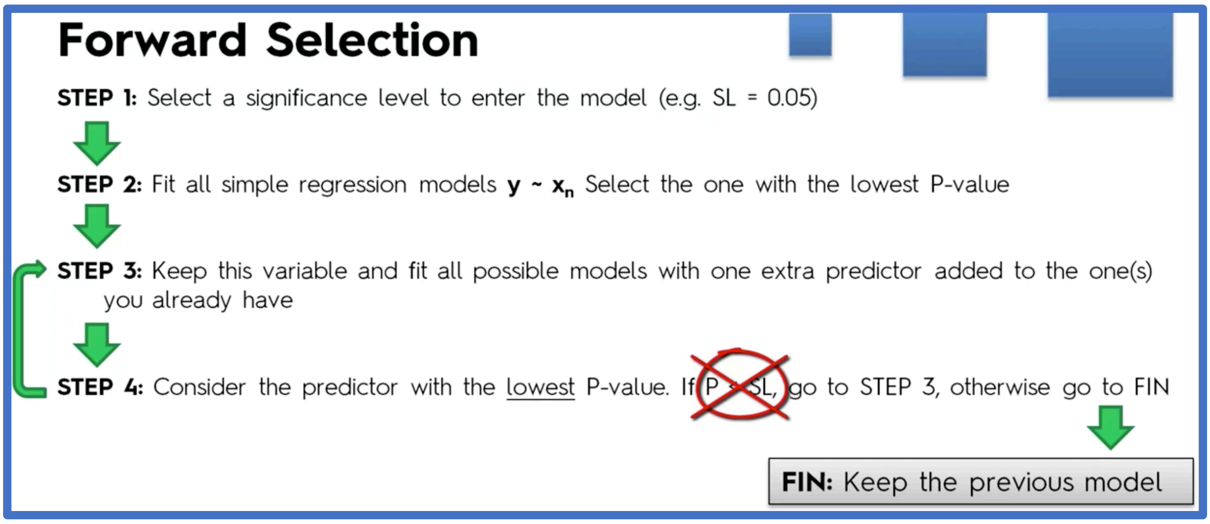
有迴歸係數變成負數，此時就可知道自變項中存在著足以導致錯誤結論的共線性。

C. 虛擬變數(Dummy Variable)

a. 使用時機
   I. 如果自變項是類別的變項，我們可以將這些類別一一建構成為虛擬變項。
b. 模型帶入原則
   I. 依照類別數目（k），我們只需建構 k-1 個虛擬變項即可。
c. 範例
   I. 如性別有兩類，因此我們只需建構一個「男性」的虛擬變項。如果受訪者為男性，則其「男性」變項為 1，如為女性，則其「男性」變項為 0。
   II. 如果一個類別變項有四類，如台灣地區別是分成北、中、南、東等四區，則我們可將此類別變項建構成「中部」、「南部」及「東部」等三個虛擬變項。當受訪者是在北部時，其在此三虛擬變項的值會都是 0。
d. 虛擬變數陷阱(dummy variable trap)
   I. 產生原因
      1. 將所有產生的虛擬變數全放入模型中
   II. 影響
      1. 產生變數間完全線性。
      2. X 產生 Singular Matrix，矩陣無法解出 $\boldsymbol{\beta} = (X^T X)^{-1} X^T Y$。
      3. 參考：http://www.algosome.com/articles/dummy-variable-trap-regression.html

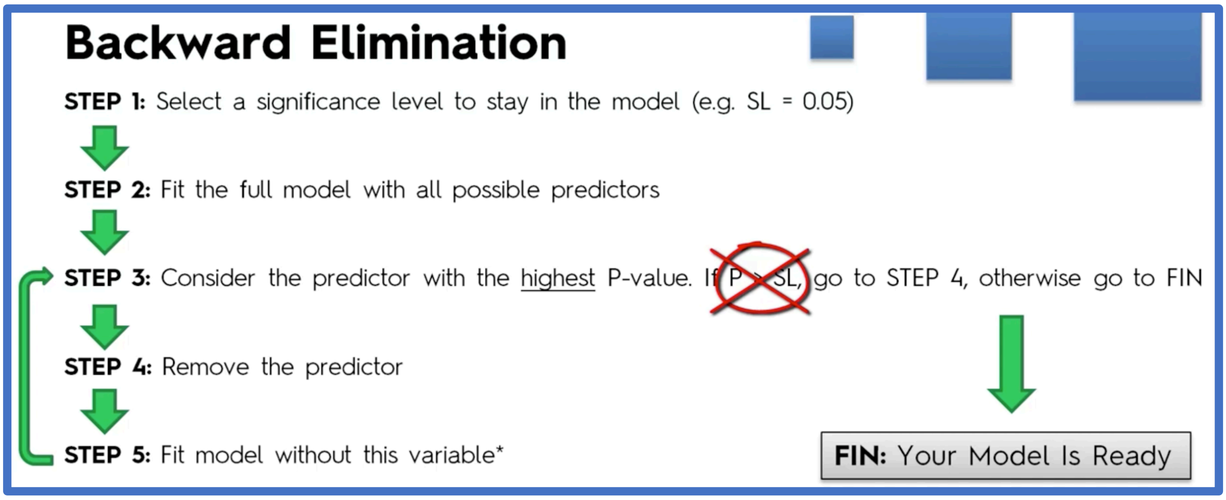
D. 選取因子方式

a. 向前選取法(Forward selection)
   I. 向前選取法是逐一增加自變項，直到任何一個自變項的額外貢獻量已經沒有統計意義。



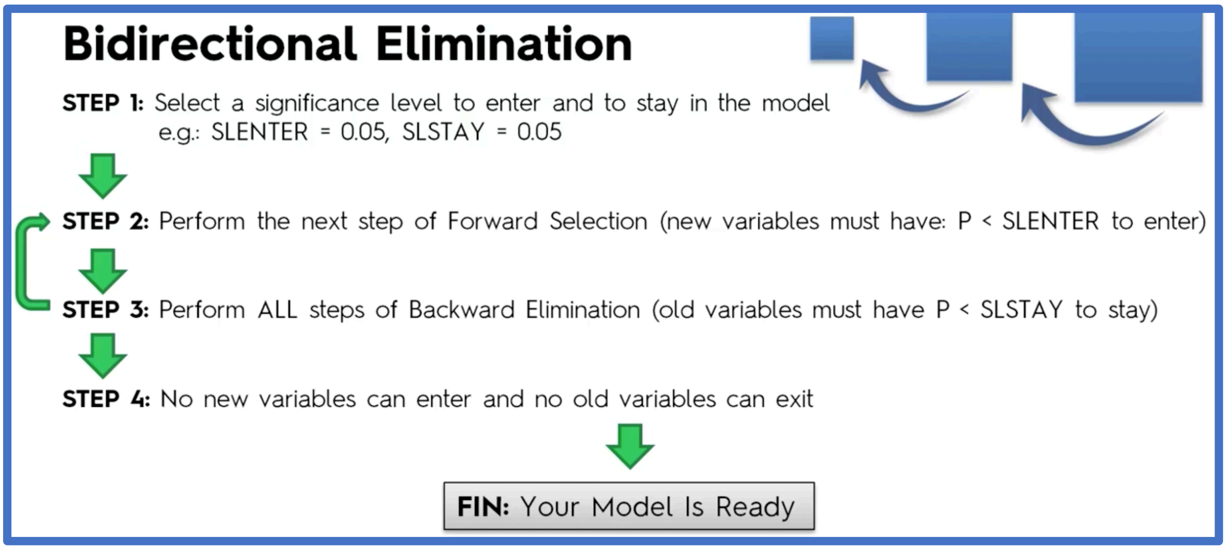
b. 向後選取法(Backward selection)

I. 向後選取法則是逐一剔除自變項，直到當剔除任何一個自變項時，模式會損失過多的解釋力，此時即停止篩選變項。

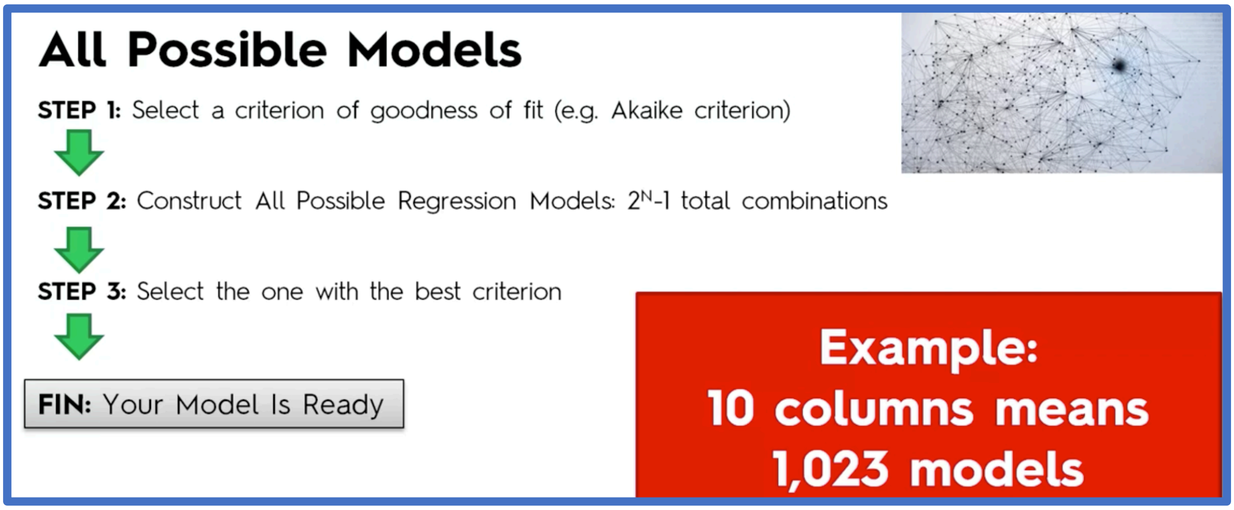


c. 雙向選取法(Bidirectional selection)

I. 逐步選取法是同時結合了向前選取及向後選取兩種方法，最大不同處是逐步選取法可以允許被排除的變項又被選進模式，也允許被選進的變項最後又被模式排除。



d. 所有可能 Model(All Possible Models)



I. 計算每一可能模型

$$C_{p^*} = \frac{SSE_{p^*}}{MSE_T} - (n - 2p^*)$$

MSE_T = 含所有自變數模型之殘差均方

n = 樣本大小

p = 模型中自變數個數

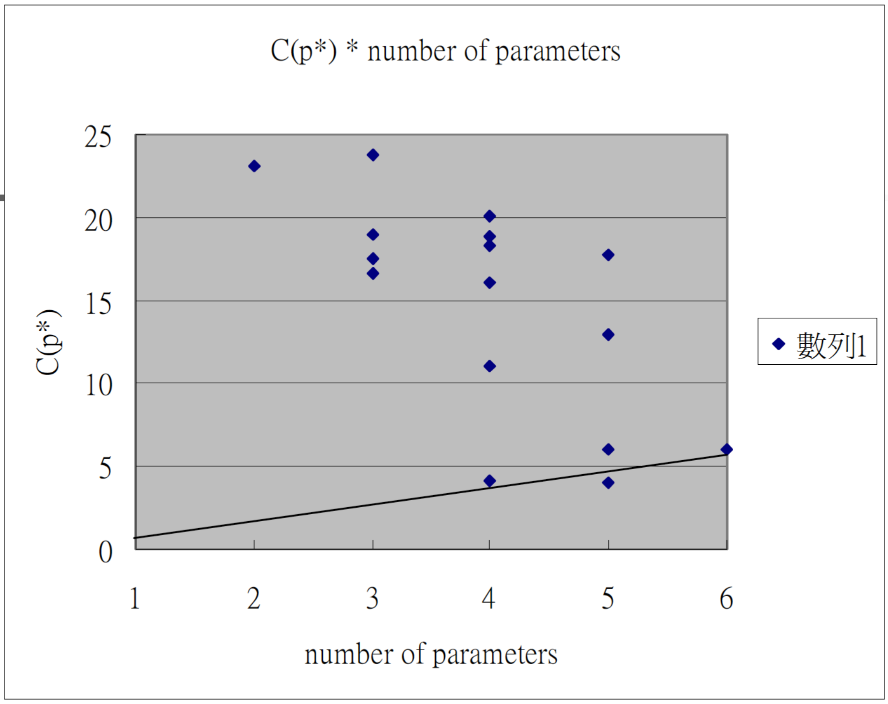
p^* = 模型中參變數個數 (p+1)

SSE_{p^*}= 含 p 個自變數模型之殘差平方和

II. 繪製(p^*, C_{p^*})之散佈圖

III. 找出最靠 C_{p^*}= p^* 之 (p^*, C_{p^*})且C_{p^*}值盡可能小

number of parameters	C(p*)
2	23.0929
3	16.6803
3	17.5332
3	19.0187
3	23.7484
4	4.0843
4	11.0255
4	16.1047
4	18.2648
4	18.8284
4	20.1014
5	4.0071
5	6.0106
5	12.8956
5	17.7694
6	6



C_{p*} 方法之最佳模似乎爲 {X1, X3, X5} 或 {X1, X2, X3, X5}。