# Lecture 4 To Do

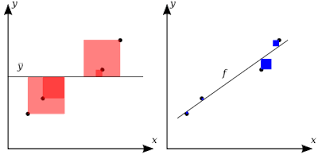
Google Meet [會議google Meet <https://meet.google.com/qjv-fvrx-rka>

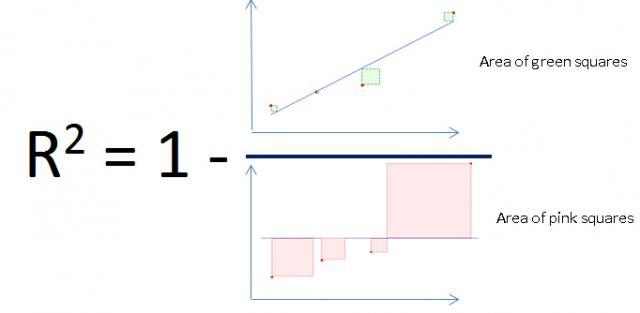
## Review 上次內容

* 線性回歸,我們介紹什麼是Excel操作並用幾個例子說明線性回歸的應用, 感測器較正
* 在用Python 來實現=>google Coble
* ML 實作方法熟悉一下follow CRISP-DM (跨領域資料分析作業流程) 方法論來建立ML 模型
  + Step 1: import library, load data
  + Step 2: prepare “feature” All-in今天多元回歸 (multilinear regression) 會著重select feature 方法, 還有包括一些前處理 (missing data=> imputation, normalization, data type 例如 categorical data🡺 onehot encoding, Label\_Encoding,資料型別的 transformation)
  + Step 3: Build model
  + Step 4: Evaluate model
    - MSE ,MAE, R2

MSE

R2:





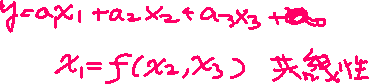
* + Step 5: Deploy 應用 predict new value

1. **多元線性回歸**



**問題特色:**

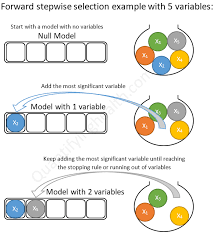
1. Linearity (因變數xi 跟y之間有線性關係)
2. Homoscendasticity (每一個xi 方差都一樣)
3. Mulivariate normality (每一個xi normal distribution)
4. Independence of error (誤差獨立)
5. Lack of multicollinearity (不可有共線性)

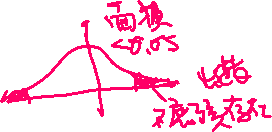


實作上沒辦法一一檢驗, 所以用下面這種工程方法來挑選適合的特徵組

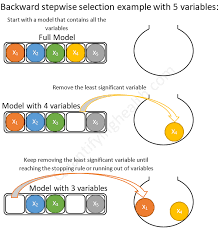
1. All-in (何時可用🡺 你是專家, 滿足上面五個要求, Con: 當特徵很多時 100個以上, 擬就算是專家, 你也很難完全保證)
2. Stepwise Selection (greedy method)
   1. Backward selection (最常用)

篩選標準 p-value (p-value < 0.05 顯著, <0.01非常顯著) 出錯的機率





* 1. Forward selection



* 1. Bidirection selection

1. **Score selection** (feature selection optimization)

假設有 10 特徵可選, 共有2^10=1024 特徵集的選擇

暴力破解可找到最佳 (sklearn Grid Serarch)

可是壞處就是沒時間因此要訴諸所謂的**搜尋演算法**

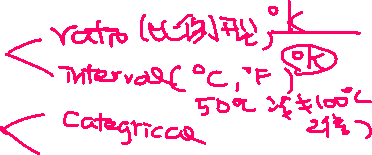
進階的搜尋演算法使用Meta-heuristic

參考 聯合大學陳士杰 <http://debussy.im.nuu.edu.tw/sjchen/ML_final.html>

主要有名: Hill climbing algorithm, GA, PSO, SA, TS…



1. **data 型態**



Nominal data 需要做

OneHotEncoding (1 to n ) 產生數個DUMMY VARIABLE (特徵)

這時又產生另一個問題 Dummy variable trap 陷阱(如果n個都用的話一錠最後一個是線性相依 xn=1-x1-x2-….-xn-1)

Solution🡺新產生n 個只能放n-1進去式子裡面

Label Encoding 動作 (1 to 1)== Label

看p-value sample code

https://stackoverflow.com/questions/27928275/find-p-value-significance-in-scikit-learn-linearregression

**from** **sklearn.feature\_selection** **import** SelectFromModel