第一屆金融科技應用與創新校際競賽

AI@FinTech-2019

學生講座系列二

概率 (Probability)、線性與邏輯迴歸 (Linear & Logistic Regression)及 Lab Session

29th June, 2019

離散隨機變量、概率、期望和方差

1) 隨機變量 (Random Variable)

定義：任何變量其結果為隨機的

符號：**X** 表示隨機變量，**x** 表示可能的值。

在這裡，我們考慮的離散(discrete)隨機變量，只包括有限數量的可能結果。

例子：明天的天氣可能是晴天、多雨、多雲; 擲骰子的結果可以是{1, 2, ..., 6}中的任何數字；借款人在償還貸款之日的狀態可以是已償還，也可以是未償還（違約）。

反例：等待巴士的時間、一包糖的重量。此類屬於連續隨機變量，本課程不作考慮。

2) 概率分佈 (Probability Distribution)

離散隨機變量 X 的概率分佈是 X 的所有可能值及其概率：

$$f(x) = P[X = x]$$

概率分佈必須滿足的條件：

$$f(x) \geq 0 \quad and \quad \sum_{all\ x} f(x) = 1$$

(DP1a) 擲硬幣：
設 X 是擲硬幣的結果，這是隨機的。
其可能的值是正面(H)或反面(T)。
如果使用公平的硬幣，X 的概率分佈是什麼？

解： 樣本空間 Ω = {H, T}
由於使用公平的硬幣，得到正面或反面的機會是相同的
所以概率分佈是 P{H} = 0.5, P{T} = 0.5

(DP1b) 擲三次硬幣：
設 X 為擲三次硬幣產生正面的總數。
假設使用公平的硬幣。 H 的概率分佈是什麼？

解： 樣本空間 Ω = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}
由於使用公平的硬幣，得到每個結果的機會都是 1/8
X = 0 代表擲三次硬幣產生 0 次正面, 有{TTT}
X = 1 代表擲三次硬幣產生 1 次正面, 有{HTT, THT, TTH}

X = 2 代表擲三次硬幣產生 2 次正面, 有{HHT, HTH, THH}
X = 3 代表擲三次硬幣產生 3 次正面, 有{HHH}
所以概率分佈是 P(X=0) = P{TTT} = 1/8,
P(X=1) = P{HTT, THT, TTH} = 3/8,
P(X=2) = P{HHT, HTH, THH} = 3/8,
P(X=3) = P{HHH} = 1/8

亦可以用表列方式表達：

x	0	1	2	3
P(X=x)	1/8	3/8	3/8	1/8

3)　隨機變量的期望值(Expectation)。

離散隨機變量的均值，又稱期望值(Expectation)，是其中一種集中量數(measure of central tendency)。計算公式如下

$$\mu = \sum_x xf(x) = E[x]$$

(DP1c) 承上題，平均而言，擲三次硬幣出現多少次正面？

解：　期望值 E[X]=0 *1/8 + 1 *3/8 + 2 *3/8 + 3 *1/8=12/8=1.5

4)　隨機變量的方差 (Variance)及標準差(Standard Deviation)

方差是一種離散量數，用以描述X的可能值中（或資料中）各數值和中心值間之平均變化或差異程度 (deviation)。當X的可能值的分佈較廣時，則測量差異的數值也越大。

方差定義及計算

$$\begin{aligned}\sigma^2 = V[X] &= E[(X-\mu)^2] \\ &= \sum_x (x-\mu)^2 f(x) \\ &= \sum_x x^2 f(x) - \mu^2 \\ &= E[X^2] - (E[X])^2\end{aligned}$$

標準差

$$\sigma = \sqrt{\sigma^2} = [V(X)]^{1/2}$$

(DP1d) 承上錯，擲三次硬幣出現正面次數的方差及標準差是多少？

解： 我們已知 E[H] = 1.5。
計算 $\mathbf{E[X^2]}$:

$$\mathrm{E}[X^2] = \sum_{x=0}^{3} x^2 P(X = x) = 0^2 * \frac{1}{8} + 1^2 * \frac{3}{8} + 2^2 * \frac{3}{8} + 3^2 * \frac{1}{8} = \frac{24}{8} = 3$$

所以方差$\sigma^2 = \mathrm{E[X^2]} - (\mathrm{E[X]})^2 = 3 - 1.5^2 = 0.75$ 及
$\sigma = \sqrt{\mathrm{V(X)}} = \sqrt{0.75} = 0.8660$

(DP2) 投資例子:
假設你有 100 元用作投資，希望賺取一些回報。
現在有兩個投資計劃

計劃 A
有 0.1%機會賺取 5000 元
有 0.5%機會賺取 1000 元
有 99.4%機會賺取 0 元

計劃 B
有 30%機會賺取 20 元
有 20%機會賺取 10 元
有 50%機會賺取 4 元

你會選擇哪個投資計劃？
計算每個計劃的預期收益
計算每個計劃的方差

解： 計劃 A 的預期收益$= 5000 * 0.001 + 1000 * 0.005 + 0 * 0.994 = 10$元
計劃 B 的預期收益$= 20 * 0.3 + 10 * 0.2 + 4 * 0.5 = 10$元

計劃 A 的方差$= (5000^2 * 0.001 + 1000^2 * 0.005 + 0^2 * 0.994) - 10^2 = 29900$
計劃 B 的方差$= (20^2 * 0.3 + 10^2 * 0.2 + 4^2 * 0.5) - 10^2 = 48$

(如何選擇?)

5) **獨立隨機變量**

考慮兩個隨機變量 X 和 Y。如果知道一個隨機變量的結果不影響另一個隨機變量的概率分佈，則它們是獨立的。

例子：考慮兩個人分別擲硬幣。 兩次擲硬幣的結果是獨立的。

（想想反例？）

當考慮 n 個獨立的隨機變量時，它們總和的期望值和方差有以下關係：

$$E[X_1 + X_2 + \cdots + X_n] = E[X_1] + E[X_2] + \cdots + E[X_n]$$
$$V[X_1 + X_2 + \cdots + X_n] = V[X_1] + V[X_2] + \cdots + V[X_n]$$

(DP3) 隨機變量 X 的分佈如下:

$$f(x) = \begin{cases} kx, & x = 1,2,3,4 \\ 0, & \text{其他值} \end{cases}$$

(a) 找 k 的值。
(b) 表列 X 的分佈。
(c) 找 $P(X \leq 3)$

解：

(DP4) X的概率分佈如下：

x	1	2	3	4	5
f(x)	0.1	0.3	0.2	0.3	0.1

(a) 找E(X)。
(b) 找 $E(X^2)$。
(c) 找 Var(X)。
(d) 找X的標準差。
((a) 3, (b) 10.4, (c) 1.4, (d) 1.1832)

解：

線性迴歸

1) 基本統計知識

(i) **mean** $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

(ii) **sample variance** $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

population variance $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

(iii) **sample s.d.** $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

population s.d. $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

2) 統計學上，我們常常探討兩樣（或多樣）東西之間的關係。我們可抽取樣本及量化為數據，看看兩組數據有沒有關係。如果有關係，我們往往會追問它們有否線性關係，因此便會使用線性迴歸分析。

3) 所謂「迴歸」，就是探討一個變數對另一變數的影響。

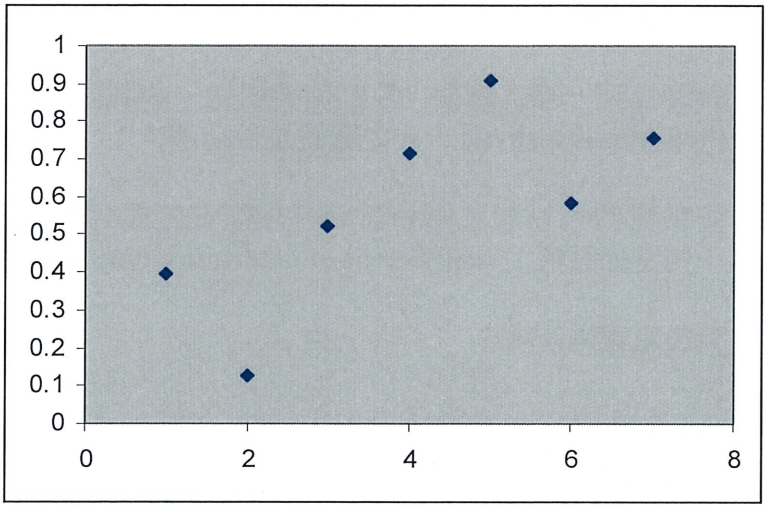
4) 模型構作

(I) 假設我們想研究 **x** 和 **y** 之間的線性關係，即想找參數(**parameter**) **α** 和 **β**使以下關係成立：

$$y = \alpha + \beta x$$

(II) 為找 **α** 和 **β**，我們需要收集一些 **x** 和 **y** 的抽本。假設我們找到了 **n** 對樣本：$(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)$

(III) 如果把這些樣本用 **xy-plane** 表示，我們會見到一些離散的點在圖上：



(IV) 找 α 和 β就相對於在圖中找一條「最好」的直線，使每點與該直線的距離最短。數學上，我們可用「最小二乘法」。

(V) 最小二乘法 **(Least Squares Method)**

如果把每對樣本(x_i, y_i)放入直線方程，應該會有一定的誤差ε_i：

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

如果把這些誤差的平方加起來，我們得出

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

所謂 **"Least Square"**，就是找出找 α 和 β，使 L「最小」。

由於解釋過程涉及微積分，就此略過，並只列出結果。

首先定義兩個項：

$$SS_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

$$SS_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

當中$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ 及$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ 。則 α 和 β 的估值如下：

$$\begin{cases} \hat{\beta} = \dfrac{SS_{xy}}{SS_{xx}} \\ \hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} \end{cases}$$

(VI) 其實找到一條直線後，我們尚要透過一些統計測試，如假設檢定 **(hypothesis testing)** ，才能驗證是否合理。

(VII) 如果我們可以想了解這個線性模型究竟是否適合，我們可看看「判定係數值」**(coefficient of determination)** 或 **R^2**：

首先定義兩個項：

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \ ;$$

及 $SS_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$

則 $R^2 = \dfrac{SS_{yy} - SSE}{SS_{yy}} = 1 - \dfrac{SSE}{SS_{yy}}$

R^2 的意思就是看看利用這個線性模型，x 能解釋 y 多少。R^2 為 0 至 1 之間的數字，如果 R^2=0.95，表示 x 能解釋 95%的 y，所以這樣線性關係頗合理。

5) **注意：即使我們能為 x 和 y 找到一線性關係，亦不表示我們找到一個因果關係 (causal relation) 。例如，我們相信鞋帶越長，智商(IQ)便越高。這是因為當人越長大／高，所需穿的鞋的尺碼亦增大，因而需要一對較長的鞋帶。但是，你相信自己馬上去更換一對較長的鞋帶後，能使你變得更聰明嗎？**

(LR1) 一間家品店想研究宣傳 (advertising) 商品是否有助增加利潤 (sales revenue)。它抽取了最近五個月的數據 ：

Month	Advertising Expenditure x (in hundreds of dollors)	Salles Revenue y (in thousands of dollors)
1	1	1
2	2	1
3	3	4
4	4	4
5	5	8

(a) 為兩組數據找出最優擬合線 (best-fit line)。

(b) 找出該線的判定係數值。

(c) 利用(a)估算當這間家品店花 6 百元作為宣傳費時的利潤。

解(a)：$\sum_{i=1}^{5} x_i = 15$ ，$\sum_{i=1}^{5} y_i = 18$ ，$\sum_{i=1}^{5} x_i^2 = 55$ ，$\sum_{i=1}^{5} x_i y_i = 71$

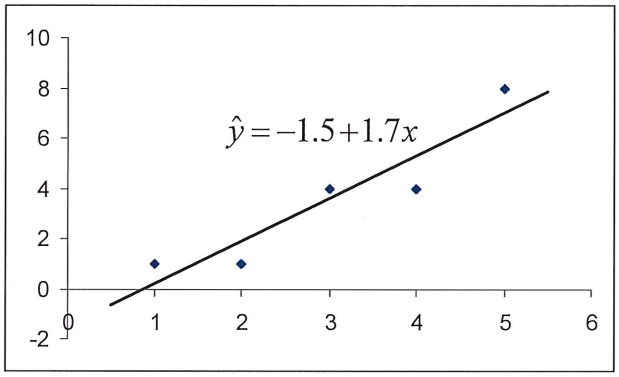
$$SS_{xy} = \sum_{i=1}^{5} x_i y_i - \frac{\left(\sum_{i=1}^{5} x_i\right)\left(\sum_{i=1}^{5} y_i\right)}{5} = 71 - \frac{(15)(18)}{5} = 17$$

$$SS_{xx} = \sum_{i=1}^{5} x_i^2 - \frac{\left(\sum_{i=1}^{5} x_i\right)^2}{5} = 55 - \frac{(15)^2}{5} = 10$$

$$\therefore \quad \hat{\beta} = \frac{SS_{xy}}{SS_{xx}} = \frac{17}{10} = 1.7$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \frac{\sum_{i=1}^{5} y_i}{5} - \hat{\beta}\frac{\sum_{i=1}^{5} x_i}{5} = \frac{18}{5} - (1.7)\frac{15}{5} = -1.5$$

我們得出最優擬合線： $\hat{y} = -1.5 + 1.7x$



解(b)：先計算剛才定義的 SS_{yy} 和 SSE：

$$SS_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{5} y_i^2 - \frac{\left(\sum_{i=1}^{5} y_i\right)^2}{5} = 98 - \frac{(18)^2}{5} = 33.2$$

$$SSE = \sum_{i=1}^{5}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{5}(y_i - 1.5 + 1.7x_i)^2 = 4.3$$

$$\therefore \quad R^2 = 1 - \frac{1.1}{6} = 0.87$$

根據這線性模型，家品店的利潤約八成七可被宣傳費解釋。

解(c)： $\hat{y} = -1.5 + 1.7(6) = 8.7$

所以當這家品店用 6 百元宣傳費時，預計利潤可達 8700 元。

(LR2) 一間保險公司相信它的推銷員的每月推銷利潤 (monthly sales) 可隨經驗 (months on job) 而增加，以下是該公司隨意抽取的有關樣本：

Months on Job x	**Monthly Sales** y (thousands of dollars)
3	8.6
5	11.8
2	4.9
8	19.3
6	16.4
9	23.2
3	7.3
4	10.9

(a) 試為樣本找一條迴歸直線 (regression line) 。

(b) 把樣本和 (a) 部分的直線劃在圖上。

(c) 試估計一位有 9 個月經驗和一位有 6 個月經驗的推銷員的每月利潤相差多少。

(LR3) 以下記錄了 20 位同學的數學科期中試和期末試的成績：

Midterm Exam	Final Exam
85	62
52	80
84	21
43	22
85	85
71	22
88	87
81	60
87	62
76	58
54	64
91	98
77	66
71	69
75	39
95	67
86	42
82	61
40	23
80	100

(a) 請找一條最優擬合線，使我們能靠期中試的分數來估算期末試的成績。

(b) 根據這條線，估算一位期中試拿 84 分的同學的期末試分數。

(c) 計算判定係數值。根據其他科目的經驗，老師認為如果迴歸線大概能解釋七成的期末試分數才算可靠，那麼他覺得這條線可靠嗎？

6) 使用 Excel 去進行線性迴歸

(此部份於 Lab Session 補充)

邏輯迴歸

在實際應用中，y 變量可能只有兩個可能的結果。例子如下

例 1 假設我們關心一個候選人當選的概率受那些因素影響。這時候 y 變量是 binary 0/1 類型的：當選或沒有當選。影響當選概率的因素有競選投入的資金，是否有負面新聞，是否在任等等。

例 2 我們關心 GRE 考試成績，大學本科 GPA 以及本科學校聲譽如何影響一個學生被研究生項目接受的概率。這裡 y 變量同樣是 binary 0/1 類型的：接受或被拒絕。

邏輯回歸(Logistic or Logit Regression)用來分析以上類型的問題。該模型特性如下：

(a) Y 變量只有 0 或 1，而且是隨機的。通常 Y=1 設定為「成功」(Success)。

(b) 假設用來預測的變量 $X_1, X_2, \cdots, X_p$ 是線性關係。

(c) (a)和(b)的關係以 Logit 函數連繫：$Logit(x) = \log\left(\frac{x}{1-x}\right)$

綜合以上得出：

$$Logit(P(Y=1)) = \log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

當中 $\frac{P(Y=1)}{1-P(Y=1)}$ 可視作「勝算比」(odds)。因此，邏輯回歸模型可用來預測某事件的「成功」機會。

稍作推演，邏輯迴歸可用以下公式表達：

$$\boxed{P(Y=1) = f\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right)}$$

其中 $f(x) = \frac{1}{1+e^{-x}}$ ，稱為 sigmoid 函數，其取值總在 0 和 1 之間。

這裡有 p 個因素影響 Y=1 的概率。這裡我們不討論如何利用數據估計係數 $\beta_0, \beta_1, \ldots, \beta_p$ 的方法。下面我們通過一個例子利用 EXCEL 來完成估計。

(LogR1) 假設某總統當選(Y=1)與否(Y=0)，和選舉經費(x ，以百萬為單位)掛鉤，並得出以下關係：

$$P(Y=1)=\frac{1}{1+\exp(7.055-0.027x)}$$

試比較 100(百萬)、200(百萬)和 300(百萬)選舉經費的勝出機會。

解：

x	P(Y=1)	增加幅度
100	0.0127	------
200	0.1604	11.7 倍
300	0.7398	3.6 倍

(LogR2) 考慮例 2，假設我們的數據如下：

student	GRE	GPA	Rank	admit
1	380	3.61	3	0
2	660	3.67	3	1
3	800	4	1	1
4	640	3.19	4	1
5	520	2.93	4	0
6	760	3	2	1
7	633	3.5	2	1
8	572	3.3	2	1
9	465	2.5	4	0
10	337	1.8	3	0
11	670	3.3	1	1
12	520	2.9	3	1
13	712	3.8	1	1
14	250	1.5	4	0
15	600	3.4	2	1

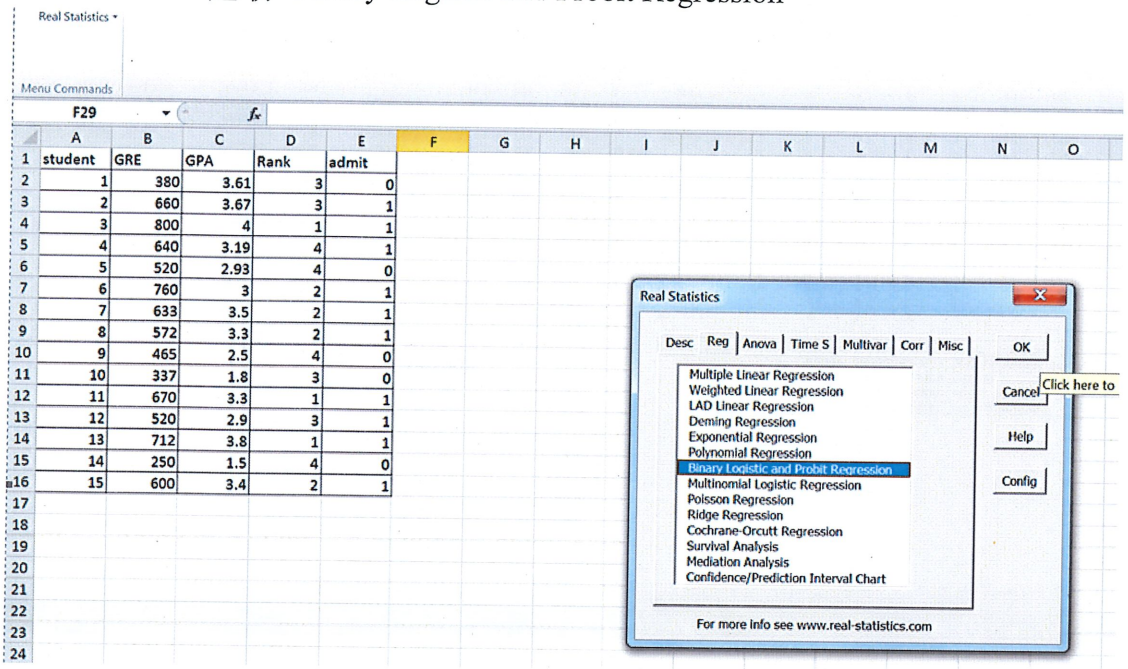
admit=1 表示被錄取，admit=0 表示被拒絕。Rank 指學校聲譽，1 為最好。

試利用邏輯迴歸，預測以下同學的取錄機會。

student	GRE	GPA	Rank
1	550	3.61	2
2	600	3.67	3
3	590	3.9	4
4	280	3	1
5	600	2.8	4

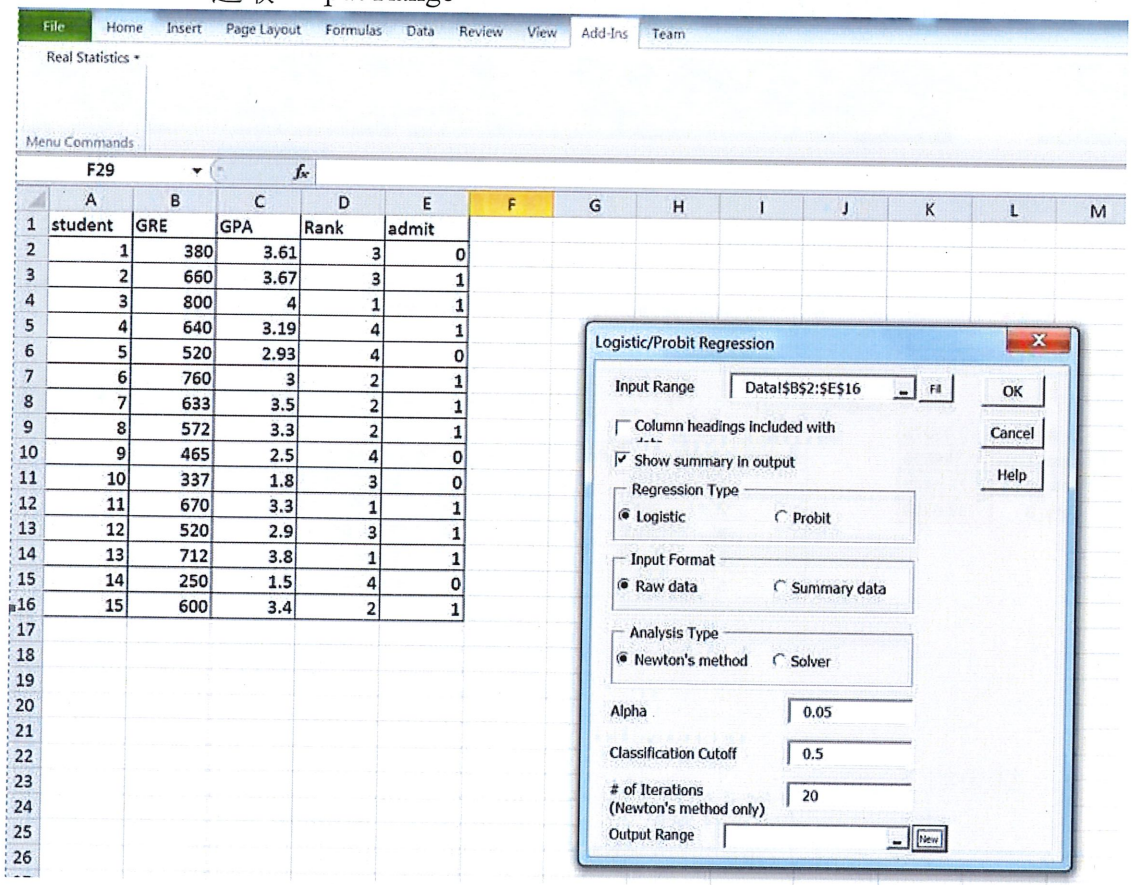
解： 1. 在 Excel 中先裝載 real statistics resource pack. (參見 “installation guide.docx”)

2. 選取 “Binary Logistic and Probit Regression”



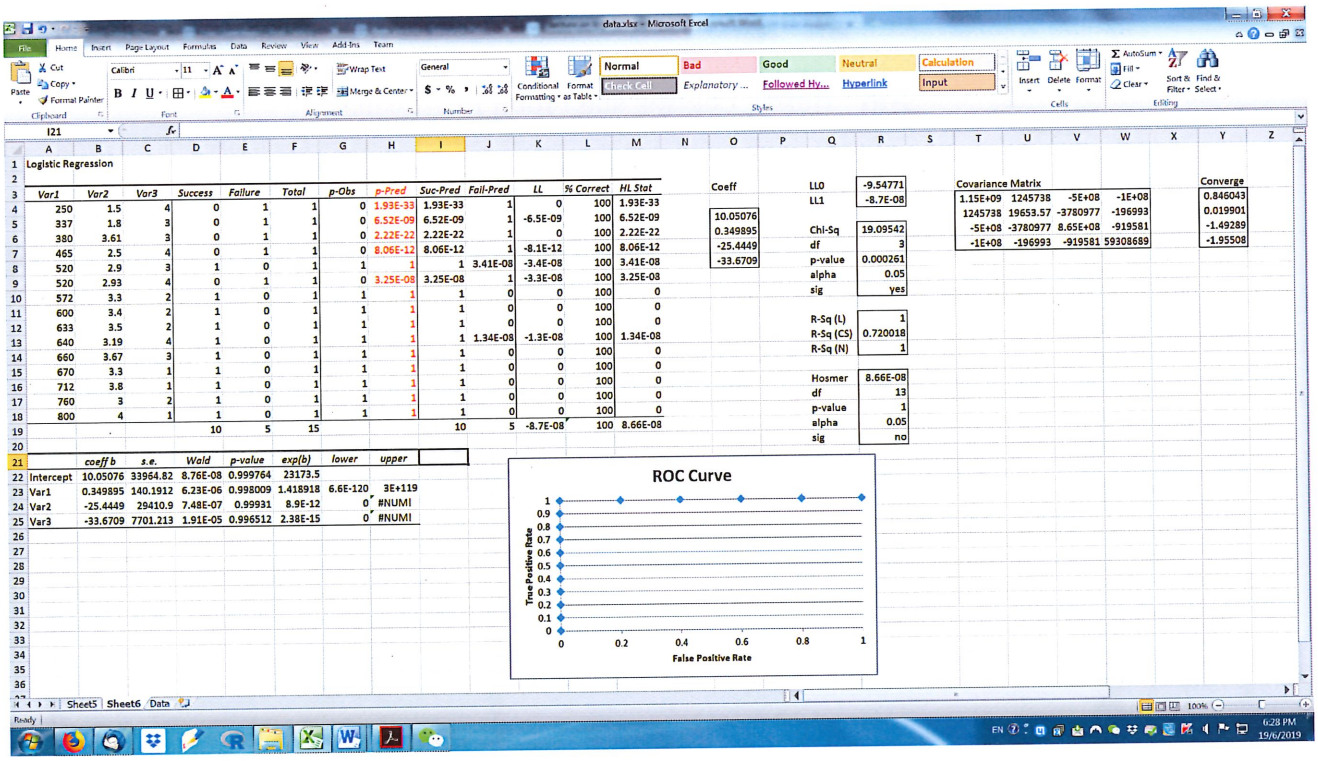
	A	B	C	D	E
1	student	GRE	GPA	Rank	admit
2	1	380	3.61	3	0
3	2	660	3.67	3	1
4	3	800	4	1	1
5	4	640	3.19	4	1
6	5	520	2.93	4	0
7	6	760	3	2	1
8	7	633	3.5	2	1
9	8	572	3.3	2	1
10	9	465	2.5	4	0
11	10	337	1.8	3	0
12	11	670	3.3	1	1
13	12	520	2.9	3	1
14	13	712	3.8	1	1
15	14	250	1.5	4	0
16	15	600	3.4	2	1

3. 選取 “Input Range”

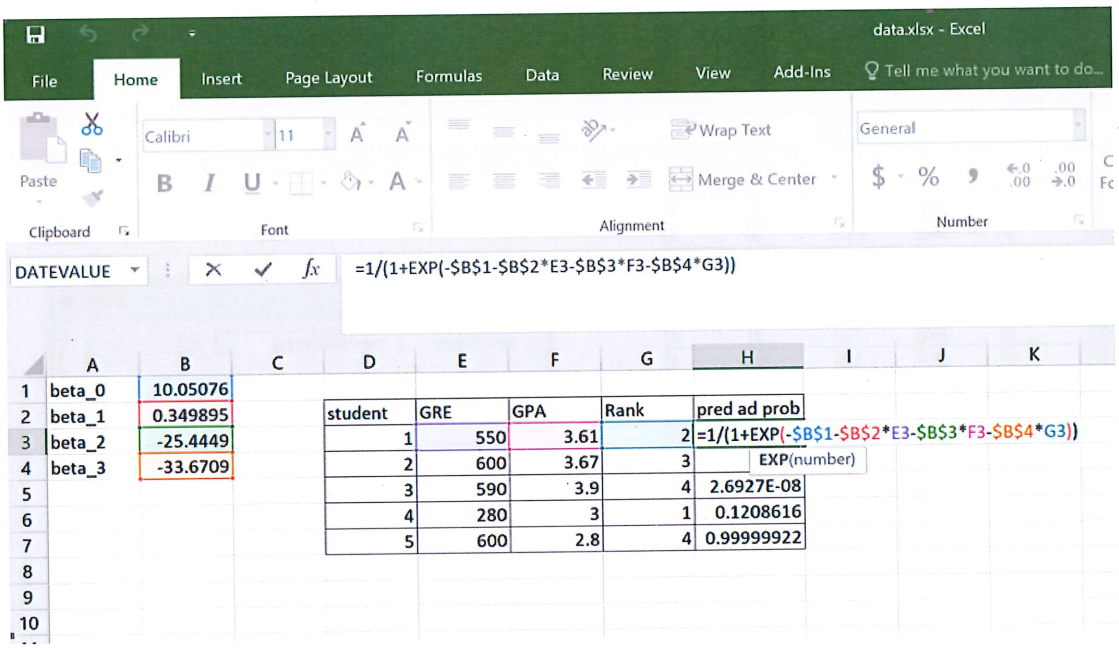


	A	B	C	D	E
1	student	GRE	GPA	Rank	admit
2	1	380	3.61	3	0
3	2	660	3.67	3	1
4	3	800	4	1	1
5	4	640	3.19	4	1
6	5	520	2.93	4	0
7	6	760	3	2	1
8	7	633	3.5	2	1
9	8	572	3.3	2	1
10	9	465	2.5	4	0
11	10	337	1.8	3	0
12	11	670	3.3	1	1
13	12	520	2.9	3	1
14	13	712	3.8	1	1
15	14	250	1.5	4	0
16	15	600	3.4	2	1

4. 按 OK 後，便能顯示結果。



現在我們考慮利用估計出來的模型來預測另外五位同學的錄取概率。



結果如下

student	GRE	GPA	Rank	pred ad prob
1	550	3.61	2	1
2	600	3.67	3	1
3	590	3.9	4	2.6927E-08
4	280	3	1	0.1208616
5	600	2.8	4	0.99999922