Logistic Regression 在社會科學研究上的應用

主講人：羅文輝（政大新聞所教授）

整　理：張璨文

這次演講的主要目的，是介紹 Logistic Regression在社會科學研究上的應用，內容包括：一、Logistic Regression 和 Linear Regression 的不同；二、Odds Ratio 的意義；三、The Simple Logistic Regression；四、The Multiple Logistic Regression；五、如何評估模式是否合適；六、Spsspc＋ 指令說明。

一、Logistic Regression和Linear Regression的不同？

這個問題可從㈠模式的選擇；㈡變項關係的分佈情形；㈢對隨機誤差的假設等方面來看：

㈠模式的選擇

Linear Regression 所處理的依變項（dependent variable）必須是等距或等比變項（interval or ratio），它不能處理名目（nominal）變項；Logistic Regression 則主要處理依變項是有兩個類目的名目變項。比如，在醫學研究上，可以從各種狀況來預測一個人是否會患心臟病？在傳播研究上，則可以預測個人是讀者或非讀者（reader or nonreader）？在政治研究上，則可以預測民眾是選民或非選民（voter or nonvoter）？所以，Logistic regression 可以解決 Linear Regression 不能處理依變項是名目變項的問題。在社會科學研究中，最適合採用 Logistic Regression 的情形，是所有的獨立變項都是等距或等比變項，或是有的獨立變項是等距或等比變項，有的是名目變項，而依變項是只有兩個類目的名目變項。

如果我們研究年齡（AGE）和患心臟病（CHD）之間的關係，若用 Linear Regression 進行分析，我們會先用 Scatterplot 來顯示這兩個變項的關係。

（圖一）

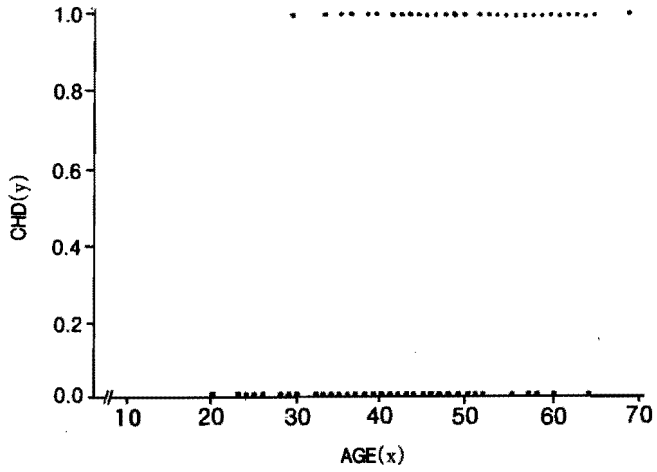

圖 一

圖一顯示：(1)年輕者患心臟病的比例較低，年長者患心臟病的比例較高。

(2)依變項呈現兩個類目（有心臟病、沒有心臟病）。

(3)圖中兩個變項的關係並不十分清楚。

(4)圖中顯示，患心臟病的變異量 (Variability) 在所有的年齡層都很大。

遇到這種情形，研究者可能會把獨立變項（年齡）分成幾組，並計算每一組患心臟病的平均數。這樣做可以減少依變項（是否患心臟病）的Variation ，同時維持兩個變項間關係的結構。把年齡分組後的情形請參考表一，把年齡分組後，再把年齡和患心臟病的關係，用 Scatterplot 顯示（見圖二）。

表一

		CHD		
Age Group	n	Absent	Present	Mean (Proportion)
20－29	10	9	1	0.10
30－34	15	13	2	0.13
35－39	12	9	3	0.25
40－44	15	10	5	0.33
45－49	13	7	6	0.46
50－54	8	3	5	0.63
55－59	17	4	13	0.76
60－69	10	2	8	0.80
Total	100	57	43	0.43

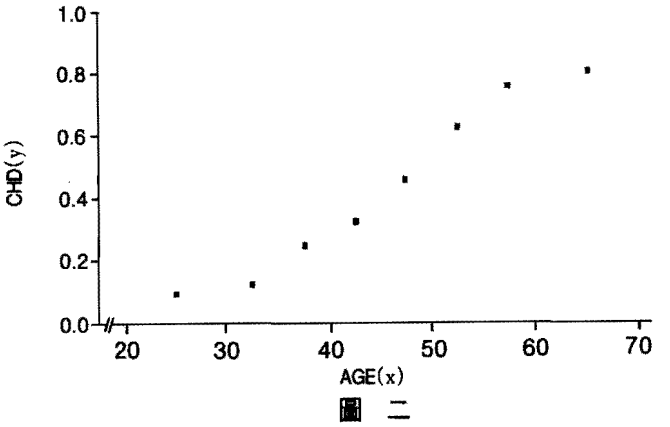

圖 二

表一及圖二顯示出 Logistic Regression 與 Linear Regression 的主要不同。

用Logistic Regression進行分析時，由於依變項是只有兩個類目的名目變項，因此依變項 Y 的期望值：E（Y／X）必須介於 0 和 1 之間；而在 Linear Regression 中，依變項是等距或等比變項，因此 E（Y／X）不會介於 0 和 1 之間。

例：當我們研究年齡和患心臟病間的關係時，年齡為獨立變項 x，是否患心臟病為依變項 y。

Logistic Regression 方程式為 $E(Y/X)=B_0+B_1X$

假設 $B_0=0.055$

$B_1=0.20$

當 X＝1(20-29 歲)時　$E(Y/X)=0.055+0.20(1)=0.255$

當 X＝5(45-49 歲)時　$E(Y/X)=0.055+0.20(5)=1.055$

Y 的期望值大於 1，由於依變項是兩個類目的名目變項，期望值應介於 0 和 1 之間，因此求出的期望值與實際狀況並不相符，即迴歸方程式所預測的母數並不存在，可見 Linear Regression 在依變項是名目變項的情況下並不適用。

㈡變項關係分佈的形狀不同

(1)Linear Regression

顧名思義，Linear Regression 變項關係的分佈形狀是直線形的。迴歸係數是指當獨立變項 X 變動 1 單位時，依變項 Y 隨之變動的單位。當 X 從 1 變動到 2 時，或是由 5 變動到 6 時，Y 隨之變動的單位相同。

(2)Logistic Regression

如圖二所示，X 軸代表年齡，Y 軸 1 表示患心臟

病，0 表示沒有患心臟病。在 Logistic Regression 中，變項關係分佈的型態呈 S 形。在 20 到 34 歲之間，隨著年齡的增加，患病的機率並沒有顯著的增加；但 35 歲之後，隨年齡的增加，患病的機率驟增，一直到 55 歲之後，才緩和下來。也就是說，當 X（年齡）從 1 變動到 2 時，或由 5 變動到 6 時，Y 隨之變動的比例並不相等；在 X 愈趨向 X 軸的兩極時，Y 的變動量較小。

㈢對隨機誤差（Random Error）的假設不同

(1)Linear Regression

公式為：$E(X/Y)=B_0+B_1X+\varepsilon$

其中對 ε 有以下的假設：

①呈常態分配（normal distribution）；

②其平均數（mean）等於 0；

③其變異量（variance）是常數（constant）；

(2)Logistic Regression

公式為：$E(X/Y)=B_0+B_1X+\varepsilon$

$Y=E(Y/X)+\varepsilon$

因為在 Logistic Regression 中，Y 不是等於 1，就是等於 0，所以 ε 也會有兩種值：

①當 $Y=1$，$\varepsilon=1-E(Y/X)$

②當 $Y=0$，$\varepsilon=-E(Y/X)$

因此，ε 的分佈不是常態分配，而是雙項分配（binomial distribution），這和 Linear Regression 假設 ε 呈常態分配明顯不同。

二、Odds Ratio

在 Linear Regression 中，研究者用未標準化迴歸係數、標準化迴歸係數和R square 來解釋變項間的關係；在 Logistic Regression 中，則主要用 Odds Ratio 來加以解釋。以下便針對 Odds Ratio 做進一步說明：

㈠簡介The odds ratio

其定義是一事件會發生的或然率除以不會發生的或然率（The odds is simply the ratio of the probability of an event occurring versus the probability of it not occurring）。公式為：$\frac{\text{Prob(event)}}{\text{Prob(no event)}}$

舉例說明如下：（見表二）

如果我們分析性別與會不會把票投給女性這兩個變項間的關係，就可以用 odds ratio 來加以解釋。

表二

〈性別〉〈是否會投票給婦女〉

	男	女	合計
會	A:47	B:81	128
不會	C:69	D:39	108
合計	116	120	236

(1)不論性別，受訪者會把票投給婦女的比例，是不會把票投給婦女的 1.18 倍（128／108＝1.18）。

(2)把性別考慮進去，此時計算的是 conditional odds：

①男人會把票投給婦女的比例，是不會投給婦女的比例的 0.68 倍（47／69＝0.68）。

②女人會把票投給婦女的比例，是不會投給婦女的比例的 2.08 倍（81／39＝2.08）。

(3)再考慮性別間的比較，此時計算的是 the ratio of 2 conditional odds：

①男性會把票投給女性的比例，是女性會把票投給女性的

比例的 0.328 倍。

$$\frac{A}{C} \div \frac{B}{D} = \frac{AD}{BC} = \frac{47 \times 39}{81 \times 69} = 0.328$$

②女性會把票投給女性的比例，是男性會把票投給女性的比例的 3.049 倍。

$$\frac{B}{D} \div \frac{A}{C} = \frac{BC}{AD} = \frac{81 \times 69}{47 \times 39} = 3.049$$

㈡利用迴歸係數求出 odds ratio（ Ψ ）

在 Logistic Regression 中，有一常用的名詞叫 Log of the odds，數學符號是 In（ Ψ ）。

In（ Ψ ）= B_1；B_1 是迴歸係數，

即Ψ的對數等於 Linear Regression 中的迴歸係數；

因此，$\hat{\Psi}360 = e^{B1}$，e 是指數（ exponential ）。

例：有無抽煙為獨立變項 X（ 有、無 ）

有無感染肺癌為依變項 Y（ 有、無 ）

當$\hat{\Psi}$＝2時，意指抽煙者得肺癌的比例，是不抽煙者的兩倍。

三、The Simple Logistic Regression

在進行 Simple Logistic Regression 時，要考慮以下三種情形：

㈠當獨立變項是兩個類目的名目變項（ dichotomous ）

例：（ 見表三及表四 ）

獨立變項為年齡（ 分成55歲以上及以下兩組 ），依變項為是否得心臟病（ 是或否 ）。

則 odds ratio 有兩種求法

①$\hat{\Psi} = \frac{21 \times 51}{22 \times 6} = 8.11$

②$\hat{\Psi}=e^{B1}=e^{2.094}=8.11$

解釋：55 歲以上得心臟病的比例，是 55 歲以下者的 8.11 倍。

表三

	AGE(X)		
CHD(y)	55(1)	<55(0)	Total
Present(1)	21	22	43
Absent(0)	6	51	57
Total	27	73	100

表四

Variable	Estimated Coefficient	Standard Error	Coeff./SE	$\hat{\Psi}$
AGE	2.094	0.529	3.96	8.1
Constant	−0.841	0.255	−3.30	

㈡當獨立變項是兩個類目以上的名目變項(polytomous)

例：（見表五）

獨立變項為種族，分成白人、黑人、西裔美人、其他四個類目，依變項為是否得心臟病（是或否）。

這時就須要用 design variable（也稱 indicator variable）的方式，將獨立變項的各類目重新處理。首先，必須從四個類目中，選擇一個當參考團體（reference group），再將各類目用類似 Linear Regression 中的 Dummy variable 重新登錄，（見表六，即是以白人當參考團體）。

表五及表七中數據有下列意義：

①以白人為參考團體，所以白人患病的 odds ratio（Ψ）為 1。

②西裔美人患病的$\Psi=\frac{15\times20}{5\times10}=6.00$（或$\Psi=e^{B1}=e^{1.386}=6.00$），

即西裔美人得心臟病的比例是白人的六倍。

③黑人患病的$\Psi=\frac{20\times20}{5\times10}=8.00$（或$\Psi=e^{B1}=e^{1.792}=8.00$），即黑人得心臟病的比例是白人的八倍。

表五

CHD Status	White	Black	Hispanic	Other	Total
Present	5	20	15	10	50
Absent	20	10	10	10	50
Total	25	30	25	20	100
Odds Ratio ($\hat{\Psi}$)	1.0	8.0	6.0	40	
95% CI		(2.3,27.6)	(1.7,21.3)	(1.1,14.9)	
ln ($\hat{\Psi}$)	0.0	2.08	1.79	1.39	

表六

	Design Variables		
RACE(Code)	D_1	D_2	D_3
White(1)	0	0	0
Black(2)	1	0	0
Hispanic(3)	0	1	0
Other(4)	0	0	1

表七

Variable	Estimated Coefficient	Standard Error	Coeff./SE	$\hat{\Psi}$
RACE(1)	2.079	0.633	3.29	8.0
RACE(2)	1.792	0.646	2.78	6.0
RACE(3)	1.386	0.671	2.07	4.0
Constant	−1.386	0.500	−2.77	

㈢當獨立變項是連續（ continuous ）變項

例：表八顯示，獨立變項為年齡，依變項為是否患心臟病。此時要先求出發生事件（ 患心臟病 ）的機率，

其公式為：$\text{Prob (Event)}=\frac{e^{B0+B1X}}{1+e^{B0+B1X}}=\frac{1}{1+e^{-(B0+B1X)}}$ 。

Logistic Regression 的方程式為

$g(X)=\text{In}[\frac{\pi(X)}{1-\pi(X)}]=B_0+B_1X=-5.310+0.111(\text{age})$

如果年齡＝20歲，

則 $g(X)=-5.310+0.111\ (20)=-3.09$

$$\text{Prob（ 得心臟病 ）}=\frac{1}{1+e^{-(B0+B1X)}}$$

$$=\frac{1}{1+e^{-(-3.09)}}$$

$$=\ .0435$$

表八

Variable	Estimated Coefficient	Standard Error	Coeff./SE
AGE	0.111	0.024	4.61
Constant	−5.310	1.134	−4.68

即 20 歲患心臟病的機率是 0.0435（ 4.35% ），幾乎是不太可能。

年齡變動一歲對患病機率的影響很小，如果以 10 年為單位，則 $\Psi=e^{10\times.111}=3.03$，即年齡每增加 10 年，得心臟病的危險就增加 3.03 倍。

四、The Multiple Logistic Regression

當獨立變項不只一個的時候，就會用到 The Multiple Logistic Regression。

例：依變項為 Nodes：淋巴切片結果（1 是惡性，0 是良性）

獨立變項有五個，分別是：

1. 年齡：Age——連續變項

2. 血清中的磷酸指數：Acid——連續變項

3. x 光檢驗結果：Xray——1 是陽性，0 是陰性

4. 腫瘤屬惡性或良性：Grade——1 是惡性，0 是良性

5. 腫瘤的發展階段：Stage——0 代表第一階段，1 代表第二階段

電腦 Spsspc＋ 跑出的 Logistic Regression 分析結果如下：

表九

LOGISTIC REGRESSION NODES WITH AGE ACID XRAY GRADE STAGE.

----------Variables in the Equation----------

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
AGE	−.0693	.0579	1.4320	1	.2314	.0000	.9331
ACID	.0243	.0132	3.4229	1	.0643	.1423	1.0246
XRAY	2.0453	.8072	6.4207	1	.0113	.2509	7.7317
GRADE	.7614	.7708	.9758	1	.3232	.0000	2.1413
STAGE	1.5641	.7740	4.0835	1	.0433	.1722	4.7783
Constant	.0618	3.4599	.0003	1	.9857		

根據表九，Logistic Regression 的方程式為：

$$g(X) = 0.0618 - 0.0693(Age) + 0.0243(Acid) + 2.0453(Xray) + 0.7614(Grade) + 1.5641(Stage)$$

〈問題一〉如果一個人 66 歲，Acid＝48，Xray＝0，Stage＝0，Grade＝0，則①g(X)＝？ ②淋巴切片結果為惡性的機率有多大？③其 odds ratio(Ψ)＝？

答：①g(X)＝0.0618－0.0693(66)＋0.0243(48)＝－3.346

②Prob（惡性）$=\frac{1}{1+e^{-(B0+B1X\cdots)}}=\frac{1}{1+e^{-(-3.346)}}$

$=0.034$

解釋：這個人淋巴切片結果呈惡性的機率只有 0.034

③$\Psi=\frac{\text{Pro(event)}}{\text{Pro(no event)}}=\frac{0.034}{1-0.034}=0.035$

解釋：這個人淋巴切片結果呈惡性的比例是呈良性比例的 .035倍。

$\hat{\Psi}=\frac{1-0.034}{0.034}=28.41$

換言之，這個人淋巴切片結果呈良性的比例是呈惡性比例的 28.41 倍。

〈問題二〉如果一個人 60 歲，Acid＝62，Xray＝1，Stage＝0，Grade＝0，則當 Grade 由 0 變成 1 時（即腫瘤由良性變成惡性），其淋巴切片結果為惡性的 odds ratio 會增加幾倍？

當 grade＝0，Logistic Regression 的方程式為：

g(X) ＝ 0.0618 － 0.0693(60) ＋ 0.0243(62) ＋ 2.0453(1) ＋ 0.7614(0)＋1.5641(0)＝－0.54

Prob（惡性）$=\frac{1}{1+e^{-(-.54)}}=0.37$

odds ratio(Ψ)$=\frac{0.37}{1-0.37}=0.59$

當 Grade＝1，Logistic Regression 的方程式為：

$g(X) = 0.0618 = 0.0693(60) + 0.0243(62) + 2.0453(1) + 0.7614(1) + 1.5641(0) = 0.22$

$$\text{Prob（惡性）} = \frac{1}{1+e^{-.22}} = 0.555$$

$$\text{odds ratio}(\Psi) = \frac{.555}{1-.555} = 1.25$$

$$\frac{1.25}{.59} = 2.12$$

答：當 Grade 由 0 變成 1 時，其淋巴切片結果為惡性的 odds ratio 會增加 2.12 倍。這個數據可以從表九的 Exp(B) 看出，Exp(B) 其實就是 odds ratio，Grade 這一行中的 Exp(B) 是 2.1413，和上面運算出的 2.12 很相近。

五、如何評估模式是否適合？（Assessing the Goodness of Fit of the Model）

在 Logistic Regression 中，評估模式是否適合的方法有兩種，茲分述如下：

㈠利用 Wald Statistics 做假設驗證

Wald Statistics $= \left(\frac{B}{SE}\right)^2$ ————此數值呈 Chi-Square(X^2)分佈

零假設 H_o：$b=0$　即獨立變項不能預測依變項

研究假設 H_a：$b \neq 0$　即獨立變項可以預測依變項

以表九為例：

——Age 的 Wald Statistics＝1.4320，顯著程度是 .2314 ($p > .05$)，因此在顯著水準為 .05 的情況下，不能拒絕零假設。即年齡不能預測淋巴切片的結果是否會呈惡性。所以在此便可將年齡從獨立變項中去除。

——表九所列的變項中，只有 Stage 和 Xray 達顯著水準，所以在這個迴歸模式中，只需保留 Stage 和 Xray 兩個獨立變項。

㈡利用 Log Likelihood 做假設驗證

用 Wald Statistics 做假設驗證有一顧慮，當迴歸係數 B 很大時，迴歸係數的標準差 SE 也會很大，如此一來，便不容易拒絕零假設。在這種狀況下，Log likelihood 是比較可靠的方法。

Likelihood 表示獨立變項預測之母數發生的可能性，其值通常小於 1，所以統計學家便將 Likelihood 的值取對數，並乘上－2，成為－2 Log Likelihood（簡稱－2LL），這樣的數值與 Chi-Square distribution 相近，比較好瞭解。

用－2 Log Likelihood 來評估模式，是比較兩個模式－2LL 的差異，其中一個模式未包括考慮的變項，另一個模式則包括考慮的變項，這兩個模式的－2LL之差距，則代表了包括考慮變項的模式，其預測力是否比未包括考慮變項的模式好；兩個模式之－2LL的差距用G值表示，公式如下：

$$G = -2\ \mathrm{In}\left[\frac{\text{Likelihood without the variable}}{\text{Likelihood with the variable}}\right]$$

$$= -2\ [\text{Log Likelihood without the variable} - \text{Log Likelihood with the variable}]$$

以G值（improvement）做假設驗證：

零假設為 H_0：Model n－1＝Model n，Model n 和 Model n－1 的預測力並沒差別。

研究假設為 H_a：Model n－1≠Model n，Model n 的預測力有顯著的改進。

以表十為例說明，模式 1.2.3.4.5. 逐次納用新增的預測變項（電腦會選擇顯著程度最低，即最顯著的變項，先進入模式），模式 5.（Saturated）則表示包括所有的變項，茲分別說明如下：

表十

	Model	−2LL	Improvement	d.f.	sig.
1.	Constant	70.25			
2.	Xray	59.00	11.25	1	.001
3.	Xray,Stage	53.35	5.65	1	.0175
4.	Xray, Stage, Acid	51.05	2.30	1	.354
5.	Saturated	48.126	2.92	2	.425

(1)模式1.只包括常數（Constant），−2LL＝70.25。

(2)模式2.中輸入第一個變項 Xray ，模式中包括 Xray 和 Constant 。−2LL 的值從模式1.的 70.25 降至 59.00，改善（improvement）11.25，也就是說，G值為 11.25。表十中的 sig 意指用G值進行假設驗證，在顯著水準 .001 時拒絕零假設，顯示第二個模式比第一個模式的預測力好。

(3)模式3.中輸入第二個變項 Stage，此時模式中包括 Xray、Stage 和 Constant。−2LL 的值 59.00 降至 53.35，改善 5.65。sig 顯示，在顯著水準 .0175($p<.05$) 時拒絕零假設，我們有足夠的證據說，第三個模式比第二個模式的預測力好。

(4)模式4.中輸入第三個變項 Acid，此時模式中有三個獨立變項及常數。−2LL 值由 53.35 降至 51.05，只改善 2.30。sig＝.354($p>.05$)，顯示在顯著水準 .05 時，無法拒絕零假設。也就是說，我們沒有足夠的證據顯示，第四個模式比第三個模式的預測力好。

(5)模式5.是 Saturated 模式，所有的獨立變項均進入模式中。－2LL 值降至 48.126，只改善了 2.92。sig＝.425(p＞.05)，顯示在顯著水準 .05 時，無法拒絕零假設。模式5.和模式4.並無差異。

迴歸分析評估模式有兩個主要的原則：一是該模式必須是最適合的模式（ the best fitting model ），二是該模式必須最為精簡（ the most parsimonious model ）。

在 Linear Regression 中，研究者通常選擇變項最少、但能解釋最多變異量的模式；而在 Logistic Regression 中，也應遵循這種原則，因此在表十中，我們選擇第三個模式，這個模式所用的預測變項最少，而預測力和第四個模式並無差異。

六、Spsspc十指令說明

㈠基本指令

Spsspc＋的版本，要 4.0 以上才能處理 Logistic Regression。如果只進行變項間關係的分析，用下列指令，就能跑出像表十般的表格：

Logistic Regression Nodes with Age Acid Xray Grade Stage. 上述指令中，Logistic Regression 告訴電腦進行 Logistic Regression 分析，Nodes 是依變項，with 以後是獨立變項。如果獨立變項中，有兩個類目以上的名目變項，則需要另外的指令。

㈡評估模式的指令

在 Spsspc＋中，評估模式時，可以選擇 Forward 和 Backward 兩種方法，在此介紹的是 Forward Stepwise

Selection。其程序如下：

(1)第一個模式只包含常數，不過，只包含常數的模式通常沒有意義，研究者可以下指令省略這道程序。

(2)電腦會用表十一中的 Score，選擇顯著程度最低（最顯著）的變項（occu）進入第二個模式。

(3)用 Residual Chi-Square 來做假設驗證。

零假設 H_o：所有未進入迴歸方程式（未放進模式）之變項的迴歸係數都等於零。（即這些變項都不具顯著的預測力，原有的模式不須修正。）

研究假設 H_a：未進入迴歸方程式之變項的迴歸係數中，至少有一個不等於零。（即至少有一個變項具顯著預測力，原有的模式至少要再加入一個新的變項。）

(4)電腦會繼續把顯著程度最低的變項選入下一個模式中，直到所有顯著的獨立變項都被選入迴歸方程式中，而 Residual Chi-Square 的值不再顯著為止。

㈢實例說明

如果我們用性別(sex)、職業(occu)、年齡(age)、地區（manarea）、及職級（rank）、五個變項為獨立變項，來預測依變項「在電視新聞上被引述的程度」（form）。則 Spsspc＋的 Forward Stepwise Selection 印出的 Logistic Regression 結果如表十一所示：

表十一 Spsspc^{+} Printout

〈第一個模式〉

	Chi-Square	df	Significance
-2 Log Likelihood	750.559	858	.9965
Goodness of Fit	859.000	858	.4840

Variables in the Equation

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
Constant	-1.6708	.0935	319.5282	1	.0000		

SPSS/PC^{+}

Variables not in the Equation

Residual Chi Square 36.549 with 5 df Sig=.0000

Variable	Score	df	Sig	R
SEX	.9639	1	.3262	.0000
MANAREA	18.2019	1	.0000	.1469
AGE	1.6070	1	.2049	.0000
OCCU	22.6631	1	.0000	.1659
RANK	11.5617	1	.0007	.1129

〈第二個模式〉

	Chi-Square	df	Significance
-2 Log Likelihood	726.987	857	.9995
Model Chi-Square	23.572	1	.0000
Improvement	23.572	1	.0000
Goodness of Fit	858.995	857	.4744

Variables in the Equation

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
OCCU	.9593	.2065	21.5708	1	.0000	.1615	2.6098
Constant	-2.2668	.1727	172.2687	1	.0000		

SPSS/PC^{+}

Variables not in the Equation

Residual Chi Square 15.249 with 4 df Sig=.0042

Variable	Score	df	Sig	R
SEX	.0041	1	.9489	.0000
MANAREA	14.4841	1	.0001	.1290
AGE	1.1306	1	.2876	.0000

〈第三個模式〉

	Chi-Square	df	Significance
-2 Log Likelihood	710.124	856	.9999
Model Chi-Square	40.435	2	.0000
Improvement	16.862	1	.0000
Goodness of Fit	853.313	856	.5195

Variables in the Equation

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
MANAREA	1.1508	.3167	13.2044	1	.0003	.1222	3.1607
OCCU	.8717	.2087	17.4524	1	.0000	.1435	2.3910
Constant	-3.1799	.3249	95.7767	1	.0000		

SPSS/PC+

Variables not in the Equation

Residual Chi Square .744 with 3 df Sig=.8558

Variable	Score	df	Sig	R
SEX	.0009	1	.9757	.0000
AGE	.0914	1	.7624	.0000
RANK	.6435	1	.4224	.0000

表十一顯示，在第一個模式中，Spsspc＋ 會首先把常數（ constant ）選入迴歸方程式中，所有的獨立變項均未在迴歸方程式中（ variables not in the equation ），其中 Residual Chi Square＝36.549，sig＝.0000(p＜.001)，顯示在顯著水準 .001時，拒絕零假設，即未在迴歸方程式中的獨立變項，至少有一個的迴歸係數不等於零。此時，電腦會用 Score 來選擇顯著程度最低（ 最顯著 ）的變項職業（ Occu ）進入下一個模式。第一個模式的－2 Log Likelihood 為 750.559 。

第二個模式顯示，迴歸方程式中的變項（ variables in the equation ）包括 occu 與 constant，而未在迴歸方程式中的變項有 sex、manarea、age 及 rank。Residual Chi Square 為 15.249，sig＝.0042，顯示在顯著水準 .0042 (p＜.01) 時，拒絕零假設，即在 sex、manarea、age 及 rank 四個變項中，至少

有一個對依變項具有預測力。電腦會用 Score 把這四個變項中，顯著程度最低的變項 manarea 選入下一個模式。此時，本模式的－2 Log Likelihood 已從第一個模式的 750.559 降為 726.987，改進了 23.572。

第三個模式顯示，迴歸方程式中的變項包括 occu、manarea 與 constant，而未在迴歸方程式中的變項有 sex、age 及 rank。Residual Chi Square＝.774，sig＝ .8558(p＞.05)，顯示在顯著水準 .05 時，無法拒絕零假設，即 sex、age 及 rank 無法幫助預測依變項。本模式的 －2 Log Likelihood 為 710.124，比上一個模式的 726.987，改進了16.862。

由於未進入迴歸方程式的三個獨立變項 sex、age 及 rank 都無法幫助預測依變項，Spsspc＋程式不再逐步選擇變項，運算就此停止，電腦告訴我們，最適合預測依變項的模式包括職業與地區兩個獨立變項。

七、結語

這場演講只介紹 Logistic Regression 的一些基本概念，一些比較複雜的問題，如獨立變項間的互動、依變項如果有兩個以上的類目等問題，則請各位參考附上的參考書目。

此外，我必須特別聲明，在演講中引用的圖表中，圖一、圖二、表一、表三至表八係出自 David W. Hosmer & Stanley Lemeshow, Applied Logistic Regression, New York: John Wiley & Sons, 1989；而表九則出自 Marija J. Norusis, Spsspc＋ Advanced Statistics Version 5.0, Chicago, IL: Spss Inc., 1992。

總之，Logistic Regression 是一種相當有用的統計方法，希望今天的演講能幫助各位對Logistic Regression有最初步的瞭解。

八、參考書目

Agresti, A. (1990). Categorical data analysis. New York: Wiley.

Aldrich, J. H.,& Nelson, F. D. (1984). Linear probability, logit, and probit models. Guantitative Applications in the Social Sciences (series no. 07-045). Beverly Hills and London: Sages Pubns.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). Discrete multivariate analysis. Cambridge, MA: MIT.

Brand, R.,& Keirse, M. J. N. C. (1990). Using Logistic Regression in perinatal epidemiology:An introduction for clinical researchers. Part Ⅰ: Basic concepts. Paediatric and Perinantal epidemiology, 4, 22-38.

Christensen, R. (1990). Log-linear models. New York: Springer-Verlag.

Cleary, P. D., & Angel, R. (1984). The analysis of relationships involving dichotomous dependent variable. Journal of Health and Social Behavior, 25, 334-348.

Fienberg, S. E. (1981). The analysis of cross-classified Categorical data (2nd ed). Cambridge, MA: MIT.

Fingleton, B. (1984). Models of category counts. Cam-

bridge, England: Cambridge University Press.

Fleiss, J. L., Williams, J. B., & Dubro, A. F. (1986). The logistic regression analysis of psychiatric data. Journal of Psychiatric research, 20, 195-209.

Haberman, S. J. (1978). Analysis of quantitative data (Vols. 1 & 2). New York: Academic Press.

Hanushek, E. A., & Jackson, J. E. (1977). Statistical methods for social scientists (ch. 7). New York: Academic Press.

Hosmer, D. W., & Lemeshow, D. (1989). Applied logistic regression (chapter 1-5). New York: John Wiley & sons.

King, G. (1989). Unifying political methodology: The likelihood theory of statistical inference. Cambridge, England: Cambridge University Press.

Kmenta, J. (1986). Elements of econometrics (2nd ed., section 11-5). New York: MacMillan.

Landwehr, J. M., Pregibon, D., & Shoemaker, A. C. (1984). Graphical methods for assessing logistic regression models. Journal of the American Statistical Association, 79, 61-71.

McCullagh, P., & Nelder, J. A. (1989). Generalized Linear models (2nd ed). London: Chapman and Hall.

Upton, G. J. (1978). The analysis of cross-tabulated data. Chichester, England: Wiley.