人口學刊
第 37 期，2008 年 12 月，頁 87-113
Journal of Population Studies
No. 37, December 2008, pp. 87-113

研究論文

設計矩陣與人口資料分析：線性代數之應用

陳寬政* 楊靜利**

* 長庚大學醫務管理學系暨研究所教授
** 成功大學醫學院老年學研究所副教授

收稿日期：2008.01.17；接受刊登：2008.12.07

中文摘要

本文引入設計矩陣的概念，使用幾個簡化的例子說明傳統的迴歸分析與變異數分析乃是同一個分析方法，也指出次數分配的分析可以參數化，適用同一套最小平方法的分析，打破傳統方法與計量教科書有關計量尺度與分析方法的迷思。本文同時指出，設計矩陣的安排對應著模型與參數設定，儘管矩陣的階序相同，不同的矩陣內容代表不同的模型與參數設定及研究設計，則進行資料分析以前，一但有了分析模型與參數設定的構想，研究者宜乎審慎考量其設計矩陣與模型設立的對應關係，確認設計矩陣正確代表其研究設計。正交的設計矩陣保證參數估計「分別」為之，互不相干，不會產生解釋變項間牽扯不清的問題，乃研究者需要多加注意的設計。在討論中，本文也指出傳統方法與計量教科書對於分析單元的主張未必能適應學術的進展，尤其在多層次分析與機率或次數分析的架構下，此一主張有需要修飾之處。

關鍵詞：迴歸分析、設計矩陣、計量尺度、分析單元、概化線性模型

壹、前言

本文使用線性代數（linear algebra），先就平均數之分析引入分組迴歸（subgroup regressions）的概念，說明分組迴歸其實就是一個統合迴歸，只要設計矩陣（design matrix）安排妥當，是否分別估計並不重要。接著，我們就次數分佈適用最小平方法（least squares method）估計事件發生的機率，說明只要設計矩陣的安排妥當，次數分佈的分析與迴歸分析其實沒有兩樣。我們使用的例子指出，無論是事件發生機率、總平均數或是分組平均數之計算，都適用迴歸分析方法，只要設計矩陣安排妥當，同一套方法適用各類尺度資料（scales of measurement）之分析（McCullagh and Nelder 1999），而分析單元（unit of analysis）的定位也愈來愈模糊（Raudenbush and Bryk 2002），則基礎研究法與統計學教科書的內容有進一步予以檢討的必要，至少在傳統內容之後，應再補充說明線性模型的涵蓋性，才趕得上學術的發展。本文所有計算均直截了當，套用 Microsoft Office 的 Excel 配件就能完成，事實上 Open Office 也有對應的套件，所以本文數據均連結著工作表運算。

各種社會科學所使用的資料分析方法其實大同小異，基本上是同一套邏輯與數學演化出來的程序，也許敘述詞語略有不同，程序上卻是一致的，人口資料的分析也不例外。舉例而言，台灣自 1951 年以來生育率長期大幅下跌，自 1983 年以後維持低於平均每位育齡婦女一生生育一個女兒的替換水準，1997 年以後更為全球少數幾個最低生育率的國家之一。超低生育率的問題引起各方關注，表 1 顯示此一生育率降低的過程係由兩個相反方向的運動所組成，其一為 1951 至 1980 年間，中高育齡婦女停止生育的發展，其二為 1980 年以後由於中上教育制度大幅擴張，大量 25 歲以下的育齡婦女進入學校就讀而停止生育的變化。我們在表 1 尾端加了一列平均值，列出歷年來各年齡組

表 1　台灣育齡婦女年齡別生育率與總生育率，1951-2006

單位：千分比

年期	年齡別生育率							總生育率
	15-19	20-24	25-29	30-34	35-39	40-44	45-49	
1951	68	287	350	311	226	132	34	7,040
1952	53	272	342	294	220	113	29	6,615
1953	48	265	336	292	218	108	27	6,470
1954	48	263	334	292	218	104	26	6,425
1955	50	273	341	295	219	103	25	6,530
1956	51	264	340	296	222	105	23	6,505
1957	45	249	325	275	197	92	17	6,000
1958	43	248	336	281	199	90	14	6,055
1959	46	258	334	270	190	86	14	5,990
1960	48	253	333	255	169	79	13	5,750
1961	45	248	342	245	156	71	10	5,585
1962	45	255	338	235	145	65	10	5,465
1963	41	252	337	231	139	60	10	5,350
1964	37	254	335	214	120	52	8	5,100
1965	36	261	326	195	100	41	6	4,825
1966	40	274	326	188	91	38	6	4,815
1967	39	250	295	158	70	28	4	4,220
1968	41	256	309	161	68	26	4	4,325
1969	40	245	298	151	63	23	4	4,120
1970	40	238	293	147	59	20	3	4,000
1971	36	224	277	134	51	16	3	3,705
1972	35	208	257	117	41	13	2	3,365
1973	33	203	250	105	37	12	2	3,210
1974	34	197	235	96	35	10	2	3,045
1975	37	194	215	83	27	8	2	2,830
1976	38	213	241	88	28	8	1	3,085
1977	37	194	206	73	23	6	1	2,700
1978	36	194	213	73	20	5	1	2,710
1979	35	194	209	72	18	4	0	2,660
1980	33	180	200	69	16	4	1	2,515

（續下表）

表 1　台灣育齡婦女年齡別生育率與總生育率，1951-2006（續）

單位：千分比

年期	年齡別生育率							總生育率
	15-19	20-24	25-29	30-34	35-39	40-44	45-49	
1981	31	176	197	69	14	3	1	2,455
1982	29	166	186	66	14	3	0	2,320
1983	26	154	174	62	13	2	0	2,155
1984	23	144	168	60	13	2	0	2,050
1985	20	129	158	56	12	2	0	1,885
1986	18	112	139	52	12	2	0	1,675
1987	16	109	147	54	12	2	0	1,700
1988	16	111	164	64	13	2	0	1,850
1989	16	98	145	61	14	2	0	1,680
1990	17	100	159	68	15	2	0	1,805
1991	17	92	149	68	16	2	0	1,720
1992	17	91	148	72	16	2	0	1,730
1993	17	91	149	75	18	2	0	1,760
1994	17	87	148	79	18	2	0	1,755
1995	17	86	148	82	20	2	0	1,775
1996	17	83	145	84	21	2	0	1,760
1997	15	80	147	87	22	3	0	1,770
1998	14	66	116	73	21	3	0	1,465
1999	13	66	126	82	21	3	0	1,555
2000	14	72	132	90	24	3	0	1,675
2001	13	61	106	75	21	3	0	1,395
2002	13	57	102	73	20	3	0	1,340
2003	11	52	92	68	20	3	0	1,230
2004	10	49	86	68	20	3	0	1,180
2005	8	44	79	67	21	3	0	1,110
2006	7	41	78	71	23	3	0	1,115
平均值	30	171	223	134	69	28	5	3,350

資料來源：歷年台灣人口統計。總生育率的概念為年齡別生育率加總，適用五歲年齡組生育率時則因各年齡組生育率實為各該組內單歲年齡組生育率之平均，所以加總為總生育率時乘 5 累計。

育齡婦女的平均生育率。一般計算平均數的等式為 $\bar{y}=\Sigma_i y_i/n$，但是平均數計算的原理其實是最小平方法，本文檢討同一套方法之運用於各種模型之運算。

我們計算表 1 各年齡組之平均生育率時，直接使用分組迴歸方法，設立線性模型為

$$\begin{bmatrix} \underline{y}_{15-19} \\ \underline{y}_{20-24} \\ \underline{y}_{25-29} \\ \underline{y}_{30-34} \\ \underline{y}_{35-39} \\ \underline{y}_{40-44} \\ \underline{y}_{45-49} \end{bmatrix} = \begin{bmatrix} \underline{1} & \underline{0} & \underline{0} & \underline{0} & \underline{0} & \underline{0} & \underline{0} \\ \underline{0} & \underline{1} & \underline{0} & \underline{0} & \underline{0} & \underline{0} & \underline{0} \\ \underline{0} & \underline{0} & \underline{1} & \underline{0} & \underline{0} & \underline{0} & \underline{0} \\ \underline{0} & \underline{0} & \underline{0} & \underline{1} & \underline{0} & \underline{0} & \underline{0} \\ \underline{0} & \underline{0} & \underline{0} & \underline{0} & \underline{1} & \underline{0} & \underline{0} \\ \underline{0} & \underline{0} & \underline{0} & \underline{0} & \underline{0} & \underline{1} & \underline{0} \\ \underline{0} & \underline{0} & \underline{0} & \underline{0} & \underline{0} & \underline{0} & \underline{1} \end{bmatrix} \begin{bmatrix} b_{15-19} \\ b_{20-24} \\ b_{25-29} \\ b_{30-34} \\ b_{35-39} \\ b_{40-44} \\ b_{45-49} \end{bmatrix} + \begin{bmatrix} \underline{e}_{15-19} \\ \underline{e}_{20-24} \\ \underline{e}_{25-29} \\ \underline{e}_{30-34} \\ \underline{e}_{35-39} \\ \underline{e}_{40-44} \\ \underline{e}_{45-49} \end{bmatrix},$$

其中 1 與 0 底下均加橫線，表示為一行 1 與一行 0，以表 1 資料而言乃各 56 個 1 與 0；同樣地，年齡別生育率 y 與離均差 e 也都加了底線，也是各為 56 個 y 與 e；平均數 b 則未加底線，表示單一數值。我們可以進一步簡化模型等式為

$$\underline{y}=X\underline{b}+\underline{e},$$

表 1 共有七個年齡組，所以 $\underline{y}$ 表示一行 56×7＝392 個年齡別生育率，X 表示 392×7 的矩陣，有 392 列 7 行共 2,744 個 0 與 1；$\underline{b}$ 則表示一行 7 個平均數，$\underline{e}$ 表示一行 56×7＝392 個離均差。

計算各年齡組平均生育率時，只要在等式兩邊均乘以 X'（表示 X 矩陣橫擺），得 $X'\underline{y}=X'X\underline{b}+X'\underline{e}$，採用最小平方法之設定，令 $X'\underline{e}=\underline{0}$，則 $\underline{b}=(X'X)^{-1}X'\underline{y}$；由於

$$X'X=\begin{bmatrix}56&0&0&0&0&0&0\\0&56&0&0&0&0&0\\0&0&56&0&0&0&0\\0&0&0&56&0&0&0\\0&0&0&0&56&0&0\\0&0&0&0&0&56&0\\0&0&0&0&0&0&56\end{bmatrix}，X'\underline{y}=\begin{bmatrix}\Sigma_i y_{i,15-19}\\\Sigma_i y_{i,20-24}\\\Sigma_i y_{i,25-29}\\\Sigma_i y_{i,30-34}\\\Sigma_i y_{i,35-39}\\\Sigma_i y_{i,40-44}\\\Sigma_i y_{i,45-49}\end{bmatrix}，$$

取得我們的平均數估計值為 $\underline{b}'$＝（30，171，223，134，69，28，5），與分年齡組各別累加生育率後除以 56 的結果完全相同。簡而言之，分別計算或分組計算的動作並不必然是一一處理；就像找來七個學生各自負責一個年齡組同時計算般，我們以矩陣陳列分組迴歸的模型，七個平均數一次完成計算而互不干擾，本就是省時省事而能照顧全局的方法，需要一個特別的腦袋才會認為這是將分組資料混合一起，是違背教科書而離經叛道的方法。以下的討論為避免動輒數千甚至數萬個數字之處理，我們使用簡縮的例子以節省篇幅，重點在邏輯與數學原理，而不在於所使用的例子是否具有某一學科「意涵」。由於平均生育率的計算不涉及抽樣分配（sampling distribution），以上討論並未區分母體與樣本，但以下其他例子的討論涉及抽樣與檢證時，我們將使用希臘小寫字母表示母體參數，英文小寫字母表示樣本參數。

貳、平均數

以 8 個學生的體重為例，y_i＝（78，70，65，67，50，45，52，54）平均數為 60.13 公斤，我們可以設立一個單一參數的模型 $y_i=\alpha+\varepsilon_i$，則對應每個 y_i 的 x_i 值均為 1，使用最小平方法來估計參數 α，得 $a=(\underline{x}'\underline{x})^{-1}\underline{x}'\underline{x}=(1/8)*481=60.13$，符合平均數為最小平方參數的瞭解。現在如果這 8 個學生有 4 個男生，4 個女生，使用 1 表示男

生，0 表示女生，則 8 個學生的性別為 z_i =（1，1，1，1，0，0，0，0），而若目標為估計兩性的平均體重

$\bar{y}_i$ =（70.00，50.25），

可以設立兩個參數的模型 $y_{ij}=\beta_j+\varepsilon_{ij}$，其中 $i=1，2，\cdots，n_j$，$j=1，2$。我們的模型可以矩陣表示為

$$\begin{bmatrix} y_{11} \\ y_{21} \\ y_{31} \\ y_{41} \\ y_{12} \\ y_{22} \\ y_{32} \\ y_{42} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \varepsilon_{31} \\ \varepsilon_{41} \\ \varepsilon_{12} \\ \varepsilon_{22} \\ \varepsilon_{32} \\ \varepsilon_{42} \end{bmatrix}$$，或是

$$\begin{pmatrix} \underline{y}_1 \\ \underline{y}_2 \end{pmatrix} = \begin{bmatrix} \underline{1} & \underline{0} \\ \underline{0} & \underline{1} \end{bmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \underline{\varepsilon}_1 \\ \underline{\varepsilon}_2 \end{pmatrix}$$，或是

$\underline{y}=X\underline{\beta}+\underline{\varepsilon}$；

只要矩陣的階序（rank order）正確，三種表示方式都是正確的。從第三個表式，我們知道

$$X'X=\begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}，X'\underline{y}=\begin{pmatrix} 280 \\ 201 \end{pmatrix}，$$

所以令 $X'\underline{e}=\underline{0}$，$\underline{\beta}$ 的估計值乃為

$$\underline{b}=(X'X)^{-1}X'\underline{y}=\begin{pmatrix} 0.25 & 0 \\ 0 & 0.25 \end{pmatrix}\begin{pmatrix} 280 \\ 201 \end{pmatrix}=\begin{pmatrix} 70.00 \\ 50.25 \end{pmatrix}，$$

等於我們事先分別計算取得的平均數。這個練習告訴我們，是否分組計算的關鍵不在於是不是「分別」計算，而在於設計矩陣X的安排；我們的 X'X 矩陣顯示，兩個參數之估計是互不相干或正交（orthogonal）的形式，這也是 X 矩陣被稱為設計矩陣的原因。

其次，我們考量計算總平均數與分組平均數的可能，此一計算需要三個參數的模型 $y_{ij}=\alpha+\beta_j+\varepsilon_{ij}$，使用矩陣來表示則為

$$\begin{bmatrix} y_{11} \\ y_{21} \\ y_{31} \\ y_{41} \\ y_{12} \\ y_{22} \\ y_{32} \\ y_{42} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \varepsilon_{31} \\ \varepsilon_{41} \\ \varepsilon_{12} \\ \varepsilon_{22} \\ \varepsilon_{32} \\ \varepsilon_{42} \end{bmatrix} 。$$

由於X矩陣的第二與第三個直行相加等於第一行，我們不能以這形式估計三個參數，也就是 X' X 矩陣不可逆（non-invertible），或者說是這個聯立等式系統無解。為了解決這個問題，我們可以刪減X矩陣對應 β_2 的一行，也就是設定 $\beta_2=0$，則

$$X'X=\begin{pmatrix} 8 & 4 \\ 4 & 4 \end{pmatrix},\ (X'X)^{-1}=\begin{pmatrix} 0.25 & -0.25 \\ -0.25 & 0.25 \end{pmatrix},\ X'\underline{y}=\begin{pmatrix} 481 \\ 280 \end{pmatrix},$$

等式中的第二個下標 j 也就不再需要，而 i＝1，2，…，8 包含全體學生，α與β的估計值 a 與 b 為

$$\underline{b}=\begin{pmatrix} a \\ b \end{pmatrix},\ (X'X)^{-1}X'\underline{y}=\begin{pmatrix} 0.25 & -0.25 \\ -0.25 & 0.25 \end{pmatrix}\begin{pmatrix} 481 \\ 280 \end{pmatrix}=\begin{pmatrix} 50.25 \\ 19.75 \end{pmatrix}$$

換句話說，$y_i=a+bx_i+e_i$，當 $x_i=1$ 表示男性學生時，其體重估計為 $y_i=a+bx_i=50.25+19.75=70.00$ 公斤，乃男性學生的平均體重；當

$x_i=0$ 表示女性學生時，其體重估計為 $y_i=a+bx_i=50.25$ 公斤，乃女性學生的平均體重。這個練習引入了虛擬變項（dummy variable）的概念，但是我們的計算顯示與其使用虛擬變項來估計兩個參數，不若使用分組迴歸估計兩性平均體重來得直截了當。

我們也可以取三個參數模型X矩陣對應β_1與β_2的兩行相減，一樣只估計兩個參數，得總平均體重為60.13公斤，男生平均體重為60.125＋9.875＝70.00公斤，女生平均體重為 60.125－9.875＝50.25公斤，不但事實上同時估算出總平均體重與男女平均體重，而且X'X矩陣為正交形式，兩個參數之估計互不相干。此地，我們引入心理計量經常使用的虛擬變異（dummy variate），事實上估計單變項變異數分析（one-way analysis of variance）的模型

$$y_{ij}=\bar{y}+(\bar{y}_j-\bar{y})+\varepsilon_{ij}\ ;$$

$$\begin{bmatrix} y_{11} \\ y_{21} \\ y_{31} \\ y_{41} \\ y_{12} \\ y_{22} \\ y_{32} \\ y_{42} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & -1 \\ 1 & -1 \\ 1 & -1 \\ 1 & -1 \end{bmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \varepsilon_{31} \\ \varepsilon_{41} \\ \varepsilon_{12} \\ \varepsilon_{22} \\ \varepsilon_{32} \\ \varepsilon_{42} \end{bmatrix} ;$$

$$X'X=\begin{pmatrix} 8 & 0 \\ 0 & 8 \end{pmatrix},\ (X'\underline{y})=\begin{pmatrix} 481 \\ 79 \end{pmatrix},\ b=(X'X)^{-1}X'\underline{y}=\begin{pmatrix} 60.125 \\ 9.875 \end{pmatrix}$$

由於 $\dfrac{\Sigma_j n_j \bar{y}_j}{n}=\bar{y}$，

n_j為已知數值而且 j 只有兩組，既已估計了總平均數，就只能估計兩性平均數與總平均數的差額，所以三個參數的模型無法估計。這個例子說明傳統教科書上的迴歸分析與變異數分析其實是同一個方法，都

是標準線性模型（standard linear model）適用最小平方法求解的結果。

不同於虛擬變項只取 0 與 1 的數值，虛擬變異可取 0、1 與－1 的數值，更由於產生正交的X' X矩陣，估計的參數互不相干，沒有共變（covariation）的問題，使得變異數的拆解成為可能。將總平均數納入估計帶來了一個好處，y_{ij}的變異量可以定義為 $s^2=\Sigma_j\Sigma_i\,(y_{ij}-\bar{y})^2$，由於$\Sigma_j\Sigma_i\,(\bar{y}_j-\bar{y})\,e_{ij}=0$，$\Sigma_j\Sigma_i\,(y_{ij}-\bar{y})^2=\Sigma_j\Sigma_i[(a+bx_{ij}+e_{ij})-\bar{y}]^2=\Sigma_j\Sigma_i\,(\bar{y}_j-\bar{y}+e_{ij})^2=\Sigma_j n_j\,(\bar{y}_j-\bar{y})^2+\Sigma_j\Sigma_i e_{ij}^2$，形成變異數（量）的分解。由於 $e_{ij}=(y_{ij}-\bar{y})-(\bar{y}_j-\bar{y})=y_{ij}-\bar{y}_j$，表示個別體重與組平均的差異，乃為模型所界定的個別差異，而 $y_j-\bar{y}$ 則表示組平均數與總平均數的差異，界定了男女性別差異，所以 y_{ij} 變異數分解為性別差異與個別差異的成份，兩個成份所解釋的變異量相加正好等於總變異量，不多也不少；前者稱為組間變異（between group variation），後者稱為組內變異（within group variation）。如果我們使用虛擬變項，由於a不等於$\bar{y}$，而且X' X矩陣非正交形式（non-diagonal），a與b產生共變，變異數分解就不再有清晰的定義了。

參、多層次模型（Hierarchical Linear Models）

接著，我們考量$\underline{y}'_1$＝（78，67，65，54，52，48）與$\underline{y}'_2$＝（70，65，60，50，48，45）分別為前後兩個時點同一群學生體重之簡單隨機抽樣（simple random samples），分為兩性各三位，均為男性排列在前，女性在後；排列秩序在計算上不重要，但方便閱讀。我們希望瞭解兩性平均體重之變化，可以設立四個參數的模型 $y_{ijk}=\alpha_{jk}+\varepsilon_{ijk}$，j＝1，2 表示性別，k＝1，2 表示兩個時點，以矩陣陳列為

$$\begin{bmatrix} y_{111} \\ y_{211} \\ y_{311} \\ y_{121} \\ y_{221} \\ y_{321} \\ y_{112} \\ y_{212} \\ y_{312} \\ y_{122} \\ y_{222} \\ y_{322} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} \alpha_{11} \\ \alpha_{21} \\ \alpha_{12} \\ \alpha_{22} \end{pmatrix} + \begin{bmatrix} \varepsilon_{111} \\ \varepsilon_{211} \\ \varepsilon_{311} \\ \varepsilon_{121} \\ \varepsilon_{221} \\ \varepsilon_{321} \\ \varepsilon_{112} \\ \varepsilon_{212} \\ \varepsilon_{312} \\ \varepsilon_{122} \\ \varepsilon_{222} \\ \varepsilon_{322} \end{bmatrix}$$

同樣使用最小平方法求解，得α_{11}、α_{21}、α_{12}、α_{22}的估計值，顯示兩個時點間，男女生的平均體重均有下降，減重成功

$$\underline{b} = \begin{pmatrix} a_{11} \\ a_{21} \\ a_{12} \\ a_{22} \end{pmatrix} = (X'X)^{-1}X'\underline{y} = \begin{pmatrix} 1/3 & 0 & 0 & 0 \\ 0 & 1/3 & 0 & 0 \\ 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 1/3 \end{pmatrix} \begin{pmatrix} 210 \\ 164 \\ 200 \\ 143 \end{pmatrix} = \begin{pmatrix} 70.00 \\ 54.67 \\ 66.67 \\ 47.67 \end{pmatrix} 。$$

或許有人以為這是混合兩次抽樣的資料分析（pooled analysis），事實上X矩陣的安排已經明確指出這是分性別分時點的「分別」計算，乃四個平均數之估計彼此互不相干的正交設計（orthogonal design）。我們認為使用矩陣形式來排列資料的好處是研究設計一目瞭然，更可以進一步討論多組資料間的殘差相關（Theil 1971）。

但並不是所有研究設計都是有解的，前一個計算男女總平均與兩性平均的例子就是無解的例子，除非我們增加一些設定條件。同樣估計四個參數，我們的兩個時點資料若估計兩個性別參數與兩個時點參數，$y_{ijk} = \alpha_j + \beta_k + \varepsilon_{ijk}$，或

$$\begin{bmatrix} y_{111} \\ y_{211} \\ y_{311} \\ y_{121} \\ y_{221} \\ y_{321} \\ y_{112} \\ y_{212} \\ y_{312} \\ y_{122} \\ y_{222} \\ y_{322} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{bmatrix} \varepsilon_{111} \\ \varepsilon_{211} \\ \varepsilon_{311} \\ \varepsilon_{121} \\ \varepsilon_{221} \\ \varepsilon_{321} \\ \varepsilon_{112} \\ \varepsilon_{212} \\ \varepsilon_{312} \\ \varepsilon_{122} \\ \varepsilon_{222} \\ \varepsilon_{322} \end{bmatrix} 。$$

則 X 矩陣顯示，由於第一與第二行相加等於第三與第四行相加的結果，我們不能以此形式估計四個參數。使用虛擬變項求解，比照傳統的參數設定，我們設定一個截距 α 對應著一行 1，一個性別參數 β 對應於男生 $x_{ijk}=1$，女生 $x_{ijk}=0$；一個時點參數 γ 對應於第一時點 $z_{ijk}=1$，第二時點 $z_{ijk}=0$；再加個交互作用參數 λ，對應於第一時點的男生體重 $w_{ijk}=x_{ijk}z_{ijk}=1$，其餘 $w_{ijk}=x_{ijk}z_{ijk}=0$；我們的虛擬變項模型乃為

$$y_{ijk}=\alpha+\beta x_{ijk}+\gamma z_{ijk}+\lambda w_{ijk}+\varepsilon_{ijk}，$$

也可去除 jk 下標，表示為

$y_i=\alpha+\beta x_i+\gamma z_i+\lambda w_i+\varepsilon_i$，$i=1，2，\cdots，n$，$n=\Sigma_j n_j=\Sigma_k n_k=\Sigma_j\Sigma_k n_{jk}$；或以矩陣陳列為

$$\begin{bmatrix} y_{111} \\ y_{211} \\ y_{311} \\ y_{121} \\ y_{221} \\ y_{321} \\ y_{112} \\ y_{212} \\ y_{312} \\ y_{122} \\ y_{222} \\ y_{322} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \\ \lambda \end{pmatrix} + \begin{bmatrix} \varepsilon_{111} \\ \varepsilon_{211} \\ \varepsilon_{311} \\ \varepsilon_{121} \\ \varepsilon_{221} \\ \varepsilon_{321} \\ \varepsilon_{112} \\ \varepsilon_{212} \\ \varepsilon_{312} \\ \varepsilon_{122} \\ \varepsilon_{222} \\ \varepsilon_{322} \end{bmatrix} 。$$

令 $X'\underline{e}=\underline{0}$，以最小平方法求解，

$$X'X=\begin{pmatrix} 12 & 6 & 6 & 3 \\ 6 & 6 & 3 & 3 \\ 6 & 3 & 6 & 3 \\ 3 & 3 & 3 & 3 \end{pmatrix}，X'\underline{y}=\begin{pmatrix} 717 \\ 307 \\ 374 \\ 210 \end{pmatrix}，\underline{b}=(X'X)^{-1}X'\underline{y}=\begin{pmatrix} 47.67 \\ 19.00 \\ 7.00 \\ -3.67 \end{pmatrix}；$$

α 表示不分性別與時點的截距，β 表示性別參數，γ 表示時點參數，而λ則表示交互參數。代入估計式則第一時點的男生體重估計由於 $x_i=1$、$z_i=1$、$w_i=1$，得平均體重為 47.67＋19.00＋7.00－3.67＝70.00 公斤，女生體重估計由於 $x_i=0$、$z_i=1$、$w_i=0$，得平均體重為 47.67＋7.00＝54.67 公斤；第二時點的男生體重估計由於 $x_i=1$、$z_i=0$、$w_i=0$，得平均體重為 47.67＋19.00＝66.67 公斤，女生則 $x_i=0$、$z_i=0$、$w_i=0$，得平均體重為 47.67 公斤。同樣使用四個參數，虛擬變項估計結果經過詮釋後，正好等於分組迴歸四個平均數，但是產生非正交的 X' X 矩陣，不若採用分組迴歸來得直截了當。

晚近以來，利用虛擬變項與分組迴歸的關係，多層次線性模型可

以先設立性別平均數的模型 $y_{ij}=\alpha_j+\beta_j x_{ij}+\varepsilon_{ij}$，進一步則設立性別平均數在兩個時點的變化為 $\alpha_j=\alpha_0+\alpha_1 z_j+\omega_j$，$\beta_j=\beta_0+\beta_1 z_j+\upsilon_j$，其中 ε_{ij}、ω_j 與 υ_j 均為殘差項，x_{ij} 與 z_j 則均為虛擬變項。將時點變化模型代入性別模型中，

$$\begin{aligned} y_{ij} &= (\alpha_0+\alpha_1 z_j+\omega_j)+(\beta_0+\beta_1 z_j+\upsilon_j)x_{ij}+\varepsilon_{ij} \\ &= \alpha_0+\alpha_1 z_j+\beta_0 x_{ij}+\beta_1 x_{ij} z_j+(x_{ij}\upsilon_j+\omega_j+\varepsilon_{ij}) \\ &= \alpha_0+\alpha_1 z_j+\beta_0 x_{ij}+\beta_1 x_{ij} z_j+\eta_{ij}, \end{aligned}$$

η_{ij} 為新的殘差項，α_0、α_1、β_0、與 β_1為四個可以估計的參數；由於參數設定完全相同，多層次模型的參數估計等同於虛擬變項迴歸的結果。當 $x_{ij}=1$ 代表男性而 $x_{ij}=0$ 代表女性，$z_j=1$ 代表第一時點而 $z_j=0$ 代表第二時點時，第一時點的男性平均數為 $a_0+a_1+b_0+b_1=70.00$ 公斤，女性平均數為 $a_0+a_1=54.67$ 公斤，第二時點的男性平均數為 $a_0+b_0=66.67$ 公斤，女性平均數為 $a_0=47.67$ 公斤。這個模型有助於跨越「分析單元」，連結不同層次的變項與參數，所以稱為多層次線性模型。事實上，由於「層次」交錯而產生了時點與性別的「交互作用」，此一模型雖與虛擬變項迴歸雷同，卻因使用「層次」的概念而擴大應用的範圍。

肆、研究設計

我們可以設想此地的兩個時點為實驗前後的兩個時點，而除減重實驗之變項外，性別變項也是重要的考量，套用變異數分析的模型則

$$y_{ijk}-\bar{y}=(\bar{y}_j-\bar{y})+(\bar{y}_k-\bar{y})+(\bar{y}_{jk}-\bar{y}_j-\bar{y}_k+\bar{y})+\varepsilon_{ijk}$$

表示除總平均數 $\bar{y}$ 外，共需估計 y_j、y_k、與 y_{jk} 八個參數，事實上解開模型等式後

$$y_{ijk}-\bar{y}=(\bar{y}_{jk}-\bar{y})+\varepsilon_{ijk}$$

只能估計四個參數。但若只估計四個參數，我們可以考慮原有八個參數的某些組合，不必侷限於使用實驗前後的兩性平均數為參數；利用虛擬變異來解決問題，

$$\begin{bmatrix} y_{111} \\ y_{211} \\ y_{311} \\ y_{121} \\ y_{221} \\ y_{321} \\ y_{112} \\ y_{212} \\ y_{312} \\ y_{122} \\ y_{222} \\ y_{322} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \\ \lambda \end{pmatrix} + \begin{bmatrix} \varepsilon_{111} \\ \varepsilon_{211} \\ \varepsilon_{311} \\ \varepsilon_{121} \\ \varepsilon_{221} \\ \varepsilon_{321} \\ \varepsilon_{112} \\ \varepsilon_{212} \\ \varepsilon_{312} \\ \varepsilon_{122} \\ \varepsilon_{222} \\ \varepsilon_{322} \end{bmatrix}$$

指出可以估計一個總平均數參數 α，一個實驗參數 β，一個性別參數 γ，與一個交互作用參數 λ。如前所述，由於

$$\frac{\Sigma_j n_j \bar{y}_j}{n}=\frac{\Sigma_k n_k \bar{y}_k}{n}=\bar{y}$$，而且 $$\frac{\Sigma_j n_{jk} \bar{y}_{jk}}{n_j}=\bar{y}_j$$，$$\frac{\Sigma_k n_{jk} \bar{y}_k}{n_k}=\bar{y}_k$$，

n_j、n_k、n_{jk}、與 n 均為已知的條件，除總平均數外，我們只需要估計一個實驗前後平均數與總平均數的差額、一個性別平均數與總平均數的差額、以及一個交互作用項與前兩項差額之差額，就足以完整代表變異數分析的模型。將實驗前後學生體重資料代入計算，同樣以最小平方法求解，設定 $X'\underline{e}=\underline{0}$，

$$X'X=\begin{pmatrix}12&0&0&0\\0&12&0&0\\0&0&12&0\\0&0&0&12\end{pmatrix},\ X'\underline{y}=\begin{pmatrix}717\\31\\103\\-11\end{pmatrix},$$

$$\underline{b}=(X'X)^{-1}X'\underline{y}=\begin{pmatrix}a\\b\\c\\d\end{pmatrix}=\begin{pmatrix}59.75\\2.58\\8.58\\-0.92\end{pmatrix}。$$

換句話說，實驗前男生平均體重為 $a+b+c+d=70.00$ 公斤，女生平均體重為 $a+b-c-d=54.67$ 公斤；實驗後男生平均體重為 $a-b+c-d=66.67$ 公斤，女生平均體重為 $a-b-c+d=47.67$ 公斤。同樣使用四個參數，分組迴歸、虛擬變項、與變異數分析所得到的參數估計值均不相同，根據參數所計算出來的分組平均數卻均相同。只是不同於虛擬變項之運用，分組迴歸與變異數分析模型產生正交的設計矩陣，保證迴歸等式中，等號右邊的解釋變項彼此互不相關，解釋變項與被解釋變項的相關係數等於標準化迴歸係數，RSQ 與變異數分解也不會有解釋變項彼此相關而造成「剪不斷理還亂」的結果。這在實驗設計中是非常重要的原理，在社會科學與管理科學的研究設計中卻經常被有意或無意地忽視，更經常以「效果」來稱呼其所估算的參數，完全無視於研究設計與統計理論的基礎原理，值得進一步的檢討。即使參數數量相同使得設計矩陣階序相同，不同的設計產生不同的參數估計，其意義也對應不同，設計矩陣之排列乃為研究計劃的核心工作。

但是此地也需要指出，使用虛擬變異設定變異數分析模型時，$X'X$ 矩陣為對角線矩陣（正交形式）還需要滿足各分組樣本規模 n_{jk} 相等的條件（Iversen and Norpoth 1987; Turner and Thayer 2001），否則 X 矩陣任意不同兩行相乘不會產生正負相抵而累加為零的結果，引進了實驗設計的考量。在變異數分解時，既然 $s^2=\Sigma_i\Sigma_j\Sigma_k(y_{ijk}-\bar{y})^2=$

$\Sigma_i\Sigma_j\Sigma_k y^2_{ijk} - n\bar{y}^2$，而$\Sigma_i\Sigma_j\Sigma_k y^2_{ijk} = \underline{b}'\,X'\,X\underline{b} + \underline{b}'\,X'\,\underline{e} + \underline{e}'\,X\underline{b} + \underline{e}'\,\underline{e}$，其中$X'\,\underline{e} = (\underline{e}'X)' = \underline{0}$係最小平方法求解時必有的條件，則

$$s^2 = \underline{b}'\,X'\,X\underline{b} + \underline{e}'\,e - n\bar{y}^2。$$

我們的例子顯示，當X' X矩陣為正交的對角線矩陣時，主對角線上的數值均為 n，其餘為 0，則 $\underline{b}'\,X'\,X\underline{b}$ 實為 $na^2 + nb^2 + nc^2 + nd^2$而已；既然 $a = \bar{y}$，

$$s^2 = nb^2 + nc^2 + nd^2 + \Sigma_i\Sigma_j\Sigma_k e^2_{ijk}。$$

$\Sigma_i\Sigma_j\Sigma_k e^2_{ijk}$ 為最小平方法所取得的最小平方，在變異數分析架構下，我們稱之為個別差異；nb^2 為減重實驗的「效果」，nc^2 為性別差異的影響，nd^2 為減重實驗對於性別差異的影響。等號兩邊均除以 s^2形成 R^2 之分解，等號左邊為 1，右邊的（$nb^2 + nd^2 + nc^2$）/s^2 為 R^2，可以更細分為實驗、性別、與交互作用之 R^2，均為大於 0 小於 1 的正值小數，$\Sigma_i\Sigma_j\Sigma_k e^2_{ijk}/s^2 = 1 - R^2$。

除了變異數與RSQ分解外，正交的設計矩陣在構成抽樣分配時也非常重要。標準線性模型設定

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}，E(\underline{\varepsilon}) = 0，E(\underline{\varepsilon}\underline{\varepsilon}') = \sigma^2 I$$

的條件，我們的 $\underline{\beta}$ 估計值 $\underline{b}$ 乃據之而有抽樣分配為

$$E(\underline{b} - \underline{\beta})(\underline{b} - \underline{\beta})' = E[(X'X)^{-1}X'(X\underline{\beta} + \underline{\varepsilon}) - \underline{\beta}][(X'X)^{-1}X'(X\underline{\beta} + \underline{\varepsilon}) - \underline{\beta}]'$$
$$= E(X'X)^{-1}X'\underline{\varepsilon}\underline{\varepsilon}'X(X'X)^{-1} = \sigma^2(X'X)^{-1}。$$

當X' X為對角線矩陣時，$\underline{b}' = (a，b，c，d)$數列中的每一個參數均有獨立的抽樣分配，互不相干，則驗證假設 $\underline{b} = \underline{\beta}$ 時，可以個別參數為之而無問題；但若X' X矩陣不是對角線矩陣，個別參數適用統計分配

$$z=\frac{b-\beta}{\sqrt{\sigma^2/n}}$$，或是 $$t=\frac{b-\beta}{\sqrt{\hat{\sigma}^2/n}}$$

就有問題了。

為了說明 n_{jk} 不等的結果，我們取兩個時點的女生樣本各減掉一位女生，得 $\underline{y}'_1$＝（78，67，65，54，52）與 $\underline{y}'_2$＝（70，65，60，50，48）。從 $\underline{y}'_1$ 我們知道實驗前男女生平均體重分別為 70.00 與 53.00 公斤，實驗後 $\underline{y}'_2$ 顯示男女生體重分別為 66.67 與 49.00 公斤。變異數分析的設計矩陣為

$$X=\begin{bmatrix}1&1&1&1\\1&1&1&1\\1&1&1&1\\1&-1&1&-1\\1&-1&1&-1\\1&1&-1&-1\\1&1&-1&-1\\1&1&-1&-1\\1&-1&-1&1\\1&-1&-1&1\end{bmatrix}，X'X=\begin{pmatrix}10&2&0&0\\0&2&10&0\\0&0&10&2\\0&0&2&10\end{pmatrix}，$$

雖然有解，卻不是正交的解，也不是正確的解。估算得實驗前男女生平均體重分別為 70.00 與 52.67 公斤，實驗後則為 66.33 與 49.00 公斤，顯然 α 與 β、及 λ 與 γ 參數間的共變影響了參數估計，參數檢證時所使用的抽樣分配也受到影響，更不能適用上述變異數分析 R^2 分解的結論，即使 $n_{jk}=n_jn_k/n$ 令次數分佈滿足 $\chi^2=0$ 的條件。

由於 X' X 矩陣涵蘊著迴歸變項間的相關係數，此一問題經常被解釋為共線性（colinearity）的問題，則共線性是有無，而非大小的問題。但若我們採用分組迴歸的設計，

$$X=\begin{bmatrix}1&0&0&0\\1&0&0&0\\1&0&0&0\\0&1&0&0\\0&1&0&0\\0&0&1&0\\0&0&1&0\\0&0&1&0\\0&0&0&1\\0&0&0&1\end{bmatrix}，X'X=\begin{pmatrix}3&0&0&0\\0&2&0&0\\0&0&3&0\\0&0&0&2\end{pmatrix}，$$

則設計矩陣仍為正交形式，由於參數之間未有共變問題，可以取得正確的結果，實驗前男女生平均體重分別為 70.00 與 53.00 公斤，實驗後則分別為 66.67 與 49.00 公斤，參數檢證可以使用各自獨立的抽樣分配，R^2分解也沒有剪不斷理還亂的問題，也可透過刪減參數再估計模型的方式建構 F 統計數，檢定實驗與性別各自獨立的影響（Fisher 1970）。

伍、次數分佈

我們需要考察某事件出現的比例 q，共 n 次觀察中出現的次數為 m，設每出現一次給 1 分，未出現給 0 分，事實上是分析名目尺度的資料，每次觀察得 $y_i=1$ 或 0，

$$\bar{y}=\frac{\sum_{i=1}^{n}y_i}{n}=q=\frac{m}{n}1+\frac{n-m}{n}0=\frac{m}{n}\ 。$$

母體中觀察該事件出現的機率為 $\mu=E(y_i)=\Sigma_i p(y_i)y_i$，$E(y_i)$表示 y_i 的統計期望值，$p(y_i)$則表示特定 y_i 值的出現機率；在簡單隨機抽樣的條件下，適用抽樣分配的理論，可以建構 q 與 μ 間的推論程序。既然已知樣本平均數為 $\bar{y}$，重複抽樣取得無數個 $\bar{y}$ 形成抽樣分配後，這

個分配可以驗證為一常態分配，其平均數等於母體平均數，而其變異數為母體變異數σ^2_y除以樣本規模n，n愈大則抽樣分配的變異數愈小，樣本平均數趨近於母體平均數μ，

$$\lim_{n\to\infty}\frac{\sigma_y^2}{n}=0，$$

我們稱之為中限定理（central limit theorem）。

以8次觀察而有5次事件出現而言，我們的y_i＝（1，0，0，1，0，1，1，1），樣本平均數為 q＝5/8＝0.625，假設母體平均數μ＝0.5，則

$$\sigma_y^2=E(y_i^2)-\mu^2=0.5-0.25=0.25，$$

所以檢測樣本比例q與母數μ的差距，$z=(0.625-0.5)/(0.25/8)^{1/2}=0.7071$；查常態分配表得 $p(-1.96\leqq z\leqq 1.96)=0.95$，不能在0.05的顯著水準下接受母數$\mu\neq 0.5$而拒斥$\mu=0.5$的假設。

一般列聯表（contingency table）分析的重點在於統計獨立（statistical independence）的檢定，例如以p_i表示父親教育程度的比例分佈，p_j表示兒子教育程度的比例分佈，i，j＝1，2，…，k，統計獨立的條件是兒子教育程度並不因父親教育程度而有不同，套用貝氏定理（Bayes' Theorem）表示為$p(j/i)=p_{ij}/p_i=p_j$，也就是$p_{ij}=p_ip_j$。表2使用統計獨立的設定條件，設計出兩代間教育程度變化的分佈。由於表1的基礎設定是$p_{ij}=p_ip_j$，所以傳統的列聯表統計如卡方分析等是沒有用處的，我們可以直接指出$\chi^2=0$。但是表2雖然設定了兒子教育程度不因父親教育而不同的條件，兩代間教育程度分佈卻發生了劇烈的變化，我們特意設定子代教育程度分佈翻轉父代教育程度分佈的情況，從原本集中低階教育的分佈翻轉為集中高階，而此地的任務乃是找到合適的模型，使用參數充分表現這兩個設定條件。

既然$p_{ij}=p_ip_j$，則$\log p_{ij}=\log p_i+\log p_j+\varepsilon_{ij}=\alpha_i+\beta_j+\varepsilon_{ij}$乃是最合

表 2　父子兩代間的教育程度變遷

父親教育程度	兒子教育程度				
	國小以下	初中職	高中職	大專以上	小計
國小以下	0.042	0.083	0.125	0.250	0.500
初中職	0.021	0.042	0.063	0.125	0.250
高中職	0.014	0.028	0.042	0.083	0.167
大專以上	0.007	0.014	0.021	0.042	0.083
小計	0.083	0.167	0.250	0.500	1.000

適的模型。由於我們同時設定 $\alpha_1=\beta_4$、$\alpha_2=\beta_3$、$\alpha_3=\beta_2$、與 $\alpha_4=\beta_1$的條件，此一模型可以矩陣表示為

$$\underline{y}=\begin{bmatrix}\log\ p_{11}\\ \log\ p_{12}\\ \log\ p_{13}\\ \log\ p_{14}\\ \log\ p_{21}\\ \log\ p_{22}\\ \log\ p_{23}\\ \log\ p_{24}\\ \log\ p_{31}\\ \log\ p_{32}\\ \log\ p_{33}\\ \log\ p_{34}\\ \log\ p_{41}\\ \log\ p_{42}\\ \log\ p_{43}\\ \log\ p_{44}\end{bmatrix}=X\underline{\beta}+\underline{\varepsilon}=\begin{bmatrix}1&0&0&1\\0&1&0&1\\0&0&1&1\\0&0&0&2\\1&0&1&0\\0&1&1&0\\0&0&2&0\\0&0&1&1\\1&1&0&0\\0&2&0&0\\0&1&1&0\\0&1&0&1\\2&0&0&0\\1&1&0&0\\1&0&1&0\\1&0&0&1\end{bmatrix}\begin{pmatrix}\beta_1\\ \beta_2\\ \beta_3\\ \beta_4\end{pmatrix}+\begin{bmatrix}\varepsilon_{11}\\ \varepsilon_{12}\\ \varepsilon_{13}\\ \varepsilon_{14}\\ \varepsilon_{21}\\ \varepsilon_{22}\\ \varepsilon_{23}\\ \varepsilon_{24}\\ \varepsilon_{31}\\ \varepsilon_{32}\\ \varepsilon_{33}\\ \varepsilon_{34}\\ \varepsilon_{41}\\ \varepsilon_{42}\\ \varepsilon_{43}\\ \varepsilon_{44}\end{bmatrix}。$$

使用表 2 數據代入運算，適用最小平方法，得

$$(X'X)=\begin{pmatrix}10 & 2 & 2 & 2\\ 2 & 10 & 2 & 2\\ 2 & 2 & 10 & 2\\ 2 & 2 & 2 & 10\end{pmatrix},\ X'\underline{y}=\begin{pmatrix}-32.591\\ -27.046\\ -23.803\\ -18.257\end{pmatrix},$$

$$\underline{\beta}\text{之估計值為 }\underline{b}=(X'X)^{-1}X'\underline{y}=\begin{pmatrix}-2.485\\ -1.792\\ -1.386\\ -0.693\end{pmatrix},$$

也就是 $p_j=\exp(b_j)=(0.083，0.167，0.250，0.500)$，所以 $p_i=(0.500，0.250，0.167，0.083)$。既然表 1 比例值均為根據上述兩個條件而設立的數據，無殘差項來表現其他條件與抽樣誤差所產生的影響，所以 $\varepsilon_{ij}=0$，而 RSQ＝1.000。

由於 $p_{ij}=n_{ij}/n$，$p_i=n_i/n$，$p_j=n_j/n$，$n=\Sigma_i n_i=\Sigma_j n_j=\Sigma_i\Sigma_j n_{ij}$，我們所使用的方法其實不限於比例或機率，將上列矩陣模型中的 $\log p_{ij}$ 改為 $\log n_{ij}$，直接使用次數分佈從事分析，β_j 參數估計值 b_j 併入 n 的乘數作用，$n_j=n^{1/2}\exp(b_j)$，其餘結果相同。但是既然 $n_{ij}=n_i n_j/n$，

$$\begin{aligned}\log n_{ij}&=\log n_i+\log n_j-\log n+\varepsilon_{ij}\\ &=\alpha_i+\beta_j+\gamma+\varepsilon_{ij}，\end{aligned}$$

理論上應可設立模型來估計 n 的乘數作用參數 γ；如同前述平均數分析中總平均數與組平均數之估計般，這個模型是無解的，我們無法在不增加設定的條件下同時估計 α_i、β_j 與 γ 參數。如同總平均數與組平均數的估計般，增加設定則改變模型，也改變設計矩陣 X 的安排，突顯 X 矩陣的「設計」特性。

當然，表 2 數據若非我們特意設計的結果，而是來自抽樣調查的數據，就必需容許 $\varepsilon_{ij}\neq 0$ 而 $0\leqq RSQ\leqq 1$ 的估計，模型設定（也就是參數設定，或設計矩陣的安排）也就成為研究計劃的關鍵工作了。我們的兩代教育程度變遷例子指出，對於資料分析而言，重要的並不是計

量尺度與分析方法之選擇，而是設計矩陣之安排，這也是概化線性模型發展的主要成果。同時，這個例子也指出執著於「分析單元」可能產生誤導的結果；從資料構造看來，此一分析似以子代個體為單元，從表 2 及其分析模型看來，則分析單元似為成對的父子，而不是子代個體。進一步而言，個體屬性並無「機率」可言，需在特定條件下形成「可重複試驗」的事件才有機率可言，則科學分析的單元不可能為特定個體。更重要地，這個例子指出最小平方法仍然是「最好」（最小平方和）的方法，而次數或比例資料也沒有一定要使用邏輯迴歸（logit regression）的道理。

陸、討論

本文使用同一套最小平方法適用比例尺度與名目尺度的資料，順序尺度的資料視為有順序差別的分類，其分析也可適用同一套方法（Goodman 1984），指出不同尺度資料適用不同方法的主張是莫需有的，而分組或分期資料不能混合分析則更是削足適履，自尋煩惱。我們認為資料分析的關鍵在於設計矩陣，X'X 相對於 X 矩陣更有縮小階序的作用，研究設計一目瞭然，捨之則標的不明，自然陷入與電腦打迷糊戰的境地。我們建議從事資料分析的研究者在形成研究假設時，務必就參數設定進一步考量其設計矩陣的各種可能安排，則至少不會再發生有參數估計與文字說明背道而馳的現象。我們覺得研究與分析工作是思考性的工作，而不是將資料輸入電腦抄出數據的機械性工作；仔細思考規劃設計矩陣以後，只要計算一群乘積和就能取得 X'X 與 X'y 的矩陣，輸入工作表就能從事各種矩陣運算取得參數估計與假設驗證的結果，並不受限於資料規模，程序簡便而明瞭。

我們的討論限於線性代數之應用，模型設立以線性代數來陳述，所引用的理論完全是線性代數的文獻，而一般統計分析也適用同一套程序與文獻。只是這並不表示非線性模型不能適用這套方法，例如本

文所討論的流動表分析基本上乃是非線性模型，$n_{ij}=\gamma\alpha_i\beta_j\varepsilon_{ij}$等式兩邊均取對數後，即可代入線性模型求解。同樣的，人口學文獻中有名的Brass Equation 指出經驗率 $p(x)$ 與標準率 $q(x)$ 的關係可以表示為

$$\frac{p(x)}{1-p(x)}=\alpha\left[\frac{q(x)}{1-q(x)}\right]^{\beta}，$$

等式兩邊取對數後可以線性代數求解。比較麻煩的如已婚率為年齡的邏輯函數（logistic function），

$$0\le G(x)=\frac{k}{1+Ae^{-\alpha x}}\le 1，$$

$$\lim_{x\to\infty} G(x)=k，$$

k、A、與α均為必需估算的參數，無法以簡單的數據換算（transformation）來取得線性的估計。已知的方法是猜測 k 值，代入等式中取得

$$\frac{k-G(x)}{G(x)}=Ae^{-\alpha x}，$$

兩邊取對數求解。我們可以不斷猜測 k 值取得對應的最小平方和或RSQ，以產生最大RSQ或最小的最小平方和之一組k、A、與α為「最佳」的一組解，文獻上稱為重複計算的最小平方法（iterative least squares）。換句話說，雖然本文程序與理論立基於線性代數之應用，並不限於線性模型之求解。

參考文獻

Fisher, Franklin. 1970. "Tests of Equality between Sets of Coefficients in-Two Linear Regressions: An Expository Notes." *Econometrica* 38 (March): 361-366.

Goodman, L. 1984. *The Analysis of Cross-Classified Data Having Ordered Categories*. Cambridge: Harvard University Press.

Iversen,Gudmund R. and Helmut Norpoth. 1987. *Analysis of Variance*. Thousand Oaks, CA: Sage Publications, Inc.

McCullagh, P. and J. A. Nelder. 1999. *Generalized Linear Models*, Second Ed. Boca Raton, FL: Chapman & Hall, CRC Press.

Raudenbush, S. W. and A. S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*, Second Ed. Thousand Oaks, CA: Sage Publications, Inc.

Theil, H. 1971. *Principles of Econometrics*. New York: John Wiley & Sons.

Turner, J. Rick and Julian Thayer. 2001. *Introduction to Analysis of Variance: Design, Analysis & Interpretation*. Thousand Oaks, CA: Sage Publications, Inc.

Design Matrix and Data Analysis: An Application of Linear Algebra

Kuanjeng Chen* Chingli Yang**

Abstract

Introducing the concept of a "design matrix", this paper relates the regression analysis, the analysis of variance, and the analysis of frequency counts to the use of the least squares method. It is demonstrated with some simplified examples showing that, given a properly arranged design matrix, the analyses of means, probabilities, and frequencies share the same least squares approach as that of the regression analysis. Special attention is called to the "orthogonal" design matrix which results in independent parameter estimations jointly. Researchers are urged to review the possible arrangements of the design matrix before actually doing on the estimation.

Keywords: regression analysis, design matrix, measurement scales, unit of analysis, generalized linear model

* Professor, Department of Healthcare Management, Chang Gung University. Taoyuan, Taiwan.
** Associate Professor, Institute of Gerontology, College of Medicine, National Cheng Kung University. Tainan, Taiwan.