GEE之敏感度分析—偵測高影響之觀察值

張玉坤

高影響觀察值 (high influential observations) 的確認 (identification) 在統計迴歸模型 (regression model) 的應用上有其不容置疑重要性。在早期的一般線性模型 (general linear model)及近幾年來被多位學者廣泛探討的廣義線性模型 (generalized linear model) 中，對此問題已有多篇論文發表。但是，處理長期資料 (longitudinal data) 線性模型的統計方法 -- GEE (generalized estimating equation) [1]，對此問題至今尚未見任何有關之論文刊載。本文對此問題提出一個簡單可行的圖形判讀法，並將原來的SAS/IML Macro程式，GEE1，加以修改後納入此項功能，以利原使用者之應用。我們也成功地將此方法應用在台灣省立新竹醫院眼科的一組資料上。(中華衛誌 1996；**15**(5)：403-410)

關鍵字：敏感度分析，高影響觀察值，廣義線性模型，長期資料。

前　言

在迴歸分析的方法中，如何檢查所選定模型的適切性，是一項不容忽視的重要工作。蓋因，迴歸分析如果未經任何適切性檢查或殘餘數分析 (residual analysis) 等模型診斷 (model diagnostics) 的檢驗過程，其所有相關之統計推論總是難免會令人有所置疑。一般所謂的模型診斷，所含蓋的範圍甚廣，例如：(1) 自變項 (independent variables) 之選用是否適當？是否需轉換(transformation)？是否有遺漏重要項目？．．．等 (2) 依變項 (dependent variable) 所選用的連繫函數 (link function) 是否適當？(3) 所收集資料中是否有異常值 (outliers) 或具高影響之觀察值 (high influential observations) 存在？．．．等等。

進行檢查的方式，大致可分為非正式 (informal) 及正式 (formal) 兩大類[2]。前者大都以繪圖方式，利用視覺來判讀是否具有某種趨勢 (pattern)或異常 (abnormal) 現象，故檢查成效有賴個人豐富經驗。最常見的有：residuals plots, half-normal plots of residuals，．．．等等。後者則是經由增項，亦即是將現有的模型 (current model) 中增加一項新的變項及未知參數(parameter)，來進行。進一步說明如下：假設此一新增未知參數為 θ，而 θ_0為現有模型 (current model) 中對應於該新增項之參數值。則正式 (formal)法往往在擴大後之參數空間 (parameter space)，即增項後之模型中迴歸係數之所有可能值，中尋求 θ的"最佳"估計量，$\hat{\theta}$。然後比較 $\hat{\theta}$與 θ_0以"檢驗 (testing)"此新增項是否有顯著地改善 (improve) 現有模型 (current model)。此類方法的典型代表有 added variable plots[3], partial residuals plots[4]及 constructed variable plots[5]等。O'Hara Hines and Carter[6]又將前兩者改進後用來偵測廣義線性模型 (generalized linear

聯絡單位：中央研究院統計科學研究所
聯絡人：張玉坤
聯絡地址：台北市南港區研究院路二段128號
聯絡電話：(02)783-5611 轉 305
傳真：(02)783-1523
投稿日期：84年10月
接受日期：85年 3 月

models)中高影響觀察值 (high influential observations)。最近，Hall, Zeger and Bandeen-Roche[7]更將 O'Hara Hines and Carter 的方法略加修正，用來處理相關性資料 (dependent data) 中高影響力 (influence) 及高位能 (leverage) 的問題。但因所用方法在觀念上較難讓非統計領域者理解，且文中並未提及所用之程式為何，僅在所附圖形之文字中提到以added variable plot 的方式描點。對國內公衛及醫學界人士在使用上較為困難。對於相關性資料，如長期性資料及家族資料…等之分析，國內學者較熟悉者為 Liang & Zeger 之方法。Liang & Zeger[1]係將具相關性的一組觀察值 (observations) 當成一個 cluster (如長期資料的同一人之多筆觀察值及來自同一家庭所有成員的多筆觀察值之家族資料等)，並借由一個虛構的工作相關矩陣 (working correlation matrix) 來代表在同一 cluster 內觀察值間的相關性。然後，再將這些理論架構納入他們所提的廣義估計函數式 (generalized estimating equation，簡稱 GEE) 中，以數值分析的方法解得線性模型中的係數估計量，β。目前，此方面可供利用的程式有 SAS/IML 版及 S-Plus 版兩種。其中較常使用的是前者，它是利用 SAS/IML 所寫的一個 Macro 程式，稱為 GEE1.SAS，並備有詳細之使用說明檔，GEE1.DOC，可供參考。上列程式因原作者開放提供給外界使用，有興趣者可向他們或筆者函索。然而遺憾的是，Liang & Zeger 之法在模型診斷方面，至今仍未有具說服力之方法提出。

本文針對此一缺失，提出一種簡易圖形判讀法，來處理長期資料中高影響觀察值的問題。我們將原來分析長期資料廣義線性模型程式，GEE1，修改後加入此一功能。原使用者僅需增加一個 Argument 及幾道繪圖指令，即能獲得此圖以供判讀。第二節中我們先回顧幾種類似的方法，於第三節中詳述本文所提方法，第四節中我們將此方法應用在台灣省立新竹醫院眼科的一組資料上，結語及相關之討論則放在第五節。

方法回顧

一組資料 (data set) 中，如果將某一觀察值 (observation) 由資料中去除後，會造成迴歸結果顯著差異而導致迴歸分析上的重大改變，此一觀察值我們稱之為高影響觀察值 (high influential observation)。由此定義可知，一種最直觀而具說服力的偵測方法為每次去除一個觀察值，然後"測量"去除前後迴歸模型的差異大小。此"差異量"的大小即可定義為該觀察值的"影響力"大小。

重點是，如何定義此一具說服力之量度標準？在此方面較具代表性的有 Cook's distance[8]。目前已有多種統計套裝軟體能直接或間接求得 Cook's distance。其中較常用者有 GLIM, Genstat, SAS,及 BMDP 等〔5〕。但是，上列各種軟體僅能處理具統計上獨力性 (independence) 之資料 (data)，對長期資料 (longitudinal data) 等不具統計上獨力性之資料則不適用。

為便於爾後討論之進行，我們先簡單介紹 Cook's distance。考慮下列的線性迴歸問題：

$$E(Y) = X\beta \ ; \ Var(Y) = \sigma^2 I_n,$$

其中 Y 是 $n \times 1$ 的依變項 (dependent variable) 所成的隨機向量 (random vector)，X 為 $n \times p$ 的設計矩陣 (design matrix)，σ^2 為共同變異數，I_n 為 $n \times n$ 的單位矩陣。則 β 的最小平方估計統計量 (least squares estimator) 為

$$\hat{\beta} = (X^T X)^{-1} X^T Y .$$

現假設 $\hat{\beta}_{(i)}$ 為去除第 i 個觀測值後，β 的最小平方估計統計量。則 Cook's distance 定義為

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T (X^T X) (\hat{\beta}_{(i)} - \hat{\beta})}{p\hat{\sigma}^2} \qquad (2.1)$$

其中 $\hat{\sigma}^2 = RSS/(n-p)$，且 $RSS = Y^T Y - \hat{\beta}^T (X^T X) \hat{\beta}$。此式提供了一個直觀且具說服力的測量"影響力"的量度標準。因為迴歸分析之重心在於如何估計迴歸係數，β。故影響力的大小應架構在迴歸係數因第 i 個觀測值去除後所產生的變化量的大小上。變化量越

大，表示該觀測值的影響力越大。除此之外，迴歸係數估計量的變異數(variance) 大小，亦應列入考量的準則。上式，(2.1)，中也具備此一性質。

Belsley, Kuh, and Welsch[9]對此問題亦提出類似的統計量，稱為 $DFBETAS_i$。其定義如下：

$$(DFBETAS_i)^2 = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T (X^T X)(\hat{\beta}_{(i)} - \hat{\beta}_{(i)})}{\hat{\sigma}^2_{(i)}} \quad (2.2)$$

由上二式可看出，此兩種統計量用來測量觀察值之"影響力"的結果大致相同。但是 D_i 及 $DFBETAS_i$ 僅適用在傳統的迴歸分析上。對於長期資料方面，則需將資料間之相關性列入考量。

方　法

現考慮在長期資料 (longitudinal data) 的架構下，來自同一個 cluster的觀察值，彼此間不具統計上之獨力性 (independence)。假設 $\hat{\beta}_g$ 為利用所有觀察值，以 GEE1 的方法，所得之廣義線性模型中迴歸係數之估計量。而 $\hat{\beta}_{g(ij)}$ 為去除第 i 個 cluster 中第 j 個觀察值後，以 GEE1 的方法所得之 β 的估計量，$i = 1, \cdots, K$；$j = 1, \cdots, n_i$。我們定義測量該去除觀察值之"影響力"為

$$DFBETA_{(ij)} = \sqrt{(\hat{\beta}_g - \hat{\beta}_{g(ij)})^T \hat{V}^{-1} (\hat{\beta}_{g(ij)})(\hat{\beta}_g - \hat{\beta}_{g(ij)})} \quad (3.1)$$

其中 $\hat{V}^{-1}(\hat{\beta}_{g(ij)})$ 為利用 GEE1 程式所得之 $\hat{\beta}_{g(ij)}$ 的估計變異矩陣 (estimated covariance matrix)。此 $\hat{V}^{-1}(\hat{\beta}_{g(ij)})$，如 Liang & Zeger[1] 文中所述，對前言中所提的工作相關矩陣 (working correlation matrix) 具有統計上之強韌性 (robustness)。簡言之，即此估計量對該虛擬之結構假設具有很強之"包容性"。同一組資料在不同結構的工作相關矩陣下，所得的 $\hat{V}^{-1}(\hat{\beta}_{g(ij)})$ 值，差異並不大，具有很好的穩定性。

上式使用時，先分別算出每一觀察值 (observation) 的 $DFBETA_{(ij)}$，再以 $DFBETA_{(ij)}$ 為縱座標，而以該觀察值在整個 data set 的存放次序 (order) 為橫座標，畫二度空間的 scatter plots。由所繪圖形中，直接判讀是否存在有"影響力"之觀察值，以作為進一步討論的依據。基本上，(3.1) 式與 (2.2) 式類似，唯獨 (3.1) 式是利用 GEE1 的方法求得，如前言所述，故已將長期資料間之相關性列入考量。

同時，為方便 GEE1 使用者之應用，我們將原來的 SAS/IML Macro 程式，GEE1，加以修改，納入此項計算 $DFBETA_{(ij)}$ 之值的功能，使用上與原程式相同，僅須增加"OUT1 = *xxx*"的參數，以宣告欲進行"影響力"評估即可。下節中，我們將此方法應用在台灣省立新竹醫院眼科的一組資料上。

應用及結果

台灣省立新竹醫院眼科，柯美蘭醫師，在探討"正常人與不同視網膜病變等級之糖尿病病患間瞳孔參數值之比較"時，面臨了一個統計分析上的問題：資料來自同一人雙眼之兩筆資料彼此間不具"統計上之獨立性"。此處所謂的"瞳孔參數"定義為

$$\text{瞳孔參數} = \frac{\text{瞳孔直徑 (Pupil Diameter)}}{\text{內膜直徑 (Corneal Diameter)}} \times 100\%$$

除此之外，柯醫師亦想探討年齡、性別 (SEX = 0 表男性; 1 表女性)、糖尿病病史 (DM 年數) 及糖尿病網膜病變等級 (DR = 1, 2, 3, 4) 等因素與瞳孔參數的關係為何？其中，DR = 1 表示 DM 病人但無網膜病變，DR = 2 為 Background DR，DR = 3 為 Preproliferative DR, 而 DR = 4為 Proliferative DR。

基於上述理由，分析上採用 GEE1 的方法。樣本收集結果如表一所示。去除單眼人數後，實際分析人數為 234 人。另外，為了分析上的需要，我們利用四個啞變數 (dummy variables)，DR1，DR2，DR3，DR4，表示五種可能的 DR 值 (DR = 0 表正常組)，詳如表二所示。例如：以 (DR1, DR2, DR3, DR4) = (1, 0, 0, 0) 表 DR = 1 之組，而以 (DR1, DR2, DR3, DR4) = (0, 1, 0, 0) 表 DR = 2 之組，其餘類推。

分析結果如下：先分別考慮各單獨因素(factor)時，男、女性之瞳孔參數無顯著差異($p = 0.1611$)；年齡則有顯著差異($z = -2.19$; $p = 0.0143$)，且平均每增加一歲瞳孔參數值減少 0.11%；糖尿病組比正常組之瞳孔參數值平均小 6.06%，且顯著性非常高($z = -6.92$; $p \approx 0$)。其次，經調整(adjust) DM 年數效應後，DR = 1 之組比正常組瞳孔參數值平均顯著小 3.78%，DR = 2 組則平均顯著小 7.80%，但 DR = 3 及 DR = 4 比正常組則平均分別顯著小 13.26% 及 13.68% (表三)。最後，同時調整年齡、性別、DM年數之效應後 DR = 1，DR = 2，DR = 3 及 DR = 4 各組比正常組之瞳孔參數值平均顯著小 3.70%，7.93 %，13.38% 及 13.84% (表四)。

現以表四之線性模型為最終之分析結果，來進一步探討各觀察值之 "影響力"時，圖一顯示，第 135 號病人其 DFBETA 值顯著高於其他病人之值。換句話說，此位病人的資料對迴歸分析之影響力遠大於其他所有病人。因此，在利用 GEE1 進行迴歸分析之推論(inference)時，對此組資料需特別注意。表五即是將此組資料去除後，所得之迴歸分析結果。值得注意的是，去除該組資料後，年齡、性別、 DM年數之迴歸係數影響並不大。但是，DR = 1，DR = 2，DR = 3 及 DR = 4 各項之值則分別更改為 -3.71%，-8.23%，-12.35% 及 -12.99%。更值得一提的是，其對應之"s.e.-Robust"值，除 DR1 外，減少很多。換言之，去除該組資料後所得之估計值較準確。圖二則為去除第 135 號病人後，所得之"影響力"評估圖。圖中並無特異之 DFBETA 值呈現。

表一　樣本相關數據

	人數	眼數	單眼數	年齡	DM年數
糖尿病組	146	286	7	41-86	0-20
正常組	102	197	7	40-85	-

表二　啞變數 DR1, DR2, DR3, 及 DR4 定義之對照表

		DR1	DR2	DR3	DR4
DR	0	0	0	0	0
	1	1	0	0	0
	2	0	1	0	0
	3	0	0	1	0
	4	0	0	0	1

討　論

本文對長期資料之廣義線性模型提出一種圖示每一觀察值"影響力"大小的簡易方法。最重要的是，對原來 GEE1 的使用者而言，無需重新學習，即能馬上使用。附錄一

表三　DR's 經調整(adjust)DM年數後之 GEE1 迴歸分析結果

	Estimate	s.e.-Naive	s.e.-Robust	z-Robust
INT	46.442991	0.664	0.619	75.07
DM YEAR	0.085353	0.119	0.115	0.74
DR1	-3.784426	1.232	1.244	-3.04
DR2	-7.800558	1.432	1.358	-5.75
DR3	-13.260000	1.645	1.654	-8.02
DR4	-13.684240	1.878	1.741	-7.86

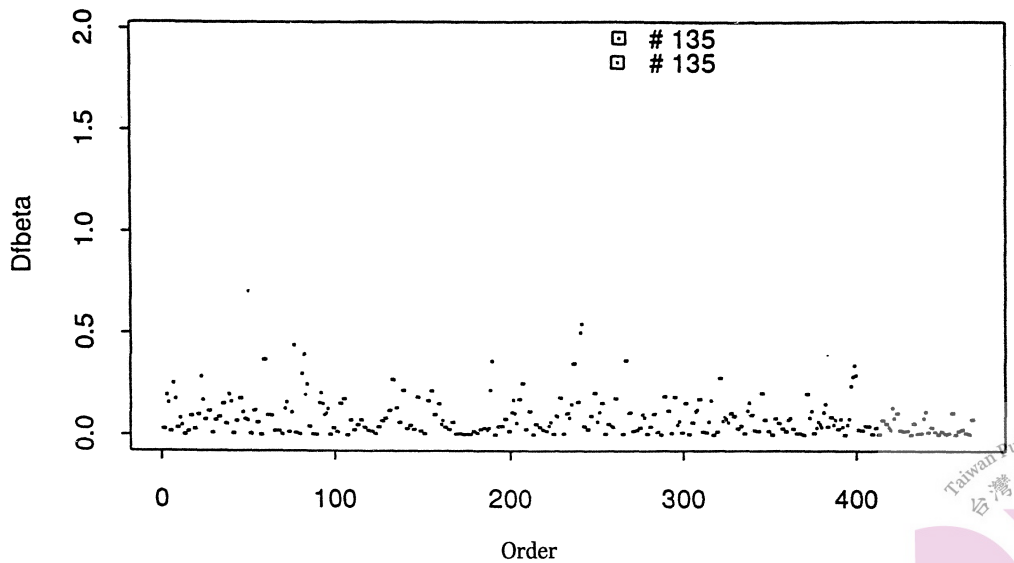
表四　GEE1 迴歸分析結果 (所有 Data)

	Estimate	s.e.-Naive	s.e.-Robust	z-Robust
INT	48.839268	2.868	2.930	16.67
SEX	1.240548	0.854	0.834	1.49
AGE	-0.052433	0.046	0.047	-1.11
DM YEAR	0.114860	0.120	0.116	0.99
DR1	-3.704868	1.224	1.231	-3.01
DR2	-7.931709	1.421	1.350	-5.88
DR3	-13.380860	1.635	1.652	-8.10
DR4	-13.837910	1.869	1.734	-7.98

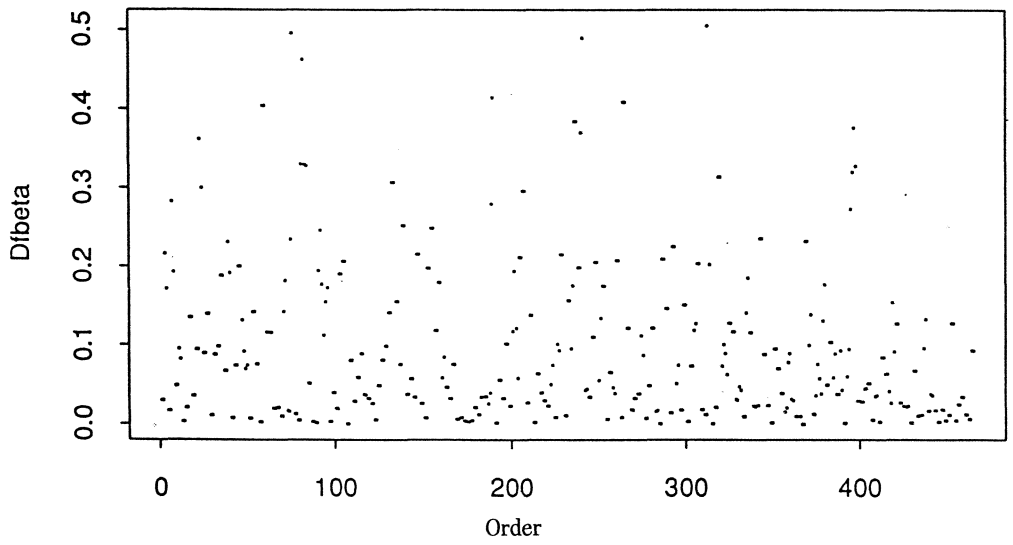
表五　GEE1 迴歸分析結果 (去除 # 135 Data)

	Estimate	s.e.-Naive	s.e.-Robust	z-Robust
INT	48.945861	2.873	2.924	16.74
SEX	1.210996	0.856	0.836	1.45
AGE	-0.053974	0.046	0.047	-1.14
DM YEAR	0.116479	0.120	0.117	1.00
DR1	-3.712319	1.225	1.232	-3.01
DR2	-8.228189	1.429	1.329	6.19
DR3	-12.352530	1.707	1.482	-8.33
DR4	-12.992000	1.909	1.673	-7.76

圖一　Scatter Plot for Nodrug.dat



圖二　Scatter Plot for Nodrug without #135



展示本文之應用實例的使用方法，供讀者參考。程式，GEE1-1.SAS，可直接函索。

此方法與 Hall, Zeger and Bandeen-Roche [7]，簡稱 HZB，所提方法比較，HZB 的方法從統計學的角度來看較好。因為他們引用了 Added Variable Plots 的觀念，即利用了 Score Test 的統計觀念。但是從使用者的角度來看則較不實際。因為，除了過程複雜難懂外，無法讓外界簡單地直接應用在實際資料上為其主要缺失。雖然，我們相信 HZB 應有程式可用，但使用者必需重新學習。

HZB 除了對每一觀察值提出"影響力"評估外，亦對每一 cluster 作"影響力"評估。針對此點，本文所提方法僅需略加修改即可有類似之功能。例如，定義第 i 個 cluster 的"影響力"為

$$\mathrm{DFBETA}_{(i)} = \sum_{j=1}^{n_i} \mathrm{DFBETA}_{ij}$$

然後對 $\mathrm{DFBETA}_{(i)}$, $i = 1, \cdots K$，作 scatter plots，亦可有相似之功能。

另外，雖有文獻[10]指出，資料中觀察值的影響力有時可能具有互動關係，即彼此間有時可能會相互遮掩 (mask) 對方的影響力，若資料中確實存在此問題，則此現象不易經由上式方法，即每次去除一個觀察值，偵測出。因此，有人提議每次去除多個觀察值 (multiple-case methods) 可能較妥當。對此，筆者認為，上述情形在某些資料也許可能存在。但 multiple-case 的方法卻也會延生其他複雜問題，例如，每次應去除幾個？2 個？3 個？或者更多。簡單的排列組合常識可預知，每增加一個，整個問題的複雜度會增加非常非常地多。現階段，本文對此問題尚未進行探討。由圖一及圖二的結果看來，此組資料似乎不會存有上述 masking 的問題。

Liang & Zeger[1]雖以"長期資料"(longitudinal data)為題，但其理論架構實已包含家族資料的範圍。故對該類資料，本文所提方法仍可適用。有興趣的學者，可向筆者函索程式進一步探討。當然，另一種常被數理統計學者採用的"模擬實驗"(simulation study) 亦可列入考慮進行研究。另外，GEE 的敏感度分析仍存有很多問題值得後續探討，如前言中所提：模型中自變項的增減項問題及所用連繫函數的適當與否…等等，均有待我們不斷地努力去探索解答。

誌　謝

作者感謝審查委員的寶貴意見，也謝謝助理黃冠華先生在程式上的協助及台灣省立新竹醫院眼科柯美蘭醫師允諾使用該筆糖尿病網膜病變資料。本文在國科會 NSC 84 - 2121 - M - 001 -018 計劃的資助下完成，在此一併誌謝。

參考文獻

1. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986; **73**:13-22.
2. McCullagh P, Nelder J. Generalized linear models, 2nd ed. London: Chapman and Hall, 1989; 392.
3. Wang PC. Adding a variable in generalized linear models. Technometrics 1985; **27**:273-276.
4. Landwehr JM, Pregibon D, Shoemaker AC. Graphical methods for assessing logistic regression models (with discussion). JASA 1984; **79**:61-83.
5. Collett D. Modelling binary data. London; Chapman and Hall 1991.
6. O'Hara Hines RJ, Carter EM. Improved added variable and partial residual plots for detection of influential observations in generalized linear models (with comment). Applied Statistics 1993; **42**:3-20.
7. Hall CB, Zeger SL, Bandeen-Roche KJ. Added variable plots for regression with dependent data. Technical Report # P-792 1994; The Johns Hopkins Univ., School of Hygiene and Public Health, Department of Biostatistics.
8. Cook RD. Detection of influential observations in linear regression. Technometrics 1977; **19**:15-18.
9. Belsley DA, Kuh E, Welsch RE. Regression diagnostics: identifying influential data and sources of collinearity. New York 1980; Wiley.
10. Weisberg S. Applied linear regression, 2nd ed. New York 1985; Wiley:125.

附錄一

```
DATA ttt;
INFILE 'b:nodrug1.dat';
INPUT ID age sex year osize cdiate dr dm;
        int = 1;
        order = _n_;
        ocratio = osize / cdiate * 100;
        csex = 0;
        if sex = 2 then csex = 1 ;
        dr1 = 0; dr2 = 0; dr3 = 0; dr4 = 0;
        if dr = 1 then dr1 = 1;
        if dr = 2 then dr2 = 1;
        if dr = 3 then dr3 = 1;
        if dr = 4 then dr4 = 1;
%INCLUDE 'b:gee 1 - 1 _pc.sas';
%GEE ( DATA = ttt,
        YVAR = ocratio,
        XVAR = int csex age year dr1 dr2 dr3 dr4,
        ID = ID,
        LINK = 1,
        VARI = 1,
        CORR = 4, OUT1 = sss );
data uuu;
        set sss;
        file 'b:dfnod1.out';
        put order id noclu ocratio dfbete;
        run;
```

SENSITIVITY ANALYSIS IN GEE—IDENTIFICATION OF HIGH INFLUENTIAL OBSERVATIONS

YUE-CUNE CHANG

The importance of identification of high influential observations in the applications of regression model is indubitable. There are a lot of related papers published for the general linear model and the generalized linear model as well. However, for the longitudinal data analysis, we haven't seen any literature published yet. In this paper, we proposed a simple graphic method to handle this sensitivity analysis problem. We also modified the original longitudinal data analysis SAS/IML macro program, GEE1, to include the proposed graphic method. For those GEE1 user, the modified macro program is easy to use. We successfully applied this graphic method to analyze a real data set which was conducted by the Provincial Hsin-chu hospital in Taiwan. *(Chin J Public Health (Taipei): 1996; **15**(5): 403-410)*

Key words: sensitivity analysis, high influential observations, generalized linear models, longitudinal data.

Institude of Statistical Science Academia Sinica