返回总目录

目　　录

第四部分
回 归 分 析

第 17 章　SAS 系统内七种回归分析程序概述

17.1　七种回归分析程序

在 SAS 系统中，适用于回归分析的统计程序有许多，其中常用到的有 REG，PROBIT，LOGISTIC，ORTHOREG，RSREG，GLM 及 NLIN 等程序。此外还有 AUTOREG，SYSLIN，PDLREG 及 MODEL 等程序。下面简单地描述这几个程序：

REG　执行普通线性回归分析，适用于各式的输入／输出格式，并有诊断性以及简化模型的功能。

PROBIT　执行概率回归分析或逻辑斯谛的回归分析。这个程序所处理的数据通常含二分 (或二分以上) 的因变量以及数个连续的自变量。

LOGISTIC　执行逻辑斯谛的回归分析，分析方式含逐步回归分析以及各式的诊断统计值；是新的 6.06 版中添加的程序。

ORTHOREG　使用 Gentleman-Givens 的计算程序来估计回归模型中的参数值。适用于估计值之标准误差较大的数据，详情较第 21 章的说明。

RSREG　建立二项式反应面 (Response-Surface) 的回归模型。

GLM　最普通的线性分析，自变量可以是类别变量或多项式。

NLIN　建立非线性的回归模型。

AUTOREG　利用时间系列的数据导出回归模型。此法中各误差 (Errors) 之间可以是相关的 (此程序并不包括在本书讨论中，有兴趣的读者请自行参阅 SAS/ETS 手册)。

SYSLIN　用于经济学的模型 (本书不讨论此程序，有兴趣的读者请自行参阅 SAS/ETS 手册)。

PDLREG　本书不讨论此程序，有兴趣的读者请自行参阅 SAS/ETS 手册。

MODEL　处理非线性的联立方程序，适合经济学中讨论的模型，有兴趣的读者请自行参阅 SAS/ETS 手册。

其它更不常用的回归分析程序，则必须在 “SUGI Supplemental Library User's Guide ” 中才可找到。

由于 GLM 程序又可以用来执行回归分析又可以用来执行变异数分析，所以在第六部分第 31 章内将详加介绍。其余的六种 (即 REG， PROBIT， LOGISTIC， ORTHOREG，RSREG 以及 NLIN) 在第四部分第 18 章至第 23 章内逐一说明。

值得读者注意的是，过去在第 5 版环境下所习用的 RSQUARE 及 STEPWISE 程序，如今都纳入了 PROC REG 程序（见第 18 章的第 18.1 节以及附录 D 中有关 REG 程序在新版中的改进）。此外，各程序都可在交谈式的环境下执行，如此，读者可以更有效地修正每一个测试的回归模型。

■ PROC REG 程序

这是最通俗的回归分析程序，其功能如下：

* 可以同时测试好几个不同的回归模型。
* 有九种不同的方法可简化回归模型。
* 输入数据可以是相关系数矩阵或是向量内乘积 (Cross Product) 的矩阵
* 印出预测值、误差、信赖区间及向量内乘积矩阵等。并可将这些分析好的数据存在一个 SAS 文件中，使它成为其它统计程序的输入文件。
* 印出影响度的值、相关系数以及 (半) 净相关系数。
* 估计参数据检验线性回归模型。
* 提供共线性 (Collinearity) 的诊断。
* 取代第 5 版中的 RSQUARE 及 STEPWISE 两程序。
* 提供九种筛选回归模型的方法，即 NONE， FORWARD， BACKWARD，STEPWISE， MAXR， MINR， RSQUARE， CP 以及 ADJRSQ 等。这九种方法的详细介绍收录在第 18.2 节的指令 #2 MODEL 部分。

■ PROC PROBIT 程序

本程序主要是利用最大可能率估计法找出一个回归模型的参数估计值，或生物实验数据以及类别数据中的底线率。在估计这些参数值的过程中，PROBIT 程序容许读者选择各式不同的模型如：概率单位 (Probit)、对数奇数比 (Logit)、次序逻辑斯谛 (Ordinal Logistic)， 以及成长曲线 (Gompit) 等模型。

■ PROC LOGISTIC 程序

此程序适合处理二分或二分以上的类别数据。统计模型的形式可以是概率模型或逻辑斯谛模型。当模型中的自变量数目过多时，LOGISTIC 程序可提供逐步排除的方法来挑选最精简的模型。报表的输出资料含回归模型的诊断以及预测值，预测误差等。

■ PROC ORTHOREG 程序

这个程序最适用于参数估计值的标准误差差较大的数据。 在这种情况下，REG 或 GLM 程序分析的结果只能算是最小误差平方解 (LS) 的趋近值，而非真正的 LS 解。不过，读者仍可借 REG 程序对数据作初步的分析，看看自变量之间是否有极高的关系 (此由共线性的诊断值可看出来)，然后，再决定有没有必要继续执行 ORTHOREG 程序的分析。

■ PROC RSREG 程序

此程序适用于反应面的分析，其优点包括：

* 自动印出自变量的平方与三次方值，并将它们包括在回归模型中。
* 检验模型的精确值。
* 解出反应面的临界值 (Critical Value)。
* 计算出特征值 (Eigen Value) 的值及其平方值。

■ PROC GLM 程序（归入第六部分第 31 章）

此程序可用来执行线性回归分析、变异数分析与共变量分析。若用来执行回归分析，此程序有以下的特色：

* 适于处理类别数据。
* 可直接建立多元多项式的回归模型。

■ PROC NLIN 程序

此程序采用最小误差平方法 (Least Squares Method) 及循环推测法 (Iterative EstimationMethod) 来建立一个非线性模型。一般而言，读者必须自订参数的名字、参数的启动值 (StartingValue)、非线性的模型与循环推测法所用的准则。若读者不指明，则 NLIN 程序自动以高斯-牛顿迭代法 (Gauss-Newton Iterative Procedure) 为估计参数的方法。另外，此程序也备有扫描 (Grid Search) 的功能来帮助读者选择合适的参数启动值。由于非线性回归分析十分不易处理，NLIN 程序不保证一定可以算出符合最小误差平方法之标准的参数估计值。

17.2　七种回归分析程序的比较

本节就七种最常见的 SAS 回归分析程序的输出资料类型及诊断功能做比较。这七个程序是：REG， PROBIT， LOGISTIC， ORTHOREG， RSREG， GLM 及 NLIN。

■相同类型的输出数据

七个程序都提供下列几种的输出数据：

* 用最小误差平方法所估计的参数值 (如：b_0，b_1，....)。
* 误差变异数的估计值。
* 参数估计值的标准误差差或变异数。
* 有关参数的假设 (如 H_0：β_0=0) 检验。
* 各种预测值及其误差。
* 对整个回归公式有效度的检验。

■相异的诊断功能

REG，LOGISTIC，PROBIT 与 RSREG 等程序提供下列的诊断功能，其它程序则无：

* REG 程序提供共线性 (Collinearity) 的诊断，这个诊断探讨自变量间相关的程度及可能造成的影响。
* REG， LOGISTIC， 及 RSREG 三个程序提供影响度诊断以决定各观察体对参数估计值、误差的平方和 (SSE) 及预测值等的影响。LOGISTIC 程序也有这种功能，不过其分析原理是采最大可能率法。
* PROBIT 与 RSREG 两程序提供回归模型精确度 (Accuracy) 的诊断，所用的方法是比较误差的变异数及其估计值。

* REG 程序提供时间序列分析 (Time Series Analysis) 的诊断，特别是有关时间的误差以及误差间彼此的相关。

17.3　有关回归分析的基本统计概念

上面所提的七个程序都适用于回归分析，现在来讨论一下回归分析的基本概念：回归分析的目的是借一个回归公式来做预测。回归公式等号左边的值是因变量，等号右边是一系列的自变量及参数（又称回归系数，它是一个常数）的线性组合。

■回归公式

假如我们希望推测某个观察体的因变量数据，则下面的公式涵盖回归分析的原理：

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i$$

其中，Y_i 是因变量；

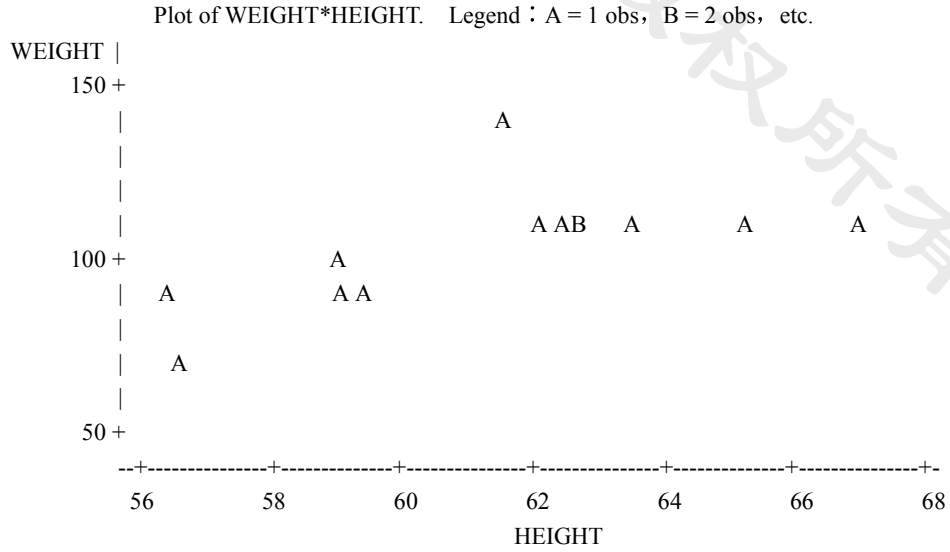
X_{ij} 是自变量；

β_0 及 β_j 均是参数，分别代表 Y 的截距及回归线的系数，它们的值由统计估计而来；

$\in_i$ 是误差。

比方说，根据上列公式，我们用身高来推测学生的体重。所以身高是自变量，而体重是因变量。我们取一个样本 (十三位小学三年级的学生)，测量出他们的身高及体重，把这些值用平面坐标图表示如下：

报表 17.1　以身高推测学生的体重



根据线性回归的测量，β_0=-138.2，β_1=3.95。所以这群小学生的身高与体重的线性关

系可用下列公式表明：

体重的估计值=(-138.2) + 3.95*(身高)

除此例外，线性回归分析也可用来寻找一个未知的线性关系。比如：教育程度与年收入所得，学生智商与成绩，气体的体积与压力等关系均可借回归分析的方法表示出来。

■不能证明因果关系的存在

线性回归分析的结果并不表示因果关系的存在。因果关系的存在只有借着纯科学性的实验法（比如说实验组与对照组的观察比较）才可证明。

■最小均方差法（Least Squares Method）

在回归分析中，参数的值一般是按照最小误差平方法 (Least Squares Method) 导出。此法的目的在减少因变量预测值与实际值之间的平方误差。在英文中，由此法导出的参数称为 Least Square Estimates，而其平方误差称为 SSE。若以数学符号来表示，则最小误差平方法的精义如下：

$$SSE = Min\left[\sum_{i=1}^{n}(Y_i - b_0 - \sum_{j=1}^{p} b_j X_{ij})^2\right]$$

在此式中，b_0 与 b_j 是参数 β_0 与 β_j 的估计值。若读者欲深入研究最小误差平方法，请自行参阅下列书目：Draper 及 Smith（1981），Daniel 及 Wood（1980）或 Johnston (1972)。

17.4　SAS 程序的运算原则

以下讨论 SAS 程序在执行回归分析时所采用的运算原则。

■矩阵的表示

一个线性回归模型可以用矩阵表示如下：

$$Y = X * \beta + E$$

在此，X代表一个 n*k 的矩阵，其横列 (即 n) 代表观察体，纵行 (即 k) 代表自变量。一般而言，X矩阵的第一纵行皆为 1，以推测截距的值；β 是一个 k*1 的向量，代表参数；而E是一个 n*1 的误差向量。根据矩阵运算的原理，Y也会是一个 n*1 的向量。

■线性回归的假设

线性回归的重要假设如下：

a. 所有自变量是固定的，或由实验结果导出。
b. 回归模型是正确的。
c. 自变量的测量没有误差。
d. 误差的平均值是 0。
e. 误差之间的变异数是常数，其值以 σ^2 表示。
f. 误差与误差之间没有相关。

当我们要检验回归模型的有效度 (Significance) 时，我们必须附加另外一个假设：

g. 误差值在母群内形成一个常态分配。

■统计的模型

当上述 a 到 f 的假设都成立时，由最小误差平方法所推测出来的参数估计值也就是最佳线性不偏估计值 (Best Linear Unbiased Estimates，简称 B.L.U.E.)。也就是说，此估计值是最精确的。如果 g 的假设也成立，则我们可做以下的结论：

* 所有统计参数的估计与检验所必须的理论基础 (即抽样分配) 成立。

* 参数的估计值形成一个常态分配。

* 各个离差平方和 (Sum of Squares of Deviations) 形成一个类似 χ^2 的分配。

* 参数估计值与其标准误差差 (Standard Error) 之间的比例形成一个 t 分配。

若上述 a 到 g 的假设不能全部成立，则你必须谨慎地解释上述的结论。有关回归分析的假设条件与结论之间的资料可参阅 Box (1966)，Mosteller 及 Tukey (1977，12-13 章) 等参考书。

■估计各参数

(1) 参数之最小平方误差的估计值是由解正规方程序 (Normal Equations) 而导出：

$b=(X'X)^{-1}X'Y$

假如 (X'X) 是一个满秩 (Full Rank) 的矩阵，则误差的变异数 (σ^2) 可由下列的公式间接算出：

(2) $S^2=MSE=SSE/(n-k)=\sum(Y_i-X_ib)^2/(n-k)$

此处 X_i 是指自变量矩阵 X 的第 i 列。由于两个估计值都是不偏的估计值，所以

$E(b)=\beta$

$E(S^2)=\sigma^2$

(3) b 值的变异数 (Variance) 可由下式算出：

$Var(b)=(X'X)^{-1}\sigma^2$，或 $Var(b)=(X'X)^{-1}S^2$

所以 b 值的标准误差差就是：

(4) $STDERR(b_i)=\sqrt{D(X'X)_i^{-1}S^2}$

其中，$D(X'X)_i^{-1}$ 代表 $(X'X)^{-1}$ 矩阵中对角斜线上第 i 个元素。

有了估计值与其标准误差差之后，我们可以检验这些估计值如下：

(5) $t=\dfrac{b_i}{STDERR(b_{i)}}$

这个 t 值是各程序输出数据的一部分。其统计显著度也会被印出来，以便读者判断估计值的有效度，虚无假设是 H_0：$\beta_i=0$。

(6) 两种平方和 (SSI，SSⅡ)

回归分析程序计算两种平方和：第一种平方和 (SSⅠ) 代表每一个自变量对整个回归模型的贡献，此值与该自变量进入回归模型的顺序有关。第二种平方和 (SSⅡ) 则代表自变量从模型中剔除之后对整个平方总和的影响。SSⅡ 与 GLM 程序

中的第三及第四种平方和 (亦即 SSⅢ 及 SSⅣ) 相当。详情请见本书第 31、32 章的说明。

(7) 标准参数估计值

「标准参数估计值」是指标准化后的参数估计值。这些标准值均以 0 为平均数，1 为标准差。标准化的过程是一个线性转换：将原值减去平均数，然后除以标准差。

(8) 容忍量与变异数膨胀值

另外两个统计数：容忍量 (Tolerance) 与变异数膨胀值 (Variance Inflation) 是用来表示模型中自变量之间的相关程度。容忍量 (简称 TOL) 的值等于 $1-R^2$；在此，R^2 是 y 与 y 的估计值之间相关系数的平方。变异数膨胀值 (简称 VIF) 是容忍量的倒数。所以当所有自变量之间无相关时，TOL=VIF=1。但相反的，如果自变量之间有很强的相关时，TOL 趋近于 0，而 VIF 趋向一个极大的数值。

■预测值、误差值与其它相关的统计数

当我们估计参数后，可将这些参数估计值带回到回归公式以便导出预测值与预测误差(以下简称误差)，这些概念的数学公式表示法如下：

预测值　$Y_i = X_i b$

预测值的标准误差　$STDERR(\hat{Y}_i) = \sqrt{X_i (X'X)^{-1} X_i' S^2}$

误差　$\varepsilon_i = Y_i - \hat{Y}_i$

误差的标准误差　$STDERR(\varepsilon_i) = \sqrt{[1 - X_i (X'X)^{-1} X_i'] S^2}$

标准化误差　$student = \varepsilon_i / STDERR(\varepsilon_i)$

预测值平均数的信赖区间 (关键字 CLM)
$$\begin{cases} LowerM = X_i b - t_{\alpha/2} * [STDERR(Y_i)] \\ UpperM = X_i b + t_{\alpha/2} * [STDERR(Y_i)] \end{cases}$$

预测值的信赖区间 (关键字 CLI)
$$\begin{cases} LowerY_i = X_i b - t_{\alpha/2} * [STDERR(\hat{Y}_i) + \sqrt{MSE}] \\ UpperY_i = X_i b + t_{\alpha/2} * [STDERR(\hat{Y}_i) + \sqrt{MSE}] \end{cases}$$

在此，MSE 是 Mean Square Error 的简称，中文翻译成误差的均方。

COOKD 值　$COOKD = student^2 [STDERR(Y_i)/STDERR(\varepsilon_i)]^2 / k$

最后一个公式是 Cook 于 1977，1979 所提出的概念，其用途在测量各观察体对参数估计值的影响力。

■线性假设的检验

有关参数的线性假设，一般可以下式表示：

H_0：$L\beta = C$

C 是一个常数，其值多半设为 0。

欲检验此虚无假设，首先必须导出一个线性函数：

(Lb-C)

此函数的变异数是

$Var(Lb-C)=L\ Var(b)\ L'=L(X'X)^{-1}\ L'\ \sigma^2$　　此处，b 是参数 β 的估计值。

从上述函数与函数变异数中，可以导出第三个概念。这个概念称为二项式 (或作平方总和 SS)：

$SS(Lb-C)=(Lb-C)'[L(X'X)^{-1}\ L']^{-1}\ (Lb-C)$

若这个二项式的值愈大，则回归分析所求出的函数便愈有效。有效的程度是以 F 检定来测验的：

$F=[SS(Lb-C)/k]/S^2$

这个 F 检定的两个自由度分别是 k 与 (n-k+1)。在此，k 代表自变量的数目，而 n 代表样本的大小，S^2 的定义见估计各参数的第 (2) 个公式。

■ 因变量的统计检验

复因变量的检验牵涉一个以上的因变量。我们若修改上节的虚无假设，就可使它成为复因变量检验的假设：

$H_0：L\beta M = d$

此处，d 是一个常数，其值通常以零表示，M 代表因变量的矩阵。

欲检验此虚无假设，我们必须另外定义两个矩阵 H 与 E。它们的定义如下：

$H=M'(L\beta-c_j)'[L(X'X)^{-1}L']^{-1}\ (L\beta-c_j)M$　　此处，j=含 1 的列向量

c=常数的行向量

$E=M'[Y'Y-\beta'(X'X)\beta]M$

利用 H 与 E 的值，下述四个检定可测出 H 值是否比 E 值显着大。当 H 与 E 值相近时，虚无假设成立；反之，虚无假设被推翻。

(1) Wilks' Lambda (Λ)

$\Lambda = \det(E)/\det(H+E)$

其中，det 代表矩阵的行列式值 (Determinant)。

(2) Pillai's Trace (V)

$V = Trace\ [H\ (H+E)^{-1}]$

其中，Trace 代表矩阵的轨迹 (Trace)

(3) Hotelling-Lawley's Trace (U)

$U=Trace\ (E^{-1}H)$

(4) Roy's Maximum Root (Q)

$Q=\lambda_1=(E^{-1}H)$ 矩阵的最大特征值

上述四个检定都以 F 分配为基础，有关临界值 (Critical Value) 的表可在 Pillai (1960) 书中找到。

有兴趣的读者，请研读 Morrison (1976)， Timm (1975)， Mardia，Kent，Bibby (1979) 或 Bock，R. D. (1975) 等书。

第 18 章　一般性回归统计分析：统计程序 PROC REG

18.1　PROC REG 程序概述

REG 程序将参数估计值带入线性回归模型中以便执行回归分析的预测。这些参数估计值是由最小误差平方法所导出的。

REG 程序是 SAS 所有回归分析程序中，用途最广泛的一种 (好比 GLM 程序是 SAS 所有变异数分析程序中用途最广泛的一种)。其它回归分析的程序各有其特殊的用途，第 17 章内对它们作了一个简单的介绍。值得读者注意的是，新版 (6.06 与 PC-6.03 版) 的 REG 程序已经取代旧版 STEPWISE 与 RSQUARE 两程序的功能。此外，分析的结果可直接绘制成点状图 (Scatter Plot)，并在图上强调某个或某些观察个体。

现列举 REG 程序的功能如下：

1. 能同时考虑数个线性回归模型，并以交谈式执行回归分析。
2. 输入数据可以是相关系数或变量的向量内乘积 (Cross Product) 等。
3. 可印出因变量的预测值、误差、标准化误差、信赖区间等。这些统计值可被储存在一个输出文件内，或利用指令 PLOT 绘成点状图。
4. 印出各种影响力的值。
5. 绘制净回归图 (Partial Regression Leverage Plots)。
6. 由最小误差平方法估计参数。
7. 检验线性回归系数。
8. 检验多变量的假设。
9. 将自变量的向量积矩阵纳入输出文件。
10. 诊断自变量之间线性相关的程度。
11. 有九种不同的方法可简化模型。
12. 利用 PAINT 指令可特别强调数据中某个或某些观察个体。

18.2　如何撰写 PROC REG 程序

PROC REG 含十八道指令，它们的格式如下：

不可省略	PROC REG	选项串；
不可省略，可用交谈式执行	MODEL	因变量名称串=自变量名称串 /选项串；
必须在 RUN；指令之前	VAR	变量名称串；
	FREQ	变量名称；
	WEIGHT	变量名称；
	ID	变量名称；
	BY	变量名称串；
这些指令可放在 MODEL 指令后的任何一处而且可在交谈式环境下执行	ADD	变量名称串；
	DELETE	变量名称串；
	RESTRICT	等式 1，等式 2，...；
	TEST	等式 1，等式 2，... /选项；
	MTEST	等式 1，等式 2，... /选项串；
	OUTPUT	OUT=输出文件名关键字=变量名称串；
	REWEIGHT	加权条件式 ALLOBS /选项串；
	(或 REWEIGHT STATUSUNDO；)	
	REFIT；	
	PAINT	强化条件式 ALLOBS /选项串；
	(或 PAINTSTATUSUNDO；)	
	PLOT	图形指令串 /选项串；
	PRINT	选项串 ANOVA MODELDATA；

其中，PROC REG 与 MODEL 两道指令是必需的，不可省略。一个 REG 程序中可含多个 MODEL 指令。在每一个 MODEL 指令之后，可有一个 OUTPUT 指令及多个 RESTRICT， TEST， MTEST 等指令。至于 WEIGHT， FREQ 及 ID 指令则可有可无，而且只须使用一次，其效力即可贯穿整个 REG 程序。

指令 #1　PROC REG 选项串；

有下列十个选项可供选择 [其中 (1)-(4) 选项与输出 / 输入文件的界定有关，(5)-(10) 选项与报表打印或其它的控制有关]：

(1) DATA=输入文件名称

指明到底对哪一个文件执行分析。若省略此选项，则 SAS 会自动找出在本程序之前最后形成的 SAS 文件，对它执行回归分析。新版下执行 REG 程序可采用六种不同的输入文件形式。第一种是原始数据文件，第二种是相关系数矩阵（TYPE=CORR），第三种是共变异数矩阵（TYPE=COV），第四种是未对平均数矫正过的相关系数矩阵 （TYPE=UCORR），第五种是未矫正过的共变异数矩阵（TYPE=UCOV），第六种是向量的内乘积（TYPE=SSCP)。若读者输入的数据形属于上述第二到第六种中的任何一种形式，则下列几道指令不再适用：FREQ， ID，OUTPUT， PAINT， PLOT， REWEIGHT 及 WEIGHT 等。其它指令如 MODEL 与 PRINT 也有若干选项不采用，请读者仔细研读这两个指令的基本语法。

(2) OUTEST=输出文件名称

此文件专门用来储存所有的参数估计值。

(3) COVOUT

要求选项 OUTEST=输出的文件中也包含参数估计值之间的共变异数矩阵。

(4) OUTSSCP=输出文件名称

这个 TYPE=SSCP 的文件矩阵专门用来储存变量的平方和与内乘积。

(5) NOPRINT

所有分析的结果皆不印出。若 REG 程序含三个 MODEL 指令，则在 PROC REG 指令中选用一次 NOPPINT 选项，就相当于在各 MODEL 指令中分别选用 NOPRINT 选项，比较精简。

(6) SIMPLE

印出所有参与分析之变量的简单描述性统计值 (如：平均数，标准差)。

(7) USSCP

印出所有参与分析之变量的平方和与内乘积的矩阵。

(8) ALL

要求印出所有的分析结果。若选用 ALL，就可以省略 SIMPLE 和 USSCP 两选项，因 ALL 已包含 SIMPLE 和 USSCP 的值。另外，如同上述的 NOPRINT 选项，读者若想取得所有的分析结果，则只要在 PROC REG 指令中使用一次 ALL，其效力即可及于所有的 MODEL 指令，如此比较精简。

(9) CORR

要求打印在 MODEL 指令或 VAR 指令中界定之变量间的相关系数矩阵。

(10) SINGULAR=正实数

这是一个很少用到的选项，其功能在于检验各矩阵是否为满秩矩阵，内设值是 10 的 -7 次方。

指令 #2　MODEL 因变量名称串＝自变量名称串 / 选项串;

删除号 (/) 前的部分很直接了当，无须赘述。唯一值得交待的是，每一个模型指令可冠以一个代号，其长度不超过八个字母，如：

```
PROC REG;
    M1 : MODEL Y=X1;
    M2 : MODEL Y=X2;
```

删除号 (/) 后的选项可分成六类：第一类选项与报表的打印有关；第二类选项控制计算的过程的打印；第三类选项界定参数估计值的有关事宜；第四类选项与预测值、预测误差有关；第五类选项界定回归模型的选择；第六类选项与 SELECTION=RSQUARE，ADJRSQ，CP 的设定有关。以下分述各类别下的选项：

第一类选项　此处有三个选项与报表的打印有关：

(1) NOPRINT

不印出 MODEL 指令所界定的分析结果。

(2) ALL

印出 MODEL 指令所有分析的结果。这个选项将印出下列的统计值：XPX， SS1，SS2， STB， TOL， VIF， COVB， CORRB， SEQB， I， P， R， CLI， CLM，SPEC， ACOV， PCORR1， PCORR2， SCORR1， SCORR2。这些统计值的定义，请参阅本指令的第三及第四类选项。

(3) NOINT

规定回归模型中不包含截距。

第二类选项　控制计算过程的打印，有两个选项：

(1) XPX

印出回归模型的 (X'X) 向量积矩阵。在此 X 代表自变量的矩阵。

(2) I

印出上述 (X'X) 的反矩阵，即 $(X'X)^{-1}$。

第三类选项　界定有关参数估计值的有关事宜，有十六个：

(1) SS1

按回归模型中各参数估计值之顺序，印出其第一型离差平方和。

(2) SS2

按回归模型中各参数估计值之顺序，印出其第二型离差平方和。有关 SS1 与 SS2 的详细介绍，请参阅本书第 32 章。

(3) STB

印出标准化后的回归系数。

(4) TOL

印出各参数估计值的容忍量 (Tolerance)。容忍量 (简称 TOL) 的定义是 $1-R^2$。在此，R^2 是以此参数对应的变量为因变量，而模型中所有其它的变量为自变量所导出的复相关平方。

(5) VIF

是上述容忍量的倒数，称为变异数的膨胀值 (Variance Inflation)。

(6) COVB

印出参数估计值的共变异数矩阵，其定义是 $(X'X)^{-1}S^2$，在此 S^2 是误差的平均方。

(7) CORRB

即 $(X'X)^{-1}$ 经过标准化后的矩阵。

(8) SEQB

以自变量进入回归模型的先后为顺序，印出其对应的参数估计值。这些参数估计值会以一个下三角形矩阵表示；此矩阵的每一横列代表一系列的参数估计值。

(9) COLLIN

对自变量之间的共线性 (Collinearity，即变量间线性相关的程度) 执行分析。其分析结果包括特性根、线性相关的系数及参数估计值的变异数分析。

(10) COLLINOINT

与上述 COLLIN 选项的作用几乎一样，但 Y 截距不包括在共线性分析之内。

(11) ACOV

在不同变异数的假设之下，印出参数估计值的近似共变异数矩阵。这个矩阵的定义如下：

$(X'X)^{-1}[X'diag(ei^2)X](X'X)^{-1}$

其中，$e_i = Y_i - X_i * \beta$，是回归分析的误差。

(12) SPEC

对回归模型中的第一和第二动差 (Moment) 作统计检验。

(13) PCORR1

根据第一型离差平方和的定义印出净相关平方矩阵。

(14) PCORR2

根据第二型离差平方和的定义印出净相关平方矩阵。

(15) SCORR1

根据第一型离差平方和的定义印出半净相关矩阵 (Semi-Partial Correlation Matrix)。

(16) SCORR2

根据第二型离差平方和的定义印出半净相关矩阵。

第四类选项　此类选项有七个，均与预测值、预测误差有关：

(1) P

由输入数据及回归模型预测因变量的值。这个选项将产生包含原数据、因变量的实际值与预测值以及预测误差的报表。

(2) R

要求 REG 程序对预测误差做进一步的分析。这个选项所产生的报表将包括上述 P 选项所产生的所有数据，再加上预测值和误差的标准误、标准化误差、以及库格氏 (Cook)的 D 值。D 值可用来测量每一观察体对参数估计的影响力。

(3) CLM

印出预测值平均数的 95% 信赖区间之上限与下限 ，这个区间只考虑到参数估计值的抽样误差，但不考虑到观察体的抽样误差。公式列在第 17 章。

(4) CLI

印出各个预测值的 95% 信赖区间之上限与下限。这个区间同时考虑到参数估计值的抽样误差以及观察体的抽样误差，因此，比上述 (3) 的区间更宽，更不精确。公式列在第 17 章。

(5) DW

计算杜本-华生氏 (Durbin-Watson) 统计值。这个统计值旨在测量误差之间线性相关的程度，最适用于与时间顺序有关的回归分析。

(6) INFLUENCE

针对每一观察体对参数估计值与因变量预测值的影响力做分析。

(7) PARTIAL

对每一个自变量作净回归图 (Partial Regression Leverage Plots)。

第五类选项　界定回归模型的选择，有下列十个选项：

(1) SELECTION=FORWARD (或 F) [顺向选择法]
SELECTION=BACKWARD (或 B)[反向淘汰法]
SELECTION=STEPWISE　[逐步排除法]
SELECTION=MAXR　[最大相关法]
SELECTION=MINR　[最小相关法]
SELECTION=RSQUARE　[复相关系数平方法]
SELECTION=ADJRSQ　[矫正后的复相关系数平方法]
SELECTION=CP　[Cp 法]
SELECTION=NONE　[内设值，进行全型的回归分析]

这九种选择 "最佳" 模型的方法均有其统计的理论基础(见 Draper and Smith，1981)。前五种方法，亦即 FORWARD，BACKWARD，STEPWISE，MAXR，MINR 等在旧版内由 PROCSTEPWISE 执行，第六及第七种方法 (即 RSQUARE 与 ADJRSQ 法) 由 PROC RSQUARE 执行。如今，它们只是 PROC REG 的一个选项而已。

顺向选择法 (FORWARD 或 F) 依自变量的重要性逐渐增加模型中自变量的个数，直至模型达到最精简。反之，反向淘汰法 (BACKWARD 或 B) 将全型的模型中 "不重要" 的自变量剔除，直到所有剩下的自变量都是不可少的。逐步排除法 (即 STEPWISE) 是前述两种方法的大成：一方面进行顺向法，一方面回头检验模型中的自变量是否该剔除。最大相关法是对一个固定的自变量个数，找出一组自变量使 R^2 的值为最大。最小相关法是针对一个固定的自变量个数，找出一组自变量使得 R^2 的增进降为最低。复相关平方法是找出含一个到 P 个 (自变量的总数) 的几组模型，每组的自变量组合都产生最高的 R^2 值。ADJRSQ 法与 RSQUARE 法相似，目的是产生最高的矫正过的 R^2 值。Cp 法是依 Mallows (1973) 的著作。NONE 则要求将所有自变量均包括在回归模型里 (又称全型的模型)；NONE 是本选项的内设值。

(2) DETAILS

在顺向选择，反向淘汰，和逐步排除法中，这个指令规定印出回归分析的每一步细节。

(3) INCLUDE=正整数 (如 3)

这个选项规定将 MODEL 指令的前几 (如 3) 个自变量纳入每一个回归模型里；此选项不可与上述 SELECTION=NONE 的设定联用。

(4) START=正整数 (如 2)

规定分析的第一个回归模型内至少应包含的自变量之数目 (如两个)，与上述选项 (1)SELECTION=MAXR， MINR 以及 STEPWISE 的设定联用，内设值等于 0。若与 SELECTION=RSQUARE， ADJRSQ， 或 CP 的设定联用时，内设值等于 1。

(5) STOP=正整数 (如 5)

这个指令指示 REG 程序搜寻出一个含 STOP= 个数 (如 5) 的最佳回归模型后即停止。当此选项与 SELECTION=RSQUARE， ADJRSQ， 以及 CP 等回归模型的选择法联用时，STOP=的值界定任何一组模型所含之自变量个数的上限。然而，

当此选项与 SELECTION=MAXR 以及 MINR 选择法并用时，STOP= 的值界定最佳回归模型可含之自变量个数的上限。此选项的内设值等于 MODEL 指令中所界定的自变量之总个数。此选项只可与 RSQUARE，ADJRSQ，CP，MAXR，MINR 等选择法联合使用。

(6) SLENTRY (或 SLE)= 统计显着的程度

在顺向选择与逐步排除法中，这个指令可用来决定某一个变量是否有资格被纳入回归模型中。内设值分别是 .50 (顺向选择法) 和 .15 (逐步排除法)。

(7) SLSTAY (或 SLS)=统计显着的程度

在反向淘汰与逐步排除法中，这个指令可用来决定某一个变量是否应继续被保留在回归模型中。内设值分别是 .10 (反向淘汰法) 和 .15 (逐步排除法)。

(8) BEST=正整数 (如 2)

这个选项只可与 SELECTION=RSQUARE，ADJRSQ，或 CP 等联用。若 SELECTION=ADJRSQ 或 CP，则 BEST= 界定的值代表最好的几个 (如 2 个) 回归模型；这几个模型的 AdjR^2 或 Cp 值都是最佳的。

当 SELECTION=RSQUARE 时，选项 BEST= 的值仍代表最好的几个 (如 2 个) 的回归模型。不过此时 REG 程序会在 3 个自变量或 4 个或 ... P 个自变量的排列组合中每次找出 (2 个) 最佳的模型。

(9) GROUPNAMES='变量组名' ... '变量组名'

这个选项的功能是为变量组合数的小组命名。只可与 FORWARD，BACKWARD 及 STEPWISE 等方法联用。同一小组的变量名称串以大括号括起，在分析过程中视为一个自变量。举例来说，下式的指令会让 REG 程序认为只有两个自变量，即 'HT 与 WT' 或'IQ_SCORE'：

```
PROC REG;
    MODEL Y={HT WT} IQ_SCORE/SELECTION=F;
    GROUPNAMES='HTWT' 'IQ_SCORE';
```

变量组名不可超过八个字母。若读者使用大括号将变量串括起，然而未使用 GROUPNAMES=的指令，则 REG 程序自动依 GROUP1， GROUP2... 等顺序来命名。

(10) NOINT

规定回归模型中不包括截距。

第六类选项　与 SELECTION=RSQUARE， ADJRSQ， CP 的设定有关，有十四个选项：

(1) ADJRSQ

要求计算出每一模型 (针对自由度) 矫正过后的 R^2 值。这个矫正过后的 R^2 值是母群内 R^2 值的不偏估计值。

(2) AIC

要求计算出每一模型的赤池资讯量指标 (Akaike's Information Criterion) 。有关这个指标的详细说明，请见赤池氏 1969 年的原着。

(3) BIC

要求计算出每一个模型的贝叶斯信息指标 (Bayesian Information Criterion)。详情请参考苏氏 (Sawa) 1978 年的原作。

(4) CP

要求计算出每一个模型的玛氏 (Mallow)的 Cp 统计值。详情请参考玛氏的原作(1973)。

(5) GMSEP

要求计算出样本的估计误差之平均方 (Mean Square)。这时 RSQUARE 程序假设自变量与因变量的分配均属常态分配。

(6) JP

要求计算出 Jp 值。Jp 值是在 (甲) 自变量的值固定且 (乙) 模型是正确的两假设下的误差均方值。Jp 值又称终极预测误差 (FPE=Final Prediction Error)。

(7) MSE

要求计算出每一回归模型的误差平均方，此指令不需任何假设。

(8) RMSE

要求印出上述 MSE 的平方根。

(9) PC

要求计算出每一个模型的阿美氏 (Amemiya) 的预测指标。请参考其 1976 年的原作。

(10) SBC

要求计算出每一个模型的休瓦氏 (Schwarz) 的贝氏指标。请参考其 1978 年的原作。

(11) SIGMA=正有理数

规定以误差的 "真正" 标准差来计算 Cp 及 BIC 值 (见上面的解释) ，否则 REG 程序会自动以误差的估计标准差来计算 Cp 及 BIC。

(12) SP

要求计算出每一个模型的 Sp 统计值。请参看 Hocking (1976) 的原作。

(13) SSE

要求计算出每一个模型的误差平方和。

(14) B

要求计算出每一个模型的回归系数之估计值。

指令 #3　VAR 变量名称串；

此指令的功用是要求将那些在 MODEL 指令中未提到的数值变量也一起包括在向量内乘积矩阵里，此选项须与选项 OUTSSCP= 并用。

指令 #4　FREQ 变量名称；

FREQ 变量的值表示各观察体重复出现的次数。

指令 #5　WEIGHT 变量名称；

WEIGHT 变量的值代表各观察体的加权比重。若此加权值和观察体的变异数倒数成比例，则所求得的参数估计值称为最佳线性不偏估计值 (Best Linear Unbiased Estimates，简称 B.L.U.E.)。

指令 #6　ID 变量名称；

指明一个变量，其功用在于识别观察体。

指令 #7　BY 变量名称串；

REG 程序依据此指令所列举的变量将文件分成几个小的文件，然后对每一个小的文件分别执行分析。当读者选用此指令时，文件内的数据必须先按照 BY 变量串的值做由小到大的重新排列，这个步骤可藉 PROC SORT 达成。

指令 #8　ADD 变量名称串；

此指令可要求 REG 程序在交谈式的分析法中将额外的自变量添加在模型理。这个指令所提到的变量串必须是 VAR 指令中也提到的，或已在 MODEL 指令中出现过的，甚至是 (指令 #9) DELETE 已剔除的变量方可。

指令 #9　DELETE 变量名称串；

此指令的作用与上述 ADD 指令刚好相反；要求 REG 程序在交谈式的分析法中将某些自变量从模型中排除。

指令 #10　RESTRICT 等式 1，等式 2，…；

本指令的功用在于以等式限制 MODEL 指令中参数的估计。这些等式之间可以逗号分开。等式的写法如下：

±转换变量 { ±转换变量} { = ±转换变量 { ±转换变量...}}

转换变量可以是一个变量，一个常数或是两者的乘积。这些变量必须是在 MODEL 指令中提过的自变量 (回归模型的截距也算是一个有效的变量)。若等式中不含等号，如：A，则它相当于 A=0。大括号 {} 可有可无。下面举一些 RESTRICT 指令的例子：

```
RESTRICT A+B=1;
RESTRICT A=B=C;
RESTRICT A=B, B=C;
RESTRICT 2*F=G + H, INTERCEPT=0;
RESTRICT F=G=H=INTERCEPT;
```

请读者注意：同一个 RESTRICT 指令内各等式之间必须一致，不可有冲突。请看下一个不正确的 RESTRICT 指令：

```
RESTRICT F - G=0,
         F - INTERCEPT=0,
```

```
G - INTERCEPT=1;
```

上面指令的错误在于这三个等式之间有不一致的情形。若此指令被用在回归模型上，则其中一个参数的值必须等于 0。其自由度也是 0，这是因为 REG 程序无法执行这些限制条件。

现在，我们来解释这些等式的意义。比方说：RESTRICT A+B=1；表示 A 与 B 这两个自变量所对应的参数，它们的估计值之和必须是 1。另外，假如 RESTRICT A=B=C；则表示 A，B 与 C 这三个自变量所对应的参数估计值必须相等。

指令 #11　TEST 等式 1，等式 2，…／选项；

本指令的等式与上述 RESTRICT 指令里的等式写法非常类似。同样的，回归模型的截距也可用在等式里。但在此等式中的变量必须是自变量。此指令的目的在于检验前述 MODEL 指令中有关参数的假设。每一个等式界定一个线性假设，以供检验。下面是一些示范的例子：

```
MODEL Y=A1 A2 B1 B2;
TEST  A1+A2=1;
TEST  B1=0, B2=0;
TEST  B1, B2;
```

第二和第三等式是完全相同的，因为假如等式中不含等号，则该项的内设值为 0。删除号 (/) 后的选项只有一个 (即 PRINT)，这个选项要求印出统计检验的过程。

指令 #12　MTEST 等式 1，等式 2，…／选项串；

本指令的等式与前述 RESTRICT 指令里的等式写法完全相同，其目的是检验多变量回归模型中的假设。每一个等式界定一个线性假设以供检验。请看下面四个例子：

例 1

```
MODEL Y1 Y2=X1 X2 X3;
MTEST X1, X2;
```

这个 MTEST 指令检验 X1 与 X2 的参数在预测 Y1 与 Y2 这两个因变量上是否皆为 0。

例 2

```
MODEL Y1 Y2=X1 X2 X3;
MTEST Y1-Y2, X1;
```

这个 MTEST 指令检验 X1 的参数在预测 Y1 与 Y2 上是否相同。

例 3

```
MODEL Y1 Y2=X1 X2 X3;
MTEST Y1-Y2;
```

这个 MTEST 指令检验除截距以外的参数在预测 Y1 与 Y2 上是否相同。

例 4

```
MODEL Y1 Y2=X1 X2 X3;
MTEST ;
```

例 3 和例 4 都是用来检验除截距以外的参数在预测 Y1 与 Y2 的过程中是否具同等的值。

删除号 (/) 后的选项有三个：

(1) PRINT

印出 F 检验的分子与分母矩阵。

(2) CANPRINT

印出自变量向量与因变量向量之间的典型相关系数。若读者在 MTEST 指令后选用 CANPRINT 选项，则 REG 程序自动印出所有因变量与所有自变量之间的典型相关系数。

(3) DETAILS

印出因变量矩阵与其有关的统计值。

指令 #13　OUTPUT OUT＝输出文件名称　关键字=变量名称串；

本指令包括两个部分：OUT= 与 关键字=，分别介绍如下：

OUT=输出文件名称

这个文件含原输入文件的所有变量，以及本指令中所提到的变量 (如：PREDICTED，RESIDUAL 等，详情见下面的说明)。

关键字=变量名称串;

下列是十六种关键字及其定义：

(1) PREDICTED (或 P)= 预测值。

(2) RESIDUAL (或 R)= 预测误差。

(3) L95M= 因变量平均数的 95% 信赖区间之下限。

(4) U95M= 因变量平均数的 95% 信赖区间之上限。

(5) L95= 因变量预测值的 95% 信赖区间之下限。这个值考虑了抽样误差及参数估计值的变异数。

(6) U95= 因变量预测值的 95% 信赖区间之上限。这个值考虑了抽样误差及参数估计值的变异数。

(7) STDP= 预测值平均数的标准误。

(8) STDR= 误差的标准误。

(9) STDI= 个别预测值的标准误。

(10) STUDENT= 经过标准化后的误差。

(11) COOKD= 库格氏影响力的统计值。

(12) H= 影响力，定义是 $X_i(X'X)^{-1}X_i'$。

(13) PRESS= 定义是误差除以 (1-H)。

(14) RSTUDENT= 除去一个观察体后所求得的该观察体的标准化误差。

(15) DFFITS= 将观察体对预测值的影响力加以标准化。

(16) COVRATIO= 将观察体对回归系数的共变异数之影响力加以标准化。

下面举一个例子说明这些关键字的撰写：

```
PROC REG DATA=A;
     MODEL Y Z=X1 X2;
     OUTPUT OUT=B
           P=YHAT ZHAT
           R=YRESID ZRESID;
```

这些指令最后产生的输出文件叫 B。B 除了包括输入文件 (A) 的原有数据外，还包括了下列四个变量：YHAT，ZHAT (Y 与 Z 的预测值)，YRESID，ZRESID (Y 与 Z 的预测误差)。

指令 #14　REWEIGHT 加权条件式 ALLOBS / 选项串;(或 REWEIGHT STATUSUNDO;)

这个指令的目的是重新界定观察体的加权值，有两种语法：(甲) REWEIGHT 加权条件式 ALLOBS / 选项串；或 (乙) REWEIGHT STATUSUNDO;。现分述如下：

(甲) REWEIGHT 加权条件式 ALLOBS / 选项串;

分三部分，即加权条件式、ALLOBS， 以及选项串。

加权条件式

加权条件式无非是界定一个或一个以上的加权范围，在这个 (些) 范围内，观察体的加权值增加、减少或等于零 (即将观察体暂时自分析中排除，打入冷宫也!)。以下是三个条件式的示范写法：

例 1　指明加权的对象

```
REWEIGHT NAME='Lee';                    (重新界定 Lee 的加权值)
REWEIGHT NAME='LEE';                    (重新界定 LEE 的加权值，与上式不同)
REWEIGHT TEMP=38;                       (重新界定变量 TEMP=38 的加权值)
REWEIGHT NAME='Sue'NAME='Steve';        (重新界定 Sue 或者 Steve 的加权值)
```

例 2　指明加权的范围

```
REWEIGHT OBS. LE 10;      (若观察体的识别号小于或等于 10，则其加权值重新
                           界定)
REWEIGHT RESIDUAL.  >2;   (若预测误差大于 2，则其对应的观察体之加权值会改
                           变)
```

例 3　多重条件式的混合撰写

```
REWEIGHT OBS. LE 10 AND RESIDUAL.>2;     (在例 2 的双重条件成立下，观察
                                          体的加权值才改变)
```

```
REWEIGHT P.>=100 OR P.<=-100;        (当预测值大于或等于 100，或小于或
                                      等于-100 时，其对应的观察体之加权
                                      值会改变)
```

从前面的三个例子里，我们可以归纳出下列三点撰写条件式的规则：

1. 加权的对象可用一个变量名称 (如 NAME， TEMP) 或观察体的识别号 (即 OBS) 或 OUTPUT 指令中所提及的关键字 (如 RESIDUAL.， P.) 来标明。
2. 加权的范围可用六种比较的运算符号来划分。这六种符号的解释列表如下：

界定加权条件式的符号	定　　义
LT 或 <	(小于)
GT 或 >	(大于)
EQ 或 =	(等于)
LE 或 <=	(小于或等于)
GE 或 >=	(大于或等于)
NE 或 ^=	(不等于)

3. 多重条件式的连接词不外乎 AND (或作 &，同时成立)、OR (或作，亦即任何一个条件成立即可)。

ALLOBS

这个关键字的意思是说所有的观察体之加权值都要改变。因此，不可与上述的条件式同时使用，这也就是竖号 | 所代表的意义。请看下列两个例子：

```
REWEIGHT ALLOBS/RESET;   (将所有的加权值恢复成 1)
REWEIGHT ALLOBS;         (将所有的加权值设定为零，如此会令分析停止—这是一
                          个无意义的加权程序)
```

选项串

删除号（/）后的选项有三个，列举如下：

(1) RESET

将加权值恢复至起初所界定的值。这个起初的加权值随 WEIGHT 指令的有无而不同。若 WEIGHT 指令不存在，则起初的加权值等于 1；若 WEIGHT 指令已存在，则起初的值就是各观察体在 WEIGHT 变量上的值。

(2) WEIGHT= 正实数或 0

这个选项直接了当地将加权值改变成一个新的正实数或 0，如：

```
REWEIGHT/WEIGHT=.5;        (将所有的加权值定为 1/2)
REWEIGHT R.>2/WEIGHT=0.1;(将预测误差大于 2 的那些观察体之加权值定为 1/10)
```

(3) NOLIST

要求 REG 程序勿将加权值改变的那些观察体的识别号 (亦即 OBS 的号码) 打印在日志窗口下 (Log Window) 或日志文件 (Log File) 内。

若读者不选用以上任何一个选项，则新的加权值自动变成零，如 REWEIGHT

ALLOBS；的例子。根据这个内设的加权原则，前述加权条件式一节中所列举的例 1～例 3 的程序都会导致加权值等于零的后果。

（乙）REWEIGHT STATUS UNDO;

这种语法含两个关键字：STATUS 或 UNDO，二者不可同时联用。它们的功能如下所阐释：

STATUS

要求 REG 程序将各加权值有变动的观察体识别号 (亦即 OBS 的号码) 及其 "新" 的加权值打印在日志窗口下 (Log Window) 或日志文件 (Log File) 内。

若加权值变成零，则该观察体自分析中被去除 (Deleted) 然而它 (们) 仍存留在输入文件内。

UNDO

宣告在此之前最后一个 REWEIGHT 指令所界定的加权值无效。下面是一个 UNDO 的例子：

```
REWEIGHT NAME='Yang'/WEIGHT=100;    (将 Yang 的加权值变成 100)
REWEIGHT;                           (将 Yang 的加权值贬为零)
REWEIGHT UNDO;                      (将 Yang 的加权值恢成 100)
```

注意事项

当读者界定 REWEIGHT 指令时，应注意以下的几种限制：

- 这个指令不适用于非原始数据的文件 (亦即 TYPE=CORR， COV， UCORR，UCOV， SSCP 的文件不可与 REWEIGHT 联用)。
- 这个指令不会自动导出新的参数估计值或因变量的预测值、预测误差等。然而，若在 REWEIGHT 指令后紧接着的指令是 REFIT，PLOT，或 PRINT，则 PROC REG 会重新计算回归模型里一切受到新加权值影响的统计量 (参考第 18.4 节有关适用于交谈式环境的指令之范例)。

指令 #15　REFIT;

这个指令的功用在于将回归模型的参数估计值以及 Y 的预测值重新计算一次。其位置最好紧接在 REWEIGHT 或 PAINT 指令之后，这是因为这两个指令只能改变观察体的加权值 (REWEIGHT 指令) 或强化某些观察在点状图上的视觉效果 (PAINT 指令)，然而却不能要求 REG 程序重新计算参数的估计值、Y 的预测值或相关统计量。因此，指令 REFIT 就是用来弥补这个缺陷的。请看下面程序的示范：

```
PAINT NAME='HWANG'/SYMBOL='H';
PLOT R.*P.;
REWEIGHT R.>20;
REFIT;
PAINT NAME='HWANG'/SYMBOL='H';
PLOT R.*P.;
```

上述的程序会产生两个点状图：第一个点状图是根据原数据组的资料所产生的；第二个点状图则是将预测误差过大 (超过 20) 的观察体自分析中排除后，重新计算各观察体之误差与预测值所产生的。因此，这两个图形的形状应是不同的；它们相同的地方是：两个图形上姓 "HWANG" 的点都以 "H" 来表示。

REFIT 指令的功能可被 PLOT 或 PRINT 取代。因此，当 REWEIGHT 或 PAINT 指令之后已有了 PLOT 或 PRINT 的指令，则读者可以不必再使用 REFIT 的指令。此时的参数值或因变量的预测值、预测误差等已在界定 PLOT 或 PRINT 的同时，重新被计算过了。

指令 #16　PAINT 强化条件式 ALLOBS / 选项串；(或 PAINT STATUSUNDO；)

这个指令的功能旨在强调数据中的某些观察体 (或点状图上所对应的点)。强调的方式是采用一些特殊的符号，如人名的第一个字母或键盘上不常用的键，将某些观察体的独特性显在点状图上。由于 PAINT 指令没有绘图的功能，因此每一个 PAINT 指令之后必须立刻接上 PLOT 的指令，否则无法显出强调的效果。PLOT 指令的说明见下一段 (指令 #17)。

撰写 PAINT 指令的语法与指令 #14 REWEIGHT 十分接近，也分两种：(甲) PAINT 强化条件式 ALLOBS / 选项串；或 (乙) PAINT STAT | USUNDO;。现分述如下：

(甲) PAINT 强化条件式 | ALLOBS / 选项串

分三部分，即强化条件式、| ALLBOS， 以及选项串。

强化条件式

强化条件式无非是界定一个或一个以上的强化范围，在这个 (些) 范围内的观察体都会被特殊的符号强调。以下是条件式的三个示范写法：

例 1　指明强化的对象

```
PAINT NAME='Lee';                    (强调名叫 Lee 的观察体)
PAINT NAME='LEE';                    (强调名叫 LEE 的观察体，与上式不同)
PAINT TEMP=38;                       (强调变量 TEMP=38 的观察体)
PAINT NAME='Sue' | NAME='Steve';     (强调名叫 Sue 或 Steve 的观察体)
```

例 2　指明强化的范围

```
PAINT OBS. LE 10;        (强调那些识别号 OBS 小于或等于 10 的观察体)
PAINT RESIDUAL. >2;      (强调预测误差大于 2 的观察体)
```

例 3　多重条件式的混合撰写

```
PAINT OBS. LE 10 AND RESIDUAL，>2;   (强调那些符合例 2 之双重条件的观察体)
PAINT P.>=100 OR P，<=-100;          (当预测误差大于或等于 100，或小于 / 等于
                                      -100 时，其对应的观察体应被强调)
```

从前面的三个例子里，我们可以归纳出下列三点撰写条件式的规则：

1. 强化的对象可用一个变量名称 (如 NAME， TEMP) 或观察体的识别号 (即 OBS)

或 OUTPUT 指令中所提及的关键字 (如 RESIDUAL.， P.) 来标明。

2. 强化的范围可用六种比较的运算符号来划分。这六种符号的解释列表如下：

界定加权条件式的符号	定　　义
LT 或 <	(小于)
GT 或 >	(大于)
EQ 或 =	(等于)
LE 或 <=	(小于或等于)
GE 或 >=	(大于或等于)
NE 或 ^=	(不等于)

3. 多重条件式的连接词不外乎 AND (或作 &，同时成立)、OR（或作 | ，任何一个条件成立即可)。

ALLOBS

这个关键字的意思是说所有的观察体都要被强调。因此，不可与上述的条件式同时使用，这也就是竖号 | 所代表的意义。请看下面两个例子：

```
PAINT ALLOBS/RESET; (将所有的强化符号恢复成 PLOT 内设的符号，亦即取消强化的功能)
PAINT ALLOBS;    (将所有的强化符号恢复成 PAINT 内设的符号，即选项 SYMBOL=所界
                  定的符号或"@" ー第一次使用 PAINT 指令的内设符号)
```

选项串

删除号 (/) 后的选项有三个，列举如下：

(1) SYMBOL='强化的符号'

这个选项界定强化的特殊符号，它必须是一个字元，如 '#' 或 'P'。若省略此选项，则 REG 自动采用此指令之前最后一个 PAINT 指令所界定的强化符号或 SAS 内设的符号，即 @。

一般而言，读者不应采用 1 到 9 的阿拉伯数字或星号 (*)，因为它们就是 PLOT 指令的内设符号而且代表图形上点重叠的次数。'*' 代表 10 点或 10 点以上重叠在一起。若 SYMBOL="，则强化的效果不存在；若 SYMBOL=' '，则所有的点都自动在图形上消失ー这是因为读者不小心将强化的符号设定为空白的缘故。

(2) RESET

规定将所有强化的符号恢成 PLOT 指令内设的绘图符号，亦即取消强化的功能。PLOT 指令的内设绘图符号由选项 SYMBOL= 决定；否则采用 1 到 9 的阿拉伯数字或星号 (*)。

(3) NOLIST

要求 REG 程序勿将被强调的观察体之识别号 (亦即 OBS 的号码)、强化的符号、以及被强调的总观察体数都印在日志窗口 (Log Window) 或日志文件 (Log File) 内。若你不采用此选项，则以上三种信息都会被显现出来。

（乙）PAINT STATUSUNDO;

这种语法含两个关键字：STATUS 或 UNDO，二者不可同时联用。它们的功能如下

所阐释：

STATUS

要求 REG 程序将被强调的观察体之识别号 (亦即 OBS 的号码) 及其强化符号一种打印在日志窗口 (Log Window) 下或日志文件 (Log File) 内。

UNDO

宣告在此之前最后一个 PAINT 指令执行后的结果无效。下面是一个 UNDO 的范例：

```
PAINT OBS.<=5/SYMBOL='A';  (前五个观察体以 'A' 表示)
PAINT OBS.=1/SYMBOL='T';   (第一个观察体以 'T' 表示)
PAINT UNDO;                (第一个观察体以 'A' 表示)
```

上述程序中最后一个 PAINT 指令若改写成

```
PAINT/RESET;
```

则第一个观察体之强化效果立即消失。

注意事项

当读者界定 PAINT 指令时，应注意以下的几种限制：

- 这个指令不适用于非原始数据的文件 (亦即 TYPE=CORR， COV， UCORR，UCOV， SSCP 的文件不可与 PAINT 联用)。
- PAINT 指令之后应紧接 PLOT 指令，否则强化的效果不会彰显出来。
- 在任何一个 PLOT 指令之前的 PAINT 指令都是有效的，其效力贯穿整个 REG 程序。
- 若数个被强化的观察体在图形上是同一个点，则最后一个被强调的观察体所对应的符号决定点状图上的绘图记号。

指令 #17　PLOT 图形指令串 / 选项串；

这个指令直接控制报表上图形的呈现方式。有“图形指令串”与“选项串”两种控制语句。现分别解释如下：

图形指令串

图形指令串界定 X，Y 轴成为前一个 PLOT 指令重新规划。它的语法不外乎下面三种格式：

图形指令串的格式	举　例
(1) Y 轴之变量名 *X 轴之变量名；	PLOT RESIDUAL.*OBS.;
(2) Y 轴之变量串 *X 轴之变量串；	PLOT(RESIDUAL.STUDENT.)*(X1 X2 PREDICTED.);
(3) 重新规划前一个 PLOT 指令	PLOT;

根据以上第 (1) 种格式的写法，报表上只产生一个图形，其 Y 轴的界定在前，X 轴的界定在后。此外，Y 轴或 X 轴的定义可以是 VAR 指令或 MODEL 指令中提过的变量名称，或是 OUTPUT 指令中有关统计量的关键字，甚至观察体的识别号 (即 OBS. 变量) 也可用来界定图形的参考轴。

根据以上第 (2)种格式的写法，则你在报表上会看到多重的图形。图形的数目就是定义 Y 轴与 X 轴两两变量的组合数。因此，上述的例子会产生六个图形。它们分别由 (RESIDUAL. *X1)、(RESIDUAL.*X2)、(RESIDUAL.*PREDICTED.)、(STUDENT.*X1)、(STUDENT.*X2)， 以及 (STUDENT.*PREDICTED.) 等组合而来。

上述第 (1) 与第 (2) 种的格式可同时选用绘图的符号，请看下面的示范：

```
PLOT Y*X='*';
```

或

```
PLOT Y*(P. R.)='#';
```

第 (3) 种格式的写法是用来重新界定，在此 PLOT 指令之前，最后一个 PLOT 指令的执行一亦即将最后一次执行的 PLOT 指令之任何选项去除，重新绘制点状图。所以，这种写法不需提到 Y 轴或 X 轴是如何定义的。

选项串

删除号 (/) 之后的选项有七个，分述如下：

(1) SYMBOL='绘图的符号"

这个选项的功能、语法与 PAINT 指令中同名的选项完全相同 。单括号内的符号必须是一个字元，如 '#' 或 '$'。一般而言，读者不应采用 1 到 9 的阿拉伯数字或星号 (*)，因为它们就是 PLOT 指令的内设符号而且代表图形上点重叠的次数。因此，'1' 代表一个点 (或 1 个观察体)，'2' 代表两点；余此类推，'*' 代表 10 点或 10 点以上重叠。

若 SYMBOL="，则 PLOT 采内设的符号，若 SYMBOL=' '，则所有的点会从图形上消失，因为绘图的符号在此语法中是一个空白。

注意事项

- PAINT 指令中界定的绘图符号取代 PLOT 指令所界定的符号或内设的绘图符号。
- 删除号 (/) 前所界定的绘图符号，其效力超过 SYMBOL= 所界定的符号。因此，根据下面的程序

```
PLOT Y*X Y*Z='#'/SYMBOL='*';
```

Y 与 X 的图形点会以 '*' 来表示，然而 Y 与 Z 的图形则以 '#' 来显示。

- 若读者决定采用 OVERLAY 或 COLLECT 的选项，则最好同时界定绘图的符号以免混淆。有关 OVERLAY， COLLECT 选项的说明，请看以下 (2)、(3) 节。

(2) OVERLAY

此选项要求将两个或两个以上的图形重叠在一起。这个重叠的图形会以第一图的变量来定义 Y 轴与 X 轴。当读者选择此选项时，最好在 SAS 程序的起首宣告图形将会重叠地打印。这个宣告的语句是：

```
OPTIONS=OVP;
```

如此宣告之后，重叠图上即使有一点代表不止一个图形的数据，它的绘图符号也将揉合原来图形所有的绘图符号。否则，这类数据将以第一图所用的符号来表示。

请读者特别注意，OVERLAY 选项是将同一个 PLOT 指令所界定的图形重叠地画(见本章之例一的示范)。下面 (3) 所介绍的 COLLECT 选项则是将不同的 PLOT 指令所界定的图形重叠地画出来，两者的作用迥异。

(3) COLLECT

如上节所述，COLLECT 的作用是将不同之 PLOT 指令所界定的图形串连起来，并且技巧地刻划坐标轴。比方说，执行下列的程序后产生两个图形：

```
PLOT RESIDUAL.*PREDICTED. Y*X/COLLECT;
RUN;
```

若在这个程序之后再附加下面的一个程序，

```
PLOT RESIDUAL.*X;
RUN;
```

则 (RES1DUAL.*X) 的图会覆盖在 (RESIDUAL.*PREDICTED.) 的图形上——这是第一图。第二图则是不重叠的部分，亦即 (Y*X) 界定的图。因此，这四行指令执行的结果仍是两个图。若一旦选用了 COLLECT 的选项，则其效力贯穿整个 REG 程序。唯一可以中止这个指令的方法是界定 NOCOLLECT 的指令 [参见下面 (4) 的说明]。

(4) NOCOLLECT

中止上述 COLLECT 的效力—这是 PLOT 指令的内设值。

(5) CLEAR

适用于一个新的 COLLECT 选项之前，其作用是将以前旧的 COLLECT 指令所造成的绘图效果自记忆体空间内清除。如此，新的 COLLECT 选项所导出的贯穿效果可在宣告 CLEAR 选项之后立即奏效。

(6) VPLOTS= 正整数 (如 3)

此选项界定一页报表纸上从上端至下端可容纳多少图形。比方说，你想将下列程序所产生的六个图形两两并排，分三段印在同一页的报表纸上，则可定 VPLOTS=3，HPLOTS=2 (亦即两个图形并排打印出)：

```
PLOT (Y1 Y2)*(X1 X2 X3)/VPLOTS=3 HPLOTS=2;
RUN;
```

若一次界定此选项，则其效力贯穿整个 REG 程序。此选项的内设值是 1。

(7) HPLOTS= 正整数 (如 2)

此选项界定一页报表纸上从左边到右边所可容纳的图形数目，请回头参考选项 VPLOTS 的例子与解释。

关于 PLOT 指令的其它注意事项

- PLOT 指令不适用于非原始数据的文件 (亦即 TYPE=CORR，COV，UCORR，UCOV， SSCP 的文件不可与 PLOT 指令联用)。
- PLOT 指令会自动考虑在它之前透过 REWEIGHT 指令所界定的 "新" 加权值。所

以，PLOT 的报表反应了 "新" 模型所产生的统计量如参数估计值或预测误差等。

●SYMBOL=，COLLECT，VPLOTS=，HPLOTS= 四个选项的效力会贯穿整个 REG 程序，除非读者前后重复地界定。若重复地界定这些选项，则在程序之后的值取代以前的选项值。

●图形上表示负数的点，以问号 "?" 来表示，这是 PLOT 指令内设的规定。

●每次界定 PAINT 指令后，应立即加上 PLOT 的指令。

指令 #18　PRINT 选项串 ANOVA MODELDATA;

这个指令适用于交谈式的操作环境。其功能是在萤幕上打印出 MODEL 指令中所用到的选项(串)，变异数分析表，以及分析的数据内容。

语法分三部分，即选项串，ANOVA，及 MODELDATA。现说明如下：

选项串

即 MODEL 指令里所界定的二十六种选项：

ACOV，ALL，CLI，CLM，COLLIN，COLLINOINT，CORRB，COVB，DW，I，INFLUENCE，P，PARTIAL，PCORR1，PCORR2，R，SCORR1，SCORR2，SEQB，SPEC，SS1，SS2，STB，TOL，VIF 以及 XPX。

当读者列出其中几个选项时，PRINT 指令会在萤幕上立即显示它们的数值。

ANOVA

要求印出回归分析后所产生的变异数分析表。由于 PRINT 指令会考虑任何置于 MODEL 指令后的 ADD，DELETE，REWEIGHT 等指令效果，所以变异数分析表所依据的回归模型总是，在此指令之前最后一次修正的模型。

MODELDATA

要求印出与上述 ANOVA 表有关联的所有数据资料。

18.3　范　　例

例一：人口成长趋势

本文件 (USPOP) 的数据是美国自 1790 年到 1970 年间，每十年的人口总数，以千人为单位。这个人口数 (因变量 POP) 可以用时间 (自变量 YEAR) 的线性与二次式函数 (即 YEARSQ=YEAR2) 解释。回归分析则包括影响力的诊断与回归分析图。请注意程序中加注了 OPTIONS 指令以帮助控制图形的视觉效果，并且 PLOT，ADD，PRINT 等指令都是在交谈式的环境下执行的。

程　序

```
DATA USPOP;
     INPUT POP @@;
     RETAIN YEAR 1780;
     YEAR=YEAR+10;
     YEARSQ=YEAR*YEAR;
```

```
    POP=POP/1000;
    CARDS;
3929 5308 7239 9638 12866 17069 23191 31443 39818 50155
62947 75994 91972 105710 122775 131669 151325 179323 203211
. . .
;
PROC REG DATA=USPOP;
    VAR YEARSQ;
    MODEL POP=YEAR/P CLI CLM INFLUENCE DW;
    PLOT R.*P.;
    ADD YEARSQ;
    PRINT;
    PLOT;
    RUN;
    PLOT POP*YEAR='A'  P.*YEAR='P'
        U95.*YEAR='U' L95.*YEAR='L'/OVERLAY;
RUN;
```

结　果

利用一次式的模型 (即 MODEL1 POP=YEAR)，我们得到如下的回归公式：

POP 的预测值 = (-1958.366302)+(1.078795)*YEAR

这个公式说明了人口数 (POP) 与时间 (YEAR) 之间的正相关。因此，美国人口数在 1790 年到 1970 年间是逐年增加。上述的模型经 F 检定测试后达到 0.0001 的显著度 (F=201.873，DF=1，17)。复相关平方等于 0.9233，矫正过后的复相关平方等于 0.9178，这两个相关系数同时验证了人口数与时间之间的相关强度。不过， Durbin-Watson 的 DW 值偏低 (0.180) 而且误差间的一次相关性 (1st Order Autocorrelation) 偏高 (0.704) 。证明误差间并不彼此独立 (也参见误差对 PRED 的图形)。因此，这个模型不符合统计理论的要求。

接下来，我们看二次式模型 (即 POP=YEAR YEARSQ) 所得的公式：

POP 的预测值 = 20450+(-22.780606)*YEAR+(0.006346)*YEAR^2

这个模型经 F 检定测试后达到 0.0001 的显著度 (F=4641.719，DF=2，16)。复相关平方 (=0.9983) 以及矫正过的复相关平方 (=0.9981) 均十分理想。Durbin-Watson DW 值 (=1.264) 比前理想，但仍不十分接近虚无假设下的期待值 (=2)。故此模型仍有改进的必要。利用 PLOT 指令绘图后，我们可以清楚地看出 POP 与 YEAR 的关系是二次式的关系 (与一次式无关)。因此，下一步的分析可根据这个结论再考虑一个较佳的回归模型，如 POP=YEARSQ。

报表 18.1　人口成长趋势

Occurrence of Vaso-Constriction

Model: MODEL1

Dependent Variable: POP

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	663364692.26	663364692.26	201.873	0.0001
Error	17	55862925.291	3286054.4289		
C Total	18	719227617.55			

Root MSE	1812.74776	R-square	0.9223
Dep Mean	6976.74737	Adj R-sq	0.9178
C.V.	25.98271		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-1958.366302	142.80454644	-13.714	0.0001
YEAR	1	1.078795	0.07592765	14.208	0.0001

Occurrence of Vaso-Constriction

Durbin-Watson D 0.180

(For Number of Obs.) 19

1st Order Autocorrelation 0.704

Obs	Dep Var POP	Predict Value	Std Err Predict	Lower95% Mean	Upper95% Mean	Lower95% Predict	Upper95% Predict	Residual	Hat Diag Rstudent	H	Cov Ratio	Dffits
1	392.9	-2732.4	799.947	-4420.1	-1044.7	-6912.8	1448.0	3125.3	2.1066	0.1947	0.8592	1.0359
2	530.8	-1653.6	736.146	-3206.7	-100.5	-5781.5	2474.3	2184.4	1.3502	0.1649	1.0894	0.6000
3	723.9	-574.8	674.860	-1998.6	849.0	-4655.8	3506.2	1298.7	0.7624	0.1386	1.2203	0.3058
4	963.8	504.0	616.839	-797.4	1805.4	-3535.9	4543.9	459.8	0.2623	0.1158	1.2658	0.0949
5	1286.6	1582.8	563.095	394.8	2770.8	-2422.0	5587.6	-296.2	-0.1669	0.0965	1.2451	-0.0545
6	1706.9	2661.6	514.966	1575.1	3748.0	-1314.3	6637.4	-954.7	-0.5377	0.0807	1.1848	-0.1593
7	2319.1	3740.4	474.168	2740.0	4740.8	-212.9	7693.6	-1421.3	-0.8038	0.0684	1.1196	-0.2178
8	3144.3	4819.2	442.730	3885.1	5753.2	882.2	8756.1	-1674.9	-0.9501	0.0596	1.0757	-0.2393
9	3981.8	5898.0	422.747	5006.0	6789.9	1970.8	9825.1	-1916.2	-1.0932	0.0544	1.0336	-0.2622
10	5015.5	6976.7	415.873	6099.3	7854.2	3052.9	10900.6	-1961.2	-1.1198	0.0526	1.0247	-0.2639
11	6294.7	8055.5	422.747	7163.6	8947.5	4128.4	11982.7	-1760.8	-0.9988	0.0544	1.0578	-0.2395
12	7599.4	9134.3	442.730	8200.3	10068.4	5197.4	13071.3	-1534.9	-0.8668	0.0596	1.0952	-0.2183
13	9197.2	10213.1	474.168	9212.7	11213.5	6259.9	14166.3	-1015.9	-0.5690	0.0684	1.1642	-0.1542

14	10571.0	11291.9	514.966	10205.4	12378.4	7316.1	15267.8	-720.9	-0.4045	0.0807	1.2033	-0.1198
15	12277.5	12370.7	563.095	11182.7	13558.7	8365.9	16375.5	-93.2202	-0.0525	0.0965	1.2490	-0.0172
16	13166.9	13449.5	616.839	12148.1	14750.9	9409.6	17489.4	-282.6	-0.1610	0.1158	1.2726	-0.0583
17	15132.5	14528.3	674.860	13104.5	15952.1	10447.3	18609.3	604.2	0.3497	0.1386	1.2907	0.1403
18	17932.3	15607.1	736.146	14054.0	17160.2	11479.2	19735.0	2325.2	1.4482	0.1649	1.0567	0.6436
19	20321.1	16685.9	799.947	14998.2	18373.6	12505.5	20866.3	3635.2	2.5798	0.1947	0.6992	1.2686
20	.	17764.7	865.708	15938.2	19591.2	13526.4	22003.0	.	.	0.2281	.	.
21	.	18843.5	933.015	16875.0	20812.0	14542.1	23144.9	.	.	0.2649	.	.
22	.	19922.3	1001.554	17809.2	22035.4	15552.8	24291.7	.	.	0.3053	.	.

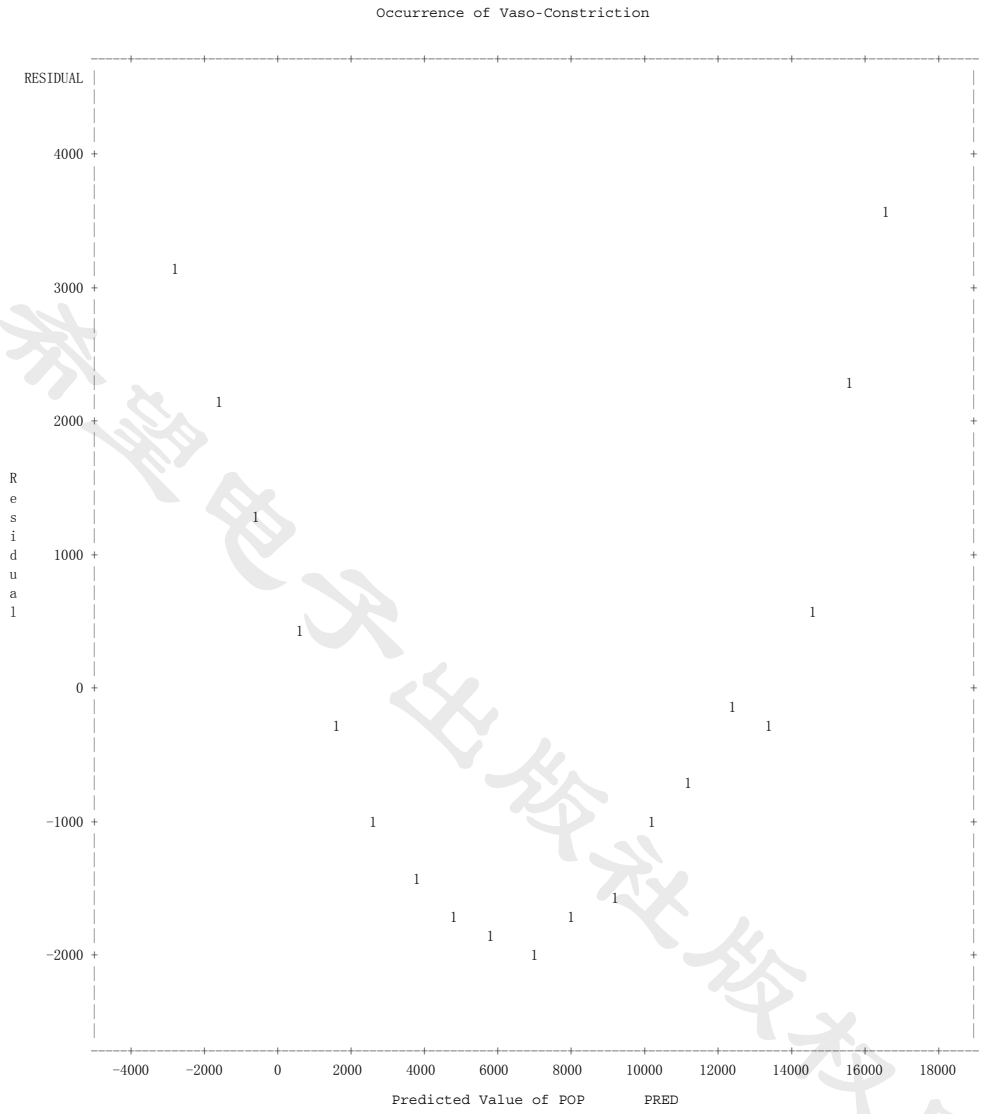
Obs	INTERCEP Dfbetas	YEAR Dfbetas
1	0.9002	-0.8849
2	0.5048	-0.4951
3	0.2462	-0.2408
4	0.0719	-0.0701
5	-0.0379	0.0368
6	-0.0977	0.0940
7	-0.1102	0.1046
8	-0.0886	0.0821
9	-0.0546	0.0471
10	-0.0077	0.0000
11	0.0361	-0.0430
12	0.0689	-0.0749
13	0.0701	-0.0741
14	0.0678	-0.0707
15	0.0112	-0.0116
16	0.0419	-0.0430
17	-0.1079	0.1105
18	-0.5202	0.5310
19	-1.0641	1.0837
20	.	.
21	.	.
22	.	.

Sum of Residuals 0

Occurrence of Vaso-Constriction

Sum of Squared Residuals 55862925.291

Predicted Resid SS (Press) 76199035.421



Model: MODEL1

Dependent Variable: POP

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	717990161.9	358995080.95	4641.719	0.0001
Error	16	1237455.6509	77340.978181		
C Total	18	719227617.55			

Root MSE	278.10246	R-square	0.9983
Dep Mean	6976.74737	Adj R-sq	0.9981
C.V.	3.98613		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	20450	843.47532634	24.245	0.0001
YEAR	1	-22..780606	0.89784904	-25.372	0.0001
YEARSQ	1	0.006346	0.00023877	26.576	0.0001

Occurrence of Vaso-Constriction

Durbin-Watson D 1.264

(For Number of Obs.) 19

1st Order Autocorrelation 0.299

Obs	Dep Var POP	Predict Value	Std Err Predict	Lower95% Mean	Upper95% Mean	Lower95% Predict	Upper95% Predict	Residual	Rstudent	Hat Diag H	Cov Ratio	Dffits
1	392.9	503.8	172.886	137.3	870.3	-190.3	1198.0	-110.9	-0.4972	0.3865	1.8834	-0.3946
2	530.8	503.9	139.086	209.0	798.7	-155.3	1163.1	26.9101	0.1082	0.2501	1.6147	0.0625
3	723.9	630.8	113.037	391.2	870.5	-5.5395	1267.2	93.0534	0.3561	0.1652	1.4176	0.1584
4	963.8	884.7	95.711	681.8	1087.6	261.2	1508.2	79.0849	0.2941	0.1184	1.3531	0.1078
5	1286.6	1265.5	87.208	1080.6	1450.4	647.6	1883.4	21.1047	0.0774	0.0983	1.3444	0.0256
6	1706.9	1773.2	85.784	1591.3	1955.0	1156.2	2390.1	-66.2872	-0.2431	0.0951	1.3255	-0.0788
7	2319.1	2407.8	88.351	2220.5	2595.1	1789.2	3026.4	-88.6908	-0.3268	0.1009	1.3214	-0.1095
8	3144.3	3169.3	92.018	2974.2	3364.4	2548.3	3790.3	-25.0061	-0.0923	0.1095	1.3605	-0.0324
9	3981.8	4057.7	94.873	3856.6	4258.9	3434.8	4680.6	-75.9331	-0.2820	0.1164	1.3519	-0.1023
10	5015.5	5073.1	95.925	4869.7	5276.4	4449.4	5696.7	-57.5718	-0.2139	0.1190	1.3650	-0.0786
11	6294.7	6215.3	94.873	6014.2	6416.4	5592.4	6838.2	79.3778	0.2949	0.1164	1.3499	0.1070
12	7599.4	7484.5	92.018	7289.4	7679.6	6863.5	8105.5	114.9	0.4265	0.1095	1.3144	0.1496
13	9197.2	8880.6	88.351	8693.3	9067.9	8262.0	9499.1	316.6	1.2189	0.1009	1.0168	0.4084
14	10571.0	10403.5	85.784	10221.7	10585.4	9786.6	11020.5	167.5	0.6207	0.0951	1.2430	0.2013
15	12277.5	12053.4	87.208	11868.6	12238.3	11435.6	12671.3	224.1	0.8407	0.0983	1.1724	0.2776
16	13166.9	13830.2	95.711	13627.4	14033.1	13206.8	14453.7	-663.3	-3.1845	0.1184	0.2924	-1.1673
17	15132.5	15734.0	113.037	15494.3	15973.6	15097.6	16370.4	-601.5	-2.8433	0.1652	0.3989	-1.2649
18	17932.3	17764.6	139.086	17469.8	18059.5	17105.4	18423.8	167.7	0.6847	0.2501	1.4757	0.3954
19	20321.1	19922.1	172.886	19555.6	20288.6	19228.0	20616.3	399.0	1.9947	0.3865	0.9766	1.5831
20	.	22206.6	213.482	21754.0	22659.2	21463.4	22949.8	.	.	0.5893	.	.
21	.	24618.0	260.191	24066.4	25169.5	23810.6	25425.3	.	.	0.8753	.	.
22	.	27156.2	312.571	26493.6	27818.9	26269.3	28043.2	.	.	1.2632	.	.

Obs	INTERCEP Dfbetas	YEAR Dfbetas	YEARSQ Dfbetas
1	-0.2842	0.2810	-0.2779
2	0.0376	-0.0370	0.0365
3	0.0666	-0.0651	0.0636
4	0.0182	-0.0172	0.0161
5	-0.0030	0.0033	-0.0035
6	0.0296	-0.0302	0.0307
7	0.0609	-0.0616	0.0621
8	0.0216	-0.0217	0.0218
9	0.0743	-0.0745	0.0747
10	0.0586	-0.0587	0.0587
11	-0.0784	0.0783	-0.0781

12	-0.1018	0.1014	-0.1009
13	-0.2357	0.2338	-0.2318
14	-0.0811	0.0798	-0.0784
15	-0.0427	0.0404	-0.0380
16	-0.1531	0.1636	-0.1747
17	-0.4843	0.4958	-0.5076
18	0.2240	-0.2274	0.2308
19	1.0902	-1.1025	1.1151
20	.	.	.
21	.	.	.
22	.	.	.

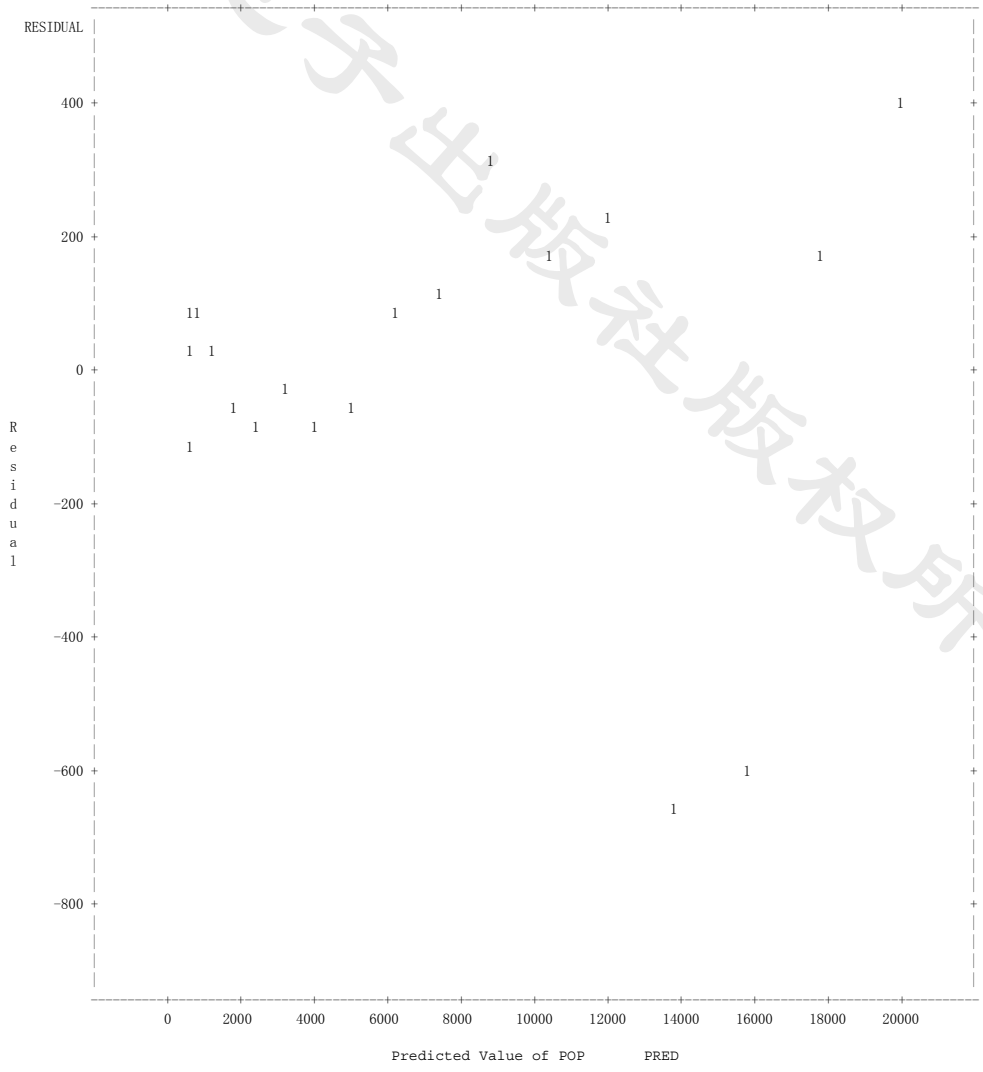
Sum of Residuals -2.53E-9

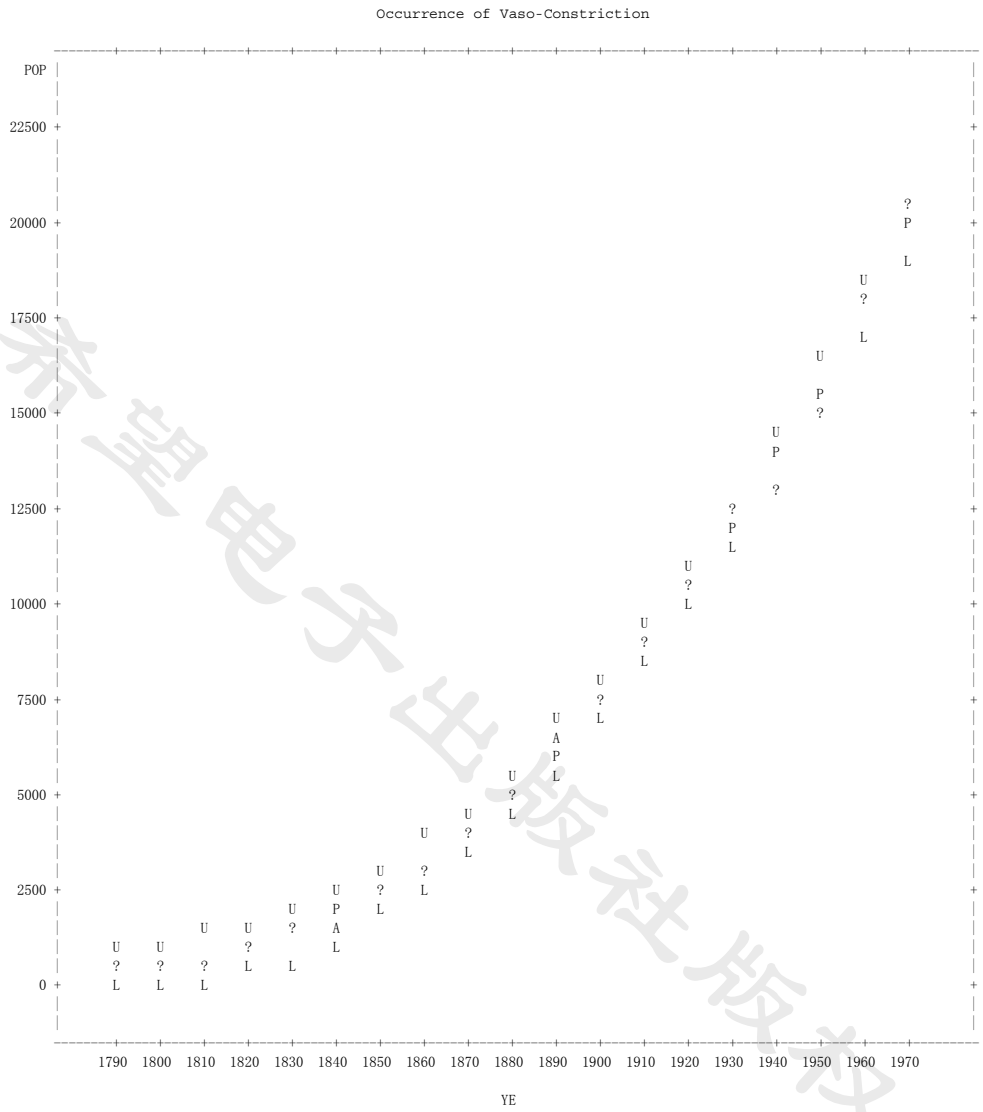
Occurrence of Vaso-Constriction

Sum of Squared Residuals 1237455.6508

Predicted Resid SS (Press) 1885492.4307

Occurrence of Vaso-Constriction





例二：预测人体吸入氧气的效率

本资料档（FITNESS）的数据来自一群中年男士的健康状态资料，由 Linnerud 提供。每一名男士提供七个数据，分别是：年龄（AGE），体重（WEIGHT），吸氧的效率（OXY），跑 1.5 英里所需的时间——以分钟计（RUNTIME），休息时的心跳（RSTPULSE），跑步时的心跳率（RUNPULSE），和最高心跳率（MAXPULSE）。其中，吸氧效率（OXY）是因变量，另外六个均是自变量。

分析的过程是先用逐步排除法，再用最大相关法，以便找出一个又精简又有效的递归模型。

程　序

```
DATA FITNESS;
     INPUT AGE WEIGHT OXY RUNTIME RSTPULSE RUNPULSE MAXPULSE @@;
     CARDS;
44 89.47 44.609 11.37 62 178 182    51 69.63 40.836 10.95 57 168 172
40 75.07 45.313 10.07 62 185 185    51 77.91 46.672 10.00 48 162 168
44 85.84 54.297  8.65 45 156 168    48 91.63 46.774 10.25 48 162 164
42 68.15 59.571  8.17 40 166 172    49 73.37 50.388 10.08 67 168 168
38 89.02 49.874  9.22 55 178 180    57 73.37 39.407 12.63 58 174 176
47 77.45 44.811 11.63 58 176 176    54 79.38 46.080 11.17 62 156 165
40 75.98 45.681 11.95 70 176 180    56 76.32 45.441  9.63 48 164 166
43 81.19 49.091 10.85 64 162 170    50 70.87 54.625  8.92 48 146 155
44 81.42 39.442 13.08 63 174 176    51 67.25 45.118 11.08 48 172 172
38 81.87 60.055  8.63 48 170 186    54 91.63 39.203 12.88 44 168 172
44 73.03 50.541 10.13 45 168 168    51 73.71 45.790 10.47 59 186 188
45 87.66 37.388 14.03 56 186 192    57 59.08 50.545  9.93 49 148 155
45 66.45 44.754 11.12 51 176 176    49 76.32 48.673  9.40 56 186 188
47 79.15 47.273 10.60 47 162 164    48 61.24 47.920 11.50 52 170 176
54 83.12 51.855 10.33 50 166 170    52 82.78 47.467 10.50 53 170 172
49 81.42 49.156  8.95 44 180 185
;
PROC REG DATA=FITNESS OUTEST=EST;
     MODEL OXY=AGE WEIGHT RUNTIME RUNPULSE MAXPULSE RSTPULSE
      /SELECTION=STEPWISE;
     MODEL OXY=AGE WEIGHT RUNTIME RUNPULSE MAXPULSE RSTPULSE
      /SELECTION=MAXR;
RUN;
```

结　果

根据逐步排除法的选择标准，依次进入递归模型的自变量是 RUNTIME、AGE、RUNPULSE 与 MAXPULSE。这四个变量的组合可解释 84.3% 的 OXY 之变异数。模型的形式是：

吸氧效率（OXY）的预测值=100.07909519+(-0.21265570)*AGE+(-2.76824065)* RUN-TIME+(-0.33956528)*RUNPULSE+(0.25535199)*MAXPULSE

此模型的 F 检定值高达 34.90; P=0.0001,是一个有效的递归公式。

根据最大相关法的选择标准，分析结果可归纳如下：

自变项数目	模型	R2	F 值	显著度
1	OXY=82.4+(-3.3)*RUNTIME	0.74	84.01	0.0001
2	OXY=89.2+(-0.2)*AGE+(-3.2)*RUNTIME	0.77	46.92	0.0001
3	OXY=113.1+(-0.3)*AGE+(-2.8)*RUNTIME+(-0.1)*RUNPULSE	0.82	41.09	0.0001
4	OXY=100.1+(-0.2)*AGE+(-2.8)*RUNTIME+(-0.3)*RUNPULSE+(0.3)*MAXPULSE	0.84	34.90	0.0001
5	OXY=104.0+-0.2)*AGE+(-0.1)*WEIGHT+(-2.7)*RUNTIME+(-0.4)*RUNPULSE+(0.3)*MAXPULSE	0.85	29.30	0.0001
6	OXY=104.9+(-0.2)*AGE+(-0.1)*WEIGHT+(2.6)*RUNTIME+(-0.4)*RUNPULSE+(0.3)*MAXPULSE+(-0.03)*RSTPULSE	0.86	23.62	0.0001

若以 R^2 的改变来衡量模型的精简法，则 MAXR 的分析结果以 3 各自变量的递归公式最好。因此，这两种分析所导出来的结论不尽相同！

报表 18.2　预测人体吸入氧气的效率

The SAS System

Stepwise Procedure for Dependent Variable OXY

Step 1　Variable RUNTIME Entered　R-square = 0.74338010　C(p) = 15.52440481

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	1	632.90009985	632.90009985	84.01	0.0001
Error	29	218.48144499	7.53384293		
Total	30	851.38154484			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	82.42177268	3.85530378	3443.36654076	457.05	0.0001
RUNTIME	-3.31055536	0.36119485	632.90009985	84.01	0.0001

Bounds on condition number:　1,　1

Step 2　Variable AGE Entered　R-square = 0.77019043　C(p) = 13.08167294

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	2	655.72592146	327.86296073	46.92	0.0001
Error	28	195.65562338	6.98770084		
Total	30	851.38154484			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	89.17681435	5.26828724	2002.16248393	286.53	0.0001
AGE	-0.16474787	0.09115358	22.82582161	3.27	0.0815

RUNTIME	-3.20466496	0.35275614	576.70059582	82.53	0.0001

Bounds on condition number:　1.028367,　4.11347

Step 3　Variable RUNPULSE Entered　R-square = 0.82033615　C(p) = 6.77204159

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	698.41905906	232.80635302	41.09	0.0001
Error	27	152.96248578	5.66527725		
Total	30	851.38154484			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	113.05757768	9.90850509	737.57170464	130.19	0.0001
AGE	-0.26916519	0.09046159	50.15687717	8.85	0.0061
RUNTIME	-2.82451970	0.34650163	376.44351785	66.45	0.0001
RUNPULSE	-0.13506550	0.04920123	42.69313760	7.54	0.0106

Bounds on condition number:　1.347312,　11.46113

The SAS System

Step 4　Variable MAXPULSE Entered　R-square = 0.84297752　C(p) = 5.02014687

	DF	Sum of Squares	Mean Square	F	Prob>F
egression	4	717.69550475	179.42387619	34.90	0.0001
Error	26	133.68604009	5.14177077		
Total	30	851.38154484			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	100.07909519	11.57738678	384.21858339	74.72	0.0001
AGE	-0.21265570	0.09098843	28.08629280	5.46	0.0274
RUNTIME	-2.76824065	0.33138140	358.80966599	69.78	0.0001
RUNPULSE	-0.33956528	0.11555136	44.40267890	8.64	0.0068
MAXPULSE	0.25535199	0.13188096	19.27644569	3.75	0.0638

Bounds on condition number:　8.522205,　77.34387

All variables left in the model are significant at the 0.1500 level.
No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Procedure for Dependent Variable OXY

Step	Entered	Removed	Variable In	Number R**2	Partial R**2	Model C(p)	F	Prob>F
1	RUNTIME		1	0.7434	0.7434	15.5244	84.0076	0.0001
2	AGE		2	0.0268	0.7702	13.0817	3.2666	0.0815
3	RUNPULSE		3	0.0501	0.8203	6.7720	7.5359	0.0106
4	MAXPULSE		4	0.0226	0.8430	5.0201	3.7490	0.0638

The SAS System

Model: MODEL1

Maximum R-square Improvement for Dependent Variable OXY

Step 1　Variable RUNTIME Entered　R-square = 0.74338010　C(p) = 15.52440481

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	1	632.90009985	632.90009985	84.01	0.0001
Error	29	218.48144499	7.53384293		
Total	30	851.38154484			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	82.42177268	3.85530378	3443.36654076	457.05	0.0001
RUNTIME	-3.31055536	0.36119485	632.90009985	84.01	0.0001

Bounds on condition number:　1,　1

The above model is the best　1-variable model found.

Step 2　Variable AGE Entered　R-square = 0.77019043　C(p) = 13.08167294

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	2	655.72592146	327.86296073	46.92	0.0001
Error	28	195.65562338	6.98770084		
Total	30	851.38154484			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	89.17681435	5.26828724	2002.16248393	286.53	0.0001
AGE	-0.16474787	0.09115358	22.82582161	3.27	0.0815
RUNTIME	-3.20466496	0.35275614	576.70059582	82.53	0.0001

Bounds on condition number:　1.028367,　4.11347

The above model is the best 2-variable model found.

Step 3　Variable RUNPULSE Entered　R-square = 0.82033615　C(p) = 6.77204159

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	698.41905906	232.80635302	41.09	0.0001
Error	27	152.96248578	5.66527725		
Total	30	851.38154484			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	113.05757768	9.90850509	737.57170464	130.19	0.0001
AGE	-0.26916519	0.09046159	50.15687717	8.85	0.0061
RUNTIME	-2.82451970	0.34650163	376.44351785	66.45	0.0001
RUNPULSE	-0.13506550	0.04920123	42.69313760	7.54	0.0106

Bounds on condition number:　1.347312,　11.46113

The above model is the best 3-variable model found.

Step 4　Variable MAXPULSE Entered　R-square = 0.84297752　C(p) = 5.02014687

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	4	717.69550475	179.42387619	34.90	0.0001
Error	26	133.68604009	5.14177077		
Total	30	851.38154484			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	100.07909519	11.57738678	384.21858339	74.72	0.0001
AGE	-0.21265570	0.09098843	28.08629280	5.46	0.0274
RUNTIME	-2.76824065	0.33138140	358.80966599	69.78	0.0001
RUNPULSE	-0.33956528	0.11555136	44.40267890	8.64	0.0068
MAXPULSE	0.25535199	0.13188096	19.27644569	3.75	0.0638

Bounds on condition number:　8.522205,　77.34387

The above model is the best　4-variable model found.

Step 5　Variable WEIGHT Entered　R-square = 0.85424362　C(p) = 5.15324519

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	5	727.28725277	145.45745055	29.30	0.0001
Error	25	124.09429207	4.96377168		
Total	30	851.38154484			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	103.99695538	11.71918502	390.89329919	78.75	0.0001
AGE	-0.23225266	0.09050434	32.68844975	6.59	0.0167
WEIGHT	-0.07240864	0.05208917	9.59174802	1.93	0.1768
RUNTIME	-2.68690881	0.33081007	327.46147710	65.97	0.0001
RUNPULSE	-0.36461305	0.11495463	49.93709671	10.06	0.0040
MAXPULSE	0.28996527	0.13194885	23.97135301	4.83	0.0375

```
Bounds on condition number:      8.836899,      105.3442
------------------------------------------------------------------------------------------

The above model is the best  5-variable model found.

Step 6   Variable RSTPULSE Entered  R-square = 0.85516840   C(p) =  7.00000000

                        DF         Sum of Squares      Mean Square          F   Prob>F

          Regression     6           728.07459433      121.34576572      23.62   0.0001
          Error         24           123.30695051        5.13778960
          Total         30           851.38154484

          Parameter       Standard           Type II
                                The SAS System

          Variable        Estimate            Error    Sum of Squares          F   Prob>F

           INTERCEP    104.86493455      12.12725479      384.15977836      74.77   0.0001
           AGE          -0.24073621       0.09459302       33.27675879       6.48   0.0178
           WEIGHT       -0.07455833       0.05327812       10.06168928       1.96   0.1745
           RUNTIME      -2.62442388       0.37249127      255.04222229      49.64   0.0001
           RUNPULSE     -0.35991816       0.11756561       48.15298878       9.37   0.0054
           MAXPULSE      0.28765957       0.13437098       23.54631568       4.58   0.0427
           RSTPULSE     -0.02531626       0.06467048        0.78734155       0.15   0.6989

Bounds on condition number:      8.853911,      137.8171
------------------------------------------------------------------------------------------

The above model is the best  6-variable model found.

No further improvement in R-square is possible.
```

例三：利用身高与年龄测试学生的体重

本资料档（HTWT）的数据由 Lowis 及 Taylor（1967）提供。每一学生提供四个数据：性别，年龄——以月计（AGE），身高——以英寸计（HEIGHT），及体重——以英镑计（WEIGHT）。其中，年龄和身高是自变量，体重是因变量。

本范例最主要是在示范如何利用 BY 指令将资料档按性别分成两个小资料档（男和女），然后对每一个小资料档分别进行递归分析。

程　序

```
DATA HTWT;
    INPUT SEX $ AGE :3.1 HEIGHT WEIGHT @@;
    CARDS;
F 143 56.3  85.0 F 155 62.3 105.0 F 153 63.3 108.0 F 161 59.0  92.0
F 191 62.5 112.5 F 171 62.5 112.0 F 185 59.0 104.0 F 142 56.5  69.0
M 164 66.5 112.0 M 189 65.0 114.0 M 164 61.5 140.0 M 167 62.0 107.5
M 151 59.3  87.0
;
TITLE '--------DATA ON AGE,WEIGHT, AND HEIGHT OF CHILDREN---------';
PROC REG OUTEST=EST1 OUTSSCP=SSCP1;
    BY SEX ;
    EQ1:MODEL WEIGHT=HEIGHT;
    EQ2:MODEL WEIGHT=HEIGHT AGE;
PROC PRINT DATA=SSCP1;
    TITLE2 'SSCP TYPE DATA SET';
PROC PRINT DATA=EST1;
TITLE2 'EST TYPE DATA SET';
RUN;
```

结　果

对青春期前后的女生而言，体重与身高呈几乎线性的关系（R^2=0.7867, F= 22.127, P=0.0033）。然而对同年龄的男生而言，这个线性的关系并不存在（R^2=0.0736, F=0.238, P=0.6589）。

含年龄（AGE）的复递归模型并不比单递归模型更好。所以，仍是女生组的 EQ1 最有效；EQ1 或 EQ2 对男生而言均无效（R^2=0.1249，F=0.143，P=0.8751）。

报表　18.3　利用身高与年龄预测学生的体重

--------DATA ON AGE,WEIGHT, AND HEIGHT OF CHILDREN---------

---------- SEX=F ----------

Model: EQ1

Dependent Variable: WEIGHT

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	1286.78827	1286.78827	22.127	0.0033

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-189.054900	61.17709168	-3.090	0.021

```
Error     6   348.93048     58.15508                        HEIGHT      1      4.777605    1.01566515         4.704       0.003
C Total   7  1635.71875

     Root MSE    7.62595      R-square  0.7867
     Dep Mean   98.43750      Adj R-sq  0.7511
     C.V.        7.74699

                    --------DATA ON AGE,WEIGHT, AND HEIGHT OF CHILDREN---------
                ---------------------------------- SEX=F ----------------------------------
Model: EQ2
Dependent Variable: WEIGHT

               Analysis of Variance                                    Parameter Estimates

                 Sum of        Mean                                  Parameter     Standard   T for H0:
Source   DF     Squares      Square  F Value  Prob>F       Variable  DF     Estimate        Error  Parameter=0  Prob > |T

Model     2  1469.03635   734.51817   22.033  0.0033       INTERCEP   1  -184.270863   46.36371103       -3.974       0.010
Error     5   166.68240    33.33648                        HEIGHT     1     3.849268    0.86543356        4.448       0.006
C Total   7  1635.71875                                    AGE        1     3.140887    1.34332371        2.338       0.066

     Root MSE    5.77378      R-square  0.8981
     Dep Mean   98.43750      Adj R-sq  0.8573
     C.V.        5.86542

                    --------DATA ON AGE,WEIGHT, AND HEIGHT OF CHILDREN---------
                ---------------------------------- SEX=M ----------------------------------
Model: EQ1
Dependent Variable: WEIGHT

               Analysis of Variance                                    Parameter Estimates

                 Sum of        Mean                                  Parameter     Standard   T for H0:
Source   DF     Squares      Square  F Value  Prob>F       Variable  DF   Estimate        Error  Parameter=0  Prob > |T|

Model     1   105.44134   105.44134     0.238  0.6589       INTERCEP   1  -0.106582   230.07806753       -0.000       0.9997
Error     3  1327.75866   442.58622                        HEIGHT     1   1.785024    3.65710505        0.488       0.6589
C Total   4  1433.20000

     Root MSE   21.03773      R-square  0.0736
```

Dep Mean 112.10000 Adj R-sq -0.2352

C.V. 18.76693

--------DATA ON AGE,WEIGHT, AND HEIGHT OF CHILDREN---------

---------- SEX=M ----------

Model: EQ2

Dependent Variable: WEIGHT

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	179.05153	89.52577	0.143	0.8751
Error	2	1254.14847	627.07423		
C Total	4	1433.20000			

Root MSE 25.04145 R-square 0.1249

Dep Mean 112.10000 Adj R-sq -0.7501

C.V. 22.33849

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	6.854588	274.61714302	0.025	0.9824
HEIGHT	1	0.625680	5.51356748	0.113	0.9200
AGE	1	3.947018	11.52018888	0.343	0.7645

--------DATA ON AGE,WEIGHT, AND HEIGHT OF CHILDREN---------

SSCP TYPE DATA SET

OBS	SEX	_TYPE_	_NAME_	INTERCEP	HEIGHT	WEIGHT	AGE
1	F	SSCP	INTERCEP	8.0	481.40	787.50	130.10
2	F	SSCP	HEIGHT	481.4	29024.62	47657.15	7845.43
3	F	SSCP	WEIGHT	787.5	47657.15	79155.25	12944.35
4	F	SSCP	AGE	130.1	7845.43	12944.35	2139.15
5	F	N		8.0	8.00	8.00	8.00
6	M	SSCP	INTERCEP	5.0	314.30	560.50	83.50
7	M	SSCP	HEIGHT	314.3	19789.99	35292.10	5258.53
8	M	SSCP	WEIGHT	560.5	35292.10	64265.25	9396.35
9	M	SSCP	AGE	83.5	5258.53	9396.35	1402.03
10	M	N		5.0	5.00	5.00	5.00

EST TYPE DATA SET

OBS	SEX	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	INTERCEP	HEIGHT	WEIGHT	AGE
1	F	EQ1	PARMS	WEIGHT	7.6259	-189.055	4.77761	-1	.

```
2   F    EQ2   PARMS   WEIGHT   5.7738 -184.271 3.84927   -1    3.14089
3   M    EQ1   PARMS   WEIGHT  21.0377   -0.107 1.78502   -1     .
4   M    EQ2   PARMS   WEIGHT  25.0415    6.855 0.62568   -1    3.94702
```

18.4　注 意 事 项

遗漏数据的处理

观察体只要在任何一个自变量上含遗漏数据，则 REG 程序自动将此观察体剔除于递归分析过程之外。

六种不同形态的输入资料档

（1）观察体*变量之矩阵
（2）TYPE=CORR
（3）TYPE=COV
（4）TYPE=SSCP
（5）TYPE=UCORR
（6）TYPE=UCOV

大多数 REG 程序的输入资料档是一个长方型的原始数据矩阵。其中横列代表观察体，直行代表变量。其实相关系数的矩阵（即 TYPE=CORR），变异数/共变异数矩阵（即 TYPE=COV）、平方和以及内乘积矩阵（即 TYPE=SSCP），未经平均数纠正过的相关系数矩阵（即 TYPE=UCORR）或未经平均数纠正过的变异数/共变异数矩阵（即 TYPE=UCOV）同样也可以成为 REG 程序的输入资料档。其中 TYPE=CORR、UCORR 或 COV、COV 的资料档是由 CORR 程序产生的，它包括相关系数、平均数、标准差等统计值。而 TYPE=SSCP 的资料档则由另一个 REG 程序产生，它包括变异数与变量间的内乘积。

下面的例子是将本章第二个范例在第一次递归分析后所产生的输出资料档（TYPE=SSCP）以第二次递归分析以及较精简的模型（只含 RUNTIME、AGE、WEIGHT 等自变量）在加以处理。

程　序

```
PROC CORR DATA=FITNESS OUTP=R;
     VAR OXY RUNTIME AGE WEIGHT RUNPULSE MAXPULSE RSTPULSE;
PROC PRINT DATA=R;
PROC REG DATA=R;
     MODEL OXY=RUNTIME AGE WEIGHT;
RUN;
```

结　果

利用 TYPE=SSCP 分析的结果十分理想（R^2=0.7774, F=31.432, P=0.0001）。RUNTIME

与 AGE 这两个自变量，经 t 检定测试均达 0.10 的显著度，WEIGHT 则未达此标准。所以，下一次的递归分析里，似可再简化模型。

另外，读者应已注意到，利用 TYPE=SSCP 数据所分析的结果，是无法产生任何 Y 的个别预测值的。

采用特殊资料档的好处

采用前述第（2）~（6）钟矩阵位输入资料档有一个好处：会节省电脑作业时间（即 CPU 时间）。所节省的时间可高达 99%；特别当观察体的数目在千个以上，而变量的数目超过百个以上时，CPU 时间的缩短就更加重要。

采用特殊资料档的限制

TYPE=CORR 或 TYPE=SSCP 的输入资料档也有其限制：

- 一定要在 PROC REG 指令中，用小括号注明资料档的形态，如：PROC REG DATA=A（TYPE=CORR）;
- OUTPUT 指令，MODEL 指令，以及 PRINT 指令中的 P、R、CLM、CLI、DW、INFLUENCE、ACOV、SPEC、及 PARTIAL 选项都会失效，不能选用。
- 牵涉到原始数据的指令如：FREQ、ID、PAINT、PLOT、REWEIGHT、及 WEIGHT 等无效。

参数的估计以及有关的统计值

用本章例二的数据，下面的指令可列出参数估计过程中所能获得的全部统计值：

```
PROC REG DATA=FITNESS;
     MODEL OXY=RUNTIME AGE WEIGHT
               RUNPULSE MAXPULSE RSTPULSE
                /SS1 SS2 STB TOL VIF COVB CORRB;
```

当递归模型不是一个满秩的模型是，参数的最小误差平方估计值将有无限多个解。（一般而言，每一个参数的最小误差平方估计值应该只有一个。）

在这种情况下，SAS 采用通用倒数（Generalized Inverse）来解决这个问题。通用倒数的解等于一般最小误差平方解的前面乘以（X' X）$^{-1}$（X' X）。其定义如下：

b=（X' X）$^{-1}$（X' X）（X' X）$^{-1}$（X' Y）

其中，（X' X）$^{-1}$（X' Y）是最小误差平方解。

诊断自变量之间的相关性

当某一自变量与其它自变量之间有高度线性相关时，参数的估计值将会不稳定，而且会含偏高的标准误。这个现象称为共线性（Collinearity）或多变量共线性（Multicollinearity）。

针对这一个问题，我们可采用 COLLIN 选项来诊断到底哪些自变量之间有共线性。这个诊断的理论基础来自 Belsley，Kuh 及 Welsch 于 1980 年所发表的论文。

诊断的步骤如下：

第一：将（X' X）矩阵标准化，使其对角线上的值都成为 1。若读者选用 COLLINOINT 选项，则 Y 结局将由矩阵中删除。

第二：计算出（X' X）矩阵的特性根与特性向量。

第三：以最大的特性根为分子，其它特性根分别为分母，形成几个不同的比例。这些比例的平方根便是共线性指标。若指标的值较大时，则表示变量之间的共线性情形可能极为严重。在这种情况下，参数的估计值较不准确。

下面是一个示范的例子；数据来自本章例二的资料档：

程　序

```
PROC REG DATA=FITNESS;
    MODEL OXY=RUNTIME AGE WEIGHT
              RUNPULSE MAXPULSE RSTPULSE
             /TOL VIF COLLIN;
RUN;
```

结　果

共线性比较严重的变量是 RUNPULSE 与 MAXPULSE。此外，RUNTIME 与 RSTPULSE 以及 WEIGHT 与 RSTPULSE 之间也有中等程度的共线性。

报表 18.5 利用　　PROC　REG 诊断自变项之间的相关性

Model: MODEL1

Dependent Variable: OXY

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	6	728.07459	121.34577	23.618	0.0001
Error	24	123.30695	5.13779		
C Total	30	851.38154			

Root MSE	2.26667	R-square	0.8552
Dep Mean	47.37581	Adj R-sq	0.8190
C.V.	4.78444		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	104.864935	12.12725479	8.647	.0001
RUNTIME	1	-2.624424	0.37249127	-7.046	0.0001
AGE	1	-0.240736	0.09459302	- 2.545	0.0178
WEIGHT	1	-0.074558	0.05327812	-1.399	0.1745
RUNPULSE	1	-0.359918	0.11756561	-3.061	0.0054
MAXPULSE	1	0.287660	0.13437098	2.141	0.0427
RSTPULSE	1	-0.025316	0.06467048	-0.391	0.6989

Tolerance and Variance Inflation

Variable	DF	Tolerance	Variance Inflation
INTERCEP	1	.	0.00000000
RUNTIME	1	0.64122533	1.55951420
AGE	1	0.66393116	1.50618026
WEIGHT	1	0.86979463	1.14969668
RUNPULSE	1	0.11789039	8.48245529
MAXPULSE	1	0.11294444	8.85391072
RSTPULSE	1	0.70533662	1.41776278

Collinearity Diagnostics

Number	Eigenvalue	Condition Number	Var Prop INTERCEP	Var Prop RUNTIME	Var Prop AGE	Var Prop WEIGHT	Var Prop RUNPULSE	Var Prop MAXPULSE	Var Prop RSTPULSE
1	6.94916	1.00000	0.0000	0.0002	0.0002	0.0002	0.0000	0.0000	0.0003
2	0.01922	19.01594	0.0019	0.0219	0.1750	0.0052	0.0000	0.0000	0.3516
3	0.01511	21.44841	0.0008	0.1318	0.1372	0.2425	0.0012	0.0013	0.0498
4	0.00916	27.54875	0.0059	0.6315	0.0302	0.1685	0.0014	0.0012	0.2075
5	0.00614	33.63435	0.0018	0.1145	0.1058	0.4627	0.0147	0.0082	0.3647
6	0.00104	81.80750	0.7853	0.0858	0.4776	0.0987	0.0703	0.0053	0.0195
7	0.0001773	197.95206	0.2043	0.0143	0.0742	0.0222	0.9125	0.9841	0.0066

预测值与预测误差

控制预测值与预测误差的选项包括 P、R、CLM、CLI 等（详情请见前面的介绍）。下例的程序以本章例一的数据为基础，用 P、R、CLM、CLI、OUTPUT、PLOT 等指令控制报表上有关预测值及其误差的计算与绘图：

程　序

```
PROC REG DATA=USPOP;
     VAR YEARSQ;
     MODEL POP=YEAR/P CLI CLM INFLUENCE DW;
     PLOT R.*P.;
     ADD YEARSQ;
     PRINT;
     PLOT;
     RUN;
     PLOT POP*YEAR='A'  P.*YEAR='P'
     U95.*YEAR='U' L95.*YEAR='L'/OVERLAY;
RUN;
```

结　果

报表的输出与李一完全相同，故不再列出。

观察体对递归分析的影响力

INFLUENCE 选项可用来诊断观察体对递归分析的影响力。在解释影响力的诊断之前，首先让我们对一些名词做如下的定义：

b（i）=删除观察体 i 后所得的递归系数估计值。

s（i）=删除观察体 i 后所得的样本标准差。

X（i）=删除观察体 i 后所得的自变量矩阵。

Y（i）=删除观察体 i 后所得的该个体之 Y 的预测值。

$e_i = Y_i - \hat{Y}_i$　　观察体 i 的预测误差

$h(i) = h_i = X_i (X'X)^{-1} X_i'$　　又称 HaT　Matrix

$RSTUDENT = e_i/[s(i)*\sqrt{1-h(i)}]$，标准化误差。

$COVRATIO = \det\{S^2(i)[X(i)'X(i)]^{-1}\}$，删除观察体 i 后所得的参数的共变量数矩阵的行列式（既 det）。

因此，影响力的值=DIFFITS，其定义如下：

$$[Y_i - Y(i)]/[s(i)*\sqrt{h(i)}]$$

其中，$\hat{Y}_i$ 表示含观察体 i 所得的 Y 预测值，而 Y（i）表示删除 i 观察体后所得的该个体预测值。

下面的程序简单扼要的示范选项 INFLUENCE 的撰写，所使用的数据仍与例一相同：

程　序

```
PROC REG DATA=FITNESSS;
    MODEL OXY=RUNTIME WEIGHT AGE/INFLUENCE;
RUN;
```

结　果

这个程序的结果与例一完全一致，故不再重复。

与时间有关的数据处理法

当递归分析所处理的数据与时间有关时，其预测误差往往是前后相关联的。与检验这种相关程度的大小，我们可选用 DW 选项以计算杜本-华生氏（Durbin-Watson）的统计值。其定义如下：

当误差之间完全没有线性相关时，DW 值应该十分靠近 2。当 DW 值靠近 0（下限）或 4（上限）时，误差之间有负或正的线性相关。利用杜本-华生氏所发表的抽样分配，我们还可以进一步检验此 DW 值的统计显著程度。下面是一个有关 DW 值的例子，数据的来源是本章例一的范例：人口成长的趋势。

程　序

```
PROC REG DATA=USPOP;
    MODEL POP=YEAR YEARSQ/DW;
RUN;
```

结　果

执行下面程序后，报表中有关 DW 值的检验如下所示：

```
Durbin-Watson D                 1.264
(For Number of  Obs.)           19
1st  Order  Autocorrelation     0.299
```

由于 DW=1.264 不十分接近虚无假设下的期待值（2），故我们可下结论说：误差之间的相关似乎存在。

多变项的统计检验

当递归分析中含一个以上的因变量时，读者可利用 MTEST 指令来进行多变量的统计检验。SAS 提供四个统计检验的值，它们分别是：Wilk's Lambda，Pillai's Trace，

Hetelling-Lawley's Trace，及 Roy's 最大特性根。这四个值都是由 E 与 H 矩阵间接导出的（参见第 17.4 节的定义）。E 矩阵是误差（或分母）的矩阵；而 H 矩阵是与假设有关的分子矩阵。请读者参考下页的例子：

程　序

```
DATA A;
  INPUT SEX $ DRUG $ @;
  DO REP=1 TO 4;
  INPUT Y1 Y2 @;
  OUTPUT;
  END;
  CARDS;
M A  5  6  5  4  9  9  7  6
M B  7  6  7  7  9 12  6  8
M C 21 15 14 11 17 12 12 10
F A  7 10  6  6  9  7  8 10
F B 10 13  8  7  7  6  6  9
  C 16 12 14  9 14  8 10  5
;
DATA B;
    SET A;
    SEXCODE=(SEX='M')-(SEX='F');
    DRUG1=(DRUG='A')-(DRUG='C');
    DRUG2=(DRUG='B')-(DRUG='C');
    SEXDRUG1=SEXCODE*DRUG1;
    SEXDRUG2=SEXCODE*DRUG2;
PROC REG;
    MODEL Y1 Y2=SEXCODE DRUG1 DRUG2 SEXDRUG1 SEXDRUG2;
    SEX:     MTEST SEXCODE;
    DRUG:    MTEST DRUG1,DRUG2;
    SEXDRUG:  MTEST SEXDRUG1,SEXDRUG2;
    1MY2:    MTEST Y1-Y2;
    Y1Y2DRUG: MTEST Y1=Y2,DRUG1,DRUG2;
    DRUGSHOW: MTEST DRUG1,DRUG2 /PRINT CANPRINT;
 RUN;
```

结　果

报表　18.6 PROC REG 对多变项的统计分析

Model: MODEL1

Dependent Variable: Y1

Analysis of Variance			Parameter Estimates		
Sum of	Mean		Parameter	Standard	T for H0:

Source	DF	Squares	Square	F Value	Prob>F
Model	5	316.00000	63.20000	12.038	0.0001
Error	18	94.50000	5.25000		
C Total	23	410.50000			

Root MSE	2.29129	R-square	0.7698
Dep Mean	9.75000	Adj R-sq	0.7058
C.V.	23.50039		

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	9.750000	0.46770717	20.846	0.0001
SEXCODE	1	0.166667	0.46770717	0.356	0.7257
DRUG1	1	-2.750000	0.66143783	-4.158	0.0006
DRUG2	1	-2.250000	0.66143783	-3.402	0.0032
SEXDRUG1	1	-0.666667	0.66143783	-1.008	0.3269
SEXDRUG2	1	-0.416667	0.66143783	-0.630	0.5366

Dependent Variable: Y2

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	5	69.33333	13.86667	2.189	0.1008
Error	18	114.00000	6.33333		
C Total	23	183.33333			

Root MSE	2.51661	R-square	0.3782
Dep Mean	8.66667	Adj R-sq	0.2055
C.V.	29.03782		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	8.666667	0.51370117	16.871	0.0001
SEXCODE	1	0.166667	0.51370117	0.324	0.7493
DRUG1	1	-1.416667	0.72648316	-1.950	0.0669
DRUG2	1	-0.166667	0.72648316	-0.229	0.8211
SEXDRUG1	1	-1.166667	0.72648316	-1.606	0.1257
SEXDRUG2	1	-0.416667	0.72648316	-0.574	0.5734

Multivariate Test: SEX

Multivariate Statistics and Exact F Statistics

S=1　M=0　N=7.5

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.99253694	0.0639	2	17	0.9383
Pillai's Trace	0.00746306	0.0639	2	17	0.9383
Hotelling-Lawley Trace	0.00751918	0.0639	2	17	0.9383
Roy's Greatest Root	0.00751918	0.0639	2	17	0.9383

Multivariate Test: DRUG

Multivariate Statistics and F Approximations

S=2　M=-0.5　N=7.5

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.16862952	12.1991	4	34	0.0001
Pillai's Trace	0.88037810	7.0769	4	36	0.0003
Hotelling-Lawley Trace	4.63953666	18.5581	4	32	0.0001
Roy's Greatest Root	4.57602675	41.1842	2	18	0.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

NOTE: F Statistic for Wilks' Lambda is exact.

Multivariate Test: SEXDRUG

Multivariate Statistics and F Approximations

S=2　M=-0.5　N=7.5

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.77436234	1.1593	4	34	0.3459
Pillai's Trace	0.22694905	1.1520	4	36	0.3481
Hotelling-Lawley Trace	0.28969161	1.1588	4	32	0.3473
Roy's Greatest Root	0.28372273	2.5535	2	18	0.1056

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

NOTE: F Statistic for Wilks' Lambda is exact.

Multivariate Test: Y1MY2

Multivariate Statistics and Exact F Statistics

S=1　M=1.5　N=8

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.27497936	9.4919	5	18	0.0001
Pillai's Trace	0.72502064	9.4919	5	18	0.0001
Hotelling-Lawley Trace	2.63663664	9.4919	5	18	0.0001
Roy's Greatest Root	2.63663664	9.4919	5	18	0.0001

Multivariate Test: Y1Y2DRUG

Multivariate Statistics and Exact F Statistics

S=1　M=0　N=8

Statistic	Value	F	Num DF	Den DF	Pr > F

Wilks' Lambda	0.28053917	23.0811	2	18	0.0001
Pillai's Trace	0.71946083	23.0811	2	18	0.0001
Hotelling-Lawley Trace	2.56456456	23.0811	2	18	0.0001
Roy's Greatest Root	2.56456456	23.0811	2	18	0.0001

Multivariate Test: DRUGSHOW

E, the Error Matrix　　H, the Hypothesis Matrix

94.5	76.5	301	97.5
76.5	114	97.5	36.333333333

	Canonical Correlation	Adjusted Canonical Correlation	Approx Standard Error	Squared Canonical Correlation
1	0.905903	0.899927	0.040101	0.820661
2	0.244371	.	0.210254	0.059717

Eigenvalues of INV(E)*H
= CanRsq/(1-CanRsq)

	Eigenvalue	Difference	Proportion	Cumulative
1	4.5760	4.5125	0.9863	0.9863
2	0.0635	.	0.0137	1.0000

Test of H0: The canonical correlations in the current row and all that follow are zero

	Likelihood Ratio	Approx F	Num DF	Den DF	Pr > F
1	0.16862952	12.1991	4	34	0.0001
2	0.94028273	1.1432	1	18	0.2991

Multivariate Statistics and F Approximations

S=2　M=-0.5　N=7.5

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.16862952	12.1991	4	34	0.0001

Pillai's Trace	0.88037810	7.0769	4	36	0.0003
Hotelling-Lawley Trace	4.63953666	18.5581	4	32	0.0001
Roy's Greatest Root	4.57602675	41.1842	2	18	0.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

NOTE: F Statistic for Wilks' Lambda is exact.

适用于交谈式环境的指令

随着新版 SAS 的改进，REG 程序变得更富弹性：它如今可容许读者在交谈式的环境下修改递归的模型甚至原始的数据。根据第 18.2 节的介绍，我们可知适用于交谈式的指令有：ADD、DELETE、RESTRICT、TEST、MTEST、OUTPUT、REWEIGHT、REFIT、PAINT、PLOT、以及 PRINT 等。不过值得读者注意的是，若 REG 程序中含 BY 的指令，则上述是一个指令都会无效。

下面举一例来说明如何灵活地运用这些交谈式指令。

第一部分分析的程序

这个数据文件（CLASS）含学生的姓名（NAME）、身高（HEIGHT）、体重（WEIGHT）、以及年龄（AGE）。我们想利用身高与年龄来预测学生的体重。

```
DATA CLASS;
INPUT NAME $ HEIGHT WEIGHT AGE;
CARDS;
Alfred  69.0 112.5 14
Alice   56.5  84.0 13
Barbara 65.3  98.0 13
Carol   62.8 102.5 14
Henry   63.5 102.5 14
James   57.3  83.0 12
Jane    59.8  84.5 12
Janet   62.5 112.5 15
Jeffrey 62.5  84.0 13
John    59.0  99.5 12
Joyce   51.3  50.5 11
Judy    64.3  90.0 14
Louise  56.3  77.0 12
Mary    66.5 112.0 15
Philip  72.0 150.0 16
Robert  64.8 128.0 12
Ronald  67.0 133.0 15
Thomas  57.5  85.0 11
William 66.5 112.0 15
;
/*第一步分析*/
```

```
PROC REG;
MODEL WEIGHT=AGE HEIGHT;
ID NAME;
/*第二步分析*/
DELETE AGE;
PRINT;
RUN;
/* 第三步分析 */
PLOT R.*P.;
RUN;
/* 第四步分析*/
REWEIGHT R.>15 OR R.<-15;
PLOT;
RUN;
```

第一步分析的结果

以身高及年龄来预估学生的体重似乎十分理想，F 值高达 27.228 (P<0.0001)而且纠正过后的复相关平方也近乎 75%（0.7445）。不过，我们想进一步了解是否模型可简化到只含一个自变量的模型。因此，第二步的分析就把年龄去掉了。

```
Model: MODEL1
Dependent Variable: WEIGHT

                        Analysis of Variance

                           Sum of          Mean
Source          DF        Squares        Square       F Value    Prob>F

Model            2     7215.63710     3607.81855        27.228    0.0001
Error           16     2120.09974      132.50623
C Total         18     9335.73684

    Root MSE          11.51114    R-square      0.7729
    Dep Mean         100.02632    Adj R-sq      0.7445
    C.V.              11.50811

                     Parameter Estimates

                   Parameter     Standard      T for H0:
Variable    DF      Estimate        Error     Parameter=0    Prob > |T|

INTERCEP     1   -141.223763     33.38309350      -4.230         0.0006
AGE          1      1.278393      3.11010374       0.411         0.6865
HEIGHT       1      3.597027      0.90546072       3.973         0.0011
```

第二步分析的程序

下面的指令是紧接着上述的结果之后在交谈式的环境下执行

```
DELETE AGE;
PRINT;
RUN;
```

第二步分析的结果

若将这个结果与第一步分析所得的做比较，则你不难发现此次的结果更理想：因为 F 值更高（F=57.076，P<0.0001），而且纠正过后的复相关平方更优（=0.7570）。所以我们决定采用这个简化的模型（亦即，体重的预测值=-143.026918+(3.899)*身高）。

接下来我们进行误差值的检验（见第三步的分析）

```
Model: MODEL1
Dependent Variable: WEIGHT

                        Analysis of Variance

                                Sum of           Mean
Source                  DF      Squares         Square    F Value    Prob>F

Model                    1   7193.24912     7193.24912     57.076    0.0001
Error                   17   2142.48772      126.02869
C Total                 18   9335.73684

    Root MSE       11.22625    R-square      0.7705
    Dep Mean      100.02632    Adj R-sq      0.7570
    C.V.           11.22330

                     Parameter Estimates

                                Parameter       Standard     T for H0:
Variable            DF          Estimate          Error    Parameter=0    Prob > |T|

INTERCEP             1       -143.026918     32.27459130         -4.432        0.0004
HEIGHT               1          3.899030      0.51609395          7.555        0.0001
```

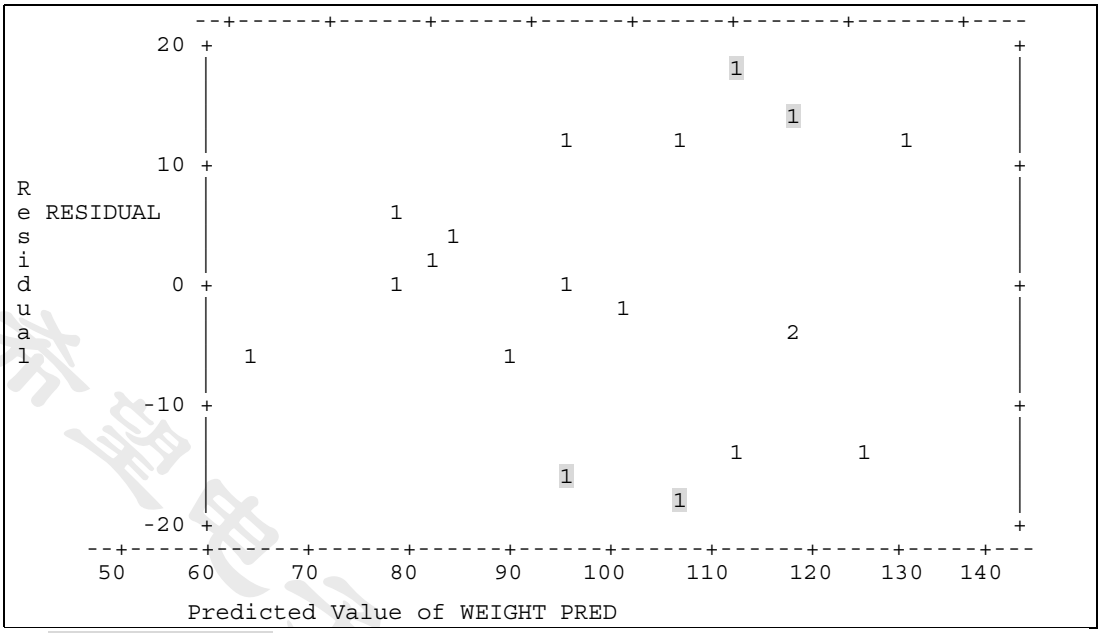
第三步分析的程序

将预测误差对体重预测值作图：

```
PLOT R.*P.;
RUN;
```

第三步分析的结果

报表中有少数点似乎代表了劣质数据，因此我们将他们剔除并重新执行递归的分析。



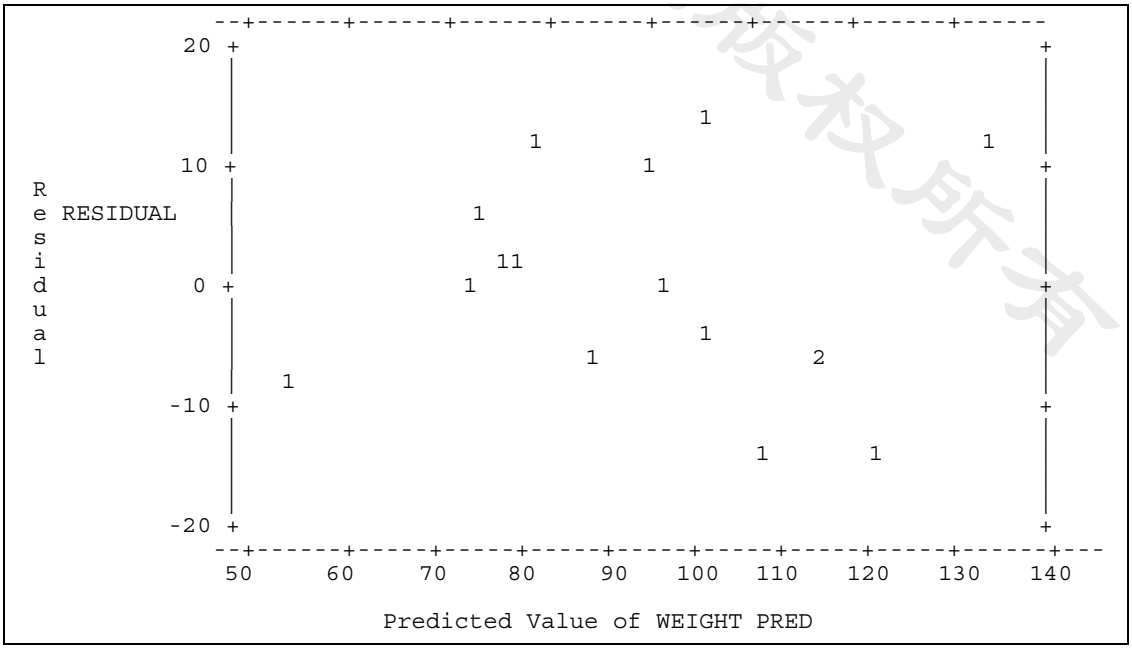
第四步分析的程序

剔除劣质数据的方法是将它们的加权值重新设定为零。然后再执行一次递归分析。劣质数据的标准是预测误差（R.）之绝对值大于 15 者。

```
REWEIGHT R.>15 OR R.<-15;
PLOT;
RUN;
```

第四步分析的结果

此时点状图上所有的误差都介于（-15，15）之间，因此分析至此，可算大功告成：



第 19 章　二分数据的预估：统计程序 PROC PROBIT

19.1　PROC PROBIT 程序概述

本程序主要是利用最大可能率估计法找出一个回归模型的参数估计值，或生物实验数据以及类别数据中的底线率 (Threshold Response Rate)。在估计这些参数值的过程中，PROC PROBIT 容许读者选择各式不同的模型如：概率单位 (Probit)、对数奇数比 (Logit)、次序逻辑斯谛 (Ordinal Logistic)，以及成长曲线 (Gompit) 等模型。概率单位模型也就是下一章所介绍的常态单位模型；其理论基础是累积常态分配，有兴趣的读者可以同时参考第 20.1 节与 20.2 节的内容。

上述数学模型以及其理论的导出源自于最小平方误差 (LS) 的方法不适合用来估计以二分数据 (或类别数据) 为因变量的模型参数。由此发展出以最大可能率法来估计参数的一系列回归模型。以下是这些模型的公式 (适用于二分数据的分析)：

p=Pr(Y=0)=C+(1-C)F(X'b)

在此，p= 某种反应结果 (亦即 Y=0 的结果) 之概率

C= 底线率 (或称截距)

F= 累积分配函数

X= 自变量 (串) 的值

b= 参数 (串) 的估计值。

上述公式中的累积分配函数 (F) 决定模型的特殊形式如：概率单位、对数奇数比或成长曲线等。概率单位 (又称常态数单位) 的理论基础是累积常态分配；对数奇数比的理论基础是累积的逻辑斯谛分配。这两种模型所导出的参数估计值或概率值十分接近，这是因为这两种模型的概率分配的形状类似，都呈钟形，而且以零为中央点左右对称。

然而成长曲线的模型是以刚氏分配 (Gompertz) 为理论基础。刚氏分配是一个左右不对称的分布，因此只适用于符合这种形态的二分数据。

自变量 (串) X 可以是连续变量或定值变量 (如变异数分析中的自变量)。唯一值得注意的是这些自变量必须是彼此线性独立的。若它们之间存有绝对的或近似的线性相依关系，则参数的估计会不稳定。

公式中的参数 C 代表底线率，其值可由读者自设为一个常数或由 PROBIT 估计。此参数的内设值是 0。

如前所述，参数的估计法采用最大可能率法，其执行程序是依据修正过的高斯牛顿法 (ModifiedGauss-Newton Method)。函数值与数据间的适合度 (Goodness-of-Fit) 以两种卡平方 (χ^2) 值来表示：(甲) 皮尔森 χ^2 检定，以及 (乙) 对数可能比卡方检定。此外，读者也可要求 PROBIT 计算各自变量之 95% 的信赖区间。有关这方面的指令语法，请参见下一节的内容。

19.2　如何撰写 PROC PROBIT 程序

PROC PROBIT 含六道指令，它们的格式如下：

```
PROC PROBIT       选项串;
     CLASS        变量名称串;
     MODEL        反应变量=自变量串 / 选项串;
     OUTPUT       OUT= 输出文件名称  关键字串;
     WEIGHT       变量名称;
     BY           变量名称串;
```

指令 #1　PROC PROBIT 选项串;

这道指令有十二个选项，分属四类：第一类选项与输入 / 输出的文件有关；第二类选项与模型的界定有关，第三类选项与适合度检定有关，第四类选项与报表的打印有关。

第一类选项　与输入 / 输出文件有关，含三道选项：

(1) DATA= 输入文件名称

指明到底对那一个文件进行分析。若省略此选项，则 SAS 会自动找出在此程序之前最后形成的 SAS 文件，对它执行分析。

(2) OUTEST= **输出文件名称**

这个文件含模型的参数估计值，以及它们之间的共变异数。有关这个输出文件的进一步说明，请参阅第 19.4 节的内容。

(3) COVOUT

要求将参数估计值之间的共变异数纳入上述 OUTEST= 输出文件内。

第二类选项　与模型的界定有关，有下列几个选项：

(4) ORDER=DATA

ORDER=FORMATTED (内设值)

ORDER=FREQ

ORDER=INTERNAL

这个选项界定分类变量 [亦即 CLASS 的变量 (串)] 下各组别的先后次序。当 ORDER=DATA 时，组别的定义 (亦即第一组，第二组，...等) 由输入文件内的数据来决定。当 ORDER=FORMATTED 时，组别的次序由外在格式决定。当 ORDER=FREQ 时，组别的次序以各组内的成员数 (或频率) 的多少来决定，人数最多的那一组被视为第一组，人数次多的是第二组，以此类推。若 ORDER=INTERNAL，则组别的次序以其数字代号的小 (第一组) 大 (最后一组) 或文字代号的第一个字母的前后顺序来决定。

(5) OPTC　　或

(6) C=正小数

这两个选项均与底线率的设定 (或估计) 有关。若只选用 OPTC，则 PROBIT 程序会自动估计参数 C 的值。若只选用 C= 正小数，则读者必须自行界定此值，而且这个值必须是合理的。若读者同时省略这两个选项，则 C=0，也就是说参数 C 不存在 。若同时界定这两个选项，则 C= 正小数的值成为参数 C 的初步估计值。

(7) LOG (或 LN)

要求 PROBIT 程序将自变量串 (X) 中第一个连续变量作自然对数的转换，然后再进行分析。在一般生物学的实验数据中，这个连续变量多半是治疗某种疾病的用药量 (Dosage)。报表上会打印出这个连续变量 (如用药量) 的估计值以及它的 95% 参考区间 (Fiducial Limit)。若读者同时选用上述的 OPTC 选项，则参数 C 的估计值来自控制组内产生反应结果 (亦即 Y=0) 的百分比。而控制组成员的定义就是凡在连续变量 (如 Dosage) 上的值小于或等于零的受试者。

(8) LOGIO

与选项 LOG(LN) 的用法一致，只是转换的函数是以 10 为底的对数函数。

(9) INVERSECL

要求 PROBIT 程序计算第一个连续自变量 (如用药量的多少) 的信赖区间。若在计算的过程中，程序无法收敛，则 PROBIT 会在报表上以遗漏数据取代信赖区间的上限与下限。

第三类选项　与适合度的检定有关，有下列两个选项：

(10) LACKFIT

这个选项要求 PROBIT 执行两个适合度的检定：第一个是皮尔森的卡平方检定，第二个是对数可能比检定 (也是根据卡平分分配)。值得读者特别注意的是：适合度检定只适用于整理过的数据文件 (亦即数据已经按照各自变量的值作由小至大的排列) ，而且各自变量类别下的数据点不宜过少。若皮尔森的卡平方检定达到显著水准，则 PROBIT 程序会自动调整共变异数与标准误的估计值。

这个选项也可同时出现在 MODEL 的指令中。

(11) HPROB=极小的概率值 (内设值等于 .10)

这是用来界定适合度检定的显著水准，必须与上述 LACKFIT 选项同时界定。若皮尔森检定所导出的实际显著水准大于 HPROB= 的值，则 PROBIT 程序会自动以 1.96 为临界值建立一个 95% 的参考区间。若实际的显著水准低于 HPROB= 的值，则 PROBIT 程序会自 t 分配中导出一对临界值与 95% 的双尾信赖区间相对应。t 分配的自由度等于(k-1)*m-q，在此 k= 反应变量的类别数，m=自变量 (串) 值的不同组合数，q=模型中参数的总数。

读者应当注意，选项 HPROB= 可同时出现在指令 #1 (PROC PROBIT) 或指令 #3 (MODEL)中。在这种情况下，指令 #3 的界定取代指令 #1 的界定。

第四类选项　与报表的打印有关，含一个选项：

(12) NOPRINT

要求 PROBIT 程序抑止一切分析结果的打印。

指令 #2　CLASS 变量名称串；

此指令界定分析中所用到的分类变量 (串) ，必须放在 MODEL 指令之前。即使 MODEL 指令只提到一个反应变量 (也是分类变量)，它的名称也必须出现在 CLASS 指令中。

指令 #3　MODEL 反应变量=自变量串 / 选项串；

这个指令界定一个数学模型，可用来解释反应变量与自变量之间的关系。反应变量通常是一个类别变量，也是 CLASS 指令所提到的分类变量。

若反应变量是一个二分的变量 [如 CURE，下分 YES (治愈)，NO (尚未治愈) 两个结果]，则 MODEL 指令也可用下式来表示，如：

```
MODEL NO_YES/NO_CURE=DOSE AGE;
```

在此，NO_YES 是一个变量名称，其值等于治愈的病人总数，NO_CURE 也是一个变量名称，其值等于总病人数 (含治愈的与未治愈的)。因此，这个 MODEL 指令所影射的模型是指某种疾病治愈的成功率 (以百分比表之) 与用药量的轻重 (DOSE) 以及病人的年龄 (AGE) 有关。

同一个 PROBIT 程序里，读者可界定好几个模型指令。欲辨认这些指令的分析结果，我们可在 MODEL 之前加标签，如：

```
M1：MODEL SUCCESS=EDUC AGE;
M2：MODEL SUCCESS=LUCK;
```

标签名 (如 M1，M2) 的长度不可超过八个字元，而且第一个字元必须是字母。标签后加冒号 (：)，以便与 MODEL 指令的其它语句分开。

删除号 (/) 后的选项有十三个，分成三类。第一类选项与模型的界定有关，第二类选项与模型的适合度有关，第三类选项与报表的打印有关。下面分述各类的选项：

第一类选项　与模型的界定有关，含四个选项：

(1) D=NORMAL (内设值) 或
　D=LOGISTIC　或
　D=GOMPERTZ
　界定一个累积分配函数 (亦即第 19.1 节公式所提的 F 函数) 是用来解释反应变量的类别频率与一连串自变量之间的关系。若 D=NORMAL，则模型的形态是累积常态分配，因此所导出的反应变量值以常态数为单位。若 D=LOGISTIC ，则模型的特殊形式采用累积的逻辑斯谛分配，其反应变量值是对数的奇数比。当 D=GOMPERTZ 时，模型采用刚氏的成长曲线分配。这个选项的内设值是 NORMAL。

(2) INTERCEPT=截距的初值
　这个选项设定截距 (亦即第 19.1 节通式中参数 C) 的初值。其内设值等于零。

(3) INITIAL=各参数的初值

这个选项设定除截距 (又称底线率或 INTERCEPT) 以外各参数的初值。所有初值的界定顺序必须按删除号 (/) 前 MODEL 内自变量列举的顺序。下面举几个示范的写法：

```
INITIAL=3 4 5              (以一个空白隔开相邻两个的值)
INITIAL=3，4，5            (以逗点隔开相邻两个的值)
INITIAL=3 TO 5             (表 3，4，5 三个初值)
INITIAL=3 TO 7 BY 2        (表 3，5，7 三个初值)
INITIAL=1，3 TO 5，17      (表 1，3，4，5，与 17 等五个初值)
```

以上所举的最后一个例子表示一种混合的撰写方式，也是一种适合法的界定方法。值得读者注意的是：针对任何一个类别自变量 (下含 k 个组别)，则你需要界定 (k-1) 个参数的初值。

若读者不使用此选项，则 PROBIT 视所有参数的初值为零。

(4) NOINT

要求 PROBIT 程序将截距 (又称底线率) 的参数自模型中剔除。当反应变量是一个二分变量时，这种界定方式最合适。然而，当反应变量含 k 个组别时，则你需要界定 (k-1) 个截距 (否则，PROBIT 程序视它们均为零)。

第二类选项　与模型的适合度有关，有下列几个选项：

(5) LACKFIT

这个选项要求 PROBIT 程序执行两个适合度的检定：第一个是皮尔森的卡平方检定，第二个是对数可能比检定 (也是根据卡平方分配)。这个选项也可出现在指令 #1 PROCPROBIT 语法中，请参见指令 #1 第 (10) 个选项的解释。若读者重复界定这个选项，则此处 MODEL 指令的界定取代 PROC PROBIT 指令的同一界定。

(6) HPROB=极小的概率值 (内设值等于 .10)

必须与上述 LACKFIT 的选项联用。这个选项也可出现在指令 #1 PROC PROBIT 中，请参见指令 #1 第 (11) 个选项的解释。若读者重复界定这个选项，则此处 MODEL 指令的界定取代 PROC PROBIT 指令的同一界定。

(7) MAXITER=正整数 (如 60)

这个选项界定参数的循环推测法的最高循环次数 (如 60 次)。内设值等于 50 次。

(8) CONVERGE=极小的正实数 (如 0.00001)

界定循环推测过程的收敛指标。当前后两次循环推测所导出的参数估计值的变动 (以绝对值为准) 均小于这个收敛指标时，推测的过程即停止。不过，当某个 (或某些) 参数估计值小于 0.01 时，收敛指标自动改变为相对的变动。这个选项的内设值等于 0.001。

(9) SINGULAR=极小的正实数 (如 10 的负 12 次方)

这个选项的值决定自变量之间线性相依的程度 (PROBIT 程序比较自变量的内乘积矩阵之行列式值与此选项的值，若前者小于后者，则 PROBIT 程序认定线性

相依的关系存在)。内设值等于 10 的负 12 次方。

第三类选项　与报表的打印结果有关：

(10) CORRB

要求 PROBIT 程序打印参数估计值间的相关系数矩阵。

(11) COVB

要求 PROBIT 程序打印参数估计值间的变异数 / 共变异数矩阵。

(12) INVERSECL

要求 PROBIT 程序计算第一个连续变量 (如用药量的多少) 的信赖区间。若在计算的过程中，程序无法收敛，则 PROBIT 会在报表上以遗漏数据取代信赖区间的上限与下限。这个选项也可出现在指令 #1 PROC PROBIT 的语法中，请参见指令 #1 第 (9) 个选项的解释。若读者重复界定这个选项，则此处 MODEL 指令的界定取代 PROC PROBIT 指令的同一界定。

(13) ITPRINT

要求 PROBIT 在报表上打印循环推测的过程以及相关的统计量。

指令 #4　OUTPUT OUT= 输出文件名称　关键字串；

这个指令会产生一个输出文件，内含所有输入文件的数据，以及反应变量等于某个数值的概率估计值 (参见第 19.1 节通式中的 p 值)，估计值的标准误差等。以下分别解释这个指令的两个选项：

(1) OUT=输出文件名称

这个选项界定输出文件的文件名。若你欲将输出文件储存为永久磁盘的文件，则必须界定一个含文件名与文件型的二段式文件名。否则用一段式文件名即可。若省略此选项，则 PROBIT 程序将以内设的命名方式，自动给予一个 DATAn 的文件名 (如：DATA1， DATA2， ...等)，n 值按照输出文件产生的先后顺序，由 1 向上累加。

(2) 关键字串

这个选项要求 PROBIT 将以下三种统计值包括在输出文件内：

关键字	意义
PROB (或 P)	估计出来的累积概率
XBETA	估计出来的概率=a_j+X'β
STD	上述概率的标准误差

这三个关键字可重新命名，见下面的例子：

```
OUTPUT  OUT=EXAMPLE
        P=CUMP
        XBETA=PREDP
        STD=SEPREDP;
```

除了上述提到的关键字之外，另一个内设的关键字 _LEVEL_ 也可能会掺杂在输出

文件内。这是因为当反应变量含三个或三个以上的类别时，PROBIT 程序必须为每一个类别估计其可能发生的概率。不过，因为这些概率值加起来必须等于 1，所以若反应变量下分 k 组，则 PROBIT 程序将估计出 (k-1) 组的参数。所以内设的关键字 _LEVEL_ 就是用来区分这些组别的关键字，其值不外乎 1，2，...，(k-1)。一般而言，第 k 组也就是第一组 (或反应变量值最小的一组)，其参数不会包含在输出文件内。

指令 #5　WEIGHT 变量名称；

这个变量必须是输入文件中的一个数值变量，其值代表每一观察体的加权值。若某一观察体的加权值小于零，等于零，或是一个遗漏数据，则该观察体将不纳入分析中。

指令 #6　BY 变量名称串；

PROBIT 程序依据此指令所列举的变量将文件分成几个小的文件，然后对每一个小的文件分别执行分析。当读者选用此指令时，文件内的数据必须先按照 BY 变量串的值做由小到大的重新排列，这个步骤可藉 PROC SORT 达成。

19.3　范　　例

例一：最新药物 AZT 对爱滋病的影响

本文件 (A) 的数据是用来探讨最新药物 AZT 的用药量 (以 DOSE 表之) 对稳定爱滋病病情 (以 RESPONSE 表之) 的影响。另外，N 代表服药量不等的各组之总人数。

PROBIT 程序的指令界定 LOG10 的选项，此代表将 DOSE 作一个 (以 10 为底) 的对数转换，然后再用 PROBIT 及 LOGIT 两个模型来解释药物的用药量对稳定控制爱滋病之成功率。模型的适合度以及概率值的信赖区间分别由选项 LACKFIT 以及 INVERSECL 来界定。

最后，根据 LOGIT 模型所估计出来的概率值再与实际的估计值作比较，以 PLOT 的图形来表示这两者间吻合的程度。

程　序

```
DATA A;
     INFILE CARDS EOF=EOF;
     INPUT DOSE N RESPONSE;
     PHAT=RESPONSE/N;
     OUTPUT;
     RETURN;
     EOF: DO DOSE=.5 TO 7.5 BY .25;
          OUTPUT;
        END;
     CARDS;
1 10 1
2 12 2
3 10 4
```

```
    4 10 5
    5 12 8
    6 10 8
    7 10 10
    ;
    PROC PROBIT LOG10;
        M1:MODEL RESPONSE/N=DOSE/LACKFIT INVERSECL
        M2:MODEL RESPONSE/N=DOSE/D=LOGISTIC INVERSE
        OUTPUT OUT=B P=PROB STD=STD XBETA=XBETA;
        TITLE 'Output from Probit Procedure';
    PROC PLOT;
        PLOT PHAT*DOSE='X' PROB*DOSE='P'/OVERLAY;
        TITLE 'Plot of Observed and Fitted Probabilities';
    RUN;
```

结　果

PROBIT 程序的分析分两个模型来进行：首先考虑 PROBIT 的模型 (M1) 然后再考虑 LOGIT 的模型 (M2)。估计参数的方法采循环搜索法，经过四次循环之后，参数的估计值即稳定下来。

在第四次循环后，参数 b_0 (见下式的说明)=-1.812704962，参数 b_1=3.418117919。因此，AZT 的用量对稳定爱滋病病情的成功率 (P) 可用下面的函数关系来表示：

P=F(b0+b1*log10(DOSE))，b_0= 报表上的 INTERCEPT，

b_1= 报表上的 LOG 10(DOSE)。

根据 M1 的模型，F 是累积常态分配，其平均数等于 0.53023 (见报表上的 MU)，标准差等于 0.292559 (见报表上的 SIGMA)。数据与 M1 模型吻合的程度由 χ^2 检定 (df=5，p=0.6009) 及对数可能比检定 (df=5，p=0.4616) 来鉴定。这两个检定的结果均支持 M1 的模型。

根据 M2 的模型，F 是累积的逻辑斯谛分配，参数 b_0=-3.2246442，b_1=5.97017999。95% 的信赖区间，在 M1 与 M2 模型分析之后，分别以原数据 (DOSE) 或对数转换的数据 [Log10(DOSE)] 打印在报表上。

最后，PLOT 程序将实际的概率 (X) 与估计的概率 (P) 同时对 DOSE 作图。图形显示，数据与函数的估计值配合地相当密切；不过极端值如 DOSE=1.0 或 7.0 时，X 与 P 的值有显著的出入。

报表 19.1　最新药物 AZT 对爱滋病的影响

Output from Probit Procedure

Probit Procedure

Iter	Ridge	LogLikelihood	INTERCPT	Log10(DOSE)
0	0	-51.29289136144	0	0
1	0	-37.88116628	-1.355817008	2.635206083
2	0	-37.28616865823	-1.764939171	3.3408954936
3	0	-37.28038879265	-1.812147863	3.4172391614
4	0	-37.2803880211	-1.812704962	3.418117919

Data Set　　=WORK.A

Dependent Variable=RESPONSE

Dependent Variable=N

Number of Observations=　7

Number of Events　　=　　38　Number of Trials =　　74

Observations with Missing Values=　29

Log Likelihood for NORMAL -37.28038802

Last Evaluation of the Gradient

INTERCPT	Log10(DOSE)
0.000000343	-2.09809E-8

Last Evaluation of the Hessian

	INTERCPT	Log10(DOSE)
INTERCPT	36.005280	20.152676
Log10(DOSE)	20.152676	13.078826

Goodness-of-Fit Tests

Statistic	Value	DF	Prob>Chi-Sq
Pearson Chi-Square	3.6497	5	0.6009
L.R.　Chi-Square	4.6381	5	0.4616

Response Levels：　2　Number of Covariate Values：　7

NOTE：Since the chi-square is small (p > 0.1000),fiducial limits will be calculated using a t value of　1.96.

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	1	-1.812705	0.449341	16.27431	0.0001	Intercept
Log10(DOS)	1	3.41811792	0.745546	21.01963	0.0001	

Estimated Covariance Matrix

	INTERCPT	Log10(DOSE)
INTERCPT	0.201907	-0.311111
Log10(DOSE)	-0.311111	0.555839

Probit Model in Terms of Tolerance Distribution

MU	SIGMA
0.530323	0.292559

Estimated Covariance Matrix for Tolerance Parameters

	MU	SIGMA
MU	0.002418	-0.000409
SIGMA	-0.000409	0.004072

Probit Analysis on DOSE

Probability	Log10(DOSE)	95 Percent Fiducial Limits		DOSE	95 Percent Fiducial Limits	
		Lower	Upper		Lower	Upper
0.01	-0.15027	-0.69520	0.07710	0.70750	0.20174	1.19428
0.02	-0.07052	-0.55768	0.13475	0.85012	0.27690	1.36381
0.03	-0.01992	-0.47066	0.17157	0.95517	0.33833	1.48445
0.04	0.01814	-0.40535	0.19941	1.04266	0.39323	1.58275
0.05	0.04911	-0.35235	0.22218	1.11971	0.44428	1.66794
0.06	0.07546	-0.30733	0.24165	1.18976	0.49280	1.74444
0.07	0.09857	-0.26794	0.25882	1.25478	0.53959	1.81474
0.08	0.11926	-0.23275	0.27426	1.31600	0.58513	1.88043
0.09	0.13807	-0.20081	0.28837	1.37427	0.62978	1.94253
0.10	0.15539	-0.17148	0.30142	1.43019	0.67379	2.00182
0.15	0.22710	-0.05087	0.35631	1.68696	0.88948	2.27148
0.20	0.28410	0.04368	0.40124	1.92353	1.10582	2.51907
0.25	0.33299	0.12342	0.44116	2.15276	1.32868	2.76162
0.30	0.37690	0.19348	0.47857	2.38180	1.56126	3.01001
0.35	0.41759	0.25658	0.51505	2.61573	1.80541	3.27375
0.40	0.45620	0.31428	0.55183	2.85893	2.06198	3.56308
0.45	0.49356	0.36754	0.58999	3.11573	2.33096	3.89040
0.50	0.53032	0.41693	0.63057	3.39096	2.61173	4.27141
0.55	0.56709	0.46296	0.67451	3.69051	2.90372	4.72622
0.60	0.60444	0.50618	0.72271	4.02199	3.20757	5.28094
0.65	0.64305	0.54734	0.77603	4.39594	3.52649	5.97082
0.70	0.68374	0.58745	0.83551	4.82770	3.86764	6.84712
0.75	0.72765	0.62776	0.90265	5.34134	4.24384	7.99198
0.80	0.77655	0.66999	0.98009	5.97787	4.67723	9.55182
0.85	0.83354	0.71675	1.07280	6.81617	5.20898	11.82500
0.90	0.90525	0.77313	1.19192	8.03992	5.93102	15.55685
0.91	0.92257	0.78645	1.22098	8.36704	6.11581	16.63355
0.92	0.94139	0.80083	1.25266	8.73752	6.32162	17.89203
0.93	0.96208	0.81653	1.28760	9.16385	6.55428	19.39079
0.94	0.98519	0.83394	1.32673	9.66463	6.82242	21.21933

0.95	1.01154	0.85367	1.37150	10.26925	7.13946	23.52336
0.96	1.04250	0.87669	1.42425	11.02811	7.52812	26.56140
0.97	1.08056	0.90479	1.48930	12.03830	8.03145	30.85292
0.98	1.13116	0.94189	1.57603	13.52585	8.74757	37.67327
0.99	1.21092	0.99987	1.71322	16.25233	9.99702	51.66816

Output from Probit Procedure

Probit Procedure

Data Set =WORK.A
Dependent Variable=RESPONSE
Dependent Variable=N
Number of Observations= 7
Number of Events = 38 Number of Trials = 74
Observations with Missing Values= 29

Log Likelihood for LOGISTIC -37.11065336

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	1	-3.2246442	0.886057	13.24465	0.0003	Intercept
Log10(DOS)	1	5.97017999	1.449172	16.97208	0.0001	

Estimated Covariance Matrix

	INTERCPT	Log10(DOSE)
INTERCPT	0.785097	-1.215470
Log10(DOSE)	-1.215470	2.100098

Probit Model in Terms of Tolerance Distribution

MU	SIGMA
0.540125	0.167499

Estimated Covariance Matrix for Tolerance Parameters

	MU	SIGMA
MU	0.002378	-0.000381
SIGMA	-0.000381	0.001653

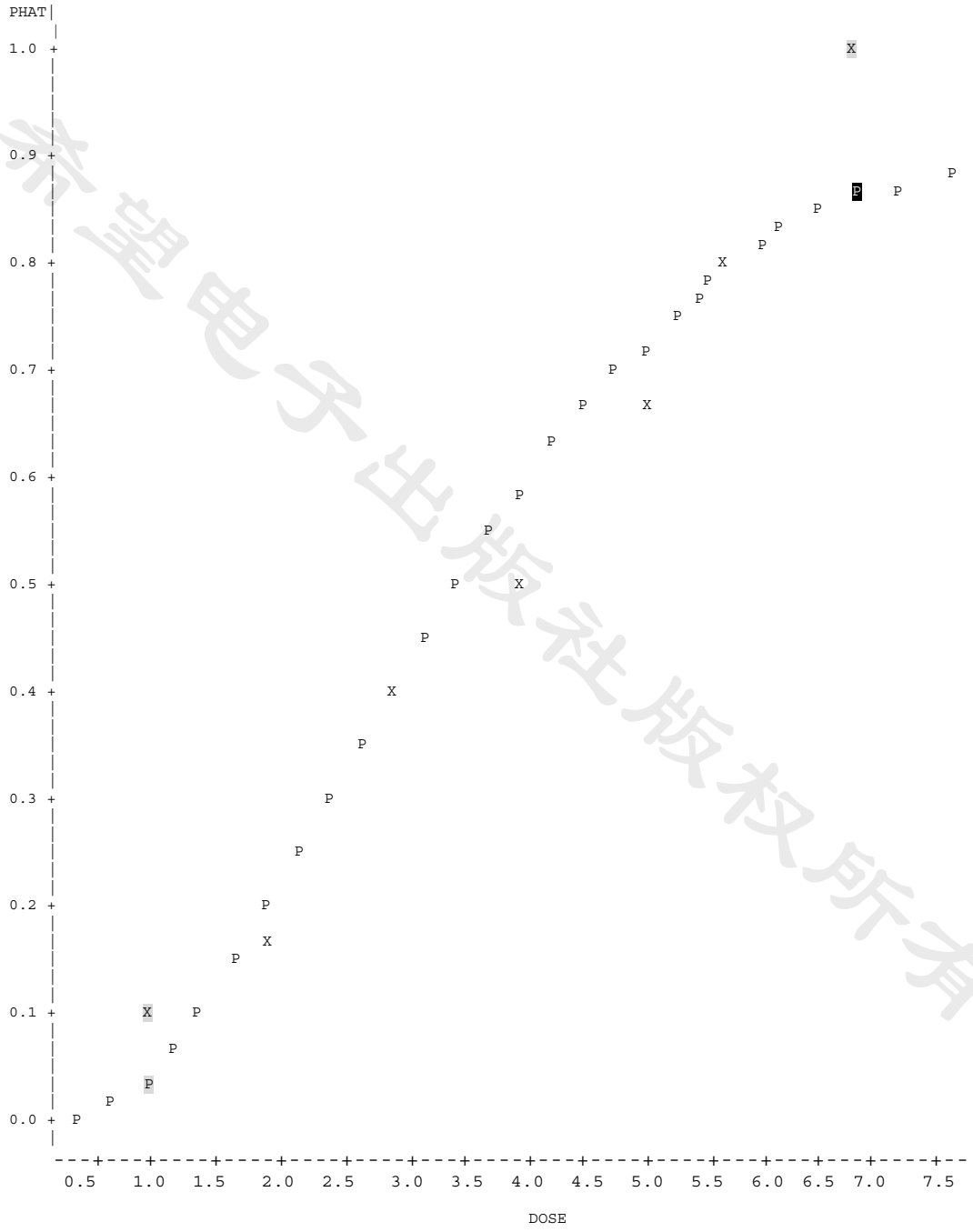
Probit Analysis on DOSE

Probability	Log10(DOSE)	95 Percent Fiducial Limits		DOSE	95 Percent Fiducial Limits	
		Lower	Upper		Lower	Upper
0.01	-0.22955	-0.97443	0.04234	0.58945	0.10606	1.10241
0.02	-0.11175	-0.75160	0.12404	0.77312	0.17717	1.33059
0.03	-0.04212	-0.62020	0.17266	0.90757	0.23977	1.48818
0.04	0.00780	-0.52620	0.20771	1.01813	0.29772	1.61328
0.05	0.04693	-0.45266	0.23533	1.11413	0.35264	1.71923
0.06	0.07925	-0.39207	0.25827	1.20018	0.40545	1.81245
0.07	0.10686	-0.34039	0.27796	1.27896	0.45668	1.89655
0.08	0.13103	-0.29522	0.29530	1.35218	0.50673	1.97380
0.09	0.15259	-0.25503	0.31085	1.42100	0.55586	2.04573
0.10	0.17209	-0.21876	0.32498	1.48625	0.60428	2.11340
0.15	0.24958	-0.07553	0.38207	1.77656	0.84036	2.41031
0.20	0.30792	0.03091	0.42645	2.03199	1.07377	2.66962
0.25	0.35611	0.11742	0.46451	2.27043	1.31043	2.91417
0.30	0.39820	0.19143	0.49933	2.50152	1.55391	3.15737
0.35	0.43644	0.25684	0.53275	2.73172	1.80650	3.40997
0.40	0.47221	0.31587	0.56619	2.96627	2.06954	3.68293
0.45	0.50651	0.36985	0.60090	3.21006	2.34343	3.98929
0.50	0.54013	0.41957	0.63807	3.46837	2.62766	4.34580
0.55	0.57374	0.46559	0.67895	3.74746	2.92137	4.77469
0.60	0.60804	0.50846	0.72475	4.05546	3.22449	5.30576
0.65	0.64381	0.54895	0.77673	4.40366	3.53960	5.98046
0.70	0.68205	0.58815	0.83638	4.80891	3.87389	6.86087
0.75	0.72414	0.62752	0.90583	5.29836	4.24153	8.05054
0.80	0.77233	0.66915	0.98877	5.92009	4.66819	9.74470
0.85	0.83067	0.71631	1.09243	6.77126	5.20363	12.37174
0.90	0.90816	0.77561	1.23344	8.09391	5.96506	17.11758
0.91	0.92766	0.79014	1.26932	8.46559	6.16797	18.59179
0.92	0.94922	0.80607	1.30913	8.89644	6.39834	20.37650
0.93	0.97339	0.82378	1.35392	9.40575	6.66466	22.59024
0.94	1.00100	0.84384	1.40524	10.02317	6.97974	25.42373
0.95	1.03332	0.86713	1.46548	10.79732	7.36425	29.20649
0.96	1.07245	0.89511	1.53866	11.81534	7.85434	34.56649
0.97	1.12237	0.93053	1.63230	13.25466	8.52168	42.88406
0.98	1.19200	0.97952	1.76331	15.55972	9.53935	57.98471
0.99	1.30980	1.06166	1.98571	20.40815	11.52540	96.76344

Plot of Observed and Fitted Probabilities

Plot of PHAT*DOSE. Symbol used is 'X'.

Plot of PROB*DOSE. Symbol used is 'P'.



NOTE：29 obs had missing values. 9 obs hidden.

例二：多重反应的分析

在这个范例中，我们示范两个自变量 (PRE=杀虫剂的种类，DOSE=杀虫剂的用药量) 对昆虫所产生的影响。杀虫剂的种类分两类：标准剂 (Stand) 以及试验品 (Test)。杀虫剂的用量分四类：轻 (10)、稍轻 (20)、稍重 (30)，与重 (40) 等。昆虫的反应 (SYMPTOMS) 分三类：无反应 (None)、轻微 (Mild) 以及极重 (Severe) 等。

分析的过程中，首先考虑一个不平行的模型 (其标签是 Nonpara)。根据这个不平行的模型，标准杀虫剂的用药量与试验品对昆虫所引起的反应症状是不同的。因此，MODEL 指令的撰写必须添加第三个自变量 (PREPDOSE)；第三个自变量的功用是用来区分两种杀虫剂所界定的两组数据。

其次，程序中再考虑一个平行的模型 (其标签是 Parallel)。根据平行的模型，昆虫的反应症状不受两类杀虫剂的影响，只受杀虫剂多少的影响。因此，在 MODEL 指令的等号右边不再含第三个自变量 (PREPDOSE)。

值得读者注意的是：无论是平行或不平行的模型，杀虫剂的用药量都先经过以 10 为底的对数转换 (LDOSE)，然后才进行分析。因此，所求得的函数关系必须再经过指数函数的转换才能真正表现出 DOSE 与 SYMPTOMS 之间的关系。

程　序

```
DATA MULTI;
    INPUT PREP $ DOSE SYMPTOMS $ N @@;
    LDOSE=LOG10(DOSE);
    IF PREP='TEST' THEN PREPDOSE=LDOSE;
    ELSE PREPDOSE=0;
    CARDS;
STAND  10  None    33     TEST   10  None    44
STAND  10  Mild     7     TEST   10  Mild     6
STAND  10  Severe  10     TEST   10  Severe   0
STAND  20  None    17     TEST   20  None    32
STAND  20  Mild    13     TEST   20  Mild    10
STAND  20  Severe  17     TEST   20 Severe   12
STAND  30  None    14     TEST   30  None    23
STAND  30  Mild     3     TEST   30  Mild     7
STAND  30  Severe  28     TEST   30  Severe  21
STAND  40  None     9     TEST   40  None    16
STAND  40  Mild     8     TEST   40  Mild     6
STAND  40  Severe  32     TEST   40  Severe  19
;
PROC PROBIT ORDER=DATA;
    CLASS PREP SYMPTOMS;
    NONPARA:MODEL SYMPTOMS=PREP LDOSE PREPDOSE/LACKFIT;
    WEIGHT N;
    PARALLEL:MODEL SYMPTOMS=PREP LDOSE/LACKFIT;
    WEIGHT N;
    TITLE 'Probit Models for Symptom Severity';
RUN;
```

结　果

不平行的模型所求得的结果显示：第三个自变量 PREPDOSE 所对应的参数未达统计显着的程度 (其 Wald 氏卡平方统计值等于 0.039)。不过，两个模型的适合度检验未显示出何者较优，何者较劣。因此，从精简的标准看来，第二个模型 (亦即平行的模型) 会较第一个模型 (亦即不平行的模型) 来得有效。

根据第二个模型的参数估计值，LDOSE 的值愈增加，昆虫产生极严重症状 (Severe) 的概率也相对地增加。换句话说，其不产生反应的概率 (None 或 Mild 的反应) 相对地减低。此外，标准杀虫剂 (Stand)，一般而言，较试验品 (Test) 的效力更强烈。这个结论不受杀虫剂药量的影响，乃是各用药量组一致观察到的现象。

报表 19.2　多重反应的分析

```
                Probit Models for Symptom Severity

                      Probit Procedure

                   Class Level Information

               Class    Levels    Values
               SYMPTOMS      3    None Mild Severe
               PREP          2    STAND TEST
                 Number of observations used = 23
Data Set          =WORK.MULTI
Dependent Variable=SYMPTOMS
Weight Variable   =N

     Weighted Frequency Counts for the Ordered Response Categories

                        Level     Count
                         None       188
                         Mild        60
                       Severe       139

Log Likelihood for NORMAL -345.9401767

                     Goodness-of-Fit Tests

        Statistic               Value      DF   Prob>Chi-Sq
        ------------------       --------    --   -----------
        Pearson Chi-Square       757.9822    41       0.0000
        L.R.    Chi-Square       691.8804    41       0.0000

        Response Levels:   3  Number of Covariate Values:   23
```

```
WARNING: All variances and covariances have been multiplied by the
       heterogeneity factor H= 18.487. Please check to be sure that the
       large chi-square (p < 0.0001) is not caused by systematic departure
       from the model. A t value of  2.0195 will be used in computing
       fiducial limits.

 Variable  DF   Estimate  Std Err ChiSquare  Pr>Chi Label/Value

 INTERCPT   1 3.80802509 2.688023  2.006936  0.1566 Intercept

 PREP       1                      0.127483  0.7211
            1 -1.2572764 3.521313  0.127483  0.7211  STAND
            0          0       0         .       .   TEST

 LDOSE      1 -2.1511952 1.680684  1.638277  0.2006
 PREPDOSE   1 -0.5072196 2.556132  0.039375  0.8427
 INTER.2    1 0.46843813 0.240404                    Mild

              Estimated Covariance Matrix

                   INTERCPT      PREP.1       LDOSE

 INTERCPT         7.225470    -7.218672   -0.017052
  PREP.1         -7.218672    12.399645   -3.752596
   LDOSE         -0.017052    -3.752596    2.824700
 PREPDOSE        -5.132611     8.893583   -2.802939
  INTER.2         0.030594    -0.017753   -0.032212

                          PREPDOSE          INTER.2

        INTERCPT        -5.132611         0.030594
         PREP.1          8.893583        -0.017753
          LDOSE         -2.802939        -0.032212
        PREPDOSE         6.533813        -0.006831
         INTER.2        -0.006831         0.057794

          Probit Models for Symptom Severity

                Probit Procedure
             Class Level Information

          Class    Levels    Values

          SYMPTOMS      3    None Mild Severe

          PREP          2    STAND TEST
```

Number of observations used = 23

Data Set =WORK.MULTI
Dependent Variable=SYMPTOMS
Weight Variable =N

Weighted Frequency Counts for the Ordered Response Categories

Level	Count
None	188
Mild	60
Severe	139

Log Likelihood for NORMAL -346.306141

Goodness-of-Fit Tests

Statistic	Value	DF	Prob>Chi-Sq
Pearson Chi-Square	758.6189	42	0.0000
L.R. Chi-Square	692.6123	42	0.0000

Response Levels: 3 Number of Covariate Values: 23

WARNING: All variances and covariances have been multiplied by the heterogeneity factor H= 18.062. Please check to be sure that the large chi-square (p < 0.0001) is not caused by systematic departure from the model. A t value of 2.0181 will be used in computing fiducial limits.

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	1	3.41481724	1.753563	3.792208	0.0515	Intercept
PREP	1			1.125567	0.2887	
	1	-0.5675155	0.534924	1.125567	0.2887	STAND
	0	0	0	.	.	TEST
LDOSE	1	-2.372131	1.253524	3.581063	0.0584	
INTER.2	1	0.46780285	0.237314			Mild

Estimated Covariance Matrix

	INTERCPT	PREP.1	LDOSE	INTER.2
INTERCPT	3.074984	-0.214464	-2.142736	0.024505

PREP.1	-0.214464	0.286144	0.053568	-0.008211
LDOSE	-2.142736	0.053568	1.571322	-0.034202
INTER.2	0.024505	-0.008211	-0.034202	0.056318

例三：逻辑斯谛的回归分析

这个例子旨在示范如何利用 PROC PROBIT 执行逻辑斯谛的回归分析 (见第 20 章的解说)。数据的来源是四十位成年人，他们的年龄 (AGE) ，性别 (SEX) 以及是否订阅一种新报纸的倾向 (SUBS)。若 SUBS=0，则表示该受试者倾向于订阅；反之，若 SUBS=1，则该受试者无此意愿。因此，本例题所考虑的数学模型如下：

$p=Pr(SUBS=0)=F(b_0 + b_1*SEX + b_2*AGE)$

此外，F 表累积的逻辑斯谛分配。

程　序

```
DATA NEWS;
    INPUT SEX $ AGE SUBS @@;
    CARDS;
Female  35  1   Female    48  1    Male    50  0   Female  39  1
Male    44  1   Female    56  0    Female  45  1   Male    34  1
Male    45  0   Male      46  0    Female  47  1   Female  52  0
Female  47  0   Female    59  0    Female  30  0   Female  46  1
Female  51  1   Female    46  0    Female  39  1   Male    58  0
Female  47  1   Male      59  0    Female  51  1   Female  50  0
Male    54  0   Male      38  0    Female  45  1   Female  32  1
Male    47  0   Female    39  1    Female  43  0   Female  52  0
Female  35  1   Male      49  0    Male    39  0   Female  35  1
Female  34  1   Male      42  0    Male    31  1   Female  51  1
;
PROC PROBIT;
    CLASS SUBS SEX;
    MODEL SUBS=SEX AGE/D=LOGISTIC ITPRINT;
    TITLE 'Logistic Regression of Subscription Status';
RUN;
```

结　果

分析的结果显示：b_0=-5.7620267，b_1=-2.4224077，b_2=0.1649503。所以，订阅新报纸的可能性随年龄而增加。然而，同一年龄组内，男性较女性更倾向于订阅!

报表 19.3　逻辑斯谛的回归分析

```
Logistic Regression of Subscription Status
            Probit Procedure
        Class Level Information
  Class     Levels    Values
  SUBS           2    0 1
  SEX            2    Female Male
  Number of observations used = 40
```

```
Iter Ridge   LogLikelihood     INTERCPT        SEX.1          AGE
   0     0  -27.7258872224            0            0            0
   1     0 -20.14265929083 -3.634567629 -1.648455751 0.1051634384
   2     0 -19.52245047938 -5.254865196 -2.234724956 0.1506493473
   3     0  -19.4904387863 -5.728485385 -2.409827238 0.1639621828
   4     0 -19.49030280973  -5.76187293 -2.422349862 0.1649007124
   5     0 -19.49030280687   -5.7620267 -2.422407743 0.1649050312
  Data Set           =WORK.NEWS
  Dependent Variable=SUBS

       Weighted Frequency Counts for the Ordered Response Categories
                                  Level      Count
                                      0         20
                                      1         20

  Log Likelihood for LOGISTIC -19.49030281
                       Last Evaluation of the Gradient
                         INTERCPT              SEX.1                AGE
                    -5.95482E-12       8.768327E-10       -1.636697E-8
                       Last Evaluation of the Hessian
                           INTERCPT              SEX.1                AGE
          INTERCPT         6.459740           4.604222          292.040518
             SEX.1         4.604222           4.604222          216.208295
               AGE       292.040518         216.208295               13487
        Variable  DF   Estimate  Std Err ChiSquare  Pr>Chi Label/Value
        INTERCPT   1 -5.7620267  2.76345  4.347576  0.0371 Intercept
        SEX        1                      6.422  0.0113
                   1 -2.4224077 0.955899     6.422  0.0113 Female
                   0          0        0         .       .  Male
        AGE        1 0.16490503 0.065188  6.399204  0.0114
                       Estimated Covariance Matrix
                           INTERCPT              SEX.1                AGE
          INTERCPT         7.636658           0.518768          -0.173672
             SEX.1         0.518768           0.913743          -0.025881
               AGE        -0.173672          -0.025881           0.004250
```

19.4　注 意 事 项

■选项 OUTEST= 输出文件的进一步说明

此选项所导出的输出文件含参数的估计值以及各模型所导出的对数可能数(LogLikelihood)。针对 MODEL 指令所界定的每一个统计模型，OUTEST= 文件内会自动包括一组与其有关的观察体。因此，当读者界定两个或两个以上的统计模型时，最好在模型的指令中加上标签 (如：M1，M2) 以便区分各组的参数估计值(有兴趣的读者，可参

考例一的程序)。

以下列举 OUTEST= 文件内所含的变量：

(1) BY　　指令中的变量名称串。

(2) _MODEL_，　是一个文字变量，其值代表各统计模型的标签 (如：M1，M2)，标签的长度以八个字元为限。

(3) _NAME_，　是一个文字变量，其值代表模型中反应变量的名称（为了识别各组的参数估计值)，或 (参数估计值之) 共变异数矩阵的横列。

(4) _TYPE_，　是一个文字变量，其值是 "PARMS" 或 "COV"。PARMS 表示与上述反应变量相对应的参数估计值。COV 则表示与上述矩阵之横列相对应的共变异数。

(5) _DIST_，　是一个文字变量，其值代表统计模型中的累积分配。

(6) _LNLIKE_，　是一个数值变量，其值代表循环估计的最后一个过程中所产生的对数可能数。

(7) INTERCEP，　是一个数值变量，包含截距的估计值及共变异数。

第 20 章　逻辑斯谛回归分析：统计程序 PROC LOGISTIC

20.1　PROC LOGISTIC 程序概述

本程序适合用来预测一个二分的或次序变量的值。其统计理论基础是逻辑斯谛回归分析 (Logistic Regression)，这个分析所用的参数估计法是最大可能率法。

二分的因变量 (或称反应变量) 值可以是上榜、落榜的结果，或疾病经过治疗后治愈、复发的两种可能。不论其定义如何，逻辑斯谛分析的目的是为了找出这个因变量值与一组连续变量 (或称自变量) 之间的线性关系。这个线性关系的表示可用因变量的对数奇数比单位 (Logit)，常态数单位 (Normit)、或双对数单位 (Log-Log) 等。如此，就产生了三类的线性函数表示法 (由选项 LINK= 来界定)。

此外，LOGISTIC 程序也可利用三种简化模型的方式来帮助读者找到一个更精简的函数关系，这三种简化模型的方式由 MODEL 指令中的选项 SELECTION= 来界定。

同理，次序变量的值 （如 1=私营企业，2= 公私合营企业，3= 国营企业）亦可经由上述的指令控制来找出其与一组连续变量之间的线性关系。

若读者有意深入探讨这种回归分析的理论基础，可参考 Cox 与 Snell (1989) ——针对二分的反应变量分析，或 Freeman（1987），Hosmer 与 Lemeshow（1989）合着的作品。

20.2　逻辑斯谛回归模型的种类

在上节里，我们已经提到：反应变量可以是一个二分的变量或次序变量，模型的量化单位则可以是 Logit，Normit 或 Log-Log 等三种。在本节内，我们就这几种可能情况的排列组合再深入地介绍逻辑斯谛回归模型的种类：

二分反应变量的模型

若反应变量的值只可以是二分的 (如：1= 正向的结果，2= 负向的结果)，则任何一个观察体在此变量上得 1 (即正向结果) 的概率，p=Prob(Y =1|X)，可用对数奇数比的单位来表示，如：

$$\text{logit}(p)=\log[p/(1-p)]=\alpha+\beta'X$$

在此，X 代表一组自变量，α 是模型中的截距，β 是一组与 X 对应的回归系数 (也是待估计的一组参数)。

这种回归模型与一般的线性模型无异，都代表因变量 Y 的平均数，即 Prob(Y=1)，与一组连续变量间的函数对应关系。也正因为这种通性，Nelder 与 Wedderburn 在 1972 年的文献里将这种函数对应的模型称作附会函数 (Link Function)。

附会函数的数值除了可以用对数奇数比来表示外，也可用常态数或双对数单位来表示。这些量化单位的选择都由选项 LINK= 来控制。

次序变量的模型

若变量的数值有大、小或高低之分，如 1= 高中毕业，2=大 (专) 学毕业，3=研究所肄业或以上的学历...，则我们可用 1，..，k，k+1 的整数来代表这些组别。由于组别数可能不止 2，因此 LOGISTIC 程序得将上述的函数表示法改写成：

$$g[\text{Prob}(Y \leqq i|X) = \alpha^i + \beta'X \quad \text{在此，} 1 \leqq i \leqq k$$

所以，(k+1) 个组只需 k 个截距参数再加上一组 (k 个)与斜率有关的参数即可解释次序变量上反应分布的情况。

20.3　LOGISTIC 程序的基本语法与报表形式

其实，这个程序无论在功能上或基本语法上，都和 SAS 系统中其它的回归程序十分类似。所以，假设 Y 代表一个二分的反应变量，X1，X2 分别是两个自变量，则 SAS 程序的写法在 LOGISTIC 程序以及 REC 程序内完全一致：

```
PROC LOGISTIC;
     MODEL Y=X1 X2;
RUN;
```

上式中 Y 变量的组别可用数值 (如 1，2) 或文字 (如 F，M) 来表示。若组别以数值表示，则数值的大小代表组别的先后次序。若以文字表示组别，则其第一个字母就决定组别先后的排序。

对于二分的变量，读者也可利用频率次数来界定回归分析的模型。比方说，N 代表总实验的次数 (或样本的大小)，R 代表样本中表现出研究者有兴趣之反应的观察体个数，则 LOGISTIC 程序的 MODEL 指令可修正如下：

```
PROC LOGISTIC;
     MODEL R/N=X1 X2;
RUN;
```

上述 R，N 的定义方法与二项式分配中样本数及 "成功" 之个数的定义完全相同。这种程序写法所导出来的报表含各反应项目的剖面图 (Profile) 以及样本在各自变量上分布的情形(如平均数、标准差、最小值、最大值等)。

LOGISTIC 程序对参数估计的方法采用 IRLS 解法。IRLS 的全名是循环加权最小误差平方法 (Iteratively Reweighed Least Squares Algorithm)。分析所得的结果含 (标准化后的) 参数估计值，参数估计值的 χ^2 检定等。

若因变量是一个次序变量，则报表上也会打印等斜率假设的检定结果。

模型整体的有效度以对数可能率来表示，其值等于 [-2 log likelihood]。不论模型的形式是简单的 (只含一个或数个截距) 还是复杂的 (含 k 个截距与 k 个斜率参数)，这个对数可能率的检定都是针对模型中所有参数的联合有效度而设计的。因此，每个参数个别对

模型的影响力则必须看其它的统计量，如 Wald 氏的 χ^2 检定。

报表上除了推论性统计值之外，还打印了四个描述性的统计值，它们分别是：素摩尔系数（Somer's D）、甘玛系数 （Gamma）、陶系数 （Tau-a） 以及 C 系数等。

以上种种有关语法及报表形式的举例说明，请参见第 20.5 节的例一。

20.4　如何撰写 PROC LOGISTIC 程序

PROC LOGISTIC 含五道指令，它们的格式如下：

```
PROC LOGISTIC 选项串;
    MODEL   反应变量= 自变量名称串 / 选项串;
    OUTPUT OUT= 输出文件名称
          关键字=变量名称串 / ALPHA= 概率值;
    WEIGHT 变量名称;
    BY      变量名称串;
```

其中，PROC LOGISTIC 与 MODEL 两道指令是必需的，不可省略。其余三道指令在程序里出现的次序，可由读者自行决定。

指令 #1　PROC LOGISTIC 选项串;

有六个选项，其中 (1) 与 (2) 是界定输入文件；(3) 与 (4) 用来界定输出文件；(5) 与 (6) 可用来控制报表打印的统计值。

(1) DATA= 输入文件名称

指明到底对那一个输入文件执行 LOGISTIC 回归分析。若省略此选项，则 SAS 会自动找出在此程序之前最后形成的 SAS 文件，对它执行分析。

(2) ORDER=DATA

ORDER=INTERNAL (内设值)

ORDER=FORMATTED

这个选项界定反应变量下组别的先后次序。若 ORDER=DATA，则组别的先后次序以输入文件内各组出现的次序来决定。若 ORDER=INTERNAL，则组别以反应变量值的小大或字母排列的先后次序来决定。若 ORDER=FORMATTED，组别次序由外在格式决定。当省略此选项时，内设值是 ORDER=INTERNAL。

(3) COVOUT

要求 LOGISTIC 程序将共变异数矩阵也并入 OUTEST= 输出文件内。此选项必须与下一个选项 OUTEST= 同时界定。

(4) OUTEST= 输出文件名称

这个 TYPE=EST 的文件含回归模型的参数估计值，或外加共变异数矩阵 (如果读者界定前述 COVOUT 选项)。有关这一个文件的进一步说明，请参见第 20.6 节。

(5) NOSIMPLE

抑止所有有关自变量之描述性统计量的打印。

(6) NOPPINT

不印出任何回归分析的结果。

指令 #2　MODEL 反应变量=自变量名称串 / 选项串；

删除号（/）前的部分

反应变量可以是一个二分的名义变量或次序变量。若是一个二分变量，则读者可用 R/N 的形式来取代反应变量名称。有关这种界定形式，请参见第 20.3 节的举例。

自变量串的值必须是连续的数值。若读者不提任何自变量的名称，亦即等号右边是空白的，则 LOGISTIC 程序假设回归模型应只含截距参数，而不含斜率参数。

删除号（/）后的部分

有二十四个选项，可分成六大类：第一类选项界定回归的模型，第二类选项与模型界定的过程有关，第三类选项与模型适合度的检验有关，第四类选项与报表上打印的分析过程有关，第五类选项与分析结果的表现形式有关，第六类选项则与回归分析中的诊断统计量有关。以下就这六个分类一一介绍其所属的选项：

第一类选项　下面四选项与回归模型的界定有关：

(1) LINK=LOGIT (或 NORMIT 或 CLOGLOG)

这个选项的目的是界定模型的量化单位。有关 LOGIT 的定义，在前面第 20.2 节内已有介绍，故不再赘述。

NORMIT 分数的导出，则根据下列的反应函数：

$g(p)=\Phi^{-1}(p)$

在此，Φ^{-1} 代表累积常态概率函数的反函数。一般文献中，将 Normit 的分数也称作 Probit (见第 19 章的 19.1 节介绍)。不过，根据定义，Probit 分数应等于 Normit 加上 5。

LINK=CLOGLOG 所导出的反应变量值是双对数，其对应的反应函数如下：

$g(p)=\log(-\log(1-p))$

这个反应函数是极端值之累积函数的反函数。原累积函数的表示式如下：

$F(X)=1-\exp(-\exp(X))$

这三个反应函数所造成的分数分布，其变异数及平均数皆不同 (参见下页的表 20.1)。

表 20.1　三个反应函数之分布的平均数与变异数

反应函数值	平均数据	变异数
Normit	0	1
Logit	0	π 2/3
Log-Log	- ɣ *	π 2/6

* ɣ = 一个 Euler 常数

由上表可知，由这三个反应函数所得的参数估计值不能直接比较，这是由于测量单位之不同所致。不仅如此，双对数的平均数与常态数及对数奇数比均不同。因此，除了测量单位相异之外，分数的正负号也不完全代表相同的意义。

这个选项的内设值是 LINK=LOGIT。

(2) NOINT

要求 LOGISTIC 程序在塑造回归模型时不考虑截距参数 (如果反应变量是一个二分的变量) 或不考虑第一组的截距参数 (如果反应函数是一个次序变量)。

(3) NOFIT

要求 LOGISTIC 程序将分析的重点放在所有自变量与反应变量之间的整体关系，而非参数的估计值或 χ^2 检定。

(4) SELECTION=NONE　　(或 N，内设值)

SELECTION=FORWARD　(或 F)

SELECTION=BACKWARD (或 B)

SELECTION=STEPWISE (或 S)

SELECTION＝SCORE

这个选项界定五种选择 "最佳" 回归模型的方法。NONE 要求将所有自变量均包括在回归模型里 (又称全型的模型)。当，LOGISTIC 程序逐次加增模型中参数的个数，直至模型以外的自变量均不能达到 [SLENTRY=概率值] 的显著度或当 [STOP=正整数] 的选项已符合。读者也可利用 [START=，INCLUDE= 正整数]两个选项来要求模型中至少应包括几个自变量。SELECTION=BACKWARD (反向淘汰法) 的分析策略与 FORWARD 刚好相反。根据反向分析的原理，模型的界定首先是全型的，然后，LOGISTIC 程序逐次将 "不重要" 的自变量剔除，直至模型内的变量均达到[SLSTAY=概率值] 的显著度或当 [STOP=正整数] 的选项已符合。此外，读者也可利用 [START=正整数] 的选项来控制模型中至少应包括几个自变量。

当 SELECTION=STEPWISE (逐步排除法) 的分析原理是顺向与反向两种方法的综合。换句话说，逐步排除法按照顺向选择法的逻辑不断挑选 "重要" 的自变量，将其纳入回归模型里。但同时，它也依据反向淘汰法的原则，对模型中既存的自变量一一作检定，看看它们当中是否有些自变量如今是多余的。若是，则逐步排除法仍有体会将这些 "不重要" 的自变量从模型中剔除出去。从上述的说明里，读者不难领悟，这三种界定模型的方法，仍以逐步排除法 (STEPWISE) 为最好。

当 SELECTION＝SCORE，LOGISTIC 程序根据最大可能分数（Likelihood Score，转换成卡平方值）来界定最佳的模型。模型中自变量的个数可以是 1，2，…，至总自变量数，或由 START＝（最低个数），STOP＝（最高个数），BEST＝（最佳个数）等选项控制。

第二类选项　下列十一个选项与模型界定的过程有关：

(1) DETAILS

要求将模型界定的过程详细地打印出来，包括模型里与模型外之自变量检定以及四个相关系数 (见第 20.3 节的说明)之数值。当读者选用 SELECTION=NONE

时，此选项无效。

(2) INCLUDE= 正整数 (如 3)

要求将自变量的前 (3) 个包括在每一个回归模型里。当 SELECTION= NONE 时，此选项无效。

(3) START= 正整数 (如 2)

这个指令规定回归模型里至少要含 MODEL 指令中前几个 (如 2 个) 自变量。若 SELECTION=FORWARD 或 STEPWISE，则 START 的内设值是零。若 SELECTION=BACKWARD，则指令的内设值等于 MODEL 指令中所提到之自变量的总个数。

值得读者注意的是， START= 与 INCLUDE= 两指令的意义不同。指令 START= 正整数的效力只及于第一个回归模型，以后模型的改变则不受此限制 。相对而言，INCLUDE= 正整数的效力贯穿整个回归分析的过程，所以，由此指令所界定的那几个自变量会从头到尾地留在每一步的回归模型中。

(4) STOP= 正整数 (如 5)

这个指令规定在顺向选择与反向淘汰法中，若找到最佳的 (5) 个自变量，则停止寻找。这个选项对 SELECTION=NONE 或 STEPWISE 不发生作用。当 SELECTION=FORWARD 时，这个选项的内设值是 MODEL 指令中所提到的自变量总数。当 SELECTION=BACKWARD 时，这个指令内设值等于零。

(5) BEST＝正整数（如 3）

界定最佳模型内自变量个数，必须与 SELECTION=SCORE 选项联用。

(6) SLENTRY (或 SLE)= 统计显著的程度

在顺向选择与逐步排除法中，这个指令可用来决定某一个变量是否有资格被纳入回归模型中。内设值是 0.05。

(7) SLSTAY (或 SLS)= 统计显著的程度

在反向淘汰与逐步排除法中，这个指令可用来决定某一个变量是否应继续被保留在回归模型中。内设值是 0.05。

(8) STOPRES (或 SR)

这个选项根据回归模型中误差的 χ^2 值来决定一个自变量是否应被纳入模型或自模型中排除。若 SELECTION=FORWARD，则 LOGISTIC 程序会不断将自变量添加在模型中直等到误差的 χ^2 值不再达到显著的程度(或其显著度大于 SLENTRY=的值) 时为止。当 SELECTION=BACKWARD 时，LOGISTIC 程序会不断地自模型中将自变量剔除，直等到误差的 χ^2 值达到显著的程度 (或其显著度小于 SLSTAY= 的值) 时停止。此选项对 SELECTION=NONE 或 STEPWISE 均无效。

(9) MAXSTEP= 正整数 (如 12)

这个选项必须与 SELECTION=STEPWISE 联用，其作用是限制自变量进出回归模型的总次数。内设值等于自变量总个数的两倍。

此选项对 SELECTION=NONE，FORWARD 或 BACKWARD 等方法无效。

(10) SEQUENTIAL (或 SEQ)

这个选项要求 LOGISTIC 程序在执行顺向选择法，反向淘汰法或逐步排除法时以 MODEL 指令中列举自变量的顺序来执行。此选项对 SELECTION=NONE 无影响力。

(11) FAST

此选项与 SELECTION=BACKWARD 或 STEPWISE 联用。其作用是要求 LOGISTIC 程序根据 Lawless 与 Singhal (1978) 所发展出来的计算程序，决定每一步骤中被剔除的自变量，其所对应的斜率参数的确达不到统计显著的程度。

第三类选项　有三个选项，与模型适合度的检验有关：

(1) CONVERGE= 极小的正实数

此选项界定最大可能率估计法的收敛指标。内设值等于 10 的 -4 次方。

(2) MAXITER= 正整数 (如 30)

此选项决定参数估计过程的循环估计之次数 (如 30 次)。内设值等于 25 次。

(3) SINGULAR= 极小的正实数

此选项判断数据矩阵的第二偏微分矩阵 (又称 Hessian 矩阵) 是否达到非满秩 (或奇异性) 的标准。这个选项的内设值等于 10 的 -12 次方。

第四类选项　与报表上打印的分析过程有关，有下列三个选项：

(1) ITPRINT

要求将分析过程每一步骤中所产生的统计值印出来。

(2) CORRB

要求打印出参数估计值间的相关系数矩阵。

(3) COVB

要求打印出参数估计值间的共变异数矩阵。

第五类选项　与分析结果的表现形式有关，含两个选项：

(1) CTABLE

根据模型内参数的估计值，将观察回归属于一个二分反应变量下两个类别的分类结果印在报表上。这个选项不适合用在含三组以上的反应变量上。

(2) PPROB= 概率值 (如 0.6)

与上述选项 CTABLE 联用。若某个观察体隶属于反应变量下第一组的概率达到这个标准(如大于或等于 0.6)，则 LOGISTIC 程序将这个观察体分到第一组。否则，这个观察体被分到第二组。所谓的第一组就是二项式分配中定义为“成功”的那一个组别。PPROB=的内设值是 0.5。

第六类选项　与回归分析中的诊断统计量有关，下含两个选项：

(1) INFLUENCE

这个选项可帮助读者将数据内对分析结果具过重之影响力的观察体找出来，只适用于二分的反应变量上。凡观察体对分析结果的影响力愈大，则报表上打印的统计量也愈大。

(2) IPLOTS

利用上述的统计量对每一个观察体作图。凡影响力过高的观察体在图形上都会显

得特别突出。请参见后面例四的示范。

指令 #3　OUTPUT OUT= 输出文件名称字=变量名称串 / ALPHA= 概率值；

删除号（/）前的部分

本指令包括两个部分：OUT= 与 关键字=，分别介绍如下：

OUT= 输出文件名称

这个文件含原输入文件的所有变量，以及本指令中所提到的变量（如：PREDICTED，RESCHI 等，详情见下面的说明）。若反应变量的组别以数值代表，如 (0，1) 或 (0，1，2)，则 OUT= 输出文件会含一个特殊变量 _LEVEL_，其值代表分析后观察体所属的组别 (请参见报表 20.2 最后的部分)。

关键字=变量名称串

下列是十三种关键字及其定义：

(1) PREDICTED (或 P)=观察体属于某一反应组别的预测概率。

(2) RESCHI　=皮尔森残差，其功能是找出与模型不太符合的观察体。

(3) RESDEV　=上述 RESCHI 的标准化值，其功能与 (2) 同。

(4) DIFCHISQ　=若将某个观察体自数据文件中剔除，则其对皮尔森 χ^2 适合度之数值的影响。

(5) DIFDEV　=若将某个观察体自数据文件中剔除，则其对偏激统计量 (Deviance) 的影响。

[上述 (4) 与 (5) 的值均可用来辨认与模型不符的观察体。]

(6) C　=信赖区间的错位诊断量 (Displacement Diagnostic)，其功能在于计算各观察体对参数估计值的影响。

(7) CBAR　=另一种信赖区间错位诊断量，其功能在于计算每一观察体自输入文件内剔除后对参数整体估计的影响。

(8) DFBETAS　=当某个观察体自输入资料内剔除后，造成每一参数估计值改变的标准化值。若读者选用此选项时，报表上所打印的标准化值以截距为首 (第一个 DFBETAS 值)，其次才是第一自变量的斜率参数改变之值 (第二个 DFBETAS 值)，以此类推。若模型中未包括某一自变量，则其对应的 DFBETAS 值是一个遗漏数据。

(9) H　=用来辨认实验设计空间中极端点的 Hat 矩阵的对角线元素。

(10) XBETA　=回归模型中参数与自变量值之线性组合的总值[亦即 $\alpha_i + X'\beta$ 之值，在此，i=次序反应变量上的组别，由变量 _LEVEL_ 而来]。

(11) STDXBETA　=上述 XBETA 值的标准误差。

(12) UPPER (或 U)　=前述 (1)P (或预测概率) 的上限。

(13) LOWER (或 L)　=前述 (1)P (或预测概率) 的下限。

下面举一例说明这些关键字的撰写：

```
PROC LOGISTIC DATA=A;
    MODEL  Y=X1 X2;
    OUTPUT OUT=B
           P=YHAT
           U=UPPER L=LOWER/ALPHA=.10;
```

这些指令最后产生的文件叫 B，它除了包括输入文件 (A) 的原有数据外，还包括了下列三个变量：YHAT (观察体属于某组别的预测概率)，UPPER (前述 YHAT 之 90% 信赖区间的上限)，LOWER (YHAT 之 90% 信赖区间的下限)。[ALPHA= 的说明见下面]

删除号 (/) 后的部分有一个选项，其说明如下：

(1) ALPHA= 概率值

界定前述之 P 值 (或预测概率) 的信赖度。内设值等于 0.05，根据这个内设值所产生的信赖区间是 95% 的信赖区间。若定 ALPHA=.10，则产生 90% 的信赖区间。

指令 #4　WEIGHT 变量名称；

这个指令所界定的变量值代表观察体在回归分析中的加权值。只有含正加权值的观察体才可进入回归模型中。值得注意的是：当数据中含加权值而反应变量是一个二分变量时，读者应使用 MODEL R/N=X1 X2...；的方式来界定模型而非 MODEL Y=X1 X2...；的方式。

指令 #5　BY 变量名称串；

LOGISTIC 程序依据此指令所列举的变量将文件分成几个小的文件，然后对每一个小的文件分别执行分析。当读者选用此指令时，文件内的数据必须先按照 BY 变量串的值做由小到大的重新排列， 这个步骤可藉 PROC SORT 来达成。或者，文件可先经由 PROC DATASETS 处理，将分组的代号附加在每一个观察体旁。如此，读者可直接使用这个分组代号来撰写 BY 的指令。有关这种处理法的详细介绍，可参考附录 C.9 节。

20.5　范　　例

例一：金属条的分析

本文件 (INGOTS) 含四个变量，它们分别是样本的总数 (N)，样本中尚不够软化的金属条个数 (R)，金属条加热的时间 (HEAT) 以及金属条浸在化学溶液中的时间 (SOAK)。分析的目的是找出 R/N 的比例与加热时间及溶液浸泡时间之间的关系。所有数据由 Cox 及 Snell (1989，第 10-11 页) 提供。

程　序

```
DATA INGOTS;
    INPUT HEAT SOAK R N @@;
    CARDS;
```

```
 7 1.0 0 10   7 1.7 0 17   7 2.2 0  7   7 2.8 0 12   7 4.0 0  9
14 1.0 0 31  14 1.7 0 43  14 2.2 2 33  14 2.8 0 31  14 4.0 0 19
27 1.0 1 56  27 1.7 4 44  27 2.2 0 21  27 2.8 1 22  27 4.0 1 16
51 1.0 3 13  51 1.7 0  1  51 2.2 0  1  51 4.0 0  1
;
PROC LOGISTIC DATA=INGOTS;
     MODEL R/N=HEAT SOAK;
RUN;
```

结　果

首先请看报表上标为 "Criterion for Assessing Model Fit" 的部分：两个自变量 (即 HEAT 与 SOAK，在报表上则称作共变量——Covariates) 与 R/N 的关系，若以 SCORE 表之，则其值达非常显著的程度 (P=0.0005)。这个关系，若以对数可能率来表示，则其值也达到相当显著的程度 (P=0.0030)。此外，对数可能率的检定也可应用在截距 (Intercept Only) 或截距加上自变量 (Intercept and Covariates) 的混合模型上。至于赤池资讯量指标 (AIC) 或萧氏指标 (SC 或 Schwartz Criterion) 在这个例子中则没有太大的意义。这是因为这两个指标最好用来比较模型的优劣；愈是优良的模型，其所对应的这两个指标值也愈小。

其次，请看报表上标为 "Analysis of Maximum Likelihood Estimates" 的部分。根据这一部分参数的估计值，我们可说：

logit(p)=-5.5592 + (0.082)*HEAT + (0.0568)*SOAK

因此，若 HEAT=7，SOAK=1，则 logit(p)=-4.9284。由于 logit(p) 代表 P 值的对数奇数比。所以，P 值应等于

$P=e^{-4.9284}/(1+e^{-4.9284})=0.0072$

这个 P 值就是一块金属条在 HEAT=7，SOAK=1 的情况下仍然不能软化的预测概率 (亦即输出文件内的 P 关键字)。

最后，报表所打印的统计值是关于上述模型的预测力，四个相关系数的值及若干计算过程所用到的统计量。它们均包括在 "Association of Predicted Probabilities and Observed Responses" 一栏内。

报表 20.1　金属条的分析

```
                        The LOGISTIC Procedure
Data Set：WORK.INGOTS
Response Variable (Events)：R
Response Variable (Trials)：N
Number of Observations：19
Link Function：Logit

                      Response Profile
                Ordered  Binary
                  Value  Outcome      Count
```

1	EVENT	12
2	NO EVENT	375

Simple Statistics for Explanatory Variables

Variable	Mean	Standard Deviation	Minimum	Maximum
HEAT	19.875969	9.936071	7.00000	51.0000
SOAK	2.033333	0.942794	1.00000	4.0000

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	108.988	101.346	.
SC	112.947	113.221	.
-2 LOG L	106.988	95.346	11.643 with 2 DF (p= 0.0030)
Score	.	.	15.109 with 2 DF (p= 0.0005)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-5.5592	1.1197	24.6504	0.0001	.	0.004
HEAT	1	0.0820	0.0237	11.9453	0.0005	0.449368	1.085
SOAK	1	0.0568	0.3312	0.0294	0.8639	0.029509	1.058

Association of Predicted Probabilities and Observed Responses

Concordant	= 64.4%	Somers' D	= 0.460
Discordant	= 18.4%	Gamma	= 0.555
Tied	= 17.2%	Tau-a	= 0.028
(4500 pairs)		c	= 0.730

例二：癌症复发率的逐步回归分析

本文件 (REMISS) 含二十七位癌症病人的病历 (以 CELL，SMEAR，INFIL， LI，BLAST 以及 TEMP 等六个变量表之) 以及他们是否复发的记录 (以 REMISS 表之：若 REMISS=1，则该病人的病状已进入潜伏期；否则，癌症仍有随时复发的可能)。数据的来源是 Lee （1974）年的论文。分析时，采用逐步排除法来建立回归的模型，结果则输入两个 SAS 文件：第一个含参数的估计值与变量间的共变异数矩阵；第二个文件则含各观察体病况不稳定（即属于 REMISS=0 组别）的预测概率以及该概率的 95% 信赖区间。

程　序

```
DATA REMISS;
     INPUT REMISS CELL SMEAR INFIL LI BLAST TEMP;
     LABEL REMISS='Complete remission';
     CARDS;
1 .8 .83 .66 1.9 1.1 .996
1 .9 .36 .32 1.4 .74 .992
0 .8 .88 .7 .8 .176 .982
0 1 .87 .87 .7 1.053 .986
```

```
1 .9 .75 .68 1.3 .519 .98
0 1 .65 .65 .6 .519 .982
1 .95 .97 .92 1 1.23 .992
0 .95 .87 .83 1.9 1.354 1.02
0 1 .45 .45 .8 .322 .999
0 .95 .36 .34 .5 0 1.038
0 .85 .39 .33 .7 .279 .988
0 .7 .76 .53 1.2 .146 .982
0 .8 .46 .37 .4 .38 1.006
0 .2 .39 .08 .8 .114 .99
0 1 .9 .9 1.1 1.037 .99
1 1 .84 .84 1.9 2.064 1.02
0 .65 .42 .27 .5 .114 1.014
0 1 .75 .75 1 1.322 1.004
0 .5 .44 .22 .6 .114 .99
1 1 .63 .63 1.1 1.072 .986
0 1 .33 .33 .4 .176 1.01
0 .9 .93 .84 .6 1.591 1.02
1 1 .58 .58 1 .531 1.002
0 .95 .32 .3 1.6 .886 .988
1 1 .6 .6 1.7 .964 .99
1 1 .69 .69 .9 .398 .986
0 1 .73 .73 .7 .398 .986
TITLE 'Stepwise Regression on Cancer Remission Data';
PROC LOGISTIC DATA=REMISS OUTEST=BETAS COVOUT;
     MODEL REMISS=CELL SMEAR INFIL LI BLAST TEMP
                 /SELECTION=STEPWISE
                  SLENTRY=0.3
                  SLSTAY=0.3
                  DETAILS;
     OUTPUT OUT=PRED P=PHAT LOWER=LCL UPPER=UCL;
RUN;
PROC PRINT DATA=BETAS;
     TITLE2 'Parameter Estimates and Covariance Matrix';
RUN;
PROC PRINT DATA=PRED;
     TITLE2 'Predicted Probabilities and 95% Confidence Limits';
RUN;
```

结　果

根据逐步排除的分析方法，第一个被纳入回归模型的变量是 LI ，接下来次序是 TEMP 及 CELL。 此时，除了这三个自变量之外，没有任何其它变量达到 SLENTRY=.30 的标准。因此，最佳的模型是：

logit(P)=(-67.6339)+(-9.6522)*CELL+(-3.8671)*LI+(82.0738)*TEMP

其次，OUTEST=BETAS 经 PROC PRINT 打印后显示：此文件含上述模型中 CELL， LI， TEMP 等的参数估计值。然而，其余的变量如 INFIL， BLAST， SMEAR

等因不是模型的一部分，故它们所对应的参数估计值空缺。

第二个输出文件 OUT=PRED 以观察体为数据单位，含所有原文件 (REMISS) 的数据以及病情不稳定的预测概率以及它的 95% 信赖区间的上下限 (分别以 UCL，LCL 表之)。以第一位病人 (OBS=1) 为例，其 PHAT=0.277，这表示他 (她) 的癌症复发的概率就是 0.277 (_LEVEL_=0 也就是 REMISS=0 的意思)。这个概率值的下限是 0.02907，上限是 0.83108。

报表 20.2　癌症复发率的逐步回归分析

Stepwise Regression on Cancer Remission Data

The LOGISTIC Procedure

Data Set: WORK.REMISS
Response Variable: REMISS　Complete remission
Response Levels: 2
Number of Observations: 27
Link Function: Logit

Response Profile

Ordered Value	REMISS	Count
1	0	18
2	1	9

Simple Statistics for Explanatory Variables

Variable	Mean	Standard Deviation	Minimum	Maximum
CELL	0.881481	0.186645	0.200000	1.00000
SMEAR	0.635185	0.214052	0.320000	0.97000
INFIL	0.570741	0.237567	0.080000	0.92000
LI	1.003704	0.467795	0.400000	1.90000
BLAST	0.688852	0.535804	0.000000	2.06400
TEMP	0.997000	0.014861	0.980000	1.03800

Step 0. Intercept entered:

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	0.6931	0.4082	2.8827	0.0895	.	2.000

Residual Chi-Square = 9.4609 with 6 DF (p=0.1493)

Analysis of Variables Not in the Model

Variable	Score Chi-Square	Pr > Chi-Square
CELL	1.8893	0.1693
SMEAR	1.0745	0.2999
INFIL	1.8817	0.1701

LI	7.9311	0.0049
BLAST	3.5258	0.0604
TEMP	0.6591	0.4169

Step 1. Variable LI entered:

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	36.372	30.073	.
SC	37.668	32.665	.
-2 LOG L	34.372	26.073	8.299 with 1 DF (p=0.0040)
Score	.	.	7.931 with 1 DF (p=0.0049)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	3.7771	1.3786	7.5064	0.0061	.	43.691
LI	1	-2.8973	1.1868	5.9594	0.0146	-0.747230	0.055

Association of Predicted Probabilities and Observed Responses

Concordant = 84.0%	Somers' D = 0.710
Discordant = 13.0%	Gamma = 0.732
Tied = 3.1%	Tau-a = 0.328
(162 pairs)	c = 0.855

Residual Chi-Square = 3.1174 with 5 DF (p=0.6819)

Analysis of Variables Not in the Model

Variable	Score Chi-Square	Pr > Chi-Square
CELL	1.1183	0.2903
SMEAR	0.1369	0.7114
INFIL	0.5715	0.4497
BLAST	0.0932	0.7601
TEMP	1.2591	0.2618

Step 2. Variable TEMP entered:

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	36.372	30.648	.
SC	37.668	34.535	.
-2 LOG L	34.372	24.648	9.724 with 2 DF (p=0.0077)
Score	.	.	8.365 with 2 DF (p=0.0153)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-47.8559	46.4416	1.0618	0.3028	.	0.000
LI	1	-3.3020	1.3594	5.9005	0.0151	-0.851626	0.037
TEMP	1	52.4331	47.4934	1.2188	0.2696	0.429597	999.000

Association of Predicted Probabilities and Observed Responses

Concordant = 87.0%	Somers' D = 0.747
Discordant = 12.3%	Gamma = 0.752
Tied = 0.6%	Tau-a = 0.345
(162 pairs)	c = 0.873

Residual Chi-Square = 2.1431 with 4 DF (p=0.7095)

Analysis of Variables Not in the Model

Variable	Score Chi-Square	Pr > Chi-Square
CELL	1.4701	0.2253
SMEAR	0.1730	0.6775
INFIL	0.8275	0.3630
BLAST	1.1014	0.2940

Step 3. Variable CELL entered:

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	36.372	29.953	.
SC	37.668	35.137	.
-2 LOG L	34.372	21.953	12.418 with 3 DF (p=0.0061)
Score	.	.	9.250 with 3 DF (p=0.0261)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-67.6339	56.8875	1.4135	0.2345	.	0.000
CELL	1	-9.6522	7.7511	1.5507	0.2130	-0.993231	0.000
LI	1	-3.8671	1.7783	4.7290	0.0297	-0.997359	0.021
TEMP	1	82.0738	61.7124	1.7687	0.1835	0.672450	999.000

Association of Predicted Probabilities and Observed Responses

Concordant = 88.9%	Somers' D = 0.778
Discordant = 11.1%	Gamma = 0.778
Tied = 0.0%	Tau-a = 0.359
(162 pairs)	c = 0.889

Residual Chi-Square = 0.1831 with 3 DF (p=0.9803)

Analysis of Variables Not in the Model

Variable	Score Chi-Square	Pr > Chi-Square
SMEAR	0.0956	0.7572
INFIL	0.0844	0.7714
BLAST	0.0208	0.8852

NOTE: No (additional) variables met the 0.3 significance level for entry into the model.

Summary of Stepwise Procedure

Step	Variable Entered	Removed	Number In	Score Chi-Square	Wald Chi-Square	Pr > Chi-Square
1	LI		1	7.9311	.	0.0049
2	TEMP		2	1.2591	.	0.2618
3	CELL		3	1.4701	.	0.2253

Parameter Estimates and Covariance Matrix

OBS	_LINK_	_TYPE_	_NAME_	INTERCPT	CELL	SMEAR	INFIL	LI	BLAST	TEMP	_LNLIKE_
1	LOGIT	PARMS	ESTIMATE	-67.63	-9.652	.	.	-3.8671	.	82.07	-10.9767
2	LOGIT	COV	INTERCPT	3236.19	157.097	.	.	64.5726	.	-3483.23	-10.9767
3	LOGIT	COV	CELL	157.10	60.079	.	.	6.9454	.	-223.67	-10.9767
4	LOGIT	COV	SMEAR	.	.	.	.	.	.	.	-10.9767
5	LOGIT	COV	INFIL	.	.	.	.	.	.	.	-10.9767
6	LOGIT	COV	LI	64.57	6.945	.	.	3.1623	.	-75.35	-10.9767
7	LOGIT	COV	BLAST	.	.	.	.	.	.	.	-10.9767
8	LOGIT	COV	TEMP	-3483.23	-223.669	.	.	-75.3513	.	3808.42	-10.9767

Predicted Probabilities and 95% Confidence Limits

OBS	REMISS	CELL	SMEAR	INFIL	LI	BLAST	TEMP	_LEVEL_	PHAT	LCL	UCL
1	1	0.80	0.83	0.66	1.9	1.100	0.996	0	0.27735	0.02907	0.83108
2	1	0.90	0.36	0.32	1.4	0.740	0.992	0	0.42126	0.16238	0.73212
3	0	0.80	0.88	0.70	0.8	0.176	0.982	0	0.89540	0.36581	0.99219
4	0	1.00	0.87	0.87	0.7	1.053	0.986	0	0.71742	0.34317	0.92502
5	1	0.90	0.75	0.68	1.3	0.519	0.980	0	0.28582	0.05124	0.74782
6	0	1.00	0.65	0.65	0.6	0.519	0.982	0	0.72911	0.31049	0.94148
7	1	0.95	0.97	0.92	1.0	1.230	0.992	0	0.67844	0.40484	0.86745
8	0	0.95	0.87	0.83	1.9	1.354	1.020	0	0.39277	0.04713	0.89428
9	0	1.00	0.45	0.45	0.8	0.322	0.999	0	0.83368	0.43877	0.96982
10	0	0.95	0.36	0.34	0.5	0.000	1.038	0	0.99843	0.31038	1.00000
11	0	0.85	0.39	0.33	0.7	0.279	0.988	0	0.92715	0.50018	0.99386
12	0	0.70	0.76	0.53	1.2	0.146	0.982	0	0.82714	0.12794	0.99363
13	0	0.80	0.46	0.37	0.4	0.380	1.006	0	0.99654	0.53470	0.99999
14	0	0.20	0.39	0.08	0.8	0.114	0.990	0	0.99982	0.03518	1.00000
15	0	1.00	0.90	0.90	1.1	1.037	0.990	0	0.42878	0.16027	0.74697
16	1	1.00	0.84	0.84	1.9	2.064	1.020	0	0.28530	0.02811	0.84638
17	0	0.65	0.42	0.27	0.5	0.114	1.014	0	0.99938	0.37335	1.00000
18	0	1.00	0.75	0.75	1.0	1.322	1.004	0	0.77711	0.36330	0.95517
19	0	0.50	0.44	0.22	0.6	0.114	0.990	0	0.99846	0.20356	1.00000
20	1	1.00	0.63	0.63	1.1	1.072	0.986	0	0.35089	0.09445	0.73695
21	0	1.00	0.33	0.33	0.4	0.176	1.010	0	0.98307	0.49525	0.99971
22	0	0.90	0.93	0.84	0.6	1.591	1.020	0	0.99378	0.43938	0.99997

23	1	1.00	0.58	0.58	1.0	0.531	1.002	0	0.74739	0.36403	0.93863
24	0	0.95	0.32	0.30	1.6	0.886	0.988	0	0.12989	0.01519	0.59089
25	1	1.00	0.60	0.60	1.7	0.964	0.990	0	0.06868	0.00427	0.55886
26	1	1.00	0.69	0.69	0.9	0.398	0.986	0	0.53949	0.21471	0.83388
27	0	1.00	0.73	0.73	0.7	0.398	0.986	0	0.71742	0.34317	0.92502

例三：癌症复发率的反向回归分析

这个例题沿用例二的数据，不过分析的方法由逐步分析改为反向淘汰法，而且程序中引用了 FAST 与 CTABLE 指令。因此分析的过程不但十分快速而且结果可呈现在一个 2×2 的列联表内，将正确与不正确的预测清楚地表示出来。

程　序

```
DATA REMISS;
    INPUT REMISS CELL SMEAR INFIL LI BLAST TEMP;
    LABEL REMISS='Complete remission';
    CARDS;
1 .8 .83 .66 1.9 1.1 .996
1 .9 .36 .32 1.4 .74 .992
0 .8 .88 .7 .8 .176 .982
0 1 .87 .87 .7 1.053 .986
1 .9 .75 .68 1.3 .519 .98
0 1 .65 .65 .6 .519 .982
1 .95 .97 .92 1 1.23 .992
0 .95 .87 .83 1.9 1.354 1.02
0 1 .45 .45 .8 .322 .999
0 .95 .36 .34 .5 0 1.038
0 .85 .39 .33 .7 .279 .988
0 .7 .76 .53 1.2 .146 .982
0 .8 .46 .37 .4 .38 1.006
0 .2 .39 .08 .8 .114 .99
0 1 .9 .9 1.1 1.037 .99
1 1 .84 .84 1.9 2.064 1.02
0 .65 .42 .27 .5 .114 1.014
0 1 .75 .75 1 1.322 1.004
0 .5 .44 .22 .6 .114 .99
1 1 .63 .63 1.1 1.072 .986
0 1 .33 .33 .4 .176 1.01
0 .9 .93 .84 .6 1.591 1.02
1 1 .58 .58 1 .531 1.002
0 .95 .32 .3 1.6 .886 .988
1 1 .6 .6 1.7 .964 .99
1 1 .69 .69 .9 .398 .986
0 1 .73 .73 .7 .398 .986
TITLE 'Backward Elimination on Cancer Remission Data';
PROC LOGISTIC DATA=REMISS NOSIMPLE;
    MODEL REMISS=TEMP CELL SMEAR INFIL LI BLAST
```

```
                        /SELECTION=BACKWARD
                         FAST
                         SLSTAY=0.2
                         CTABLE PPROB=.5;
    RUN;
```

结　果

根据反向淘汰的分析原理以及标准 (SLSTAY=0.2)，只有一个自变量 LI 被保留在模型内。若加上截距，则模型的结构如下：

logit(P)=3.7771 + (-2.8973)*LI

报表最后的部分是一个 2×2 的列联表，表的横列是原数据中 REMISS=0 (或 EVENT) 及 REMISS=1(或 NO EVENT) 的人数；表的直行则是上述回归模型执行后将观察体分组的结果。二十七位病人中有二十位，其病情被 LOGISTIC 程序正确地推测出来。所以，上述回归模型的预测有效度等于 74.1% (亦即 20/27×100%)。其它相关的统计量如敏感度 (Sensitivity=16/18×100%)、精确度(Specificity=4/9×100%) 等则打印在 2×2 列联表的下端。

报表 20.3　癌症复发率的反向回归分析

```
                    Backward Elimination on Cancer Remission Data
                                The LOGISTIC Procedure
Data Set：WORK.REMISS
Response Variable：REMISS    Complete remission
Response Levels：2
Number of Observations：27
Link Function：Logit
                                Response Profile
                              Ordered
                                Value  REMISS     Count
                                    1       0        18
                                    2       1         9
                              Backward Elimination Procedure
Step  0. The following variables were entered:
        INTERCPT  TEMP      CELL      SMEAR     INFIL     LI        BLAST
                                Criteria for Assessing Model Fit
                                                 Intercept
                                    Intercept        and
                  Criterion          Only       Covariates    Chi-Square for Covariates
                  AIC                36.372        35.751          .
                  SC                 37.668        44.822          .
                  -2 LOG L           34.372        21.751       12.621 with 6 DF (p=0.0495)
                  Score                .             .          9.461 with 6 DF (p=0.1493)

Step  1. Fast Backward Elimination:
                        Analysis of Variables Removed by Fast Backward Elimination
                                                                          Pr >
                  Variable                    Pr >        Residual        Residual
```

Removed	Chi-Square	Chi-Square	Chi-Square	DF	Chi-Square
BLAST	0.0044	0.9471	0.0044	1	0.9471
INFIL	0.0971	0.7554	0.1015	2	0.9505
SMEAR	0.0844	0.7714	0.1859	3	0.9798
CELL	1.3619	0.2432	1.5478	4	0.8181
TEMP	0.6533	0.4189	2.2010	5	0.8207

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	36.372	30.073	.
SC	37.668	32.665	.
-2 LOG L	34.372	26.073	8.299 with 1 DF (p=0.0040)
Score	.	.	7.931 with 1 DF (p=0.0049)

Residual Chi-Square = 3.1174 with 5 DF (p=0.6819)

Summary of Backward Elimination Procedure

Step	Variable Removed	Number In	Wald Chi-Square	Pr > Chi-Square
1	BLAST	5	0.00440	0.9471
1	INFIL	4	0.0971	0.7554
1	SMEAR	3	0.0844	0.7714
1	CELL	2	1.3619	0.2432
1	TEMP	1	0.6533	0.4189

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	3.7771	1.3786	7.5064	0.0061	.	43.691
LI	1	-2.8973	1.1868	5.9594	0.0146	-0.747230	0.055

Association of Predicted Probabilities and Observed Responses

Concordant	= 84.0%	Somers' D	= 0.710
Discordant	= 13.0%	Gamma	= 0.732
Tied	= 3.1%	Tau-a	= 0.328
(162 pairs)		c	= 0.855

下面报表是在 PC-6.04 中执行的结果：

Classification Table

		Predicted EVENT	Predicted NO EVENT	Total
Observed	EVENT	16	2	18
	NO EVENT	5	4	9
	Total	21	6	27

Sensitivity= 88.9%　Specificity= 44.4%　Correct= 74.1%

```
                                False Positive Rate= 23.8%  False Negative Rate= 33.3%
NOTE: An EVENT is an outcome whose ordered response value is 1.
下面报表是在主机或 WINDOWS 中执行的结果：
                                The LOGISTIC Procedure
                                 Classification Table
                  Correct       Incorrect                 Percentages
                ------------   ------------   ------------------------------------
                Prob            Non-           Non-                Sensi-   Speci-   False   False
                Level  Event    Event  Event   Event    Correct     tivity   ficity   POS     NEG
                ------------------------------------------------------------------------------------
                0.500     16        4      5       2       74.1       88.9     44.4    23.8    33.3
```

例四：四肢皮肤收缩的研究

这个例题示范气体的流通速度 (RATE) 与体积 (VOLUME) 是否引起四肢皮肤的收缩 (RESPONSE)。若皮肤有收缩的现象 (这也就是二项式分配的公式里一个 "成功" 的事件)，则 RESPONSE=1 (第一种反应)。否则，RESPONSE=0 (第二种反应)。文件 (VASO) 内的数据由 Finney (1947) 提供。分析时，首将 VOLUME 与 RATE 两个自变量转换成自然对数 (即 LOGVOL 与 LOGRATE)。然后，在模型的指令里要求对每一观察体执行诊断力分析 (INFLUENCE) 以及作图 (IPLOTS)。

程　序

```
DATA VASO;
     INPUT VOLUME RATE RESPONSE @@;
     LOGVOL=LOG(VOLUME);
     LOGRATE=LOG(RATE);
     CARDS;
3.7   .825 1  3.5  1.09  1  1.25 2.5   1   .75 1.5  1
 .8  3.2   1   .7  3.5   1   .6   .75  0  1.1  1.7  0
 .9   .75  0   .9   .45  0   .8   .57  0   .55 2.75 0
 .6  3.0   0  1.4  2.33  1   .75 3.75  1  2.3  1.64 1
3.2  1.6   1   .85 1.415 1  1.7  1.06  0  1.8  1.8  1
 .4  2     1   .95 1.36  0  1.35 1.35  0  1.5  1.36 0
1.6  1.78  1   .6  1.5   0  1.8  1.5   1   .95 1.9  0
1.9   .95  1  1.6   .4   0  2.7   .75  1  2.35  .03 0
1.1  1.83  1  1.1  2.2   1  1.2  2.0   1   .8  3.33 1
 .95 1.9   0   .75 1.9   0  1.3  1.625 1
;
TITLE 'Occurrence of Vaso-Constriction';
PROC LOGISTIC DATA=VASO ORDER=DATA;
     MODEL RESPONSE=LOGRATE LOGVOL/INFLUENCE IPLOTS;
RUN;
```

结　果

根据皮尔森残差以及偏激残差的图形，我们不难发现第 4、第 18 个与第 21 个观察体与模型 (或者其它观察体) 不太符合。若检视 Hat 矩阵的对角线元素的图形，则我

们可说第 31 个观察体过于偏激。 由其它的影响力指标，如 DFBETA 等来看，第 4、第 18 个与第 21 个观察体在样本中显然与其它观察体格格不入。因此，下一阶段的分析可以考虑将这三个 (或四个) 观察体剔除。然后，再检视类似的图形，看看是否所有的数据均与模型相调和。(绘图时限制其宽度为 100 行，长度为 30 列，以指令 OPTIONS LINESIZE=100；OPTIONS PAGESIZE=30；表示)

报表 20.4　四肢皮肤收缩的研究

Occurrence of Vaso-Constriction

The LOGISTIC Procedure

Data Set: WORK.VASO

Response Variable: RESPONSE

Response Levels: 2

Number of Observations: 39

Link Function: Logit

Response Profile

Ordered Value	RESPONSE	Count
1	1	22
2	0	17

Simple Statistics for Explanatory Variables

Variable	Mean	Standard Deviation	Minimum	Maximum
LOGRATE	0.317839	0.828558	-3.50656	1.32176
LOGVOL	0.159636	0.537657	-0.91629	1.30833

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	55.423	40.512	.
SC	57.086	45.502	.
-2 LOG L	53.423	34.512	18.911 with 2 DF (p=0.0001)
Score	.	.	13.853 with 2 DF (p=0.0010)

Analysis of Maximum Likelihood Estimates

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCPT	-1.5650	0.8784	3.1746	0.0748	.
LOGRATE	3.2739	1.3072	6.2726	0.0123	1.495553
LOGVOL	3.2238	1.2045	7.1631	0.0074	0.955610

Association of Predicted Probabilities and Observed Responses

Concordant = 89.3%	Somers' D = 0.789
Discordant = 10.4%	Gamma = 0.791
Tied = 0.3%	Tau-a = 0.398
(374 pairs)	c = 0.894

Occurrence of Vaso-Constriction

Regression Diagnostics

Case Number	Covariates LOGRATE	LOGVOL	Pearson Residual Value	(1 unit = 0.38) −8 −4 0 2 4 6 8	Deviance Residual Value	(1 unit = 0.27) −8 −4 0 2 4 6 8	Hat Matrix Diagonal Value	(1 unit = 0.01) 0 2 4 6 8 12 16
1	−0.1924	1.3083	0.3637	|*|	0.4984	|*|	0.1415	|*|
2	0.0862	1.2528	0.2521	|*|	0.3511	|*|	0.0776	|*|
3	0.9163	0.2231	0.3406	|*|	0.4685	|*|	0.0523	|*|
4	0.4055	−0.2877	1.7905	|*|	1.6950	|*|	0.0937	|*|
5	1.1632	−0.2231	0.4668	|*|	0.6279	|*|	0.0879	|*|
6	1.2528	−0.3567	0.4999	|*|	0.6680	|*|	0.1115	|*|
7	−0.2877	−0.5108	−0.1253	|*|	−0.1765	|*|	0.0438	|*|
8	0.5306	0.0953	−1.2709	|*|	−1.3866	|*|	0.0419	|*|
9	−0.2877	−0.1054	−0.2409	|*|	−0.3359	|*|	0.0888	|*|
10	−0.7985	−0.1054	−0.1044	|*|	−0.1473	|*|	0.0405	|*|
11	−0.5621	−0.2231	−0.1272	|*|	−0.1791	|*|	0.0482	|*|
12	1.0116	−0.5978	−0.9137	|*|	−1.1018	|*|	0.1453	|*|
13	1.0986	−0.5108	−1.2122	|*|	−1.3446	|*|	0.1339	|*|
14	0.8459	0.3365	0.3184	|*|	0.4394	|*|	0.0495	|*|
15	1.3218	−0.2877	0.3995	|*|	0.5443	|*|	0.0998	|*|
16	0.4947	0.8329	0.2541	|*|	0.3538	|*|	0.0519	|*|
17	0.4700	1.1632	0.1554	|*|	0.2185	|*|	0.0391	|*|
18	0.3471	−0.1625	1.6100	|*|	1.5992	|*|	0.0862	|*|
19	0.0583	0.5306	−1.1831	|*|	−1.3232	|*|	0.0972	|*|
20	0.5878	0.5878	0.3240	|*|	0.4468	|*|	0.0509	|*|
21	0.6931	−0.9163	3.0797	|*|	2.1679	|*|	0.1184	|*|
22	0.3075	−0.0513	−0.6964	|*|	−0.8892	|*|	0.0792	|*|
23	0.3001	0.3001	−1.2122	|*|	−1.3446	|*|	0.0549	|*|
24	0.3075	0.4055	−1.4541	|*|	−1.5073	|*|	0.0555	|*|
25	0.5766	0.4700	0.3989	|*|	0.5435	|*|	0.0497	|*|
26	0.4055	−0.5108	−0.3898	|*|	−0.5319	|*|	0.1026	|*|
27	0.4055	0.5878	0.4366	|*|	0.5908	|*|	0.0571	|*|
28	0.6419	−0.0513	−1.2038	|*|	−1.3385	|*|	0.0478	|*|
29	−0.0513	0.6419	0.8453	|*|	1.0384	|*|	0.1265	|*|
30	−0.9163	0.4700	−0.2177	|*|	−0.3043	|*|	0.1216	|*|
31	−0.2877	0.9933	0.7064	|*|	0.8998	|*|	0.2117	|*|
32	−3.5066	0.8544	−0.00583	|*|	−0.00824	|*|	0.000762	|*|
33	0.6043	0.0953	0.6974	|*|	0.8904	|*|	0.0413	|*|
34	0.7885	0.0953	0.5159	|*|	0.6870	|*|	0.0488	|*|

35	0.6931	0.1823	0.5241		*		0.6967		*		0.0447	*
36	1.2030	-0.2231	0.4374		*		0.5917		*		0.0891	*
37	0.6419	-0.0513	-1.2038	*			-1.3385	*			0.0478	*
38	0.6419	-0.2877	-0.8224	*			-1.0165	*			0.0784	*
39	0.4855	0.2624	0.6472		*		0.8365		*		0.0426	*

Occurrence of Vaso-Constriction

Case	INTERCPT Dfbeta (1 unit = 0.1)			LOGRATE Dfbeta (1 unit = 0.08)			LOGVOL Dfbeta (1 unit = 0.14)		
Number	Value	-8 -4 0 2 4 6 8		Value	-8 -4 0 2 4 6 8		Value	-8 -4 0 2 4 6 8	
1	0.00922	*		0.00304	*		0.1138	*	
2	-0.0133	*		0.0222	*		0.0654	*	
3	-0.0374	*		0.0669	*		0.0531	*	
4	0.5408	*		-0.3899	*		-0.4802	*	
5	-0.0531	*		0.1134	*		0.0313	*	
6	-0.0602	*		0.1326	*		0.0171	*	
7	-0.0260	*		0.0250	*		0.0232	*	
8	-0.1283	*		-0.00266	*		0.00114	*	
9	-0.0776	*		0.0738	*		0.0571	*	
10	-0.0210	*		0.0213	*		0.0156	*	
11	-0.0283	*		0.0280	*		0.0222	*	
12	-0.1048	*		-0.0439	*		0.2247	*	
13	-0.0201	*		-0.1822	*		0.1665	*	
14	-0.0339	*		0.0588	*		0.0543	*	
15	-0.0612	*		0.1123	*		0.0335	*	
16	-0.0218	*		0.0347	*		0.0548	*	
17	-0.0140	*		0.0190	*		0.0302	*	
18	0.4825	*		-0.3478	*		-0.3715	*	
19	-0.2548	*		0.1729	*		-0.0746	*	
20	-0.0254	*		0.0470	*		0.0641	*	
21	0.8313	*		-0.6021	*		-1.1084	*	
22	-0.2017	*		0.1433	*		0.1315	*	
23	-0.1892	*		0.0847	*		-0.0291	*	
24	-0.1503	*		0.0297	*		-0.1290	*	
25	-0.0214	*		0.0523	*		0.0707	*	
26	-0.1203	*		0.0942	*		0.1234	*	
27	-0.00780	*		0.0393	*		0.0821	*	
28	-0.1167	*		-0.0206	*		0.0513	*	
29	0.2059	*		-0.1539	*		0.0739	*	
30	-0.0799	*		0.0819	*		0.0411	*	
31	0.1792	*		-0.1488	*		0.1582	*	
32	-0.00014	*		0.000159	*		0.000087	*	
33	0.0393	*		0.0343	*		0.0211	*	
34	-0.0209	*		0.0754	*		0.0487	*	
35	-0.0123	*		0.0644	*		0.0548	*	
36	-0.0557	*		0.1111	*		0.0342	*	
37	-0.1167	*		-0.0206	*		0.0513	*	
38	-0.1637	*		0.0677	*		0.1608	*	

39	0.0367		*		0.0240		*		0.0495		*	

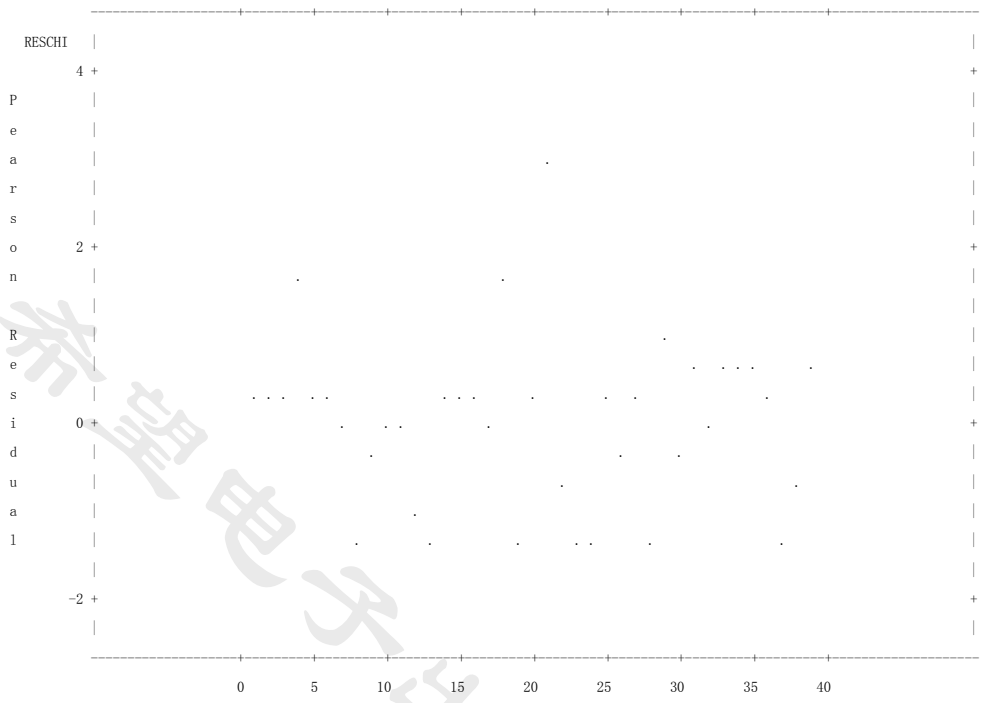
Occurrence of Vaso-Constriction

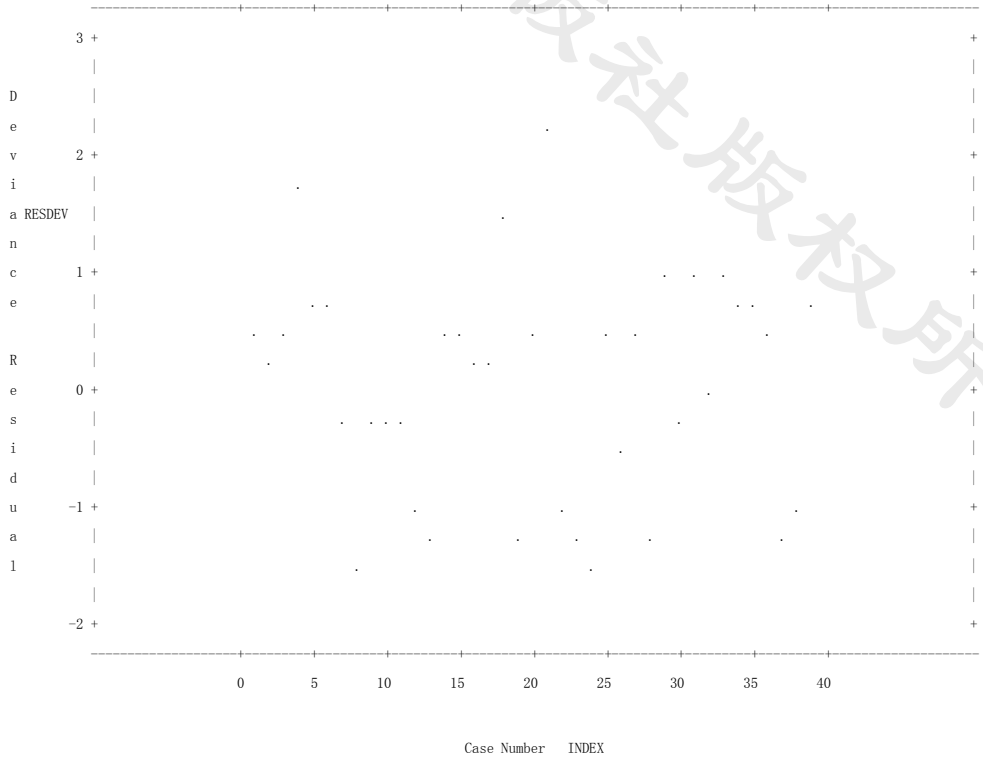
Case	C			CBAR			DIFDEV		
Number	Value	(1 unit = 0.09) 0 2 4 6 8 12 16		Value	(1 unit = 0.08) 0 2 4 6 8 12 16		Value	(1 unit = 0.37) 0 2 4 6 8 12 16	
1	0.0254	*		0.0218	*		0.2702	*	
2	0.00580	*		0.00535	*		0.1286	*	
3	0.00675	*		0.00640	*		0.2259	*	
4	0.3657		*	0.3314		*	3.2044		*
5	0.0230	*		0.0210	*		0.4153	*	
6	0.0353	*		0.0314	*		0.4775	*	
7	0.000753	*		0.00072	*		0.0319	*	
8	0.0738	*		0.0707	*		1.9934		*
9	0.00621	*		0.00566	*		0.1185	*	
10	0.000479	*		0.00046	*		0.0221	*	
11	0.00086	*		0.000818	*		0.0329	*	
12	0.1660	*		0.1419	*		1.3559		*
13	0.2624		*	0.2272		*	2.0353		*
14	0.00555	*		0.00528	*		0.1984	*	
15	0.0197	*		0.0177	*		0.3139	*	
16	0.00373	*		0.00354	*		0.1287	*	
17	0.00102	*		0.000983	*		0.0487	*	
18	0.2674		*	0.2444		*	2.8019		*
19	0.1669	*		0.1507	*		1.9015		*
20	0.00593	*		0.00563	*		0.2052	*	
21	1.4447			*	1.2736			*	5.9734			*
22	0.0453	*		0.0417	*		0.8324	*	
23	0.0903	*		0.0854	*		1.8934		*
24	0.1316	*		0.1243	*		2.3964		*
25	0.00876	*		0.00832	*		0.3037	*	
26	0.0194	*		0.0174	*		0.3002	*	
27	0.0122	*		0.0116	*		0.3605	*	
28	0.0764	*		0.0727	*		1.8642		*
29	0.1185	*		0.1035	*		1.1817	*	
30	0.00747	*		0.00656	*		0.0991	*	
31	0.1700	*		0.1340	*		0.9436	*	
32	2.591E-8	*		2.589E-8	*		0.000068	*	
33	0.0219	*		0.0210	*		0.8137	*	
34	0.0144	*		0.0137	*		0.4857	*	
35	0.0135	*		0.0129	*		0.4983	*	
36	0.0205	*		0.0187	*		0.3688	*	
37	0.0764	*		0.0727	*		1.8642		*
38	0.0625	*		0.0576	*		1.0908	*	
39	0.0195	*		0.0186	*		0.7183	*	

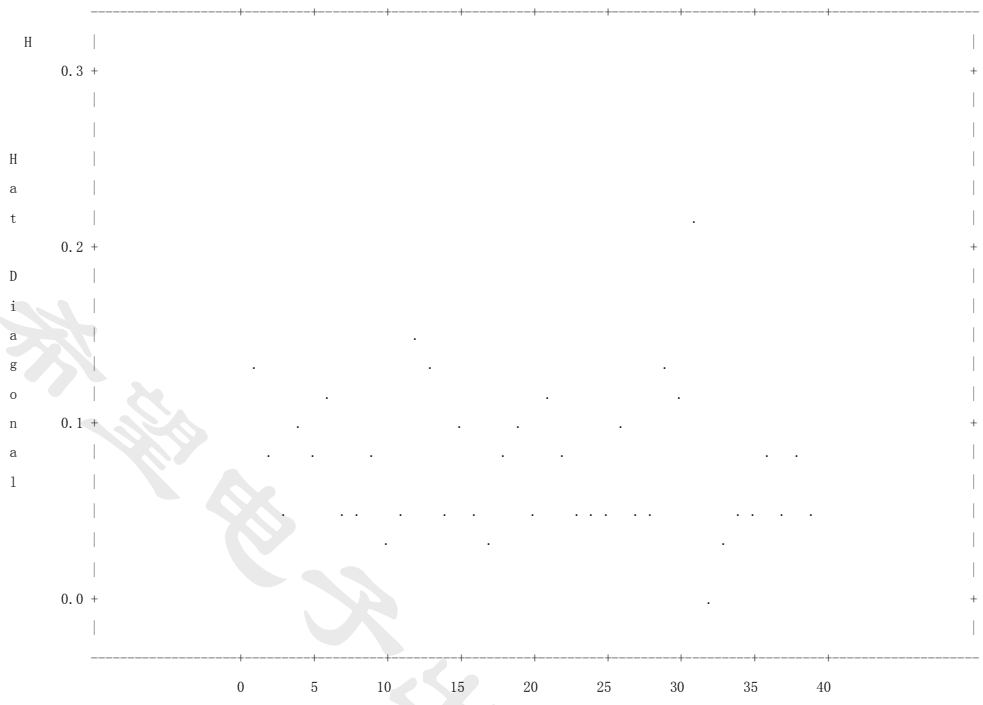
Regression Diagnostics

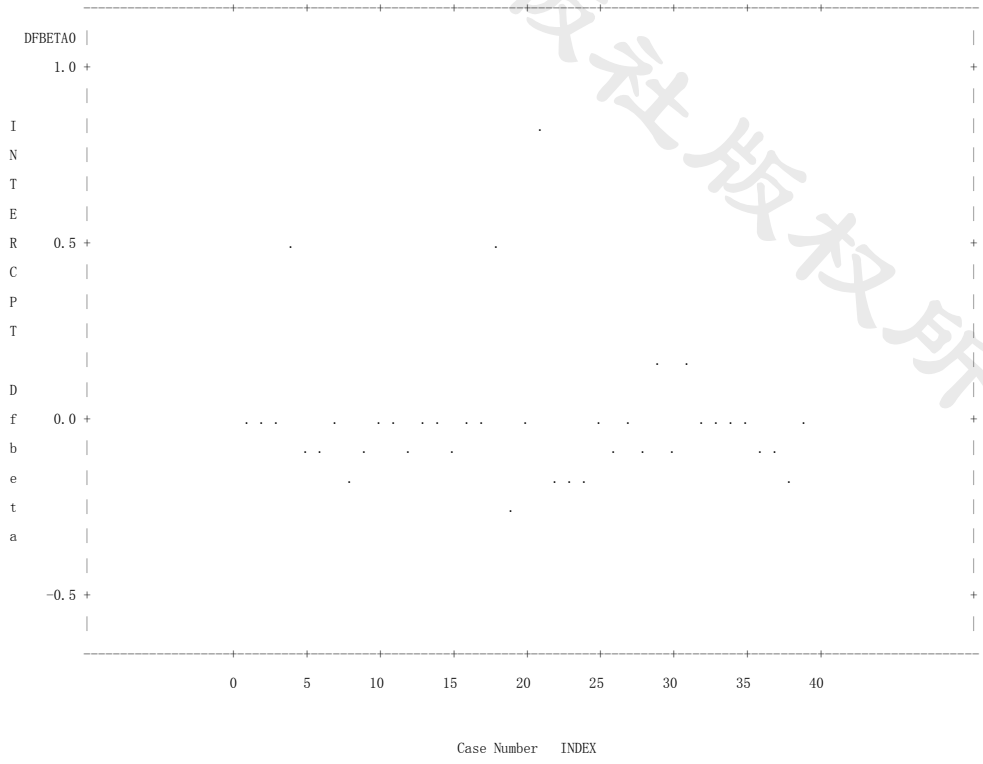
DIFCHISQ

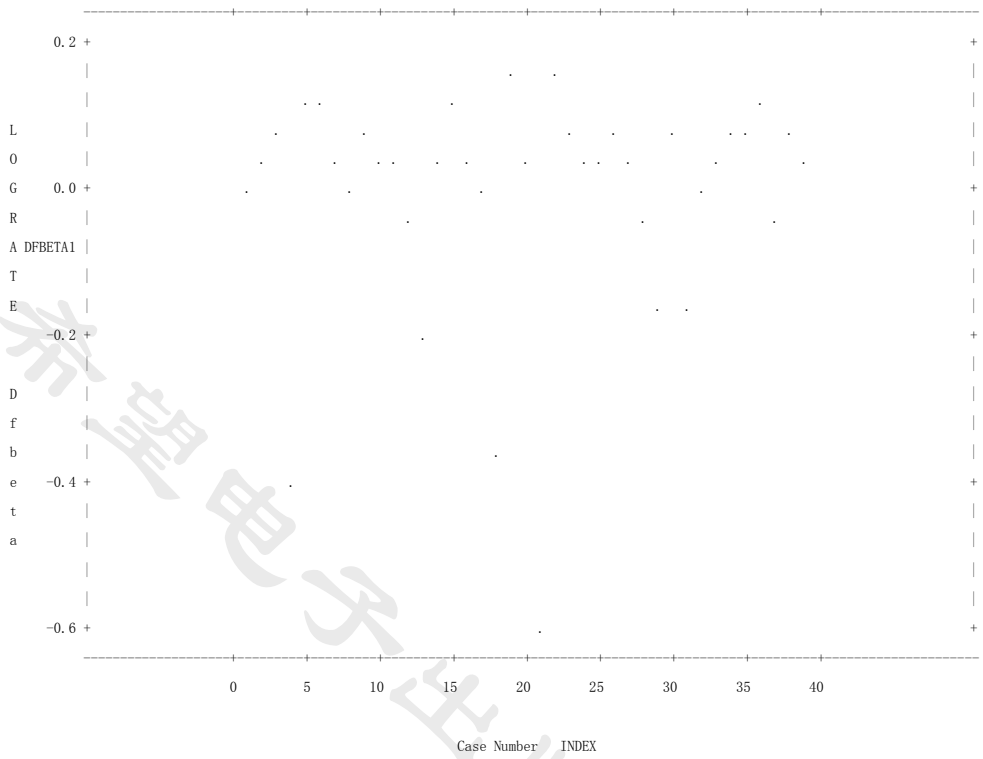
Case Number	Value	(1 unit = 0.67) 0 2 4 6 8 12 16
1	0.1541	|* |
2	0.0689	|* |
3	0.1224	|* |
4	3.5373	| * |
5	0.2389	|* |
6	0.2813	|* |
7	0.0164	|* |
8	1.6859	| * |
9	0.0637	|* |
10	0.0114	|* |
11	0.0170	|* |
12	0.9768	| * |
13	1.6968	| * |
14	0.1066	|* |
15	0.1773	|* |
16	0.0681	|* |
17	0.0251	|* |
18	2.8365	| * |
19	1.5505	| * |
20	0.1106	|* |
21	10.7580	| *|
22	0.5266	| * |
23	1.5549	| * |
24	2.2387	| * |
25	0.1675	|* |
26	0.1693	|* |
27	0.2022	|* |
28	1.5219	| * |
29	0.8180	| * |
30	0.0539	|* |
31	0.6330	| * |
32	0.000034	|* |
33	0.5074	| * |
34	0.2798	|* |
35	0.2876	|* |
36	0.2100	|* |
37	1.5219	| * |
38	0.7339	| * |
39	0.4375	| * |

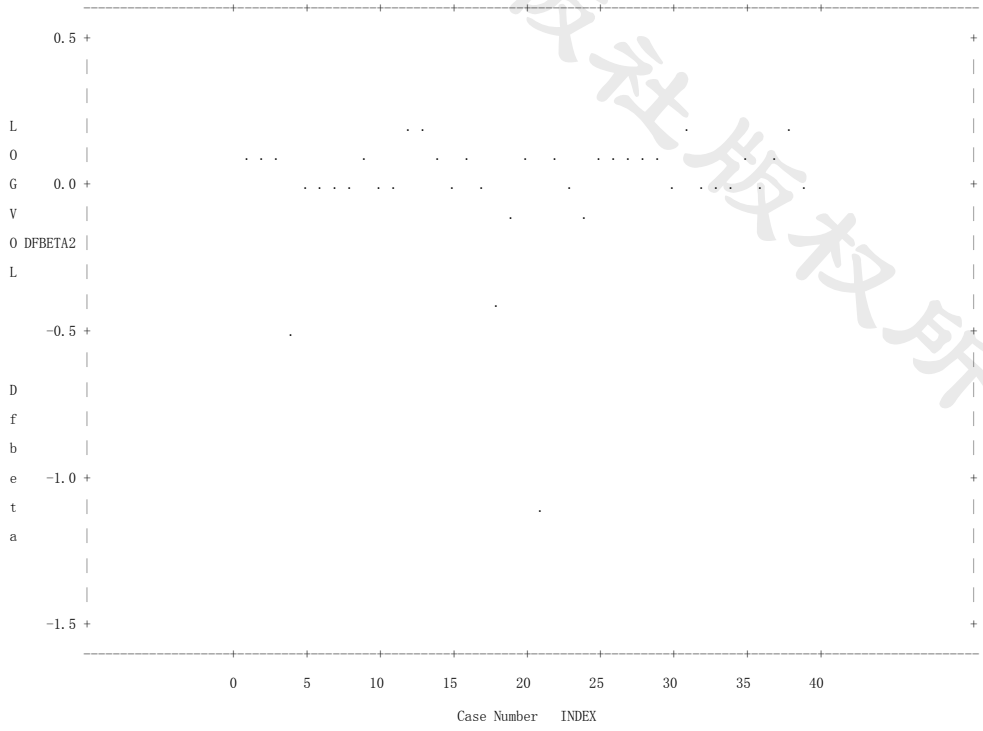


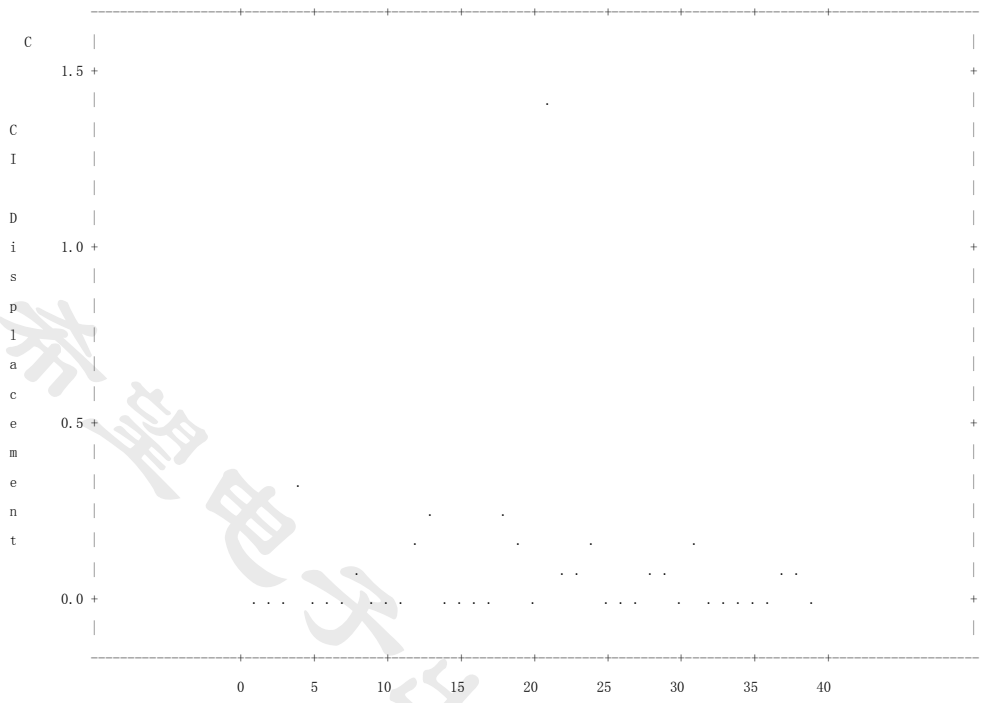


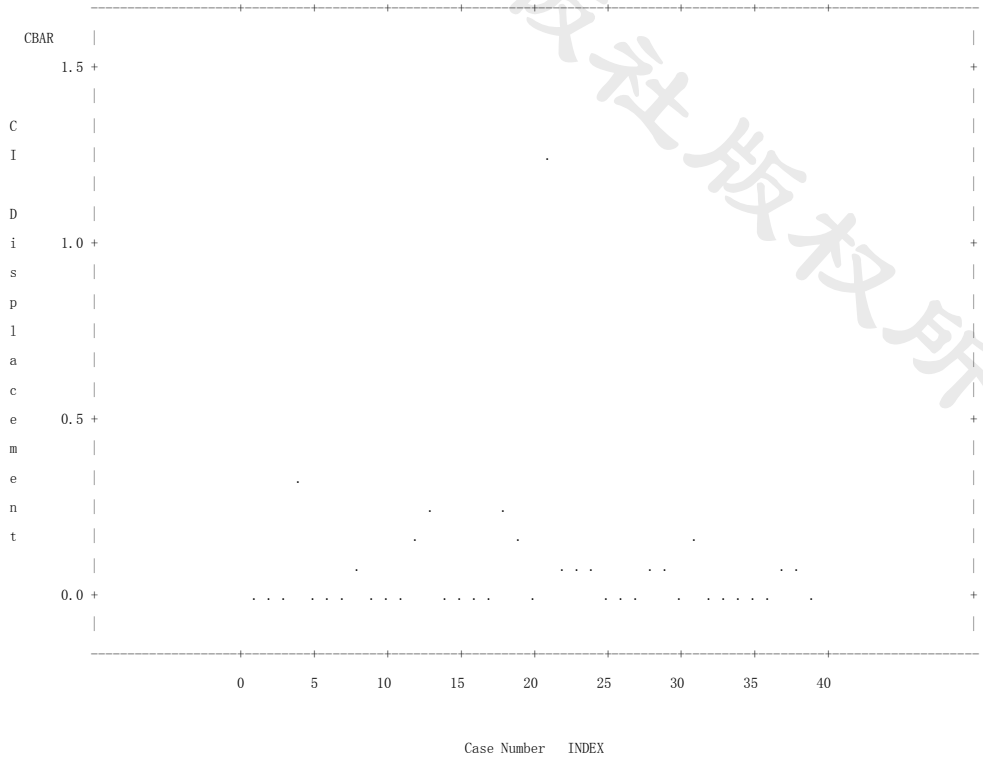


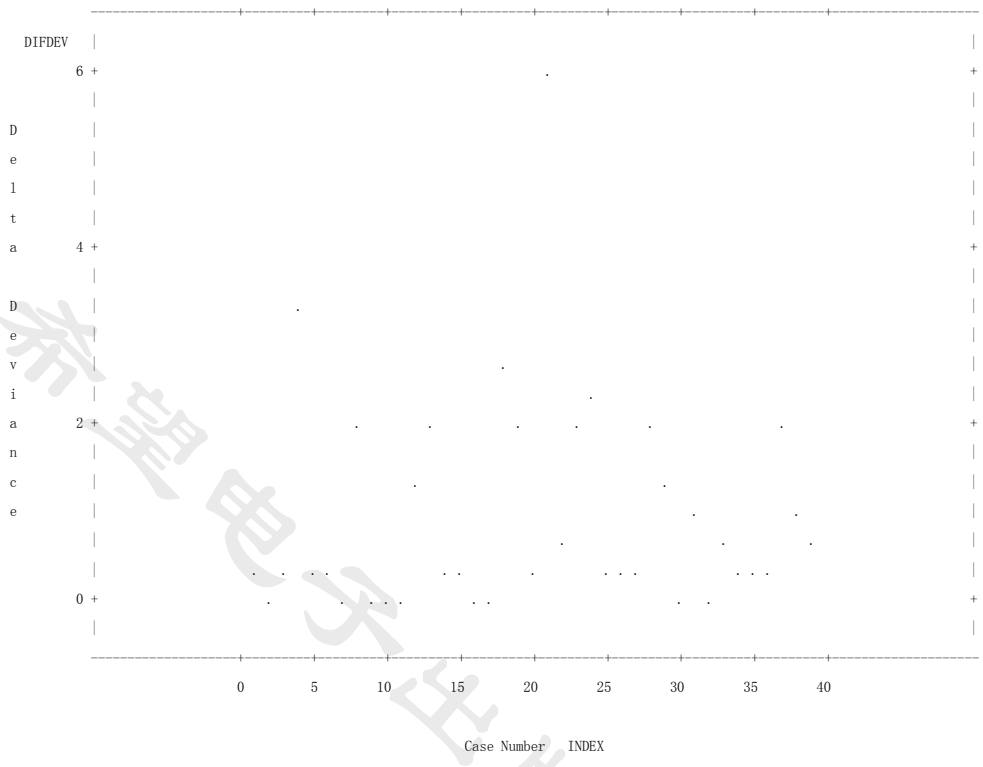


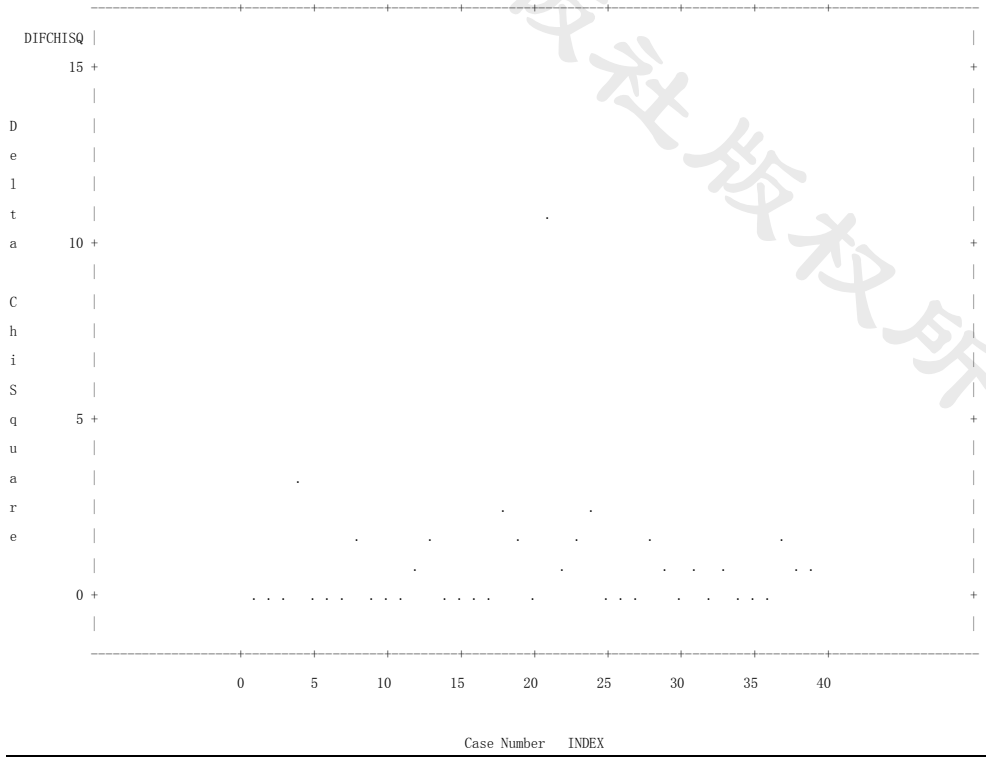












20.6　注 意 事 项

■遗漏数据的处理

观察体只要在反应变量或任何一个自变量上含遗漏数据，则 LOGISTIC 程序会自动将此观察体自分析过程中剔除，此外，若观察体的加权值 (由 WEIGHT 指令界定) 是负值、零、或遗漏值，则此观察体也会从分析过程里排除。

■ 选项 OUTEST= 输出文件的进一步说明

此选项所产生的文件是以参数的估计值为主 (即 TYPE=EST)。若读者同时使用 COVOUT 选项，则参数间的共变异数矩阵亦纳入此输出文件内。此文件所含的所有变量如下：

(1) BY 指令中的变量名称串。

(2) _TYPE_ 变量，其值不是 PARMS (参数估计值) 就是 COV (参数的共变异数)。

(3) 表示截距参数的变量名称，以 INTERCP1，...，INTERCP9，INTERC10，INTERC11，...等命名。请读者特别注意第 9 个与第 10 个截距参数命名原则的更动。

(4) 表示斜率参数的变量名称，以原自变量名称命名。若某个自变量未被包括在模型内，则其对应的参数估计值为遗漏数据。

(5) _LINK_，代表反应变量的量化单位。只有三种可能的文字值：LOGIT (对数奇数比)、NORMIT (常态数) 或 CLOGLOG (双对数)。

(6) _NAME_，是一个文字变量，其值不外乎是 ESTIMATE (参数估计值)、INTERCEP (截距)及各自变量的名称。 有关 OUTEST= 文件的内容，读者可回头仔细研究报表 20.2 后面的部分。

■ 相关统计量的定义

在本节我们仅就几个重要的相关统计量作简单的说明。这些统计量背后的理论基础，则请读者参考第 20.1 节所提的几篇文献。

如何利用最大可能率法将观察体所属的组别预测出来?

根据第 20.4 节指令 #2 的说明，一个反应变量的量化单位有三，即对数奇数比 (Logit)、常态数 (Normit) 以及双对数 (Log-Log)。这三种分数所对应的累积常态分配 F(x) 均不同，现列表如下页：

量化单位	F(x) 的形式
logit	$F(x)=1/(1+\exp(-x))$
normit	$F(x)=\Phi(x)=(2\pi)^{-1/2}\int_{-\infty}^{x}\exp(-z^2/2)dz.$
log-log	$F(x)=1-\exp(-\exp(x))$.

在上述 F(x) 的表达式里，x 其实代表一组自变量的加权组合，也就是自变量的值乘

以其对应的斜率参数再加上一个截距参数。这个加权组合的一般写法是

$$\alpha_i + \beta'X \qquad 1 \leqq i=\text{组别数} \leqq k$$

利用循环加权最小误差平方法 (也是 IRLS 解法)，LOGISTIC 程序可循序渐进地推算出 α_i 以及 β 向量的值。根据这些参数估计值的大小，以及观察体 (j) 在自变量上的分数 (以 x_j 表示)，我们可以求得该观察体隶属于各组别的率，以 $Prob(Y_j=I|x_j)$ 表示。这个概率的计算方法如下：

$$Prob(Y_j=i|x_j) = \begin{cases} F(\alpha_1 + \beta'x_j) & i=1 \text{ (第一组)} \\ F(\alpha_i + \beta'x_j) - F(\alpha_{i-1} + \beta'x_j) & 1< i \leqq k \text{ (中间的组别)} \\ 1-F(\alpha_k + \beta'x_j) & i=k+1 \text{ (最后一组)} \end{cases}$$

将这些率值算出来之后，LOGISTIC 程序会将此观察体分派到概率值最大的那一组去。

回归模型的适合度

LOGISTIC 程序根据 三个指标来鉴定回归模型的优劣：

(1) -2 log likelihood

这个指标的定义如下：

$$-2LogL=-2\sum_j w_j \log(p_j)$$

在此，w_j 是第 j 个观察体的加权值，

p_j 就是上节所定义的 $Prob(Y_j=ix_j)$ 之最大值。

(2) 赤池资讯量指标 (Akaike Information Criterion，又作 AIC)

$$AIC=-2\ LogL + 2(k+s)$$

在此，L 的定义如 (1) 所示，k 代表反应变量之组别数减 1，s 代表自变量的个数。

(3) 萧氏指标 (Schwartz Criterion，又作 SC)

$$SC=-2\ LogL + (k+s)\log(N)$$

在此，L， k， s 的定义如 (1)、(2) 所示，N 则代表样本数的大小。

上述的三个指标都是对数可能率 (Log Likelihood) 的函数。其中，(1) 的统计量可用 χ^2 的抽样分配来鉴定其值是否达到统计显着的程度 (报表上以 p 值来表示)。其余两个统计量则是对数可能率的衍生值——分别对自变量数目或样本数作矫正。AIC 与 SC 最主要的功能是比较各个模型的劣；愈是优秀的模型，其所对应的 AIC 与 SC 值都 (相对地) 愈小。

SCORE 统计量与统计检验

另一个与对数可能率息息相关的统计量称为 SCORE，其取样分配也是 χ^2，自由度与上节定义的 (1)-2 LogL 相同。SCORE 的表示式如下：

$$U'(\gamma_0)I^{-1}(\gamma_0)U(\gamma_0)$$

在此，γ_0 是参数向量 γ 的定值向量 (亦即当 $\gamma=\gamma_0$)，$U(\gamma_0)$是对数可能率对参数 γ_0，的偏微分向量；$I(\gamma_0)$ 则是对数可能率对参数 γ_0 的第二偏微分的负矩阵，$I^{-1}(\gamma_0)$ 是 $I(\gamma_0)$ 的反矩阵。

残差的卡平方值 (Residual Chi-square)

这个统计量只有当读者选用 SELECTION=FORWARD， BACKWARD， 或 STEPWISE 等方法建立回归模型时才打印出来。它的定义牵涉到饱和的回归模型以及简

化的回归模型之间参数向量的不同。

假设，一个含 s 个自变量及 (k+1) 反应变量的饱和模型，其参数的向量如下所示：

$\gamma=(\alpha_1, ..., \alpha_k, \beta_1, ..., \beta_s)'$

假如简化的模型从 s 个自变量中只选了 t 个 (t<s)，则残差的卡平方值就是上述 γ 向量定值在 γ_0 向量上的 SCORE 统计量。γ_0 的定义如下：

$\gamma_0=(\alpha_1, ..., \alpha_k, \beta_1, ... \beta_t, 0, ..., 0)'$

在此，$\alpha_1, ..., \alpha_k$ 是根据简化模型所求得之截距的估计值，$\beta_1, ..., \beta_t$ 是斜率的估计值。

残差的卡平方值之自由度= s-t。

四个等级相关系数

LOGISTIC 程序之报表上所打印的四个等级相关系数都是由两个统计量导出，即 nc 与 nd。nc 表分类一致的观察体对的个数 (No. of Concordant Pairs)，nd 表分类不一致的观察体对的个数 (No.of Discordant Pairs)。所谓分类一致是指一对观察体若根据原数据是分属两个不同的组别，则逻辑斯谛的分析亦将其正确地分派到不同的组内，而且组别的顺序与原数据相同。如此，这一对观察体被称作协调对 (Concordant Pair)。否则，这一对观察体被称作不协调对(Discordant Pair)。

由以上的定义看来，凡是模型良好的分析结果，nc 的值应远远地超过 nd 的值。以下四种相关系数的原理就是根据这一个简单的理念而来：

c=(nc+0.5(t-nc-nd))/t　　(C 系数)

Somers' D=(nc-nd)/t　　(素摩尔系数)

Goodman-Kruskal Gamma=(nc-nd)/(nc+nd)　　(甘玛系数)

Kendall's Tau-a=(nc-nd)/(0.5N(N-1))　　(陶系数)

在此，N= 样本数大小，t= 不能断定是协调对或不协调对的配对个数。

二分的分类列联表

当反应变量的组别只有两个时，读者可指定 CTABLE 的选项。这个选项包含在 MODEL 指令内 (参见第 20.4 节指令 #2)，其作用是将观察体已知的组别与分析后所属的组别作比较。若这两者完全一致，则模型是有效的。否则，模型有待改进。有关这个报表的形式与解释，请参见第 20.5 节例三报表的解释。

影响力统计量

这一节所介绍的统计量，都由选项 INFLUENCE 的界定而来。它们的功能都在于找出样本中具极端 (过度) 影响力的观察体。有时候，这一类的观察体在测量时有误差，因此与其它数据格格不入，造成它们对分析结果有不成比例的影响力。也有时候，测量的数字没有问题，只因这些观察体在本质上与其它的数据迥异，是所谓的劣质数据 (Outliers)。研究者可考虑将它们自分析中剔除。

不论原因为何，影响力的分析结果值得读者分神去探讨。在介绍各统计量之前，首先介绍一些通用的符号：

j　　观察体的识别代号。

W_j　　第 j 个观察体的加权值 (通常 =1)。

R_j　　在 n_j 次尝试中，第 j 个观察体成功的次数 (r_j)。

若二分反应变量下的组别以数值 1，2 代表，并且组别=1，则 r_j=1。反之，若组别=2，则 r_j=0。

P_j　　第 j 个观察体隶属于第一组的预测概率。

B　　模型中参数的估计值。

V_b　　参数间的共变异数矩阵。

B_j　　将第 j 个观察体剔除后，参数的新估计值。

以下介绍各影响力的指标：

Hat 矩阵的对角线元素 （h_{jj}）

定义如下：

$h_{jj}=w_j n_j p_j(1-p_j)(1，x_j')V_b(1，x_j')'$

在此，p_j 是预测概率的估计值。h_{jj} 愈大时，表示其对应的第 j 个观察体之影响力愈强。

皮尔森残差（Pearson Residual）与偏激残差（Deviance Residual）

这两个统计量都是用来鉴别样本中与模型不甚符合的观察体。皮尔森残差的定义如下：

$$x_i=\frac{(r_j-n_j p_j)\sqrt{w_j}}{\sqrt{n_j p_j(1-p_j)}}$$

偏激残差的定义如下：

$$d_i=\begin{cases}-\sqrt{-2w_j n_j\log(1-p_j)} & \text{如果 } r_i=0\\ \pm\sqrt{2w_j(r_j\log(r_j/(n_j p_j))+((n_j-r_j)\log(n_j-r_j)/(n_j(1-p_j))))} & \text{如果 } 0<r_j<n_j\\ \sqrt{-2w_j n_j\log(p_j)} & \text{如果 } r_j=n_j\end{cases}$$

上式 ± 的取法是这样的：

如果 r_j/n_j 大于 p_j，则取 + 号，

如果 r_j/n_j 小于 p_j，则取 - 号。

C 系数与 CBAR 值

C 与 CBAR 都代表信赖区间的错位诊断量，其功能在于计算第 j 个观察体对参数估计值的影响力，计算公式如下：

C 系数　　$C_j=X^2_j h_{jj}/(1-h_{jj})^2$

CBAR 值　　$C_j=X^2_j h_{jj}/(1-h_{jj})$

在此，x^2_j 是皮尔森残差的平方。

DIFDEV 与 DIFCHISQ

这两个统计量也是用来区分数据中的劣质观察体。此处的劣质观察体特指那些容易导致原始数据与预测概率之间不协调的观察个体。在这个理念下，LOGISTIC 程序将 DIFDEV 及 DIFCHISQ 定义为：

DIFDEV=$\triangle_j D=d_j^2+C_j$ (若将第 j 个观察体剔除，其对偏激统计量所造成的改变)。

DIFCHISQ=$\triangle_j x^2=C_j/h_{jj}$ (若将第 j 个观察体剔除，其对皮尔森 x^2 适合度的影响)。

若这两个数值愈大，则相对应的影响力也愈大。

第 21 章　正交回归分析：统计程序 PROC ORTHOREG

21.1　PROC ORTHOREG 程序的简介

这个程序最适用于参数 (估计值) 的标准误差较大的数据。在这种情况下，一般的线性模型解(如 REG 与 GLM 程序分析的结果) 只能算是最小误差平方解 (LS) 的趋近值，而非真正的 LS 解。 因此，本程序采用 Gentleman 及 Givens 两位统计学家在 1970 年代文献中所提出的转换公式，将分析过程中所产生的 R 矩阵不断地修正并重新估计，由此可避免参数值的估计不准确。R 矩阵的产生是根据输入资料矩阵的 QR 分解而来，其本身是一个右上角的正方矩阵。有关这个分析方法的论文，可参见 Gentleman 在 1972 年提出的两篇文章。

21.2　如何撰写 PROC ORTHOREG 程序

PROC ORTHOREG 含四道指令，其格式如下：

PROC ORTHOREG	选项串;
MODEL	因变量=自变量名称串 / 选项;
WEIGHT	变量名称;
BY	变量名称串;

PROC ORTHOREG 与 MODEL 指令是必须的， 不可省略的。 每一个 ORTHOREG 程序中只可含一个 MODEL 指令。

指令 #1　PROC ORTHOREG 选项串;

有四个选项：

(1) NOPRINT
不印出任何分析的结果。

(2) SINGULAR= 正实数
这个选项的功能在于鉴定上节所提到的 R 矩阵是否为一个满秩的矩阵，其内设值是 10 的 -12 次方。

(3) DATA= 输入文件名称
指明到底对那一个 SAS 文件执行分析。若省略此选项，则 SAS 会自动找出在此程序之前最后形成的 SAS 文件，并对它执行分析。
输入文件的数据形式不可以是 TYPE=CORR，COV 或 SSCP 等矩阵。

(4) OUTEST= 输出文件名称
此文件专门用来储存分析的结果，如参数估计值，分组变量值 (由指令 #4 来界

定) 及特殊变量如 _TYPE_ (其值均等于 PARMS，代表参数估计值)、_NAME_ (其值是空白的)、_RMSE_ (其值等于误差的均方根) 以及 INTERCEP (其值就是模型中的截距)。有关这个输出文件的进一步说明，请参见本章第 21.4 节。

指令 #2　MODEL 因变量＝自变量名称串 / 选项；

这个指令的功能在定义一个回归分析的模型，所有自变量的值必须先在 DATA 程序内产生。因此，交互作用如 X1*X2 不可列在 MODEL 等号的右边。删除号 (/) 后的选项只有一个，说明如下：

(1)　NOINT

规定回归模型中不含截距。

指令 #3　WEIGHT 变量名称；

这个指令所界定的变量值代表观察体在回归分析中的加权值。若加权值就是各观察体变异数的倒数，则参数的加权估计值将会是所谓的 "最佳线性不偏估计值 (B.L.U.E)"。

指令 #4　BY 变量名称串；

ORTHOREG 程序依此指令所列举的变量值将文件分成几个小的文件，然后对每一个小文件分别执行分析。当读者选用此指令时，文件内的数据必须先按照 BY 变量串的值做由小到大的重新排列，这个步骤可藉 PROC SORT 达成。或者，文件可先经由 PROC DATASETS 处理，将分组的代号附加在每一个观察体旁。如此，读者可直接使用这个分组代号来撰写 BY 的指令。有关这种处理法的详细介绍，可参考附录 C.9 节。

21.3　范　　例

例一：朗氏数据的回归分析

本文件 (LONGLEY) 撷取自朗氏 1960 年代的一篇论文 (Longley，1967)。经由 ORTHOREG 与 GLM 程序分析后，理论上应该证实 GLM 程序提供的最小误差平方的解只能算是趋近值，因为该程序认为数据的矩阵达到奇异性 (Singularity)；然而，当 ORTHOREG 程序分析时，却能很轻易地将这个奇异性的问题避开而获得一个精准的 LS 解。因此，在这种情况下，ORTHOREG 程序分析的结果应该比较好。

程　序

```
DATA LONGLEY;
    INPUT Y X1 X2 X3 X4 X5 X6;
    CARDS;
60323  83.0 234289 2356 1590 107608 1947
61122  88.5 259426 2325 1456 108632 1948
60171  88.2 258054 3682 1616 109773 1949
61187  89.5 284599 3351 1650 110929 1950
63221  96.2 328975 2099 3099 112075 1951
```

```
63639  98.1 346999 1932 3594 113270 1952
64989  99.0 365385 1870 3547 115094 1953
63761 100.0 363112 3578 3350 116219 1954
66019 101.2 397469 2904 3048 117388 1955
67857 104.6 419180 2822 2857 118734 1956
68169 108.4 442769 2936 2798 120445 1957
66513 110.8 444546 4681 2637 121950 1958
68655 112.6 482704 3813 2552 123366 1959
69564 114.2 502601 3931 2514 125368 1960
69331 115.7 518173 4806 2572 127852 1961
70551 116.9 554894 4007 2827 130081 1962
;
PROC ORTHOREG DATA=LONGLEY OUTEST=LONGOUT;
    MODEL Y=X1-X6;
RUN;
PROC PRINT DATA=LONGOUT;
    FORMAT _NUMERIC_ 20.14;
RUN;
PROC GLM DATA=LONGLEY;
    MODEL Y=X1-X6;
RUN;
```

结　果

上述的程序经过 IBM-PC 上的 6.04 版以及 VAX 8*** 主机上的 6.08 版分析后，并未显出参数估计的不适性。这个结果与英文版使用手册所刊登的报表内容不一致；根据原使用手册第 1215-1216 页的分析结果，截距，X2， X3， X5， X6 等回归系数的估计值旁边有 "B" 的记号，表示这些估计值不够准确。原文版的分析是在 IBM 主机上执行。

我们猜想这两种相异的分析结果可能源于不同机型的设计以及其内设的数值精确度所致。因此，读者若在 IBM 主机上运行此程序，可能会得到一个近乎原使用手册的结果。[不过若将 X6 自 MODEL 指令中去掉，则 X2， X3 与 X5 的参数估计值的旁边就出现 "B" 的记号，与原文使用手册的结果完全一致。]

报表 21.1　朗氏数据的回归分析

ORTHOREG Regression Procedure

Dependent Variable Y

Sum of Squared Errors	836424.05551
Degrees of Freedom	9
Mean Squared Error	92936.006167
Root Mean Sqr Error	304.85407356
R-square	0.9954790046

Variable	DF	Parameter Estimate	Std Error	T-Ratio	Prob>|t|
INTERCEP	1	-3482258.63459581	890420.38361	-3.91	0.0036
X1	1	15.0618722713731	84.914925775	0.18	0.8631

X2	1	-0.03581917929259	0.0334910078	-1.07	0.3127
X3	1	-2.02022980381682	0.4883996817	-4.14	0.0025
X4	1	-1.03322686717359	0.2142741632	-4.82	0.0009
X5	1	-0.05110410565358	0.2260732001	-0.23	0.8262
X6	1	1829.15146461354	455.47849914	4.02	0.0030

OBS	_TYPE_	_NAME_	_RMSE_	INTERCEP	X1	X2
1	PARMS		304.85407356196400	-3482258.63459581000	1.50618722713733	-0.03581917929259

OBS	X3	X4	X5	X6
1	-2.02022980381682	-1.03322686717359	-0.05110410565358	1829.15146461355000

General Linear Models Procedure

Number of observations in data set = 16

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	184172401.94449400	30695400.32408230	330.29	0.0001
Error	9	836424.05550590	92936.00616732		
Corrected Total	15	185008826.00000000			

R-Square	C.V.	Root MSE	Y Mean
0.995479	0.466730	304.85407356	65317.00000000

`L1 `SC48 `SS0.01 `SD0.4 `GEPBE48

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X1	1	174397449.77912700	174397449.77912700	1876.53	0.0001
X2	1	4787181.04444969	4787181.04444969	51.51	0.0001
X3	1	2263971.10981840	2263971.10981840	24.36	0.0008
X4	1	876397.16186109	876397.16186109	9.43	0.0133
X5	1	348589.39964975	348589.39964975	3.75	0.0848
X6	1	1498813.44958735	1498813.44958735	16.13	0.0030

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X1	1	2923.97636102	2923.97636102	0.03	0.8631
X2	1	106306.25889540	106306.25889540	1.14	0.3127
X3	1	1590137.97172242	1590137.97172242	17.11	0.0025
X4	1	2160905.48176653	2160905.48176653	23.25	0.0009
X5	1	4748.94813184	4748.94813184	0.05	0.8262
X6	1	1498813.44958735	1498813.44958735	16.13	0.0030

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	-3482258.635	-3.91	0.0036	890420.3836
X1	1.506	0.18	0.8631	8.4915
X2	-0.036	-1.07	0.3127	0.0335
X3	-2.020	-4.14	0.0025	0.4884
X4	-1.033	-4.82	0.0009	0.2143
X5	-0.051	-0.23	0.8262	0.2261
X6	1829.151	4.02	0.0030	455.4785

例二：文氏数据的回归分析

本文件 (WAMPLER) 取自文氏 1970 年的论文 (Wampler，1970) ，其目的在于自创三组数据，其参数的值是已知的。然后，分别以 PROC ORTHOREG 及 PROC GLM 来分析这三组数据。PROCRTHOREG 的结果见报表 21.2a，GLM 分析的结果则见报表 21.2b。

三组资料的参数值如下所示：

因变量	自变项的回归系数						
	X1	X2	X3	X4	X5	截距	误差的平均方
Y1	1	1	1	1	1	-9999	0
Y2	0.1	0.01	0.001	0.0001	0.00001	-999	0
Y3	1	1	1	1	1	-9999	>0

程　序

```
DATA WAMPLER;
    DO X=0 TO 20;
       INPUT D @@;
       X1=X; X2=X*X; X3=X2*X; X4=X2*X2; X5=X3*X2;
       X1=X1+10000; X2=X2+10000; X3=X3+10000;
       X4=X4+10000; X5=X5+10000;
       Y1=1+X+X2+X3+X4+X5;
       Y2=1+.1*X+.01*X2+.001*X3+.0001*X4+.00001*X5;
       Y3=Y1+D; Y4=Y1+100*D; Y5=Y1+10000*D;
       OUTPUT;
    END;
    CARDS;
```

```
759 -2048 2048 -2048 2523 -2048 2048 -2048 1838 -2048 2048
-2048 1838 -2048 2048 -2048 2523 -2048 2048 -2048 759
;
PROC ORTHOREG DATA=WAMPLER;
    MODEL Y1=X1 X2 X3 X4 X5;
RUN;
PROC ORTHOREG DATA=WAMPLER;
    MODEL Y2=X1 X2 X3 X4 X5;
RUN;
PROC ORTHOREG DATA=WAMPLER;
    MODEL Y3=X1 X2 X3 X4 X5;
RUN;
PROC GLM DATA=WAMPLER;
    MODEL Y1-Y3=X1-X5;
RUN;
```

结　果

对这一组数据而言，ORTHOREG 与 GLM 程序分析的结果一样好，均与预知的参数估计值十分接近。从这个例子以及上例的分析结果，我们知道 ORTHOREG 程序可适用的数据范围较 GLM 程序更广。特别是当数据的抽样误差过大，或最小误差平方的解无法以传统方式得到时，ORTHOREG 程序会更有用。

报表 21.2a　文氏数据的 ORTHOREG 分析结果

ORTHOREG Regression Procedure

Dependent Variable Y1

Sum of Squared Errors	0
Degrees of Freedom	15
Mean Squared Error	0
Root Mean Sqr Error	0
R-square	1

Variable	DF	Parameter Estimate	Std Error	T-Ratio	Prob>|t|
INTERCEP	1	-9999.0000001	0	9999.99	0.0001
X1	1	1.00000000001227	0	9999.99	0.0001
X2	1	0.99999999999761	0	9999.99	0.0001
X3	1	1.00000000000011	0	9999.99	0.0001
X4	1	0.99999999999999	0	9999.99	0.0001
X5	1	1	0	9999.99	0.0001

Dependent Variable Y2

Sum of Squared Errors	0
Degrees of Freedom	15
Mean Squared Error	0
Root Mean Sqr Error	0
R-square	1

Variable	DF	Parameter Estimate	Std Error	T-Ratio	Prob>|t|
INTERCEP	1	-998.999999999997	0	9999.99	0.0001
X1	1	0.09999999999999	0	9999.99	0.0001
X2	1	0.01	0	9999.99	0.0001
X3	1	0.00099999999999	0	9999.99	0.0001
X4	1	0.0001	0	9999.99	0.0001
X5	1	9.9999999999999E-6	0	9999.99	0.0001

Dependent Variable Y3

Sum of Squared Errors	83554268
Degrees of Freedom	15
Mean Squared Error	5570284.5333
Root Mean Sqr Error	2360.1450238
R-square	0.999995559

Variable	DF	Parameter Estimate	Std Error	T-Ratio	Prob>|t|
INTERCEP	1	-9999.00000007087	17133638.649	-0.00	0.9995
X1	1	1.00000000000591	2363.5517347	0.00	0.9997
X2	1	1.0000000000017	779.34352433	0.00	0.9990
X3	1	0.99999999999943	101.47550755	0.01	0.9923
X4	1	1.00000000000003	5.6456651217	0.18	0.8618
X5	1	0.99999999999999	0.1123248547	8.90	0.0001

报表 21.2b　文氏数据的 GLM 分析结果

General Linear Models Procedure

Number of observations in data set = 21

Dependent Variable: Y1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	18814317208116.6000000	3762863441623.3300000	99999.99	0.0
Error	15	0.0000000	0.0000000		
Corrected Total	20	18814317208116.6000000			

R-Square	C.V.	Root MSE	Y1 Mean
1.000000	0	0	663960.33333333

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X1	1	12607306605877.6000000	12607306605877.6000000	99999.99	0.0
X2	1	5322302930223.9600000	5322302930223.9600000	99999.99	0.0
X3	1	840541375443.4290000	840541375443.4290000	99999.99	0.0
X4	1	43724801709.9072000	43724801709.9072000	99999.99	0.0
X5	1	441494861.7117720	441494861.7117720	99999.99	0.0

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X1	1	0.99714247	0.99714247	99999.99	0.0
X2	1	9.17099361	9.17099361	99999.99	0.0
X3	1	540.94767534	540.94767534	99999.99	0.0
X4	1	174762.04114806	174762.04114806	99999.99	0.0
X5	1	441494835.15222200	441494835.15222200	99999.99	0.0

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	-9998.993833	-9999.99	0.0	0
X1	0.999999	9999.99	0.0	0
X2	1.000000	9999.99	0.0	0
X3	1.000000	9999.99	0.0	0
X4	1.000000	9999.99	0.0	0
X5	1.000000	9999.99	0.0	0

Dependent Variable: Y2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	6602.91858365	1320.58371673	99999.99	0.0
Error	15	0.00000000	0.00000000		
Corrected Total	20	6602.91858365			

R-Square	C.V.	Root MSE	Y2 Mean
1.000000	0	0	125.88093333

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X1	1	4961.42515101	4961.42515101	99999.99	0.0001

X2	1	1471.78848977	1471.78848977	99999.99	0.0001
X3	1	163.60891771	163.60891771	99999.99	0.0001
X4	1	6.05187567	6.05187567	99999.99	0.0001
X5	1	0.04414949	0.04414949	99999.99	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X1	1	0.00997119	0.00997119	99999.99	0.0001
X2	1	0.00091711	0.00091711	99999.99	0.0001
X3	1	0.00054095	0.00054095	99999.99	0.0001
X4	1	0.00174762	0.00174762	99999.99	0.0001
X5	1	0.04414949	0.04414949	99999.99	0.0001

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	-998.9999932	-9999.99	0.0	0
X1	0.1000000	9999.99	0.0	0
X2	0.0100000	9999.99	0.0	0
X3	0.0010000	9999.99	0.0	0
X4	0.0001000	9999.99	0.0	0
X5	0.0000100	9999.99	0.0	0

Dependent Variable: Y3

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	18814317208116.6000000	3762863441623.3300000	99999.99	0.0001
Error	15	83554268.0002136	5570284.5333476		
Corrected Total	20	18814400762384.6000000			

R-Square	C.V.	Root MSE	Y3 Mean
0.999996	0.355465	2360.14502382	663960.33333333

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X1	1	12607306605877.6000000	12607306605877.6000000	99999.99	0.0001
X2	1	5322302930223.9600000	5322302930223.9600000	99999.99	0.0001
X3	1	840541375443.4290000	840541375443.4290000	99999.99	0.0001
X4	1	43724801709.9072000	43724801709.9072000	7849.65	0.0001
X5	1	441494861.7117720	441494861.7117720	79.26	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X1	1	0.99711762	0.99711762	0.00	0.9997
X2	1	9.17106539	9.17106539	0.00	0.9990
X3	1	540.94721188	540.94721188	0.00	0.9923
X4	1	174762.06137411	174762.06137411	0.03	0.8618
X5	1	441494861.71177100	441494861.71177100	79.26	0.0001

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	-9998.993833	-0.00	0.9995	17133638.53
X1	0.999999	0.00	0.9997	2363.55
X2	1.000000	0.00	0.9990	779.34
X3	1.000000	0.01	0.9923	101.48
X4	1.000000	0.18	0.8618	5.65
X5	1.000000	8.90	0.0001	0.11

21.4　注 意 事 项

■遗漏数据的处理

观察体只要在任何一个变量上含遗漏数据，则 ORTHOREG 程序会自动将此观察体剔除于分析过程之外。

■选项 OUTEST＝输出文件的进一步说明

此选项所导出的文件是一个含参数估计值的文件 (即 TYPE=EST)。此文件所包含的变量如下：

(1) BY 指令中的变量串的值。

(2) 在 MODEL 指令中所提的因变量及自变量串。

(3) 特殊变量 _TYPE_，其值一律是 PARMS。

(4) 特殊变量 _NAME_，其值是空白。

(5) 特殊变量 _RMSE_，其值等于误差的平均方根。

(6) 特殊变量 INTERCEP，其值等于预估的截距。若读者选用 NOINT，则此变量不会出现在此输出文件内。

第 22 章　多项式的回归分析：统计程序 PROC RSREG

22.1　PROC RSREG 程序概述

RSREG 程序的特殊之处在于它可以包括自变量的自乘积 (如 X_1^2，X_1^3) 或相乘积 (如：X_1*X_2)。因此这种回归模型称为多项式的回归模型，或称为反应面分析 (Response Surface Analysis)。

一个含两个自变量 (X_1，X_2) 的多项式回归分析的模型可以写成：

```
PROC RSREG;
     MODEL Y=X1 X2;
```

上述分析的结果可用来解答下列的问题：

1. 到底多项式中的一次式，二次式，或相乘积对因变量 (Y) 的变异数的解释量最大?
2. 这种多项式的模型是否合理?
3. 多项式中哪些项是多余的?
4. 多项式中哪些项的组合是最精简的?
5. 多项式模型的几何表示方法是一个平面，一个抛物线，还是一个马鞍的形状?
6. 到底 Y 的预测值是多少?

诚然，SAS 也有其它的统计程序可以用来执行多项式的回归分析，但它们都不如 RSREG 程序来得精简。比方说一个模型含三个自变量 (X_1，X_2 和 X_3)，若用 RSREG 来执行多项式回归分析，则我们只要写

```
PROC RSREG;
     MODEL Y=X1 X2 X3;
```

但同例若以 GLM 来执行多项式回归分析，则我们就要写：

```
PROC GLM;
     MODEL Y=X1 X1*X1
            X2 X1*X2 X2*X2
            X3 X1*X3 X2*X3 X3*X3;
```

才能达到同样的效果。所以，在执行多项式的回归分析时，只要自变量的数目够多，则我们就需要用 RSREG 程序来简化撰写程序的工作了。

22.2　如何撰写 PROC RSREG 程序

PROC RSREG 含六道指令，它们的格式如下：

```
PROC RSREG   选项串;
    MODEL    因变量名称串=自变量名称串 /选项串;
    RIDGE    选项串;
    WEIGHT   变量名称;
    ID       变量名称串;
    BY       变量名称串;
```

只有 PROC RSREG 指令及 MODEL 指令是必须的，其余则视读者的需要再添加进去。

指令 #1　PROC RSREG 选项串;

这个指令有三个选项：

(1) DATA=输入文件名称

指明到底对那一个 SAS 文件执行多项式的回归分析。 若省略此选项，则 SAS 会自动找出在此程序之前最后形成的 SAS 文件，对它执行分析。

(2) OUT=输出文件名称

这个文件包括下列的统计值代号：

代号 (_TYPE_=)	定义
ACTUAL	原数据值
PREDICT	Y 的预测值
RESIDUAL	原数据值减去预测值
U95M	Y 的平均预测值之 95 % 信赖区间的上限
L95M	Y 的平均预测值之 95 % 信赖区间的下限
U95	个别的 Y 预测值之 95 % 信赖区间的上限
L95	个别的 Y 预测值之 95 % 信赖区间的下限
D	库格氏 (Cook) 的影响力指标
BYOUT	利用 BY 变量串将原文件分成小的文件，如：男、女或老、幼等。然后只根据第一个小文件 (即：男、老) 的数据来估计回归系数，其余小组的分析结果则纳入输出文件内。

(3) NOPRINT

不印出任何分析的结果。

指令 #2　MODEL 因变量名称串=自变量名称串 / 选项串;

删除号 (/) 后的选项有十六个：

(1) LACKFIT

要求对回归模型执行不适合度检定 (Lack-of-Fit Test)。当读者选用此选项时，必须先将文件内的数据按自变量的值做由小到大的排列，如此相等的数据就会被排在一起。

(2) BYOUT

此选项必须与 BY 指令联用，其目的是要求 RSRGE 程序只选用第一个分组（亦即 BY 变量的值最小的那一组） 的数据来进行分析。

(3) COVAR=正整数 (如 4)

这个选项宣告 MODEL 指令中的前几 (4) 个自变量为共变量，所以它们只以一次式的形态进入回归模型中。

[以下八个选项与输出文件内的统计值有关：]

(4) ACTUAL

界定输入文件内的原始数据。

(5) PREDICT

界定因变量的预测值。

(6) RESIDUAL

界定预测误差。

(7) L95

界定因变量预测值 (以观察体为单位) 的 95% 信赖区间的下限。

(8) U95

界定因变量预测值 (以观察体为单位) 的 95% 信赖区间的上限。

(9) L95M

界定因变量预测平均数的 95% 信赖区间的下限。

(10) U95M

界定因变量预测平均数的 95% 信赖区间的上限。

(11) D

界定库格氏的影响力统计值 (亦即 Cook's D)。

[以下三个选项控制报表的打印：]

(12) NOANOVA (或 NOAOV)

抑止变异数分析及参数估计值的打印。

(13) NOOPTIMAL (或 NOOPT)

抑止二项式反应面之典型分析的打印。

(14) NOPRINT

是上述选项 (12) 与 (13) 的联合效果。

[其它的选项：]

(15) NOCODE

要求 RSREG 在执行典型分析或脊梁分析 (Ridge Analysis) 时，采用每一个自变量的原始数值，而非标准化后的变量值。

(16) PRESS

针对每一个因变量，计算并打印预测误差的平方和 (SS)。若读者已经选用了 NOANOVA 或 NOPRINT 选项，则 PRESS 选项无效。

指令 #3　RIDGE 选项串；

这个指令的功用是执行脊梁分析。一般而言，若模型所代表的几何反应面内已经包含了一个最优解 (Optimum)，则我们不必再考虑脊梁分析。然而，当最优解可能位于现有之反应面的范围以外时，则脊梁分析可帮助我们指出寻找最优解的方向。

含六个选项，分述如下：

(1) MINIMUM (或 MIN)

计算最小反应面的脊梁。

(2) MAXIMUM (或 MAX)

计算最大反应面的脊梁。

(3) CENTER=m1，m2，...，mp (脊梁中心点的坐标，p=MODEL 指令中自变量的数目) 或 CENTER=M1 (中心点的第一个坐标值) TO m1+(p-1)(p=MODEL 指令中自变量的数目) 或

CENTER=m1 TO mp BY i (i=相邻两坐标的差额)

这个选项旨在界定脊梁中心点的坐标值。不论那一种写法，坐标的数目应等于 MODEL 指令中自变量的数目；而且，其顺序应由 MODEL 指令上自变量列举的次序相对应。现举几个例子说明：

```
MODEL Y=X1 X2 X3;
RIDGE CENTER=11 12 13;
或 RIDGE CENTER=11 TO 13;
或 RIDGE CENTER=11 TO 13 BY 1;
```

上述程序所界定的中心点就是在三维空间里，对应至 (11，12，13) 坐标的点。

此选项的内设值是每一自变量上的中数。

(4) RADIUS=R_1，R_2，...，R_P (脊梁的半径，p=MODEL 指令中自变量的数目) 或

RADIUS=R_1 TO R_1+(p-1)　(p=MODEL 指令中自变量的数目) 或

RADIUS=R_1 TO R_p BY i　(i= 相邻两半径的差额)

这个选项旨在界定脊梁在几何空间的范围。这个范围的界定是根据中心点的坐标与半径。

因此，上列的 R 值都必须是正实数。不论读者采用何法界定半径，其数目应等于 MODEL 指令中自变量的数目；并且，R 的先后顺序应按照自变量在 MODEL 指令中排列的次序。此选项的内设值是 RADIUS=O TO 1 BY 0.1。

(5) OUTR= 输出文件名称

这个文件含脊梁分析的结果。具体地说，它包括：

●BY 变量上所有的值

●文字变量，_DEPVAR_，其值等于因变量的名称。

●文字变量，_TYPE_，其值不外乎 MIN (见选项 1) 或 MAX (见选项 2)。若读者同时界定选项 (1) 与 (2)，则输出文件内会将 MIN 分析的结果列在 MAX 结果之前。

●数值变量，_RADIUS_，其值等于脊梁的半径。
●与最优解相对应的点之坐标 (亦即各自变量上的变量值)
●数值变量，_PRED_，其值等于最优解之点在因变量上的数值。
●数值变量，_STDERR_，其值就是上述 _PRED_ 值的标准误差。

(6) NOPRINT

要求 RSREG 程序只进行脊梁分析并产生一个含统计值的输出文件，而不在报表上打印任何分析结果。

指令 #4　WEIGHT 变量名称；

此变量的值代表观察体的加权值，所以它们必须是正实数。

指令 #5　ID 变量名称串；

此指令所指的变量表示除一般输出资料外，读者希望在输出文件内包括的名义变量。

指令 #6　BY 变量名称串；

RSREG 程序依据此指令所列举的变量将文件分成几个小的文件，然后对每一个小的文件分别执行多项式的回归分析。当读者选用此指令时，文件内的数据必须先按照 BY 变量串的值做由小到大的重新排列。这个步骤可藉 PROC SORT 达成。

22.3　范　例

例一：三元二次的多项式回归分析

本文件 (A) 的数据由 John (1971) 提供。Schneider 与 Stockett 于 1963 年做了一个实验。这个实验的目的在降低一个化学药品的臭气 (ODOR)。他们检查了三个有关的自变量：温度(X1)，瓦斯与水的比率 (X2)，以及装箱的高度 (X3)。每一个自变量以一次式，二次式，及两两变量的相乘积纳入回归模型中 (见报表 22.1a)。

接下来第二次的分析以绘图为重点。在此，我们固定 X3 值等于 1.77，然后分别以 X1，X2 为自变量，Y 的预测值为因变量纳入输出数据组内。最后以 PROC PLOT 绘出反应面的平面图形(见报表 22.1b)。

程　序

```
DATA A;
     INPUT Y X1-X3 @@;
     LABEL Y='ODOR'
          X1='TEMPERATURE'
          X2='GAS-LIQUID RATIO'
          X3='PACKING HEIGHT';
     CARDS;
66  -1   -1   0    39   1   -1   0     43   -1   1   0    49   1   1   0
```

```
58  -1   0  -1    17   1   0  -1     -5   -1   0   1    -40   1   0   1
65   0   -1  -1     7   0   1  -1     43    0  -1   1    -22   0   1   1
-31  0   0   0    -35   0   0   0     -26    0   0   0
;
PROC SORT ; BY X1-X3;
PROC RSREG;
    MODEL Y=X1-X3/LACKFIT;
RUN;
DATA B;
  *-----THE ACTUAL VALUES-----;
   SET A END=EOF;
   OUTPUT;
  *-----FOLLOWED BY AN  X1*X2 GRID FOR PLOTTING-----;
    IF EOF THEN DO; Y=.; X3=1.77;
     DO X1=-1.5 TO 1.5 BY .1;
        DO X2=-2 TO 2 BY .1;
           OUTPUT;
           END;
        END;
     END;
PROC RSREG DATA=B OUT=C NOPRINT;
    MODEL Y=X1-X3/PREDICT NOPRINT;
DATA D; SET C; IF X3=1.77;
PROC PLOT DATA=D;
    PLOT X1*X2=Y/CONTOUR=6 HPOS=50 VPOS=36;
RUN;
```

结　果

一次式中以 X3 最有预测力 (T=-2.690，P=0.0433)。二次式中以 X1*X1，X2*X2 达 0.05 的统计显著程度，其余的变量则无显著的预测力。

报表 22.1a　三元二次的多项式回归分析

Coding Coefficients for the Independent Variables

Factor	Subtracted off	Divided by
X1	0	1.000000
X2	0	1.000000
X3	0	1.000000

Response Surface for VariableY

Response Mean	15.200000
Root MSE	22.478508
R-Square	0.8820
Coef. Of Variation	147.8849

Regression	Degrees of Freedom	Type I Sum of Squares	R-Square	F-Ratio	Prob>F
Linear	3	7143.250000	0.3337	4.712	0.0641
Quadratic	3	11445	0.5346	7.550	0.0264
Crossproduct	3	293.500000	0.0137	0.194	0.8965
Total Regress	9	18882	0.8820	4.152	0.0657

Residual	Degrees of Freedom	Sumof Squares	Mean Square	F-Ratio	Prob>F
Lack of Fit	3	2485.750000	828.583333	40.750	0.0240
Pure Error	2	40.666667	20.333333		
Total Error	5	2526.416667	505.283333		

Parameter	Degrees of Freedom	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob>|T|	Estimate from Coded Data
INTERCEPT	1	-30.666667	12.977973	-2.363	0.0645	-30.666667
X1	1	-12.125000	7.947353	-1.526	0.1876	-12.125000
X2	1	-17.000000	7.947353	-2.139	0.0854	-17.000000
X3	1	-21.375000	7.947353	-2.690	0.0433	-21.375000
X1*X1	1	32.083333	11.698187	2.743	0.0407	32.083333
X2*X1	1	8.250000	11.239254	0.734	0.4959	8.250000
X2*X2	1	47.833333	11.698187	4.089	0.0095	47.833333
X3*X1	1	1.500000	11.239254	0.133	0.8990	1.500000
X3*X2	1	-1.750000	11.239254	-0.156	0.8824	-1.750000
X3*X3	1	6.083333	11.698187	0.520	0.6252	6.083333

Factor	Degrees of Freedom	Sumof Squares	Mean Square	F-Ratio	Prob>F	
X1	4	5258.016026	1314.504006	2.602	0.1613	TEMPERATURE
X2	4	11045	2761.150641	5.465	0.0454	GAS-LIQUID RATIO
X3	4	3813.016026	953.254006	1.887	0.2510	PACKING HEIGHT

Canonical Analysis of Response Surface
(based on coded data)

Factor	Critical Value Coded	Uncoded	
X1	0.121913	0.121913	TEMPERATURE
X2	0.199575	0.199575	GAS-LIQUID RATIO
X3	1.770525	1.770525	PACKING HEIGHT

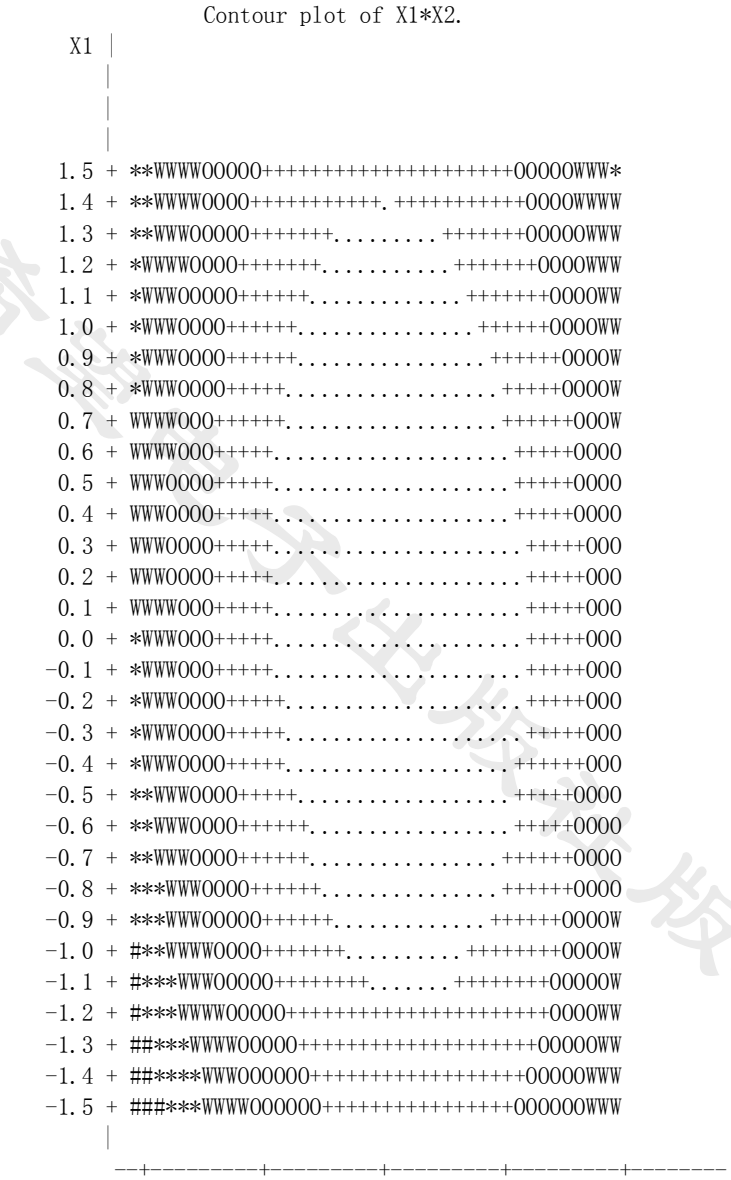
Predicted value at stationary point　　-52.024631

Canonical Analysis of Response Surface
(based on coded data)

Eigenvalues	Eigenvectors X1	X2	X3
48.858807	0.238091	0.971116	-0.015690
31.103461	0.970696	-0.237384	0.037399
6.037732	-0.032594	0.024135	0.999177

Stationary point is a minimum.

报表 22.1b　三元二次的多项式回归分析之几何表示



Symbol	Y	Symbol	Y	Symbol	Y
.....	-60 - 0	00000	60 - 120	*****	180 - 240
+++++	0 - 60	WWWWW	120 - 180	#####	240 - 300

第 23 章　非线性回归分析：统计程序 PROC NLIN

23.1　PROC NLIN 程序概述

本程序的名字 NLIN 来自英文 NonLINear regression。顾名思义，这一套统计分析程序是用来执行非线性的回归分析的。也就是说，因变量与参数之间的关系是二次或二次以上的。下面的模型是一个例子：

$Y=\beta_0(1-e^{-\beta 1X})$　在此，β_0，β_1 为参数。

NLIN 程序依据最小误差平方法或加权最小误差平方法，估计非线性回归模型中的参数值。由于非线性模型较线性模型难处理，所以当非线性模型输入 SAS 时，读者必须同时标明参数的名称，参数的起始值 (Starting Value)。然后利用下列五种搜索法之一，以循环式的搜索过程 (Iterative Procedure) 找出参数的估计值。下面依序介绍这五种估计参数的方法：

1. 改良高斯-牛顿法 (Modified Gauss-Newton Method)
   其对应的选项是 METHOD=GAUSS。
2. 玛克底特法 (Marquardt Method)
   其对应的选项是 METHOD=MARQUARDT。
3. 梯度法 (又称最速下降法，Gradient)
   其对应的选项是 METHOD=GRADIENT。
4. 多元正割法 (又称错位法，Secant)
   其对应的选项是 METHOD=DUD。
5. 牛顿法
   其所对应的选项是 METHOD=NEWTON。

在使用 NLIN 程序时，读者必须提供以下的资料：

●参数名称与其起始值。
●非线性的回归模型。
●模型对每一个参数的微分方程序。

另外，读者也可以选择是否要：

●限制参数估计值的上限与下限。
●改变参数估计值搜索法的收敛指标。
●形成一个含预测值、误差、参数估计值及误差平方和的输出文件。
●自订一个目标函数，以供 NLIN 程序对这个函数求得最佳解。

23.2　如何撰写 PROC NLIN 程序

PROC NLIN 含九道指令，它们的格式如下：

```
PROC NLIN 选项串;
        PARAMETERS (或 PARMS) 参数名称=起始值...;
        BOUNDS 参数的极限;
        MODEL 因变量名称=自变量名称串;
        DER.参数=回归模型的微分方程序;
        DER.参数.参数=回归模型的微分方程序;
        OUTPUT OUT=输出文件名称关键字=变量名称串;
        ID 变量名称串;
        BY 变量名称串;
```

其中，PROC NLIN，PARMS，与 MODEL 三指令是必须的，不可省略。

指令 #1　PROC NLIN 选项串;

本指令的选项可分为四大部分来说明：第一类选项界定输出 / 输入文件，第二类选项界定参数的估计值的打印，第三类选项界定参数估计的搜索法，第四类选项调整统计检定的精确性，现逐项说明如下：

第一类选项　下列两个选项界定文件：

(1) DATA=输入文件名称

指明到底对那一个文件执行非线性的回归分析。若省略此选项，则 SAS 会自动找出在此程序之前最后形成的 SAS 文件，对它执行分析。

(2) OUTEST=输出文件名称

这一个输出文件包含参数的估计值。

第二类选项　下列的选项与参数估计值的打印有关：

(1) BEST=N (正整数)

要求 NLIN 程序只印出由最佳 N 组参数起始值所导出的误差平方和。若省略此选项，则每一组参数起始值所导出的误差平方和都将被印出。

第三类选项　下列十个选项界定参数估计值搜索法的有关事宜：

(1) METHOD=GAUSS 或
METHOD=MARQUARDT 或
METHOD=GRADIENT 或
METHOD=DUD 或
METHOD=NEWTON

指明到底采用那一种搜索法来估计参数的值。若省略此选项，内设值是 METHOD=GAUSS（如果读者采用 DER. 指令），或是 METHOD=DUD （如果读者未采用 DER. 指令）。

(2) NOHALVE

要求搜索法停止分割的步骤。这个选项可用在某些种类的加权的回归分析中，当 SMETHOD=HALVE 时，此选项才有效。

(3) SIGSQ=正实数

要求计算参数标准误差时，以此值取代样本原有的误差均方。这个选项应与参数的最大可能率估计法 (Maximum Likelihood Estimation) 联合使用。

(4) G4

要求在估计参数的过程中，包含 G4 (又称 Moore-Penrose 反比) 的值。

(5) G4SINGULAR

当 Jacobian 矩阵是一个非满秩的矩阵时，要求在估计参数的过程中，包含 G4 (又称 Moore-Penrose 反比) 的值。

(6) TAU=正实数

界定搜索法中搜索区间的初始长度 (Initial Interval Length)，内设值是 1。但当 METHOD=MARQUARDT 时，内设值等于 0.01。

(7) RHO=正实数

界定搜索法中搜索区间的缩短长度 (Subinterval Length)。内设值是 .1。[不过当 METHOD=MARQUARDT 时，PHO=10]。PHO 的值愈小，则搜索的过程愈细致，所花费的时间也愈长。

(8) SMETHOD=HALVE (或 GOLDEN 或 ARMGOLD 或 CUBIC)

这个选项可用在界定循环搜索的过程里，使函数值改进的方法。有四个值可供选择：HALVE [内设值，其理论基础是折半法 (Step-Halving Method)]、GOLD [黄金分割法 (Golden Section Search)]、ARMGOLD [G-A 法，代表 Goldstein-Armijo 法]，以及 CUBIC [三次方内插法 (Cubic Interpolation Method)]。

(9) STEP= 正整数

控制上述折半法的折半次数，内设值等于 20。

(10) SAVE

要求将参数估计值自动收集在 OUTEST= 的输出文件内。

第四类选项　下列四个选项可用来调整统计检定的精确性：

(1) EFORMAT

要求所有数值以科学符号 E 表示。当参数值的差距很大时，这个选项特别有用。

(2) MAXITER=正整数

界定搜索法的最高循环次数，内设值是 50。

(3) CONVERGE(或 CONVERGEOBJ)=收敛指标

收敛指标 (C) 是一个极小的正实数，内设值是 10 的 -8 次方。当下式成立时，我们说循环搜索已达到收敛标准，所以可以停止了：

$(SSE_{i-1} - SSE_i)/(SSE_i + 10^{-6}) < C$

其中，i 表示循环的次数，SSE 表示误差的平方和。

(4) CONVERGEPARM=收敛指标

收指标 (C) 是一个极小的正实数，内设值是 10 的 -8 次方。当下式成立时，我

们说循环搜索已达到收敛标准，所以可以停止了：

$\max_j(|\ \beta_j^{i-1} - \beta_j^{i}\ |/|\ \beta_j^{i-1}\ |) < C$

在此，β_j^{i} 是第 j 个参数在第 i 次循环中估计出来的值。

指令 #2　PARAMETERS（或 PARMS）参数名称=起始值…；

读者可以在同一个 PARMS 指令中界定多个参数的名称与其起始值 。这些参数的名称不可以和输入文件中变量的名称相同。一般而言，一个参数只有一个起始值，但你也可为同一个参数界定多个起始值。以下是各种界定的写法：

参数名称= m　　界定一个起始值

参数名称= m1，m2，...，mn　界定 n 个起始值

参数名称= m TO n　　界定由 m 到 n 的一系列连续整数

参数名称= m TO n BY i　　界定由 m 到 n 的一系列整数，相邻两起始值的差额是 i。

参数名称= m1，m2 TO m3 上述语法的混合型态

下面例子里 ，我们用 PARMS 指令一口气界定了五个参数名称，及它们的起始值：

```
PARMS B0=0
      B1=4 TO 8
      B2=0 TO .6 BY .2
      B3=1，10，100
      B4=0，.5，1 TO 4;
```

这一个 PARMS 指令代表了下面的启动值：

参数据	起始值
B0	0
B1	4 5 6 7 8
B2	0 .2 .4 .6
B3	1 10 100
B4	0 .5 1 2 3 4

从这些起始值里，NLIN 程序算出 1×5×4×3×6=360 组可能的排列组合。若读者没有选用 BEST= 选项则 NLIN 程序会自动印出 360 组起始值的误差平方和。

指令 #3　BOUNDS 参数的极限；

首先，请读者注意：在你所限制的极限下所求得的参数估计值并不一定是最好的。这是读者必须自行负责的地方。

参数极限的设立包括参数名称，(不) 等号，常数值。下面是一个例子：

```
BOUNDS A<=20，0<=B<=10，20>C;
```

由上式可知，在一个 BOUNDS 指令中可同时设立多个参数的极限，其间以逗号分隔。参数极限可以同时含上限与下限或是只含上或下限。

若极限的设立与两个或两个以上的参数有关，如：A+B<1; 则读者必须重新安排模型的写法，使 A+B 成为一个一体的参数而不是两个分开的参数。

指令 #4　MODEL 因变量名称＝自变量名称串；

这个指令与前面几章介绍的回归分析程序中的 MODEL 指令语法完全相同。唯一比较特殊的地方是在 NLIN 程序中读者可将回归模型以下列的方法表示：

MODEL.Y=自变量名称串；

如此，MODEL.Y 将自动成为 Y 的预测值名称。

指令 #5　DER. 参数＝回归模型的微分方程序；

假设有一个非线性的回归模型如下：$Y=\beta_0(1-e^{-\beta 1X})$；则本指令要求 NLIN 程序将此方程序对参数 β_0，β_1 分别求得一次微分。结果如下所示：

```
PROC NLIN;
    PARMS B0=0 TO 10
         B1= .01 TO .09 BY .005;
    MODEL Y=B0*(1-EXP(-B1*X));
    DER.B0=1-EXP(-B1*X);
    DER.B1=B0*X*EXP(-B1*X);
```

若想节省电脑时间，则最后的三个指令可更改为：

```
TEMP=EXP(-B1*X);
MODEL Y=B0*(1-TEMP);
DER.B0=1-TEMP;
DER.B1=B0*X*TEMP;
```

指令 #6　DER. 参数. 参数=回归模型的微分方程序；

这个指令要求 NLIN 程序将一个非线性的回归方程序分别对参数求得二次的微分。二次的微分方程序是牛顿法所必须的。其撰写方式与指令 #5 相似，现举一例说明：

```
PROC NLIN METHOD=NEWTON;
    PARMS B0=0 TO 10
         B1=.01 TO .09 BY 0.005;
    TEMP=exp (-B1*X);
    MODEL Y=B0*(1-TEMP);
    DER.B0=1-TEMP; DER.B1=B0*X*TEMP;
    DER.B0.B0=0;
    DER.B0.B1=X*TEMP;
    DER.B1.B1= -DER.B1*X;
```

指令 #7　OUTPUT OUT＝输出文件名称 关键字=变量名称串；

本指令包括两个部分：OUT= 与关键字= 。现分别将这两部分说明如下：

OUT=输出文件名称

这个文件含原输入文件的所有变量，以及本指令中所提到的变量 (如：PREDICTED，RESIDUAL 等，详情见下段)。

关键字=变量名称串

下列是十四种关键字及其意义：

(1) PREDICTED (或 P)= 预测值
(2) RESIDUAL (或 R) = 预测误差值
(3) L95M = 因变量平均数的 95% 信赖区间之下限
(4) U95M = 因变量平均数的 95% 信赖区间之上限
(5) L95 = 因变量预测值的 95% 信赖区间之下限，它包括误差及参数估计值的标准误差
(6) U95 = 因变量预测值的 95% 信赖区间之上限，它包括误差及参数估计值的标准误差
(7) STDI = 各个观察体预测值的标准误差
(8) STDP = 预测值平均数的标准误差
(9) STDR = 误差的标准误差
(10) STUDENT = 标准化后的误差
(11) H = 影响力 (Leverage) 统计值，定义是 $X_i(X'X)^{-1}X_{i'}$，其中 X= e F/ e β。
(12) PARMS = 参数的估计值
(13) SSE (或 ESS) = 误差的平方总和。对每一个观察体而言，这个变量的值是常数。
(14) WEIGHT = 含 _WEIGHT_ 加权值的变量名称。

注：当 METHOD=DUD 时，H， L95， U95， L95M， U95M， STDP， STDR，以及 STUDENT 等关键字无效。

指令 #8　ID 变量名称串；

这些变量必须是上面各指令中从未提及的变量。它们将一并纳入输出文件内，而且印在观察体代号的左边。

指令 #9　BY 变量名称串；

NLIN 程序依据此指令所列举的变量值将文件分成几个小的文件，然后对每一固小文件分别执行分析。当读者选用此指令时，文件内的数据必须先按照 BY 变量串的值作由小到大的重新排列，这个步骤可藉 PROC SORT 来达成。

23.3　范　　例

例一：负指数函数的回归分析

本例的非线性回归模型可以下式表示：

$$Y=\beta_0*(1-e^{-\beta_1 X})$$

其中，X 是自变量，Y 是因变量，β_0 与 β_1 是参数。β_0 与 β_1 有多个起始值，用玛克底特法搜索这两个参数的估计值。Y 的预测值与预测误差将被包含在输出文件（B）中。

程　序

```
TITLE 'NEGATIVE EXPONENTIAL:Y=B0*(1-EXP(-B1*X))';
DATA A;
    INPUT X Y @@;
    CARDS;
020 0.57 030 0.72 040 0.81 050 0.87 060 0.91 070 0.94
080 0.95 090 0.97 100 0.98 110 0.99 120 1.00 130 0.99
140 0.99 150 1.00 160 1.00 170 0.99 180 1.00 190 1.00
200 0.99 210 1.00
;
PROC NLIN BEST=10 METHOD=MARQUARDT;
    PARMS B0=0 TO 2 BY .5  B1=.01 TO .09 BY .01;
    MODEL Y=B0*(1-EXP(-B1*X));
    DER.B0=1-EXP(-B1*X);
    DER.B1=B0*X*EXP(-B1*X);
    OUTPUT OUT=B P=YHAT R=YRESID;
PROC PLOT DATA=B;
    PLOT Y*X='A' YHAT*X='P' /OVERLAY VPOS=25;
    PLOT YRESID*X /VREF=0 VPOS=25;
RUN;
```

结　果

参数的估计十分顺利；在四次循环搜索后即达内设的收敛指标。所求得的非线性模型的均方(8.83586) 显著地大于误差的平均方。另外，Y 的预测误差对自变量 X 的描图显示出随机的形状。所以，我们可下结论说：参数的估计值令人满意。

报表 23.1　负指数函数的回归分析

NEGATIVE EXPONENTIAL:Y=B0*(1-EXP(-B1*X))

Non-Linear Least Squares Grid Search　　Non-Linear Least Squares Iterative Phase

Dependent Variable Y　　Dependent Variable Y　Method: Marquardt

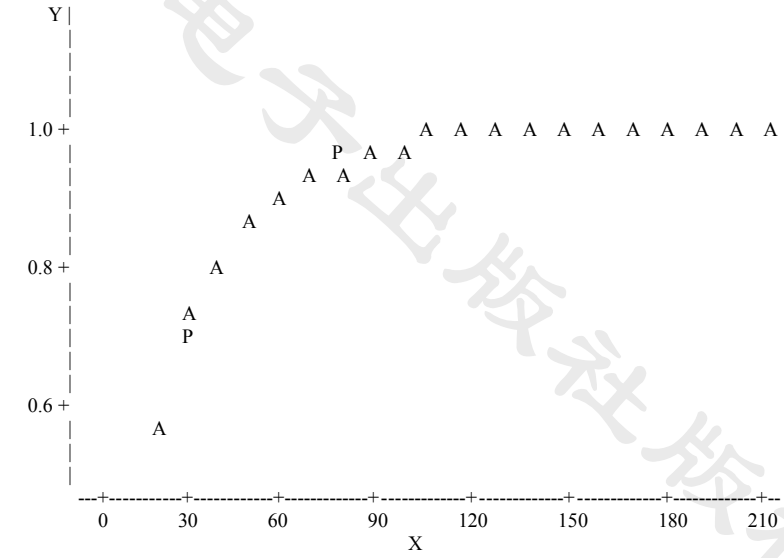
B0	B1	Sum of Squares	Iter	B0	B1	Sum of Squares
1.000000	0.040000	0.001404	0	1.000000	0.040000	0.001404
1.000000	0.050000	0.016811	1	0.996139	0.041857	0.000580
1.000000	0.060000	0.055155	2	0.996192	0.041952	0.000577
1.000000	0.030000	0.066571	3	0.996189	0.041954	0.000577
1.000000	0.070000	0.097284	4	0.996189	0.041954	0.000577
1.000000	0.080000	0.136536	NOTE: Convergence criterion met.			
1.000000	0.090000	0.170839				
1.000000	0.020000	0.419285				
1.500000	0.010000	0.975724				
1.000000	0.010000	2.165290				

```
Non-Linear Least Squares Summary Statistics
Dependent Variable Y
Source                     DF   Sum of Squares     Mean Square
Regression                  2    17.671723189     8.835861595
Residual                   18     0.000576811     0.000032045
Uncorrected Total          20    17.672300000
(Corrected Total)          19     0.243855000
Parameter            Estimate      Asymptotic                Asymptotic 95 %
                                   Std. Error             Confidence Interval
                                                        Lower           Upper
B0          0.9961885657   0.00161380015   0.99279811976   0.99957901159
B1          0.0419538868   0.00039822900   0.04111724424   0.04279052935

                Asymptotic Correlation Matrix
 Corr                B0              B1
 B0                   1    -0.555895746
 B1    -0.555895746                 1
```

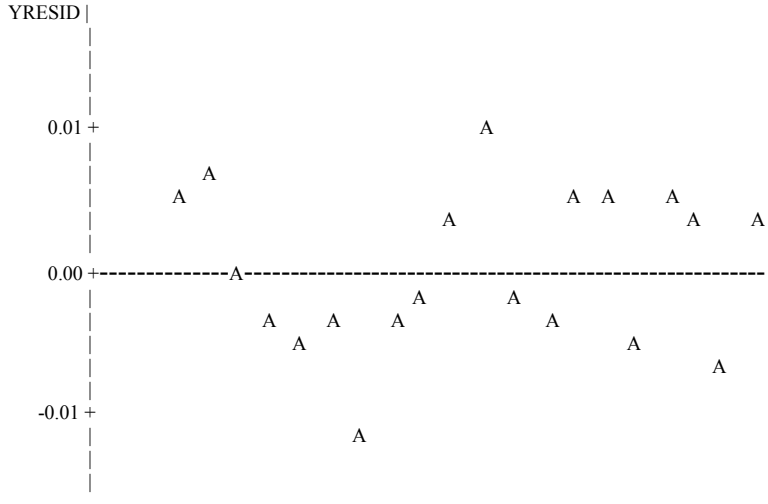
Plot of Y*X. Symbol used is 'A'.

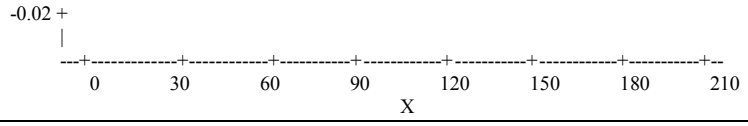
Plot of YHAT*X. Symbol used is 'P'.



NOTE: 18 obs hidden.

Plot of YRESID*X. Legend: A = 1 obs,B = 2 obs,etc.





例二：经济学上的 CES 生产指标

经济学上的 CES 生产函数代表生产量 (Q)，投资 (K)，以及劳力 (L) 等之间的关系；这个函数关系首先由 Arrow，Chenery，Minhas 及 Solow 提出，并将它命名为 CES 函数 (全名为 Constant Elasticity of Substitution，恒常弹性取代函数)。本文件 (CES) 的数据由 Lutkepohl 提供 (见 Judge 等人，1980)，变量 A 是效度参数，D 是分配 (或分享) 参数，R 是取代参数，LOGQ 是生产量 Q 的自然对数。

程　序

```
TITLE 'CES MODEL:LOGQ=B0+A*LOG(D*L**R+(1-D)*K**R)';
DATA
     INPUT L K LOGQ @@;
     CARDS;
.228 .802 -1.359 .258 .249 -1.695
.821 .771   .193 .767 .511  -.649
.495 .758  -.165 .487 .425  -.270
.678 .452  -.473 .748 .817   .031
.727 .845  -.563 .695 .958  -.125
.458 .084 -2.218 .981 .021 -3.633
.002 .295 -5.586 .429 .277  -.773
.231 .546 -1.315 .664 .129 -1.678
.631 .017 -3.879 .059 .906 -2.301
.811 .223 -1.377 .758 .145 -2.270
.050 .161 -2.539 .823 .006 -5.150
.483 .836  -.324 .682 .521  -.253
.116 .930 -1.530 .440 .495  -.614
.456 .185 -1.151 .342 .092 -2.089
.358 .485  -.951 .162 .934 -1.275
;
PROC NLIN DATA=CES;
     PARMS B0=1 A=-1 D=.5 R=-1;
     LR=L**R;
     KR=K**R;
     Z=D*LR+(1-D)*KR;
     MODEL LOGQ=B0+A*LOG(Z);
     DER.B0=1;
     DER.A=LOG(Z);
```

```
        DER.D=(A/Z)*(LR-KR);
        DER.R=(A/Z)*(D*LOG(L)*LR+(1-D)*LOG(K)*KR);
    RUN;
```

结　果

经过十五次的循环分析后，参数的估计值趋向稳定。所求得的非线性回归模型平均方 (32.50) 远超过误差的平均方 (0.07)。因此，我们可下结论说：参数的估计值有意义。

报表 23.2　经济学上的 CES 生产指标

CES MODEL:LOGQ=B0+A*LOG(D*L**R+(1-D)*K**R

Non-Linear Least Squares Iterative Phase

Dependent Variable LOGQ　Method: Gauss-Newton

Iter	B0	A	D	R	Sum of Squares
0	1.000000	-1.000000	0.500000	-1.000000	37.096512
1	0.533488	-0.481091	0.450601	-1.499936	35.486564
2	0.320516	-0.307656	0.383160	-2.309682	22.690597
3	0.124790	-0.287428	0.301408	-3.418181	1.845468
4	0.124044	-0.307921	0.317150	-3.204351	1.833362
5	0.122933	-0.355632	0.349730	-2.800352	1.820337
6	0.125085	-0.324295	0.330214	-3.089113	1.774004
7	0.124011	-0.342505	0.340530	-2.951604	1.762108
8	0.124713	-0.332754	0.334596	-3.038983	1.761177
9	0.124346	-0.338244	0.337849	-2.993735	1.761057
10	0.124563	-0.335197	0.336024	-3.020171	1.761043
11	0.124446	-0.336890	0.337035	-3.005870	1.761040
12	0.124512	-0.335947	0.336471	-3.013966	1.761040
13	0.124476	-0.336471	0.336785	-3.009505	1.761039
14	0.124496	-0.336179	0.336610	-3.012002	1.761039
15	0.124485	-0.336341	0.336707	-3.010617	1.761039

NOTE: Convergence criterion met.

Non-Linear Least Squares Summary Statistics　Dependent Variable LOGQ

Source	DF	Sum of Squares	Mean Square
Regression	4	130.00369371	32.50092343
Residual	26	1.76103929	0.06773228
Uncorrected Total	30	131.76473300	
(Corrected Total)	29	61.28965430	

Parameter	Estimate	Asymptotic Std. Error	Asymptotic 95 % Confidence Interval Lower	Upper
B0	0.124485105	0.0783429642	-0.0365498914	0.2855201005
A	-0.336341238	0.2721800618	-0.8958109440	0.2231284680
D	0.336707458	0.1360850556	0.0569828319	0.6164320846
R	-3.010617415	2.3229032585	-7.7853756933	1.7641408635

Asymptotic Correlation Matrix

Corr	B0	A	D	R
B0	1	0.2964899511	-0.176549933	-0.32669583
A	0.2964899511	1	-0.783557332	-0.999129892
D	-0.176549933	-0.783557332	1	0.7833628736
R	-0.32669583	-0.999129892	0.7833628736	1

例三：概率单位与数值微分方程序

本文件 (USPOP) 的数据是美国自 1780 年来每十年的人口总数。其回归模型以累积常态分配的反函数为主。这个反函数又称作概率单位 (PROBIT)。由于概率单位的微分方

程序不易获得，本例采取数值微分方程序 (Numerical Derivative) 的步骤来估计参数。

程　序

```
TITLE 'U.S. POPULATION GROWTH';
TITLE2 'PROBIT MODEL WITH NUMERICAL DERIVATIVES';
DATA USPOP;
    INPUT POP : 6.3 @@;
    RETAIN YEAR 1780;
    YEAR=YEAR+10;
    YEARSQ=YEAR*YEAR;
    CARDS;
3929 5308 7239 9638 12866 17069 23191 31443 39818 50155
62947 75994 91972 105710 122775 131669 151325 179323 203211
;
PROC NLIN DATA=USPOP;
    PARMS A=-2.4 B=.012 C=400;
    DELTA=.0001;
    X=YEAR-1790;
    POPHAT=C*PROBNORM(A+B*X);
    MODEL POP=POPHAT;
    DER.A=(POPHAT-C*PROBNORM((A-DELTA)+B*X))/DELTA;
    DER.B=(POPHAT-C*PROBNORM(A+(B-DELTA)*X))/DELTA;
    DER.C=POPHAT/C;
    OUTPUT OUT=P P=PREDICT;
PROC PLOT DATA=P;
    PLOT POP*YEAR PREDICT*YEAR='P' /OVERLAY VPOS=30;
RUN;
```

结　果

参数估计值在五次循环搜索后，即稳定下来。所求得的非线性模型的平均方 (54742.63) 显著地大于误差的平均方值 (11.086)。另外，人口数的实际值 (以 A，B 表示) 与预测值 (以 P 表示) 十分接近。这个结论可由 PLOT 的重叠图的检视而导出。

报表 23.3　概率单位与数值微分方程序

U.S. POPULATION GROWTH

PROBIT MODEL WITH NUMERICAL DERIVATIVES

Non-Linear Least Squares Iterative Phase

Dependent Variable POP　Method: Gauss-Newton

Iter	A	B	C	Sum of Squares
0	-2.400000	0.012000	400.000000	7174.590805

1	-2.271908	0.012623	399.066499	209.327927
2	-2.302425	0.012661	404.804742	177.392064
3	-2.302788	0.012628	407.072751	177.370044
4	-2.302819	0.012629	407.079801	177.369804
5	-2.302818	0.012629	407.082668	177.369803

NOTE: Convergence criterion met.

Non-Linear Least Squares Summary Statistics　Dependent Variable POP

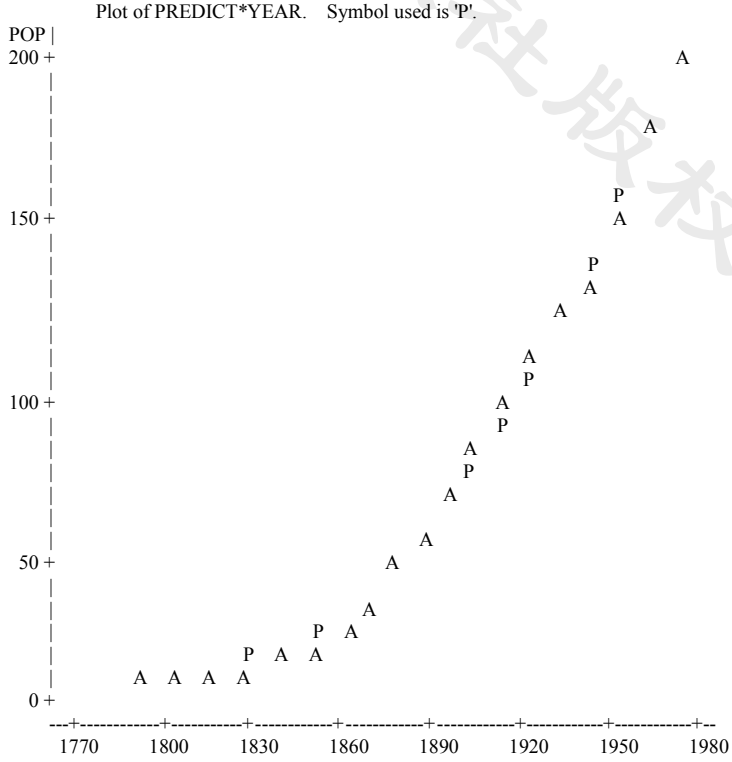
Source	DF	Sum of Squares	Mean Square
Regression	3	164227.89925	54742.63308
Residual	16	177.36980	11.08561
Uncorrected Total	19	164405.26906	
(Corrected Total)	18	71922.76175	

Parameter	Estimate	Asymptotic Std.Error	Asymptotic 95 % Confidence Interval Lower	Upper
A	-2.3028183	0.032832711	-2.37242015	-2.23321637
B	0.0126285	0.000956986	0.01059980	0.01465722
C	407.0826677	61.784898470	276.10518493	538.06015048

Asymptotic Correlation Matrix

Corr	A	B	C
A	1	-0.007910181	-0.219797799
B	-0.007910181	1	-0.972273297
C	-0.219797799	-0.972273297	1

Plot of POP*YEAR.　Legend: A = 1 obs,B = 2 obs,etc.
Plot of PREDICT*YEAR.　Symbol used is 'P'.



```
                                    YEAR
NOTE: 12 obs hidden.
```

23.4　注 意 事 项

■ 遗漏数据的处理法

当观察体的因变量含遗漏数据时，NLIN 程序仍然可以求出其预测值，并将之纳入输出文件中。然而，若观察体的任何一个自变量含遗漏数据，则 NLIN 程序会将此观察体排除在分析之外。

■ 电脑分析可能遭遇的问题

(1) CPU 时间不够

这个问题的起源有两种可能：

一、 它可能源于参数的起始值。当参数起始值的组合过多时，电脑分析所需的时间便相对地增加。

二、 它可能源于参数估计值的搜索过程。若循环搜索的次数过多，则电脑分析所需的时间便可能太长。

(2) 参数之间的相关

当参数之间有线性相关时，微分方程序的矩阵会是一个非满秩矩阵 (Singular Matrix)。如此模型可能导致没有独特解或是无解。请看下面的例子：

```
PARMS B0=0 B1=.022 B2=0;
MODEL POP=B0*EXP(B1*(YEAR-1790)+B2);
DER.B0=EXP(B1*(YEAR-1790)+B2);
DER.B1=(YEAR-1790)*B0*EXP(B1*(YEAR-1790)+B2);
DER.B2=B0*EXP(B1*(YEAR-1790)+B2);
```

若 B0 以 0 为起始值，则 B1 在搜索过程的第一回合会自动成为 0。在第一次循环后，B0 与 B2 的参数估计值彼此互换而无法稳定下来。因而将导致收敛指标无法逐次递减，以致这个问题无解。

(3) 函数值无法再改善

在参数估计值的搜索过程当中，有时遇到一些函数值是该函数的局部最佳解，然而这并不是全面最佳解。若局部最佳解使搜索的过程停止 (因已达到收敛指标)，则 NLIN 程序会在主机的报表上印出 'CONVERGENCE ASSUMED' (收敛似乎达成) 或在 PC 报表上印出 'PROC NLIN failed to converge' (NLIN 程序无法将函数指标值再作收敛) 的字样。读者必须仔细检查误差的平方总和还有微分方程序以决定是否应该继续尝试搜索的过程。若欲继续搜索，读者可以试另一组新的起始值，另一种搜索参数估计值的方法，或采 G4 选项。

(4) 扩散的函数

有时参数估计值的搜索过程会导致函数的扩散而非收敛。请见下例：

```
PARMS B=0;
MODEL Y=X/B;
```

假设所有的 Y 值等于 0，而所有的 X 值不等于 0。在此情形下，很明显的无最小平方的解。然而 NLIN 的分析方法仍会导致搜索过程的收敛。这是因为收敛的标准以前后两次搜索所估计出来的参数值的改变为准 ，所以不一定代表真正的收敛。若将模型改写成下式，则问题就可以解决：

```
MODEL Y=A*X;
```

(5) 局部解

局部解也会导致 NLIN 程序做错误的收敛决定。请看下例：

```
PARMS A=1 B=-1;
MODEL Y=(1-A*B)*(1-B*X);
DER.A=-X*(1-B*X);
DER.B=-X*(1-A*X);
```

(6) 函数的不连贯性

所有参数估计值的搜索法都假设模型是参数的连续函数。若此假设不成立，则将导致无解，所以下面的两个模型无解：

```
MODEL Y=A+INT(B*X);
MODEL Y=A+B*X+4*(Z>C);
```

■ 经验谈——野人献曝

NLIN 程序不保证你第一次就能找出最好的解。读者在程序中所界定的参数起始值对整个分析的成败有很大的关系，请务必注意。另外，如果分析过程中遭遇到问题，可以试着换一组参数起始值，或换一种搜索参数估计值的方法。

■ PROC NLIN 程序所产生的一些特殊变量

下面几个变量是 NLIN 程序自动产生的。读者可以在 NLIN 程序中用它们，但不可更改这些变量的值。

_N_　　文件内观察体的总数。

ERROR　若程序中有不合理的指令，或错误的 (不) 等式，则此变量的值为 1。否则，其值是 0。

OBS　　文件内有效的观察体数目。

ITER　　搜索过程中循环的次数。

MODEL　若分析的过程只需 Y 的预测值，则此变量的值为 1。若分析的过程中需要 Y 的预测值以及微分方程序，则此变量的值为 0。若微分方程序

的计算过于复杂，读者可在 MODEL 指令后，微分方程序前加上下面这一行，以节省电脑计算的时间：

```
IF _MODEL_ THEN RETURN;
```

SSE　　前一次循环搜索中误差的标准误差平方总和。

以下两个特殊变量是用来控制收敛指标的，这两个变量的值由读者自定。

HALVE　搜索过程的每一循环里，函数值折半 (Step Halvings) 的最高次数。

LOSS　在最大可能率估计法 (Maximum Likelihood Method) 中，目标函数的值。其值由读者自定。

■ PROC NLIN 与 DATA 程序部分的交递作用

NLIN 程序是一个十分特别的统计程序，因为它容许读者在分析的过程中同时界定新变量或执行若干 DATA 程序的指令。比方说，下面的程序会一方面进行非线性的回归分析，一方面产生新变量：

```
PROC NLIN; PARMS B0=0 TO 10
                 B1=1 TO 9;
    TEMP=EXP(-B1*X);
    MODEL Y=B0*(1-TEMP);
RUN;
```

除了界定新的变量外，NLIN 程序还可接纳下列数据 DATA 程序的指令：

1. ARRAY 的界定，向量的名称可采用两节式命名 (Compound Names)
2. 任何指派语句（参阅附录 C.2 节）
3. CALL 指令
4. DO 指令以及循环的 DO 语句
5. DO UNTIL 以及 DO WHILE
6. END 指令
7. FILE 指令
8. GO TO 指令
9. IF-THEN/ELSE 指令
10. LINK-RETURN 指令
11. PUT 指令，其结果纳入日志文件，若读者想同时存一份在报表文件内，则每次执行 PUT 指令前，应加上下列的指令：

```
FILE PRINT;
```

12. RETAIN 指令
13. RETURN 指令
14. SELECT 指令
15. sum. 指令