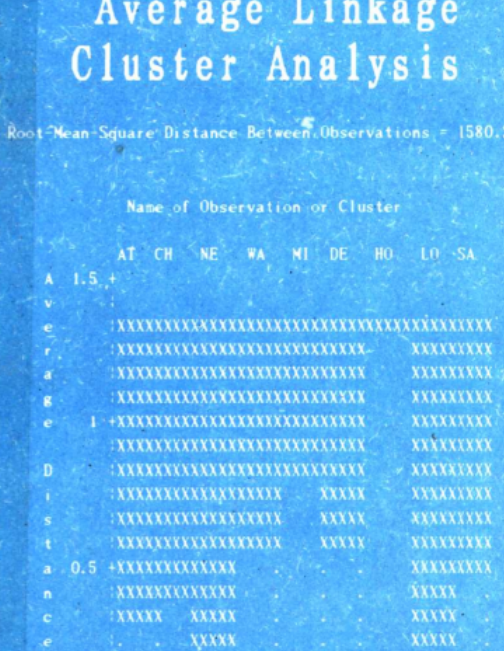
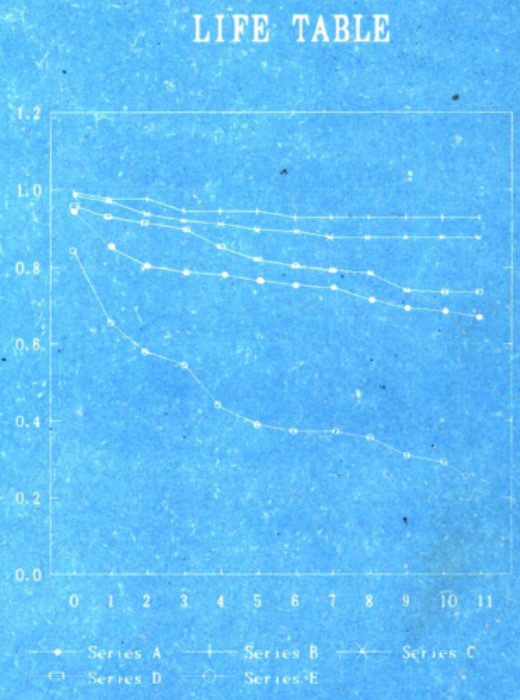
董大钧　张尔强　何武　等译

SAS 统计过程指导





辽宁科学技术出版社

SAS统计过程指导

[美] SAS 研究所 著

董大钧 张尔强 何 武等译

刘延龄 审校

辽宁科学技术出版社

(辽) 新登字 4 号

SAS 统计过程指导
SAS TONGJI GUOCHENG ZHIDAO
[美]SAS研究所著
董大钧 张尔强 何 武等译

辽宁科学技术出版社出版发行
(沈阳市和平区北一马路108号 邮政编码110001)
中国医科大学计算中心激光照排
中国医科大学印刷厂印刷

开本:787×1092 1/16 印张:32.5 字数:748,000
1992年2月第1版 1992年2月第1次印刷

责任编辑: 枫岚 版式设计: 李夏
封面设计: 太文 责任校对: 王莉

印数:1-1500
ISBN 7-5381-1551-X/TP.22 定价:19.5元

内 容 简 介

SAS (Statistical Analysis System)是一个用来分析数据和编写报告的软件系统。它是美国 SAS研究所经过十年研制，于1976年推出的。是世界上最受欢迎的统计软件包之一。目前 SAS已由一个流行的功能强大的统计分析软件发展成为用途广泛的第四代高级编程语言。它广泛用于医学、农林、财经、社会科学等一切从事数据管理与数据分析处理的领域中。它使用简单，几乎能用极简单的命令去做你想作的一切数据整理和分析工作。

SAS语言实用性强，容易学习，它极适于科研人员和各种从事数据处理和信息管理的人员使用。该软件在国外极流行，近两年在我国正迅速普及。

本书译自美国SAS 公司的 <<SAS User's Geide: Statistics>> 。本书通过大量实例详细介绍了 SAS统计软件包中各种多变量分析等实用统计过程的使用方法。

本书不仅是SAS软件使用手册，而且可作为学习SAS软件课程的教材，还可作为计算机应用人员和统计工作者使用SAS的参考资料。

前　言

在信息时代的今天，人们在工作实践中会获取到大量的信息。如何存贮、整理和分析处理它们是一件极重要的工作。由于对数据的分析大都是基于基本的统计原理进行的，国内外学者多年来编制了许多统计软件包。SAS软件包则是诸多统计软件包中的佼佼者。

SAS (Statistical Analysis System)是一个用来分析数据和编写报告的软件系统。它是美国 SAS 研究所经过十年研制，于1976年推出的。最初 SAS只能运行于大型机上，1985年被移植到微机，从而得到迅速推广。SAS 公司每年都在改进SAS 系统，目前 SAS 已由一个流行的功能强大的统计分析软件包发展成为用途广泛的第四代高级编程语言。它广泛用于医学、农林、财经、社会科学等一切从事数据管理与数据分析处理的领域中。它使用简单，几乎能用极简单的命令去做你想作的一切数据整理和分析工作。SAS 语言实用性强，容易学习，它极适于科研人员和各种从事数据处理和信息管理的人员使用。该软件在国外极流行，近两年在我国正迅速普及。

SAS 软件包可对数据进行一般描述的统计分析、分类统计检验、分布评价、可信区间计算、方差分析、因子分析、回归分析(包括Logistic回归分析) 等多因素统计分析，也可进行时间系列分析。它使用简单，几乎能用极简单的命令去作你想作的一切数据整理和分析工作。

SAS系统具有积木式的结构，在SAS/BASE软件的基础上，可以任意增加象SAS/GRAPH, SAS/STAT, SAS/IML 等功能块，组成完善的SAS 系统。 与其它几个世界上流行的统计软件包相比，SAS 系统在数据预处理、中间结果的存贮与调用以及数据管理等方面具有独特之处。因此，除了广泛应用于统计工作之外，SAS在商业界也得到越来越多的应用。

本书通过大量实例详细介绍了SAS 统计软件包中各种多变量分析等实用统计过程的使用方法。

该书的第 1至第8 章、12章、13章、第17至19章、第26至27章、第33至35章、第38至39章、第41章由张尔强同志翻译；第 9章、16章、28章、36、 37章是陈智同志翻译；第10章、15章、第29至32章是谢红同志翻译；第11章、20章是李传俊同志翻译；第14章是董大钧同志翻译；第21章、22章是于石成同志翻译；第23章、24章、40章是景立臣同志翻译；第25章是何武同志翻译；由刘延龄教授审阅。

编　者

1990年3月

目　录

第1章 回归过程

§1 概述

本章综合叙述了可用于回归分析的各SAS过程，其中包括REG、RSQUARE、STEPWISE、NLIN及RSREG过程。

许多SAS过程都可以进行回归分析，但又各有特色:

REG: 进行通用的回归分析，并且带有许多诊断和输入输出能力;

RSQUARE: 建立模型，并对所有可能的模型显示其拟合程度;

STEPWISE: 提供几种逐步选择模型的方法;

NLIN: 建立非线性回归模型;

RSREG: 建立二次响应曲面回归模型。

此外还有:

GLM: 进行广义线性模型分析，包括含有分类项及多项式的模型在内。(此过程在第2章方差分析中介绍)。

以上这些过程执行回归分析，即对一组值进行拟合，得到一个方程。调整参数使拟合最优，然后利用这个方程即可根据自变量预测响应变量。例如，第i个观测值可以是:

$$Y_i=\beta_0+\beta_1 x_i+\varepsilon_i$$

此处Y_i是响应变量，X_i是回归变量，β_0和β_1是未知的估计参数，而ε_i是误差项。例如，可以利用回归分析，根据一个人的身高去推测他的体重。假定你搜集了20个学生的身高和体重的数据，(数据略)你只要估计下式中的截距β_0和斜率β_1:

$$WEIGHT=\beta_0+\beta_1\ HEIGHT+\varepsilon$$

此处: WEIGHT 是响应变量，又称因变量;

β_0, β_1 是未知参数;

HEIGHT 是回归变量，又称自变量;

ε 是未知误差。

此题的回归估计是 $b_0=-125.6$及$b_1=3.7$，因而回归方程是:

$$WEIGHT=-125.6+3.7*HEIGHT$$

回归经常用来寻找经验公式，例如求WEIGHT和HEIGHT之间的关系。用来估计参数的方法是使实测值与由方程得到的预测值之差的平方和最小，因此也叫最小平方法，判别值被称为误差平方和:

$$SSE=\Sigma\ (y_i-b_0-b_1x_i)^2$$

此处b_0和b_1是使误差平方和SSE最小的参数值。

典型的回归分析产生以下信息:

(1) 利用最小二乘法得到的参数估计。

(2) 误差项的方差估计。

(3) 参数估计的标准差或方差估计。

(4) 参数假设检验。

(5) 利用估计得出预测值及残差。

(6) 评价拟合优劣的统计量。

除了对回归产生的通常的拟合统计量之外，SAS 回归过程还可以产生许多其它指定的诊断统计量，其中包括：

(7) 共线性诊断，度量回归变量与其它回归变量的依赖程度，以及对估计值的稳定性和方差的影响(REG)。

(8) 影响诊断，度量每个单个的观测值对决定参数估计、SSE、拟合值的贡献(REG、RSREG)。

(9) 拟合不佳诊断，通过比较误差方差估计和不依赖模型的纯误差方差度量回归模型的拟合不佳程度。

(10) 对等间隔时序数据进行时序诊断，度量误差与相邻观测值的关系。这些诊断还可以度量按回归或响应排序的数据的拟合优度(REG)。

§2 过程的比较

1. REG过程

REG过程是一个通用回归过程，具有以下特点:

(1) 可以有多个MODEL(模型)语句;

(2) 输入可以是相关或叉积;

(3) 打印预测值，残差，STUDENT化残差和可信限，并且可以将这些项输出到一个输出数据集中;

(4) 打印特定的影响统计量;

(5) 打印偏及半偏相关系数;

(6) 产生偏回归影响(效力)图;

(7) 估计服从线性限制的参数;

(8) 检验线性假设;

(9) 检验多元假设;

(10) 将估计值写入输出数据集;

(11) 将叉积矩阵写入输出数据集;

(12) 计算特殊的共线性诊断。

2. RSQUARE过程

RSQUARE过程对在MODEL语句中列出的变量的所有可能的组合进行拟合，并且打印参数估计及评价拟合的几个统计量。当你要寻找备择模型时，RSQUARE过程是非常有用的。由于可能的模型的数目随着变量的增多迅速增大（2^n，n为变量数），因此只有当可

以考虑的回归变量少于20时，才可以使用RSQUARE过程。

3. STEPWISE过程

STEPWISE过程通过各种逐步的方法为模型选择回归变量。为了选择一个好模型，可以要求 5种不同的方法。FORWARD 法又称向前选入法，从空模型开始，每步选入一个最大限度的拟合的变量。BACKWARD法又称后退剔除法，从满模型出发，每步移出一个对拟合贡献最小的变量。此外还有 3种方法：STEPWISE, MAXR, MINR。如果选择了 FORWARD, BACKWARD 或 SETPWISE法，过程将在模型建立的每一步，打印移进或移出变量及统计量。PROC STEPWISE过程还产生方差分析和参数估计，但不能给出预测值和残差。

4. NLIN过程

NLIN过程提供几种迭代法以便为非线性模型寻找最小平方估计值。缺省的方法是高斯－牛顿(Gauss-Newton)法。你必须指定参数名和起始值，模型表达式，模型关于参数的导数的表达式(除了METHOD=DUD)。格点搜索也可以用来选择参数的起始值。由于非线模型的估计常常是困难的，因此NLIN并不是总能找到最小二乘估计。

5. RSREG过程

RSREG 过程拟合二次响应面模型，这对于寻找优化响应的因子值是很有用的。

RSREG 的以下特色使得它在分析响应面方面优于其它回归过程：

(1) 自动产生二次效应；

(2) 拟合不佳检验；

(3) 面的临界值解；

(4) 与二次型关联的特征值；

6. GLM过程

GLM 过程对线性模型能进行回归，方差分析，协方差分析。对于回归来说，它与其它回归过程的区别是：

(1) 容易指明分类效应(GLM自动地为分类变量产生虚拟变量)。

(2) 直接指定多项式效应。

§3 统计基础

本章余下部分列出许多 SAS回归过程计算各种回归量的一般方法。例外和补充说明在各过程中介绍。用矩阵代数符号，线性模型可写作：

$$y=X\beta+\varepsilon$$

此处X是n*k设计矩阵(行是观测值，列是回归变量)，β 是K*1未知参数向量，ε 是n*1未知误差向量。X的第一列通常是用来估计截距项的1的向量。

线性模型的统计理论基于某些严格的经典假设。理想地说，响应是由因子度量的，

而因子又是受实验确定环境控制的。或者，如果不能控制实验因子，则必须假定因子相对于响应变量来说是固定的。其它的假设是：

. 模型的形式是正确的
. 回归变量可以无误差地度量
. 误差的期望值是零
. 所有观测值的误差的方差是一个常数，称为σ^2。
. 误差与观测值不相关

当检验假设时，还要假设：

. 误差是正态分布的。

统计模型

如果模型满足所有必要的假设，则最小平方估计量是最佳线性无偏估计量(BLUE)。换句话说，在响应的各种线性估计值之间，这些估计量有最小方差。如果误差是正态分布的假设也满足的话，那么：

(1) 被计算的统计量对假设检验来说有适当的样本分布；
(2) 参数估计将有正态分布；
(3) 各种平方和正比于卡方分布，至少在适当的假设之下；
(4) 估计值对标准误差的比值在一定假设下呈STUDENT分布；
(5) 平方和的近似比值在一定假设下呈F分布。

当回归分析用来对不满足假设的数据求模型时，结果将被警告，并降低显著概率的可信度。

§4 参数估计及相应的统计量

参数估计应用最小平方准则来进行的。由于：

$$(X'X)b=X'y$$

因而：

$$b=(X'X)^{-1}X'y\ .$$

设$(X'X)$是满秩的(以后这个条件可放宽)，误差的方差σ^2可以用下式求得：

$$s^2=MSE=SSE/(n-k)=\Sigma\ (y_i-x_ib)^2/(n-k)$$

这里x_i是回归变量的第i行。

参数估计是无偏的：

$$E(b)=\beta$$

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

于是有方差—协方差矩阵：

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

上面公式中，方差矩阵的估计值是用s^2代替σ^2

$$COVB=(X'X)^{-1}s^2$$

估计的相关通过使对角线变换为1来求出，设:

$$S=\mathrm{diag}((X'X)^{-1})^{-.5}$$

$$CORRB=S(X'X)^{-1}S$$

估计值的标准误由下式计算:

$$STDERR(bi)=\sqrt{(X'X)^{-1}_{ii}S^2}$$

此处$(X'X)^{-1}_{ii}$是$(X'X)^{-1}$的第i个对角线元素，比值t=bi/stderr(bi)在β i为0的假设下是STUDENT t分布。回归过程打印比值t及显著概率。当概率小于某个水平时，假设被拒绝。

Ⅰ类及Ⅱ类SS(平方和)用来度量变量对变化SSE(误差平方和)所做的贡献。Ⅰ类SS度量一个变量被顺序地加进模型中SSE的减少，而Ⅱ类SS是从满模型中移走一个变量所导致的SSE的增加。Ⅱ类SS等同于GLM过程中的Ⅲ类及Ⅳ类SS。如果Ⅱ类SS被用在F检验的分子，那么这种检验等价于参数为0的t检验。在多项式模型中，Ⅰ类SS度量每个多项式的项在它被正交化(对模型中前面的项)的贡献。四种类型的SS在GLM过程及第9章四种估计函数中作了详细说明。

标准化估计被定义为所有的变量都被标准化为均值为0方差为1时的估计值。这可以通过把原估计值乘以回归变量的标准差然后除以因变量的样本标准差获得。

容许差及方差膨胀因子度量模型中回归变量之间相互关系的强度。如果所有变量都是彼此正交的，则容许差及方差膨胀因子都是1。如果某一个变量与其它变量密切相关，则容许差为0而方差膨胀因子变得非常大。容许差(TOL)是1减去R^2，R^2是模型中的其它变量对前述变量回归产生的。如果$(X'X)$被换算成相关方式。方差膨胀(VIF)是$(X'X)^{-1}$的对角线，统计量的关系如下:

$$VIF=1/TOL$$

如果模型是非满秩的，则通常使用如下的广义逆变，以使SSE最小:

$$b=(X'X)^{-}X'y$$

然而，这些估计值并不唯一，因为使用不同的广义逆变，结果有无数个。REG及其它回归过程对所有线性独立的变量选择一个非零解，对其余变量选择一个零解。这对应于在正规方程中使用广义逆变，而且估计的期望值是X'X的Hermite正规方式乘以参数:

$$E(b)=(X'X)^{-}(X'X)\beta$$

零参数估计的自由度是零，不可检验的假设打印为缺项。模型非满秩的信息包括矩阵中有关的输出。

§5 预测值及残差

在模型拟合后，预测值和残值差通常被计算且输出。预测值根据估计的回归方程算出，而残差是实测值与预测值之差。某些过程可以计算标准误。

设Xi为第i个回归量，b是参数估计的向量，s^2是均方差，考虑第i个观测值:

$$hi=x_i(X'X)^{-1}x'_i \qquad \text{(影响)}$$

那么: $\hat{y}_i=X_ib$ (预测值)

$STDERR(\hat{y}_i)=\sqrt{h_i s^2}$ (预测值的标准误)

$resid_i=y_i-x_i b$ (残差)

$STDERR(resid_i)=\sqrt{(1-hi)s^2}$ (残差的标准误)

残差与标准误的比值，称为STUDENT化残差，有时写成:

student=resid/STDERR(resid)

这里有两种预测值的可信区间。一种是响应的期望值的可信区间，另一种是响应的实测值的可信区间，它等于期望值加上误差。

例如，可以为第i个观测值写出一个概率为1-α的含有真期望值的可信区间，其上下限为:

下限 $M=x_i b-t_{\alpha/2}\sqrt{(h_i s^2)}$

上限 $M=x_i b+t_{\alpha/2}\sqrt{(h_i s^2)}$

而一个实测的单个响应的可信区间为:

下限 $I=x_i b-t_{\alpha/2}\sqrt{(h_i s^2+s^2)}$

上限 $I=x_i b+t_{\alpha/2}\sqrt{(h_i s^2+s^2)}$

统计量COOKD测量由于删除每个观测值而使估计结果的变化:

$$COOKD=Student^2(STDERR(\hat{Y})/STDERR(resid))^2/K$$

第i个观测值预测残差被定义为从参数估计中去掉第i个观测值所产生第i 个观测值的残差。预测残差的平方和叫作Press统计量。

$$Presid_i=resid_i/(1-h_i)$$

$$Press=\sum presid_i^2$$

§ 6 线性假设检验

参数的线性假设一般形式是:

$$H_0: L\beta=C$$

此处L是qxk，β是kx1；C是qx1向量，为了检验该假设，考虑以下参数的线性函数:

$$(Lb-c)$$

其方差为: $Var(Lb-c)=LVar(b)L'=L(X'X)^{-}L'\sigma^2$

此处b 是β的估计值。

在上述假设下的二次型:

$$SS(Lb-c)=(Lb-c)'(L(X'X)^{-}L')^{-1}(Lb-c)$$

假设检验是可进行的，则SS可以作为F检验的分子:

$$F=SS(Lb-c)/q/s^2$$

这称为带有自由度q和dfe的F分布，此处dfe是剩余误差的自由度。

§7 多元检验

多元假设在公式中包含几个因变量：

H_0: L β M=d

此处L是回归自变量侧的线性函数，β是参数矩阵，M是因变量侧的线性函数，d是常数矩阵。对每个因变量常数都相同的特殊情况(被REG处理)则可写作：.

(L β -cj) M=0

此处C是常数的列向量而j是1的行向量。当常数皆为零时，

L β M=0

为了检验该假设，要用2个矩阵H及E，它们分别对应于一元F检验的分子和分母。

$H=M'(LB-cj)'(L(X'X)^{-1}L')^{-1}(LB-cj)M$

$E=M'(Y'Y-B'(X'X)B)M$

根据$E^{-1}H$或$(E+H)^{-1}H$的特征值，可计算4个检验统计量。设λi是$E^{-1}H$的有序特征值(如果逆矩阵存在的话)，并且ξi是$(E+H)^{-1}H$的有序特征值，则有：

$\xi i=\lambda i/(1+\lambda i)$

及 $\lambda i=\xi i/(1-\xi i)$

且 $\rho i=\sqrt{\xi i}$是第i个典型相关。

设ρ是(H+E)的秩，它小于或等于M的列数，设q是$L(X'X)^{-1}L'$的秩，v是误差自由度，设：$S=\min(p,q)$，$m=0.5(|p-q|-1)$，$n=0.5(v-p-1)$则下面的统计量近似于F统计量：

Wilks' lambda (威尔克斯λ)

$\wedge=\det(E)/\det(H+E)=\prod 1/(1+\lambda i)=\prod(1-\xi i)$

$F=(1-\wedge^{1/t})/(\wedge^{1/t})(rt-2u)/pq$

近似于F，此处：

$r=v-(p-q+1)/2$，$u=(pq-2)/4$，

如果 $(p^2+q^2-5)>0$ 则 $t=\sqrt{(p^2q^2-4)/(p^2q^2-5)}$ 否则：$t=1$

自由度是pq及rt-2u。如果$\min(p,q)<2$，这个近似值是准确的。

Pillai's trace (迹)：

$V=\text{trace}(H(H+E)^{-1})=\sum\lambda i/(1+\lambda i)=\sum\xi i$

$F=(2n+s+1)/(2m+s+1)$，$V/(s-V)$是自由度为：

S(2m+s+1)和s(2n+s+1)的近似F值。

Hotelling-Lanley trace:

$U=\text{trace}(E^{-1}H)=\sum\lambda i=\sum\xi i/(1-\xi i)$

$F=2(sn+1)U/(S^2(2m+S+1)$是自由度为S(2m+S+1)及2(sn+1)的F近似值。

Roy's maximam root (Roy 最大根)：

$\theta=\lambda 1$

$F=\theta(V-r-1)/r$，$r=\max(p,q)$

第2章 方差分析

§1 概述

本章概括介绍可用于方差分析的过程，其中包括 GLM, ANOVA, NESTED, VARCOMP, NPAR1WAY, TTEST和PLAN。应用最广的方差分析过程是GLM, 它能处理大多数问题。其它过程也各有特色。

GLM　进行方差分析、回归、协方差分析、多元方差分析;

ANOVA　对平衡设计进行方差分析;

NESTED　纯嵌套随机模型的方差分析;

VARCOMP　估计方差成分;

NPAR1WAY　秩次分的非参数单向分析;

TTEST　比较两组观测值的均值;

PLAN　为实验计划产生随机配置。

方差分析是一种分析实验数据常用的技术。通过分类变量指明各种实验条件，测得相应的连续响应变量。响应中的变异可解释为按随机误差分类计算剩余变异效应。

对每个观测值，ANOVA模型常常通过样本均值预测响应。实测值和预测值之差称为残差。方差分析过程拟合参数。以使残差的平方和最小，因此该方法又称最小二乘法。随机误差的方差σ^2由均方差(MSE或s^2)来估计。

§2 ANOVA (平衡设计)

决定使用哪个过程的因素之一是看数据是平衡的还是非平衡的。当设计一个实验的时候，对分类水平的每一个组合(或单元)，选择多少个实验单位呢？为了获得好的统计特性并简化统计算法，典型的做法是对实验的每种组合安排相同数目的实验单位，这种设计称为平衡设计。

如果数据是平衡的，则平方和的计算可以大大简化。在SAS中，可以用ANOVA过程，而不必用开销太多的GLM过程。可把平衡概念推广，用平衡设计的算法去处理每个单元含有的观测值数不等的设计。对所有单向模型都可以用平衡算法而不必管各单元计数是否平衡。甚至可以用平衡算法计算，尽管并不是所有单元中都有数据。

然而，如果你使用ANOVA过程去分析不平衡设计，可能得到不正确的结果，甚至平方和的值是负的。

方差分析过程通过在无效假设下，比较相对于它们期望值的均方进行ANOVA 检验。在一个固定的方差分析模型中，均方有一个期望值，它是由两部分组成；固定参数的二次函数和随机变异。 对于一个称为A的固定效应，它的均方期望值被写为:

$$E(MS(A))=Q(\beta)+\sigma^2$$

在无效假设下，期望值的固定部分Q(β)为0，接着与另外的均方比较，比如说MS(E)与第一部分无关，且有期望值σ^2，它们的比值F在无效假设下服从F分布：

$$F=MS(A)/MS(E)$$

当无效假设失效时，分子项有很大的期望值，但分母不变。因此F值大，导致拒绝无效假设。检验通过控制第一类错误率，即拒绝一个真实无效假设的概率来判断结论。如果这个概率小，比如说低于0.05或0.01，那么你犯错误的概率分别是0.05或0.01。如果你不能拒绝假设，则假设可能是真的，也可能是你没有足够的数据。

§3 广义线性模型

如果你的数据不满足平衡设计，那么你可能需要GLM过程中的线性模型体系。一个方差分析模型可以写成线性模型，即把响应看作是参数和设计变量的线性函数。一般地写作：

$$y_i=\beta_0X_{0i}+\beta_1X_{1i}+\ldots\ldots+\beta_kX_{ki}+\varepsilon_i$$

$$i=1\ldots\quad n$$

此处y_i是第i个观测值的响应，β_k是未知的待估计的参数，X_{ij}是设计变量。对方差分析来说，设计变量是指示变量，它们的值不是0就是1。

最简单的模型是用单独一个均值去拟合所有观测值。这种情况下，只有一个参数β_0和一个值总是1的设计变量X_{0i}：

$$Y_i=\beta_0X_{0i}+\varepsilon_i=\beta_0+\varepsilon_i$$

β_0的最小二乘法估计是Y的均值。这个简单模型是所有更复杂模型的基础，并且所有更大的模型都与这个简单均值模型相比较。

通过为分类变量的每个水平引进一个指示变量，可以写出单向模型。假设变量A有4个水平，每个水平有2个观测值，那么指示变量可建立如下：

intercept	a1	a2	a3	a4
1	1	0	0	0
1	1	0	0	0
1	0	1	0	0
1	0	1	0	0
1	0	0	1	0
1	0	0	1	0
1	0	0	0	1
1	0	0	0	1

线性模型可写成：

$$Y_i=\beta_0+a1_i\beta_1+a2_i\beta_2+a3_i\beta_3+a4_i\beta_4$$

为了构成交叉及嵌套效应，可以简单地列出主效应的所有组合。详见GLM过程中参数化一节。