试验设计基础

石 磊　王学仁　孙文爽　编著

重庆大学出版社

试验设计基础

石 磊 王学仁 孙文爽 编著

重庆大学出版社

内容提要

本书从理论和实践应用上介绍试验设计的原理和方法，内容包括：线性模型理论，多项式回归及正交多项式，模型诊断及最优统计设计准则，单因素及双因素试验的方差分析模型，析因试验，回归的正交设计，直接择优试验及稳定性择优试验并介绍了正交表的使用。

本书适合作为大学概率论与数理统计、统计学、管理学等专业教材。也适合工科类及财经类有关专业的本科生及研究生阅读。同时可供工程、农林、生物及医学等领域的教师和科技人员参考。

试验设计基础

石　磊　王学仁　孙文爽　编著

责任编辑　肖顺杰

*

重庆大学出版社出版发行

新　华　书　店　经　销

重庆建筑大学印刷厂印刷

*

开本：850×1168　1/32　　印张：9.125 字数：245 千

1997 年 2 月第 1 版　　1997 年 2 月第 1 次印刷

印数：1—2 000

ISBN 7-5624-1200-6/O・134　　定价：12.50 元

（川）新登字 020 号

前　言

试验设计是一门研究如何正确地安排试验和分析数据并以较少的试验得到最佳的试验结果的科学。它是统计学中一个最基本的分支，可以说统计学的最初发展与试验设计的出现是分不开的。

随着科学技术的进步，试验设计的理论和方法由最初的农业、生物及医学试验逐渐发展成一个在各个领域都具有广泛应用价值的统计学科。特别是在产品设计、质量管理及经济学等领域中的广泛应用越来越受到人们的重视。曾有这样一种说法，战后日本经济的腾飞，百分之十的功劳应归功于试验设计及正交表方法在经济及工业领域中的推广使用。由此可见试验设计方法在国民经济中所起的作用。

在国外，试验设计作为一门课程是大学统计学专业学生的必修课，其它专业如医学、生物、化学及工程等专业的学生也把它作为一门选修课。在国内，一些大学的统计学专业也相继开设了这门课程。在应用领域，正交试验法在全国得到了一定的普及和推广，取得了许多应用成果。但是，目前既能适合本科学生的教学要求，又能为工程技术等领域的科技人员提供较好参考的教材还不多。近年来，作者多次为统计学专业的本科学生和研究生讲授试验设计的课程，同时在一些应用领域向科技人员介绍过这方面的知识。在此基础上，通过不断地修改和补充，本着理论和实际相结合的原则，编写了这本书。

本书共分十章。第一章介绍线性模型的基本理论；第二章介绍多项式回归及正交多项式；第三章介绍模型假定的偏离及诊断、最优统计设计问题，并给出了几种最优的统计设计准则；第四章及第五章是方差分析的内容；第六章介绍析因试验，讲述两水平因素的全面试验、部分试验及区组试验的分析方法；第七章是回归设计的

内容；第八章及第九章介绍直接择优试验，其中第九章介绍了三次设计及正交表的使用；第十章是有关稳定性择优的内容。

前八章在理论上比较成熟，后两章的内容在应用上非常活跃，其理论的研究正在探讨之中，因而我们主要介绍其方法和应用实例，理论上不过多涉及。对非统计学专业的工程技术人员可根据需要选读部分章节。本书最后给出了附录 A、B、C、D、E。分别是矩阵知识、正交投影知识、随机变量及其分布、常用的正交表、常用统计用表。

本书在选材上尽量简明扼要，略去了一些较复杂的内容，但保留了一些重要的理论结果，使之自成体系。在每一章之后，附有一定量的习题，旨在巩固和加深所学知识。

在写作过程中，我们参阅了该领域国内外的一些著述及文献，引用了部分实例，在此向原作者、译者表示衷心的感谢。由于作者水平有限，疏漏和错误在所难免，恳请广大读者批评指正。

编著者
于云南大学统计系
云南省应用数学研究所
1996 年 3 月 28 日

目 录

第一章 线性模型及其理论 …… 1
§1.1 线性回归模型 …… 1
§1.2 最小二乘估计理论 …… 4
§1.3 线性等式约束下的最小二乘估计理论 …… 12
§1.4 降秩线性模型参数的估计理论 …… 14
§1.5 线性模型极大似然估计理论 …… 18
§1.6 置信椭球及联合(同时)置信区间 …… 21
§1.7 线性假设的检验问题 …… 25
综合练习一 …… 29
第二章 多项式回归及正交多项式 …… 32
§2.1 多项式回归 …… 32
§2.2 多项式回归存在的问题 …… 33
§2.3 正交多项式 …… 34
§2.4 正交多项式的构造 …… 36
§2.5 正交多项式回归的应用实例 …… 42
综合练习二 …… 44
第三章 模型诊断及最优统计设计准则 …… 45
§3.1 模型假定的偏离 …… 45
§3.2 协方差矩阵假定的偏离 …… 47
§3.3 正态假定的偏离 …… 49
§3.4 残差分析及模型诊断 …… 51
§3.5 纯误差及拟合损失 …… 60
§3.6 最优统计设计准则 …… 62
综合练习三 …… 64
第四章 单因素试验的方差分析模型 …… 67
§4.1 单因素试验的回归表示 …… 67
§4.2 单因素试验的方差分析模型 …… 70
§4.3 参数的估计及平方和分解 …… 71
§4.4 效应的显著性检验 …… 73
§4.5 等重复试验时的一些结果 …… 76
§4.6 多重比较法 …… 79
综合练习四 …… 83
第五章 双因素试验的方差分析模型 …… 85
§5.1 效应的定义及模型的引入 …… 85
§5.2 可加主效应模型的等重复试验 …… 87
§5.3 可加主效应模型的不等重复试验 …… 92
§5.4 交互效应模型试验 …… 96
§5.5 相关问题的讨论 …… 99

综合练习五 …… 102
第六章 析因试验 …… 104
§6.1 两水平因子试验 …… 105
§6.2 两水平因子试验分析 …… 107
§6.3 两水平因子试验的方差模型 …… 112
§6.4 2^m 型试验的回归模型及效应的区间估计 …… 114
§6.5 两水平区组试验 …… 116
§6.6 2^m 型部分试验及其分析 …… 120
综合练习六 …… 134
第七章 回归设计 …… 137
§7.1 什么是回归设计 …… 137
§7.2 一次回归的正交设计 …… 138
§7.3 回归的复合设计 …… 146
§7.4 变换模型及其回归设计 …… 157
综合练习七 …… 163
第八章 直接择优——快速登高法 …… 167
§8.1 经验建模及响应曲面 …… 167
§8.2 响应曲面的直接择优:快速登高法 …… 170
§8.3 寻求最优设计点的实际应用 …… 172
§8.4 带有约束条件下的快速登高法 …… 177
§8.5 二次回归曲面的直接择优 …… 181
§8.6 关于快速登高法的几点注释 …… 183
综合练习八 …… 183
第九章 直接择优试验及三次设计 …… 185
§9.1 正交表介绍 …… 185
§9.2 如何运用正交表安排试验 …… 189
§9.3 三次设计及正交表的使用 …… 190
§9.4 实例分析 …… 198
综合练习九 …… 203
第十章 稳定性择优试验 …… 205
§10.1 稳定性择优及衡量指标 …… 205
§10.2 静态特性的参数设计 …… 208
§10.3 动态特性的参数设计 …… 217
综合练习十 …… 230
附录 …… 232
附录 *A* 矩阵知识初步 …… 232
附录 *B* 正交投影 …… 235
附录 *C* 随机变量及其分布 …… 236
附录 *D* 常用正交表 …… 240
附录 *E* 常用统计用表 …… 252
参考文献 …… 282

第一章　线性模型及其理论

§1.1　线性回归模型

在科学试验及统计分析中，人们常常需要研究两个量之间的关系。例如，人的身高与体重，收入与智商，一定体积气体的温度与压力等等，以此了解其中一个量的变化对另一个量所产生的影响。如果设这两个量分别为 x 和 y，且有 n 对这样的观测值 (x_i, y_i)，$i=1,\cdots\cdots,n$。一个自然的想法是把这些点点在坐标纸上，画出所谓的散点图，这样从图中可以找出它们应该服从的某种直线或曲线关系。但通常情况下，这种关系并不能完全拟合数据点，由于变量的观测受到许多不可控因素的影响，它们的观测值都会发生随机波动，加之测量误差的存在，这种波动是无法避免的，因此，x 和 y 表现出来的关系往往是非确定性的(近似的)。

在有些情况下，x 和 y 的关系可以通过其理论根据来确定，也可以通过 x 与 y 的散点图进行选择。以下用几个例子来说明。

例 1.1.1　理论化学指出，对于保持恒温的一定量的气体，其体积 V 和压力 P 近似满足关系：$PV^{\gamma}=C$，其中 $\gamma\neq 1$。若取对数则可以得到：

$$\log P = \log C - \gamma \log V$$

令 $y=\log P$，$x=\log V$，$\beta_0=\log C$，$\beta_1=-\gamma$，则 y 与 x 应满足如下的线性关系：

$$y = \beta_0 + \beta_1 x$$

如果做了 n 次试验，得到 n 对观测数据：(y_i, x_i)，$i=1,\cdots,n$，则可以用如下的统计模型来描述 x_i 与 y_i 之间的关系：

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1,\cdots,n$$

这里 ε_i 表示随机误差。只要能从上述模型中估计出 β_0 及 β_1,γ 及 C 的值也就可以确定了。

例 1.1.2　在地质数据的分析中，了解某矿体关系密切的 Pt＋Pd 与 Cu＋Ni 的关系，以便根据 Cu＋Ni 的变化规律研究贵金属 Pt＋Pd 的变化规律。为此取了 20 个样品，其数据散点图如图 1.1 所示。显然数据表明简单的直线回归是不合理的。令 y＝Pt＋Pu，x＝Cu＋Ni，由经验下述模型是比较合理的：

$$y_i = \beta_0 + \beta_1/x_i + \varepsilon_i$$

这里 ε_i 表示误差所引起的波动。用上述模型进行统计分析及预测也取得了较好的效果。这是一个通过数据及经验建立模型的例子。

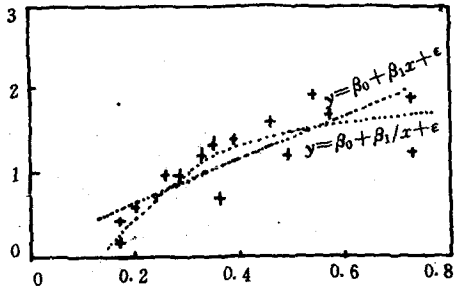

图 1.1　例 1.1.2 数据散点图及拟合曲线

在试验设计中，x 常常称之为因素，y 称之为试验指标观测值。人们所要了解的是因素对指标值的影响大小，因此也就要确定因素和指标值之间的函数关系模型。在通常情况下，因素不仅仅只有一个。例如，在研究作物产量与作物品种，播种密度及播种日期的关系的农业试验中，作物产量是试验指标，作物品种，播种密度及播种日期均是试验因素。

用数学的语言来描述，令 $\xi=(\xi_1,\cdots,\xi_K)'$ 表示 K 个因素，η 表示指标的理论值，则可以假定 η 与 ξ 之间存在如下的函数关系：

$$\eta = F(\xi) \qquad (1.1.1)$$

这里 $F(\cdot)$ 是某种函数关系。在实际中，$F(\cdot)$ 常常是未知的，为

了理论处理及计算上的方便，假定 $F(\xi)$ 属于某个带有未知参数 $\beta=(\beta_0,\beta_1,\cdots,\beta_P)'$ 的函数族，也即 $\eta=F(\xi,\beta)$。如果 $F(\xi,\beta)$ 的函数形式已知，则确定 $F(\xi,\beta)$ 的问题就转化为确定 β 的问题。为此，取定因素 ξ 的取值进行试验，就得到指标 η 的观测值 y。由于试验受到随机因素的干扰，观测值 y 常常不等于指标值 η，它们之间存在一定的误差，即：

$$y = F(\xi,\beta) + \varepsilon \qquad (1.1.2)$$

其中 ε 称为试验误差或误差变量。(1.1.1)称为理论模型，而(1.1.2)称为数据结构模型或统计模型。

如果 $F(\xi,\beta)$ 可用一组已知函数

$$\psi_0(\xi),\psi_1(\xi),\cdots,\psi_p(\xi)$$

近似展开，即

$$F(\xi,\beta) \approx \sum_{j=0}^{p}\psi_j(\xi)\beta_j \qquad (1.1.3)$$

则此时由(1.1.2)及(1.1.3)确定的模型就是一个线性模型。

例如，在单因素情形(即 $K=1$)，若取

$$\psi_0(\xi)=1,\psi_1(\xi)=\xi,\cdots,\xi_p(\xi)=\xi^P$$

则由(1.1.3)确定的近似模型就是 ξ 的 p 阶多项式近似。特别地，若取 $p=1$，则 $F(\xi,\beta)=\beta_0+\xi\beta_1$ 就是例 1.1 中所使用的最简单的线性模型。需要提醒注意的是：这里所说的线性是对 $\beta_j(j=0,\cdots,p)$ 而言，而不是对因素变量 ξ。

为了对模型进行验证及给出其未知参数 β 的估计，需要在不同的试验点上取值进行试验，从而得到一系列不同的观测值。设 $\xi_1,\cdots,\xi_n$ 代表不同的试验条件，相应的观测值记为 $y_1,\cdots,y_n$，令 $x_i=(\psi_0(\xi_i),\psi_1(\xi_i)\cdots,\psi_p(\xi_i)),i=1,\cdots,n$，则由(1.1.2)及(1.1.3)得到如下的 n 个方程组：

$$y_i = x'_i\beta + \varepsilon_i \qquad (1.1.4)$$

这里 $\varepsilon_1,\cdots,\varepsilon_n$ 代表 n 个随机误差，如令 $X'=(x_1,\cdots,x_n),Y=(y_1,\cdots,y_n)',\varepsilon=(\varepsilon_1,\cdots,\varepsilon_n)'$，则(1.1.4)可写成如下的矩阵形式：

$$Y = X\beta + \varepsilon \qquad (1.1.5)$$

这就是通常研究的线性回归模型的矩阵形式。这里 X 称为模型设计阵，它在试验设计中占有重要的地位。X 的结构反映了模型的布点情况，同时也反映出设计的好坏。

§1.2 最小二乘估计理论

这一节中，讨论线性模型(1.1.5)中未知参数 β 的估计问题。为了进一步研究 β 的估计的性质，对误差变量 $\varepsilon=(\varepsilon_1,\cdots,\varepsilon_n)'$ 做一定的假定，即假设 $\varepsilon_i(i=1,\cdots,n)$ 是一组相互独立同分布的随机变量，其前两阶矩存在，分别为

$$E\varepsilon_i = 0, V(\varepsilon_i) = \sigma^2, i = 1,\cdots,n \qquad (1.2.1)$$

其中 $\sigma^2>0$ 为未知参数。

由于误差变量 $\varepsilon_i(i=1,\cdots,n)$ 是由试验中一些无法控制的随机因素产生的波动引起的，因此一种估计 β 的想法是使试验误差 $\varepsilon_i(i=1,\cdots,n)$ 尽可能的达到极小。令

$$Q(\beta) = \sum_{i=1}^{n}\varepsilon_i^2 = \varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta)$$

则 $Q(\beta)$ 度量了 $\varepsilon_i(i=1,\cdots,n)$ 的大小。$Q(\beta)$ 越小，表明模型拟合得越好，所谓最小二乘估计法，就是通过极小化 $Q(\beta)$ 求出 β 的值。

1.2.1 参数的估计及其性质

定义 1.2.1 在模型(1.1.5)中，若 Y 的线性函数 $\hat{\beta}$ 满足：

$$Q(\hat{\beta}) = \min_{\beta} Q(\beta)$$

则称 $\hat{\beta}$ 为 β 的最小二乘估计，记为 LSE。

定理 1.2.1 在模型(1.1.5)中，若秩$(X)=p+1$，则

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

证明 由于 $\partial y'X\beta/\partial\beta = X'Y$，　$\partial\beta'X'X\beta/\partial\beta = 2X'X\beta$

从而有：$\partial Q(\beta)/\partial\beta = -2X'y + 2X'X\beta$，令其为 0，可得

$$X'X\beta = X'Y \qquad (1.2.2)$$

由于 $(X'X)$ 可逆，从而有 $\hat{\beta}=(X'X)^{-1}X'Y$。(1.2.2)称为求解 β 的

正规方程。以下证明 $\hat{\beta}$ 确实使 $Q(\beta)$ 达到极小：

$$Q(\beta)=(Y-X\beta)'(Y-X\beta)=(Y-X\hat{\beta})'(Y-X\hat{\beta})$$
$$+(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta)\geqslant(Y-X\hat{\beta})'(Y-X\hat{\beta})$$
$$=Q(\hat{\beta})$$

故 $\hat{\beta}$ 确实使 $Q(\beta)$ 达到极小。从而定理得证。

$\hat{Y}=X\hat{\beta}$ 称为回归值的拟合值，$e=Y-\hat{Y}$ 称为残差向量。令 $P_x=X(X'X)^{-1}X'$，则由附录 B.6 知，P_x 是 $\mathscr{R}(X)$ 空间的投影矩阵。这里 $\mathscr{R}(X)$ 表示由 X 的列向量所张成的线性空间。因此 $\hat{Y}=P_xY$ 是 Y 在 $\mathscr{R}(X)$ 上的正交投影，而残差 $e=(I_n-P_x)Y$ 则是 Y 在 $\mathscr{R}^{\perp}(X)$ 上的正交投影。从而 Y 有如下的正交分解：

$$Y=\hat{Y}+e \tag{1.2.3}$$

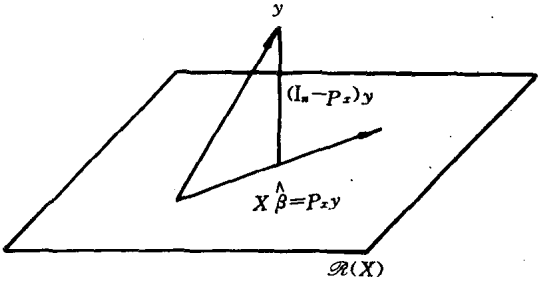

图 1.2　LSE 的几何表示

(1.2.3)的分解过程可以用图 1.2 直观地表示出来。由(1.2.3)可得：

$$Y'Y=\hat{Y}'\hat{Y}+e'e$$

$e'e=\hat{Y}'(I_n-P_x)Y$ 称为残差平方和，可以用来度量模型拟合的好坏。$Y'Y$ 称为总平方和，$\hat{Y}'\hat{Y}$ 称为回归平方和。令 $SS_T=Y'Y$，SS_R

$=\hat{Y}'\hat{Y}=\hat{\beta}(X'X)\hat{\beta}$，则有：

$$SS_T = SS_R + SS_e \tag{1.2.4}$$

(1.2.4)式即通常所说的平方和分解。总平方和可以分解为两部分，残差平方和及回归平方和。SS_R 反映了回归项的显著大小，而 SS_e 则反映了误差项的贡献大小。

在模型(1.1.5)及(1.2.1)中，除回归系数 β 之外，还有另外一个参数，即误差方差 σ^2，由于

$$\begin{aligned} E(SS_e) &= E((Y - X\hat{\beta})'(Y - X\hat{\beta})) = E(Y'(I_n - P_x)Y) \\ &= E[tr((I_n - P_x)YY')] \\ &= tr[(I_n - P_x)EYY'] \\ &= tr[(I_n - P_x)(\sigma^2 I_n + X\beta\beta'X')] \end{aligned}$$

因为 $(I_n - P_x)X = 0$，从而(由附录 A.2)有

$$\begin{aligned} E(SS_e) &= tr[(I_n - P_x)\sigma^2] \\ &= (n - p - 1)\sigma^2 \end{aligned} \tag{1.2.5}$$

如令 $\hat{\sigma}^2 = SS_e/(n-p-1) = Y'(I_n - P_x)Y/(n-p-1)$，则有 $E(\hat{\sigma}^2) = \sigma^2$。这说明 $\hat{\sigma}^2$ 是 σ^2 的一个无偏估计。回过头来，$\hat{\beta}$ 也具有无偏的性质：

$$E(\hat{\beta}) = (X'X)^{-1}X'EY = (X'X)^{-1}X'X\beta = \beta$$

因此 β 的 LSE 也是 β 的一个无偏估计。容易求得 $\hat{\beta}$ 的协方差结构为：

$$\begin{aligned} \mathrm{Cov}(\hat{\beta}) &= \mathrm{Cov}((X'X)^{-1}X'Y) \\ &= \sigma^2(X'X)^{-1}(X'X)(X'X)^{-1} = \sigma^2(X'X)^{-1} \end{aligned}$$

因此，若 $\hat{\beta}_i$ 为 $\hat{\beta}$ 的第 i 个元素，则有

$$V(\hat{\beta}_i) = \sigma^2(X'X)^{-1}_{ii}$$

这里 $(X'X)^{-1}_{ii}$ 表示矩阵 $(X'X)^{-1}$ 的第 i 个对角元素。$[\sigma^2(X'X)^{-1}_{ii}]^{\frac{1}{2}}$ 称为 $\hat{\beta}_i$ 的标准差，它反映了 β_i 的估计的精度。如果 σ^2 未知，可用 σ^2 的一个估计，例如 $\hat{\sigma}^2$ 代替 σ^2，就得到 $\hat{\beta}_i$ 的标准差的估

计，标准差的估计也称为标准误差，简记为 $s.e.$。$s.e.$ 越小，说明该估计的精度越高。

现在提出的一个问题是：为什么选择 $\hat{\beta}$ 作为 β 的估计，而不选择其它的估计呢？这是因为 $\hat{\beta}$ 作为 β 的估计具有很好的优良性质，即 $\hat{\beta}$ 作为 β 估计，将使得它在 β 的一切线性无偏估计类中具有最小的方差（当然要假定模型(1.1.5)是正确的）。

定理 1.2.2　设 $\psi=c'\beta$ 为 β 的任一线性函数，c 为 $(p+1)$ 维常数向量。则在 $\psi=c'\beta$ 的一切线性无偏估计类中，$\hat{\psi}=c'\hat{\beta}$ 是唯一的具有最小方差的估计，称 $c'\hat{\beta}$ 为 $c'\beta$ 的最佳线性无偏估计，简记为 BLUE。

证明　设 $a'Y$ 为 ψ 的任一无偏估计，a 为 $(p+1)$ 维向量，则有：

$$V(a'Y)=a'\mathrm{Cov}(Y)a=\sigma^2a'a,E(a'Y)=a'X\beta=c'\beta$$

因此　$c'=a'X$，从而有

$$\begin{aligned}V(a'Y)-V(c'\hat{\beta})&=\sigma^2a'a-\sigma^2c'(X'X)^{-1}c\\&=\sigma^2a'a-\sigma^2a'P_xa=\sigma^2(a'a-a'P_xa)\\&=\sigma^2a'(I_n-P_x)a\end{aligned}$$

由于 (I_n-P_x) 为非负定矩阵，因而 $a'(I_n-P_x)a\geqslant 0$。

从而有：$V(a'Y)-V(c'\hat{\beta})\geqslant 0$。定理得证。

在定理 1.2.2 中，令 $c_i=(0,\cdots,0,1,0,\cdots 0)'$ 表示第 i 个元素为 1，其余元素为 0 的 $p+1$ 维向量，则 $c_i'\beta=\beta_i,i=1,\cdots,n$。因此在线性无偏估计类里，$\hat{\beta}_i$ 是 β_i 的最小方差估计。

1.2.2　平方和分解及方差分析表

在前面讲到了回归模型的平方和及其分解式(1.2.4)，平方和常常与一个称为自由度的数值联系在一起。自由度表示平方和中相互独立的变量个数，它对确定平方和的分布是一个很重要的量。在以后的各章节还要逐步讨论。在(1.2.4)的分解式中，SS_T，SS_R，及 SS_e 的自由度分别为 n，$p+1$，及 $n-p-1$。很明显，如果用 f_T，

f_R 及 f_e 分别表示 SS_T,SS_R 及 SS_e 的自由度,则对应于(1.2.4),自由度也有相应的分解,即

$$f_T = f_R + f_e$$

单位自由度各平方和的贡献称之为均方和,因而各平方和的均方和定义为:

$$MS_T = \frac{SS_T}{f_T}, MS_R = \frac{SS_R}{f_R}, MS_e = \frac{SS_e}{f_e}$$

如果所拟合的模型比较显著,则 MS_R 与 MS_e 相比,应该比较显著。因此可以用 MS_R 与 MS_e 的比值来刻画模型的显著程度,即

$$F = \frac{MS_R}{MS_e}$$

上式就是以后将要提到的 F 统计量。将平方和及 F 值制成一个表格,称为方差分析表,如表 1.1 所示。在表 1.1 中,为了确定回归模型是否显著,需要确定 F 统计量的一个临界值,这就要求给出 F 统计量的分布,这一内容将在 § 1.7 中讨论。

表 1.1 回归模型的方差分析表

方差来源	平方和	自由度	均方和	F 值
回归	SS_R	$p+1$	MS_R	$F=\frac{MS_R}{MS_e}$
残差	SS_e	$n-p-1$	MS_e	
总和	SS_T	n	MS_T	

在上面所进行的平方和分解中,如果回归模型中存在常数项,比如在(1.1.4)中,$\psi_0(\xi_i)=1, i=1,\cdots,n$。则回归模型变为:

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i \qquad (1.2.6)$$

其中 $x_{ij}=\psi_j(\xi_i), i=1,\cdots,n, j=1,\cdots,p$。此时由(1.2.4)定义的回归平方和 SS_R 中包含了 $\beta_0,\beta_1,\cdots\beta_p$ 效应的影响。在实际中,人们主要关心的是 $\psi_1(\xi_i),\cdots,\psi_p(\xi_i)$ 回归项的显著性程度。因此,希望回归平方和中只含有 $\beta_1,\cdots\beta_p$ 效应的影响,也即要把 β_0 效应从中分解出来。

令 $X=(1_n, X_1)$，其中 $1_n=(1,\cdots,1)'$，$X_1=(x_{ij})_{n\times p}$，将模型(1.1.5)重新表成如下形式：

$$Y = 1_n\nu_0 + Z\nu + \varepsilon \tag{1.2.7}$$

其中 $\nu_0=\beta_0+\bar{x}'\nu$，$Z=X_1-1_n\bar{X}'$，$\bar{X}=X'_1 1_n/n$，$\nu=(\beta_1,\cdots,\beta_p)'$，由于 $1_nZ=0$，容易证明此时(1.2.7)中参数的 LSE 为：

$$\hat{\nu}_0 = \hat{\beta}_0 + \bar{x}'\hat{\nu} = \bar{y}$$

$$\hat{\nu} = (Z'Z)^{-1}Z'Y \tag{1.2.8}$$

其中 $\bar{y}=\frac{1}{n}\sum_{j=1}^{n} y_j$，令 $\hat{Y}$ 表示其拟合值，则有

$$\hat{Y} = X\hat{\beta} = 1_n\hat{\nu}_0 + Z\hat{\nu} = 1_n\bar{y} + (X_1 - 1_n\bar{X}')\hat{\nu} \tag{1.2.9}$$

因此

$$(\hat{Y} - 1_n\bar{y}) = Z\hat{\nu} \tag{1.2.10}$$

这样(1.2.10)的右边只与 $\hat{\beta}_1,\cdots,\hat{\beta}_p$ 有关。残差平方和为：

$$(Y - \hat{Y})'(Y - \hat{Y})$$

$$= (Y - 1_n\bar{y})'(Y - 1_n\bar{y}) - \hat{\nu}'Z'Z\hat{\nu}$$

也即：

$$(Y - 1_n\bar{y})'(Y - 1_n\bar{y})$$

$$= \hat{\nu}'Z'Z\hat{\nu} + (Y - \hat{Y})'(Y - \hat{Y}) \tag{1.2.11}$$

(1.2.11)式左边一项称之为校正的总平方和，$\hat{\nu}'Z'Z\hat{\nu}$ 称为校正的回归平方和，把它们分别记为 SS_T^* 及 SS_R^*，则平方和分解式为：

$$SS_T^* = SS_R^* + SS_e \tag{1.2.12}$$

在校正的回归平方和 SS_R^* 中，消除了 $\hat{\beta}_0$ 的影响，实质上这一项被并入到总平方和中。在计算上，可以采用如下的方便形式：

$$SS_T^* = Y'Y - n\bar{y}^2, SS_R^* = \hat{\beta}'X'X\hat{\beta} - n\bar{y}^2 \tag{1.2.13}$$

如果在模型(1.1.5)中，$1'_nX_1=0$，则此时参数的估计形式就变得很简单：

$$\hat{\beta}_0 = \bar{y}, \hat{\nu} = (\hat{\beta}_1,\cdots,\hat{\beta}_p)' = (X_1X_1)^{-1}X'_1Y$$

校正的回归平方和 $SS_R^* = \hat{\nu} X'_1 X_1 \hat{\nu}$。在使用上，一般先将 $x_i = (\psi_i(\xi_1), \cdots \psi_i(\xi_n))$ 减去其平均值 $\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} \psi_i(\xi_j)$ 后参加回归，则此时就有 $1'_n X_1 = 0$。由(1.2.12)得到的平方和及 F 值相应地可以制成一方差分析表，如表 1.2 所示：

表 1.2　具有常数项回归模型的方差分析表

方差来源	平方和	自由度	均方和	F 值
回归	$\hat{\beta} X' X \hat{\beta} - n\bar{y}^2$	p	$\frac{(\hat{\beta} X' X \hat{\beta} - n\bar{y}^2)}{p}$	$F = \frac{(\hat{\beta}' X' X \hat{\beta} - n\bar{y})^2/p}{(Y-\hat{Y})'(Y-\hat{Y})/(n-p-1)}$
残差	$(Y-\hat{Y})'(Y-\hat{Y})$	$n-p-1$	$\frac{(Y-\hat{Y})'(Y-\hat{Y})}{(n-p-1)}$	
总和	$Y'Y - n\bar{y}^2$	$n-1$		

在以后的分析中，如果回归模型中有常数项，其方差分析表均采用表 1.2 的形式。如果回归项中没有常数项，如在(1.1.5)中回归项有 $p+1$ 项，则采用表 1.1 的形式。

例 1.2.1　单变量的直线回归

在某些试验中，影响试验指标值的因素只有一个，记为 x。这时指标值的观测值 y 与 x 之间最常用的回归模型是直线回归。设在 $x_1, \cdots, x_n$ 上进行试验，相应的观测值为 $y_1, \cdots, y_n$。则有如下的直线回归模型：

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \cdots, n \tag{1.2.14}$$

容易算出 β_0 及 β_1 的 LSE 分别为

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) / \sum_{i=1}^{n} (x_i - \bar{x})^2 \end{aligned} \tag{1.2.15}$$

其中 $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j, \bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_i$。$\hat{\beta}_1$ 的标准误差为

$$s.e.(\hat{\beta}_1) = S / (\sum_i (x_i - \bar{x})^2)^{\frac{1}{2}}$$