SPT

非数学专业

实用统计方法

西安交通大学
梅长林 周家良 编著

科学出版社

21世纪高等院校教材(非数学专业)

实用统计方法

西安交通大学

梅长林　周家良　编著

科学出版社

2002

内 容 简 介

本书是在高等院校非数学类专业概率论与数理统计必修课内容基础上编写的以介绍实用数理统计方法为目的的教材.内容包括:多元回归分析、主成分分析及典型相关分析、判别分析、聚类分析、非参数秩方法、列联表的独立性分析、试验设计、抽样调查.并针对本书内容,在附录中对国际先进的SAS软件作了简要介绍.

本书可作为工程类、医学类、财经及管理类各有关专业高年级学生的选修课教材或作为非数学类硕士研究生的数理统计教材,也可作为数理统计应用工作者的参考书籍.

图书在版编目(CIP)数据

实用统计方法/梅长林,周家良编著. —北京:科学出版社,2002
(21世纪高等院校教材(非数学专业))
ISBN 7-03-009756-4

Ⅰ.实… Ⅱ.①梅…②周… Ⅲ.数理统计-高等学校-教材 Ⅳ.0212

中国版本图书馆CIP数据核字(2001)第076299号

科学出版社 出版
北京东黄城根北街16号
邮政编码:100717
http://www.sciencep.com
双青印刷厂 印刷
科学出版社发行 各地新华书店经销
*
2002年2月第 一 版 开本:720×1000 1/16
2002年2月第一次印刷 印张:22 1/4
印数:1—3 000 字数:424 000
定价: 30.00元
(如有印装质量问题,我社负责调换〈环伟〉)

前　言

数理统计是研究如何有效地收集、整理和分析受随机因素影响的数据，对所考察的问题作出推断，进而为制定决策和采取行动提供科学依据的一门学科.随着计算机的普及和发展，数理统计方法的应用不但日益深入到工业、农业、国防、经济、管理、医学、社会、地质、气象等各个领域，而且也越来越受到各行各业的普遍重视.目前，我国高校的许多非数学类专业也开设了概率论与数理统计必修课，但重点只在介绍概率论和数理统计的基础知识和基本方法，与将统计方法广泛应用于实际尚有一定的距离.因此，让学生进一步学习和掌握一些近代实用的统计方法是面向21世纪培养人才的一个重要方面.为此目的，在原国家教委“面向21世纪工科数学教学内容和课程体系改革”立项课题中要求为非数学类专业本科生编写一本“以方法为主，不追求理论的系统性和完整性，方法要注意实用性和先进性，结构模块化，便于选用”的数理统计方法选修课教材.本书正是应此要求而编写的.

在本书的内容选择和写作手法上力求体现上述编写要求，各章自成体系.另外，在附录中对当前国际上先进的SAS软件作了简明介绍，并针对书中内容，叙述了调用各方法的SAS主要语句及对输出结果的适当解释，使学生能初步了解SAS系统的使用方法，为今后进一步深入学习开启了窗口.各章的习题一般也有两类，一类是基本习题，目的在于理解掌握所学的基本内容和方法，另一类是需要借助SAS软件利用计算机进行计算和分析的综合性习题.经过试用，我们给出各章的参考教学时数如下：

第一章：多元回归分析，8学时.

第二章：主成分分析及典型相关分析，6学时.

第三章：判别分析，6学时.

第四章：聚类分析，4学时.

第五章：非参数秩方法，8学时.

第六章：列联表的独立性分析，6学时.

第七章：试验设计，4学时.

第八章：抽样调查，4学时.

讲完全部内容大约需要46学时，附录内容可结合课后作业以学生自学为主.由于内容的模块化结构，不同类型的专业可选择不同的模块.下面是几个

供参考的模块：

工程类：第一、三、四、五、七章或再加第二章，共计约 30 或 36 学时.

经济、管理类：第一、三、四、五、六、八章，共计约 36 学时.

社科类：第一、五、六、八章，共计约 26 学时.

医学类：第一、三、四、五、六章，共计约 32 学时.

以上教学时数及内容模块仅供参考，各专业可适当予以调整，使教学内容与时数安排更适合本专业的特点及要求.

本书的第一至第六章及附录由梅长林编写，第七、八章由周家良编写，并由梅长林统稿.限于作者水平，书中难免有不妥和错误之处，敬请广大读者提供宝贵的批评和建议，使本教材不断得以完善.另外，对于引用了其中习题或例题的有关书籍，我们均列入书末的参考文献中，在此特向其作者表示衷心的感谢.

编者

2001 年 2 月

目　录

第一章　多元回归分析

回归分析是应用极其广泛的数理统计方法之一.它基于观测数据建立变量间适当的依赖关系,以分析数据的内在规律,并可用于预报、控制等问题.在数理统计基础部分,我们已学习了一元线性回归分析的基本内容,即当影响因变量 Y 的因素只有一个(记为 X)时,如何建立 Y 与 X 的适当的线性回归关系.在实际问题中,影响 Y 的因素往往很多,设有 $X_1, X_2, \cdots, X_{p-1}$ 共 $p-1$ 个,建立这 $p-1$ 个因素与 Y 的依赖关系将具有更广泛的应用价值.本章讨论多元线性回归模型的系统建模方法,主要包括模型的参数估计、假设检验、残差分析以及最优回归方程的选取等.

§1.1　多元线性回归模型

1.1.1　多元线性回归模型及其矩阵表示

设 Y 是一个可观测的随机变量,它受到 $p-1$ 个非随机因素 $X_1, X_2, \cdots, X_{p-1}$ 和随机因素 ε 的影响.若 Y 与 $X_1, X_2, \cdots, X_{p-1}$ 有如下线性关系:

$$Y = \beta_0 + \beta_1 X_1 + \beta X_2 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon, \tag{1.1.1}$$

其中 $\beta_0, \beta_1, \cdots, \beta_{p-1}$ 是未知参数;ε 是均值为零方差为 $\sigma^2 > 0$ 的不可观测的随机变量,称为误差项,并通常假定 $\varepsilon \sim N(0, \sigma^2)$.该模型称为多元线性回归模型,且称 Y 为因变量,$X_1, X_2, \cdots, X_{p-1}$ 为自变量.

要建立多元线性回归模型,首先要估计未知参数 $\beta_0, \beta_1, \cdots, \beta_{p-1}$,为此我们进行 $n(n \geqslant p)$ 次独立观测,得到 n 组数据(称为样本)

$$(X_{i1}, X_{i2}, \cdots, X_{i,p-1}; Y_i), \qquad i = 1, 2, \cdots, n.$$

它们应满足(1.1.1),即有

$$\begin{cases} Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \cdots + \beta_{p-1} X_{1,p-1} + \varepsilon_1, \\ Y_2 = \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \cdots + \beta_{p-1} X_{2,p-1} + \varepsilon_2, \\ \quad \vdots \\ Y_n = \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \cdots + \beta_{p-1} X_{n,p-1} + \varepsilon_n. \end{cases} \tag{1.1.2}$$

其中 $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$ 相互独立且均服从 $N(0,\sigma^2)$ 分布.

令

$$\boldsymbol{Y}=\begin{bmatrix}Y_1\\Y_2\\\vdots\\Y_n\end{bmatrix}_{n\times 1},\quad \boldsymbol{X}=\begin{bmatrix}1 & X_{11} & X_{12} & \cdots & X_{1,p-1}\\1 & X_{21} & X_{22} & \cdots & X_{2,p-1}\\\vdots & \vdots & \vdots & & \vdots\\1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1}\end{bmatrix}_{n\times p},$$

$$\boldsymbol{\beta}=\begin{bmatrix}\beta_0\\\beta_1\\\vdots\\\beta_{p-1}\end{bmatrix}_{p\times 1},\quad \boldsymbol{\varepsilon}=\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_n\end{bmatrix}_{n\times 1}.$$

则(1.1.2)式可简写为如下的矩阵形式:

$$\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon},\tag{1.1.3}$$

其中 $\boldsymbol{Y}$ 称为观测向量,$\boldsymbol{X}$ 称为设计矩阵,它们是由观测数据得到的,是已知的.并假定 $\boldsymbol{X}$ 为列满秩的,即 $\mathrm{rank}(\boldsymbol{X})=p$.$\boldsymbol{\beta}$ 是待估计的未知参数向量,$\boldsymbol{\varepsilon}$ 是不可观测的随机误差向量.

(1.1.3)式称为多元线性回归模型的矩阵形式.

1.1.2 $\boldsymbol{\beta}$ 及 $\boldsymbol{\sigma^2}$ 的估计

1.$\boldsymbol{\beta}$ 的最小二乘估计

如果 Y 与 $X_1,X_2,\cdots,X_{p-1}$ 满足线性回归模型(1.1.1),则误差 ε 应是比较小的.因此,我们选择 $\boldsymbol{\beta}$ 使误差项的平方和

$$S(\boldsymbol{\beta})\triangleq\sum_{i=1}^{n}\varepsilon_i^2=\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}=(\boldsymbol{Y}-\boldsymbol{X\beta})^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{X\beta})=\sum_{i=1}^{n}\Big(Y_i-\sum_{j=0}^{p-1}X_{ij}\beta_j\Big)^2\tag{1.1.4}$$

达到最小,其中 $X_{i0}=1(i=1,2,\cdots,n)$.为此,将(1.1.4)分别对 $\beta_0,\beta_1,\cdots,\beta_{p-1}$ 求偏导并令其等于零,得

$$\frac{\partial S(\boldsymbol{\beta})}{\partial\beta_k}=-\sum_{i=1}^{n}\Big(Y_i-\sum_{j=0}^{p-1}X_{ij}\beta_j\Big)X_{ik}=0,\quad k=0,1,\cdots,p-1,$$

即

$$\sum_{i=1}^{n} Y_i X_{ik} = \sum_{i=1}^{n}\sum_{j=0}^{p-1} X_{ij}X_{ik}\beta_j = \sum_{j=0}^{p-1}\left(\sum_{i=1}^{n} X_{ij}X_{ik}\right)\beta_j, \quad k = 0,1,\cdots,p-1,$$

进一步可写为矩阵形式

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}, \tag{1.1.5}$$

称此方程为正规方程.

因为 $\mathrm{rank}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}) = \mathrm{rank}(\boldsymbol{X}) = p$,故$(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$存在.解正规方程即得 $\boldsymbol{\beta}$ 的最小二乘估计 $\hat{\boldsymbol{\beta}}$ 为

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}. \tag{1.1.6}$$

2. $\boldsymbol{\beta}$ 的最大似然估计

由于 $\varepsilon_i (i = 1,2,\cdots,n)$相互独立且均服从正态分布 $N(0,\sigma^2)$,则可由最大似然估计法估计 $\boldsymbol{\beta}$.这时 $Y_i (i = 1,2,\cdots,n)$相互独立,且 $Y_i \sim N(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}, \sigma^2)$.从而 $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_n)^{\mathrm{T}}$ 的似然函数为

$$L(\boldsymbol{\beta}) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1}X_{i,p-1})^2\right\}$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}S(\boldsymbol{\beta})\right\}.$$

求 $\hat{\boldsymbol{\beta}}$ 使 $L(\boldsymbol{\beta})$达到最大等价于使 $S(\boldsymbol{\beta})$达到最小.因此在 $\varepsilon_i \sim N(0,\sigma^2)(i = 1,2,\cdots,n)$之下,$\boldsymbol{\beta}$ 的最大似然估计和最小二乘估计是相同的,均为 $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$.

由(1.1.3)式可知,$\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$,故 $\mathrm{E}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{\beta}$,即 $\hat{\boldsymbol{\beta}}$ 为 $\boldsymbol{\beta}$ 的一个无偏估计.

当给出 $\boldsymbol{\beta}$ 的估计 $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_{p-1})^{\mathrm{T}}$ 后,将其代入(1.1.1)式并略去误差项,则得 $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_{p-1} X_{p-1}$为回归方程.利用回归方程,可由自变量 $X_1, X_2, \cdots, X_{p-1}$的观测值求出因变量 Y 的估计值.

3. 误差方差 σ^2 的估计

将自变量的各组观测值代入回归方程,可得因变量的各估计值(称为拟合

值)为 $\hat{\boldsymbol{Y}} \triangleq (\hat{Y}_1, \hat{Y}_2, \cdots, \hat{Y}_n)^{\mathrm{T}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. 称

$$\boldsymbol{e} \triangleq \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = [\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}]\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} \tag{1.1.7}$$

为残差向量,其中 $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$ 为 n 阶对称幂等矩阵,$\boldsymbol{I}$ 为 n 阶单位阵. 称数

$$\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e} = \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$$

为残差平方和.

由于 $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$ 且 $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X} = \boldsymbol{0}$,则 $\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e} = [\boldsymbol{Y} - \mathrm{E}(\boldsymbol{Y})]^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{H})[\boldsymbol{Y} - \mathrm{E}(\boldsymbol{Y})] = \boldsymbol{\varepsilon}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\varepsilon}$,由此可得

$$\begin{aligned}
\mathrm{E}(\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}) &= \mathrm{E}\{\mathrm{tr}[\boldsymbol{\varepsilon}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\varepsilon}]\} \\
&= \mathrm{tr}[(\boldsymbol{I} - \boldsymbol{H})\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}})] \\
&= \sigma^2 \mathrm{tr}[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}] \\
&= \sigma^2\{n - \mathrm{tr}[(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}]\} \\
&= \sigma^2(n - p),
\end{aligned}$$

其中 $\mathrm{tr}(\cdot)$ 表示矩阵的迹. 从而

$$\hat{\sigma}^2 \triangleq \frac{1}{n - p}\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e} \tag{1.1.8}$$

为 σ^2 的一个无偏估计.

1.1.3 有关的统计推断

1. 回归关系的统计推断

给定因变量 Y 与自变量 $X_1, X_2, \cdots, X_{p-1}$ 的 n 组观测值,利用前述方法可得到未知参数 $\boldsymbol{\beta}$ 和 σ^2 的估计,从而可给出 Y 与 $X_1, X_2, \cdots, X_{p-1}$ 之间的线性回归方程. 但所求得的回归方程是否有意义,也就是说 Y 与 $X_1, X_2, \cdots, X_{p-1}$ 之间是否存在显著的线性关系,还需要对回归方程进行检验.

(1) *建立方差分析表*

(i) 离差平方和的分解

我们知道观测值 $Y_1, Y_2, \cdots, Y_n$ 之所以有差异，是由下述两个原因引起的，一是当 Y 与 $X_1, X_2, \cdots, X_{p-1}$ 之间确有线性关系时，由于 $X_1, X_2, \cdots, X_{p-1}$ 取值的不同，而引起 Y_i 值的变化；另一方面是除去 Y 与 $X_1, X_2, \cdots, X_{p-1}$ 的线性关系以外的因素，如 $X_1, X_2, \cdots, X_{p-1}$ 对 Y 的非线性影响及随机因素的影响等. 记 $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$，则数据的总的离差平方和(Total Sum of Squares)

$$\mathrm{SST} \triangleq \sum_{i=1}^{n}(Y_i - \overline{Y})^2 \tag{1.1.9}$$

反应了数据 $Y_1, Y_2, \cdots, Y_n$ 波动性的大小.

残差平方和(Error Sum of Squares)

$$\mathrm{SSE} \triangleq \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{1.1.10}$$

反映了除去 Y 与 $X_1, X_2, \cdots, X_{p-1}$ 之间的线性关系(即 $\hat{Y}_i$)以外的因素引起的数据 $Y_1, Y_2, \cdots, Y_n$ 的波动. 若 SSE=0，则每个观测值可由线性关系精确拟合，SSE 越大，观测值和线性拟合值间的偏差也越大.

对于回归平方和(Regression Sum of Squares)

$$SSR \triangleq \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2, \tag{1.1.11}$$

由于可证明 $\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i = \overline{Y}$，故 SSR 反映了线性拟合值与它们的平均值的总偏差，即由变量 $X_1, X_2, \cdots, X_{p-1}$ 的变化所引起的 $Y_i(i=1,2,\cdots,n)$ 的波动. 若 SSR=0，则每个拟合值均相等，即 $\hat{Y}_i(i=1,2,\cdots,n)$ 不随 $X_1, X_2, \cdots, X_{p-1}$ 的变化而变化，这实质上反映了 $\beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$. 另一方面，经过代数运算及正规方程(1.1.5)可证明(证明从略)

$$\mathrm{SST} = \mathrm{SSE} + \mathrm{SSR}. \tag{1.1.12}$$

因此，SSR 越大，说明由线性回归关系所描述的 $Y_i(i=1,2,\cdots,n)$ 的波动性的比例就越大，即 Y 与 $X_1, X_2, \cdots, X_{p-1}$ 的线性关系就越显著.

另外，通过矩阵运算可证明 SST，SSE 和 SSR 有如下形式的矩阵表示：

$$\mathrm{SST} = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} - \frac{1}{n}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{J}\boldsymbol{Y} = \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Y}, \tag{1.1.13}$$

$$\mathrm{SSE} = \boldsymbol{e}^{\mathrm{T}}\boldsymbol{e} = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} - \hat{\boldsymbol{\beta}}\,\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}, \tag{1.1.14}$$

$$\mathrm{SSR} = \hat{\boldsymbol{\beta}}\,\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} - \frac{1}{n}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{J}\boldsymbol{Y} = \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Y}, \tag{1.1.15}$$

其中 $\boldsymbol{J}$ 表示一个元素全为 1 的 n 阶方阵.

(ii) 自由度的分解

对应于 SST 的分解(1.1.12),其自由度也有相应的分解.这里的自由度是指平方和中独立变化项的数目.在 SST 中,由于有一个关系式 $\sum_{i=1}^{n}(Y_i - \overline{Y}) = 0$,即 $Y_i - \overline{Y}(i = 1,2,\cdots,n)$彼此不是独立变化的,故其自由度为 $n-1$.

可以证明,SSE 的自由度为 $n-p$,SSR 的自由度为 $p-1$.因此对应于 SST 的分解(1.1.12),它们的自由度之间也有如下关系:

$$n - 1 = (n - p) + (p - 1). \tag{1.1.16}$$

(iii) 方差分析表

基于以上 SST 和其自由度的分解式(1.1.12)和(1.1.16)可建立如下的方差分析表:

表 1.1.1 方差分析表

方差来源	平方和(SS)	自由度(f)	均方(MS)
回 归	$SSR = \hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} - \frac{1}{n}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{J}\boldsymbol{Y}$	$p-1$	$MSR = \frac{SSR}{p-1}$
误 差	$SSE = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} - \hat{\boldsymbol{\beta}}\,\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$	$n-p$	$MSE = \frac{SSE}{n-p}$
总 和	$SST = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} - \frac{1}{n}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{J}\boldsymbol{Y}$	$n-1$	

其中回归平方和及残差平方和与各自的自由度之比分别称为均方回归(Regression Mean Square)及均方残差(Error Mean Square).利用方差分析表,可对回归方程的显著性做检验.

(2) 线性回归关系的显著性检验

为检验 Y 与 $X_1, X_2, \cdots, X_{p-1}$ 之间是否存在显著的线性回归关系,即检验假设

$$\begin{cases} H_0: \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0, \\ H_1: \text{至少有某个 } \beta_i \neq 0, \quad 1 \leqslant i \leqslant p-1. \end{cases} \tag{1.1.17}$$

这是因为若 H_0 成立，则 $Y=\beta_0+\varepsilon$，即 Y 与 $X_1,X_2,\cdots,X_{p-1}$ 之间不存在线性回归关系. 基于上述方差分析表，构造如下检验统计量

$$F \triangleq \frac{\mathrm{MSR}}{\mathrm{MSE}}, \tag{1.1.18}$$

当 H_0 为真时，可以证明 $F\sim F(p-1,n-p)$，这里 $F(p-1,n-p)$ 表示自由度为 $p-1$ 和 $n-p$ 的 F 分布. 由上述对回归平方和 SSR 的讨论可知，若 H_0 不真，F 的值有偏大的趋势. 因此，给定显著性水平 α，查 F 分布表得临界值 $F_\alpha(p-1,n-p)$（即 F 分布的上侧 α 分位数），计算 F 的观测值 F_0，若 $F_0\leqslant F_\alpha(p-1,n-p)$，接受 H_0，即在显著性水平 α 之下，认为线性回归关系不显著；若 $F_0>F_\alpha(p-1,n-p)$，拒绝 H_0，即认为 Y 与 $X_1,X_2,\cdots,X_{p-1}$ 之间存在显著的线性回归关系.

(3) 检验的 p 值法

在现代统计应用中，由于计算机的广泛使用，计算各种常见分布的分布函数值已是一件十分容易的事情. 因此显著性检验问题一般不再通过查表求其临界值，而是通过计算所谓的统计量的 p 值（p-value）来考察检验的显著性. 简单地说，一个检验统计量的 p 值是当 H_0 成立时，检验统计量取其观测值及更有利于备择假设 H_1 的值的概率. 具体地说，设检验统计量为 T，通过样本求得其观测值为 T_0，若大的 T 值意味着拒绝 H_0（或等价地有利于接受 H_1），则其 p 值为 $P_{H_0}(T\geqslant T_0)$；反之，若小的 T 值有利于接受 H_1，则 p 值为 $P_{H_0}(T\leqslant T_0)$；若大的 $|T|$ 值有利于接受 H_1，则 p 值为 $P_{H_0}(|T|\geqslant|T_0|)$，其中 P_{H_0} 表示在 H_0 为真时的概率. 有了 p 值后，对于给定的显著水平 α，任何检验准则均为

$$\begin{cases}\text{若 } p<\alpha,\text{拒绝 } H_0,\\ \text{若 } p\geqslant\alpha,\text{接受 } H_0.\end{cases} \tag{1.1.19}$$

这样不需要查相应分布的分位数表，而直接根据 p 和 α 的大小便可判断是拒绝还是接受 H_0，在 SAS 及其它一些统计软件中，对显著性检验问题，其输出结果通常是统计量的 p-值.

对于线性回归关系的显著性检验问题，其 p 值为

$$p=P_{H_0}(F\geqslant F_0).$$

检验准则为 $p\geqslant\alpha$，接受 H_0，否则拒绝 H_0. 当然，这两个准则是等价的，因为