全国高等农业院校教材

回归分析与试验设计

徐中儒 主编

农科类各专业用

中国农业出版社

全国高等农业院校教材

回归分析与试验设计

徐中儒　主编

农科类各专业用

中国农业出版社

全国高等农业院校教材
回归分析与试验设计
徐中儒　主编

责任编辑　徐建华
出　　版　中国农业出版社
（北京市朝阳区农展馆北路2号）
发　　行　新华书店北京发行所
印　　刷　中国农业出版社印刷厂

*　　*　　*

开　　本　787mm×1092mm16开本
印　　张　11.75　　字　数　265千字
版　次　1998年5月第1版
1998年5月北京第1次印刷
印　　数　1～2,500册　　定价 13.00元

书　　号　ISBN 7-109-04878-0/S·3028

ISBN 7-109-04878-0
9 787109 048782 >

前　言

随着工农业生产的发展和科学研究的深入，需要科学的分析方法，而数学分析方法是其核心，只有掌握这种方法，才能进行试验设计，以最少的试验，获取最有价值的信息，才能把大量的试验数据，有效地整理归纳，比较科学地反映出各因素的作用及相互联系的规律性。回归分析及试验设计是数理统计中的一个分支，在农业生产和科学研究中有着广泛的应用，它是解决农业科学试验中多变量问题的数据处理、建立数学模型、制定最佳农艺措施、获得优化决策的有效的统计数学方法。

本书是在我们多年运用回归设计方法所取得的一些应用成果和收集国内外有关资料的基础上，经过系统整理而写成的。编写中着重以实际应用为主，辅以数学的理论阐述，全书包含古典回归分析、回归正交设计、回归旋转设计、回归最优设计、回归模型的选择准则及回归方程的优化分析等六章，通过在农业试验中已取得成果的30多个实例，详细说明如何运用各种最优设计方法和它们的分析与应用。

本书的有关内容，我们曾先后在北京、武汉、安阳、西宁、哈尔滨、佳木斯等地为农业科研和教学工作者举办的讲习班上讲授，后来多所高等农业院校为本科生、研究生，以这一内容开设了有关课程，通过这些学术活动和课程讲授，对有些内容又进行了一些修改，这次以教科书形式出版，希望能为农业院校本科生、研究生学习回归分析和最优试验设计方法有所助益；对于从事科学试验的科研人员，本书对他们亦有重要的参考价值，全书所需的教学时数为36学时。

本书由徐中儒教授任主编，参加本书各章编写工作的有张嘉林、邓华玲（第一章）、葛家麒（第二章）、徐中儒（第三、五章）、姚晓敏（第四章）、孟军（第六章），袁志发教授审阅了全书，其中邓华玲、孟军老师还做了大量的稿件抄写、图表整理等工作。

由于编者水平所限，书中难免有不少不妥之处，敬请读者批评指正。

编　者

1997年3月

目　录

第一章　古典回归分析

在农学和生物学研究中，往往归结为弄清楚一些有关变量之间的联系，因为同时出现的多个变量，通常不是各自孤立地在变化，而是相互依赖，相互联系的，这种联系具体反映为寻求其中一个变量 y 通过其它的变量 x_1，x_2，…，x_m 表达出来，这一表达可分为两种基本类型：第一类的特征是，只要知道了变量 x_1，x_2，…，x_m 所取的值，变量 y 的值就完全确定了，这种关系称做确定性关系，即数学上的所谓函数关系。例如，圆的面积（设以 y 表示）和半径（设以 x 表示）的关系为 $y=\pi x^2$，根据这一关系可以知道对应于任一 x 值必有一个确定的 y 值。另一类是非确定性关系。例如降水量与农作物产量之间的关系是非确定性的，即使在同样的雨水条件下，作物产量也不是唯一确定的。再如，犁的耕深与牵引阻力的关系。这一类变量之间的关系，称为非确定性关系，亦称相关关系。比如人的身高与体重之间的关系，虽然由一个人的身高并不能确定体重，但是通过大量观测，平均说来，身高者，体也重，即身高与体重这两个变量具有相关关系。

一般把讨论随机变量与非随机变量之间的关系问题称为回归分析；把讨论随机变量之间关系的问题，称为相关分析。为了简单起见，我们不区分回归与相关，而统称为相关。若从变量的个数来区分，把两个变量之间的相关，称为简单相关；把 3 个变量以上的相关，称为复相关；把在复相关的条件下，仅研究两个变量的相关，称为偏相关。

§1　一元线性回归

回归分析是研究变量间相关关系的有力工具。它不仅提供了建立变量间关系的数学表达式（通常称为经验公式）的一般方法，而且进行分析讨论判明所建立的经验公式的有效性，以及如何利用所得到的经验公式去达到预测、控制等目的。回归分析要解决的问题可归纳为：

（1）确定二组或二组以上相对应的变量之间的相关关系，找出这些变量之间的定量关系式。

（2）对这个定量关系式的可靠性进行统计检验。

（3）根据变量之间的定量关系式作出预报和控制，进行优化分析。

（4）对多因素问题进行因素分析，确定各因素之间的主次关系。

（5）应用回归分析的原理，作出试验处理少，统计性质好的试验设计。

由以上各点可见，回归分析在工农业生产及科学技术研究中有着广泛的应用价值。

Expectation

一、一元线性回归方程

1. 回归方程　一元回归是处理两个变量之间的相关关系。假定两个随机变量 X 和 Y，它们之间相互依赖存在着相关关系，由于 Y 的随机性，对于确定的 X，Y 也没有确定的值与之对应，而只有确定的概率分布与之对应。为了寻求 Y 和 X 之间的关系，很自然的想法是，当 $X=x_0$ 时，用随机变量 Y 的数学期望代替 Y 的值，即

$$Y\mid_{X=x_0}=E(Y\mid X=x_0)$$

这里 $E(Y|X=x_0)$ 表示 $X=x_0$ 条件下 Y 的条件数学期望。当 X 取不同值时，Y 的数学期望取不同值，即 $y=E(Y|X=x)$ 是变量 X 取值 x 的函数，可记为

$$y=E(Y\mid X=x)=\mu(x) \tag{1.1}$$

用该式来描述变量 X 和 Y 之间的变化规律称为 Y 对 X 的回归方程。它所对应的曲线称为回归曲线。例如，上面提到的体重（Y）和身高（X）具有相关关系。这里 $\mu(x)$ 就是身高为 $X=x$ 的所有人（某特定范围）体重的平均值。回归方程是确定性关系，用它可以近似代替相关关系。

可以证明，对任意的函数 $\varphi(x)$ 恒有

$$E(Y-\varphi(x))^2\geqslant E(Y-\mu(x))^2$$

其中 $E(Y)=\mu(x)$

证明：
$$\begin{aligned}E(Y-\mu(x))^2&=E[Y-E(Y)]^2\\&=E(Y-\varphi(x)-E(Y)+\varphi(x))^2\\&=E(Y-\varphi(x))^2-2E(Y-\varphi(x))(E(Y)-\varphi(x))+E(E(Y)\\&\quad-\varphi(x))^2\\&=E(Y-\varphi(x))^2-E(E(Y)-\varphi(x))^2\end{aligned}$$

即有 $E(Y-\mu(x))^2\leqslant E(Y-\varphi(x))^2$

这里

$$E(Y-\varphi(x))(E(Y)-\varphi(x))=0$$

即在一切 x 的函数中，用回归函数 $\mu(x)$ 估计 Y，使估计的偏差平方和达到最小。因此回归方程成为研究变量相关关系的重要工具。

在实际问题中，回归函数是不知道的，要从任意函数 $\varphi(x)$ 中找出 $\mu(x)$ 也比较困难，需要由试验或观察数据来估计判断 $\mu(x)$ 是属于哪一类型的函数，并估计它，这样得出的方程叫经验方程。回归分析的基本任务是根据已知的数据，估计 $\mu(x)$，并利用此结果作预测和控制，估计 $\mu(x)$ 又称求 y 对 x 的回归方程。

例 1. 大豆栽培试验中，测得株龄（周）x 与株高（cm）y 的数据如下：

x_i	1	2	3	4	5
y_i	5	17	24	33	41

这里 x 是一般变量，y 是随机变量，求 y 对 x 的回归。

为研究这些数据所蕴含的规律性，我们把株龄作为横坐标，株高作为纵坐标在坐标系下描出图，称为散点图（图 1—1）。从该图可以看出 (x_i, y_i) 大致呈一条直线，故可设

Y 和 X 具有线性关系，用 $\alpha+\beta x$ 来作为 $\mu(x)$ 估计 Y 的数学期望，$E(Y)=y=\alpha+\beta x$

用线性函数 $\alpha+\beta x$ 来估计 Y 的数学期望的问题，称为一元线性回归问题。对每个 x，$y=E(Y)=\alpha+\beta x$ 都假设 $Y\sim N(\alpha+\beta x, \sigma^2)$，其中 α，β 及 σ^2 都是未知参数，并且都不依赖于 x，对 Y 作这样的正态假设，相当于设

$$y=\alpha+\beta x+\varepsilon \qquad \varepsilon\sim N(0, \sigma^2) \quad (1.2)$$

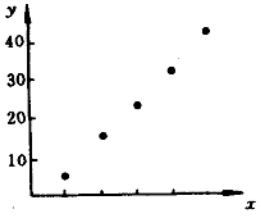

图 1—1　散点图

上式即称为一元线性回归模型。

回归分析的主要问题是根据 x、y 的观测值 (x_i, y_i) $(i=1, 2, \cdots, n)$ 给出 α、β 的估计值 a、b，同时对 a、b 做统计检验，以便确定它们的可靠程度。$\alpha+\beta x$ 的估计为 $a+bx$ 记作 $\hat{y}$，而方程

$$\hat{y}=a+bx$$

称为 Y 对 X 的一元线性回归方程或简称为回归方程，其图形称为回归直线。

2. α、β 的最小二乘估计　对于已知具有线性关系的两个变量 X 和 Y，求它们之间的定量表达式，实际上就是对 (1.2) 中的 α、β 进行估计。α、β 的任意一对估计值 a、b 决定了一条回归直线，

$$\hat{y}=a+bx$$

那么在这些直线中哪一条更能反映 X，Y 的线性相关关系呢？即怎样取 α、β 的估计值最合适？可采用最小二乘法来估计 α、β。

设给定 n 个点 (x_1, y_1)，(x_2, y_2)，$\cdots$ (x_n, y_n)，那么，对于平面上任意一条直线 l：$y=a+bx$，我们用 $[y_i-(a+bx_i)]^2$ 来刻划点 (x_i, y_i) 到直线 l 的远近程度，于是

$$\sum_{i=1}^{n}[y_i-(a+bx_i)]^2$$

就定量地描述了直线跟这 n 个点的总的远近程度，这个量是随不同的直线而变化的，也可以说是随不同的 a、b 而变化的，即可看成 a、b 的二元函数，记为 $Q(a, b)$

$$Q(a, b)=\sum_{i=1}^{n}[y_i-(a+bx_i)]^2 \qquad (1.3)$$

于是，要找一条直线，使该直线总的来看与这 n 个点最“接近”，也就是求一对 a，b 使 $Q(a, b)$ 达到最小。

由于 $Q(a, b)$ 是 n 个平方之和，所以使 $Q(a, b)$ 最小的原则称为平方和最小原则，习惯上称为最小二乘法原则。根据数学分析中求极值的原理，分别求 Q 对 a、b 的偏导数，并令它们等于零。

$$\begin{aligned}\frac{\partial Q}{\partial a}&=-2\sum_{i=1}^{n}[y_i-(a+bx_i)]=0\\ \frac{\partial Q}{\partial b}&=-2\sum_{i=1}^{n}x_i[y_i-(a+bx_i)]=0\end{aligned} \qquad (1.4)$$

由以上二式可解得

$$\begin{cases} b=\dfrac{\sum x_iy_i-\dfrac{1}{n}\sum x_i\sum y_i}{\sum x_i^2-\dfrac{1}{n}(\sum x_i)^2}=\dfrac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}=\dfrac{l_{xy}}{l_{xx}} \\ a=\bar{y}-b\bar{x} \end{cases} \tag{1.5}$$

其中$\sum$表示 i 从 1 到 n 求和，由（1.5）确定的直线 $\hat{y}=a+bx$ 叫 Y 对 X 的回归直线，a、b 叫回归系数，该直线过（$\bar{x}$，$\bar{y}$）点。

根据（1.5），对例 1 列表 1—1 计算如下：

表 1—1　大豆生长时间与株高的回归计算表

x	y	x^2	y^2	xy
1	5	1	25	5
2	17	4	289	34
3	24	9	576	72
4	33	16	1089	132
5	41	25	1681	205
$\sum x=15$	$\sum y=120$	$\sum x^2=55$	$\sum y^2=3660$	$\sum xy=448$

将表格中的有关数据代入公式计算：

$$\sum x_i=15 \quad \sum y_i=120 \quad n=5$$

$$\bar{x}=3 \quad \bar{y}=24$$

$$\sum x_i^2=55 \quad \sum y_i^2=3660 \quad \sum x_iy_i=448$$

$$l_{xx}=\sum(x_i-\bar{x})^2=\sum x_i^2-n\bar{x}^2=55-5\times3^2=10$$

$$l_{yy}=\sum(y_i-\bar{y})^2=\sum y_i^2-n\bar{y}^2=3660-5\times24^2=780$$

$$l_{xy}=\sum(x_i-\bar{x})(y_i-\bar{y})=\sum x_iy_i-n\bar{x}\bar{y}=88$$

$$\therefore b=\frac{l_{xy}}{l_{xx}}=\frac{88}{10}=8.8$$

$$a=\bar{y}-b\bar{x}=24-8.8\times3=-2.4$$

得回归方程　$\hat{y}=-2.4+8.8x$

二、回归方程的方差分析

用最小二乘法求出的回归方程，并不需要事先假定 y 与 x 具有线性关系。就方法而言，对任意一组数据（x_i，y_i）（$i=1$，2，…，n）都可以按最小二乘法的原则得到一条直线，所以对求得的回归直线是否有意义，还需进行检验，即 y 和 x 之间是否真的存在着线性关系。

如果变量 y 和 x 之间存在着线性关系，则 $\beta\neq0$，否则 $\beta=0$，所以要检验 y 与 x 之间是否有线性关系，也就是检验 β 是否为零。这里介绍 F 检验法和相关系数检验法。

1.F 检验法　我们知道，观测值 y_1，y_2，…，y_n 之间的差异可以认为是由两方面的因素引起的：一是自变量 x；二是其它因素（包括试验误差）。如果由 x 引起的差异比较大，则 $\beta\neq0$。因此可先将 y_1，y_2，…，y_n 之间的总差异分解成两部分。

几个观测值之间的差异，可用观测值 y_i 与其平均值 $\bar{y}$ 的离差平方和来表示，称为总的离差平方和，记作 $D_{总}$

$$D_{总}=\sum(y_i-\bar{y})^2=l_{yy}$$

且

$$D_{总}=\sum(y_i-\bar{y})^2=\sum[(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y})]^2$$

$$=\sum(\hat{y}_i-\bar{y})^2+\sum(y_i-\hat{y}_i)^2+2\sum(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})$$

其中

$$\sum(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})=\sum(y_i-\hat{y}_i)(a+bx_i-\bar{y})$$

$$=(a-\bar{y})\sum(y_i-\hat{y}_i)+b\sum x_i(y_i-\hat{y}_i)=0$$

即

$$\sum(y_i-\bar{y})^2=\sum(\hat{y}_i-\bar{y})^2+\sum(y_i-\hat{y}_i)^2 \tag{1.6}$$

或写成

$$D_{总}=D_{回}+D_{剩}$$

$$D_{回}=\sum(\hat{y}_i-\bar{y})^2=bl_{xy}=b^2l_{xx}$$

称为回归平方和，它表示 $\hat{y}_1$，$\hat{y}_2$，…，$\hat{y}_n$ 对平均值 $\bar{y}$ 的分散程度是由回归直线上它们所对应的横坐标 x_1，x_2，…x_n 的变化引起的。

$$D_{剩}=\sum(y_i-\hat{y}_i)^2=l_{yy}-bl_{xy}=Q$$

称为剩余平方和，是观测数据与回归值的差异，具有相同的 x 值，因此是由其它因素或误差引起的。

现在回到统计检验问题上来，在原假设 H_0：$\beta=0$ 成立时，可以证明 $D_{总}$、$D_{回}$、$D_{剩}$ 都是 χ^2 变量，其自由度分别为 $n-1$、1、$n-2$，于是构造统计量

$$F=\frac{D_{回}}{D_{剩}/n-2} \tag{1.7}$$

F 的大小表示由 x 引起的差异与误差差异的比较，$F\sim F(1,n-2)$。给定水平 α，找出 F_α，可用 F 与 $F\alpha$ 的比较来判定 x 的重要性。

回归方程显著性检验一般程序如下：

(1) 建立统计假设 H_0：$\beta=0$；

(2) 计算统计量，先计算 $D_{回}$、$D_{剩}$，再按公式 (1.7) 计算 F 值；

(3) 给定信度 α，查自由度为 1，$n-2$ 的 F 分布表，得临界值 F_α；

(4) 判断　如果 $F>F_\alpha$，则拒绝假设 H_0：$\beta=0$，即认为 x 与 y 之间具有线性相关关系；

如果 $F\leqslant F_\alpha$，则接受假设 H_0：$\beta=0$，即认为 x 与 y 之间不存在线性相关关系，回归直线无实际意义。

由此列出方差分析表 1—2。

现在用该方法对例 1 中回归方程的显著性进行检验：

表 1—2　一元线性回归的方差分析表

方差来源	平方和	自由度	均　方	F　值	临界值
回归	$D_{回}=\sum(\hat{y}_i-\bar{y})^2$	$f_{回}=1$	$S^2_{回}=D_{回}/f_{回}$	$F=\frac{D_{回}}{D_{剩}/n-2}$	$F\alpha(f_{回},f_{剩})$
剩余	$D_{剩}=\sum(y_i-\hat{y}_i)^2$	$f_{剩}=n-2$	$S^2_{剩}=D_{剩}/f_{剩}$		
总和	$D_{总}=\sum(y_i-\bar{y})^2$	$f_{总}=n-1$			

$$D_{回}=b^2l_{xx}=8.8^2\times10=774.4$$

$$D_{总}=l_{yy}=780$$

$$D_{剩}=D_{总}-D_{回}=780-774.4=5.6$$

列出回归分析表 1—3。

表 1—3　回归方差分析表

方差来源	平方和	自由度	均　方	F　值	临界值
回归	774.4	1	774.4	414.1**	$F_{0.05}=10.1$
剩余	5.6	3	1.87		$F_{0.01}=34.1$
总和	780.0	4			

查 F 表，$F_{0.05}(1,3)=10.1$，$F_{0.01}(1,3)=34.1$。现求得 $F=414.1$，$F>F_{0.01}$，说明假设 H_0：$\beta=0$ 不成立。可认为线性相关关系极显著。

应该指出，上面用剩余平方和去检验回归平方和所作出的“回归方程显著”这一判断，只是表明相对于其他因素及试验误差来说，因素 x 的一次项对指标 y 的影响是主要的，但并没有告诉我们：影响 y 的除 x 外，是否还有一个或几个不可忽视的其他因素，以及 x 和 y 的关系确是线性关系，也就是说，在上述意义下的“回归方程显著”，并不表明这个回归方程是拟合得很好的。

2. 相关系数检验法　回归方程的显著性也可采用相关系数检验法，构造统计量 r

$$r=\sqrt{\frac{D_{回}}{D_{总}}}=\sqrt{\frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\bar{y})^2}}=\frac{l_{xy}}{\sqrt{l_{xx}l_{yy}}} \tag{1.8}$$

它反映了回归平方和在总平方和中的比例，显然 $|r|$ 愈大，说明回归效果愈好，r 称为变量 x 与 y 的相关系数，它的取值范围是

$$-1\leqslant r\leqslant1$$

当 $r=0$ 时，$l_{xy}=0$，而 $b=\frac{l_{xy}}{l_{xx}}=0$ 说明 x 与 y 无线性关系。

当 $0<|r|<1$ 时，x 与 y 存在着一定的线性关系，当 $r>0$ 时，$b>0$ 散点图呈 y 随 x 增加而增加的趋势，称 x 与 y 正相关〔图 1—2 (1)〕；当 $r<0$ 时，$b<0$，散点图呈 y 随 x 的增加而减小的趋势，称 y 与 x 负相关〔图 1—2 (2)〕。

当 $|r|=1$ 时，所有的点都在同一条直线上，称 x 与 y 完全线性相关，这时 x 与 y 之间存在着确定的线性函数关系。

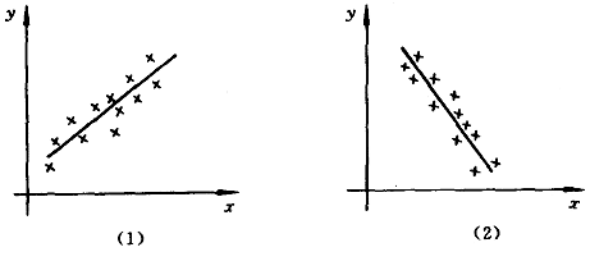

图 1—2　线性相关关系图
(1) 正相关　(2) 负相关

从上面的分析可以看出，相关系数 r 确实反映 x 与 y 之间线性相关的密切程度，$|r|$ 愈大，愈接近 1，x 与 y 的线性相关愈密切；$|r|$ 愈接近零，x 与 y 的线性相关程度愈小。对一个具体问题，只有当 $|r|$ 大到一定程度时，才可认为 x 与 y 有线性相关关系。一般用求得的 $|r|$ 与书末的相关系数检验表给出的不同显著水平 α（0.05 及 0.01）下相关系数达到显著的最小值 r_α 进行比较，当 $|r|>r_\alpha$ 时，才可认为 x 与 y 之间具有线性相关关系。

例 1 中大豆生长时间与株高的相关系数为：

$$r=\frac{l_{xy}}{\sqrt{l_{xx}l_{yy}}}=\frac{88}{\sqrt{10\times 780}}=0.996$$

查自由度为 $n-2=3$ 的相关系数表有 $r_{0.05}=0.878$，$r_{0.01}=0.959$，而 $r=0.996>0.959$，说明大豆生长时间与株高线性相关性极显著。

对回归方程进行检验，可用 F 检验，亦可用 r 进行检验，而两种检验本质上是一样的，可以证明：

$$F=\frac{D_{回}}{D_{剩}/n-2}=\frac{(n-2)\ D_{回}}{D_{总}-D_{回}}=\frac{(n-2)\ D_{回}/D_{总}}{1-D_{回}/D_{总}}=\frac{(n-2)\ r^2}{1-r^2} \tag{1.9}$$

F 较大时，$|r|$ 也较大。

三、有重复的回归方程的检验

检验回归方程显著，不一定方程拟合的就好，还需检验除 x 对 y 的影响外，是否还有其它因素对 y 有较大的影响，即剩余平方和只是由误差引起的，还是有其它未加控制的因素影响，通常采用重复试验的方法，重复试验可以对部分试验点进行，亦可对全部试验点进行。

1. 部分试验有重复的方程检验　假设对第 n 号试验进行 m 次重复，得到 $n+m-1$ 个数据，y_1，y_2，…，y_n，y_{n+1}…，y_{n+m-1}，其中前 $n-1$ 个无重复，第 n 个重复 m 次，对这 $n+m-1$ 个数据求得：

$$\begin{cases} D_{总}=\sum_{a=1}^{n+m-1}(y_a-\bar{y})^2 & f_{总}=(n+m-1)-1 \\ D_{回}=\sum_{a=1}^{n+m-1}(\hat{y}_a-\bar{y})^2 & f_{回}=1 \\ D_{剩}=\sum_{a=1}^{n+m-1}(y_a-\hat{y}_a)^2 & f_{剩}=n+m-3 \end{cases}$$

其中 $\bar{y}=\frac{1}{n+m-1}\sum_{a=1}^{n+m-1}y_a$

利用后 m 个数据得误差平方和

$$D_{误}=\sum_{a=n}^{n+m-1}(y_a-\bar{y}_n)^2 \quad f_{误}=m-1$$

$$\bar{y}_n=\frac{1}{m}\sum_{a=n}^{n+m-1}y_a$$

$D_{误}$ 是在 x 不变的情况下得到的 y_a 之间的差异，是由误差造成的。

$$D_{剩}-D_{误}=D_{拟}\quad f_{拟}=n+m-3-m+1=n-2$$

称 $D_{拟}$ 为失拟平方和，反映未加控制因素对 y 的影响，它的大小反映回归方程拟合的好坏程度。

$$D_{总}=D_{回}+D_{误}+D_{拟}$$

在假设 H_0：$\beta=0$ 成立的条件下，

$$\frac{D_{回}}{\sigma^2}\sim\chi^2\ (1),\ \frac{D_{拟}}{\sigma^2}\sim\chi^2\ (n-2),$$

$$\frac{D_{误}}{\sigma^2}\sim\chi^2\ (m-1)$$

并且它们之间相互独立，于是可用统计量

$$F_1=\frac{D_{拟}/f_{拟}}{D_{误}/f_{误}}\sim F\ (f_{拟},\ f_{误})$$

来检验回归方程拟合的是好还是坏。

对于给定的信度 α，假如算得 $F_1\leqslant F_\alpha\ (f_{拟},\ f_{误})$，则 F 检验结果不显著，说明失拟平方和基本上是由试验误差等偶然因素引起的，这时可把 $D_{拟}$ 与 $D_{误}$ 合并，并用来检验 $D_{回}$，即

$$F_2=\frac{D_{回}/f_{回}}{(D_{拟}+D_{误})\ /\ (f_{拟}+f_{误})}\sim F\ (f_{回},\ f_{拟}+f_{误})$$

如果第二次 F 检验结果显著，就称回归方程拟合得好。

如果第二次 F 检验结果不显著，这时有如下两种可能：①没有什么因素对 y 有系统的影响；②试验误差过大。当然这时的回归方程不理想。

对于给定的信度 α，如 $F_1>F_\alpha$，即第一次 F 检验结果显著，则说明失拟平方和中除试验误差影响外，还有其它因素影响，有以下几种可能：①影响 y 的除 x 外至少还有一个不可忽略的因素；②y 与 x 是曲线关系；③y 与 x 无关。这时，即使第二次 F 检验结果显著，所得的回归方程有一定的作用，也不能说方程拟合得好，仍需查明原因，改变模型，做进一步的研究。

2．全部试验点有重复的方程检验　假设对全部 n 个试验点各重复 m 次，得到 $n\cdot m$ 个数据，

$$y_{\alpha j}=\alpha+\beta x_\alpha+\varepsilon_{\alpha j}\qquad \begin{matrix}\alpha=1,\ 2,\ \cdots,\ n\\ j=1,\ 2,\ \cdots,\ m\end{matrix}$$

这就是有重复试验情况下的一元线性回归模型，其中 $\varepsilon_{\alpha j}$ 是相互独立的随机变量，$\varepsilon_{\alpha j}\sim N\ (0,\ \sigma^2)$。

同样用最小二乘法估计 α、β

$$a=\bar{y}-b\bar{x}$$

$$b=\frac{\sum\limits_\alpha x_\alpha\bar{y}_\alpha-\frac{1}{n}\ (\sum\limits_\alpha x_\alpha)\ (\sum\limits_\alpha\bar{y}_\alpha)}{\sum\limits_\alpha x_\alpha^2-\frac{1}{n}(\sum\limits_\alpha x_\alpha)^2}$$

其中 $\bar{x}=\frac{1}{n}\sum_{\alpha}x_{\alpha}$，$\bar{y}=\frac{1}{nm}\sum_{\alpha}\sum_{j}y_{\alpha j}$，$\bar{y}_{\alpha}=\frac{1}{m}\sum_{j}y_{\alpha j}$

即用所有的数据估计 α、β，与原来的回归方程一样，$\bar{y}_{\alpha}$ 代替 y_{α}。

对回归方程的检验，同样有

$$D_{总}=D_{回}+D_{拟}+D_{误}$$

其中

$$D_{总}=\sum_{\alpha}\sum_{j}(y_{\alpha j}-\bar{y})^2 \quad f_{总}=n\cdot m-1$$

$$D_{回}=\sum_{\alpha}\sum_{j}(\hat{y}_{\alpha}-\bar{y})^2=m\sum_{\alpha}(\hat{y}_{\alpha}-\bar{y})^2 \quad f_{回}=1$$

$$D_{误}=\sum_{\alpha}\sum_{j}(y_{\alpha j}-\bar{y}_{\alpha})^2 \quad f_{误}=n(m-1)$$

$$D_{拟}=\sum_{\alpha}\sum_{j}(\bar{y}_{\alpha}-\hat{y}_{\alpha})^2=m\sum_{\alpha}(\bar{y}_{\alpha}-\hat{y}_{\alpha})^2 \quad f_{拟}=n-2$$

构造统计量

$$F_1=\frac{D_{拟}/f_{拟}}{D_{误}/f_{误}}\sim F(f_{拟},f_{误})$$

对给定的信度 α，$F_1>F_{\alpha}$，说明失拟平方和 $D_{拟}$ 中除有试验误差外，还有其他因素影响，需进一步查明原因，再作分析；如 $F_1<F_{\alpha}$，说明 $D_{拟}$ 基本上是由误差等偶然因素引起，可将 $D_{误}$ 与 $D_{拟}$ 合并，并构造统计量

$$F_2=\frac{D_{回}/f_{回}}{(D_{拟}+D_{误})/(f_{拟}+f_{误})}\sim F(1,n\cdot m-2)$$

假如对给定的信度 α，$F_2>F_{\alpha}$，则回归方程显著，可以说方程拟合得好；如 $F_2<F_{\alpha}$，则回归方程不显著，这可能是由于试验中误差过大引起的；也可能是由于并不存在对 y 有显著影响的因素。

四、根据回归方程进行预报和控制

若回归方程拟合得好，它在一定程度上反映了两个变量之间的内在规律，可用它来进行预报和控制。

预报问题，对于

$$y=\alpha+\beta x+\varepsilon$$

所谓的预报问题即对任意给定的 x_0，推断 y_0 的值大致在什么范围，可用 $\hat{y}_0=a+bx_0$ 来作为 $y_0=\alpha+\beta x_0+\varepsilon_0$ 的一个点估计值，亦可做一个区间估计，即在一定的信度 α 下，寻找一个 $\delta>0$，使实际观察值以 $1-\alpha$ 的概率落在 $(\hat{y}_0-\delta,\hat{y}_0+\delta)$ 内，即

$$P\{\hat{y}_0-\delta<y_0<\hat{y}_0+\delta\}=1-\alpha$$

或者：

$$P\{|y-\hat{y}_0|<\delta\}=1-\alpha$$

已知 $y_0-\hat{y}_0$ 服从正态分布且：

$$E(y_0-\hat{y}_0)=E(y_0)-E(\hat{y}_0)=0$$

$$D(y_0-\hat{y}_0)=\sigma^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right]$$

所以

$$y_0-\hat{y}_0\sim N\left(0,\ \sigma\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}}\right)$$

由于上式中的 σ 是未知的，因而还不能立即求得 δ，这时可用 $D_{剩}/n-2$ 代替 σ^2，而统计量

$$\frac{(y_0-\hat{y}_0)^2}{\left\{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right\}\frac{D_{剩}}{n-2}}$$

服从自由度为（1，$n-2$）的 F 分布，由此不难求出 y_0 的置信概率为 $1-\alpha$ 的置信区间为

$$\hat{y}_0-\delta<y_0<\hat{y}_0+\delta$$

而

$$\delta=\sqrt{F_\alpha\frac{D_{剩}}{n-2}\left\{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right\}} \tag{1.10}$$

上式说明，y 的置信区间不仅与 α 有关，与 n 有关，而且与观测值 x_0 有关，当 x_0 靠近 $\bar{x}$ 时，δ 就小，区间就小；当 x_0 远离 $\bar{x}$ 时，δ 就大，区间也大，所以 $\delta=\delta(x_0)$ 是 x_0 的一个函数。做出 $y_1=\hat{y}-\delta$ 和 $y_2=\hat{y}+\delta$ 的图形（图 1—3），则它们把回归直线夹在中间，两头都呈喇叭形。

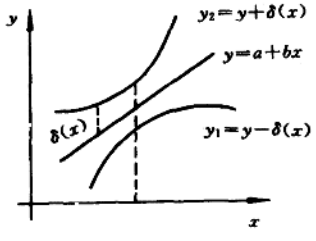

图 1—3　回归方程对 y 的预报区间

关于控制问题，实际是预报的反问题，即：若要求观测值 y 在一定范围 $y_1<y<y_2$ 内取值，应该把自变量 x 控制在何处，也就是说，要寻找 x_1，x_2，使得

$$\hat{y}-\delta(x_1)>y_1$$
$$\hat{y}+\delta(x_2)<y_2$$

如果 x_1，x_2 存在，问题就解决了。

前式（1.10）在计算时十分麻烦，应用中可进行简化。当 x_0 取值在 $\bar{x}$ 附近，n 又比较大时，有：

$$1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}\approx 1$$

$$\sigma^2\approx\delta^2=\frac{D_{剩}}{n-2}$$

在这种情况下，可近似地认为

$$y_0-\hat{y}_0\sim N(0,\ \delta^2)$$

则

$$P\{\hat{y}_0-2\delta<y_0<\hat{y}_0+2\delta\}=95\%$$
$$P\{\hat{y}_0-3\delta<y_0<\hat{y}_0+3\delta\}=99\% \tag{1.11}$$

于是在实际应用中经常是用（1.11）式来进行预报和控制。

如图 1—4 所示，在平面上作两条平行于回归直线的直线，

$$y=a-2\delta+bx$$
$$y=a+2\delta+bx$$