安徽理工大學

ANHUI UNIVERSITY OF SCIENCE & TECHNOLOGY

研究生课程

试验设计与分析

主讲　闵凡飞　教授

材料科学与工程学院

2020年5月

第四章 试验结果的分析
-回归分析

4.3 均匀试验设计结果的回归分析

1. 回归分析基本方法

若变量之间存在关系，这种关系一般有两种。

（1）确定性关系：变量之间存在完全确定的函数关系。一般可表达为：$y=f(x_1, x_2\cdots x_n)$，并称$x_1, x_2\cdots x_n$为自变量，y为因变量。例如众所周知的欧姆定律，I=V/R，三个变量中只要知道其中的两个，剩余的一个变量可准确计算出来。

（2）相关关系：在许多工程中，由于关系复杂或受试验误差影响，很难得到精确的数学表达式，从而使变量之间的关系存在某种不确定性，但又服从某种统计规律，可以用统计方法进行研究，称之为相关关系。例如：随着煤炭密度的增加，其对应的基元灰分在增加，但这种增加的规律是不确定的。同一矿区不同煤层的煤炭，相同密度级煤炭的灰分一般不同，不同矿区同一煤层的煤炭，相同密度级煤炭的灰分一般也不同，但随密度增加灰分增加的规律相同。

利用统计方法研究这种相关关系称为回归分析，有时也称为相关分析。回归分析主要处理连续型随机变量之间的相关关系，利用它来建立经验模型，检验模型的显著性，估计模型中的参数，确定最佳条件，实现预测和控制。

2. 一元线性回归

N组数据x_i，y_i，x是确定性变量，y为服从正态分布的随机变量，假定它们之间存在线性关系，则可以用一个回归方程表示：$\hat{y} = a + bx$ 。

1、回归系数的确定

回归系数的确定采用最小二乘法，即在精确度相等而误差呈正态分布的许多试验数据中求得最优概值的方法，其判断标准为各数据的偏差平方和为最小。

$$Q=\sum_{i=1}^{N}(y_i-\hat{y}_i)^2=\sum_{i=1}^{N}(y_i-a-bx_i)^2$$

根据最小二乘法的基本原理，回归系数 a，b 的大小应使偏差平方和 Q 为最小。根据极值原理，有

$$\frac{\partial Q}{\partial a}=-2\sum_{i=1}^{N}(y_i-a-bx_i)=0$$

$$\frac{\partial Q}{\partial b}=-2\sum_{i=1}^{N}(y_i-a-bx_i)x_i=0$$

记 $\bar{x}=\sum_{i=1}^{N}x_i/N$，$\bar{y}=\sum_{i=1}^{N}y_i/N$，

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

$$b=\frac{\sum_{i=1}^{N}x_iy_i-\frac{1}{N}\sum_{i=1}^{N}x_i\sum_{i=1}^{N}y_i}{\sum_{i=1}^{N}x_i^{\,2}-\frac{1}{N}(\sum_{i=1}^{N}x_i)^2}$$

令　$$lxx=\sum_{i=1}^{N}(x_i-\bar{x})^2=\sum_{i=1}^{N}x_i^2-\frac{1}{N}(\sum_{i=1}^{N}x_i)^2=\sum_{i=1}^{N}x_i^2-N\bar{x}^2$$

$$lyy=\sum_{i=1}^{N}(y_i-\bar{y})^2=\sum_{i=1}^{N}y_i^2-\frac{1}{N}(\sum_{i=1}^{N}y_i)^2=\sum_{i=1}^{N}y_i^2-N\bar{y}^2$$

$$lxy=\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{N}x_iy_i-\frac{1}{N}\sum_{i=1}^{N}x_i\sum_{i=1}^{N}y_i=\sum_{i=1}^{N}x_iy_i-N\bar{x}\bar{y}$$

则　$$b=\frac{lxy}{lxx}$$

2、模型显著性检验

在确定系数时，假设x，y间呈线性关系，这种假设是否正确，就要对回归模型的显著性进行检验。一元线性回归模型显著性用相关系数检验。

在因变量y的偏差平方和中，有两方面的因素引起y_i变化。

（1）当x取不同的x_i时，引起y_i的变化；

（2）试验误差及其它随机因素的影响。

$$lyy=\sum_{i=1}^{N}(y_i-\bar{y})^2=\sum_{i=1}^{N}[(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y})]^2$$

$$=\sum_{i=1}^{N}(y_i-\hat{y}_i)^2+2\sum_{i=1}^{N}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})+\sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2$$

可以证明，上式中第二项为0：

$$\sum_{i=1}^{N}(y_i-a-bx_i)(a+bx_i-\bar{y})=\sum_{i=1}^{N}(y_i-a-bx_i)(bx_i-b\bar{x})$$

$$=b\sum_{i=1}^{N}(y_i-a-bx_i)x_i-b\bar{x}\sum_{i=1}^{N}(y_i-a-bx_i)=0-0=0$$

$$lyy==\sum_{i=1}^{N}(y_i-\hat{y}_i)^2+\sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2=Q+U$$

$$U=\sum_{i=1}^{N}(\hat{y}_i-\bar{y})=\sum_{i=1}^{N}[(a+bx_i)-(a+b\bar{x})]^2=b^2\sum_{i=1}^{N}(x_i-\bar{x})^2=b^2lxx=b\frac{lxy}{lxx}lxx=blxy$$

回归平方和U由x的变化引起，反映由于x与y 间线性关系引起y的变化，它的大小说明了自变量x 的重要性。

剩余平方和（残差平方和）Q，反映了试验误差和其他因素的影响。

在总偏差平方和一定的条件下，剩余平方和Q越小，变量x与y 之间的线性关系越明确，U就越接近lyy，其比值接近于1。反之，接近于0。

令　$R^2 = \frac{U}{lyy}$

则　$R^2 = \frac{blxy}{lyy} = \frac{lxy}{lxx}\frac{lxy}{lyy}$

$R = \frac{lxy}{\sqrt{lxx \cdot lyy}}$

R称为相关系数，数值在-1到1之间。R的大小反映着变量x 与y 之间的线性密切程度。R与b同号，负值称负相关，说明y随x增加而减小，正值称正相关，说明y随x增加而增加。

用相关系数R判断x 与y 之间的线性密切程度准则是：

（1）｜R｜ =1，确定性线性关系

（2）｜R｜ =0，无线性关系。无关系或是其它关系

（3）0<｜R｜<1，查相关系数表。表中给出了不同自由度和显著性水平α下的临界相关系数，对一元线性回归，自由度f=N-2。如果R值大于表中某水平下的数值，则有1-α 的把握说回归方程是显著的，或有1-α 的把握说x与y间存在线性关系。自由度越小，其临界相关系数值越大，虽然越接近于1越好，但要注意自由度。

3、回归方程的精度

通过计算，可以得到一元线性回归的剩余均方差

$$S=\sqrt{\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{N-2}}$$

从统计学角度可以证明均方差 S 代表着变量 y 偏离回归直线的误差，对应任一固定的 x_i，相应观测值 y_i 将以 1-α 的概率落在区间$(\hat{y}_i-Z_{\alpha/2}S,\hat{y}_i+Z_{\alpha/2}S)$，其中 $Z_{\alpha/2}$是标准正态分布上 $\alpha/2$ 分位点。

如图 4-1 所示，在回归直线$\hat{y}=a+bx$ 的上下做两条平行线（虚线），

$$L_1:\ y=a+bx-Z_{\alpha/2}S$$

$$L_2:\ y=a+bx+Z_{\alpha/2}S$$

它表明，在全部可能出现的观察值 y_i 中，大约有 100（1-α）%的点落在这两条线 L_1 与 L_2 之间的范围内。

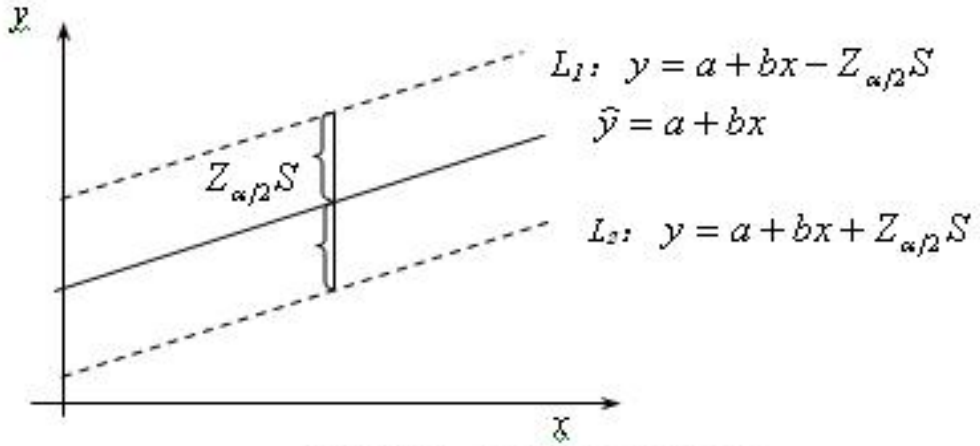


图 4-1 一元线性回归的预测

实例：煤中的无机硫主要是硫铁矿硫。从矿物学的角度来说，总有一些亲硫元素如Pb、As、Ni、Co等会以类质同象赋存于硫铁矿中或呈单独硫化矿与硫铁矿共存于煤中。下表是某原煤不同密度级中S，Ni的含量，试对它们进行线性回归，求模型参数与相关系数。

解：将计算过程列于下表中

线性回归计算表

编号	x（$S_{p,d}$）	y（Ni）	x^2	y^2	xy	$\hat{y}$	$\hat{y}-y$
1	.46	2	.2116	4	.92	5.3	3.3
2	1.3	3.3	1.69	10.89	4.29	6.6	3.3
3	3.78	8.2	14.2884	67.24	30.996	10.6	2.4
4	5.17	13.2	26.7289	174.24	68.244	12.8	-.4
5	7.84	22	61.4656	484	172.48	17.0	-5.0
6	8.69	24.6	75.5161	605.16	213.774	18.3	-6.3
7	30.49	50.2	929.6401	2520.04	1530.598	52.9	2.7
$\sum$	57.73	123.5	1109.5407	3865.57	2021.302		
l_{ij}	lxx=633.4331　lyy=1686.677 lxy=1002.7799					Q=99.19 S=4.45	
a=4.5869　b=1.5831　R=.9702							

本回归数据总数N=7，自由度f=5，查相关系数表$R_{0.001}$=0.95074，$R>R_{0.001}$，所以说煤中Ni含量与硫铁矿硫之间存在显著的线性关系。

3.可线性化曲线的线性回归

有一些非线性函数可先通过代数变换转换成线性关系，再利用线性回归方法求得函数参数。

曲线的直线化回归分为3步

（1）用试验数据绘制散点图，结合专业知识和经验选择适宜的函数，将函数线性化，同时将原始数据也按同样方式进行转化；

（2）用线性回归方法对转换后的试验数据进行回归，求得回归系数和线性回归相关系数；

（3）对线性回归系数进行反变换，得到曲线函数的回归系数，计算曲线的回归精度和相关系数。

1、可线性化曲线的基本类型与线性转化

下面的初等函数，是构成经验数学模型的基本单元，也是最常用的可线性化模型。

1. 双曲线

$$\frac{1}{y}=a+\frac{b}{x}$$

令 $y'=\frac{1}{y}$，$x'=\frac{1}{x}$，则有 $y'=a+bx'$

2. 幂函数

$$y=dx^{b}$$

令 $y'=\ln y$，$x'=\ln x$，$a=\ln d$ 则有 $y'=a+bx'$

3. 指数函数 1

$$y=de^{bx}$$

令 $y'=\ln y$，$a=\ln d$ 则有 $y'=a+bx$

4. 指数函数 2

$$y=de^{\frac{b}{x}}$$

令 $y'=\ln y$，$x'=\frac{1}{x}$，$a=\ln d$ 则有 $y'=a+bx'$

5. 对数函数

$$y=a+b\ln x$$

令 $x'=\ln x$，则有 $y=a+bx'$

6. S 型曲线

$$y=\frac{1}{a+be^{x}}$$

令 $y'=\frac{1}{y}$，$x'=e^{x}$，则有 $y'=a+bx'$

2、模型的检验

一元线性回归中，用相关系数检验回归方程的显著性。对于呈可直线化曲线关系的变量，也可以借用相关系数的定义进行检验。

$$R=\sqrt{\frac{\sum_{i=1}^{N}(\hat{y}-\bar{y})^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2}}=\sqrt{1-\frac{\sum_{i=1}^{N}(y_i-\hat{y})^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2}}$$

但此时必须用未变换的原始数据和呈曲线形态的函数计算值进行计算，计算出的相关系数也不同于线性回归的相关系数。其原因在于线性回归依据的是剩余平方和为最小，若对曲线进行了变换，则要求的是变换后的剩余平方和为最小，因此，在曲线的直线化时，最好多选择几个模型，进行比较，比较时可以比较剩余平方和Q，曲线的相关系数R和剩余方差S。R大者为优，Q、S小者为优。

实例：描述矿粒粒度分布的数学模型一般称为粒度特性方程。下表前2列给出了筛分试验数据，试选择多种模型对其关系进行回归。

解：虽然提出了多种粒度特性方程，但尚无一种方程能适合所有的粒度分布。常用的可以线性化的粒度特性方程有：（3？）

1. Rosin-Rammler 模型

$R(d)=100e^{-ad^b}$，d-粒度，$R(d)$-正累积产率

线性化过程为

$$Ln(Ln\frac{100}{R(d)})=Ln(a)+bLn(d)$$

2. Gauidn-Schuhmann 模型

$Y(d)=ad^b$，d-粒度，$Y(d)$-负累积产率

线性化过程为

$$Ln(Y(d))=Ln(a)+bLn(d)$$

3. 别洛格拉佐夫模型

$$Y(d)=\frac{ad^b}{1+ad^b}\times 100$$　d-粒度，$Y(d)$-负累积产率

线性化过程为：

$$Ln\frac{1}{1/Y(d)-1}=Ln(a)+bLn(d)$$

粒度曲线直线化回归计算表

x_i	y_i	$\ln x_i$	$-\ln(100/y_i-1)$	$-\ln(100/y_i-1)\ln x_i$	$\hat{y}_i$	$\hat{y}_i-y_i$
50	90.09	3.912023	2.207265	8.634871	89.63	-.46
25	81.18	3.218876	1.461749	4.705188	81.4	.22
13	67.68	2.564949	0.739104	1.895766	69.725	2.045
6	52.62	1.791759	0.104896	0.187949	51.878	-.742
3	36.08	1.098612	-0.57189	-0.628289	35.312	-.768
0.5	8.56	-0.693147	-2.36858	1.641777	8.592	.032
$\sum()$		11.89307	1.57254	16.43726		
$\sum()^2$		37.14185	13.50325			
l_{ij}		13.56766	13.09111	13.32021		

$\ln a$=-1.6839，a=.1856，b=.9818，$R_{回}$=.99947，$R_{总}$=.99939，Q=5.584，S=1.182

不同模型对粒度曲线直线化回归结果

编号	方程	线性	曲线	
		相关系数	剩余方差	相关系数值
1	$y = a + bx$	0.8380	18.49	0.8380
2	$y = 100 - ae^{bx}$	-0.9635	13.49	0.9174
3	$y = a + b\ln x$	0.9972	2.52	0.9972
4	$y = ax^{b}$	0.9597	16.93	0.8662
5	$y = 100*(1 - e^{-ax^{b}})$	0.9905	4.41	0.9915
6	$y = 100*(e^{-ax^{b}})$	-0.9871	4.59	0.9908
7	$y = ae^{b/x}$	-0.9715	13.85	0.9126
8	$y = a + b/x$	-0.8503	17.83	0.8503
9	$y = \frac{1}{a + b/x}$	0.9999	1.32	0.9992
10	$y = \frac{100}{a + be^{x}}$	-0.8491	16.20	0.8782
11	$y = 100\frac{ax^{b}}{1 + ax^{b}}$	0.9995	1.18	0.9994

根据煤炭筛分试验数据，用包括线性方程在内的11种模型来进行回归。从表中可以看出，线性回归的相关系数与曲线的相关系数或剩余方差不完全一致。例如模型9与11，模型4与10，模型2与7，所以在曲线直线化回归时，一定要计算剩余方差或曲线相关系数，并对多个模型进行比较。

4. 一元多项式回归分析

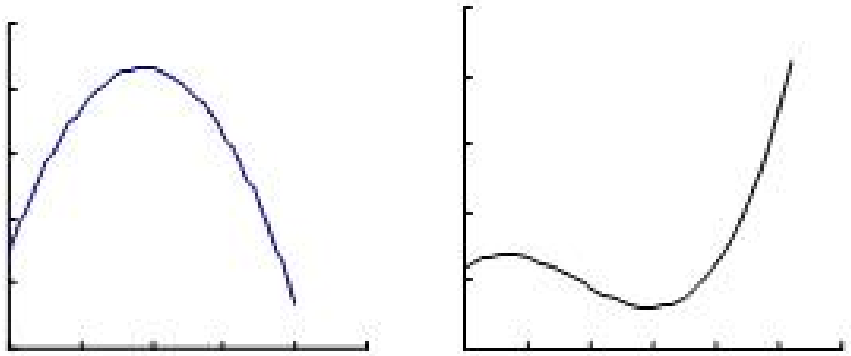
1、多项式阶数的判断

依据微积分理论，相当广泛的一类曲线可以用分段多项式来表示。当因变量存在极值，曲线形态为抛物线形态时，必须用多项式函数来描述变量间的关系。另外，当在初等函数中找不到满意的可直线化非线性函数，或事先不能确定出函数的类型时，也可以采用多项式函数。多项式函数的一般形式为：$y=b_0+b_1x+b_2x^2+\ldots b_px^p$

虽然任意曲线都可以近似的用多项式表示，增加多项式的阶数在一般情况下可以减小回归误差，提高精度，但也可能使试验点外的回归曲线振荡，导致预测精度下降。当选择多项式时，需要对多项式的阶数进行判断。

简单的方法可以根据试验点绘制散点图，通过散点图的形态确定多项式的项次。

例如，当曲线只有一个极大值或最小值时，可以选择二次抛物线，如下图所示；当曲线既有极大值又有极小值时，可以选择三次或三次以上抛物线，如下图所示。



多项式曲线形态图

当自变量x_i为等差数列，间隔为常数h时，可以用差分判别多项式的最高项次。

一阶差分　$\Delta y_i = y_{i+1} - y_i$

二阶差分　$\Delta^2 y_i = \Delta y_{i+1} - \Delta y_i$

三阶差分　$\Delta^3 y_i = \Delta^2 y_{i+1} - \Delta^2 y_i$

⋮

将计算结果列成差分表，见下表。

判别的原则是，若 p 阶差分是常数，$p+1$ 阶差数为 0 时，则函数应是 p 阶多项式。

差分表

x_i	y_i	Δy_i	$\Delta^2 y_i$	$\Delta^3 y_i$	……
x_1	y_1	Δy_1	$\Delta^2 y_1$	$\Delta^3 y_1$	……
x_2	y_2	Δy_2	$\Delta^2 y_2$	$\Delta^3 y_2$	……
x_3	y_3	Δy_3	$\Delta^2 y_3$	$\Delta^3 y_3$	……
x_4	y_4	Δy_4	$\Delta^2 y_4$	$\Delta^3 y_4$	……
x_5	y_5	Δy_5	$\Delta^2 y_5$	$\Delta^3 y_5$	……
⋮	⋮	⋮	⋮	⋮	……

实例：已知某选煤厂精煤灰分与吨原煤产值的对应关系，其差分计算表如下，判断多项式的阶数。

产值 C 与精煤灰分 A 差分计算表

精煤灰分 A	产值 C	ΔC	$\Delta^2 C$	$\Delta^3 C$
8.5	33.5	1.3	-0.4	-0.1
9.0	34.8	0.9	-0.5	0.1
9.5	35.7	0.4	-0.4	-0.1
10.0	36.1	0	-0.5	0.1
10.5	36.1	-0.5	-0.4	
11.0	35.6	-0.9	⋮	
11.5	34.5			

从表中可以看出，二阶差分已接近常数，三阶差分趋近于0，所以产值C与精煤灰分A间的关系可以用二阶多项式表示。

2、模型参数的估计

当选择多项式时，即使是二次抛物线也不能转换成线性关系，进而用一元线性回归的方法求出模型参数。但同样可以用最小二乘法，建立方程组，求出模型参数。

$$Q=\sum_{i=1}^{N}(y_i-\hat{y}_i)^2=\sum_{i=1}^{N}(y_i-b_0-b_1x_i-b_2x_i^{\ 2}\cdots-b_px_i^{\ p})^2$$

由最小二乘法

$$\begin{cases}\dfrac{\partial Q}{\partial b_0}=-2\sum_{i=1}^{N}(y_i-b_0-b_1x_i-b_2x_i^{\ 2}\cdots-b_px_i^{\ p})=0\\ \dfrac{\partial Q}{\partial b_j}=-2\sum_{i=1}^{N}(y_i-b_0-b_1x_i-b_2x_i^{\ 2}\cdots-b_px_i^{\ p})x_j=0 \quad (j=1,2\cdots p)\end{cases}$$

得到 p+1 个方程组

$$\begin{cases} b_0 N + b_1 \sum x_i + b_2 \sum x_i^2 + \cdots + b_p \sum x_i^p = \sum y_i \\ b_0 \sum x_i + b_1 \sum x_i^2 + b_2 \sum x_i^3 + \cdots + b_p \sum x_i^{p+1} = \sum y_i x_i \\ b_0 \sum x_i^2 + b_1 \sum x_i^3 + b_2 \sum x_i^4 + \cdots + b_p \sum x_i^{p+2} = \sum y_i x_i^2 \\ \cdots \\ b_0 \sum x_i^p + b_1 \sum x_i^{p+1} + b_2 \sum x_i^{p+2} + \cdots + b_p \sum x_i^{2p} = \sum y_i x_i^p \end{cases}$$

可以用高斯消元法解线性方程，得到多项式参数。

5. 多元线性回归分析

1、多元线性回归

当遇到多个自变量，一个因变量，每个自变量和因变量之间均为线性关系时，可以用多元线性函数来表示：

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

对比多元线性函数和一元多项式的表达式，可以发现多元线性函数是自变量下标在变化，多项式是自变量的幂次在变化，将多项式的幂次移到下标位置，即将每个幂次看作一个新的自变量，则一元多项式就转化为多元线性函数式。

$$Q=\sum_{i=1}^{N}(y_i-\hat{y}_i)^2=\sum_{i=1}^{N}(y_i-b_0-b_1x_{1i}-b_2x_{2i}\cdots-b_px_{pi})^2$$

$$令\begin{cases}\dfrac{\partial Q}{\partial b_0}=-2\sum_{i=1}^{N}(y_i-b_0-b_1x_{1i}-b_2x_{2i}\cdots-b_px_{pi})=0\end{cases}$$

$$记\ \bar{x}_j=\frac{1}{N}\sum_{i=1}^{N}x_{ij},\ \bar{y}=\frac{1}{N}\sum_{i=1}^{N}y_j,\ \ j=1,2,\ldots p$$

$$b_0=\bar{y}-b_1\bar{x}_1-b_2\bar{x}_2-\cdots--b_p\bar{x}_p$$

$$Q=\sum_{i=1}^{N}(y_i-\hat{y}_i)^2=\sum_{i=1}^{N}[(y_i-\bar{y})-b_1(x_{1i}-\bar{x}_1)-b_2(x_{2i}-\bar{x}_2)\cdots-b_p(x_{pi}-\bar{x}_p)]^2$$

$$令\frac{\partial Q}{\partial b_j}=0,\ \ j=1,2,\ldots p$$

$$
\begin{cases}
\sum_{i=1}^{N}(x_{1i}-\bar{x}_1)^2 b_1+\sum_{i=1}^{N}(x_{2i}-\bar{x}_2)(x_{1i}-\bar{x}_1)b_2+\cdots+\sum_{i=1}^{N}(x_{pi}-\bar{x}_p)(x_{1i}-\bar{x}_1)b_p \\
\qquad\qquad =\sum_{i=1}^{N}(y_i-\bar{y})(x_{1i}-\bar{x}_1) \\
\sum_{i=1}^{N}(x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2)b_1+\sum_{i=1}^{N}(x_{2i}-\bar{x}_2)^2 b_2+\cdots+\sum_{i=1}^{N}(x_{pi}-\bar{x}_p)(x_{2i}-\bar{x}_2)b_p \\
\qquad\qquad =\sum_{i=1}^{N}(y_i-\bar{y})(x_{2i}-\bar{x}_2) \\
\qquad\qquad \cdots \\
\sum_{i=1}^{N}(x_{1i}-\bar{x}_1)(x_{pi}-\bar{x}_p)b_1+\sum_{i=1}^{N}(x_{2i}-\bar{x}_2)(x_{pi}-\bar{x}_p)b_2+\cdots+\sum_{i=1}^{N}(x_{pi}-\bar{x}_p)^2 b_p \\
\qquad\qquad =\sum_{i=1}^{N}(y_i-\bar{y})(x_{pi}-\bar{x}_p)
\end{cases}
$$

$$l_{jk}=\sum_{i=1}^{N}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)$$

令　　，j=1,2,…p

$$l_{jy}=\sum_{i=1}^{N}(x_{ij}-\bar{x}_j)(y_i-\bar{y})$$

则方程组为

$$\begin{cases} l_{11}b_1+l_{12}b_2+\cdots+l_{1p}b_p=l_{1y} \\ l_{21}b_1+l_{22}b_2+\cdots+l_{2p}b_p=l_{2y} \\ \cdots \\ l_{p1}b_1+l_{p2}b_2+\cdots+l_{pp}b_p=l_{py} \end{cases}$$

若以矩阵形式表示

$$L=\begin{pmatrix} l_{11} & l_{12} & \cdots & l_{1p} \\ l_{21} & l_{22} & \cdots & l_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ l_{p1} & l_{p2} & \cdots & l_{pp} \end{pmatrix},\quad B=\begin{pmatrix} b_1 \\ b_2 \\ \cdots \\ b_p \end{pmatrix},\quad F=\begin{pmatrix} l_{1y} \\ l_{2y} \\ \cdots \\ l_{py} \end{pmatrix}$$

则 $L\cdot B=F$，　$B=L^{-1}F$

若将矩阵 L^{-1} 元素记为 c_{jk}，j,k=1,2,…p

则回归系数　$b_j=\sum_{k=1}^{p}c_{jk}l_{ky}$

2、回归方程的显著性检验

1）复相关系数

类似于一元线性回归分析，总偏差平方和仍可以分解成剩余平方和和回归平方和。

$$l_{yy}=\sum_{i=1}^{N}(y_i-\bar{y})^2=\sum_{i=1}^{N}[(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y})]^2=U+Q$$

$$U=\sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2=\sum_{k=1}^{p}b_k l_{ky}$$

借鉴相关系数的概念来评价多元线性回归方程的显著型，由于是与多个自变量之间的相关关系，所以称之为复相关系数。

$$\text{所以}\quad R=\sqrt{\frac{U}{l_{yy}}}=\sqrt{\frac{l_{yy}-Q}{l_{yy}}}=\sqrt{1-\frac{Q}{l_{yy}}}$$

当$|R|$接近于 1 时，说明因变量与诸个自变量组成的线性方程线性关系密切，反之，线性关系不密切甚至不存在线性关系。

由于复相关系数不能明确指出每个变量的作用，而且*R*不仅与试验数据数量有关，而且与自变量的数量有关，使用时没有一元线性方程的相关系数方便，而理论上又可以证明复相关系数检验方法实质上与F检验法相同，因此在多元回归分析中一般用F检验法检验回归方程的显著性。

2）F检验

在多元线性回归中，总平方和的自由度为*N*-1，回归平方和的自由度等于自变量个数*p*，剩余平方和的自由度则等于*N*-*p*-1。可以证明，在满足矩阵L满秩与假设*H*0（*y*与诸*x*之间无线性关系）成立的条件下，回归均方与剩余均方相互独立，构成:

$$F=\frac{U/p}{Q/(N-p-1)}$$

服从第一自由度为 p，第二自由度为 N-p-1 的 F 分布。

对于给定的置信度α，相应的自由度 p 和 N-p-1，可查 F 分布表，得到 F_α。如果 $F>F_\alpha$，否定原假设，即认为 y 与诸 x 之间存在线性关系，回归方程具有实际意义；反之，则接受原假设，y 与诸 x 之间无线性关系。

将上述分析归纳成方差分析表

方差分析表

方差来源	平方和	自由度	均方差	F 值	显著性
回归平方和（U）	$U=\sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2=\sum_{k=1}^{p}b_k l_{ky}$	p	$\frac{U}{p}$	$\frac{U/p}{Q/(N-p-1)}$	$F>F_\alpha$ y 与诸 x 之间存在显著的线性关系
剩余平方和（Q）	$Q=\sum_{i=1}^{N}(y_i-\hat{y}_i)^2=l_{yy}-U$	N-p-1	$\frac{Q}{N-p-1}$		
总平方和（l_{yy}）	$lyy=\sum_{i=1}^{N}(y_i-\bar{y})^2$	N-1			

3）回归系数的检验

回归平方和是所有自变量对y变差的总贡献。所考虑的自变量越多，回归平方和越大。因此，若从自变量总数中去掉一个自变量x_k，回归平方和只会减小，不会增加，减小的程度越大，说明被去掉的自变量在回归模型中起的作用越大，该因素越重要。称取消一个自变量后回归平方和的减少值为y对这个变量的偏回归平方和，用p_k表示

$$p_k = U^p - U^{p-1}$$

可以证明 $$p_k = \frac{b_k^2}{c_{kk}}$$

式中 b_k-x_k对应的偏回归系数

c_{kk}-正规方程组系数矩阵的逆矩阵主对角线上第 k 个元素。

p_k越大，该变量对 y 的影响越大，其定量判断指标可以通过引入 F 检验来解决。

$$F_k = \frac{p_k}{Q/(N-p-1)}$$

F_k 为第一自由度为 1，第二自由度为 N-p-1 的 F 分布。因此，对于给定的置信度α和相应的自由度，查 F 分布表，得到 F_α。如果 $F > F_\alpha$，该自变量对 y 有显著影响，应当保留。反之，则接受原假设，该自变量对 y 无显著影响，应当从回归模型中剔除。

6. 均匀试验设计试验结果的回归分析

实例：在啤酒生产的某项试验中，选了如下表所示的因素和水平。

因素 \ 水平	1	2	3	4	5	6	7	8	9
x_1 底水/g	136.5	137	137.5	138	138.5	139	139.5	140	140.5
x_2 吸氨时间/min	170	180	190	200	210	220	230	240	250

我们选用$U_9(9^5)$来安排，由它的使用表应选1,3两列，试验结果吸氧量y列于下表。

x_1	x_2	y	x_1	x_2	y
(1) 136.5	(4) 200	5.8	(6) 139.0	(6) 220	4.5
(2) 137.0	(8) 240	6.3	(7) 139.5	(1) 170	3.0
(3) 137.5	(3) 190	4.9	(8) 140.0	(5) 210	3.6
(4) 138.0	(7) 230	5.4	(9) 140.5	(9) 250	4.1
(5) 138.5	(2) 180	4.0			

相关数据计算结果如下：

$$\sum_{i=1}^{9} x_{i1} = 1246.5 \quad \sum_{i=1}^{9} x_{i2} = 1890 \quad \sum_{i=1}^{9} y_i = 41.6$$

$$\overline{x_1} = 138.5 \quad \overline{x_2} = 210 \quad \overline{y} = 4.62$$

$$l_{11}=15 \quad l_{12}=30 \quad l_{21}=30 \quad l_{22}=6000$$

$$l_{1y}=-9.8 \quad l_{2y}=110 \quad l_{yy}=9.24$$

得方程组：

$$\begin{cases} 15b_1 + 30b_2 = -9.8 \\ 30b_1 + 6000b_2 = 110 \end{cases}$$

解方程得：b_1=-0.697　　b_2=0.0218

$b_0=\bar{y}-b_1\overline{x_1}-b_2\overline{x_2}=96.44$

得回归方程：

y=96.44-0.697x_1+0.0218x_2

回归平方和：

$U=b_1l_{1y}+b_2l_{2y}$=9.2188

残差平方和：

$Q=l_{yy}-U$=0.0162

复相关系数为：

$$R=\sqrt{\frac{U}{l_{yy}}}=0.9991$$　由此可知回归方程是显著的。

残差标准差为：

$$s_y=\sqrt{\frac{Q}{n-m-1}}=0.052$$

$$\begin{pmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \end{pmatrix}^{-1}=\begin{pmatrix} 15 & 30 \\ 30 & 6000 \end{pmatrix}^{-1}=\begin{pmatrix} 0.06734 & -0.00034 \\ -0.00034 & 0.000168 \end{pmatrix}$$

$$F_1=\frac{b_1^2}{C_{11}Q/(n-m-1)}=2664.288$$

$$F_2=\frac{b_1^2}{C_{22}Q/(n-m-1)}=1045.528$$

对 $\alpha=0.01$，由 F 分布查得 $F_{0.01}(1,6)=13.7$，所以两个变量都是高度显著的，由回归方程我们可知 y 随 x_1 的增加而减少，y 随 x_2 的增加而增加。

操千曲而后晓声

观千剑而后识器


你的进步，我的快乐！