试验数据的回归分析

4.1 基本概念

（1） 相互关系

①确定性关系：

- **变量之间存在着严格的函数关系**

②相关关系：

- **变量之间近似存在某种函数关系**

（2） 回归分析（**regression analysis**）

处理变量之间相关关系的统计方法

- **确定回归方程：变量之间近似的函数关系式**
- **检验回归方程的显著性**
- **试验结果预测**

4.2 一元线性回归分析

4.2.1 一元线性回归方程的建立

（1）最小二乘原理

- 设有一组试验数据（如表），若$\boldsymbol{x}$，$\boldsymbol{y}$符合线性关系

x	x_1	x_2	……	x_n
y	y_1	y_2	……	y_n

- 一元线性回归方程 ：

$$\hat{y}_i = a + bx_i$$

- ***a***，***b***——**回归系数（regression coefficient）**

  $\hat{y}_i$ ——**回归值/拟合值，由x_i代入回归方程计算出的y值。**

- 计算值 $\hat{y}_i$与试验值$\boldsymbol{y}_{\mathrm{i}}$不一定相等
- $\hat{y}_i$与$\boldsymbol{y}_{\mathrm{i}}$之间的偏差称为残差：

$$e_i = y_i - \hat{y}_i$$

- 残差平方和：

$$SS_e = Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

- 残差平方和最小时，回归方程与试验值的拟合程度最好

求残差平方和极小值:

$$\begin{cases} \dfrac{\partial Q}{\partial a} = -2\sum_{i=1}^{n} (y_i - a - bx_i) = 0 \\ \dfrac{\partial Q}{\partial b} = -2\sum_{i=1}^{n} (y_i - a - bx_i)x_i = 0 \end{cases}$$

- 正规方程组（**normal equation**）：

$$\begin{cases} na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

- 解正规方程组：

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n(\overline{x})^2}$$

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

- 简算法：

$$L_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n(\overline{x})^2$$

$$L_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}$$

$$b = \frac{L_{xy}}{L_{xx}}$$

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

4.2.2　一元线性回归效果的检验

（1）相关系数检验法

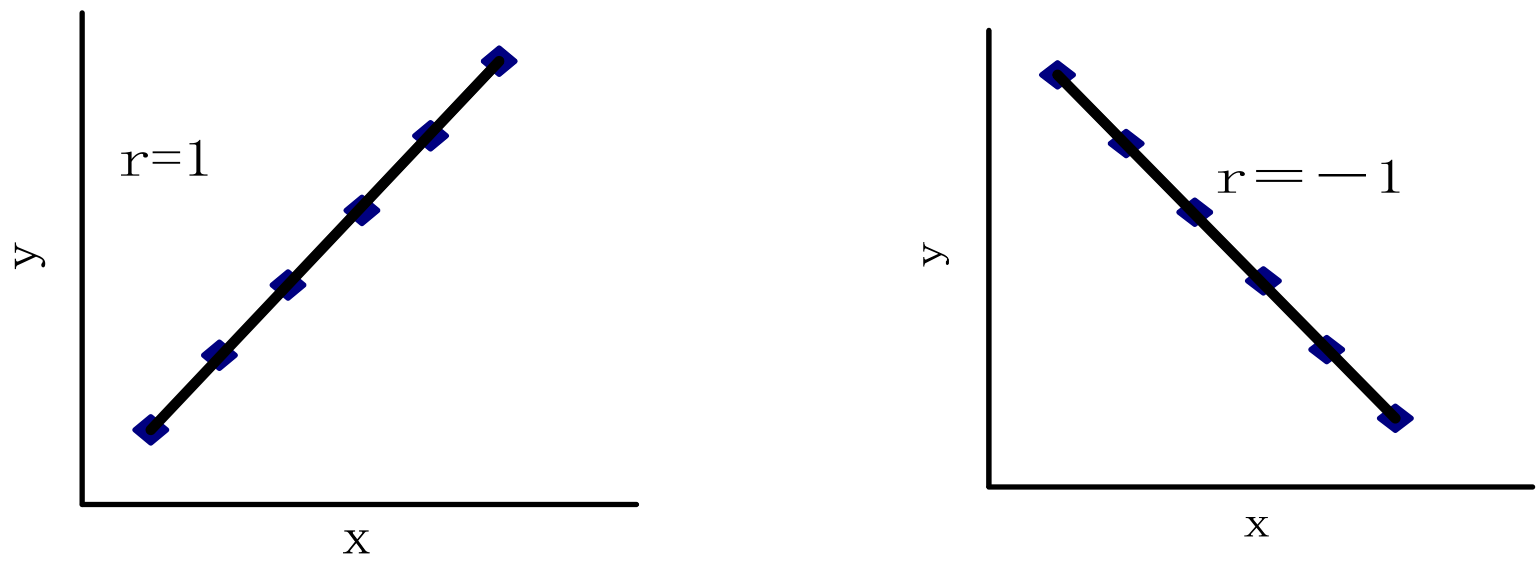
①相关系数（**correlation coefficient**）：

- 描述变量x与y的线性相关程度
- 定义式：

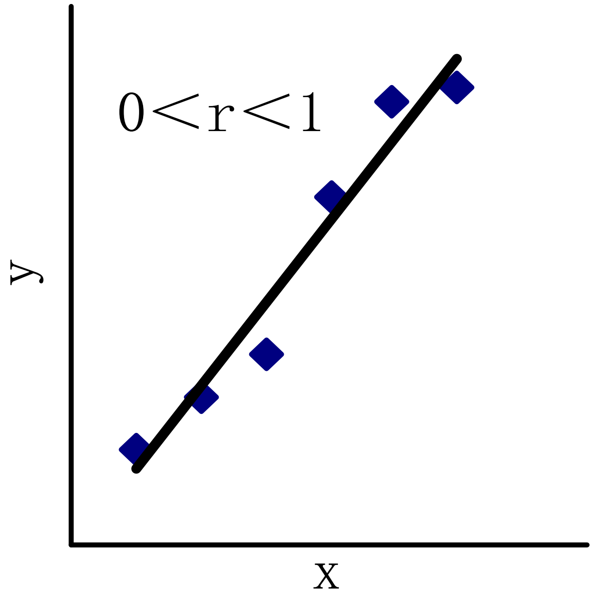
$$r = \frac{L_{xy}}{\sqrt{L_{xx} L_{yy}}}$$

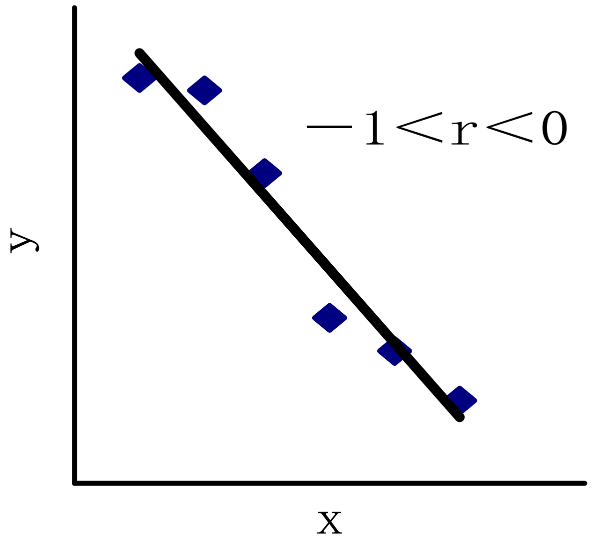
②相关系数特点：

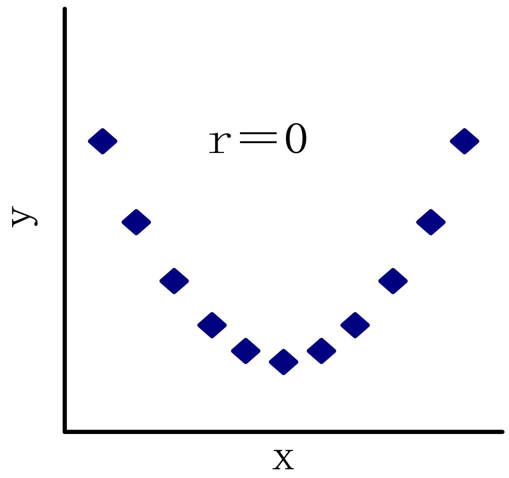
- $-1 \leq r \leq 1$
- **r＝±1：*x*与*y*有精确的线性关系**

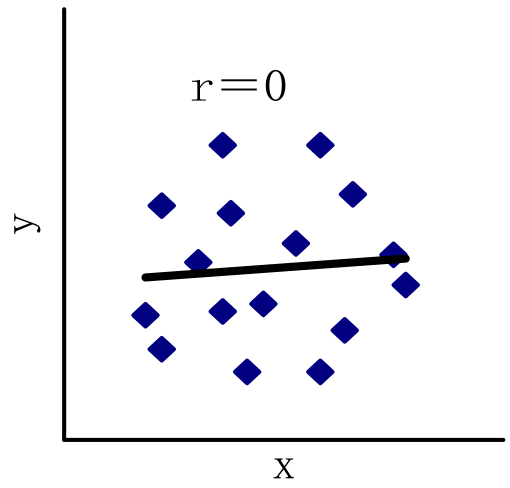


- $r<0$：x与y负线性相关（**negative linear correlation**）
- $r>0$：x与y正线性相关（**positive linear correlation**）









- ***r*≈0**时，***x***与***y***没有线性关系，但可能存在其它类型关系
- 相关系数***r***越接近**1**，***x***与***y***的线性相关程度越高
- 试验次数越少，***r***越接近**1**

③相关系数检验

- 对于给定的显著性水平**α**，查相关系数临界值$\mathbf{r}_{\mathbf{min}}$
- 当 $|r| > r_{\min}$ ，说明$\boldsymbol{x}$与$\boldsymbol{y}$之间存在显著的线性关系

相关系数 r 与 R 的临界值表

n-m-1	α	自变量的个数（m）			
		1	2	3	4
1	0.05	0.997	0.999	0.999	0.999
2	0.05	0.950	0.975	0.983	0.987
3	0.05	0.878	0.930	0.950	0.961
4	0.05	0.811	0.881	0.912	0.930
5	0.05	0.754	0.863	0.874	0.898
6	0.05	0.707	0.795	0.839	0.867
7	0.05	0.666	0.758	0.807	0.838
8	0.05	0.632	0.726	0.777	0.811
9	0.05	0.602	0.697	0.750	0.786
10	0.05	0.567	0.671	0.726	0.763
11	0.05	0.553	0.648	0.703	0.741
12	0.05	0.532	0.627	0.683	0.722
13	0.05	0.514	0.608	0.664	0.703

（2）F检验

①离差平方和

- 总离差平方和： $$SS_T = \sum_{i=1}^{n}(y_i - \overline{y})^2 = L_{yy}$$
- 回归平方和（**regression sum of square**）：

$$SS_R = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = b^2 L_{xx} = bL_{xy}$$

- 残差平方和：

$$SS_e = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- 三者关系：

$$SS_T = SS_R + SS_e$$

②自由度

- SS_T的自由度： $df_{\mathrm{T}}=\mathrm{n}-1$
- SS_R的自由度： $df_{\mathrm{R}}=1$
- SS_e的自由度： $df_{\mathrm{e}}=\mathrm{n}-2$
- 三者关系： $df_{\mathrm{T}}= df_{\mathrm{R}}+df_{\mathrm{e}}$

③均方

$$MS_R=\frac{SS_R}{df_R}$$

$$MS_e=\frac{SS_e}{df_e}$$

④F检验

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

- F服从自由度为（1，n－2）的F分布
- 给定的显著性水平α下，查得临界值：　F_{α}（1，n－2)
- 若F＞ F_{α}（1，n－2)，则认为x与y有明显的线性关系，所建立的线形回归方程有意义

⑤方差分析表

一元线性回归方差分析表

差异源	SS	df	MS	F	显著性
回归	SS_R	1	$MS_R = SS_R$	$F = MS_R / MS_e$	
误差	SS_e	n－2	$MS_e = SS_e / (n-2)$		
总和	SS_T	n－1			

4.3 多元线性回归分析

4.3.1 多元线性回归方程的建立

（1）多元线性回归形式

- 试验指标（因变量）y与m个试验因素（自变量）x_j（j=1,2,…,m）
- 多元线性回归方程：

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + \ldots + b_m x_m$$

- 偏回归系数：

$$b_1,\quad b_2,\quad . . . ,\quad b_m$$

（2）回归系数的确定

- 根据最小二乘法原理：求偏差平方和最小时的回归系数
- 偏差平方和：

$$Q=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-a-b_1x_1-b_2x_2-\ldots-b_mx_m)^2$$

- 根据：

$$\frac{\partial Q}{\partial a}=0 \qquad \frac{\partial Q}{\partial b_j}=0$$

得到正规方程组，正规方程组的解即为回归系数。

4.3.2 多元线性回归方程显著性检验

（1） **F检验法**

- 总平方和：

$$SS_T = L_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2$$

- 回归平方和：

$$SS_R = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = b_1 L_{1y} + b_2 L_{2y} + \ldots + b_m L_{my}$$

- 残差平方和：

$$SS_e = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = SS_T - SS_R$$

- 方差分析表:

多元线性回归方差分析表

差异源	SS	df	MS	F	显著性
回归	SS_R	m	$MS_R = SS_R/m$	$F = MS_R/MS_e$	
残差	SS_e	n－m－1	$MS_e = SS_e/(n-m-1)$		
总和	SS_T	n－1			

- **F**服从自由度为（$\boldsymbol{m}$，$\boldsymbol{n-m-1}$）的分布
- 给定的显著性水平α下，若$F > F_\alpha(m,\ n-m-1)$，则y与x_1，x_2，…，x_m间有显著的线性关系

（2）相关系数检验法

- 复相关系数（**multiple correlation coefficient**）**R**：

反映了一个变量y与多个变量（x_1，x_2，…，x_m）之间线性相关程度

- 计算式：
$$R^2=\frac{SS_R}{SS_T}=1-\frac{SS_e}{SS_T}$$

- **R**一般取正值，**0≤R≤1**

- $R=1$时，y与变量x_1，x_2，…，x_m之间存在严格的线性关系
- $R\approx 0$时，y与变量x_1，x_2，…，x_m之间不存在线性相关关系
- 当$0<R<1$时，变量之间存在一定程度的线性相关关系
- $R>R_{min}$时，y与x_1，x_2，…，x_m之间存在密切的线性关系

4.3.3 因素主次的判断

（1）偏回归系数的标准化

- 设偏回归系数b_j的标准化回归系数为P_j:

$$P_j = \left|b_j\right| \sqrt{\frac{L_{jj}}{L_{yy}}}$$

- P_j越大，则对应的因素（x_j）越重要

（2） 偏回归系数的显著性检验

- 计算每个偏回归系数的偏回归平方和SS_j：

$$SS_j = b_j L_{jy}$$

- SS_j的大小表示了因素x_j对试验指标y影响程度，对应的自由度$df_j = 1$

- $$F_j = \frac{MS_j}{MS_e} = \frac{SS_j}{MS_e}$$ 服从自由度为（1，$n-m-1$）的F分布

- 如果若$F < F_\alpha(1, n-m-1)$，，则说明x_j对y的影响是不显著的，这时可将它从回归方程中去掉，变成（$m-1$）元线性方程

（3）偏回归系数的t检验

- **计算偏回归系数 b_j 的标准差：**

$$s_{b_j} = \sqrt{\frac{SS_e}{L_{jj}}}$$

- **t值的计算：**

$$|t_j| = \frac{|b_j|}{s_{b_j}} = \frac{|b_j|}{\sqrt{MS_e / L_{jj}}} = \sqrt{\frac{b_j^2 L_{jj}}{MS_e}} = \sqrt{\frac{SS_j}{MS_e}} = \sqrt{F_j}$$

- **检验：**

单侧t分布表 $\rightarrow t_{\frac{\alpha}{2}}(1, n-m-1)$

如果 $|t_j| > t_{\frac{\alpha}{2}}(1, n-m-1)$

说明x_j对y的影响显著，否则影响不显著，

4.4 非线性回归分析

4.4.1 一元非线性回归分析

- 通过线性变换，将其转化为一元线性回归问题：
- **直角坐标中画出散点图；**
- **推测y与x之间的函数关系；**
- **线性变换；**
- **用线性回归方法求出线性回归方程；**
- **返回到原来的函数关系，得到要求的回归方程**

4.4.2 一元多项式回归

- 任何复杂的一元连续函数都可用高阶多项式近似表达：

$$y = a + b_1 x + b_2 x^2 + \ldots + b_m x^m$$

- 可以转化为多元线性方程：

$$\hat{y} = a + b_1 X_1 + b_2 X_2 + \ldots + b_m X_m$$

4.4.3 多元非线性回归

- 如果试验指标***y***与多个试验因素$\boldsymbol{x_j}$之间存在非线性关系，如二次回归模型 ：

$$\hat{y} = a + \sum_{j=1}^{m} b_j x_j + \sum_{i=1}^{m} b_{jj} x_j^2 + \sum_{j<k} b_{jk} x_j x_k$$

4.5 Excel在回归分析中的应用

4.5.1 “规划求解”在回归分析中应用

- 解方程组
- 最优化

4.5.2 Excel内置函数在回归分析中应用

4.5.3 Excel图表功能在回归分析中的应用

4.5.4 分析工具库在回归分析中应用