第八章 一元线性回归和相关分析

揚州大學 YANGZHOU UNIVERSITY

- 8.1 线性回归和线性相关的概念
- 8.2 线性回归方程和离回归标准误
- 8.3 线性回归方程的假设测验
- 8.4 线性回归的区间估计
- 8.5 线性相关分析
- 8.6 线性回归和相关的内在关系及应用注意事项

8.1 线性回归和线性相关的概念

8.1.1 变量间的函数关系与相关关系

函数关系：即当一个变量每取一个值时，相应的另一个变量必然有一个确定值与之对应 。 $y = f(x)$

例：（1） 某种商品的销售额(y)与销售量(x)之间的关系可表示为

$y = p\,x$ (p 为单价)

（2）圆的面积(S)与半径(R)之间的关系可表示为

$S = \pi R^2$

8.1 线性回归和线性相关的概念

揚州大學 YANGZHOU UNIVERSITY

8.1.1 变量间的函数关系与相关关系

相关关系：一定范围内，一个变量的任意观察值，虽然没有另一个变量的确定值与之对应，但却有一个特定的条件概率分布与之对应。

$$y = f(x)+e$$

例：(1)粮食亩产量与施肥量之间的关系

(2)玉米穗长与穗粗的关系

8.1 线性回归和线性相关的概念

揚州大學 YANGZHOU UNIVERSITY

8.1.2 散点图

1.定义

散点图（scatter diagram）：将两个变量的n对观察值（X_1，Y_1），（X_2，Y_2）….（X_n，Y_n）分别以坐标点的形式标记于同一直角坐标的平面上。

8.1 线性回归和线性相关的概念

揚州大學 YANGZHOU UNIVERSITY

8.1.2 散点图

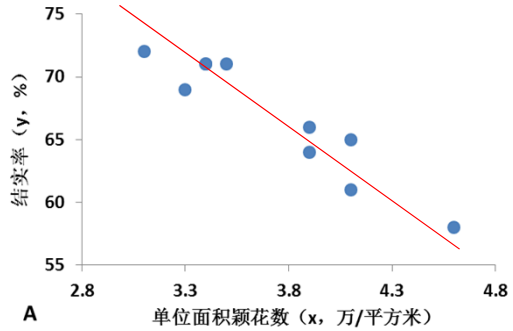
2.作用

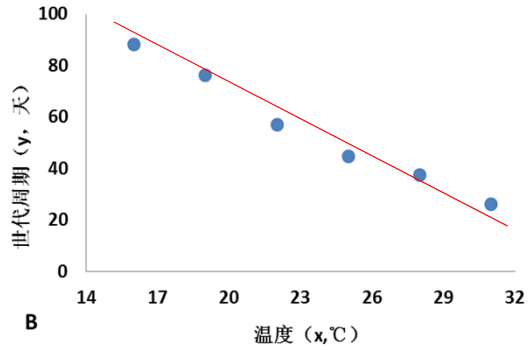
① 判断两变量间是否为线性关系

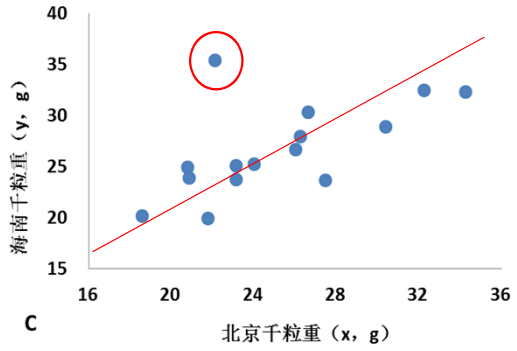
② 判断两变量间的关系是正向还是负向

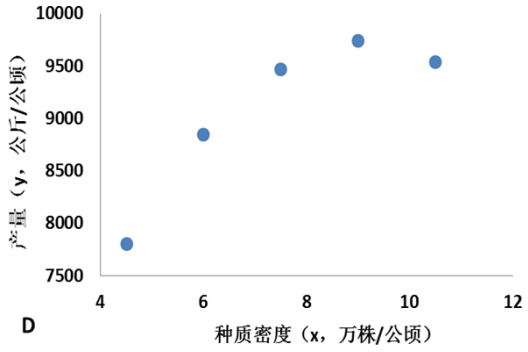
③ 初步判断两变量间的相关密切程度

④ 去除极端值









8.1 线性回归和线性相关的概念

揚州大學 YANGZHOU UNIVERSITY

8.1.3 自变量与依变量

两个变量间若具有原因和结果的关系，则称这两个变量间存在因果关系。

原因变量为自变量（independent variable），以*X*表示；

结果变量为依变量（dependent variable），以*Y*表示。

粮食亩产量与施肥量之间的关系

穗长与穗粗的关系

8.1 线性回归和线性相关的概念

揚州大學 YANGZHOU UNIVERSITY

8.1.4 统计分析的任务

1. 有自变量依变量之分： 回归分析

$\hat{y}$ =f（x）:表示Y依X的回归方程;

$\hat{y}$：在给定x时，y的估计值。

$\hat{y} = b_0 + b_1 x$为一元线性回归方程。

8.1 线性回归和线性相关的概念

揚州大學 YANGZHOU UNIVERSITY

8.1.4 回归分析和相关分析

2. 无自变量依变量之分： 相关分析

目的：计算表示Y和X线性相关密切程度和性质的统计数（相关系数, 记为*r*）, 并测验其显著性。

第八章 一元线性回归和相关分析

揚州大學 YANGZHOU UNIVERSITY

- 8.1 线性回归和线性相关的概念
- **8.2 线性回归方程和离回归标准误**
- 8.3 线性回归方程的假设测验
- 8.4 线性回归的区间估计
- 8.5 线性相关分析
- 8.6 线性回归和相关的内在关系及应用注意事项

8.2 线性回归方程和离回归标准误

8.2.1 线性回归方程及其参数估计

（1）线性回归方程

$$\hat{y} = a + bx$$ （a、b为两个参数）

式为变量Y依X的一元线性回归方程；

$\hat{y}$：在给定x时，y的估计值。

a表示回归截距，是x=0时的$\hat{y}$值;

b表示回归系数，是x每增加一个单位数时，$\hat{y}$将要平均增加（b>0）或减少（b<0）的单位数，体现x对y影响的性质和程度。

8.2 线性回归方程和离回归标准误

8.2.1 线性回归方程及其参数估计

$$Q=\sum_{1}^{n}(y-\hat{y})^2=\sum_{1}^{n}(y-a-bx)^2 \text{为最小}$$

分别对a和b求偏导数并令其为0，可得：

$$\begin{cases} an+b\sum x=\sum y \\ a\sum x+b\sum x^2=\sum xy \end{cases}$$

得

$$a=(\sum y-b\sum x)/n=\overline{y}-b\overline{x}$$

8.2 线性回归方程和离回归标准误

揚州大學 YANGZHOU UNIVERSITY

8.2.1 线性回归方程及其参数估计

$$b=\frac{\sum xy-\frac{1}{n}\sum x\sum y}{\sum x^2-\frac{1}{n}(\sum x)^2}=\frac{\sum(x-\bar{x})\ (y-\bar{y})}{\sum(x-\bar{x})^2}=\frac{SP}{SS_x}$$

$$\hat{y}=(\bar{y}-b\bar{x})+bx=\bar{y}+b(x-\bar{x})$$

8.2 线性回归方程和离回归标准误

8.2.1 线性回归方程及其参数估计

最小二乘法

基本性质：

性质1：$\sum (y-\hat{y})^2$=最小

性质2：$\sum (y-\hat{y})$=0

性质3：回归直线必然经过（$\bar{x}$，$\bar{y}$）

8.2 线性回归方程和离回归标准误

揚州大學 YANGZHOU UNIVERSITY

8.2.1 线性回归方程及其参数估计

（2）线性回归方程的运算

例题 1

x	y
1	3
2	4
3	5
4	5
5	7

（1）计算6个一级数据 （2）计算5个二级数据

（1）一级数据（由原始数据直接求出）

$$\sum x \quad \sum x^2 \quad \sum y \quad \sum y^2 \quad \sum xy \quad n$$

8.2 线性回归方程和离回归标准误

8.2.1 线性回归方程及其参数估计

（2）线性回归方程的运算

（2）二级数据

$$\bar{x} = \sum x / n \quad \bar{y} = \sum y / n$$

$$SS_x = \sum x^2 - (\sum x)^2 / n$$

$$*SS_y = \sum y^2 - (\sum y)^2 / n$$

$$SP = \sum xy - \sum x \sum y / n$$

因而有：

$$b = SP / SS_x$$

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

$$\hat{y} = 2.1 + 0.9x$$

8.2 线性回归方程和离回归标准误

揚州大學 YANGZHOU UNIVERSITY

例题 2 8.2.1 线性回归方程及其参数估计

一些夏季害虫盛发期的早迟和春季温度高低有关。如果可以根据春季温度高低预测害虫盛发的时期，则可以适时采取预防措施。资料显示，某地10年测定3月下旬至5月上旬旬平均温度累积值（x，旬·度）和一代桑螟盛发期（y，以5月30日为0）的关系。试计算其线性回归方程。

x累积温	y盛发期
73.0	7
71.6	11
74.3	2
70.0	10
77.7	-2

$\sum x \quad \sum x^2 \quad \sum y \quad \sum y^2 \quad \sum xy \quad n$

$\overline{x} \quad \overline{y} \quad SS_x \quad SSy \quad SP$

8.2 线性回归方程和离回归标准误

8.2.1 线性回归方程及其参数估计

（3）线性回归方程图示构建

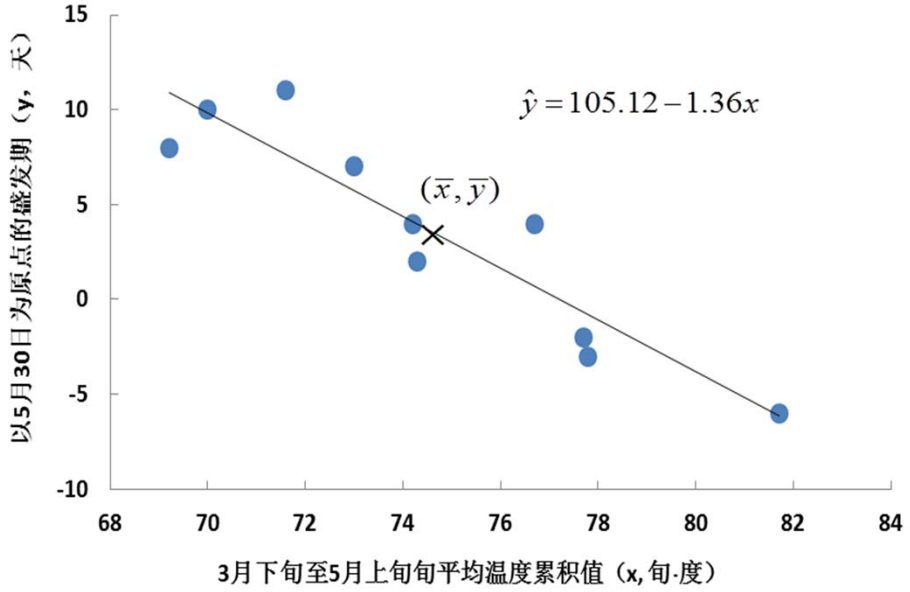
a、建立一个直角坐标系，x为横坐标，y为纵坐标

b、构建散点图

c、取x的最小值带入方程得 $\hat{y}1$ ； 取x的最大值带入方程得$\hat{y}2$

d、直线连接（x_{min}, $\hat{y}1$ ）（x_{max}, $\hat{y}2$ ）

e、空白处列出方程



8.2 线性回归方程和离回归标准误

揚州大學 YANGZHOU UNIVERSITY

8.2.2 离回归标准误

为确定回归方程的精确度，必须估计分布的变异度，这个变异度的统计数叫做线性回归的估计标准误（standard error of estimate），或离回归标准差（standard deviation from regression），记作$S_{y/x}$，定义为：

$$s_{y/x}=\sqrt{\frac{Q}{n-2}}=\sqrt{\frac{\Sigma(y-\hat{y})^2}{n-2}}$$

Q为离回归平方和或回归离差平方和（sum of squares due to deviation from regression）；*n*-2为与Q对应的自由度。

8.2 线性回归方程和离回归标准误

8.2.2 离回归标准误

$$Q=\Sigma(y-\hat{y})^2=SS_y-\frac{(SP)^2}{SS_x}$$

计算例题1中 $\mathbf{S_{y/x}}$

$$Q=SS_y-\frac{(SP)^2}{SS_x}=8.8-81/10=0.7$$

$$s_{y/x}=\sqrt{\frac{Q}{n-2}}=\sqrt{\frac{0.7}{3}}=0.483$$

第八章 一元线性回归和相关分析

揚州大學 YANGZHOU UNIVERSITY

- 8.1 线性回归和线性相关的概念
- 8.2 线性回归方程和离回归标准误
- **8.3 线性回归方程的假设测验**
- 8.4 线性回归的区间估计
- 8.5 线性相关分析
- 8.6 线性回归和相关的内在关系及应用注意事项

8.3 线性回归方程的假设测验

揚州大學 YANGZHOU UNIVERSITY

8.3.1 线性回归的基本假定

3个基本假定

（1）在x和y两变量中，x是固定变量，而y是一个有误差的随机变量。

（2）在可能取值区间内，x变量上任一值,都存在着y变量的一个条件正态总体。

（3）这一系列y的条件正态总体遵循N~（$\mu_{y/x}$，$\sigma^2_{y/x}$），$\sigma^2_{y/x}$不因x的不同而不同，$\mu_{y/x}$随x的改变呈线性改变，关系可表示为:

$$\mu_{Y/X} = \alpha + \beta X$$

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

$\mu_{Y/X}$: 条件总体平均数

8.3 线性回归方程的假设测验

8.3.2 线性回归方程的假设测验

目的：测验*X*和*Y*有无线性关系。 $H_0: \beta = 0$ $H_A: \beta \neq 0$

可以通过*F*测验或*t*测验作出。

（1）*F*测验

总变异=离回归变异+线性回归变异

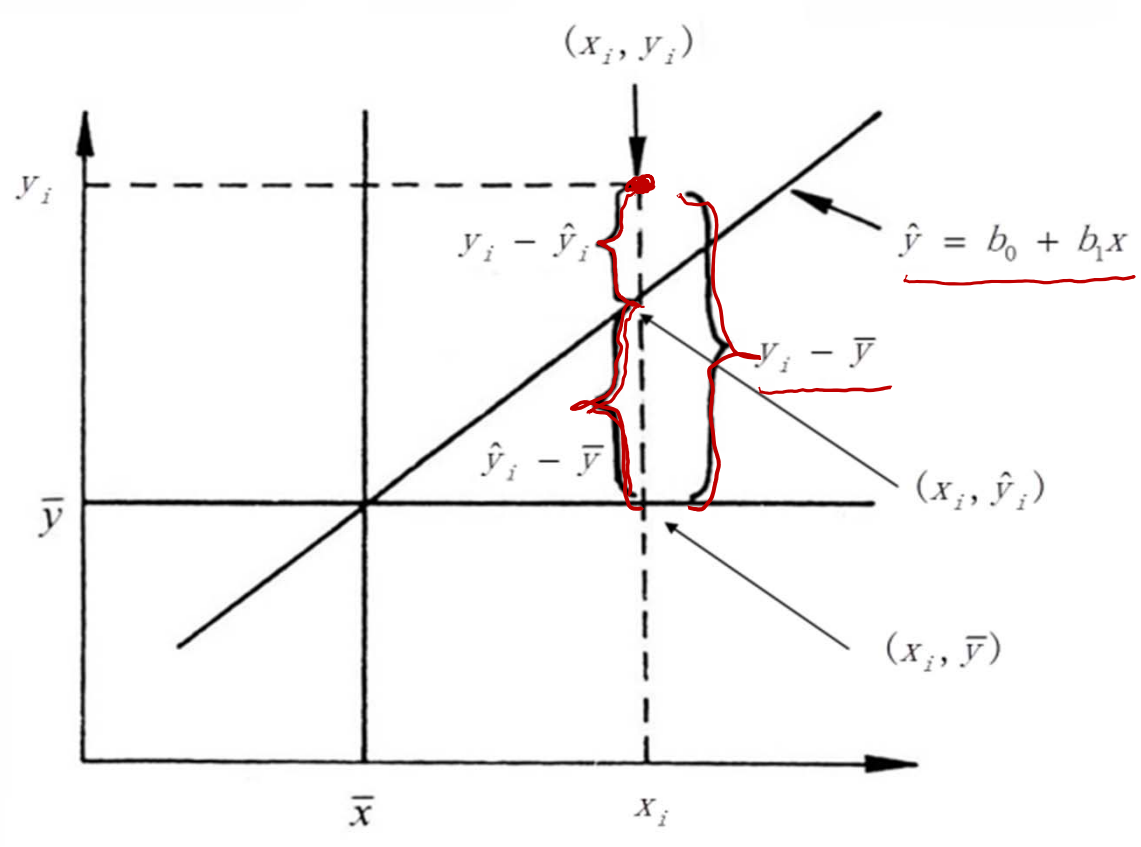
总变异平方和： $SS_y = \sum (y - \bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n}$； $df_y = n-1$

离回归平方和： $Q = SS_y - \frac{SP^2}{SS_X}$； $df_Q = n-2$

线性回归平方和： $U = SS_y - Q = \frac{SP^2}{SS_X}$； $df_U = 1$

8.3 线性回归方程的假设测验

8.3.2 线性回归方程的假设测验



8.3 线性回归方程的假设测验

8.3.2 线性回归方程的假设测验

（1）*F*测验

例题1, 已知 X: 1、2、3、4、5 ； Y: 3、4、5、5、7，其线性回归方程为 $\hat{y}=2.1+0.9x$，试测验该线性回归的显著性。

S.O.V	*df*	*SS*	*MS*	*F*
线性回归	1	8.1	8.1	34.76**
离回归	3	0.7	0.233	
总变异	4	8.8		

H_0: $\beta = 0$　H_A: $\beta \neq 0$

8.3 线性回归方程的假设测验

8.3.2 线性回归方程的假设测验

（2）*t*测验

H_0: $\beta = 0$ H_A: $\beta \neq 0$

$t = \dfrac{b - \beta}{s_b}$ 测验线性回归的显著性。

$$s_b = \frac{s_{y/x}}{\sqrt{ss_x}} = \sqrt{\frac{Q}{(n-2)ss_x}}$$

*t*遵循自由度为*n*-2的*t*分布。

8.3 线性回归方程的假设测验

揚州大學 YANGZHOU UNIVERSITY

8.3.2 线性回归方程的假设测验

（2）*t*测验

已知 X: 1、2、3、4、5 ； Y: 3、4、5、5、7，其线性回归方程为$\hat{y}$=2.1+0.9x，试测验该线性回归的显著性。

$$H_0:\ \beta=0 \quad H_A:\ \beta\neq 0$$

df=n-2=3　　$t_{0.05}$=3.182　　$t_{0.01}$=5.841

$$s_b=\frac{s_{y/x}}{\sqrt{ss_x}}=\sqrt{\frac{Q}{(n-2)ss_x}}=\sqrt{\frac{0.7}{3\times 10}}=0.153 \qquad t=\frac{b-\beta}{s_b}=\frac{b}{s_b}=\frac{0.9}{0.153}=5.882$$

$|t| > t_{0.01}$ 所以否定H_0，接受H_A

8.3 线性回归方程的假设测验

8.3.2 线性回归方程的假设测验

*F*测验和*t*测验的关系

小结：$$F=\frac{MS_U}{MS_Q}=\frac{U}{Q\ (n-2)}=\frac{sp^2/ss_x}{S^2_{y/x}}=\frac{(SP/SS_x)^2}{S^2_{y/x}/SS_x}=\frac{b_1^2}{S^2_{b_1}}=t^2$$

第八章 一元线性回归和相关分析

揚州大學 YANGZHOU UNIVERSITY

- 8.1 线性回归和线性相关的概念
- 8.2 线性回归方程和离回归标准误
- 8.3 线性回归方程的假设测验
- **8.4 线性回归的区间估计**
- 8.5 线性相关分析
- 8.6 线性回归和相关的内在关系及应用注意事项

8.4 线性回归方程的区间估计

揚州大學 YANGZHOU UNIVERSITY

α , β , $\mu_{Y/X}$

a , b , $\hat{y}$

1.条件总体平均数$\mu_{Y/X}$的区间估计

置信区间:$\left[L_1=\hat{y}-t_{\alpha,n-2}s_{\hat{y}},\ L_2=\hat{y}+t_{\alpha,n-2}s_{\hat{y}}\right]$

$$s_{\hat{y}}=\sqrt{s_{\bar{y}}^2+s_b^2(x-\bar{x})^2}=\sqrt{\frac{s_{y/x}^2}{n}+\frac{s_{y/x}^2}{SS_x}(x-\bar{x})^2}=s_{y/x}\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{SS_x}}$$

$$\hat{y}=\bar{y}+b(x-\bar{x})$$

8.4 线性回归方程的区间估计

揚州大學 YANGZHOU UNIVERSITY

1.条件总体平均数$\mu_{Y/X}$ 的区间估计

例: X: 1、2、3、4、5 ； Y: 3、4、5、5、7，其线性回归方程为$\hat{y}$=2.1+0.9x，$\hat{y}$=2.1+0.9x ，当x=4时，$\mu_{y/x}$的95%的置信区间？

$$\left[L_1=\hat{y}-t_{\alpha,n-2}s_{\hat{y}},\ L_2=\hat{y}+t_{\alpha,n-2}s_{\hat{y}}\right]$$

$t_{0.05}$=3.182

$$s_{\hat{y}}=s_{y/x}\sqrt{\frac{1}{n}+\frac{(x-\overline{x})^2}{SS_x}}$$

$$s_{y/x}=\sqrt{\frac{Q}{n-2}}=\sqrt{\frac{0.7}{3}}=0.483$$

$$[4.86,\ 6.54]$$

8.4 线性回归方程的区间估计

揚州大學 YANGZHOU UNIVERSITY

2.回归截距α的区间估计

$$\text{置信区间}:\left[L_1 = a - t_{\alpha,n-2}s_a,\ L_2 = a + t_{\alpha,n-2}s_a\right]$$

$$s_a = s_{y/x}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{SS_x}} \quad \xleftarrow{x=0} \quad s_{\hat{y}} = s_{y/x}\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{SS_x}}$$

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

例1中α的95%的置信区间？ $\hat{y}$=2.1+0.9x $[0.49,\ 3.71]$

8.4 线性回归方程的区间估计

揚州大學 YANGZHOU UNIVERSITY

3.回归系数β的区间估计

置信区间：$\left[L_1 = b - t_{\alpha,n-2}s_b,\ L_2 = b + t_{\alpha,n-2}s_b\right]$

$$s_b = \frac{s_{y/x}}{\sqrt{ss_x}} = \sqrt{\frac{Q}{(n-2)ss_x}}$$

第八章 一元线性回归和相关分析

揚州大學 YANGZHOU UNIVERSITY

- 8.1 线性回归和线性相关的概念
- 8.2 线性回归方程和离回归标准误
- 8.3 线性回归方程的假设测验
- 8.4 线性回归的区间估计
- **8.5 线性相关分析**
- 8.6 线性回归和相关的内在关系及应用注意事项

8.5 线性相关分析

揚州大學 YANGZHOU UNIVERSITY

8.5.1 相关系数和决定系数

1. 相关系数

确定一对成直线关系的两变量，若不需要用x来预测y，仅需了解两者的相关程度和性质，则需计算表示两者密切程度和性质的统计数——相关系数。一般以ρ表示总体相关系数，以r表示样本相关系数（correlation coefficient）。

$$r=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}=\frac{SP}{\sqrt{SS_xSS_y}}$$

8.5 线性相关分析

8.5.1 相关系数和决定系数

2.决定系数

y总变异中由x的改变而呈线性改变的平方和称（$U_{y/x}$）占总平方和SS_y的比率，称为决定系数（determination coefficient），记作r^2：

$$r^2=\frac{(SP)^2}{SS_xSS_y}$$

8.5 线性相关分析

揚州大學 YANGZHOU UNIVERSITY

8.5.1 相关系数和决定系数

2.相关系数和决定系数的区别与联系

① 决定系数r^2的取值区间$[0,1]$，相关系数r的取值区间$[-1,1]$。

② 相关系数r可表示两个变量相关的性质（正相关或者负相关），决定系数r^2则不可以，r^2能表示两个变量之间的相关密切程度。

8.5 线性相关分析

揚州大學 YANGZHOU UNIVERSITY

8.5.2 相关系数的假设测验

1. 测验相关系数取自无线性关系总体的概率

*r*的抽样误差 $S_r=\sqrt{\frac{1-r^2}{n-2}}$ ；

df=n-2 ；

$t=\frac{r}{S_r}$

2. 可通过查表的方法得出; df=n-2 的*r*的临界值 r_α，实测值与其比较即可。

8.5 线性相关分析

揚州大學 YANGZHOU UNIVERSITY

8.5.2 相关系数的假设测验

例题1，已知 X: 1、2、3、4、5 ； Y: 3、4、5、5、7，其线性回归方程为 $\hat{y}$=2.1+0.9x，试测验该资料的*r*值的显著性。

假设 H_0: ρ =0，H_A： $\rho \neq 0$

取 α=0.05　*df*=n-2=3　$t_{0.05}$=3.182

α=0.01　*df*=n-2=3　$t_{0.01}$=5.841

8.5 线性相关分析

8.5.2 相关系数的假设测验

$$t=\frac{r-\rho}{S_r}=\frac{r}{S_r} \qquad r=\frac{SP}{\sqrt{SS_x SS_y}}=\frac{9}{\sqrt{10*8.8}}=0.9564$$

$$S_r=\sqrt{\frac{1-r^2}{n-2}}=\sqrt{\frac{1-0.9594^2}{3}} \qquad t=\frac{r}{S_r}=5.894$$

$|t|>t_{0.01}$ 否定H_0，接受H_A

所以该双变量资料是有极显著的线性正相关关系的。