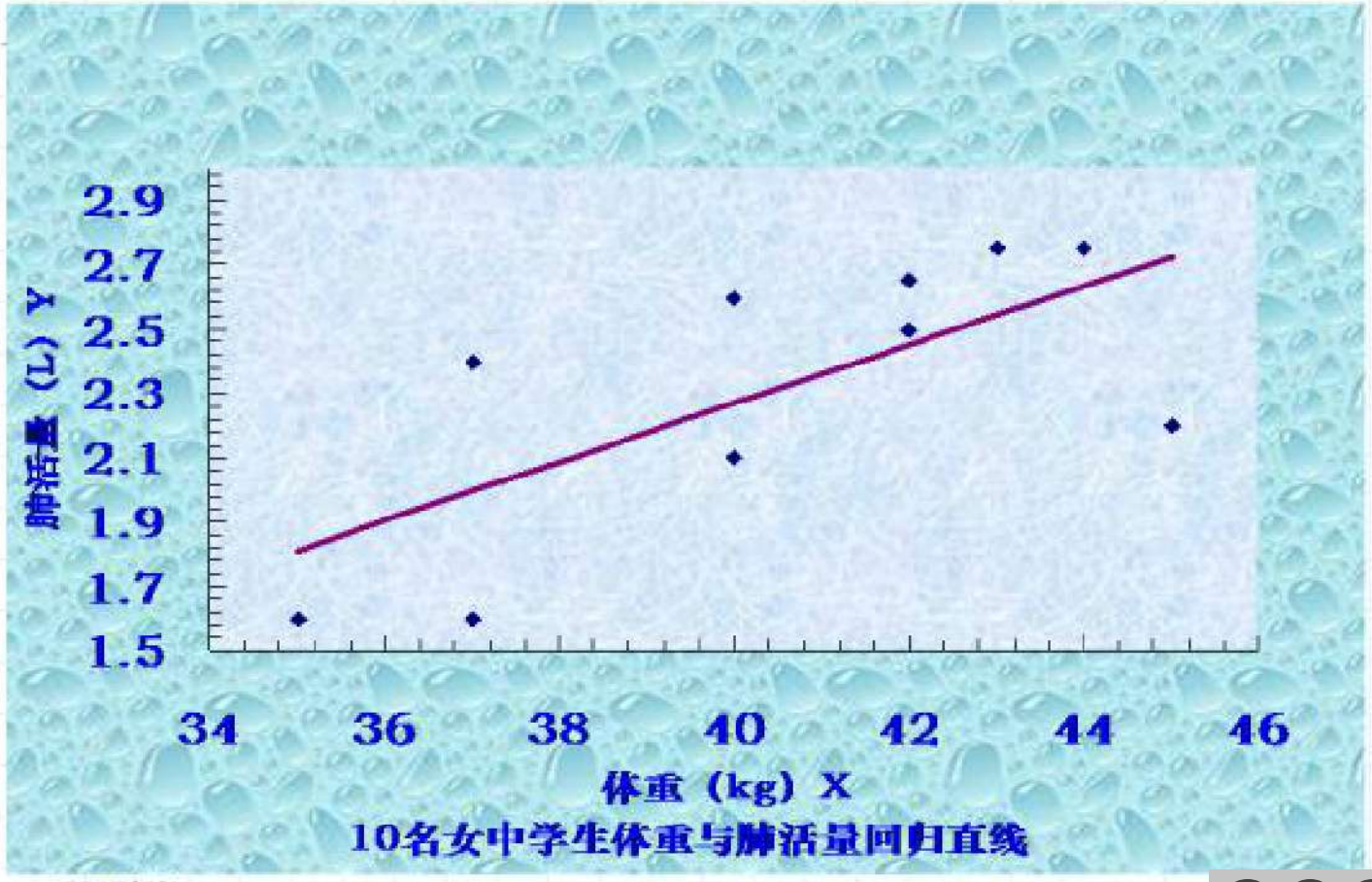
第九章 直线相关与回归分析

Page154~170

第一节　直线相关

Page154~170

一、相关系数



10名女中学生体重与肺活量回归直线

（一）两变量间关系的图示

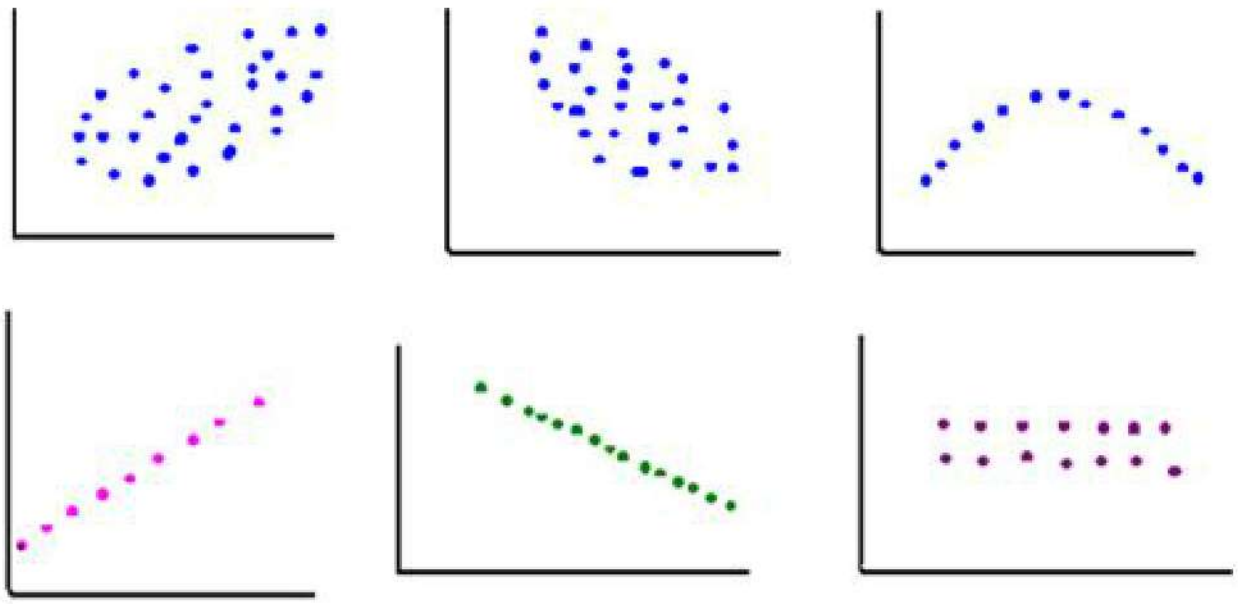


图9-1　相关关系示意图

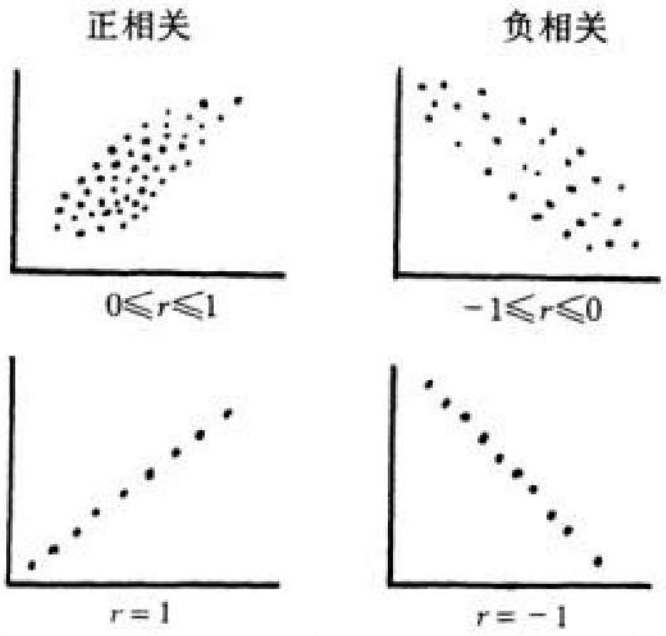


图9—6　相关系数示意图

散点呈椭圆形分布，

X、*Y* 同时增减——正相关（positive correlation）；

X、*Y* 此增彼减——负相关（negative correlation）。

散点在一条直线上，

X、*Y* 变化趋势相同——完全正相关；

反向变化——完全负相关。

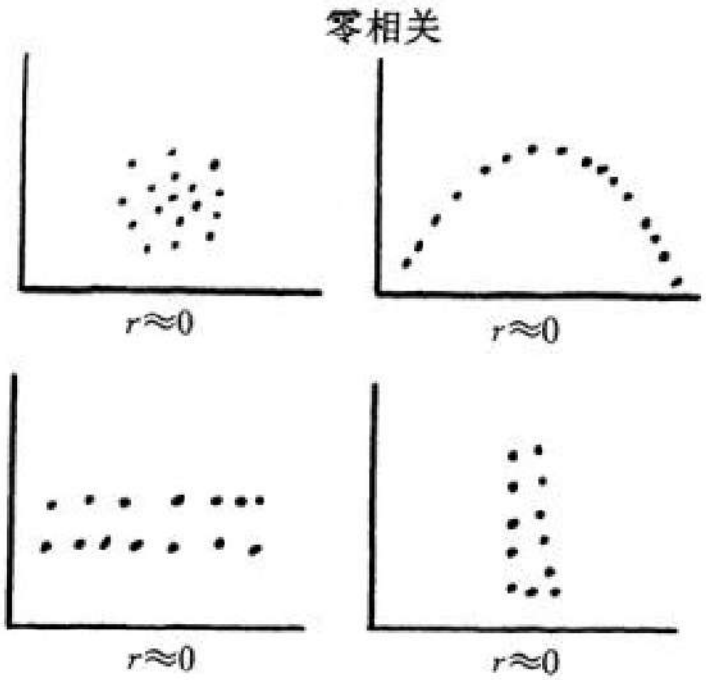


图9—6　相关系数示意图

X、*Y* 变化互不影响或无直线相关关系——零相关

(zero correlation)

直线相关的种类及分析的任务

- 直线相关的种类：正相关，$0<r\leq 1$

  零相关，$r=0$

  负相关，$-1\leq r<0$

- 直线相关分析：用相关系数（r）描述两变量间是否有直线关系以及直线关系的方向和密切程度的分析方法。

（二）相关系数的定义和意义

定义：说明具直线关系的两个变量间，相关关系的密切程度与相关方向的指标。

意义：r的大小表示密切程度，r的正负表示相关方向。

相关系数的计算公式

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2}\sqrt{\sum (Y - \bar{Y})^2}} = \frac{l_{XY}}{\sqrt{l_{XX} \cdot l_{YY}}}$$

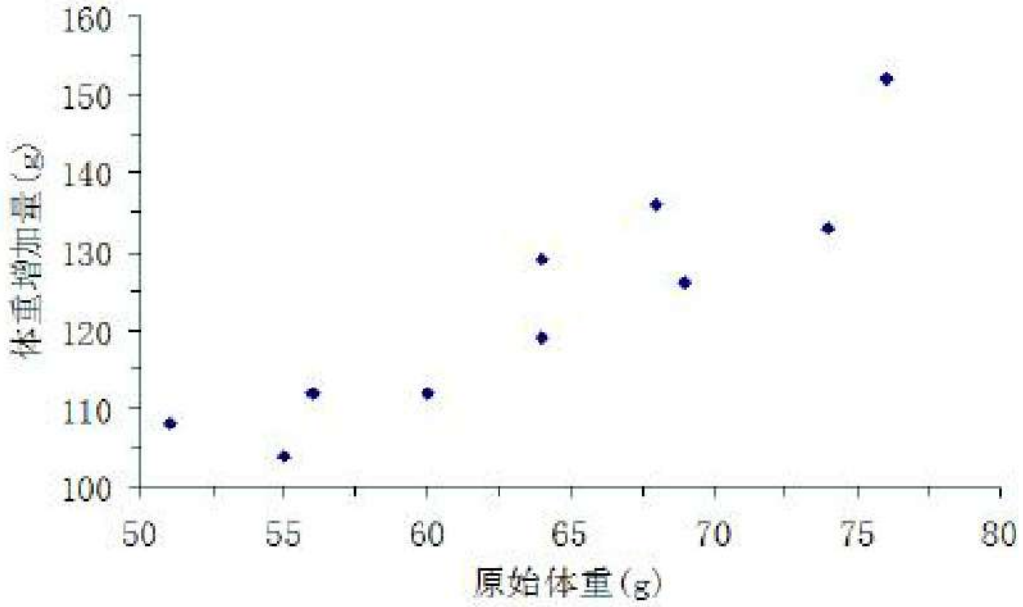
（三）相关系数的计算

例9-1　某研究人员测得10只鼠龄24~64天的雌性老鼠的原始体重（g），以及喂养高蛋白饲料后的体重增加量（g）。资料见表9-1。试分析老鼠体重增加量与原始体重的关系。

表9-1　10只老鼠的原始体重与体重增加量的关系

鼠号 (1)	原始体重X	(2) 体重增加量Y (3)
1	55	104
2	64	129
3	76	152
4	64	119
5	74	133
6	60	112
7	69	126
8	68	136
9	56	112
10	51	108

分析..... 1.绘制散点图



10只老鼠的原始体重与体重增加量的散点图

注　　意

- 当两变量的散点图呈直线趋势时，用直线相关分析。
- 当两变量的散点图不呈直线趋势，而呈某种曲线趋势时，用曲线相关分析

本例：　可用直线相关分析

分析…… 2.计算相关系数

①由原始数据求

$$\sum X \quad \sum Y \quad \sum X^2 \quad \sum Y^2 \quad \sum XY$$

表9-1　10只老鼠的原始体重与体重增加量的关系

鼠号 (1)	原始体重X (2)	体重增加量 Y (3)	X^2 (4)	Y^2 (5)	XY (6)
1	55	104	3025	10816	5720
2	64	129	4096	16641	8256
3	76	152	5776	23104	11552
4	64	119	4096	14161	7616
5	74	133	5476	17689	9842
6	60	112	3600	12544	6720
7	69	126	4761	15876	8694
8	68	136	4624	18496	9248
9	56	112	3136	12544	6272
10	51	108	2601	11664	5508
合计	637	1231	41191	153535	79428

分析…… 2.计算相关系数

②计算 l_{xx} l_{yy} l_{xy}

$$l_{XX}=\sum(X-\bar{X})^2=\sum X^2-\frac{(\sum X)^2}{n}=41191-\frac{637^2}{10}=614.1$$

$$l_{YY}=\sum(Y-\bar{Y})^2=\sum Y^2-\frac{(\sum Y)^2}{n}=153535-\frac{1231^2}{10}=1998.9$$

$$l_{XY}=\sum(X-\bar{X})(Y-\bar{Y})=\sum XY-\frac{\sum X\sum Y}{n}=79428-\frac{637\times1231}{10}=1013.3$$

分析..... 2. 计算相关系数

③按公式计算r

$$r=\frac{l_{XY}}{\sqrt{l_{XX}l_{YY}}}=\frac{1013.3}{\sqrt{614.1\times1998.9}}=0.9146$$

此处相关系数r>0，说明两变量有正的直线相关关系.

分析..... 2. 计算相关系数

注意：此处相关系数r>0，并不表示两变量一定有正的直线相关关系，因为只是样本相关系数

二、相关系数的统计推断

（一）相关系数的假设检验

目的：是判断两变量的总体是否有相关关系

t检验：样本相关系数r与总体相关系数ρ的比较

$$t=\frac{r-0}{S_r}=\frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \qquad \upsilon=n-2$$

表法：直接查相关系数界值表得到相应的概率P。

$r \neq 0$原因：①由于抽样误差引起，$\rho=0$

②存在相关关系，　　$\rho \neq 0$

公式　$$t_r = \frac{|r-0|}{S_r} = \frac{|r|}{\sqrt{\frac{1-r^2}{n-2}}}$$，$\upsilon = n-2$

S_r——相关系数的标准误

H_0：$\rho=0$，即两变量间无直线相关关系

H_1：$\rho\neq0$，即两变量间有直线相关关系

α=0.05

本例n=10，r=0.9146，按公式（9-2）

$$t=\frac{r}{\sqrt{\frac{1-r^2}{n-2}}}=\frac{0.9146}{\sqrt{\frac{1-0.9146^2}{10-2}}}=6.397$$

t界值表 得 p<0.001，按 α=0.05 水准拒绝H_0，接受 H_1

老鼠的原始体重与体重增加量间呈正相关关系

（二）总体相关系数的估计

当 $\rho=0$时，r的抽样分布对称，近似正态，但是$\rho\neq 0$，r的抽样分布呈偏态分布，故不能按t或z分布的原理来估计总体相关系数。

$$z=\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \Rightarrow z\pm z_{\alpha}\frac{1}{\sqrt{n-3}}$$

$$r=\frac{e^{2z}-1}{e^{2z}+1}$$

（二）总体相关系数的估计……本例

$$z=\frac{1}{2}\ln\left(\frac{1+0.9146}{1-0.9146}\right)=1.5550$$

$$1.5550\pm1.96\times\frac{1}{\sqrt{10-3}}=(0.8142,2.2958)$$

下限：$r=\dfrac{e^{2\times0.8142}-1}{e^{2\times0.8142}+1}=0.6719$　上限同理

得总体相关系数的95%置信区间为(0.6719，0.9799)

三、进行直线相关分析时的注意事项

- 进行直线相关分析要有实际意义
- 相关表示两个变量之间的相关关系是双向的
- 相关系数的计算适用双变量正态分布资料
- 据公式计算出的相关系数仅是样本相关系数
- 不把相关系数的假设检验结果误认为两事物或两现象间相关的密切程度
- 相关分析是用相关系数来描述两个变量间相关关系的密切程度和方向，而两个事物之间的关系，可能是依存因果关系，也可能是相互伴随关系.

第二节　等级相关（秩相关）

适用资料：(1)不服从双变量正态分布

(2) 总体分布类型未知

(3) 原始数据用等级表示

等级相关系数 r_s（即Spearman Correlation Coefficient）——反映两变量间相关的密切程度与方向。

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

1. n为样本含量，d为x，y的秩次之差

2. 其值一定介于-1与1之间

3. 秩相关系数的意义与积相关系数相同

表 9-3　某省 1995 年到 1999 年居民死因构成与 WYPLL 构成

死因类别	死因构成(%)		WYPLL 构成(%)		d	d^2	PQ
(1)	X(2)	P(3)	Y(4)	Q(5)	(6)=(3)-(5)	(7)=(6)^2	(8)=(3)(5)
1	0.03	1	0.05	1	0	0	1
2	0.14	2	0.34	2	0	0	4
3	0.20	3	0.93	6	-3	9	18
4	0.43	4	0.69	4	0	0	16
5	0.44	5	0.38	3	2	4	15
6	0.45	6	0.79	5	1	1	30
7	0.47	7	1.19	8	-1	1	56
8	0.65	8	4.74	12	-4	16	96
9	0.95	9	2.31	9	0	0	81
10	0.96	10	5.95	14	-4	16	140
11	2.44	11	1.11	7	4	16	77
12	2.69	12	3.53	11	1	1	132
13	3.07	13	3.48	10	3	9	130
14	7.78	14	5.65	13	1	1	182
15	9.82	15	33.95	18	-3	9	270
16	18.93	16	17.16	17	-1	1	272
17	22.59	17	8.42	15	2	4	255
18	27.96	18	9.33	16	2	4	288
合 计	—	171	—	171	—	92	2063

$$r_s = 1 - \frac{6\sum d^2}{n^3 - n} = 1 - \frac{6 \times 92}{18^3 - 18} = 0.905$$

注意：相同秩次较多时应校正 r_s。

$$r_s' = \frac{\left[(n^3 - n)/6\right] - (T_X + T_Y) - \sum d^2}{\sqrt{\left[(n^3 - n)/6\right] - 2T_X}\sqrt{\left[(n^3 - n)/6\right] - 2T_Y}}$$

$$T_X(\text{或}T_Y) = \Sigma(t^3 - t)/12$$

◆ 秩相关系数的假设检验

$$H_0 : \rho_S = 0$$

$$H_1 : \rho_S > 0$$

$$\alpha = 0.05$$

$$r_s = 1 - \frac{6\sum d^2}{n^3 - n} = 1 - \frac{6 \times 92}{18^3 - 18} = 0.905$$

N=18, 查附表13, 得p<0. 0005, 拒绝 H_0, 接受H_1

有正相关存在

分类变量的关联性分析

◆适用条件

对定性变量之间的联系通用的方法是根据两个定性变量交叉分类基数所得的频数资料（列联表）作关联性分析，即关于两种属性独立性的卡方检验

计算公式

$$\chi^2 = \sum \frac{(A-T)^2}{T}$$

交叉分类2×2表的关联分析

- 例　为观察婴儿腹泻是否与喂养方式有关，某医院儿科随机收集了消化不良的婴儿82例，对每个个体分别观察腹泻与否和喂养方式两种属性，2×2种结果分类记述如表11-3所示，试分析两种属性的关联性。

婴儿腹泻与喂养方式的关系

喂养方式	腹泻		合计
	有	无	
人工	30	10	40
母乳	17	25	42
合计	47	35	82

◆假设检验

H_0 :喂养方式与婴儿腹泻之间相互独立

H_1 :喂养方式与婴儿腹泻之间有关联

$\alpha = 0.05$

将表中各数据代入公式(7-7)得

$\chi^2 = 9.98$

$\chi^2 > \chi^2_{0.005,1} = 7.78, p < 0.005$，拒绝原假设,说明婴儿腹泻与喂养方式之间存在关联性.

婴儿腹泻与喂养方式的关系

喂养方式	腹泻		合计
	有	无	
人工	30	10	40
母乳	17	25	42
合计	47	35	82

二 2×2配对资料的关联性分析

- 例 有56份咽喉涂抹标本，把每份标本一分为二，依同样的条件分别接种于甲乙两种白喉杆菌培养基上，观察白喉菌生长的情况，结果如表11-5，问两种培养基的结果有无关联？

两种白喉杆菌培养结果

甲培养基	乙培养基		合计
	+	-	
+	22	18	40
-	2	14	16
合计	24	32	56

◆假设检验

H_0 :两种培养基的结果之间互相独立

H_1 :两种培养基的结果之间有关联

$\alpha = 0.05$

将本例数据代入公式(7-7)得

$$\chi^2 = 9.98 > 3.84, p < 0.05$$

有理由拒绝零假设,可以认为甲、乙两种培养基之间存在关联性

三　R×C表分类资料的关联性分析

◆例　某地居民主要有三种祖籍，均流行甲状腺肿。为探索甲状腺肿类型与祖籍是否有关联，现根据居民甲状腺肿筛查结果，按甲状腺肿类型与祖籍两种属性交叉分类，得表11-6的资料。问甲状腺肿类型与祖籍见有否关联？

某地居民按甲状腺肿类型与祖籍两种属性的交叉分类表

祖籍	甲状腺肿类型			合计
	弥漫型	结节型	混合型	
甲	486	2	4	492
乙	133	260	51	444
丙	100	315	85	500
合计	719	577	140	1436

◆ 假设检验

H_0：甲状腺类型与祖籍无关

H_1：甲状腺类型与祖籍有关联

$\alpha = 0.05$

同样作检验得

$$\chi^2 = 9.98$$

由$\nu = (3-1)(3-1) = 4$, 查χ^2 界值表$\chi^2 > \chi^2_{0.005,4} = 18.55, p < 0.005$ ，拒绝零假设，说明甲状腺肿类型与祖籍之间有关联性

计算列联系数

$$r = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{723.783}{723.783 + 1436}} = 0.579$$

第三节　直线回归

Page160~166

双变量计量资料： 每个个体有两个变量值

总体： 无限或有限对变量值

样本： 从总体随机抽取的n对变量值

(X_1,Y_1)，(X_2,Y_2)，…，(X_n,Y_n)

目的： 研究X和Y的数量关系

方法： 回归与相关

简单、基本——直线回归、直线相关

“回归”名称的由来

F·Galton（1822——1911年）和他的学生、现代统计学的奠基者之一K·Pearson(1856——1936年)在研究父母身高与其子女身高的遗传问题时，观察了1078对夫妇，以每对夫妇中父亲的身高作为解释变量X，而取他们的一个成年儿子的身高作为被解释变量Y（应变量），将结果在平面直角坐标系上绘成散点图，发现趋势近乎一条直线。计算出的回归直线方程为：

儿子身高（Y，英寸）与父亲身高（X，英寸）存在线性关系：

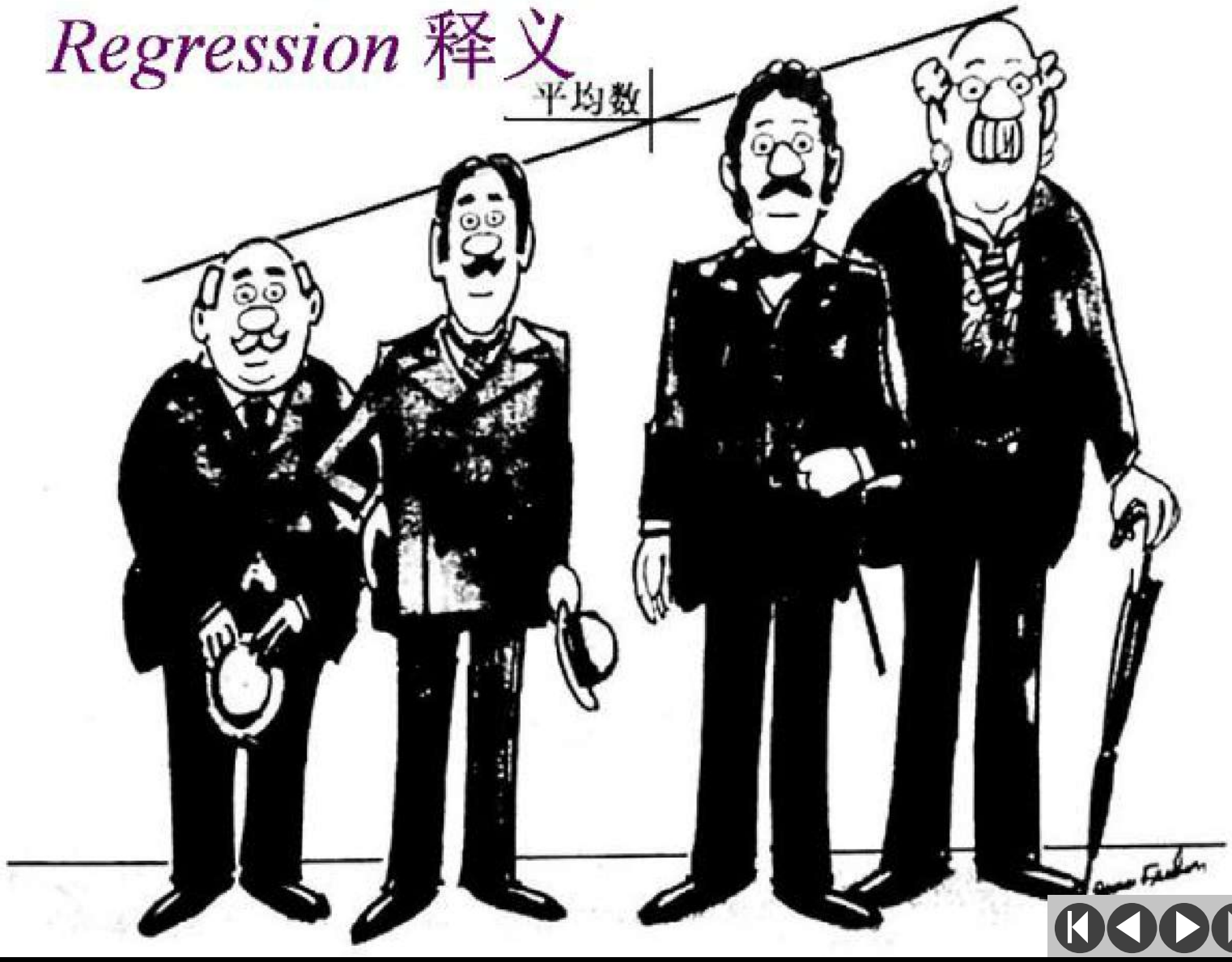
$$\hat{Y} = 33.73 + 0.516X$$

也即高个子父代的子代在成年之后的身高平均来说不是更高，而是稍矮于其父代水平，而矮个子父代的子代的平均身高不是更矮，而是稍高于其父代水平。*Galton*将这种趋向于种族稳定的现象称之“回归”

“回归”已成为表示变量之间某种数量依存关系的统计学术语，相关并且衍生出“回归方程”“回归系数”等统计学概念。如研究糖尿病人血糖与其胰岛素水平的关系，研究儿童年龄与体重的关系等。

Regression 释义

平均数



一、直线回归方程

直线回归的概念

目的： 研究应变量Y对自变量X的数量依存关系。

特点： 统计关系。X值和Y的均数的关系，不同于一般数学上的X和Y的函数关系

〈一〉直线回归模型

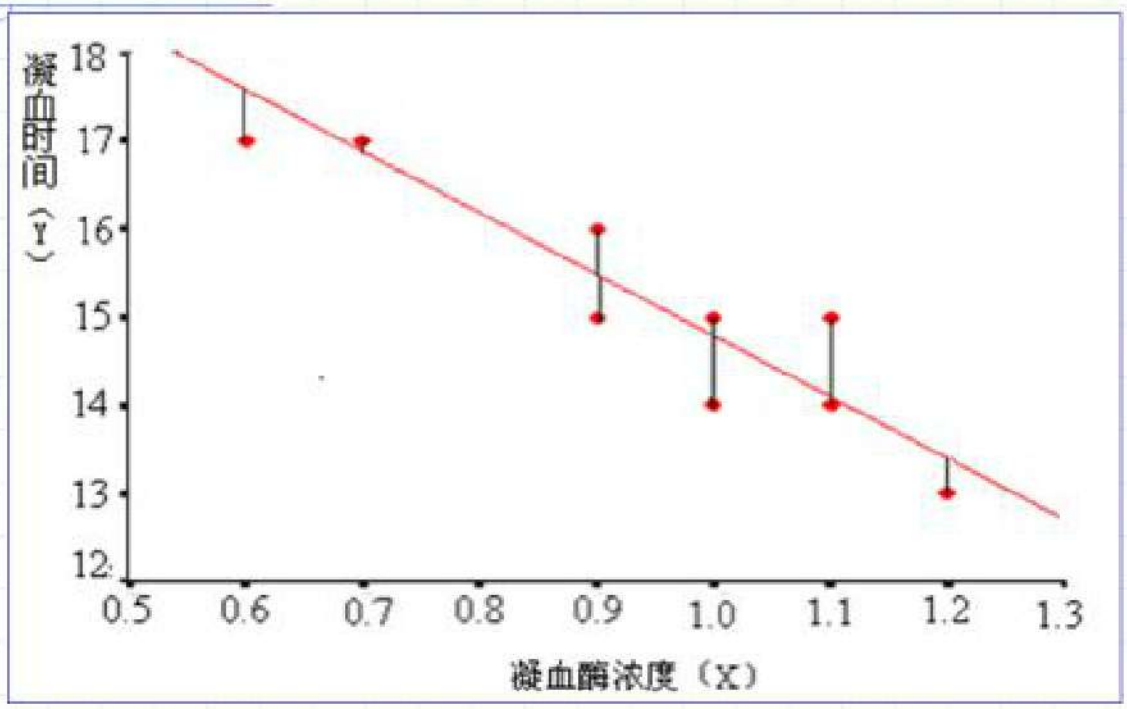
1、 资料数据格式

例号	X（自变量）	Y（应变量）
1	X_1	Y_1
2	X_2	Y_2
.	.	.
n	X_n	Y_n

为了直观地说明直线回归的概念，以15名健康人凝血酶浓度（X）与凝血时间(Y)数据进行回归分析，得到图所示散点图（scatter plot）

No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X	1.1	1.2	1.0	0.9	1.2	1.1	0.9	0.6	1.0	0.9	1.1	0.9	1.1	1.0	0.7
Y	14	13	15	15	13	14	16	17	14	16	15	16	14	15	17

在定量描述健康人凝血酶浓度（*X*）与凝血时间(*Y*)数据的数量上的依存关系时，将凝血酶浓度称为自变量(independent variable)，用 *X* 表示；凝血时间称为应变量(dependent variable)，用 *Y* 表示


凝血时间（Y）
18
17
16
15
14
13
12
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
凝血酶浓度（X）

凝血时间随凝血酶浓度的增加而减低且呈直线趋势，但并非所有点子恰好全都在一直线上，此与两变量间严格的直线函数关系不同，称为**直线回归**（linear regression），其方程叫**直线回归方程**，以区别严格意义的直线方程。是回归分析中最基本、最简单的一种，故又称简单回归。

2、变　量

- Y(应变量，结果变量)：一般是难测（或不可测）的变量，如血压、腹中胎儿体重、肺活量、医疗费用等(要求呈正态分布的随机变量)
- X(自变量，原因变量)：一般是可测（或易测）的变量，如年龄、孕妇宫底高度，体重、年龄等（可是随机变量或人为给定的量）

3、直线回归方程的一般形式为：

$$\hat{Y} = a + bX$$

a 为回归直线在 Y 轴上的截距

- $a > 0$，表示直线与纵轴的交点在原点的上方
- $a < 0$，则交点在原点的下方
- $a = 0$，则回归直线通过原点

*b*为回归系数，即直线的斜率

- *b*<0，直线从左上方走向右下方，*Y*随 *X* 增大而减小；
- *b*>0，直线从左下方走向右上方，*Y*随 *X* 增大而增大；
- *b*=0，表示直线与 *X* 轴平行，*X*与*Y*无直线关系

b 的统计学意义是：*X*每增加(减)一个单位，*Y*平均改变*b*个单位

4、直线回归分析一般可分为三个步骤

- 将原始数据在坐标图上绘散点图
- 根据样本数据求得估计值a、b

即得样本回归方程 $\hat{Y} = a + bX$

- 对回归方程作假设检验，并对方程的拟合效果作出评价

注　　意

- 当两变量的散点图呈直线趋势时，用直线回归分析。
- 当两变量的散点图不呈直线趋势，而呈某种曲线趋势时，用曲线回归分析

例9-1：　可用直线回归分析

回归参数的估计

——最小二乘法原则

- 残差(residual)或剩余值，即实测值Y与假定回归线上的估计值$\hat{Y}$的纵向距离 $Y-\hat{Y}$
- 求解a、b实际上就是“合理地”找到一条能最好地代表数据点分布趋势的直线。

原则：最小二乘法(least sum of squares)，即可保证各实测点(Y)至直线的纵向距离的平方和最小 $\sum(Y-\hat{Y})^2$

回归参数的估计方法

$$b=\frac{l_{XY}}{l_{XX}}=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2}$$

$$a=\bar{Y}-b\bar{X}$$

式中 l_{XY} 为 X 与 Y 的离均差乘积和：

$$l_{XY}=\sum(X-\bar{X})(Y-\bar{Y})=\sum XY-\frac{(\sum X)(\sum Y)}{n}$$

例9-1的计算

$$\sum X \quad \sum Y \quad \sum X^2 \quad \sum Y^2 \quad \sum XY$$

$$l_{XY} = \sum (X - \bar{X})(Y - \bar{Y}) = \sum XY - \frac{(\sum X)(\sum Y)}{n}$$

$$\hat{Y} = 17.9886 + 1.6501X$$

解题步骤

1.由原始数据及散点图观察两变量间是否有直线趋势

2. 计算 X 、 Y 的均数 $\bar{X}$ 、 $\bar{Y}$ ，离均差平方和 l_{XX} 、 l_{YY} 与离均差积和 l_{XY} 。

3、计算有关指标的值

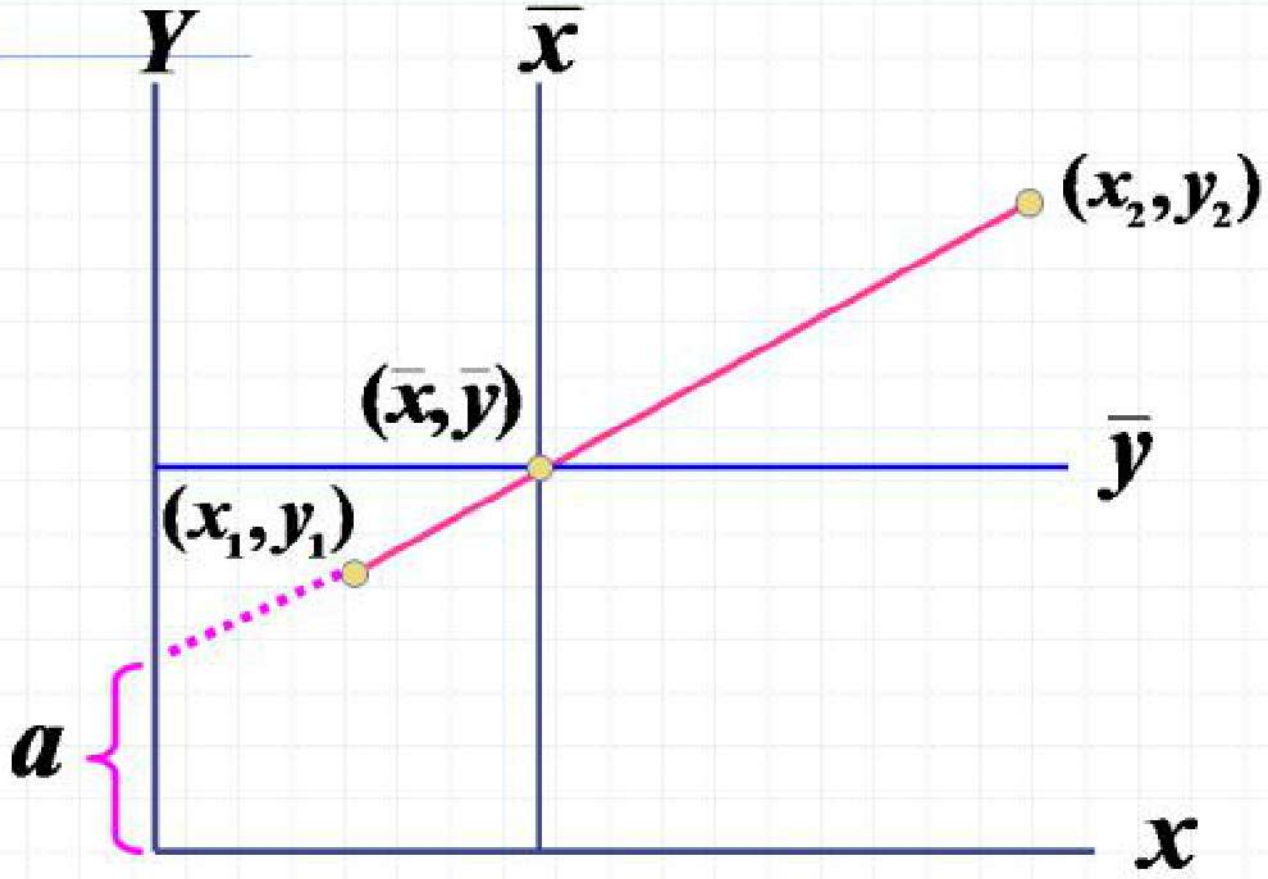
4、计算回归系数和截距

5、列出回归方程

绘制回归直线

此直线必然通过点（$\bar{X}$，$\bar{Y}$）且与纵坐标轴相交于截距a。在自变量实测范围内远端取易于读数的 X 值代入回归方程得到一个点的坐标，连接此点与点（$\bar{X}$，$\bar{Y}$）也可绘出回归直线。

直线回归方程的图示

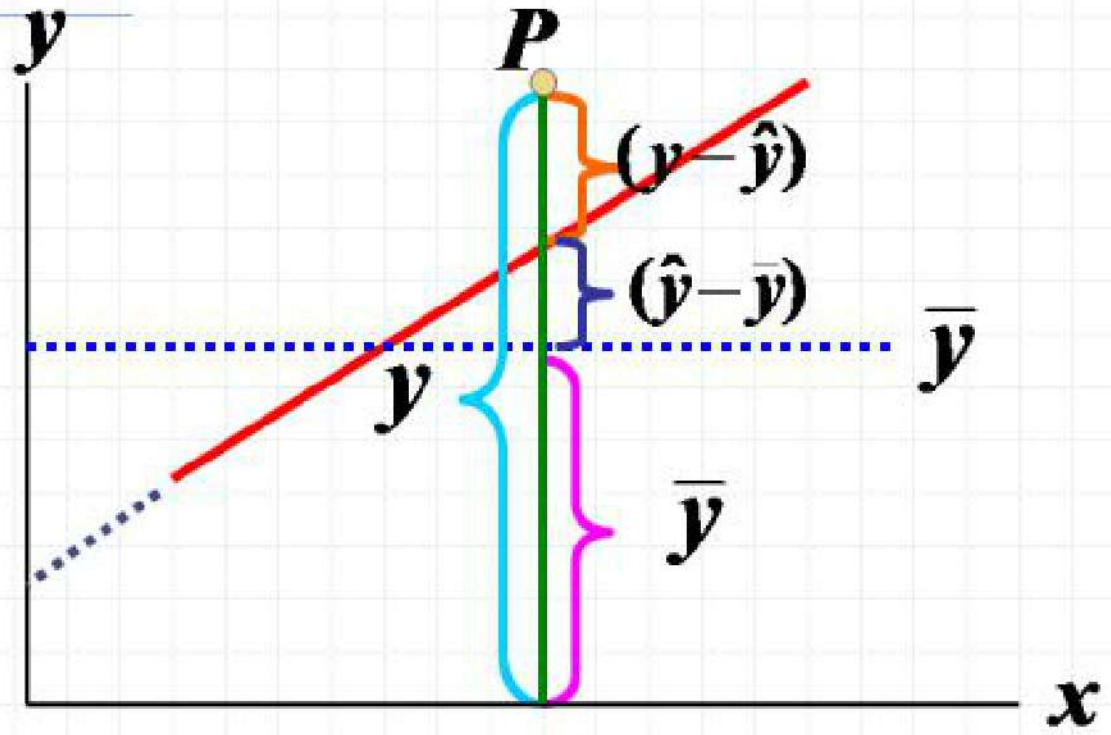


二、回归系数的统计推断

建立样本直线回归方程，只是完成了统计分析中两变量关系的统计描述，研究者还须回答它所来自的总体的直线回归关系是否确实存在，即是否对总体有 $\beta \neq 0$？

- 回归方程的假设检验，其实质是检验样本回归系数b所代表的总体回归系数 β 是否为0。（方法：方差分析，t检验）

1、l_{yy} 的分析



第一段$(y-\hat{y})$：表示P点与回归直线的纵向距离，称为剩余或残差。

以3号大白鼠为例：

$X=76 \qquad \overline{X}=63.7$

$Y=152 \qquad \overline{Y}=123.1$

$\hat{Y}=17.9886+1.6501\times 76=143.3962$

$Y-\hat{Y}=152-143.3962=8.6038$

第二段$(\hat{y}-\bar{y})$：即估计值$\hat{y}$与均数$\bar{y}$之差，与回归系数有关

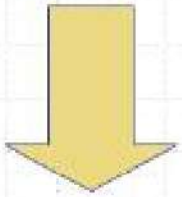
$$\hat{Y}-\bar{Y}=143.3962-123.1=20.2962$$

$$X-\bar{X}=76-63.7=12.3$$

第三段$\bar{y}$：是应变量y的均数。

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

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



$$SS_{总}=SS_{回}+SS_{剩}$$

$SS_{总}$即$\sum(y-\bar{y})^2$，为y的总变异，由很多因素引起，如进食量、消化吸收功能等。

$SS_{回}$即$\sum(\hat{y}-\bar{y})^2$，为回归平方和，它反映y的总变异中由x引起的那部分。$SS_{回}$越大，说明回归效果越好。

$SS_{剩}$即$\sum(y-\hat{y})^2$，为剩余平方和，反映x影响之外的其它一切因素对y的作用。$SS_{剩}$越小，说明直线回归的估计误差越小，回归估计的精度越高。

三者的关系

$$SS_{总} = SS_{回} + SS_{剩} \qquad \upsilon_{总} = \upsilon_{回} + \upsilon_{剩}$$

$$\upsilon_{总} = n-1 \qquad \upsilon_{回} = 1$$

$$\upsilon_{剩} = n-2$$

2、方差分析

基本思想： 将总变异$SS_{总}$分解为$SS_{回}$和$SS_{剩}$然后利用F检验来判断回归程是否成立。

$$F=\frac{SS_{回}/\nu_{回}}{SS_{剩}/\nu_{剩}}=\frac{MS_{回}}{MS_{剩}}$$

$$\upsilon_{回}=1 \qquad \upsilon_{剩}=n-2$$

本　例

$$l_{XX}=614.1 \qquad l_{XY}=1013.3$$

$$l_{YY}=1998.9$$

$$SS_{总}=l_{YY}=1998.9$$

$$SS_{回}=bl_{XY}=b^2l_{XX}=1672.05$$

$$SS_{剩}=SS_{总}-SS_{回}=326.85$$

方差分析结果表

变异来源	SS	υ	MS	F	P
回归	1672.05	1	1672.05	40.92	<0.01
剩余	326.85	8	40.86		
总变异	1998.9	9			

3、t检验

基本思想：是利用样本回归系数 b与总体回归系数 β 进行比较来判断回归方程是否成立。

$$t=\frac{b-0}{s_b} \qquad s_b=\frac{S_{Y.X}}{\sqrt{l_{XX}}}$$

$$S_{Y.X}=\sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}=\sqrt{\frac{SS_{剩}}{n-2}}$$

$S_{Y.X}$

- **剩余标准差**：表示固定了X（即扣除了进食量的影响）后，Y（所增体重）方面仍有变异，是由X以外的其它因素（如消化吸收功能、个体差异等）引起的。

注意：

$\sqrt{F}=t$，即直线回归中对回归系数的t检验与F检验等价，类似于两样本均数比较可以作t检验亦可作方差分析。

4、总体回归系数的估计

由样本数据计算所得的样本回归方程中,回归系数b是总体回归系数β的点估计值，存在着抽样误差，因此需对总体回归系数β作出区间估计。

$$b \pm t_{\alpha/2(n-2)} S_b \qquad S_b = \frac{S_{Y.X}}{\sqrt{l_{XX}}}$$

$$S_{Y.X} = \sqrt{\frac{\sum (y-\hat{y})^2}{n-2}} = \sqrt{\frac{SS_{剩}}{n-2}} = \sqrt{MS_{剩余}}$$

本　　例

$$\hat{Y} = 17.9886 + 1.6501X$$

$$S_b = \frac{S_{Y.X}}{\sqrt{l_{XX}}} = \frac{\sqrt{40.86}}{\sqrt{614.1}} = 0.2579$$

$t_{0.05/2,8} = 2.306$　　代入即得

总体回归系数β95%的可信区间为

(1.0554,2.2448)

三、决定系数

$$R^2 = r^2 = \frac{l_{XY}^2}{l_{XX} l_{YY}} = \frac{SS_{回}}{SS_{总}}$$

本例：$R^2 = \frac{1672.05}{1998.9} = 0.8365$

意义：其体重增加量的**83.65%**可由原始体重解释，另外约**26%**由其他因素引起。

四、直线回归方程的应用

1、描述两变量间的依存关系

2、给定X值时，其个体Y值的预测区间。

3、给定X值时，其Y的总体均数（1-a）可信区间。

个体Y值的预测区间

$$\hat{Y} \pm t_{\alpha/2(n-2)} S_Y$$

$$S_Y = S_{Y.X} \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X - \bar{X})^2}}$$

意义：总体中，X为某特定值X_0时，个体Y的波动范围。

以3号鼠为例:

$$X_0 = 76 \quad \hat{Y} = 143.3962 \quad S_{Y.X} = \sqrt{40.86} = 6.392$$

$$t_{0.05(10-2)} = 2.306$$

$$S_{\hat{Y}} = 6.392\sqrt{1+\frac{1}{10}+\frac{(76-63.7)^2}{614.1}} = 7.4168$$

得95%容许区间为（126.293，160.499）g

意　义

表示所有原始体重为6g的大白鼠中，有95%的大白鼠体重增加量在上述区间内。

Y的总体均数的可信区间

X为某特定值X_0时，Y的总体均数的点估计值为

$\hat{Y} = a + bX_0$，其$1-\alpha$可信区间为

$$\hat{Y} \pm t_{\alpha/2(n-2)} S_{\hat{Y}}$$

$$S_{\hat{Y}} = S_{Y.X} \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X - \bar{X})^2}}$$

本　　例

如果$X_0 = 76$,则$\hat{Y} = 143.3962$

$$S_{\hat{Y}} = S_{Y.X}\sqrt{\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum(X-\bar{X})^2}} = 6.392\sqrt{\frac{1}{10}+\frac{(76-63.7)^2}{614.1}}$$

$$= 0.5885$$

$t_{0.05/2,(10-2)} = 2.306$　　代入即得

95%的可信区间为(142.039,144.753)g

意　义

表示所有原始体重为6g的大白鼠，其平均所增体重（即$\mu_{\bar{Y}}$）有95%的可能性在上述区间内。

应用直线相关、回归的注意事项

1、进行相关与回归分析要有实际意义
2、先绘制散点图，根据散点图的趋势再作分析
3、相关分要求双变量正态分布资料，而回归分析只要求Y呈正态分布
4、须作假设检验，或求其置信区间，且$t_r=t_b$
5、P的大小不代表事物或现象间相关的密切程度
6、不能把两变量的相关关系误认为因果关系
7、直线回归方程的应用只能“内插”，不能“外延”

直线相关与回归的区别与联系

两者的联系

1、对于既可作相关又可作回归分析的同一组数据，计算出的r与b正负号一致。

2、相关系数与回归系数的假设检验等价，即对于同一样本，$t_r=t_b$。

3、同一组数据的相关系数和回归系数可以相互换算，$r=b\times S_X/S_Y$

直线相关与回归的区别与联系

两者的区别

1、资料要求上：相关要求服从双变量正态分布，这种资料进行回归分析称为Ⅱ型回归；回归要求在Y服从正态分布，X是可以精确测量和严格控制的变量，称为Ⅰ型回归。

2、应用上：说明两变量间相互关系用相关，此时两变量的关系是平等的；而说明两变量间依存变化的数量关系用回归，用以说明Y如何依赖于X而变化。

直线相关与回归的区别与联系

两者的区别

3、意义上：r说明具有直线关系的两变量间相互关系的方向与密切程度；b表示X每变化一个单位所导致Y的平均变化量。

4、计算上：$r = l_{XY} / \sqrt{l_{XX} l_{YY}}$　　$b = l_{XY} / l_{XX}$

5、取值范围：$-1 \le r \le 1$　$-\infty < b < \infty$

6、单位：r没有单位，b有单位。