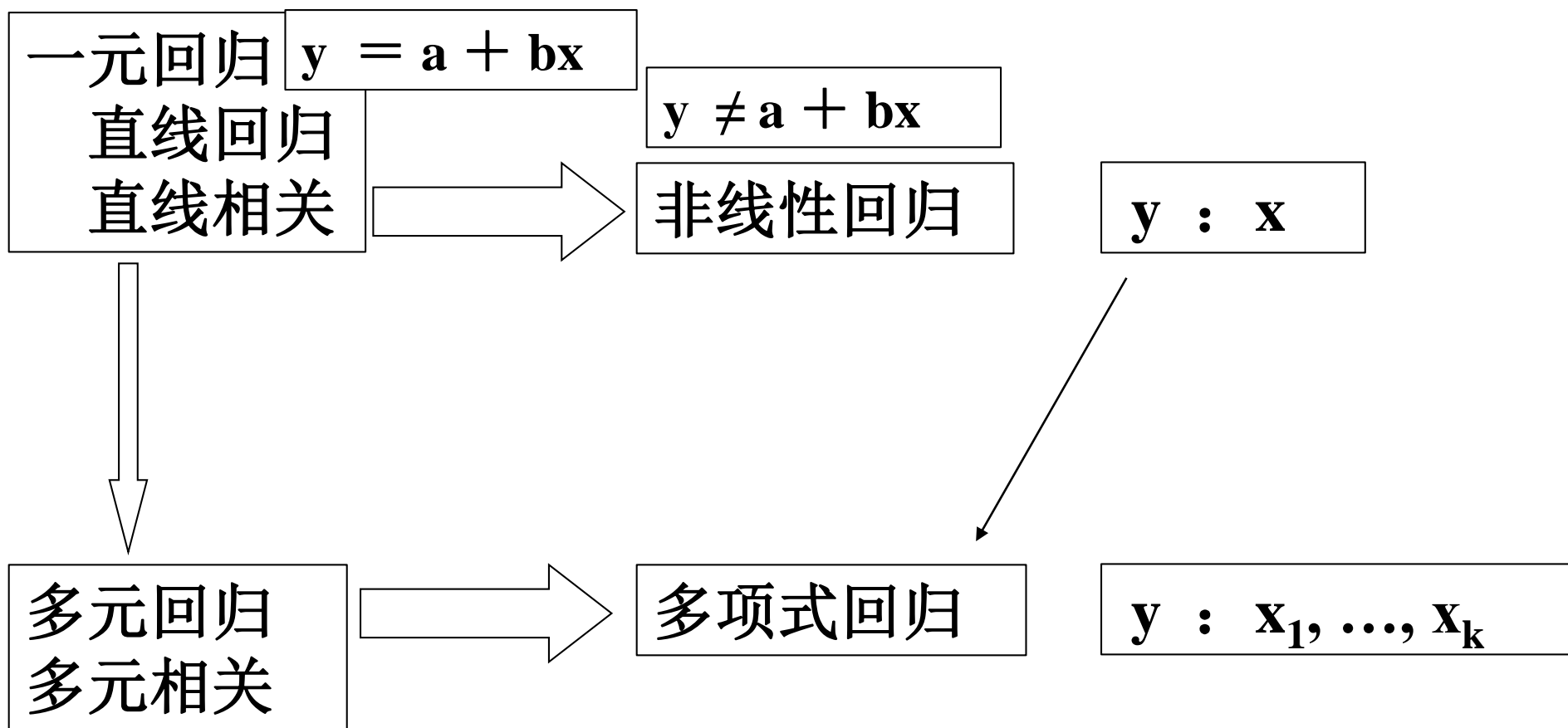


回归分析—相关分析

内容关联结构



第八章

一元回归与相关分析

回归----相关，研究变量之间的关系

为确认这种关系，用到假设检验---- 此仅为研究“关系”服务

变量与变量之间的关系有两类：

函数关系：确定性关系，班上人数=男生数+女生数

面积=长×宽

体积=长×宽×高



相关关系：非确定性关系，也称为统计相关关系

树：高度与茎粗

人：血压与年龄

图 7-5 利用树围预测树高

8—1 直线回归

- 一、直线回归方程的建立
- 二、直线回归的数学模型与基本假定
- 三、直线回归方程的假设检验
- 四、直线回归的区间估计
- 五、两条乃至多条回归直线的比较
- 六、应用实例

8—2 直线相关

8—3 可直线化的非线性回归分析



8-1 直线回归

一、直线回归方程的建立

1、直线回归方程的一般形式：

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

式中，

\hat{y} 是与自变量 x 相对应的依变量 y 的点估计值

a 是当 $x=0$ 时的 \hat{y} 值，即直线在 y 轴上的截距

b 是回归直线的斜率，叫回归系数

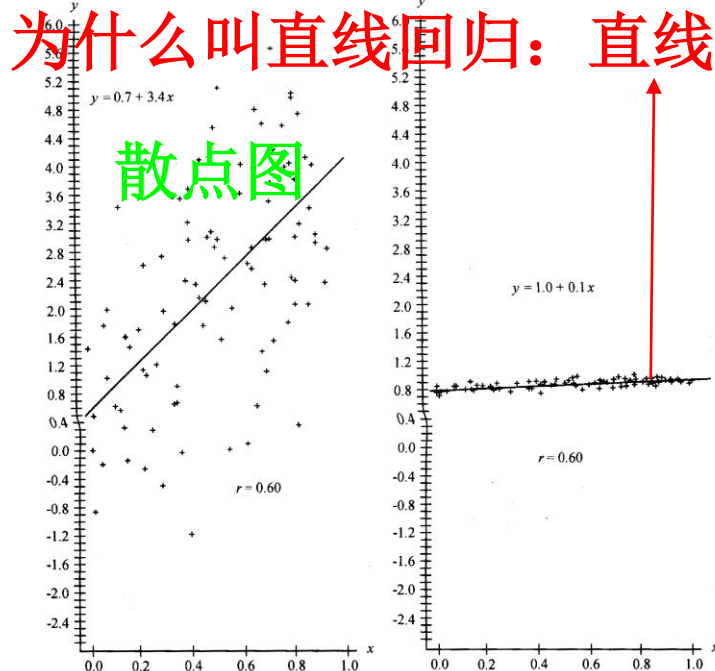
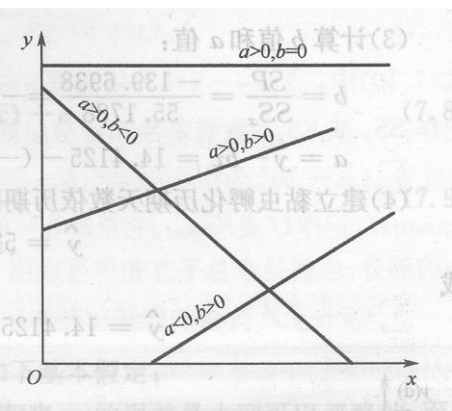


图 10.7 相关系数相同、而回归直线不同的两个数据集的散点图。

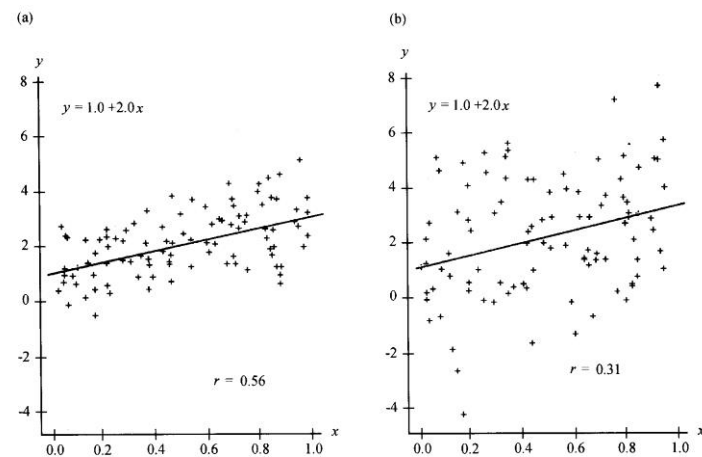


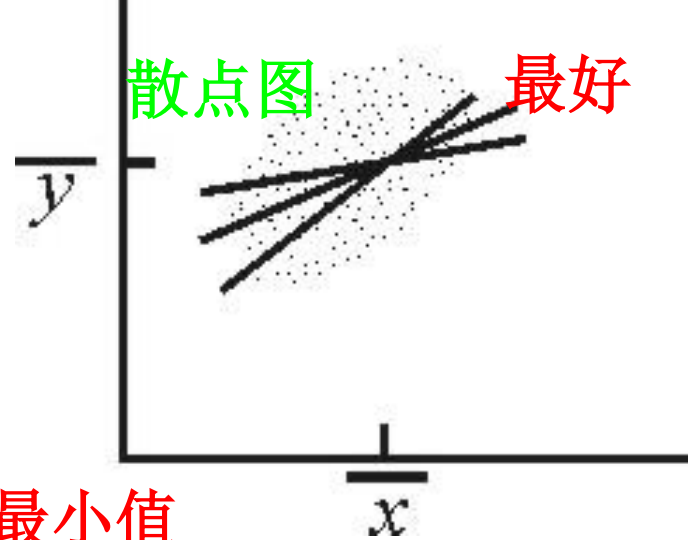
图 10.8 相关系数不同、而回归直线相同的两个数据集的散点图。

2、直线回归方程的最小二乘估计

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

离回归平方和最小：

$$Q = \sum (y - \hat{y})^2 = \sum (y - a - bx)^2 = \text{最小值}$$



将 Q 对 a 、 b 求偏导数，并令其等于 0，得：

正规方程组：P124

$$an + b\sum x = \sum y$$

$$a\sum x + b\sum x^2 = \sum xy$$

解方程组，得

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

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

离均差乘积和 = 0 ?

美国统计学家斯蒂格勒认为：

最小二乘法之于数理统计学，有如微积分之于数学。

现今许多出版物把最小二乘法的发明归于德国数学家高斯，但第一个公开以书面形式发表的是法国数学家勒让德(1805)。

很久很久以前--老皇历：

最小二乘法提出之前，采用合并后解方程组：
待估参数为 k 个，就合并出 k 个方程

当待估参数 k 较大时，计算任务很大：

地图：1858年英国绘制本国地图， $k=920$ 、 $n=1554$ ，用两组人员独立计算，花了两年半才完成。

炼钢：1958年某研究所计算炼钢课题， $k=14$ ，采用电动计算机30多人夜以继日花了一个多月才完成。

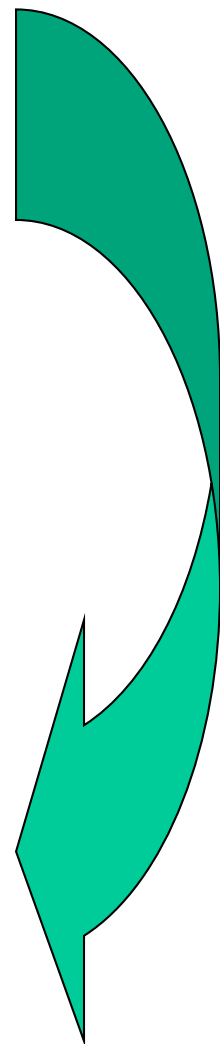
3、直线回归方程的三个基本性质：教材 P125

性质1、 $Q = \sum (y - \hat{y})^2 = \text{最小值}$

性质2、 $\sum (y - \hat{y}) = 0$

性质3、回归直线通过中心点 (\bar{x}, \bar{y})

平均数的两个基本性质

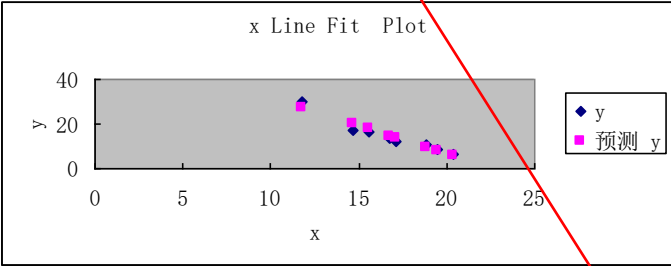


4、实例：P125

昆虫孵化历期天数与平均温度

$y=a+bx$
 $=57.04-2.53x$

x 温度	y 天数
11.8	30.1
14.7	17.3
15.6	16.7
16.8	13.6
17.1	11.9
18.8	10.7
19.5	8.3
20.4	6.7



注：1.98376²=3.93531

回归统计	
R	0.9682
R Square	0.93741
Adjusted	0.92698
标准误差	1.98376
观测值	8

方差分析				
	df	SS	MS	F
回归分析	1	353.657	353.657	89.8675
残差	6	23.6119	3.93531	
总计	7	377.269		
	Coefficient	标准误差	t Stat	P-value
Intercept	57.0393	4.55094	12.5335	1.6E-05
x	-2.5317	0.26706	-9.4798	7.8E-05

方差分析的运用

- 多个样本平均数的比较
- 多个因素间互作的分析 ---- 交互作用

其它用途:

- 回归方程的假设检验
- 多元相关系数的检验

方差分析				
	df	SS	MS	F
回归分析	1	353.657	353.657	89.8675
残差	6	23.6119	3.93531	
总计	7	377.269		
	Coefficient	标准误差	t Stat	P-value
Intercept	57.0393	4.55094	12.5335	1.6E-05
x	-2.5317	0.26706	-9.4798	7.8E-05

二、直线回归的数学模型与基本假定

1、数学模型

在直线回归中， y 总体的每一个观察值可分解为三部分：

- y 的总体平均数 μ_y 、
- x 的变异引起 y 的变异 $\beta(x-\mu_x)$ 、
- ε 随机误差

因此，**直线回归的数学模型**

变 $y = \mu_y + \beta(x - \mu_x) + \varepsilon$

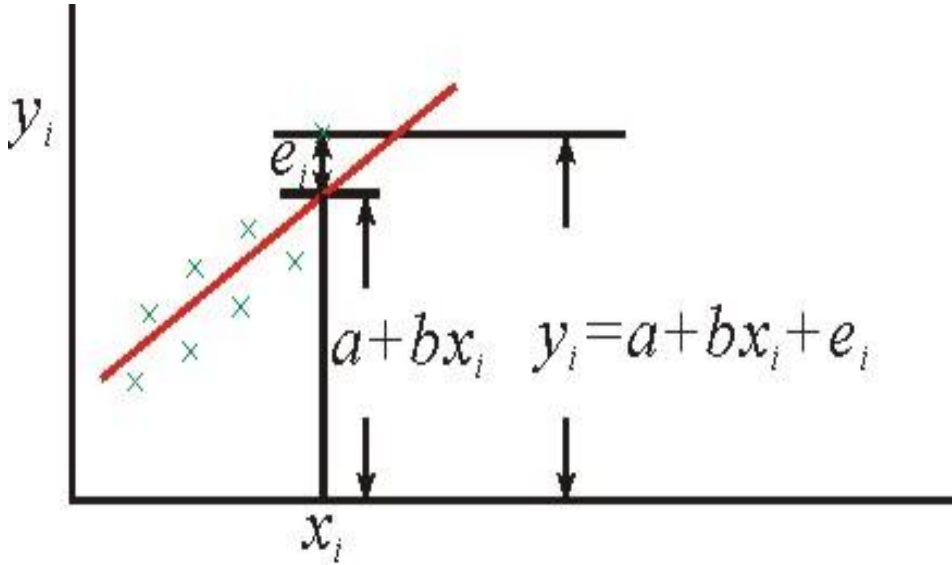
$y = \alpha + \beta x + \varepsilon$

$\alpha = \mu_y - \beta \mu_x$

对于样本资料，数学模型

变 $y = \bar{y} + b(x - \bar{x}) + e$

$y = a + bx + e$



2、基本假定：P127

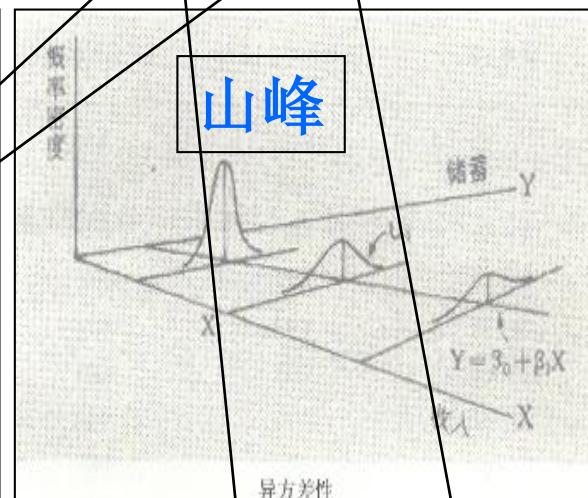
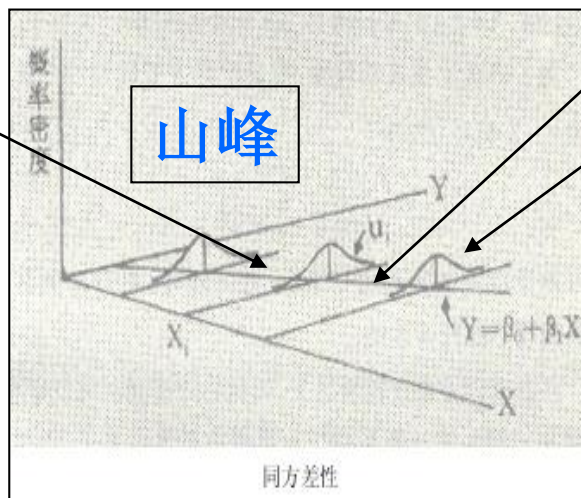
(1)、 x 是没有误差的固定变量(有时称为可控变量), 或者说 x 的误差小到可以忽略, 而 y 是随机变量, 且具有随机误差;

(2)、 x 的任一值都对应一个 y 总体且作正态分布 $N(\mu_{y/x}, \sigma_{y/x}^2)$, 平均数($\mu_{y/x}$), 方差 ($\sigma_{y/x}^2$) 不因 x 的变化而改变;

平均数: $\mu_{y/x} = \alpha + \beta x$

方差: $\sigma_{y/x}^2$

家庭收入 X 与家庭消费 Y



(3)、随机误差 ε 是相互独立的且作正态分布, 具有 $N(0, \sigma_\varepsilon^2)$

注: 由 $y = \alpha + \beta x + \varepsilon$

$$V(y) = V(\alpha + \beta x + \varepsilon) = V(\alpha) + V(\beta x) + V(\varepsilon) = V(\varepsilon)$$

$$\text{得 } \sigma_{y/x}^2 \longleftarrow \sigma_\varepsilon^2$$

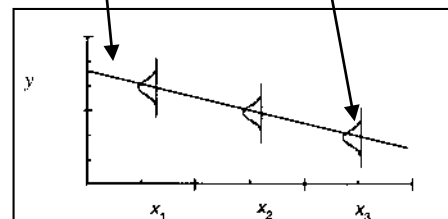


图 8.7 线性回归中的正态性和方差假设

三、直线回归方程的假设检验

任何一组成对数据都可建立一个回归方程，关键在于建立的方程是否有意义——回归是否达到显著水平。

检验内容：

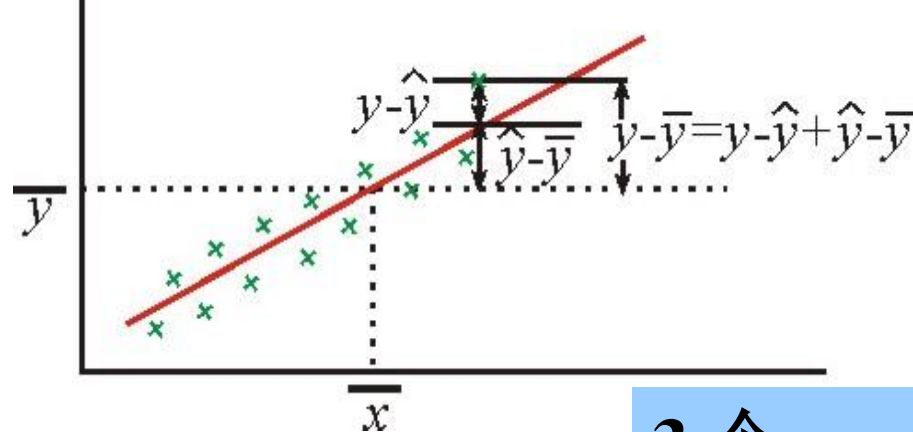
- 1、直线回归的变异来源：为检验做准备
- 2、F测验——回归关系的检验
- 3、t检验——回归系数的检验
- 4、t检验——回归截距的检验
- 5、补充说明

前面提到的方差分析

假设检验：帮助确认是否真有关系

1、直线回归的变异来源

P128



平方和分解:

$$\begin{aligned} SS_y &= \sum (y - \bar{y})^2 = \sum (y - \bar{y} + \hat{y} - \hat{y})^2 \\ &= \sum (\hat{y} - \bar{y})^2 + \sum (y - \hat{y})^2 \end{aligned}$$

$$\text{回归平方和 } U = \sum (\hat{y} - \bar{y})^2$$

$$= b \sum (x - \bar{x})(y - \bar{y}) = b SP$$

$$\text{离回归平方和 } Q = \sum (y - \hat{y})^2 = SS_y - U$$

$$\text{自由度分解: } (n-1) = 1 + (n-2)$$

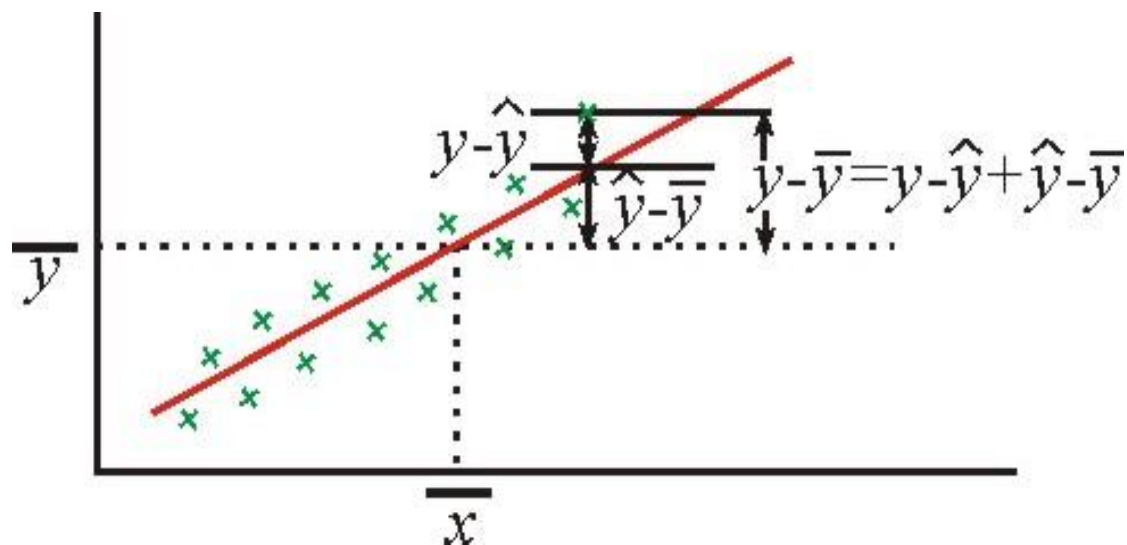
$$\text{两个方差: 回归方差} = U / 1$$

$$\text{离回归方差} = s^2_{y/x} = Q / (n-2)$$

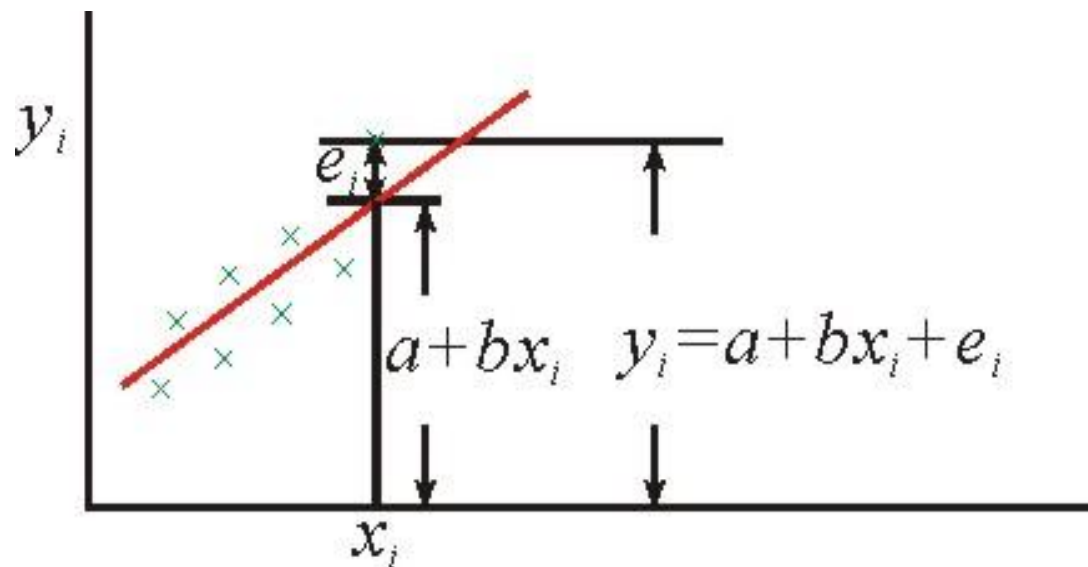
$$\text{离回归标准差, 或称回归估计标准差: } s_{y/x} = \sqrt{[Q / (n-2)]}$$

变异分解的作用: 回归关系检验、参数检验、区间估计

变异剖分



线性模型



2、F测验——回归关系的检验

方差分析:

H_0 : 两变量间无线性关系

H_A : 两变量间有线性关系

$F = \text{回归方差} / \text{离回归方差}$

$$= (U/1) / [Q / (n-2)] \sim F(1, n-2)$$

方差分析				
	df	SS	MS	F
回归分析	1	353.657	353.657	89.8675
残差	6	23.6119	3.93531	
总计	7	377.269		
	Coefficient	标准误差	t Stat	P-value
Intercept	57.0393	4.55094	12.5335	1.6E-05
x	-2.5317	0.26706	-9.4798	7.8E-05

3、t检验——回归系数的检验，此特例0也是检验直线回归关系

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

$$\text{回归系数的标准误 } s_b = s_{y/x} / \sqrt{[SS_x]}$$

$$t = (b - \beta) / s_b = b / s_b \sim t(n - 2)$$

4、t检验——回归截距的检验

$$H_0: \alpha = 0; \quad H_A: \alpha \neq 0$$

$$\text{回归截距的标准误 } s_a = s_{y/x} \times \sqrt{[1/n + \bar{x}^2 / SS_x]}$$

$$t = (a - \alpha) / s_a = a / s_a \sim t(n - 2)$$

	Coefficient	标准误差	t Stat	P-value
Intercept	57.0393	4.55094	12.5335	1.6E-05
x	-2.5317	0.26706	-9.4798	7.8E-05

5、补充说明：回归系数与回归截距的检验的一般情形 ($\alpha_0, \beta_0 \neq 0$)

$$H_0: \beta = \beta_0; \quad H_A: \beta \neq \beta_0$$

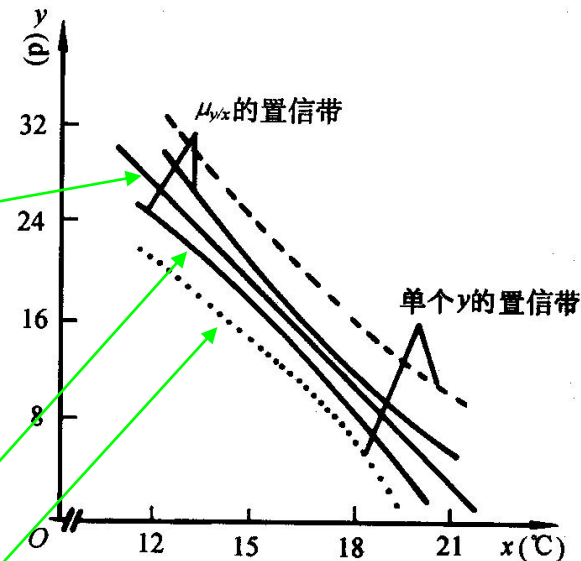
$$t = (b - \beta_0) / s_b \sim t(n - 2)$$

$$H_0: \alpha = \alpha_0; \quad H_A: \alpha \neq \alpha_0$$

$$t = (a - \alpha_0) / s_a \sim t(n - 2)$$

四、直线回归的区间估计

- 1、回归截距的置信区间
- 2、回归系数的置信区间
- 3、 $\mu_{y/x}$ 的置信区间 ----- 直线
- 4、单个 y 的预测区间 ----- 散点



回归截距的标准误: $s_a = s_{y/x} \times \sqrt{[1/n + x^{-2}/SSx]}$ (7.19)

回归系数的标准误: $s_b = s_{y/x} / \sqrt{[SSx]}$ (7.16)

对于给定的 x , 预测总体的平均数 $\mu_{y/x}$ 时的标准误: $s_{y^{\wedge}}$ (7.23)

对于给定的 x , 预测单个 y 观测值的标准误: s_y (7.25)

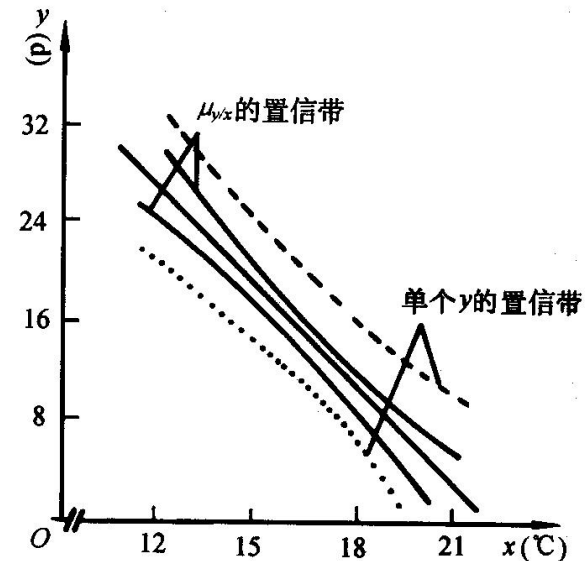
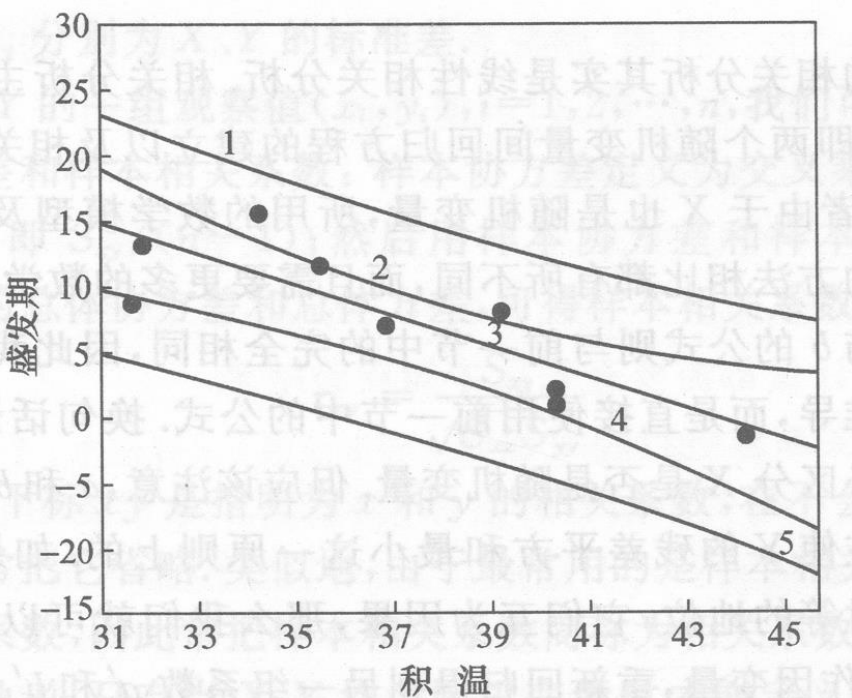
利用 $t_{(n-2)\alpha}$ 的值, 得到置信区间或预测区间:

$$L = a \pm t_{\alpha} s_a \quad (7.21)$$

$$L = b \pm t_{\alpha} s_b \quad (7.22)$$

$$L = y^{\wedge} \pm t_{\alpha} s_{y^{\wedge}} \quad (7.24)$$

$$L = y^{\wedge} \pm t_{\alpha} s_y \quad (7.26)$$



a、b 的置信区间

	Coefficient	标准误差	t Stat	P-value	lower 95%	upper 95%
Intercept	18.0421	4.73081	3.81375	0.01888	4.90727	31.177
G基因个数	20.1519	1.22149	16.4978	7.9E-05	16.7604	23.5433

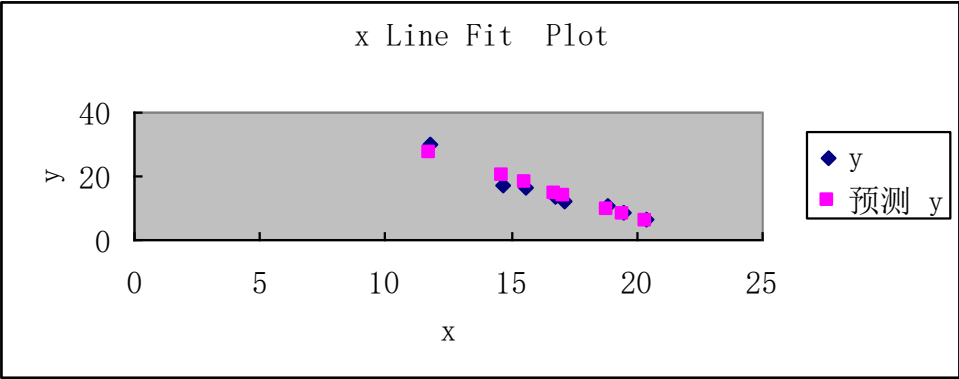
六、实例

1、P125

昆虫孵化历期天数与平均温度

$y=a+bx=57.04-2.53x$

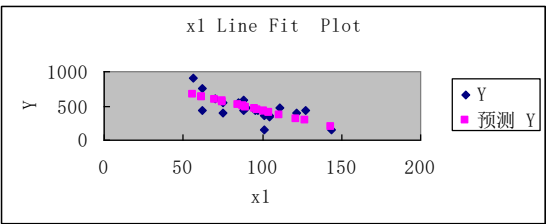
x 温度	y 天数
11.8	30.1
14.7	17.3
15.6	16.7
16.8	13.6
17.1	11.9
18.8	10.7
19.5	8.3
20.4	6.7



回归统计	
R	0.9682
R Square	0.93741
Adjusted	0.92698
标准误差	1.98376
观测值	8

方差分析				
	df	SS	MS	F
回归分析	1	353.657	353.657	89.8675
残差	6	23.6119	3.93531	
总计	7	377.269		
Coefficient		标准误差	t Stat	P-value
Intercept	57.0393	4.55094	12.5335	1.6E-05
x	-2.5317	0.26706	-9.4798	7.8E-05

2、实例 2





藕产量 y 与莲籽产量 x

$$y = a + bx = 966.414 - 5.4342x$$

1斤莲子5斤半藕

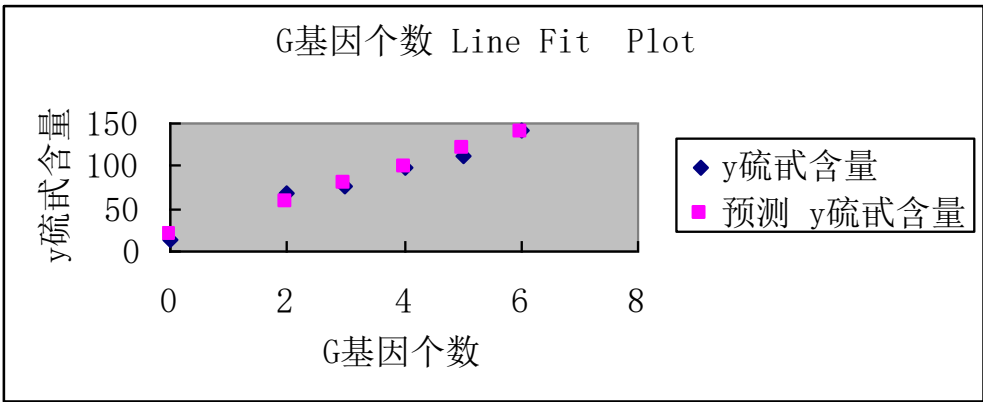
藕产量	壳莲产量	现蕾数	莲蓬数
Y	x1	x2	x3
600	70.2	8	4
157	101.1	6	5
152	143.6	10	6
355	104.7	9	7
443	127	8	6
764	62.3	6	4
401	122	7	5
390	75.5	8	4
437	97	7	5
553	75.3	7	3
430	95.3	8	6
432	88.3	9	5
334	104.8	8	6
476	110.8	11	7



回归统计						
Multiple R	0.70981					
R Square	0.50384					
Adjusted R Square	0.47627					
标准误差	125.78					
观测值	20					
方差分析						
	df	SS	MS	F		
回归分析	1	289173	289173	18.2783		
残差	18	284771	15820.6			
总计	19	573945				
Coefficients		标准误差	t Stat	P-value	lower 95%	upper 95%
Intercept	966.414	121.596	7.94773	2.7E-07	710.95	1221.88
x1	-5.4342	1.27107	-4.2753	0.00046	-8.1047	-2.7638

3、油菜品质（硫甙含量）的遗传分析：
基因个数 G （没有误差）决定硫甙含量
作物学报（1989年）
 $y=a+b G=18.04+20.15 G$

y硫甙量	G基因数
14.62	0
67	2
75.95	3
98.35	4
112.96	5
142.41	6



遗传假定：
① 3个座位6个基因
② 效应：独立、等效、累加

回归统计	
R	0.99273
R Square	0.98552
Adjusted	0.9819
标准误差	5.90036
观测值	6

方差分析						
	df	SS	MS	F	ificance	F
回归分析	1	9475.6	9475.6	272.176	7.9E-05	
残差	4	139.257	34.8143			
总计	5	9614.86				
Coefficient		标准误差	t Stat	P-value	lower 95%	pper 95%
Intercept	18.0421	4.73081	3.81375	0.01888	4.90727	31.177
G基因个数	20.1519	1.22149	16.4978	7.9E-05	16.7604	23.5433

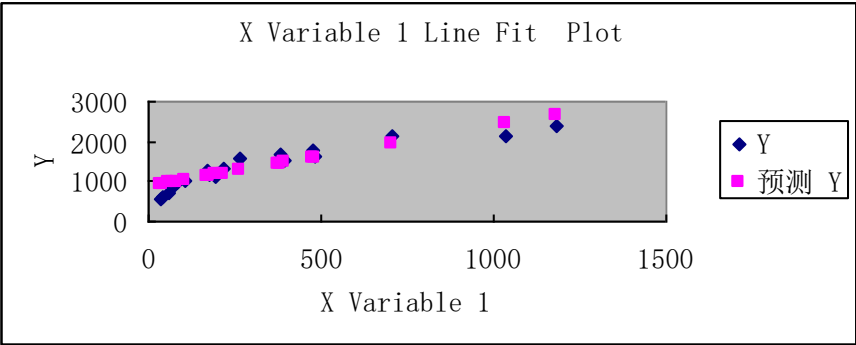
4、

GR.Iversen		
M.Gergen	统计学	P299
吴喜之译		

$y=a+bx=862.46+1.5x$

x=结婚人数
y=离婚人数

年份	结婚人数	离婚人数
1890	570	33
1895	620	40
1900	709	56
1905	842	68
1910	948	83
1915	1008	104
1920	1274	170
1925	1188	175
1930	1127	196
1935	1327	218
1940	1596	264
1945	1613	485
1950	1667	385
1955	1531	377
1960	1523	393
1965	1800	479
1970	2159	708
1975	2153	1036
1980	2413	1182



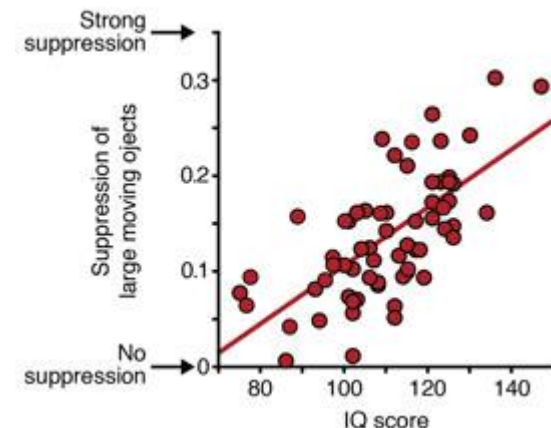
回归统计	
R	0.92512
R Square	0.85584
Adjusted	0.84736
标准误差	207.684
观测值	19

方差分析					
	df	SS	MS	F	nificance
回归分析	1	4353263	4353263	100.927	1.4E-08
残差	17	733255	43132.6		
总计	18	5086518			
Coefficient		标准误差	t Stat	P-value	lower 95%
Intercept	862.463	69.5885	12.3938	6.1E-10	715.644
X Variabl	1.5005	0.14936	10.0463	1.4E-08	1.18538

预测智商新方法：视觉性测试

2013-05-26 来源：生物360

- 美国罗切斯特大学的科学家找到了一种方法测量智商：简单测量大脑下意识过滤运动图像噪声的能力。论文刊登在《当代生物学》。
- 科学家发现，**快速运动过滤**和 **IQ** 之间存在出乎意料的联系，因此利用可视化测试可以得到类似经典 **IQ** 测试的结果。
- **IQ** 与运动抑制之间的关系涉及到智商背后的基本认知过程，大脑受到了海量感官信息的轰炸，其**信息处理效率**不仅仅基于神经网络如何快速处理信号，还基于如何高效过滤掉无意义的信息。
- 研究人员认为，**可视化 IQ 测试**可以消除**标准 IQ 测试**中被广泛批评的**文化偏见**。



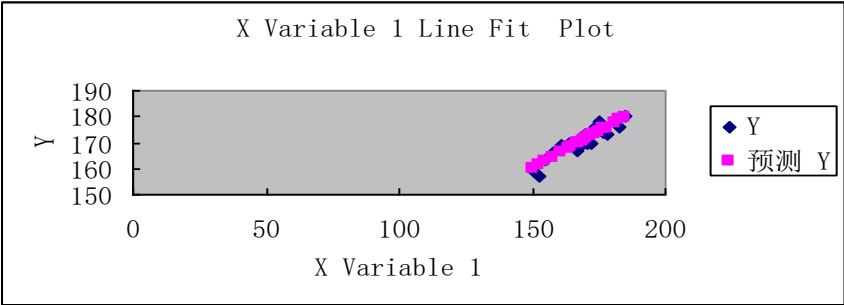
5、回归的来历：儿子 y 与老子 x 的身高

regression to the mean

$y=a+bx=74.16+0.57x$



父高 x	子高 y
150	159
153	157
155	163
158	166
161	169
164	170
165	169
167	167
168	169
169	170
170	173
171	170
172	170
174	176
175	178
177	174
178	173
181	178
183	176
185	180



方差分析						
	df	SS	MS	F	ificance F	F
回归分析	1	603.662	603.662	114.513	3.1E-09	
残差	18	94.888	5.27156			
总计	19	698.55				
Coefficient		标准误差	t Stat	P-value	Lower 95%	Upper 95%
Intercept	74.1652	9.00298	8.23786	1.6E-07	55.2507	93.0798
X Variab	0.56981	0.05325	10.7011	3.1E-09	0.45794	0.68169

回归统计	
Mult R	0.9296
R Square	0.86416
Adjusted	0.85662
标准误差	2.29599
观测值	20

8—2 直线相关

考试成绩 x
再考试成绩 y

	生物统计	英语上	数学上	化学上	英语下	数学下	化学下
生物统计	1						
英语上	0.45112	1					
数学上	0.62444	0.1735	1				
化学上	0.6609	0.46405	0.61595	1			
英语下	0.32721	0.33596	0.17874	0.19134	1		
数学下	0.60633	0.16906	0.54766	0.59749	0.31017	1	
化学下	0.64019	0.35079	0.65372	0.72746	0.39932	0.56986	1

8-2 直线相关

- 一、相关系数与决定系数
- 二、相关系数的假设检验
- 三、相关系数的区间估计
- 四、慎用相关——注意相关的实际意义

两个变量之间的关系：
因果关系：身高与年龄
 血压与年龄
平行关系：身高与体重
 身高与腰围

具有因果关系的变量可用回归分析；

具有平行关系的变量虽然也可用回归分析，但能作出两条回归直线，且两条回归直线并不重叠。

一、相关系数与决定系数

双变量总体的相关系数： P135公式(7.27)

$$\rho = \frac{\sum (x - \mu_x)(y - \mu_y)}{\sqrt{[\sum (x - \mu_x)^2 \sum (y - \mu_y)^2]}}$$

样本相关系数：

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{[\sum (x - \bar{x})^2 \sum (y - \bar{y})^2]}}$$
$$= SP / \sqrt{[SS_x \times SS_y]}$$

r 的取值范围： $[-1, 1]$

相关关系的图示

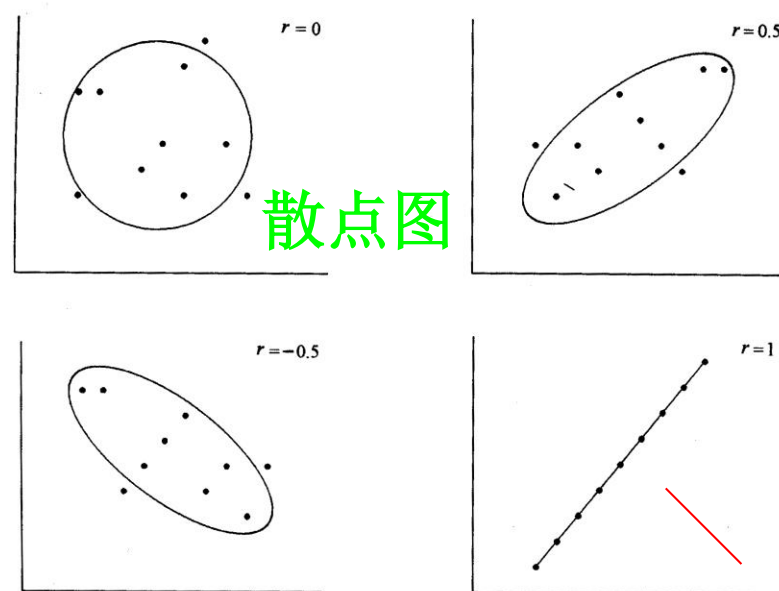
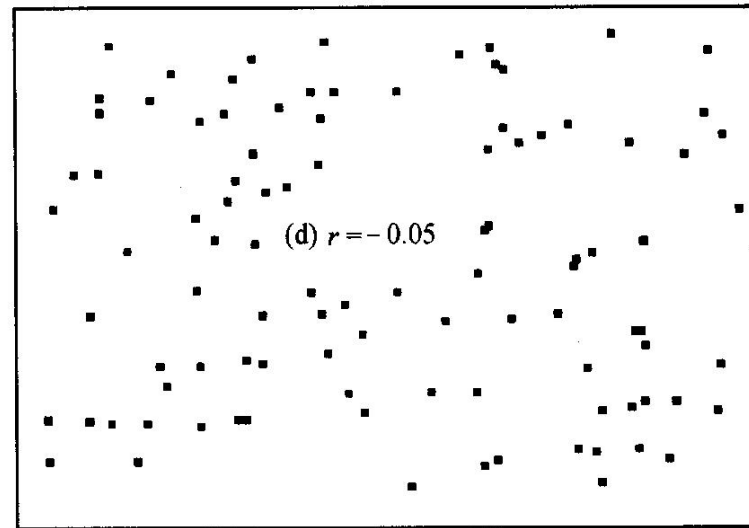
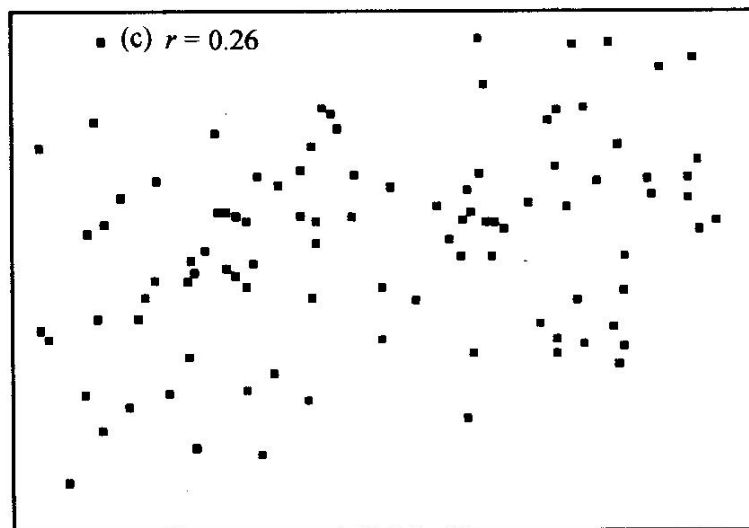
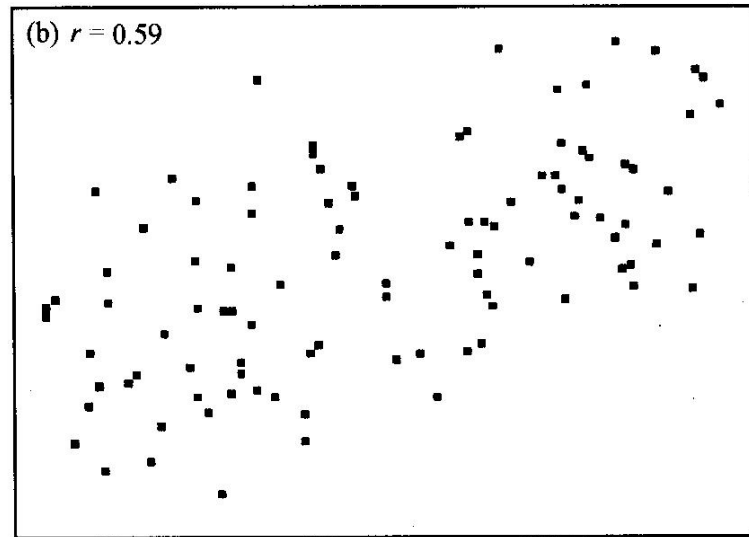
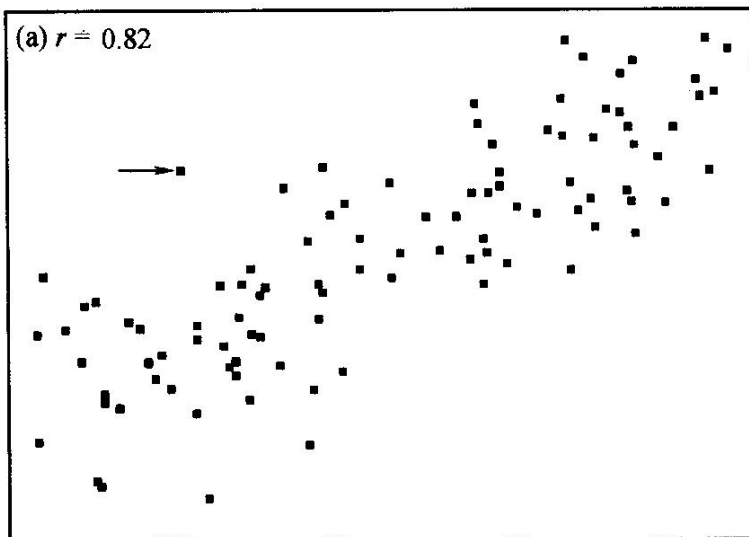


图 10-15 不同 r 值的散点图

图示：

相关关系从强到弱



样本相关系数:

r 的取值范围: $[-1, 1]$

决定系数:

r^2 : 变量 x 引起 y 变异的回归平方和占 y 变异总平方和的比率

r^2 的取值范围: $[0, 1]$

调整的 (修正的) 决定系数:

$$\text{adj-} R^2 = 1 - (1 - R^2) (n-1) / (n-m-1)$$



消除自变量个数影响

见实例1: 昆虫孵化数据

回归统计	
R	0.9682
R Square	0.93741
Adjusted	0.92698
标准误差	1.98376
观测值	8

例1

离均差乘积

						
x	$(x - \bar{x})$	$(x - \bar{x})^2$	y	$(y - \bar{y})$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
1	-2	4	2	-4	16	8
2	-1	1	4	-2	4	2
3	0	0	6	0	0	0
4	1	1	8	2	4	2
5	2	4	10	4	16	8
Σ		10			40	20

$$r = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{\sqrt{[\Sigma (x - \bar{x})^2 \Sigma (y - \bar{y})^2]}}$$

$$= 20 / \sqrt{(10 * 40)} = 20 / 20 = 1$$

散点都落在直线上
 $y = 0 + 2x$

二、相关系数的假设检验——特例0是对直线回归关系的检验 任何一组成对数据，都可算出相关系数，关键是有无意义。

假设检验

假设检验：帮助确认是否真有关系

$$H_0: \rho = 0$$

$$H_A: \rho \neq 0$$

相关系数的标准误： $s_r = \sqrt{[(1 - r^2)/(n - 2)]}$

t 值： $t = (r - \rho) / s_r = r / s_r \sim t(n - 2)$

此检验也是前述“直线回归关系的显著性检验”方法之三。

$$t = b / s_b = r / s_r$$

前面例1：当 $r=1$ 时， $s_r = 0$

则 $t = 1 / 0 = \infty$

结论：相关系数与 0 有极显著的差异。

P136 例7.9:

$$r = -0.9682$$

$$s_r = \sqrt{[(1 - r^2)/(8 - 2)]} = 0.1021$$

$$t = -9.48$$

$$t_{6, 0.01} = 3.707$$

新郎年龄—新娘年龄

GR.Iversen	
M.Gergen	统计学
吴喜之译	P299

$$r=0.91702$$

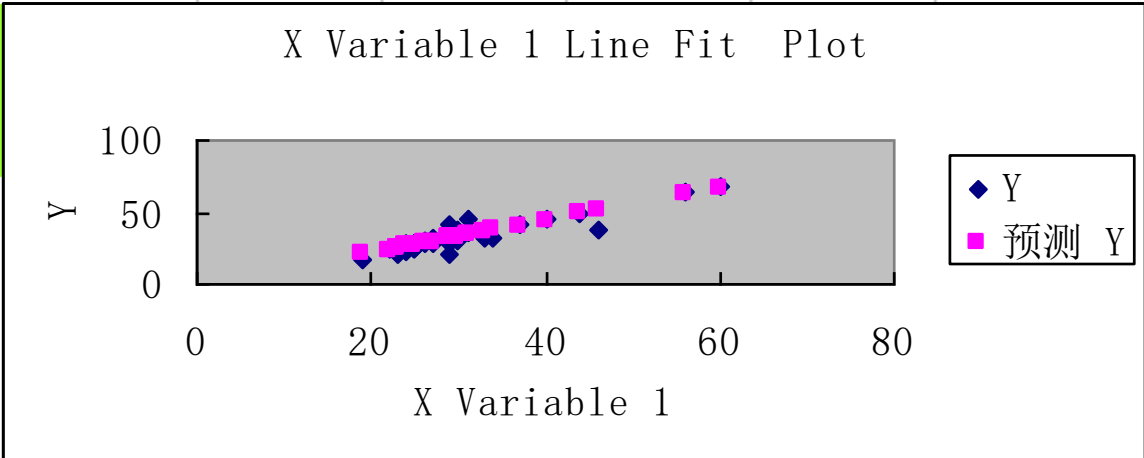
$$s_r=\sqrt{[(1-r^2)/(37-2)]}=?$$

$$t=?$$

回归统计			
Multiple R	0.91703		
R Square	0.84095		
Adjusted R Square	0.8364		
标准误差	4.4941		
观测值	37		

1	新郎年龄	新娘年龄
2	17	19
3	21	23
4	21	29
5	22	23
6	23	24
7	24	22
8	24	22
9	24	23
10	24	24
11	24	25
12	25	25
13	26	23
14	28	25
15	28	25
16	28	26
17	28	27
18	28	29
19	29	24
20	30	26
21	30	27
22	30	30
23	32	27
24	32	29
25	32	30
26	32	33
27	33	34
28	36	31
29	36	33
30	37	30
31	37	46
32	42	29
33	42	37
34	45	31
35	45	40
36	49	44
37	65	56
38	68	60

姐弟恋
老夫少妻



三、相关系数的区间估计：教材 P201

只要总体的相关系数 ρ 不等于0，则 r 的抽样分布就**不服从** t 分布或 u 分布，

---- 所以 r 的区间估计也就**不能直接使用** t 分布或 u 分布。

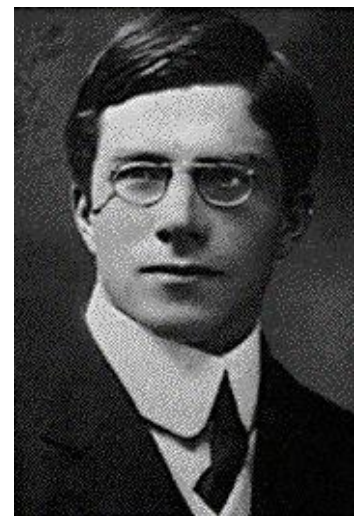
将 r 转换成 z 值：

$$z = 0.5 \ln [(1+r) / (1-r)]$$

z 值近似服从正态分布，其**标准误**为：

$$\sigma_z = 1 / \sqrt{(n - 3)}$$

太聪明了！



R. A. Fisher (1915, 25岁, 剑桥大学) 证明： z 值近似服从正态分布

z 值:

$$z = 0.5 \ln [(1+r) / (1-r)]$$

$$\sigma_z = 1 / \sqrt{(n - 3)}$$

z 值的用途:

(1)、区间估计: P137

$$z \text{ 的区间估计: } L = z \pm t_{\alpha} \sigma_z$$

$$r \text{ 的区间估计: } r = (e^{2L} - 1) / (e^{2L} + 1)$$

**(2)、检验相关系数与任意给定（相关系数）值的差异显著性
采用u 统计量**

$$H_0: \rho = \rho_0; \quad H_A: \rho \neq \rho_0$$

(3)、两个相关系数间的差异显著性检验

$$H_0: \rho_1 = \rho_2; \quad H_A: \rho_1 \neq \rho_2$$

$$\text{统计量 } u = (z_1 - z_2) / \sqrt{[1/(n_1 - 3) + 1/(n_2 - 3)]}$$

例:

[美]Rosner, B 著 《Fundamentals of Biostatistics》

测量了100个父亲与其大儿子的体重，计算他们间的相关系数为0.38, 问这个样本相关系数与潜在的从基因出发的相关系数0.5相一致吗？

样本相关系数（表型相关系数）： $r = 0.38$

理论相关系数（遗传相关系数）： $r = 0.5$

群体遗传学例子

解： $H_0: \rho = 0.5$

$H_A: \rho \neq 0.5$

$$z_0 = 0.549$$

$$z = 0.4$$

$$n = 100$$

$$u = (z - z_0) / \sqrt{[1/(n - 3)]} = -1.47$$

结论：接受 H_0

$$z = 0.5 \ln [(1+r) / (1-r)]$$

例2: 两个相关系数间的差异显著性检验

随机选男女生各11名，计算身高与体重的相关系数：

男生： $r=0.993478$

女生： $r=0.974755$

检验二者是否差异显著。

解： $H_0: \rho_1 = \rho_2; \quad H_A: \rho_1 \neq \rho_2$

$$z_1 = 2.86123$$

$$z_2 = 2.17979$$

$$n_1 = n_2 = 11$$

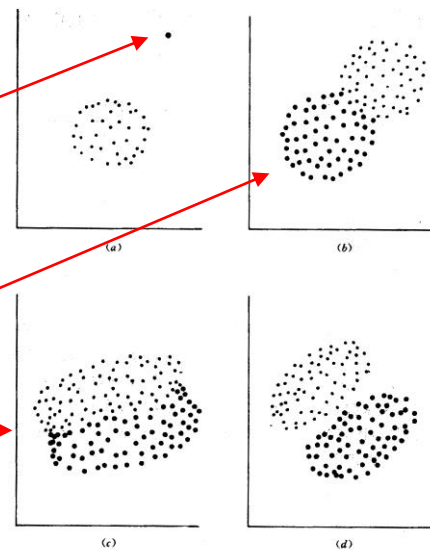
$$u = (z_1 - z_2) / \sqrt{[1/(n_1 - 3) + 1/(n_2 - 3)]} = 1.3628$$

结论：两个相关系数间无显著差异。

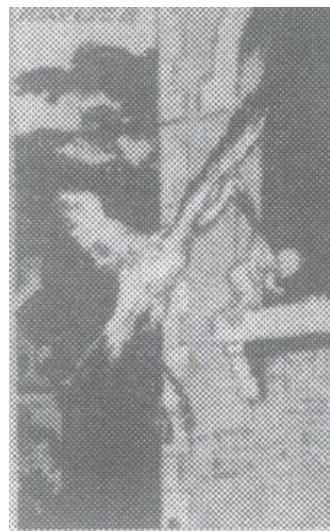
$$z = 0.5 \ln [(1+r) / (1-r)]$$

四、慎用相关——注意相关的**实际意义** (前面的假设检验是考虑其**统计学意义**)

- 1、**无相关**时未必真的无内在联系
- 2、**有相关**时未必真的有内在联系——因果检验
- 3、出现异常值时慎用相关
- 4、分层资料盲目合并时易出现假象



慎用相关的情形
(a) 异常值; (b), (c), (d) 分层资料不可合并



相关的实例：

20世纪20年代提出的裙边理论：

经济决定裙子长短——长筒丝袜时代
女人的裙子越短，经济形势就越好

长筒丝袜

武汉晨报 2003. 2. 19



裙子长短不再决定经济好坏

晨报讯 过去，美国流行这样一句话：“女人的裙子越短，经济形势就越好。”现在，美国迷你裙都快卖疯了，是不是美国经济和股市会像火箭般攀升呢？专家说，不一定。

经济曾决定裙子长短

20世纪20年代，有学者提出了所谓的“裙边理论”。当时沃顿商业学校的经济学家乔治·泰勒发现，在

经济状况良好的时候，女人就会穿短裙子，以炫耀自己穿的长筒丝袜；在经济状况糟糕的时候，女人无钱买丝袜，只得把裙边放低，这样人们就看不到她们穿没穿长筒丝袜了。

“裙边理论”不准了

流行观察家们说，即使“裙边理论”在某个方面来说是正确的，那也是长袜时代的事。现在，各种不同的

面料和低廉的生产费用，使设计师可以设计出各种衣服。今天的消费者比过去有了更多选择。

武汉晨报 2003. 2. 19

人们总希望穿着与众不同

市场分析指出，现在并不是经济决定人们最终买什么。经济和世界政治越糟糕，就会有更多的消费者选择流行服饰，他们觉得这样能远离纷扰，当然前提是有多余钱。

举个例子，去年秋天在2003年米兰春夏服装展示会上，模特们在T型台上，身着的是既短又性感的时装。意大利Diesel服饰公司的总裁安德里亚·鲁斯说：“从心理上来说，这没有任何意义，并不表明现在的经济有多么景气。它只是表明人们希望穿与众不同的东西而已。”

上图：16日，两名儿童在2003年伦敦秋冬时装周上看名模表演。

相关的实例

相处越久长得越像？

英科学家揭秘“夫妻相”

作者:王高山 发表时间:2006-2-16 摘自:新华网

我们都喜欢和自己长得像的人，因为他们往往和我们具有相似的性格特点。员的一项最新研究显示，与一个人相处的时间越长，那么两人在外貌上也就拥有相似点。

据“生活科学”2月14日报道，日常生活中，人们不难发现一些长时间生活在往往在外貌上很像，这就是人们所谓的“夫妻相”。

相同生活造就“夫妻相”？

英国科学家最近着手对“夫妻相”产生的原因进行调查研究。研究过程中，他男性参与者和11名女性参与者通过照片对160对夫妇的年龄、魅力和性格特点。由于丈夫和妻子的照片是分开进行观看的，因此这些参与者并不知道其中究竟谁夫妻。

研究者发现，参与者对事实上是夫妻的男人和女人的外貌和性格特点的评价而且，相处时间越长的夫妻，人们对他们的评价也就越相似。对此，研究人员推的生活经历可能会对夫妇的外貌产生潜移默化的影响。

外貌与性格具有某种关系

此项研究的参与者之一、英格兰利物浦大学的托尼·利特尔说，一个人的外貌两者间具有某种关系的理论，开始听起来或许会有些奇怪。但是，这其中可能存学上的原因。

恋爱越长婚姻越长久

据新华社电 新加坡社会学家调查发现，恋爱时间较长的情侣婚后不合甚至离婚的几率比恋爱时间较短的夫妻要低。

这项有1026名已婚者及827名离婚者参加的调查显示，除恋爱时间长短外，家务分工、伴侣工作超时的频率以及是否生养子女等都对婚姻有很大影响。其中，丈夫每分担多一项家务，婚姻持久的几率就增加1.8倍。

夫妻“魔鬼定律”

文/梦云

没衣穿定律：当妻子说没衣服穿的时候，她的意思是没有新衣服穿；当丈夫说没衣服穿的时候，他的意思是没有干净的衣服可穿。

炒菜定律：经常炒菜的多是妻子，炒菜好吃的多是丈夫。

买菜定律：一到菜市场就不知买什么菜好的多是妻子，一到菜市场见啥菜买啥菜的多是丈夫。

成熟定律：越是被妻子爱着的丈夫越是成熟，越是被丈夫宠着的妻子就越是不成熟。

说话定律：夫妻之间谁说的话越多，谁的话就越没分量。

伤害定律：夫妻之间，一方对另一方付出得越多，分手时所得到的伤害越大。

抱怨定律：经常抱怨的总是妻子，经常被抱怨的总是丈夫。

干活定律：在丈夫的眼里，家里总是没有什么活；在妻子的眼里，家里总是有干不完的活。

劝说定律：夫妻之间一旦发生矛盾，出面劝说的人越多，矛盾越是不容易解决。

做事定律：做事见好就收的总是丈夫，做事想好上加好的总是妻子。

出门定律：最着急出门的是妻子，最后一个出门的也是妻子。

爱你一生的前提

一段八年前就情终于结束，而这件事真正发生的然感到失望。港星琪的恋情再一次成热点，只是这一次女主角梁咏琪。在事的真相穷根究底透露出一点现代恋回归的价值取向是一边倒地向着那

若这样的三角一个传统的社会当责任，谴责第三者恋，那就一点都不这样。一个价值观多感情重感觉经责任，高歌“不求曾经拥有”的复杂社会当中，感不清，道不明的一件事情。人人至期望复合的旧人邵美琪和“间”的新人梁咏琪都说了同样的事情没有人会知道。”

每一段感情都会有激动人，没有人在对深爱的人说“因为照顾你一生一世”的时候，会想着天分手。但我们不要过于依赖，为有了这句话便一切永远不变。提首先得你一生都可爱，如果了，或者你仍然可爱，只是他已经要照顾你一生或被你照顾一生累。要让自己一生都可爱真的是难，据说邵美琪为了不想给邵伊闹不闹不问，只懂逃避，弄得自己泪流连夜店，以酒消愁。许多人对方善意的处事态度很可爱，这论支持的原因之一。但郑伊健来，见到熟睡中的 Maggie，突然缺少沟通，失去了爱恋的感觉，只离开的理由似乎也不错。

渴望自由地爱人，又渴望一，现代人的爱情智慧还需要究竟是什么呢？

最近有科学家研究出爱情是人的大脑中一种“化学鸡尾酒”的物质激发出来的，它们包括多巴胺、苯乙胺和催产素。时间一长，人会对这些令人对异性产生激动情绪的化学物质产生抗体，经过2年左右的时间，它们就逐渐失效了。当第三者出现时，这三种化学物质还会“死灰复燃”，只是它们同样不会超过30个月。这对爱情抱美好理想的人可能是个不小的打击。爱情除了要有“爱情物质”，它更多的是一

爱你一生的前提

问世间情为何物？

科学家说是化学物质

获得美国“人类工程学”博士学位的土耳其人类工程学家佐嘉，在回答记者提出“爱情是何物？”的问题时表示，“爱情”是人体脑部的激素化学物质经过一连串运动后产生的一种“结果”，只要脑部功能正常，就会有化学分子的变化，分泌出爱情激素，产生所谓的“七情六欲”，因此无论年龄大小，每个人都会对“爱情”有所感觉。

佐嘉分析爱情的产生时说，在经由五官感觉对方头发颜色、说话声音、身体气味和外在形体的同时，人脑也开始分析收集到的资料，并与出生以来就储存在脑中的其他记忆比较，比较结果

如果切合“期望”，脑中的视招下部，也就是间脑底部开始分泌

出一种化合物“苯乙胺”，这种物质涉及内脏功能的无意识调节、情绪、神经等。

分泌的“苯乙胺”愈多，发生的效力也就愈激烈，产生的“爱意”也就愈浓，最后成为“苯乙胺”的俘虏，也就成为所谓坠入情网的“爱情俘虏”。

(摘自《世界科技译报》)

科学家说是化学物质：与含量呈正相关

爱情是一种病

“爱情”也会致病

就像以前的医学研究得出的结论一样,坠入爱河的过程足以给人的内脏器官带来一系列的生理反应:当你看到你的爱人时,肾上腺激素分泌会增多,瞳孔扩大,心跳加速,汗腺分泌也随之增多。从技术的角度来看,这些机体反应再次证实,爱情带来的紧张和压力,甚至超过了人们面对考试或求职面试的压力。

五工止亦 叶1 自什照后从在

Love is a grave mental disease.
爱是一种害得不轻的“精神病”

在比萨大学，名坠人爱，对自己的，为，不羽叶一的，都是，一些，从十七名里，出了血样，失調，也，到血样进行，生血液复，公理性迷，者血液，百分之四，科学，常说，道埋，初恋的，液中复合，正常水平，编译/秋凌

作者:雪山 发表时间:2005-6-3 摘自:新华网

美专业杂志报道 人类脑部产生激情的区域促使人渴望接近爱人——

据《纽约时报》5月31日报道，最新一期的《神经生理学杂志》刊登了美国顶尖神经科学家对人类为爱痴迷的最新诠释：人类为爱发疯的行为就跟吃饭、睡觉、眨眼睛一样再平常不过。

生父被男友活活打死 女儿竟助凶手潜逃

一家酒店和一家中餐馆工作。17岁的女儿梁唐妮就读于曼哈顿数学及科技学校。梁氏夫妇得知女儿爱上了20岁的无业青年路易斯·安特瓦，极力反对。

热恋中的梁康妮和路易斯逐渐失去理智。11月2日夜間，梁父下班回家在家中看电视，路易斯从风衣上拍下腰带，在梁康妮协助下，将梁父活活勒死。晚10时许，梁母从餐馆回家，被埋伏一室的两人以相同的手法杀害。两人事后被警方抓获。

如果罪名成立，路易斯和萊康妮有可能面臨終身監禁。

亲生父亲不满其女儿和男友在家中同居,为此多次找女儿及其男友,结果被“准女婿”殴人活活打死,但女儿和其前妻不但阻止且向青岛法院控诉青年追逃5日。青高院中级人民法院作出一个判决:判处杀害前妻的手的女儿有罪,判处三年,判处犯死“准女婿”的前妻有期徒刑一年,缓刑四年。

王美香生活,因没有房子母女俩和父亲孙文泰同住一处,1999年18岁的孙菲香毕业后没找到工作,后与23岁的无业青年孙文鑫在家中同居。对此,母亲王美香有之任之,甚至喜留孙文鑫及装狐假虎势多人共居一室。同住一个屋檐下的父亲实在看不惯,多次为此责骂女儿及孙菲香,并为之发生争吵。

马回家责难的女儿及其男友文鑫。孙文鑫立即打电话找“小舅”周涛,分别用菜刀和手枪将孙有亮活活打死。女儿孙菲香和母亲王美香只是躲到阳台上并关上窗户。文鑫等人行动,而是趁孙菲香助孙文鑫出逃至银川,并资助孙文鑫在当地开了个酒吧。孙菲香先后两次到银川报复孙文鑫。

又为其租房并与之同居,不久母亲王美香还和孙菲一起为“准女婿”孙文鑫举办生日。去年8月31日,孙文鑫在李村桥摇车时被交警抓获,用菜刀帮孙文鑫砍杀“岳父”周涛,也在家人的劝说下投案自首。

(崔振开 刘世杰)
本报青岛消息

李岛市女青年孙鑫在父 2000年9月9日晚10 2001年6月,孙文鑫以 九州纪事
母于1998年病故后随其 时计,孙有充值爱酒劲又麻 为风自己过区同有站,孙菲

(姜淑芹 刘世杰)
本报青岛消息

九州纪事

黄江日报 2002.4.9



网上民调结果 令人大跌眼镜

男人爱宝钗 女人爱八戒

本报上海消息 最近有这样两则网上民意调查十分有趣。一则:《红楼梦》里美女如云,你愿意娶谁为妻? 一则:《西游记》的师徒四人当中,你愿意嫁给谁?

调查结果如下:以温文尔雅著称的薛宝钗以 5679 票名列榜首,精明强干的王熙凤以 2839 票屈居第二,俏丽灵巧的晴雯获 1894 票,名列第三,忠心耿耿的袭人获 985 票,就连恬不知耻的鲍二家的那位也得了 103 票,而

“天上掉下来”的林妹妹却只得了 5 票。在 98 名参与调查的女性中,选唐僧的无,选孙悟空的 10 人,选沙和尚的 14 人,选猪八戒的 74 人,猪八戒名列榜首。

跨入新世纪,男人最爱是宝钗,女人最爱是八戒,让你惊诧了吧?

为什么不喜欢林黛玉? 男人会说:她有病,到时候是我伺候她还是她伺候我? 她太爱吃醋,她太嫩,动不动吟百首诗作个面之乎

者也的,我的那半桶水哪里经得起她晃荡? 她太清高,她太傲,她太理想,她太单纯。当然还是宝钗好,她懂人情世故,她会八面玲珑,她肯妥协忍耐。我喝醉了,我回家睡了,只要是给领导、陪客户,她保准没意见。林黛玉能做到吗?

为什么喜欢猪八戒? 女人说:唐僧有什么好? 遇事百无一用,只会阿弥陀佛,还不阴不阳,全无一点男子汉气魄,女生给他

一个媚眼,还浑身打哆嗦。假正经! 孙悟空有什么好? 猴里猴气的,一出差就十万八千里,不闹家,一点人情味也没有。沙和尚人太老实,太听话,没情调。他会带我参加 party 吗? 他会给我买新型文胸吗? 八戒多好呵! 别看他长得丑,可他很温柔,也能干,能顾家,懂生活,会享受,有点油嘴滑舌,有点偷奸取巧,有点贪图实惠。这年头,这样的男人不吃亏,说不定领导还很喜欢呢!

《现代女报》近日在 98 位读者中做了一次民意调查,调查内容:如让你在唐僧师徒四人中选择一位做你的恋人,你将选择谁? 调查结果令人大吃一惊:选唐僧 0 人,选孙悟空 10 人,选沙僧 14 人,选猪八戒 74 人,占总人数的 72.7%。

点评:猪八戒以绝对优势压倒同门诸人,其原因可以用排除法分析——

首先,所有女性不喜欢唐僧,天天有十大罪状:①虚伪,总是叫徒弟们去化缘,自己双手合十,盘腿大坐,念念有

悟空可以做朋友 沙僧可以做老公 恋人还是八戒好

词,贪图安逸,②自私,自己骑马让别人走路,③贪生怕死,撒腿就跑,对自己徒弟横眉立目,真遇上妖魔鬼怪反面立刻装模作样,大喊救命。④没有领导风范,做事哼哼唧唧,还不阴不阳,没一点男子气概。⑤优柔寡断,见到美女不能斩断绝,与女儿国国王半推半就,畏葸不前。⑥贪仗人势,治不了孙悟空,去求观音教他紧箍咒,卑鄙小人,却说是为了救救猴性,选

唐僧无能,除了念经什么也不会做。⑦翻脸无情,悟空几次救他性命,对他有100个再造之恩,看他怎么对待悟空的?⑧不明是非,三打白骨精妇孺皆知,不须赘言。

其次,女性读者们认为,孙悟空是可以做好朋友、做“哥们”的那种人,如果和这种人交朋友,有什么七灾八难,只要打声招呼,包管上天入地替自己讨回公道,不过作为恋人,做哥

一点反叛精神。最后,相比之下,就猪八戒过得最可爱,他性情随和,为人宽宏,感情丰富,在高老庄做女婿时对丈人一片虔诚,又有金钱观念,知



战争数量与气温呈负相关

相关的实例

气温下降 — 产量减少 — 战争增多 — 改朝换代

《科学通报》ISSN:0023-074X 2004年 第49卷 第23期

2468-2474

气候变化与中国的战争、社会动乱和朝代变迁

章典^① 詹志勇^① 林初升^① 何元庆^② 李峰^①

(^①香港大学地理系, 香港; ^②中国科学院寒区旱区环境与工程研究所, 兰州 730000. Email: zhangd@hkucc.hku.hk)

摘要 人类进化史始终与气候变化有密切的关系, 这种观点已被科学家们所接受. 尽管人们目前都在预测气候变化对我们将来社会的影响, 但是至今还没有对有历史记载以来气候变化对社会发展和演化之影响进行系统和定量的研究. 利用过去 1.15 ka 来的古气候记录, 对中国唐末到清朝的战争、社会动乱和社会变迁进行了系统地对比分析. 结果发现冷期战争率显著高于暖期, 70%~80% 的战争高峰期, 大多数的朝代变迁和全国范围动乱都发生在气候的冷期. 研究表明, 由于冷期温度下降导致土地生产力下降, 从而引起生活资料的短缺. 在这种生态压力和一定的社会背景下, 战争高峰期和全国范围内的社会动乱随之产生. 在许多情况下, 最终导致王朝灭亡和新朝代的建立. 进一步分析还发现战争数量与温度距平有显著的负相关关系. 在不同的气候带, 由于土地承载力的不同, 战争与温度的相关程度也存在着差别. 因此我们认为所谓中国历史的朝代循环, 以及大乱和大治的交替, 气候的波动变化是决定性因素之一. |

因与果

- 无巧不成书：故事往往总是很玄
- 少了一个铁钉，失了一个马掌；
- 少了一个马掌，失了一匹战马；
- 少了一匹战马，丢了一个国王；
- 丢了一个国王，输了一场战争；
- 输了一场战争，失了一个国家。

回归分析

相关分析

步骤:

(1)、求:

回归方程
 $y = a + bx$

相关系数
 r

(2)、检验:

方程意义 (F)
a (t)
b (t)

r (t)

(3)、使用:

预测

适用对象: 因果关系

平行关系

8-3 可直线化的非线性回归分析

直线关系是双变量间最简单的一种关系。
双变量间的非线性关系更普遍。

目的：P172

- ① 确定规律：两个变量之间数量变化的规律
- ② 估计参数：回归系数、极大（小）值、渐近值
- ③ 预测
- ④ 修匀试验数据：避免异常观测值的影响

步骤：

- (1) 作散点图（判断所符合的曲线方程类型）直线化
- (2) 求直线回归方程： a 、 b 值；检验回归关系
- (3) 转换成曲线回归方程

曲线



直线



曲线

8-3 可直线化的非线性回归分析

一、非线性回归的直线化

二、倒数函数曲线

三、对数函数曲线

四、指数函数曲线

五、幂函数曲线

六、Logistic生长曲线

一、非线性回归的直线化

$$y = a + bx$$

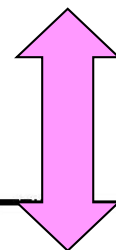
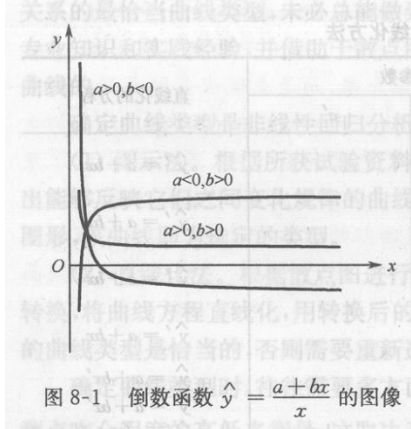


表 10-1 常用曲线模型的直线化方法

曲线回归方程	经尺度转换的新变量及参数			直线化的方程
	y'	x'	a'	
$\hat{y} = \frac{a + bx}{x}$	$y' = yx$			$\hat{y}' = a + bx$
$\hat{y} = \frac{1}{a + bx}$	$y' = \frac{1}{y}$			$\hat{y}' = a + bx$
$\hat{y} = \frac{x}{a + bx}$	$y' = \frac{x}{y}$			$\hat{y}' = a + bx$
$\hat{y} = ax + bx^2$	$y' = \frac{y}{x}$			$\hat{y}' = a + bx$
$\hat{y} = a + b \ln x$		$x' = \ln x$		$\hat{y} = a + bx'$
$\hat{y} = a + b \lg x$		$x' = \lg x$		$\hat{y} = a + bx'$
$\hat{y} = ax^b$	$y' = \ln y$	$x' = \ln x$	$a' = \ln a$	$\hat{y}' = a' + bx'$
$\hat{y} = ae^{bx}$	$y' = \ln y$		$a' = \ln a$	$\hat{y}' = a' + bx$
$\hat{y} = axe^{bx}$	$y' = \ln \frac{y}{x}$		$a' = \ln a$	$\hat{y}' = a' + bx$
$\hat{y} = \frac{1}{ax^b}$	$y' = \ln \frac{1}{y}$	$x' = \ln x$	$a' = \ln a$	$\hat{y}' = a' + bx'$

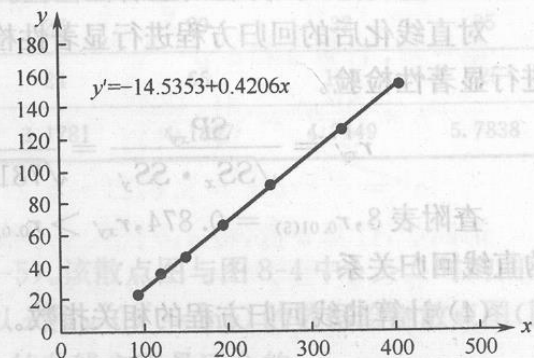
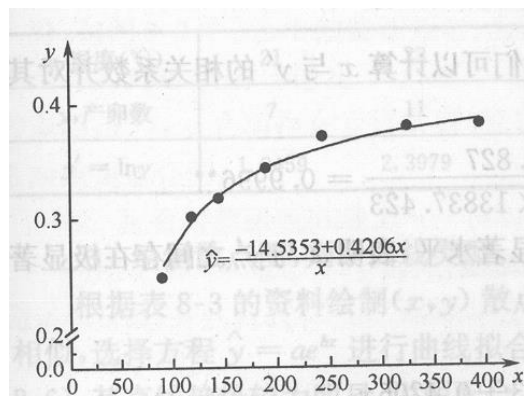
二、倒数函数曲线:



曲线回归方程	经尺度转换的新变量及参数			直线化的方程
	y'	x'	a'	
$\hat{y} = \frac{1}{a+bx}$	$y' = \frac{1}{y}$			$\hat{y}' = a + bx$
$\hat{y} = \frac{x}{a+bx}$	$y' = \frac{x}{y}$			$\hat{y}' = a + bx$

第四版，P142，
例8.1

玉米株重 x
经济系数 y



三、对数函数曲线:

$y=a+b\lg x$

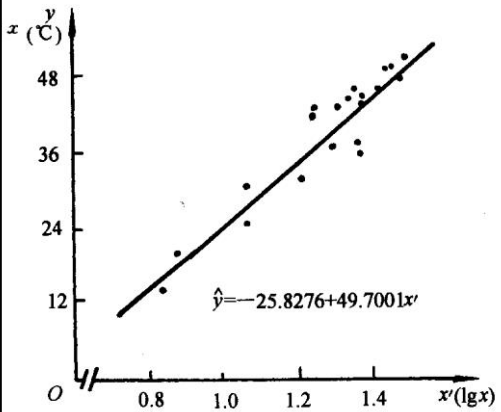
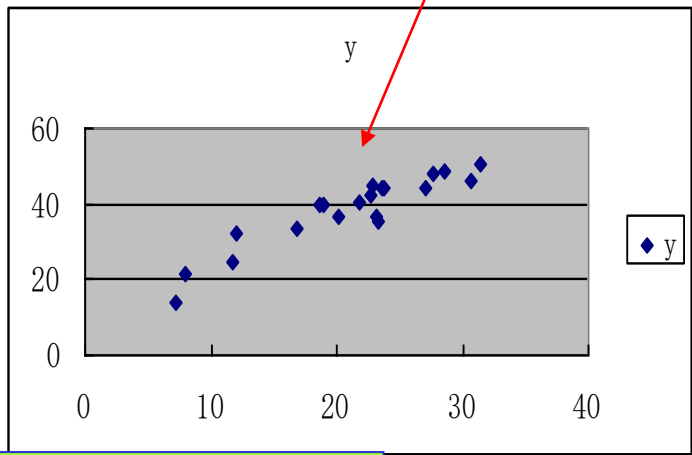
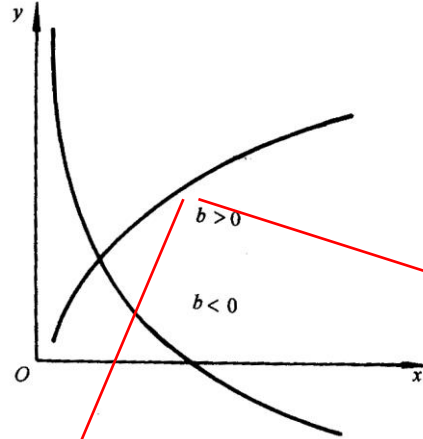


图 10.3 例 10.1 资料 x' 与 y 之间的直线关系

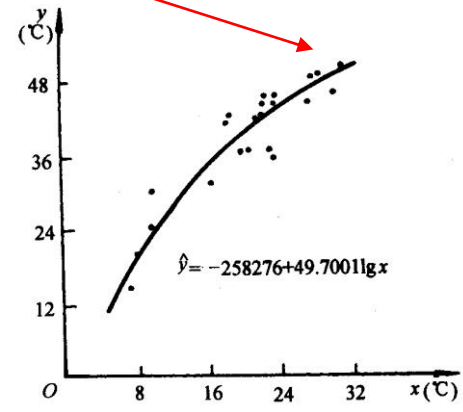


图 10.2 苗床内最高气温 y 与空气最高气温 x 的关系

水稻薄膜育秧:
膜内苗床空气温度 y
膜外空气温度 x

表8-6 x' 与 y 回归关系的假设检验

变异来源	df	SS	s^2	F	$F_{0.05}$	$F_{0.01}$
回 归	1	1549.2466	1549.2466	161.78 **	4.41	8.29
离回归	18	172.3709	9.5762			
总变异	19	1721.6175				

四、指数函数曲线: $y = a e^{bx}$ $\ln y = \ln a + bx$
 $y = a b^x$ $\ln y = \ln a + (\ln b)x$

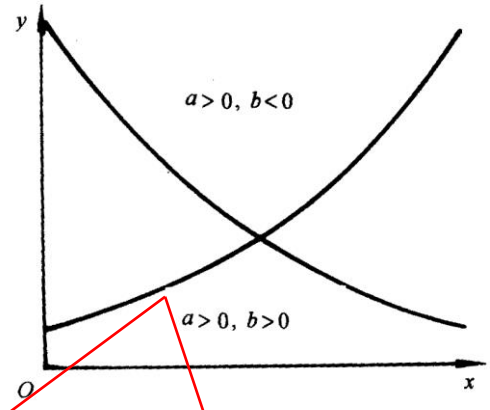


图 10.4 指数曲线 $\hat{y} = ae^{bx}$ 的图象

例P145: 棉花红铃虫产卵数 y 与温度 x 的关系

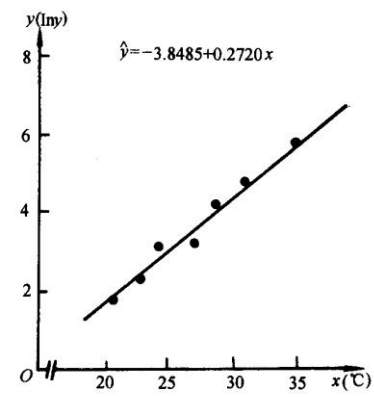
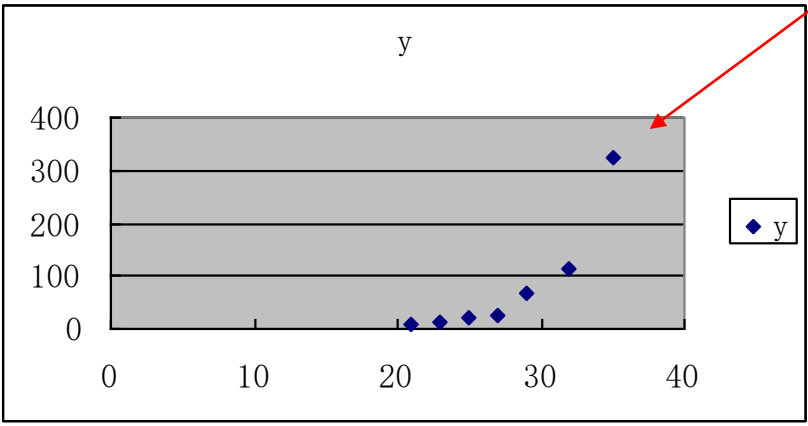


图 10.6 例 10.2 中 $y'(\ln y)$ 与 x 之间的直线关系

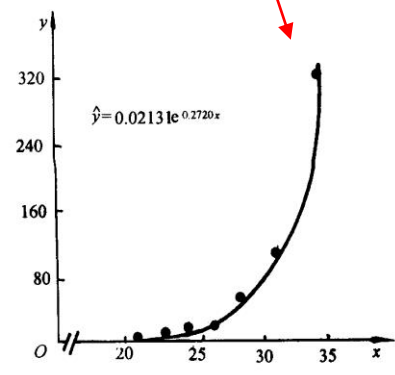


图 10.5 红铃虫产卵数 y 与温度 x 的关系

五、幂函数曲线:

$$y = a x^b \quad \lg y = \lg a + b \lg x$$

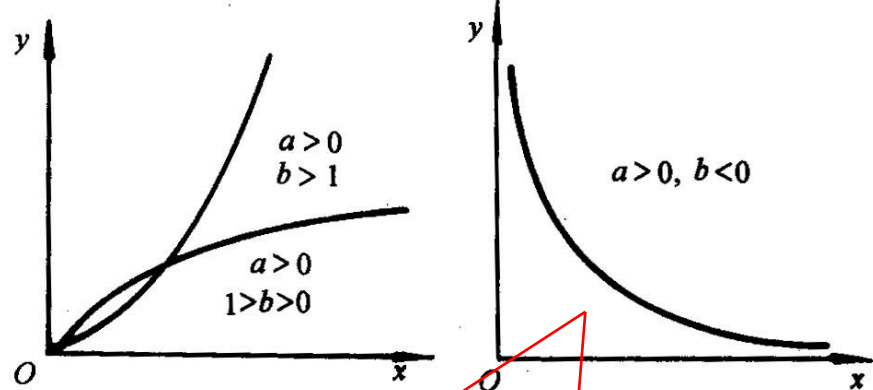


图 10.7 幂函数曲线 $\hat{y} = ax^b$ 的图象

例P149: 烟叶的叶绿素含量 y 与烘烤时间 x 的关系

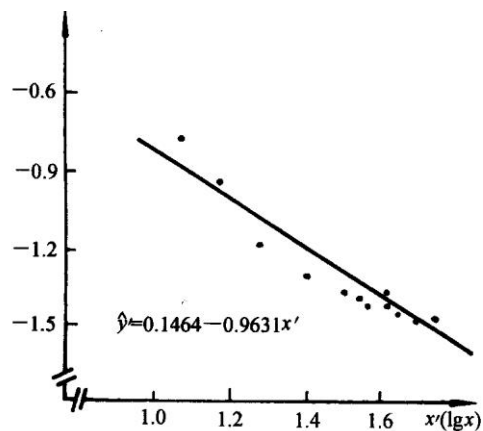
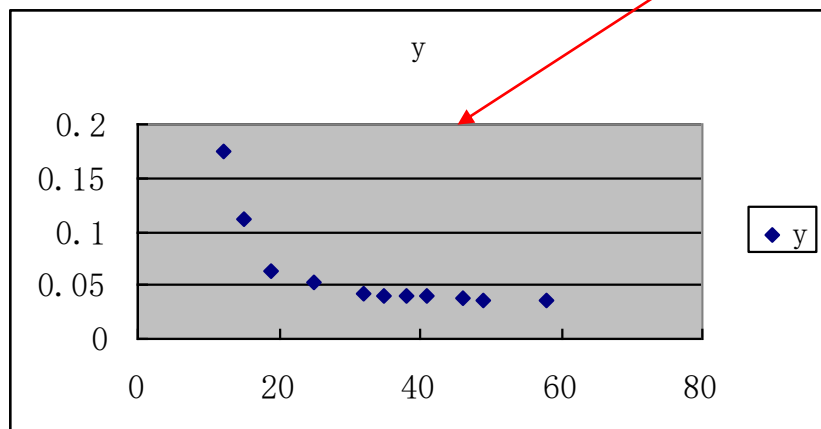


图 10.9 例 10.3 中 $y'(\lg y)$ 与 x 之间的直线关系

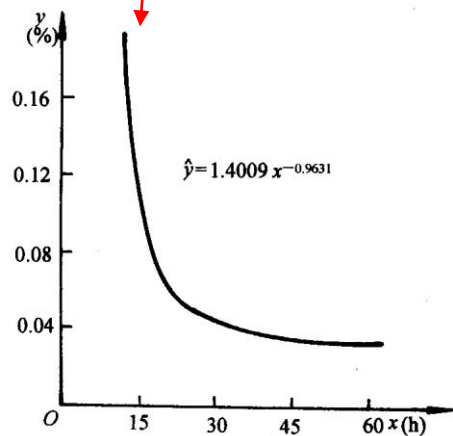


图 10.8 烟叶烘烤时间 x 与叶绿素含量 y 的关系

六、Logistic生长曲线

S 曲线

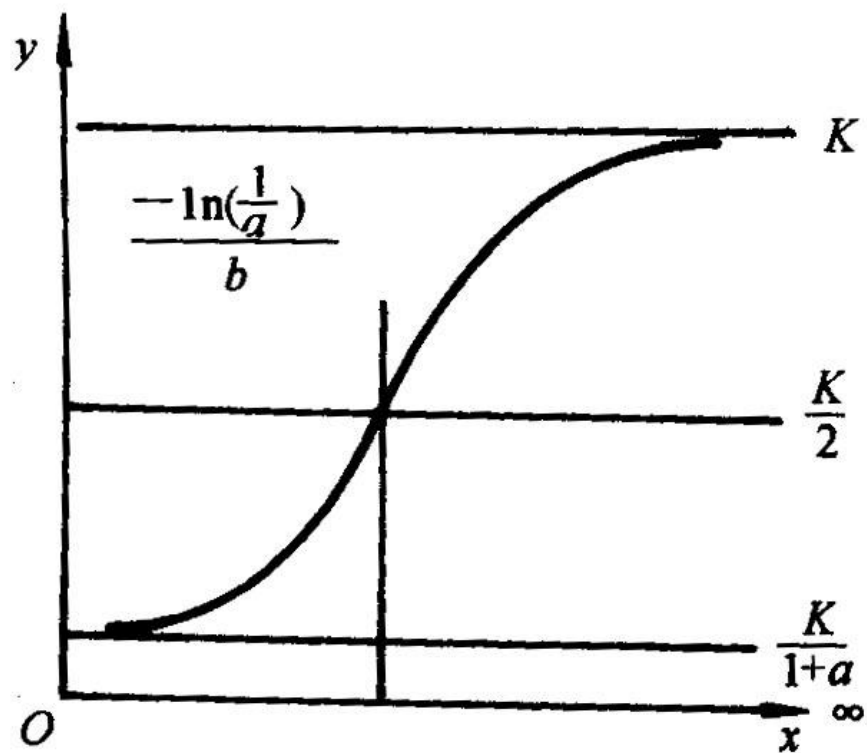


图 10.10 Logistic 生长曲线 $\hat{y} = \frac{K}{1 + ae^{-bx}}$ 的图象

作业：共四题

第一题、第二题：第四版P138：习题7.2和习题7.4

第三题、第四题：第四版P154：习题8.1和习题8.2