第九章 相关分析和回归分析

§1 相关关系

在许多场合下我们经常需要研究一些事物之间的关系，如果从定量角度来研究，就归结为某些变量之间关系的研究。

从大的方面说，变量之间的关系有两大类。一类是确定性的关系，如给定正方形的边长 a ，则正方形的面积为 a^2 。另一类是不确定的关系，即两者有一定的制约关系，但还不能由一个变量来决定另一个的取值。这种关系称为相关关系。例如一个人的身高和体重之间的关系以及商品的价格和产量之间的关系都是相关关系。

产生相关关系的原因很多，大体上有如下几个原因：

1) 给出变量的数值时会产生误差。如物体重量(y)＝比重× 体积(V)，体积测量存在误差时，同一物体在每次测量中得到不同体积，故y和V就没有确定的关系，
2) 影响变量取值的因素不止一个变量。设y 是“果”,x是“因”。例如y为亩产量, x为施肥量。x和y有因果关系，但影响y的因素不止施肥x一个，还有种子品种，日照时间等等。这时x与y的关系只能是相关关系。
3) x和y的关系是通过其他因素反映出来的。例如以x表示某人一年旅游的支出，y表示某人一年的饮食支出。则x和y会有关系(当x大时, 一般y也大)，但它们之间不是确定的因果关系。因为这两者都由该人的收入所确定。 社会学研究中会发现收入高的家庭小孩人数越少。这两者之间也没有因果关系，其后面有其他复杂的社会因素影响着这两个变量。如何描述这种相关关系？统计上用相关系数来描述这种相关关系。

§2 相关系数(Pearson coefficient of correlation)

1. 比例刻度数据相关系数 $(x_1,y_1),(x_2,y_2),\cdots;(x_n,y_n)$,

对变量x和y观察到了n组数据

则变量和变量的相关系数定义为

$$r=\frac{n\sum x_i y_i-\sum x_i\sum y_i}{\sqrt{n\sum x_i^2-(\sum x_i)^2)(n\sum y_i^2-(\sum y_i)^2)}}$$

若记 $$S_{xx}=\sum x_i^2-\frac{1}{n}(\sum x_i)^2=\sum x_i^2-n\bar{x}^2=\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$S_{yy}=\sum y_i^2-\frac{1}{n}(\sum y_i)^2=\sum_{i=1}^{n}(y_i-\bar{y})^2$$

$$S_{xy}=\sum(x_i-\bar{x})(y_i-\bar{y})=\sum x_i y_i-\frac{1}{n}\sum x_i\sum y_i$$

则 $$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \qquad (1)$$

如果把 $(x_1,\cdots,x_n)$和 $(y_1,\cdots,y_n)$看成是从总体X和总体Y中得到的一组样本,则 r 即为这两个总体的样本相关系数，因此很容易推知

$$|r| \le 1$$

$r = 1(-1)$	变量x与y正(负)线性相关;
$r \in (0,1)$	变量x与y正相关;
$r = 0$	变量x与y线性无关;
$r \in (-1,0)$	变量x与y负相关。

注意: r 是刻画两组数据是否线性相关的一个度量，因此 $r=1$ 意味着y与x同步增长，$r\in(0,1)$意味着数据x越大，y也越大，但两者无严格的线性关系。$r=0$ 表示两组数据线性无关，即x的变化与y是否变化没有明确的关系，如果把数据 $(x_i,y_i),\ i=1,\cdots,n$ 标在坐标纸上，r 表示描述两者相关程度的直线与x轴夹角正切的倍数。而 $r=\pm1$ 表明组数据在一条直线上。

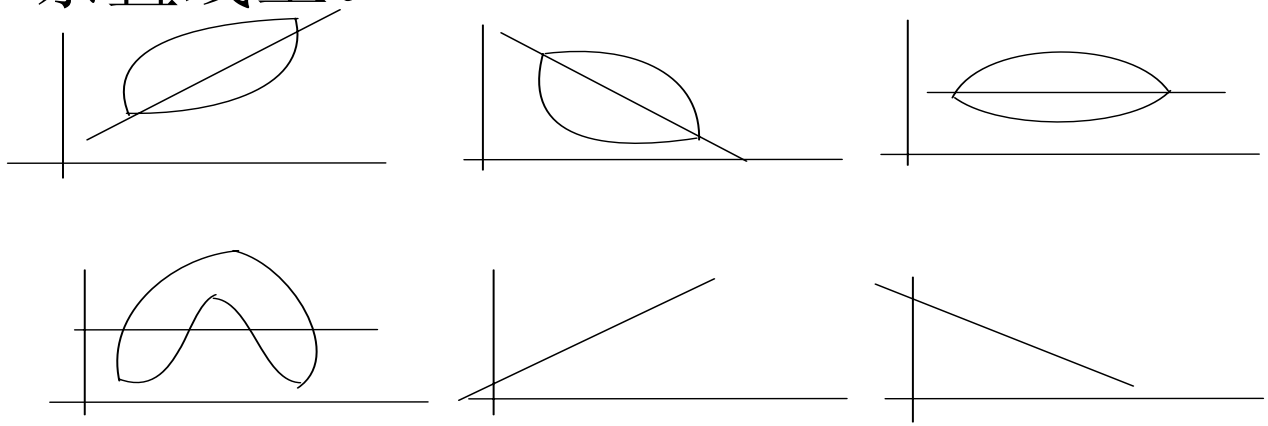


图 9.1 各 种 相 关 关 系 示 意 图

例1.分析 *Jasbeerajan&Co.* 销售额y与推销费x的关系

年.季度	销售额（万元）	推销费（万元）
1.1	169	50
1.2	52	18
1.3	140	58
1.4	733	76
2.1	224	56
2.2	114	45
2.3	181	60
2.4	753	69
3.1	269	61
3.2	214	32
3.3	210	58
3.4	860	95
4.1	345	63
4.2	203	31
4.3	233	67
4.4	922	123
5.1	324	76
5.2	224	40
5.3	284	85
5.4	822	124
6.1	352	86
6.2	280	39
6.3	295	81
6.4	930	135
7.1	345	61
7.2	320	50
7.3	390	79
7.4	978	140
8.1	483	90
8.2	320	91

由（1）可算得销售额y与推销费x的相关系数为r=0.8378，由此知：

1）销售额与推销费正相关，即推销费多，则销售额也大。

2）销售额与推销费不是严格的线性关系，但r=0.8378表明两者高度相关。r^2 也称为可决系数。它反映由于相关关系，y的变化可以由x来解释的百分比，上例中 ，$r^2 = 0.70198$ 故销售总量中的70%可以由推销费来解释,而其余的30%要由其他因素来解释。

2.非线性相关比例刻度数据相关的度量

两个变量之间可能没有线性关系，但有某种曲线关系。如图9.2，数据大体在某条指数曲线附近。如果对y的值取对数，则数据 $(x_i, \log_{10} y_i)$ 在某条直线附近。因此可以计算x和lg y的相关系数。 我们把 $z = \log_{10} y$ 称为线性性变换，进而研究x与z之间的线性相关性。

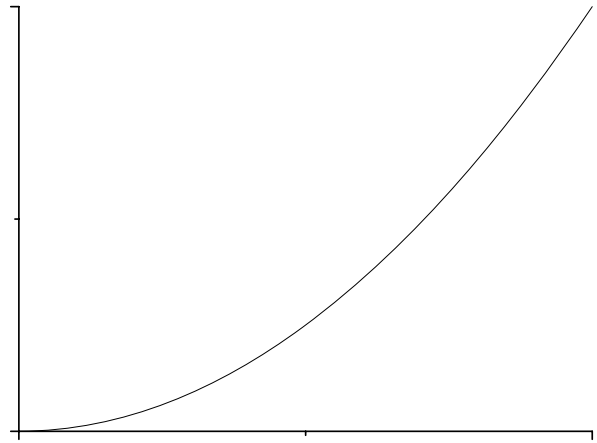
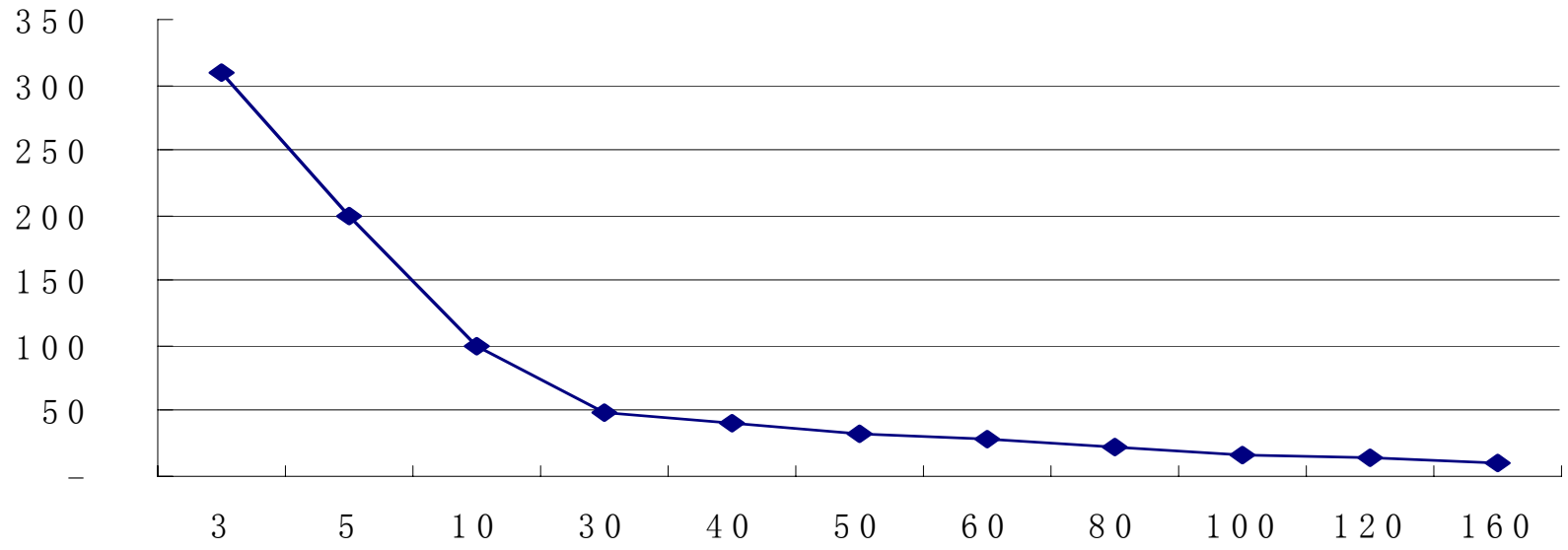


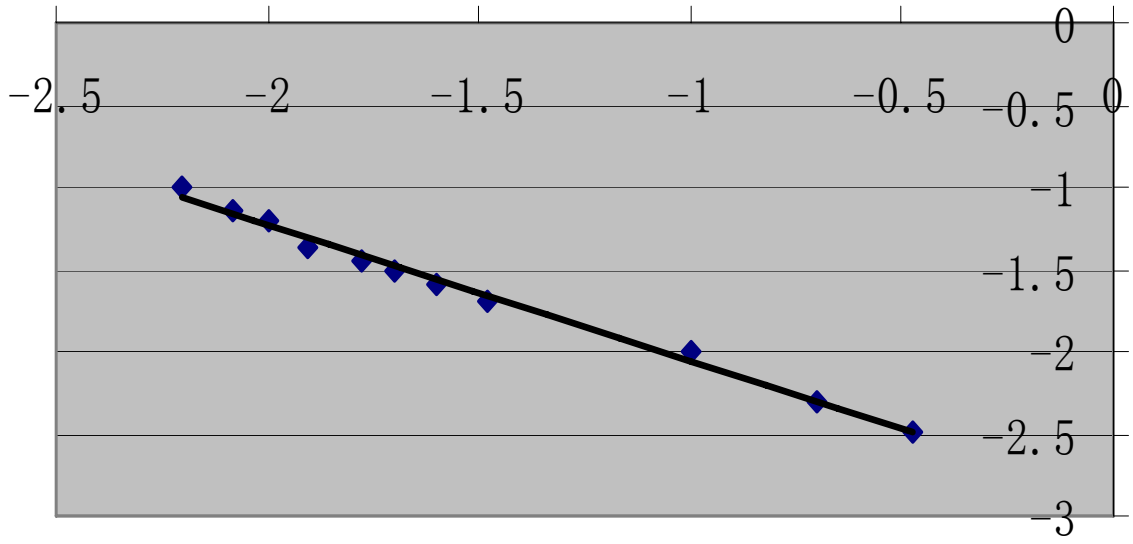
图9.2

例2. 某厂表面处理车间试验将铬后污水同电解污泥混合，使之生成无毒溶液，效果很好。但实际排污水浓度不同，而一定浓度的定量铬后污水只有同定量的电解污泥混合，才能反应完全。现通过试验，找出铬后污水用量与电解污泥用量之比对于铬后污水浓度之间的关系，数据如下：

序号	铬后污水浓度（g/ℓ） x	$\log x$	铬后污水用量/电解污泥用量 y	$\log y$
1	3	(.4771)	310	(2.491)
2	5	(.6990)	200	(2.301)
3	10	(1.0)	100	(2.000)
4	30	(1.477)	49	(1.690)
5	40	(1.602)	40	(1.602)
6	50	(1.699)	32	(1.505)
7	60	(1.778)	28	(1.447)
8	80	(1.903)	23	(1.362)
9	100	(2.0)	16	(1.204)
10	120	(2.079)	14	(1.146)
11	160	(2.204)	10	(1.000)

对这批数据，先作散点图，由散点图我们可以用幂函数 $y = dx^b$ 来描述铬后污水用量与电解污泥用量之比y和铬后污水浓度x的关系。即用 $\lg y = \lg d + b \lg x$ ，来近似表示。





$y = -0.8263x - 2.8843$

作线性性变换 $u=\lg y, t=\lg x$ ， 则数据变换($(\lg x_i, \lg y_i)$, 由此算得变换后数据的相关系数 $r=-0.9967$ 。这说明变换后数据高度相关，这些数据比较集中地落在某直线（斜率为负）的附近，这也说明原数据不是线性相关，而是大体在曲线附近 $y=dx^b$ 。

假设检验 ， 原假设r=0, 统计量

$t=r\sqrt{\frac{n-2}{1-r^2}}$ 服从自由度为n-2的t分布。

3.有序数据相关系数

有序数据是由数据在一个有序名单中的位置值组成。如排名的名次，对某事物的表态（很支持，支持，无所谓，不支持，极不支持 。有序数据不同于刻度数据。对有序数据，可表达为 $1, 2, \cdots, n$ (排序)，对这种序，可以定义“ Sperman ”相关系数。

例3．*Jasbeer , Rajan Co*. 10个产品销售情况的排序（去年和今年）。

产品	去年排名	今年排名	去年-今年排名
饼干类食品	1	3	-2
游戏绳	2	4	-2
帽子	3	1	2
假面具	4	2	2
游戏食品	5	6	-1
气球	6	10	-4
口哨	7	9	-2
饰带	8	7	1
旗帜	9	8	1
微型趣味小书	10	5	5

定义 *Sperman* 秩相关系数为

$$r = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

其中 $d_i =$ 第i个产品的去年名次--今年名次，
n: 总产品个数。本例中，

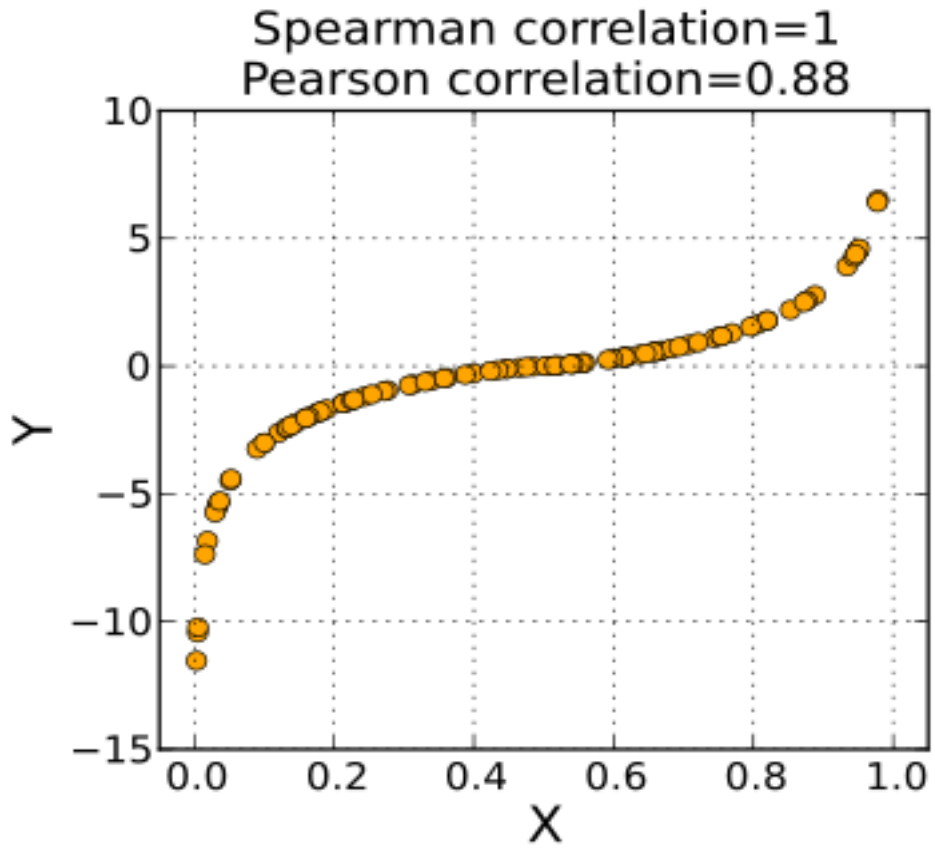
$$\sum d_i^2 = 5\times 2^2 + 3\times 1^2 + 4^2 + 5^2 = 64,$$

$$n(n^2 - 1) = 10(100 - 1) = 990$$

故

$$r = 1 - \frac{6\times 64}{990} = 1 - 0.387878 = 0.61212$$

这说明今年产品销售名次与去年有很大关系，原来名次在前的，今年大体也在前面。


Spearman correlation=1
Pearson correlation=0.88
10
5
0
−5
−10
−15
Y
0.0
0.2
0.4
0.6
0.8
1.0
X

Kendall tau 秩相关系数

Let (x_1, y_1), (x_2, y_2), …, (x_n, y_n) be a set of joint observations from two random
variables *X* and *Y* respectively, such that all the values of (x_i) and (y_i) are unique.
Any pair of observations (x_i, y_i) and (x_j, y_j) are said to be *concordant*（一致的） if the
ranks for both elements agree: that is, if both $x_i > x_j$ and $y_i > y_j$ or if both
$x_i < x_j$ and $y_i < y_j$. They are said to be *discordant*（不一致的）, if $x_i > x_j$ and $y_i < y_j$ or
if $x_i < x_j$ and $y_i > y_j$. If $x_i = x_j$ or $y_i = y_j$, the pair is neither concordant nor discordant.
The Kendall tau coefficient is defined as:

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\frac{1}{2}n(n-1)}.$$

Under a null hypothesis of *X* and *Y* being independent, the sampling distribution of τ will have an expected value of zero. The precise distribution cannot be characterized in terms of common distributions for larger samples, it is common to use an approximation to the normal distribution, with mean zero and variance

$$\frac{2(2n+5)}{9n(n-1)}$$

cor.test(x, y, alternative = c("two.sided", "less", "greater"), method = c("pearson", "kendall", "spearman"), exact = NULL, conf.level = 0.95, continuity = FALSE, ...)

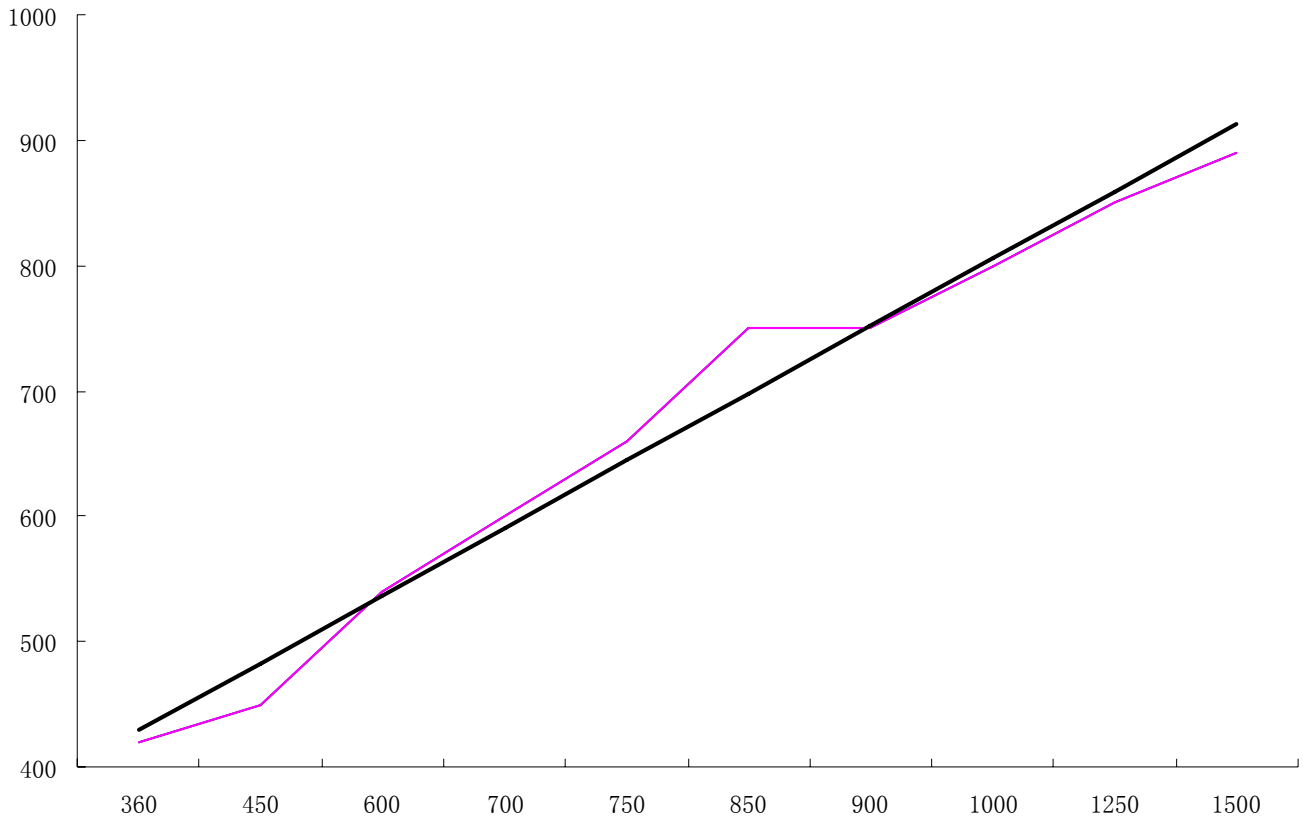
```
x <- c(44.4, 45.9, 41.9, 53.3, 44.7, 44.1, 50.7, 45.2, 60.1)
y <- c( 2.6, 3.1, 2.5, 5.0, 3.6, 4.0, 5.2, 2.8, 3.8)
cor.test(x, y, method = "kendall", alternative = "greater")
cor.test(x, y, method = "kendall", alternative = "greater", exact = FALSE)
cor.test(x, y, method = "spearm", alternative = "g")
cor.test(x, y, alternative = "g")
```

§3 回归分析

回归分析(*regression analysis*)是研究两组变量之间相互关系的统计分析方法。其中最重要的是研究一个因变量和一个或几个自变量之间的线性关系。首先我们研究一个因变量和一个自变量之间的线性关系。设因变量y（通常是随机变量）和一个非随机变量x之间有某种相关关系。在x的不全相同的取值点 $x_1,\cdots,x_n$ 作独立观察得到y的n个观察值 $y_1,\cdots,y_n$ ，记为 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$ 。我们的目的是从这n组数据中寻求x和y之间的关系。

例4：观察家庭月收入与月支出之间的关系，随机抽取10个家庭作调查得如下结果：

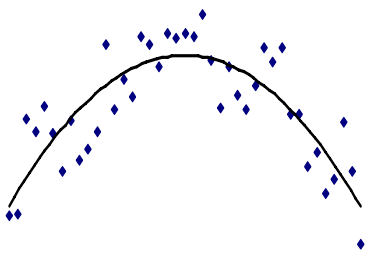
收入（x）	600	450	700	850	1250	1500	1000	900	750	360
支出 (y)	540	450	600	750	850	890	800	750	660	420

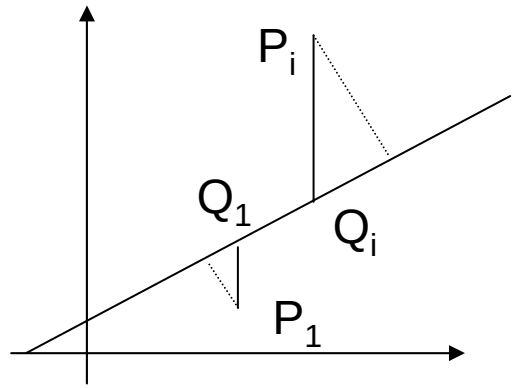


把 (x_i, y_i), $i = 1,2,\cdots,10$ 。标在图上，称为散点图。发现这10个点在一条直线附近，即x,y 的关系可以用一条直线近似表出，这条直线称为回归直线，此时y和x的关系可记为

$$y_i = a + bx_i + e_i$$

其中，e_i 称为误差，由于 e_i 的存在，使x和y不是确定性关系，而是相关关系。当然，一般散点图不一定能用直线来很好地表示出来，如下面左边的散点图可以用抛物线来较好地表达，此时这条抛物线称为回归曲线。我们的任务是根据n组数据的值来求出回归曲线。





回归曲线中最简单的是直线，设x和y有如下关系

$$y = a + bx + e \qquad (2)$$

其中 e 为误差（统计上把它看作随机变量），有如下性质：

$$Ee = 0, \quad \mathrm{E}e^2 = \sigma^2$$

$$Cov(e_i, e_j) = 0 \quad （不同数据对的误差不相关）$$

设我们已有了n组观察数据 。问题是如何定出直线，即定出a和b。一般而言直线不可能同时过n个点。我们希望这n个点和直线距离之和越小越好，这等价于右上图中n个散点与直线上相同横坐标的点的距离平方和最小，即希望求出a,b 使

$$f(a,b) \hat{=} \sum_{i=1}^{n} (y_i - a - bx_i)^2 = min$$

用多元函数求极小值方法可求得

$$\hat{a} = \bar{y} - \hat{b}\bar{x}, \qquad \bar{y} = \frac{1}{n}\sum_{i-1}^{n} y_i \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{b} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) / \sum_{i=1}^{n} (x_i - \bar{x})^2$$

记

$$S_{xx} = \sum (x_i - \bar{x})^2, S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x}) y_i$$

则

$$\hat{a} = \bar{y} - \hat{b}\bar{x}, \qquad \hat{b} = \frac{S_{xy}}{S_{xx}}$$

如用带线性回归功能的计算器运算，$\hat{a},\hat{b}$ 都能在计算器上直接得到。如果没有线性回归功能则可利用公式

$$S_{xy}=\sum_{i=1}^{n}x_i y_i-\frac{1}{n}\sum_{i=1}^{n}x_i\sum_{j=1}^{n}y_j$$

求出 $\hat{b}$ 和 $\hat{a}$ 。对例4而言,可得到

$$\hat{a}=296.47 \qquad \hat{b}=0.4480$$

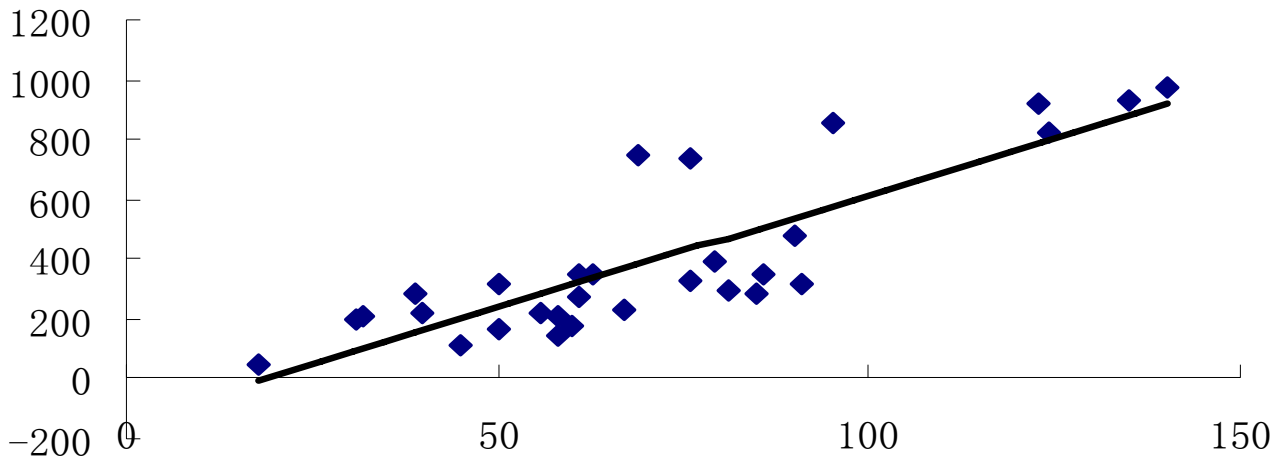
用(2)得到估计的方法称为最小二乘法，记为LS 估计(Least Square)，在本例中得到回归方程为

$$y=296.47+0.4480\,x$$

例5 . 考虑某公司销售量与推销费的关系，数据见下表。

年.季度	销售量	奖金
1.1	166	50
1.2	52	18
1.3	140	58
1.4	733	76
2.1	224	56
2.2	114	45
2.3	181	60
2.4	753	69
3.1	269	61
3.2	214	32
3.3	210	58
3.4	860	95
4.1	345	63
4.2	203	31
4.3	233	67
4.4	922	123
5.1	324	76
5.2	224	40
5.3	284	85
5.4	822	124
6.1	352	86
6.2	280	39
6.3	295	81
6.4	930	135
7.1	345	61
7.2	320	50
7.3	390	79
7.4	978	140
8.1	483	90
8.2	320	91

解：先作散点图，发现这些点大体在一条直线附近。我们来寻求销售额与推销费之间的线性关系，即给出回归直线。由计算得：



$$\sum x_i = 2139 \qquad \sum y_i = 11966 \qquad \sum x_i^2 = 179291$$

$$\sum y_i^2 = 694797 \qquad \sum x_i y_i = 105539$$

$$\hat{b} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = 7.550894$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = \frac{11966}{30} - 7.55089 \times \frac{2139}{30} = -139.51208$$

故

$$\hat{y} = -139.512058 + 7.550894\ x$$

即销售额= $-139.51 + 7.55 \times$ 推销费。

这就是说推销费每增加1元, 销售额平均增加7.55元。

由(2)式，误差 $e_i = y_i - a - bx_i$

一个重要的问题是估计误差的方差 σ^2 ，因为它涉及到预测精度的估计。可以证明

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2 \hat{=} \frac{1}{n-2}RSS \quad (\frac{Q}{n-2})$$

是 σ^2 的无偏估计(即 $\mathrm{E}\,\hat{\sigma}^2 = \sigma^2$)。以后我们将一再用到这一估计。

$RSS = \sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2$ 称为残差平方和，它描述了回归直线与n个点拟合好坏的程度。

误差方差 σ^2 的大小与预测有很大关系。因为标准差 σ 是反映波动大小的一个量。如果 σ 大，则预测精度就差。

前面已定义 $(x_i, y_i), i = 1, \cdots, n$ 的样本相关系数为

$$\hat{r} = S_{xy} \Big/ \sqrt{S_{xx} S_{yy}}$$

在线性回归分析中回归系数估计为 $\hat{b} = S_{xy} / S_{xx}$

故 $\hat{b} = \hat{r}\sqrt{S_{yy} / S_{xx}}$ 或 $\hat{r} = \hat{b}\sqrt{S_{xx} / S_{yy}}$

此外，可得到

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

$$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{a} + \hat{b}x_i - \bar{y})^2$$

$\sum_{i=1}^{n}(y_i - \bar{y})^2$ ：总平方和，　$\sum_{i=1}^{n}(y_i - \hat{y})^2$ ：误差平方和

$\sum(\hat{y}_i-\bar{y})^2$ ：回归平方和，　　$\hat{y}_i=\hat{a}+\hat{b}x_i$

因此　$r^2=1-\dfrac{\text{误差平方和}}{\text{总平方和}}=\dfrac{\text{回归平方和}}{\text{总平方和}}$

简单计算可得：回归平方和$=b^2S_{xx}$

由上式知 r^2（称为**可决系数**）表示由相关关系可解释的变差占总变差的百分比。在没有回归计算功能的计算器上，可以用下式来计算误差平方和：

$$RSS=(1-r^2)S_{yy}\qquad S_{yy}=(n-1)S_y^2$$

其中 S_y 为样本 y 的标准差，这在下一节的预测中是非常有用的公式。

§4 $\hat{a},\hat{b}$ 的检验和预测

1.检验

对n组数据 (x_i, y_i), $i = 1.\cdots, n$, 形式上都可以由最小二乘法画一条直线作为这组数据的回归直线。但这条直线是否有意义?这是一个值得讨论的问题。其中在应用上最有意义的是检验回归系数是否为零的问题：

$$\mathrm{H}_0 : b = 0 \leftrightarrow H_1 : b \neq 0$$

如果b=0,则回归直线为y=a, 即y的取值与x的值无关，如果 $b \neq 0$, 则表明x,y之间有一定的线性关系。因此在根据LS作出a,b估计后，希望检验是否为0。如果原假设成立，则$|\hat{b}|$应倾向于取小的值, 时$|\hat{b}|$应倾向于取大值，故有如下检验规则：对适当的常数C，当

$|\hat{b}| \le c$ 接受 H_0

$|\hat{b}| > c$ 拒绝 H_0

可以证明在水平α 下，如果我们假定误差服从正态分布，则c 可取为

$$c = St_{\alpha/2}(n-2)\left(\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^{-1} = \hat{\sigma}\, t_{\alpha/2}(n-2)/\sqrt{S_{xx}}$$

这儿

$$S = \sqrt{\frac{1}{n-2}RSS} = \hat{\sigma}$$

以下我们都是在误差服从正态分布的假定下进行的。在SPSS等软件包上也给出检验的p 值：

$$p = \mathrm{P}(|\, t_{n-2}\,| > \hat{b})$$

由于 $r = b\sqrt{S_{xx}/S_{yy}}$ 故对 r 的检验问题：

$$\mathrm{H}_0 \quad : \quad r = 0 \leftrightarrow \quad \mathrm{H}_1 : \quad r \neq 0$$

可以化为对b的检验问题 。经过推导可以证明上述检验可以用t-分布来检验，其中

$$t = \frac{r}{\sqrt{n-2}(1-r^2)}$$

当 $|t| > t_{\alpha/2}(n-2)$ 时拒绝 H_0（即 $r \neq 0$ ）。

2. 区间估计：

（1） a，b 的区间估计

在置信水平 $(1-\alpha)\times 100\%$ 下回归方程中的截a和回归系数b的区间估计为

$$b \in \hat{b} \pm St_{\alpha/2}(n-2)/\sqrt{S_{xx}}$$

$$a \in \hat{a} \pm St_{\alpha/2}(n-2)\left[n^{-1}+\bar{x}^2/S_{xx}\right]^{1/2}$$

（2）回归函数 $a+bx$ 的区间估计

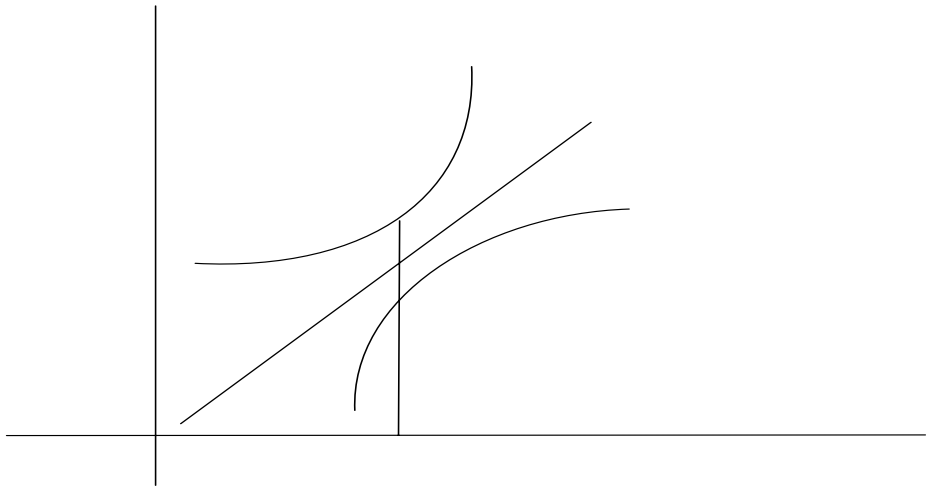
在给出回归方程后，经常会遇到自变量x在一个新的取值下对应的因变量y取值的点估计问题和因变量y平均值的区间估计问题。

关于点估计，显然我们可以用 $\hat{a}+\hat{b}x$ 作为y取值的一个点估计，即 $\hat{y}=\hat{a}+\hat{b}x$ 。关于因变量y的平均值在水平α下的区间估计,可以证明由下式给出：

$$a+bx\in\hat{a}+\hat{b}x\pm St_{\alpha/2}(n-2)\left[n^{-1}+(x-\bar{x})^2/S_{xx}\right]^{1/2}$$

其中s为误差标准差 σ 的估计。

由上式知区间的长度（精度）与所处位置有关，当 $x=\bar{x}$ 时。区间长度最小，即精度最高。当$x-\bar{x}$大时，区间长度增加，精度变差。故在试验点列中心（重心）$\bar{x}$ 附近，对回归函数 的估计更精确些。因此对 $\bar{x}$ 附近处的x 对应的y作预测，精度要高一点。（见下面图）



例6. 考虑例4中的b检验以及a和b的区间估计 ($\alpha = 0.05$)

解： 在例5中 $n=10$ $t_{\alpha/2}(n-2) = t_{0.025}(8) = 2.306$

$$S_{xx} = 9 \times 349.7046^2 = 1100640$$

$$S_{yy} = 9 \times 164.4148^2 = 243290$$

$$r = 0.9530$$

$$RSS = (1-r^2)S_{yy} = (1-0.9082) \times 243290 = 22339.4080$$

$$S = \sqrt{RSS/(10-2)} = \sqrt{22339.408/8} = 52.8434$$

$$c = St_{\alpha/2}(n-2)/\sqrt{S_{xx}} = 52.8434 \times 2.306/1049.1139 = 0.1162$$

$\because \hat{b} = 0.448 > 0.1162$, 因此拒绝原假设 H_0 ： $b=0$

即x与y有关系。

b 的区间估计为

$$\hat{b} \pm St_{\alpha/2}(n-2)/\sqrt{S_{xx}} = 0.4480 \pm 0.1162 = (0.3318, 0.5642)$$

这就是说家庭月收入每多增加一元，平均说来要多消费0.33到0.56元之间。(这个区间太长，因为样本大小n=10太小)。

如要估计月收入为750元家庭的平均月支出，由

$$y = 296.47 + 0.448x$$

代入 $x = 750$ 得 $y = 296.47 + 0.448 \times 750 = 632.47$ (点估计)

这类家庭平均月支出的区间估计为

$$\hat{a} + \hat{b}x \pm St_{\alpha/2}(n-2)\left[n^{-1} + (x-\bar{x})^2 / S_{xx}\right]^{1/2}$$

$$= 632.47 \pm 52.8434 \times 2.306 \times \left[10^{-1} + (750-836)^2 / 1100640\right]^{1/2}$$

$$= 632.47 \pm 39.81 = (592.66,\ 672.28)$$

同样对“身高1.7米的人，平均体重在什么范围内？”“每亩施肥50斤，平均亩产在什么范围内？”等问题，也可以在得到样本后做出回答。

（3） 因变量x的值在 $x = x_0$ 处的预测.

现在的问题是，如果知道某个家庭月收入为750元，问该家庭的月支出是多少? (注意，这儿不是平均值)。可以证明对 $y = a + bx$ 的预测区间 (置信水平为 $1 - \alpha$) 为

$$\hat{a}+\hat{b}x \pm St_{n-2}(\alpha/2)\left[1+n^{-1}+(x-\bar{x})^2/S_{xx}\right]^{1/2} \qquad (9)$$

例7. （续例4） 如果随机抽取一个家庭,该家庭的月收入为750元。预测该家庭当月的月支出 (注意,不是收入为750元这类家庭的平均支出)($\alpha = 0.05$)

解：由上面(9)式

$$St_{\alpha/2}(n-2)[1+n^{-1}+(x-\bar{x})^2/S_{xx}]^{1/2}$$

$$=\ 52.8434\times 2306\times\sqrt{1+10^{-1}+(750-836)^2/1100640}=128.19$$

故 y 的预测区间为 $(632.47\pm 128.19)=(504.28,\ 760.66)$

(注意，这与在 $x=750$ 时y平均值的估计不相同，关键是由 $y=\hat{a}+\hat{b}x+e$ 知与回归方程相差一个误差 e 。

§5 另一回归直线

对散点图，我们把x看作自变量，y看作应变量，由此得回归直线 $y=\hat{a}+\bar{b}x$ 。在有些场合，x和y也说不清谁是因，谁是果，因此也可以通过把y 看作自变量，x看作应变量来作回归直线，即

$$\hat{x}=c+my$$

由前面推理(把x, y位置互换)，得

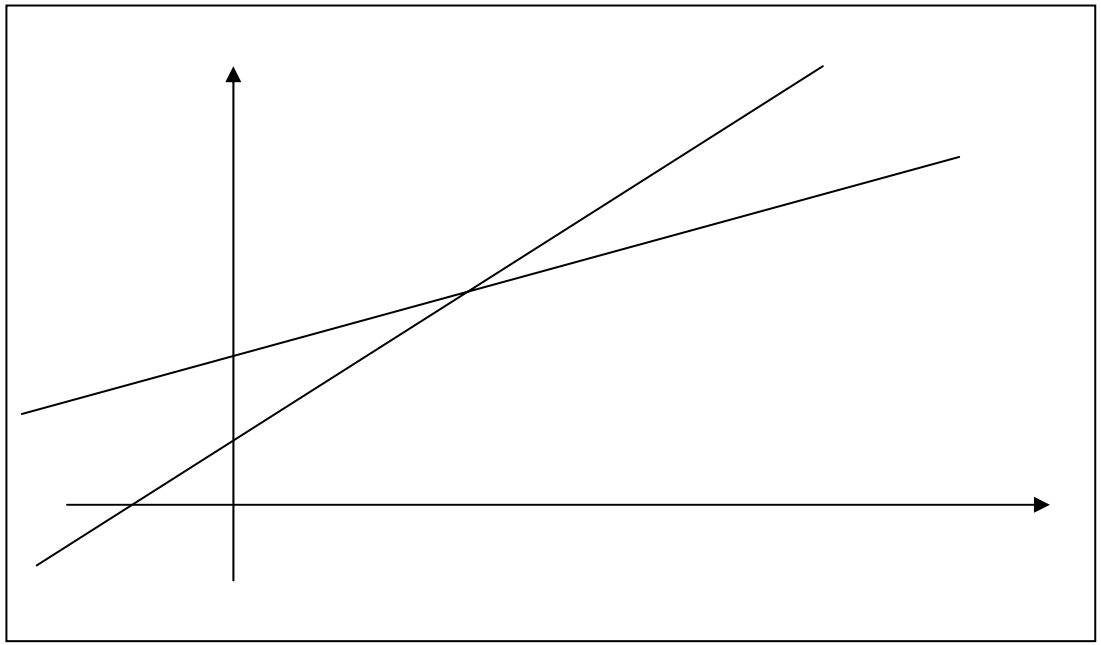
$$\hat{m}=\frac{n\sum x_i y_i-\sum x\sum y_i}{n\sum y_i^{\,2}-(\sum y_i)^2}$$

$$\hat{c}=\bar{x}-\hat{m}\bar{y}$$

这称为x关于y的回归直线，注意到该直线也通过 $(\bar{x},\bar{y})$，点 但在大多数场合下，与 $\hat{y}=\hat{a}+\hat{b}x$ 直线并不重合。

例8，考虑一产品的广告费x和销售量y之间的关系。数据见下表

x	10	20	30	40	50	60	70	80	90	100
y	30	25	61	57	60	64	55	64	85	75



由此可算得

$$y = 29.667 + 0.50788x$$

$$x = 1.40397y - 25.869$$

§6 可化为线性函数的非线性关系

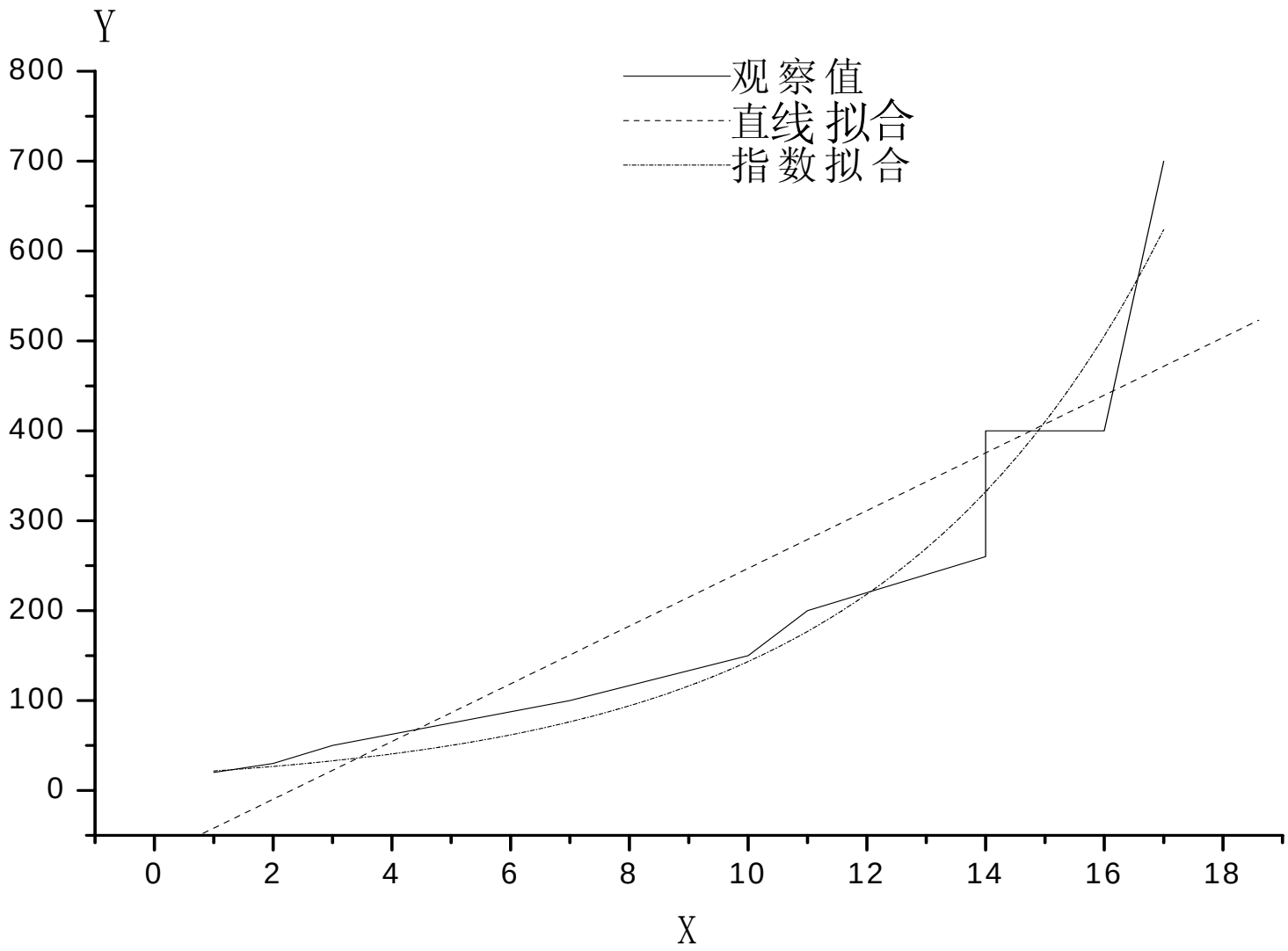
变量x和y之间的关系仅有一部分能用线性关系来描述，大量的是非线性的相关关系，但在非线性关系中，有一部分可以通过变量的替换化为线性回归函数来做。

例9. 设 x, y 有如下数据

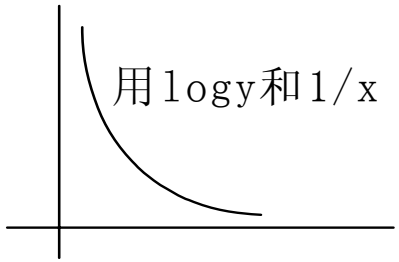
x	1	2	3	7	10	11	14	14	16	17
y	20	30	50	100	150	200	260	400	400	700
$\log_{10} y$	1.30	1.48	1.70	2.00	2.18	2.30	2.42	2.60	2.60	2.85

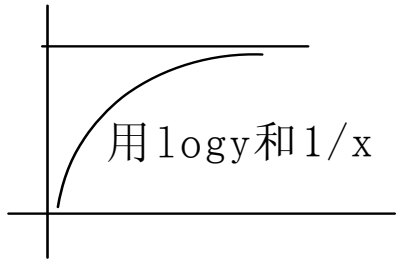
由散点图，x与y关系明显不是直线关系，而象指数关系，对y取常用对数后 $(x, \log_{10} y)$ 散点图明显在一条直线附近。由LS估计可得

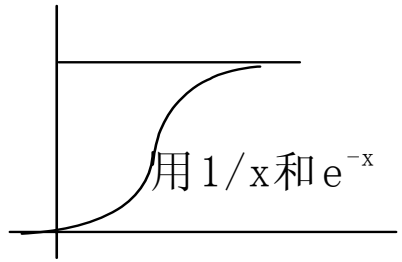
$$\hat{\log}_{10} y = 1.33 + 0.0854x$$

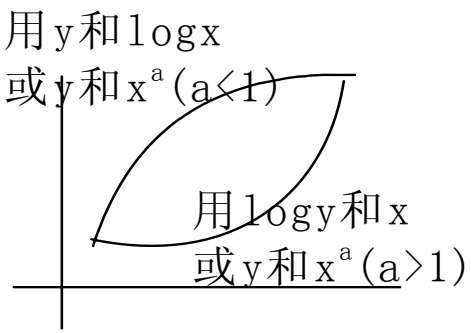

Y
800
700
600
500
400
300
200
100
0
观察值
直线拟合
指数拟合
0
2
4
6
8
10
12
14
16
18
X

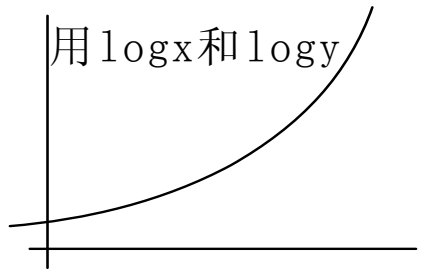
容易算得x与 $\log_{10} y$ 的相关系数为 0.9866。其他一些常见的散点图可作相应变换，如

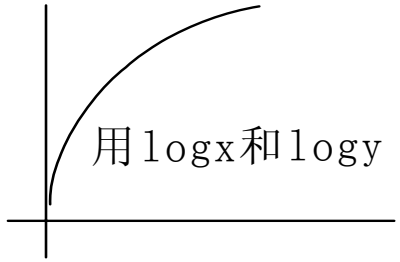


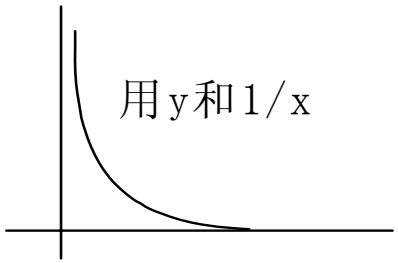


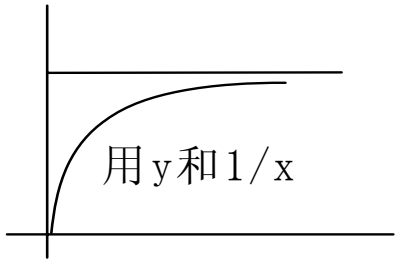








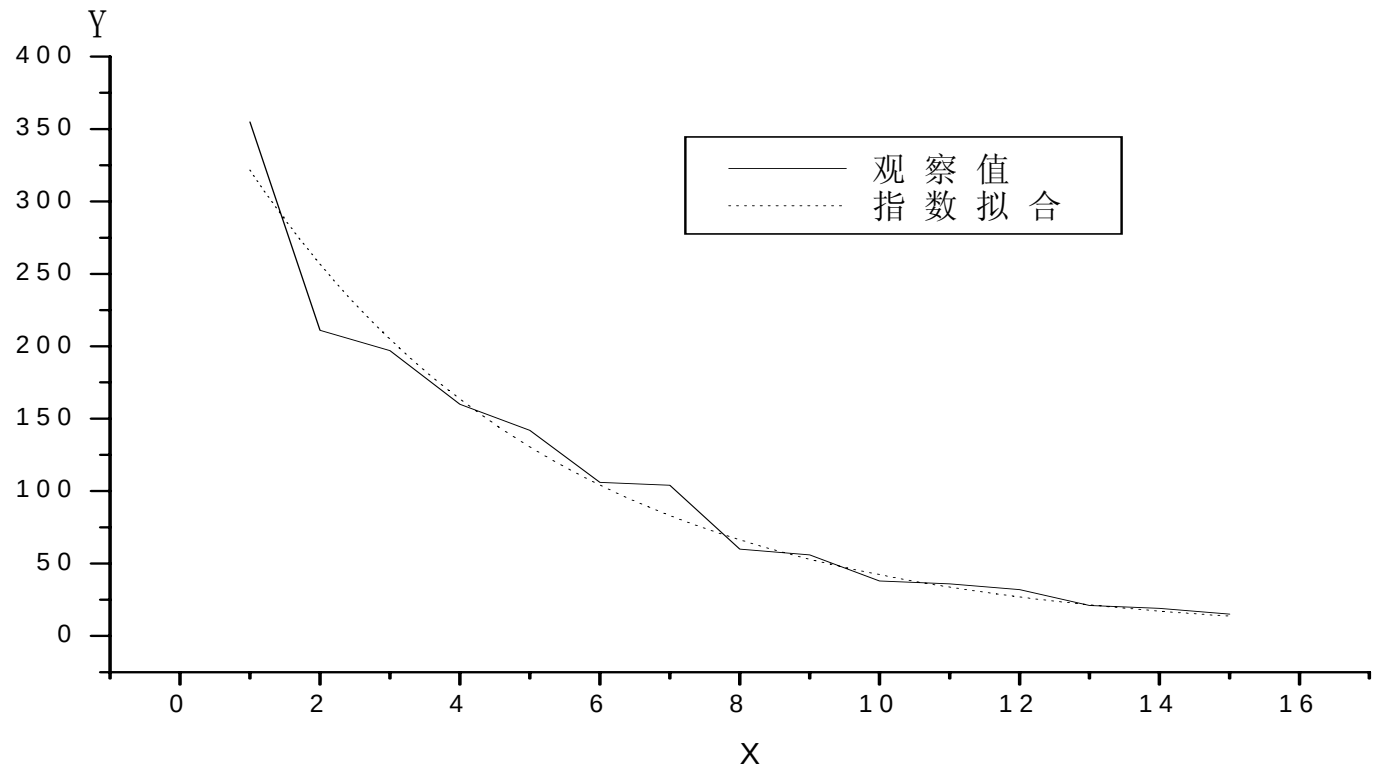




例10. 为检验X射线的杀菌作用，用200千伏的X射线来照射细菌，每次照射6分钟，记照射次数为t，共照射15次，每次照射后所剩细菌数如下:

y	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
t	355	211	197	160	142	106	104	60	56	38	36	32	21	19	15

其散点图如下：



由图看出t和y可以看作 $y=\alpha e^{\beta t}$ 或 t 与 $\log y$ 的关系，由此求出

$$\ln \hat{y} = 5.97316 - 0.21843t$$

$$\hat{y} = e^{5.97316}\, e^{-21843\ t} = 3.9275\ e^{-0.2184} \qquad (10)$$

$(t, \ln y)$ 的相关系数为-0.9942，这表明两者高度线性相关。由图可知，也可用y和 t^{α} 拟合，可得（等价于 $(\ln t, \ln y)$）

$$\ln y = 6.415 - 1.177 \ln t$$

即 $$\hat{y} = 610.941 t^{-1.177} \qquad (11)$$

$\ln t$ 和 $\ln y$ 的相关系数为 -0.9365

这儿有两个模型，哪个更好？ 由前面分析知道看哪个残差平方和小即可。记

$$Q=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \qquad R^2=1-\frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$$

这儿 R^2 称为相关指数（注意在线性回归时 R^2 即为 r^2（可决系数））。 Q的值小等价于 R^2 的值大。

在模型（10）下

$$Q_1=415.315 \quad \hat{\sigma}=S_1=\sqrt{Q_1/(n-2)}=17.575$$

$$R_1^2=1-Q_1/S_{yy}=0.969$$

在模型（11）下， $Q_2=77745.956$

$$\hat{\sigma}=S_2=\sqrt{\frac{Q_2}{n-2}}=77.333 \qquad R_2^2=1-\frac{Q_2}{S_{yy}}=0.397$$

因此模型（10）好一点。

Generalized linear model（GLM）

$$\mathrm{E}(\mathbf{Y}) = \boldsymbol{\mu} = g^{-1}(\mathbf{X}\boldsymbol{\beta})$$

Distribution		
Normal	,indentity	$\mu = \mathbf{X}\boldsymbol{\beta}$
Exponential(Gamma)	,inverse	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1}$
Inverse Gaussian	,inverse squared	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1/2}$
Poisson	,log	$\mu = \exp(\mathbf{X}\boldsymbol{\beta})$
Biomial	,logit	$\mu = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})} = \frac{1}{1+\exp(-\mathbf{X}\boldsymbol{\beta})}$
Biomial	,probit	$\Pr(Y = 1 \mid X) = \Phi(X'\beta),$

glm(formula, family = gaussian, data, weights, subset, na.action, start = NULL, etastart, mustart, offset, control = list(...), model = TRUE, method = "glm.fit", x = FALSE, y = TRUE, contrasts = NULL, ...)

Family:

binomial(link = "logit")

binomial(link = “probit")

gaussian(link = "identity")

Gamma(link = "inverse")

inverse.gaussian(link = "1/mu^2")

poisson(link = "log")

quasi(link = "identity", variance = "constant")

quasibinomial(link = "logit")

quasipoisson(link = "log")

```
counts <- c(18,17,15,20,10,20,25,13,12)
outcome <- gl(3,1,9)
treatment <- gl(3,3)
print(d.AD <- data.frame(treatment,
outcome, counts))
glm.D93 <- glm(counts ~ outcome +
treatment, family=poisson())
anova(glm.D93)
summary(glm.D93)
```