文科概率统计

3.9 线性回归分析

3.9.1 回归分析的基本思想

在客观世界中，我们经常会遇到一些相互依赖、相互制约的变量．它们之间有着一定的关系，这种关系在量上主要有两种类型．一类是确定性关系，也就是我们所熟悉的函数关系，其特点就是给定自变量的值，能确定因变量的对应值．另一类是非确定性关系，其特点是给定自变量的值，不能确定因变量的值．例如，身高和体重之间的关系，一般来说，人高一些，体重大一些，但同样高度的人体重往往不尽相同；又如在气候、土质、水利、种子和栽培技术等条件基本相同时，水稻亩产量 y 与施肥量 x 有密切关系，但是施肥量相同，亩产量不一定相同等．这种既存在着密切的关系，但又不能由一个（或一组）变量的值来确定另一个变量的值，在数理统计中，我们把这类变量之间的非确定性的关系称为**统计关系或相关关系**．回归分析就是研究这类关系的统计方法．

3.9.2 一元线性回归的数学模型

我们先从一个例子入手.

例 3.9.1 某种合成纤维的强度与其拉伸倍数之间有一定关系,下表是实测 24 个纤维样品的强度 y 与相应的拉伸倍数 x 的数据记录. 试求出它们之间的关系.

编　号	拉伸倍数 x	强度 y	编　号	拉伸倍数 x	强度 y
1	1.9	1.4	13	5.0	5.5
2	2.0	1.3	14	5.2	5.0
3	2.1	1.8	15	6.0	5.5
4	2.5	2.5	16	6.3	6.4
5	2.7	2.8	17	6.5	6.0
6	2.7	2.5	18	7.1	5.3
7	3.5	3.0	19	8.0	6.5
8	3.5	2.7	20	8.0	7.0
9	4.0	4.0	21	8.9	8.5
10	4.0	3.5	22	9.0	8.0
11	4.5	4.2	23	9.5	8.1
12	4.6	3.5	24	10.0	8.1

解　从表中可以看出，y 有随着 x 增加而增加的趋势，但它们之间的关系又是不确定的. 为了研究 x 与 y 之间的内在联系，我们以 x 为横坐标，以 y 为纵坐标，在直角坐标系中将表中的 24 对数据 (x_i, y_i)，$(i = 1, 2, \cdots, 24)$ 描成图 3.53，在回归分析中，这种图称为**散点图**. 散点图有助于我们粗略地了解两个变量之间大致上存在怎样的相关关系.

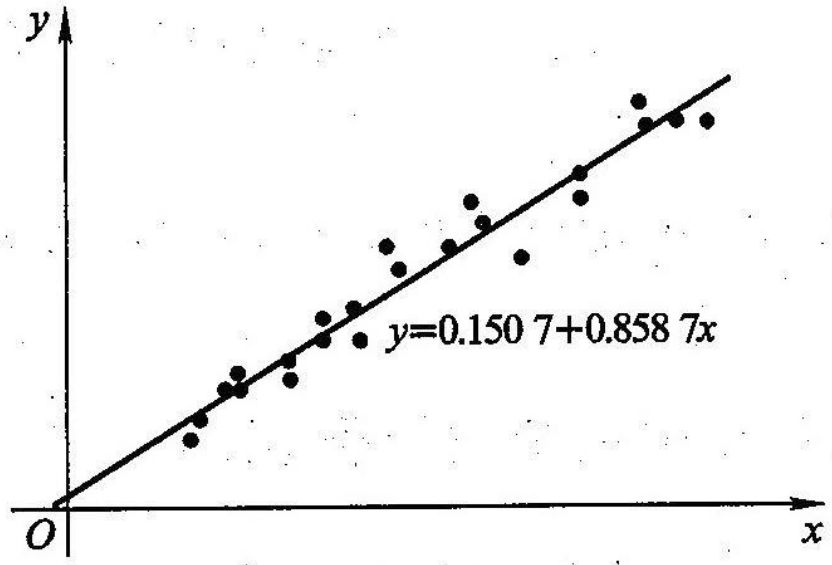


图 3.53

如图 3.53 所示，这些点大致分布在一条直线的附近. 变量 x 和 y 之间的关系基本上可看做是线性的. 但这些点与直线还有一定的偏离，这是因为除了因素 x 以外，还有许多其他随机因素在影响着 y，使 y 与直线产生了误差 ε. 因此，y 与 x 应满足下列关系：

$$y = a + bx + \varepsilon,$$

3.9.3一元线性回归中未知参数的最小二乘估计

求未知参数 a 与 b 的估计值 $\hat{a}$ 和 $\hat{b}$，而使回归直线方程 $\hat{y}=\hat{a}+\hat{b}x$ 与所有的观测点 $(x_i,y_i)(i=1,2,\cdots,n)$ 拟合得最好.

对任一给定的 x_i，y_i 的估计值为

$$\hat{y}_i=a+bx_i \quad (i=1,2,\cdots,n),$$

这些回归值同实际观测值 y_i 之间的离差（或随机误差）为

$$\varepsilon_i=y_i-\hat{y}_i=y_i-a-bx_i.$$

于是，离差平方和

$$Q=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}(y_i-(a+bx_i))^2$$

定量地描述了直线 $\hat{y}=a+bx$ 与所有散点 $(x_i,y_i)(i=1,2,\cdots,n)$ 的拟合程度. Q 的值随着 a 和 b 的不同而变化，它是 a 和 b 的二元函数，要找一条与这 n 个散点拟合最好的直线，就是找出使得 Q 达到最小值的 $\hat{a}$ 和 $\hat{b}$，即 $Q(\hat{a},\hat{b})=\min\limits_{a,b}Q(a,b)$（图 3.54）. 我们可以利用微积分中的极值求法求得 a,b 的估计值 $\hat{a}$ 和 $\hat{b}$.

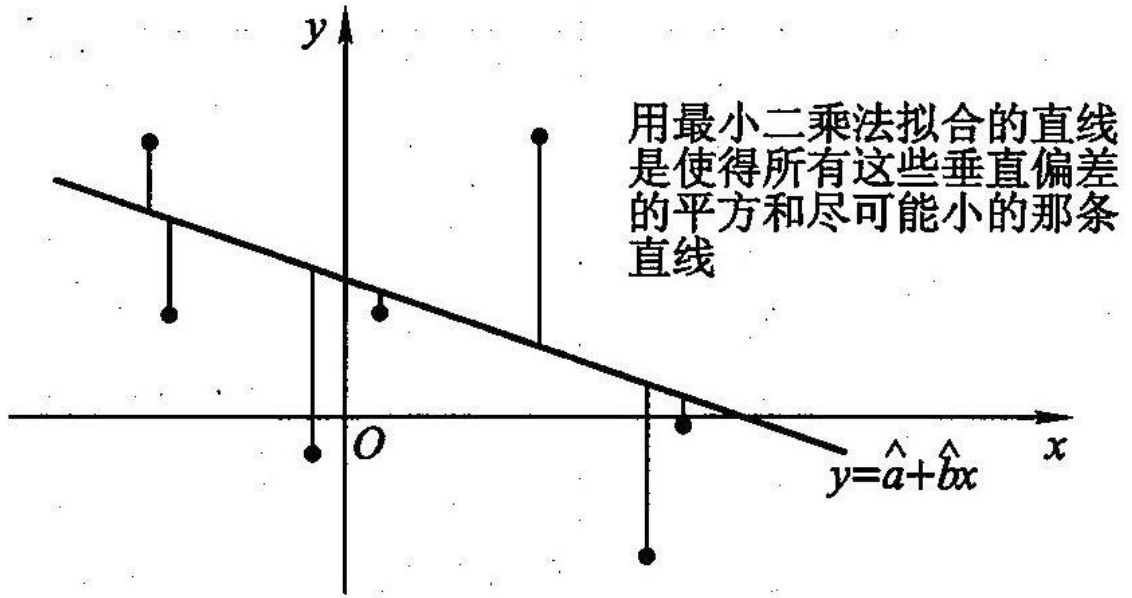


图 3.54

$$\begin{cases} \hat{b} = \dfrac{L_{xy}}{L_{xx}}, \\ \hat{a} = \bar{y} - \hat{b}\bar{x}, \end{cases}$$

其中 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$，$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$，

$$L_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2,$$

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

另记 $L_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$.

可以证明，$\hat{a}$、$\hat{b}$确能使离差平方和 Q 达到最小. 用这种方法求出的估计值 $\hat{a}$、$\hat{b}$称为 a、b 的**最小二乘估计值**.

例 3.9.2 计算例 3.9.1 中的回归直线方程.

回归直线方程的计算步骤(Ⅰ)

编　　号	x	y	x^2	y^2	xy
1	1.9	1.4	3.61	1.96	2.66
2	2.0	1.3	4.00	1.69	2.60
3	2.1	1.8	4.41	3.24	3.78
4	2.5	2.5	6.25	6.25	6.25
5	2.7	2.8	7.29	7.84	7.56
6	2.7	2.5	7.29	6.25	6.75
7	3.5	3.0	12.25	9.00	10.50
8	3.5	2.7	12.25	7.29	9.45
9	4.0	4.0	16.00	16.00	16.00
10	4.0	3.5	16.00	12.25	14.00
11	4.5	4.2	20.25	17.64	18.90
12	4.6	3.5	21.16	12.25	16.10

13	5.0	5.5	25.00	30.25	27.5
14	5.2	5.0	27.04	25.00	26.00
15	6.0	5.5	36.00	30.25	33.00
16	6.3	6.4	39.69	40.96	40.32
17	6.5	6.0	42.25	36.00	39.00
18	7.1	5.3	50.41	28.09	37.63
19	8.0	6.5	64.00	42.25	52.00
20	8.0	7.0	64.00	49.00	56.00
21	8.9	8.5	79.21	72.25	75.65
22	9.0	8.0	81.00	64.00	72.00
23	9.5	8.1	90.25	65.61	76.95
24	10.0	8.1	100.0	65.61	81.00
$\sum$	127.5	113.1	829.61	650.93	731.6

回归直线方程的计算步骤(Ⅱ)

$$\sum_{i=1}^{24} x_i = 127.5, \sum_{i=1}^{24} y_i = 113.1, n = 24,$$

$$\bar{x} = 5.312\,5, \quad \bar{y} = 4.712\,5,$$

$$\sum_{i=1}^{24} x_i^2 = 829.61, \sum_{i=1}^{24} y_i^2 = 650.93, \sum_{i=1}^{24} x_i y_i = 731.6,$$

$$L_{xx} = \sum_{i=1}^{24} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{24} x_i\right)^2 = 152.266\,3,$$

$$L_{xy} = \sum_{i=1}^{24} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{24} x_i\right)\left(\sum_{i=1}^{24} y_i\right) = 130.756\,3,$$

$$L_{yy} = \sum_{i=1}^{24} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{24} y_i\right)^2 = 117.946\,3,$$

$$\hat{b} = \frac{L_{xy}}{L_{xx}} = \frac{130.756\ 3}{152.266\ 3} = 0.858\ 7,$$

$$\hat{a} = 4.712\ 5 - 0.858\ 7 \times 5.312\ 5 = 0.150\ 7.$$

所求回归直线方程为

$$\hat{y} = 0.150\ 7 + 0.858\ 7x.$$

注 在计算回归直线方程时，并不需要 L_{yy} 的值，但在进一步分析中经常要用到，因此顺便计算出来.

3.9.4 一元线性回归效果的显著性检验（相关系数检验法）

由一元线性回归的数学模型可知，一元线性回归的数学模型是

$$y = a + bx + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2),$$

若在 $y = a + bx + \varepsilon$ 中 $b = 0$，说明 x 的变化对 y 没有影响，这时回归方程 $\hat{y} = \hat{a} + \hat{b}x$ 就不能近似地描述变量 x 与 y 之间的关系. 因此为了判断 x 与 y 之间是否存在线性关系，只需检验假设

$$H_0 : b = 0.$$

此问题也称为**线性回归方程的显著性检验问题**.

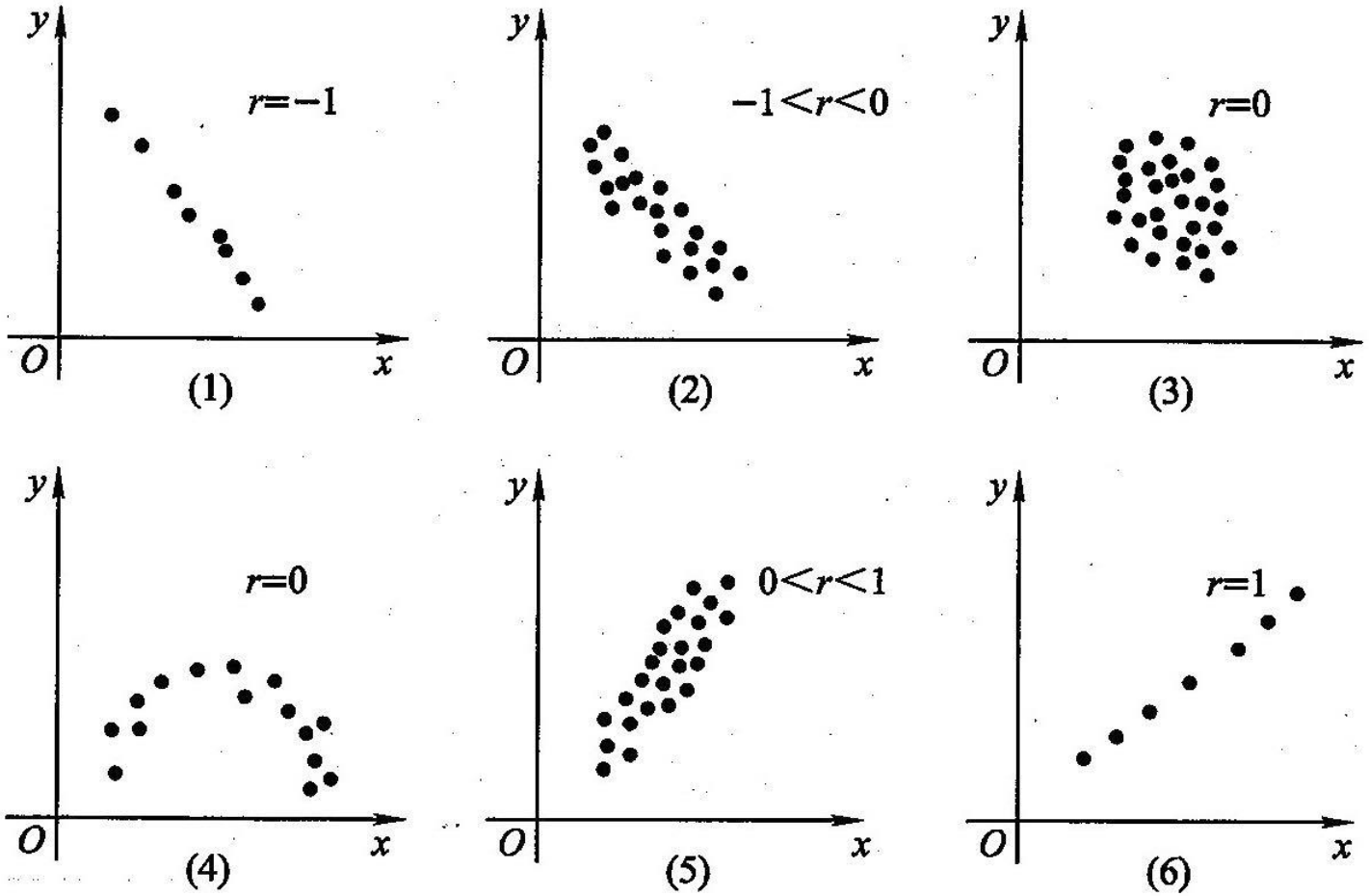
根据观测数据 (x_i, y_i) $(i = 1, 2, \cdots, n)$ 作出拒绝或接受原假设 $b = 0$ 的判断. 拒绝原假设才能确认线性回归模型是合理的，接受原假设表示不能认为 x、y 之间有线性相关关系.

令 $r = \dfrac{L_{xy}}{\sqrt{L_{xx} L_{yy}}}$

此统计量称为相关系数.

$0 \leqslant |r| \leqslant 1$. $|r|$越接近 1,回归方程对样本数据的拟合程度越好;反之,$|r|$越接近 0,回归方程对样本数据的拟合程度越差.

下面利用散点图具体说明,当 r 取各种不同数值时,散点分布的情形



当 $0<|r|<1$ 时，x 与 y 线性相关. 但只有当 r 的绝对值大到一定程度时，才能认为 x 与 y 线性关系密切. 此时，我们认为相关系数是显著的，所求的回归直线方程才有意义，否则无意义. $|r|$究竟大到什么程度时，才算 x 与 y 的线性关系为密切呢？

对于给定的显著性水平 α，查相关系数显著性检验表，可得临界值 $r_\alpha(n-2)$，使得

$$P(|r|>r_\alpha(n-2))=\alpha.$$

因此其拒绝域是 $W=\{|r|>r_\alpha(n-2)\}$.

由样本观测值计算统计量 r 的观测值 r_0，若 $|r_0|\geqslant r_\alpha(n-2)$，则应拒绝 H_0，即 x 与 y 之间线性关系显著；否则认为 x 与 y 之间的线性关系不显著或根本不存在线性关系，回归方程没有实用价值. 这种检验方法称为**相关系数检验法**.

例 3.9.3 试用相关系数检验法检验例 3.9.2 中回归直线方程的效果($\alpha=0.05$).

解 根据题意,要检验的假设为

$$H_0: b=0, \quad H_1: b\neq 0.$$

例 3.9.2 的回归直线方程为

$$\hat{y}=0.150\ 7+0.858\ 7x.$$

又 $n=24, \alpha=0.05,$

$$L_{xx}=152.266\ 3, L_{xy}=130.756\ 3, L_{yy}=117.946\ 3.$$

查相关系数显著性检验表得

$$r_\alpha(n-2)=r_{0.05}(22)=0.404$$

而

$$|r_0|=\left|\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}\right|=\left|\frac{130.756\ 3}{\sqrt{152.266\ 3\times 117.946\ 3}}\right|=0.975\ 7.$$

显然

$$|r_0|=0.975\ 7>r_{0.05}(22).$$

所以拒绝 H_0,接受 H_1,即 x 与 y 之间的线性关系是显著的.

例 3.9.4 在服装标准的制定过程中，调查了许多人的身材，得到一系列的服装各部分的尺寸与身高、胸围等的关系，下表给出的是一组女青年身高 x 与裤长 y 的数据（单位：cm）.

（1）求裤长 y 对身高 x 的回归方程；

（2）在显著性水平 $\alpha=0.01$ 下检验回归方程；

i	x	y	i	x	y	i	x	y
1	168	107	11	158	100	21	156	99
2	162	103	12	156	99	22	164	107
3	160	103	13	165	105	23	168	108
4	160	102	14	158	101	24	165	106
5	156	100	15	166	105	25	162	103
6	157	100	16	162	105	26	158	101
7	162	102	17	150	97	27	157	101
8	159	101	18	152	98	28	172	110
9	168	107	19	156	101	29	147	95
10	159	100	20	159	103	30	155	99

解 （1）根据已知数据得：

回归方程计算步骤（Ⅰ）

i	x	y	x^2	y^2	xy
1	168	107	28 224	11 449	17 976
2	162	103	26 244	10 609	16 686
3	160	103	25 600	10 609	16 480
4	160	102	25 600	10 404	16 320
5	156	100	24 336	10 000	15 600
6	157	100	24 649	10 000	15 700
7	162	102	26 244	10 404	16 524
8	159	101	25 281	10 201	16 059
9	168	107	28 224	11 449	17 976
10	159	100	25 281	10 000	15 900
11	158	100	24 964	10 000	15 800
12	156	99	24 336	9 801	15 444
13	165	105	27 225	11 025	17 325
14	158	101	24 964	10 201	15 958

续表

i	x	y	x^2	y^2	xy
15	166	105	27 556	11 025	17 430
16	162	105	26 244	11 025	17 010
17	150	97	22 500	9 409	14 550
18	152	98	23 104	9 604	14 896
19	156	101	24 336	10 201	15 756
20	159	103	25 281	10 609	16 377
21	156	99	24 336	9 801	15 444
22	164	107	26 896	11 449	17 548
23	168	108	28 224	11 664	18 144
24	165	106	27 225	11 236	17 490
25	162	103	26 244	10 609	16 686
26	158	101	24 964	10 201	15 958
27	157	101	24 649	10 201	15 857
28	172	110	29 584	12 100	18 920
29	147	95	21 609	9 025	13 965
30	155	99	24 025	9 801	15 345
$\sum$	4 797	3 068	767 949	314 112	491 124

故 $\sum_{i=1}^{30} x_i = 4\ 797$, $\sum_{i=1}^{30} y_i = 3\ 068$, $n = 30$

$\bar{x} = 159.9$, $\bar{y} = 102.267$,

$\sum_{i=1}^{30} x_i^2 = 767\ 949$, $\sum_{i=1}^{30} y_i^2 = 314\ 112$, $\sum_{i=1}^{30} x_i y_i = 491\ 124$.

$$L_{xx} = \sum_{i=1}^{30} x_i^2 - \frac{1}{30}\left(\sum_{i=1}^{30} x_i\right)^2 = 767\ 949 - \frac{1}{30} \times 4\ 797^2 = 908.7$$

$$L_{xy} = \sum_{i=1}^{30} x_i y_i - \frac{1}{30}\left(\sum_{i=1}^{30} x_i\right)\left(\sum_{i=1}^{30} y_i\right) = 491\ 124 - \frac{1}{30} \times 4\ 797 \times 3\ 068$$

$$= 550.8$$

$$L_{yy} = \sum_{i=1}^{30} y_i^2 - \frac{1}{30}\left(\sum_{i=1}^{30} y_i\right)^2 = 314\ 112 - \frac{1}{30} \times 3\ 068^2 = 357.867$$

$$\hat{b} = \frac{L_{xy}}{L_{xx}} = \frac{550.8}{908.7} = 0.606\ 1,$$

$\hat{a} = \bar{y} - \hat{b}\bar{x} = 102.266\ 7 - 0.606\ 1 \times 159.9 = 5.351\ 3.$

所求回归直线方程(图 3.56)为　　$\hat{y} = 5.351\ 3 + 0.606\ 1x.$

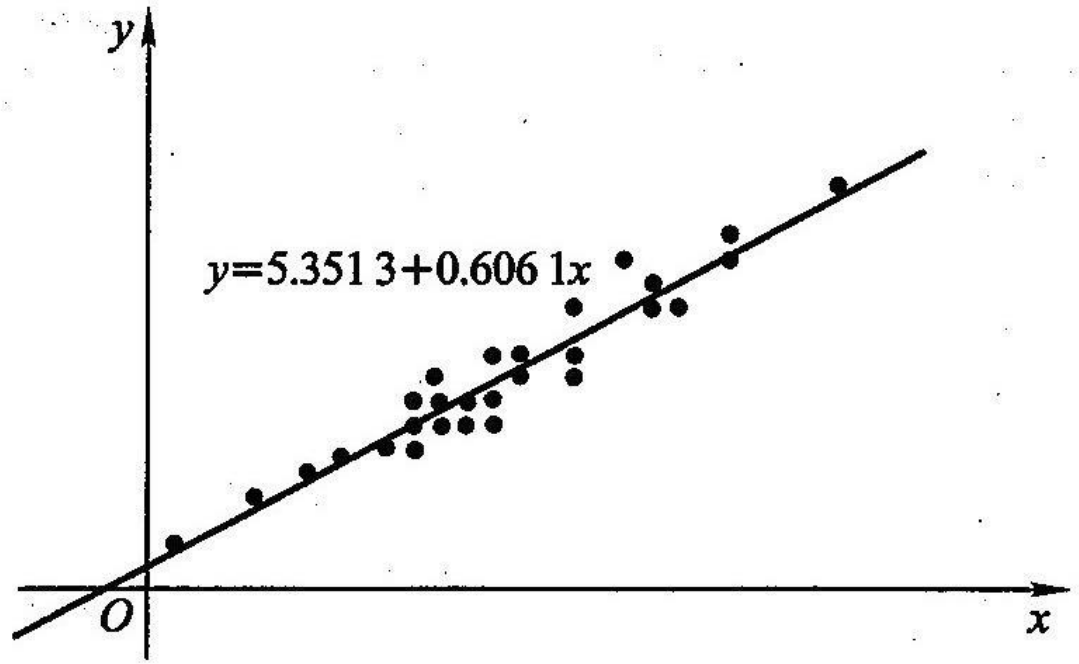


图 3.56

（2）下面我们来检验身高 x 与裤长 y 之间是否具有线性关系.

根据题意,我们作假设　$H_0: b=0.$

$$n=30, \alpha=0.01.$$

查相关系数临界值表得

$$r_\alpha(n-2)=r_{0.01}(28)=0.463$$

而根据样本观测数据计算:

$$|r_0|=\left|\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}\right|=\left|\frac{550.8}{\sqrt{908.7\times357.867}}\right|=0.965\ 9>r_\alpha(n-2)$$

故拒绝 H_0,即由相关系数检验法可知,女青年身高 x 与裤长 y 之间的线性关系是显著的,且它们之间的关系为

$$\hat{y}=5.351\ 3+0.606\ 1x.$$

以家庭为单位某商品年需求量 y 与该商品价格 x 之间的一组调查数据如下：

编号	价格（x_i）	需求量（y_i）	x_i^2	y_i^2	$x_i y_i$
1	5	1	25	1	5
2	2	3.5	4	12.25	7
3	2.5	2.4	6.25	5.76	6
4	2.8	2	7.84	4	5.6
5	3	1.5	9	2.25	4.5
6	3.5	1.2	12.25	1.44	4.2
Σ	18.8	11.6	64.34	26.7	32.3

试求需求量 y 关于价格 x 的线性回归方程，并检验回归方程的显著性（$\alpha = 0.05$）（结果保留两位小数）

相关系数显著性检验表：$r_{0.05}(4) = 0.811$、$r_{0.05}(5) = 0.754$、$r_{0.05}(6) = 0.707$

解：（1） $\sum_{i=1}^{n} x_i = 18.8, \sum_{i=1}^{n} y_i = 11.6, \sum_{i=1}^{n} x_i^2 = 64.34, \sum_{i=1}^{n} y_i^2 = 26.7.$

$$\sum_{i=1}^{n} x_i y_i = 32.3, \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \approx 3.13, \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \approx 1.93.$$

$$L_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2 = 64.34 - \frac{1}{6} \times 18.8^2 \approx 5.43,$$

$$L_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i) = 32.3 - \frac{1}{6} \times 18.8 \times 11.6 \approx -4.05,$$

$$L_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2 = 26.7 - \frac{1}{6} \times 11.6^2 \approx 4.27,$$

$$\hat{b} = \frac{L_{xy}}{L_{xx}} = \frac{-4.05}{5.43} \approx -0.75,$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 1.93 + 0.75 \times 3.13 \approx 4.28.$$

所以线性回归方程为 $\hat{y} = \hat{a} + \hat{b}x = 4.28 - 0.75x.$

（2）根据题意，做假设 $H_0 : b = 0.$

$n = 6, \alpha = 0.05, \gamma_\alpha(n-2) = \gamma_{0.05}(4) = 0.811.$

根据样本观测数据

$$|\gamma_0| = \left| \frac{L_{xy}}{\sqrt{L_{xx} L_{yy}}} \right| = \left| \frac{-4.05}{\sqrt{5.43 \times 4.27}} \right| \approx 0.84 > 0.811,$$

故拒绝 H_0，即该商品需求量与价格之间的线性关系是显著的。

它们的关系为 $\hat{y} = 4.28 - 0.75x.$

某校 10 名学生中学毕业考试成绩及刚进校时进行的能力测试成绩分别为

$y_i, x_i\ (1 \leqslant i \leqslant 10)$. 已知 $\sum_{i=1}^{10} x_i = 595$，$\sum_{i=1}^{10} y_i = 645$，$\sum_{i=1}^{10} x_i^2 = 41075$，$\sum_{i=1}^{10} y_i^2 = 45175$，$\sum_{i=1}^{10} x_i y_i = 41450$，

试求 y 对 x 的线性回归方程，并检验回归方程的显著性.（结果保留二位小数，$\alpha = 0.05$）

相关系数显著性检验表：$r_{0.05}(8) = 0.632$、$r_{0.05}(9) = 0.602$、$r_{0.05}(10) = 0.576$

解： （1） $n=10,\ \bar{x}=59.5,\ \bar{y}=64.5$

$$L_{xx}=\sum_{i=1}^{n}x_i^{\,2}-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2=41075-\frac{1}{10}\times595^2=5672.5,$$

$$L_{xy}=\sum_{i=1}^{n}x_iy_i-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)=41450-\frac{1}{10}\times595\times645=3072.5,$$

$$L_{yy}=\sum_{i=1}^{n}y_i^{\,2}-\frac{1}{n}\left(\sum_{i=1}^{n}y_i\right)^2=45175-\frac{1}{10}\times645^2=3572.5$$

$$\hat{b}=\frac{L_{xy}}{L_{xx}}\approx0.54,\quad \hat{a}=\bar{y}-\hat{b}\bar{x}=32.37$$

所以线性回归方程为 $\hat{y}=\hat{a}+\hat{b}x=32.37+0.54x.$

（2）根据题意，做假设 $H_0: b = 0.$

$$n = 10, \alpha = 0.05, \gamma_\alpha(n-2) = \gamma_{0.05}(8) = 0.632,$$

根据样本观测数据

$$|\gamma_0| = \left|\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}\right| = \left|\frac{2.385}{\sqrt{2.336 \times 5.476}}\right| \approx 0.683 > 0.632,$$

故拒绝 H_0，即 y 对 x 的线性关系是显著的。

它们的关系为 $\hat{y} = \hat{a} + \hat{b}x = 32.37 + 0.54x.$