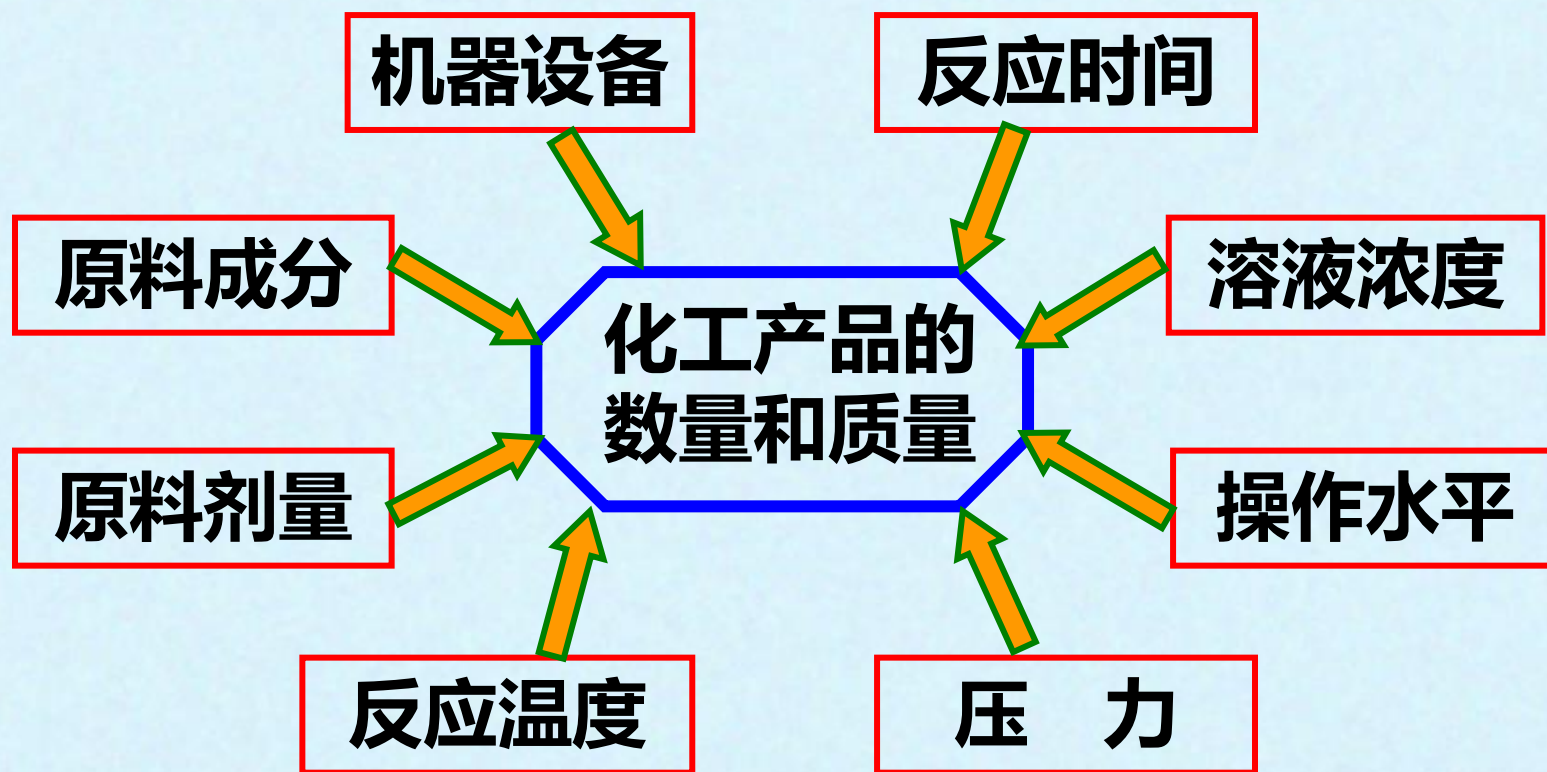


第一节 单因素试验的方差分析

- 一、单因素试验
- 二、平方和的分解
- 三、 S_A, S_E 的统计特性
- 四、假设检验问题的拒绝域
- 五、未知参数的估计
- 六、小结

一、单因素试验



方差分析——根据试验的结果进行分析,鉴别各个有关因素对试验结果影响的有效方法.

试验指标——试验中要考察的指标.

因素——影响试验指标的条件.

因素	可控因素	不可控因素
----	------	-------

水平——因素所处的状态.

单因素试验——在一项试验中只有一个因素改变.

多因素试验——在一项试验中有多个因素在改变.

例1 设有三台机器, 用来生产规格相同的铝合金薄板. 取样, 测量薄板的厚度精确至千分之一厘米. 得结果如下表所示.

表9.1 铝合金板的厚度

机器I	机器II	机器III
0.236	0.257	0.258
0.238	0.253	0.264
0.248	0.255	0.259
0.245	0.254	0.267
0.243	0.261	0.262



试验指标: 薄板的厚度 **因素:** 机器

水平: 不同的三台机器是因素的三个不同的水平.

假定除机器这一因素外, 其他条件相同,

属于**单因素试验**.

试验目的: 考察各台机器所生产的薄板的厚度有无显著的差异. 即考察机器这一因素对厚度有无显著的影响.

结论: 如果厚度有显著差异,
表明机器这一因素对厚度的影响是显著的.



例2 下表列出了随机选取的、用于计算器的四种类型的电路的响应时间（以毫秒计）

表9.2 电路的响应时间

类型I	类型II	类型III	机器IV
19 15	20 40	16 17	18
22	21	15	22
20	33	18	19
18	27	26	



试验指标：电路的响应时间

因素：电路类型

水平：四种电路类型为因素的四个不同的水平

单因素试验

试验目的：

考察电路类型这一因素对响应时间有无显著的影响.



例3 一火箭用四种燃料, 三种推进器作射程试验. 每种燃料与每种推进器的组合各发射火箭两次, 射程如下.

表9.3 火箭的射程

推进器(B)		B_1	B_2	B_3
燃料(A)	A_1	58.2	56.2	65.3
		52.6	41.2	60.8
	A_2	49.1	54.1	51.6
		42.8	50.5	48.4
	A_3	60.1	70.9	39.2
		58.3	73.2	40.7
	A_4	75.8	58.2	48.7
		71.5	51.0	41.4



试验指标: 射程

因素: 推进器和燃料

水平: 推进器有3个, 燃料有4个

双因素试验

试验目的:

考察推进器和燃料两因素对射程有无显著的影响.



关于例1的讨论

设总体均值分别为 μ_1, μ_2, μ_3 .

检验假设 $H_0: \mu_1 = \mu_2 = \mu_3,$

$H_1: \mu_1, \mu_2, \mu_3$ 不全相等.

进一步假设各总体均为正态变量, 且各总体的方差相等, 但参数均未知.

问题 检验同方差的多个正态总体均值是否相等.

解决方法 **方差分析法**(一种统计方法)



数学模型

设因素A有 s 个水平 $A_1, A_2, \dots, A_s$, 在水平 $A_j (j = 1, 2, \dots, s)$ 下, 进行 $n_j (n_j \geq 2)$ 次独立试验, 得到如下表的结果.



表9.4

观察结果 \ 水平	A_1	A_2	$\dots$	A_s
	X_{11}	X_{12}	$\dots$	X_{1s}
	X_{21}	X_{22}	$\dots$	X_{2s}
	$\vdots$	$\vdots$		$\vdots$
	$X_{n_1 1}$	$X_{n_2 2}$	$\dots$	$X_{n_s s}$
样本总和	$T_{\bullet 1}$	$T_{\bullet 2}$	$\dots$	$T_{\bullet s}$
样本均值	$\bar{X}_{\bullet 1}$	$\bar{X}_{\bullet 2}$	$\dots$	$\bar{X}_{\bullet s}$
总体均值	μ_1	μ_2	$\dots$	μ_s



假设

1.各个水平 $A_j (j = 1, 2, \dots, s)$ 下的样本

$$X_{1j}, X_{2j}, \dots, X_{n_j j}$$

来自具有相同方差 σ^2 , 均值分别为 $\mu_j (j = 1, 2, \dots, s)$ 的正态总体 $N(\mu_j, \sigma^2)$, μ_j 与 σ^2 均未知;

2.不同水平 A_j 下的样本之间相互独立.



因为 $X_{ij} \sim N(\mu_j, \sigma^2)$, 所以 $X_{ij} - \mu_j \sim N(0, \sigma^2)$.

记 $X_{ij} - \mu_j = \varepsilon_{ij}$ 表示随机误差,

那么 X_{ij} 可写成

$$X_{ij} = \mu_j + \varepsilon_{ij},$$

$\varepsilon_{ij} \sim N(0, \sigma^2)$, 各 ε_{ij} 独立,

$i = 1, 2, \dots, n_j, j = 1, 2, \dots, s,$

μ_j 与 σ^2 均未知.

**单因素试验方差分析的
数学模型**



需要解决的问题

1. 检验假设 $H_0 : \mu_1 = \mu_2 = \cdots = \mu_s,$
 $H_1 : \mu_1, \mu_2, \cdots, \mu_s$ 不全相等.
2. 估计未知参数 $\mu_1, \mu_2, \cdots, \mu_s, \sigma^2.$



数学模型的等价形式

记 $n = \sum_{j=1}^s n_j$,

$$\mu = \frac{1}{n} \sum_{j=1}^s n_j \mu_j.$$

总平均

水平 A_j 的效应, 表示水平 A_j 下的总体平均值与总平均的差异.

$$\delta_j = \mu_j - \mu, j = 1, 2, \dots, s.$$

$$n_1 \delta_1 + n_2 \delta_2 + \dots + n_s \delta_s = 0.$$



原数学模型

$$\left. \begin{aligned} X_{ij} &= \mu_j + \varepsilon_{ij}, \\ \varepsilon_{ij} &\sim N(0, \sigma^2), \text{ 各 } \varepsilon_{ij} \text{ 独立}, \\ i &= 1, 2, \dots, n_j, j = 1, 2, \dots, s, \\ \mu_j &\text{ 与 } \sigma^2 \text{ 均未知.} \end{aligned} \right\}$$

改写为

$$\left. \begin{aligned} X_{ij} &= \mu + \delta_j + \varepsilon_{ij}, \\ \varepsilon_{ij} &\sim N(0, \sigma^2), \text{ 各 } \varepsilon_{ij} \text{ 独立}, \\ i &= 1, 2, \dots, n_j, j = 1, 2, \dots, s, \\ \sum_{j=1}^s n_j \delta_j &= 0. \end{aligned} \right\}$$



原检验假设 $H_0 : \mu_1 = \mu_2 = \cdots = \mu_s,$
 $H_1 : \mu_1, \mu_2, \cdots, \mu_s$ 不全相等.

等价于检验假设

$$H_1 : \delta_1 = \delta_2 = \cdots = \delta_s = 0,$$
$$H_0 : \delta_1, \delta_2, \cdots, \delta_s \text{不全为零}.$$



二、平方和的分解

$$\bar{X} = \frac{1}{n} \sum_{j=1}^s \sum_{i=1}^{n_j} X_{ij}$$

数据的总平均

$$S_T = \sum_{j=1}^s \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2$$

总偏差平方和 (总变差)

$$\bar{X}_{\cdot j} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij}$$

水平 A_j 下的样本平均值



$$\begin{aligned} S_T &= \sum_{j=1}^s \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2 \\ &= \sum_{j=1}^s \sum_{i=1}^{n_j} [(X_{ij} - \bar{X}_{\cdot j}) + (\bar{X}_{\cdot j} - \bar{X})]^2 \\ &= \sum_{j=1}^s \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_{\cdot j})^2 + \sum_{j=1}^s \sum_{i=1}^{n_j} (\bar{X}_{\cdot j} - \bar{X})^2 + \\ &\quad + 2 \sum_{j=1}^s \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_{\cdot j})(\bar{X}_{\cdot j} - \bar{X}) \end{aligned}$$



其中 $2 \sum_{j=1}^s \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_{\cdot j})(\bar{X}_{\cdot j} - \bar{X})$

$$= 2 \sum_{j=1}^s (\bar{X}_{\cdot j} - \bar{X}) \left[\sum_{i=1}^{n_j} (X_{ij} - \bar{X}_{\cdot j}) \right]$$

$$= 2 \sum_{j=1}^s (\bar{X}_{\cdot j} - \bar{X}) \left[\sum_{i=1}^{n_j} X_{ij} - n_j \bar{X}_{\cdot j} \right]$$

$$= 0$$

于是 S_T 可分解为 $S_T = S_E + S_A$,



其中
$$S_E = \sum_{j=1}^s \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_{\cdot j})^2$$

$$\begin{aligned} S_A &= \sum_{j=1}^s \sum_{i=1}^{n_j} (\bar{X}_{\cdot j} - \bar{X})^2 = \sum_{j=1}^s n_j (\bar{X}_{\cdot j} - \bar{X})^2 \\ &= \sum_{j=1}^s n_j \bar{X}_{\cdot j}^2 - n \bar{X}^2 \end{aligned}$$

$S_T = S_E + S_A$ 称为平方和分解式。

误差平方和

效应平方和



三、 S_E , S_A 的统计特性

$$S_E = \sum_{j=1}^s \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_{\cdot j})^2$$

$$= \sum_{i=1}^{n_1} (X_{i1} - \bar{X}_{\cdot 1})^2 + \cdots + \sum_{i=1}^{n_s} (X_{is} - \bar{X}_{\cdot s})^2,$$

$\sum_{i=1}^{n_j} (X_{ij} - \bar{X}_{\cdot j})^2$ 是 $N(\mu_j, \sigma^2)$ 的样本方差的 $n_j - 1$ 倍,

$$\frac{\sum_{i=1}^{n_j} (X_{ij} - \bar{X}_{\cdot j})^2}{\sigma^2} \sim \chi^2(n_j - 1).$$



又由于各 X_{ij} 独立, 所以由 χ^2 分布的可加性知

$$\frac{S_E}{\sigma^2} \sim \chi^2 \left(\sum_{j=1}^s (n_j - 1) \right),$$

即 $\frac{S_E}{\sigma^2} \sim \chi^2(n - s),$ 其中 $n = \sum_{j=1}^s n_j.$

根据 χ^2 分布的性质可以得到,

$$S_E \text{ 的自由度为 } n - s; \quad E(S_E) = (n - s)\sigma^2.$$



$$S_A = \sum_{j=1}^s \sum_{i=1}^{n_j} (\bar{X}_{\cdot j} - \bar{X})^2 = \sum_{j=1}^s [\sqrt{n_j}(\bar{X}_{\cdot j} - \bar{X})]^2$$

因为
$$\sum_{j=1}^s \sqrt{n_j} [\sqrt{n_j}(\bar{X}_{\cdot j} - \bar{X})] = \sum_{j=1}^s n_j (\bar{X}_{\cdot j} - \bar{X})$$

$$= \sum_{j=1}^s \sum_{i=1}^{n_j} X_{ij} - n\bar{X} = 0$$

所以 S_A 的自由度为 $s-1$.

又因为 $\mu = \frac{1}{n} \sum_{j=1}^s n_j \mu_j$, $\bar{X} = \frac{1}{n} \sum_{j=1}^s \sum_{i=1}^{n_j} X_{ij}$, X_{ij} 相互独立,

所以
$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$



$$\begin{aligned} E(S_A) &= E\left[\sum_{j=1}^s n_j \bar{X}_{\cdot j}^2 - n \bar{X}^2\right] = \sum_{j=1}^s n_j E(\bar{X}_{\cdot j}^2) - n E(\bar{X}^2) \\ &= \sum_{j=1}^s n_j \left[\frac{\sigma^2}{n_j} + (\mu + \delta_j)^2 \right] - n \left[\frac{\sigma^2}{n} + \mu^2 \right] \\ &= (s-1)\sigma^2 + 2\mu \sum_{j=1}^s n_j \delta_j + n\mu^2 + \sum_{j=1}^s n_j \delta_j^2 - n\mu^2 \\ &= (s-1)\sigma^2 + \sum_{j=1}^s n_j \delta_j^2 \end{aligned}$$

S_A 与 S_E 独立, H_0 为真时, $\frac{S_A}{\sigma^2} \sim \chi^2(s-1)$.



四、假设检验问题的拒绝域

检验假设 $H_1 : \delta_1 = \delta_2 = \cdots = \delta_s = 0,$

$H_0 : \delta_1, \delta_2, \cdots, \delta_s$ 不全为零.

$$H_0 \text{为真时,} \quad E\left(\frac{S_A}{s-1}\right) = \sigma^2,$$

$$H_1 \text{为真时,} \quad \sum_{j=1}^s n_j \delta_j^2 > 0,$$

$$E\left(\frac{S_A}{s-1}\right) = \sigma^2 + \frac{1}{s-1} \sum_{j=1}^s n_j \delta_j^2 > \sigma^2.$$



不管 H_0 是否为真, $S_E/(n-s)$ 都是 σ^2 的无偏估计.

$$F = \frac{S_A/(s-1)}{S_E/(n-s)}.$$

1. 分子和分母相互独立;
2. 分母 $S_E/(N-S)$ 的数学期望始终是 σ^2 ;
3. H_0 为真时, 分子的期望为 σ^2 , H_0 不真时, 分子取值有偏大的趋势.

拒绝域的形式为 $F = \frac{S_A/(s-1)}{S_E/(n-s)} \geq k.$

其中 k 由预先给定的显著水平 α 确定.

因为 H_0 为真时,

$$\frac{S_E}{\sigma^2} \sim \chi^2(n-s), \quad \frac{S_A}{\sigma^2} \sim \chi^2(s-1),$$

$$\frac{S_A/(s-1)}{S_E/(n-s)} = \frac{S_A/\sigma^2}{(s-1)} \bigg/ \frac{S_E/\sigma^2}{(n-s)} \sim F_\alpha(s-1, n-s).$$

检验假设 $H_0 : \delta_1 = \delta_2 = \cdots = \delta_s = 0,$
 $H_1 : \delta_1, \delta_2, \cdots, \delta_s$ 不全为零.

拒绝域为 $F = \frac{S_A/(s-1)}{S_E/(n-s)} \geq F_\alpha(s-1, n-s).$



9.5 单因素试验方差分析表

方差来源	平方和	自由度	均方	F 比
因素A	S_A	$s - 1$	$\bar{S}_A = \frac{S_A}{s - 1}$	$F = \bar{S}_A / \bar{S}_E$
误差	S_E	$n - s$	$\bar{S}_E = \frac{S_E}{n - s}$	
总和	S_T	$n - 1$		

表中 $\bar{S}_A = \frac{S_A}{s - 1}$ 和 $\bar{S}_E = \frac{S_E}{n - s}$ 称为 S_A 和 S_E 的均方。



S_T 、 S_A 和 S_E 的简便计算公式：

$$\text{记 } T_{\cdot j} = \sum_{i=1}^{n_j} X_{ij}, \quad j = 1, \cdots, s, \quad T_{\cdot\cdot} = \sum_{j=1}^s \sum_{i=1}^{n_j} X_{ij},$$

$$S_T = \sum_{j=1}^s \sum_{i=1}^{n_j} X_{ij}^2 - n\bar{X}^2 = \sum_{j=1}^s \sum_{i=1}^{n_j} X_{ij}^2 - \frac{T_{\cdot\cdot}^2}{n},$$

$$S_A = \sum_{j=1}^s n_j \bar{X}_{\cdot j}^2 - n\bar{X}^2 = \sum_{j=1}^s \frac{T_{\cdot j}^2}{n_j} - \frac{T_{\cdot\cdot}^2}{n},$$

$$S_E = S_T - S_A.$$



例4 设有三台机器, 用来生产规格相同的铝合金薄板. 取样, 测量薄板的厚度精确至千分之一厘米. 得结果如下表所示. 取 ($\alpha = 0.05$) :

表9.1 铝合金板的厚度

机器I	机器II	机器III
0.236	0.257	0.258
0.238	0.253	0.264
0.248	0.255	0.259
0.245	0.254	0.267
0.243	0.261	0.262

检验假设 $H_0 : \mu_1 = \mu_2 = \mu_3$, $H_1 : \mu_1, \mu_2, \mu_3$ 不全相等



解 $s=3, n_1=n_2=n_3=5, n=15$

$$S_T = \sum_{j=1}^3 \sum_{i=1}^5 X_{ij}^2 - \frac{T_{..}^2}{15}$$

$$= 0.963912 - \frac{3.8^2}{15} = 0.00124533,$$

$$S_A = \sum_{j=1}^3 \frac{T_{.j}^2}{n_j} - \frac{T_{..}^2}{n}$$

$$= \frac{1}{5}(1.21^2 + 1.28^2 + 1.31^2) - \frac{3.8^2}{15} = 0.00105333,$$

$$S_E = S_T - S_A = 0.000192.$$



例4的方差分析表

方差来源	平方和	自由度	均方	F 比
因素	0.00105333	2	0.00052667	32.92
误差	0.000192	12	0.000016	
总和	0.00124533	14		

$F = 32.92 > F_{0.05}(2, 12) = 3.89$. 在水平0.05下拒绝 H_0
各机器生产的薄板厚度有显著差异.



在MATLAB中的求解

函数:*anova1*

格式:*p=anova1(x)*

说明:对样本 X 中的多列数据进行单因素方差分析,比较各列的均值,返回“零假设”成立的概率值,如果概率值接近于零,则零假设值得怀疑,表明各列的均值事实上是不同的.

源程序: $x=[0.236,0.238,0.248,0.245,0.243;$
 $0.257,0.253,0.255,0.254,0.261;$
 $0.258,0.264,0.259,0.267,0.262];$
 $p=anova1(x')$

程序运行结果

方差分析表

Box 图检验

帮

助



五、未知参数的估计

不论 H_0 是否为真, $\hat{\sigma}^2 = \frac{S_E}{n-s}$ 是 σ^2 的无偏估计.

$$E(\bar{X})=\mu, \quad E(\bar{X}_{\cdot j})=\mu_j, \quad j=1,2,\cdots,s$$

故 $\hat{\mu} = \bar{X}$, $\hat{\mu}_j = \bar{X}_{\cdot j}$ 分别是 μ 和 μ_j 的无偏估计.

若拒绝 H_0 , 意味着效应 $\delta_1, \delta_2, \cdots, \delta_s$ 不全为零,

$$\delta_j = \mu_j - \mu, \quad j=1,2,\cdots,s$$

于是 $\hat{\delta}_j = \bar{X}_{\cdot j} - \bar{X}$ 是 δ_j 的无偏估计.

由于 $\delta_j = \mu_j - \mu, j = 1, 2, \dots, s.$

$$\hat{\delta}_j = \bar{X}_{\cdot j} - \bar{X}$$

$$\sum_{j=1}^s n_j \hat{\delta}_j = \sum_{j=1}^s n_j \bar{X}_{\cdot j} - n \bar{X} = 0$$

$$E(\bar{X}_{\cdot j} - \bar{X}_{\cdot k}) = \mu_j - \mu_k,$$

$$D(\bar{X}_{\cdot j} - \bar{X}_{\cdot k}) = \sigma^2 \left(\frac{1}{n_j} + \frac{1}{n_k} \right),$$

$\bar{X}_{\cdot j} - \bar{X}_{\cdot k}$ 与 $\hat{\sigma}^2 = \frac{S_E}{n-s}$ 独立,



$$\begin{aligned}
 & \text{于是} \frac{(\bar{X}_{\cdot j} - \bar{X}_{\cdot k}) - (\mu_j - \mu_k)}{\sqrt{\bar{S}_E \left(\frac{1}{n_j} + \frac{1}{n_k} \right)}} \\
 &= \frac{(\bar{X}_{\cdot j} - \bar{X}_{\cdot k}) - (\mu_j - \mu_k)}{\sigma \sqrt{1/n_j + 1/n_k}} \bigg/ \sqrt{\frac{S_E}{\sigma^2} / (n-s)} \sim t(n-s).
 \end{aligned}$$

均值差 $\mu_j - \mu_k = \delta_j - \delta_k$ 的置信水平为 $1-\alpha$ 的置信区间为

$$\left(\bar{X}_{\cdot j} - \bar{X}_{\cdot k} \pm t_{\alpha/2}(n-s) \sqrt{\bar{S}_E \left(\frac{1}{n_j} + \frac{1}{n_k} \right)} \right).$$



例5 求例4中的未知参数 $\sigma^2, \mu_j, \delta_j (j=1,2,3)$ 的点估计及均值差的置信水平为0.95的置信区间.

解 $\hat{\sigma}^2 = S_E / (n - s) = 0.000016,$

$$\hat{\mu}_1 = \bar{x}_{\cdot 1} = 0.242, \quad \hat{\mu}_2 = \bar{x}_{\cdot 2} = 0.256,$$

$$\hat{\mu}_3 = \bar{x}_{\cdot 3} = 0.262. \quad \hat{\delta}_1 = \bar{x}_{\cdot 1} - \bar{x} = -0.011,$$

$$\hat{\mu} = \bar{x} = 0.253, \quad \hat{\delta}_2 = \bar{x}_{\cdot 2} - \bar{x} = 0.003,$$

$$\hat{\delta}_3 = \bar{x}_{\cdot 3} - \bar{x} = 0.009.$$

因为 $t_{0.025}(n - s) = t_{0.025}(12) = 2.1788,$



$$t_{0.025}(12) \sqrt{\bar{S}_E \left(\frac{1}{n_j} + \frac{1}{n_k} \right)} = 0.006,$$

所以 $\mu_1 - \mu_2$ 的置信水平为0.95的置信区间为

$$(0.242 - 0.256 \pm 0.006) = (-0.020, -0.008),$$

$\mu_1 - \mu_3$ 的置信水平为0.95的置信区间为

$$(0.242 - 0.262 \pm 0.006) = (-0.026, -0.014),$$

$\mu_2 - \mu_3$ 的置信水平为0.95的置信区间为

$$(0.256 - 0.262 \pm 0.006) = (-0.012, 0).$$



例6 设在例2的四种类型电路的响应时间的总体均为正态, 且各总体的方差相同, 但参数均未知. 又设各样本相互独立, 试取水平 $\alpha = 0.05$ 检验各类型电路的响应时间是否有显著差异.

解 分别以 $\mu_1, \mu_2, \mu_3, \mu_4$ 记类型 I、II、III、IV 四种电路响应时间的总体的平均值. 检验($\alpha = 0.05$)

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4,$$

$$H_1 : \mu_1, \mu_2, \mu_3, \mu_4 \text{ 不全相等.}$$



$$n=18, \quad s=4, \quad n_1=n_2=n_3=5, \quad n_4=3$$

$$S_T = \sum_{j=1}^4 \sum_{i=1}^{n_j} x_{ij}^2 - \frac{T_{..}^2}{18} = 8992 - \frac{386^2}{18} = 714.44$$

$$S_A = \sum_{j=1}^4 \frac{T_{.j}^2}{n_j} - \frac{T_{..}^2}{18}$$

$$= \left[\frac{1}{5} (94^2 + 141^2 - 92^2) + \frac{59^2}{3} \right] - \frac{386^2}{18}$$

$$= 318.98,$$



$$S_E = S_T - S_A = 395.46$$

S_T, S_A, S_E 的自由度依次为17,13,14, 结果载于下表

例6的方差分析表

方差来源	平方和	自由度	均方	F 比
因素	318.98	3	106.33	3.76
误差	395.46	14	28.25	
总和	714.44	17		

$F_{0.05}(3,14) = 3.34 < 3.76$, 故在水平0.05下拒绝 H_0 , 认为各类型电路的响应时间有显著差异.



在MATLAB中求解

```
x=[19,22,20,18,15,20,40,21,33,27,16,17,15,18,26,18,  
    22,19];
```

```
y=[1,1,1,1,1,2,2,2,2,2,3,3,3,3,3,4,4,4];
```

```
p=anova1(x,y)
```

程序运行结果

方差分析表

Box 图检验

帮

助



六、小结

1. 随机试验: **单因素试验、多因素试验**
2. **单因素试验方差分析步骤**
 - (1) 建立数学模型;
 - (2) 分解平方和;
 - (3) 研究统计特性;
 - (4) 进行假设检验;
 - (5) 估计未知参数.



第二节 双因素试验的方差分析

- 一、双因素等重复试验的方差分析
- 二、双因素无重复试验的方差分析
- 三、小结

一、双因素等重复试验的方差分析

因素A: $A_1, A_2, \dots, A_r$. 因素B: $B_1, B_2, \dots, B_s$.

表 9.8

因素A \ 因素B	B_1	B_2	$\dots$	B_s
A_1	$X_{111}, X_{112}, \dots, X_{11t}$	$X_{121}, X_{122}, \dots, X_{12t}$	$\dots$	$X_{1s1}, X_{1s2}, \dots, X_{1st}$
A_2	$X_{211}, X_{212}, \dots, X_{21t}$	$X_{221}, X_{222}, \dots, X_{22t}$	$\dots$	$X_{2s1}, X_{2s2}, \dots, X_{2st}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
A_r	$X_{r11}, X_{r12}, \dots, X_{r1t}$	$X_{r21}, X_{r22}, \dots, X_{r2t}$	$\dots$	$X_{rs1}, X_{rs2}, \dots, X_{rst}$



假设 $X_{ijk} \sim N(\mu_{ij}, \sigma^2)$, $i = 1, \dots, r$, $j = 1, \dots, s$,
 $k = 1, \dots, t$. 各 X_{ijk} 独立, μ_{ij}, σ^2 均为未知参数.

$$\left. \begin{aligned} X_{ijk} &= \mu_{ij} + \varepsilon_{ijk}, \\ \varepsilon_{ijk} &\sim N(0, \sigma^2), \text{ 各 } \varepsilon_{ijk} \text{ 独立}, \\ i &= 1, 2, \dots, r; \quad j = 1, 2, \dots, s; \quad k = 1, 2, \dots, t. \end{aligned} \right\} (2.1\text{式})$$

引入记号

$$\mu = \frac{1}{rs} \sum_{i=1}^r \sum_{j=1}^s \mu_{ij} \quad \mu \text{ 为总平均,}$$



$$\mu_{i\cdot} = \frac{1}{s} \sum_{j=1}^s \mu_{ij}, \quad i = 1, \dots, r$$

$$\mu_{\cdot j} = \frac{1}{r} \sum_{i=1}^r \mu_{ij}, \quad j = 1, \dots, s$$

$$\alpha_i = \mu_{i\cdot} - \mu, \quad i = 1, \dots, r$$

α_i 为水平 A_i 的效应,

$$\beta_j = \mu_{\cdot j} - \mu, \quad j = 1, \dots, s$$

β_j 为水平 B_j 的效应,

$$\text{显然 } \sum_{i=1}^r \alpha_i = 0, \quad \sum_{j=1}^s \beta_j = 0.$$



则
$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu)$$
$$= \mu + \alpha_i + \beta_j + \gamma_{ij},$$

其中 $\gamma_{ij} = \mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu$

γ_{ij} 称为水平 A_i 和水平 B_j 的交互效应,

$$\sum_{i=1}^r \gamma_{ij} = 0, \quad j = 1, \dots, s$$

$$\sum_{j=1}^s \gamma_{ij} = 0, \quad i = 1, \dots, r$$



(2.1式)可改写如下:

$$\left. \begin{aligned} X_{ijk} &= \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \\ \varepsilon_{ijk} &\sim N(0, \sigma^2), \quad \text{各 } \varepsilon_{ijk} \text{ 独立}, \\ i &= 1, 2, \dots, r; j = 1, 2, \dots, s; \quad k = 1, 2, \dots, t, \\ \sum_{i=1}^r \alpha_i &= 0, \quad \sum_{j=1}^s \beta_j = 0, \quad \sum_{i=1}^r \gamma_{ij} = 0, \quad \sum_{j=1}^s \gamma_{ij} = 0, \end{aligned} \right\} (2.2\text{式})$$

其中 $\mu, \alpha_i, \beta_j, \gamma_{ij}$ 及 σ^2 均为未知参数。

(2.2式)为双因素试验方差分析的数学模型



以上模型需检验以下三个假设：

$$\begin{cases} H_{01} : \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0, \\ H_{11} : \alpha_1, \alpha_2, \cdots, \alpha_r \text{不全为零}. \end{cases}$$

$$\begin{cases} H_{02} : \beta_1 = \beta_2 = \cdots = \beta_s = 0, \\ H_{12} : \beta_1, \beta_2, \cdots, \beta_s \text{不全为零}. \end{cases}$$

$$\begin{cases} H_{03} : \gamma_{11} = \gamma_{12} = \cdots = \gamma_{rs} = 0, \\ H_{13} : \gamma_{11}, \gamma_{12}, \cdots, \gamma_{rs} \text{不全为零}. \end{cases}$$

与单因素情况类似，
检验方法建立在平方和的分解上。



引入记号

$$\bar{X} = \frac{1}{rst} \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^t X_{ijk},$$

$$\bar{X}_{ij\cdot} = \frac{1}{t} \sum_{k=1}^t X_{ijk}, \quad i = 1, 2, \dots, r; \quad j = 1, 2, \dots, s,$$

$$\bar{X}_{i..} = \frac{1}{st} \sum_{j=1}^s \sum_{k=1}^t X_{ijk}, \quad i = 1, 2, \dots, r;$$

$$\bar{X}_{\cdot j\cdot} = \frac{1}{rt} \sum_{i=1}^r \sum_{k=1}^t X_{ijk} \quad j = 1, 2, \dots, s,$$



引入总偏差平方和(称为**总变差**)

$$\begin{aligned}
 S_T &= \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^t (X_{ijk} - \bar{X})^2 \\
 &= \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^t [(X_{ijk} - \bar{X}_{ij\cdot}) + (\bar{X}_{i\cdot\cdot} - \bar{X}) + \\
 &\quad (\bar{X}_{\cdot j\cdot} - \bar{X}) + (\bar{X}_{ij\cdot} - \bar{X}_{i\cdot\cdot} - \bar{X}_{\cdot j\cdot} + \bar{X})]^2 \\
 &= \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^t (X_{ijk} - \bar{X}_{ij\cdot})^2 + st \sum_{i=1}^r (\bar{X}_{i\cdot\cdot} - \bar{X})^2 + \\
 &\quad + rt \sum_{j=1}^s (\bar{X}_{\cdot j\cdot} - \bar{X})^2 + t \sum_{i=1}^r \sum_{j=1}^s (\bar{X}_{ij\cdot} - \bar{X}_{i\cdot\cdot} - \bar{X}_{\cdot j\cdot} + \bar{X})^2
 \end{aligned}$$



可得平方和的分解式

$$S_T = S_E + S_A + S_B + S_{A \times B}$$

误差
平方和

因素A的
效应平方和

因素B的
效应平方和

因素A,B的交
互效应平方和

其中 $S_E = \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^t (X_{ijk} - \bar{X}_{ij\cdot})^2$, $S_A = st \sum_{i=1}^r (\bar{X}_{i\cdot\cdot} - \bar{X})^2$,

$$S_B = rt \sum_{j=1}^s (\bar{X}_{\cdot j\cdot} - \bar{X})^2,$$

$$S_{A \times B} = t \sum_{i=1}^r \sum_{j=1}^s (\bar{X}_{ij\cdot} - \bar{X}_{i\cdot\cdot} - \bar{X}_{\cdot j\cdot} + \bar{X})^2.$$



统计特性

	自由度	数学期望
S_T	$rst - 1$	
S_E	$rs(t - 1)$	$rs(t - 1)\sigma^2$
S_A	$r - 1$	$(r - 1)\sigma^2 + st \sum_{i=1}^r \alpha_i^2$
S_B	$s - 1$	$(s - 1)\sigma^2 + rt \sum_{j=1}^s \beta_j^2$
$S_{A \times B}$	$(r - 1)(s - 1)$	$(r - 1)(s - 1)\sigma^2 + t \sum_{i=1}^r \sum_{j=1}^s \gamma_{ij}^2$



且有

$$E\left(\frac{S_E}{rs(t-1)}\right) = \sigma^2, \quad E\left(\frac{S_A}{r-1}\right) = \sigma^2 + \frac{st \sum_{i=1}^r \alpha_i^2}{r-1},$$

$$E\left(\frac{S_B}{s-1}\right) = \sigma^2 + \frac{rt \sum_{j=1}^s \beta_j^2}{s-1},$$

$$E\left(\frac{S_{A \times B}}{(r-1)(s-1)}\right) = \sigma^2 + \frac{t \sum_{i=1}^r \sum_{j=1}^s \gamma_{ij}^2}{(r-1)(s-1)},$$



拒绝域情况

当 $H_{01} : \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$ 为真时,

$$F_A = \frac{S_A / (r - 1)}{S_E / (rs(t - 1))} \sim F(r - 1, rs(t - 1)).$$

取显著性水平为 α , 得假设 H_{01} 的拒绝域为

$$F_A = \frac{S_A / (r - 1)}{S_E / (rs(t - 1))} \geq F_\alpha(r - 1, rs(t - 1)).$$



类似地，取显著性水平为 α ，
得假设 H_{02} 的拒绝域为

$$F_B = \frac{S_B / (s - 1)}{S_E / (rs(t - 1))} \geq F_\alpha(s - 1, rs(t - 1)).$$

取显著性水平为 α ，得假设 H_{03} 的拒绝域为

$$\begin{aligned} F_{A \times B} &= \frac{S_{A \times B} / ((r - 1)(s - 1))}{S_E / (rs(t - 1))} \\ &\geq F_\alpha((r - 1)(s - 1), rs(t - 1)). \end{aligned}$$

上述结果可汇总为方差分析表.



双因素试验的方差分析表

方差来源	平方和	自由度	均方	F 比
因素A	S_A	$r - 1$	$\bar{S}_A = \frac{S_A}{r - 1}$	$F_A = \frac{\bar{S}_A}{\bar{S}_E}$
因素B	S_B	$s - 1$	$\bar{S}_B = \frac{S_B}{s - 1}$	$F_B = \frac{\bar{S}_B}{\bar{S}_E}$
交互作用	$S_{A \times B}$	$(r - 1)(s - 1)$	$\bar{S}_{A \times B} = \frac{S_{A \times B}}{(r - 1)(s - 1)}$	$F_{A \times B} = \frac{\bar{S}_{A \times B}}{\bar{S}_E}$
误差	S_E	$rs(t - 1)$	$\bar{S}_E = \frac{S_E}{rs(t - 1)}$	
总和	S_T	$rst - 1$		



记 $T_{...} = \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^t X_{ijk},$

$$T_{ij\bullet} = \sum_{k=1}^t X_{ijk}, \quad i=1,2,\dots,r; \quad j=1,2,\dots,s,$$

$$T_{i\bullet\bullet} = \sum_{j=1}^s \sum_{k=1}^t X_{ijk}, \quad i=1,2,\dots,r;$$

$$T_{\bullet j\bullet} = \sum_{i=1}^r \sum_{k=1}^t X_{ijk}, \quad j=1,2,\dots,s,$$

则各平方和 S_T 、 S_E 、 S_A 、 S_B 和 $S_{A \times B}$ 的公式可改写为：



平方和 S_T 、 S_E 、 S_A 、 S_B 和 $S_{A \times B}$ 的公式:

$$\left. \begin{aligned} S_T &= \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^t X_{ijk}^2 - \frac{(T_{...})^2}{rst}, \\ S_A &= \frac{1}{st} \sum_{i=1}^r (T_{i..})^2 - \frac{(T_{...})^2}{rst}, \\ S_B &= \frac{1}{rt} \sum_{j=1}^s (T_{.j.})^2 - \frac{(T_{...})^2}{rst}, \\ S_{A \times B} &= \left(\frac{1}{t} \sum_{i=1}^r \sum_{j=1}^s (T_{ij.})^2 - \frac{(T_{...})^2}{rst} \right) - S_A - S_B, \\ S_E &= S_T - S_A - S_B - S_{A \times B}. \end{aligned} \right\}$$



例1 接上一节例3

假设符合双因素方差分析模型所需的条件,在水平0.05下,检验不同燃料(因素A)、不同推进器(因素B)下的射程是否有显著差异? 交互作用是否显著?

解 $T_{...}$ 、 $T_{ij\cdot}$ 、 $T_{i..}$ 、 $T_{.j\cdot}$ 的计算见下表9.10,

表中括弧内的数字是 $T_{ij\cdot}$,

现在 $r = 4, s = 3, t = 2$.



表9.10

$A \backslash B$	B_1	B_2	B_3	$T_{i..}$
A_1	58.2 52.6 (110.8)	56.2 41.2 (97.4)	65.3 60.8 (126.1)	334.3
A_2	49.1 42.8 (91.9)	54.1 50.5 (104.6)	51.6 48.4 (100)	296.5
A_3	60.1 58.3 (118.4)	70.9 73.2 (144.1)	39.2 40.7 (79.9)	342.4
A_4	75.8 71.5 (147.3)	58.2 51.0 (109.2)	48.7 41.4 (90.1)	346.6
$T_{.j.}$	468.4	455.3	396.1	1319.8



根据平方公式有

$$\begin{aligned} S_T &= (58.2^2 + 52.6^2 + \cdots + 41.4^2) - \frac{1319.8^2}{24} \\ &= 2638.29833, \end{aligned}$$

$$\begin{aligned} S_A &= \frac{1}{6}(334.3^2 + 296.5^2 + 342.4^2 + 346.6^2) - \frac{1319.8^2}{24} \\ &= 261.67500, \end{aligned}$$

$$\begin{aligned} S_B &= \frac{1}{8}(468.4^2 + 455.3^2 + 396.1^2) - \frac{1319.8^2}{24} \\ &= 370.98083, \end{aligned}$$



$$S_{A \times B} = \frac{1}{2}(110.8^2 + 91.9^2 + \cdots + 90.1^2) - \frac{1319.8^2}{24} - S_A - S_B \\ = 1768.69250,$$

$$S_E = S_T - S_A - S_B - S_{A \times B} = 236.95000.$$

方差分析表见下页.



表9.11 例1的方差分析表

方差来源	平方和	自由度	均方	F 比
因素A (燃料)	261.67500	3	87.2250	$F_A=4.42$
因素B (推进器)	370.98083	2	185.490	$F_B=9.39$
交互作用 ($A \times B$)	1768.69250	6	294.7821	$F_{A \times B}=14.9$
误差	236.95000	12	19.7458	
总和	2638.29833	23		



需检验假设 H_{01} 、 H_{02} 、 H_{03} ,

因为 $F_{0.05}(3,12) = 3.49 < F_A$,

$$F_{0.05}(2,12) = 3.89 < F_B,$$

所以在显著水平0.05下, 拒绝假设 H_{01} 、 H_{02} ,

即燃料和推进器对射程的影响是显著的.

又因为 $F_{0.05}(6,12) = 3.00 < F_{A \times B}$, 故拒绝假设 H_{03} ,

注意到 $F_{0.001}(6,12) = 8.38 < F_{A \times B}$,

故交互作用效应是高度显著的.



在MATLAB中求解

函数:*anova2*

格式:*p=anova2(x, reps)*

说明:执行平衡的双因素试验的方差分析来比较*x*中两个或多个列或行的均值.不同列的数据代表某一因素的差异,不同行的数据代表另一因素的差异.如果每行列对有多于一个的观察值,则变量*reps*指出每一单元观察点的数目,每一单元包含*reps*行.



源程序: $a=[58.2,56.2,65.3;52.6,41.2,60.8;$
49.1,54.1,51.6;42.8,50.5,48.4;
60.1,70.9,39.2;58.3,73.2,40.7;
75.8,58.2,48.7;71.5,51.0,41.4];
 $p=\text{anova2}(a,2)$

程序运行结果

方差分析表

帮

助



例2 在某种金属材料的生产过程中, 对热处理温度(因素 B)与时间(因素 A)各取两个水平, 产品强度的测定结果(相对值)如表9.12所示. 在同一条件下每个实验重复两次. 设各水平搭配下强度的总体服从正态分布且方差相同. 各样本独立. 问热处理温度、时间以及这两者的交互作用对产品强度是否有显著的影响? (取显著性水平为 0.05)



表9.12

$A \backslash B$	B_1	B_2	$T_{i..}$
A_1	38.0 38.6 (76.6)	47.0 44.8 (91.8)	168.4
A_2	45.0 43.8 (88.8)	42.4 40.8 (83.2)	172
$T_{.j.}$	165.4	175	340.4

解 做如下计算

$$S_T = (38.0^2 + 38.6^2 + \cdots + 40.8^2) - \frac{340.4^2}{8} = 71.82,$$

$$S_A = \frac{1}{4}(168.4^2 + 172^2) - \frac{340.4^2}{8} = 1.62,$$

$$S_B = \frac{1}{4}(165.4^2 + 175^2) - \frac{340.4^2}{8} = 11.52,$$

$$S_{A \times B} = 14551.24 - 14484.02 - 1.62 - 11.52 = 54.08,$$

$$\begin{aligned} S_E &= S_T - S_A - S_B - S_{A \times B} \\ &= 71.82 - 1.62 - 11.52 - 54.08 = 4.6, \end{aligned}$$

需检验假设 H_{01} 、 H_{02} 、 H_{03} ,

因为 $F_{0.05}(1,4) = 7.71$,

所以认为时间对强度的影响不显著,而温度的影响显著,且交互作用的影响显著.



表9.13 例2的方差分析表

方差来源	平方和	自由度	均方	F 比
因素A	1.62	1	1.62	$F_A = 1.4$ $F_B = 10.0$ $F_{A \times B} = 47.0$
因素B	11.52	1	11.52	
$S_{A \times B}$	54.08	1	54.08	
误差	4.6	4	1.15	
总和	71.82	7		



在MATLAB中求解

源程序 `a=[38.0,47.0;
 38.6,44.8;
 45.0,42.4;
 43.8,40.8];
 p=anova2(a,2)`

程序运行结果

方差分析表

帮

助



二、双因素无重复试验的方差分析

检验两个因素的交互效应, 对两个因素的每一组合至少要做两次试验.

如果已知不存在交互作用, 或已知交互作用对试验的指标影响很小, 则可以不考虑交互作用.

对两个因素的每一组合只做一次试验, 也可以对各因素的效应进行分析——双因素无重复试验的方差分析.



表 9.14

因素A \ 因素B	B_1	B_2	$\dots$	B_s
A_1	X_{11}	X_{12}	$\dots$	X_{1s}
A_2	X_{21}	X_{22}	$\dots$	X_{2s}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
A_r	X_{r1}	X_{r2}	$\dots$	X_{rs}

假设

$$X_{ij} \sim N(\mu_{ij}, \sigma^2), \quad i = 1, \dots, r; \quad j = 1, \dots, s.$$



各 X_{ij} 独立, μ_{ij}, σ^2 均为未知参数.

$$\left. \begin{array}{l} X_{ij} = \mu_{ij} + \varepsilon_{ij}, \\ i = 1, 2, \dots, r; j = 1, 2, \dots, s, \\ \varepsilon_{ij} \sim N(0, \sigma^2), \text{各 } \varepsilon_{ij} \text{ 独立,} \end{array} \right\} (2.3\text{式})$$

由于不存在交互作用, $\gamma_{ij} = 0$, $\mu_{ij} = \mu + \alpha_i + \beta_j$.

(2.3式)可改写为下式

$$\left. \begin{aligned} X_{ij} &= \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \\ \varepsilon_{ij} &\sim N(0, \sigma^2), \text{各 } \varepsilon_{ij} \text{ 独立}, \\ i &= 1, 2, \dots, r; \quad j = 1, 2, \dots, s, \\ \sum_{i=1}^r \alpha_i &= 0, \quad \sum_{j=1}^s \beta_j = 0. \end{aligned} \right\}$$

**双因素无重复
试验方差分析
的数学模型**

检验假设 $\left\{ \begin{aligned} H_{01} : & \alpha_1 = \alpha_2 = \dots = \alpha_r = 0, \\ H_{11} : & \alpha_1, \alpha_2, \dots, \alpha_r \text{ 不全为零.} \end{aligned} \right.$

$\left\{ \begin{aligned} H_{02} : & \beta_1 = \beta_2 = \dots = \beta_s = 0, \\ H_{12} : & \beta_1, \beta_2, \dots, \beta_s \text{ 不全为零.} \end{aligned} \right.$



表9.15 双因素无重复试验的方差分析表

方差来源	平方和	自由度	均方	F 比
因素A	S_A	$r - 1$	$\bar{S}_A = \frac{S_A}{r - 1}$	$F_A = \frac{\bar{S}_A}{\bar{S}_E}$
因素B	S_B	$s - 1$	$\bar{S}_B = \frac{S_B}{s - 1}$	$F_B = \frac{\bar{S}_B}{\bar{S}_E}$
误差	S_E	$(r - 1) \times (s - 1)$	$\bar{S}_E = \frac{S_E}{(r - 1)(s - 1)}$	
总和	S_T	$rs - 1$		



取显著性水平为 α ,得假设 H_{01} 的拒绝域为

$$F_A = \frac{\bar{S}_A}{\bar{S}_E} \geq F_{\alpha}(r-1, (r-1)(s-1)).$$

取显著性水平为 α ,得假设 H_{02} 的拒绝域为

$$F_B = \frac{\bar{S}_B}{\bar{S}_E} \geq F_{\alpha}(s-1, (r-1)(s-1)).$$



表9.15中的平方和可按下式计算:

$$S_T = \sum_{i=1}^r \sum_{j=1}^s X_{ij}^2 - \frac{(T_{..})^2}{rs}, \quad S_A = \frac{1}{s} \sum_{i=1}^r (T_{i.})^2 - \frac{(T_{..})^2}{rs},$$

$$S_B = \frac{1}{r} \sum_{j=1}^s (T_{.j})^2 - \frac{(T_{..})^2}{rs}, \quad S_E = S_T - S_A - S_B.$$

其中 $T_{..} = \sum_{i=1}^r \sum_{j=1}^s X_{ij}, \quad T_{i.} = \sum_{j=1}^s X_{ij}, \quad i = 1, 2, \dots, r;$

$$T_{.j} = \sum_{i=1}^r X_{ij}, \quad j = 1, 2, \dots, s,$$

例3 下面给出了在某 5 个不同地点、不同时间空气中的颗粒状物 (以 mg/m^3 计) 的含量的数据:

		因素B(地点)					$T_{i\bullet}$
		1	2	3	4	5	
因素A (时间)	1975年10月	76	67	81	56	51	331
	1976年 1 月	82	69	96	59	70	376
	1976年 5 月	68	59	67	54	42	290
	1996年 8 月	63	56	64	58	37	278
$T_{\bullet j}$		289	251	308	227	200	1275



设本题符合模型中的条件, 试在显著性水平为 0.05 下检验: 在不同时间下颗粒状物含量的均值有无显著差异, 在不同地点下颗粒状物含量的均值有无显著差异.



解 按题意检验假设, $T_{i\cdot}, T_{\cdot j}$ 的值已算出载于上表, 现在 $r = 4, s = 5$. 由平方和表达式得到,

$$S_T = 76^2 + 67^2 + \cdots + 37^2 - \frac{1275^2}{20} = 3571.75$$

$$S_A = \frac{1}{5}(331^2 + 376^2 + 290^2 + 278^2) - \frac{1275^2}{20} = 1182.95$$

$$S_B = \frac{1}{4}(289^2 + 251^2 + \cdots + 200^2) - \frac{1275^2}{20} = 1947.50$$

$$S_E = 3571.75 - (1182.95 + 1947.50) = 441.30$$



表9.16 例3的方差分析表

方差来源	平方和	自由度	均方	F 比
因素A	$S_A = 1182.95$	3	394.32	$F_A = 10.72$
因素B	$S_B = 1947.50$	4	486.88	$F_B = 13.24$
误 差	$S_E = 441.30$	12	36.78	
总 和	$S_T = 3571.75$	19		

由于 $F_{0.05}(3,12) = 3.49 < 10.72$,

$F_{0.05}(4,12) = 3.26 < 13.24$, 故拒绝域 H_{01} 及 H_{02} , 即认为不同时间下颗粒物含量的均值有显著差异, 也认为不同地点下颗粒状物含量的均值有显著差异.



在MATLAB中求解

源程序 `a=[76,67,81,56,51;
82,69,96,59,70;
68,59,67,54,42;
63,56,64,58,37];
p=anova2(a)`

程序运行结果

方差分析表

帮

助



三、小结

1. 双因素等重复试验的方差分析步骤

- (1) 建立数学模型;
- (2) 分解平方和;
- (3) 研究统计特性;
- (4) 确定拒绝域.

2. 双因素无重复试验的方差分析步骤

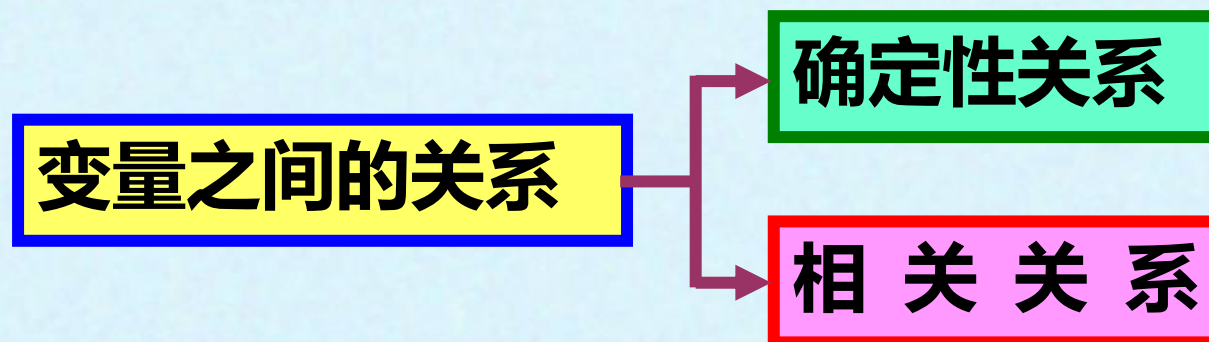
- (1) 建立数学模型;
- (2) 分解平方和;
- (3) 研究统计特性;
- (4) 确定拒绝域.



第三节 一元线性回归

- 一、回归分析的基本的思想
- 二、一元回归的数学模型
- 三、可化为一元线性回归的问题
- 四、小结

一、回归分析的基本思想



$$S = \pi r^2$$

确定性关系

身高和体重

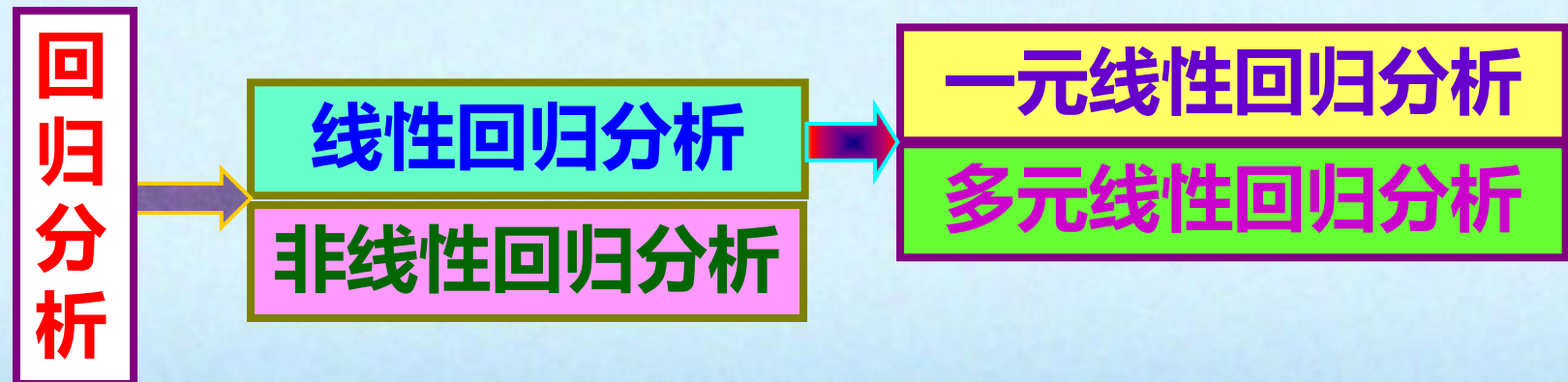
相关关系

相关关系的特征是：

变量之间的关系很难用一种精确的方法表示出来。

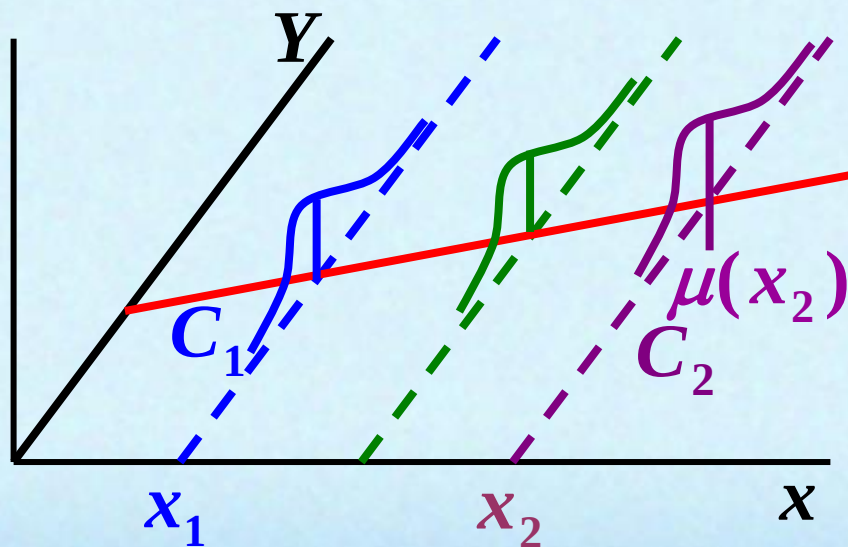
由于存在测量误差等原因, 确定性关系在实际问题中往往通过相关关系表示出来; 另一方面, 当对事物内部规律了解得更加深刻时, 相关关系也有可能转化为确定性关系.

回归分析—— 处理变量之间的相关关系的一种数学方法, 它是最常用的数理统计方法.



二、一元线性回归的数学模型

设随机变量 Y (因变量)与普通变量 x (自变量)之间存在着相互关系. 由于 Y 是随机变量, 对于 x 的确定值, Y 有它的分布(如图)



图中 C_1, C_2 分别是 x_1, x_2 处 Y 的概率密度曲线.



$F(y|x)$ 表示当 x 取确定的值 x 时, 所对应的 Y 的分布函数 $E(Y) = \mu(x)$.

因为对随机变量 η , 当 $c = E(\eta)$ 时, $E[(\eta - c)^2]$ 达到最小.

所以在一切 x 的函数中以回归函数 $\mu(x)$ 作为 Y 的近似, 均方误差 $E[(Y - \mu(x))^2]$ 为最小.

实际问题中的 $\mu(x)$ 一般未知.



回归分析的任务:

- (1)根据试验数据估计回归函数;**
- (2)讨论回归函数中参数的点估计、区间估计;**
- (3)对回归函数中的参数或者回归函数本身进行假设检验;**
- (4)利用回归函数进行预测与控制等等.**

特别对随机变量 Y 的观察值做出点预测和区间预测.



问题的一般提法

对 x 的一组不完全相同的值 $x_1, x_2, \dots, x_n$, 设 $Y_1, Y_2, \dots, Y_n$ 分别是在 $x_1, x_2, \dots, x_n$ 处对 Y 的独立观察结果.

称 $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$ 是一个样本.

对应的样本值记为 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.

利用样本来估计 Y 关于 x 的回归函数 $\mu(x)$.
首先推测 $\mu(x)$ 的形式.



在一些问题中,可以由专业知识知道 $\mu(x)$ 的形式.
否则,可将每对观察值 (x_i, y_i) 在直角坐标系中描出它的相应的点,这种图称为散点图.

求解步骤

1.推测回归函数的形式

方法一 根据专业知识或者经验公式确定;

方法二 作散点图观察.



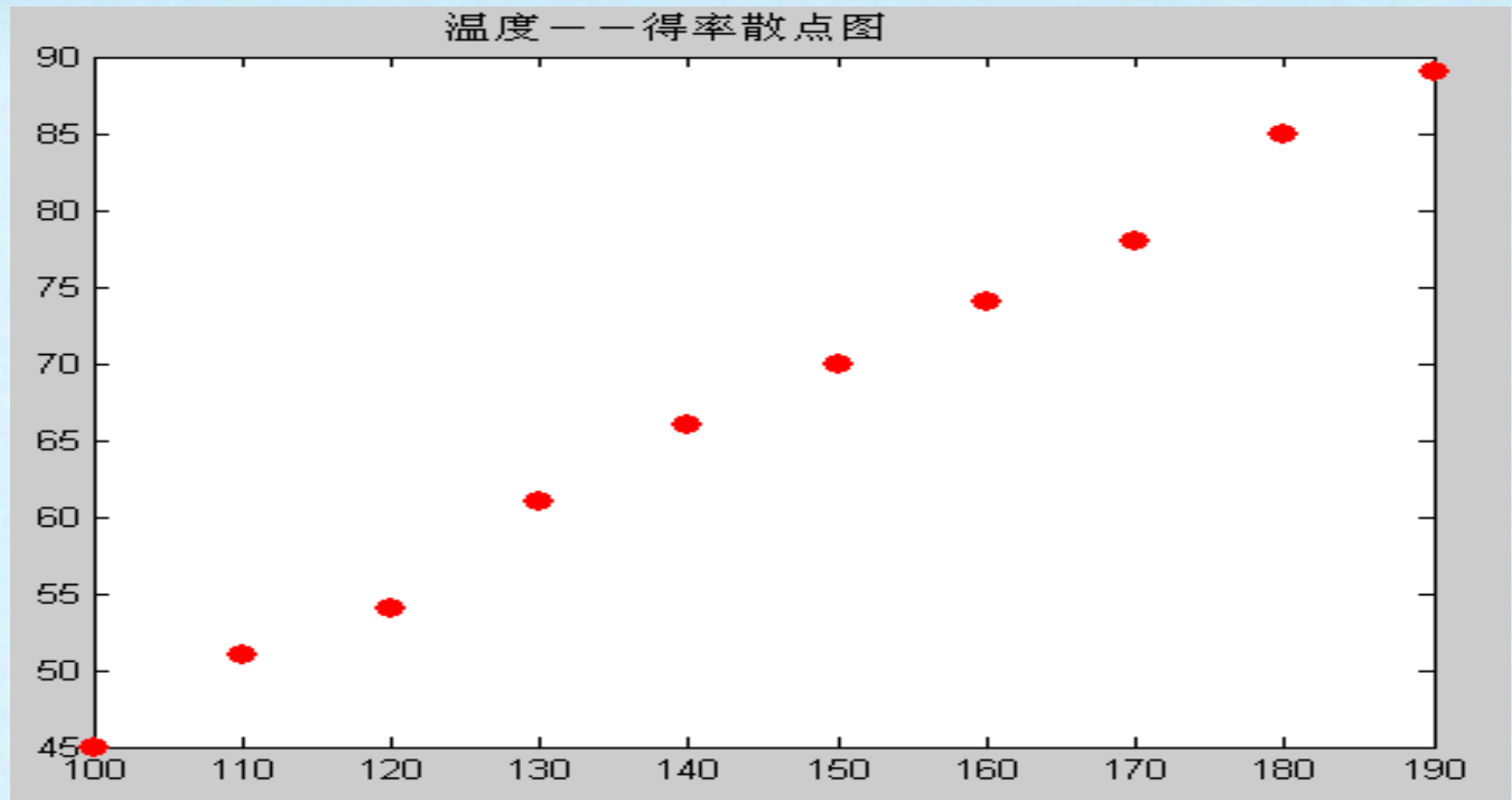
例1 为研究某一化学反应过程中, 温度 $x(^{\circ}\text{C})$ 对产
品得率 $Y(\%)$ 的影响, 测得数据如下.

温度 $x(^{\circ}\text{C})$	100	110	120	130	140	150	160	170	180	190
得率 $Y(\%)$	45	51	54	61	66	70	74	78	85	89

这里自变量 x 是普通变量, Y 是随机变量.

画出散点图如下,





观察散点图, $\mu(x)$ 具有线性函数 $a + bx$ 的形式.



设 Y 关于 x 的回归函数为 $\mu(x)$.

利用样本来估计 $\mu(x)$ 的问题称为求 Y 关于 x 的回归问题.

特别, 若 $\mu(x)$ 为线性函数: $\mu(x) = a + bx$,

此时估计 $\mu(x)$ 的问题称为求一元线性回归问题.



2. 建立回归模型

$$\mu(x) = a + bx \quad \text{一元线性回归问题}$$

假设对于 x 的每一个值有 $Y \sim N(a + bx, \sigma^2)$,
 a, b, σ^2 都是不依赖于 x 的未知参数.

记 $\varepsilon = Y - (a + bx)$, 那么

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

a, b, σ^2 是不依赖于 x 的未知参数.

一元线性回归模型

x 的线性函数 随机误差

假设对于 x (在某个区间内的)的每一个值

$$Y \sim N(a + bx, \sigma^2)$$

其中 a, b, σ^2 都是不依赖于 x 的未知参数. 记

$$\varepsilon = Y - (a + bx)$$

对 Y 作这样的正态假设, 相当于假设

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

其中未知参数 a, b 及 σ^2 都不依赖于 x . 上式称为一元线性回归模型, 其中 b 称为回归系数.



3. 未知参数 a, b 的估计

取 x 的 n 个不全相同的值 $x_1, x_2, \dots, x_n$ 做独立试验, 得到样本 $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$.

$Y_i = a + bx_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$, 各 ε_i 相互独立. 于是 $Y_i \sim N(a + bx_i, \sigma^2)$, $i = 1, 2, \dots, n$. 由 $Y_1, Y_2, \dots, Y_n$ 的独立性, 知 $Y_1, Y_2, \dots, Y_n$ 的联合密度函数为

$$L = \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} \exp \left[-\frac{1}{2\sigma^2} (y_i - a - bx_i)^2 \right]$$



$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - a - bx_i)^2\right].$$

用最大似然估计估计未知参数 a, b .

对于任意一组观察值 $y_1, y_2, \dots, y_n$,

L 就是样本的似然函数,

$$L = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - a - bx_i)^2\right] \text{要取最大值,}$$

只要上式右端方括弧中的平方和部分最小,



即只需函数

$$Q(a,b) = \sum_{i=1}^n (y_i - a - bx_i)^2 \text{取最小值.}$$

根据

$$\left. \begin{aligned} \frac{\partial Q}{\partial a} &= -2 \sum_{i=1}^n (y_i - a - bx_i) = 0 \\ \frac{\partial Q}{\partial b} &= -2 \sum_{i=1}^n (y_i - a - bx_i) x_i = 0 \end{aligned} \right\}$$



得方程组

$$\begin{cases} na + \left(\sum_{i=1}^n x_i\right)b = \sum_{i=1}^n y_i \\ \left(\sum_{i=1}^n x_i\right)a + \left(\sum_{i=1}^n x_i^2\right)b = \sum_{i=1}^n x_i y_i \end{cases}$$

由于 x_i 不全相同,方程组的系数行列式

$$\begin{vmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{vmatrix} = n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2 = n \sum_{i=1}^n (x_i - \bar{x})^2 \neq 0$$



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

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

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

在得到 a, b 的估计 $\hat{a}, \hat{b}$ 后, 对于给定的 x ,



取 $\hat{a} + \hat{b}x$ 作为回归函数 $\mu(x) = a + bx$ 的估计, 即

$$\widehat{\mu(x)} = \hat{a} + \hat{b}x$$

称为 Y 关于 x 的经验回归函数, 记 $\hat{a} + \hat{b}x = \hat{y}$, 方程

$$\hat{y} = \hat{a} + \hat{b}x$$

称为 Y 关于 x 的经验回归方程, 简称**回归方程**, 其图形称为**回归直线**.

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

对于样本值 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, 回归直线通过散点图的几何中心 $(\bar{x}, \bar{y})$.



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

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

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

$$= \sum_{i=1}^n x_i y_i - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right),$$

$$\hat{b} = \frac{S_{xy}}{S_{xx}}, \quad \hat{a} = \frac{1}{n} \sum_{i=1}^n y_i - \left(\frac{1}{n} \sum_{i=1}^n x_i \right) \hat{b}.$$



例2 例1中的随机变量 Y 符合一元线性回归模型所述的条件, 求 Y 关于 x 的线性回归方程.

	x	y	x^2	y^2	xy
	100	45	10000	2025	4500
	110	51	12100	2601	5610
	120	54	14400	2916	6480
	130	61	16900	3721	7930
	140	66	19600	4356	9240
	150	70	22500	4900	10500
	160	74	25600	5476	11840
	170	78	28900	6084	13260
	180	85	32400	7225	15300
	190	89	36100	7921	16910
Σ	1450	673	218500	47225	101570



$$S_{xx} = 218500 - \frac{1}{10} \times 1450^2 = 8250$$

$$S_{xy} = 101570 - \frac{1}{10} \times 1450 \times 673 = 3985$$

$$\hat{b} = S_{xy} / S_{xx} = 0.48303$$

$$\hat{a} = \frac{1}{10} \times 673 - \frac{1}{10} \times 1450 \times 0.48303 = -2.73935$$

回归直线方程 $\hat{y} = -2.73935 + 0.48303x$

或 $\hat{y} = 67.3 + 0.48303(x - 145)$



在MATLAB中求解

源程序 `x=100:10:190;`
`y=[45,51,54,61,66,70,74,78,85,89];`
`polytool(x,y,1,0.05)`

程序运行结果

回归图形

参数传送

置信区间

帮

助



4、未知参数 σ^2 的估计

$$E\{[Y - (a + bx)]^2\} = E(\varepsilon^2) = D(\varepsilon) + [E(\varepsilon)]^2 = \sigma^2.$$

σ^2 越小,用回归函数 $\mu(x) = a + bx$ 作为 Y 的近似导致的均方误差就越小. 利用回归函数

$$\mu(x) = a + bx$$

去研究随机变量 Y 与 x 的关系就愈有效

为了估计 σ^2 , 引入残差平方和

$$\hat{y}_i = \hat{y}|_{x=x_i} = \hat{a} + \hat{b}x_i,$$

$y_i - \hat{y}_i$ 为 x_i 处的残差.



残差平方和

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

是经验回归函数在 x_i 处的函数值 $\widehat{\mu(x)} = \hat{a} + \hat{b}x$ 与 x_i 处的观察值 y_i 的偏差的平方和.

$$\begin{aligned} Q_e &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n [y_i - \bar{y} - \hat{b}(x_i - \bar{x})]^2 \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 - 2\hat{b} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + (\hat{b})^2 \sum_{i=1}^n (x_i - \bar{x})^2 \end{aligned}$$



$$= S_{yy} - 2\hat{b}S_{xy} + (\hat{b})^2 S_{xx}$$

由 $\hat{b} = S_{xy} / S_{xx}$

$$Q_e = S_{yy} - \hat{b}S_{xy}.$$

b, a 的估计量为

$$\hat{b} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

$$\hat{a} = \frac{1}{n} \sum_{i=1}^n Y_i - \frac{\hat{b}}{n} \sum_{i=1}^n x_i = \bar{Y} - \hat{b}\bar{x}$$

其中 $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$. 在 S_{yy}, S_{xy} 的表达式中

将 y_i 改为 $Y_i (i = 1, 2, \dots, n)$, 并把它们分别记为 S_{YY}, S_{xY}

$$S_{YY} = \sum_{i=1}^n (Y_i - \bar{Y})^2,$$

$$S_{xY} = \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}).$$



残差平方和 Q_e 的相应的统计量为

$$Q_e = S_{YY} - \hat{b}S_{xY}$$

残差平方和 Q_e 服从分布

$$\frac{Q_e}{\sigma^2} \sim \chi^2(n-2),$$

$$E\left(\frac{Q_e}{\sigma^2}\right) = n-2, \quad E\left(\frac{Q_e}{n-2}\right) = \sigma^2.$$

σ^2 的无偏估计量为

$$\sigma^2 = \frac{Q_e}{n-2} = \frac{1}{n-2} [S_{YY} - \hat{b}S_{xY}].$$



例3 求例2中方差的无偏估计.

解

$$\begin{aligned} S_{yy} &= \sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 \\ &= 47225 - \frac{1}{10} \times 673^2 \\ &= 1932.1 \end{aligned}$$

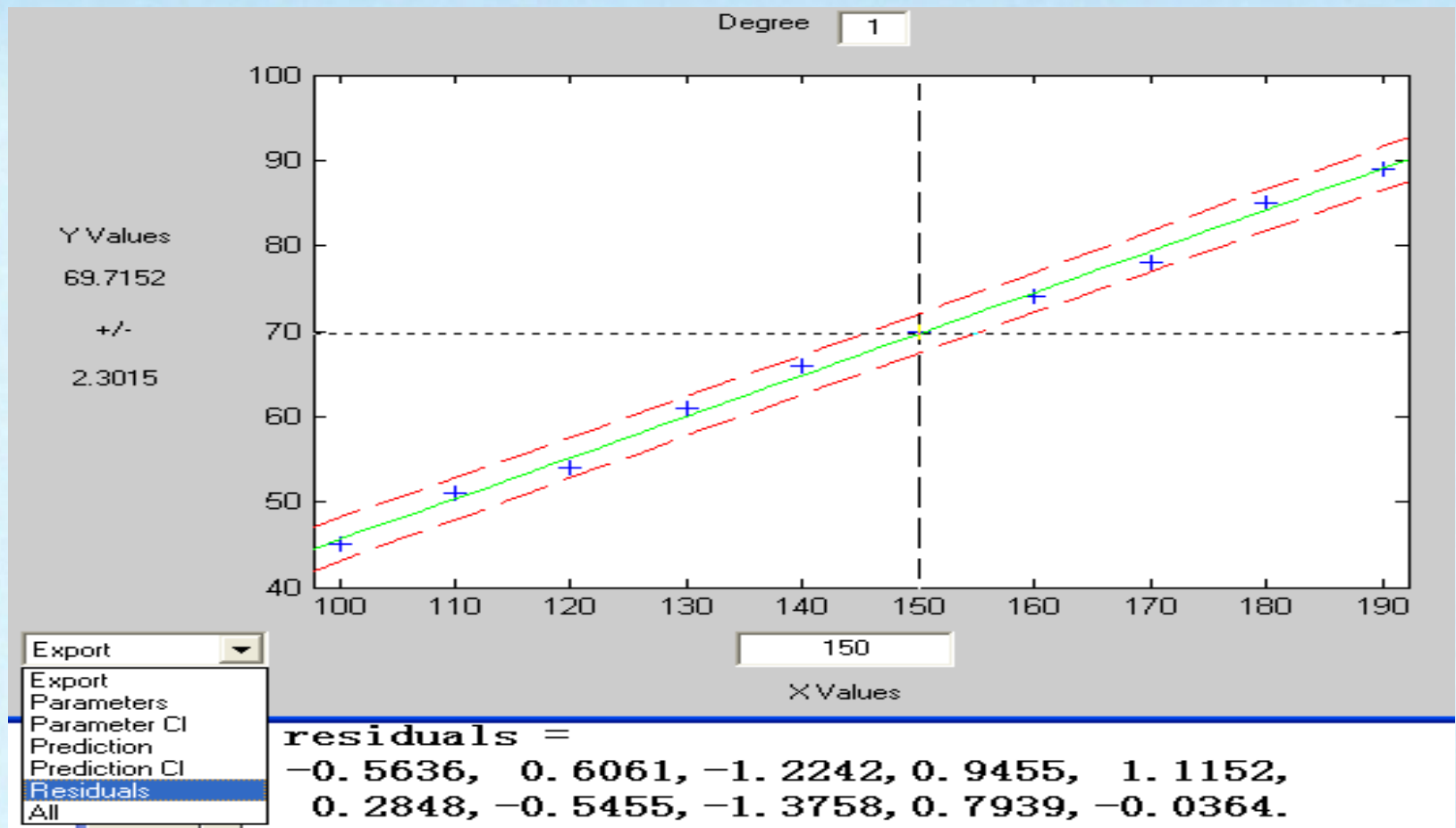
又知 $S_{xy} = 3985$, $\hat{b} = 0.48303$

$$Q_e = S_{yy} - \hat{b} S_{xy} = 7.23$$

$$\hat{\sigma}^2 = Q_e / (n - 2) = 7.23 / 8 = 0.90 .$$



9.3 一元线性回归



$$Q_e = \sum_{i=1}^{10} (\text{residuals})_i^2 = 7.2236, \hat{\sigma}^2 = \frac{7.2236}{8} = 0.9030.$$

5. 线性假设的显著性检验

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

检验假设: $H_0: b = 0,$

$$H_1: b \neq 0.$$

使用 t 检验法来进行检验

$$\hat{b} \sim N(b, \sigma^2 / S_{xx}),$$

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{Q_e}{\sigma^2} \sim \chi^2(n-2).$$

并且 $\hat{b}, Q_e$ 相互独立,



$$\hat{b} - b / \sqrt{\sigma^2 / S_{xx}} / \sqrt{\frac{(n-2)\hat{\sigma}^2}{\sigma^2} / (n-2)} \sim t(n-2),$$

即
$$\frac{\hat{b} - b}{\hat{\sigma}} \sqrt{S_{xx}} \sim t(n-2).$$

当 H_0 为真时 $b = 0$, 此时

$$t = \frac{\hat{b}}{\hat{\sigma}} \sqrt{S_{xx}} \sim t(n-2),$$

并且 $E(\hat{b}) = b = 0$, 得 H_0 的拒绝域为

$$|t| = \frac{|\hat{b}|}{\hat{\sigma}} \sqrt{S_{xx}} \geq t_{\alpha/2}(n-2).$$



拒绝 $H_0 : b = 0$, 认为回归效果显著.

接受 $H_0 : b = 0$, 认为回归效果不显著.

回归效果不显著的原因分析:

- (1) 影响 Y 取值的, 除 x 及随机误差外还有其他不可忽略的因素;
- (2) $E(Y)$ 与 x 的关系不是线性的; 而存在其他关系.
- (3) Y 与 x 不存在关系.



例4 检验例 2 中的回归效果是否显著, 取显著性水平为 0.05 .

解 已知 $\hat{b} = 0.48303$, $S_{xx} = 8250$, $\hat{\sigma}^2 = 0.90$,

查表得 $t_{0.05/2}(n-2) = t_{0.025}(8) = 2.3060$.

假设 $H_0 : b = 0$ 的拒绝域为

$$|t| = \frac{|\hat{b}|}{\hat{\sigma}} \sqrt{S_{xx}} \geq 2.3060,$$

$$|t| = \frac{0.48303}{\sqrt{0.90}} \times \sqrt{8250} = 46.25,$$

拒绝 $H_0 : b = 0$, 认为回归效果显著.



6. 系数 b 的置信区间

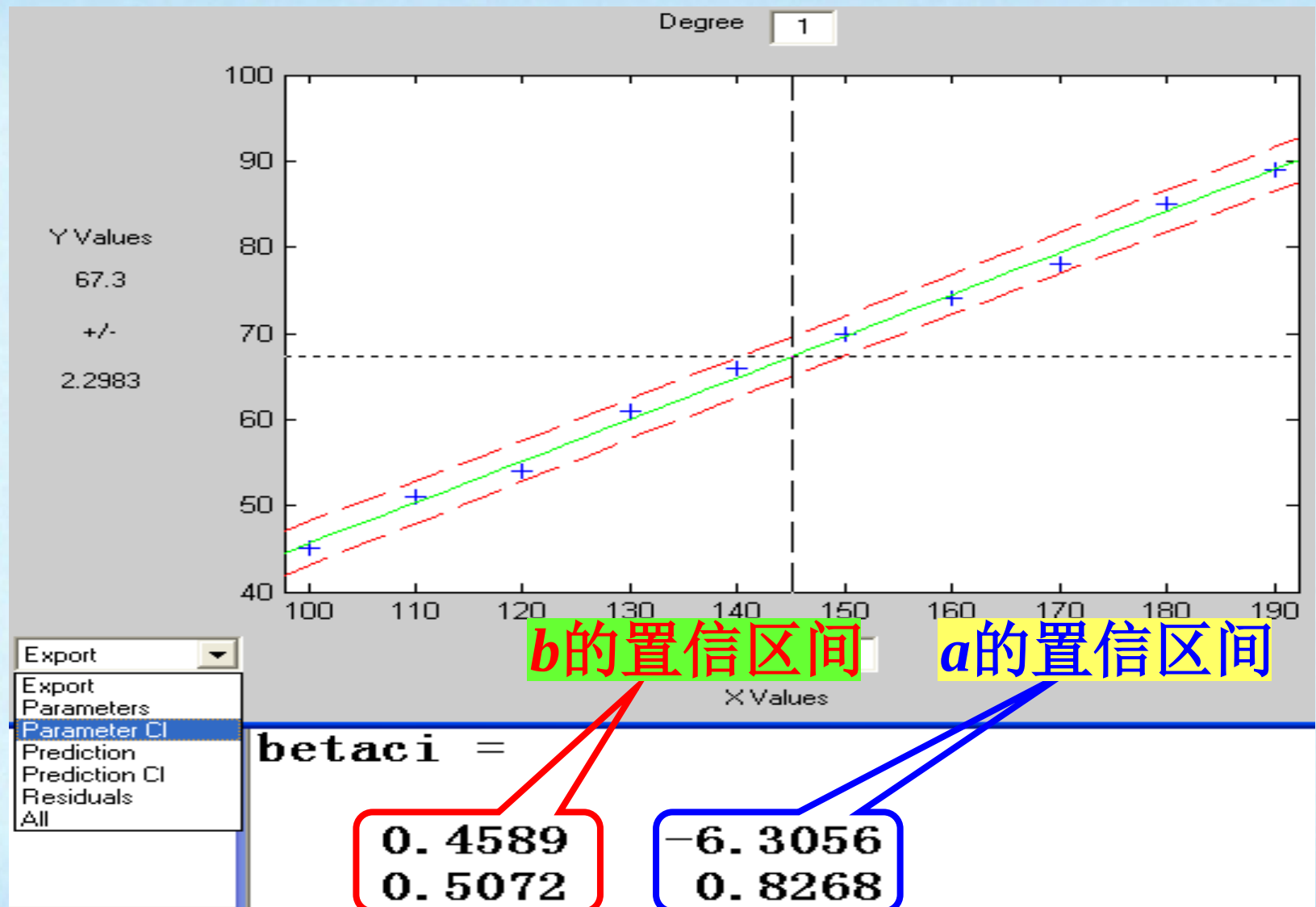
当回归效果显著时, 对系数 b 作区间估计.
系数 b 的置信水平为 $1-\alpha$ 的置信区间为

$$\left(\hat{b} \pm t_{\alpha/2}(n-2) \times \frac{\hat{\sigma}}{\sqrt{S_{xx}}} \right).$$

例如, 求例1中 b 的置信水平为0.95的置信区间.

$$\left(0.48303 \pm 2.3060 \times \sqrt{\frac{0.90}{8250}} \right) = (0.45894, 0.50712).$$

9.3 一元线性回归



7. 回归函数 $\mu(x) = a + bx$ 函数值的点估计和置信区间

设 x_0 是自变量 x 的某一指定值.用经验回归函数 $\hat{y} = \hat{\mu}(x) = \hat{a} + \hat{b}x$ 在 x_0 的函数值 $\hat{y}_0 = \hat{\mu}(x_0) = \hat{a} + \hat{b}x_0$ 作为 $\hat{\mu}(x_0) = \hat{a} + \hat{b}x_0$ 的点估计.

$$\hat{y}_0 = \hat{\mu}(x_0) = \hat{a} + \hat{b}x_0,$$

考虑相应的估计量 $\hat{Y}_0 = \hat{a} + \hat{b}x_0$.

因为 $E(\hat{Y}_0) = a + bx_0$, 所以估计量是无偏的.

求 $\hat{\mu}(x_0) = \hat{a} + \hat{b}x_0$ 的置信区间.



$$\frac{\hat{Y}_0 - (a + bx_0)}{\sigma \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim N(0,1),$$

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{Q_e}{\sigma^2} \sim \chi^2(n-2),$$

$Q_e, \hat{Y}_0$ 相互独立.

$$\frac{\hat{Y}_0 - (a + bx_0)}{\sigma \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \bigg/ \sqrt{\frac{(n-2)\hat{\sigma}^2}{\sigma^2} / (n-2)} \sim t(n-2),$$



$$\frac{\hat{Y}_0 - (a + bx_0)}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t(n-2),$$

得到 $\mu(x_0) = a + bx_0$ 的置信水平为 $1 - \alpha$ 的置信区间为

$$\left(\hat{Y}_0 \pm t_{\alpha/2}(n-2) \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right)$$

或
$$\left(\hat{a} + \hat{b}x_0 \pm t_{\alpha/2}(n-2) \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right).$$



8. Y 的观察值的点预测和预测区间

设 Y_0 是在 $x = x_0$ 处对 Y 的观察结果.

$$Y_0 = a + bx_0 + \varepsilon_0, \quad \varepsilon_0 \sim N(0, \sigma^2).$$

x_0 处的经验回归函数值 $\hat{Y}_0 = \hat{\mu}(x_0) = \hat{a} + \hat{b}x_0$

作为 $Y_0 = a + bx_0 + \varepsilon_0$ 的点预测

Y_0 是将来做一次独立试验的结果,

它与已经得到的试验结果 $Y_1, Y_2, \dots, Y_n$ 相互独立.

$\hat{b}$ 是 $Y_1, Y_2, \dots, Y_n$ 的线性组合,



故 $\hat{Y}_0 = \bar{Y} + \hat{b}(x_0 - \bar{x})$ 是 $Y_1, Y_2, \dots, Y_n$ 的线性组合,
 Y_0 与 $\hat{Y}_0$ 相互独立

$$\hat{Y}_0 - Y_0 \sim N\left(0, \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right] \sigma^2\right),$$

$$\frac{\hat{Y}_0 - Y_0}{\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim N(0, 1)$$

$Y_0, \bar{Y}_0, Q_e$ 的相互独立性知



$$\frac{\hat{Y}_0 - Y_0}{\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \bigg/ \sqrt{\frac{(n-2)\hat{\sigma}^2}{\sigma^2}} \bigg/ (n-2) \sim t(n-2),$$

$$\frac{\hat{Y}_0 - Y_0}{\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t(n-2), \text{ 给定置信水平为 } 1-\alpha,$$

$$P \left\{ \left| \hat{Y}_0 - Y_0 \right| \bigg/ \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \leq t_{\alpha/2}(n-2) \right\} \\ = 1 - \alpha$$



$$P\left\{\hat{Y}_0 - t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} < Y_0 \right. \\ \left. < \hat{Y}_0 + t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\right\} = 1 - \alpha$$

区间 $\left(\hat{Y}_0 \pm t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right)$

或 $\left(\hat{a} + \hat{b}x_0 \pm t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right)$

 Y_0 的置信水平为 $1 - \alpha$ 的预测区间



例5 (续例2)(1) 求回归函数 $\mu(x)$ 在 $x = 125$ 处的值 $\mu(125)$ 的置信水平为0.95 的置信区间, 求在 $x = 125$ 处 Y 的新观察值 Y_0 的置信水平为0.95 的预测区间;
(2) 求在 $x = x_0$ 处 Y 的新观察值 Y_0 的置信水平为0.95 的预测区间.

解 (1)已知 $\hat{b} = 0.48303$, $\hat{a} = -2.7394$, $S_{xx} = 8250$,
 $\hat{\sigma}^2 = 0.90$, $\bar{x} = 145$. 查表得 $t_{0.05/2}(n-2) = t_{0.025}(8)$
 $= 2.3060$.



$$\hat{Y}_0 = \hat{Y}|_{x=125} = [-2.73935 + 0.48303x]_{x=125} = 57.64,$$

$$t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$$= 2.3060 \times \sqrt{0.90} \times \sqrt{\frac{1}{10} + \frac{(125 - 145)^2}{8250}}$$

$$= 0.84,$$

$$t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} = 2.34.$$



回归函数 $\mu(x)$ 在 $x = 125$ 处的值 $\mu(125)$ 的置信水平为0.95的置信区间为

$$(57.64 \pm 0.84).$$

在 $x_0 = 125$ 处得率 Y_0 的置信水平为0.95的预测区间为

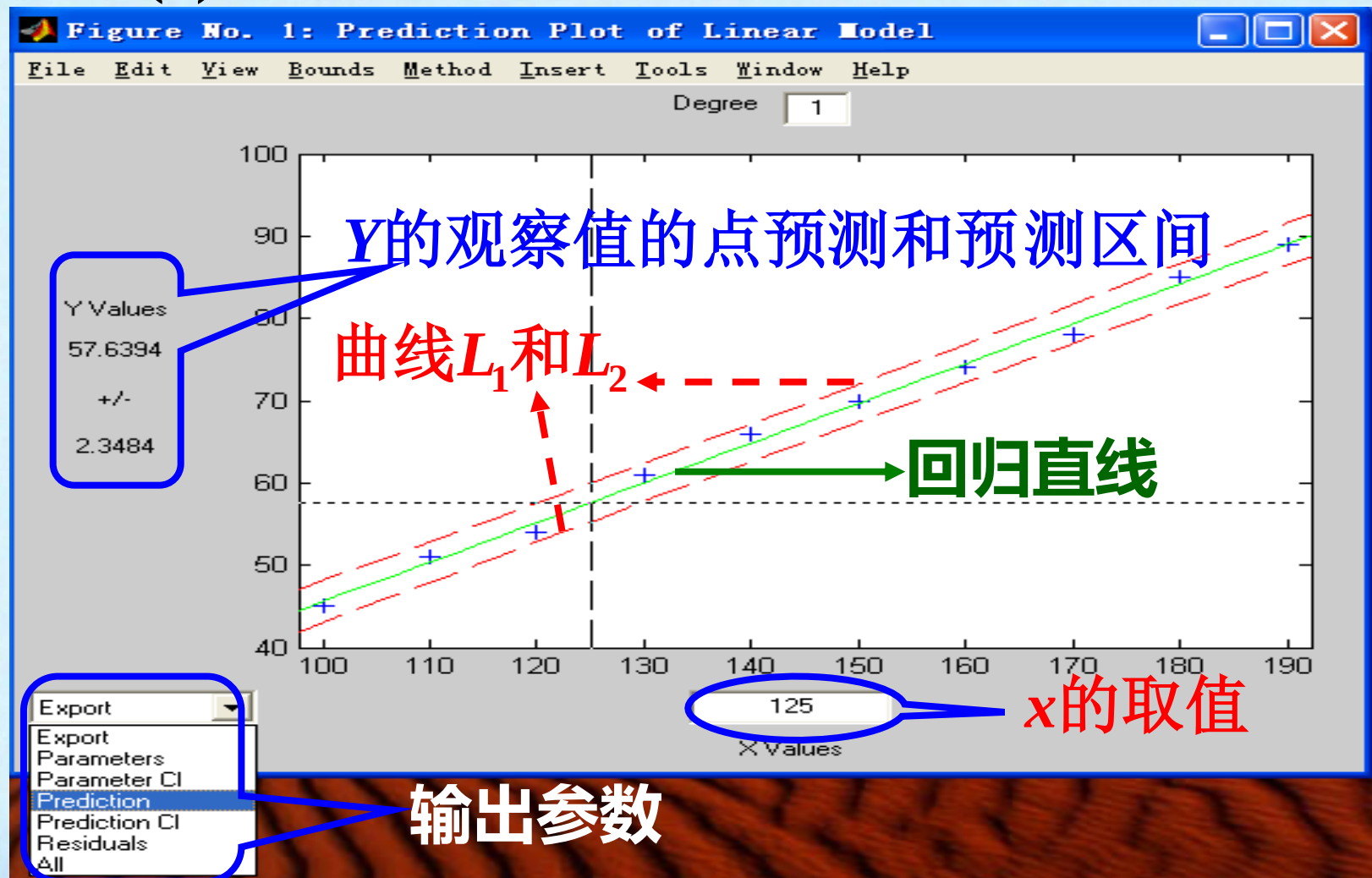
$$(57.64 \pm 2.34).$$

(2)在 $x = x_0$ 处 Y 的新观察值 Y_0 的一个置信水平为0.95的预测区间为

$$\left(\hat{Y}_0 \Big|_{x=x_0} \pm t_{0.025}(8) \hat{\sigma} \sqrt{1 + \frac{1}{10} + \frac{(x_0 - 145)^2}{8250}} \right)$$



(2)在MATLAB中求解



三、可化为一元线性回归的例子

方法——通过适当的变量变换,化成一元线性回归问题进行分析处理.

几种常见的可转化为一元线性回归的模型

$$1. Y = \alpha e^{\beta x} \cdot \varepsilon, \quad \ln \varepsilon \sim N(0, \sigma^2).$$

其中 α, β, σ^2 是与 x 无关的未知参数.

将 $Y = \alpha e^{\beta x} \cdot \varepsilon$ 两边取对数,

得
$$\ln Y = \ln \alpha + \beta x + \ln \varepsilon.$$

令 $\ln Y = Y', \ln \alpha = a, \beta = b, x = x', \ln \varepsilon = \varepsilon'$



转化为一元线性回归模型:

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

$$2. Y = \alpha x^\beta \cdot \varepsilon, \quad \ln \varepsilon \sim N(0, \sigma^2).$$

其中 α, β, σ^2 是与 x 无关的未知参数.

将 $Y = \alpha e^{\beta x} \cdot \varepsilon$ 两边取对数,

得
$$\ln Y = \ln \alpha + \beta \ln x + \ln \varepsilon.$$

令 $\ln Y = Y', \ln \alpha = a, \beta = b, \ln x = x', \ln \varepsilon = \varepsilon'$

转化为一元线性回归模型:

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



$$3. Y = \alpha + \beta h(x) + \varepsilon, \quad \ln \varepsilon \sim N(0, \sigma^2).$$

其中 α, β, σ^2 是与 x 无关的未知参数.

$h(x)$ 是 x 的已知函数,

$$\text{令 } \alpha = a, \quad \beta = b, \quad h(x) = x',$$

转化为一元线性回归模型:

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



例6 下表是 1957 年美国旧轿车价格的调查资料, 今以 x 表示轿车的使用年数, Y 表示相应的平均价格(以美元计), 求 Y 关于 x 的回归方程.

表

年数 x	1	2	3	4	5	6	7	8	9	10
价格 Y	2651	1943	1494	1087	765	538	484	290	226	204



解

在MATLAB中求解

首先作散点图

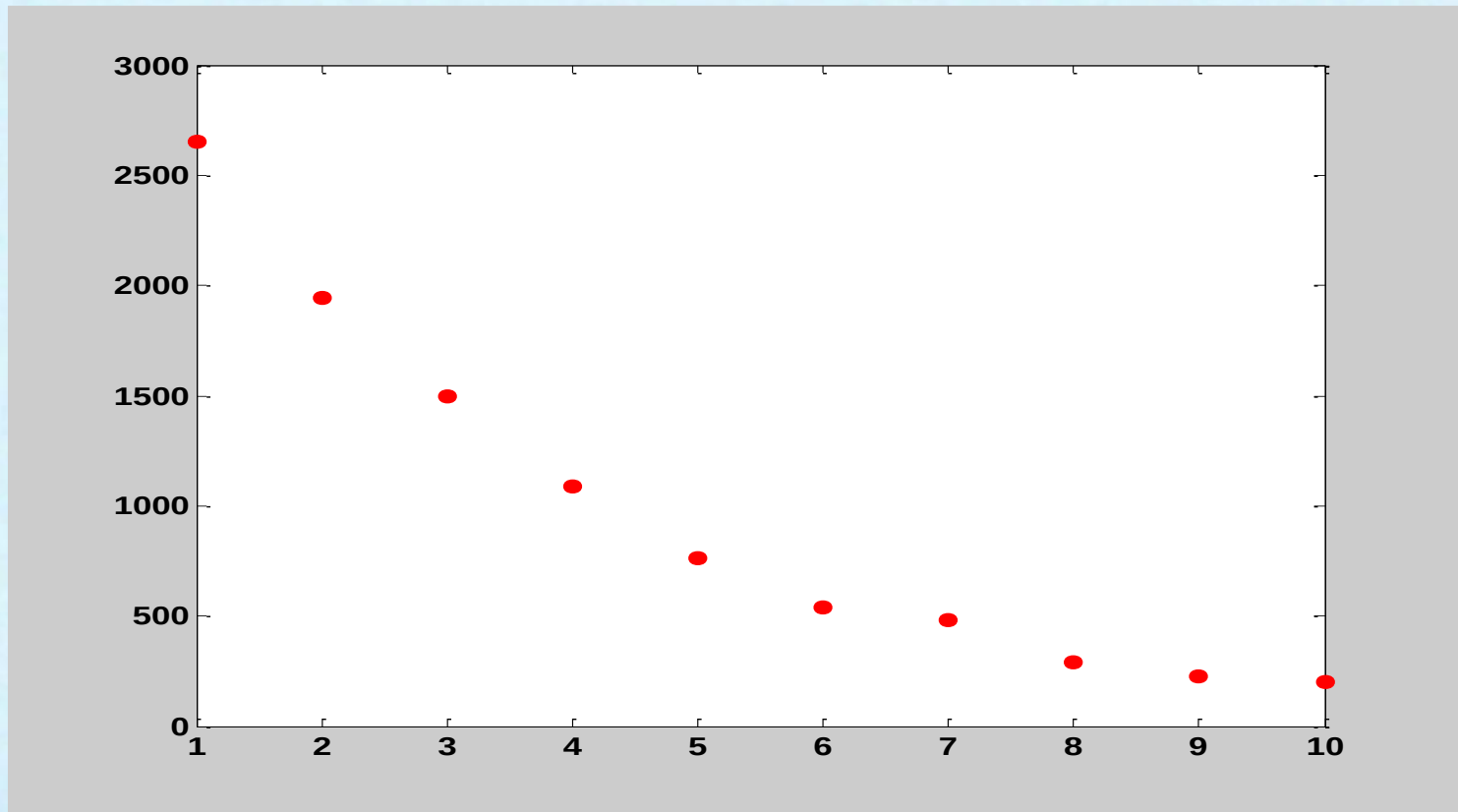
```
x=1:1:10;
```

```
y=[2651,1943,1494,1087,765,538,484,290,226,204];
```

```
plot(x,y,'.r')
```



Y 与 x 呈指数关系,



选择模型 $Y = \alpha e^{\beta x} \cdot \varepsilon, \quad \ln \varepsilon \sim N(0, \sigma^2).$

变量变换 $Y' = a + bx' + \varepsilon', \quad \varepsilon' \sim N(0, \sigma^2).$

其中 $\ln Y = Y', \quad a = \ln \alpha, \quad b = \beta, \quad x' = x, \quad \varepsilon' = \ln \varepsilon$

数据变换后得

$x' = x$	1	2	3	4	5
$y' = \ln y$	7.8827	7.5720	7.3092	6.9912	6.6399

$x' = x$	6	7	8	9	10
$y' = \ln y$	6.2879	6.1821	5.6699	5.4205	5.3181



经计算 $\hat{b} = -0.2977$, $\hat{a} = 8.1646$.

$$\hat{y}' = -0.2977x + 8.1646.$$

线性假设的显著性检验

$$|t| = \frac{\hat{b}}{\hat{\sigma}} \sqrt{S_{xx}} = 32.3693 > t_{0.05/2}(8) = 2.3060.$$

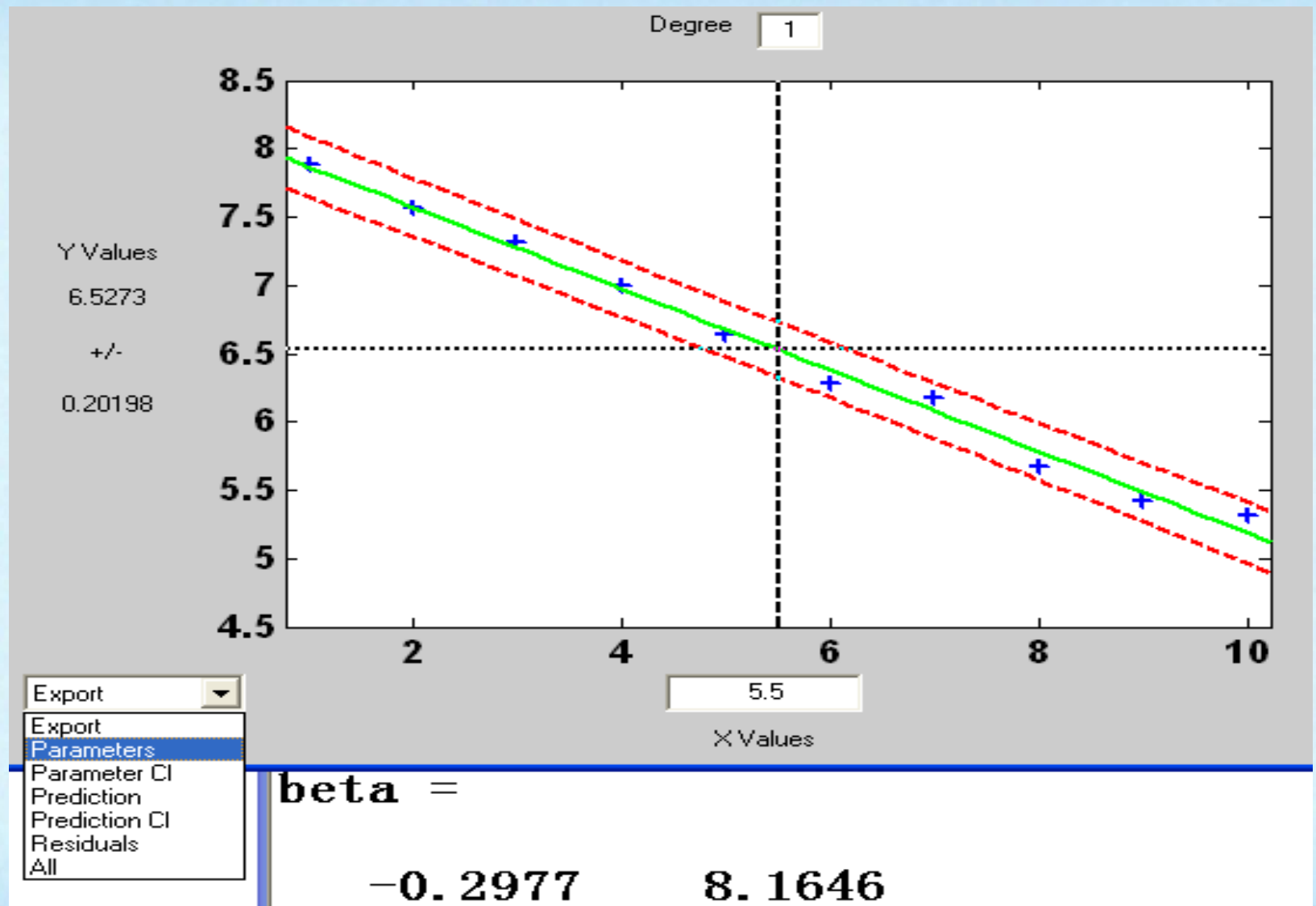
线性回归效果高度显著.

代回原变量, 得曲线回归方程

$$\begin{aligned}\hat{y} &= \exp(\hat{y}') = \exp(-0.2977x + 8.1646) \\ &= 3514.3e^{-0.2977x}.\end{aligned}$$



9.3 一元线性回归



一般情况下, 一元回归模型为

$$Y = \mu(x; \theta_1, \theta_2, \dots, \theta_p) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

其中 $\theta_1, \theta_2, \dots, \theta_p, \sigma^2$ 是与 x 无关的未知参数.

(1) 如果回归函数 $\mu(x; \theta_1, \theta_2, \dots, \theta_p)$ 是参数 $\theta_1, \theta_2, \dots, \theta_p$ 的线性函数 (不必是 x 的线性函数),

则称 $Y = \mu(x; \theta_1, \theta_2, \dots, \theta_p) + \varepsilon$ 为线性回归模型;



(2)如果回归函数 $\mu(x; \theta_1, \theta_2, \dots, \theta_p)$ 是参数 $\theta_1, \theta_2, \dots, \theta_p$ 的非线性函数,

则称 $Y = \mu(x; \theta_1, \theta_2, \dots, \theta_p) + \varepsilon$ 为非线性回归模型

例如 $Y = \theta_1 e^{\theta_2 x} + \varepsilon, \varepsilon \sim N(0, \sigma^2)$

是非线性回归模型.



不能经过变量变换转化为线性回归模型的称为**本质的非线性回归模型**.

例如

$$Y = (\theta_1 + \theta_2 x + \theta_3 x^2)^{-1} + \varepsilon$$

Holliday 模型

$$Y = \frac{\theta_1}{1 + \exp(\theta_2 - \theta_3^x)} + \varepsilon$$

Logistic 模型

四、小结

1.回归分析的任务

研究变量之间的相关关系

2.一元线性回归的步骤

- (1) 推测回归函数; (2) 建立回归模型;
- (3) 估计未知参数; (4) 进行假设检验;
- (5) 预测与控制.

3.可化为一元线性回归的问题

关键:选择适当的变量代换.



第四节 多元线性回归

- 一、多元线性回归的数学模型
- 二、数学模型的分析与求解
- 三、MATLAB中回归分析的实现
- 四、小结



一、多元线性回归的数学模型

实际问题中的随机变量 Y 通常与多个普通变量 $x_1, x_2, \dots, x_p$ ($p > 1$) 有关.

对于自变量 $x_1, x_2, \dots, x_p$ 的一组确定值, Y 具有一定的分布, 若 Y 的数学期望存在, 则它是 $x_1, x_2, \dots, x_p$ 的函数.

$$\mu_{Y|x_1, x_2, \dots, x_p} = \mu(x_1, x_2, \dots, x_p)$$

Y 关于 x 的回归函数



$\mu(x_1, x_2, \dots, x_p)$ 是 $x_1, x_2, \dots, x_p$ 的线性函数.

$$Y = b_0 + b_1 x_1 + \dots + b_p x_p + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

$b_0, b_1, \dots, b_p, \sigma^2$ 是与 $x_1, \dots, x_p$ 无关的未知参数.



二、数学模型的分析与求解

设 $(x_{11}, x_{12}, \dots, x_{1p}, y_1), \dots, (x_{n1}, x_{n2}, \dots, x_{np}, y_n)$
是一个样本.

用最大似然估计法估计参数.

取 $\hat{b}_0, \hat{b}_1, \dots, \hat{b}_p$, 当 $b_0 = \hat{b}_0, b_1 = \hat{b}_1, \dots, b_p = \hat{b}_p$ 时,

$$Q = \sum_{i=1}^n (y_i - b_0 - b_1 x_{i1} - \dots - b_p x_{ip})^2$$

达到最小.



求 Q 分别关于 $b_0, b_1, \dots, b_p$ 的偏导数,

并令它们等于零, 得

$$\left. \begin{aligned} \frac{\partial Q}{\partial b_0} &= -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_{i1} - \dots - b_p x_{ip}) = 0, \\ \frac{\partial Q}{\partial b_j} &= -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_{i1} - \dots - b_p x_{ip}) x_{ij} = 0, \\ & j = 1, 2, \dots, p \end{aligned} \right\}$$

化简可得

$$b_0 n + b_1 \sum_{i=1}^n x_{i1} + b_2 \sum_{i=1}^n x_{i2} + \cdots + b_p \sum_{i=1}^n x_{ip} = \sum_{i=1}^n y_i,$$

$$b_0 \sum_{i=1}^n x_{i1} + b_1 \sum_{i=1}^n x_{i1}^2 + b_2 \sum_{i=1}^n x_{i1} x_{i2} + \cdots + b_p \sum_{i=1}^n x_{i1} x_{ip}$$

$$= \sum_{i=1}^n x_{i1} y_i,$$

...

$$b_0 \sum_{i=1}^n x_{ip} + b_1 \sum_{i=1}^n x_{ip} x_{i1} + b_2 \sum_{i=1}^n x_{ip} x_{i2} + \cdots + b_p \sum_{i=1}^n x_{ip}^2$$

$$= \sum_{i=1}^n x_{ip} y_i.$$

正规方程组



引入矩阵

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad B = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix}.$$

$$X^T X = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ \vdots & \vdots & & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix},$$



$$= \begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \cdots & \sum_{i=1}^n x_{ip} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \cdots & \sum_{i=1}^n x_{i1} x_{ip} \\ \vdots & \vdots & & \vdots \\ \sum_{i=1}^n x_{ip} & \sum_{i=1}^n x_{ip} x_{i1} & \cdots & \sum_{i=1}^n x_{ip}^2 \end{pmatrix}$$



$$X^T Y = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ \vdots & \vdots & & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1} y_i \\ \vdots \\ \sum_{i=1}^n x_{ip} y_i \end{pmatrix}$$

正规方程组的矩阵形式

$$X' X B = X' Y$$



$$\hat{B} = \begin{pmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \vdots \\ \hat{b}_p \end{pmatrix} = (X'X)^{-1} X'Y \quad \text{最大似然估计值}$$

取 $\hat{b}_0 + \hat{b}_1 x_1 + \cdots + \hat{b}_p x_p$ 记成 $\hat{y}$

作为 $\mu(x_1, x_2, \cdots, x_p) = b_0 + b_1 x_1 + \cdots + b_p x_p$ 的估计,

$$\text{方程 } \hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \cdots + \hat{b}_p x_p$$

称为 **P 元经验线性回归方程**, 简称**回归方程**.



三、*MATLAB*中回归分析的实现

多元线性回归

1.确定回归系数的点估计值, 用命令:

b=regress(Y, X)

2.求回归系数的点估计和区间估计, 并检验回归模型, 用命令:

[b, bint, r, rint, stats]=regress(Y, X, alpha)

3.画出残差及其置信区间, 用命令:

rcoplot(r, rint)



符号说明

$$(1) \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$$b = \hat{B} = \begin{pmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \vdots \\ \hat{b}_p \end{pmatrix} = (X'X)^{-1} X'Y.$$

一元线性回归, 取 $p = 1$.



- (2) α 为显著性水平,默认为 0.05;
- (3) $bint$ 为回归系数的区间估计;
- (4) r 与 $rint$ 分别为残差及其置信区间;
- (5) $stats$ 是用于检验回归模型的统计量, 有三个数值,第一个是相关系数 r^2 , 其值越接近于 1, 说明回归方程越显著; 第二个是 F 值, $F > F_{1-\alpha}(p, n-p-1)$ 时, 拒绝 H_0 , F 越大, 说明回归方程越显著; 第三个是与 F 对应的概率 p , $p < \alpha$ 时拒绝, 回归模型成立.

例1 测得16名女子的身高和腿长如下(单位:cm):

身高	143	145	146	147	149	150	153	154
腿长	88	85	88	91	92	93	93	95
身高	155	156	157	158	159	160	162	164
腿长	96	98	97	96	98	99	100	102

试研究这些数据之间的关系.

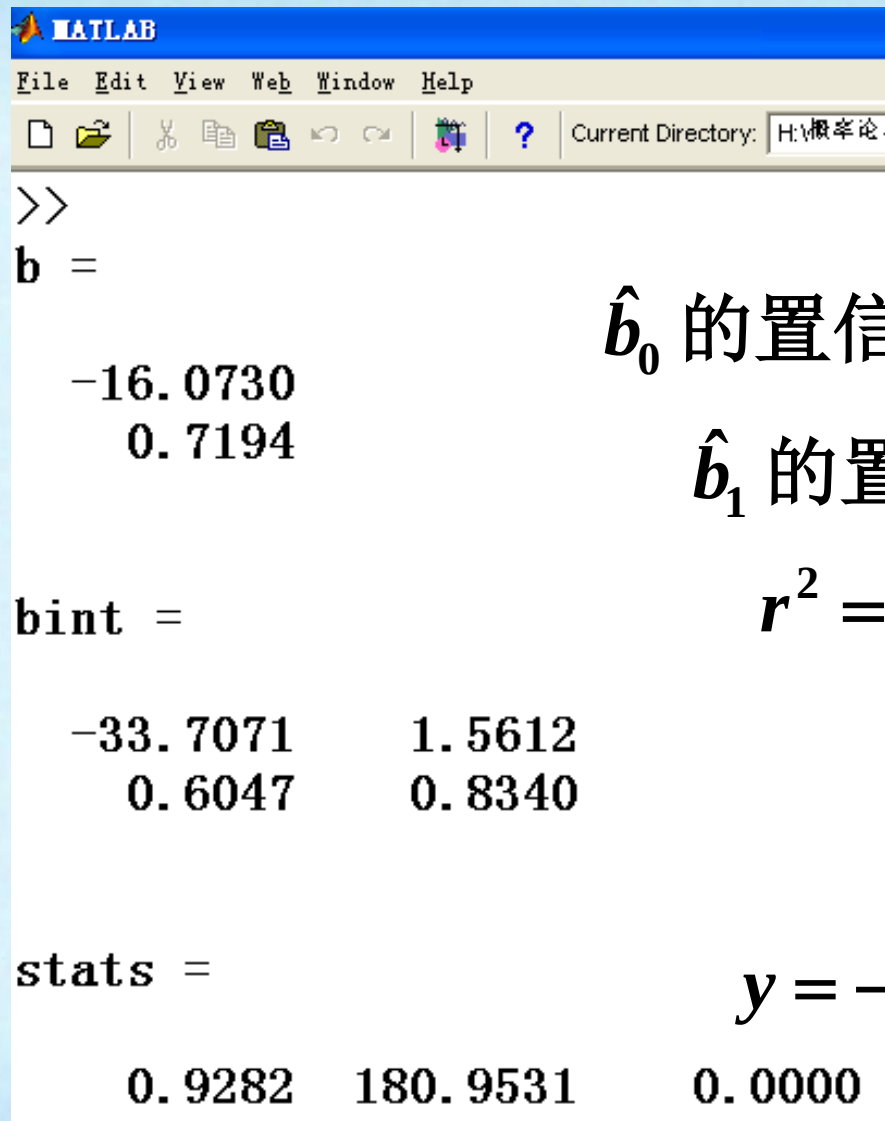
输入数据

```
x=[143,145,146,147,149,150,153,154,155,156,157,  
    158,159,160,162,164]';  
X=[ones(16,1),x];  
Y=[88,85,88,91,92,93,93,95,96,98,97,96,98,99,100,  
    102]';
```

回归分析及检验

```
[b,bint,r,rint,stats]=regress(Y,X);  
b,bint,stats
```





```

MATLAB
File Edit View Web Window Help
[Icons] ? Current Directory: H:\根率论!

>>
b =

    -16.0730
     0.7194

bint =

    -33.7071    1.5612
     0.6047    0.8340

stats =

    0.9282    180.9531    0.0000

```

$$\hat{b}_0 = -16.0730,$$

$$\hat{b}_1 = 0.7194.$$

$\hat{b}_0$ 的置信区间(-33.7071, 1.5612).

$\hat{b}_1$ 的置信区间(0.6047, 0.8340).

$$r^2 = 0.9282, \quad F = 180.9531,$$

$$p = 0.0000.$$

$p < 0.05$, 回归模型

$y = -16.0730 + 0.7194x$ 成立.

残差分析

`rcoplot(r,rint)`

预测及作图

$z=b(1)+b(2)*x$

`plot(x,Y,'k+',x,z,'r')`

数据比较

程序运行结果

残差图形

预测图形

数据比较

帮

助



一元多项式回归

1.确定多项式系数, 用命令:

$[p,S]=polyfit(x,y,m)$

$x = (x_1, x_2, \dots, x_n), \quad y = (y_1, y_2, \dots, y_n).$

$p = (a_1, a_2, \dots, a_{m+1})$ 确定多项式

$y = a_1 x^m + a_2 x^{m-1} + \dots + a_m x + a_{m+1},$

S 是一个矩阵, 用来估计预测误差.

也可使用命令: $polytool(x,y,m)$

结果产生一个交互式的画面, 画面中有拟合曲



线和 y 的置信区间,左下方的 *Export* 可以输出参数.

2.预测和预测误差估计用命令:

$$Y = \text{polyval}(p, x)$$

求回归多项式在 x 处的预测值 Y .

$$[Y, DELTA] = \text{polyconf}(p, x, S, alpha)$$

求回归多项式在 x 处的预测值 Y 以及预测值的显著性为 $1-alpha$ 的置信区间 $Y \pm DELTA$, $alpha$ 的默认值是 0.05.

一元多项式回归可化为多元线性回归求解.



例2 某件产品每件平均单价 Y (元)与批量 x (件)之间的关系的一组数据

x	20	25	30	35	40	50	55	60	65	70	80	90
y	1.81	1.70	1.65	1.55	1.48	1.40	1.30	1.26	1.24	1.21	1.20	1.18

解 $Y = b_0 + b_1x + b_2x^2 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$

令 $x_1 = x, \quad x_2 = x^2,$

$$Y = b_0 + b_1x_1 + b_2x_2 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

$$X = \begin{pmatrix} 1 & 20 & 400 \\ 1 & 25 & 625 \\ 1 & 30 & 900 \\ 1 & 35 & 1225 \\ 1 & 40 & 1600 \\ 1 & 50 & 2500 \\ 1 & 60 & 3600 \\ 1 & 65 & 4225 \\ 1 & 70 & 4900 \\ 1 & 75 & 5625 \\ 1 & 80 & 6400 \\ 1 & 90 & 8100 \end{pmatrix} \quad Y = \begin{pmatrix} 1.81 \\ 1.70 \\ 1.65 \\ 1.55 \\ 1.48 \\ 1.40 \\ 1.30 \\ 1.26 \\ 1.24 \\ 1.21 \\ 1.20 \\ 1.18 \end{pmatrix} \quad B = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix}$$



$$X^T X = \begin{pmatrix} 12 & 640 & 40100 \\ 640 & 40100 & 2779000 \\ 40100 & 2779000 & 204702500 \end{pmatrix}$$

$$(X^T X)^{-1} =$$

$$\frac{1}{\Delta} \begin{pmatrix} 4.8572925 \times 10^{11} & -1.95717 \times 10^{10} & 170550000 \\ -1.95717 \times 10^{10} & 848420000 & -7684000 \\ 170550000 & -7684000 & 71600 \end{pmatrix}$$

$\Delta = 1.41918 \times 10^{11}$, 正规方程组的解为



$$\hat{B} = \begin{pmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \hat{b}_2 \end{pmatrix} = (X'X)^{-1} X'Y = (X^T X)^{-1} \begin{pmatrix} 16.98 \\ 851.3 \\ 51162 \end{pmatrix}$$
$$= \begin{pmatrix} 2.19826629 \\ -0.02252236 \\ 0.00012507 \end{pmatrix}$$

得到回归方程

$$\hat{y} = 2.19826629 - 0.02252236x + 0.00012507x^2$$



试用一元二次多项式进行回归分析，输入数据

$x=[20,25,30,35,40,50,60,65,70,75,80,90];$

$y=[1.81,1.70,1.65,1.55,1.48,1.40,1.30,1.26,1.24,1.21,$
 $1.20,1.18];$

作二次多项式回归

$[p,S]=polyfit(x,y,2)$

预测及作图

$Y=polyconf(p,x,y)$

$plot(x,y,'b+',x,Y,'r')$

程序运行结果

回归结果

残差图形

预测图形

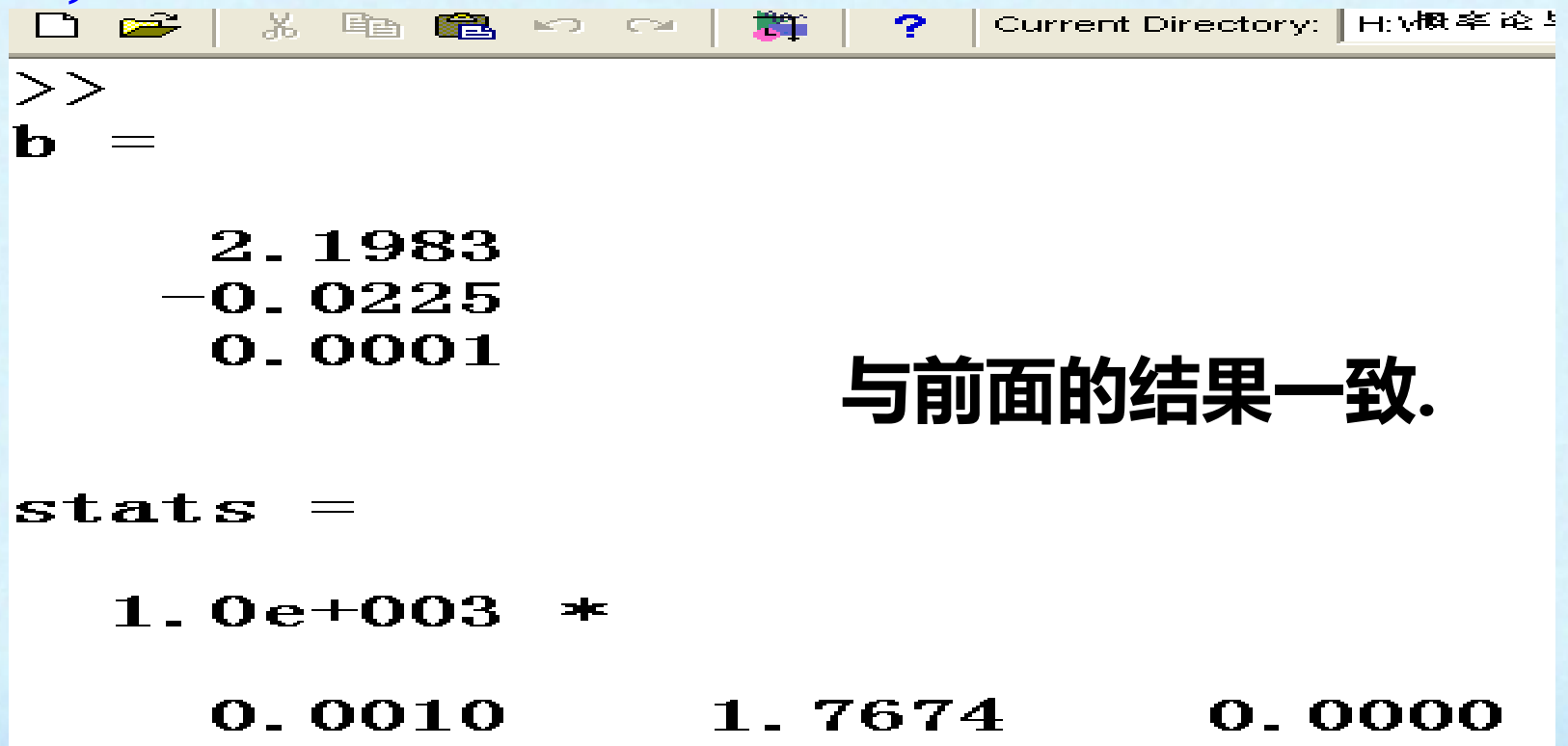
帮

助



化为多元线性回归

```
X=[ones(12,1) x' (x.^2)'];  
[b,bint,r,rint,stats]=regress(y',X);  
b,stats
```



The screenshot shows a MATLAB command window with the following content:

```
>>  
b =  
  
    2.1983  
   -0.0225  
    0.0001  
  
stats =  
  
    1.0e+003 *  
  
    0.0010    1.7674    0.0000
```

The window title bar includes icons for file operations and a status bar showing the current directory as H:\概率论与统计.

与前面的结果一致.

多元二项式回归

rstool(*x*,*y*, 'model', *alpha*)

其中,输入数据 *x*, *y* 分别为 $n \times m$ 矩阵和 n 维列向量; *alpha* 为显著性水平,默认为 0.05; *model* 为下列四种模型中的一种,输入相应的字符串,默认为线性模型.

linear (线性): $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$

purequadratic (纯二次):

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \sum_{j=1}^m \beta_{jj} x_j^2$$

interaction(交叉):

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \sum_{1 \leq j \neq k \leq m}^m \beta_{jk} x_j x_k$$

quadratic(完全二次):

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \sum_{1 \leq j, k \leq m}^m \beta_{jk} x_j x_k$$

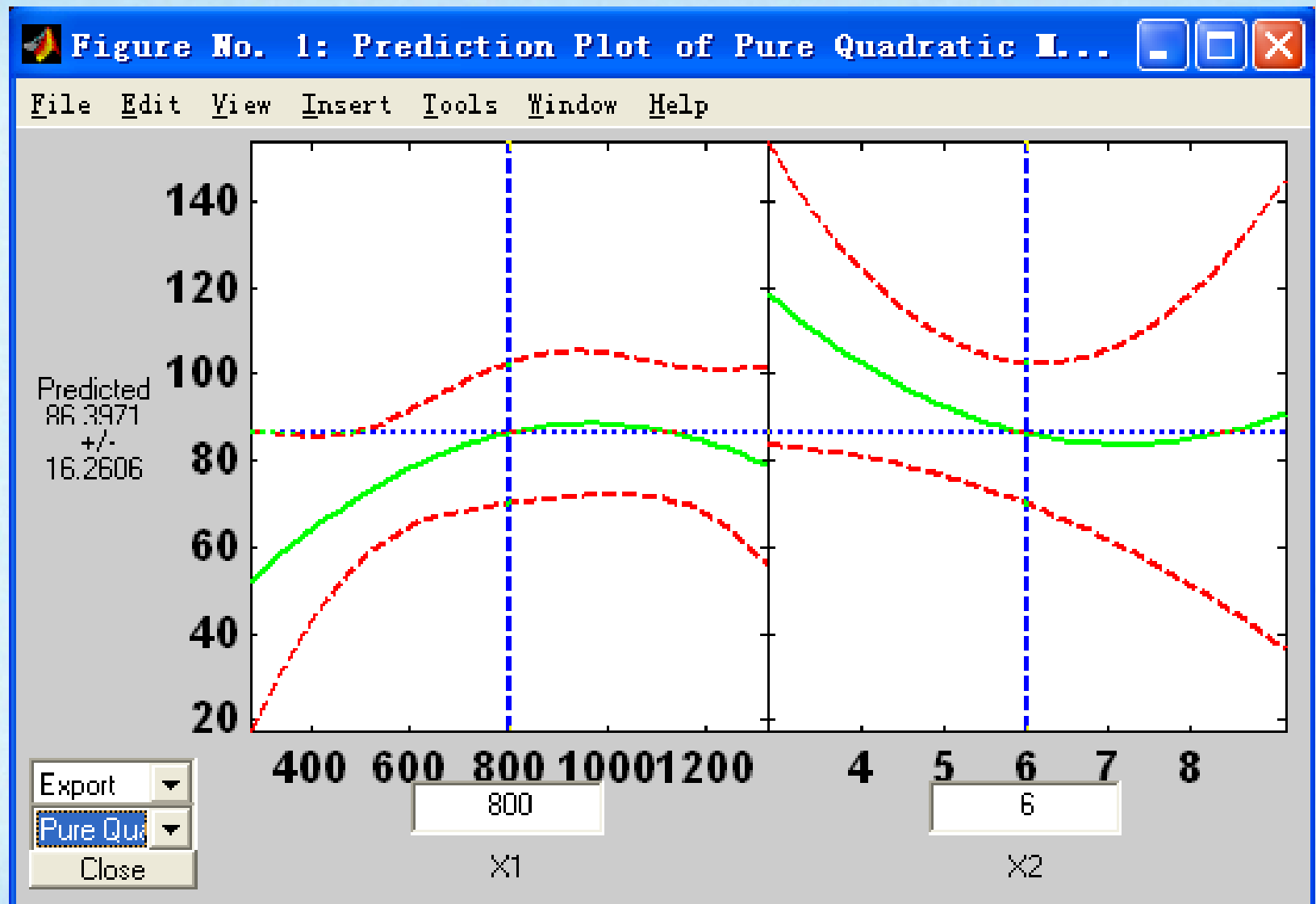
*rstool*的输出是一个交互式画面, 画面中有 m 个图形, 分别给出了一个独立变量 x_i 与 y 的拟合曲线, 以及 y 的置信区间, 此时其余 $m-1$ 个变量取固定值.



可以输入不同的变量的不同值得到 y 的相应值.

图的左下方有两个下拉式菜单, 一个用于传送回归系数、剩余标准差、残差等数据; 另一个用于选择四种回归模型中的一种, 选择不同的回归模型, 其中剩余标准差最接近于零的模型回归效果最好.





例3 设某商品的需求量与消费者的平均收入、商品价格的统计数据如下,建立回归模型,预测平均收入为 1000, 价格为 6 时的商品需求量。

需求量	100	75	80	70	50
收 入	1000	600	1200	500	300
价 格	5	7	6	6	8
需求量	65	90	100	110	60
收 入	400	1300	1100	1300	300
价 格	7	5	4	3	9



选择纯二次模型,即

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2$$

数据输入

```
x1=[1000,600,1200,500,300,400,1300,1100,1300,300];
```

```
x2=[5,7,6,6,8,7,5,4,3,9];
```

```
y=[100,75,80,70,50,65,90,100,110,60]';
```

```
x=[x1' x2'];
```

回归、检验与预测

```
rstool(x,y,'purequadratic')
```

程序运行结果

回归图形

回归结果

帮

助



化为多元线性回归求解

```
x1=[1000,600,1200,500,300,400,1300,1100,1300,300];  
x2=[5,7,6,6,8,7,5,4,3,9];  
y=[100,75,80,70,50,65,90,100,110,60]';  
X=[ones(10,1) x1' x2' (x1.^2)' (x2.^2)'];  
[b,bint,r,rint,stats]=regress(y,X)
```



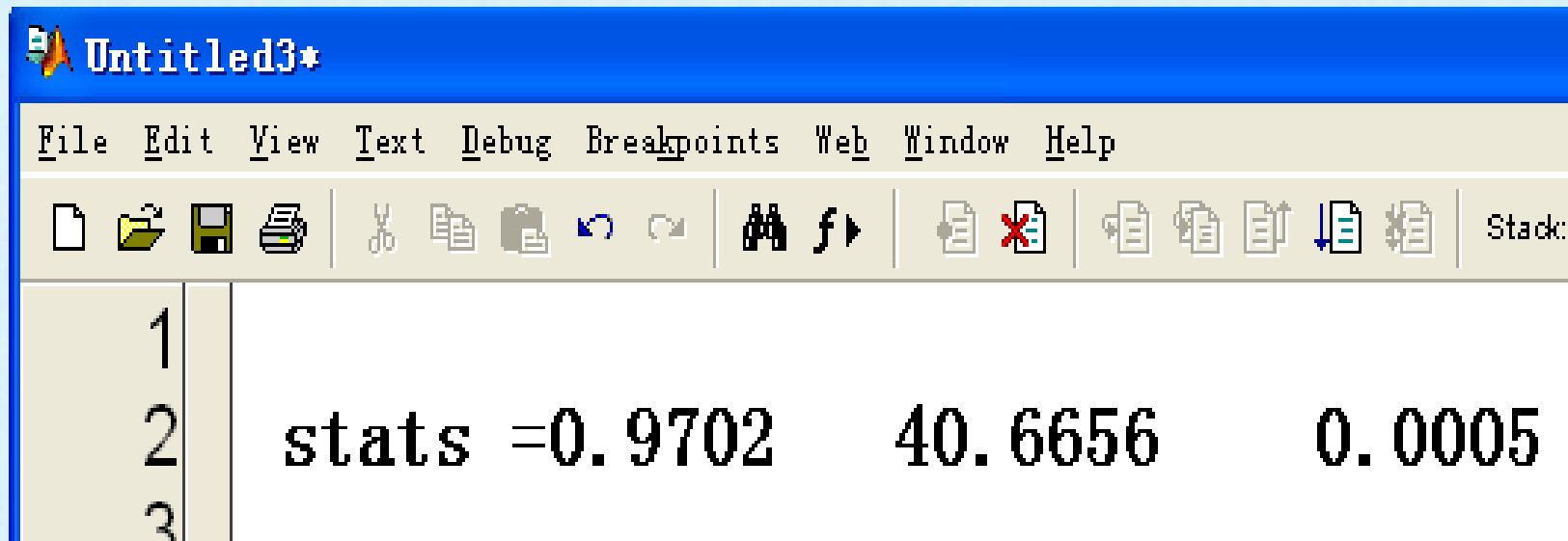
回归系数的点估计以及区间估计

Untitled3*				
File Edit View Text Debug Breakpoints Web Window Help				
                  Stack: Base				
1	b =110.5313 0.1464 -26.5709 -0.0001 1.8475			
2	bint =			
3	57.2602	163.8024		
4	0.0408	0.2521		
5	-43.2247	-9.9171		
6	-0.0001	-0.0000		
7	0.3745	3.3205		

残差及其置信区间

Untitled3*			
File Edit View Text Debug Breakpoints Web Window Help			
                 			
1	r =	rint =	
2	5.2724	-2.8991	13.4438
3	-0.7162	-10.7426	9.3103
4	-4.5158	-11.2788	2.2472
5	-1.9390	-11.3778	7.4997
6	-3.3315	-12.3214	5.6583
7	3.4566	-5.9980	12.9111
8	3.4843	-3.5514	10.5200
9	-3.4452	-13.0340	6.1437
10	-0.0976	-6.3831	6.1878
11	1.8320	-3.3221	6.9862

检验回归模型的统计量



The screenshot shows a software window titled 'Untitled3*' with a menu bar (File, Edit, View, Text, Debug, Breakpoints, Web, Window, Help) and a toolbar. Below the toolbar, a table displays regression statistics. The table has three columns: the first column contains line numbers 1, 2, and 3; the second column contains the text 'stats ='; the third column contains the value '0.9702'; the fourth column contains the value '40.6656'; and the fifth column contains the value '0.0005'.

1				
2	stats =	0.9702	40.6656	0.0005
3				

相关系数 $r^2 = 0.9702$, 接近于1, 回归方程显著;

$F = 40.6656 > F_{0.95}(4,5) = 6.26$, 回归方程显著;

$P = 0.0005 < \alpha = 0.05$, 回归模型成立

逐步回归分析

在实际问题中,影响因变量的因素很多,而这些因素之间可能存在多重共线性. 为得到可靠的回归模型,需要一种方法能有效地从众多因素中挑选出对因变量贡献大的因素.

如果采用多元线性回归分析,回归方程稳定性差,每个自变量的区间误差积累将影响总体误差,预测的可靠性差、精度低;另外,如果采用了影响小的变量,遗漏了重要变量,可能导致估计量产生偏倚和



不一致性.

“最优” 的回归方程应该包含所有有影响的变量而不包括影响不显著的变量.

选择 **“最优”** 回归方程的方法

- 1.从所有可能的变量组合的回归方程中选择最优者;
- 2.从包含全部变量的回归方程中逐次剔除不显著因子;
- 3.从一个变量开始,把变量逐个引入方程;



4. “有进有出”的逐步回归分析.

逐步回归分析法在筛选变量方面比较理想,是目前较常用的方法.它从一个自变量开始,根据自变量作用的显著程度,从大到小地依次逐个引入回归方程,但当引入的自变量由于后面变量的引入而变得不显著时,要将其剔除掉.引入一个自变量或从回归方程中剔除一个自变量,为逐步回归的一步,对于每一步,都进行检验,以确保每次引入新的显著性变量前回归方程中只包含作用显著的变量.



反复进行上面的过程,直到没有不显著的变量从回归方程中剔除,也没有显著变量可引入到回归方程.



函数: *stepwise*

用法: *stepwise(x,y,inmodel,alpha)*

符号说明:

x—自变量数据, 为 $n \times m$ 矩阵;

y—因变量数据, 为 $n \times 1$ 矩阵;

inmodel—由矩阵*x*列的指标构成, 表明初始模型中引入的自变量, 默认为全部自变量;

alpha—判断模型中每一项显著性的指标, 默认相当于对回归系数给出95%的置信区间.



例4 水泥凝固时放出的热量 y 与水泥中的四种化学成分 x_1, x_2, x_3, x_4 有关, 今测得一组数据如下, 试用逐步回归法确定一个线性模型.

序号	1	2	3	4	5	6	7
x_1	7	1	11	11	7	11	3
x_2	26	29	56	31	52	55	71
x_3	6	15	8	8	6	9	17
x_4	60	52	20	47	33	22	6
y	78.5	74.3	104.3	87.6	95.9	109.2	102.7
序号	8	9	10	11	12	13	
x_1	1	2	21	1	11	10	
x_2	31	54	47	40	66	68	
x_3	22	18	4	23	9	8	
x_4	44	22	26	34	12	12	
y	72.5	93.1	115.9	83.8	113.3	109.4	



输入数据

```
x1=[7,1,11,11,7,11,3,1,2,21,1,11,10]';  
x2=[26,29,56,31,52,55,71,31,54,47,40,66,68]';  
x3=[6,15,8,8,6,9,17,22,18,4,23,9,8]';  
x4=[60,52,20,47,33,22,6,44,22,26,34,12,12]';  
y=[78.5,74.3,104.3,87.6,95.9,109.2,102.7,72.5,93.1,  
    115.9,83.8,113.3,109.4]';  
x=[x1,x2,x3,x4];
```

逐步回归分析

```
stepwise(x,y)
```

程序运行结果

逐步回归

回归平面

帮

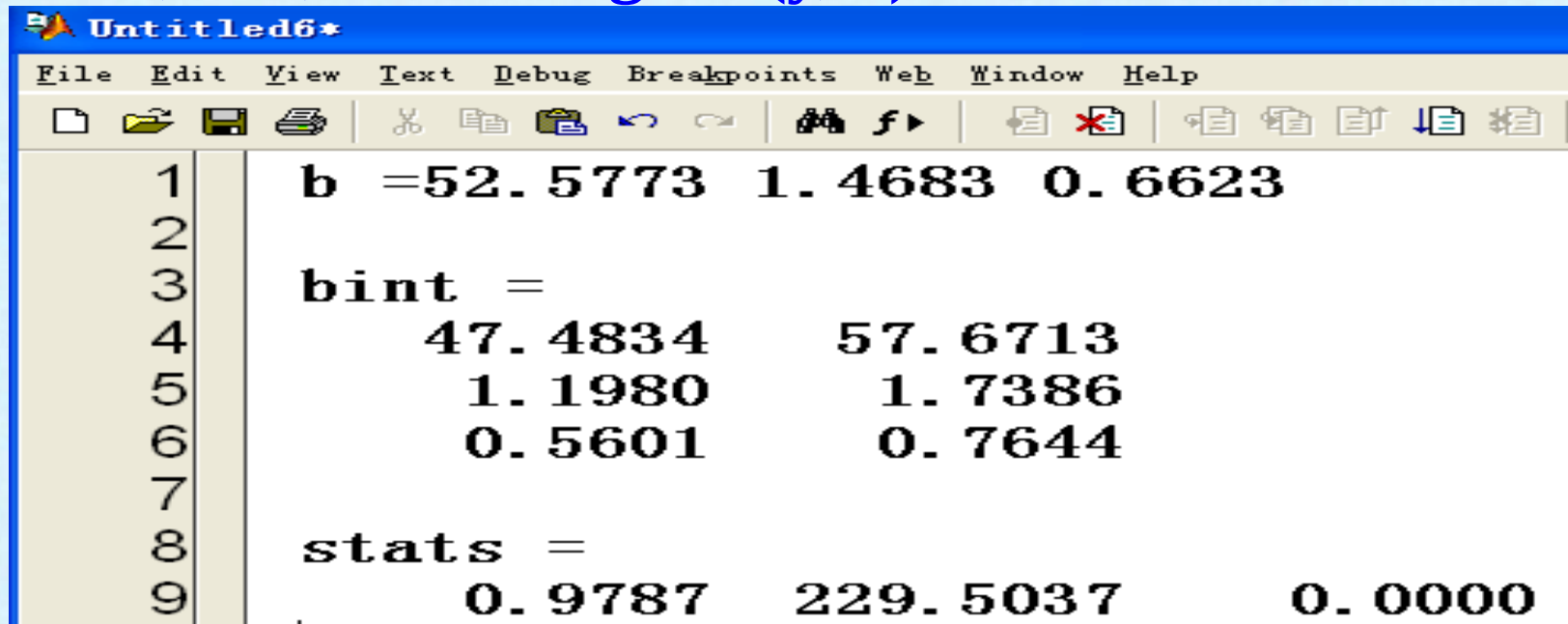
助



对变量 y 和 x_1, x_2 作线性回归.

$X=[\text{ones}(13,1),x1,x2];$

$[b,\text{bint}, \text{stats}]=\text{regress}(y,X)$



```
Untitled6*
File Edit View Text Debug Breakpoints Web Window Help
b = 52.5773 1.4683 0.6623
bint =
    47.4834    57.6713
    1.1980    1.7386
    0.5601    0.7644
stats =
    0.9787   229.5037    0.0000
```

回归模型为 $y = 52.5773 + 1.4683x_1 + 0.6623x_2$,
三个统计量表明: 回归效果显著.

四、小结

1.多元线性回归的数学模型

$$Y = b_0 + b_1 x_1 + \cdots + b_p x_p + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

$b_0, b_1, \cdots, b_p, \sigma^2$ 是与 $x_1, \cdots, x_p$ 无关的未知参数.

2.数学模型的分析与求解

$$\hat{B} = (X' X)^{-1} X' Y,$$

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \cdots + \hat{b}_p x_p.$$



3. MATLAB中回归分析的实现

(1)多元线性回归

$b = \text{regress}(Y, X)$

(2)一元多项式回归

$[p, S] = \text{polyfit}(x, y, m)$

(3)多元二项式回归

$\text{rstool}(x, y, 'model', \alpha)$

(4)逐步回归分析

$\text{stepwise}(x, y, \text{inmodel}, \alpha)$

