第九章 方差分析和回归分析

- 单因素试验的方差分析
- 双因素试验的方差分析
- 一元线性回归
- 多元线性回归

§9.1 单因素试验的方差分析

- 方差分析(Analysis of variance, 简称:ANOVA)，是由英国统计学家费歇尔(Fisher)在20世纪20年代提出的，可用于推断两个或两个以上总体均值是否有差异的显著性检验.
- 方差分析就是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.

- 在方差分析中, 通常把研究对象的特征值, 即所考察的试验结果称为试验指标.
- 对试验指标产生影响的原因称为因素.
- 因素中各个不同状态称为水平.
- 如果在一项试验的过程中只有一个因素在改变称为单因素试验，如果多于一个因素在改变称为多因素试验.

例1 为了比较三种不同类型日光灯管的寿命(小时), 现将从每种类型日光灯管中抽取 **8**个，总共 **24** 个日光灯管进行老化试验，根据下面经老化试验后测算得出的各个日光灯管的寿命(小时)，试判断三种不同类型日光灯管的寿命是不是有存在差异.

日光灯管的寿命(小时)

类型	寿命(小时)							
类型I	5290	6210	5740	5000	5930	6120	6080	5310
类型II	5840	5500	5980	6250	6470	5990	5470	5840
类型III	7130	6660	6340	6470	7580	6560	7290	6730

引起日光灯管寿命不同的原因有二个方面:

- 由于日光灯类型不同,而引起寿命不同.
- 同一种类型日光灯管,由于其它随机因素的影响, 也使其寿命不同.

本例中,

- 试验指标：日光灯管的寿命.
- 因素：日光灯管类型
- 水平：日光灯管三个不同的类型

§9.1.1 单因素试验

- 单因素方差分析：仅考虑有一个因素A对试验指标的影响. 假如因素 A有r个水平, 分别在第 i 水平下进行了多次独立观测, 所得到的试验指标的数据

$A_1: N(\mu_1, \sigma^2)$	$A_2: N(\mu_2, \sigma^2)$	$\cdots$	$A_r: N(\mu_r, \sigma^2)$
X_{11}	X_{12}	$\cdots$	X_{1r}
X_{21}	X_{22}	$\cdots$	X_{2r}
$\vdots$	$\vdots$	$\cdots$	$\vdots$
$X_{n_1 1}$	$X_{n_2 2}$	$\cdots$	$X_{n_r r}$

假定：各个水平A_j(j=1,2,…,r)下的样本$X_{1j}, X_{2j}, \cdots, X_{n_j j}$来自具有相同方差$\sigma^2$，均值分别为$\mu_j$(j=1,2,…,r)的正态总体$N(\mu_j, \sigma^2)$，$\mu_j$与$\sigma^2$未知.且设不同水平$A_j$下的样本之间相互独立. 因此, 可写成如下的数学模型:

$$\left.\begin{array}{c} X_{ij} = \mu_j + \varepsilon_{ij} \\ \varepsilon_{ij} \sim N(0, \sigma^2), \text{各}\varepsilon_{ij}\text{独立} \\ i = 1, 2, \cdots, n_j, \ \ j = 1, 2, \cdots, r \end{array}\right\}$$

单因素试验方差分析的数学模型

方差分析的目的就是要比较因素A的r个水平下试验指标理论均值的差异, 问题可归结为比较这r个总体的均值差异.

（1）检验r个正态总体$N(\mu_j, \sigma^2)$ (j=1,2,…,r)的均值是否相等，即

检验假设　$H_0: \mu_1 = \mu_2 = ... = \mu_r$

$H_1: \mu_1, \mu_2, ..., \mu_r$不全相等。

（2）作出未知参数μ_j, σ^2的估计.

记 $\mu=\frac{1}{n}\sum_{j=1}^{r}n_j\mu_j$ ——总平均， 其中$\sum_{j=1}^{r}n_j=n$

$\delta_j=\mu_j-\mu$——水平A_j的效应, $j=1,2,...,r$

此时有 $n_1\delta_1+n_2\delta_2+...+n_r\delta_r=0$

模型为：$X_{ij}=\mu+\delta_j+\varepsilon_{ij}$

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

$i=1,2,\cdots,n_j,\ \ j=1,2,\cdots,r$

$n_1\delta_1+n_2\delta_2+...+n_r\delta_r=0$

假设等价于

$H_0:\delta_1=\delta_2=\cdots=\delta_r=0$

$H_1:\delta_1,\delta_2,\cdots,\delta_r$不全为零。

§9.1.2 平方和的分解

为给出上面的检验，主要采用的方法是平方和分解。即假设数据总的差异用总偏差平方和S_T表示，可分解为二个部分：

一部分是由于因素A引起的差异，即效应平方和S_A；

另一部分则由随机误差所引起的差异，即误差平方和S_E.

总偏差平方和 $S_T = \sum_{j=1}^{r}\sum_{i=1}^{n_j}\left(X_{ij} - \overline{X}\right)^2$

效应平方和 $S_A = \sum_{j=1}^{r} n_j \left(\overline{X}_{\bullet j} - \overline{X}\right)^2$

$$= \sum_{j=1}^{r} n_j \overline{X}_{\bullet j}^{\,2} - n\overline{X}^2$$

误差平方和 $S_E = \sum_{j=1}^{r}\sum_{i=1}^{n_j}\left(X_{ij} - \overline{X}_{\bullet j}\right)^2$

其中$\overline{X}._j$为水平A_j下的样本平均值，即 $\overline{X}._j = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij}$

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

性质1：$S_T = S_A + S_E$

证明：
$$S_T = \sum_{j=1}^{r}\sum_{i=1}^{n_j}\left(X_{ij} - \bar{X}\right)^2 = \sum_{j=1}^{r}\sum_{i=1}^{n_j}\left(X_{ij} - \bar{X}_{\bullet j} + \bar{X}_{\bullet j} - \bar{X}\right)^2$$

$$= \sum_{j=1}^{r}\sum_{i=1}^{n_j}\left(X_{ij} - \bar{X}_{\bullet j}\right)^2 + \sum_{j=1}^{r}\sum_{i=1}^{n_j}\left(\bar{X}_{\bullet j} - \bar{X}\right)^2 + 2\sum_{j=1}^{r}\sum_{i=1}^{n_j}\left(X_{ij} - \bar{X}_{\bullet j}\right)\left(\bar{X}_{\bullet j} - \bar{X}\right)$$

$$= S_A + S_E$$

$$\sum_{j=1}^{r}\sum_{i=1}^{n_j}\left(X_{ij} - \bar{X}_{\bullet j}\right)\left(\bar{X}_{\bullet j} - \bar{X}\right) = \sum_{j=1}^{r}\left(\bar{X}_{\bullet j} - \bar{X}\right)\sum_{i=1}^{n_j}\left(X_{ij} - \bar{X}_{\bullet j}\right) = 0$$

§9.1.3 S_E，S_A的统计特性

性质2：$E(S_T)=\sum_{j=1}^{r} n_j\delta_j^2+(n-1)\sigma^2$

$$E(S_A)=\sum_{j=1}^{r} n_j\delta_j^2+(r-1)\sigma^2$$

$$E(S_E)=(n-r)\sigma^2$$

证明：各个水平A_j(j=1,2,…,r)下的样本$X_{1j}, X_{2j}, \ldots, X_{n_j j}$

服从正态总体$N(\mu_j, \sigma^2)$，即 $X_{ij} \sim N(\mu_j, \sigma^2)$

则A_j下的样本均值$\overline{X}._j \sim N(\mu_j, \sigma^2/n_j)$

由正态分布线性可加性知

$$\sum_{j=1}^{r} n_j \overline{X}._j \sim N(\sum_{j=1}^{r} n_j \mu_j, \sum_{j=1}^{r} n_j \sigma^2)$$

$$\overline{X} = \frac{1}{n}\sum_{j=1}^{r} n_j \overline{X}._j \sim N(\frac{1}{n}\sum_{j=1}^{r} n_j \mu_j, \frac{1}{n^2}\sum_{j=1}^{r} n_j \sigma^2)$$

即 $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

记$\mu = \frac{1}{n}\sum_{j=1}^{r} n_j \mu_j$，$n = \sum_{j=1}^{r} n_j$

证明：

$$E(S_T)=E\left(\sum_{j=1}^{r}\sum_{i=1}^{n_j}(X_{ij}-\bar{X})^2\right):$$

$$=E\left(\sum_{j=1}^{r}\sum_{i=1}^{n_j}X_{ij}{}^2-n\bar{X}^2\right)=\sum_{j=1}^{r}\sum_{i=1}^{n_j}E(X_{ij}{}^2)-nE(\bar{X}^2)$$

$$=\sum_{j=1}^{r}\sum_{i=1}^{n_j}[\sigma^2+(\mu+\delta_j)^2]-n[\frac{\sigma^2}{n}+\mu^2]$$

$$=n\sigma^2+n\mu^2+2\mu\underbrace{\sum_{j=1}^{r}n_j\delta_j}_{=0}+\sum_{j=1}^{r}n_j\delta_j{}^2-\sigma^2-n\mu^2$$

$$=\sum_{j=1}^{r}n_j\delta_j^2+(n-1)\sigma^2$$

$$E(S_E) = \sum_{j=1}^{r} E\left\{ \sum_{i=1}^{n_j} \left(X_{ij} - \bar{X}_{\bullet j} \right)^2 \right\}$$

$$= \sum_{j=1}^{r} (n_j - 1)\sigma^2 = (n - r)\sigma^2$$

$$E(S_A) = E(S_T - S_E) = \sum_{j=1}^{r} n_j \delta_j^2 + \left(r - 1 \right)\sigma^2$$

性质3：

$$(1)\frac{S_E}{\sigma^2}\sim\chi^2(n-r);$$

$$(2)\text{当}H_0\text{为真时},\frac{S_A}{\sigma^2}\sim\chi^2(r-1);$$

且有，S_A与S_E相互独立。

(3) 当H_0为真时，$F=\dfrac{S_A/(r-1)}{S_E/(n-r)}\sim F(r-1,n-r).$

§9.1.4 假设检验问题的拒绝域

当H_0为真时，$$F=\frac{S_A/(r-1)}{S_E/(n-r)}\sim F(r-1,n-r).$$

且 $$E(\frac{S_A}{r-1})=\sigma^2+\frac{1}{(r-1)}\sum_{j=1}^{r}n_j\delta_j^2=\sigma^2 \qquad E(\frac{S_E}{n-r})=\sigma^2$$

拒绝域具有形式 $$F=\frac{S_A/(r-1)}{S_E/(n-r)}\geq k$$

检验的拒绝域为

$$F=\frac{S_A/(r-1)}{S_E/(n-r)}\geq F_\alpha(r-1,n-r)$$

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	r-1	$\overline{S_A}=\frac{S_A}{r-1}$	$F=\frac{\overline{S_A}}{\overline{S_E}}$
误差	S_E	n-r	$\overline{S_E}=\frac{S_E}{n-r}$	
总和	S_T	n-1		

计算S_T, S_A, S_E的简便公式：

$$记 T_{\bullet j} = \sum_{i=1}^{n_j} X_{ij},\ \ j = 1, 2, \cdots, r, \quad T_{\bullet\bullet} = \sum_{j=1}^{r}\sum_{i=1}^{n_j} X_{ij}$$

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

$$S_A = \sum_{j=1}^{r} n_j \bar{X}_{\bullet j}^{\ 2} - n\bar{X}^2 = \sum_{j=1}^{r} \frac{T_{\bullet j}^2}{n_j} - \frac{T_{\bullet\bullet}^{\ 2}}{n}$$

$$S_E = S_T - S_A$$

例1 设有5种治疗荨麻疹的药，要比较它们的疗效。假设将30个病人分成5组，每组6人，令同组病人使用一种药，并记录病人从使用药物开始到痊愈所需时间，得到下面的记录：(α=0.05)

药物类型	治愈所需天数x
1	5，8，7，7，10，8
2	4，6，6，3，5，6
3	6，4，4，5，4，3
4	7，4，6，6，3，5
5	9，3，5，7，7，6

分析：这里药物是因素，共有5个水平，这是一个单因素方差分析问题，要检验的假设是“所有药物的效果都没有差别”。

解：检验假设 $H_0:\mu_1=\mu_2=\mu_3=\mu_4=\mu_5$

$H_1:\mu_1,\mu_2,...,\mu_5$不全相等。

$$r=5,n_1=n_2=n_3=n_4=n_5=6,n=30,$$

$$\sum_{j=1}^{r}\sum_{i=1}^{n_j}X_{ij}^{\ 2}=1047,\quad T_{\cdot 1}=45,T_{\cdot 2}=30,T_{\cdot 3}=26,T_{\cdot 4}=31,$$

$$T_{\cdot 5}=37,T_{\cdot\cdot}=169$$

方差来源	平方和	自由度	均方	F比
因素A	36.467	4	9.117	3.90
误差	58.500	25	2.334	
总和	94.967	29		

因$F_{0.05}(4,25)= 2.76<3.90$，故在显著性水平0.05下拒绝$H_0$，认为疗效有显著差异.

例2 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.

机器1	机器2	机器3
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($\alpha=0.05$)：

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

H_1: μ_1、μ_2、μ_3不全相等.

§9.1.5 未知参数的估计

(1) σ^2的估计$\hat{\sigma}^2 = \dfrac{S_E}{n-r}$；　(2)$\mu$的估计$\hat{\mu} = \bar{X}$；

(3)μ_j的估计$\hat{\mu}_j = \bar{X}_{\bullet j}$；　　(4)$\delta_j$的估计$\hat{\delta}_j = \bar{X}_{\bullet j} - \bar{X}$。

容易证明，以上估计均为相应参数的无偏估计。

当拒绝H_0时，进一步比较$N(\mu_j, \sigma^2)$和$N(\mu_k, \sigma^2)$的差异，可以作　$\mu_j - \mu_k = \delta_j - \delta_k (j \neq k)$的区间估计。

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

且$\bar{X}_{\bullet j}-\bar{X}_{\bullet k}$与$\hat{\sigma}^2=MS_E$相互独立。

故$\dfrac{(\bar{X}_{\bullet j}-\bar{X}_{\bullet k})-(\mu_j-\mu_k)}{\sqrt{MS_E(1/n_j+1/n_k)}}=\dfrac{(\bar{X}_{\bullet j}-\bar{X}_{\bullet k})-(\mu_j-\mu_k)}{\sigma\sqrt{(1/n_j+1/n_k)}}\Big/\sqrt{\dfrac{S_E}{\sigma^2}\Big/(n-r)}\sim t(n-r)$

得$(\mu_j-\mu_k)$的水平为$1-\alpha$的置信区间 $\left(\bar{X}_{\bullet j}-\bar{X}_{\bullet k}\pm t_{\alpha/2}(n-r)\sqrt{MS_E(1/n_j+1/n_k)}\right)$

例3 求例1中未知参数$\sigma^2, \mu_j, \delta_j (j=1,2,3,4,5)$的点估计，并求$\mu_1-\mu_3$，$\mu_1-\mu_2$，$\mu_3-\mu_5$的置信度为0.95的置信区间。

药物类型	治愈所需天数x
1	5，8，7，7，10，8
2	4，6，6，3，5，6
3	6，4，4，5，4，3
4	7，4，6，6，3，5
5	9，3，5，7，7，6

解：σ^2的估计$\hat{\sigma}^2 = \dfrac{S_E}{n-r} = 2.3334$；$\mu$的估计$\hat{\mu} = \bar{X} = 5.6333$；

μ_j的估计分布为:7.5，5，4.3333，5.1667，6.1667；

δ_j的估计分布为:1.8667，-0.6333，-1.3，-0.4666，0.5334

查表得$t_{0.025}(25) = 2.0595$, $\sqrt{MS_E(1/n_j + 1/n_k)} = 0.8819$

$\mu_1 - \mu_3$，$\mu_1 - \mu_2$，$\mu_3 - \mu_5$的置信度为0.95的置信区间分别为：

(1.3504，4.983)，(0.6837，4.3163)，(-3.6497，-0.0171)

说明μ_1与μ_3，μ_1与μ_2，μ_3与μ_5的差异都显著。

例4 求例2中未知参数σ^2，μ_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.

机器1	机器2	机器3
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

§9.2 双因素试验的方差分析

例如 检验a,b两种药物的抗癌效果，要做动物试验。
作法是将患有某种癌的白鼠随机地分成三组。
第一组：注射a物质；第二组：注射b物质；第三组：不做处理.
经过一段时间观察后，得到寿命数据。

在试验中，考虑白鼠的性别有可能对其寿命有显著的影响。
将“性别”作为另一个因素——“双因素试验”。
因素A：药物，三个水平；因素B：性别，二个水平；
两个因素共有2×3＝6种组合。

§9.2.1 等重复试验

因素A有r个水平$A_1, A_2, \cdots, A_r$，因素B有s个水平$B_1, B_2, \cdots, B_s$.

现对因素A，B的水平的每对组合$(A_i, B_j)\left(i=1,\cdots,r; j=1,\cdots,s\right)$

都作$t(t \geq 2)$次试验（称为等重复试验），得到如下结果：

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

$$
\left.
\begin{array}{c}
\text{设} \quad X_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \\
\varepsilon_{ijk} \sim N(0, \sigma^2), \text{各}\varepsilon_{ijk}\text{独立}, \\
i = 1, \ldots, r, j = 1, \ldots, s, k = 1, \ldots, t. \\
\mu_{ij}, \sigma^2 \text{均为未知参数.}
\end{array}
\right\}
$$

记 $\mu = \dfrac{1}{rs}\sum_{i=1}^{r}\sum_{j=1}^{s}\mu_{ij}$，——总平均　$\mu_{i\bullet} = \dfrac{1}{s}\sum_{j=1}^{s}\mu_{ij}, i = 1, \ldots, r,$

$$\mu_{\bullet j} = \frac{1}{r}\sum_{i=1}^{r}\mu_{ij}, j = 1, \ldots, s,$$

$\alpha_i = \mu_{i\bullet} - \mu$，——水平$A_i$的效应，$i = 1, \ldots, r,$

$\beta_j = \mu_{\bullet j} - \mu$，——水平$B_j$的效应，$j = 1, \ldots, s.$

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

$$\mu_{ij}=\mu+\alpha_i+\beta_j+(\mu_{ij}-\mu_{i\bullet}-\mu_{\bullet j}+\mu), i=1,\ldots,r,\ j=1,\ldots,s.$$

记 $\gamma_{ij}=\mu_{ij}-\mu_{i\bullet}-\mu_{\bullet j}+\mu$, ——水平$A_i$和水平$B_j$的交互效应,

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

模型可写成 $X_{ijk}=\mu+\alpha_i+\beta_j+\gamma_{ij}+\varepsilon_{ijk}$,

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

$i=1,\ldots,r,\ j=1,\ldots,s,k=1,\ldots,t.$

$$\sum_{i=1}^{r}\alpha_i=0,\sum_{j=1}^{s}\beta_j=0,\sum_{i=1}^{r}\gamma_{ij}=0,\sum_{j=1}^{s}\gamma_{ij}=0.$$

$\mu,\alpha_i,\beta_j,\gamma_{ij},\sigma^2$均未知.

分别检验假设：

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

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

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

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

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

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

$$\bar{X}_{\bullet j\bullet}=\frac{1}{rt}\sum_{i=1}^{r}\sum_{k=1}^{t}X_{ijk},j=1,\ldots,s.$$

总偏差平方和（总变差）$S_T=\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}\left(X_{ijk}-\bar{X}\right)^2$

误差平方和 $S_E=\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}\left(X_{ijk}-\bar{X}_{ij\bullet}\right)^2$

因素A的效应平方和 $S_A=st\sum_{i=1}^{r}\left(\bar{X}_{i\bullet\bullet}-\bar{X}\right)^2$

因素B的效应平方和 $S_B=rt\sum_{j=1}^{s}\left(\bar{X}_{\bullet j\bullet}-\bar{X}\right)^2$

A,B交互效应平方和 $S_{AB}=t\sum_{i=1}^{r}\sum_{j=1}^{s}\left(\bar{X}_{ij\bullet}-\bar{X}_{i\bullet\bullet}-\bar{X}_{\bullet j\bullet}+\bar{X}\right)^2$

性质：　(1) $S_T = S_A + S_B + S_{A\times B} + S_E$

$$(2)E\left(\frac{S_E}{rs(t-1)}\right)=\sigma^2,$$

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

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

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

当H_{01}成立时，$F_A=\dfrac{S_A/(r-1)}{S_E/(rs(t-1))}\sim F\left(r-1,rs\left(t-1\right)\right)$

拒绝域为：$W_A=\{F_A\geq F_\alpha\left(r-1,rs\left(t-1\right)\right)\}$

当H_{02}成立时，$F_B=\dfrac{S_B/(s-1)}{S_E/(rs(t-1))}\sim F\left(s-1,rs\left(t-1\right)\right)$

拒绝域为：$W_B=\{F_B\geq F_\alpha\left(s-1,rs\left(t-1\right)\right)\}$

当H_{03}成立时，$F_{AB}=\dfrac{S_{AB}/((r-1)(s-1))}{S_E/(rs(t-1))}\sim F\left((r-1)(s-1),rs\left(t-1\right)\right)$

拒绝域为：$W_{AB}=\{F_{AB}\geq F_\alpha\left((r-1)(s-1),rs\left(t-1\right)\right)\}$

计算：$T_{\bullet\bullet\bullet}=\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}X_{ijk}$，$T_{ij\bullet}=\sum_{k=1}^{t}X_{ijk},i=1,...,r,j=1,...,s,$

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

$S_T=\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}X_{ijk}^{\ 2}-\frac{T_{\bullet\bullet\bullet}^{\ 2}}{rst},$ $S_A=\frac{1}{st}\sum_{i=1}^{r}T_{i\bullet\bullet}^{\ 2}-\frac{T_{\bullet\bullet\bullet}^{\ 2}}{rst}$

$S_B=\frac{1}{rt}\sum_{j=1}^{s}T_{\bullet j\bullet}^{\ 2}-\frac{T_{\bullet\bullet\bullet}^{\ 2}}{rst},$ $S_{AB}=(\frac{1}{t}\sum_{i=1}^{r}\sum_{j=1}^{s}T_{ij\bullet}^{2}-\frac{T_{\bullet\bullet\bullet}^{\ 2}}{rst})-S_A-S_B$

$S_E=S_T-S_A-S_B-S_{AB}.$

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	$r-1$	$MS_A = S_A/(r-1)$	$F_A = \frac{MS_A}{MS_E}$
因素B	S_B	$s-1$	$MS_B = S_B/(s-1)$	$F_B = \frac{MS_B}{MS_E}$
交互作用	S_{AB}	$(r-1)(s-1)$	$MS_{AB} = S_{AB}/(r-1)(s-1)$	$F_{AB} = \frac{MS_{AB}}{MS_E}$
误差	S_E	$rs(t-1)$	$MS_E = S_E/rs(t-1)$	
总和	S_T	$rst-1$		

例1 一火箭使用四种燃料，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，得射程表如下（以海里计）：

A \ B	B_1	B_2	B_3	$T_{i..}$
A_1	58.2 52.6	56.2 41.2	65.3 60.8	334.3
A_2	49.1 42.8	54.1 50.5	51.6 48.4	296.5
A_3	60.1 58.3	70.9 73.2	39.2 40.7	342.4
A_4	75.8 71.5	58.2 51.0	48.7 41.4	346.6
$T_{.j.}$	468.4	455.3	396.1	1319.8

设本题符合模型中的条件. 试在显著性水平0.05下，检验不同燃料（因素A）、不同推进器（因素B）下射程是否有显著差异？交互作用是否明显？

解

方差来源	平方和	自由度	均方	F比
因素A	261.67500	3	87.2250	F_A=4.42
因素B	370.98083	2	185.4904	F_B=9.39
交互作用AB	1768.69250	6	294.7821	F_{AB}=14.9
误差	236.95000	12	19.7458	
总和	2638.29833	23		

$F_{0.05}(3,12)=3.49<F_A$，故拒绝H_{01}.
$F_{0.05}(2,12)=3.89<F_B$，故拒绝H_{02}.
$F_{0.05}(6,12)=3.00<F_A$，故拒绝H_{03}.

即认为不同燃料或不同推进器下的射程有显著差异，且交互作用显著.

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）如下表所示.在同一条件下每个试验重复两次.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0.05）？

A \ B	B_1	B_2	$T_{i..}$
A_1	38.0 38.6	47.0 44.8	168.4
A_2	45.0 43.8	42.4 40.8	172
$T_{.j.}$	165.4	175	340.4

§9.2.2 无重复试验

因素A有r个水平$A_1, A_2, \cdots, A_r$，因素B有s个水平$B_1, B_2, \cdots, B_s$.

现对因素A，B的水平的每对组合$(A_i, B_j)\left(i=1,\cdots,r; j=1,\cdots,s\right)$

只作一次试验（此时无法分离交互作用与误差），得到如下结果：

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

$$\left.\begin{array}{c} \text{设} \quad X_{ij}=\mu_{ij}+\varepsilon_{ij}, \\ \varepsilon_{ij}\sim N(0,\sigma^2),\text{各}\varepsilon_{ij}\text{独立}, \\ i=1,\ldots,r,\ j=1,\ldots,s. \\ \mu_{ij},\sigma^2\text{均为未知参数}. \end{array}\right\}$$

记 $\mu=\dfrac{1}{rs}\sum_{i=1}^{r}\sum_{j=1}^{s}\mu_{ij}$，——总平均　$\mu_{i\bullet}=\dfrac{1}{s}\sum_{j=1}^{s}\mu_{ij}, i=1,\ldots,r,$

$$\mu_{\bullet j}=\frac{1}{r}\sum_{i=1}^{r}\mu_{ij}, j=1,\ldots,s,$$

$\alpha_i=\mu_{i\bullet}-\mu$，——水平$A_i$的效应，$i=1,\ldots,r,$

$\beta_j=\mu_{\bullet j}-\mu$，——水平$B_j$的效应，$j=1,\ldots,s.$

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

即 $\mu_{ij}=\mu+\alpha_i+\beta_j, i=1,...,r, j=1,...,s.$

$$\left.\begin{array}{l} \text{模型可写成}\quad X_{ijk}=\mu+\alpha_i+\beta_j+\varepsilon_{ij}, \\ \varepsilon_{ij}\sim N(0,\sigma^2),\text{各}\varepsilon_{ij}\text{独立}, \\ i=1,...,r, j=1,...,s. \\ \sum_{i=1}^{r}\alpha_i=0,\sum_{j=1}^{s}\beta_j=0. \\ \mu,\alpha_i,\beta_j,\sigma^2\text{均未知}. \end{array}\right\}$$

分别检验假设

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

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

记号:

$$\bar{X}=\frac{1}{rs}\sum_{i=1}^{r}\sum_{j=1}^{s}X_{ij},\quad \bar{X}_{i\bullet}=\frac{1}{s}\sum_{j=1}^{s}X_{ij},i=1,...,r,\quad \bar{X}_{\bullet j}=\frac{1}{r}\sum_{i=1}^{r}X_{ij},j=1,...,s.$$

总偏差平方和（总变差）$S_T=\sum_{i=1}^{r}\sum_{j=1}^{s}\left(X_{ij}-\bar{X}\right)^2$

因素A的效应平方和 $S_A=s\sum_{i=1}^{r}\left(\bar{X}_{i\bullet}-\bar{X}\right)^2$

因素B的效应平方和 $S_B=r\sum_{j=1}^{s}\left(\bar{X}_{\bullet j}-\bar{X}\right)^2$

误差平方和 $S_E=\sum_{i=1}^{r}\sum_{j=1}^{s}\left(\bar{X}_{ij}-\bar{X}_{i\bullet}-\bar{X}_{\bullet j}+\bar{X}\right)^2$

性质： (1) $S_T = S_A + S_B + S_E$

$$(2)E\left(\frac{S_E}{(r-1)(s-1)}\right)=\sigma^2,$$

$$E\left(\frac{S_A}{r-1}\right)=\sigma^2+\frac{s\sum_{i=1}^{r}\alpha_i^2}{r-1},$$

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

$$H_{01}:\alpha_1=\alpha_2=\cdots=\alpha_r=0, H_{11}:\alpha_1,\ldots,\alpha_r\text{不全为零}$$

当H_{01}成立时，$F_A=\dfrac{S_A/(r-1)}{S_E/(r-1)(s-1)}\sim F\left(r-1,(r-1)(s-1)\right)$,

拒绝域为：$W_A=\{F_A\geq F_\alpha\left(r-1,(r-1)(s-1)\right)\}$

$$H_{02}:\beta_1=\beta_2=\cdots=\beta_s=0, H_{12}:\beta_1,\ldots,\beta_s\text{不全为零}$$

当H_{02}成立时，$F_B=\dfrac{S_B/(s-1)}{S_E/(r-1)(s-1)}\sim F\left(s-1,(r-1)(s-1)\right)$

拒绝域为：$W_B=\{F_B\geq F_\alpha\left(s-1,(r-1)(s-1)\right)\}$

计算:

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

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

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

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	$r-1$	$MS_A = S_A/(r-1)$	$F_A = \frac{MS_A}{MS_E}$
因素B	S_B	$s-1$	$MS_B = S_B/(s-1)$	$F_B = \frac{MS_B}{MS_E}$
误差	S_E	$(r-1)(s-1)$	$MS_E = S_E/(r-1)(s-1)$	
总和	S_T	$rs-1$		

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：

时间 \ 地点	1	2	3	4	5	$T_{i.}$
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
$T_{.j}$	289	251	308	227	200	1275

设本题符合模型中的条件.试在显著性水平0.05下，检验不同时间和地点下颗粒状物含量的均值是否有显著差异？

§9.3 一元线性回归

变量与变量之间的关系 $\begin{cases} \text{确定性关系} \\ \text{相关性关系} \end{cases}$

确定性关系：

当自变量给定一个值时，就确定应变量的值与之对应。如：在自由落体中，物体下落的高度h与下落时间t之间有函数关系：

$$h = \frac{1}{2} g t^2$$

相关性关系：

变量之间的关系并不确定，而是表现为具有随机性的一种“趋势”。即对自变量x的同一值，在不同的观测中，因变量Y可以取不同的值，而且取值是随机的，但对应x在一定范围的不同值，对Y进行观测时，可以观察到Y随x的变化而呈现有一定趋势的变化。

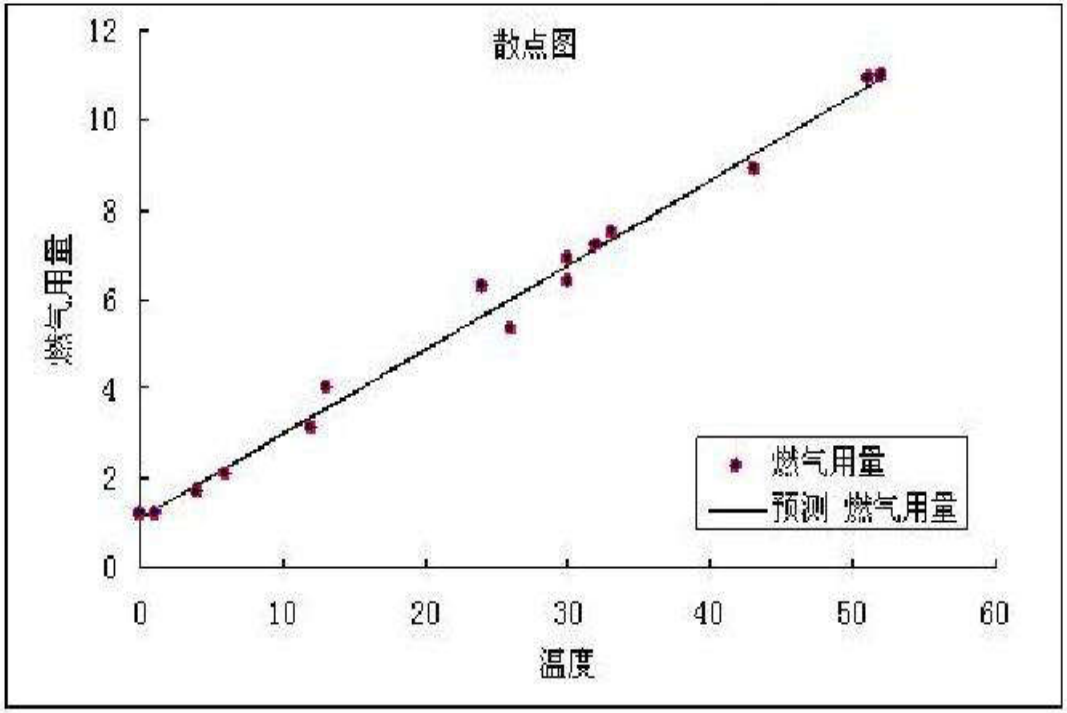
- 如：身高与体重，不存在这样的函数可以由身高计算出体重，但从统计意义上来说，身高者，体也重。
- 如：父亲的身高与儿子的身高之间也有一定联系，通常父亲高，儿子也高。

§9.3.1 一元线性回归

例：某户人家打算安装太阳能热水器. 为了了解加热温度与燃气消耗的关系, 记录了16个月燃气的消耗量, 数据见下表.

月份	平均加热温度	燃气用量	月份	平均加热温度	燃气用量
Nov.	24	6.3	Jul.	0	1.2
Dec.	51	10.9	Aug.	1	1.2
Jan.	43	8.9	Sep.	6	2.1
Feb.	33	7.5	Oct.	12	3.1
Mar.	26	5.3	Nov.	30	6.4
Apr.	13	4	Dec.	32	7.2
May.	4	1.7	Jan.	52	11
Jun.	0	1.2	Feb.	30	6.9

- 我们称“燃气消耗量”为响应变量记为**Y**,“加热温度”为自变量记为**X**.
- 如果以加热温度作为横轴, 以消耗燃气量作为纵轴,得到散点图的形状大致呈线性.



- 加热温度 X 的变化是引起燃气消耗量 Y 变化的主要因素,还有其他一些因素对燃气消耗量 Y 也起着影响, 但这些因素是次要的.
- 从数学形式来考虑, 由于加热温度 X 的变化而引起燃气消耗量 Y 变化的主要部分记为 $\alpha+\beta X$ ，其中 α,β 是未知参数，
- 另一部分是由其他随机因素引起的记为 ε ，即 $Y=\alpha+\beta X+\varepsilon$.

设随机变量Y与x之间存在某种相关关系。

这里，x是可以控制或精确观测的变量（不是随机变量），如年龄、试验时的温度、施加的压力、电压与时间等。

由于Y是随机变量，对于x的每个确定值，Y有相应的分布，记其分布函数为$F(y|x)$。因此如果掌握了$F(y|x)$随着x的取值而变化的规律，也就完全掌握了Y与x之间的关系了。

然而这样做，实际中往往很难实现。作为一种近似，考察Y的数学期望$E(Y)$（假设存在），其值随x的取值而定，它是x的函数，将其记为$\mu(x)$，称为Y关于x的回归函数。于是将讨论Y与x相关关系问题转换为讨论$E(Y)=\mu(x)$与x的关系问题了。

在实际问题中，回归函数μ(x)一般是未知的，需要根据试验数据去估计。

对于x取定一组不完全相同的值$x_1, x_2, ..., x_n$，设分别在x_i处对Y作独立观察得到样本(x_i, Y_i)，$i = 1, 2, ..., n$，对应的样本观察值$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$.

将每对观察值(x_i, y_i)在直角坐标系中描出它相应的点（称为散点图），可以粗略看出$\mu(x)$的形式。

假设$\mu(x)$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。

一元线性回归模型：

$$\text{基本假设：}\begin{cases} Y = a + bx + \varepsilon \\ \varepsilon\text{是随机误差，不可控制，} \\ E(\varepsilon) = 0, D(\varepsilon) = \sigma^2, \\ a, b\text{(回归系数)}, \sigma^2\text{未知.} \end{cases}$$

正态假设：$\varepsilon \sim N\left(0,\ \sigma^2\right).$

一元线性回归要解决的问题：

(1) α, β的估计；

(2) σ^2的估计；

(3) 线性假设的显著性检验；

(4) 回归系数β的置信区间；

(5) 回归函数$\mu(x) = \alpha + \beta x$的点估计和置信区间；

(6) Y的观察值的点预测和区间预测。

§9.3.2 α，β的估计

对x的一组不全相同的值,得到样本$(x_1,Y_1),(x_2,Y_2),...,(x_n,Y_n)$

一元线性回归模型：

$$\begin{cases} Y_i=\alpha+\beta x_i+\varepsilon_i, i=1,2,...,n, \\ \varepsilon_i\sim N\left(0,\ \sigma^2\right),\text{且相互独立}, \\ \alpha,\beta(\text{回归系数}),\sigma^2\text{未知}. \end{cases}$$

$Y_i\sim N(\alpha+\beta x_i,\sigma^2)$，i=1,2,…,n.由$Y_1,Y_2,\ldots,Y_n$的独立性可知，$Y_1,Y_2,\ldots,Y_n$的联合概率密度为

$$L(\alpha,\beta)=\frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\alpha-\beta x_i\right)^2\right\}$$

采用最大似然估计法来估计未知参数α和β.

$$Q(\alpha,\beta)=\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2$$

求估计$\hat{\alpha},\hat{\beta}$,使$Q\left(\hat{\alpha},\hat{\beta}\right)=\min_{\alpha,\beta}Q(\alpha,\beta)$。

$$\frac{\partial Q}{\partial \alpha}=-2\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)=0,$$

$$\frac{\partial Q}{\partial \beta}=-2\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)x_i=0.$$

整理得正规方程系数行列式

$$n\alpha+(\sum_{i=1}^{n}x_i)\beta=\sum_{i=1}^{n}y_i,$$

$$(\sum_{i=1}^{n}x_i)\alpha+(\sum_{i=1}^{n}x_i^{\,2})\beta=\sum_{i=1}^{n}x_iy_i.$$

记：

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

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

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

将正规方程整理得：$\hat{\alpha}+\bar{x}\hat{\beta}=\bar{y},\quad s_{xx}\hat{\beta}=s_{xy}.$

α和β的估计值可写成：

$$\hat{\alpha}=\bar{y}-\bar{x}\hat{\beta},\ \hat{\beta}=s_{xy}/s_{xx}$$

例1 为研究某一化学反应过程中，温度x(℃)对产品得率Y(%)的影响，测得数据如下：

x	100	110	120	130	140	150	160	170	180	190
Y	45	51	54	61	66	70	74	78	85	89

求Y关于x的线性回归方程.

§9.3.3 σ²的估计

误差ε_i的估计

定义：残差$e_i = y_i - \hat{y}_i,\ \ i = 1,2,\cdots,n,$

$$\text{残差平方和}Q_e = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = Q\left(\hat{a},\hat{b}\right)$$

则 $(1)Q_e = S_{yy} - \hat{b}S_{xy},$

$(2)\hat{\sigma}^2 = \dfrac{Q_e}{n-2}$ 是 σ^2 的无偏估计（证略）.

证明：（1）$e_i = y_i - \hat{y}_i = y_i - \hat{a} - \hat{b}x_i = y_i - \bar{y} - \hat{b}(x_i - \bar{x})$

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

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

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

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

例2 为研究某一化学反应过程中，温度x(℃)对产品得率Y(%)的影响，测得数据如下：

x	100	110	120	130	140	150	160	170	180	190
Y	45	51	54	61	66	70	74	78	85	89

求σ^2的无偏估计.

§9.3.4 线性假设的显著性检验

μ(x)是否为x的线性函数，一要根据专业知识和实践来判断，二要根据实际观察得到的数据用假设检验方法来判断。

即要检验假设 $H_0: b = 0, H_1: b \neq 0,$

若原假设被拒绝，说明回归效果是显著的；否则，若接受原假设，说明Y与x不是线性关系，回归方程无意义。回归效果不显著的原因可能有以下几种：

（1）影响Y取值的，除了x，还有其他不可忽略的因素；
（2）E(Y)与x的关系不是线性关系，而是其他关系；
（3）Y与x不存在关系。

检验假设 $H_0: b = 0, H_1: b \neq 0,$

拒绝域形式：$\left|\hat{b}\right| \geq c$。

假定：ε_i独立同服从$N\left(0, \sigma^2\right)$分布$\left(i = 1, 2, \cdots, n\right)$。

则可以证明 (1) $\hat{b} \sim N\left(b, S_{xx}^{-1}\sigma^2\right)$;

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

(3) $\hat{b}$与Q_e独立。

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

当H_0为真即$b=0$时，$t=\dfrac{\hat{b}}{\hat{\sigma}}\sqrt{S_{xx}} \sim t(n-2),$

水平为α的检验拒绝域：$\left|t\right|=\dfrac{1}{\hat{\sigma}}\left|\hat{b}\right|\sqrt{S_{xx}} \geq t_{\alpha/2}(n-2).$

例3 为研究某一化学反应过程中，温度x(℃)对产品得率Y(%)的影响，测得数据如下：

x	100	110	120	130	140	150	160	170	180	190
Y	45	51	54	61	66	70	74	78	85	89

检验Y关于x的线性回归方程（例1）的回归效果是否显著，取α=0.05.

§9.3.5 系数b的置信区间

当回归效果显著时，常需要对回归系数b作区间估计.

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

所以 $\dfrac{|\hat{b}-b|}{\hat{\sigma}}\sqrt{S_{xx}}\le t_{\alpha/2}(n-2)$。

即 b的置信水平$1-\alpha$的置信区间：

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

例4 为研究某一化学反应过程中，温度x(℃)对产品得率Y(%)的影响，测得数据如下：

x	100	110	120	130	140	150	160	170	180	190
Y	45	51	54	61	66	70	74	78	85	89

求Y关于x的线性回归方程中b的置信水平为0.95的置信区间.

§9.3.6 回归函数a+bx函数值的点估计和置信区间

对给定的x_0，$\mu(x_0)=a+bx_0$的点估计为$\hat{y}_0=\hat{\mu}(x_0)=\hat{a}+\hat{b}x_0$.

则有(1)相应的估计量$\hat{Y}_0=\hat{a}+\hat{b}x_0$是$\mu(x_0)=a+bx_0$无偏估计，

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

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

证明： (1)因为$E(\hat{b}) = b, E(\hat{a}) = a,$

所以$E(\hat{Y}_0) = E(\hat{a} + \hat{b}x_0) = a + bx_0$.即为无偏估计

（2）可以证明：$\hat{Y}_0 \sim N(a+bx_0, \left(\frac{1}{n}+\left(x_0-\bar{x}\right)^2 S_{xx}^{-1}\right)\sigma^2)$.

又有，$\frac{(n-2)\hat{\sigma}^2}{\sigma^2}=\frac{Q_e}{\sigma^2}\sim\chi^2(n-2)$; 且$\hat{Y}_0$与$Q_e$独立。

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

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

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

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

§9.3.7 Y的观察值的点预测和预测区间

考虑对指定点$x=x_0$处因变量Y的观察值Y_0的预测问题。由于在$x=x_0$处并未进行观察，或暂时无法观察。经验回归函数的重要应用是，可利用它对因变量Y的新观察值Y_0进行点预测和区间预测。

设Y_0是在$x=x_0$处对Y的观察结果。则

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

(1) Y_0的点预测为：$\hat{Y}_0=\hat{a}+\hat{b}x_0$.

(2) Y_0的置信水平为$1-\alpha$的预测区间为：

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

证明：因Y_0是将要做的独立试验结果，因此，它与已得到的试验结果$Y_1, Y_2, ..., Y_n$相互独立。

又$\hat{Y}_0 = \bar{Y} + \hat{b}(x_0 - \bar{x})$是$Y_1, Y_2, ..., Y_n$的线性组合，

故Y_0与$\hat{Y}_0$相互独立。

$$Y_0 \sim N(a + bx_0, \sigma^2), \hat{Y}_0 \sim N(a + bx_0, \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\sigma^2).$$

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

又$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{Q_e}{\sigma^2} \sim \chi^2(n-2);$$ 且Y_0， $\hat{Y}_0$， Q_e相互独立。

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

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

所以，Y_0的置信水平为$1-\alpha$的预测区间为：

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

注1，这一预测区间的长度随$|x_0-\bar{x}|$的增加而增加，当$x_0=\bar{x}$时最短。

注2，在相同的置信水平下，$\mu(x_0)$的置信区间要比Y_0的预测区间短。这是因为$Y_0 = a+bx_0+\varepsilon_0$比$\mu(x_0)=a+bx_0$多了一项$\varepsilon_0$的缘故。

例5 为研究某一化学反应过程中，温度x(℃)对产品得率Y(%)的影响，测得数据如下：

x	100	110	120	130	140	150	160	170	180	190
Y	45	51	54	61	66	70	74	78	85	89

(1)求回归函数μ(x)在x=125处的值μ(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；

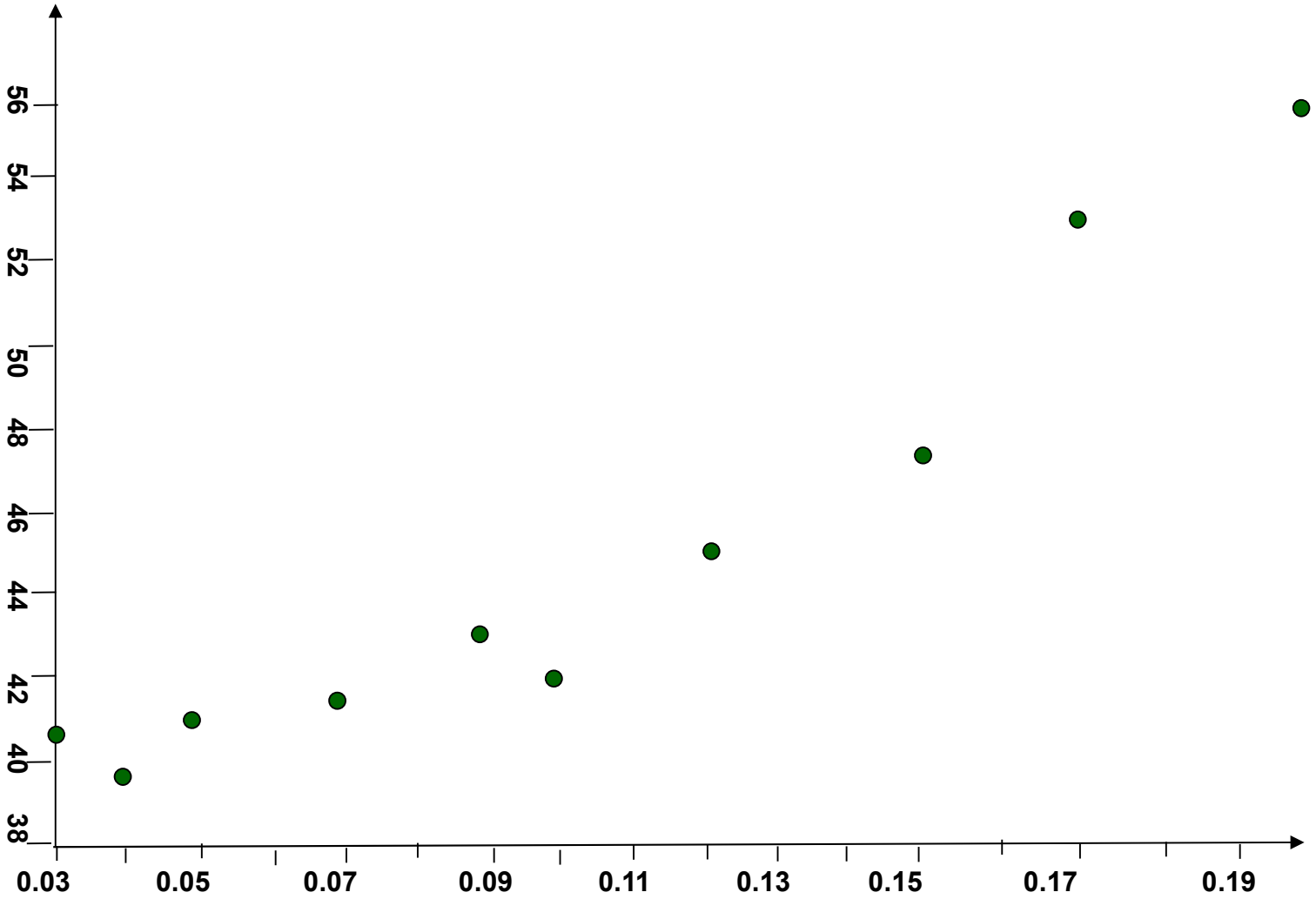
(2)求在$x=x_0$处Y的新观察值Y_0的置信水平为0.95的预测区间.

例6 合金钢的强度y与钢材中碳的含量x有密切关系。为了冶炼出符合要求强度的钢常常通过控制钢水中的碳含量来达到目的，为此需要了解y与x之间的关系。其中x：碳含量（%） y：钢的强度（kg/mm^2）数据见下：

x	0.03	0.04	0.05	0.07	0.09	0.10	0.12	0.15	0.17	0.20
y	40.5	39.5	41.0	41.5	43.0	42.0	45.0	47.5	53.0	56.0

（1）画出散点图；
（2）设 $\mu(x)=a+bx$, 求a+b的估计；
（3）求误差方差的估计，画出残差图；
（4）检验回归系数 β 是否为零（取 $\alpha=0.05$）；
（5）求回归系数 β 的95%置信区间；
（6）求在x=0.06点，回归函数的点估计和95%置信区间；
（7）求在x=0.06点，Y的点预测和95%区间预测。

（1）合金钢的强度y与钢材中碳的含量x的散点图



(2)计算得：

$$\sum_i y_i = 449, \sum_i x_i = 1.02,$$

$$\sum_i x_i^2 = 0.1338, \sum_i x_i y_i = 48.555,$$

$$S_{xx} = 0.02976, S_{xy} = 2.757.$$

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

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

$$\hat{a} = 35.4506,\ \hat{b} = 92.6411$$

回归方程：$\hat{y} = 35.4506 + 92.6411x.$

或写成：$\hat{y} = 44.9 + 92.6411(x - 0.102).$

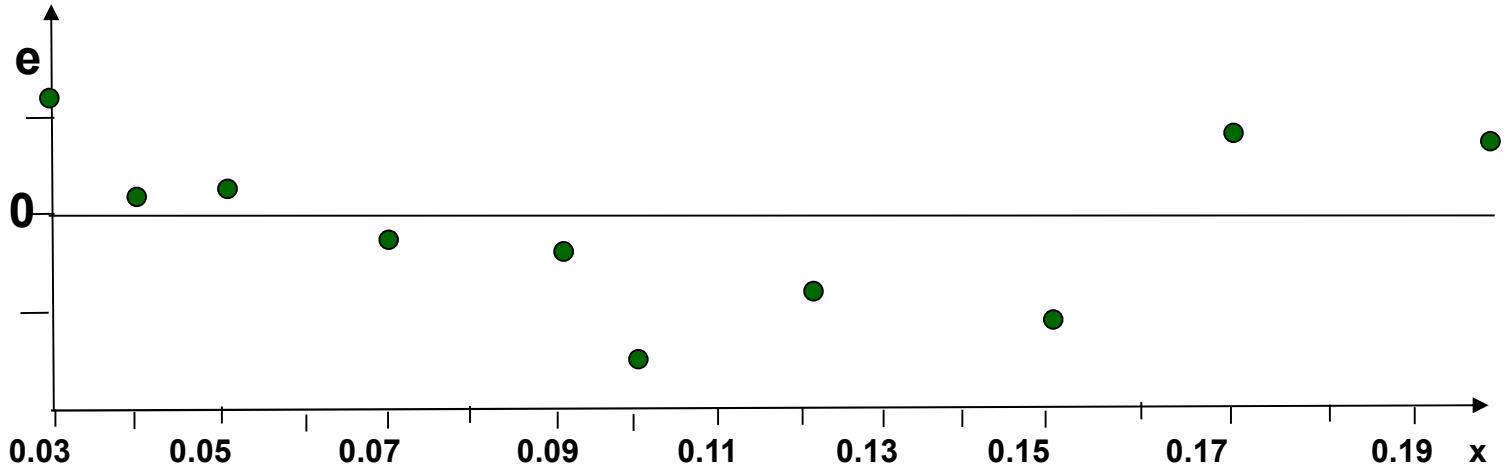
(3)计算得：

$$\sum_i y_i = 449, \sum_i y_i^2 = 20443, S_{yy} = 282.9.$$

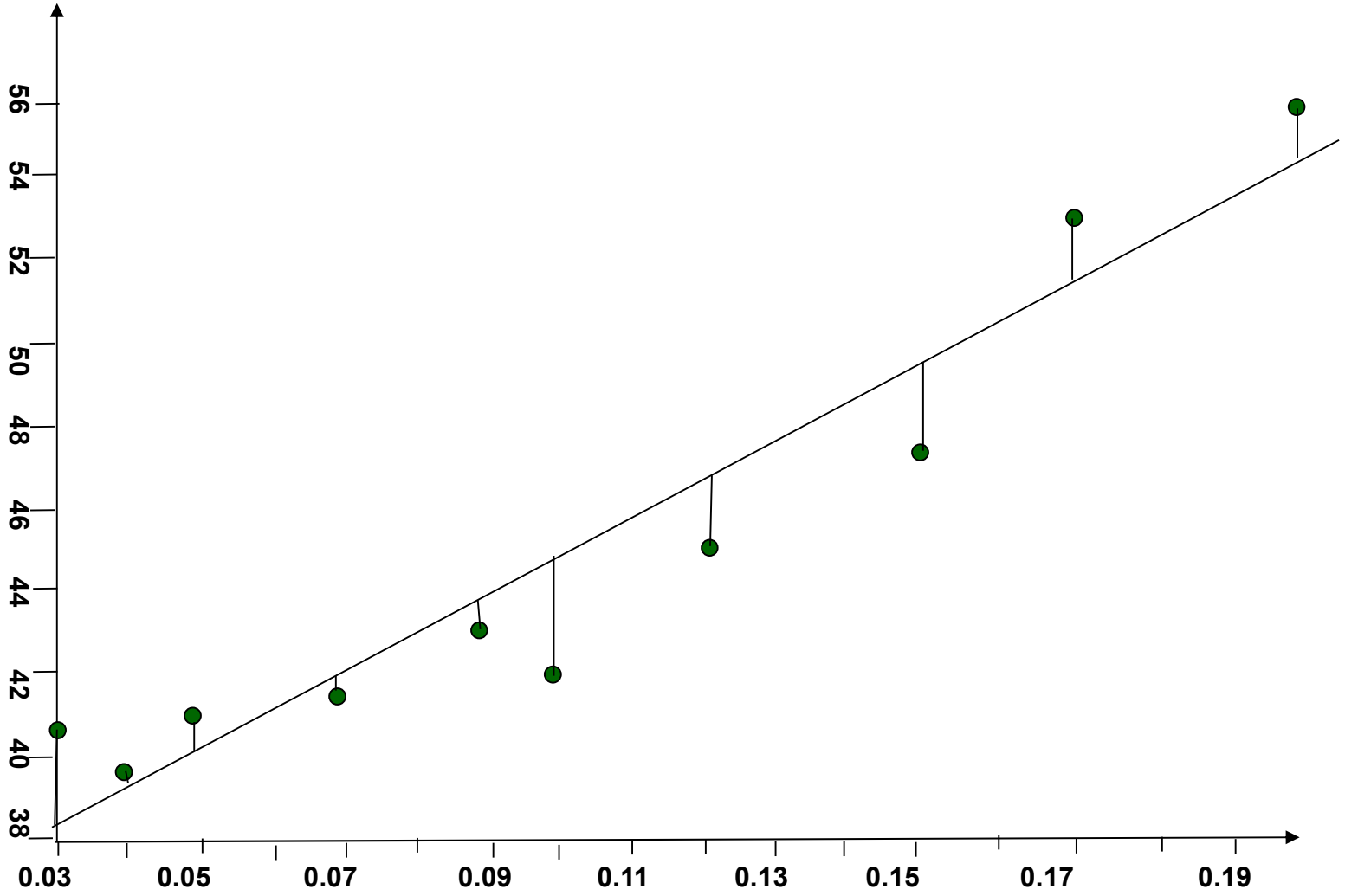
又已知 $S_{xy} = 2.757, \hat{b} = 92.6411.$

$$Q_e = S_{yy} - \hat{b}S_{xy} = 27.4884,$$

所以，σ^2的无偏估计 $\hat{\sigma}^2 = Q_e/(n-2) = 3.436.$



合金钢的强度y与钢材中碳的含量x的回归直线图



(4)检验假设$H_0: b=0, H_1: b\neq 0$的显著性水平

为α的检验拒绝域：　　$|t|=\dfrac{|\hat{b}|}{\hat{\sigma}}\sqrt{S_{xx}}\geq t_{\alpha/2}(n-2)$。

经计算

$$|t|=\frac{92.6411}{\sqrt{3.436}}\sqrt{0.02976}=8.6217\geq t_{0.025}(8)=2.306,$$

拒绝原假设，认为合金钢强度与炭含量的回归效果显著。

(5) 回归系数b的 置信水平95%的置信区间：

$$\left(\hat{b}\pm t_{\alpha/2}(n-2)\times\frac{\hat{\sigma}}{\sqrt{S_{xx}}}\right)=(67.8629,\ 117.4193).$$

（6）当$x_0 = 0.06$时，$\hat{y}_0 = \hat{a} + \hat{b}x_0 = 41.0091$

$$t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}} = 2.306\times\sqrt{3.436}\sqrt{\frac{1}{10}+\frac{(0.06-0.102)^2}{0.02976}} = 1.706$$

所以，$\mu(0.06)$的0. 95的置信区间为:$(39.303, 42.715)$.

(7)$x_0 = 0.06$时，Y_0的置信水平为0.95的预测区间为:

$(36.407,\ 45.611)$.

其中$t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}} = 4.602.$

§ 9.3.8 可化为一元线性回归的例子

实际中常会遇到很复杂的回归问题，但在某些情况下，通过适当的变量变换，可将其化为一元线性回归来处理。下面是三种常见的可转化为一元线性回归的模型。

(1) $Y=\alpha e^{\beta x}\varepsilon, \ln\varepsilon\sim N(0,\sigma^2)$,其中$\alpha,\beta,\sigma^2$为未知参数。

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

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

即可转化为一元线性回归模型：

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

(2) $Y=\alpha x^{\beta}\varepsilon, \ln\varepsilon\sim N(0,\sigma^2)$,其中$\alpha,\beta,\sigma^2$为未知参数。

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

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

即可转化为一元线性回归模型:

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

(3) $Y=\alpha+\beta h(x)+\varepsilon, \varepsilon\sim N(0,\sigma^2)$,其中$\alpha,\beta,\sigma^2$为未知参数。

这里$h(x)$是x的已知函数,

令$\alpha=a, \beta=b, h(x)=x'$,

即可转化为一元线性回归模型:

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