

# 第九章 方差分析和回归分析

单因素方差分析

一元线性回归

回归诊断

- 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.

## 9.1 单因素方差分析

**例1.1** 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:

- 其一, 由于日光灯类型不同,而引起寿命不同.
- 其二,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.

- 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( 例如日光灯管的寿命) 称为 **试验指标**.
- 对试验指标产生影响的原因称为 **因素**, “日光灯管类型” 即为**因素**.
- 因素中各个不同状态称为 **水平**, 如日光灯管三个不同的类型, 即为三个**水平**.

- 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

$A_1 : N(\mu_1, \sigma^2)$	$X_{11}$	$X_{12}$	$\cdots$	$X_{1n_1}$
$A_2 : N(\mu_2, \sigma^2)$	$X_{21}$	$X_{22}$	$\cdots$	$X_{2n_2}$
$\vdots$	$\vdots$	$\cdots$		$\vdots$
$A_r : N(\mu_r, \sigma^2)$	$X_{r1}$	$X_{r2}$	$\cdots$	$X_{rn_r}$

各个总体相互独立. 因此, 可写成如下的 数学模型:

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

- 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.



检验假设

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_r$$

$$H_1 : \mu_1, \mu_2, \dots, \mu_r \text{ 不全相等.}$$

记  $\mu = \frac{1}{n} \sum_{i=1}^r n_i \mu_i$  ——总平均, 其中  $\sum_{i=1}^r n_i = n$

$\delta_i = \mu_i - \mu$  ——水平  $A_i$  的效应,  $i = 1, 2, \dots, r$

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

$$\left. \begin{aligned} \text{模型为: } X_{ij} &= \mu + \delta_i + \varepsilon_{ij} \\ \varepsilon_{ij} &\sim N(0, \sigma^2), \text{ 各 } \varepsilon_{ij} \text{ 独立} \\ j &= 1, 2, \dots, n_i, i = 1, 2, \dots, r \\ n_1\delta_1 + n_2\delta_2 + \dots + n_r\delta_r &= 0 \end{aligned} \right\}$$

假设等价于  $H_0 : \delta_1 = \delta_2 = \dots = \delta_r = 0$   
 $H_1 : \delta_1, \delta_2, \dots, \delta_r$  不全为零.

- 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离差平方和 $SS_T$ 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 $SS_A$ ；
  - 另一部分则由随机误差所引起的差异，即误差平方和 $SS_E$ 。

总偏差平方和  $SS_T = \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$

效应平方和  $SS_A = \sum_{i=1}^r n_i (\bar{X}_{i\bullet} - \bar{X})^2$

$$= \sum_{i=1}^r n_i \bar{X}_{i\bullet}^2 - n\bar{X}^2$$

误差平方和  $SS_E = \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\bullet})^2$

性质1:  $SS_T = SS_A + SS_E$

证明: 
$$SS_T = \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2 = \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\bullet} + \bar{X}_{i\bullet} - \bar{X})^2$$
$$= \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\bullet})^2 + \sum_{i=1}^r \sum_{j=1}^{n_i} (\bar{X}_{i\bullet} - \bar{X})^2 + 2 \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\bullet})(\bar{X}_{i\bullet} - \bar{X})$$
$$= SS_A + SS_E$$

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

性质2:  $E(SS_T) = \sum_{i=1}^r n_i \delta_i^2 + (n-1)\sigma^2$

$$E(SS_A) = \sum_{i=1}^r n_i \delta_i^2 + (r-1)\sigma^2$$

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

证明:

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

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

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

$$= \sum_{i=1}^r \sum_{j=1}^{n_i} [\sigma^2 + (\mu + \delta_i)^2] - n\left[\frac{\sigma^2}{n} + \mu^2\right]$$

$$\begin{aligned} E(\bar{X}) &= \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^{n_i} E(X_{ij}) \\ &= \frac{1}{n} \sum_{i=1}^r n_i (\mu + \delta_i) = \mu \end{aligned}$$

$$= n\sigma^2 + n\mu^2 + 2\mu \sum_{i=1}^r n_i \delta_i + \sum_{i=1}^r n_i \delta_i^2 - \sigma^2 - n\mu^2$$

$$= \sum_{i=1}^r n_i \delta_i^2 + (n-1)\sigma^2$$

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

$$= \sum_{i=1}^r (n_i - 1)\sigma^2 = (n - r)\sigma^2$$



$$E(SS_A) = E(SS_T - SS_E) = \sum_{i=1}^r n_i \delta_i^2 + (r-1) \sigma^2$$

## 定理9.1.1

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

(2)  $SS_A$  与  $SS_E$  相互独立,

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

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

# 单因素试验方差分析表

方差来源	平方和	自由度	均方	F
因素A	$SS_A$	$r-1$	$MS_A = SS_A / r - 1$	$\frac{MS_A}{MS_E}$
误差	$SS_E$	$n-r$	$MS_E = SS_E / n - r$	
总和	$SS_T$	$n-1$		

当 $F \geq F_{\alpha}(r-1, n-r)$ 时，拒绝原假设.

$SS_T, SS_A, SS_E$ 的计算公式:

$$\text{记 } \bar{X}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}, \quad i = 1, 2, \dots, r, \quad \bar{X} = \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^{n_i} X_{ij}$$

$$SS_T = \sum_{i=1}^r \sum_{j=1}^{n_i} X_{ij}^2 - n\bar{X}^2$$

$$SS_A = \sum_{i=1}^r n_i \bar{X}_{i\cdot}^2 - n\bar{X}^2$$

$$SS_E = SS_T - SS_A$$

**例1.2** 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：( $\alpha=0.05$ )

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

这里药物是因子，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  $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

$H_1 : \mu_1, \mu_2, \dots, \mu_5$ 不全相等。

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

$$\sum_{i=1}^r \sum_{j=1}^{n_i} X_{ij}^2 = 1047,$$

$$\bar{X}_{1\bullet} = 7.5, \bar{X}_{2\bullet} = 5, \bar{X}_{3\bullet} = 4.33, \bar{X}_{4\bullet} = 5.17,$$

$$\bar{X}_{5\bullet} = 6.17, \bar{X} = 5.63$$



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

$> F_{0.05}(4, 25) = 2.76$ 。  
 拒绝 $H_0$ ，认为疗效  
 有显著差异。

# 未知参数的估计

(1)  $\sigma^2$  的估计  $\hat{\sigma}^2 = \frac{SS_E}{n-r}$ ;

(2)  $\mu$  的估计  $\hat{\mu} = \bar{X}$ ;

(3)  $\mu_i$  的估计  $\hat{\mu}_i = \bar{X}_{i\bullet}$ ;

(4)  $\delta_i$  的估计  $\hat{\delta}_i = \bar{X}_{i\bullet} - \bar{X}$ .

容易证明，以上估计均为相应参数的无偏估计.

注意：如果拒绝原假设 $H_0$ ，只能说明均值不全相等。接下来的问题是它们中有没有部分是相等的？仍需要进一步的推断，比较 $N(\mu_i, \sigma^2)$ 和 $N(\mu_j, \sigma^2)$ 的差异。

(1) 作  $\mu_i - \mu_j = \delta_i - \delta_j (i \neq j)$  的区间估计；

(2) 作  $H_0 : \mu_i = \mu_j, H_1 : \mu_i \neq \mu_j$  的假设检验。

# 置信区间

因为  $E(\bar{X}_{i\bullet} - \bar{X}_{j\bullet}) = \mu_i - \mu_j$ ,  $Var(\bar{X}_{i\bullet} - \bar{X}_{j\bullet}) = \sigma^2 \left( \frac{1}{n_i} + \frac{1}{n_j} \right)$   
且  $\bar{X}_{i\bullet} - \bar{X}_{j\bullet}$  与  $\hat{\sigma}^2 = MS_E$  相互独立。

$$\text{故 } \frac{(\bar{X}_{i\bullet} - \bar{X}_{j\bullet}) - (\mu_i - \mu_j)}{\sqrt{MS_E(1/n_i + 1/n_j)}} = \frac{(\bar{X}_{i\bullet} - \bar{X}_{j\bullet}) - (\mu_i - \mu_j)}{\sigma \sqrt{(1/n_i + 1/n_j)}} \bigg/ \sqrt{\frac{SS_E}{\sigma^2} / (n-r)} \\ \sim t(n-r)$$

$$\left( \bar{X}_{i\bullet} - \bar{X}_{j\bullet} \pm t_{\alpha/2}(n-r) \sqrt{MS_E(1/n_i + 1/n_j)} \right)$$

得  $(\mu_i - \mu_j)$  的水平  
为  $1-\alpha$  的置信区间

例1.3 求例1.2中未知参数 $\sigma^2, \mu_i, \delta_i (i = 1, 2, 3, 4, 5)$ 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。

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

$\mu_i$ 的估计分别为: 7.5, 5, 4.3333, 5.1667, 6.1667;

$\delta_i$ 的估计分别为: 1.8667, -0.6333, -1.3, -0.4666, 0.5334

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

$\mu_1 - \mu_3$ ,  $\mu_1 - \mu_2$ ,  $\mu_3 - \mu_5$ 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 $\mu_1$ 与 $\mu_3$ ,  $\mu_1$ 与 $\mu_2$ ,  $\mu_3$ 与 $\mu_5$ 的差异都显著。

# 假设检验

## 检验假设

$$H_0 : \mu_i = \mu_j, H_1 : \mu_i \neq \mu_j, i \neq j$$

给出检验统计量

$$t_{ij} = \frac{\bar{X}_{i\bullet} - \bar{X}_{j\bullet}}{\sqrt{MS_E (1/n_i + 1/n_j)}}$$

当 $H_0$ 成立时,  $t_{ij} \sim t(n-r)$

拒绝域 $W = \{|t_{ij}| \geq t_{\alpha/2}(n-r)\}$

观测值得到的 $t_{ij}$ 落在拒绝域内,

则拒绝原假设 $H_0$ ,

反之, 则接受原假设 $H_0$ .



例1.4 (续1.2) (1) 判断第一种、第二种药物的差异;

(2) 判断第一种、第三种药物的差异;

(3) 判断第三种、第五种药物的差异;

解: 仅检验(1), (2)和(3)留作思考题.

(1) 检验假设

$$H_0: \mu_1 = \mu_2, H_1: \mu_1 \neq \mu_2,$$

$$t_{12} = \frac{\bar{X}_{1\bullet} - \bar{X}_{2\bullet}}{\sqrt{MS_E (1/n_1 + 1/n_2)}}$$

由前面的计算知：

$$\sqrt{MS_E (1/n_i + 1/n_j)} = 0.8819$$

$$\bar{x}_{1\bullet} - \bar{x}_{2\bullet} = 7.5 - 5 = 2.5$$

$$t_{12} = \frac{\bar{x}_{1\bullet} - \bar{x}_{2\bullet}}{\sqrt{MS_E (1/n_1 + 1/n_2)}} = 2.833$$

查表得  $t_{0.025}(25) = 2.0595$ ,

$|t_{12}| > t_{0.025}(25)$ , 拒绝  $H_0$ .

# 在Excel上实现方差分析

- 先加载"数据分析" 这个模块,方法如下:
- 在**excel**工作表中点击主菜单中 “工具” 点击下拉式菜单中 “加载宏” 就会出现一个 “加载宏” 的框.
- 在 “分析工具库” 前的框内打勾点击 “确定” . 这时候再点击下拉式菜单会新出现 “数据分析” . 然后就可以进行统计分析了.

---

以下面的例子来说明用Excel进行方差分析的方法:

- 保险公司某一险种在四个不同地区一年的索赔额情况记录如表所示。试判断在四个不同地区索赔额有无显著的差异?

保险索赔记录								
地区	索赔额(万元)							
A1	1.60	1.61	1.65	1.68	1.70	1.70	1.78	
A2	1.50	1.64	1.40	1.70	1.75			
A3	1.64	1.55	1.60	1.62	1.64	1.60	1.74	1.80
A4	1.51	1.52	1.53	1.57	1.64	1.60		

- 在**Excel**工作表中输入上面的数据点击主菜单中 “工具” 点击下拉式菜单中 “数据分析” 就会出现一个 “数据分析” 的框.
- 点击菜单中 “方差分析:单因素方差分析” 点击 “确定”, 出现 “方差分析:单因素方差分析” 框.

- 在“输入区域”中标定你已经输入的数据的位置(本例为**\$B\$3:\$I\$6**),根据你输入数据分组情况(是按行分或按列分,本例点击“行”)确定分组.
- 选定方差分析中**F**检验的显著水平选定输出结果的位置点击“确定”.
- 在你指定的区域中出现如下两张表:

表一：摘要

组	观测数	求和	平均	方差
行1	7	11.72	1.674	0.0038
行2	5	7.99	1.598	0.0210
行3	8	13.19	1.649	0.0067
行4	6	9.37	1.562	0.0026



表二：方差分析表

方差来源	平方和	自由度	均方	F	P-value	F crit
组间	0.0492	3	0.0164	2.1659	0.1208	3.0491
组内	0.1666	22	0.0076			
总计	0.2158	25				

根据Excel给出的方差分析表,假设 $H_0$ 的判别有二种方法:

■  $F = 2.1658 < F_{0.05}(3, 18) = 3.16,$

接受假设  $H_0$ , 即各地区索赔额无显著差异.

■  $P\_ = 0.1208 > 0.05,$

接受假设  $H_0$ .

方差估计:  $\hat{\sigma}^2 = 0.0076$ ;

均值估计:  $\hat{\mu}_1 = 1.674, \hat{\mu}_2 = 1.598,$   
 $\hat{\mu}_3 = 1.649, \hat{\mu}_4 = 1.562.$

$\mu_1$ 的置信度为95%置信区间

$$\left( \bar{X}_{1\bullet} \pm t_{0.025}(22) \sqrt{MS_E / n_1} \right) = (1.674 \pm 2.074 \times \sqrt{0.0076/7}) \\ = (1.606, 1.742).$$

$\mu_1 - \mu_2$ 的置信度为95%置信区间

$$\left( \bar{X}_{1\bullet} - \bar{X}_{2\bullet} \pm t_{0.025}(22) \sqrt{MS_E (1/7 + 1/5)} \right) = (-0.030, 0.182).$$

$$H_0 : \mu_1 = \mu_2, H_1 : \mu_1 \neq \mu_2,$$

$$t_{12} = \frac{\bar{x}_{1\bullet} - \bar{x}_{2\bullet}}{\sqrt{MS_E (1/n_1 + 1/n_2)}} = 1.489$$

查表得 $t_{0.025}(22) = 2.074$ ,

$|t_{12}| < t_{0.025}(22)$ , 接受 $H_0$ .

# 方差分析的前提

进行方差分析必须具备三个基本的条件：

**(1) 独立性.**数据是来自 $r$ 个独立总体的简单随机样本；

**(2) 正态性.** $r$ 个独立总体均为正态总体；

**(3) 方差齐性.** $r$ 个独立总体的方差相等.

如何判断这些条件是否成立？这些条件对于方差分析的结论影响又是如何？

- 方差分析和其它统计推断一样, 样本的独立性对方差分析是非常重要的, 在实际应用中会经常遇到非随机样本的情况,
- 这时使用方差分析得出的结论不可靠. 因此, 在安排试验或采集数据的过程中, 一定要注意样本的独立性问题.

- 在实际中, 没有一个总体真正服从正态分布的, 而方差分析却依赖于正态性的假设. 不过由经验可知, 方差分析F检验对正态性的假设并不是非常敏感,
- 即, 实际所得到的数据, 若没有异常值和偏性, 或者说, 数据显示的分布比较对称的话, 即使样本容量比较小(如每个水平下的样本容量仅为5左右), 方差分析的结果仍是值得信赖的.

- 方差齐性对于方差分析是非常重要的, 因此在方差分析之前往往要进行方差齐性的诊断, 检验方差齐性假设通常采用**Barlett**检验.
- 不过, 也可采用如下的经验准则: 当最大样本标准差不超过最小样本标准差的两倍时, 方差分析**F**检验结果近似正确.



## 9.4 一元线性回归

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

### 一、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： $h = \frac{1}{2}gt^2$

## 二、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。

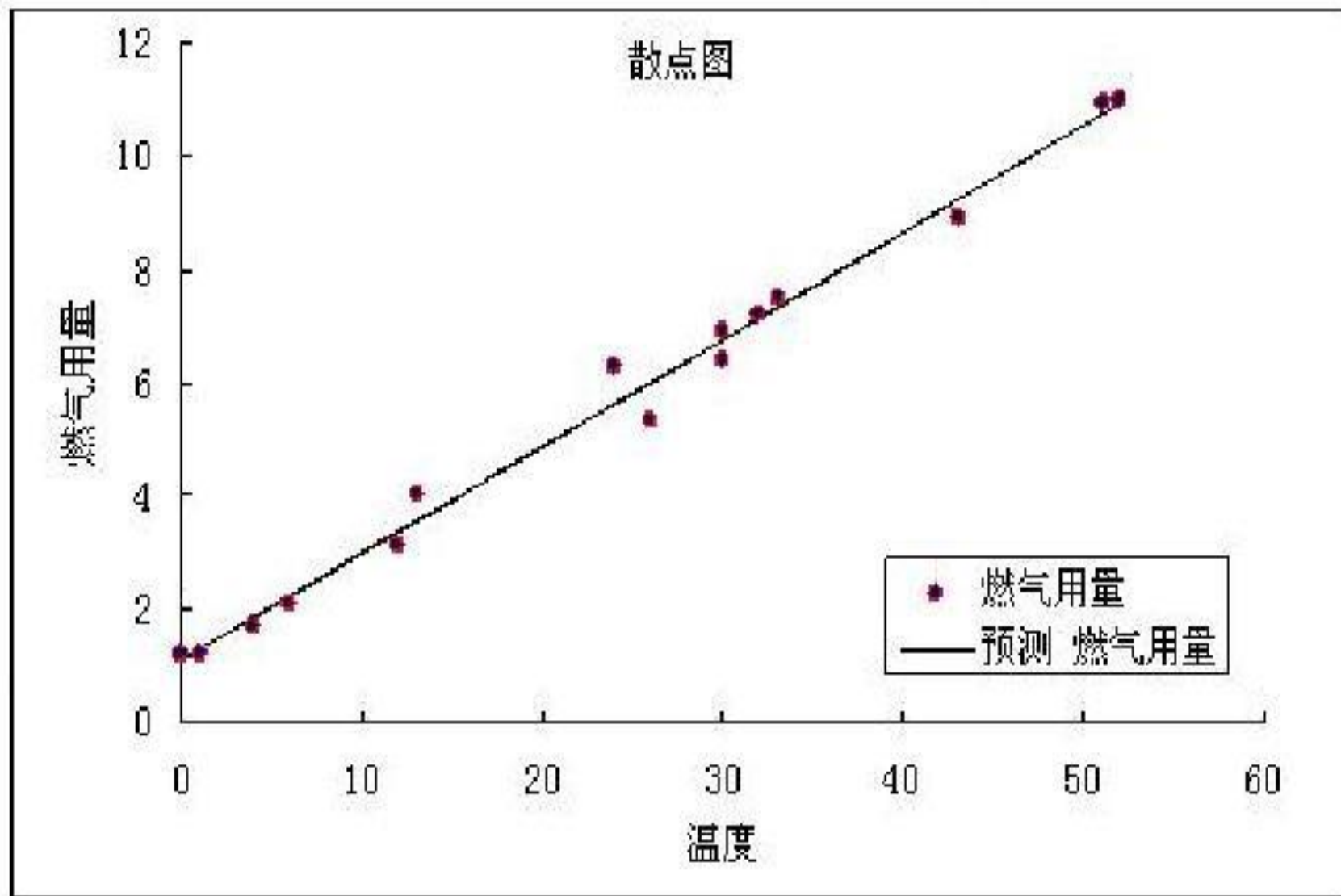
为统一记号，后面一律用 $y$ 表示因变量。

- 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。

# 我们以一个例子来建立回归模型

- 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
<b>Nov.</b>	<b>24</b>	<b>6.3</b>	<b>Jul.</b>	<b>0</b>	<b>1.2</b>
<b>Dec.</b>	<b>51</b>	<b>10.9</b>	<b>Aug.</b>	<b>1</b>	<b>1.2</b>
<b>Jan.</b>	<b>43</b>	<b>8.9</b>	<b>Sep.</b>	<b>6</b>	<b>2.1</b>
<b>Feb.</b>	<b>33</b>	<b>7.5</b>	<b>Oct.</b>	<b>12</b>	<b>3.1</b>
<b>Mar.</b>	<b>26</b>	<b>5.3</b>	<b>Nov.</b>	<b>30</b>	<b>6.4</b>
<b>Apr.</b>	<b>13</b>	<b>4</b>	<b>Dec.</b>	<b>32</b>	<b>7.2</b>
<b>May.</b>	<b>4</b>	<b>1.7</b>	<b>Jan.</b>	<b>52</b>	<b>11</b>
<b>Jun.</b>	<b>0</b>	<b>1.2</b>	<b>Feb.</b>	<b>30</b>	<b>6.9</b>



- 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.
- 如果假设中间有一条直线,这些点均匀地散布在直线的两侧.表示除了温度外还有其它的因素影响燃气消耗量.

- 在回归分析时, 我们称“燃气消耗量”为响应变量记为 $y$ , “加热温度”为解释变量记为 $x$ , 由所得数据计算相关系数得 $r=0.995$ , 表明加热温度与燃气消耗之间有非常好的线性相关性.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$



- 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 $\beta_0 + \beta_1 x$ , 其中 $\beta_0, \beta_1$ 是未知参数,
- 另一部分是由其他随机因素引起的记为 $\varepsilon$ ,  
即 $y = \beta_0 + \beta_1 x + \varepsilon$ .

对从总体 $(x, y)$ 中抽取的一个样本

$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

一元线性回归模型:

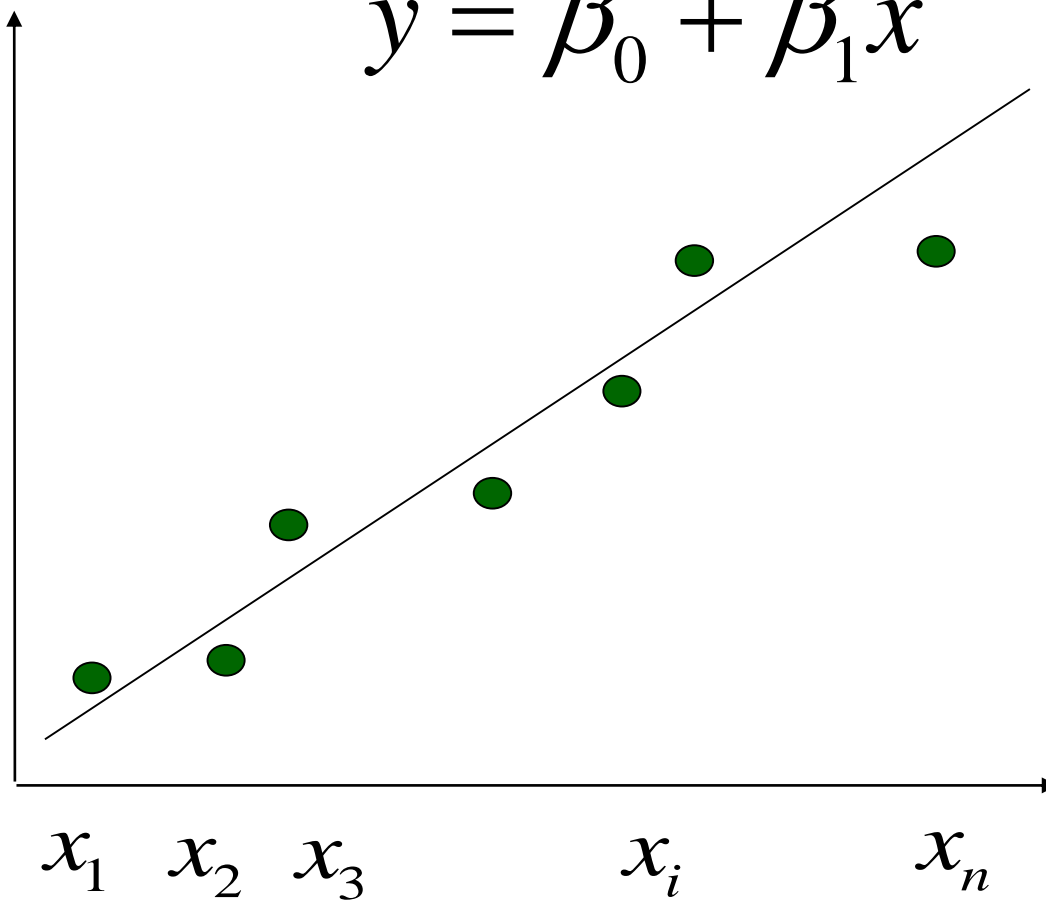
$$\begin{cases} y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, \dots, n, \\ \varepsilon_i \sim N(0, \sigma^2), \text{且相互独立}, \\ \beta_0, \beta_1 (\text{回归系数}), \sigma^2 \text{未知}. \end{cases}$$

在模型假定下 $y_i(i=1,2,\dots,n)$ 也是相互独立, 服从正态分布 $N(\beta_0 + \beta_1 x_i, \sigma^2)$ . 由所得样本可给出未知参数 $\beta_0, \beta_1$ 的点估计, 分别记为 $\hat{\beta}_0, \hat{\beta}_1$ ,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

称为 $y$ 关于 $x$ 的一元线性回归方程.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$



# 一元线性回归要解决的问题：

参数估计  $\left\{ \begin{array}{l} (1) \beta_0, \beta_1 \text{ 的估计;} \\ (2) \sigma^2 \text{ 的估计;} \end{array} \right.$

模型及参数检验  $\left\{ \begin{array}{l} (3) \text{ 线性假设的显著性检验;} \\ (4) \text{ 回归系数 } \beta_1 \text{ 的置信区间;} \end{array} \right.$

应用  $\left\{ \begin{array}{l} (5) \text{ 回归函数 } \mu(x) = \beta_0 + \beta_1 x \text{ 的点估计和置信区间;} \\ (6) y \text{ 的观察值的点预测和区间预测.} \end{array} \right.$

# 参数估计

$$Q(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

求估计  $\hat{\beta}_0, \hat{\beta}_1$ , 使  $Q(\hat{\beta}_0, \hat{\beta}_1) = \min_{\beta_0, \beta_1} Q(\beta_0, \beta_1)$ .

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0,$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i = 0.$$

## 整理得正规方程系数行列式

$$n\beta_0 + \left(\sum_{i=1}^n x_i\right)\beta_1 = \sum_{i=1}^n y_i,$$

$$\left(\sum_{i=1}^n x_i\right)\beta_0 + \left(\sum_{i=1}^n x_i^2\right)\beta_1 = \sum_{i=1}^n x_i y_i.$$

记号:  $\bar{y} = \frac{1}{n} \sum_i y_i, \bar{x} = \frac{1}{n} \sum_i x_i, S_{xx} = \sum_i (x_i - \bar{x})^2,$

$$S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y}), S_{yy} = \sum_i (y_i - \bar{y})^2.$$

将正规方程整理得:  $\hat{\beta}_0 + \bar{x} \hat{\beta}_1 = \bar{y}, S_{xx} \hat{\beta}_1 = S_{xy}.$

$\beta_0, \beta_1$  的最小二乘估计:  $\hat{\beta}_0 = \bar{y} - \bar{x} \hat{\beta}_1, \hat{\beta}_1 = S_{xy} / S_{xx}.$



(2)  $\sigma^2$  的估计,

■ 定义残差。记  $e_i = y_i - \hat{y}_i$ , 称  $e_i$  为残差。残差可以看成是不可观测的误差  $\varepsilon_i$  的估计。

■ 采用残差平方和  $\sum_{i=1}^n (y_i - \hat{y}_i)^2$  作为  $\sigma^2$  的估计。

$$s^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2},$$

可以证明  $s^2$  为  $\sigma^2$  的无偏估计。

在误差为正态分布假定下,  $\beta_0, \beta_1$  的最小二乘估计等价于极大似然估计。

$$L(\beta_0, \beta_1) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp \left\{ -\frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \right\}$$

对  $L(\beta_0, \beta_1)$  最大化等价于对  $\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$  最小化, 即最小二乘估计。

- 采用最大似然估计给出参数 $\beta_0, \beta_1$ 的估计与最小二乘法给出的估计完全一致。
- 采用最大似然估计给出误差 $\sigma^2$ 的估计如下：

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

此估计不是 $\sigma^2$ 的无偏估计。

**例3.1 K.Pearson**收集了大量父亲身高与儿子身高的资料。其中十对如下：

父亲身高 $x$ (吋)	60	62	64	65	66	67	68	70	72	74
儿子身高 $y$ (吋)	63.6	65.2	66	65.5	66.9	67.1	67.4	68.3	70.1	70

求 $y$ 关于 $x$ 的线性回归方程。

计算得：

$$\bar{y} = 67.01, \bar{x} = 66.8,$$

$$\sum_i x_i^2 = 44794, \sum_i x_i y_i = 44842.4,$$

$$S_{xx} = 171.6, S_{xy} = 79.72.$$

$$\beta_0, \beta_1 \text{的最小二乘估计: } \hat{\beta}_0 = 35.9768, \hat{\beta}_1 = 0.4646.$$

$$\text{回归方程: } \hat{y} = 35.9768 + 0.4646x.$$

$$\text{或写成: } \hat{y} = 67.01 + 0.4646(x - 66.8).$$

# 参数性质

定理 9.4.1 在模型的假设下,

$$(1) \quad \hat{\beta}_1 \sim N\left(\beta_1, \sigma^2 / S_{xx}\right)$$

$$(2) \quad \hat{\beta}_0 \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) \sigma^2\right)$$

## 证明 (1)

因为  $\hat{\beta}_1 = S_{xy} / S_{xx} = S_{xx}^{-1} \sum_i (x_i - \bar{x}) y_i$ ,

即为正态随机变量的线性组合，所以服从正态分布。

$$\begin{aligned} E(\hat{\beta}_1) &= S_{xx}^{-1} \sum_i (x_i - \bar{x}) E(y_i) = S_{xx}^{-1} \sum_i (x_i - \bar{x}) (\beta_0 + \beta_1 x_i) \\ &= \beta_1 S_{xx}^{-1} \sum_i (x_i - \bar{x}) x_i = \beta_1 S_{xx}^{-1} \sum_i (x_i - \bar{x})^2 = \beta_1 \end{aligned}$$

$$Var(\hat{\beta}_1) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{S_{xx}^2} \sigma^2 = \frac{\sigma^2}{S_{xx}}$$

**(2)** 类似可得。



### (3) 回归方程显著性检验

采用最小二乘法估计参数 $\beta_0, \beta_1$ ，并不需要事先知道 $y$ 与 $x$ 之间一定具有相关关系。

因此 $\mu(x)$ 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  $H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0,$

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。

对于假设  $H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0$ ,

可以从两个角度来看:

一是作为整体看, 检验y的回归方程是否显著。

例如, 假设y可能受到自变量 $x_1, x_2$ 的影响,

模型为

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \varepsilon,$$

则回归方程显著性检验相当于要检验:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0, H_1 : \beta_1, \beta_2, \beta_3 \text{不全为} 0,$$

另一个是作为单个参数来看，检验回归系数是否显著。

例如，在模型  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \varepsilon$

中，若  $\beta_1, \beta_2, \beta_3$  不全为0，则要进一步检验回归系数：

$H_{0i} : \beta_i = 0, H_{1i} : \beta_i \neq 0. \quad i=1,2,3$ ，即判断

$y$ 与 $x_1$ 的关系是线性的？二次的？与 $x_2$ 有关吗？

在一元线性回归  $y = \beta_0 + \beta_1 x + \varepsilon$  中

回归方程的检验与回归系数的检验合为一体，

变成检验  $H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0$ .

的两种方法。

# 回归方程的检验

采用方差分析方法:

令

$$SS_T \triangleq \sum (y_i - \bar{y})^2$$

描述  $y_1, y_2, \dots, y_n$  之间的总的差异大小,  
称  $SS_T$  为总平方和。

将总平方和分解为两部分：

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

$$SS_E \triangleq \sum (y_i - \hat{y}_i)^2 \quad \text{称为残差平方和}$$

$$SS_R \triangleq \sum (\hat{y}_i - \bar{y})^2 \quad \text{称为回归平方和}$$

可以证明：  $SS_T = SS_R + SS_E$

$$SS_T = S_{yy}, SS_E = (n-2)s^2, SS_R = \hat{\beta}_1 S_{xy} = \hat{\beta}_1^2 S_{xx}$$

定理9.4.2 在模型的假设下

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

(2)  $\hat{\beta}_1$  与  $s^2$  相互独立,  $\therefore SS_R$  与  $SS_E$  独立;

(3) 当  $H_0$  为真时,  $\frac{SS_R}{\sigma^2} \sim \chi^2(1)$ , 从而

$$\frac{\hat{\beta}_1^2 S_{xx}}{s^2} \sim F(1, n-2), \text{ 或 } \frac{\hat{\beta}_1 \sqrt{S_{xx}}}{s} \sim t(n-2).$$

## 方差分析表

来源	平方和	自由度	均方	F 比
回归	$SS_R = \frac{S_{xy}^2}{S_{xx}}$	$f_R = 1$	$SS_R / 1$	$F = \frac{SS_R / 1}{SS_E / (n - 2)}$
残差	$SS_E = (n - 2)s^2$ $= SS_T - SS_R$	$f_E = n - 2$	$SS_E / (n - 2)$	
总的	$SS_T = S_{yy}$	$f_T = n - 1$		



$$H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0,$$

当  $H_0$  为真时, 即  $\beta_1 = 0$ ,

$$F = \frac{\hat{\beta}_1^2 S_{xx}}{s^2} \sim F(1, n-2) \quad ,$$

对于给定的显著水平  $\alpha$ ,

$F$  检验的拒绝域为  $W = \{F > F_\alpha(1, n-2)\}$ .

对于模型  $Y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon$ ,

观测值为  $(y_i, x_{i1}, \dots, x_{ip}), i = 1, \dots, n$ , 检验假设:

$H_0 : \beta_1 = \dots = \beta_p = 0, H_1 : \beta_1, \dots, \beta_p$  不全为0.

方差分析表

来源	平方和	自由度	均方	F 比
回归	$SS_R$	$f_R = p$	$SS_R / p$	$F = \frac{SS_R / p}{SS_E / (n - p - 1)}$
残差	$SS_E$ $= SS_T - SS_R$	$f_E =$ $n - p - 1$	$SS_E / (n - p - 1)$	
总的	$SS_T$	$f_T = n - 1$		

当  $F\text{比} > F_\alpha(p, n - p - 1)$  时, 拒绝原假设。

## 回归系数的检验

采用 $t$ 检验

当  $H_0$  为真时，即  $\beta_1 = 0$ ，

$$t = \frac{\hat{\beta}_1 \sqrt{S_{xx}}}{s} \sim t(n-2),$$

对于给定的显著水平  $\alpha$ ，

$t$  检验的拒绝域为  $W = \{|t| > t_{\alpha/2}(n-2)\}$ ；

例3.2 检验例3.1中回归效果是否显著，取  $\alpha=0.05$ 。

$$\hat{\beta}_1 = 0.4646, \quad S_{xx} = 171.6, \quad s^2 = 0.186$$

查表得：  $t_{\alpha/2}(n-2) = t_{0.025}(8) = 2.306$ 。

假设  $H_0: \beta_1 = 0$  的检验拒绝域为  $|t| = \frac{|\hat{\beta}_1| \sqrt{S_{xx}}}{s} \geq 2.306$

计算得， $|t| = \frac{|0.4646|}{\sqrt{0.186}} \sqrt{171.6} = 14.1 > 2.306$ 。

故拒绝  $H_0: \beta_1 = 0$ ，认为回归效果是显著的。

## (4) 回归系数的置信区间

由 
$$\frac{(\hat{\beta}_1 - \beta_1)\sqrt{S_{xx}}}{s} \sim t(n-2)$$

得  $\beta_1$  的置信水平  $1-\alpha$  的置信区间：

$$\left( \hat{\beta}_1 \pm t_{\alpha/2}(n-2) \times \frac{s}{\sqrt{S_{xx}}} \right)$$

例如例3.1中 $\beta_1$ 的置信水平为0.95的置信区间为：

$$\left( 0.4646 \pm 2.306 \times \sqrt{\frac{0.186}{171.6}} \right) = (0.389, 0.541).$$

# 回归参数估计和显著性检验的Excel实现

例 3.3（续）前面我们已经分析了加热温度与燃气消耗量之间的关系，认为两者具有较好的线性关系，下面我们进一步建立燃气消耗量(响应变量)与加热温度(解释变量)之间的回归方程. 采用Excel中的“数据分析”模块.

在Excel的A1:C17输入下标:

	平均加热 温度	燃气用量	接前行	平均加热 温度	燃气用量
<b>1</b>	<b>24</b>	<b>6.3</b>	<b>9</b>	<b>0</b>	<b>1.2</b>
<b>2</b>	<b>51</b>	<b>10.9</b>	<b>10</b>	<b>1</b>	<b>1.2</b>
<b>3</b>	<b>43</b>	<b>8.9</b>	<b>11</b>	<b>6</b>	<b>2.1</b>
<b>4</b>	<b>33</b>	<b>7.5</b>	<b>12</b>	<b>12</b>	<b>3.1</b>
<b>5</b>	<b>26</b>	<b>5.3</b>	<b>13</b>	<b>30</b>	<b>6.4</b>
<b>6</b>	<b>13</b>	<b>4</b>	<b>14</b>	<b>32</b>	<b>7.2</b>
<b>7</b>	<b>4</b>	<b>1.7</b>	<b>15</b>	<b>52</b>	<b>11</b>
<b>8</b>	<b>0</b>	<b>1.2</b>	<b>16</b>	<b>30</b>	<b>6.9</b>

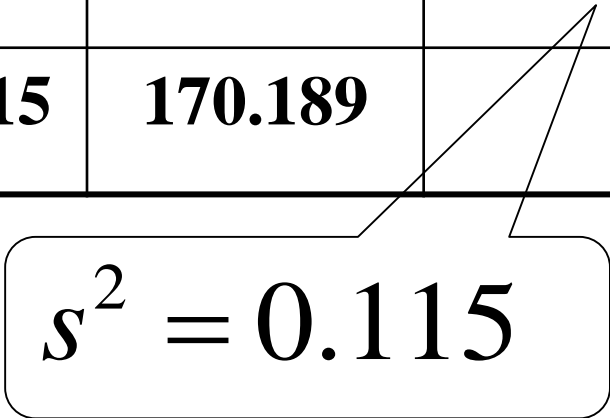


- 在Excel工作表中输入上面的数据 点击主菜单中“工具” 点击下拉式菜单中“数据分析” 就会出现一个“数据分析”的框，点击菜单中“回归”，点击“确定”，出现“回归”框。

- 在“Y值输入区域”中标定你已经输入的响应变量数据的位置（本例为\$C\$2:\$C\$17），
- 在“X值输入区域”中标定你已经输入的解釋变量数据的位置（注意：数据按“列”输入）（本例为\$B\$2:\$B\$17），“置信度”中输入你已经确定置信度的值选定输出结果的位置点击“确定”。
- 在指定位置输出相应的方差分析表和回归系数输出结果，例3.3的输出结果如下所示，

# 方差分析表

	自由度	平方和	均方	F值	P_值
回归	1	168.581	168.581	1467.551	1.415E-15 显著！
误差	14	1.608	0.115		
总的	15	170.189			


$$s^2 = 0.115$$

与方差分析中P-值一致!

	Coef.	标准误差	t Stat	P value	Lower 95%	Upper 95%
Intercept	1.089	0.139	7.841	1.729E-06	0.791	1.387
X	0.189	0.005	38.309	1.415E-15	0.178	0.200

$$\hat{\beta}_0 = 1.089$$

$$\hat{\beta}_1 = 0.189$$

$$\frac{|\hat{\beta}_1| \sqrt{S_{xx}}}{s} = 38.309$$

$$\left( \hat{\beta}_1 \pm t_{\alpha/2}(n-2) \times \frac{s}{\sqrt{S_{xx}}} \right) = (0.178, 0.200)$$

# 预测

预测一般有两种意义.

- 当给定  $x = x_0$  时, 求相应响应变量平均值即  $E[y_0]$  的点估计和区间估计, 在例 3.3 中的意义是: 求某个加热温度下, 燃气消耗量的平均值, 如平均加热温度为 10 度, 这个月份燃气消耗量的平均值.
- 当给定  $x = x_0$  时, 求  $y_0$  的预测值和预测区间, 在例 3.3 中的意义是: 求指定某个月的燃气消耗量, 如假设某个月的平均加热温度为 10 度, 预测这个月的燃气消耗量.

## (5) $E(y_0)$ 的点估计及置信区间

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \sim N\left(\beta_0 + \beta_1 x_0, \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\sigma^2\right)$$

故 $\hat{y}_0$ 作为 $E(y_0)$ 的点估计，是无偏估计.

$E(y_0)$ 的置信水平为 $1-\alpha$ 的置信区间为

$$\hat{y}_0 \pm t_{\alpha/2}(n-2) \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

## (6) $y_0$ 的点预测及区间预测

$$y_0 = \beta_0 + \beta_1 x_0 + \varepsilon$$

因此，根据观测结果，点预测为

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

所谓预测的精度是希望求出一个  $\delta$  的值，使，

$$P(|y_0 - \hat{y}_0| < \delta) = 1 - \alpha$$

其中  $\alpha$  是预先给定的一个小的正数， $0 < \alpha < 1$ 。 $\delta$  越小就表示预测的精度越高，并称：

$$(\hat{y}_0 - \delta, \hat{y}_0 + \delta)$$

为  $y_0$  的概率是  $1 - \alpha$  的预测区间。



由于  $\hat{y}_0$  与  $y_0$  独立，故知：

$$y_0 - \hat{y}_0 \sim N\left(0, \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\sigma^2\right)$$

并且  $s^2$  作为  $\sigma^2$  的无偏估计，与  $y_0 - \hat{y}_0$  独立，

所以有

$$\frac{y_0 - \hat{y}_0}{s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t(n-2)$$

对给定的  $\alpha$

$$\delta = t_{\frac{\alpha}{2}}(n-2) \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

从上式可看出，为了提高预测精度， $n$  应该足够大，并且  $x_1, x_2, \dots, x_n$  不能太集中。

例 3.4(续例 3.3)由前面的 Excel 的输出结果,  
计算设  $x_0 = 5$  时,  $E[y_0]$  的区间估计和  $y_0$   
的预测区间.

**Excel**只能输出预测值，无法输出预测区间。  
预测区间计算如下：

上例中**x**值置于**B2:B17**,**y**值置于**C2:C17**,  
在**Excel**第**18**行,  
**B18,C18,D18,E18,F18,G18**分别为

5	2.0342	22.313	299.723	4719.438	2.145
---	--------	--------	---------	----------	-------

$x_0$

$\hat{y}_0 : " = FORECAST(B18, C2 : C17, B2 : B17) "$

<b>5</b>	<b>2.0342</b>	<b>22.313</b>	<b>299.723</b>	<b>4719.438</b>	<b>2.145</b>
----------	---------------	---------------	----------------	-----------------	--------------

$\bar{x} : " = AVERAGE(B2:B17)"$

$S_{xx} : " = VAR(B2:B17)*15"$

$(x_0 - \bar{x})^2 : " = (B18 - D18)^2"$

$t_{0.025}(14) : " = T.INV(0.975,14)"$

利用以上数据，注意到， $s = \sqrt{0.115}$ ，得

$E(y_0)$ 的置信度为95%的置信区间:(1.776, 2.292),

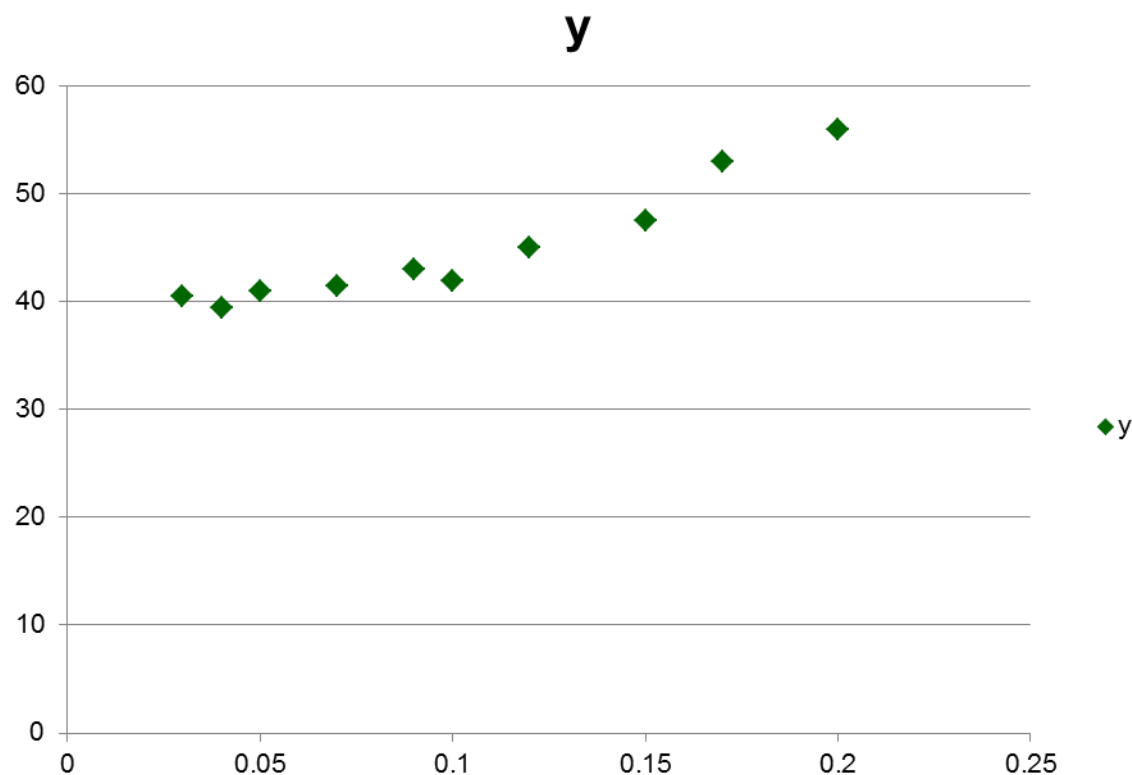
$y_0$ 的置信度为95%的预测区间:(1.263, 2.806).

例3.5 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（ $\text{kg/mm}^2$ ）数据见右表：

$y$	$x$	$x^2$
40.5	0.03	0.0009
39.5	0.04	0.0016
41	0.05	0.0025
41.5	0.07	0.0049
43	0.09	0.0081
42	0.1	0.01
45	0.12	0.0144
47.5	0.15	0.0225
53	0.17	0.0289
56	0.2	0.04

- (1) 画出散点图;
- (2) 设 $\mu(x)=\beta_0+\beta_1x$ , 求 $\beta_0, \beta_1$ 的估计;
- (3) 求误差方差的估计, 画出残差图;
- (4) 检验回归系数 $\beta_1$ 是否为零 (取  $\alpha=0.05$ );
- (5) 求回归系数 $\beta_1$ 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。
- (8) 模型还可以改进吗?

# (1) 合金钢的强度 $y$ 与钢材中碳的含量 $x$ 的散点图





## 方差分析

	df	SS	MS	F	Significance F
回归	1	255.4116	255.4116	74.33289	2.54E-05
残差	8	27.48841	3.436051		显著
总计	9	282.9			

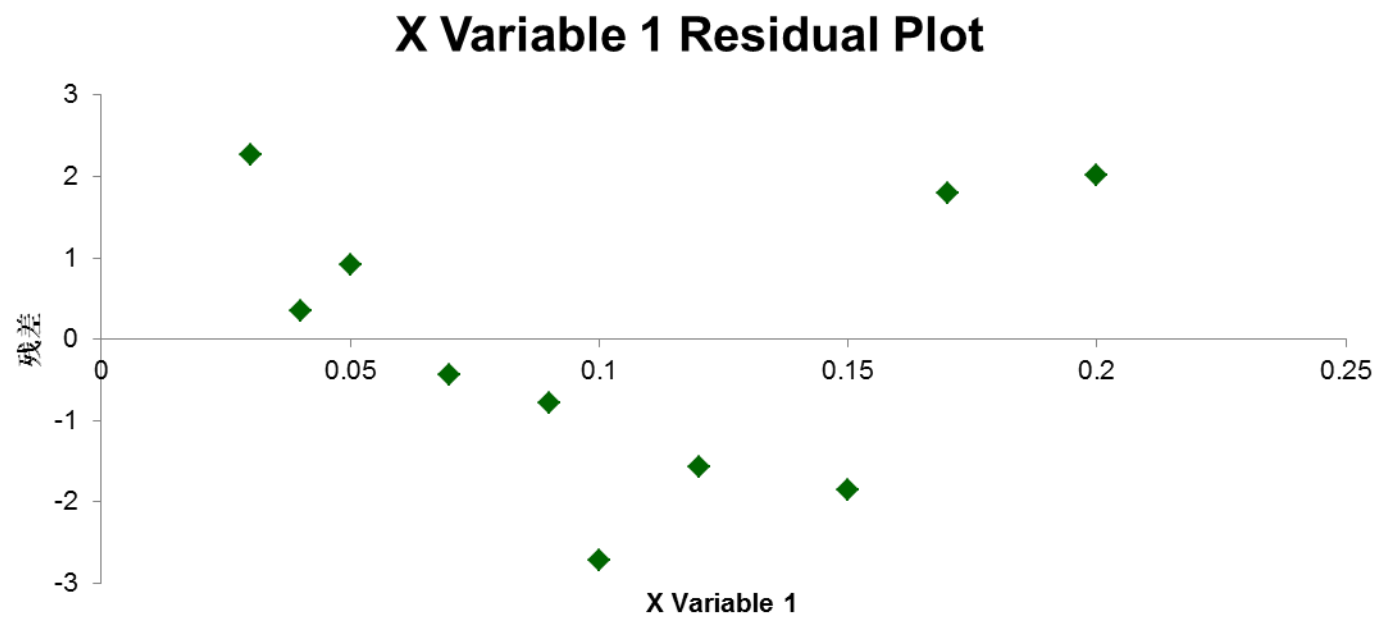
	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%
Intercept	35.4506	1.24292	28.5222	2.47E-09 显著	32.5844	38.3168
X Var. 1	92.6411	10.7452	8.62165	2.54E-05 显著	67.8627	117.420

$\beta_0, \beta_1$ 的最小二乘估计:  $\hat{\beta}_0 = 35.4506, \hat{\beta}_1 = 92.6411$

回归方程:  $\hat{y} = 35.4506 + 92.6411x$ .

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

$\sigma^2$ 的无偏估计值  $s^2 = \frac{SS_E}{n-2} = 3.436$



(4) 检验假设  $H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0$

显著水平为0.05

$$|t| = 8.6217 \geq t_{0.025}(8) = 2.306,$$

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  $\beta_1$  的 置信水平95% 的置信区间:  
(67.8629, 117.4193).

(6) 当 $x_0 = 0.06$ 时,  $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = 41.0091$

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).

(8) 进一步从残差图中发现，模型应包含二次项，即

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

*Excel*分析，将X值输入区域改为：B2:C11. 结果如下：

方差分析					
	df	SS	MS	F	Significance F
回归分析	2	276.3151	138.1576	146.8669	1.92E-06
残差	7	6.584894	0.940699		显著
总计	9	282.9			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%
Intercept	40.644	1.27932	31.7699	7.91E-09	37.6188	43.6691
X Var. 1	-30.483	26.7175	-1.14095	0.29142	-93.66	32.6935
X Var. 2	550.475	116.776	4.71394	0.00217	274.344	826.606

不显著

(8) 模型包含二次项后发现一次项不显著，修改为

$$Y = \beta_0 + \beta_2 x^2 + \varepsilon$$

*Excel*分析，将X值输入区域改为：C2:C11. 结果如下：

方差分析					
	df	SS	MS	F	Significance F
回归分析	1	275.0905	275.0905	281.8022	1.61E-07
残差	8	7.809465	0.976183		
总计	9	282.9			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%
Intercept	39.2774	0.45804	85.7509	3.82E-13	38.2212	40.3337
X Var. 1	420.223	25.0327	16.7870	1.61E-07	362.497	477.948

## 9.6 回归诊断

- 回归函数线性的诊断
- 误差方差齐性诊断
- 误差的独立性诊断
- 误差的正态性诊断



# 一、回归函数线性的诊断

## (1) 模型诊断

诊断回归函数是否是  $x_1, x_2, \dots, x_t$  的线性函

数的主要工具是以拟合值  $\hat{y}$  为横坐标的残差图。

在  $t = 1$  时不可用散点图。

$i$	$x$	$y$	$\hat{y}$	$\hat{e}$
1	80	0.6	1.6625	-1.0625
2	220	6.7	7.75	-1.05
3	140	5.3	4.27143	1.02857
4	120	4	3.40179	0.59821
5	180	6.55	6.01071	0.53929
6	100	2.15	2.53214	-0.38214
7	200	6.6	6.88036	-0.28036
8	160	5.75	5.14107	0.60893

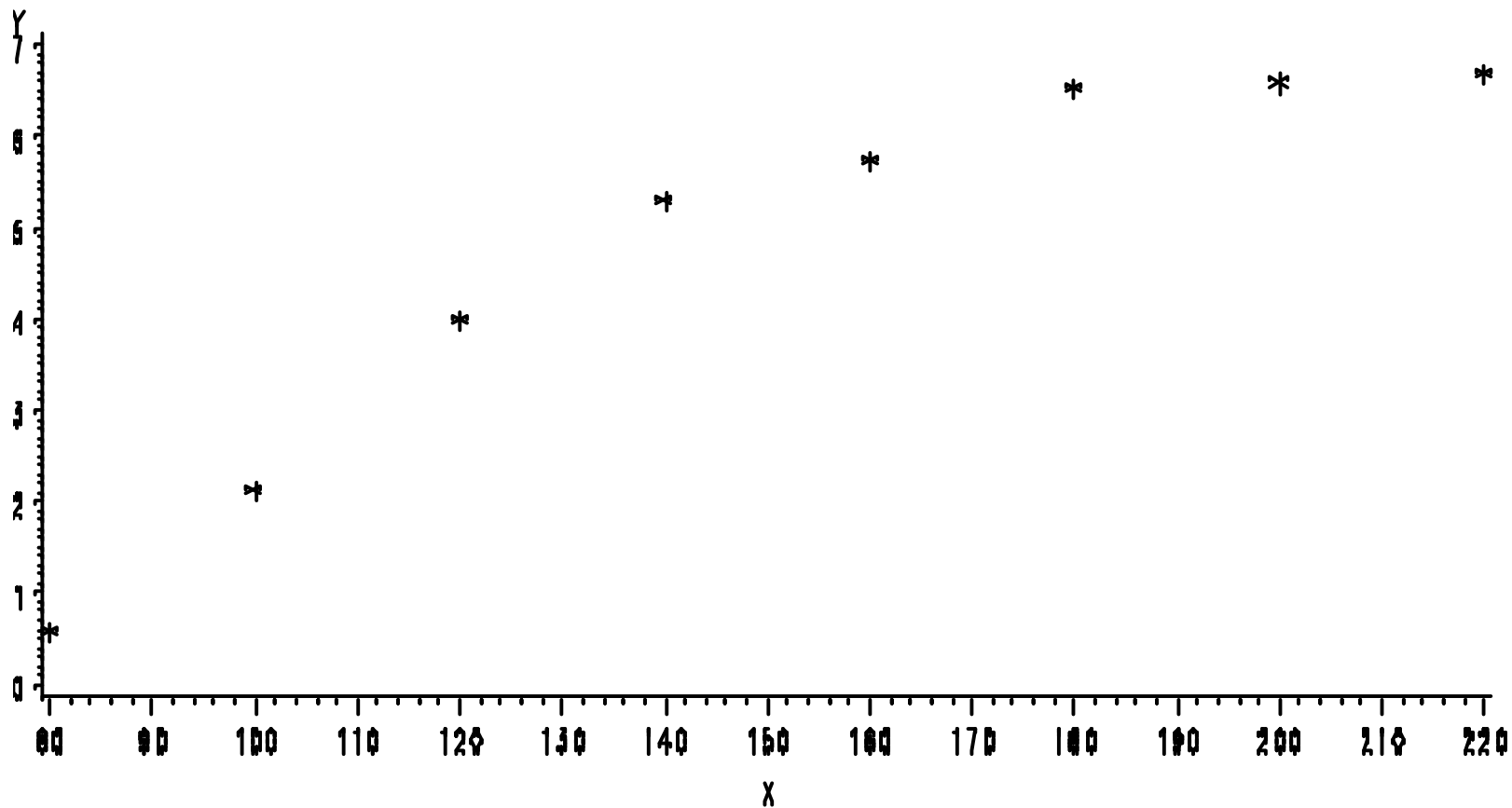
由上述数据，可得  $y$  关于  $x$  的一元线性回归方程

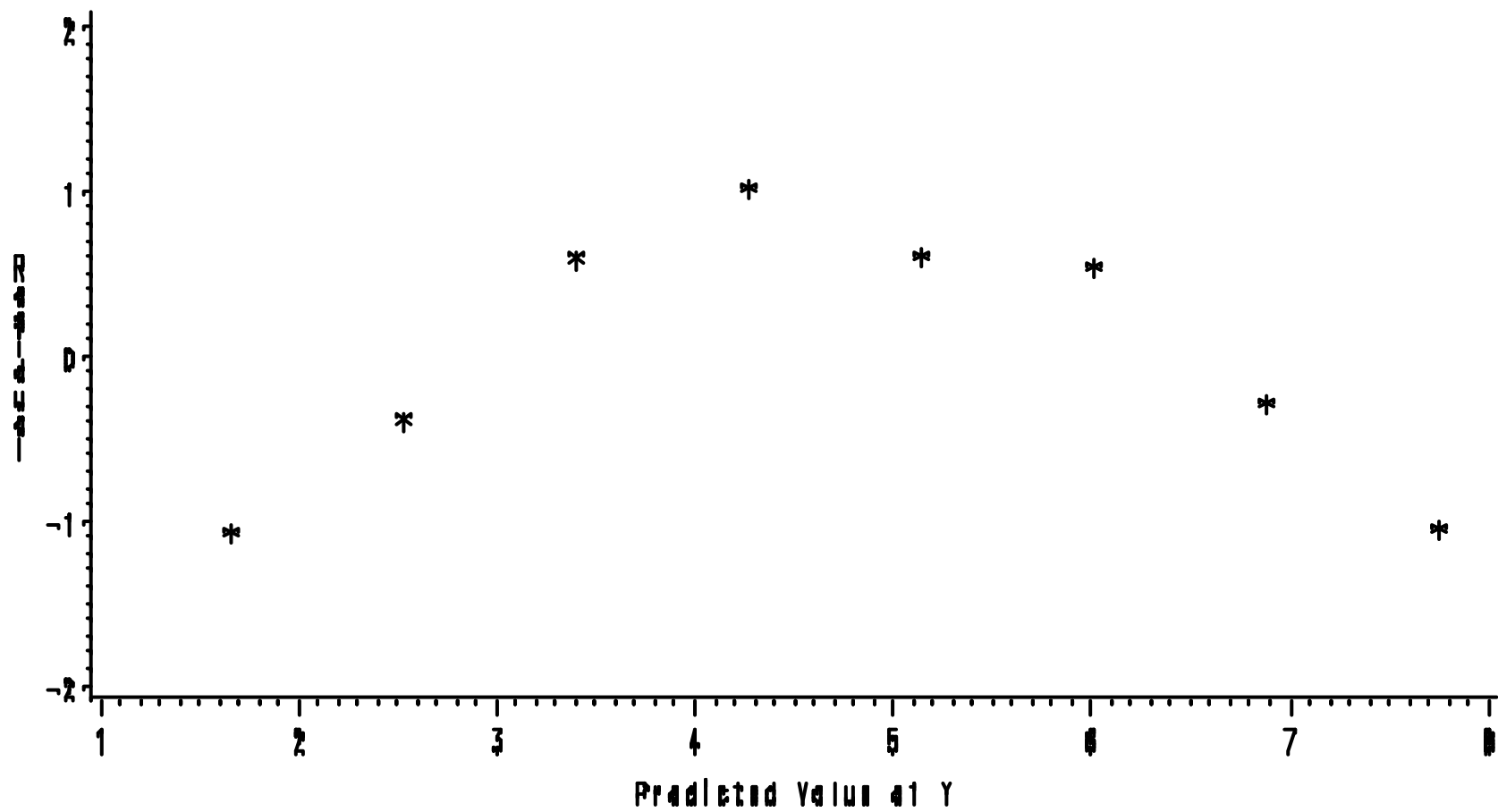
$$\hat{y} = -1.82 + 0.0435x$$

回归方程所对应的  $\sigma$  的估计值及相关系数分别为：

$$s = 0.869 \quad r = 0.935$$

其拟合值及残差见表。其散点图(a)及残差图(b)如下





- 从散点图可以看出， $x$  与  $y$  之间的关系用线性函数去描述不太适合，最好看成是  $x$  的二次函数；
  - 从残差图发现，点的散布是有规律的，对就  $\hat{y}$  小的及  $\hat{y}$  大的残差为负，而  $\hat{y}$  介于中间的点的残差是正的。
- 从上述的分析，我们有理由怀疑回归函数线性这一假定是不成立的。

## (2) 模型修正

从上图可见， $x$  较小或较大时，残差小于 0， $x$  介于中间值时，残差大于 0，从而设想改变回归模型，建立  $y$  关于  $x$ 、 $x^2$  的回归方程。其基本模型为

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, i = 1, 2, \dots, n$$

---

经计算可得回归方程如下：

$$\hat{y} = -10.028 + 0.16424x + -0.0040205x^2$$

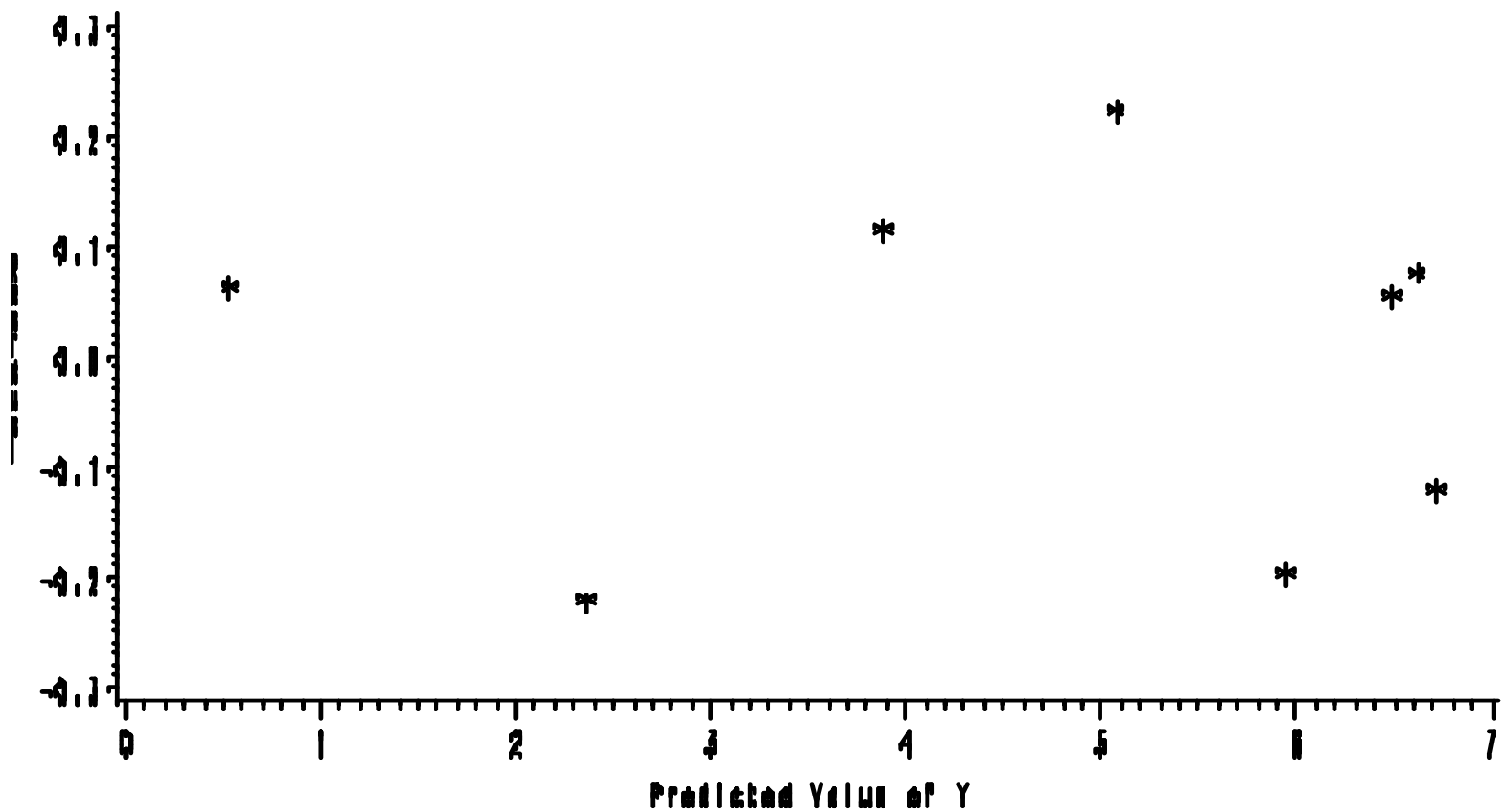
其拟合值和残差见表 3.3.2，残差图见图 3.3.3。从残差图可见点的分布已无规律，说明用二次回归方程是合适的。



# 模型修改后的预测值及残差

I	$\hat{y}$	$\hat{e}$
1	0.53542	0.06458
2	6.62292	0.07708
3	5.07649	0.22351
4	3.88482	0.11518
5	6.49375	0.05625
6	2.37113	-0.22113
7	6.71935	-0.11935
8	5.94613	-0.19613

# 模型修改后的残差图



## 二、误差方差齐性诊断

对线性回归模型，要诊断  $\sigma_i^2$  是否相等。

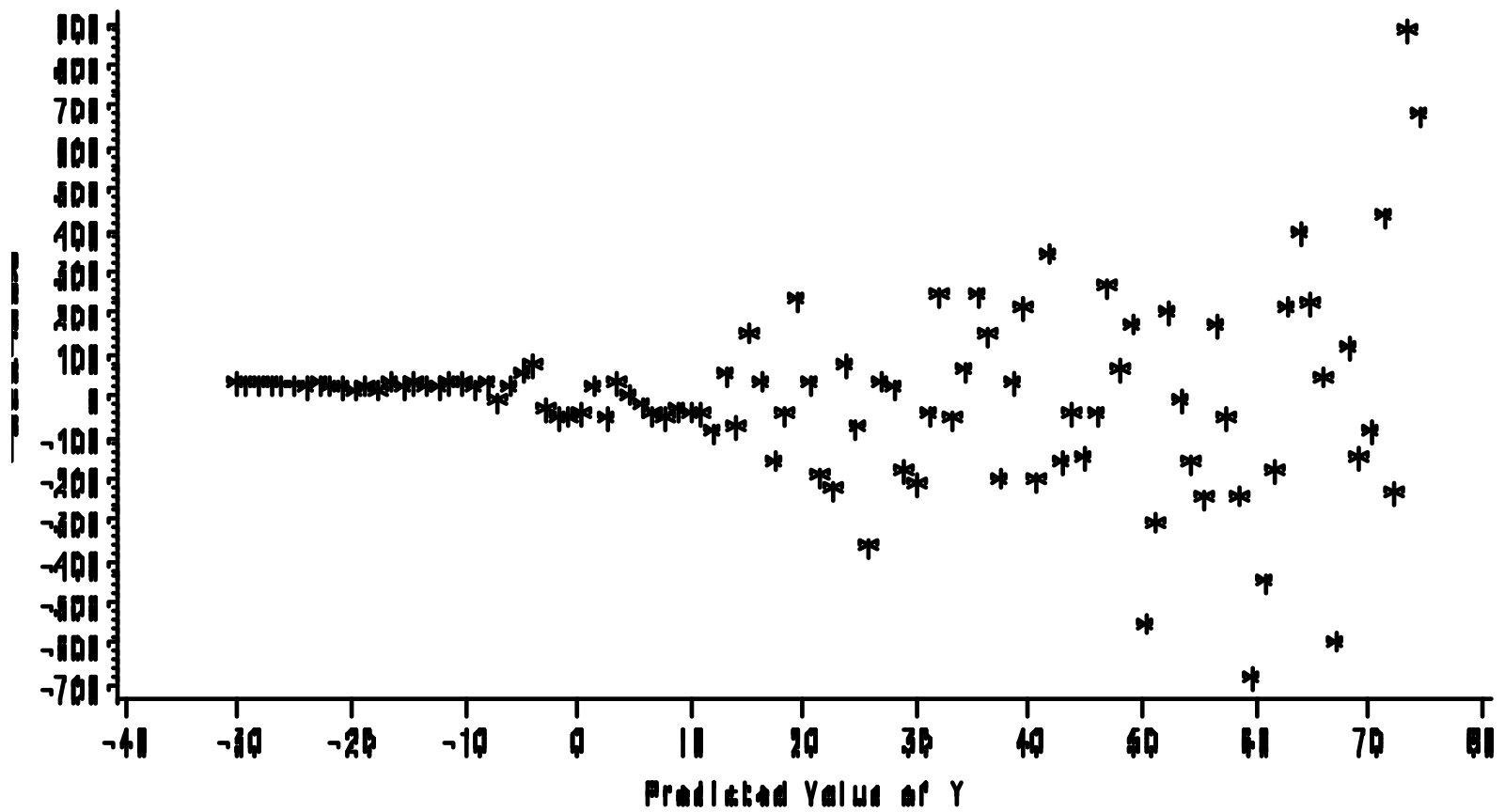
即要检验：

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2。$$

## (1) 模型诊断

常用的对误差方差的诊断通常用的方法是残差图。

误差方差非齐性时，以拟合值  $\hat{y}$  为横坐标的残差图一般是呈现喇叭型或倒喇叭型。





## (2) 模型修正

- 如果发现线性假设是不适合, 那么就需要修改模型. 在目前的回归分析的知识水平下, 不一定能很好地修改误差方差不相等这类模型, 但可以尝试响应变量的数据变换。

- 用变换后的数据, 求出线性回归方程, 求出残差, 并画出以拟合值为横坐标的残差图, 如果这里残差图已经没有任何规律, 那么说明这种变换是适合的.



常见的变换有：

$$Z = \sqrt{Y} ,$$

$$Z = \arcsin \sqrt{Y} ,$$

$$Z = \ln Y ,$$

$$Z = 1/\sqrt{Y} ,$$

$$Z = Y^{-1} .$$

### 三、误差的独立性诊断

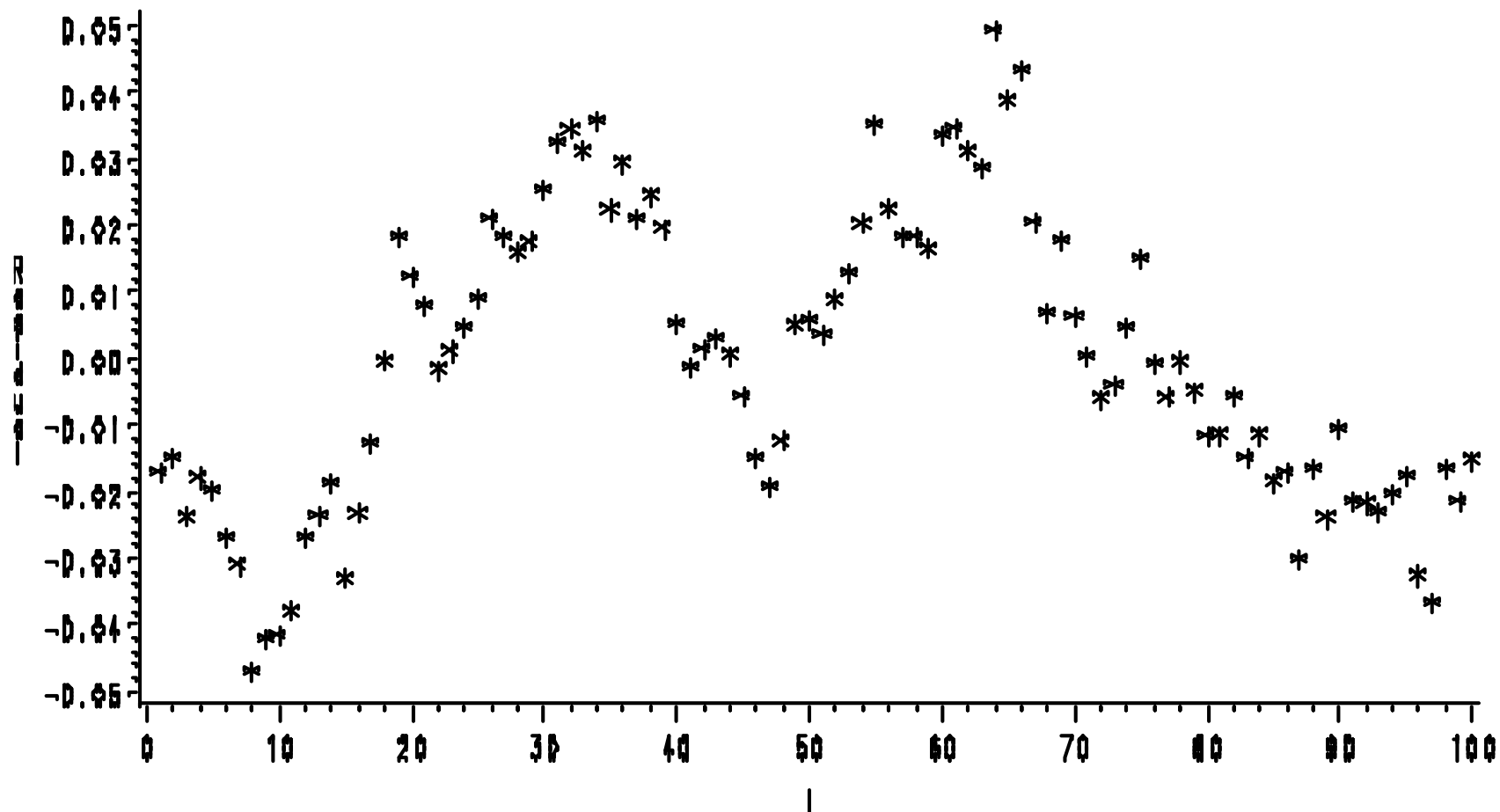
- 在不少有关时间问题中，观测值往往呈相关的趋势。如河流的水位总有一个变化过程，当一场暴雨使河流水位上涨后往往需要几天才能使水位降低，因而当我们逐日测定河流最高水位时，相邻两天的观测间就不一定独立。

## (1)模型诊断

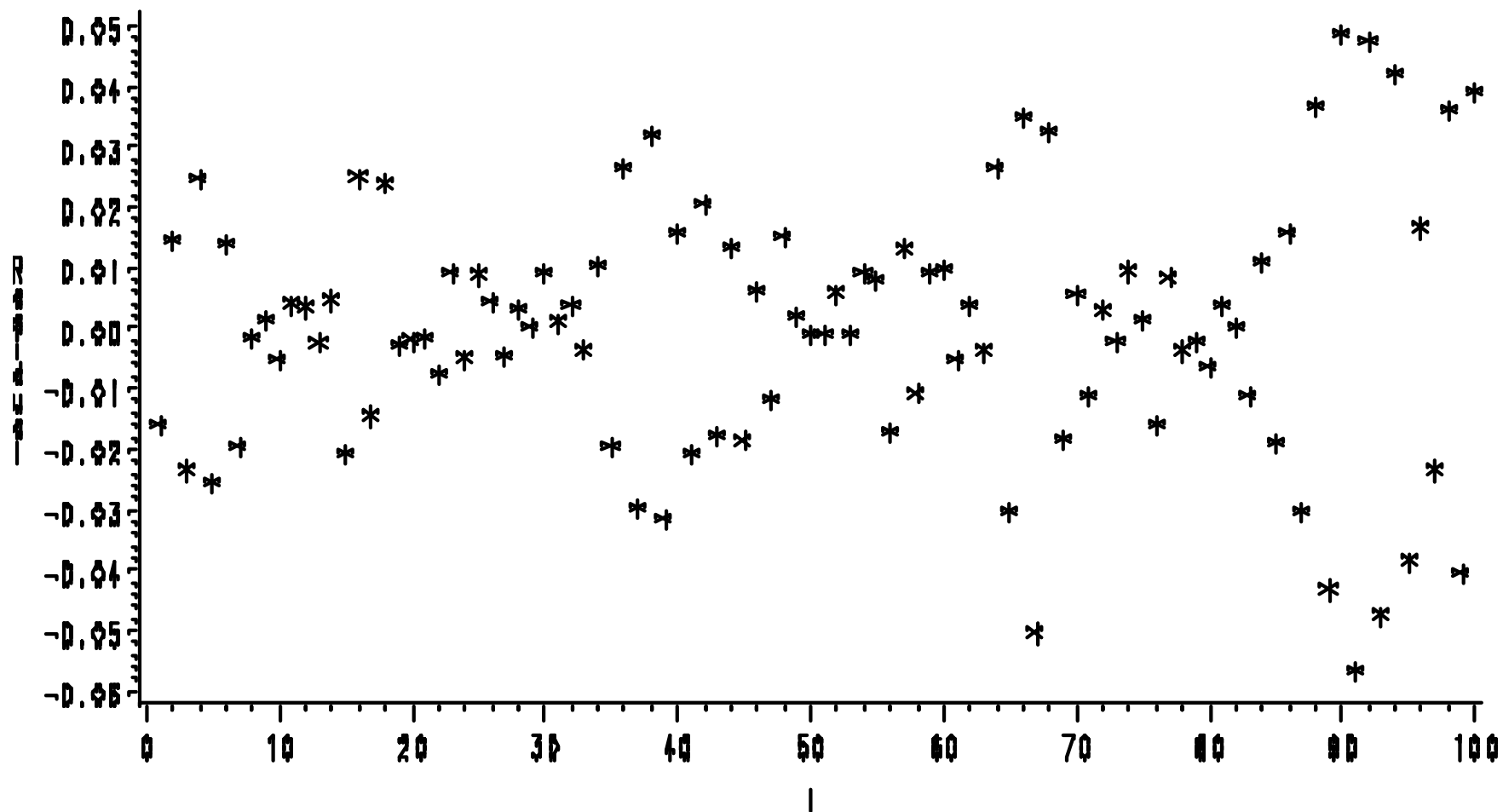
常用的残差图是以“时间”或“序号”为横坐标的残差图。相关性大约有二类。

- 一类是正相关, 随机误差之间具有正相关的话,那么残差图中残差“符号”会出现“集团性”的趋势,即连续有一段时间内残差均为“正号”,然后又一段时间内残差均为“负号”
- 另一类是负相关, 此时,残差的符号改变非常频繁,大致有正负相间的趋势.

# 残差图



# 残差图



## (2) 模型修改

可用差分法，其步骤如下：

- 先建立  $y$  关于  $x$  的回归方程，并求出相应各点的残差  $e_1, e_2, \dots, e_n$ ，若认为存在序列相关，则进行下述各步骤；

- 利用已求出的残差，求序列  $e_1, e_2, \dots, e_{n-1}$

与序列  $e_2, e_3, \dots, e_n$  的相关系数，记为  $r$ ；

- 令  $y_i^* = y_{i+1} - ry_i,$

$$x_i^* = x_{i+1} - rx_i, i = 1, 2, \dots, n-1$$



- 先建立  $y_i^*$  关于  $x_i^*$  的回归方程

$$y_i^* = \alpha + \beta x_i^*,$$

并求出相应各点的残差  $e_1^{(1)}, e_2^{(1)}, \dots, e_{n-1}^{(1)}$ ，作残差图，

如果仍不满意，继续上述步骤。

## 四、误差的正态性诊断

- 我们可采用卡方拟合检验对残差进行正态性的检验,也可以用残差画一下直方图,直观地判断残差量是不是具有正态性.
- 如果模型的误差不满足正态性时,一般可以作**Box-Cox**变换,这部分的内容这里不详细介绍,有兴趣的同学可以参考有关的回归分析的参考文献.