# 第九章 方差分析及回归分析

# 主要内容

- 9.1 单因素试验的方差分析
- 9.3 一元线性回归

# 9.1 单因素试验的方差分析

# 两个例子

- 考察不同机器（机器1、机器2、机器3）生产的薄板的厚度有无显著差异

- 考察不同电路类型的电路响应时间有无显著差异

# 两个概念：因素和水平

- 考察不同机器（机器1、机器2、机器3）生产的薄板的厚度有无显著差异
- 在这个问题中，机器就是试验的因素，这个因素有3个水平：机器1、机器2、机器3
- 方差分析就是考察不同水平的总体均数是否存在显著差异

# 方法原理(1)

| 因素水平 | $A_1$ | $A_2$ | $\cdots$ | $A_s$ |
|---|---|---|---|---|
| 样本观测值 | $X_{11}$ | $X_{12}$ | $\cdots$ | $X_{1s}$ |
| | $X_{21}$ | $X_{22}$ | $\cdots$ | $X_{2s}$ |
| | $\cdots$ | | | |
| | $X_{n11}$ | $X_{n22}$ | $\cdots$ | $X_{nss}$ |
| 样本均值 | $\overline{X}_{\cdot 1}$ | $\overline{X}_{\cdot 2}$ | $\cdots$ | $\overline{X}_{\cdot s}$ |
| 总体均值 | $\mu_1$ | $\mu_2$ | $\cdots$ | $\mu_S$ |

# 方法原理(2)

- 方差分析的任务：

(1) 检验 $s$ 个总体$N(\mu_1, \sigma^2)$，…，$N(\mu_s, \sigma^2)$的均值是否相等，即检验假设

$$H_0: \boldsymbol{\mu_1 = \mu_2 = \ldots = \mu_s}$$

$$H_1: \boldsymbol{\mu_1, \mu_2, \ldots, \mu_s}\text{不全相等}$$

(2) 作出未知参数$\boldsymbol{\mu_1, \mu_2, \ldots, \mu_s, \sigma^2}$的估计

# 方法原理(3)

- 平方和的分解

方差分析是基于变异分解的原理进行的，在单因素方差分析中，整个样本的变异（样本观测值之间的差异）由如下两个部份构成：

总偏差平方和$S_T$ = 误差平方和$S_E$ + 效应平方和$S_A$

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

总变异($S_T$) = 水平内变异($S_E$) + 水平间变异($S_A$)

总变异 = 随机变异 + 试验因素导致的变异

# 方法原理(4)

- **经分析可知：**

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

$$\frac{S_A}{\sigma^2} \sim \chi^2(s-1)\ \ (当 H_0 真时),\ 且 E(\frac{S_A}{s-1})=\sigma^2$$

$S_E$与$S_A$相互独立

- **故：**

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

# 方法原理(5)

若$H_0$真，则 $\dfrac{S_A/(s-1)}{S_E/(n-s)}$的值接近1是大概率事件

反之，$\dfrac{S_A/(s-1)}{S_E/(n-s)}$的值右向偏离1较大是小概率事件

令 $P(\dfrac{S_A/(s-1)}{S_E/(n-s)} \geq k) = \alpha$，可得 $k = F_\alpha(s-1, n-s)$

故拒绝域为 $\dfrac{S_A/(s-1)}{S_E/(n-s)} \geq F_\alpha(s-1, n-s)$

# 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 | $\overline{S}_A = \frac{S_A}{s-1}$ | $F = \frac{\overline{S}_A}{\overline{S}_E}$ |
| 误差 | $S_E$ | $n$-$s$ | $\overline{S}_E = \frac{S_A}{n-s}$ | |
| 总和 | $S_T$ | $n$-1 | | |

## $S_T$、$S_A$和$S_E$的简便计算

记 $T_{\cdot j}=\sum_{i=1}^{n_j}X_{ij}$, $j=1,2,\dots,s$ ,(第$j$列所有样本值的和)

记 $T_{\cdot\cdot}=\sum_{j=1}^{s}\sum_{i=1}^{n_j}X_{ij}$, (所有样本值的和)

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

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

$$S_E=S_T-S_A$$

# 例1(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。假设三台机器相互独立，它们生产的薄板的厚度均服从正态分布且方差相同，试检验三种机器生产的薄板的厚度有无显著性差异（$\alpha = 0.05$）。

| 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 |

$T_{\bullet j}$ 1.21 1.28 1.31

# 例1(2)

- 分析：这是单因素方差分析的问题。
- 解：$H_0$： $\mu_1 = \mu_2 = \mu_3$

  $H_1$： $\mu_1$，$\mu_2$，$\mu_3$不全相等

  现在 $s = 3$，$n_1 = n_2 = n_3 = 5$，$n = 15$

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

  即 $F = \dfrac{S_A/(s-1)}{S_E/(n-s)} \geq F_{0.05}(2,12)=3.89$

# 例1(3)

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

$$=0.963\,912-\frac{3.8^2}{15}=0.001\,245\,33$$

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

$$=\frac{1}{5}(1.21^2+1.28^2+1.31^2)-\frac{3.8^2}{15}=0.001\,053\,33$$

$$S_E=S_T-S_A=0.000\,192$$

$S_T,S_A,S_E$的自由度分别为$n-1=14,s-1=2,n-s=12$

# 例1(4)

- 方差分析表

| 方差来源 | 平方和 | 自由度 | 均方 | *F* 比 |
|---|---|---|---|---|
| 因素A | 0.001 053 33 | 2 | 0.000 526 67 | 32.92 |
| 误差 | 0.000 192 | 12 | 0.000 016 | |
| 总和 | 0.001 245 33 | 14 | | |

- *F*比32.92 > 3.89, 落入拒绝域，故在显著性水平0. 05下拒绝$H_0$，即认为各台机器生产的薄板的厚度有显著性差异。

# 未知参数的估计

$\hat{\mu}=\overline{X}, \hat{\mu}_j=\overline{X}_{\cdot j}$ 分别是$\mu, \mu_j$ 的无偏估计

$$\hat{\sigma}^2=\frac{S_E}{n-s}=\frac{\sum_{j=1}^{s}\sum_{i=1}^{n_j}(X_{ij}-\overline{X}_{\cdot j})^2}{n-s}$$ 是$\sigma^2$的无偏估计

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

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

# 例2

- 求例1中未知参数$\mu_1$，$\mu_2$，$\mu_3$，$\sigma^2$的点估计值。
- 解:

$$\hat{\sigma}^2 = \frac{S_E}{n-s} = \frac{0.000\,192}{12} = 0.000\,016$$

$$\hat{\mu}_1 = \bar{x}_{\cdot 1} = \frac{1.21}{5} = 0.242, \qquad \hat{\mu}_2 = \bar{x}_{\cdot 2} = \frac{1.28}{5} = 0.256$$

$$\hat{\mu}_3 = \bar{x}_{\cdot 3} = \frac{1.31}{5} = 0.262, \qquad \hat{\mu} = \bar{x} = \frac{3.8}{15} = 0.253$$

# 9.3 一元线性回归

# 变量间的关系(1)

变量间关系
- 确定性关系：$S=\pi r^2$
- 不确定性关系：（相关关系） 身高和体重；商品价格和销售量

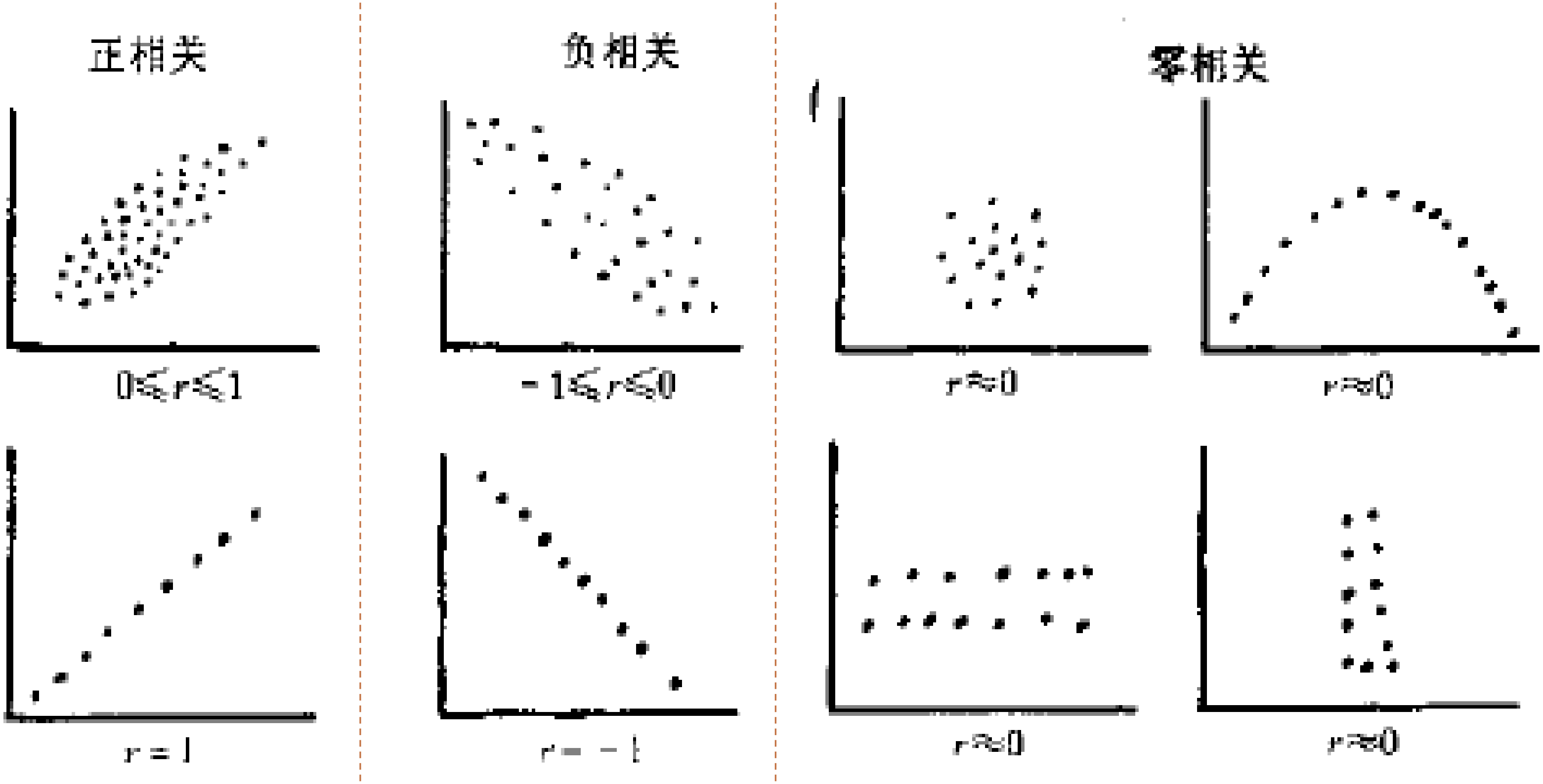
**回归分析是研究相关关系的一种工具：**

(1) 可以提供变量间相关关系的一个确定的数学表达式（经验公式）

(2) 可以检验所得到的经验公式是否有效

(3) 可以根据一个或几个变量的值，预测或控制另一个变量的取值

# 变量间的关系(2)



# 一元线性回归模型

容量为 $n$ 的二维样本：

$(x_1, y_1)$

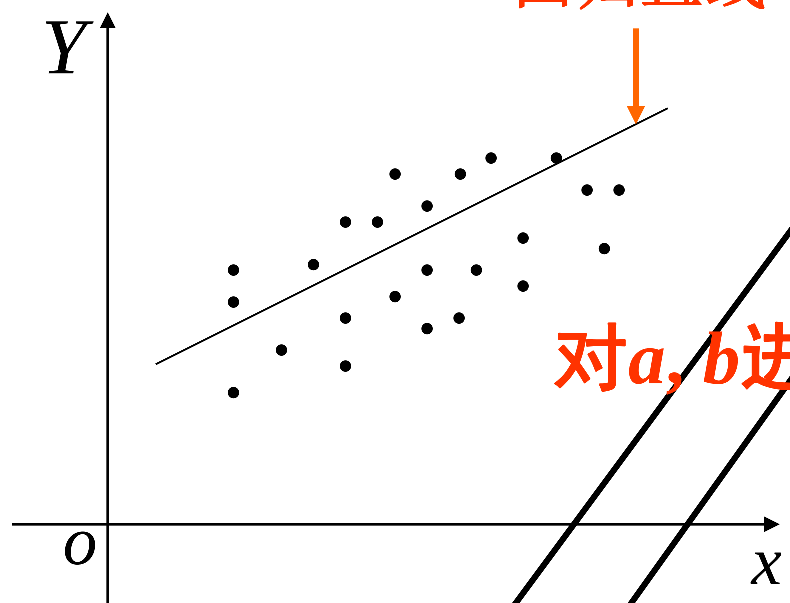
$(x_2, y_2)$

……

$(x_n, y_n)$

回归方程

回归直线 $\hat{y} = \hat{a} + \hat{b}x$




对$a, b$进行估计

一元线性回归模型 $\begin{cases} Y = a + bx + \varepsilon, \\ \varepsilon \sim N(0, \sigma^2), \quad a, b, \sigma^2 \text{为常数}. \end{cases}$

# *a*、*b*的估计(1)

- 如何估计*a*、*b*的值？
  - 最小二乘法：使各实测点距回归直线的纵向距离的平方和即残差的平方和达到最小的 *a* 和 *b* 是最优的。此平方和是关于*a*、*b* 的二元函数，记为 $Q(a, b)$。

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

使得上式取最小值的*a*、*b*就是最优的。

# $a$、$b$的估计(2)

令 $$\begin{cases} \dfrac{\partial Q}{\partial a} = -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0 \\ \dfrac{\partial Q}{\partial b} = -2\sum_{i=1}^{n}(y_i - a - bx_i)x_i = 0 \end{cases}$$

得方程组 $$\begin{cases} na + (\sum_{i=1}^{n} x_i)b = \sum_{i=1}^{n} y_i \\ (\sum_{i=1}^{n} x_i)a + (\sum_{i=1}^{n} x_i^2)b = \sum_{i=1}^{n} x_i y_i \end{cases}$$ (正规方程组)

# *a*、*b*的估计(3)

解得
$$\begin{cases} \hat{b} = \dfrac{\sum\limits_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2} \\ \hat{a} = \dfrac{1}{n}\sum\limits_{i=1}^{n} y_i - \dfrac{\hat{b}}{n}\sum\limits_{i=1}^{n} x_i = \overline{y} - \hat{b}\overline{x} \end{cases}$$

# $a$、$b$的估计(4)

令 $$S_{xx}=\sum_{i=1}^{n}(x_i-\overline{x})^2=\sum_{i=1}^{n}x_i^{\ 2}-\frac{1}{n}(\sum_{i=1}^{n}x_i)^2$$

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

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

则 $\hat{b}=\dfrac{S_{xy}}{S_{xx}}$, $\hat{a}=\overline{y}-\overline{x}\hat{b}$

从而回归方程为 $\hat{y}=\hat{a}+\hat{b}x$

# 例题(1)

- 为研究某一化学反应过程中，温度$x$（摄氏度）对产品得率$Y$（%）的影响，测得数据如下.

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

试求$Y$关于$x$的线性回归方程。

**解：** 可先画出$x$、$Y$的散点图，大致看一下两者是否有线性关系。

# 例题(2)

| | $x_i$ | $y_i$ | $x_i^2$ | $y_i^2$ | $x_i y_i$ |
|---|---|---|---|---|---|
| | 100 | 45 | 10000 | 2025 | 4500 |
| | 110 | 51 | 12100 | 2601 | 5610 |
| | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ |
| | 190 | 89 | 36100 | 7921 | 16910 |
| $\sum$ | 1450 | 673 | 218500 | 47225 | 101570 |

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

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

# 例题(3)

则 $\hat{b} = S_{xy} / S_{xx} = 3985 / 8250 = 0.48303$

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

从而回归直线方程为

$$\hat{y} = -2.73935 + 0.48303x$$

# 线性假设的显著性检验(1)

即使平面上 $n$ 个杂乱无章的样本点也可以得到回归方程，但实际上此时的回归方程毫无意义！

究竟在什么情况下所配的回归直线才有意义，回归方程真的揭示了$x$ 和$Y$ 之间存在线性关系的内在规律？

**问题：** $x$ 和$Y$ 之间是否有线性回归方程？**回归显著性检验**

# 线性假设的显著性检验(2)

作如下假设：

$$H_0: b = 0; \; H_1: b \neq 0$$

选择统计量：$t = \dfrac{\hat{b} - b}{\hat{\sigma}} \sqrt{S_{xx}} \sim t(n-2)$

其中 $\hat{\sigma} = \sqrt{\dfrac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\dfrac{S_{yy} - \hat{b} S_{xy}}{n-2}}$

拒绝域为：$|t| \geq t_{\alpha/2}(n-2)$

# 线性假设的显著性检验(3)

- **检验上例中的回归效果是否显著，取α=0.05.**

(1)作如下假设：$H_0$：$b=0$; $H_1$：$b\neq 0$

(2)选择统计量：$t=\dfrac{\hat{b}-b}{\hat{\sigma}}\sqrt{S_{xx}}\sim t(n-2)$

(3)拒绝域为：$|t|\geq t_{0.025}(8)$, 即$|t|\geq 2.306$

(4)已知$\hat{b}=0.48303, S_{xx}=8250, S_{xy}=3985$

$$S_{yy}=\sum_{i=1}^{n}y_i^2-\frac{1}{n}(\sum_{i=1}^{n}y_i)^2=47225-\frac{1}{10}\times 673^2=1932.1$$

$$\hat{\sigma}=\sqrt{\frac{S_{yy}-\hat{b}S_{xy}}{n-2}}=\sqrt{\frac{1932.1-0.48303\times 3985}{8}}=\sqrt{0.9}$$

# 线性假设的显著性检验(4)

现在 $|t| = \dfrac{0.48303}{\sqrt{0.9}} \times \sqrt{8250} = 46.25 > 2.306$

故落入拒绝域，从而拒绝$H_0$，即认为回归效果是显著的.

# 回归系数*b*的置信区间

*b*的置信水平为1-$\alpha$的置信区间为：

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

上例中*b*的置信水平为0.95的置信区间为：

$$(0.48303 \pm 2.306 \times \sqrt{\frac{0.9}{8250}}) = (0.45894, 0.50712)$$

# 回归函数$\mu(x) = a + bx$函数值的点估计和置信区间

- 取自变量$x = x_0$，则 $\mu(x_0) = a + b x_0$的点估计是：

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

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

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

注：这是对均值（$x_0$对应的$Y_0$所有取值的平均，是一个未知参数）的点估计和区间估计！

# $Y$的观察值的点预测和预测区间

- 取自变量$x = x_0$，其对应的新观测值$Y_0$的点预测是：

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

- $Y_0$的置信水平为1- $\alpha$的预测区间是：

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

- 注：这是对随机变量（$x_0$对应的随机变量$Y_0$）取值的点预测和区间预测！

# 可化为一元线性回归的例子

(1) $y = \alpha e^{\beta x}$

两边取对数：$\boxed{\ln y} = \ln \alpha + \beta \boxed{x}$

(2) $y = \alpha x^{\beta}$

两边取对数：$\boxed{\ln y} = \ln \alpha + \beta \boxed{\ln x}$

(3) $y = \dfrac{\alpha}{x - \beta}$

两边取倒数：$\boxed{\dfrac{1}{y}} = -\dfrac{\beta}{\alpha} + \dfrac{1}{\alpha} \boxed{x}$

# THE END