

第八章 一元回归及简单相关分析

第一节 回归及相关的基本概念

第二节 一元线性回归

第三节 回归方程的检验

第四节 预测及应用

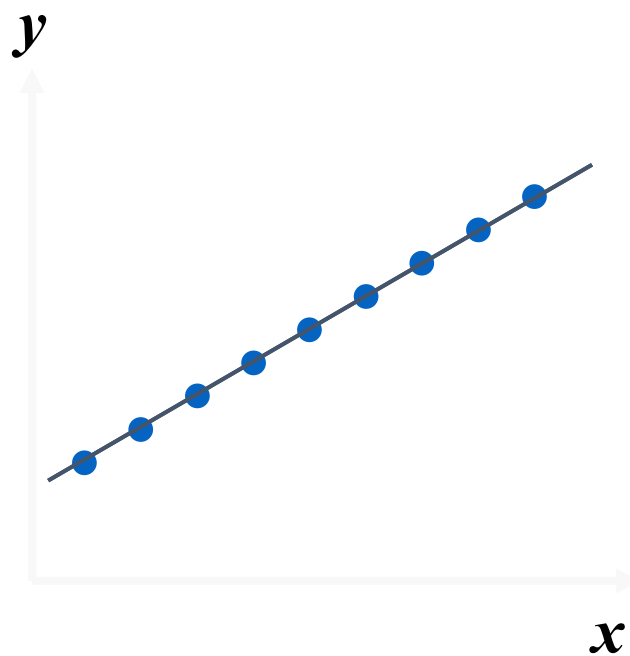
第五节 相关

第六节 可直线化的非线性回归

§ 8.1 回归与相关的基本概念

变量间的关系 (函数关系)

1. 是一一对应的确定关系
2. 设有两个变量 x 和 y ，变量 y 随变量 x 一起变化，并完全依赖于 x ，当变量 x 取某个数值时， y 依确定的关系取相应的值，则称 y 是 x 的函数，记为 $y = f(x)$ ，其中 x 称为自变量， y 称为因变量



变量间的关系

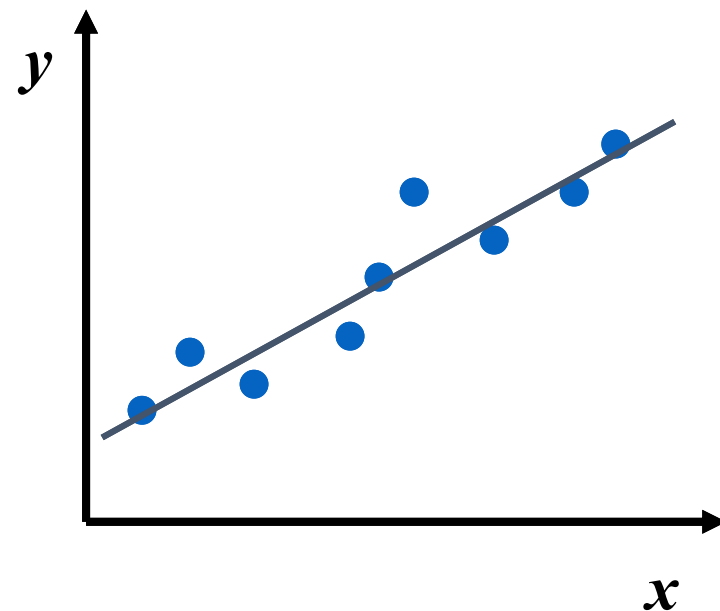
(函数关系)

➡ 函数关系的例子

- 某种商品的销售额(y)与销售量(x)之间的关系可表示为 $y = p x$ (p 为单价)
- 圆的面积(S)与半径之间的关系可表示为 $S = \pi R^2$
- 企业的原材料消耗额(y)与产量(x_1)、单位产量消耗(x_2)、原材料价格(x_3)之间的关系可表示为 $y = x_1 x_2 x_3$

变量间的关系 (相关关系)

1. 变量间关系不能用函数关系精确表达
2. 一个变量的取值不能由另一个变量唯一确定
3. 当变量 x 取某个值时，变量 y 的取值具有一个确定的分布



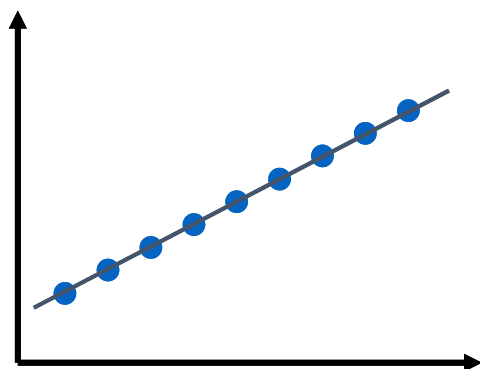
变量间的关系

(相关关系)

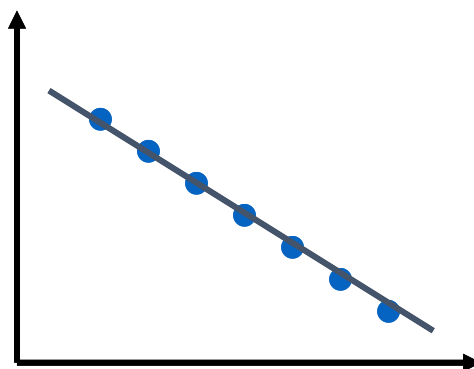
➡ 相关关系的例子

- 商品的消费量(y)与居民收入(x)之间的关系
- 商品销售额(y)与广告费支出(x)之间的关系
- 粮食亩产量(y)与施肥量(x_1)、降雨量(x_2)、温度(x_3)之间的关系
- 收入水平(y)与受教育程度(x)之间的关系
- 父亲身高(y)与子女身高(x)之间的关系

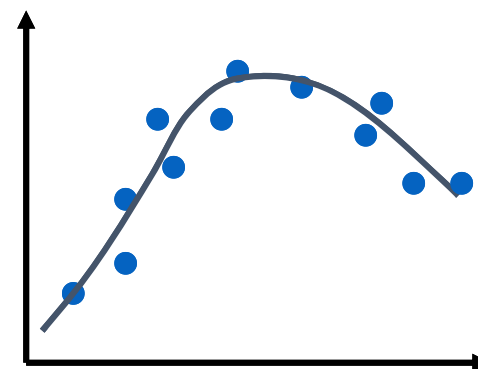
相关关系的图示



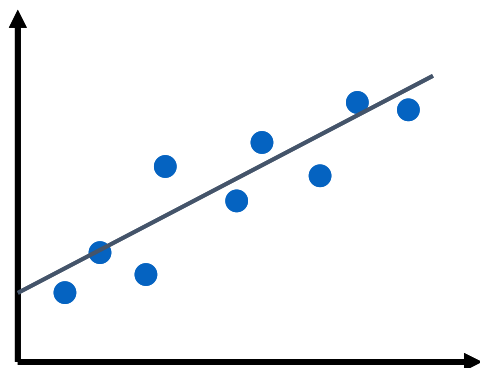
完全正线性相关



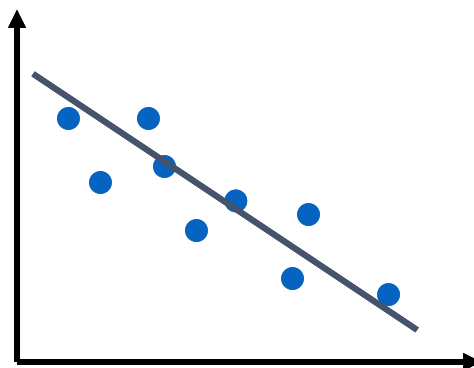
完全负线性相关



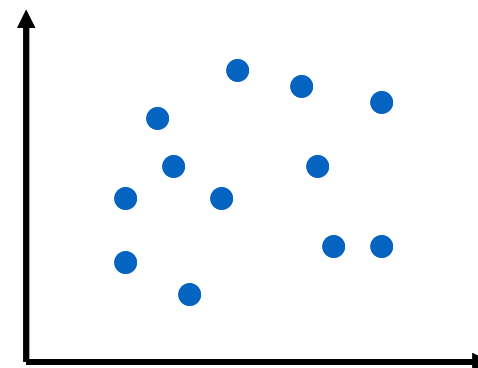
非线性相关



正线性相关



负线性相关



不相关

回归关系与相关关系的区别

1. 相关关系中，变量 x 变量 y 处于平等的地位；回归关系中，变量 y 称为因变量，处在被解释的地位， x 称为自变量，用于预测因变量的变化
2. 相关分析中所涉及的变量 x 和 y 都是随机变量；回归分析中，因变量 y 是随机变量，自变量 x 可以是随机变量，也可以是非随机的确定变量
3. 相关分析主要是描述两个变量之间线性关系的密切程度；回归分析不仅可以揭示变量 x 对变量 y 的影响大小，还可以由回归方程进行预测和控制

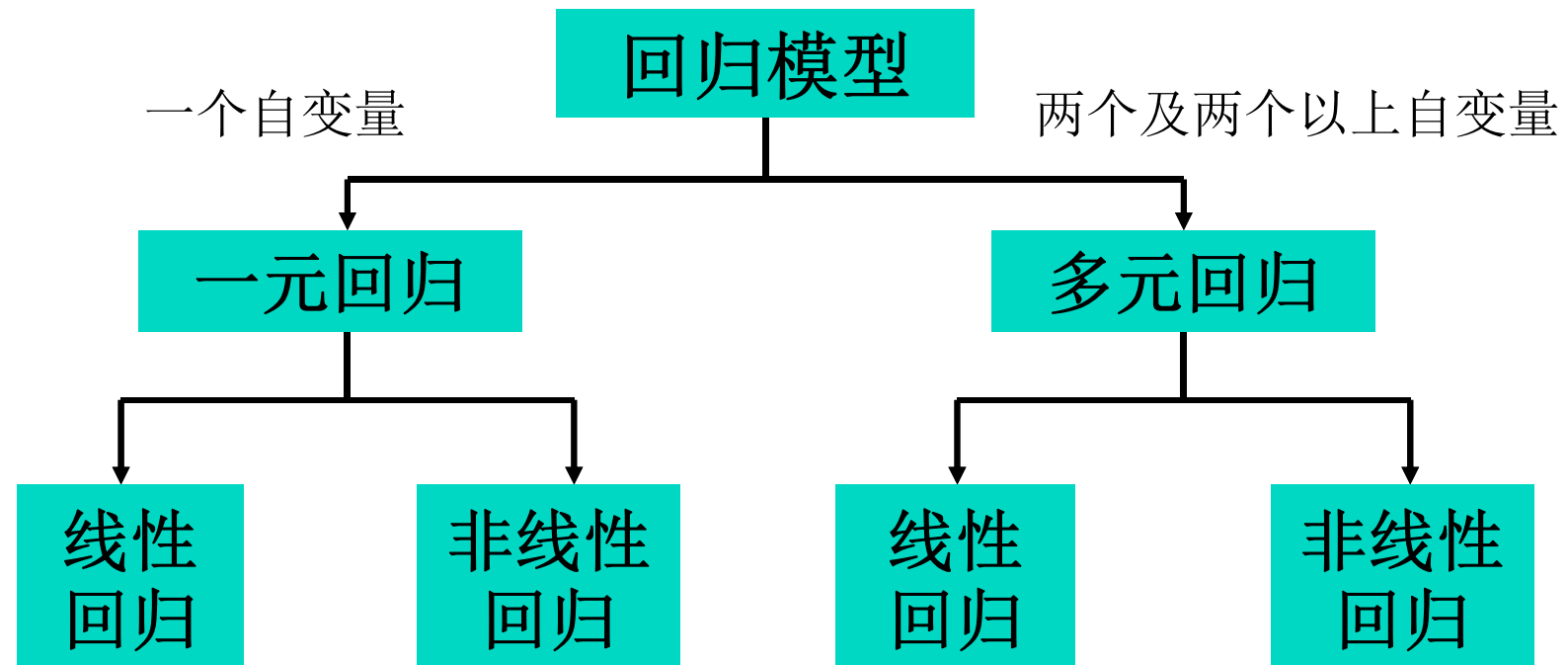
什么是回归分析？

1. 从一组样本数据出发，确定变量之间的数学关系式
2. 对这些关系式的可信程度进行各种统计检验，并从影响某一特定变量的诸多变量中找出哪些变量的影响显著，哪些不显著
3. 利用所求的关系式，根据一个或几个变量的取值来预测或控制另一个特定变量的取值，并给出这种预测或控制的精确程度

回归模型

1. 回答“变量之间是什么样的关系？”
2. 方程中运用
 - 1 个数字的因变量(响应变量)
被预测的变量
 - 1 个或多个数字的或分类的自变量 (解释变量)
用于预测的变量
3. 主要用于预测和估计

回归模型的类型

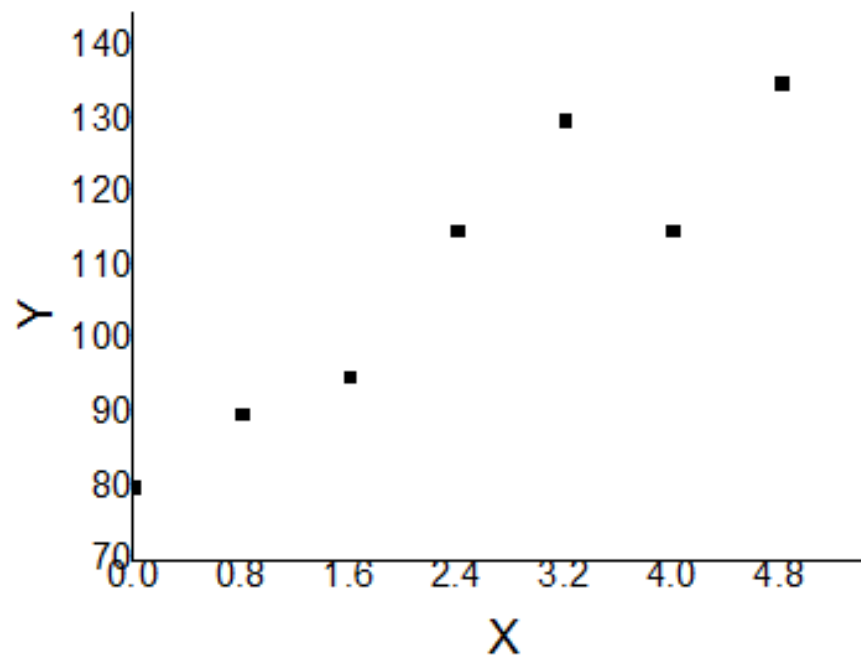


§ 8.2 一元线性回归

数据收集与整理

- 例：土壤内NaCl含量对植物的生长有很大的影响，NaCl含量过高，将增加组织内无机盐的积累，抑制植物的生长。表中的数据，是每1000g土壤中所含NaCl的不同克数(X)对植物单位叶面积干物重的影响(Y)。

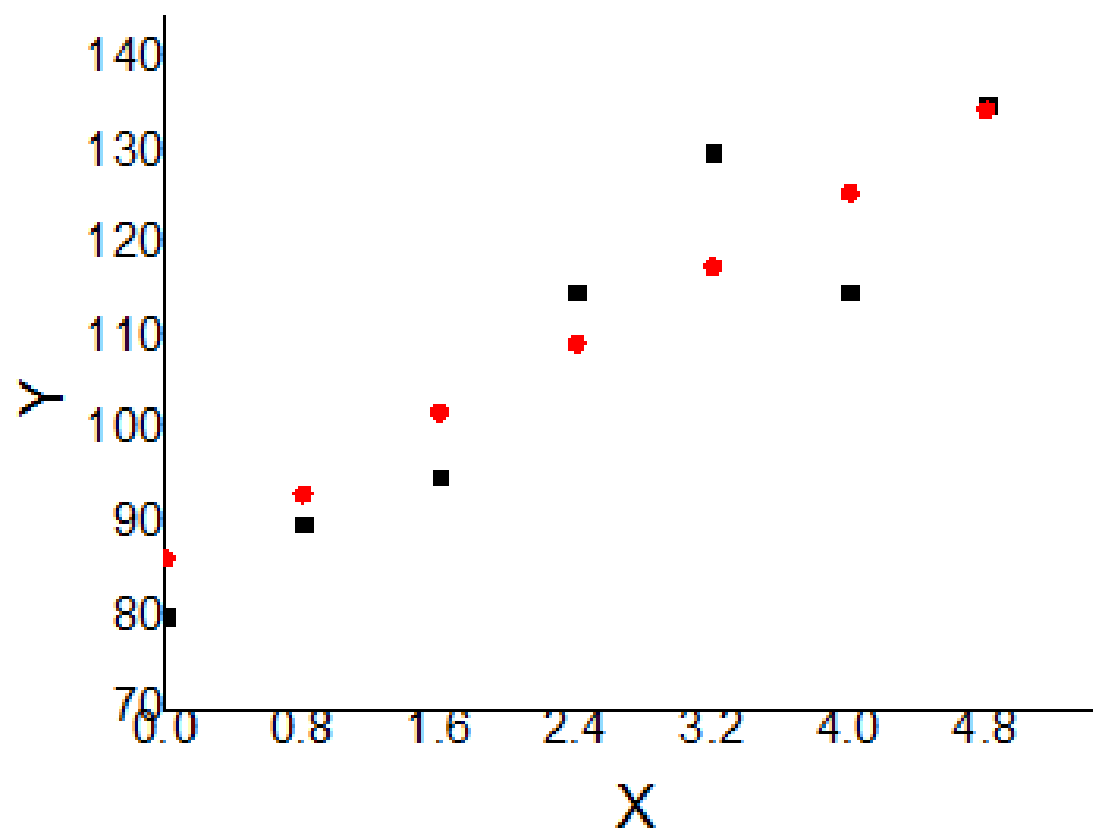
| NaCl含量X /(g·kg ⁻¹) | 0 | 0.8 | 1.6 | 2.4 | 3.2 | 4.0 | 4.8 |
|-----------------------------------|----|-----|-----|-----|-----|-----|-----|
| 干物重Y /(mg·dm ⁻²) | 80 | 90 | 95 | 115 | 130 | 115 | 135 |



判断:

- 1) 两变量之间的关系是否密切;
- 2) 两变量之间的关系是呈一条直线还是某种曲线;
- 3) 是否存在某个点偏离过大;
- 4) 是否存在其他规律。

| | | NaCL含量X /(g·kg ⁻¹) | | | | | | |
|------------------------------------------|----|--------------------------------|------|-------|-------|-------|-------|-------|
| | | 0 | 0.8 | 1.6 | 2.4 | 3.2 | 4.0 | 4.8 |
| 干物重Y /(mg·dm ⁻²) 重复观测值 | 1 | 80 | 90 | 95 | 115 | 130 | 115 | 135 |
| | 2 | 100 | 85 | 89 | 94 | 106 | 125 | 137 |
| | 3 | 75 | 107 | 115 | 103 | 103 | 128 | 128 |
| | 4 | 89 | 93 | 92 | 110 | 110 | 143 | 127 |
| | 5 | 91 | 103 | 115 | 113 | 128 | 132 | 155 |
| | 6 | 79 | 92 | 120 | 108 | 131 | 121 | 132 |
| | 7 | 101 | 78 | 95 | 121 | 117 | 129 | 148 |
| | 8 | 85 | 105 | 95 | 110 | 121 | 112 | 117 |
| | 9 | 83 | 93 | 105 | 108 | 114 | 120 | 134 |
| | 10 | 79 | 85 | 98 | 111 | 116 | 130 | 132 |
| 均值 | | 86.2 | 93.1 | 101.9 | 109.3 | 117.6 | 125.5 | 134.5 |



一元线性回归模型

1. 当只涉及一个自变量时称为一元回归，若因变量 y 与自变量 x 之间为线性关系时称为一元线性回归
2. 对于具有线性关系的两个变量，可以用一条线性方程来表示它们之间的关系
3. 描述因变量 y 如何依赖于自变量 x 和误差项 ε 的方程称为回归模型

一元线性回归模型

➡ 对于只涉及一个自变量的简单线性回归模型可表示为

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- 模型中， y 是 x 的线性函数(部分)加上误差项
- 线性部分反映了由于 x 的变化而引起的 y 的变化
- 误差项 ε 是随机变量

反映了除 x 和 y 之间的线性关系之外的随机因素对 y 的影响
是不能由 x 和 y 之间的线性关系所解释的变异性

- β_0 和 β_1 称为模型的参数

一元线性回归模型

(基本假定)

1. 误差项 ε 是一个期望值为0的随机变量，即 $E(\varepsilon)=0$ 。
对于一个给定的 x 值（ x 的误差可忽略）， y 的期望值为 $E(y)=\beta_0+\beta_1x$
2. 对于所有的 x 值， ε 的方差 σ^2 都相同
3. 误差项 ε 是一个服从正态分布的随机变量，且相互独立。即 $\varepsilon \sim N(0, \sigma^2)$
 - 独立性意味着对于一个特定的 x 值，它所对应的 ε 与其他 x 值所对应的 ε 不相关
 - 对于一个特定的 x 值，它所对应的 y 值与其他 x 所对应的 y 值也不相关

回归方程

1. 描述 y 的平均值或期望值如何依赖于 x 的方程称为回归方程
2. 简单线性回归方程的形式如下

$$E(y) = \beta_0 + \beta_1 x$$

- 方程的图示是一条直线，因此也称为直线回归方程
- β_0 是回归直线在 y 轴上的截距，是当 $x=0$ 时 y 的期望值
- β_1 是直线的斜率，称为回归系数，表示当 x 每变动一个单位时， y 的平均变动值

估计(经验)的回归方程

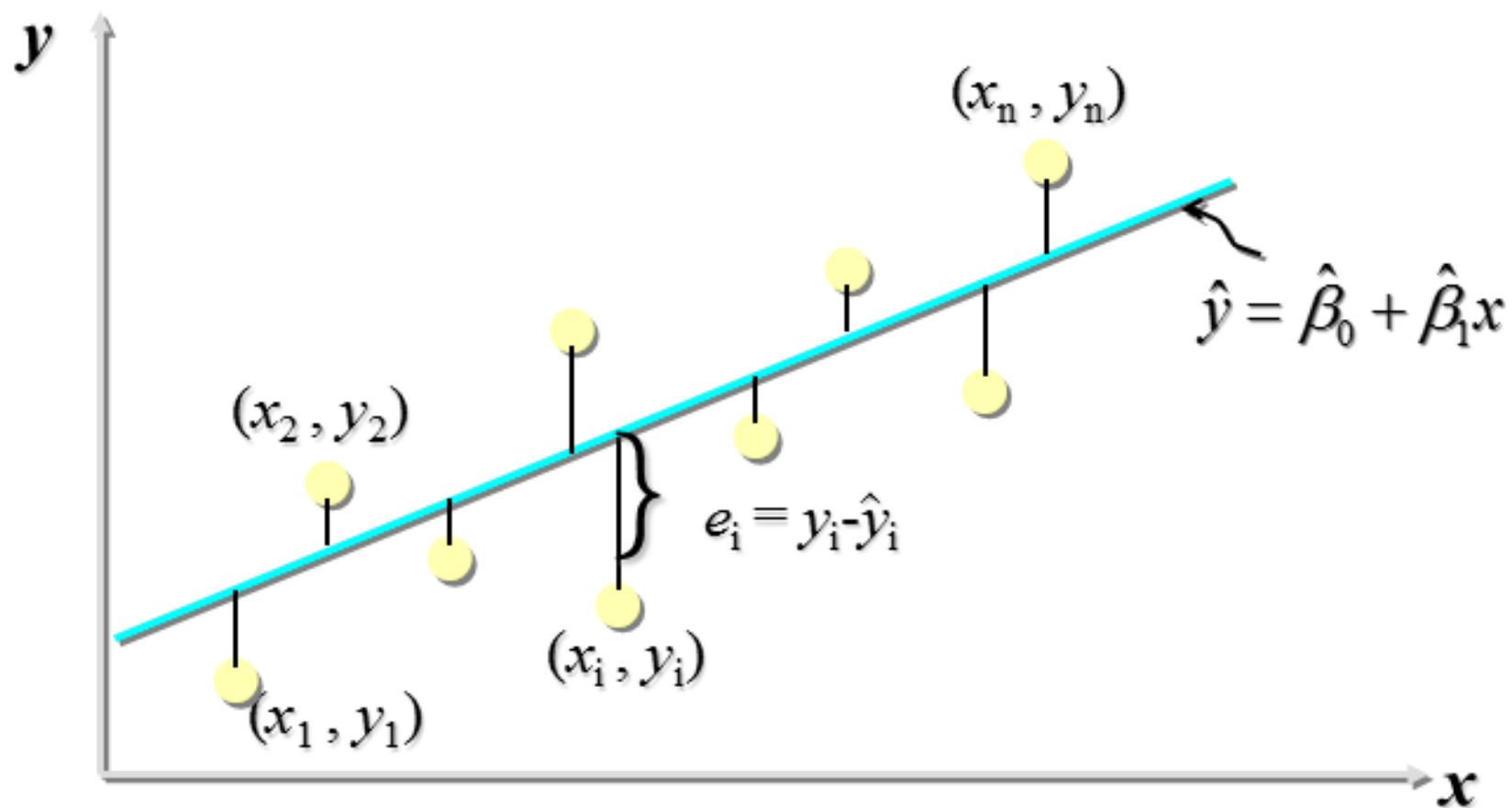
1. 总体回归参数 β_0 和 β_1 是未知的，必需利用样本数据去估计
2. 用样本统计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 代替回归方程中的未知参数 β_0 和 β_1 ，就得到了估计的回归方程
3. 简单线性回归中估计的回归方程为

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

其中： $\hat{\beta}_0$ 是估计的回归直线在 y 轴上的截距， $\hat{\beta}_1$ 是直线的斜率，它表示对于一个给定的 x 的值，是 y 的估计值，也表示 x 每变动一个单位时， y 的平均变动值

参数 β_0 和 β_1 的
最小二乘估计

最小二乘法（图示）



最小二乘法

1. 使因变量的观察值与估计值之间的离差平方和达到最小来求得 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的方法。即

$$Q(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n (y_i - \hat{y})^2 = \sum_{i=1}^n e_i^2 = \text{最小}$$

2. 用最小二乘法拟合的直线来代表 x 与 y 之间的关系与实际数据的误差比其他任何直线都小

$$L = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

$$\begin{cases} \frac{\partial L}{\partial \hat{\beta}_0} = 0 \\ \frac{\partial L}{\partial \hat{\beta}_1} = 0 \end{cases} \quad \Rightarrow \quad \begin{cases} \sum_{i=1}^n (-2)[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)] = 0 \\ \sum_{i=1}^n (-2)x_i[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)] = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \sum_{i=1}^n y_i = \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 x_i) = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i y_i = \sum_{i=1}^n [x_i(\hat{\beta}_0 + \hat{\beta}_1 x_i)] = \hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 \end{cases}$$

最小二乘法

($\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的计算公式)

➔ 根据最小二乘法的要求，可得求解 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的标准方程如下

$$\begin{cases} \hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i - \frac{1}{n} \left(\sum_{i=1}^n X_i \right) \left(\sum_{i=1}^n Y_i \right)}{\sum_{i=1}^n X_i^2 - \frac{1}{n} \left(\sum_{i=1}^n X_i \right)^2} = \frac{\sum_{i=1}^n (X_i - \bar{X}) (Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \end{cases}$$

记:

$$S_{XY} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad \text{校正交叉乘积和}$$

$$S_{XX} = \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{X的校正平方和}$$

$$S_{YY} = \sum_{i=1}^n (y_i - \bar{y})^2 \quad \text{Y的校正平方和}$$

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$$

估计方程的求法

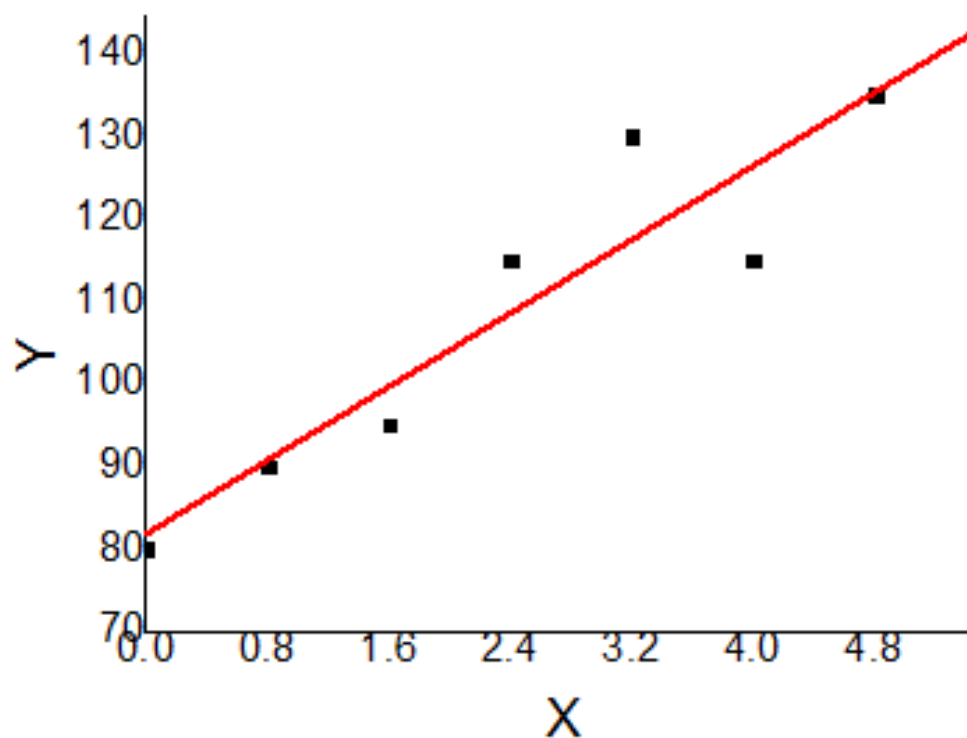
| NaCl含量X /(g·kg ⁻¹) | 0 | 0.8 | 1.6 | 2.4 | 3.2 | 4.0 | 4.8 |
|-----------------------------------|----|-----|-----|-----|-----|-----|-----|
| 干物重Y /(mg·dm ⁻²) | 80 | 90 | 95 | 115 | 130 | 115 | 135 |

$$\begin{cases} \hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \end{cases} \quad \Rightarrow \quad \begin{cases} \hat{\beta}_1 = 11.16 \\ \hat{\beta}_0 = 81.79 \end{cases}$$

估计方程

干物重对NaCl含量的回归方程为

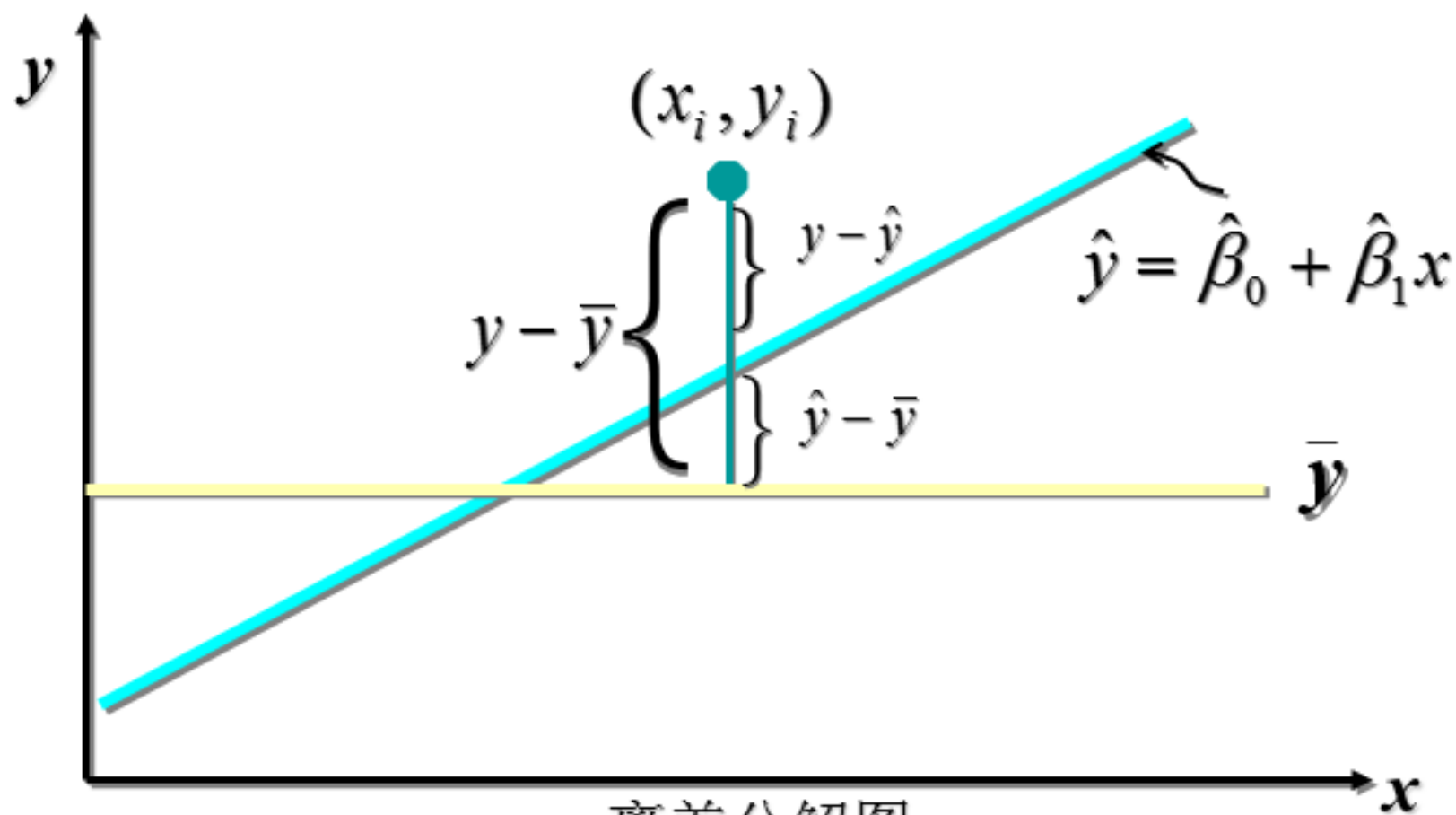
$$\hat{y} = 81.79 + 11.16x$$



§ 8.3 回归方程的显著性检验

一元回归的方差分析

离差分解



离差分解图

离差平方和的分解

1. 从图上看有


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

2. 两端平方后求和有

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


总变差平方和
(*SST*)


回归平方和
(*SSR*)


残差平方和
(*SSE*)

$$SST = SSR + SSE$$

(三个平方和的意义)

1. 总平方和(SST)

- 反映因变量的 n 个观察值与其均值的总离差

2. 回归平方和(SSR)

- 反映自变量 x 的变化对因变量 y 取值变化的影响，或者说，是由于 x 与 y 之间的线性关系引起的 y 的取值变化，也称为可解释的平方和

3. 残差平方和(SSE)

- 反映除 x 以外的其他因素对 y 取值的影响，也称为不可解释的平方和或剩余平方和

样本决定系数 (判定系数 r^2)

1. 回归平方和占总离差平方和的比例

$$r^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y})^2}{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}$$

2. 反映回归直线的拟合程度
3. 取值范围在 $[0, 1]$ 之间
4. $r^2 \rightarrow 1$, 说明回归方程拟合的越好; $r^2 \rightarrow 0$, 说明回归方程拟合的越差

回归方程的显著性检验

(线性关系的检验)

1. 检验自变量和因变量之间的线性关系是否显著
2. 具体方法是将回归离差平方和(SSR)同剩余离差平方和(SSE)加以比较，应用 F 检验来分析二者之间的差别是否显著
 - 如果是显著的，两个变量之间存在线性关系
 - 如果不显著，两个变量之间不存在线性关系

回归方程的显著性检验

(检验的步骤)

1. 提出假设

- $H_0: \beta_1=0$ (线性关系不显著)

2. 计算检验统计量 F

$$F = \frac{SSR/1}{SSE/n-2} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 / 1}{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / n-2} \sim F(1, n-2)$$

3. 确定显著性水平 α , 并根据分子自由度1和分母自由度 $n-2$ 找出临界值 F_α

4. 作出决策: 若 $F \geq F_\alpha$, 拒绝 H_0 ; 若 $F < F_\alpha$, 接受 H_0

回归方程的显著性检验

(方差分析表)

| NaCl含量X /(g·kg ⁻¹) | 0 | 0.8 | 1.6 | 2.4 | 3.2 | 4.0 | 4.8 |
|-----------------------------------|----|-----|-----|-----|-----|-----|-----|
| 干物重Y /(mg·dm ⁻²) | 80 | 90 | 95 | 115 | 130 | 115 | 135 |

| 方差来源 | 平方和SS | 自由度df | 均方MS | F 值 |
|------|---------|-------|-------|-------|
| 回归 | 2232 | 1 | 2232 | 31.55 |
| 残差 | 353.71 | 5 | 70.74 | |
| 总和 | 2585.71 | 6 | | |

因为 $F > F_{0.01}(1,5) = 16.26$ ，所以拒绝原假设。
变量具有极显著的线性回归关系。

估计标准误差 S_y

1. 实际观察值与回归估计值离差平方和的均方根
2. 反映实际观察值在回归直线周围的分散状况
3. 从另一个角度说明了回归直线的拟合程度
4. 计算公式为

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

注：上例的计算结果为8.4107

基于参数的检验

回归系数的显著性检验

1. 检验 x 与 y 之间是否具有线性关系，或者说，检验自变量 x 对因变量 y 的影响是否显著
2. 理论基础是回归系数 $\hat{\beta}_1$ 的抽样分布
3. 在一元线性回归中，等价于回归方程的显著性检验

回归系数的显著性检验

(样本统计量 $\hat{\beta}_1$ 的分布)

1. $\hat{\beta}_1$ 的分布具有如下性质

- 分布形式：正态分布
- 数学期望： $E(\hat{\beta}_1) = \beta_1$

- 标准差：
$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{\sigma}{\sqrt{S_{XX}}}$$

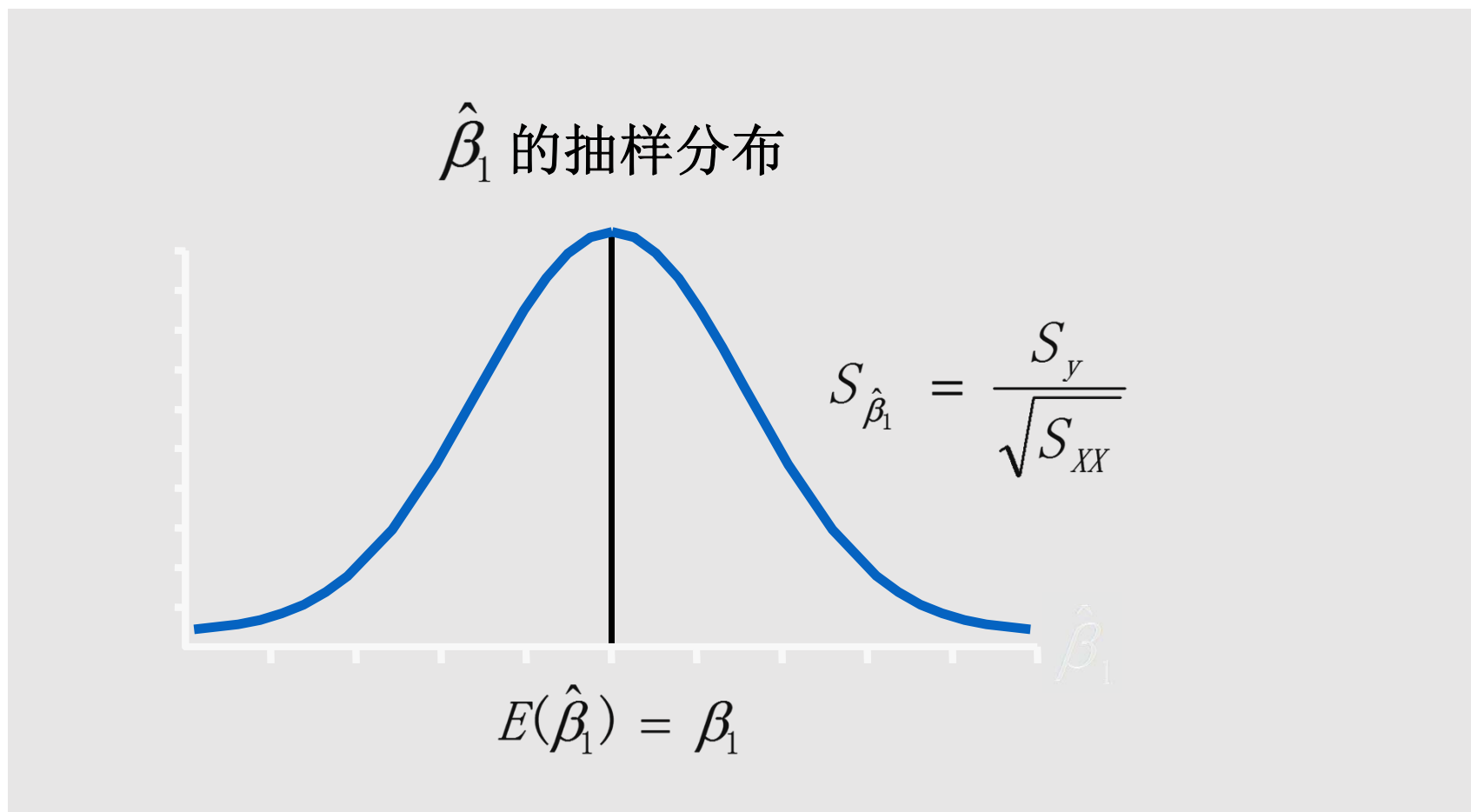
$$E(MSE) = \sigma^2$$

- 由于 σ 未知，需用其估计量 S_y 来代替，得到 $\hat{\beta}_1$ 的估计的标准差

$$S_{\hat{\beta}_1} = \frac{S_y}{\sqrt{S_{XX}}} = \sqrt{\frac{MSE}{S_{XX}}}$$

回归系数的显著性检验

(样本统计量 $\hat{\beta}_1$ 的分布)



回归系数的显著性检验

(样本统计量 $\hat{\beta}_0$ 的分布)

1. $\hat{\beta}_0$ 的分布具有如下性质

- 分布形式：正态分布

- 数学期望： $E(\hat{\beta}_0) = \beta_0$

- 标准差：

$$\sigma_{\hat{\beta}_0} = \sigma \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}}$$

用估计量 S_y 来代替 σ

$$S_{\hat{\beta}_0} = \sqrt{MSE \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right)}$$

回归系数的显著性检验

1. 提出假设

- $H_0: \beta_1 = 0$ (没有线性关系)
- $H_1: \beta_1 \neq 0$ (有线性关系)

2. 计算检验的统计量

$$t = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\sqrt{MSE / S_{XX}}} \sim t(n - 2)$$

3. 确定显著性水平 α ，并进行决策

- $|t| > t_{\alpha/2}$, 拒绝 H_0 ; $|t| < t_{\alpha/2}$, 接受 H_0

回归系数的显著性检验

👉 对前例的回归系数进行显著性检验($\alpha=0.01$)

1. 提出假设

- $H_0: \beta_1 = 0$ 干物重与NaCl含量之间无线性关系
- $H_1: \beta_1 \neq 0$ 干物重与NaCl含量之间有线性关系

2. 计算检验的统计量

$$t = \frac{11.16}{\sqrt{70.74/17.92}} = 5.61$$

3. $t=5.61 > t_{\alpha/2}(5)=4.032$, 拒绝 H_0 , 表明干物重关于NaCl含量的回归极显著。

回归系数的显著性检验

👉 检验 β_1 是不是某一给定值($\alpha=0.05$)

1. 提出假设

- $H_0: \beta_1 = 7$
- $H_1: \beta_1 \neq 7$

2. 计算检验的统计量

$$t = \frac{\hat{\beta}_1 - 7}{\sqrt{MSE / S_{XX}}} = \frac{11.16 - 7}{1.99} = 2.09$$

3. $t=2.05 < t_{\alpha/2}(5)=2.571$, 接受 H_0 。结论是 $\hat{\beta}_1$ 有可能来自 $\beta_1 = 7$ 的总体。

回归系数的显著性检验

1. 提出假设

- $H_0: \beta_0 = a$
- $H_1: \beta_0 \neq a$

2. 计算检验的统计量

$$t = \frac{\hat{\beta}_0 - \beta_0}{S_{\hat{\beta}_0}} = \frac{\hat{\beta}_0 - a}{\sqrt{MSE \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right)}} \sim t(n-2)$$

3. 确定显著性水平 α ，并进行决策

- $|t| > t_{\alpha/2}$, 拒绝 H_0 ; $|t| < t_{\alpha/2}$, 接受 H_0

对 β_0 、 β_1 的估计

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MSE / S_{XX}}} \sim t(n-2)$$

$$t = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{MSE \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} \right)}} \sim t(n-2)$$

β_1 和 β_0 的 $1 - \alpha$ 置信区间分别为：

$$\hat{\beta}_1 \pm t_{\alpha/2} \sqrt{\frac{MSE}{S_{XX}}} \quad \text{和} \quad \hat{\beta}_0 \pm t_{\alpha/2} \sqrt{MSE \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} \right)}$$

两个回归方程的比较

- 例：在优质育种工作中，为了快速筛选优良原始材料，采用染料结合（**DBC**）法测定种子中的碱性氨基酸含量。它的原理是：一种染料orange G与碱性氨基酸结合，使原来染料浓度降低。再通过测定染料减少的量，来估计碱性氨基酸的含量。已经计算出碱性氨基酸含量与**DBC**法测得结果之间有显著回归。实验测定了大麦和黑麦每千克（**kg**）试样的染料结合力（**DBC**）与碱性氨基酸含量，结果如下：

| | | | | | | | | | |
|----|---|----|----|----|----|----|-----|-----|-----|
| 大麦 | X | 91 | 93 | 94 | 96 | 98 | 102 | 105 | 108 |
| | Y | 66 | 68 | 69 | 71 | 73 | 78 | 82 | 85 |
| 黑麦 | X | 80 | 82 | 85 | 87 | 89 | 91 | 95 | |
| | Y | 55 | 57 | 60 | 62 | 64 | 67 | 71 | |

X 表示每kg试样中DBC的mmol数；
Y 表示每kg试样中碱性氨基酸的mmol数。

| | 大麦 | 黑麦 |
|-----------|----------------------------|----------------------------|
| n | 8 | 7 |
| \bar{x} | 98.4 | 87.0 |
| \bar{y} | 74 | 62.3 |
| S_{XX} | 257.9 | 162.0 |
| S_{YY} | 336.0 | 187.4 |
| S_{XY} | 294.0 | 174.0 |
| MSE | 0.140 | 0.244 |
| \hat{Y} | $\hat{Y} = -38.16 + 1.14X$ | $\hat{Y} = -31.16 + 1.07X$ |

检验两回归线有无显著差异。

解：（1）检验 MSE_1 和 MSE_2 有无显著差异：

$$H_0: \sigma_1^2 = \sigma_2^2 \quad H_1: \sigma_1^2 \neq \sigma_2^2$$

$$F = \frac{MSE_2}{MSE_1} = \frac{0.244}{0.140} = 1.74$$

$$F < F_{0.025}(5,6) = 5.99$$

所以可以认为两者具有相同的总体方差。

$$MSE = \frac{(n_1 - 2)MSE_1 + (n_2 - 2)MSE_2}{(n_1 - 2) + (n_2 - 2)} = 0.187$$

(2) 检验回归系数 β_1 和 β'_1 有无显著差异:

$$H_0: \beta_1 - \beta'_1 = 0 \quad H_1: \beta_1 - \beta'_1 \neq 0$$

$$\begin{aligned} t &= \frac{\beta_1 - \beta'_1}{\sqrt{MSE(\frac{1}{S_{XX}} + \frac{1}{S'_{XX}})}} = \frac{1.14 - 1.07}{\sqrt{0.187(\frac{1}{257.9} + \frac{1}{162})}} \\ &= 1.61 \end{aligned}$$

$$t < F_{0.05}(11) = 2.201$$

所以可以认为两者具有共同的总体回归系数。

$$\tilde{\beta}_1 = \frac{S_{XX}\beta_1 + S'_{XX}\beta'_1}{S_{XX} + S'_{XX}} = 1.11$$

(3) 检验回归系数 β_0 和 β'_0 有无显著差异:

(i) 若差异显著, 两回归线差异显著。

(ii) 若差异不显著, 两回归方程可合并。

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} \quad \bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2}$$

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1\bar{x}$$

§ 8.4 预测及应用

利用回归方程进行估计和预测

1. 根据自变量 x 的取值估计或预测因变量 y 的取值
2. 估计或预测的类型
 - 点估计
 - y 的平均值的点估计
 - y 的个别值的点估计
 - 区间估计
 - y 的平均值的置信区间估计
 - y 的个别值的预测区间估计

利用回归方程进行估计和预测 (点估计)

y 的平均值的点估计

利用估计的回归方程，对于自变量 x 的一个给定值 x_0 ，求出因变量 y 的平均值的一个估计值 $\mu_{Y \cdot X=x_0}$ ，就是平均值的点估计

$$\begin{aligned} E(\hat{y}_0) &= E(\hat{\beta}_0 + \hat{\beta}_1 x_0) \\ &= E(\hat{\beta}_0) + E(\hat{\beta}_1) x_0 \\ &= \beta_0 + \beta_1 x_0 \\ &= \mu_{Y \cdot X=x_0} \end{aligned}$$

$$\begin{aligned}
\text{var}(\hat{y}_0) &= \text{var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) \\
&= \text{var}((\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x_0) \\
&= \text{var}(\bar{y}) + (x_0 - \bar{x})^2 \text{var}(\hat{\beta}_1) \\
&= \frac{\sigma^2}{n} + (x_0 - \bar{x})^2 \frac{\sigma^2}{S_{XX}} \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right)
\end{aligned}$$

$$\hat{y}_0 \sim N \left(\mu_{Y \cdot X=x_0}, \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right) \right)$$

利用回归方程进行估计和预测 (点估计)

y 的个别值的点估计

利用估计的回归方程，对于自变量 x 的一个给定值 x_0 ，求出因变量 y 的一个个别值的估计值，就是个别值的点估计

$$\begin{aligned} E(y_0 - \hat{y}_0) &= E(\beta_0 + \beta_1 x_0 + \varepsilon - \hat{\beta}_0 - \hat{\beta}_1 x_0) \\ &= \beta_0 + \beta_1 x_0 + E(\varepsilon) - E(\hat{\beta}_0) - E(\hat{\beta}_1) x_0 \\ &= \beta_0 + \beta_1 x_0 - E(\hat{\beta}_0) - E(\hat{\beta}_1) x_0 \\ &= 0 \end{aligned}$$

$$\begin{aligned}
\text{var}(y_0 - \hat{y}_0) &= \text{var}(y_0) + \text{var}(\hat{y}_0) \\
&= \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right) \\
&= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right)
\end{aligned}$$

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

利用回归方程进行估计和预测

（区间估计）

1. 点估计不能给出估计的精度，点估计值与实际值之间是有误差的，因此需要进行区间估计
2. 对于自变量 x 的一个给定值 x_0 ，根据回归方程得到因变量 y 的一个估计区间
3. 区间估计有两种类型
 - 置信区间估计
 - 预测区间估计

利用回归方程进行估计和预测 (置信区间估计)

y 的平均值的置信区间估计

1. 利用估计的回归方程，对于自变量 x 的一个给定值 x_0 ，求出因变量 y 的平均值 $\mu_{Y \cdot X=x_0}$ 的估计区间，这一估计区间称为**置信区间**
2. $\mu_{Y \cdot X=x_0}$ 在 $1-\alpha$ 置信水平下的置信区间为

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

利用回归方程进行估计和预测

【例】根据前例，求出NaCl含量取不同值时，对应的平均干物重的0.95置信区间。

解：根据前面的计算结果

$$\bar{x}=2.4, S_{XX}=17.92, \text{MSE}=70.74, t_{\alpha/2}(5)=2.571$$

| NaCl含量X /(g·kg ⁻¹) | 0 | 0.8 | 1.6 | 2.4 | 3.2 | 4.0 | 4.8 |
|-----------------------------------|-------|-------|-------|--------|--------|--------|--------|
| 干物重Y /(mg·dm ⁻²) | 80 | 90 | 95 | 115 | 130 | 115 | 135 |
| \hat{y} | 81.79 | 90.72 | 99.65 | 108.57 | 117.50 | 126.43 | 135.36 |
| 置信区间 (±) | 14.73 | 11.56 | 9.14 | 8.17 | 9.14 | 11.56 | 14.73 |

利用回归方程进行估计和预测 (预测区间估计)

☞ y 的个别值的预测区间估计

1. 利用估计的回归方程，对于自变量 x 的一个给定值 x_0 ，求出因变量 y 的一个个别值的估计区间，这一区间称为**预测区间**
2. y_0 在 $1-\alpha$ 置信水平下的预测区间为

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

注意！# **1**

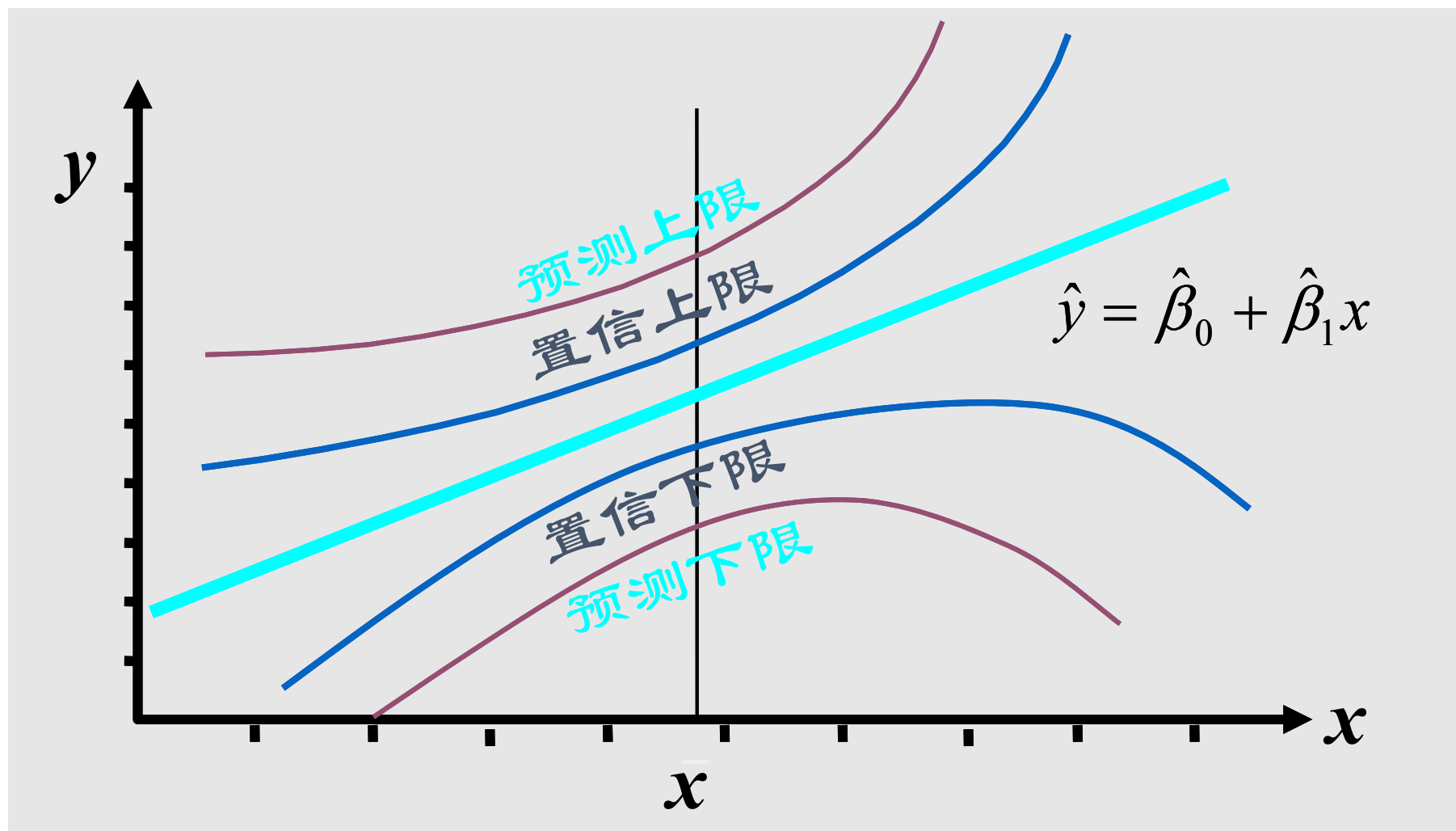
利用回归方程进行估计和预测 (置信预测区间估计:算例)

【例】根据前例，求出NaCl含量取不同值时，对应的干物重的0.95置信区间。

解：

| NaCl含量X (g·kg ⁻¹) | 0 | 0.8 | 1.6 | 2.4 | 3.2 | 4.0 | 4.8 |
|----------------------------------|-------|-------|-------|--------|--------|--------|--------|
| 干物重Y (mg·dm ⁻²) | 80 | 90 | 95 | 115 | 130 | 115 | 135 |
| \hat{y} | 81.79 | 90.72 | 99.65 | 108.57 | 117.50 | 126.43 | 135.36 |
| 预测区间 (±) | 26.17 | 24.52 | 23.48 | 23.12 | 23.48 | 24.52 | 26.17 |

置信区间、预测区间、回归方程



影响区间宽度的因素

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

1. 置信水平 ($1 - \alpha$)
区间宽度随置信水平的增大而增大
2. 数据的离散程度 (s)
区间宽度随离散程度的增大而增大
3. 样本容量
区间宽度随样本容量的增大而减小
4. 用于预测的 x_0 与 \bar{x} 的差异程度
区间宽度随 x_0 与 \bar{x} 的差异程度的增大而增大

回归分析的应用

- 描述两个变量的依存关系
- 在一定范围内对因变量进行预测
- 通过控制自变量来对因变量进行控制

- 注意问题:

- 1、回归变量的确定

- 2、回归方程应进行检验

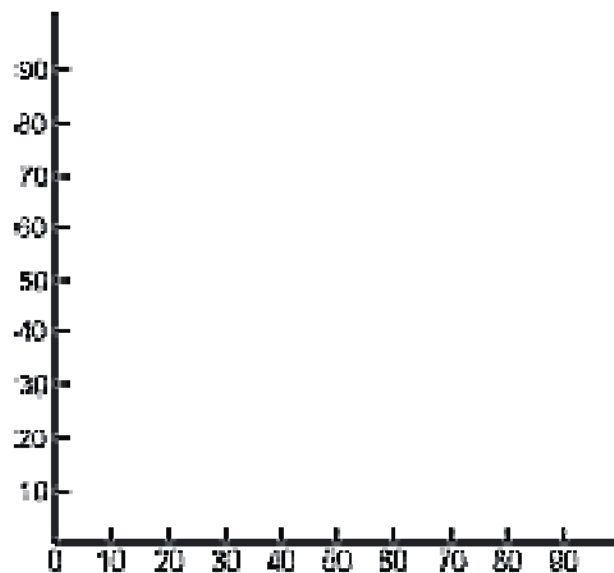
- 3、预测和外推要谨慎

§ 8.5 相关系数及其计算

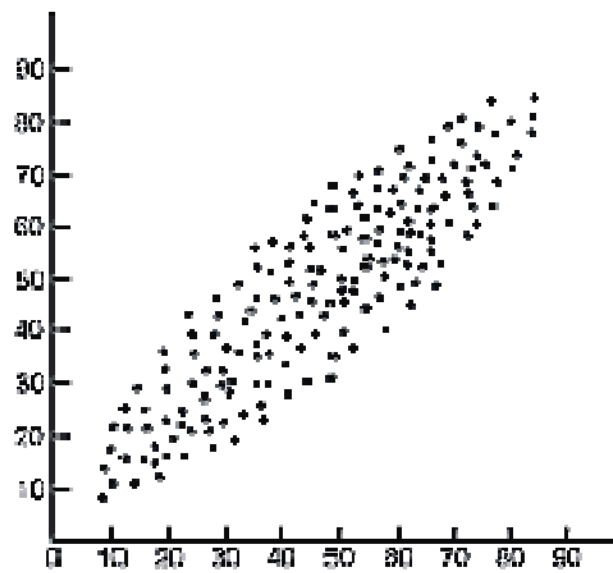
相关关系的测度

(相关系数)

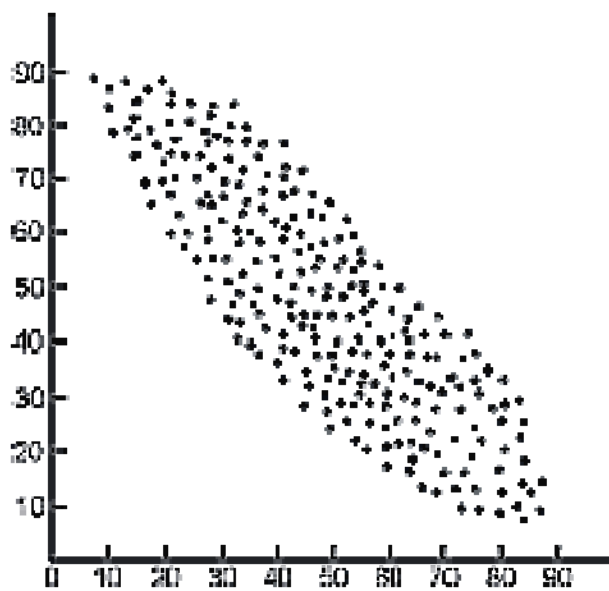
1. 对变量之间关系密切程度的度量
2. 对两个变量之间线性相关程度的度量称为简单相关系数
3. 若相关系数是根据总体全部数据计算的，称为总体相关系数，记为 ρ
4. 若是根据样本数据计算的，则称为样本相关系数，记为 r



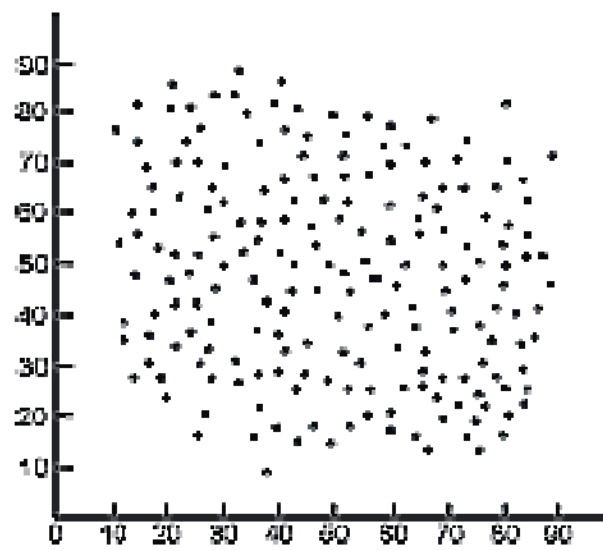
图A 散点图的建立



图B 正相关性



图C 负相关性



图D 不相关性

相关关系的测度

（相关系数：皮尔逊相关系数）

➡ 样本相关系数的计算公式

$$r = \frac{1}{N} \sum \left[\left(\frac{x - \bar{x}}{\sigma_x} \right) \left(\frac{y - \bar{y}}{\sigma_y} \right) \right]$$
$$= \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \cdot \sum (y - \bar{y})^2}}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) (y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{XY}}{S_{XX}}$$

$$\begin{aligned} SSE &= \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\ &= \sum ((y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}))^2 = SST - \hat{\beta}_1 S_{XY} \end{aligned}$$

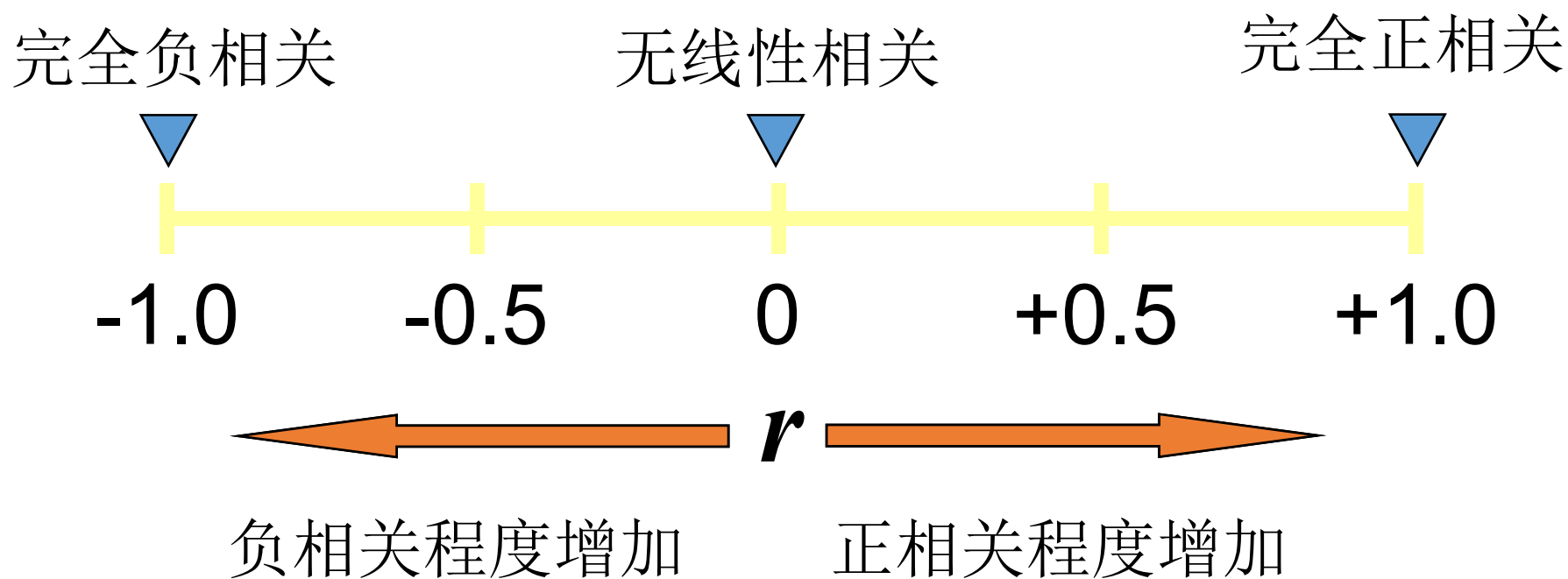
$$SSR = SST - SSE = \hat{\beta}_1 S_{XY} = \frac{S_{XY}^2}{S_{XX}}$$

$$r^2 = \left(\frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \cdot \sum (y - \bar{y})^2}} \right)^2 = \frac{S_{XY}^2}{S_{XX} SST} = \frac{SSR}{SST}$$

相关关系的测度

1. r 的取值范围是 $[-1,1]$
2. $|r|=1$ ，为完全相关
 - $r=1$ ，为完全正相关
 - $r=-1$ ，为完全负正相关
3. $r=0$ ，不存在线性相关关系
4. $-1 \leq r < 0$ ，为负相关
5. $0 < r \leq 1$ ，为正相关
6. $|r|$ 越趋于1表示关系越密切； $|r|$ 越趋于0表示关系越不密切

相关关系的测度



相关关系的测度

【例】在研究我国人均消费水平的问题中，把全国人均消费额记为 y ，把人均国民收入记为 x 。我们收集到1981～1993年的样本数据 (x_i, y_i) ， $i = 1, 2, \dots, 13$ ，数据见表1，计算相关系数。

| 表1 我国人均国民收入与人均消费金额数据 | | | | | |
|----------------------|------------|------------|------|------------|------------|
| 单位:元 | | | | | |
| 年份 | 人均 国民收入 | 人均 消费金额 | 年份 | 人均 国民收入 | 人均 消费金额 |
| 1981 | 393.8 | 249 | 1988 | 1068.8 | 643 |
| 1982 | 419.14 | 267 | 1989 | 1169.2 | 690 |
| 1983 | 460.86 | 289 | 1990 | 1250.7 | 713 |
| 1984 | 544.11 | 329 | 1991 | 1429.5 | 803 |
| 1985 | 668.29 | 406 | 1992 | 1725.9 | 947 |
| 1986 | 737.73 | 451 | 1993 | 2099.5 | 1148 |
| 1987 | 859.97 | 513 | | | |

解：根据样本相关系数的计算公式有

$$\begin{aligned} r &= \frac{n \sum xy - \sum x \sum y}{\sqrt{n \sum x^2 - (\sum x)^2} \cdot \sqrt{n \sum y^2 - (\sum y)^2}} \\ &= \frac{13 \times 9156173.99 - 12827.5 \times 7457}{\sqrt{13 \times 16073323.77 - (12827.5)^2} \cdot \sqrt{13 \times 5226399 - (7457)^2}} \\ &= 0.9987 \end{aligned}$$

人均国民收入与人均消费金额之间的相关系数为 **0.9987**，表明人均收入与人均消费之间呈正相关，即人均收入越高，人均消费越多。

$r^2=0.9974$ 表明y的变异有99.74%可用y与x之间的线性关系来解释。

相关系数的显著性检验

1. 检验两个变量之间是否存在线性相关关系
2. 等价于对回归系数 β_1 的检验
3. 采用 t 检验
4. 检验的步骤为

- 提出假设: $H_0: \rho = 0$; $H_1: \rho \neq 0$

- 计算检验的统计量:
$$t = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} \sim t(n - 2)$$

- 确定显著性水平 α , 并作出决策

- 若 $|t| > t_{\alpha/2}$, 拒绝 H_0
- 若 $|t| < t_{\alpha/2}$, 接受 H_0

➡ 对前例计算的相关系数进行显著性检验($\alpha=0.05$)

1. 提出假设: $H_0: \rho = 0$; $H_1: \rho \neq 0$
2. 计算检验的统计量

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.9987\sqrt{13-2}}{\sqrt{1-0.9987^2}} = 64.9809$$

3. 根据显著性水平 $\alpha=0.05$, 由于
 $|t|=64.9809 > t_{\alpha/2}(13-2)=2.201$, 拒绝 H_0 , 人均消费
金额与人均国民收入之间的相关关系显著

(相关系数检验表的使用)

- 注意：

- 1、两个变量都应服从正态分布。

- 2、相关系数应进行检验。

- 3、观测值尽可能多。

- 4、正确理解相关系数的含义。

相关分析的Spss过程

- 菜单式操作
- Correlate:
 - Bivariate功能项、
 - Partial功能项、
 - Distance功能项

§ 8.6 可直线化的一元非线性回归

确定曲线类型

$$R^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

决定系数也称作相关指数。反映了回归曲线拟合度的高低。

如果不确定曲线类型，最好多试几种，分别计算并比较 R^2 。

1、倒数函数

函数形式

变换

$$\hat{y} = \frac{a + bx}{x}$$

$$y' = xy$$

$$\hat{y} = \frac{1}{a + bx}$$

$$y' = 1/y$$

$$\hat{y} = \frac{x}{a + bx}$$

$$y' = x/y$$

$$\hat{y}' = a + bx$$

注意：不能有使分母为0的观测值

2、对数变换

- 例：细菌生长数量（Y）与时间（X）

$$Y = ae^{bX} \quad \text{指数函数}$$

将等式两边取对数：

$$\ln Y = \ln a + bX$$

令 $Y' = \ln Y$, $a' = \ln a$, 则变换为：

$$Y' = a' + bX$$

$$Y = ab^X \xrightarrow{Y' = \ln Y, a' = \ln a, b' = \ln b} Y' = a' + b'X$$

- 幂函数 $Y = dX^b$

将等式两边取对数: $\ln Y = \ln d + b \cdot \ln X$

令 $Y' = \ln Y$, $a' = \ln d$, $X' = \ln X$, 则有:

$$Y' = a' + bX'$$

- 对数函数 $Y = a + b \cdot \lg X$

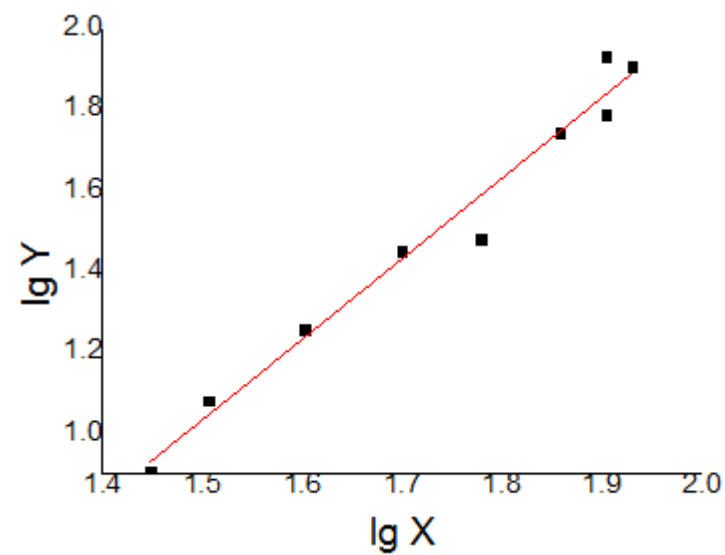
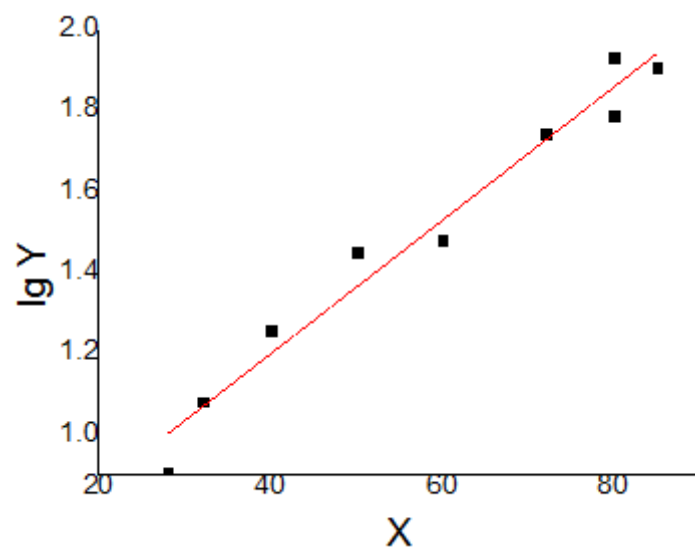
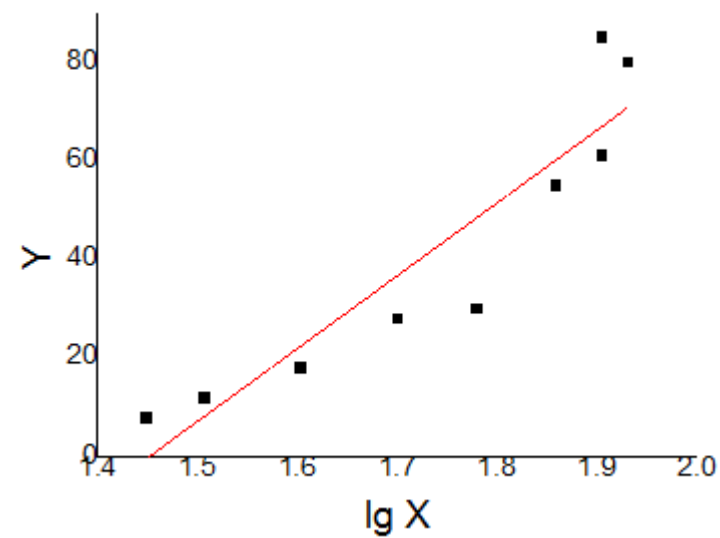
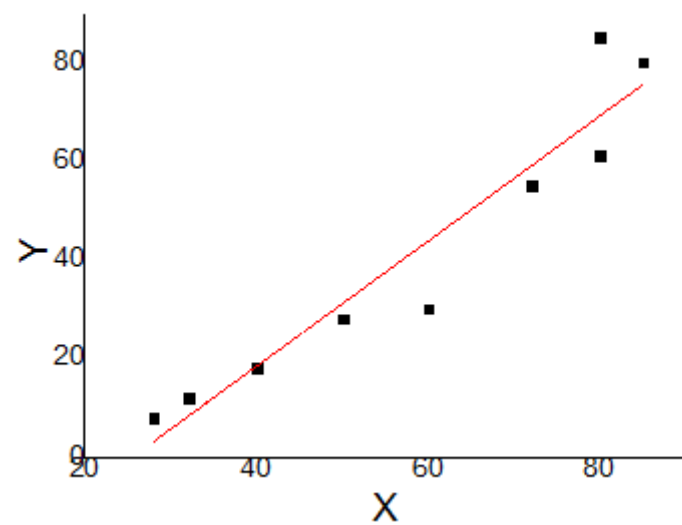
令 $X' = \lg X$, 有:

$$Y = a + bX'$$

- 例：在突变实验中，用不同剂量的射线照射植物的种子，发现苗期高度与成活株之间有一定的关系。用 X 线照射大麦的种子，记处理株第一叶平均高度占对照株高度的百分数为 X，存活百分数为 Y，得到结果：

| | | | | | | | | | |
|---|----|----|----|----|----|----|----|----|----|
| X | 28 | 32 | 40 | 50 | 60 | 72 | 80 | 80 | 85 |
| Y | 8 | 12 | 18 | 28 | 30 | 55 | 61 | 85 | 80 |

(1) 绘制散点图，判断曲线类型：



(2) 进行对数变换，得到回归方程：

令 $Y' = \lg Y$, $X' = \lg X$ ，进行线性回归分析：

| | | | | | | | | | |
|------|------|------|------|------|------|------|------|------|------|
| X' | 1.45 | 1.51 | 1.60 | 1.70 | 1.78 | 1.86 | 1.90 | 1.90 | 1.93 |
| Y' | 0.90 | 1.08 | 1.26 | 1.45 | 1.48 | 1.74 | 1.79 | 1.93 | 1.90 |

回归方程 $\hat{Y}' = -1.9582 + 1.9932X'$

$$\lg \hat{Y} = -1.9582 + 1.9932 \cdot \lg X$$

$$\hat{Y} = 0.011X^{1.9932}$$

(3) 直线化回归方程的假设检验:

$$F = \frac{SSR/1}{SSE/n-2} = \frac{\sum_{i=1}^n (\hat{y}'_i - \bar{y}')^2 / 1}{\sum_{i=1}^n (y'_i - \hat{y}'_i)^2 / 7}$$

| | | | | | | | | | |
|------------|------|------|------|------|------|------|------|------|------|
| X' | 1.45 | 1.51 | 1.60 | 1.70 | 1.78 | 1.86 | 1.90 | 1.90 | 1.93 |
| Y' | 0.90 | 1.08 | 1.26 | 1.45 | 1.48 | 1.74 | 1.79 | 1.93 | 1.90 |
| \hat{Y}' | 0.93 | 1.05 | 1.23 | 1.43 | 1.59 | 1.75 | 1.83 | 1.83 | 1.89 |

$$F = \frac{1.038/1}{0.027/7} = 268.63 > F_{0.01}(1, 7) = 12.25$$

回归关系极显著。

(4) 计算曲线回归方程的决定系数:

$$\hat{Y} = 0.011X^{1.9932}$$

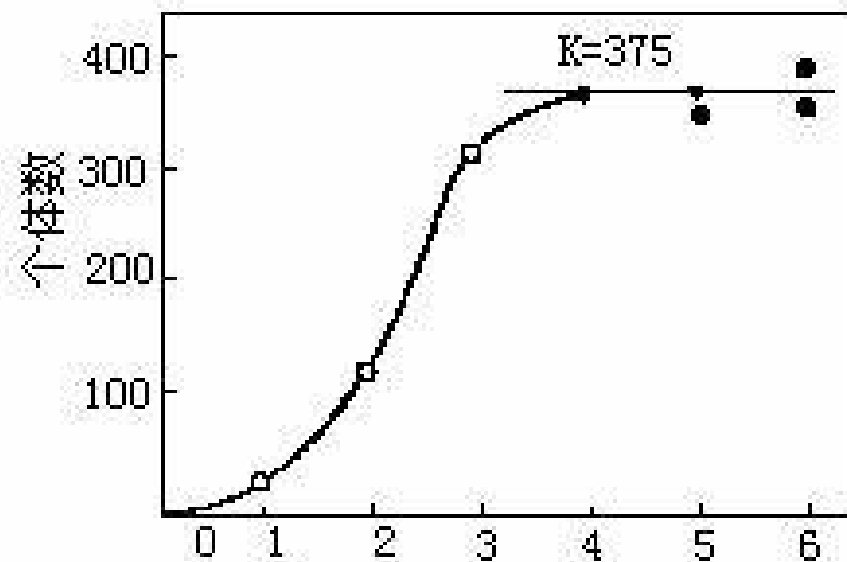
| | | | | | | | | | |
|-----------|------|-------|-------|-------|-------|-------|-------|-------|-------|
| X | 28 | 32 | 40 | 50 | 60 | 72 | 80 | 80 | 85 |
| Y | 8 | 12 | 18 | 28 | 30 | 55 | 61 | 85 | 80 |
| \hat{Y} | 8.43 | 11.00 | 17.16 | 26.78 | 38.51 | 55.39 | 68.33 | 68.33 | 77.11 |

$$R^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2} = 1 - \frac{415.90}{6794.89} = 0.9388$$

表明苗期高度与存活百分比的回归关系用该指数函数进行描述，决定系数达0.9388，该函数拟合较好。

3、Logistic生长曲线

$$y = \frac{K}{1 + ae^{-bx}}$$



大草履虫实验种群的增长和用逻辑斯谛方程拟合的结果
(仿Allee等, 1949)

$$y = \frac{K}{1 + ae^{-bx}} \quad \Rightarrow \quad \frac{K - y}{y} = ae^{-bx}$$

两边取对数：

$$\ln \frac{K - y}{y} = \ln a - bx$$

令 $Y' = \ln \frac{K - y}{y}$, $a' = \ln a$, $b' = -b$, 则有：

$$y' = a' + b'x$$

- K 值的确定:

- 1) 若y是累积频率, 可用K=100;
- 2) 若y表示生长量时, 可取3对 x 等间距的观测值 (x_1, y_1) 、 (x_2, y_2) 和 (x_3, y_3) , 代入Logistic方程可得:

$$\frac{y_2(K - y_1)}{y_1(K - y_2)} = \left[\frac{y_3(K - y_2)}{y_2(K - y_3)} \right]^{\frac{x_1 - x_2}{x_2 - x_3}}$$

若 $x_2 = \frac{x_1 + x_3}{2}$, 可得:

$$K = \frac{y_2^2(y_1 + y_3) - 2y_1y_2y_3}{y_2^2 - y_1y_3}$$

- 例：下表示测定某种肉鸡在良好的生长条件下生长过程的数据资料。试配合Logistic生长曲线方程。

| X /周次 | 2 | 4 | 6 | 8 | 10 | 12 | 14 |
|-------|-----|------|------|-----|------|------|-----|
| Y /kg | 0.3 | 0.86 | 1.73 | 2.2 | 2.47 | 2.67 | 2.8 |

解：(1) 求K值

取x为等间距的 $x_1=2$, $x_2=8$, $x_3=14$, 由公式得：

$$K = \frac{2.2^2 \times (0.3 + 2.8) - 2 \times 0.3 \times 2.2 \times 2.8}{2.2^2 - 0.3 \times 2.8} = 2.827$$

(2) 转换变量：令 $y' = \ln \frac{K-y}{y}$

| x / 周次 | 2 | 4 | 6 | 8 | 10 | 12 | 14 |
|--------|-------|-------|--------|--------|--------|--------|--------|
| Y / kg | 0.3 | 0.86 | 1.73 | 2.2 | 2.47 | 2.67 | 2.8 |
| Y' | 2.131 | 0.827 | -0.456 | -1.255 | -1.934 | -2.834 | -4.642 |

(3) 计算 x 和 y' 的相关系数：

$$r = \frac{\sum(x - \bar{x})(y - \bar{y}')}{\sqrt{\sum(x - \bar{x})^2 \cdot \sum(y - \bar{y}')^2}} = -0.9914$$

$$|r| = 0.9914 > r_{0.01}(5) = 0.874$$

说明 x 与 y' 的直线关系是极显著的，用 Logistic 生长曲线拟合原数据是合适的。

(4) 求a和b值，建立Logistic方程：

$$b' = \frac{S_{XY'}}{S_{XX}} = \frac{-58.2363}{112} = -0.520$$

$$a' = \bar{y}' - b'\bar{x} = -1.166 - (-0.520) \times 8 = 2.994$$

$$a = e^{a'} = e^{2.994} = 19.965$$

$$b = -b' = 0.52$$

所以，有：

$$\hat{y} = \frac{2.827}{1 + 19.965e^{-0.52x}}$$

The End