第二十四节：必备模型及其应用（一）

主讲老师　侯梓熙

一元线性回归

本节课程内容

- 一元线性回归
  - 1.变量间的关系
    - 函数关系
    - 相关关系
  - 2.相关关系与回归分析的区别
  - 3.一元线性回归的模型与假定
  - 4.最小二乘估计
  - 5.回归直线的拟合优度
    - 判定系数
    - 估计标准误差

变量间的关系

函数关系

1.是一一对应的确定关系

2.举例：

1)圆的面积(S)与半径(r)之间的关系可表示为 $S=\pi r^2$；

2)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);

相关关系

1.变量间关系不能用函数关系精确表达,一个变量的取值不能由另一个变量唯一确定,当变量 x 取某个值时，变量 y 的取值可能有几个.

2.举例：

1)父亲身高(y)与子女身高(x)之间的关系;

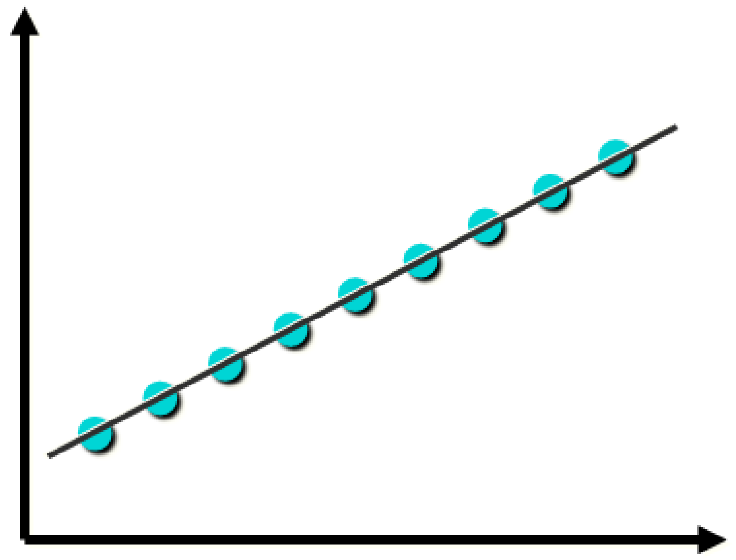
2)收入水平(*y*)与受教育程度(*x*)之间的关系;

3)商品销售额(*y*)与广告费支出(*x*)之间的关系.

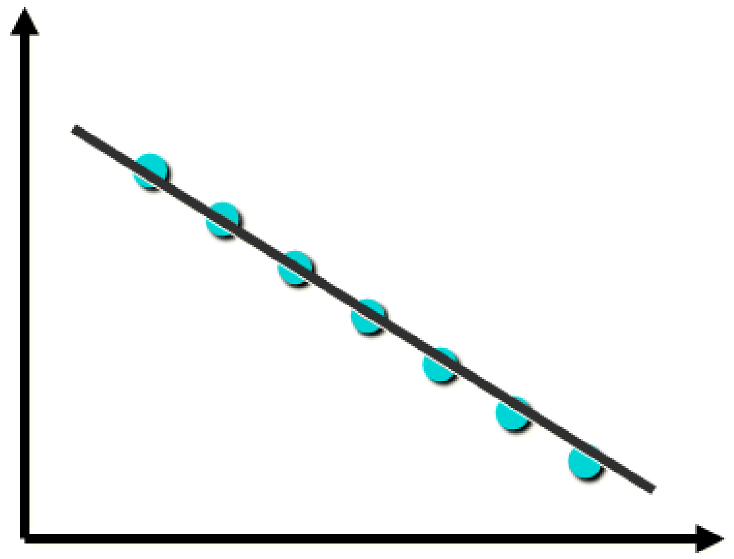
相关关系的类型

- 相关关系
  - 线性相关
    - 正线性相关
    - 负线性相关
  - 非线性相关

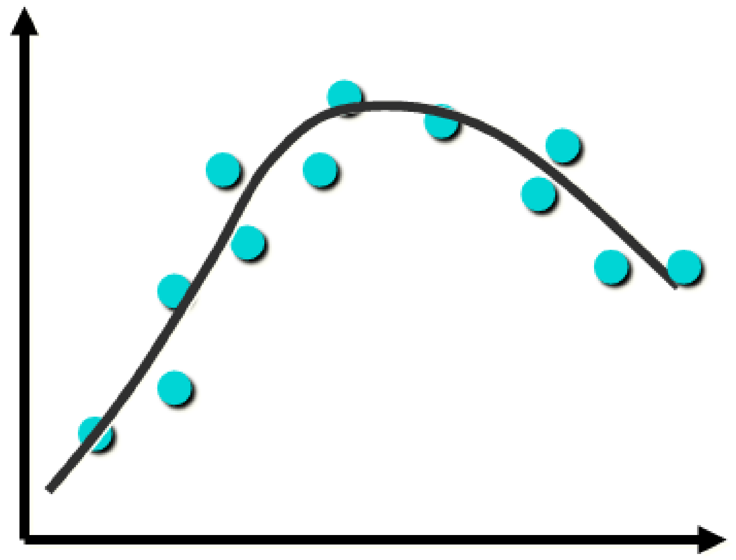
相关关系的类型



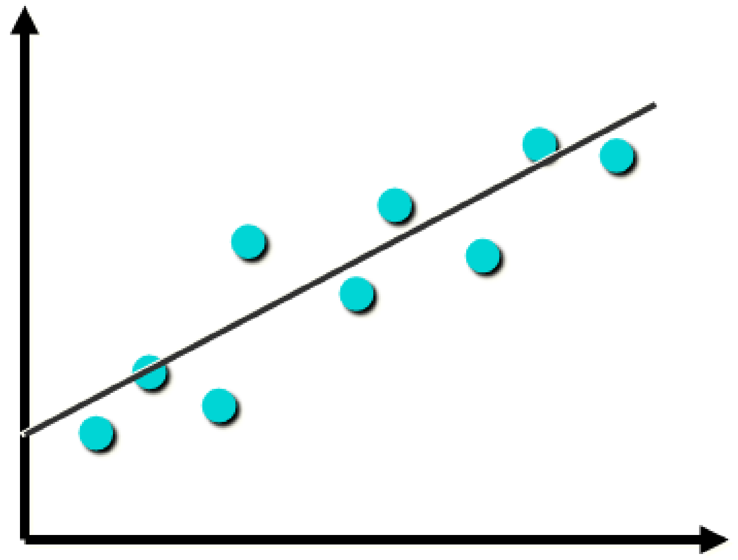
完全正线性相关



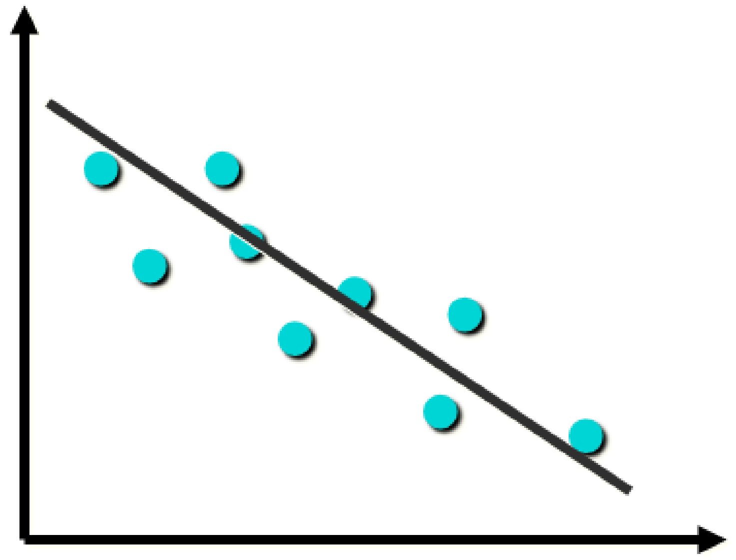
完全负线性相关



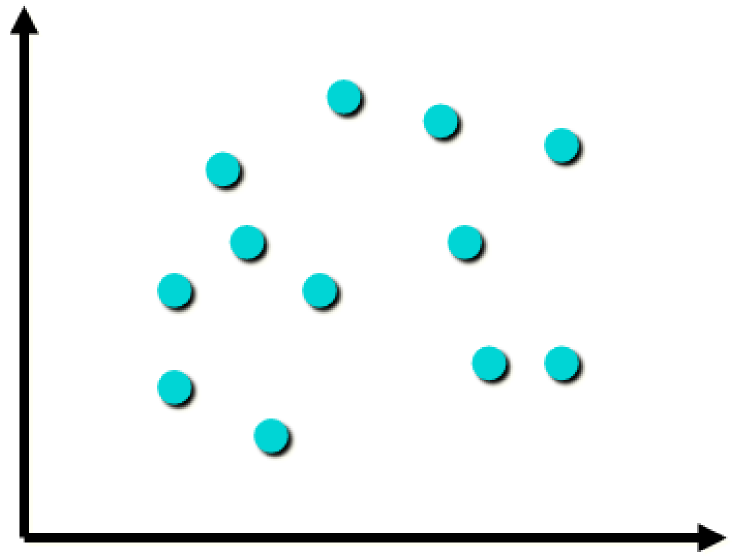
非线性相关



正线性相关



负线性相关



不相关

相关关系测度:

相关系数:
是根据样本数据计算的度量两个变量之间线性关系强度的统计量。

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2} \cdot \sqrt{n\sum y^2 - (\sum y)^2}}$$

按上述公式计算的相关系数也称为线性相关系数，或称为 Pearson 相关系数。

相关系数的性质

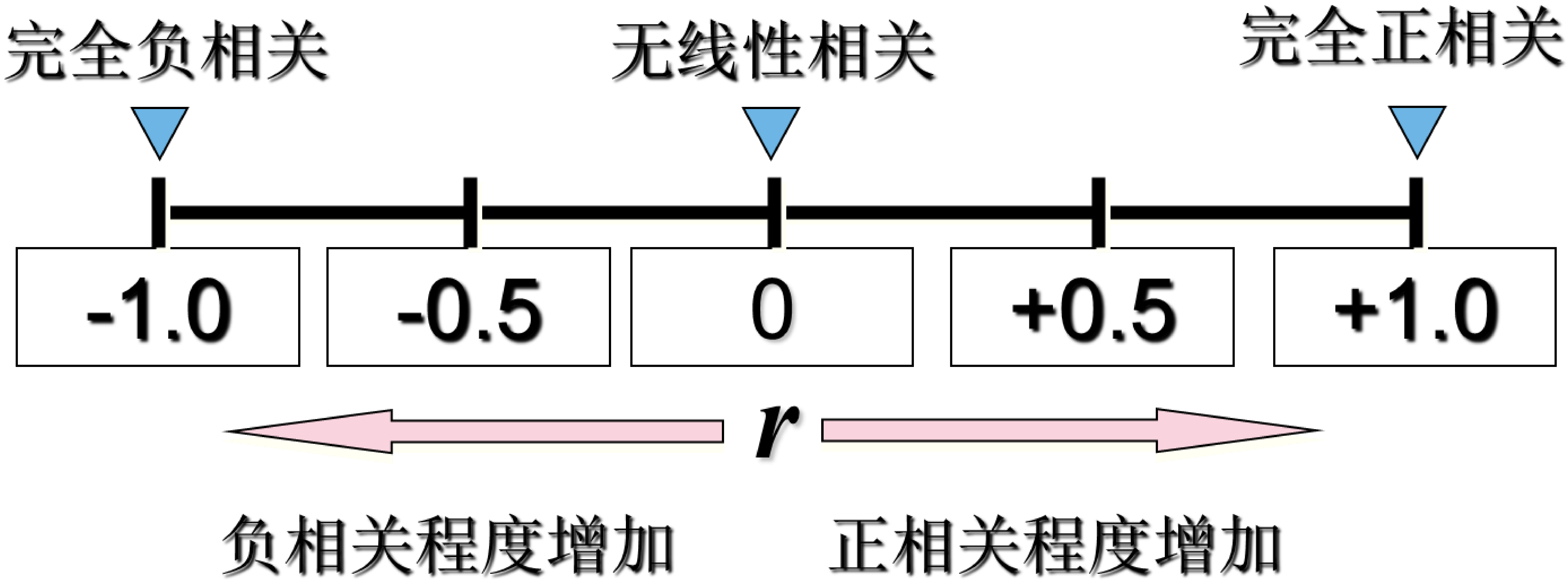
1. r 的取值范围是 [-1,1]
2. $|r|=1$，为完全相关

   $r=1$，为完全正相关

   $r=-1$，为完全负相关
3. $r=0$，不存在线性相关关系相关
4. $-1\leq r<0$，为负相关

   $0<r\leq 1$，为正相关
5. $|r|$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



回归分析是什么?

回归分析是在相关分析的基础上，考察变量之间的数量变化规律，并通过一定的数量表达式描述他们之间的关系，进而确定一个或几个变量的变化对另一特定的变量的影响程度。

回归分析与相关分析的区别

相关关系

目的在于测度变量之间关系的强度.

回归分析

侧重于考察变量之间的数量关系,还可以由回归方程进行预测和控制.

一元线性回归模型

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

β_0和β_1称为模型的参数, β_0+ β_1反映了由于x的变化而引起的y的线性变化; ε称为误差项的随机变量.

注:
1)进行回归分析时，需要确定哪个变量是因变量，哪个是变量是自变量。
2)被预测或被解释的变量称为因变量，用 y 表示。用来预测或解释因变量的一个或多个变量称为自变量，用 x 表示。

一元线性回归模型的假定

1. 因变量 y 与自变量 x 之间具有线性关系。
2. 在重复抽样中，自变量x的取值是固定的，即假设 x 是非随机的。
3. 误差项ε是一个期望值为 0 的随机变量，即$E(\varepsilon) = 0$.
4. 对于所有的 x 值，ε的方差σ^2都相同。
5. 误差项ε是一个服从正态分布的随机变量，且独立，即$\varepsilon \sim N(0, \sigma^2)$

一元线性回归方程

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

（1）一元线性回归方程的图示是一条直线，因此也称为直线回归方程。

（2）β_0是回归直线在 y 轴上的截距，是当x = 0时 y 的期望值。

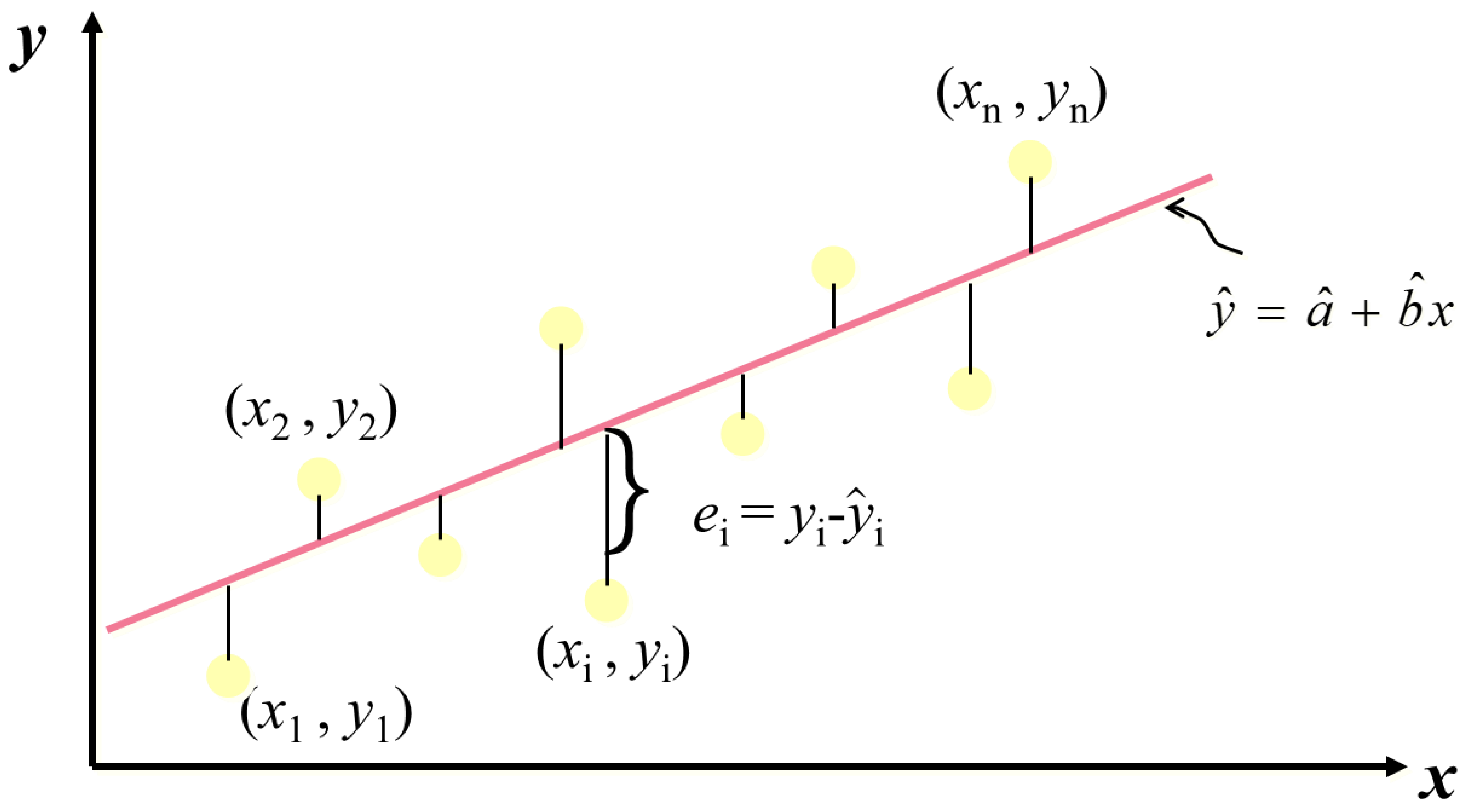
（3）β_1是直线的斜率，表示 x 每变动一个单位时，y 的平均变动。

估计的回归方程

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

总体回归参数β_0和β_1是未知的时候，利用样本数据去估计他们得到的 方程。

参数的最小二乘估计



德国科学家卡尔·高斯提出用最小化图中垂直方向的离差和来估计参数β_0和β_1，根据这一方法确定模型参数β_0和β_1的方法称为最小二乘法。

参数的最小二乘估计

根据最小二乘法，使$\sum(y_i-\hat{y}_i)^2=\sum\left(y_i-\hat{\beta}_0+\hat{\beta}_1x_i\right)^2$最小。

最小二乘法求得的参数β_0和β_1为：

$$\hat{\beta}_1=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}$$

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

回归直线的拟合优度

01 判定系数

02 估计标准误差

离差平方和的分解

1.因变量y的影响因素

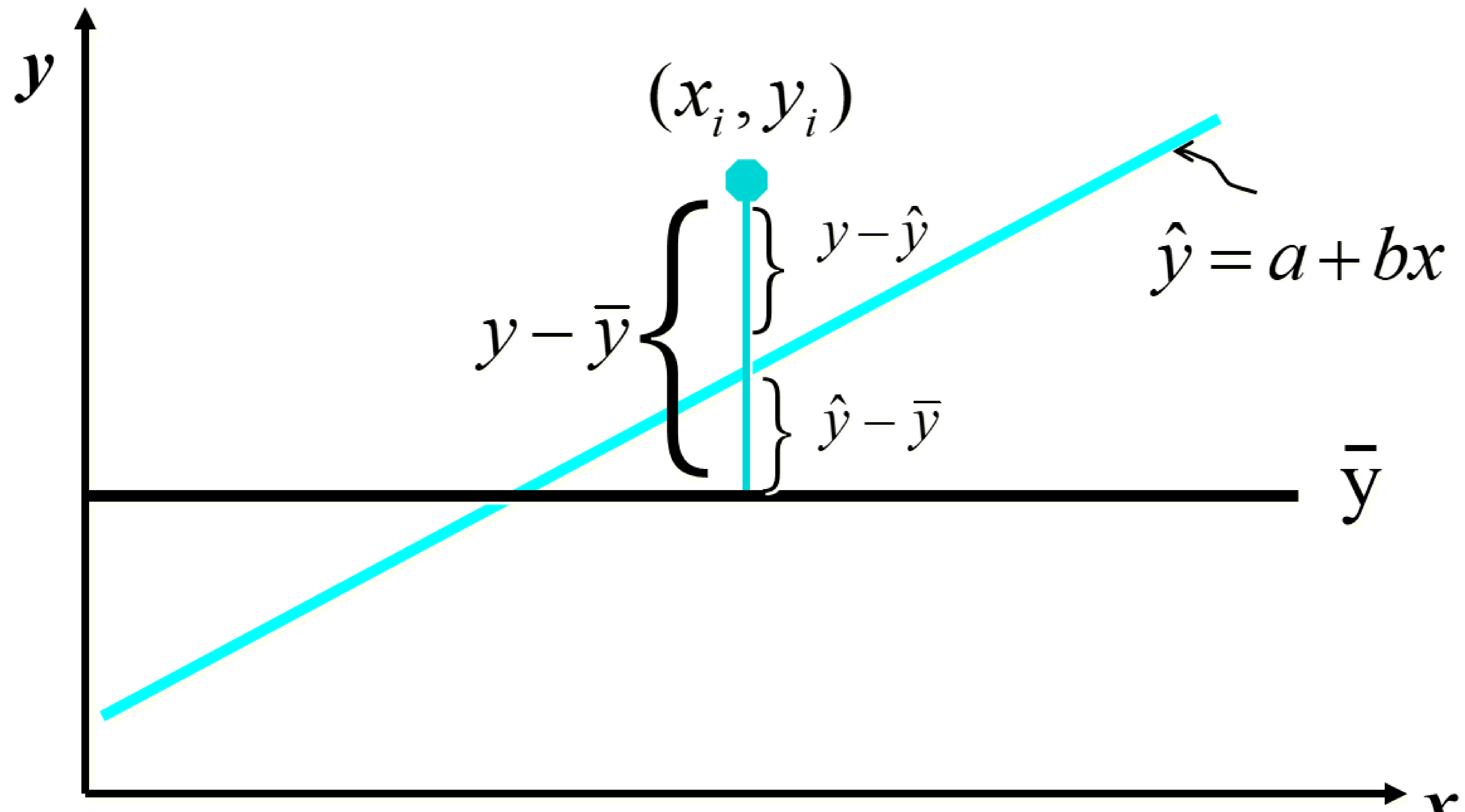
(1)由于自变量 x 的取值不同造成的

(2)除 x 以外的其他因素(如x对y的非线性影响、测量误差等)的影响

2.对于一个具体观察值

变差的大小可以通过该实际观测值与其均值之差来表示,即$(y-\bar{y})$.

离差平方和的分解



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**

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{总变差平方和（}SST\text{）}}=\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\text{回归平方和（}SSR\text{）}}+\underbrace{\sum_{i=1}^{n}(y_i-\hat{y})^2}_{\text{残差平方和（}SSE\text{）}}$$

$SST = SSR + SSE$

离差平方和的分解

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.

判定系数

$$R^2=\frac{SSR}{SST}=\frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\bar{y})^2}$$

(1) R^2反映回归直线的拟合程度;

(2) R^2取值范围在 [0 , 1] 之间;

(3) R^2越接近 1，表明回归平方和占总平方和的比例越大，回归直线与各观测点越接近，用 x 的变化来解释 y 值变差的部分就越多，回归直线的拟合程度就越好。反之R^2越接近 0，回归直线的拟合程度就越差。

估计标准误差

$$S_e = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SSE}{n-2}}$$

（1）标准误差可以看做再排除了 x 对 y 的线性影响后，y 随机波动大小的一个估计量。

（2）反映了估计的回归方程预测因变量 y 时预测误差的大小。

（3）各观测点越靠近直线，估计标准差Se越小，回归直线对各观测点的代表性就越好，根据回估计的回归方程进行预测也就越准确。

本节课程回顾

- 一元线性回归
  - 1.变量间的关系
    - 函数关系
    - 相关关系
  - 2.相关关系与回归分析的区别
  - 3.一元线性回归的模型与假定
  - 4.最小二乘估计
  - 5.回归直线的拟合优度
    - 判定系数
    - 估计标准误差

THANK YOU

感谢观看

主讲老师 侯梓熙