南京理工大学
NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY

# 回归模型

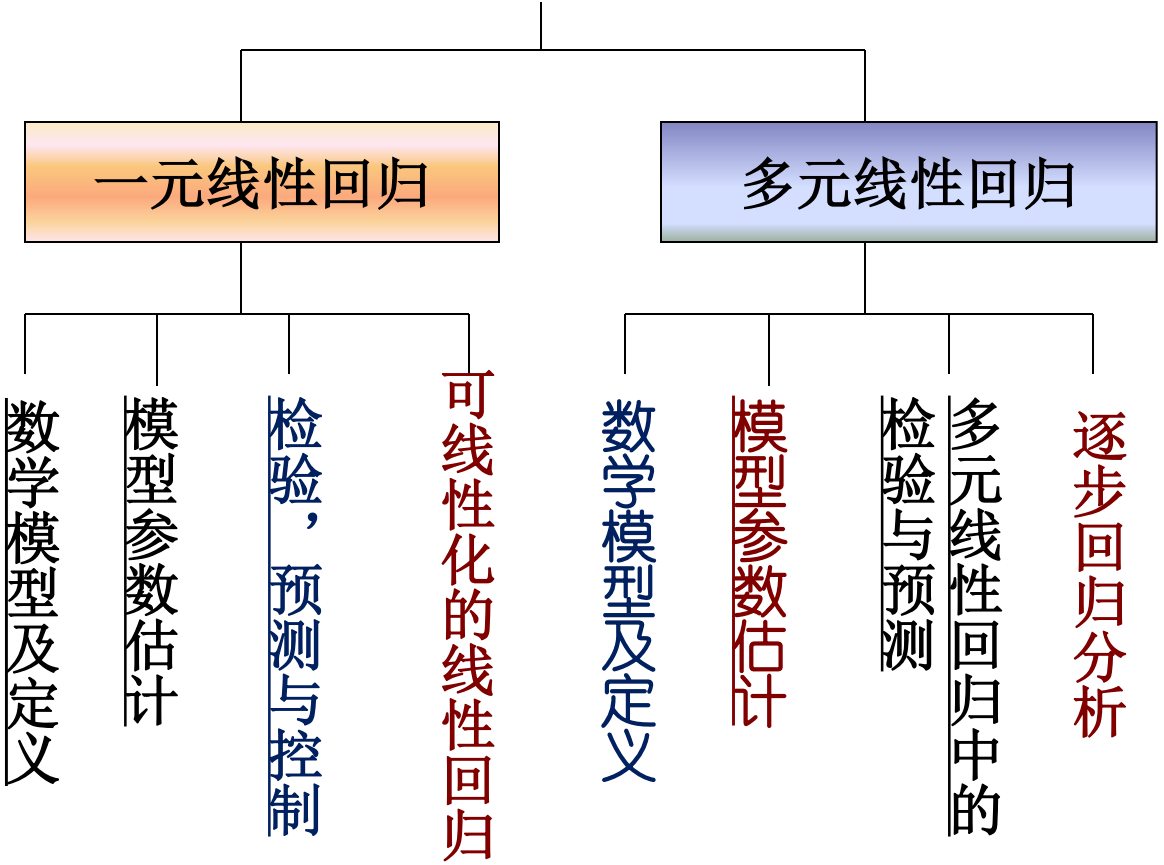
主讲人：谢建春

# 百度百科

**回归分析** 是确定两种或两种以上变数间相互依赖的定量关系的一种**统计**分析方法。

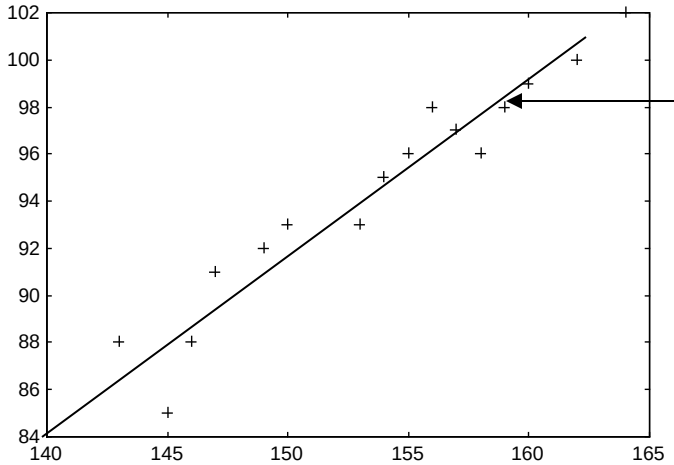
- 按照涉及的**自变量的多少**，可分为一元回归分析和多元回归分析；
- 按照**自变量和因变量之间的关系**类型，可分为线性回归分析和非线性回归分析。

# 回归分析方法


一元线性回归
多元线性回归
数学模型及定义
模型参数估计
检验，预测与控制
可线性化的线性回归
数学模型及定义
模型参数估计
检验与预测
多元线性回归中的
逐步回归分析

**例1** 测16名成年女子的身高与腿长所得数据如下：

| 身高 | 143 | 145 | 146 | 147 | 149 | 150 | 153 | 154 | 155 | 156 | 157 | 158 | 159 | 160 | 162 | 164 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 腿长 | 88 | 85 | 88 | 91 | 92 | 93 | 93 | 95 | 96 | 98 | 97 | 96 | 98 | 99 | 100 | 102 |




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

散点图

一般地，称由 $y=\beta_0+\beta_1 x+\varepsilon$ 确定的模型为一元线性回归模型，记为

$$\begin{cases} y=\beta_0+\beta_1 x+\varepsilon \\ E\varepsilon=0, D\varepsilon=\sigma^2 \end{cases} \qquad (1)$$

固定的未知参数 $\beta_0$ 、 $\beta_1$ 称为回归系数，自变量 x 也称为回归变量.

$Y=\beta_0+\beta_1 x$，称为 y 对 x 的回归直线方程.

## • 模型参数估计

有 n 组独立观测值，$(x_1, y_1)$，$(x_2, y_2)$，…，$(x_n, y_n)$

设 $\begin{cases} y_i = \beta_0 + \beta x_i + \varepsilon_i, i = 1, 2, \ldots, n \\ E\varepsilon_i = 0, D\varepsilon_i = \sigma^2 \quad \text{且} \varepsilon_1 \varepsilon_2, \ldots, \varepsilon_n \text{相互独立} \end{cases}$

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

最小二乘法就是选择 $\beta_0$ 和 $\beta_1$ 的估计 $\hat{\beta}_0$， $\hat{\beta}_1$ 使得

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

将上式分别对$\beta_0$，$\beta_1$求偏导，得

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) \\ \dfrac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) \end{cases}$$

令上两式为0，解得$\beta_0$，$\beta_1$，得

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \dfrac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} \end{cases} \quad \text{或} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^{\,2}, \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

- 可线性化的一元非线性回归--曲线回归

**例：指数曲线方程 $\hat{y}=ae^{bx}$ 的线性化**

$$\hat{y}=ae^{bx}$$

- 
- 两边取对数：$\ln\hat{y}=\ln a+bx$
- 令 $y'=\ln y$ ，可得直线回归方程：$\hat{y}'=\ln a+bx$

# 幂函数曲线方程 $\hat{y}=ax^b$的线性化

$$\hat{y}=ax^b$$

当 $y$ 和 $x$ 都大于0时可线性化为：

$$\ln\hat{y}=\ln a+b\ln x$$

若令$y'=\ln y$，$x'=\ln x$ ，即有线性回归方程：

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

通常选择的六类曲线如下：

（1）双曲线 $\frac{1}{y}=a+\frac{b}{x}$

（2）幂函数曲线 $y=ax^{b}$ ，其中 x>0,a>0

（3）指数曲线 $y=ae^{bx}$ 其中参数 a>0.

（4）倒指数曲线 $y=ae^{b/x}$ 其中 a>0，

（5）对数曲线 y=a+blogx,x>0

（6）S 型曲线 $y=\frac{1}{a+be^{-x}}$

## 2.多元线性回归

### 2. 1数学模型及定义

一般地，影响试验指标的因素不只一个，假设它们之间有如下的线性关系：

$$y=\beta_0+\beta_1 x_1+\ldots+\beta_k x_k+\varepsilon \qquad (1)$$

其中$y$为可观测的随机变量，$x_1, x_2, \ldots, x_k$为非随机的可精确观测的变量，$\beta_0$，$\beta_1$，$\ldots$，$\beta_k$为$k+1$个未知参数，$\varepsilon$为随机变量，设$E(\varepsilon)=0, D(\varepsilon)=\sigma^2>0.$

为了估计未知参数$\beta_0$，$\beta_1$，. . . . , $\beta_k$和$\sigma^2$，我们对$x_1, x_2, ..., x_k$和$y$作$n$次观测得$n$组观测值（$x_{i1}, x_{i2}, ..., x_{ik}, y_i$）（i=1, 2, 3, . . . , n）.

它们满足关系式：

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \varepsilon_i \qquad (2)$$

i=1, 2, 3, . . . n

其中$\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$相互独立且是与$\varepsilon$同分布的随机变量。

$$Y=\begin{bmatrix} y_1 \\ \cdots \\ \cdots \\ y_n \end{bmatrix},\quad X=\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix},\quad \beta=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_k \end{bmatrix},\quad \varepsilon=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{bmatrix}$$

$$Y=X\beta+\varepsilon \qquad (3)$$

$\varepsilon$满足：$E\varepsilon=0,\ COV(\varepsilon,\varepsilon)=\sigma^2 I_n.$

一般称

$$\begin{cases} Y=X\beta+\varepsilon \\ E(\varepsilon)=0, COV(\varepsilon,\varepsilon)=\sigma^2 I_n \end{cases}$$

为 k 元线性回归模型，并简记为$(Y, X\beta, \sigma^2 I_n)$

解得估计值

$$\hat{\beta} = \left(X^T X\right)^{-1}\left(X^T Y\right)$$

得到的 $\hat{\beta}_i$ 代入回归平面方程得：

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$$

称为**经验回归平面方程**. $\hat{\beta}_i$ 称为**经验回归系数**.

注意：$\hat{\beta}$服从$k+1$维正态分布，且为$\beta$的无偏估计，协方差阵为$\sigma^2(X^T X)^{-1}$.

# 2.1 牙膏的销售量

**问题**

建立牙膏销售量与价格、广告投入之间的模型

预测在不同价格和广告费用下的牙膏销售量

收集了30个销售周期本公司牙膏销售量、价格、广告费用，及同期其它厂家同类牙膏的平均售价

| 销售周期 | 本公司价格(元) | 其它厂家价格(元) | 广告费用(百万元) | 价格差（元） | 销售量(百万支) |
|---|---|---|---|---|---|
| 1 | 3.85 | 3.80 | 5.50 | -0.05 | 7.38 |
| 2 | 3.75 | 4.00 | 6.75 | 0.25 | 8.51 |
| … | … | … | … | … | … |
| 29 | 3.80 | 3.85 | 5.80 | 0.05 | 7.93 |
| 30 | 3.70 | 4.25 | 6.80 | 0.55 | 9.26 |

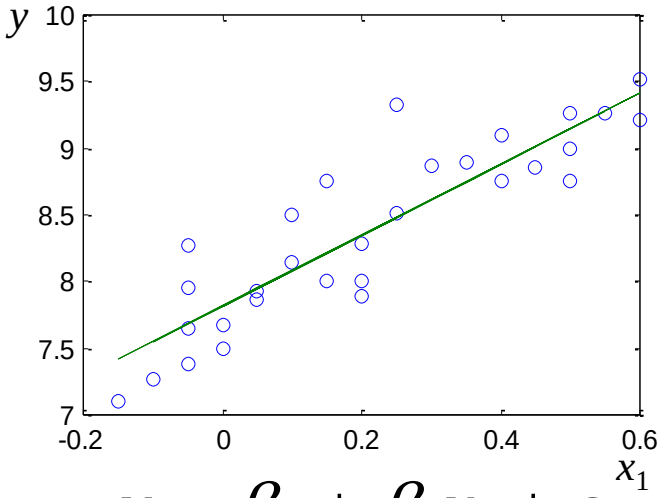
# 基本模型

$y$ ~公司牙膏销售量

$x_1$~其它厂家与本公司价格差

$x_2$~公司广告费用

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



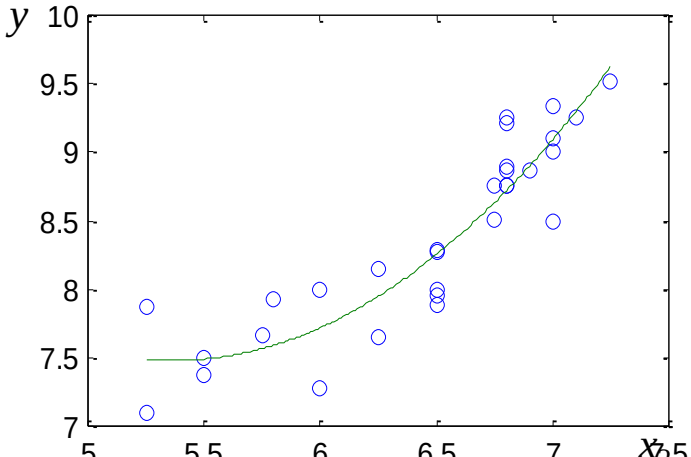

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

$y$~被解释变量（因变量）

$x_1, x_2$~解释变量(回归变量, 自变量)

$\beta_0, \beta_1, \beta_2, \beta_3$ ~回归系数

$\varepsilon$~随机误差（均值为零的正态分布随机变量）




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

# 模型求解

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \varepsilon$　由数据 $y, x_1, x_2$ 估计 $\beta$

```
[b,bint,r,rint,stats]=regress(y,x,alpha)
```

**输入**

- **y**~$n$维数据向量
- **x=** $[1 \quad x_1 \quad x_2 \quad x_2^2]$ ~$n$×4数据矩阵, 第1列为全1向量
- **alpha**(置信水平,0.05)

**输出**

- **b**~$\beta$的估计值
- **bint**~b的置信区间
- **r** ~残差向量**y-xb**
- **rint**~r的置信区间

| 参数 | 参数估计值 | 置信区间 |
|---|---|---|
| $\beta_0$ | 17.3244 | [5.7282 28.9206] |
| $\beta_1$ | 1.3070 | [0.6829 1.9311 ] |
| $\beta_2$ | -3.6956 | [-7.4989 0.1077 ] |
| $\beta_3$ | 0.3486 | [0.0379 0.6594 ] |
| $R^2$=0.9054 $F$=82.9409 $p$=0.0000 | | |

Stats~

检验统计量

$R^2, F, p$

结果分析

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

$y$的90.54%可由模型确定

$F$远超过$F$检验的临界值

$p$远小于α=0.05

模型从整体上看成立

**销售量预测** $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_2^2$

价格差$x_1$=其它厂家价格$x_3$-本公司价格$x_4$

估计$x_3$ 调整$x_4$ ⇨ 控制$x_1$ ⇨ 通过$x_1$, $x_2$预测$y$

控制价格差$x_1$=0.2元，投入广告费$x_2$=650万元

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_2^2 = 8.2933 \text{ (百万支)}$$

销售量预测区间为 [7.8230，8.7636]（置信度95%）

# 模型改进

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \beta_4 x_1 x_2 + \varepsilon$$

$x_1$和$x_2$对$y$的影响独立

$x_1$和$x_2$对$y$的影响有交互作用

| 参数 | 参数估计值 | 置信区间 |
|---|---|---|
| $\beta_0$ | 29.1133 | [13.7013 44.5252] |
| $\beta_1$ | 11.1342 | [1.9778 20.2906 ] |
| $\beta_2$ | -7.6080 | [-12.6932 -2.5228 ] |
| $\beta_3$ | 0.6712 | [0.2538 1.0887 ] |
| $\beta_4$ | -1.4777 | [-2.8518 -0.1037 ] |
| $R^2$=0.9209 $F$=72.7771 $p$=0.0000 | | |

# 两模型销售量预测比较

控制价格差$x_1$=0.2元，投入广告费$x_2$=6.5百万元

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\hat{\beta}_3x_2^2$$

$\hat{y}=8.2933$ (百万支)

区间 [7.8230，8.7636]

$$\hat{y}=\beta_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\hat{\beta}_3x_2^2+\hat{\beta}_4x_1x_2$$

$\hat{y}=8.3272$ (百万支)

区间 [7.8953，8.7592]

$\hat{y}$ 略有增加

预测区间长度更短


THANKS