# 第10章 回归

# 概念

- 回归分析：解释一组变量对另外一个变量的影响
  - 输入变量 or 自变量
  - 因变量
- 回归分析也可以看作一种有用的解释工具，用来识别对结果有较大影响的输入变量

# 用例

- 房地产：将住宅价格建模为以住宅区域为参考的函数，用来评估或者设定市场上住宅的价格。还可以包括其他参数，如卧室数量、地皮尺寸、学区排名等
- 需求预测：预测货物与服务的需求。如餐厅可以根据天气、星期几、特价商品、时间段来合理预测顾客数量
- 气温预测：根据环境信息来预测天气情况

回归分析的基本思想以及“回归”名称的由来最初是由英国生物学家兼统计学家高尔顿提出来的.

他从一千多对父母身高与其子女身高的数据分析中得出：当父亲身高很高时，儿子的身高并不像期待的那样高，而要稍矮一些，有向同龄人平均身高靠拢的现象；而当父亲身高很矮时，儿子的身高要比预期的高，也有向同龄人平均身高靠拢的现象.

正是因为儿子的身高有回到同龄人平均身高的这种趋势，才使人类的身高在一定时间内相对稳定，没有出现父辈个子高其子女更高，父辈个子矮其子女更矮的两极分化现象，说明后代的平均身高向中心靠拢了，这种现象叫回归，这就是“回归”一词的最初含义. 现在的意思是：凡是利用一个变量或一组变量来估计或预测另一个变量的情况都称之为回归。

在现实问题中处于同一个过程中的一些变量往往是相互依赖和相互制约的，它们之间的相互关系大致可分为两种：

(1)确定性关系 --函数关系

(2)非确定性关系 -- 相关关系：变量之间有一定的依赖关系，但这种关系并不完全确定。

可控变量：可以在某范围内随意地取指定数值– 自变量

不可控变量:可以观测但不可控制(随机变量)-- 因变量

例1：人的体重$y$与身高$x$之间的关系一般来说，身高高一些，体重也要重一些，但身高不能严格地确定体重，即同样身高的人，体重可能不同.

例2：人的血压$y$与年龄$x$之间的关系，不可能由一个人的年龄完全确定他的血压. 一般说人的年龄越大血压越高，但年龄相同者，血压未必相同.

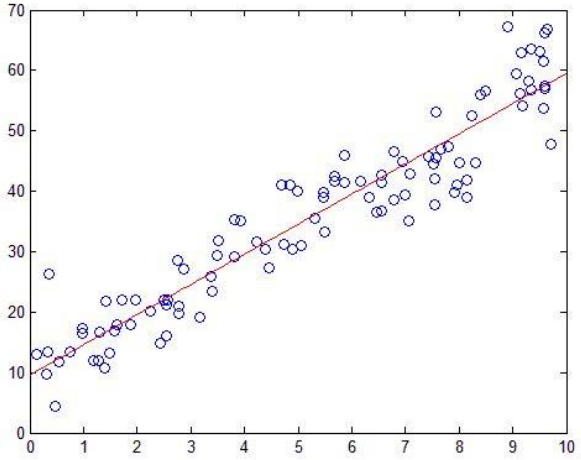
例3：水稻亩产量$y$与其施肥量$x_1$、播种量$x_2$、种子$x_3$有关系，但 $x_1$、$x_2$、$x_3$ 取相同的一组数值时，亩产量$y$可取不同数值.

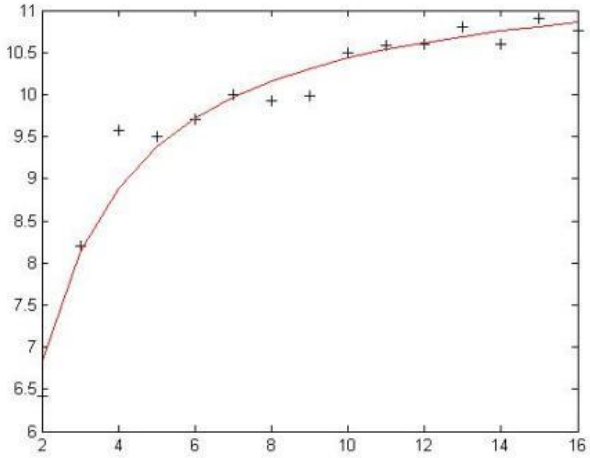
这几个例子中的两个变量之间都有一定的关系，且是一种非确定性的关系，称这类关系为相关关系.

上述例子中身高$x$，年龄$x$，施肥量$x_1$、播种量$x_2$、种子$x_3$都是可以在一定范围内随意的取指定数值，是可控变量称之为自变量，而体重$y$，血压$y$，亩产量$y$都是不可控变量称为因变量.

研究一个变量与一个(或几个)可控变量之间相关关系的统计分析方法称为回归分析.

在大数据分析中，回归分析是一种预测性的建模技术，它研究的是因变量（目标）和自变量之间的关系。这种技术通常用于预测分析、时间序列模型、变量之间的因果关系分析等领域。



# 分类

- 回归分析按照涉及变量的多少，分为一元回归分析和多元回归分析
- 按照因变量的多少，可分为简单回归分析和多重回归分析
- 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。

# 目录

- 一元线性回归
- 多元线性回归
- 非线性回归

# 一元线性回归

设 $X$ 与 $Y$ 有相关关系, 当自变量$x = x_0$ 时, 因变量 $Y$ 并不取固定的值与其对应. 如果要用函数关系近似 $X$ 与 $Y$ 的相关关系, 很自然想到, 应该以$EY_0$作为 $Y$ 与 $x = x_0$ 相对应的数值.

$$\begin{cases} Y = a + bx + \varepsilon \\ \varepsilon \sim N(0, \sigma^2) \end{cases} \qquad (1)$$

其中 $a,b,\sigma^2$ 为常数，则称 $Y$ 与 $x$ 之间存在线性相关关系，称（1）为一元正态线性回归模型，简称一元线性模型，其回归函数记为

$$\tilde{Y} = EY = a + bx$$

称为 $Y$ 对 $x$ 的线性回归，$a$ 称为回归常数，$b$ 称为回归系数。

由（1）得 $Y \sim N(a+bx,\sigma^2)$，可知 $x$ 取不同数值时，便得到不同的正态变量。

# 一元线性模型

$$\begin{cases} Y_1 = a + bx_1 + \varepsilon_1 \\ Y_2 = a + bx_2 + \varepsilon_2 \\ \cdots\cdots\cdots\cdots \\ Y_n = a + bx_n + \varepsilon_n \\ \varepsilon_1, \varepsilon_2, \cdots \varepsilon_n \text{相互独立，均服从} N(0, \sigma^2) \end{cases}$$

其中 $a, b, \sigma^2$ 为未知的常数。

由 $\varepsilon_1,\varepsilon_2,\cdots\varepsilon_n$ 独立知道 $Y_1,Y_2,\cdots Y_n$ 也相互独立，且

$$Y_i \sim N(a+bx_i,\sigma^2) \qquad i=1,2,\cdots n$$

$Y_1,Y_2,\cdots Y_n$称为来自$Y$的容量为$n$的一个独立随机样本（简称独立样本）。而

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

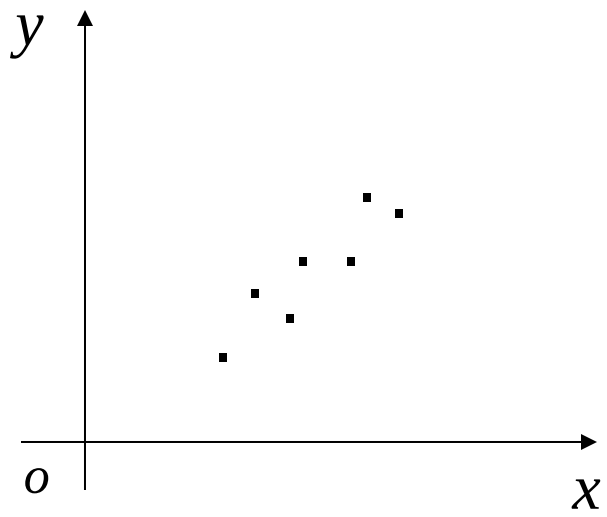
称为独立样本 $Y_1,Y_2,\cdots,Y_n$ 的一组（或一个）样本观测值，其中 $y_i,i=1,2,\cdots,n$ 为 $X$ 取固定值 $X=x_i$ 时，对 $Y_i$ 进行一次试验所得到的观测值。

利用**独立样本**及其**样本值**可得 $a,b,\sigma^2$ 的估计量及估计值 $\hat{a},\hat{b}$ 和 $\hat{\sigma}^2$, 从而得到回归函数 $\tilde{Y}=a+bx$ 的估计

$$\hat{Y}=\hat{a}+\hat{b}x$$

称为 $Y$ 对 $X$ 的经验回归方程或经验公式。

把样本值 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$ 作为平面直角坐标系的 $n$ 个点描出来，构成实验的散点图。




根据散点图，适当地选择一个函数 $\hat{y}=\hat{\mu}(x)$，使得 $(x_1,\hat{y}_1),(x_2,\hat{y}_2),\cdots(x_n,\hat{y}_n)$，在一定意义下最好地吻合于观测结果 $(x_1,y_1),(x_2,y_2),\cdots(x_n,y_n)$，常用的是最小二乘法，即

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}[y_i-\hat{\mu}(x_i)]^2=\min$$

## 二、未知参数的估计

### 1. 正规方程组、回归系数的点估计

根据最小二乘法求线性回归函数 $\tilde{y}=a+bx$ 的估计 $\hat{y}=\hat{a}+\hat{b}x$ 就是求使得

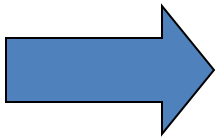
$$Q(a,b)=\sum_{i=1}^{n}[y_i-(a+bx_i)]^2$$

取得最小值的 $\hat{a},\hat{b}$，即

$$Q(\hat{a},\hat{b})=\min_{-\infty<a,b<+\infty}Q(a,b)=\min_{-\infty<a,b<+\infty}\sum_{i=1}^{n}[y_i-(a+bx_i)]^2$$

根据微分学中的二元函数极值的充分条件，将 $Q(a,b)$ 分别对 $a,b$ 求一阶偏导数并令其为零

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



正规方程组

经过整理后得到线性方程组

$$\begin{cases} na + n\bar{x}b = n\bar{y} \\ n\bar{x}a + (\sum_{i=1}^{n} x_i^2)b = \sum_{i=1}^{n} x_i y_i \end{cases}$$

其中

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

解此方程组即得使$Q(a,b)$取得最小值的$\hat{a},\hat{b}$

$$\begin{cases}\hat{a}=\bar{y}-\hat{b}\bar{x}\\ \hat{b}=\dfrac{\sum x_i y_i-n\bar{x}\bar{y}}{\sum(x_i-\bar{x})^2}=\dfrac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}=\dfrac{\sum(x_i-\bar{x})y_i}{\sum(x_i-\bar{x})^2}\end{cases}$$

分别称为$a,b$ 的最小二乘估计值. 于是，得到 $Y$ 对$x$ 的经验回归方程

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

注：用最小二乘法得到的经验回归直线通过已知$n$ 个数据点$(x_i,y_i)$ $i=1,2,\cdots,n$的几何重心$(\bar{x},\bar{y})$

事实上，最小二乘估计与极大似然估计是等价的

实际上：

在 $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$，$\beta_0$和$\beta_1$的最大似然估计为最小二乘估计

$$\hat{\beta}_0, \hat{\beta}_1$$

$y_1, y_2, \cdots, y_n$的似然函数为

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

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

例 在钢线碳含量$x$对于电阻效应$y$的研究中, 得到了以下数据:

| 碳含量（%） | 0.10 | 0.30 | 0.40 | 0.55 | 0.70 | 0.80 | 0.95 |
|---|---|---|---|---|---|---|---|
| 电阻（微欧） | 15 | 18 | 19 | 21 | 22.6 | 23.8 | 26 |

假设对于给定的 $x, y$ 为正态变量, 且方差与 $x$ 无关.

如果$x, y$满足经验公式 $y = \beta_0 + \beta_1 x$, 求线性回归方程

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

解 设 $y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon \sim N\left(0, \sigma^2\right)$

现在 $n = 7, \quad \sum x_i = 3.8, \quad \sum y_i = 145.4$

$$\sum x_i^{\,2} = 2.595 \qquad \sum x_i y_i = 85.61 \qquad \sum y_i^{\,2} = 3104.2$$

$$\Rightarrow L_{xx} = \sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2 = 2.595 - \frac{1}{7}\times 3.8^2 = 0.5321$$

$$L_{xy} = \sum x_i y_i - \frac{1}{n}\left(\sum x_i\right)\left(\sum y_i\right) = 85.61 - \frac{1}{7}\times 3.8\times 145.4 = 6.6786$$

$$L_{yy} = \sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2 = 3104.2 - \frac{1}{7}\times 145.4^2 = 84.0343$$

$$\Rightarrow \hat{\beta}_1 = \frac{L_{xy}}{L_{xx}} = \frac{6.6786}{0.5321} = 12.5503$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{1}{7}\times 145.4 - 12.5503 \times \frac{1}{7}\times 3.8 = 13.9584$$

所求的线性回归方程为

$$\hat{y} = 13.9584 + 12.5503x$$

# 多元线性回归

## 一. 多元线性回归模型

模型1

$$\begin{cases} y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \\ E(\varepsilon) = 0, D(\varepsilon) = \sigma^2 \end{cases}$$

模型2

$$\begin{cases} y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \\ \varepsilon \sim N\left(0, \sigma^2\right) \end{cases}$$

其中$\beta_0, \beta_1, \cdots, \beta_p, \sigma^2$是未知参数，$p \geq 2$

模型1和2称为$p$元线性回归模型.

$\beta_0$称为回归常数，$\beta_1, \beta_2, \cdots, \beta_p$称为回归系数，

$\varepsilon$称为随机误差. $x_1, x_2, \cdots, x_p$是自变量，

$y$是随机变量称为解释变量或因变量

在模型1下，有

$$E(y)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p, D(y)=\sigma^2$$

在模型2下，有

$$y\sim N\left(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p,\sigma^2\right)$$

观测数据$(x_{i1}, x_{i2}, \cdots, x_{ip}; y_i)$　$i = 1, 2, \cdots, n$满足

模型3
$$
\begin{cases}
y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + \varepsilon_1 \\
y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + \varepsilon_2 \\
\vdots \quad \vdots \qquad\qquad\qquad\qquad \vdots \\
y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + \varepsilon_n \\
\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n \text{相互独立, 且} \\
E(\varepsilon_i) = 0, D(\varepsilon_i) = \sigma^2, i = 1, 2, \cdots, n
\end{cases}
$$

观测数据$(x_{i1}, x_{i2}, \cdots, x_{ip}; y_i)$ $i = 1, 2, \cdots, n$满足

模型4
$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + \varepsilon_2 \\ \vdots \quad \vdots \qquad\qquad\qquad \vdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + \varepsilon_n \\ \varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n \text{独立同分布, 且} \varepsilon_i \sim N\left(0, \sigma^2\right) \end{cases}$$

**称模型3和模型4为$y$关于$x$的*p*元样本线性回归模型.**

$n>p+1$，记

$$Y=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X=\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ & & \cdots & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}_{n\times(p+1)}, \beta=\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \varepsilon=\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

**模型3和模型4可表示为如下矩阵形式**

$$\text{模型3}\begin{cases} Y = X\beta + \varepsilon \\ E(\varepsilon) = 0, D(\varepsilon) = \sigma^2 I_n \end{cases}$$

$$\text{模型4}\begin{cases} Y = X\beta + \varepsilon \\ \varepsilon \sim N(0, \sigma^2 I_n) \end{cases}$$

**其中$I_n$为 $n$ 阶单位矩阵，矩阵$X$是$n\times(p+1)$矩阵称为设计矩阵，且秩$(X)=p+1$（假设）**

# 二. 未知参数的估计

## 1. 最小二乘估计

误差平方和

$$Q(\beta_0,\beta_1,\cdots,\beta_p)=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_{i1}-\beta_2x_{i2}-\cdots-\beta_px_{ip})^2$$

最小二乘法：求 $\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_p$ 使

$$Q(\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_p)=(Y-X\hat{\beta})^{\tau}(Y-X\hat{\beta})$$

$$=\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2}-\cdots-\hat{\beta}_px_{ip})^2$$

$$=\min_{\beta_0,\beta_1,\cdots,\beta_p}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_{i1}-\beta_2x_{i2}-\cdots-\beta_px_{ip})^2$$

求 $\frac{\partial Q}{\partial \beta_0}$、$\frac{\partial Q}{\partial \beta_j}(j=1,2,\cdots,p)$ 并令其都等于0，整理后得到如下正规方程组：

$$
\begin{cases}
n\beta_0+\beta_1\sum_{i=1}^{n}x_{i1}+\beta_2\sum_{i=1}^{n}x_{i2}+\cdots+\beta_p\sum_{i=1}^{n}x_{ip}=\sum_{i=1}^{n}y_i \\
\beta_0\sum_{i=1}^{n}x_{i1}+\beta_1\sum x_{i1}^2+\beta_2\sum_{i=1}^{n}x_{i1}x_{i2}+\cdots+\beta_p\sum_{i=1}^{n}x_{i1}x_{ip}=\sum_{i=1}^{n}x_{i1}y_i \\
\beta_0\sum_{i=1}^{n}x_{i2}+\beta_1\sum_{i=1}^{n}x_{i1}x_{i2}+\beta_2\sum_{i=1}^{n}x_{i2}^2+\cdots+\beta_p\sum_{i=1}^{n}x_{i2}x_{ip}=\sum_{i=1}^{n}x_{i2}y_i \\
\cdots\quad\cdots\quad\cdots \\
\beta_0\sum_{i=1}^{n}x_{ip}+\beta_1\sum_{i=1}^{n}x_{i1}x_{ip}+\beta_2\sum_{i=1}^{n}x_{i2}x_{ip}+\cdots+\beta_p\sum_{i=1}^{n}x_{ip}^2=\sum_{i=1}^{n}x_{ip}y_i
\end{cases}
$$

正规方程组的解 $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$

就是 $\beta_0, \beta_1, \cdots, \beta_p$ 的最小二乘估计

由于

$$X^{\tau}X = \begin{pmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ip} \\ \sum_{i=1}^{n} x_{i1} & \sum x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ip} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \sum_{i=1}^{n} x_{ip} & \sum_{i=1}^{n} x_{ip}x_{i1} & \sum_{i=1}^{n} x_{ip}x_{i2} & \cdots & \sum_{i=1}^{n} x_{ip}^2 \end{pmatrix}$$

及

$$X^{\tau}Y=\begin{pmatrix}1 & 1 & 1 & \cdots & 1\\ x_{11} & x_{21} & x_{31} & \cdots & x_{n1}\\ \vdots & \vdots & \cdots & \vdots & \vdots\\ x_{1p} & x_{2p} & x_{3p} & \cdots & x_{np}\end{pmatrix}\begin{pmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{pmatrix}=\begin{pmatrix}\sum_{i=1}^{n} y_i\\ \sum_{i=1}^{n} x_{i1}y_i\\ \vdots\\ \sum_{i=1}^{n} x_{ip}y_i\end{pmatrix}$$

于是正规方程组用矩阵表示为

$$X^{\tau}X\beta=X^{\tau}Y$$

由于$Rank(X)=p+1$,因此 $X^{\tau}X$ 必存在逆阵 $(X^{\tau}X)^{-1}$

解正规方程组得到 $\beta$ 的估计为

$$\hat{\beta}=(X^{\tau}X)^{-1}X^{\tau}Y$$

称之为 $\beta$ 的最小二乘估计.

于是线性回归方程为

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\cdots+\hat{\beta}_kx_k$$

## 2. 最大似然估计

多元线性回归系数的最大似然估计与一元线性回归时求最大似然估计的想法一样

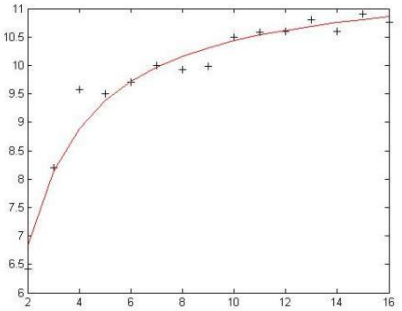
$\beta=(\beta_0,\beta_1,\cdots,\beta_p)^\tau$的最大似然估计

与最小二乘估计一样，是

$$\hat{\beta}=(X^\tau X)^{-1}X^\tau Y$$

# 非线性回归

在很多实际问题中，两个或者多个变量之间的关系不一定是线性关系. 若此时建立线性回归方程，效果肯定不会好. 而如果观测值的散点图大致呈某一曲线，又存在某种变换可将该曲线转换成直线，于是就可以选择该变换把问题转换成线性回归的问题，从而利用线性回归的一些结果解决问题.




具体做法：

1）根据样本数据，在直角坐标系中画出散点图

2）根据散点图，推测出Y与x之间的函数关系

3）选择适当的坐标变换，使之变成线性关系

4）用线性回归方法求出线性回归方程

5）返回到原来的函数关系，得到要求的回归方程

# 可线性化的一元非线性回归

1. 双曲线: $\frac{1}{y}=a+\frac{b}{x}$ 令 $x'=\frac{1}{x}$, $y'=\frac{1}{y}$

$$\Rightarrow y'=a+bx'$$

2. 幂函数: $y=ax^{b} \Rightarrow \ln y=\ln a+b\ln x$

令 $y'=\ln y,\quad \beta_0=\ln a,\quad x'=\ln x,\quad \beta_1=b$

$$\Rightarrow y'=\beta_0+\beta_1 x'$$

3. 指数曲线： $y = ae^{bx}$　　取对数得　$\ln y = \ln a + bx$

令　$y' = \ln y, \quad a' = \ln a$

$$\Rightarrow y' = a' + bx$$

4. 倒指数曲线：　$y = ae^{b/x}$　取对数得　$\ln y = \ln a + \dfrac{b}{x}$

令　$y' = \ln y, \quad a' = \ln a, \quad x' = \dfrac{1}{x}$

$$\Rightarrow y' = a' + bx'$$

5. 对数曲线：　$y = a + b\ln x$　　令　$x' = \ln x$

$$\Rightarrow y = a + bx'$$

**6、S型（Logistic）曲线** $y=\dfrac{K}{1+Ae^{-\lambda x}}$

变形 $y(1+Ae^{-\lambda x})=K \rightarrow y+Aye^{-\lambda x}=K$

$$\rightarrow Ae^{-\lambda x}=\frac{K-y}{y} \rightarrow \ln A-\lambda x=\ln\left(\frac{K-y}{y}\right)$$

令 $y'=\ln\dfrac{K-y}{y}, a=\ln A$ ⟶ $y'=a-\lambda x$

## 7、多项式模型

任意连续函数都可由多项式逼近

$$y=\beta_0+\beta_1 x+\beta_2 x^2+\cdots+\beta_p x^p$$

令$y'=y,\ \ x_1=x, x_2=x^2,\cdots,x_p=x^p$

$$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p+\varepsilon$$

**例1** 在彩色显影中, 根据以往经验, 形成**染料光学密度**与**析出银的光学密度**之间呈倒指数曲线关系:

$$y = ae^{b/x}, \quad b > 0$$

已测得11对数据见下表

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $x$ | 0.05 | 0.06 | 0.07 | 0.10 | 0.14 | 0.20 | 0.25 | 0.31 | 0.38 | 0.43 | 0.47 |
| $y$ | 0.10 | 0.14 | 0.23 | 0.37 | 0.59 | 0.79 | 1.00 | 1.12 | 1.19 | 1.25 | 1.29 |

试求出经验回归曲线方程.

解 （1）由 $y = ae^{b/x} \Rightarrow \ln y = \ln a + \frac{b}{x}$

令 $y' = \ln y, \quad \beta_0 = \ln a, \quad \beta_1 = b, \quad x' = \frac{1}{x}$

$$\Rightarrow y' = \beta_0 + \beta_1 x'$$

经计算得

$$n = 11$$

$$\sum_{i=1}^{n} x_i' = \sum_{i=1}^{n} \frac{1}{x_i} = 87.408, \qquad \overline{x}' = 7.95$$

$$\sum_{i=1}^{n} y_i' = \sum_{i=1}^{n} \ln y_i = -6.732, \qquad \overline{y}' = -0.612$$

$$\sum_{i=1}^{n} x'^{2}_{i} = \sum_{i=1}^{n} \frac{1}{x_i^2} = 1101.16$$

$$\sum_{i=1}^{n} y'^{2}_{i} = \sum_{i=1}^{n} (\ln y_i)^2 = 12.82$$

$$\sum_{i=1}^{n} x'_i \, y'_i = \sum_{i=1}^{n} \frac{1}{x_i} \ln y_i = -112.84$$

$$n = 11$$

$$\bar{x}' = 7.95$$

$$\bar{y}' = -0.612$$

$$\Rightarrow L_{x'x'} = \sum_{i=1}^{n} x'^{2}_{i} - n\bar{x}'^{2} = 406.6$$

$$L_{x'y'} = \sum_{i=1}^{n} x'_i \, y'_i - n\bar{x}'\bar{y}' = -59.35$$

$$L_{y'y'} = \sum_{i=1}^{n} y_i^2 - n\bar{y}'^{2} = 8.70$$

$$\Rightarrow \hat{\beta}_1 = \frac{L_{x'y'}}{L_{x'x'}} = \frac{-59.35}{406.6} = -0.146$$

$$\hat{\beta}_0 = \bar{y}' - \hat{\beta}_1 \bar{x}'$$
$$= -0.612 + 0.146 \times 7.95 = 0.549$$

=>线性回归方程为

$$\hat{y}' = 0.549 - 0.146x'$$

$$\Rightarrow \ln \hat{y} = 0.549 - \frac{0.146}{x}$$

=>**曲线回归方程为**

$$\hat{y} = e^{0.549 - 0.146/x} = e^{0.549} e^{-0.146/x} = 1.73e^{-0.146/x}$$

$$\Rightarrow \hat{a} = 1.73,\ \hat{b} = -0.146$$

$n = 11$

$\bar{x}' = 7.95$

$\bar{y}' = -0.612$

$L_{x'x'} = 406.6$

$L_{x'y'} = -59.35$

$L_{y'y'} = 8.70$

**例2** 测定某肉鸡的生长过程，每两周记录一次鸡的重量，数据如下表

| x/周 | 2 | 4 | 6 | 8 | 10 | 12 | 14 |
|---|---|---|---|---|---|---|---|
| y/kg | 0.3 | 0.86 | 1.73 | 2.2 | 2.47 | 2.67 | 2.8 |

由经验知鸡的生长曲线为**Logistic**曲线，且极限生长量为**k=2.827**，试求**y**对**x**的回归曲线方程。

解 由题设可建立鸡重**y**与时间**x**的相关关系为

$$y = \frac{2.827}{1 + Ae^{-\lambda x}}$$

令 $y' = \ln\dfrac{2.827 - y}{y}, a = \ln A$ 则有 $y' = a - \lambda x$

列表计算

| 序号 | x | y | y' | $X^2$ | $y'^2$ | xy' |
|---|---|---|---|---|---|---|
| 1 | 2 | 0.3 | 2.131 | 4 | 4.541 | 4.262 |
| 2 | 4 | 0.86 | 0.827 | 16 | 0.684 | 3.309 |
| 3 | 6 | 1.73 | -0.456 | 36 | 0.208 | -2.733 |
| 4 | 8 | 2.2 | -1.255 | 64 | 1.576 | -10.042 |
| 5 | 10 | 2.47 | -1.934 | 100 | 3.741 | -19.342 |
| 6 | 12 | 2.67 | -2.834 | 144 | 8.029 | -34.003 |
| 7 | 14 | 2.8 | -4.642 | 196 | 21.544 | -64.982 |
| ∑ | 56 | 13.03 | -8.162 | 560 | 40.323 | -123.531 |

所以 $\overline{x}=8.00 \quad \overline{y'}=-1.166$

$$L_{xx}=112 \quad L_{y'y'}=30.807 \quad L_{xy'}=-58.236$$

$$-\lambda=\frac{L_{xy'}}{L_{xx}}=-0.519967$$

$$a=\overline{y'}-\lambda\overline{x}=2.993762$$

$$A=e^{a}=19.96063$$

所以所求曲线方程为 $y=\dfrac{2.827}{1+19.9606e^{-0.51997x}}$

需要指出一点的是新引进的自变量只能依赖于原始变量，而与未知参数无关. 一般来说，变换的选择并不是一件容易的事. 事实上，根据散点图选择一种变换只能近似反映$y$与$x$的关系.

应该指出，对原始数据变换，把曲线回归转化为线性回归，利用线性回归的性质，即使对变换后的线性回归成立，也不能保证对原始数据的曲线回归成立，即线性回归性质经过变换后不一定能保持.

# 附录1：逐步回归

# Stepwise Regression（逐步回归）

- 逐步回归主要用于处理多个自变量的回归问题。
- 在这种技术中，自变量的选择是在一个自动的过程中完成的。模型通过观察统计量，如R-square，t-stats和AIC指标，来识别重要的变量。
- 关键是如何找到**最优回归方程**

# 逐步回归分析

- 寻找“最优回归方程”总的指导原则是：
  - 对因变量有显著作用的自变量，全部选入回归方程；
  - 对因变量无显著作用的自变量，一个也不引入回归方程。
- 选择“最优回归方程”的方法有:
  - 最优子集回归法
  - 逐步选择法
    - 向后剔除法（backward selection）
    - 向前引入法（forward selection）
    - 逐步回归法（stepwise selection）

## 1 最优子集回归法

求出所有自变量可能组合子集的回归方程的模型（共有$2^m-1$个，m为自变量的个数），按一定准则选择最优模型，常用的准则有：

- 校正后的决定系数（考虑了自变量的个数）
- Cp准则（C即criterion，p为所选模型中变量的个数；Cp接近p+1的模型为最优）
- AIC(Akaike`s Information Criterion)准则；AIC 越小越好

## 最优子集法的局限性

如果自变量个数为4，则所有的回归有$2^4-1=15$个；当自变量数个数为10时，所有可能的回归为 $2^{10}-1=$ 1023个；……..；当自变量数个数为50时，所有可能的回归为$2^{50}-1\approx10^{15}$个。

## 2 逐步选择法

- 前进法（forward selection）
- 后退法（backward elimination）
- 逐步回归法（stepwise selection）

它们的共同特点是每一步只引入或剔除一个自变量。决定其取舍则基于对偏回归平方和的$F$检验

$$F_j = \frac{SS_{回} - SS_{回(-j)}}{SS_{残}/(n-p-1)}; df_1 = 1; df_2 = n-p-1$$

## ◆前进法

**自变量从无到有、从少到多**

- Y对每一个自变量作直线回归，对回归平方和最大的自变量作$F$ 检验，有意义（$P$小）则引入。
- 在此基础上，计算其它自变量的偏回归平方和，选取偏回归平方和最大者作$F$ 检验，…。

局限性：后续变量的引入可能会使先进入方程的自变量变得不重要。

## ◆后退法

**先将全部自变量放入方程，然后逐步剔除**

- 偏回归平方和最小的变量，作*F*检验及相应的P值，决定它是否剔除（*P*大）。
- 建立新的回归方程。重复上述过程。

局限性：自变量高度相关时，可能得不出正确的结果；开始时剔除的变量即使后来变得有显著性也不能再进入方程。

## ◆逐步回归法

**双向筛选：引入有意义的变量（前进法），剔除无意义变量（后退法），即引入与剔出相互交替，而引入与剔出的根据就是自变量在回归方程中的偏回归平方和。**

小样本检验水准α一般定为0.10或0.15，大样本把*a*值定为0.05。α值越小表示选取自变量的标准越严。

- 逐步回归的基本思想：总结
  - **前进法**：从一个自变量开始，将自变量一个一个地引入方程，并且在每一次引入一个自变量时，这个自变量的偏回归平方和，经过检验应该是所有尚未引入回归方程的自变量中最为显著的那一个；
  - **后退法**：在引入一个新的自变量、建立新的线性回归方程之后，接着对早先引入方程的自变量逐个进行检验，由偏回归平方和最小的自变量开始，将偏回归平方和经过检验不显著的自变量从回归方程中逐个地剔出；
  - **双向筛选法**：引入自变量与剔出自变量交替进行，直到再也不能引入新的自变量又不能从方程中剔出已列入的自变量为止。

# 附录2：岭（Ridge）回归

# 岭回归估计的定义

## 一、普通最小二乘估计带来的问题

当自变量间存在复共线性时，回归系数估计的方差就很大，估计值就很不稳定，下面进一步用一个模拟的例子来说明这一点。

**例** 假设已知$x_1$，$x_2$与$y$的关系服从线性回归模型

$$y=10+2x_1+3x_2+\varepsilon$$

给定 $x_1$，$x_2$的 10 个值，如下表 7.1 的第 2、3 两行：

表

| | 序号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (1) | $x_1$ | 1.1 | 1.4 | 1.7 | 1.7 | 1.8 | 1.8 | 1.9 | 2.0 | 2.3 | 2.4 |
| (2) | $x_2$ | 1.1 | 1.5 | 1.8 | 1.7 | 1.9 | 1.8 | 1.8 | 2.1 | 2.4 | 2.5 |
| (3) | $\varepsilon_i$ | 0.8 | −0.5 | 0.4 | −0.5 | 0.2 | 1.9 | 1.9 | 0.6 | −1.5 | −1.5 |
| (4) | $y_i$ | 16.3 | 16.8 | 19.2 | 18.0 | 19.5 | 20.9 | 21.1 | 20.9 | 20.3 | 22.0 |

# 岭回归估计的定义

现在我们假设回归系数与误差项是未知的，用普通最小二乘法求回归系数的估计值得：

$$\hat{\beta}_0 = 11.292,\quad \hat{\beta}_1 = 11.307,\quad \hat{\beta}_2 = -6.591$$

而原模型的参数

$$\beta_0 = 10, \beta_1 = 2, \beta_2 = 3$$

看来相差太大。计算 $x_1$，$x_2$ 的样本相关系数得 $r_{12} = 0.986$，表明 $x_1$ 与 $x_2$ 之间高度相关。

# 岭回归估计的定义

## 二、岭回归的定义

岭回归(Ridge Regression,简记为RR)提出的想法是很自然的。

当自变量间存在线性关系时，$|\mathbf{X}'\mathbf{X}|\approx 0$，我们设想给$\mathbf{X}'\mathbf{X}$加上一个正常数矩阵$k\mathbf{I}$，（$k>0$)，那么$\mathbf{X}'\mathbf{X}+k\mathbf{I}$接近奇异的程度就会比$\mathbf{X}'\mathbf{X}$接近奇异的程度小得多。

考虑到变量的量纲问题，我们先对数据做标准化，为了记号方便，标准化后的设计阵仍然用$\mathbf{X}$表示

# 岭回归估计的定义

我们称 $$\hat{\boldsymbol{\beta}}(k)=(X'X+kI)^{-1}X'y \qquad (2)$$

为β的岭回归估计，其中$k$称为岭参数。

由于假设$\mathbf{X}$已经标准化，所以$\mathbf{X}'\mathbf{X}$就是自变量样本相关阵，（2）式计算的实际是标准化岭回归估计。（2）式中因变量观测向量$\mathbf{y}$可以经过标准化也可以未经标准化。显然，岭回归做为β的估计应比最小二乘估计稳定，当k=0时的岭回归估计就是普通的最小二乘估计。

$$\text{loss} = \underset{\beta \in \mathbf{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + k \underbrace{\|\beta\|_2^2}_{\text{Penalty}}$$

# 岭回归估计的定义

因为岭参数 k 不是唯一确定的，所以我们得到的岭回归估计 $\hat{\boldsymbol{\beta}}(k)$ 实际是回归参数 $\boldsymbol{\beta}$ 的一个估计族。

例如对例 1 可以算得不同 k 值时的 $\hat{\beta}_1(k)$， $\hat{\beta}_2(k)$，见表 2

**表2**

| k | 0 | 0.1 | 0.15 | 0.2 | 0.3 | 0.4 | 0.5 | 1.0 | 1.5 | 2 | 3 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\hat{\beta}_1(k)$ | 11.31 | 3.48 | 2.99 | 2.71 | 2.39 | 2.20 | 2.06 | 1.66 | 1.43 | 1.27 | 1.03 |
| $\hat{\beta}_2(k)$ | −6.59 | 0.63 | 1.02 | 1.21 | 1.39 | 1.46 | 1.49 | 1.41 | 1.28 | 1.17 | 0.98 |

# 岭回归估计的定义

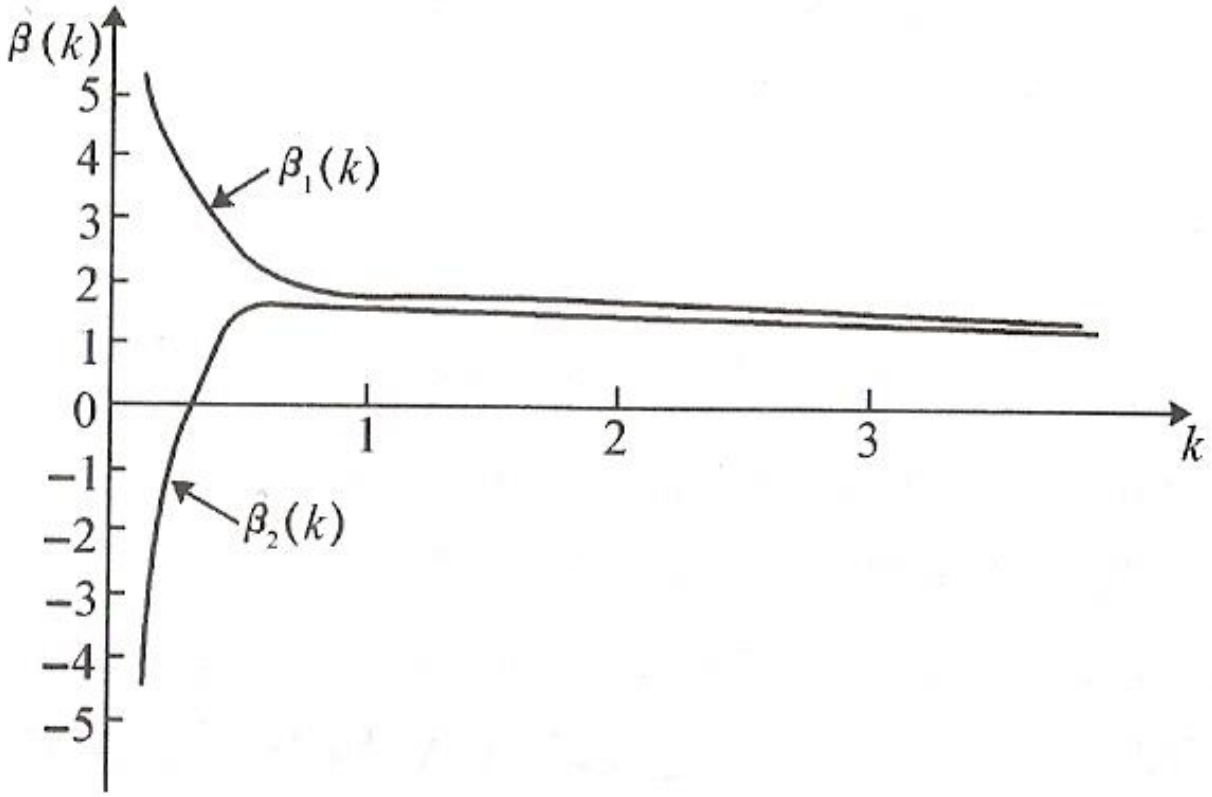



图 7.1

# 岭回归估计的性质

在本节岭回归估计的性质的讨论中，假定（2）式中因变量观测向量**y**未经标准化。

$$\hat{\boldsymbol{\beta}}(k)=(X'X+kI)^{-1}X'y \qquad (2)$$

**性质 1** $\hat{\boldsymbol{\beta}}(k)$ 是回归参数**β** 的有偏估计。

证明：
$$E[\hat{\boldsymbol{\beta}}(k)]=E[(X'X+kI)^{-1}X'y]$$
$$=(X'X+kI)^{-1}X'E(y)$$
$$=(X'X+kI)^{-1}X'X$$

显然只有当 $k=0$ 时，$E[\hat{\boldsymbol{\beta}}(0)]=\boldsymbol{\beta}$ ；当 $k\neq 0$ 时，$\hat{\boldsymbol{\beta}}(k)$ 是**β** 的有偏估计。

要特别强调的是 $\hat{\boldsymbol{\beta}}(k)$ 不再是**β** 的无偏估计了，

有偏性是岭回归估计的一个重要特性。

# 岭回归估计的性质

**性质 2**　在认为岭参数 k 是与 y 无关的常数时，$\hat{\boldsymbol{\beta}}(k)=(X'X+kI)^{-1}X'y$ 是最小二乘估计 $\hat{\boldsymbol{\beta}}$ 的一个线性变换，也是 y 的线性函数。

因为 $\hat{\boldsymbol{\beta}}(k)=(X'X+kI)^{-1}X'y=(X'X+kI)^{-1}X'X(X'X)^{-1}X'y$

$$=(X'X+kI)^{-1}X'X\hat{\boldsymbol{\beta}}$$

因此，岭估计 $\hat{\boldsymbol{\beta}}(k)$ 是最小二乘估计 $\hat{\boldsymbol{\beta}}$ 的一个线性变换，根据定义式 $\hat{\boldsymbol{\beta}}(k)=(X'X+kI)^{-1}X'y$ 知 $\hat{\boldsymbol{\beta}}(k)$ 也是 y 的线性函数。

# 岭回归估计的性质

**性质 3** 对任意 $k>0$，$\|\hat{\boldsymbol{\beta}}\|\neq 0$，总有

$$\|\hat{\boldsymbol{\beta}}(k)\|<\|\hat{\boldsymbol{\beta}}\|$$

这里 $\|\bullet\|$ 是向量的模，等于向量各分量的平方和。

这个性质表明 $\hat{\boldsymbol{\beta}}(k)$ 可看成由 $\hat{\boldsymbol{\beta}}$ 进行某种向原点的压缩，从 $\hat{\boldsymbol{\beta}}(k)$ 的表达式可以看到，当 $k\to\infty$ 时，$\hat{\boldsymbol{\beta}}(k)\to 0$，即 $\hat{\boldsymbol{\beta}}(k)$ 化为零向量。

# 岭回归估计的性质

**性质 4**　以 MSE 表示估计向量的均方误差，则存在 $k>0$，使得

$$MSE(\hat{\boldsymbol{\beta}}(k)) < MSE(\hat{\boldsymbol{\beta}})$$

即

$$\sum_{j=1}^{p} E(\hat{\beta}_j(k)-\beta_j)^2 < \sum_{j=1}^{p} D(\hat{\beta}_j)$$

# 附录3： Lasso回归

类似于岭回归，Lasso （Least Absolute Shrinkage and Selection Operator）回归也会就**回归系数向量**给出惩罚值项。回想岭回归的目标函数为

$$L_1 = arg\min \| y - X\beta \|_2^2 + \lambda \| \beta \|_2^2$$

不同之处，Lasso回归能够减少变化程度并提高线性回归模型的精度，它所采用的优化目标如下：

$$L_1 = arg\min \| y - X\beta \|_2^2 + \lambda \| \beta \|_1$$

注意这里使用的惩罚函数是$L_1$范数，而不是$L_2$范数。这导致惩罚（或等于约束估计的绝对值之和）值使一些参数估计结果等于零。使用惩罚值（$\lambda$）越大，进一步使得不重要的参数缩小趋近于零。换句话说，Lasso回归具有变量选择的作用。

如果预测的一组变量是高度相关的，Lasso 会选出其中一个变量并且将其它的收缩为零。