

7 回归（regression）分析 的前世今生



授课教师： 赵春晖

联系方式：

Email: chhzhao@zju.edu.cn

Phone: 13588312064

Room: 工控新楼308室



提纲

前世

- 线性回归（普通最小二乘OLS）
- 岭回归
 - 原理
 - 回归模型

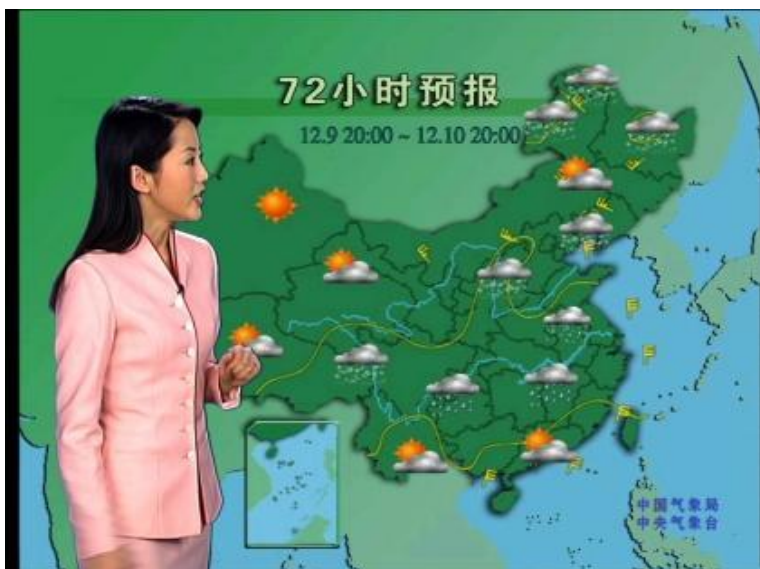
引言

如今，天气预报与人们的生活越来越密切。

你知道日常生活中的天气预报是如何实现的吗？

气象学家根据**既往的温度、湿度以及降雨**等资料，就可以预报未来一段时间某地的天气变化情况。

这要求对这些**变量之间的关系**有精确的掌握。-----**回归分析**



[浙江杭州天气预报](#) [一周天气预报](#) [中国天气网](#)

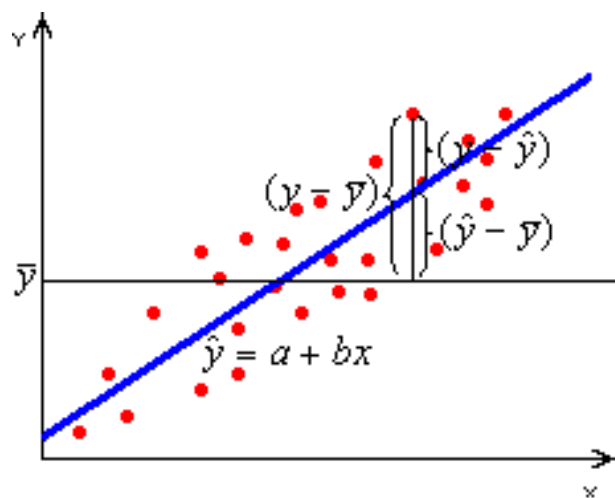


回归的基本概念

回归分析：确定**变量之间数量关系**的可能形式也即**数量模型**。

自变量是指引起**因变量**发生变化的因素或条件，因此自变量**被看作是**因变量的**原因**。

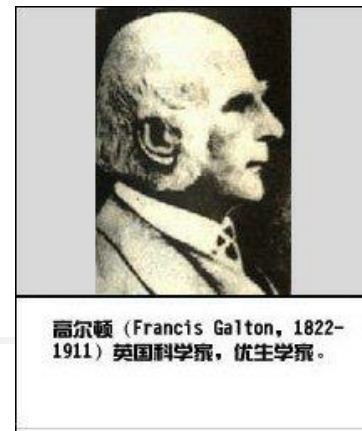
$$y = a + bx$$



变量间的数量关系可以用**散点图**来反映，图中的每个点都代表一个变量配对样本点，它是**自变量与因变量间关系**的一个具体代表。

区分 相关分析法：分析**两个变量之间关系的强度**（强、弱）

为什么叫回归分析？



- 回归分析最早是19世纪末期**高尔顿** (Sir Francis Galton) 所发展。
- 高尔顿是**生物统计学派**的奠基人, 他的表哥**达尔文**的巨著《**物种起源**》问世以后, 触动他用统计方法研究**智力进化**问题, 统计学上的“**相关**”和“**回归**”的概念也是高尔顿第一次使用的。

为什么叫回归分析？

- ◆ 1885年，高尔顿发表了一篇“**遗传的身高向平均数方向的回归**”文章：发现父母的身高可以预测子女的身高，他将子女与父母身高的这种现象**拟合出一种线形关系**。



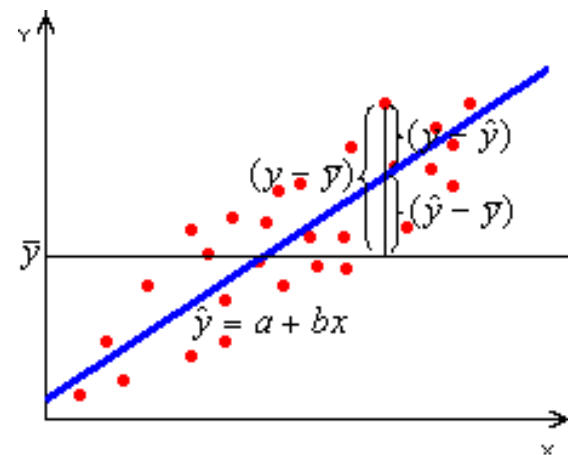
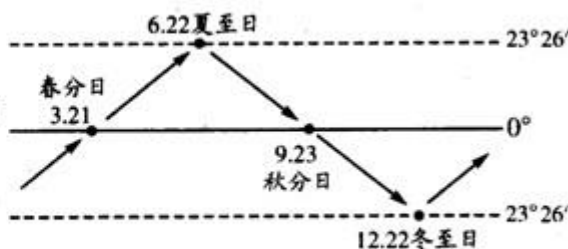
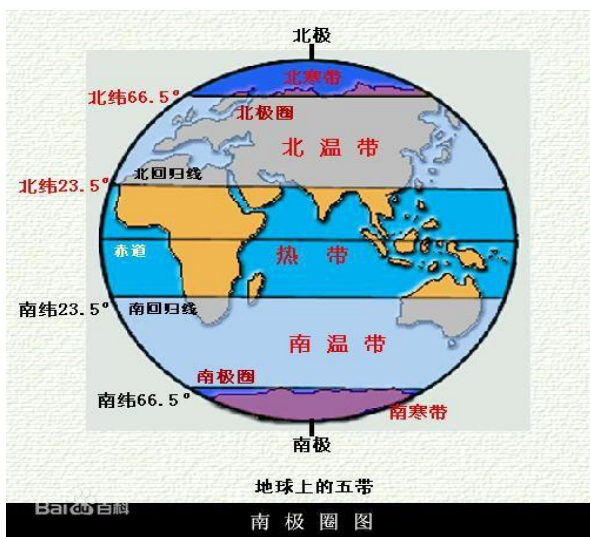
男孩 $(\text{爸爸的身高} + \text{妈妈的身高} + 13)/2$

女孩 $(\text{爸爸的身高} + \text{妈妈的身高} - 13)/2$

- ◆ 尽管这是一种拟合较好的线形关系，但仍然存在**例外**现象：矮个的人的儿子比其父要高，身材较高的父母所生子女的身高到一定程度后会往平均身高方向发生“**回归**”。这种效应被称为“**趋中回归**”。

为什么叫回归分析？

□ 地理回归线



- ◆ 太阳直射点在南北回归线之间徘徊
- ◆ 回归的含义可以具象化成一种几何过程。在这一个过程中，“中心线”两侧的值不断向“中心线”靠拢，也就是回归。



回归分析的主要内容

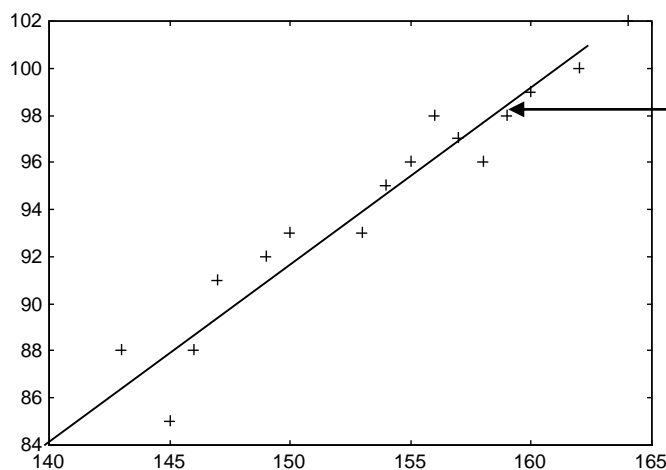
- ① 建立数学表达式：从一组数据出发确定某些变量之间的定量关系式，即建立数学模型并估计其中的未知参数。
- ② 选取重要自变量：在许多自变量共同影响着一个因变量的关系中，判断哪个（或哪些）自变量的影响是显著的，哪些自变量的影响是不显著的，将影响显著的自变量选入模型中，而剔除影响不显著的变量。
- ③ 预测应用：利用所求的关系式对某一过程进行预测。

回归分析的数学模型

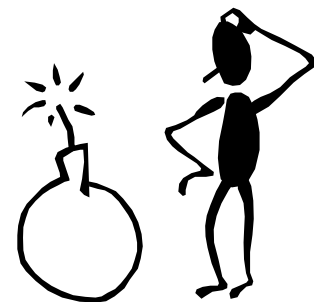
例1 测16名成年女子的身高与腿长所得数据如下：

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

以身高 x 为横坐标，以腿长 y 为纵坐标将这些数据点 (x_i, y_i) 在平面直角坐标系上标出



散点图





回归分析的数学模型

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

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

固定的未知参数 β_0 、 β_1 称为回归系数，自变量 x 也称为回归变量。

称为 **y 对 x 的回归直线方程**。

中心化处理 $y = \beta_1 x + \varepsilon$



回归分析的数学模型

若自变量为 m 个， x_j ($j=1, 2, \dots, m$)，因变量为 y ，在 y 与 x_j 间，我们可以建立多元线性回归模型，即

$$y = b_1 x_1 + b_2 x_2 + \dots + b_m x_m + e$$

在左式中， b_j 为回归系数。

$$y = \sum_{j=1}^m b_j x_j + e$$

中心化处理


在上边的叙述中，因变量为1个，而事实上可以有多个因变量。如有两个因变量 y_1 和 y_2 ，我们可以简单地写成两个线性方程：

$$y_1 = \sum_{j=1}^m b_{1,j} x_j + e_1 \quad y_2 = \sum_{j=1}^m b_{2,j} x_j + e_2$$

有截距情况下的统一表达

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \dots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \dots \\ \mathbf{x}_n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \dots \\ a_m \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1m} \\ 1 & x_{21} & \dots & x_{2m} \\ 1 & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{nm} \end{bmatrix}$$


$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{a}$$



回归分析的数学模型

若用矩阵表示，则自变量和因变量矩阵可以表示为：

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \quad \mathbf{Y} = (\mathbf{y}_1 \quad \mathbf{y}_2) = \begin{bmatrix} y_{11} & y_{12} \\ y_{12} & y_{22} \\ \cdots & \cdots \\ y_{1n} & y_{2n} \end{bmatrix}$$

回归系数矩阵和残差矩阵可以表示为：

$$\mathbf{B} = (\mathbf{b}_1 \quad \mathbf{b}_2) = \begin{bmatrix} b_{11} & b_{21} \\ b_{12} & b_{22} \\ \cdots & \cdots \\ b_{1m} & b_{2m} \end{bmatrix} \quad \mathbf{E} = (\mathbf{e}_1 \quad \mathbf{e}_2) = \begin{bmatrix} e_{11} & e_{21} \\ e_{12} & e_{22} \\ \cdots & \cdots \\ e_{1n} & e_{2n} \end{bmatrix}$$

由此得到

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}; \quad \mathbf{y}_1 = \mathbf{Xb}_1; \quad \mathbf{y}_2 = \mathbf{Xb}_2$$



回归分析模型的解法

第一代回归方法：

- 我普通但我很严谨：普通最小二乘
(ordinary least squares, OLS)
- 慈祥圆滑的回归：岭回归 (Ridge Regression RR)

第二代回归方法：

- 割裂的两步走，人无远虑必有近忧：主元回归
Principal Component Regression (PCR)
- 两手都要抓，两手都要硬：偏最小二乘
(Partial Least Squares, PLS)

回归分析大家族

PCR



PLS



RR



OLS



第二代回归

新生代回归

第一代回归

回归分析模型的解法—最小二乘

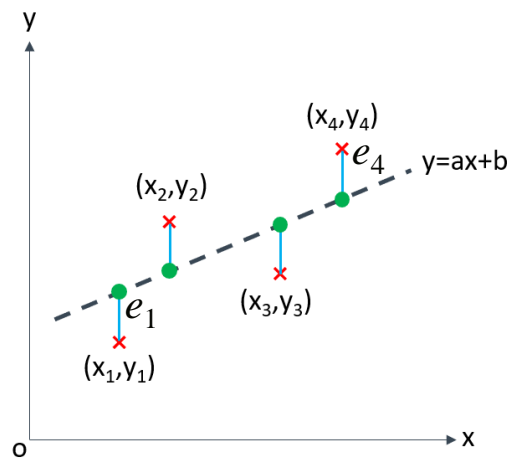
使模型拟合的好，就是找到回归系数 b 使残差矢量 e 的方差尽可能小，

$$e = y - Xb \longrightarrow \min(e'e)$$

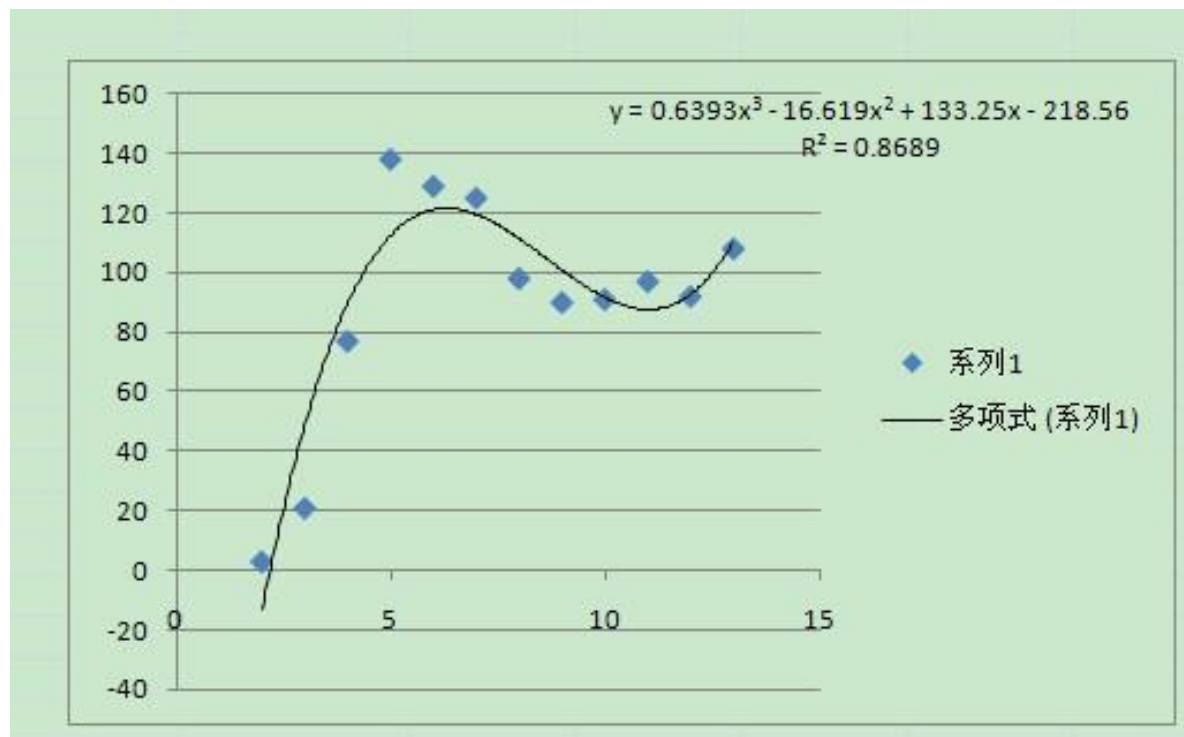
这就是我们所熟知的普通最小二乘法（ordinary least squares, OLS）。其解为：

$$b = (X'X)^{-1} X'y$$

推导过程？



一、数学模型



非线性关系



最小二乘的问题

□ 普通最小二乘的应用很广泛，因为在许多情况下该方法具有良好的性能。假若响应（即因变量）与自变量关系呈现线性，低噪声无共线性，则最小二乘方法是一种非常好的解法。

$$b = (X'X)^{-1} X'y$$

□ 但是，此种方法也有固有的缺点。当自变量间存在复共线性（multi-collinearity）时，回归系数估计的方差就很大，估计值就很不稳定。



最小二乘的问题

下面进一步用一个模拟的例子来说明这一点。

例2 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, 如下表的第 (2)、(3) 两行:

	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)	ε_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

x_1 , x_2 的样本相关系数 $r_{12}=0.986$, 表明 x_1 与 x_2 之间高度相关。



最小二乘的问题

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： $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

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

而原模型的参数

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

看来相差太大。

问题出在哪里？如何合理的解释？

x_1	1.1	1.4	1.7	1.7	1.8	1.8	1.9	2.0	2.3	2.4
x_2	1.1	1.5	1.8	1.7	1.9	1.8	1.8	2.1	2.4	2.5



回归分析模型的解法—岭回归

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

当自变量间存在复共线性时, $|X'X| \approx 0$,
我们设想给 $X'X$ 加上一个正常数矩阵 kI , ($k > 0$),
那么 $X'X + kI$ 接近奇异的程度就会比 $X'X$ 接近奇异的程度小得多。



回归分析模型的解法—岭回归

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

为 $\boldsymbol{\beta}$ 的岭回归估计，其中 k 称为岭参数。

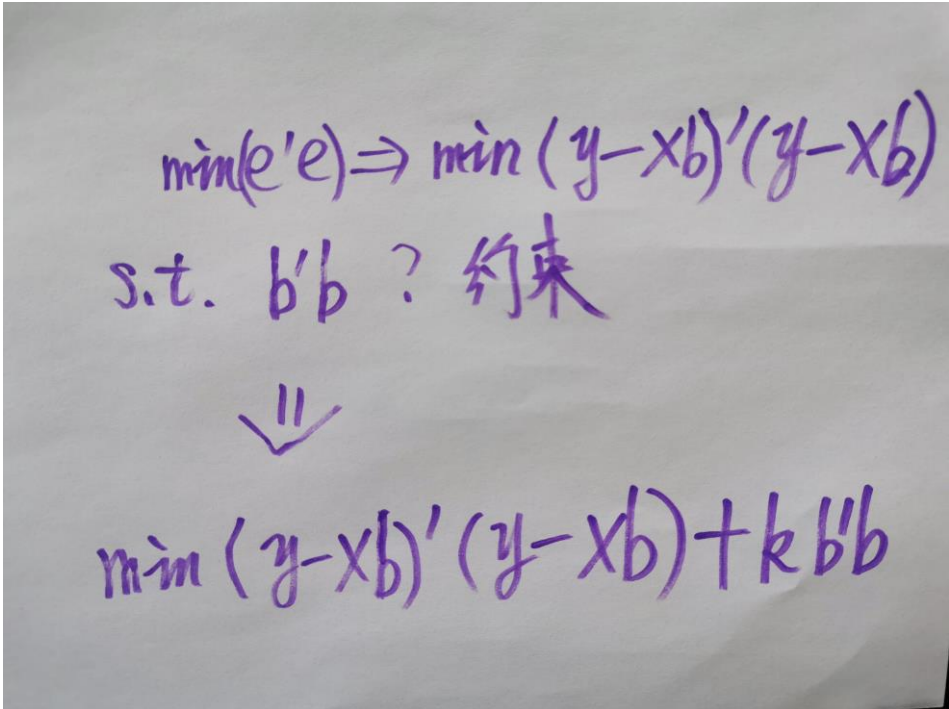
显然，岭回归做为 $\boldsymbol{\beta}$ 的估计应比最小二乘估计稳定，当 $k=0$ 时的岭回归估计就是普通的最小二乘估计。



岭回归 (Ridge regression)

问题:

岭回归参数计算公式的道理: $\hat{\beta}(k) = (X'X + kI)^{-1}X'y$


$$\begin{aligned} \min(e'e) &\Rightarrow \min(y - Xb)'(y - Xb) \\ \text{s.t. } b'b &? \text{ 约束} \\ \Downarrow \\ \min(y - Xb)'(y - Xb) + k b'b \end{aligned}$$



四、岭回归估计的定义

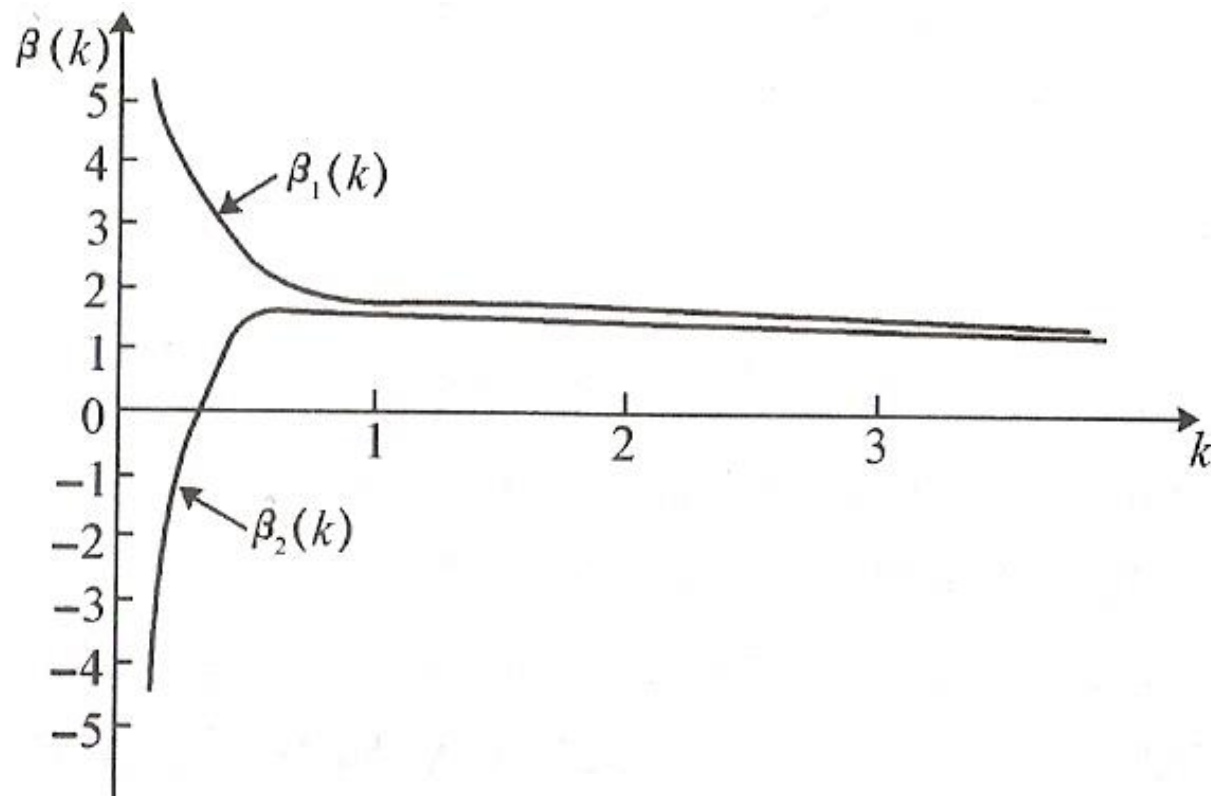
因为岭参数 k 不是唯一确定的，所以我们得到的岭回归估计 $\hat{\boldsymbol{\beta}}(k)$ 实际是回归参数 $\boldsymbol{\beta}$ 的一个估计族。

例如对之前的例子可以算得不同 k 值时的 $\hat{\beta}_1(k)$ ， $\hat{\beta}_2(k)$ ，见下表

表

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

四、岭回归估计的定义





五、岭回归估计的性质

性质 $\hat{\beta}^{(k)}$ 是回归参数 β 的有偏估计。

$$\begin{aligned}\text{证明: } E[\hat{\beta}^{(k)}] &= E[(X'X + kI)^{-1}X'y] \\ &= (X'X + kI)^{-1}X'E(y) \\ &= (X'X + kI)^{-1}X'X\beta\end{aligned}$$

显然只有当 $k=0$ 时, $E[\hat{\beta}^{(0)}] = \beta$; 当 $k \neq 0$ 时, $\hat{\beta}^{(k)}$ 是 β 的有偏估计。
要特别强调的是 $\hat{\beta}^{(k)}$ 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。



小结

- ◆ 岭回归分析实际上是一种改良的最小二乘法，是一种专门用于共线性数据分析的有偏估计回归方法。
- ◆ 当自变量间存在共线性时，解释变量相关的矩阵 $X'X$ 是奇异的，也就是说它的行列式的值接近于零，此时普通最小二乘（OLS）估计将失效。此时可采用岭回归估计。
- ◆ 岭回归就是用 $X'X+kI$ 代替正规方程中的 $X'X$ ，人为地把最小特征根提高，希望这样有助于降低均方误差。



小结

岭回归有什么问题？有何进一步思路？



主元（成分）回归



一、提出问题

国际旅游外汇收入是国民经济发展的重要组成部分，影响一个国家或地区旅游收入的因素包括自然、文化、社会、经济、交通等多方面的因素。《中国统计年鉴》把第三次产业划分为**12**个组成部分，分别为：



多重共线性

x1: 农林牧渔服务业

x2: 地质勘查水利管理业

x3: 交通运输仓储和邮电通讯业

x4: 批发零售贸易和餐食业

x5: 金融保险业

x6: 房地产业

x7: 社会服务业

x8: 卫生体育和社会福利业

x9: 教育文艺和广播

x10: 科学研究和综合艺术

x11: 党政机关

x12: 其他行业

选自**1998**年我国**31**个省、市、自治区的数据。以旅游外汇收入（百万美元）为因变量。自变量的单位为亿元人民币。数据略。



二、主成分回归方法

$$F_1 = u_{11}X_1 + u_{21}X_2 + \cdots + u_{p1}X_p$$

$$F_2 = u_{12}X_1 + u_{22}X_2 + \cdots + u_{p2}X_p$$

.....

$$F_p = u_{1p}X_1 + u_{2p}X_2 + \cdots + u_{pp}X_p$$

主成分回归: $Y_i^* = \gamma_1 F_{11} + \gamma_2 F_{12} + \cdots + \gamma_m F_{1m} + \varepsilon_i$

$$\sum_{i=1}^n \left[Y_i^* - \gamma_1 F_{i1} - \gamma_2 F_{i2} - \cdots - \gamma_m F_{im} \right]^2 = \min$$



原始数据观测矩阵

$$\mathbf{X}_0 = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

主成分系数矩阵

$$\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_p) = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix}$$



主成分得分矩阵

$$\mathbf{F} = \begin{bmatrix} F_{11} & F_{12} & \cdots & F_{1p} \\ F_{21} & F_{22} & \cdots & F_{2p} \\ \vdots & \vdots & & \vdots \\ F_{n1} & F_{n2} & \cdots & F_{np} \end{bmatrix}$$

$$\mathbf{F} = \mathbf{X}_0 \mathbf{U}$$

根据最小二乘估计，则

$$\hat{\mathbf{y}} = (\mathbf{F}'\mathbf{F})^{-1} \mathbf{F}'\mathbf{Y}$$



复习

- OLS, RR, PCR——都能表达成 $Y=XB+E$ 的形式
- 思路不同，解法不同，得到的回归系数不同
- 若数据不做中心化处理，有截距，如何解？



问题

主元回归建模方法的优点和缺点？

有什么思路、方法去改进？

变量选择、更有效的特征提取方法



变量选择



变量选择

什么是变量选择？

在回归模型中，选择最能够解释Y的解释变量的过程，称为变量选择。

- 在多元回归分析中，自变量的选择很重要。
- 遗漏了重要变量，回归分析的效果一定不好。
- 变量过多，会把对y影响不显著的变量也选入，影响回归方程的稳定性。



变量选择

- **常见的变量选择方法：** 前进法、后退法、逐步回归法、LASSO法

❖ 前进法和后退法的不足（自变量间相关时）：

⌘ 前者：只考虑引进，不考虑剔除

⌘ 后者：一旦剔除一棍子打死

- 逐步回归法：**有进有出（竞争上岗）**

什么是Lasso ?

L1范数

$$\hat{\beta}(\lambda) = \arg \min_{\beta \in R^p} \{ \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \},$$



$$(\hat{\alpha}, \hat{\beta}) = \arg \min \left\{ \sum_{i=1}^N \left(y_i - \alpha - \sum_j \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to } \sum_j |\beta_j| \leq t.$$

什么是Lasso ?

Lasso是最小二乘的一个改进

核心是加入了惩罚项 **L1范数**

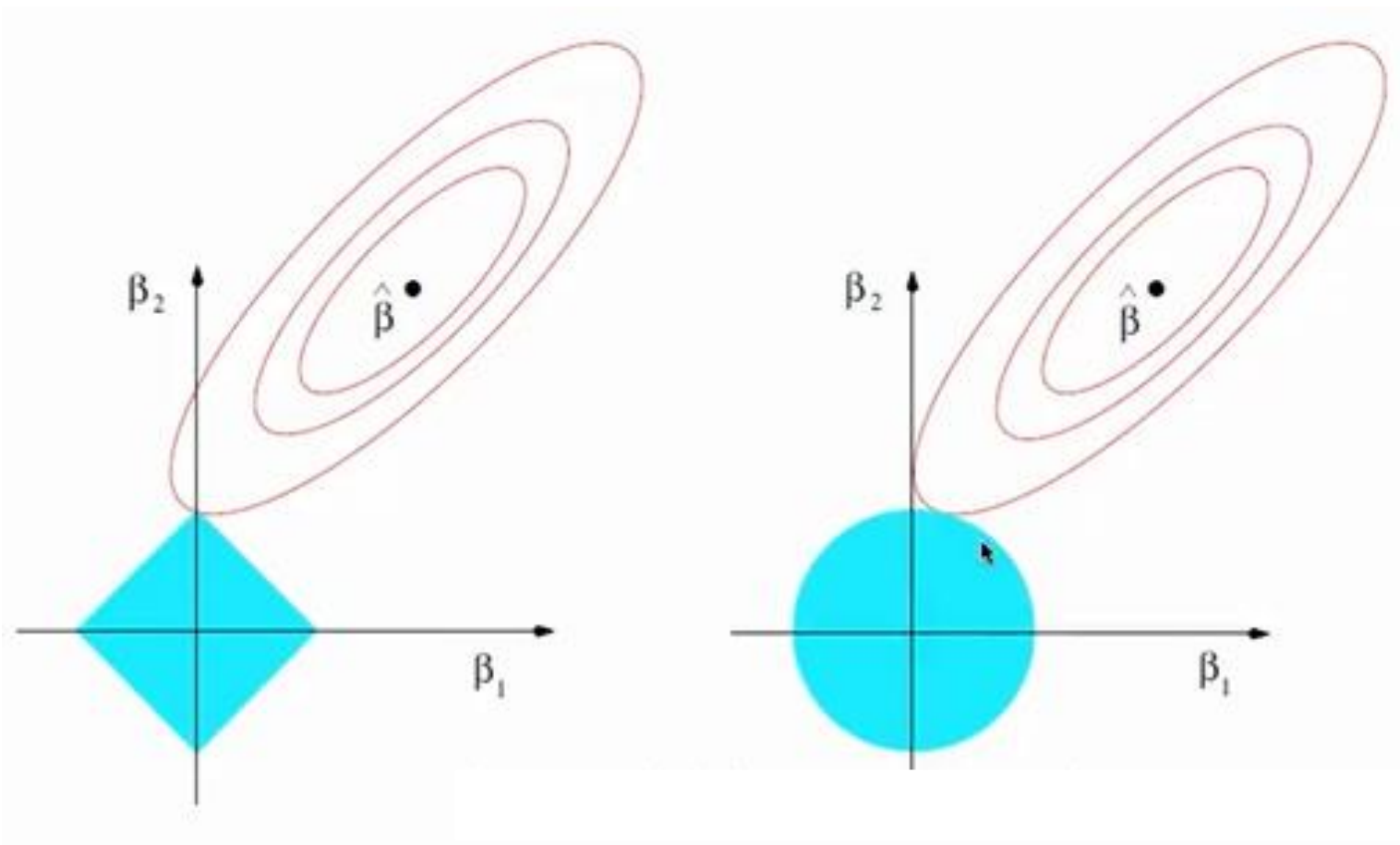
效果是变量选择

开创了一个近二十年的领域

喂饱了不少统计学家

$$(\hat{\alpha}, \hat{\beta}) = \arg \min \left\{ \sum_{i=1}^N \left(y_i - \alpha - \sum_j \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to } \sum_j |\beta_j| \leq t.$$

LASSO与岭回归





弹性网 (elastic net)

- **ElasticNet** 是一种使用L1和L2范数的线性回归模型。这种组合可以得到**只有很少的权重非零的稀疏模型**，但是又能保持正则化属性。
- 当多个特征和另一个特征相关的时候弹性网络非常有用。
Lasso 倾向于**随机选择**其中一个，而弹性网络更倾向于选择多个。

$$\widehat{\beta(\lambda)} = \arg \min_{\beta} \left\{ \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p (\alpha \beta_j^2 + (1 - \alpha) |\beta_j|) \right\}$$



更有效的特征提取

更有效的特征提取

- 割裂的两步走：人无远虑必有近忧（主元回归PCR）
- 两手都要抓，两手都要硬（偏最小二乘PLS）



这一路走来有多不容易
只有自己知道

