

智能制造过程大数据技术

Big Data Technology in Intelligent Manufacturing Process

第七讲：回归分析 Lecture 7: Regression

丁敏 dingmin@cug.edu.cn



中国地质大学(武汉) 自动化学院
School of Automation, China University of Geosciences

- 回归分析的基本概念
- 线性回归
- 高维回归系数压缩
- 非线性回归
- 回归模型的验证
- 本章小结



- 回归分析的基本概念
- 线性回归
- 高维回归系数压缩
- 非线性回归
- 回归模型的验证
- 本章小结



1 回归分析的基本概念

➤ 什么是回归？

- ❑ 起源：19世纪80年代，英国统计学家弗朗西斯.高尔顿提出
- ❑ 研究：父代身高与子代身高之间的关系
- ❑ 结论：子代的身高有向族群平均身高“回归”的趋势
- ❑ 回归模型：从输入变量到输出变量之间的映射函数（回归模型），等价于函数拟合

1 回归分析的基本概念

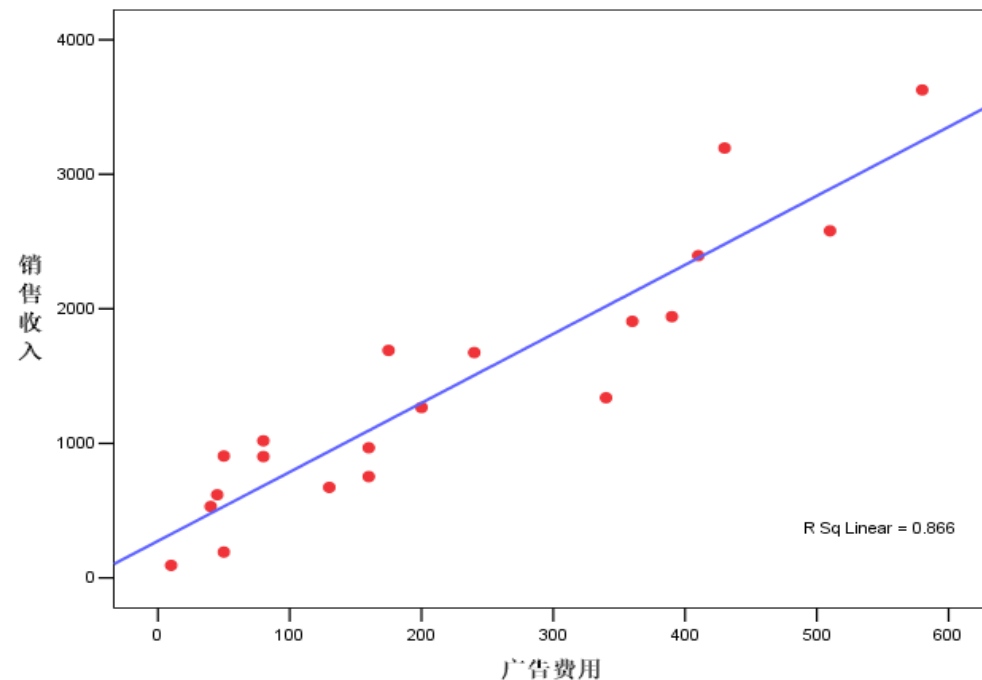
5

➤ 什么是回归问题

- 例：为研究销售收入与广告费用支出之间的关系，某医药管理部门随机抽取20家药品生产企业，得到它们的年销售收入和广告费用支出(万元)的数据如下。绘制散点图描述销售收入与广告费用之间的关系。

	A	B	C	D	E	F
1	企业编号	销售收入	广告费用	企业编号	销售收入	广告费用
2	1	618	45	11	531	40
3	2	3195	430	12	1691	175
4	3	1675	240	13	2580	510
5	4	753	160	14	93	10
6	5	1942	390	15	192	50
7	6	1019	80	16	1339	340
8	7	906	50	17	3627	580
9	8	673	130	18	902	80
10	9	2395	410	19	1907	360
11	10	1267	200	20	967	160

一元线性回归



1 回归分析的基本概念

6

➤ 数学描述—线性模型？

□ 样本： $x \in R^n$, $x = [x_1, x_2, \dots, x_n]^T$

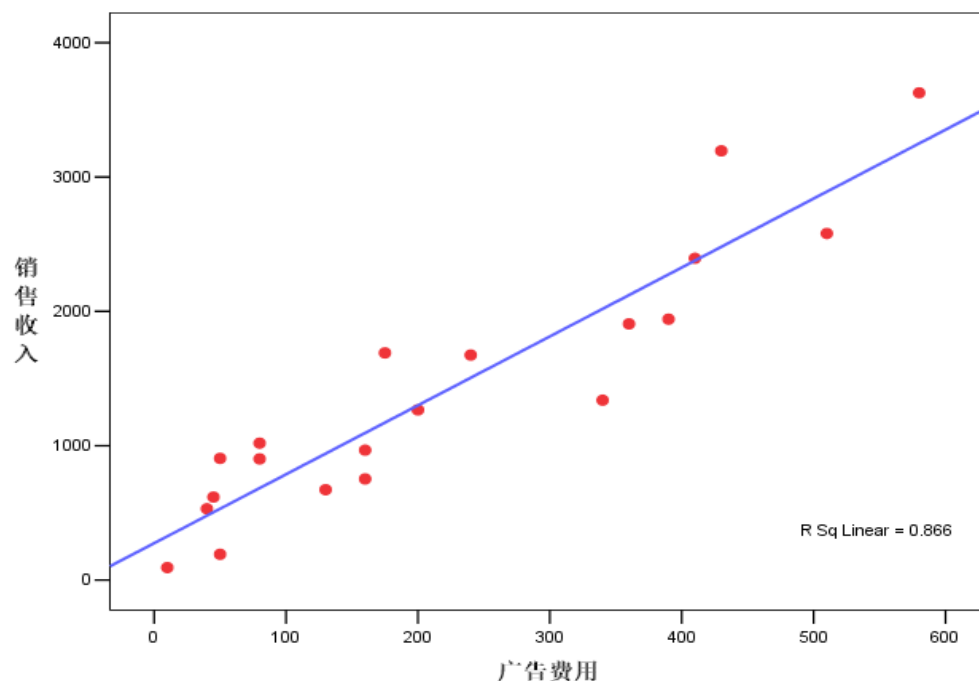
□ 参数： $\beta = [\beta_1, \beta_2, \dots, \beta_n]^T \in R^n, b \in R$

□ 线性模型： $f(x) = \beta^T x + b$

令： $x = [x_1, x_2, \dots, x_n, 1]^T \in R^{n+1}$

$\beta = [\beta_1, \beta_2, \dots, \beta_n, b]^T \in R^{n+1}$

线性模型可以简写为： $f(x) = \beta^T x$



典型的线性模型

➤ 基本概念

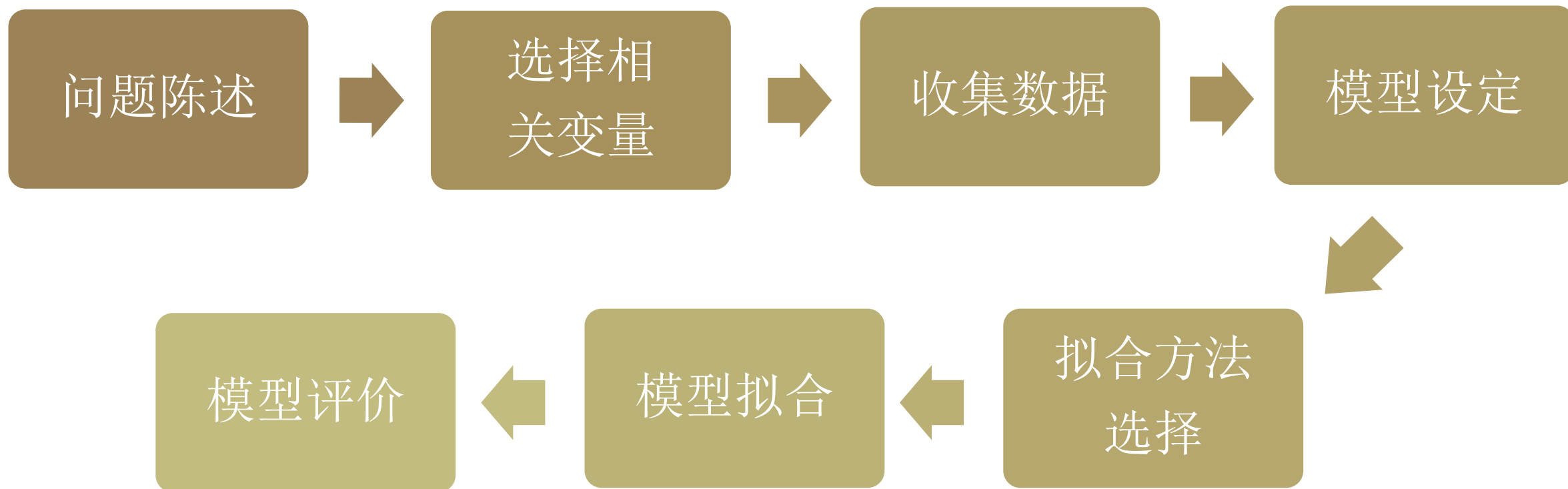
- 研究的是因变量和自变量之间的关系
- 用于预测分析，时间序列模型中发现变量之间的因果关系
- 按照输入变量个数可分为一元回归和多元回归，按照输入-输出对应关系可分为线性回归和非线性回归
- 常见的工业应用场景：
 - 预测类：工厂用电负荷预测、焦炉煤气产生量预测、余热回收量预测等
 - PHM¹类：故障预警、寿命预测等

注1:设备故障诊断与健康管理的英文缩写 (Prognostics & Health Management, PHM)

1 回归分析的基本概念

8

➤ 基本步骤



1 回归分析的基本概念

➤ 基本步骤

□ 问题陈述：确定需要分析研究哪些问题

- ✓ 例：计算未来高炉煤气利用率。应该考虑影响该利用率的相关变量的历史数据。如其他变量未发生波动，某个相关变量增大会直接导致高炉煤气利用率波动，则应该把该变量设为预测变量，高炉煤气利用率作为响应变量。

□ 选择相关变量：根据该研究领域 **专业人士的意见** 或者 **关联分析** 和 **因果分析** 等数据分析的方法选择变量集合

- ✓ 例：在高炉煤气调节中，冷风流量、冷风压力、热风压力、富氧流量、富氧压力、喷煤量、边缘矿焦比3、边缘矿焦比4、中心焦比7、中心焦比11等变量，对高炉煤气利用率 **有着较大的关系**，则选择这些操作变量作为相关变量

➤ 基本步骤

□ 收集数据：从实际中收集分析问题使用的数据

- ✓ 在每种情况下，我们收集到 n 个目标的观测数据。每个目标的观测数据都是对该目标所有潜在的相关变量的测量值

观测序号	响应变量	预测变量			
	Y	X_1	X_2	\cdots	X_p
1	y_1	x_{11}	x_{12}	\cdots	x_{1p}
2	y_2	x_{21}	x_{22}	\cdots	x_{2p}
3	y_3	x_{31}	x_{32}	\cdots	x_{3p}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
n	y_n	x_{n1}	x_{n2}	\cdots	x_{np}

➤ 基本步骤

□ 模型设定：通过模型将响应变量和预测变量联系起来

■ 线性函数： $Y = \beta_0 + \beta_1 X_1 + \varepsilon$

■ 非线性函数： $Y = \beta_0 + e^{\beta_1 X_1} + \varepsilon$

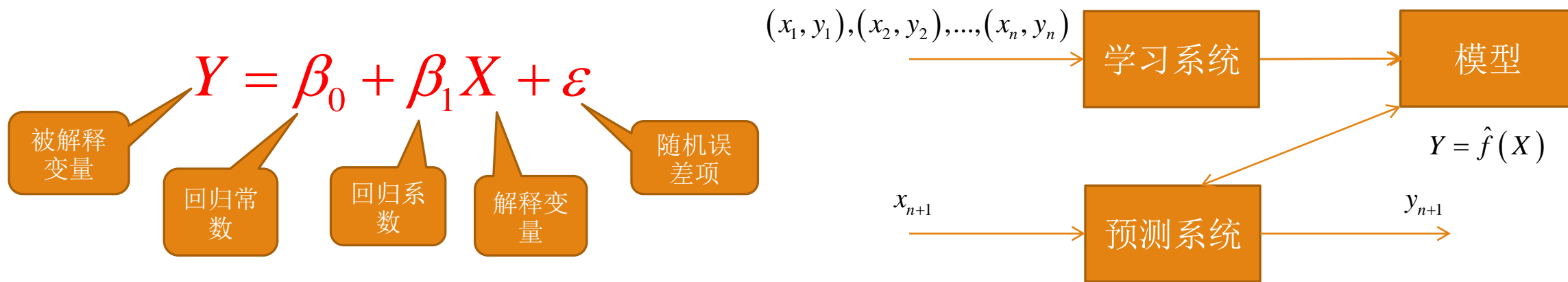
✓ 单变量回归 ：只有一个响应变量的回归分析	✓ 简单回归 ：只有一个预测变量的回归分析
✓ 多变量回归 ：有两个或两个以上响应变量的回归分析	✓ 多元回归 ：有两个或两个以上预测变量的回归分析

- 回归分析的基本概念
- 线性回归
- 高维回归系数压缩
- 非线性回归
- 回归模型的验证
- 本章小结



➤ 线性回归模型与线性回归分析

- 有训练样本： $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
- 其中： $x_i \in X, y_i \in Y, i = 1, 2, \dots, n$
- **线性回归模型**：学习一个线性模型 $Y = f(X)$ ，使得 $f(X)$ 与 Y 尽可能接近，则线性模型 $f(X)$ 称为线性回归模型。
- 线性回归分析：利用线性回归模型来对自变量（可以是向量）和因变量之间关系进行建模的过程，称为线性回归分析。



➤ 一元回归问题描述

□ 假设单个预测变量 X 与响应 Y 相关，则表示为

$$Y \approx b_0 + b_1 X$$

□ 利用一元线性回归则可以从数据中估计截距和斜率，然后用模型预测某一特定值 x 的响应：

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

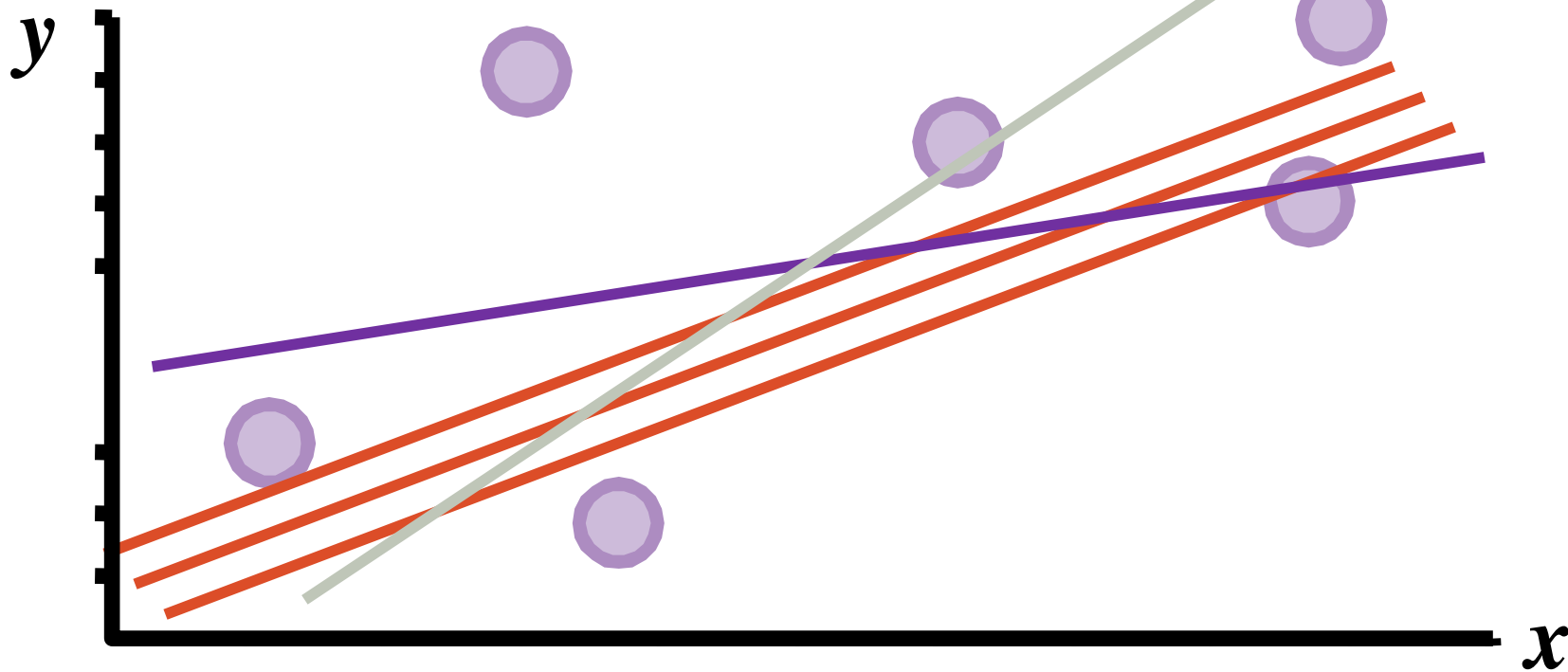
□ \hat{y} 表示对 $X = x$ 时的预测值

➤ 难点

- 是否能找到一条直线穿过所有数据点？
- 如何评估哪一条线的拟合效果最佳？

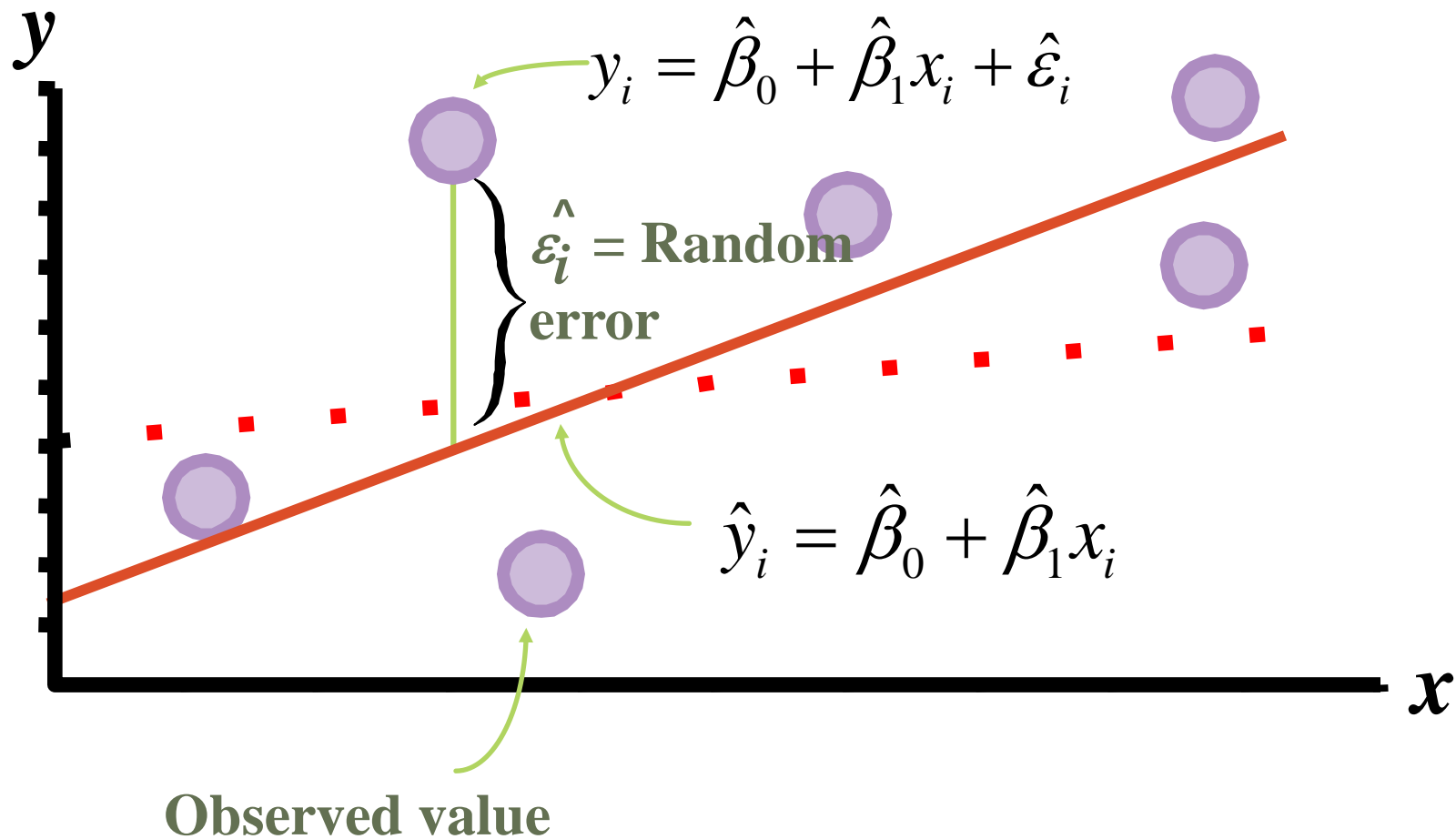


- 引用**误差平方和**（the Sum of the Squared Differences, SSE），使其最小则认为当前最佳



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

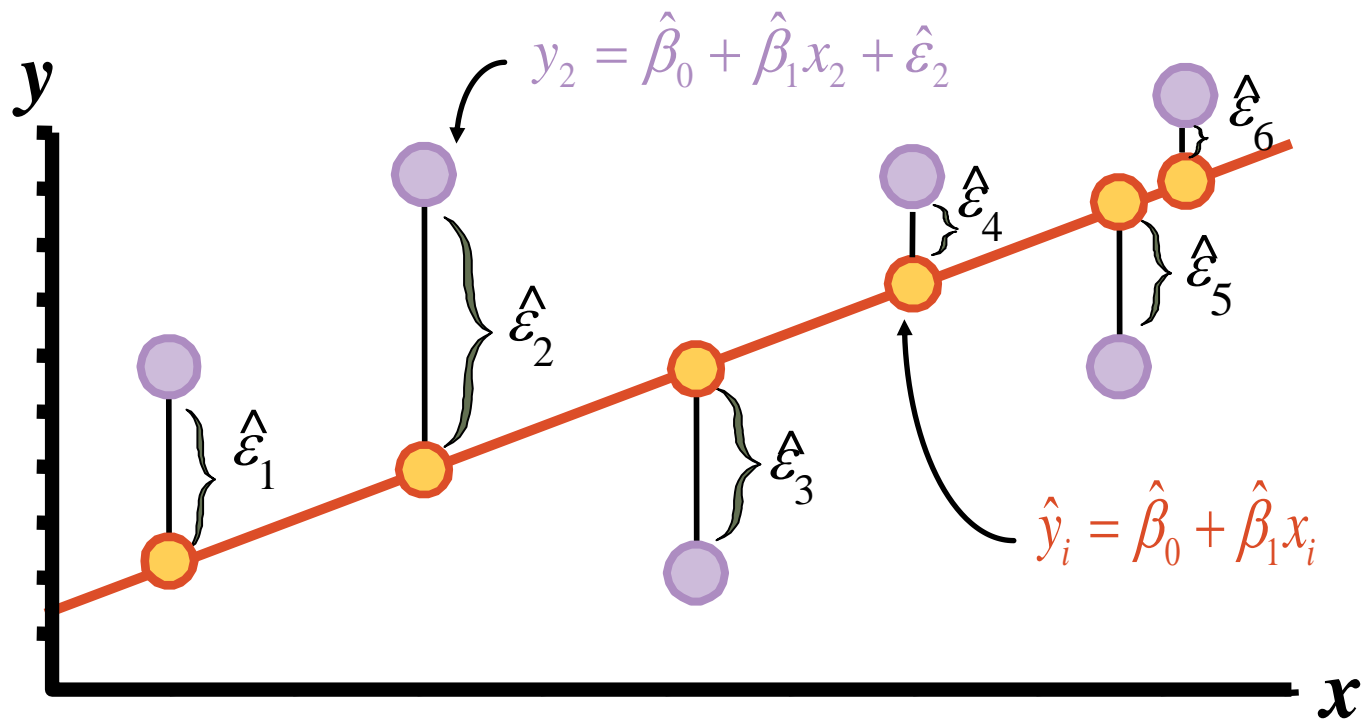
➤ 回归拟合



➤ 参数的选择

□ 使SSE最小化时，当前即为最佳拟合曲线，选择当前参数 $\hat{\beta}_0, \hat{\beta}_1$

$$\sum_{i=1}^n \hat{\varepsilon}_i^2 = \hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \hat{\varepsilon}_3^2 + \hat{\varepsilon}_4^2 + \hat{\varepsilon}_5^2 + \hat{\varepsilon}_6^2$$



➤ 最小二乘估计

□ 条件： x_1, x_2, \dots, x_p 给定的 y 的条件分布是**正态分布**，其中均值为 $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$ ，方差为 σ^2

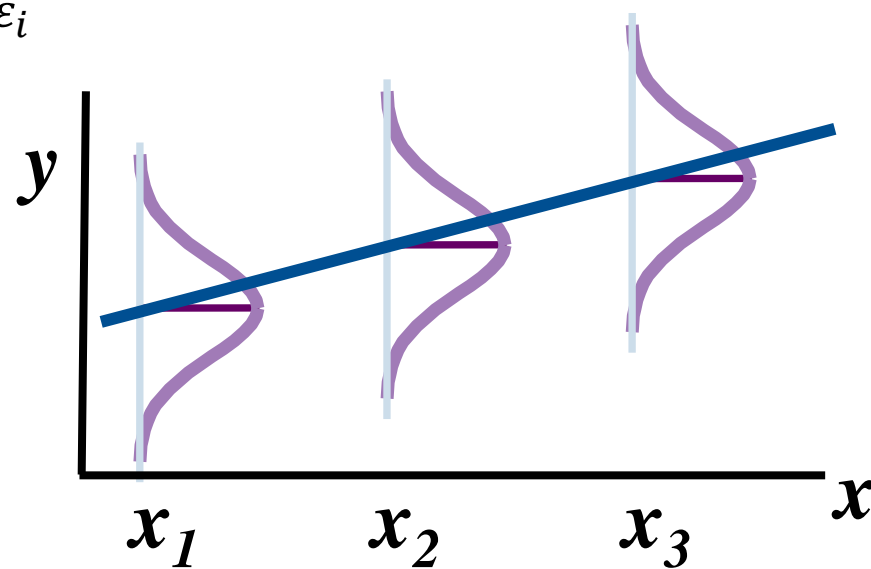
□ 模型

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + \varepsilon_i$$

□ 最小二乘函数

$$S(\beta_0, \beta_1, \dots, \beta_p) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2$$

必须对函数 S 做关于 $\beta_0, \beta_1, \dots, \beta_p$ 的最小化



➤ 最小二乘估计

□ 最小二乘正规方程为

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\text{式中, } \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

其中, 正规方程的解将是最小二乘估计量 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$

使用矩阵记号有利于表达正规方程对处理多元回归模型

➤ 最小二乘估计

□ 最小二乘法估计量向量 $\hat{\beta}$ ，最小化值为

$$S(\beta) = \sum_{i=1}^n \varepsilon_i^2 = \varepsilon' \varepsilon = (y - X\beta)'(y - X\beta) \xrightarrow{\text{满足}} \left. \frac{\partial S}{\partial \beta} \right|_{\hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0$$

↓ 简化

$$X'y = X'X\hat{\beta} \xrightarrow{\text{简化}} \hat{\beta} = (X'X)^{-1}X'y$$

□ 最终拟合的最小二乘回归方程

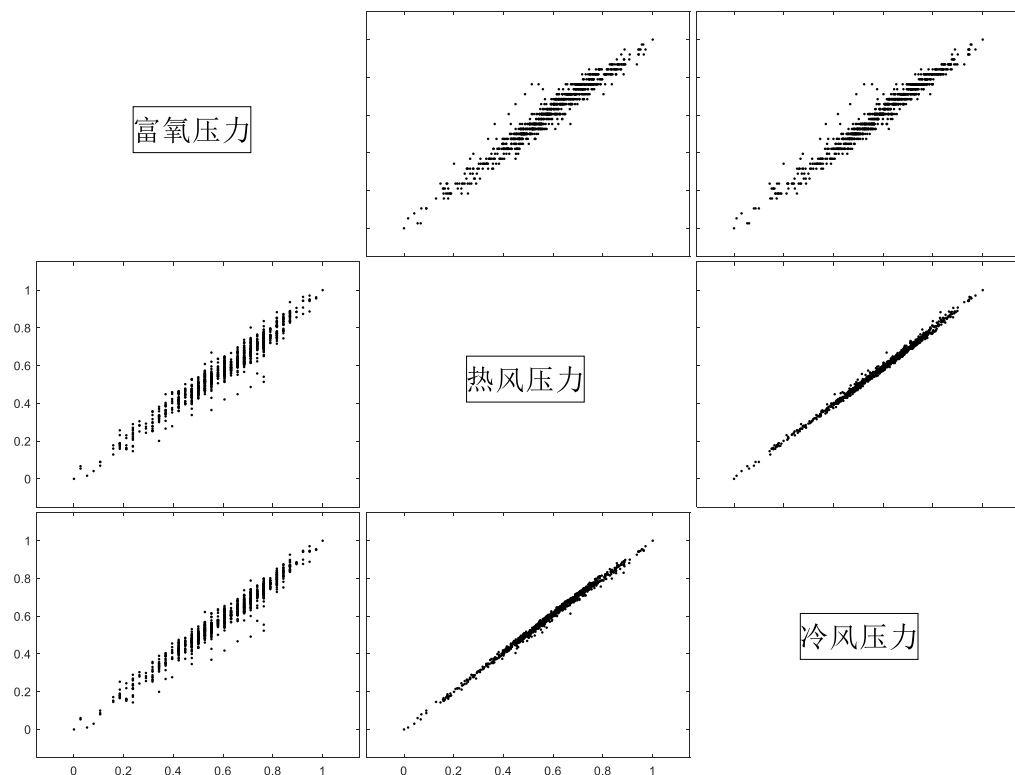
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p \xrightarrow{\text{简化}} y_i \text{ 的拟合值: } \hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_1$$

ε 的方差 σ^2 的无偏估计: $\hat{\sigma}^2 = \frac{S(\beta)}{n-p-1}$

自由度 (df) 残差平方和SSE

➤ 最小二乘估计实例

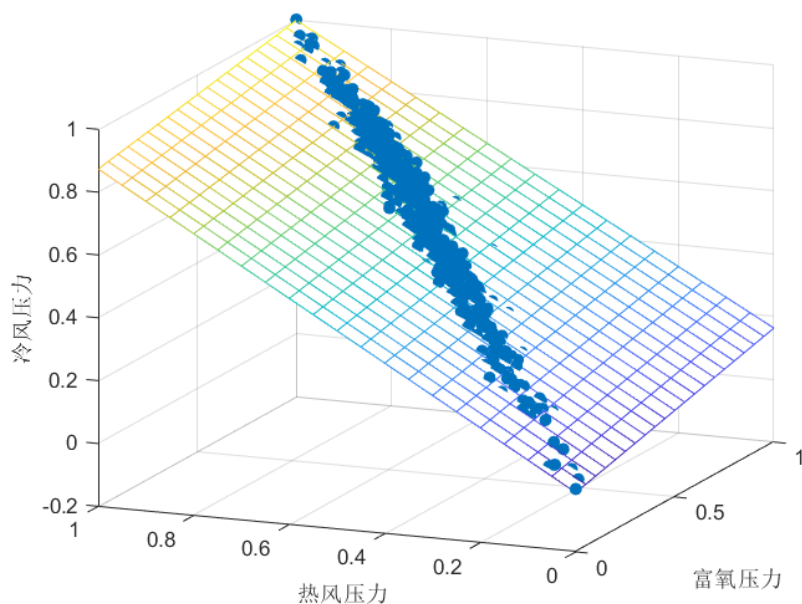
□ 用多元线性模型 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ 拟合数据。富氧压力和热风压力作为预测变量，冷风压力作为响应变量。



- ✓ 左图为二维散点图的二维阵列，其中（除对角线外）每个图框包含一个散点图
- ✓ 每个散点图可以观察到富氧压力、热风压力与冷风压力两两之间成线性的关系

➤ 最小二乘估计实例

- 三维散点图表明该多元线性回归模型可以为以富氧压力和热风压力作为预测变量，以冷风压力作为响应变量提供合理的数据拟合。



总体均方根误差为RMSE=0.0513

当前多元线性回归模型的拟合效果较为优秀

- 拟合多元回归模型

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} \\ \vdots & \vdots & \vdots \\ 1 & x_{11000} & x_{21000} \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_{1000} \end{bmatrix}$$

- β 的最小二乘估计量为

$$\hat{\beta} = (X'X)^{-1}X'y \quad \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 0.0024 \\ 0.1543 \\ 0.8491 \end{bmatrix}$$

- 最小二乘拟合为

$$\hat{y} = 0.0024 + 0.1543x_1 + 0.8491x_2$$

➤ 加权最小二乘估计

□ 当原始最小二乘不满足方差相等的假设时——**加权最小二乘**

- 当误差 ε 不相关但方差不相等， ε 的协方差矩阵为

$$\sigma^2 V = \sigma^2 \begin{bmatrix} \frac{1}{w_1} & & & 0 \\ & \frac{1}{w_2} & & \\ & & \ddots & \\ 0 & & & \frac{1}{w_n} \end{bmatrix}$$

$$\xrightarrow{\text{令 } W = V^{-1}}$$

最小二乘正规方程

$$(X'WX)\hat{\beta} = X'Wy$$

最小二乘估计量

$$\hat{\beta} = (X'WX)^{-1}X'Wy$$

方差较大的观测值将比方差较小的观测值有更小的权重

➤ 加权最小二乘估计

□ 变换后的数据集

$$B = \begin{bmatrix} 1\sqrt{w_1} & x_{11}\sqrt{w_1} & \cdots & x_{1k}\sqrt{w_1} \\ 1\sqrt{w_2} & x_{21}\sqrt{w_2} & \cdots & x_{2k}\sqrt{w_2} \\ \vdots & \vdots & & \vdots \\ 1\sqrt{w_n} & x_{n1}\sqrt{w_n} & \cdots & x_{nk}\sqrt{w_n} \end{bmatrix}, \begin{bmatrix} y_1\sqrt{w_1} \\ y_2\sqrt{w_2} \\ \vdots \\ y_n\sqrt{w_n} \end{bmatrix}$$



做最小二乘

β 的加权最小
二乘估计量

$$\hat{\beta} = (B'B)^{-1}B'z = (X'W'X)^{-1}X'W'y$$

加权最小二乘法是对原模型进行**加权**，使之成为一个新的不存在异方差性的模型

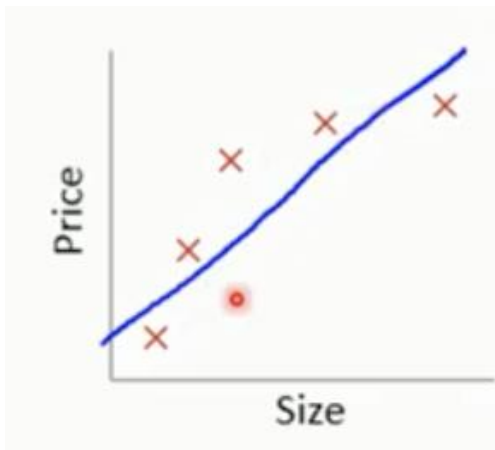
- 回归分析的基本概念
- 线性回归
- **高维回归系数压缩**
- 非线性回归
- 回归模型的验证
- 本章小结



➤ 回归中的拟合效果

□ 欠拟合：

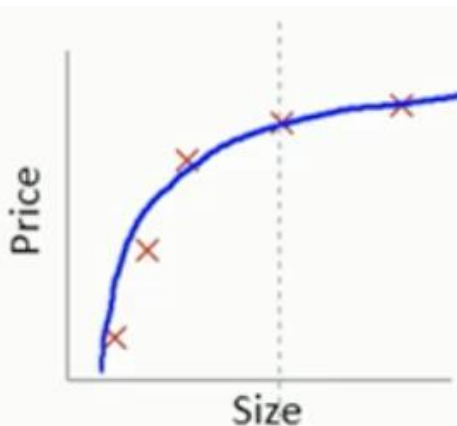
模型过于简单，对训练数据拟合效果差，模型偏差大



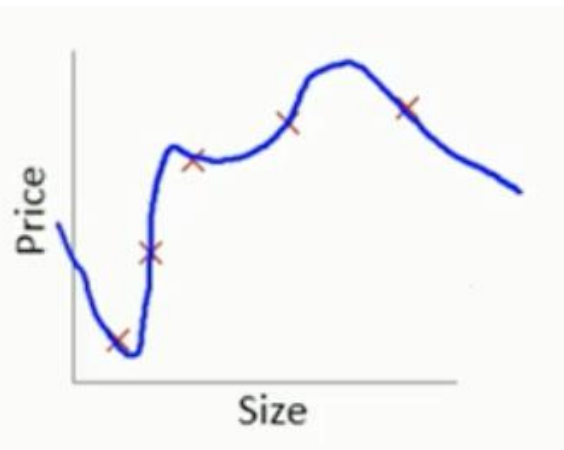
欠拟合

□ 过拟合：

模型过于复杂，对训练数据的拟合过于充分，模型方差大



正确拟合



过拟合

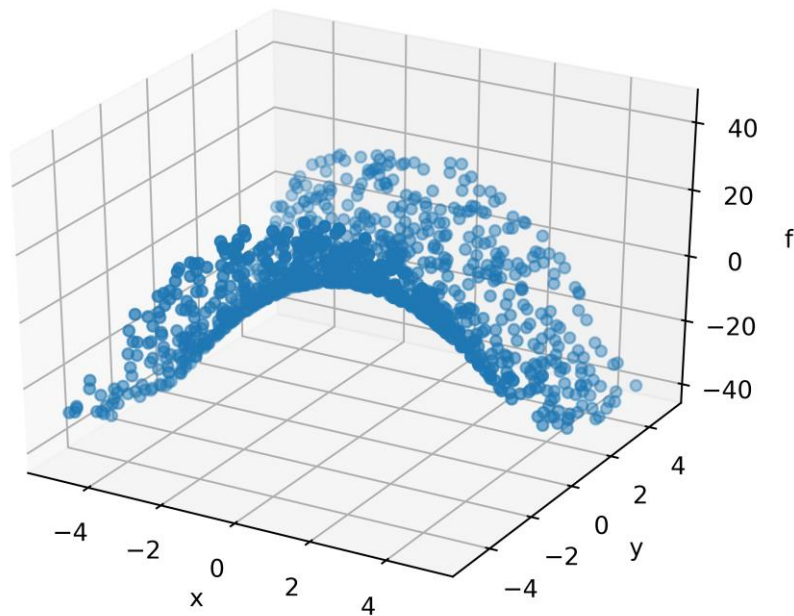
➤ 高维回归系数压缩

□ 问题：

- ✓ 很多数据存在**强耦合**等问题
- ✓ 多元线性回归模型中，若将全部数据用于回归分析，不仅导致问题难度增加，也容易造成过拟合使测试数据误差方差过大

□ 解决方案——简化模型

- ✓ 特征筛选
- ✓ 特征压缩



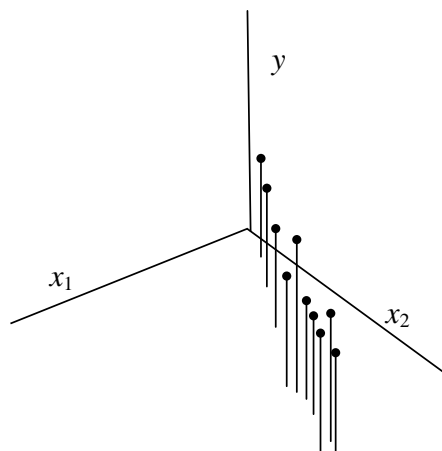
➤ 共线性的来源及影响

□ 多重共线性来源：

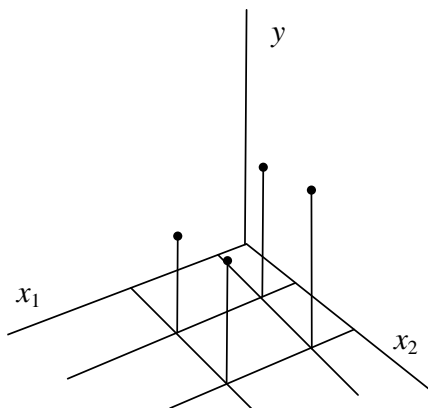
- ✓ 自变量间存在同方向的变化趋势
- ✓ 自变量之间存在着密切的关联度
- ✓ 模型中引入滞后变量也容易产生多重共线性
- ✓ 建模过程中，由于自变量选择不当，引起了变量之间的多重共线性

➤ 共线性的来源及影响

□ 多重共线性对最小二乘估计量的影响：



a) 存在多重共线性的数据集



b) 正交的回归变量

左图：平面将非常不稳定，对数据点相当小的变化敏感

右图：正交回归变数，点所拟合的平面将更为稳定

- 假设只存在两个回归变量，假设 x_1 、 x_2 与 y 都已经尺度化为单位长度。

- 模型： $y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

- 最小二乘正规方程： $(X'X)\hat{\beta} = X'y$

$$\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix}$$

$$C = (X'X)^{-1} = \begin{bmatrix} \frac{1}{1-r_{12}^2} & \frac{-r_{12}}{1-r_{12}^2} \\ \frac{-r_{12}}{1-r_{12}^2} & \frac{1}{1-r_{12}^2} \end{bmatrix}$$

$$\hat{\beta}_1 = \frac{r_{1y} - r_{12}r_{2y}}{1 - r_{12}^2}, \hat{\beta}_2 = \frac{r_{2y} - r_{12}r_{1y}}{1 - r_{12}^2}$$

x_1 与 x_2 之间强烈的多重共线性会使得回归系数的最小二乘估计具有较大的方差

➤ 多重共线性问题实例

□ 假设已知 x_1 , x_2 与 y 的关系服从线性回归模型 $y = 10 + 2x_1 + 3x_2 + \varepsilon$

序号	1	2	3	4	5	6	7	8	9	10
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
ε	0.8	-0.5	0.4	-0.5	0.2	1.9	1.9	0.6	-1.5	-1.5
y	16.3	16.8	19.2	18.0	19.5	20.9	21.1	20.9	20.3	22.0

- 假设回归系数与误差项未知，采用最小二乘估计来求回归系数


	估计值	原模型
β_0	11.292	10
β_1	11.307	2
β_2	-6.591	3

求得： x_1 , x_2 的相关系数为0.986

x_1 与 x_2 之间高度相关，即存在共线性问题，
使得最小二乘估计不能满足当前模型

➤ 岭回归

- ❑ 存在问题：最小二乘法应用于非正交数据时，可能会得出非常不良的回归系数估计值
- ❑ 岭估计量：回归系数的有偏估计量


$$(X'X + \lambda I)\hat{\beta}_{\text{岭}} = X'y$$

➤ 岭回归

- 当变量 $|X'X| \approx 0$ ，即

调整 λ 的值，可保证矩阵永远满秩，即永远存在矩阵的逆

$$X'X = \begin{vmatrix} \partial_{11} & \partial_{12} & \cdots & \partial_{1n-1} & \partial_{1n} \\ 0 & \partial_{22} & \cdots & \partial_{2n-1} & \partial_{2n} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & \partial_{n-1n-1} & \partial_{n-1n} \\ 0 & 0 & \cdots & 0 & 0 \end{vmatrix} \xrightarrow{\text{加上}\lambda I} X'X + \lambda I = \begin{vmatrix} \partial_{11} + \lambda & \partial_{12} & \cdots & \partial_{1n-1} & \partial_{1n} \\ 0 & \partial_{22} + \lambda & \cdots & \partial_{2n-1} & \partial_{2n} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & \partial_{n-1n-1} + \lambda & \partial_{n-1n} \\ 0 & 0 & \cdots & 0 & \lambda \end{vmatrix}$$

- 当 $\lambda \neq 0$ 时， $(X'X + \lambda I)$ 可逆，故 $\hat{\beta}_{\text{岭}} = (X'X + \lambda I)^{-1}X'y$

当 $\lambda=0$ 时，岭估计量是最小二乘估计量

- 由于 $\hat{\beta}_{\text{岭}} = (X'X + \lambda I)^{-1}X'y = (X'X + \lambda I)^{-1}(X'X)\hat{\beta} = Z_k\hat{\beta}$ 故，岭估计量是最小二乘估计量的线性变换

➤ 岭回归

- 最小二乘的表达式 $\hat{\beta} = (X'X)^{-1}X'y$ ，假设矩阵 X 是列正交的，即 $X'X = I$ 。有

$$\hat{\beta}_j^{ridge} = \frac{\hat{\beta}_j}{1 + \lambda}$$

- ✓ 岭回归将最小二乘的系数缩小
- ✓ 参数 λ 可以称为偏倚参数，当 λ 离开 0 时，估计的偏倚量增加

最小二乘估计的总方差受到小的特征值的影响是很大的

➤ LASSO回归法

- ❑ 岭回归存在问题：拟合的系数都非零，即岭回归达不到变量选择的目的
- ❑ LASSO回归：通过一阶惩罚项，能将一些系数恰好压缩为零，实现变量选择

● LASSO回归损失函数

$$\begin{aligned} \min_{\hat{\beta}_{\text{LASSO}}} \|X\hat{\beta}_{\text{LASSO}} - y\|_2^2 + \alpha \|\hat{\beta}_{\text{LASSO}}\|_1 & \xrightarrow{\text{求导}} \frac{\partial (RSS + \alpha \|\hat{\beta}_{\text{LASSO}}\|_1)}{\partial \hat{\beta}_{\text{LASSO}}} = \frac{\partial (\|y - X\hat{\beta}_{\text{LASSO}}\|_2^2 + \alpha \|\hat{\beta}_{\text{LASSO}}\|_1)}{\partial w} \\ & = \frac{\partial (y - X\hat{\beta}_{\text{LASSO}})'(y - X\hat{\beta}_{\text{LASSO}})}{\partial \hat{\beta}_{\text{LASSO}}} + \frac{\partial \alpha \|\hat{\beta}_{\text{LASSO}}\|_1}{\partial \hat{\beta}_{\text{LASSO}}} \\ & = 0 - 2X'y + 2X'X\hat{\beta}_{\text{LASSO}} + 2\alpha \end{aligned}$$

- ✓ LASSO无法解决（岭回归可以解决）特征间精确相关关系导致的最小二乘无法使用的问题，即 $X'X$ 不满秩

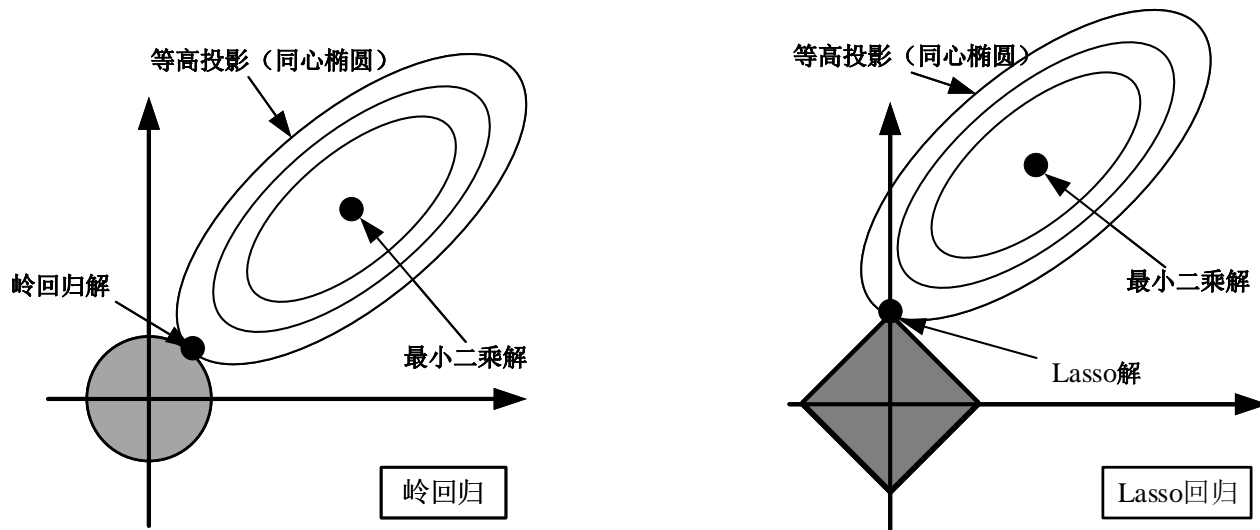
整理

$$X'X\hat{\beta}_{\text{LASSO}} = X'y - \frac{\alpha I}{2}$$

➤ LASSO回归法

● 假设 $X'X$ 的逆存在, 则有 $\hat{\beta}_{\text{LASSO}} = (X'X)^{-1} \left(X'y - \frac{\alpha I}{2} \right)$

- ✓ 增大 α , 可以为 $\hat{\beta}_{\text{LASSO}}$ 的计算增加一个负项, 限制参数估计中 $\hat{\beta}_{\text{LASSO}}$ 的大小从而防止多重共线性引起的参数被估计过大导致模型失准的问题
- ✓ **LASSO并不是从根本上解决多重共线性问题, 而是限制多重共线性带来的影响**
- ✓ L1正则化 (LASSO回归) **主导稀疏性**, 会将系数压缩到0



假设截距为0且 $\beta = (\beta_1, \beta_2)^T$ 为二维

- ✓ 岭回归的L2范数通过降低各系数的绝对值而**防止过拟合、降低模型复杂度**
- ✓ Lasso回归的L1范数惩罚项易构造稀疏矩阵, 有**特征选择作用**, 同时也有**一定程度防止过拟合**的作用

关于岭回归估计，下列说法错误的是

- ☐ A 岭回归估计为了处理自变量之间存在的多重共线性问题而引入
- ☐ B 岭回归得到的参数估计是有偏的
- ☒ C 具有稀疏化、选择变量的能力
- ☐ D 可以用岭迹法选择合适的 λ

提交

➤ 主成分回归法

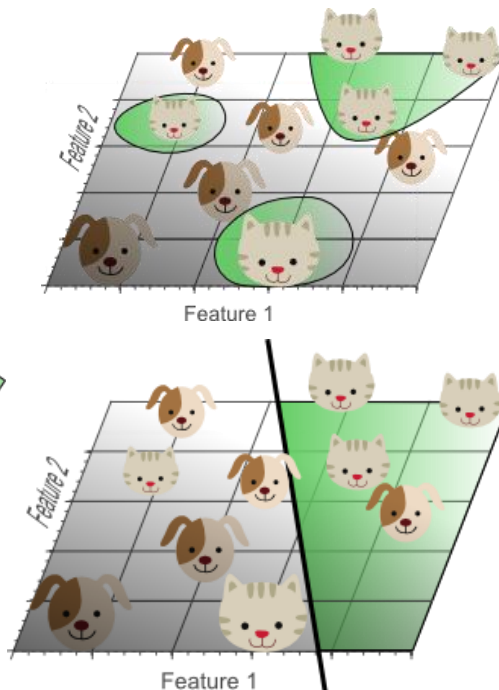
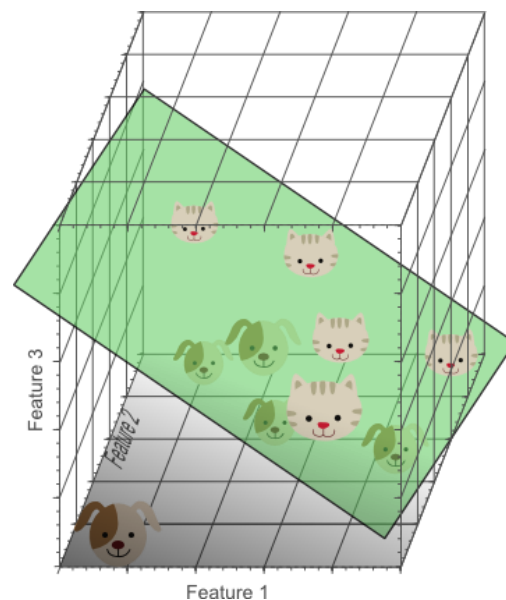
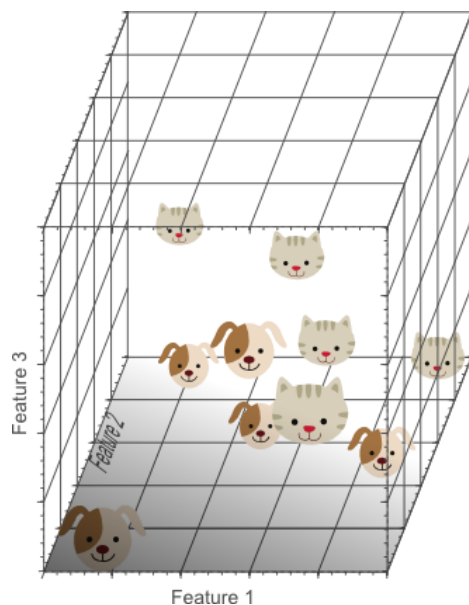
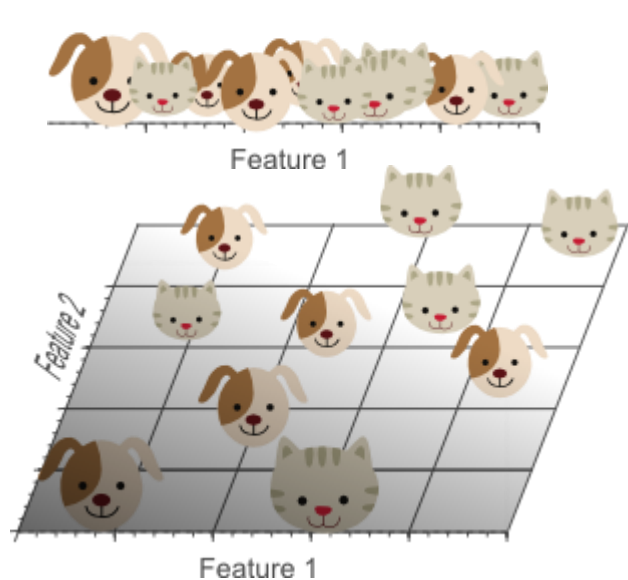
□ 主成分回归分析的背景

大数据时代回归分析面临的困难！

数据多重共线性

维数灾难

特征复杂多变



训练集上表现良好，但是对新数据缺乏泛化能力

使用更少的特征，对新数据具备更优的泛化能力

➤ 主成分回归

□ 利用线性回归公式： $\hat{\beta} = (X'X)^{-1}X'y$

可得： $y = 397 - 3.75x_1 + 5.13x_2$

根据系数可知：

年龄对血压有正作用（年纪越大，血压越高）

胆固醇对血压有反作用（胆固醇越高，血压越低）错误的结论

□ 从数据上的直接观测可以看出，血压会随着胆固醇的增加而增加

□ 产生错误的原因（1）数据量过少（2）两个参数 x_1 和 x_2 相关性过大

□ 主成分回归（PCA）：不仅可以解决变量的共线性问题，还可以有效提取数据的特征，降低数据的冗余

	Y 血压	x_1 胆固醇	x_2 年龄
	SBP	Chol	Age
1	120	126	38
2	125	128	40
3	130	128	42
4	121	130	42
5	135	130	44
6	140	132	46

➤ 主成分回归

- ❑ 为了解决上述问题，对观测数据 X 进行PCA降维
- ❑ 通过降维，可以把 X 中的样本投影到一系列正交矢量上得到一组新的观测数据 t ，且这些数据之间的相关性较小
- ❑ 利用新的数据 t ，再针对 y 进行线性回归

$$t = 0.589x_1 + 0.808x_2$$

PCA的系数

$$y = -83.9 + 1.932t$$

带入可得： $y = -83.9 + 1.14x_1 + 1.56x_2$

Y x_1 x_2
血压 胆固醇 年龄

	SBP	Chol	Age	PC1
1	120	126	38	105
2	125	128	40	108
3	130	128	42	109
4	121	130	42	111
5	135	130	44	112
6	140	132	46	115

➤ 主成分回归

□ PCA的系数，计算过程中只考虑了观测量X的分布

□ 其目的是找到X变化最广的成分

□ 在回归系统，关于数据有2点要求：

(1) 观测数据和预测数据，变化范围要大

(2) 观测数据和预测数据要有相关性

□ 为了兼顾上述两点，引入偏最小二乘回归（PLS）

PCA的系数

$$t = 0.589x_1 + 0.808x_2$$

$$y = -83.9 + 1.932t$$

PCA回归只考虑了其中1点

关于主元回归，下列说法正确的是

- ☒ A 主成分之间相互正交
- ☒ B 第一主成分是方差波动最大的方向
- ☐ C 降维后不会损失信息
- ☒ D 可以一定程度上解决多重共线性问题

提交

➤ 偏最小二乘回归

□ 为什么提出偏最小二乘回归？

案例：现在想要通过汽油的光谱分析结果，预测其辛烷含量。

一个经典的数据集是：有60个汽油样本，每个样本包含其光谱分析结果，即在401个不同波长处的光谱强度值，作为自变量（ 60×401 ）；以及其辛烷含量（ 60×1 ），作为因变量。

问题：自变量之间的多重相关性；样例很少，少于变量的维度。

如何解决？

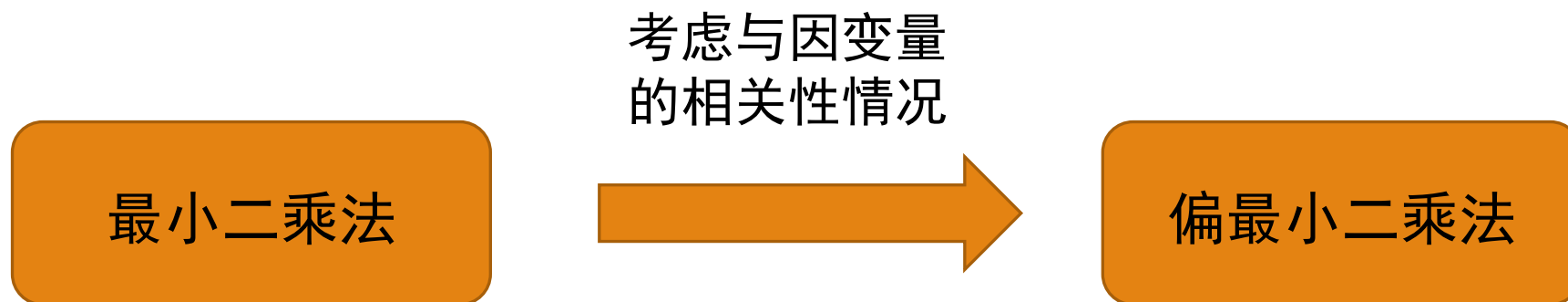
□ 偏最小二乘（PLS）：OLS+PCA+CCA 综合全面，经典融合

➤ 偏最小二乘回归

- PLS最先产生于化学领域，在利用分光镜来预测化学样本的组成时，作为解释变量的**红外区反射光谱的波长常有成百上千**，往往**超过化学样本的个数**，所造成的多重相关性使得人们很难利用传统的最小二乘法
- 基于这个应用的需要，1983年首次提出了PLS回归方法并首先在化工领域取得了广泛的应用

➤ 偏最小二乘回归

- 偏最小二乘是一种确定预测变量线性组合的技术
- 与主成分分析不同，PLS技术通过从预测变量和目标变量中依次提取因子，从而使提取因子之间的协方差最大化
- 将预测变量和观测变量投影到一个新空间，来寻找一个线性回归模型
- 作用：解决高维数据问题，即自变量的个数大于观测值的个数



➤ 偏最小二乘回归

□ 考虑 k 个因变量 y_1, y_2, \dots, y_k 与 p 个自变量 x_1, x_2, \dots, x_p 的建模问题。

□ 基本思路

- ✓ (1) 集中选择第一成分 u_1 (u_1 是 x_1, x_2, \dots, x_p 的线性组合, 且尽可能多地提取原自变量集中的变异信息)
- ✓ (2) 在因变量集中也提取第一成分 v_1 , 并要求 u_1 与 v_1 相关程度达到最大
- ✓ (3) 建立因变量 y_1, y_2, \dots, y_k 与 u_1 的回归
- ✓ (4) 如果回归方程已达到满意的精度, 则算法中止。否则继续第二对成分的提取, 直到能达到满意的精度为止
- ✓ (5) 最终建立 y_1, y_2, \dots, y_k 与 u_1, u_2, \dots, u_r (假设提取了 r 个成分) 的回归式

- 回归分析的基本概念
- 线性回归
- 高维回归系数压缩
- **非线性回归**
- 回归模型的验证
- 本章小结



➤ 非线性回归模型

□ 模型的非线性的形式

$$y = \theta_1 e^{\theta_2 x} + \varepsilon$$

参数 θ_1 与 θ_2 非线性的

$$y = f(\mathbf{x}, \boldsymbol{\theta}) + \varepsilon$$

$\boldsymbol{\theta}$ 为未知参数的 $p \times 1$ 向量

ε 为不相关的随机误差项

$E(\varepsilon) = 0$ 且 $Var(\varepsilon) = \sigma^2$

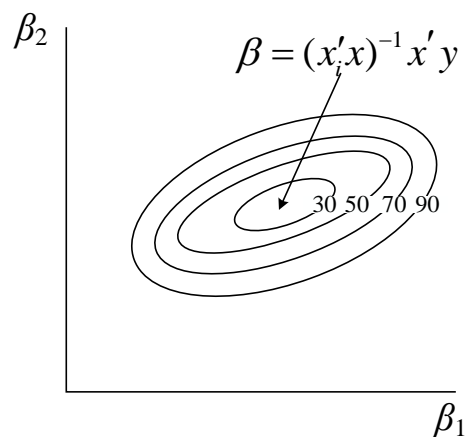
$$\frac{\partial f(\mathbf{x}, \boldsymbol{\theta})}{\partial \theta_1} = e^{\theta_2 x} \quad \text{和} \quad \frac{\partial f(\mathbf{x}, \boldsymbol{\theta})}{\partial \theta_2} = \theta_1 x e^{\theta_2 x}$$

由于导数是未知参数 θ_1 与 θ_2 的函数，
所以模型是非线性的

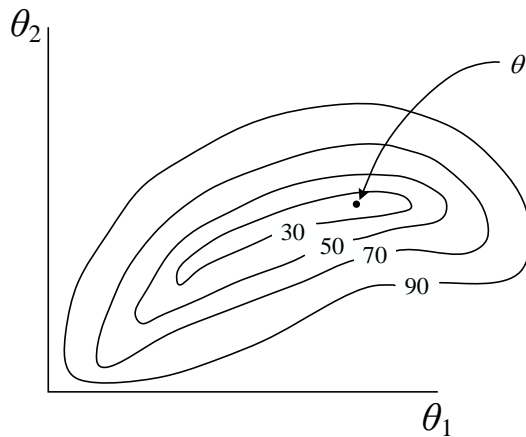
➤ 非线性最小二乘

□ 残差平方和函数的等高线

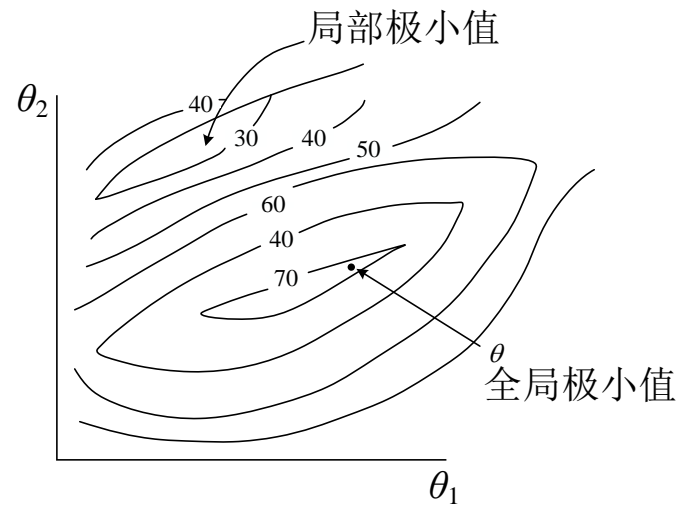
- 图a：模型为非线性回归模型时的通常状态下的等高线
- 图b：“香蕉形”的形状等高线。靠近最优值时曲面会被严重拉长，所以 θ 的许多解产生的残差平方和都会接近于全局最小值，难以求解 θ 的全局最小值
- 图c：存在一个局部极小值与一个全局最小值的情况



a) 线性模型



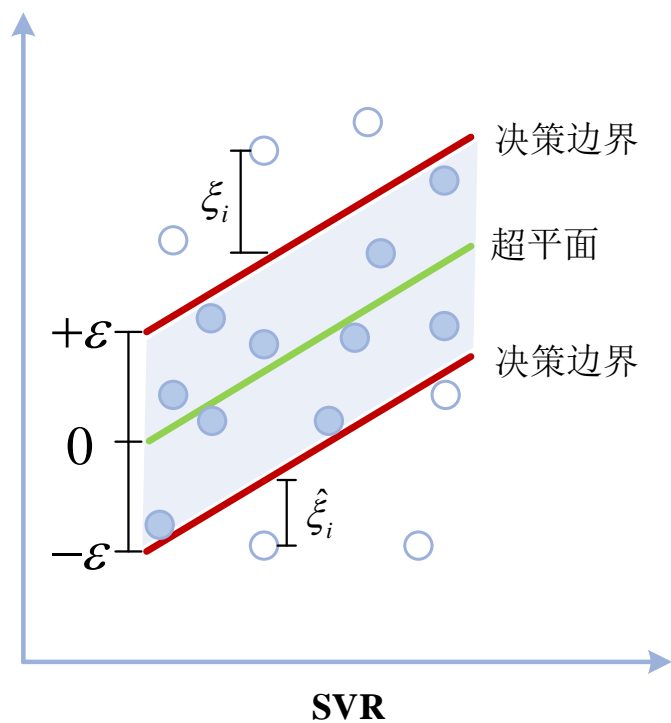
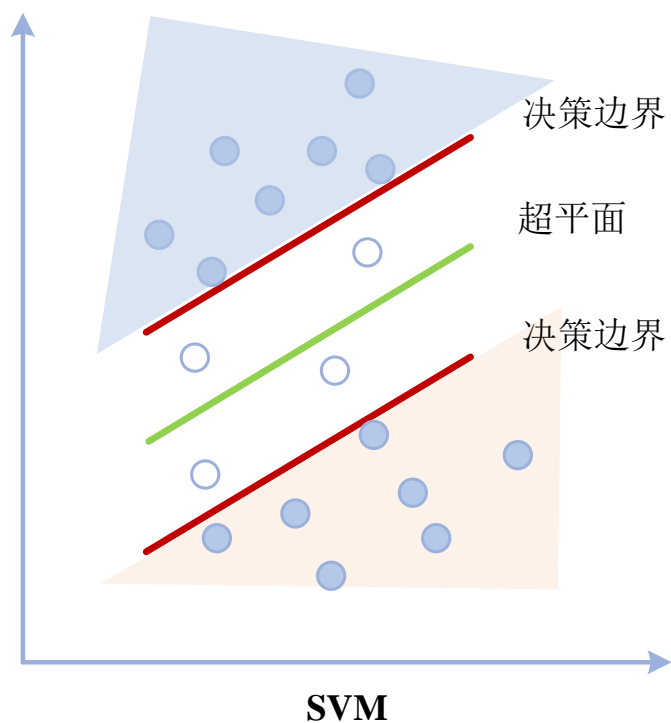
b) 非线性模型



c) 有局部极小值与全局极小值的非线性模型

➤ 支持向量回归

- ❑ SVM要通过最大化间隔，找到一个分离超平面，使得绝大多数的样本点位于两个决策边界的外侧
- ❑ SVR则考虑的是决策边界内的点，使尽可能多的样本点位于间隔内



- 回归分析的基本概念
- 线性回归
- 高维回归系数压缩
- 非线性回归
- 回归模型的验证
- 本章小结



➤ 基本概念

□ 在将模型发布之前，应当评定模型验证，以区分模型适用性检验与模型验证

- 模型适用性检验

- ✓ 残差分析、失拟检验、寻找高杠杆观测值与强影响观测值等

- 模型验证

- ✓ 在运作环境中，验证模型是否会成功运行

□ 三类程序的验证

- ✓ 模型系数与预测值的分析

- ✓ 用所收集的新数据来研究模型的预测性

- ✓ 数据分割

➤ 模型系数与预测值的分析

□ 研究最终回归模型的系数

- 系数是否稳定、系数的正负号与大小是否符合先验经验
- 所估计模型的系数与回归模型的信息进行对比

正负号不符合期望或绝对值过大的系数：

- ✓ 模型是不适用的
- ✓ 单个回归变量影响的估计值不良

➤ 收集新数据验证

□ 度量回归模型预测性能验证有效方法：收集新数据并直接对比预测值与新观测值

➤ 数据分割

- **数据验证问题**：例如数据收集的预算可能已经花光了，可能已经转变为生产其他产品，或者不能获得收集数据所需要的其他设备或资源
- **解决方法**：数据分割
 - 估计性数据（训练集）
 - 验证性数据（测试集）

} 交叉验证

➤ 模型拟合度量

□ 适用性检验可在回归过程的各个阶段，也可在得到模型后进行

● 残差图

$$e_i = y_i - \hat{y}_i (i = 1, 2, \dots, n)$$

观测值

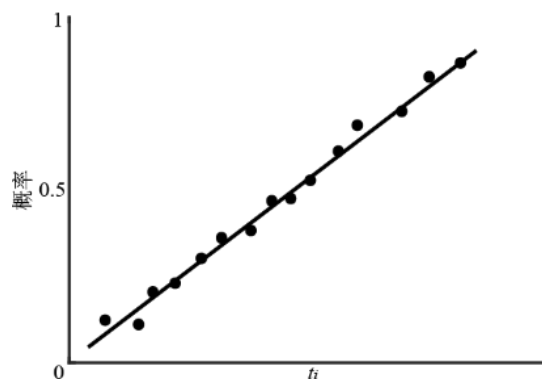
拟合值

- ✓ 残差分析是探索模型适用性的有效方法
- ✓ 残差的图形分析是研究回归模型拟合适用性与检验基本假设较为有效的方法

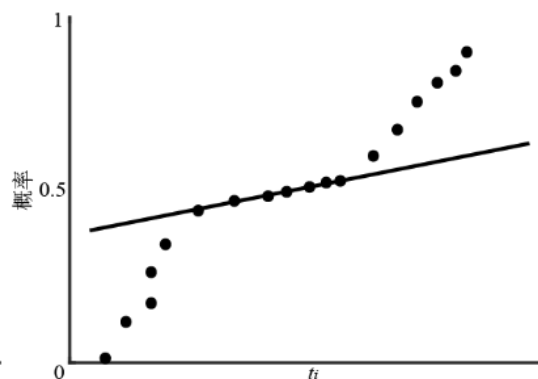
➤ 模型拟合度量

■ 正态概率图

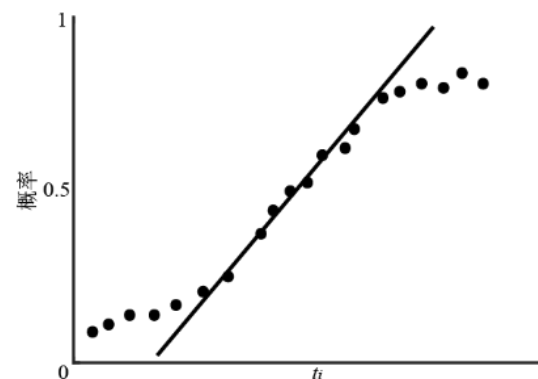
如果误差来自**厚尾分布**而不是正态分布，那么最小二乘拟合可能对数据的小型子集是敏感的。



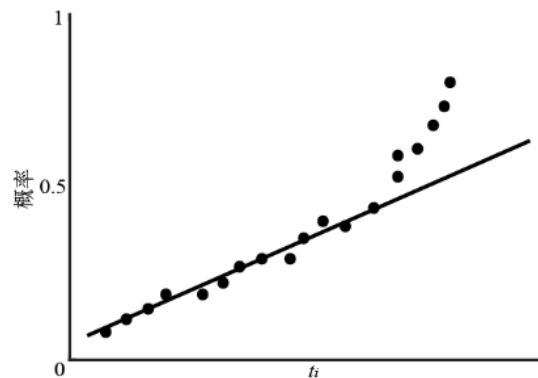
a) 理想的



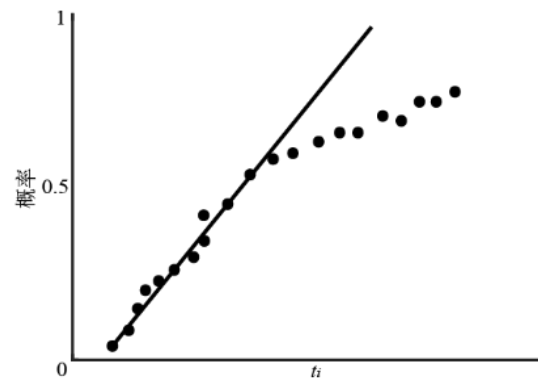
b) 轻尾分布



c) 重尾分布



d) 正的偏斜

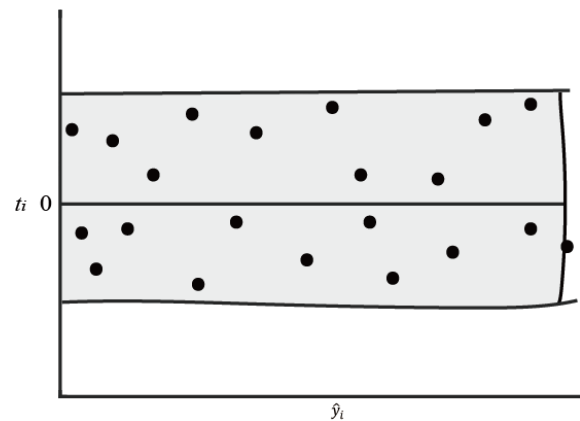


e) 负的偏斜

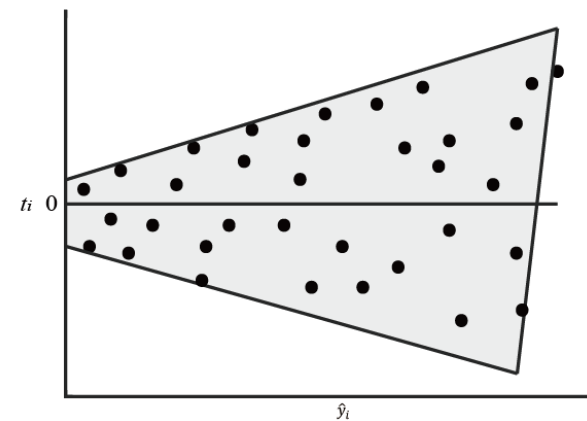
➤ 模型拟合度量

■ 残差与拟合值 \hat{y}_i 的残差图

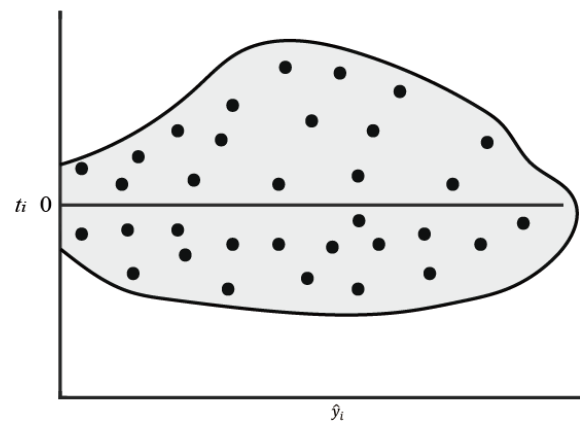
- ✓ 图a：残差包含在一条水平带中，模型不存在明显的缺点；图b-图d都是模型存在缺点的表现
- ✓ 图b-图c：误差的方差不是常数。通常利用对响应变量变换来得到稳定的方差
- ✓ 图d：曲线点表明存在非线性，可能意味着模型需要其他回归变量



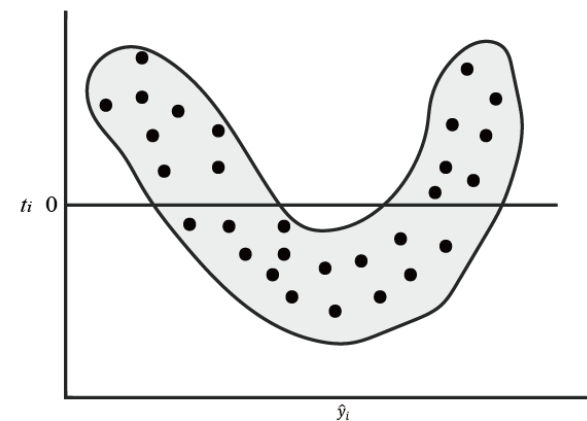
a) 令人满意的模式



b) 漏斗模式



c) 双弓模式



d) 非线性模式

➤ 模型拟合度量

● 拟合效果度量

- ✓ 方法一：考察 Y 对 \hat{Y} 的散点图，散点图上的这组点离一条直线越近， Y 与 X 之间的线性关系越强
- ✓ 方法二：获得线性模型参数的最小二乘估计后，通过某些量度量

$$SST = \sum (y_i - \bar{y})^2$$

总离差平方和

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

回归平方和

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

残差平方和

■ 总离差平方和： $SST = SSR + SSE$

■ 拟合优度指数： $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

✓ SSR度量了 X 对 Y 的**预测能力**

✓ SSE度量了**预测误差**

✓ 如果 R^2 接近1： Y 的绝大部分变化可由 X 解释

✓ R^2 ：**预测变量对响应变量的解释能力**

- 回归分析的基本概念
- 线性回归
- 高维回归系数压缩
- 非线性回归
- 回归模型的验证
- **本章小结**



□ 基本概念

- 回归分析的概念及基本步骤

□ 线性回归

- 基本的线性回归模型：最小二乘估计法、加权最小二乘估计法

□ 高维回归系数压缩

- 高维系数存在的共线性问题及其影响
- 岭回归、Lasso回归、主成分回归、偏最小二乘回归

□ 非线性回归

- 非线性最小二乘估计法
- 支持向量回归

□ 模型验证

- 拟合后模型的效果检测

- ❑ Hoerl A E, Kennard R W. Ridge regression: Biased estimation for nonorthogonal problems[J]. Technometrics, 1970, 12(1): 55-67.
- ❑ Meier L, Van De Geer S, Bühlmann P. The group lasso for logistic regression[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2008, 70(1): 53-71.
- ❑ Farrar D E, Glauber R R. Multicollinearity in regression analysis: the problem revisited[J]. The Review of Economic and Statistics, 1967: 92-107.
- ❑ Jolliffe I T. A note on the use of principal components in regression[J]. Journal of the Royal Statistical Society: Series C (Applied Statistics), 1982, 31(3): 300-303.
- ❑ Vinzi V E, Chin W W, Henseler J, et al. Handbook of partial least squares[M]. Berlin: Springer, 2010.
- ❑ Chen X, Cao W H, Gan C, Wu M. A hybrid partial least squares regression-based real time pore pressure estimation method for complex geological drilling process[J]. Journal of Petroleum Science and Engineering, 2022 210: 109771.
- ❑ Yuan X, Ge Z, Huang B, Song Z and Wang Y. Semisupervised JITL Framework for Nonlinear Industrial Soft Sensing Based on Locally Semisupervised Weighted PCR[J], IEEE Transactions on Industrial Informatics, 2017, 13(2):532-541.
- ❑ Yuan X, Li L and Wang Y. Nonlinear Dynamic Soft Sensor Modeling With Supervised Long Short-Term Memory Network[J], IEEE Transactions on Industrial Informatics, 2020, 16(5):3168-3176.