

第2章 模型评价

李高荣

北京师范大学统计学院

E-mail: *ligaorong@bnu.edu.cn*



本章纲要

1 回归模型及评价准则

- 回归模型
- 模型估计
- 回归模型精度的评价准则

2 分类模型及评价准则

3 参考文献

4 作业

微信公众号: BNULgr



- 扫二维码获取在线课件和相关教学电子资源
- 请遵守电子资源使用协议

本章纲要

1 回归模型及评价准则

- 回归模型
- 模型估计
- 回归模型精度的评价准则

2 分类模型及评价准则

3 参考文献

4 作业

本章纲要

1 回归模型及评价准则

- 回归模型
- 模型估计
- 回归模型精度的评价准则

2 分类模型及评价准则

3 参考文献

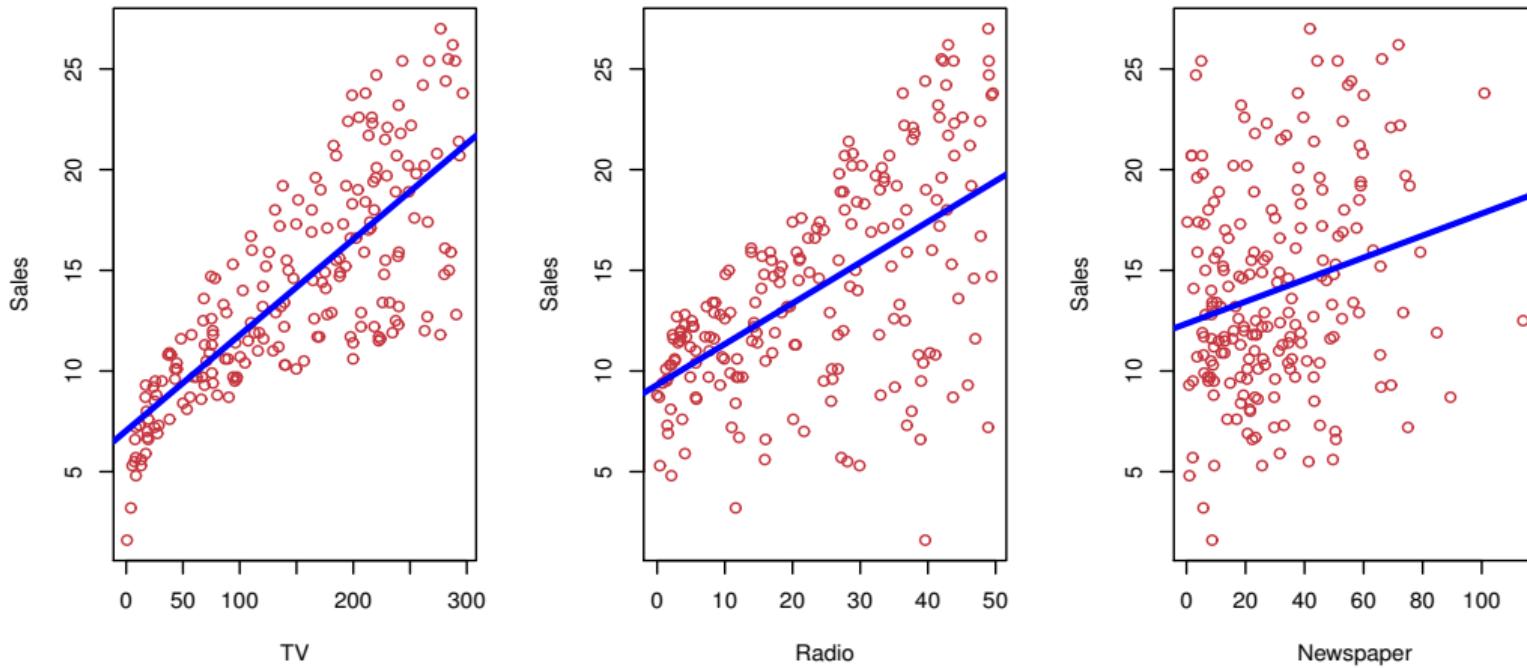
4 作业

广告数据

受客户委托做统计咨询，为某产品的销量提升提供策略咨询建议。
Advertising(广告)数据集记录了该产品在200个不同市场的销售情况及
该产品在每个市场中3类广告媒体的预算，这3类媒体分别为：

- ① 电视(TV)
- ② 广播(Radio)
- ③ 报纸(Newspaper)

回归模型



■ 目标: 建立一个基于3类广告媒体预算的精准预测销量模型.

回归模型

■ 输入变量(input variable): 广告的预算, 也称预测变量或自变量, 如

- ① X_1 表示TV的广告预算
- ② X_2 表示radio的广告预算
- ③ X_3 表示newspaper的广告预算

■ 输出变量(output variable): 产品的销售(sales), 也称响应变量或因变量, 用 Y 表示.

■ 建立的回归模型为

$$Y = g(\mathbf{X}) + \varepsilon,$$

其中 $\mathbf{X} = (X_1, X_2, X_3)^T$ 为3维的输入变量或预测变量, $g(\cdot)$ 是未知的函数, 随机误差 ε 满足: $E(\varepsilon|\mathbf{X}) = 0$.

■ **回归模型:** 通常关心 p 个协变量 X_1, \dots, X_p 与响应变量 Y 之间的相互关系, 而如下的回归模型就是研究这种变量之间相互关系的一个有力的工具, 即

$$Y = g(\mathbf{X}) + \varepsilon,$$

- Y 称为响应变量或因变量
- $\mathbf{X} = (X_1, \dots, X_p)^T$ 称为 p 维协变量, 也称为预测变量、自变量或特征变量
- $g(\mathbf{X})$ 是 p 维协变量 X_1, \dots, X_p 的未知函数
- ε 是随机误差, 通常假设与 \mathbf{X} 独立, 并满足: $E(\varepsilon|\mathbf{X}) = 0$ 和 $\text{Var}(\varepsilon) = \sigma^2 < \infty$

■ 回归函数：把给定 $X = x$ 时， Y 的条件数学期望

$$g(x) = E(Y|X=x)$$

称为随机变量 Y 对 $X = x$ 的 p 元 **回归函数**.

■ 回归函数从平均意义上刻画了协变量 X 和响应变量 Y 之间的统计规

律.

■ 回归模型的目的：

★ 预测

★ 控制

★ 统计推断

本章纲要

1 回归模型及评价准则

- 回归模型
- 模型估计
- 回归模型精度的评价准则

2 分类模型及评价准则

3 参考文献

4 作业

♠ 问题：如何根据样本观测值确定 Y 关于 X 的回归函数 $g(X)$ ？

模型估计

♠ 问题：如何根据样本观测值确定 Y 关于 X 的回归函数 $g(X)$ ？

■ 假设进行 n 次独立试验，则可得到 n 组独立同分布的观测样本，记为 $D = \{(\mathbf{x}_i, y_i), i = 1, \dots, n\}$ ，其中 $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$.

序号	Y	X_1	X_2	...	X_p
1	y_1	x_{11}	x_{12}	...	x_{1p}
2	y_2	x_{21}	x_{22}	...	x_{2p}
:	:	:	:	:	:
n	y_n	x_{n1}	x_{n2}	...	x_{np}

- 统计学习的主要任务：通过观测样本数据 D , 利用统计学习方法估计未知的回归函数 $g(\cdot)$, 即得到一个估计 $\hat{g}(\cdot)$.
- 极小化下面的**经验风险函数或目标函数** $R_n(g)$, 获得 $g(\cdot)$ 的估计, 即

$$\hat{g} = \arg \min_g R_n(g) = \arg \min_g \frac{1}{n} \sum_{i=1}^n \ell_n(y_i, g(\mathbf{x}_i)).$$

- $\ell_n(y_i, g(\mathbf{x}_i))$ 为**损失函数**, 用于度量 y_i 与 $g(\mathbf{x}_i)$ 之间的偏离程度.

■ 常用的损失函数 $\ell_n(y, g(\mathbf{x}))$ 有：

- 平方损失函数: $\ell_n(y, g(\mathbf{x})) = (y - g(\mathbf{x}))^2$;
- 绝对损失函数: $\ell_n(y, g(\mathbf{x})) = |y - g(\mathbf{x})|$;
- L_q 损失函数: $\ell_n(y, g(\mathbf{x})) = |y - g(\mathbf{x})|^q$, 其中 $q > 0$;
- 0 – 1损失函数, 负对数似然损失函数和分位数损失函数等.

■ 在实际应用中, 需要根据数据类型和已知的数据信息等使用合适的损失函数, 例如

- Y 的观测样本来自正态分布时, 可使用平方损失函数或负对数似然损失函数;
- Y 的观测样本具有离群点或来自重尾分布时, 可使用稳健的绝对损失函数或分位数损失函数.

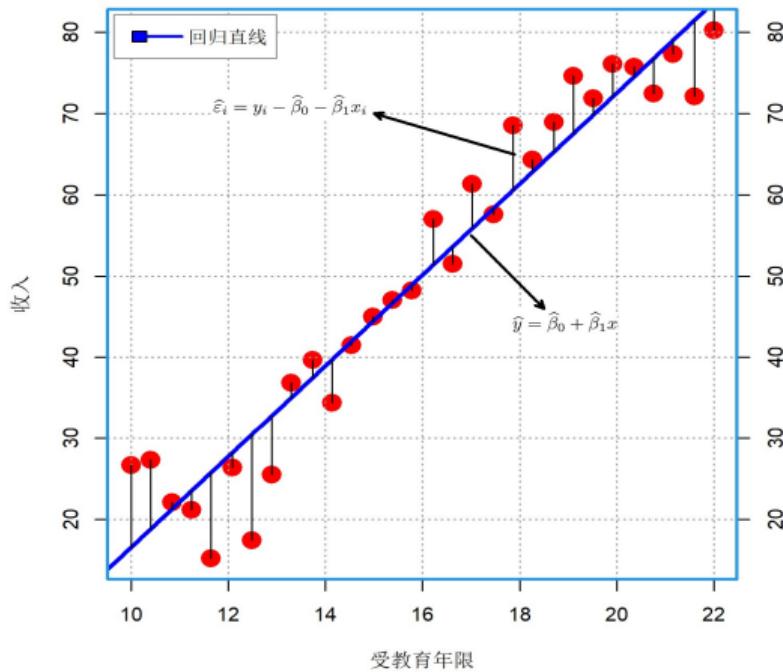
- 在统计学习和机器学习中, 估计回归函数 $g(\cdot)$ 的方法通常有: **参数方法**和**非参数方法**.
- **参数方法**是一种基于参数模型的估计方法, 通常需要根据专业知识或者经验知识假设回归函数 $g(X)$ 具有关于未知参数 β 的一定形式或形状, 一旦模型被选定后, 就需要用**训练观测样本**去拟合或训练模型.
- 最经典的参数模型为线性回归模型, 即

$$g(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p,$$

其中 $\beta_0, \beta_1, \dots, \beta_p$ 为未知的参数.

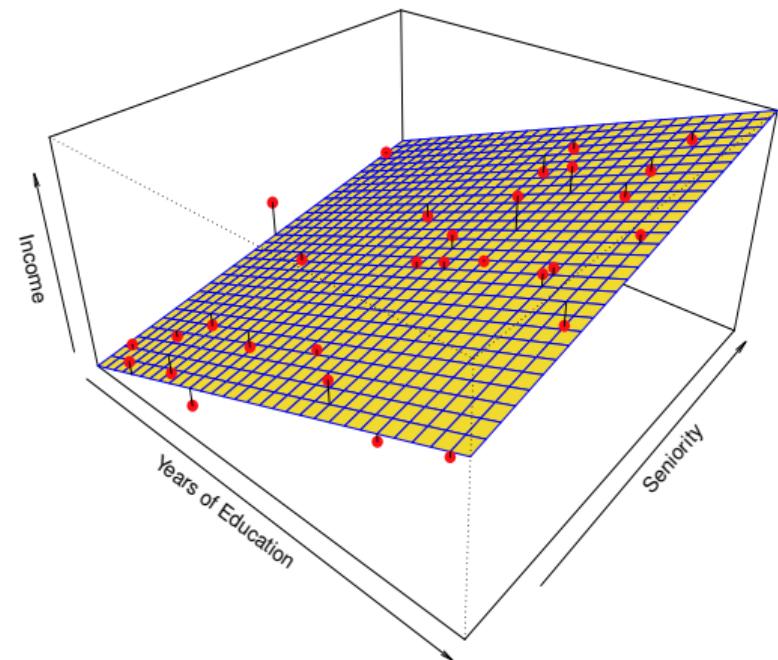
参数方法

- 线性回归模型常用的估计方法: **普通最小二乘方法(OLS)**
- 当 $p = 1$ 时, 回归函数变为 $g(X) = \beta_0 + \beta_1 X$.
- 右图展示了受教育年限对收入的散点图和普通最小二乘(OLS)拟合回归直线 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.
- 竖线表示OLS拟合回归直线与每个观测值有关的误差: $\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$.



参数方法

- 线性回归模型: $\text{income} = \beta_0 + \beta_1 \times \text{education} + \beta_2 \times \text{seniority} + \varepsilon.$
- 右图展示了当 $p = 2$ 时, 多元线性回归模型的OLS拟合平面.
- 竖线表示OLS拟合平面与每个观测值有关的误差: $\hat{\varepsilon}_i = y_i - \hat{y}_i.$



- 另外一个经典的参数回归模型是**非线性回归模型**, 即

$$Y = g(\mathbf{X}, \boldsymbol{\beta}) + \varepsilon.$$

- $g(\cdot)$ 是依赖于 p 维协变量向量 \mathbf{X} 和 q 维未知参数向量 $\boldsymbol{\beta} = (\beta_1, \dots, \beta_q)^T$ 的已知回归函数.
- 最经典的估计方法: **非线性最小二乘(nonlinear least squares, NLS)方法.**
- 在R语言中, 可用程序包**nls2**拟合非线性最小二乘问题, 更详细的讨论见Huet等(2004).

- **参数方法:** 把基于参数回归模型的估计方法统称为参数方法, 该方法仅需要对模型中的未知参数向量 β 进行估计, 不需要估计任意一个未知的回归函数 $g(\cdot)$.
- **参数方法的缺陷:** 当假定的参数回归模型错误指定时, 即与真实的回归函数 $g(\cdot)$ 偏离较大, 将导致拟合的回归函数 $\hat{g}(\cdot)$ 效果会很差, 进而不能对响应变量 Y 作出很好的预测.
- **解决方案:** 考虑更加**灵活(flexible)**的模型对数据进行拟合.
- 然而, 拟合更加灵活的模型将会导致**过拟合(overfitting)**问题.

■ **非参数方法:** 一种数据驱动的方法, 能够有效解决参数方法模型错误指定的问题, 其优点有:

- ① 不需要假设回归函数 $g(\cdot)$ 的具体形式, 可在更大的范围内选择更适合 $g(\cdot)$ 形状的估计;
- ② 函数 $g(\cdot)$ 的形式完全由数据决定, 可以更好地拟合数据.

■ 非参数方法较参数方法的缺点:

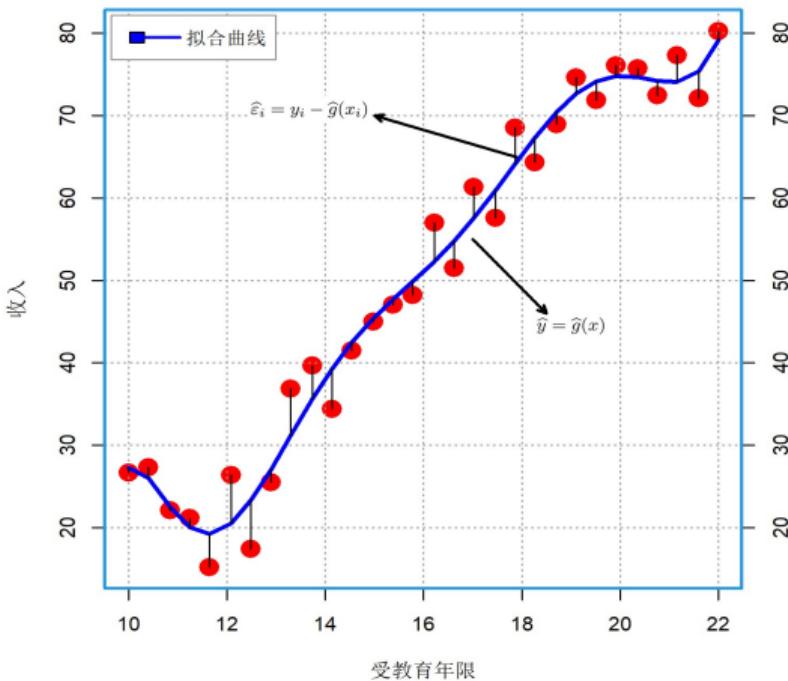
- ① 无法将估计 $g(\cdot)$ 的问题简化到仅仅对少数参数进行估计, 所以为了获得对 $g(\cdot)$ 更为精准的估计, 往往需要大量的观测点;
- ② 如果维数 p 很大时, 会遭遇“维数灾祸”问题.

非参数方法

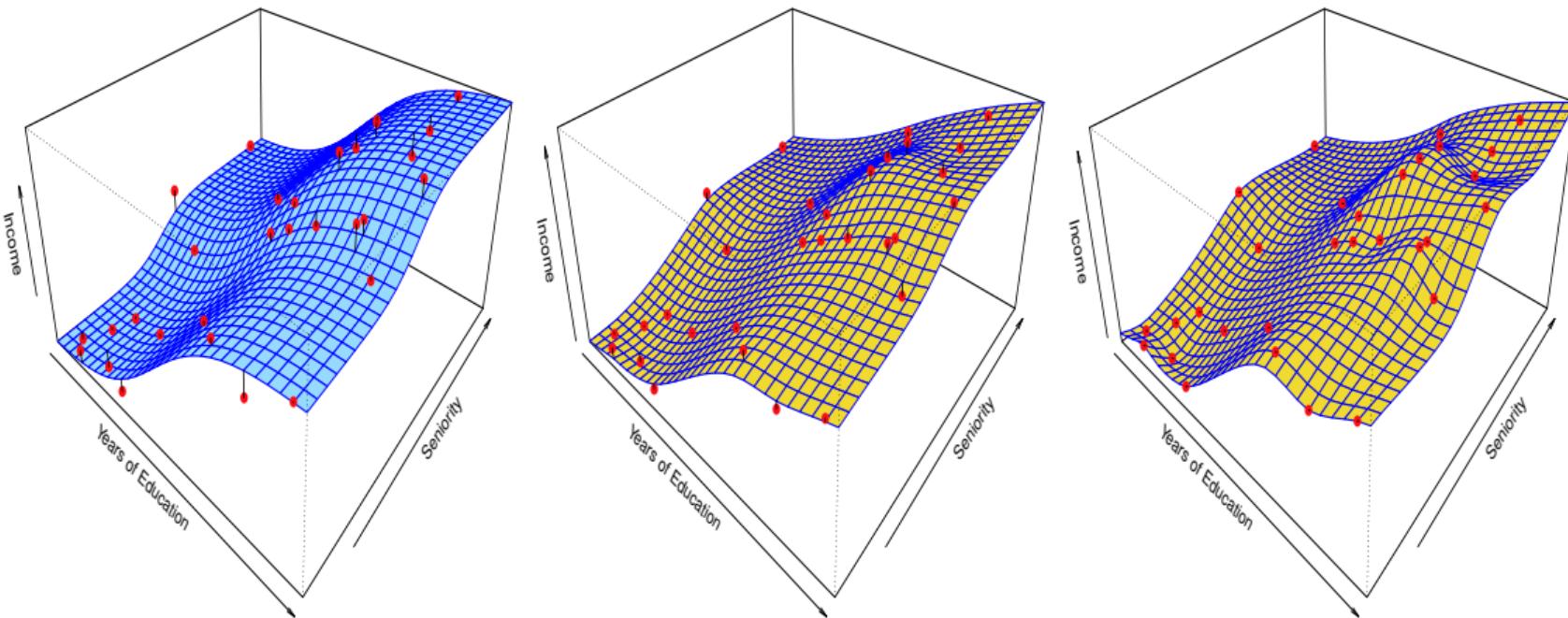
- 右图展示了利用非参数方法对受教育年限与收入关系的拟合曲线：
 $\hat{y} = \hat{g}(x)$.

- 竖线表示非参数拟合曲线与每个观测值有关的误差：
 $\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{g}(x_i)$.

- 非参数方法不对 $g(\cdot)$ 施加任何模型形式，而且可以更好捕捉数据的局部特征，能够更好拟合数据.



非参数方法



当 $p = 2$ 时的散点图, 以及三种非参数方法所得拟合曲面

■ 主要的非参数估计方法有：

- K 近邻回归方法
- 多项式回归方法
- 回归样条方法
- 光滑样条方法
- Nadaraya-Watson核光滑方法
- 局部多项式光滑方法
- 回归树方法
- 随机森林
- 神经网络
-

- 所谓“维数灾祸”问题：随着协变量 X 的维数 p 变大，利用非参数方法估计回归函数 $g(\cdot)$ 会变得越来越困难，估计的收敛速度也会越来越慢，计算的复杂度也会呈指数阶增长。
- 例如，对一个二次可微的回归函数 $g(x)$ ，其非参数估计 $\hat{g}(x)$ 的均方误差 (mean squared error, MSE) 为

$$\text{MSE}(\hat{g}(x)) \approx \frac{c}{n^{4/(4+p)}},$$

其中 $c > 0$ 为常数， n 为样本量大小， p 为协变量的维数。

- 如果要求非参数估计 $\hat{g}(\mathbf{x})$ 满足: $MSE(\hat{g}(\mathbf{x})) = \epsilon$, 其中 $\epsilon > 0$ 为给定很小的常数.
- 通过求解方程, 可解得样本量 n 为: $n \propto \left(\frac{c}{\epsilon}\right)^{p/4}$.
- 意味着: 随着维数 p 的变大, 样本量 n 需要呈指数阶增长才能保证: $MSE(\hat{g}(\mathbf{x})) = \epsilon$.
- 非参数方法需要使用 \mathbf{x} 局部邻域内的观测样本点估计回归函数 $g(\mathbf{x})$,但是在高维情形, 数据点是非常稀疏的, 进而在 \mathbf{x} 的局部邻域内样本点更少, 导致了“维数灾祸”问题.

♠ **问题:** 假设从超立方体 $[-1, 1]^p = [-1, 1] \times \cdots \times [-1, 1]$ 的均匀分布抽取 n 个样本点, 请问在子区间 $[-0.1, 0.1]^p$ 内大约有多少个样本点?

♠ **问题:** 假设从超立方体 $[-1, 1]^p = [-1, 1] \times \cdots \times [-1, 1]$ 的均匀分布抽取 n 个样本点, 请问在子区间 $[-0.1, 0.1]^p$ 内大约有多少个样本点?

□ 如果 $p = 1$ 时, 容易得到在子区间 $[-0.1, 0.1]$ 内大约有 $n/10$ 个样本点;

♠ **问题:** 假设从超立方体 $[-1, 1]^p = [-1, 1] \times \cdots \times [-1, 1]$ 的均匀分布抽取 n 个样本点, 请问在子区间 $[-0.1, 0.1]^p$ 内大约有多少个样本点?

□ 如果 $p = 1$ 时, 容易得到在子区间 $[-0.1, 0.1]$ 内大约有 $n/10$ 个样本点;

□ 如果 $p = 10$ 时, 在子区间 $[-0.1, 0.1]^{10}$ 内大约有

$$n \times \left(\frac{0.2}{2}\right)^{10} = \frac{n}{10,000,000,000}$$

个样本点.

■ 因此, 样本量 n 足够大才能保证子区间 $[-0.1, 0.1]^{10}$ 内有足够的样本点.

- **自由度:** Hastie和Tibshirani (1990) 定义了有效的**自由度**(degrees of freedom, df), 它表示模型的**复杂度或灵活度**, 对估计拟合模型的预测精度非常有用.
- 令 y_i 的预测值为 $\hat{y}_i = \hat{g}(x_i)$, 这时可定义估计 \hat{g} 的自由度为

$$df(\hat{g}) = \frac{1}{\sigma^2} \sum_{i=1}^n \text{Cov}(\hat{y}_i, y_i).$$

- 由自由度的定义可知: 估计 \hat{g} 的自由度反映的是响应变量预测值 \hat{y}_i 与观测值 y_i 之间协方差之和除以方差 σ^2 .

自由度

- 令 $\hat{\mathbf{Y}} = (\hat{y}_1, \dots, \hat{y}_n)^T \in \mathbb{R}^n$ 和 $\mathbf{Y} = (y_1, \dots, y_n)^T \in \mathbb{R}^n$, 则自由度可以写成如下矩阵形式

$$df(\hat{g}) = \frac{1}{\sigma^2} \text{tr} \left(\text{Cov}(\hat{\mathbf{Y}}, \mathbf{Y}) \right).$$

- 记 \mathbf{S} 为 $n \times n$ 的 **投影矩阵或光滑矩阵**, 不管对参数或非参数方法, 假设回归函数的估计都具有线性光滑, 即 $\hat{\mathbf{g}} = \mathbf{S}\mathbf{Y} = (\hat{g}(\mathbf{x}_1), \dots, \hat{g}(\mathbf{x}_n))^T$.
- 这时, \mathbf{Y} 的预测可以写为 $\hat{\mathbf{Y}} = \mathbf{S}\mathbf{Y}$, 且由 $\text{Cov}(\mathbf{Y}) = \sigma^2 \mathbf{I}_n$, 则自由度为:
$$df(\hat{g}) = \text{tr}(\mathbf{S}).$$

自由度

■ **自由度的直观解释:** 表示估计回归函数 $g(\cdot)$ 时, 需要估计的有效参数个数. 如果自由度变大, 模型的复杂度增加, 模型对数据的拟合变得更加灵活.

例1: 均值估计的自由度

假设 $y_1, \dots, y_n \sim i.i.d. N(\mu, \sigma^2)$, 则均值 μ 的估计为 $\hat{\mu} = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$, 且 $\hat{\mu}$ 的自由度为

$$df(\hat{\mu}) = \frac{1}{\sigma^2} \sum_{i=1}^n \text{Cov}(\bar{y}, y_i) = \frac{1}{\sigma^2} \sum_{i=1}^n \frac{\sigma^2}{n} = 1.$$

自由度

例2: 同方差模型均值向量估计的自由度

假设一个同方差模型: $\mathbf{Y} = (y_1, \dots, y_n)^T \sim (\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$, 其中 $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T$. 可得 $\boldsymbol{\mu}$ 的估计为

$$\hat{\boldsymbol{\mu}} = (y_1, \dots, y_n)^T,$$

则 $\hat{\boldsymbol{\mu}}$ 的自由度为

$$df(\hat{\boldsymbol{\mu}}) = \frac{1}{\sigma^2} \sum_{i=1}^n \text{Cov}(y_i, y_i) = n.$$

本章纲要

1 回归模型及评价准则

- 回归模型
- 模型估计
- 回归模型精度的评价准则

2 分类模型及评价准则

3 参考文献

4 作业

- 估计回归函数 $g(\cdot)$ 的主要原因有两个: 预测(prediction)和统计推断(statistical inference).
- 利用参数方法或非参数方法, 得到回归函数 $g(\cdot)$ 的一个估计 $\hat{g}(\cdot)$, 则 Y 的预测为

$$\hat{Y} = \hat{g}(X),$$

其中

- ① $\hat{g}(\cdot)$ 表示回归函数 $g(\cdot)$ 的估计, 是黑箱(black box);
- ② \hat{Y} 表示 Y 的预测值.

回归模型精度的评价准则—可约误差和不可约误差

■ \hat{Y} 作为 Y 的预测值，其精确性依赖于两个量：

- ① 可约误差 (reducible error)
- ② 不可约误差 (irreducible error)

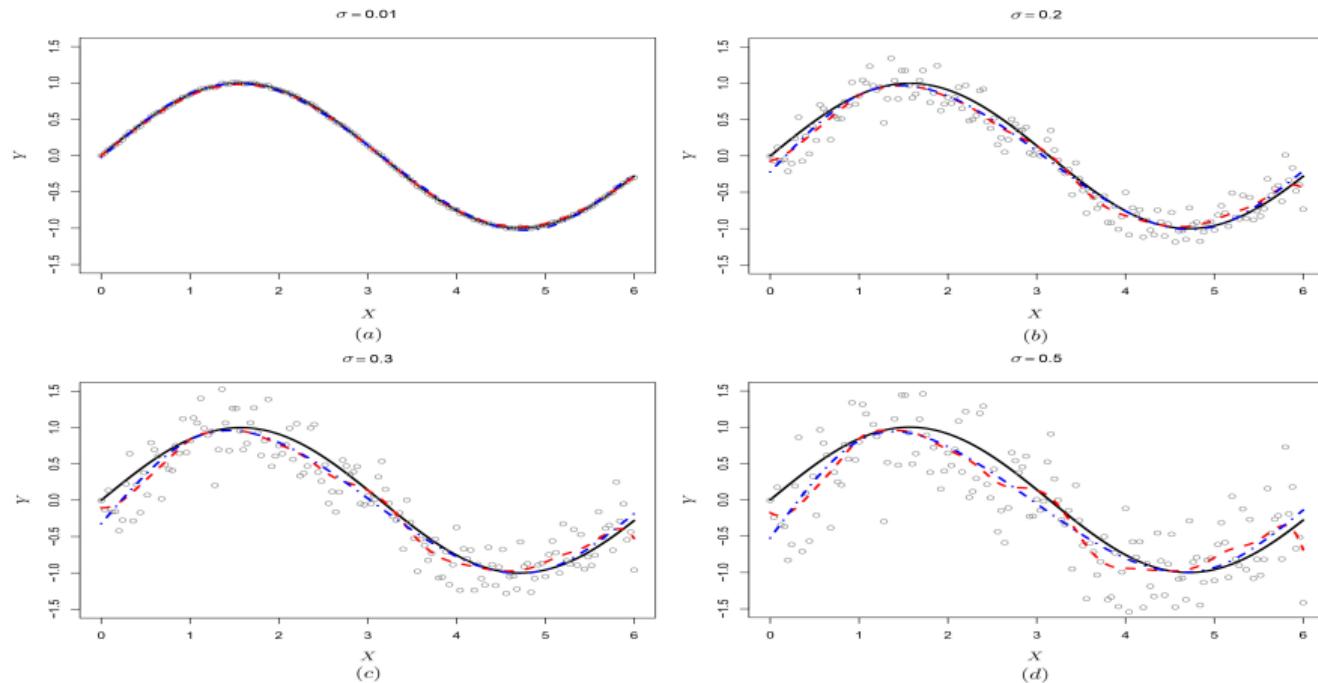
■ 给定 $X = \mathbf{x}$ 条件下，考虑均方误差(MSE)或期望预测误差(EPE)，即

$$\begin{aligned} E(Y - \hat{Y})^2 &= E[g(\mathbf{x}) + \varepsilon - \hat{g}(\mathbf{x})]^2 \\ &= \underbrace{E[g(\mathbf{x}) - \hat{g}(\mathbf{x})]^2}_{\text{可约误差}} + \underbrace{\text{Var}(\varepsilon)}_{\text{不可约误差}}, \\ &= [\text{bias}(\hat{g}(\mathbf{x}))]^2 + \text{Var}(\hat{g}(\mathbf{x})) + \text{Var}(\varepsilon). \end{aligned}$$

回归模型精度的评价准则—可约误差和不可约误差

- $\text{bias}(\hat{g}(\mathbf{x})) = E(\hat{g}(\mathbf{x})) - g(\mathbf{x})$: 表示估计 $\hat{g}(\mathbf{x})$ 的**偏差**, 反映估计 $\hat{g}(\mathbf{x})$ 对真实回归函数 $g(\mathbf{x})$ 的偏离程度;
- $\text{Var}(\hat{g}(\mathbf{x})) = E[\hat{g}(\mathbf{x}) - E(\hat{g}(\mathbf{x}))]^2$: 表示估计 $\hat{g}(\mathbf{x})$ 的**方差**, 反映的是估计 $\hat{g}(\mathbf{x})$ 对 $E(\hat{g}(\mathbf{x}))$ 的波动情况;
- $\text{Var}(\varepsilon)$: 表示**随机误差 ε 的方差**.
- 预测 \hat{Y} 的均方误差或期望预测误差分解为**可约误差**和**不可约误差**, 其中可约误差是由回归函数 $g(\mathbf{x})$ 的估计 $\hat{g}(\mathbf{x})$ 所引起, 而不可约误差是由随机误差 ε 所引起.

回归模型精度的评价准则—随机误差的影响



不同标准差 σ 情形下的散点图, 真实曲线和非参数估计拟合曲线, 其中黑色实线表示真实曲线, 蓝色点断线和红色虚线是两种非参数拟合曲线

♠ **问题:** 在统计学习中,有很多方法可以估计回归函数 $g(\cdot)$,为什么要介绍这么多不同的统计学习方法,而不直接介绍一种最优的统计学习方法用于数据拟合呢?如何评价它们对数据拟合的效果?

♠ **问题:** 在统计学习中,有很多方法可以估计回归函数 $g(\cdot)$,为什么要介绍这么多不同的统计学习方法,而不直接介绍一种最优的统计学习方法用于数据拟合呢?如何评价它们对数据拟合的效果?

- 为评价统计学习方法对某个数据集的拟合和预测效果,需要一些方法评价模型的预测结果与实际观测数据在结果上的一致性.
- 对给定的观测样本数据集 $D = \{(x_i, y_i), i = 1, \dots, n\}$,需要定量测量预测的响应值 $\hat{y}_i = \hat{g}(x_i)$ 与真实响应值 y_i 之间的接近程度.
- 常用的评价准则:**均方误差(MSE)**

■ 均方误差(MSE)定义为

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^n [y_i - \hat{g}(\mathbf{x}_i)]^2,$$

其中 $\hat{g}(\mathbf{x}_i)$ 是第*i*个观测点 \mathbf{x}_i 上拟合值或与预测值.

- ① 如果预测的响应值 $\hat{y}_i = \hat{g}(\mathbf{x}_i)$ 与真实响应值 y_i 很接近, 则MSE会非常小;
- ② 如果预测的响应值 $\hat{y}_i = \hat{g}(\mathbf{x}_i)$ 与真实响应值 y_i 之间的差异较大, 则MSE会非常大.

■ 通常的做法是把数据随机分成：

- ① 训练数据或训练样本：用统计学习方法得到 $g(\cdot)$ 的估计 $\hat{g}(\cdot)$ 后，进而计算训练均方误差(training MSE)；
- ② 测试数据或测试样本：用于测试模型的拟合效果，计算测试均方误差(test MSE).

♠ 问题：如何选择一个拟合效果好的模型呢？

■ 通常的做法是把数据随机分成：

- ① 训练数据或训练样本：用统计学习方法得到 $g(\cdot)$ 的估计 $\hat{g}(\cdot)$ 后，进而计算训练均方误差(**training MSE**)；
- ② 测试数据或测试样本：用于测试模型的拟合效果，计算**测试均方误差(test MSE)**.

♠ **问题：**如何选择一个拟合效果好的模型呢？

■ 选择学习模型使**测试均方误差最小**

■ 通常的做法是把数据随机分成：

- ① 训练数据或训练样本：用统计学习方法得到 $g(\cdot)$ 的估计 $\hat{g}(\cdot)$ 后，进而计算训练均方误差(training MSE)；
- ② 测试数据或测试样本：用于测试模型的拟合效果，计算测试均方误差(test MSE).

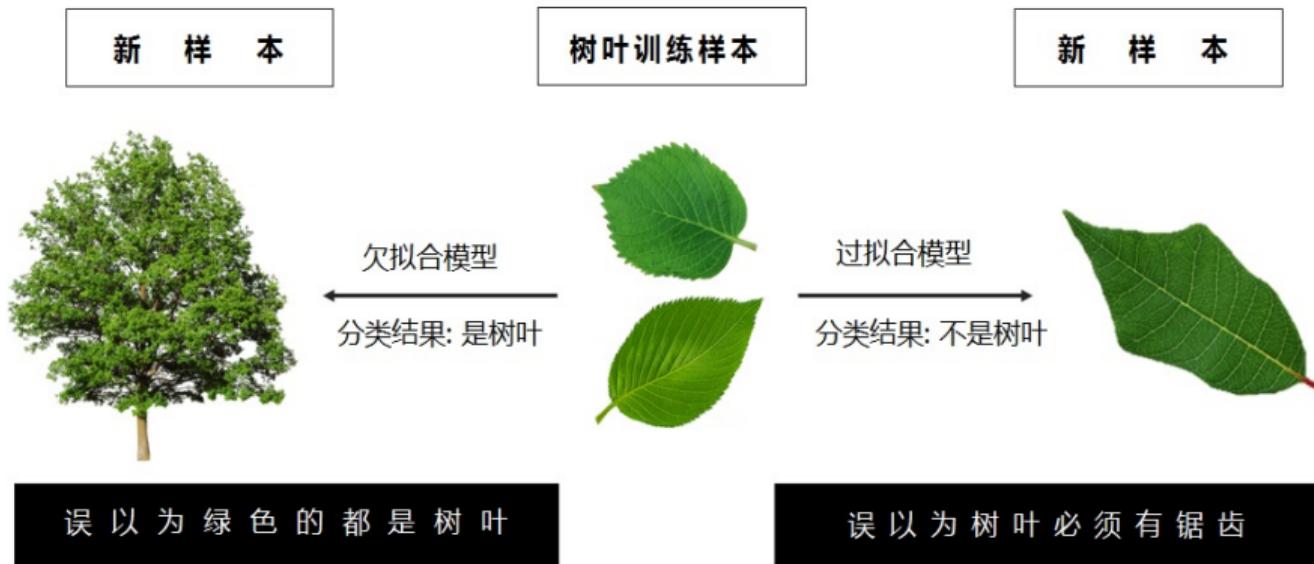
♠ 问题：如何选择一个拟合效果好的模型呢？

■ 选择学习模型使测试均方误差最小

♠ 问题：训练均方误差是否越小越好？

No! 因为会出现“**过拟合**”(overfitting)

No! 因为会出现“**过拟合**”(overfitting)



过拟合和欠拟合的直观比较

■ 在机器学习中，也把测试均方误差称为泛化误差(generalization error).

♠ **问题：**给定观测数据集，如何计算训练均方误差和测试均方误差？

■ 将观测样本数据集 $D = \{(\mathbf{x}_i, y_i), i = 1, \dots, n\}$ 随机分成互不重叠的训练集(training set)和测试集(testing set):

- 记 $D^{\text{Tr}} = \{(\mathbf{x}_i^{\text{Tr}}, y_i^{\text{Tr}}), i = 1, \dots, n_{\text{Tr}}\}$ 为训练集；
- $D^{\text{Te}} = \{(\mathbf{x}_i^{\text{Te}}, y_i^{\text{Te}}), i = 1, \dots, n_{\text{Te}}\}$ 为测试集，且 $n_{\text{Tr}} + n_{\text{Te}} = n$.

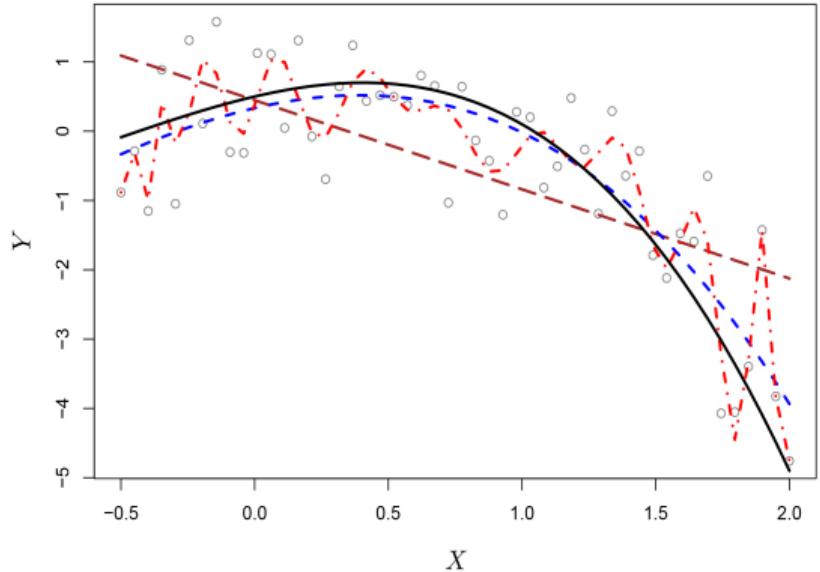
回归模型的评价准则

- 在训练集 D^{Tr} 上用统计学习方法拟合模型, 估计记为 $\hat{g}(\cdot)$, 并计算训练均方误差为

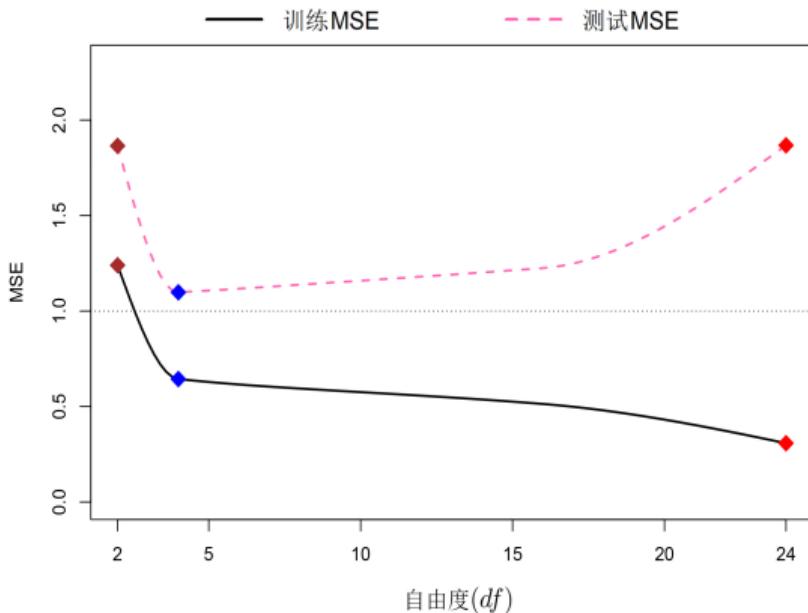
$$\text{training MSE} = \frac{1}{n_{\text{Tr}}} \sum_{i=1}^{n_{\text{Tr}}} \left[y_i^{\text{Tr}} - \hat{g}(\mathbf{x}_i^{\text{Tr}}) \right]^2.$$

- 最后, 用所得拟合模型用于测试集 D^{Te} 上预测响应变量, 记为 $\hat{y}_i^{\text{Te}} = \hat{g}(\mathbf{x}_i^{\text{Te}})$, 并计算测试均方误差为

$$\text{test MSE} = \frac{1}{n_{\text{Te}}} \sum_{i=1}^{n_{\text{Te}}} \left(y_i^{\text{Te}} - \hat{y}_i^{\text{Te}} \right)^2 = \frac{1}{n_{\text{Te}}} \sum_{i=1}^{n_{\text{Te}}} \left[y_i^{\text{Te}} - \hat{g}(\mathbf{x}_i^{\text{Te}}) \right]^2.$$



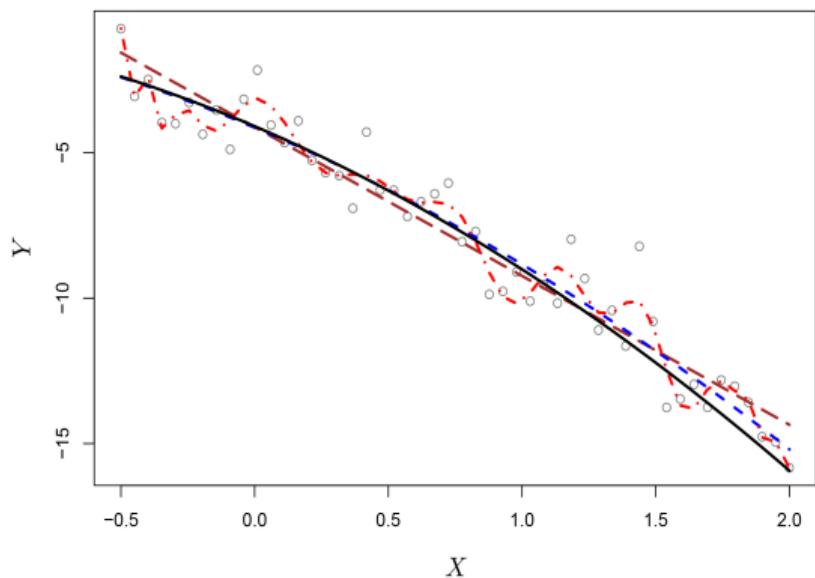
(a)



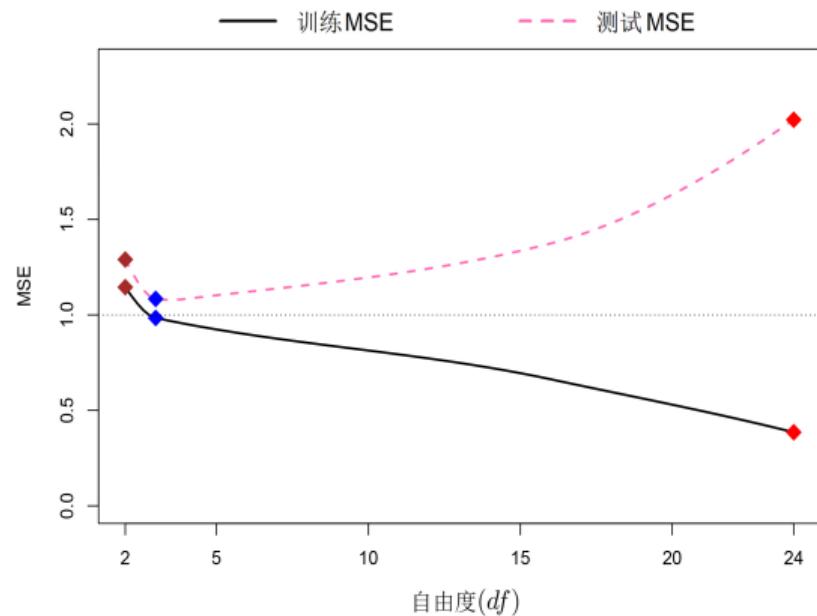
(b)

(a) 展示散点图, 真实曲线和三种估计的拟合曲线, 其中黑色实线表示真实曲线, 深红色断线表示一元线性回归模型的拟合直线, 以及另外两个是非参数拟合曲线(蓝色虚线和红色点断线); (b) 三种估计方法对应的训练均方误差(黑色曲线)和测试均方误差(红色曲线), 三种方法都已使测试均方误差尽可能最小, 三种颜色方块对应三种估计方法, 水平虚线表示不可约误差 $\text{Var}(\varepsilon)$

回归模型的评价准则



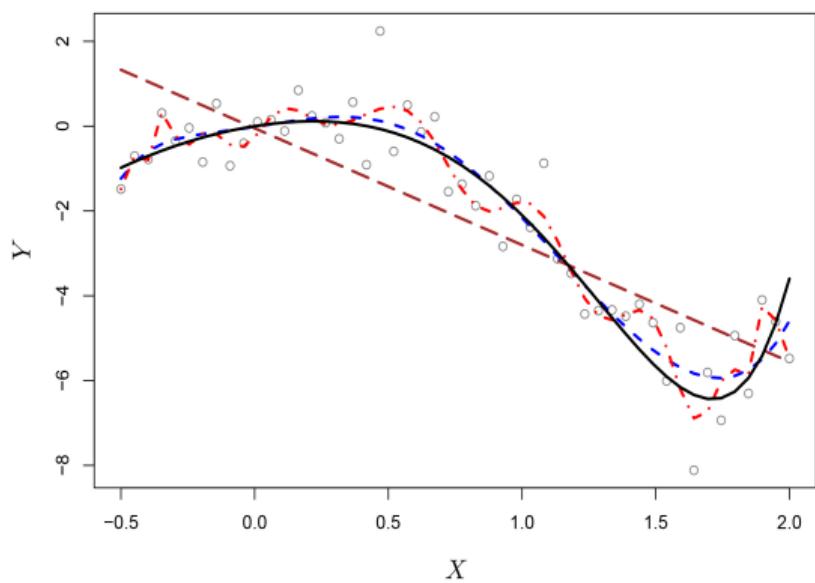
(a)



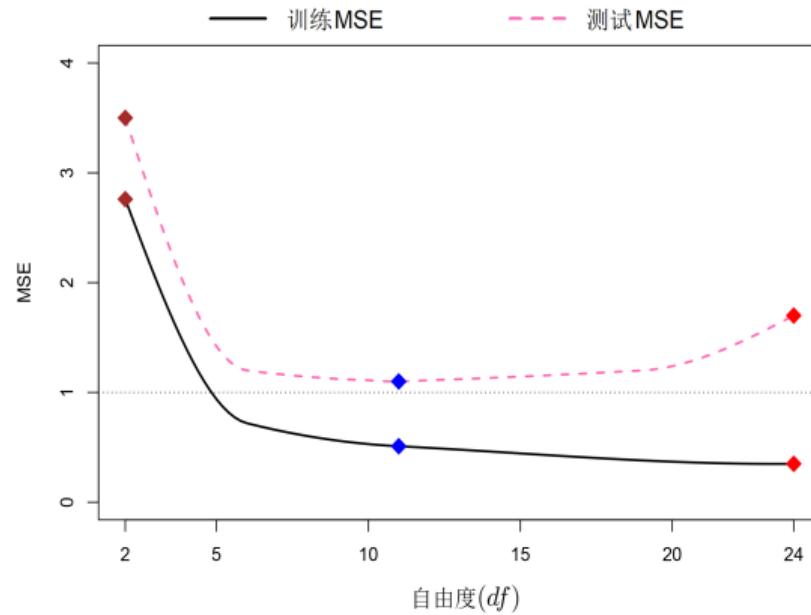
(b)

真实函数 $g(x)$ 接近于线性函数, 这时一元线性模型有较好的拟合

回归模型的评价准则



(a)



(b)

真实函数 $g(x)$ 完全不同于线性函数，这时一元线性回归模型有很差的拟合，而两种非参数方法有较好的拟合，特别是蓝色拟合曲线非常接近于真实曲线

- 随着自由度增加, 测试均方误差呈**U型曲线**, 表明统计学习方法在计算上存在两种博弈.
- 给定新的测试样本点 (\mathbf{x}_0, y_0) , 满足 $y_0 = g(\mathbf{x}_0) + \varepsilon_0$, 其中随机误差 ε_0 满足 $E(\varepsilon_0) = 0$ 和 $\text{Var}(\varepsilon_0) = \sigma^2$.
- 期望测试均方误差能分解成三个基本量的和, 即

$$E[y_0 - \hat{g}(\mathbf{x}_0)]^2 = [\text{bias}(\hat{g}(\mathbf{x}_0))]^2 + \text{Var}(\hat{g}(\mathbf{x}_0)) + \text{Var}(\varepsilon_0),$$

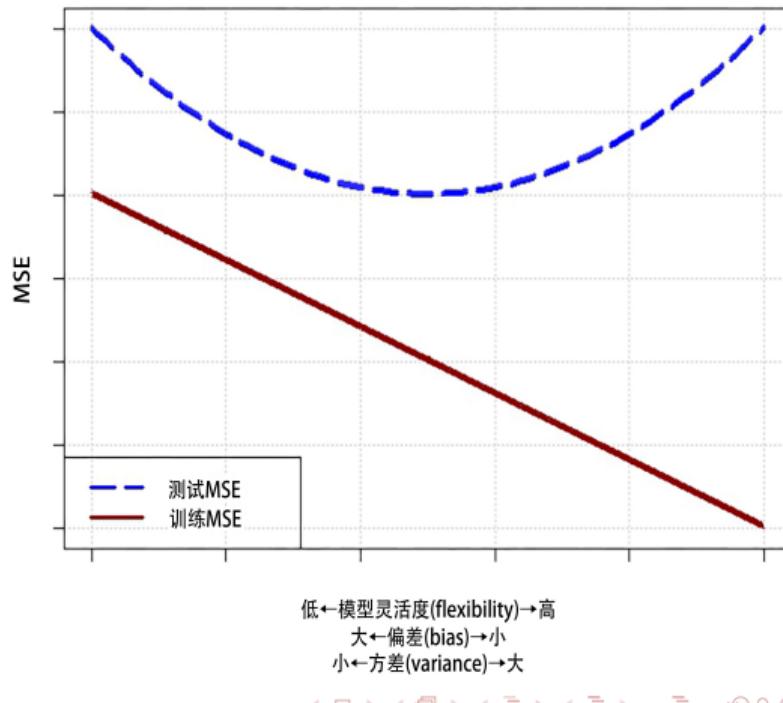
其中 $\hat{g}(\cdot)$ 为基于训练样本所得回归函数 $g(\cdot)$ 的估计.

回归模型的评价准则: 偏差一方差权衡

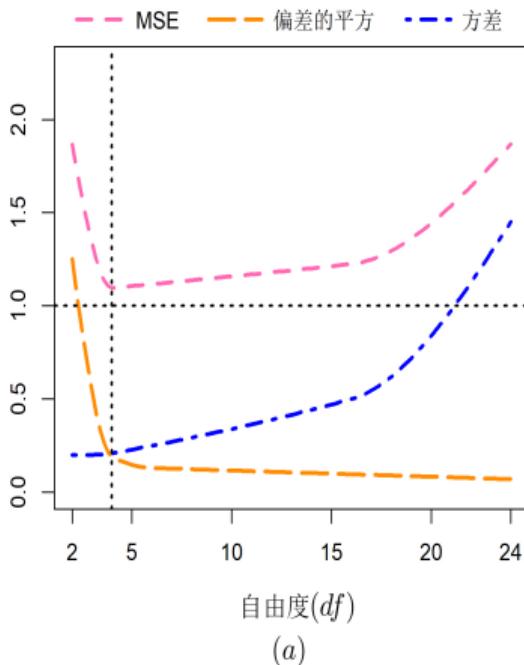
- 在实际应用中, 选择使测试均方误差最小的模型作为最终的预测模型.
- 为使测试均方误差达到最小, 需要选择一种统计学习方法使其在测试集上偏差的平方和方差同时达到最小.
 - ① 使用灵活度更高的非参数模型和方法, 尽管所得模型的偏差会减小, 但是方差会增加, 会产生过拟合问题;
 - ② 如果使用灵活度更低的模型和方法, 会使得模型的偏差增加, 而方差减小, 产生欠拟合问题.

回归模型的评价准则: 偏差一方差权衡

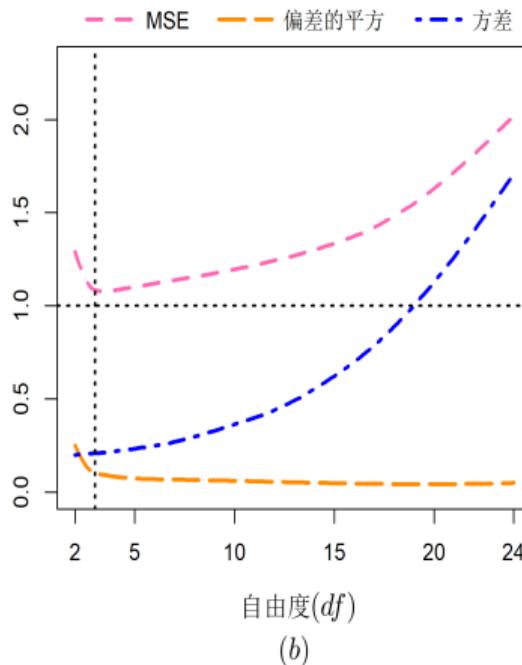
- 要使期望测试MSE达到最小, 需要选择一种统计学习方法对**偏差和方差权衡**, 使得偏差的平方和方差同时达到最小.
- 因此, 把这种寻找最优拟合模型的方法称为**偏差一方差权衡(bias-variance trade-off)**方法.
- 右图展示了模型灵活性与训练MSE和测试MSE之间的关系.



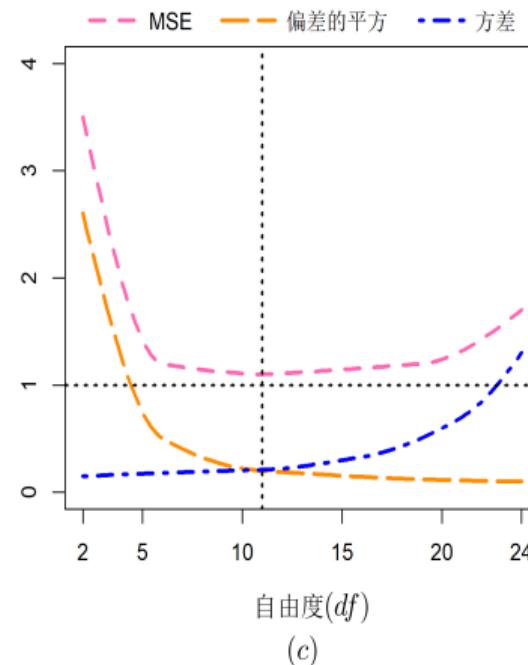
回归模型的评价准则: 偏差一方差权衡



(a)



(b)



(c)

三个例子的偏差的平方(橙色曲线), 方差(蓝色曲线), 测试均方误差(红色U型曲线), 不可约误差(水平黑色虚线), 垂直的黑色虚线表示最小测试均方误差所对应的自由度

正则化方法

■ 解决过拟合问题，可以使用**正则化**(regularization)方法，经常会在模型的拟合能力和复杂度之间进行权衡。

- * 拟合能力强的模型一般复杂度会比较高，容易导致过拟合；
- * 如果限制模型的复杂度，降低其拟合能力，又可能会导致欠拟合。

■ 正则化方法就是极小化下面的**惩罚经验风险函数**，即

$$\hat{g}_\lambda = \arg \min_g \left\{ R_n(g) + p_\lambda(g) \right\} = \arg \min_g \left\{ \frac{1}{n} \sum_{i=1}^n \ell_n(y_i, g(\mathbf{x}_i)) + p_\lambda(g) \right\},$$

- * $\ell_n(y_i, g(\mathbf{x}_i))$ 为**经验损失函数**， $p_\lambda(\cdot)$ 为**惩罚函数**， $\lambda \geq 0$ 为**正则参数或调节参数**，主要用来控制模型的复杂度或灵活度。

■ 如果回归函数 $g(\mathbf{x}_i, \boldsymbol{\beta})$ 是依赖于 p 维参数向量 $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ 的参数模型, 如多元线性回归模型, 则使用正则化方法估计参数模型中的未知参数向量, 即

$$\begin{aligned}\hat{\boldsymbol{\beta}}_\lambda &= \arg \min_{\boldsymbol{\beta}} \left\{ R_n(\boldsymbol{\beta}) + \sum_{j=1}^p p_\lambda(|\beta_j|) \right\} \\ &= \arg \min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \sum_{i=1}^n \ell_n(y_i, g(\mathbf{x}_i, \boldsymbol{\beta})) + \sum_{j=1}^p p_\lambda(|\beta_j|) \right\}.\end{aligned}$$

正则化方法

- 随着模型复杂度或灵活度的增加, 模型的拟合能力变强, 偏差减小而方差增大, 从而导致过拟合.
- 正则化方法可以通过调节参数 λ 来控制模型的复杂度或自由度.
 - ① 当 λ 变大时, 模型复杂度会降低, 可以有效地减小方差, 避免过拟合, 但偏差会上升;
 - ② 当 λ 过大时, 总的期望误差反而会上升.
- 因此, 一个好的调节参数 λ 需要在偏差和方差之间取得比较好的权衡.

本章纲要

1 回归模型及评价准则

- 回归模型
- 模型估计
- 回归模型精度的评价准则

2 分类模型及评价准则

3 参考文献

4 作业

■ 分类问题也是统计学习研究的主要内容之一, 分类学习方法有:

- 判别分析
- K 近邻分类方法
- logistic回归
- 支持向量机
- 分类树
- 随机森林
- 神经网络
-

分类模型及评价准则

■ 假设存在 n 个独立同分布的观测样本数据集, 记为 $D = \{(\mathbf{x}_i, y_i), i = 1, \dots, n\}$, 其中

- $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T \in \mathbb{R}^p$ 为观测的协变量向量;
- $y_i \in \{1, 2, \dots, J\}$ 为类别变量, 其中 $J \geq 2$.

■ 度量分类模型精度的准则为**准确率**(accuracy rate, accRate)和**错误率**(error rate, errRate), 分别定义为

$$\text{accRate} = \frac{1}{n} \sum_{i=1}^n I(y_i = \hat{y}_i), \quad \text{errRate} = \frac{1}{n} \sum_{i=1}^n I(y_i \neq \hat{y}_i),$$

其中 \hat{y}_i 为类别变量 y_i 的预测结果, $I(\cdot)$ 为示性函数.

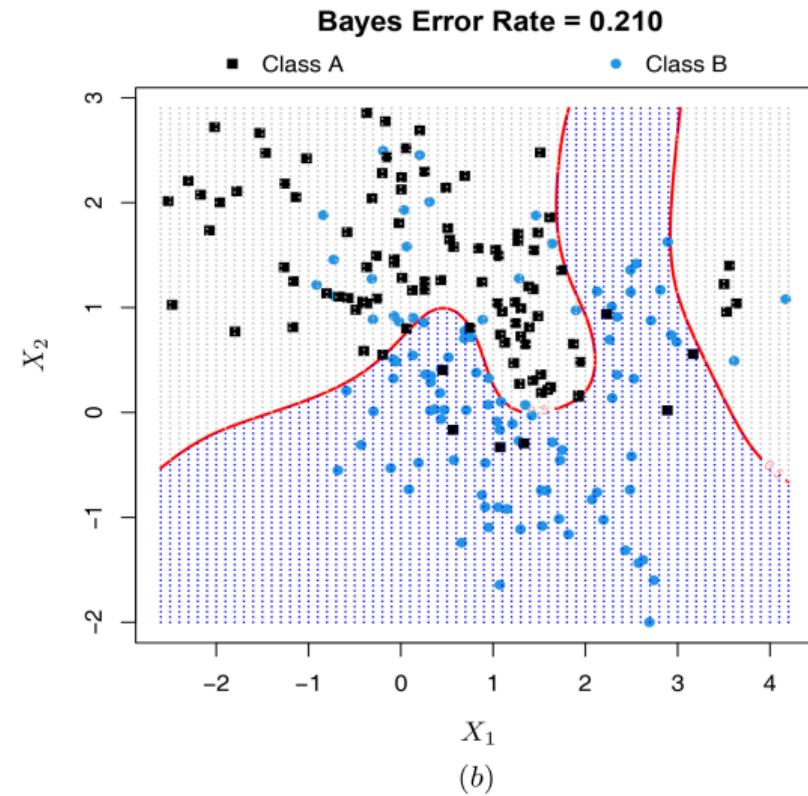
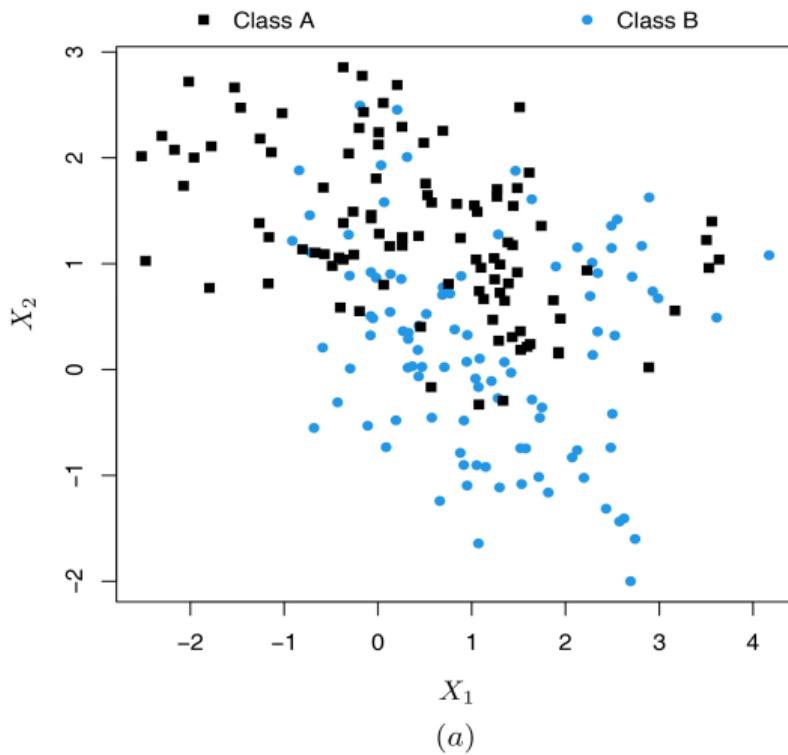
分类模型及评价准则展示

- 类似于回归模型, 分类模型依赖于**学习函数**或**学习分类器** $C(\mathbf{x})$, 并通过 $C(\mathbf{x})$ 对类别变量进行预测.
- 在统计学习中, **Bayes分类器**是一个最理想的分类方法, 其将每个观测值分配到它最大可能所在的类中, 并将这个类作为它的预测结果.
- 对 $j = 1, 2, \dots, J$, Bayes分类器的最优预测应该最大化下面**条件概率**或**后验概率**, 即

$$\max_j p_j(\mathbf{x}_i) = \max_j \mathbb{P}(y_i = j | \mathbf{X} = \mathbf{x}_i).$$

- 对于第 i 个类别变量 y_i , 如果估计的后验概率 $\hat{p}_j(\mathbf{x}_i)$ 最大, 则应选择预测结果为 $\hat{y}_i = \hat{C}(\mathbf{x}_i) = j$.
- 这种决策方式称为**Bayes最优决策**(Bayes optimal decision), 由此所得决策边界为**Bayes决策边界**(Bayes decision boundary).
- 使用**Bayes最优决策**, 所得错误率称为**Bayes错误率**(Bayes error rate).
- 以Hastie等(2009)的mixture.example模拟数据为例介绍**Bayes分类器方法**, 该数据是由协变量 X_1 和 X_2 构成二维空间的一个模拟数据.

分类模型及评价准则



■ 选择最优分类模型的标准：一个好的分类器应该是使测试错误率达到最小。

■ 解决的办法：

- 在训练集上利用统计学习方法得到学习函数或学习分类器 $C(x)$ 的估计 $\hat{C}(x)$ ，计算训练错误率(training errRate)；
- 基于测试样本或测试数据，计算测试错误率(test errRate)，也称为分类问题的泛化误差(generalization error)。

♠ 问题：如何计算训练错误率(training errRate)和测试错误率(test errRate)？

分类模型及评价准则

■ 将观测样本数据集 $D = \{(\mathbf{x}_i, y_i), i = 1, \dots, n\}$ 随机分成互不重叠的训练集和测试集，其中

- $D^{\text{Tr}} = \{(\mathbf{x}_i^{\text{Tr}}, y_i^{\text{Tr}}), i = 1, \dots, n_{\text{Tr}}\}$ 为训练集；
- $D^{\text{Te}} = \{(\mathbf{x}_i^{\text{Te}}, y_i^{\text{Te}}), i = 1, \dots, n_{\text{Te}}\}$ 为测试集，且 $n_{\text{Tr}} + n_{\text{Te}} = n$.

■ 在训练集 D^{Tr} 上用统计学习方法估计学习函数或学习分类器 $C(\mathbf{x})$ ，估计记为 $\hat{C}(\mathbf{x})$ ，并计算训练错误率为

$$\text{training errRate} = \frac{1}{n_{\text{Tr}}} \sum_{i=1}^{n_{\text{Tr}}} I(y_i^{\text{Tr}} \neq \hat{y}_i^{\text{Tr}}) = \frac{1}{n_{\text{Tr}}} \sum_{i=1}^{n_{\text{Tr}}} I(y_i^{\text{Tr}} \neq \hat{C}(\mathbf{x}_i^{\text{Tr}})),$$

其中 $\hat{y}_i^{\text{Tr}} = \hat{C}(\mathbf{x}_i^{\text{Tr}})$ 为训练集 D^{Tr} 上给定 \mathbf{x}_i^{Tr} 时的预测结果.

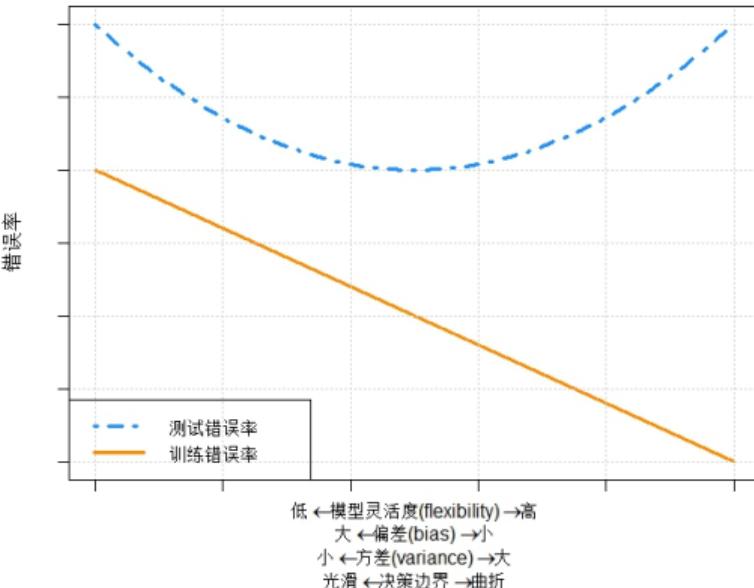
■ 最后, 用所得学习函数或学习分类器 $\hat{C}(\cdot)$ 用于测试集 D^{Te} , 并计算测试错误率

$$\text{test errRate} = \frac{1}{n_{\text{Te}}} \sum_{i=1}^{n_{\text{Te}}} I(y_i^{\text{Te}} \neq \hat{y}_i^{\text{Te}}) = \frac{1}{n_{\text{Te}}} \sum_{i=1}^{n_{\text{Te}}} I(y_i^{\text{Te}} \neq \hat{C}(\mathbf{x}_i^{\text{Te}})),$$

其中 $\hat{y}_i^{\text{Te}} = \hat{C}(\mathbf{x}_i^{\text{Te}})$ 为测试集 D^{Te} 上给定 \mathbf{x}_i^{Te} 时的预测结果.

分类模型及评价准则

- ① 如果选择灵活性很高的分类模型和方法, 可使得训练错误率最小, 决策边界会变得非常曲折, 分类结果偏差小而方差大, 产生过拟合问题;
- ② 如果选择灵活性更低的分类模型和方法, 决策边界会变得非常光滑, 会使得分类模型的偏差增加, 而方差减小, 产生欠拟合问题.



本章纲要

1 回归模型及评价准则

- 回归模型
- 模型估计
- 回归模型精度的评价准则

2 分类模型及评价准则

3 参考文献

4 作业

参考文献

- Hastie, T. and Tibshirani, R. (1990). Generalized Additive Models. London: Chapman and Hall.
- Hastie, T., Tibshirani, R. and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Edition). New York: Springer-Verlag.
- Huet, S., Bouvier, A., Poursat, M.-A. and Jolivet, E. (2004). Statistical Tools for Nonlinear Regression: A Practical Guide with S-PLUS and R Examples (Second Editon). New York: Springer-Verlag.
- James, G., Witten, D., Hastie, T. and Tibshirani, R. (2021). An Introduction to Statistical Learning with Applications in R. 2nd Ed. New York: Springer-Verlag.
- James, G., Witten, D., Hastie, T., Tibshirani, R. and Taylor, J. (2023). An Introduction to Statistical Learning with Applications in Python. New York: Springer-Verlag.

本章纲要

1 回归模型及评价准则

- 回归模型
- 模型估计
- 回归模型精度的评价准则

2 分类模型及评价准则

3 参考文献

4 作业

作业

[习题见教材: 统计学习(R语言版) — 习题2]

- **课后思考题:** 第2题、第4题、第6题、第8题
- **需要完成的课后作业:** 第1题、第3题、第5题、第7题
- **应用:** 第9题. 具体要求:
 - ① 能使用R语言产生模拟数据;
 - ② 能用学过的一些统计方法, 按照题目要求, 利用R语言对数据进行一些简单的分析, 并思考数据分析的结果.



谢谢, 请多提宝贵意见!