具体的模型选择方法

四、模型选择方法

Some popular model selection approaches

合格的模型选择方法应满足一些性质，如选择一致性（selection consistency）

假设 $\exists\ \mathcal{A}_0 \subset \mathcal{P} = \{2,3,\cdots,p\}$（去除第一个截距项），使对 $\forall j \in \mathcal{A}_0$ 有 $\beta_{[j]} \neq 0$

而 $\forall j \notin \mathcal{A}_0$ 有 $\beta_{[j]} = 0$

合格的模型选择方法应当基于数据产生一个 $\hat{\mathcal{A}}_0 \subset \{2,3,\cdots,p\}$ 满足

$$\lim_{n\to\infty} P(\hat{\mathcal{A}}_n = \mathcal{A}_0) = 1$$

一、全子集回归 All-subset Regression

考虑 $P=\{2,3,\cdots,p\}$ 的所有非空子集，使用标准 $M(\cdot)$ 选择：

$$\hat{\mathcal{A}}_n = \operatorname{argmin}_{\mathcal{A}\in\mathcal{F}} M(\mathcal{A})$$

下面描述一些记号：

(1) 记 $x_{i\mathcal{A}}$ 为 $x_i = (x_{i1},\cdots,x_{ip})^T$ 的子向量，只含 x_i 中 $\{x_{ij}: j\in\{1\}\cup\mathcal{A}\}$ 的分量

(2) 记 $\hat{\beta}_{\mathcal{A}}$ 表示数据集 $\{(y_i, x_{i\mathcal{A}}): i\in\mathcal{N}\}$ 的 OLSE $\mathcal{N}=\{1,2,\cdots,n\}$

之后，便要决定利用如何的标准 $M(\cdot)$

1.1 预测误差 prediction error

对于 $K \geq 2$，划分 $N = \{1,2,\cdots,n\}$ 为 K 个互斥子集 $N_1,\cdots,N_K$

对于 $k=1,2,\cdots,K$，记 $\hat{\beta}_A^{(-k)}$ 为使用 $\{(y_i, x_{iA}): i \notin N_k\}$ 估计出的 OLSE，则有：

$$M(A) = \sum_{k=1}^{K} \sum_{i \in N_k} (y_i - x_{iA}^T \cdot \hat{\beta}_A^{(-k)})^2$$

这便是交叉验证（cross validation）策略，同样它不限于线性模型：

$$M(A) = \sum_{k=1}^{K} \sum_{i \in N_k} (y_i - \hat{g}_A^{(-k)}(x_{iA}))^2$$

1.2 调整样本决定系数 Adjusted R^2

若以 R^2 作为 $M(\cdot)$，则对于 $A_1 \subset A_2$，必然有 $M(A_1) \leq M(A_2)$

故要对自变量个数施加惩罚

令 $R^2(A)$ 为使用 $\{(y_i, x_{iA}): i \in N\}$ 计算出的 R^2，调整为：

$$R_a^2(A) = 1 - \frac{(n-1)\cdot(1-R^2(A))}{n-|A|-1} = 1 - \frac{(n-1)\cdot SSE_A}{(n-|A|-1)\cdot SST}$$

$$1 - \frac{(n-1)\left(1 - \frac{SST - SSE_A}{SST}\right)}{n-|A|-1} = 1 - \frac{(n-1)\, SSE_A}{(n-|A|-1)\, SST}$$

其中：

$$SSE_A = \sum_{i=1}^{n} (y_i - x_{iA}^T \cdot \hat{\beta}_A)^2$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Tips: $A\uparrow \begin{cases} \Rightarrow SSE_A \downarrow \Rightarrow R_a^2(A)\uparrow \\ \Rightarrow |A|\uparrow \Rightarrow R_a^2(A)\downarrow \end{cases}$

故可将 $M(A) = -R_a^2(A)$ 作为标准

1.3 赤池信息量 Akaike information criterion, AIC

AIC 是基于 Likelihood 函数值的选择标准。

Definition AIC 定义：假设希望基于数据 Z 估计参数 $\theta \in \mathbb{R}^d$

而 Z 的似然函数为 $L(Z;\theta)$，由此得到 MLE 为 $\hat{\theta}$，则 AIC 为：

$$AIC = -2\log L(Z;\hat{\theta}) + 2d$$

以上 AIC 由 2 部分组成：第一部分表示模型拟合的好坏、$L(Z;\hat{\theta})$ 越大，拟合越好；第二部分是对模型复杂的惩罚。

在正态假设下，基于 $\{(y_i, x_{iA}): i\in N\}$ 的选模型 A 的参数 MLE

分别为 OLSE：$\hat{\beta}_A$ $\hat{\sigma}_A^2 = \frac{SSE_A}{n}$ 代入 $L(Z;\hat{\theta})$ 并去除无关项，得

$$AIC(A) = n\log SSE_A + 2\cdot|A|$$

特别地：$d = |A| + 2$ （$2,\dots,p$ + 截距 + $\hat{\sigma}_A^2$）

故最小化 $M(\mathcal{A}) = AIC(\mathcal{A})$

1.4 C_p 统计量 C_p - Statistic

若 $\mathcal{A}$ 正确，则 $x_{i\mathcal{A}}^T \cdot \hat{\beta}_{\mathcal{A}}$ 可视为 $E(y_i) = x_i^T\beta$ 的估计 $(i \in \mathcal{N})$

而估计偏差平方和与随机误差方差的比值为：

$$T_p(\mathcal{A}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(\hat{x}_{i\mathcal{A}}^T \cdot \hat{\beta}_{\mathcal{A}} - x_i^T\beta)^2$$

可以证明：

$$E[T_p(\mathcal{A})] = \frac{E(SSE_{\mathcal{A}})}{\sigma^2} - n + 2(|\mathcal{A}|+1)$$

略去无关的项得 C_p 统计量

$$C_p(\mathcal{A}) = \frac{SSE_{\mathcal{A}}}{\hat{\sigma}^2} + 2|\mathcal{A}|$$

其中 $\hat{\sigma}^2$ 为全模型 σ^2 的无偏估计：

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2$$

故选择最小化 $M(\mathcal{A}) = C_p(\mathcal{A})$

Tips. 由定理3，若 $\mathcal{A}$ 正确，有 $\beta_{[j]} = 0 \quad \forall j \notin \mathcal{A}$，则

$$E(SSE_{\mathcal{A}}) = (n-|\mathcal{A}|-1)\sigma^2$$

结合 $E(T_p(\mathcal{A}))$ 知

$$E[T_p(\mathcal{A})] = |\mathcal{A}|+1$$

故

$$M(A) = \left| SSE_A/\hat{\sigma}^2 - n + 2(|A|+1) - (|A|+1) \right|$$
$$= \left| SSE_A/\hat{\sigma}^2 - n + |A| + 1 \right|$$

也是合适的

二、逐步回归 Stepwise Regression