



统计与机器学习

第四章：变量选择

倪 蓓

DaSE@ECNU
(lni@dase.ecnu.edu.cn)



目录

① 自变量选择的影响

欠拟合

过拟合

② 自变量选择的准则

③ 逐步回归

自变量选择的影响

概述

- 预测是回归分析的主要用途之一.
- 在预测时, 我们常常希望预测值的均方误差比较小.
 - 建模时丢失了一些重要的变量, 导致模型拟合不足, 预测偏差大, 这称为**欠拟合**;
 - 建模时容纳过多不重要的变量, 导致模型过度拟合, 泛化能力差, 这称为**过拟合**;
- **如何找到最合适的回归模型?**

自变量选择的影响

全子集回归法

- 假定共有 p 个自变量 x_1, x_2, \dots, x_p , 我们可以考虑 p_1 个自变量纳入模型, 即

p_1	自变量的组合	个数
1	$\{x_1\}, \{x_2\}, \dots, \{x_p\}$	$C_p^1 = p$
2	$\{x_1, x_2\}, \dots, \{x_{p-1}, x_p\}$	$C_p^2 = \frac{p(p-1)}{2}$
\vdots	\vdots	\vdots
$p-1$	$\{x_2, \dots, x_p\}, \dots, \{x_1, \dots, x_{p-1}\}$	$C_p^{p-1} = p$
p	$\{x_1, x_2, \dots, x_p\}$	$C_p^p = 1$

- 因此, 需要考虑 $C_p^1 + \dots + C_p^p = 2^p - 1$ 个回归模型;
- 全子集回归法**是最简单直观, 但也是最繁琐的方法.

自变量选择的影响

基本定义

- 由于共有 p 个自变量纳入模型，我们将由所有 p 个自变量构造的回归模型，定义为**全模型**，即

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

其中,

- $\mathbf{y} = (y_1, y_2, \dots, y_n)'$;
- $\mathbf{X}_p = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$;
- $\boldsymbol{\beta}_p = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)'$
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)'$

自变量选择的影响

基本定义

- 从上述 p 个自变量中挑选出 p_1 个 ($p_1 < p$), 我们将由这 p_1 个自变量构造的回归模型, 定义为**选模型**, 即

$$\mathbf{y} = \mathbf{X}_{p_1} \boldsymbol{\beta}_{p_1} + \boldsymbol{\varepsilon}$$

其中,

- $\mathbf{y} = (y_1, y_2, \dots, y_n)'$;
- $\mathbf{X}_{p_1} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p_1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p_1} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np_1} \end{pmatrix}$
- $\boldsymbol{\beta}_{p_1} = (\beta_0, \beta_1, \beta_2, \dots, \beta_{p_1})'$
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)'$

自变量选择的影响

说明

- 在选模型中, p_1 个自变量 x_1, x_2, \dots, x_{p_1} 并不一定是全体 p 个自变量 x_1, x_2, \dots, x_p 中的前 p_1 个;
- **按某种规则**, 从 p 个自变量 x_1, x_2, \dots, x_p 中挑选出来的 p_1 个.
- 为了简化, 不妨认为 x_1, x_2, \dots, x_{p_1} 就是 x_1, x_2, \dots, x_p 中的前 p_1 个.

自变量选择的影响

自变量选择：全模型 vs 选模型

- 自变量的选择，看成对一个实际问题是用**全模型**，还是用**选模型**；
- 如果**全模型**是正确的，而错误地使用选模型，那么我们实际上丢失了一些重要且有用的自变量，在这种情况下认为是**欠拟合**；
- 如果**选模型**是正确的，而错误地使用全模型，那么我们实际上引入了一些不必要的自变量，在这种情况下认为是**过拟合**；

自变量选择的影响

自变量选择：全模型 vs 选模型

- 接下来，我们分别讨论欠拟合和过拟合的情况下，参数估计和拟合值会有什么变化？
- 在采用全模型的情况下，
 - β 和 σ^2 的估计分别为

$$\begin{aligned}\hat{\beta}_p &= (\mathbf{X}'_p \mathbf{X}_p)^{-1} \mathbf{X}'_p \mathbf{y} \\ \hat{\sigma}_p^2 &= \frac{1}{n - p - 1} SS_E^p \\ &= \frac{1}{n - p - 1} (\mathbf{y} - \hat{\mathbf{y}}_p)' (\mathbf{y} - \hat{\mathbf{y}}_p)\end{aligned}$$

- 在 $\mathbf{x}_{p,0} = (\mathbf{x}'_{p_1,0}, \mathbf{z}'_0)'$ 时的预测值为

$$\hat{\mathbf{y}}_0 = \mathbf{x}'_{p,0} \hat{\beta}_p$$

自变量选择的影响

自变量选择：全模型 vs 选模型

- 接下来，我们分别讨论欠拟合和过拟合的情况下，参数估计和拟合值会有什么变化？
- 在采用选模型的情况下，
 - β 和 σ^2 的估计分别为

$$\begin{aligned}\hat{\beta}_{p_1} &= (\mathbf{X}'_{p_1} \mathbf{X}_{p_1})^{-1} \mathbf{X}'_{p_1} \mathbf{y} \\ \hat{\sigma}_{p_1}^2 &= \frac{1}{n - p_1 - 1} SS_E^{p_1} \\ &= \frac{1}{n - p_1 - 1} (\mathbf{y} - \hat{\mathbf{y}}_{p_1})' (\mathbf{y} - \hat{\mathbf{y}}_{p_1})\end{aligned}$$

- 在 $\mathbf{x}_{p,0} = (\mathbf{x}'_{p_1,0}, \mathbf{z}'_0)'$ 时的预测值为

$$\hat{\mathbf{y}}_0 = \mathbf{x}'_{p_1,0} \hat{\beta}_{p_1}$$

欠拟合

模型

- 假定**全模型**为真，即

$$\begin{aligned} \mathbf{y} &= \mathbf{X}_p \boldsymbol{\beta}_p + \boldsymbol{\varepsilon} \\ &= (\mathbf{X}_{p_1} \quad \mathbf{Z}) \begin{pmatrix} \boldsymbol{\beta}_{p_1} \\ \boldsymbol{\gamma} \end{pmatrix} + \boldsymbol{\varepsilon} \\ &= \mathbf{X}_{p_1} \boldsymbol{\beta}_{p_1} + \mathbf{Z} \boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \end{aligned}$$

- 而我们错误地使用了**选模型**，即

$$\mathbf{y} = \mathbf{X}_{p_1} \boldsymbol{\beta}_{p_1} + \boldsymbol{\varepsilon}.$$

- 我们认为丢失了重要的自变量，即假定
 - $\text{rank}(\mathbf{X}_p) > \text{rank}(\mathbf{X}_{p_1})$;
 - $\boldsymbol{\gamma} \neq \mathbf{0}'_{p-p_1}$.

欠拟合

模型

- 假定**全模型**为真，而我们错误地使用了**选模型**.
- 参数估计为

$$\begin{aligned}\hat{\beta}_{p_1} &= (\mathbf{X}'_{p_1} \mathbf{X}_{p_1})^{-1} \mathbf{X}'_{p_1} \mathbf{y} \\ \hat{\sigma}_{p_1}^2 &= \frac{1}{n - p_1 - 1} (\mathbf{y} - \hat{\mathbf{y}}_{p_1})' (\mathbf{y} - \hat{\mathbf{y}}_{p_1})\end{aligned}$$

- 预测值为

$$\hat{y}_0 = \mathbf{x}'_{p_1,0} \hat{\beta}_{p_1}$$

欠拟合

参数估计—— $\hat{\beta}_{p_1}$

- 考虑 $\hat{\beta}_{p_1}$ 的期望，即

$$\begin{aligned} E\left(\hat{\beta}_{p_1}\right) &= \left(\mathbf{X}_{p_1}' \mathbf{X}_{p_1}\right)^{-1} \mathbf{X}_{p_1}' E(\mathbf{y}) \\ &= \left(\mathbf{X}_{p_1}' \mathbf{X}_{p_1}\right)^{-1} \mathbf{X}_{p_1}' E\left(\mathbf{X}_{p_1} \beta_{p_1} + \mathbf{Z} \gamma + \varepsilon\right) \\ &= \left(\mathbf{X}_{p_1}' \mathbf{X}_{p_1}\right)^{-1} \mathbf{X}_{p_1}' \left(\mathbf{X}_{p_1} \beta_{p_1} + \mathbf{Z} \gamma + E(\varepsilon)\right) \\ &= \left(\mathbf{X}_{p_1}' \mathbf{X}_{p_1}\right)^{-1} \mathbf{X}_{p_1}' \left(\mathbf{X}_{p_1} \beta_{p_1} + \mathbf{Z} \gamma\right) \\ &= \beta_{p_1} + \left(\mathbf{X}_{p_1}' \mathbf{X}_{p_1}\right)^{-1} \mathbf{X}_{p_1}' \mathbf{Z} \gamma \end{aligned}$$

- 结论：
 - 因为 $\gamma \neq 0$ ，通常来说， $\hat{\beta}_{p_1}$ 是有偏估计；
 - 偏差为 $\left(\mathbf{X}_{p_1}' \mathbf{X}_{p_1}\right)^{-1} \mathbf{X}_{p_1}' \mathbf{Z} \gamma$.

欠拟合

参数估计—— $\hat{\beta}_{p_1}$

- 注意到

$$E\left(\hat{\beta}_{p_1}\right) = \beta_{p_1} + (\mathbf{X}_{p_1}' \mathbf{X}_{p_1})^{-1} \mathbf{X}_{p_1}' \mathbf{Z} \gamma$$

- 如果 $\mathbf{X}_{p_1}' \mathbf{Z} = 0$, 那么 $(\mathbf{X}_{p_1}' \mathbf{X}_{p_1})^{-1} \mathbf{X}_{p_1}' \mathbf{Z} \gamma = 0$. 此时, $\hat{\beta}_{p_1}$ 是无偏估计.

欠拟合

参数估计—— $\hat{\sigma}_{p_1}^2$

- 考虑 SS_E^p 与 $SS_E^{p_1}$ 的关系.
- 注意到

$$SS_E^p = \mathbf{y}'(\mathbf{I} - \mathbf{H}_{\mathbf{X}_p})\mathbf{y} \quad SS_E^{p_1} = \mathbf{y}'(\mathbf{I} - \mathbf{H}_{\mathbf{X}_{p_1}})\mathbf{y}$$

- 其中

$$\begin{aligned} \mathbf{H}_{\mathbf{X}_p} &= \mathbf{X}_p(\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p' \\ &= \begin{pmatrix} \mathbf{X}_{p_1} & \mathbf{Z} \end{pmatrix} \begin{pmatrix} \mathbf{X}_{p_1}'\mathbf{X}_{p_1} & \mathbf{X}_{p_1}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X}_{p_1} & \mathbf{Z}'\mathbf{Z} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}_{p_1}' \\ \mathbf{Z}' \end{pmatrix} \end{aligned}$$

欠拟合

参数估计—— $\hat{\sigma}_{p_1}^2$

$$\begin{aligned}H_{X_p} &= (X_{p_1} \quad Z) \begin{pmatrix} X'_{p_1} X_{p_1} & X'_{p_1} Z \\ Z' X_{p_1} & Z' Z \end{pmatrix}^{-1} \begin{pmatrix} X'_{p_1} \\ Z' \end{pmatrix} \\&= (X_{p_1} \quad Z) \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} X'_{p_1} \\ Z' \end{pmatrix} \\&= H_{X_{p_1}} + H_{X_{p_1}} Z (Z' N_{X_{p_1}} Z)^{-1} Z' H_{X_{p_1}} \\&\quad - H_{X_{p_1}} Z (Z' N_{X_{p_1}} Z)^{-1} Z' - Z (Z' N_{X_{p_1}} Z)^{-1} Z' H_{X_{p_1}} + Z (Z' N_{X_{p_1}} Z)^{-1} Z' \\&= H_{X_{p_1}} + N_{X_{p_1}} Z (Z' N_{X_{p_1}} Z)^{-1} Z' N_{X_{p_1}}\end{aligned}$$

其中,

$$\begin{aligned}D &= (Z' Z - Z' X_{p_1} (X'_{p_1} X_{p_1})^{-1} X'_{p_1} Z)^{-1} = (Z' (I - H_{X_{p_1}}) Z)^{-1} = (Z' N_{X_{p_1}} Z)^{-1} \\A &= (X'_{p_1} X_{p_1})^{-1} + (X'_{p_1} X_{p_1})^{-1} X'_{p_1} Z (Z' N_{X_{p_1}} Z)^{-1} Z' X_{p_1} (X'_{p_1} X_{p_1})^{-1} \\B &= -(X'_{p_1} X_{p_1})^{-1} X'_{p_1} Z (Z' N_{X_{p_1}} Z)^{-1} \\C &= -(Z' N_{X_{p_1}} Z)^{-1} Z' X_{p_1} (X'_{p_1} X_{p_1})^{-1}\end{aligned}$$

欠拟合

参数估计—— $\hat{\sigma}_{p_1}^2$

- SS_E^p 与 $SS_E^{p_1}$ 的关系为

$$\begin{aligned}SS_E^{p_1} &= SS_E^p + \mathbf{y}'(\mathbf{H}_{X_p} - \mathbf{H}_{X_{p_1}})\mathbf{y} \\&= SS_E^p + \mathbf{y}'\mathbf{N}_{X_{p_1}}\mathbf{Z}(\mathbf{Z}'\mathbf{N}_{X_{p_1}}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{N}_{X_{p_1}}\mathbf{y}\end{aligned}$$

- $SS_E^{p_1}$ 的期望为

$$\begin{aligned}E(SS_E^{p_1}) &= E(SS_E^p) + E\left(\mathbf{y}'\mathbf{N}_{X_{p_1}}\mathbf{Z}(\mathbf{Z}'\mathbf{N}_{X_{p_1}}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{N}_{X_{p_1}}\mathbf{y}\right) \\&= (n-p-1)\sigma^2 + E(\mathbf{y})'\mathbf{N}_{X_{p_1}}\mathbf{Z}(\mathbf{Z}'\mathbf{N}_{X_{p_1}}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{N}_{X_{p_1}}E(\mathbf{y}) \\&\quad + \sigma^2\text{tr}\left(\mathbf{N}_{X_{p_1}}\mathbf{Z}(\mathbf{Z}'\mathbf{N}_{X_{p_1}}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{N}_{X_{p_1}}\right) \\&= (n-p-1)\sigma^2 + \boldsymbol{\gamma}'\mathbf{Z}'\mathbf{N}_{X_{p_1}}\mathbf{Z}\boldsymbol{\gamma} + (p-p_1)\sigma^2 \\&= (n-p_1-1)\sigma^2 + \boldsymbol{\gamma}'\mathbf{Z}'\mathbf{N}_{X_{p_1}}\mathbf{Z}\boldsymbol{\gamma}\end{aligned}$$

参数估计—— $\hat{\sigma}_{p_1}^2$

欠拟合

预测值的方差更小

- 注意到

但 σ 的参数估计更大

$$E(\hat{\sigma}_{p_1}^2) = \frac{1}{n - p_1 - 1} E(SS_E^{p_1}) = \sigma^2 + \frac{\gamma' Z' N_{X_{p_1}} Z \gamma}{n - p_1 - 1} > \sigma^2$$

- 上式中不等号是“严格的”，这是因为之前假设

$$\text{rank}(\mathbf{X}_p) > \text{rank}(\mathbf{X}_{p_1})$$

即 $N_{X_{p_1}} \mathbf{Z} \neq 0$. 因此, $\hat{\sigma}_{p_1}^2$ 不是 σ^2 的无偏估计.

欠拟合

预测值

在 $\mathbf{x}_{p,0} = (\mathbf{x}'_{p_1,0}, \mathbf{z}'_0)'$ 时, 如何预测

$$y_0 = \mathbf{x}'_{p_1,0} \boldsymbol{\beta}_{p_1} + \mathbf{z}'_0 \boldsymbol{\gamma} + \varepsilon_0 \quad ?$$

- 如果我们知道全模型是正确的, 那么就应该采用全模型. 此时, y_0 最为合理的预测为

$$\hat{y}_{0,T} = \mathbf{x}'_{p,0} \hat{\boldsymbol{\beta}}_p$$

其期望和方差分别为

$$E(\hat{y}_{0,T}) = \mathbf{x}'_{p,0} E(\hat{\boldsymbol{\beta}}_p) = \mathbf{x}'_{p_1,0} \boldsymbol{\beta}_{p_1} + \mathbf{z}'_0 \boldsymbol{\gamma}$$

$$\text{Var}(\hat{y}_{0,T}) = \sigma^2 \mathbf{x}'_{p,0} (\mathbf{X}'_p \mathbf{X}_p)^{-1} \mathbf{x}_{p,0}$$

欠拟合

预测值

在 $\mathbf{x}_{p,0} = (\mathbf{x}'_{p_1,0}, \mathbf{z}'_0)'$ 时, 如何预测

$$y_0 = \mathbf{x}'_{p_1,0} \boldsymbol{\beta}_{p_1} + \mathbf{z}'_0 \boldsymbol{\gamma} + \varepsilon_0 \quad ?$$

- 但是, 我们错误地使用了选模型. 此时, y_0 的预测为

$$\hat{y}_0 = \mathbf{x}'_{p_1,0} \hat{\boldsymbol{\beta}}_{p_1}$$

其期望为

$$\begin{aligned} E(\hat{y}_0) &= \mathbf{x}'_{p_1,0} \left(\boldsymbol{\beta}_{p_1} + (\mathbf{X}'_{p_1} \mathbf{X}_{p_1})^{-1} \mathbf{X}'_{p_1} \mathbf{Z} \boldsymbol{\gamma} \right) \\ &= \mathbf{x}'_{p_1,0} \boldsymbol{\beta}_{p_1} + \mathbf{x}'_{p_1,0} (\mathbf{X}'_{p_1} \mathbf{X}_{p_1})^{-1} \mathbf{X}'_{p_1} \mathbf{Z} \boldsymbol{\gamma} \end{aligned}$$

欠拟合

预测值

- 在全模型下, y_0 预测值的期望为

$$E(\hat{y}_{0,T}) = \mathbf{x}'_{p_1,0} \boldsymbol{\beta}_{p_1} + \mathbf{z}'_0 \boldsymbol{\gamma}$$

- 在选模型下, y_0 预测值的期望为

$$E(\hat{y}_0) = \mathbf{x}'_{p_1,0} \boldsymbol{\beta}_{p_1} + \mathbf{x}'_{p_1,0} (\mathbf{X}'_{p_1} \mathbf{X}_{p_1})^{-1} \mathbf{X}'_{p_1} \mathbf{Z} \boldsymbol{\gamma}$$

- 偏差为

$$\left(\mathbf{x}'_{p_1,0} (\mathbf{X}'_{p_1} \mathbf{X}_{p_1})^{-1} \mathbf{X}'_{p_1} \mathbf{Z} - \mathbf{z}'_0 \right) \boldsymbol{\gamma}$$

欠拟合

预测值

- 两个预测值方差的差异，即

$$\begin{aligned}\text{Var}(\hat{y}_{0,T}) &= \sigma^2 (\mathbf{x}'_{p_1,0}, z'_0) (\mathbf{X}'_p \mathbf{X}_p)^{-1} (\mathbf{x}'_{p_1,0}, z'_0)' \\ &= \sigma^2 (\mathbf{x}'_{p_1,0}, z'_0) \begin{pmatrix} \mathbf{X}'_{p_1} \mathbf{X}_{p_1} & \mathbf{X}'_{p_1} \mathbf{Z} \\ \mathbf{Z}' \mathbf{X}_{p_1} & \mathbf{Z}' \mathbf{Z} \end{pmatrix}^{-1} (\mathbf{x}'_{p_1,0}, z'_0)' \\ &= \sigma^2 (\mathbf{x}'_{p_1,0}, z'_0) \mathbf{A} (\mathbf{x}'_{p_1,0}, z'_0)'\end{aligned}$$

其中，

$$\mathbf{A} = \begin{pmatrix} (\mathbf{X}'_{p_1} \mathbf{X}_{p_1})^{-1} + \mathbf{L} \mathbf{M} \mathbf{L}' & -\mathbf{L} \mathbf{M} \\ -\mathbf{M} \mathbf{L}' & \mathbf{M} \end{pmatrix}$$

$$\mathbf{L} = (\mathbf{X}'_{p_1} \mathbf{X}_{p_1})^{-1} \mathbf{X}'_{p_1} \mathbf{Z} \quad \text{和} \quad \mathbf{M} = (\mathbf{Z}' \mathbf{N}_{\mathbf{X}_{p_1}} \mathbf{Z})^{-1}$$

欠拟合

预测值

- 两个预测值方差的差异，即

$$\begin{aligned}\text{Var}(\hat{y}_{0,T}) &= \sigma^2 (\mathbf{x}'_{p_1,0}, \mathbf{z}'_0) \mathbf{A} (\mathbf{x}'_{p_1,0}, \mathbf{z}'_0)' \\ &= \sigma^2 \mathbf{x}'_{p_1,0} (\mathbf{X}'_{p_1} \mathbf{X}_{p_1})^{-1} \mathbf{x}_{p_1,0} \\ &\quad + \sigma^2 (\mathbf{L}' \mathbf{x}_{p_1,0} - \mathbf{z}_0)' \mathbf{M} (\mathbf{L}' \mathbf{x}_{p_1,0} - \mathbf{z}_0) \\ &\geq \sigma^2 \mathbf{x}'_{p_1,0} (\mathbf{X}'_{p_1} \mathbf{X}_{p_1})^{-1} \mathbf{x}_{p_1,0} \\ &= \text{Var}(\hat{y}_0)\end{aligned}$$

- 结论：在选模型下所得到的预测方差 $\text{Var}(\hat{y}_0)$ 比“真实的”方差 $\text{Var}(\hat{y}_{0,T})$ 更小.

过拟合

模型

- 假定**选模型**为真，即

$$\begin{aligned} \mathbf{y} &= \mathbf{X}_{p_1} \boldsymbol{\beta}_{p_1} + \boldsymbol{\varepsilon} \\ &= (\mathbf{X}_{p_1} \quad \mathbf{Z}) \begin{pmatrix} \boldsymbol{\beta}_{p_1} \\ \mathbf{0} \end{pmatrix} + \boldsymbol{\varepsilon} \end{aligned}$$

- 而我们错误地使用了**全模型**，即

$$\mathbf{y} = (\mathbf{X}_{p_1} \quad \mathbf{Z}) \begin{pmatrix} \boldsymbol{\beta}_{p_1} \\ \boldsymbol{\gamma} \end{pmatrix} + \boldsymbol{\varepsilon}$$

过拟合

参数估计—— $\hat{\beta}_p$

- 考虑 $\hat{\beta}_p$ 的期望, 即

$$\begin{aligned} E(\hat{\beta}_p) &= (\mathbf{X}_p' \mathbf{X}_p)^{-1} \mathbf{X}_p' E(\mathbf{y}) \\ &= (\mathbf{X}_p' \mathbf{X}_p)^{-1} \mathbf{X}_p' \left((\mathbf{X}_{p1} \quad \mathbf{Z}) \begin{pmatrix} \beta_{p1} \\ \mathbf{0} \end{pmatrix} + E(\boldsymbol{\epsilon}) \right) \\ &= (\mathbf{X}_p' \mathbf{X}_p)^{-1} \mathbf{X}_p' (\mathbf{X}_{p1} \quad \mathbf{Z}) \begin{pmatrix} \beta_{p1} \\ \mathbf{0} \end{pmatrix} \\ &= (\mathbf{X}_p' \mathbf{X}_p)^{-1} \mathbf{X}_p' \mathbf{X}_p \begin{pmatrix} \beta_{p1} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \beta_{p1} \\ \mathbf{0} \end{pmatrix} \end{aligned}$$

- 结论: $\hat{\beta}_p$ 是无偏估计.

过拟合

参数估计—— $\hat{\sigma}_p^2$

- 考虑 SS_E^p 的期望, 即

$$\begin{aligned} E(SS_E^p) &= E(\mathbf{y}' \mathbf{N}_{X_p} \mathbf{y}) \\ &= E(\mathbf{y})' \mathbf{N}_{X_p} E(\mathbf{y}) + \sigma^2 \text{tr}(\mathbf{N}_{X_p}) \\ &= (\boldsymbol{\beta}_p' \quad \mathbf{0}') \mathbf{X}_p' \mathbf{N}_{X_p} \mathbf{X}_p \begin{pmatrix} \boldsymbol{\beta}_p \\ \mathbf{0} \end{pmatrix} + (n - p - 1)\sigma^2 \\ &= (n - p - 1)\sigma^2 \end{aligned}$$

- 结论: $\hat{\sigma}_p^2 = \frac{SS_E^p}{n-p-1}$ 是 σ^2 的无偏估计.

过拟合

预测值

在 $\mathbf{x}_{p,0} = (\mathbf{x}'_{p_1,0}, z'_0)'$ 时, 如何预测

$$y_0 = \mathbf{x}'_{p_1,0} \boldsymbol{\beta}_{p_1} + \varepsilon_0 \quad ?$$

- 如果我们知道选模型是正确的, 那么就应该采用选模型. 此时, y_0 最为合理的预测为

$$\hat{y}_{0,T} = \mathbf{x}'_{p_1,0} \hat{\boldsymbol{\beta}}_{p_1}$$

其期望和方差分别为

$$E(\hat{y}_{0,T}) = \mathbf{x}'_{p_1,0} E(\hat{\boldsymbol{\beta}}_{p_1}) = \mathbf{x}'_{p_1,0} \boldsymbol{\beta}_{p_1}$$

$$\text{Var}(\hat{y}_{0,T}) = \sigma^2 \mathbf{x}'_{p_1,0} (\mathbf{X}'_{p_1} \mathbf{X}_{p_1})^{-1} \mathbf{x}_{p_1,0}$$

过拟合

预测值

在 $\mathbf{x}_{p,0} = (\mathbf{x}'_{p_1,0}, \mathbf{z}'_0)'$ 时, 如何预测

$$y_0 = \mathbf{x}'_{p_1,0} \boldsymbol{\beta}_{p_1} + \varepsilon_0 \quad ?$$

- 但是, 我们错误地使用了全模型. 此时, y_0 的预测值为

$$\hat{y}_0 = \mathbf{x}'_{p,0} \hat{\boldsymbol{\beta}}_p$$

其期望为

$$\begin{aligned} E(\hat{y}_0) &= \mathbf{x}'_{p,0} E(\hat{\boldsymbol{\beta}}_p) = (\mathbf{x}'_{p_1,0} \quad \mathbf{z}'_0) \begin{pmatrix} \boldsymbol{\beta}_{p_1} \\ \mathbf{0} \end{pmatrix} \\ &= \mathbf{x}'_{p_1,0} \boldsymbol{\beta}_{p_1} = E(y_{0,T}) \end{aligned}$$

过拟合

预测值

- \hat{y}_0 的方差

$$\begin{aligned}\text{Var}(\hat{y}_0) &= \text{Var}\left(\mathbf{x}'_{p,0} (\mathbf{X}'_p \mathbf{X}_p)^{-1} \mathbf{X}'_p \mathbf{y}\right) \\&= \sigma^2 \mathbf{x}'_{p,0} (\mathbf{X}'_p \mathbf{X}_p)^{-1} \mathbf{x}_{p,0} \\&= \sigma^2 \mathbf{x}'_{p,0} \begin{pmatrix} (\mathbf{X}'_{p_1} \mathbf{X}_{p_1})^{-1} + \mathbf{L} \mathbf{M} \mathbf{L}' & -\mathbf{L} \mathbf{M} \\ -\mathbf{M} \mathbf{L}' & \mathbf{M} \end{pmatrix} \mathbf{x}_{p,0} \\&= \sigma^2 \mathbf{x}'_{p_1,0} (\mathbf{X}'_{p_1} \mathbf{X}_{p_1})^{-1} \mathbf{x}_{p_1,0} \\&\quad + \sigma^2 (\mathbf{L}' \mathbf{x}_{p_1,0} - \mathbf{z}_0)' \mathbf{M} (\mathbf{L}' \mathbf{x}_{p_1,0} - \mathbf{z}_0) \\&\geq \text{Var}(\hat{y}_{0,T})\end{aligned}$$

- 结论: \hat{y}_0 的方差比 $\hat{y}_{0,T}$ 的方差更大.

自变量选择的影响

结论

- 从**预测**的角度来看待变量选择的问题，一个回归模型并不是考虑的自变量越多越好.
- 在建立回归模型是，选择自变量的基本指导思想是**少而精**.
 - 在回归模型中考虑过少的自变量，虽然预测值较为稳定，但是预测值会产生较大的偏差；
 - 在回归模型中选择过多的自变量，虽然预测值无明显的偏差，但是会引起较大的波动.
- 在选择自变量时，往往需要兼顾预测**方差**和预测**偏差**，并考虑选择有实际意义的自变量.

自变量选择的影响

总结

	欠拟合	过拟合
估计	$\hat{\beta}_{p_1}$ 关于 β_{p_1} 是有偏的 $\hat{\sigma}_{p_1}^2$ 是有偏的	$\hat{\beta}_p$ 关于 $(\beta_{p_1}, \mathbf{0}')'$ 是无偏的 $\hat{\sigma}_p^2$ 是无偏的
预测	有偏预测 预测方差小	无偏预测 预测方差大

自变量选择的准则

概述

- 在一个实际问题中, 有 p 个可供选择的自变量.
- 由于每一个自变量都有**入选**和**不入选**两种情况.
- 选模型包含的自变量数目 p_1 有从 0 到 p 共有 $(p + 1)$ 种不同情况, 而对选模型中包含 p_1 个自变量对情况, 从全部 p 个自变量选出 p_1 个的方法共有 $C_p^{p_1}$ 个, 因而所有选模型的数目为

$$C_p^0 + C_p^1 + C_p^2 + \cdots + C_p^p = 2^p$$

- 这里, 将回归模型中只包含常数项的情况也考虑在内.

自变量选择的准则

概述

- 因此，在有 p 个自变量的回归模型中，一切可能的回归子集共有 2^p 个.
- 我们关心的问题：在所有的回归子集中如何选择一个最优的回归子集？
- 具体来说，
 - 在所有的回归子集中，哪个回归子集是最优的？
 - 依据怎样的标准来定义最优子集的？

自变量选择的准则

如何寻找合适的准则

- 之前，我们介绍过两个指标用于衡量模型拟合数据的好坏。
 - 残差平方和 SS_E
 - 决定系数 R^2
- 问题：这两个指标能否用于选择自变量？

自变量选择的准则

考虑残差平方和 SS_E

- 考虑 p_1 个自变量纳入线性模型，即

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p_1} x_{p_1} + \varepsilon$$

记该模型的残差平方和为 $SS_E^{p_1}$.

- 考虑 $p_1 + 1$ 个自变量纳入线性模型，即

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p_1} x_{p_1} + \beta_{p_1+1} x_{p_1+1} + \varepsilon$$

记该模型的残差平方和为 $SS_E^{p_1+1}$.

自变量选择的准则

考虑残差平方和 SS_E

- 残差平方和

$$SS_E^{p_1+1} = \mathbf{y}' (\mathbf{I}_n - \mathbf{H}_{\mathbf{X}_{p_1+1}}) \mathbf{y}$$

由于

$$\begin{aligned} \mathbf{H}_{\mathbf{X}_{p_1+1}} &= \mathbf{X}_{p_1+1} (\mathbf{X}_{p_1+1}' \mathbf{X}_{p_1+1})^{-1} \mathbf{X}_{p_1+1}' \\ &= \begin{pmatrix} \mathbf{X}_{p_1} & \mathbf{x}_{p_1+1} \end{pmatrix} \begin{pmatrix} \mathbf{X}_{p_1}' \mathbf{X}_{p_1} & \mathbf{X}_{p_1}' \mathbf{x}_{p_1+1} \\ \mathbf{x}_{p_1+1}' \mathbf{X}_{p_1} & \mathbf{x}_{p_1+1}' \mathbf{x}_{p_1+1} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}_{p_1}' \\ \mathbf{x}_{p_1+1}' \end{pmatrix} \\ &= \mathbf{H}_{\mathbf{X}_{p_1}} + \mathbf{M} \end{aligned}$$

其中, $\mathbf{M} = (\mathbf{I}_n - \mathbf{H}_{\mathbf{X}_{p_1}}) \mathbf{x}_{p_1+1} (\mathbf{x}_{p_1+1}' (\mathbf{I}_n - \mathbf{H}_{\mathbf{X}_{p_1}}) \mathbf{x}_{p_1+1})^{-1} \mathbf{x}_{p_1+1}' (\mathbf{I}_n - \mathbf{H}_{\mathbf{X}_{p_1}})$ 是一个对称矩阵.

自变量选择的准则

考虑残差平方和 SS_E

- 重要结论:

$$SS_E^{p_1+1} \leq SS_E^{p_1}$$

- 这表明了随着自变量个数的**增加**，残差平方和**减少**.

自变量选择的准则

考虑决定系数 R^2

- 由于

$$R^2 = \frac{SS_R}{SS_T} = \frac{SS_T - SS_E}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

而且对于相同的数据集, SS_T 是不变的.

- 重要结论:

$$R_{p_1+1}^2 \geq R_{p_1}^2$$

- 这表明了随着自变量个数的**增加**, 决定系数**增加**.

自变量选择的准则

如何寻找合适的准则

- 之前，我们介绍过两个指标用于衡量模型拟合数据的好坏。
 - 残差平方和 SS_E
 - 决定系数 R^2
- 问题：这两个指标能否用于选择自变量？
- 答案：不能！

自变量选择的准则

常用的变量选择的准则

- 修正后的决定系数

$$\tilde{R}^2 = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

- 显然有 $\tilde{R}^2 \leq R^2$;
- \tilde{R}^2 随着自变量的增加并不一定增大, 是因为 $\frac{n-1}{n-p-1}$ 起到惩罚作用.

自变量选择的准则

常用的变量选择的准则

- 误差项方差 σ^2 的无偏估计

$$\hat{\sigma}^2 = \frac{1}{n - p - 1} SS_E$$

- $n - p - 1$ 作为惩罚因子；
- 刚开始随着自变量个数的增加，残差平方和 SS_E 能快速减少，而作为除数的惩罚因子先 $n - p - 1$ 随之减少，但由于 SS_E 减少速度更快，因而 $\hat{\sigma}^2$ 是趋于减少的。
- 当自变量个数增加到一定程度时，重要的自变量基本都选上了，这时再增加自变量， SS_E 减少的幅度不大，以至于抵消不了除数 $n - p - 1$ 的减少，最终又导致了 $\hat{\sigma}^2$ 的增加。

自变量选择的准则

常用的变量选择的准则

- 赤池信息量准则 AIC

$$\text{AIC} = -2 \ln (\text{模型最大似然}) + 2 (\text{模型独立参数个数})$$

- 在线性模型中,

$$\begin{aligned}\text{AIC} &= n \ln(2\pi) + n \ln \left(\frac{SS_E}{n} \right) + n + 2(p + 2) \\ &\propto n \ln (SS_E/n) + 2(p + 1)\end{aligned}$$

- 对每一个回归子集计算 AIC, 而 AIC 最小者所对应的回归模型就是最优的回归模型.

自变量选择的准则

常用的变量选择的准则

- 贝叶斯信息量准则 BIC, 也称为 SBC 准则:

$$\text{BIC} = -2 \ln (\text{模型最大似然}) + \ln(n) (\text{模型独立参数个数})$$

- 在线性模型中,

$$\text{BIC} = n \ln(SS_E/n) + \ln(n)(p + 1)$$

- 对每一个回归子集计算 BIC, 而 BIC 最小者所对应的回归模型就是最优的回归模型.

自变量选择的准则

常用的变量选择的准则

- 马洛斯统计量 (Mallow's C_p): 用预测的角度来选择自变量.
- 即便全模型为真, 选模型可能得到更小的预测误差.
- 我们考虑在 n 个样本点上用选模型做回归预测, 预测值与期望值的相对偏差平方和为

$$\begin{aligned}J_{p_1} &= \frac{1}{\sigma^2}(\hat{\mathbf{y}} - E(\mathbf{y}))'(\hat{\mathbf{y}} - E(\mathbf{y})) \\&= \frac{1}{\sigma^2}(\hat{\mathbf{y}} - \mathbf{y} + \mathbf{y} - E(\mathbf{y}))'(\hat{\mathbf{y}} - \mathbf{y} + \mathbf{y} - E(\mathbf{y})) \\&= \frac{1}{\sigma^2}((\hat{\mathbf{y}} - \mathbf{y})'(\hat{\mathbf{y}} - \mathbf{y}) + 2(\hat{\mathbf{y}} - \mathbf{y})'(\mathbf{y} - E(\mathbf{y})) + (\mathbf{y} - E(\mathbf{y}))'(\mathbf{y} - E(\mathbf{y})))\end{aligned}$$

其中, 预测值 $\hat{\mathbf{y}} = X_{p_1}\hat{\beta}_{p_1}$ and $E(\mathbf{y}) = X_p\beta_p$.

自变量选择的准则

常用的变量选择的准则

- 此相对偏差平方和的期望为

$$\begin{aligned}E(J_{p_1}) &= \frac{1}{\sigma^2} (E(\hat{\mathbf{y}} - \mathbf{y})'(\hat{\mathbf{y}} - \mathbf{y}) + 2E(\hat{\mathbf{y}} - \mathbf{y})(\mathbf{y} - E(\mathbf{y})) \\&\quad + E(\mathbf{y} - E(\mathbf{y}))'(\mathbf{y} - E(\mathbf{y}))) \\&= \frac{1}{\sigma^2} (E(SS_E^{p_1}) - 2E(\mathbf{y}'\mathbf{N}_{\mathbf{X}_{p_1}}(\mathbf{y} - E(\mathbf{y}))) + E(\mathbf{y} - E(\mathbf{y}))'(\mathbf{y} - E(\mathbf{y}))) \\&= \frac{1}{\sigma^2} (E(SS_E^{p_1}) - 2(n - (p_1 + 1))\sigma^2 + n\sigma^2) \\&= \frac{E(SS_E^{p_1})}{\sigma^2} - n + 2(p_1 + 1)\end{aligned}$$

- 在欠拟合的情况下，由于

$$E(SS_E^{p_1}) = (n - p_1 - 1)\sigma^2 + \boldsymbol{\gamma}'\mathbf{Z}'\mathbf{N}_{\mathbf{X}_{p_1}}\mathbf{Z}\boldsymbol{\gamma}$$

可知， $E(J_{p_1}) > (n - p_1 - 1) - n + 2(p_1 + 1) = p_1 + 1$.

自变量选择的准则

常用的变量选择的准则

- 马洛斯 C_p 统计量为

$$\begin{aligned}C_p &= \frac{SS_E^{p_1}}{\hat{\sigma}^2} - n + 2(p_1 + 1) \\&= (n - p - 1) \frac{SS_E^{p_1}}{SS_E^p} - n + 2(p_1 + 1)\end{aligned}$$

- 马洛斯 C_p 准则：选择使 C_p 值最小的自变量子集，这个自变量子集对应的回归方程就是最优回归方程。

逐步回归

概述

- 如果自变量个数为 p , 那么所考虑的模型个数为 $2^p - 1$.
- 如果自变量个数很多, 那么尝试所有模型是十分困难.

p	$2^p - 1$
10	1,023
20	1,048,575
30	1,073,741,823

- 人们提出了一些较为简便、实用的变量选择方法.

逐步回归

前进法

- 确定一种变量选择的准则（如：AIC 最小），从最小的模型开始.
- 从 x_1, x_2, \dots, x_p 中确定 x_1 放入模型；
- 从 $(x_1, x_2), (x_1, x_3), \dots, (x_1, x_p)$ 中确定 x_2 放入模型；
- 以此类推，直到不满足准则.

逐步回归

后退法

- 确定一种变量选择的准则（如：AIC 最小），从最大的模型开始.
- 从 (x_1, x_2, \dots, x_p) 中确定 x_p 从模型中剔除，保留剩余的自变量；
- 从 $(x_1, x_2, \dots, x_{p-1})$ 中确定 x_{p-1} 从模型中剔除，保留剩余的自变量；
- 以此类推，直到不满足准则.

逐步回归

说明

- 前进法和后退法都有明显的不足.
- 前进法: 不能反映引进新自变量后的变化情况. 因为某个自变量开始可能是显著的, 当引入其他自变量后就变得不显著了, 但是也没有机会将其剔除, 即一旦引入, 就是“终身制”的.这种只考虑引入而没有考虑剔除的做法显然是不全面的.
- 后退法: 一开始把全部自变量引入回归方程, 这样计算量很大. 如果有些不太重要的自变量, 一开始就不引入, 就可以减少一些计算量. 在就是一次一旦某个自变量被剔除, 它就没有机会在进入回归方程.

逐步回归

逐步法

- 逐步回归的基本思想是**有进有出**.
- 具体做法：
 - 将自变量一个一个地引入，每引入一个自变量后，对已选入的变量要进行逐个确定，当之前引入的自变量因当前自变量引入而导致模型不再优化时，需要将其从回归方程中剔除.
 - 这个过程反复进行，直到加入其他任何一个自变量，模型并不会更优化，或者剔除模型中的任何一个自变量，模型也不会更优化
- 逐步法可以弥补了前进法和后退法各自的缺陷.