统计方法与机器学习

第三章：多重共线性 - 2

倪 葎

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

SCHOOL OF DATA
SCIENCE & ENGINEERING
数据科学与工程学院

目录

为什么要选择自变量?

概述

- **预测**是线性回归模型的主要任务之一。
- 对于预测任务而言，我们希望预测值偏离真实值的差异越小越好。
- 考虑一个问题：当我们可以获取 100 个变量时，在线性回归模型中，我们究竟放入多少个自变量，**才能使得预测结果最为准确?**
- 具体问题：
    - 如果只用少许变量作为自变量（甚至一个变量），模型预测效果会如何?
    - 如果把所有变量都当作自变量，模型预测效果会如何?

为什么要选择自变量?

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

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

其中,

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

为什么要选择自变量?

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

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

其中,

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

为什么要选择自变量?

说明

- 为了简化，不妨认为 $x_1, x_2, \cdots, x_{p_1}$ 就是 $x_1, x_2, \cdots, x_p$ 中的前 p_1 个.

为什么要选择自变量?

自变量选择，可看成全模型与选模型的比较问题

- 自变量的选择可看成用**全模型**，还是用**选模型**的问题；

- 如果**全模型**是正确的，而我们错误地使用选模型，那么，在建模时丢失了一些重要且有用的自变量，在线性回归模型中称这种情形为**欠拟合**（underfit）；

- 如果**选模型**是正确的，而错误地使用全模型，那么在建模时引入了一些不必要的自变量，在线性回归模型中称这种情形为**过拟合**（overfit）。

为什么要选择自变量?

自变量选择，可看成全模型与选模型的比较问题

- 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

$$\hat{\beta}_p = \left(\boldsymbol{X}_p'\boldsymbol{X}_p\right)^{-1}\boldsymbol{X}_p'\boldsymbol{y}$$

$$\hat{\sigma}_p^2 = \frac{1}{n-p-1}SS_E^p = \frac{1}{n-p-1}(\boldsymbol{y}-\hat{\boldsymbol{y}}_p)'(\boldsymbol{y}-\hat{\boldsymbol{y}}_p)$$

  - 在 $\boldsymbol{x}_{p,0} = (\boldsymbol{x}_{p1,0}', \boldsymbol{z}_0')'$ 时的预测值为

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

为什么要选择自变量?

自变量选择，可看成全模型与选模型的比较问题

- 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，
  - $\boldsymbol{\beta}$ 和 σ^2 的估计分别为

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

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

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

为什么要选择自变量?

自变量选择，可看成全模型与选模型的比较问题

- 在采用**选模型**的情况下，
  - $\boldsymbol{\beta}$ 和 σ^2 的估计分别为

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

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

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

欠拟合

模型

- 假定**全模型**为真，即

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

- 而我们错误地使用了**选模型**，即

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

- 两个模型的差异在于 $\boldsymbol{Z}\boldsymbol{\gamma}$。其存在需要满足
  - $\mathrm{rank}(\boldsymbol{X}_p) > \mathrm{rank}(\boldsymbol{X}_{p_1})$;
  - $\gamma \neq \mathbf{0}'_{p-p_1}$.

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

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

- 预测值为

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

欠拟合

参数估计——$\hat{\boldsymbol{\beta}}_{p_1}$

- 考虑 $\hat{\boldsymbol{\beta}}_{p_1}$ 的期望，即

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

- 结论：
  - 因为 $\gamma \neq \mathbf{0}$，所以，$\hat{\beta}_{p_1}$ 一般而言是有偏估计；
  - 偏差为 $(\boldsymbol{X}_{p_1}'\boldsymbol{X}_{p_1})^{-1}\boldsymbol{X}_{p_1}'\boldsymbol{Z}\boldsymbol{\gamma}$。

欠拟合

参数估计——$\hat{\boldsymbol{\beta}}_{p_1}$

- 注意到

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

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

参数估计——$\hat{\sigma}^2_{p_1}$

- 考虑 SS_E^p 与 $SS_E^{p_1}$ 的关系.
- 注意到

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

- 其中

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

欠拟合

参数估计——$\hat{\sigma}^2_{p_1}$

$$
\begin{aligned}
\boldsymbol{H}_{\boldsymbol{X}_p} &= \begin{pmatrix}\boldsymbol{X}_{p_1} & \boldsymbol{Z}\end{pmatrix}\begin{pmatrix}\boldsymbol{X}'_{p_1}\boldsymbol{X}_{p_1} & \boldsymbol{X}'_{p_1}\boldsymbol{Z}\\ \boldsymbol{Z}'\boldsymbol{X}_{p_1} & \boldsymbol{Z}'\boldsymbol{Z}\end{pmatrix}^{-1}\begin{pmatrix}\boldsymbol{X}'_{p_1}\\ \boldsymbol{Z}'\end{pmatrix}\\
&= \begin{pmatrix}\boldsymbol{X}_{p_1} & \boldsymbol{Z}\end{pmatrix}\begin{pmatrix}A & B\\ C & D\end{pmatrix}\begin{pmatrix}\boldsymbol{X}'_{p_1}\\ \boldsymbol{Z}'\end{pmatrix}\\
&= \boldsymbol{H}_{\boldsymbol{X}_{p_1}} + \boldsymbol{H}_{\boldsymbol{X}_{p_1}}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{N}_{\boldsymbol{X}_{p_1}}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{H}_{\boldsymbol{X}_{p_1}}\\
&\quad -\boldsymbol{H}_{\boldsymbol{X}_{p_1}}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{N}_{\boldsymbol{X}_{p_1}}\boldsymbol{Z})^{-1}\boldsymbol{Z}' - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{N}_{\boldsymbol{X}_{p_1}}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{H}_{\boldsymbol{X}_{p_1}} + \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{N}_{\boldsymbol{X}_{p_1}}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\\
&= \boldsymbol{H}_{\boldsymbol{X}_{p_1}} + \boldsymbol{N}_{\boldsymbol{X}_{p_1}}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{N}_{\boldsymbol{X}_{p_1}}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{N}_{\boldsymbol{X}_{p_1}}
\end{aligned}
$$

其中,

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

欠拟合

参数估计——$\hat{\sigma}^2_{p_1}$

- SS_E^p 与 $SS_E^{p_1}$ 的关系为

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

- $SS_E^{p_1}$ 的期望为

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

欠拟合

参数估计——$\hat{\sigma}^2_{p_1}$

- 注意到

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

- 上式中不等号是“严格的”，这是因为之前假设

$$\operatorname{rank}(\boldsymbol{X}_p) > \operatorname{rank}(\boldsymbol{X}_{p_1})$$

即 $\boldsymbol{N}_{\boldsymbol{X}_{p_1}}\boldsymbol{Z} \neq 0$. 因此，$\hat{\sigma}^2_{p_1}$ 不是 σ^2 的无偏估计.

欠拟合

预测值

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

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

- 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

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

其期望和方差分别为

$$\begin{aligned} E\left(\hat{y}_{0,T}\right) &= \boldsymbol{x}'_{p,0}E\left(\hat{\boldsymbol{\beta}}_p\right) = \boldsymbol{x}'_{p_1,0}\boldsymbol{\beta}_{p_1} + \boldsymbol{z}'_0\boldsymbol{\gamma} \\ \operatorname{Var}\left(\hat{y}_{0,T}\right) &= \sigma^2\boldsymbol{x}'_{p,0}\left(\boldsymbol{X}'_p\boldsymbol{X}_p\right)^{-1}\boldsymbol{x}_{p,0} \end{aligned}$$

欠拟合

预测值

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

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

- 但是，我们错误地使用了选模型. 此时，y_0 的预测为

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

其期望为

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

欠拟合

预测值

- 在全模型下，y_0 预测值的期望为

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

- 在选模型下，y_0 预测值的期望为

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

- 偏差为

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

预测值

- 两个预测值方差的差异，即

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

其中，

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

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

预测值

- 两个预测值方差的差异，即

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

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

过拟合

模型

- 假定**选模型**为真，即

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

- 而我们错误地使用了**全模型**，即

$$
\boldsymbol{y} = \begin{pmatrix} \boldsymbol{X}_{p_1} & \boldsymbol{Z} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_{p_1} \\ \boldsymbol{\gamma} \end{pmatrix} + \boldsymbol{\varepsilon}
$$

参数估计——$\hat{\boldsymbol{\beta}}_p$

- 考虑 $\hat{\boldsymbol{\beta}}_p$ 的期望，即

$$
\begin{aligned}
E\left(\hat{\boldsymbol{\beta}}_p\right) &= \left(\boldsymbol{X}_p'\boldsymbol{X}_p\right)^{-1}\boldsymbol{X}_p'E\left(\boldsymbol{y}\right)\\
&= \left(\boldsymbol{X}_p'\boldsymbol{X}_p\right)^{-1}\boldsymbol{X}_p'\left(\left(\boldsymbol{X}_{p_1} \quad \boldsymbol{Z}\right)\begin{pmatrix}\boldsymbol{\beta}_{p_1}\\ \mathbf{0}\end{pmatrix} + E(\boldsymbol{\varepsilon})\right)\\
&= \left(\boldsymbol{X}_p'\boldsymbol{X}_p\right)^{-1}\boldsymbol{X}_p'\left(\boldsymbol{X}_{p_1} \quad \boldsymbol{Z}\right)\begin{pmatrix}\boldsymbol{\beta}_{p_1}\\ \mathbf{0}\end{pmatrix}\\
&= \left(\boldsymbol{X}_p'\boldsymbol{X}_p\right)^{-1}\boldsymbol{X}_p'\boldsymbol{X}_p\begin{pmatrix}\boldsymbol{\beta}_{p_1}\\ \mathbf{0}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\beta}_{p_1}\\ \mathbf{0}\end{pmatrix}
\end{aligned}
$$

- 结论：$\hat{\boldsymbol{\beta}}_p$ 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

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

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

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

- 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

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

  其期望和方差分别为

$$\begin{aligned} E\left(\hat{y}_{0,T}\right) &= \boldsymbol{x}'_{p_1,0}E\left(\hat{\boldsymbol{\beta}}_{p_1}\right)=\boldsymbol{x}'_{p_1,0}\boldsymbol{\beta}_{p_1} \\ \operatorname{Var}\left(\hat{y}_{0,T}\right) &= \sigma^2\boldsymbol{x}'_{p_1,0}\left(\boldsymbol{X}'_{p_1}\boldsymbol{X}_{p_1}\right)^{-1}\boldsymbol{x}_{p_1,0} \end{aligned}$$

过拟合

预测值

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

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

- 但是，我们错误地使用了全模型. 此时，y_0 的预测值为

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

  其期望为

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

预测值

- $\hat{y}_0$ 的方差

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

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

为什么要选择自变量?

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.

为什么要选择自变量?

总结

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

选择自变量的基本逻辑

全子集回归法

- 假定共有 p 个自变量 $x_1, x_2, \cdots, x_p$，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	$\{x_1\}, \{x_2\}, \cdots, \{x_p\}$	$C_p^1 = p$
2	$\{x_1, x_2\}, \cdots, \{x_{p-1}, x_p\}$	$C_p^2 = \frac{p(p-1)}{2}$
$\vdots$	$\vdots$	$\vdots$
$p-1$	$\{x_2, \cdots, x_p\}, \cdots, \{x_1, \cdots, x_{p-1}\}$	$C_p^{p-1} = p$
p	$\{x_1, x_2, \cdots, x_p\}$	$C_p^p = 1$

- 因此，需要考虑 $C_p^1 + \cdots + C_p^p = 2^p - 1$ 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.

选择自变量的基本逻辑

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.

选择自变量的基本逻辑

概述

- 我们更为关心的问题是：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
- 提问：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

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

由于

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

其中，$\boldsymbol{M} = (\boldsymbol{I}_n - \boldsymbol{H}_{\boldsymbol{X}_{p_1}})\boldsymbol{x}_{p_1+1}(\boldsymbol{x}'_{p_1+1}(\boldsymbol{I}_n - \boldsymbol{H}_{\boldsymbol{X}_{p_1}})\boldsymbol{x}_{p_1+1})^{-1}\boldsymbol{x}'_{p_1+1}(\boldsymbol{I}_n - \boldsymbol{H}_{\boldsymbol{X}_{p_1}})$ 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：

$$R^2_{p_1+1} \geq R^2_{p_1}$$

- 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
- 提问：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 $\tilde{R}^2 \le R^2$；
- $\tilde{R}^2$ 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

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

- 在线性模型中,

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

- 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:

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

  - 在线性模型中,

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

  - 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

总结

- 常用于变量选择的指标有
  - 修正后的决定系数 $\tilde{R}^2$;
  - σ^2 的无偏估计 $\hat{\sigma}^2$;
  - 赤池信息量准则 AIC;
  - 贝叶斯信息准则 BIC;

选择自变量的路径

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。

选择自变量的路径

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 $x_1, x_2, \cdots, x_p$ 中确定 x_1 放入模型；
- 从 $(x_1, x_2), (x_1, x_3), \cdots, (x_1, x_p)$ 中确定 x_2 放入模型；
- 以此类推，直到不满足准则.

选择自变量的路径

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 $(x_1, x_2, \cdots, x_p)$ 中确定 x_p 从模型中剔除，保留剩余的自变量；
- 从 $(x_1, x_2, \cdots, x_{p-1})$ 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.

选择自变量的路径

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. 在就是一旦某个自变量被剔除，它就没有机会在进入回归方程.

选择自变量的路径

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.