第 5 章　自变量选择与逐步回归

李 杰

数据科学学院，浙江财经大学

2019 年 12 月 2 日

内容提要

1. 选择回归模型自变量是建立回归模型的重要问题；
2. 遗漏某些变量，特别是某些对因变量有重要影响的变量，模型的解释效果不好；
3. 模型中纳入过多自变量，有些变量对因变量影响可能不重要，有些变量对因变量的影响有较大重叠，这将导致计算量增大很多，且回归方程稳定性差，影响回归模型的应用。

全模型和选模型

假定研究的问题中对因变量 y 可能有影响的所有变量为 $x_1, x_2, \cdots, x_m$.

- **全模型**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon_m \tag{1}$$

- **选模型:** 从自变量集合 $\{x_1, x_2, \cdots, x_m\}$ 任选一子集 $\{x_{p1}, x_{p2}, \cdots, x_{pp}\}$, 构造模型

$$y = \beta_{p0} + \beta_{p1} x_{p1} + \beta_{p2} x_{p2} + \cdots + \beta_{pp} x_{pp} + \varepsilon_p \tag{2}$$

自变量选择问题

- 自变量选择问题就是研究实际问题时，选用全模型 (1) 还是选模型 (2) 的问题，如果使用选模型，则需包含哪些自变量。
- 该选用全模型 (1) 时误用了选模型 (2)，则说明建模时存在变量遗漏问题。
- 该选用选模型 (2) 时误用了全模型 (1)，则说明建模时引入了不必要的变量。

自变量选择与预测的影响

☞ 全模型参数估计量

$$\hat{\beta}_m = \left(\mathbf{X}_m'\mathbf{X}_m\right)^{-1}\mathbf{X}_m'\mathbf{y} \tag{3}$$

$$\hat{\sigma}_m^2 = \frac{1}{n-m-1}SSE_m \tag{4}$$

$$\hat{\mathbf{y}}_m = \mathbf{X}_m\hat{\beta}_m \tag{5}$$

☞ 选模型参数估计量

$$\hat{\beta}_p = \left(\mathbf{X}_p'\mathbf{X}_p\right)^{-1}\mathbf{X}_p'\mathbf{y} \tag{6}$$

$$\hat{\sigma}_p^2 = \frac{1}{n-p-1}SSE_p \tag{7}$$

$$\hat{\mathbf{y}}_p = \mathbf{X}_p\hat{\beta}_p \tag{8}$$

① 选模型中的变量与剔除变量相关时，选模型的估计量不是相合估计，也不是无偏估计。记 $\mathbf{X}_m = (\mathbf{X}_p, \mathbf{X}_{m-p})$, $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_p \\ \boldsymbol{\beta}_{m-p} \end{pmatrix}$，则全模型正确时，有

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_p &= \left(\mathbf{X}_p'\mathbf{X}_p\right)^{-1}\mathbf{X}_p\mathbf{y} \\
&= \left(\mathbf{X}_p'\mathbf{X}_p\right)^{-1}\mathbf{X}_p\left[(\mathbf{X}_p, \mathbf{X}_{m-p})\begin{pmatrix} \boldsymbol{\beta}_p \\ \boldsymbol{\beta}_{m-p} \end{pmatrix} + \boldsymbol{\epsilon}_m\right] \\
&= \left(\mathbf{X}_p'\mathbf{X}_p\right)^{-1}\mathbf{X}_p(\mathbf{X}_p\boldsymbol{\beta}_p + \mathbf{X}_{m-p}\boldsymbol{\beta}_{m-p} + \boldsymbol{\epsilon}_m) \\
&= \boldsymbol{\beta}_p + \left(\mathbf{X}_p'\mathbf{X}_p\right)^{-1}\mathbf{X}_p\mathbf{X}_{m-p}\boldsymbol{\beta}_{m-p} + \left(\mathbf{X}_p'\mathbf{X}_p\right)^{-1}\mathbf{X}_p\boldsymbol{\epsilon}_m \\
E\left(\hat{\boldsymbol{\beta}}_p\right) &= \boldsymbol{\beta}_p + \left(\mathbf{X}_p'\mathbf{X}_p\right)^{-1}\mathbf{X}_p\mathbf{X}_{m-p}\boldsymbol{\beta}_{m-p}
\end{aligned}$$

注：只有当 $\boldsymbol{\beta}_{m-p} = \mathbf{0}$，或模型中的变量与遗漏的变量不相关，即 $\mathbf{X}_p\mathbf{X}_{m-p} = \mathbf{0}$ 时，$\hat{\boldsymbol{\beta}}_p$ 才是无偏估计。但全模型正确时，$\boldsymbol{\beta}_{m-p} \neq \mathbf{0}$；通常情况下 $\mathbf{X}_p\mathbf{X}_{m-p} \neq \mathbf{0}$.

② 全模型正确时，选模型的预测值是有偏的。记要预测点自变量向量的观测值为 $\mathbf{x}_0 = (\mathbf{x}_{0p}, \mathbf{x}_{0,m-p})$，则全模型的预测值为

$$\hat{y}_0 = \mathbf{x}_0 \hat{\boldsymbol{\beta}}_m$$

且

$$\begin{aligned}
E(\hat{y}_0) &= E\left[\mathbf{x}_0 \left(\mathbf{X}_m'\mathbf{X}_m\right)^{-1} \mathbf{X}_m'\mathbf{y}\right] \\
&= E\left[\mathbf{x}_0 \left(\mathbf{X}_m'\mathbf{X}_m\right)^{-1} \mathbf{X}_m'(\mathbf{X}_m\boldsymbol{\beta}_m + \boldsymbol{\epsilon}_m)\right] \\
&= \mathbf{x}_0\boldsymbol{\beta}_m \\
&= \mathbf{x}_{0p}\boldsymbol{\beta}_p + \mathbf{x}_{0,m-p}\boldsymbol{\beta}_{m-p}
\end{aligned}$$

而选模型的预测值为

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

$$\begin{aligned}
E(\hat{y}_{0p}) &= E\left[\mathbf{x}_{0p} \left(\mathbf{X}_p'\mathbf{X}_p\right)^{-1} \mathbf{X}_p'\mathbf{y}\right] \\
&= E\left[\mathbf{x}_{0p} \left(\mathbf{X}_p'\mathbf{X}_p\right)^{-1} \mathbf{X}_p'(\mathbf{X}_p\boldsymbol{\beta}_p + \mathbf{X}_{m-p}\boldsymbol{\beta}_{m-p} + \boldsymbol{\epsilon}_p)\right] \\
&= \mathbf{x}_{0p}\boldsymbol{\beta}_p + \mathbf{x}_{0p}\left(\mathbf{X}_p'\mathbf{X}_p\right)^{-1}\mathbf{X}_p'\mathbf{X}_{m-p}\boldsymbol{\beta}_{m-p}
\end{aligned}$$

显然，$E(\hat{y}_0) \neq E(\hat{y}_{0p})$。

③ 选模型的参数估计量方差较小.

$$D(\hat{\beta}_{pi}) \leqslant D(\hat{\beta}_{mi}) \quad (i = 1, 2, \cdots, p)$$

其中 $\hat{\beta}_{pi}$ 是选模型中自变量 x_i 的系数估计量，$\hat{\beta}_{mi}$ 是选模型中自变量 x_i 的系数估计量。

④ 选模型预测值方差较小.

$$D(e_{0p}) \leqslant D(e_{0m})$$

⑤ 选模型预测的均方误差比全模型预测均方误差小.

$$E(e_{0p}^2) = D(e_{0p}) + [E(e_{0p})]^2 \leqslant D(e_{0m})$$

☞ 所有子集的数量

- 假设实际问题中自变量有 m 个, 则选模型的数量有 2^m 个.

$$C_m^0 + C_m^1 + \cdots + C_m^m = 2^m$$

- 复决定系数 R^2 不能作为自变量选择的标准。因为，选定 p 个变量做回归，记其残差平方和为 SSE_p，如果再增加一个解释变量，其残差平方和为 SSE_{p+1}，则必有

$$SSE_p \geqslant SSE_{p+1},$$
$$R_p^2 \leqslant R_{p+1}^2$$

即复决定系数 R^2 随着解释变量的增多而增大。故残差平方和、复决定系数都不可以作为选择变量的标准。

自变量选择的准则——$\bar{R}^2$ 最大

- 给魔性增加自变量时，复决定系数增大，但残差平方和的自由度在减小，自由度减小意味着估计和预测的可靠性低。故增加自变量后，回归模型的拟合从表面看更好，但事实上掺杂了虚假成分。
- 考虑调整的复决定系数

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

- $\bar{R}^2 \leqslant R^2$，自变量增加时 $\bar{R}^2$ 不一定增加，只有当增加的解释变量对回归的贡献比较大时 $\bar{R}^2$ 才会增大，否则反而会减小。
- 另一种解释：$\bar{R}^2$ 与扰动项方差估计量 $\hat{\sigma}^2$ 最小是一致的。 因为

$$\hat{\sigma}^2 = \frac{1}{n-{\color{red}p}-1} SSE, \quad \bar{R}^2 = 1 - \frac{n-1}{SST}\hat{\sigma}^2$$

与因变量相关的自变量可大致分为三类：对因变量有强影响、影响一般、影响较弱。当自变量个数从 0 开始增多时，一开始对因变量影响较大的变量进入模型后，$\hat{\sigma}^2$ 快速下降；然后影响一般的变量进入模型，$\hat{\sigma}^2$ 的值开始稳定；最后影响较弱的变量进入模型，$\hat{\sigma}^2$ 的值反而开始上升。

☞ AIC 最小

- 定义回归模型似然估计的似然函数为 $L(\boldsymbol{\theta}, \mathbf{y})$，其中 待估参数 $\boldsymbol{\theta}$ 的维度为 p，因变量向量 $\mathbf{y}$。定义 AIC 为

$$AIC = -2\ln L(\hat{\boldsymbol{\theta}}_L, \mathbf{y}) + 2p$$

$\hat{\boldsymbol{\theta}}_L$ 为 θ 的最大似然估计。似然函数值越大，估计量越好。上述目标函数中加入惩罚因子 $2p$，使得 AIC 最小的模型时最优模型。

- 对 AIC 变形略去与 p 的常数，得到

$$AIC = n\ln(SSE) + 2p$$

☞ Mallows 从预测的角度提出一个可用来选择自变量的 C_p 统计量

$$C_p = (n-m-1)\frac{SSE_p}{SSE_m} - n + 2p$$

☞ 所有子集回归

```
install.packages("leaps")
library(leaps)
library(foreign)

travel <- read.spss("D:\\documents\\MyDoc\\JobInZufe\\Courseware
        \\应用回归分析\\数据\\例3.1 国际旅游收入.sav",to.data.frame =T)
travel_leaps <- regsubsets(Y~.,data=travel, nbest=6)
plot(travel_leaps,scale="adjr2" )      # 调整的 R^2 最大准则
```

依据 $\bar{R}^2$ 最大准则，最优回归子集为 $Y \sim 1 + X3 + X5 + X8 + X9 + X10 + X11$.

```
plot(travel_leaps,scale="Cp" )      # 调整的 R^2 最大准则
```

依据 C_p 准则，最优回归子集为 $Y \sim 1 + X3 + X8 + X9 + X10 + X11$.

5.3 逐步回归

☞ 前进法

- **思想:** 选入的变量由少到多, 每次增加一个, 直到没有满足条件的变量选入为止.
- **方法:** 以 y 为因变量, 一开始做关于常数的回归，得到 AIC，记为 C_0，再以每个自变量为自变量做 m 个一元回归, 回归 AIC-值最小的变量选入. 然后做该变量与剩余 $m-1$ 个变量的二元回归, 选取回归 AIC 最小的进入模型. 以此类推, 直至选入新的变量后 AIC 不再减小为止.
- **问题:** "终身制", 变量一旦选入模型, 即使选入变量增多后不显著变量, 也不能剔除.

☞ 后退法

- **思想:** 选入的变量由多到少, 每次减少一个, 直到没有满足条件的变量剔除为止.
- **方法:** 以 y 为因变量, 以所有自变量为自变量做 m 元回归, 得到相应的 AIC 的值，再任意剔除一个变量做 $m-1$ 元回归, 选取 AIC 值最小的，再在其中任意剔除一个变量做 $m-2$ 元回归，以此类推, 直至没有剔除变量后 AIC 的值不再减小为止。
- **问题:** "一棍子打死", 变量一旦被剔除, 即使后期有可能变得显著, 也不能再选入模型.

☞ 逐步法

- **思想:** 变量有进有出, 直到找到 AIC 值最小的模型为止.
- **方法:** 首先估计包含 p 个变量的初始模型，计算初始模型的 AIC 值，在此模型基础上分别剔除 p 个变量和添加 $m-p$ 个变量中任何一个后模型的 AIC, 然后选择最小的 AIC 值决定是否添加或删除初始模型中的变量，如此反复，直至既不添加也不剔除模型中已有的变量时所对应的 AIC 值最小。

☞ 逐步回归

```
travel_lm <- lm(Y~.,data=travel)
travel_for <- lm(Y~1,data=travel)
                                                   # 前进法
stepAIC(travel_for,scope=list(upper=~X1+X2+X3+X4+X5+X6+
               X7+X8+X9+X10+X11+X12,lower=~1),direction="forward")

stepAIC(travel_lm,direction="backward")      # 后退法

stepAIC(travel_lm,direction="both")          # 逐步回归
```

类似地，还有

```
                                                   # 前进法
step(travel_for,scope=list(upper=~X1+X2+X3+X4+X5+X6+
               X7+X8+X9+X10+X11+X12,lower=~1),direction="forward")

step(travel_lm,direction="backward")         # 后退法

step(travel_lm,direction="both")             # 逐步回归
```