自变量的选择与逐步回归

马学俊 (主讲) 晁越 (助教)

苏州大学

数学科学学院

https://xuejunma.github.io/

引言

- 从 20 世纪 60 年代开始，关于回归自变量的选择成为统计学中研究的热点问题。统计学家们提出了许多回归选元的准则，并提出了许多行之有效的选元方法。
- 本章从回归选元对回归参数估计和预测的影响开始，
    - 介绍自变量选择常用的几个准则
    - 扼要介绍所有子集回归选元的几个方法
    - 详细讨论逐步回归方法及其应用

全模型和选模型

- 全模型　　设研究某一实际问题涉及的对因变量有影响的因素共 m 个，由因变量 y 和 m 个自变量 $x_1, x_2, \cdots, x_m$ 构成的回归模型为

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \epsilon \tag{5.1}$$

称该模型为全回归模型

- 选模型　　如果从所有可选择的 m 个变量中挑选出 p 个，记为 $x_1, x_2, \cdots, x_p$，由所选的 p 个自变量组成的回归模型为

$$y = \beta_{0p} + \beta_{1p} x_1 + \beta_{2p} x_2 + \cdots + \beta_{pp} x_p + \epsilon \tag{5.2}$$

称该模型为选模型。

模型选择不当会给参数估计和预测带来什么影响? 下面我们将分别给予讨论。

全模型和选模型

为了方便，把全模型式 (5.1) 的参数向量 $\boldsymbol{\beta}$ 和 σ^2 的估计记为

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

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

把选模型式 (5.1) 的参数向量 $\boldsymbol{\beta}$ 和 σ^2 的估计记为

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

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

自变量选择对预测的影响

假设全模型式 (5.1) 与选模型式 (5.2) 不同，即要求 $p < m$，$\beta_{p+1}x_{p+1} + \cdots + x_m x_m$ 不恒为 0。在此条件下，当全模型正确而误用了选模型时，有以下性质

- 在 x_j 与 $x_{p+1}, \cdots, x_m$ 的相关系数不全为 0 时，选模型回归系数的最小二乘估计是全模型相应参数的有偏估计，即

$$E(\hat{\beta}_{jp}) = \beta_{jp} \neq \beta_j, j = 1, 2, \cdots, p$$

- 选模型的预测是有偏的。给定新的自变量值，$\mathbf{x}_{0m} = (x_{o1}, x_{02}, \cdots, x_{0m})'$，因为新值为 $y_0 = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \cdots, \beta_m x_{0m} + \epsilon_0$，用选模型的预测值 $\hat{y}_{0p} = \hat{\beta}_{0p} + \hat{\beta}_{01} x_{01} + \cdots + \hat{\beta}_{pp} x_{0p}$ 作为 y_0 的预测值是有偏的，即 $E(\hat{y}_{0p} - y_0) \neq 0$
- 选模型的参数估计有较小的方差。选模型的最小二乘估计 $\hat{\boldsymbol{\beta}}_p = (\hat{\beta}_{0p}, \hat{\beta}_{1p}, \hat{\beta}_{2p}, \cdots, \hat{\beta}_{pp})'$，全模型的最小二乘估计为 $\hat{\boldsymbol{\beta}}_m = (\hat{\beta}_{0m}, \hat{\beta}_{1m}, \hat{\beta}_{2m}, \cdots, \hat{\beta}_{mm})'$，这一性质说明 $D(\hat{\beta}_{jp}) \leq D(\hat{\beta}_{jm}), j = 1, 2, \cdots, p$

自变量选择对预测的影响

- 选模型的预测残差有较小的方差。
  选模型的预测残差为 $e_{0p} = \hat{y}_{0p} - y_0$，
  全模型的预测残差为 $e_{0m} = \hat{y}_{0m} - y_0$，
  其中 $y_0 = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \cdots + \beta_m x_{0m} + \epsilon$，则有
  $D(e_{0p}) \leq D(e_{0m})$
- 记 $\boldsymbol{\beta}_{m-p} = (\beta_{p+1}, \cdots, \beta_m)'$，用全模型对 $\boldsymbol{\beta}_{m-p}$ 的最小二乘估计为 $\hat{\boldsymbol{\beta}}_{m-p} = (\hat{\beta}_{p+1}, \cdots, \hat{\beta}_m)'$，则在
  $D(\boldsymbol{\beta}_{m-p}) \geq \boldsymbol{\beta}_{m-p}\boldsymbol{\beta}'_{m-p}$ 的条件下，
  $E(e_{0p})^2 = D(e_{0p}) + (E(e_{0p}))^2 \leq D(e_{0m})$，即选模型预测的均方误差比全模型预测的方差小。

自变量选择对预测的影响

- 一个好的回归模型，并不是考虑的自变量越多越好。
- 在建立回归模型时，选择自变量的基本指导思想是“少而精”。哪怕我们丢掉了一些对因变量 y 还有些影响的自变量，
- 由选模型估计的保留变量的回归系数的方差，要比由全模型所估计的相应变量的回归系数的方差小。
- 对于所预测的因变量的方差来说也是如此。丢掉了一些对因变量 y 有影响的自变量后，所付出的代价是估计量产生了有偏性。然而，尽管估计量是有偏的，但预测偏差的方差会下降。
- 如果保留下来的自变量有些对因变量无关紧要，那么，方程中包括这些变量会导致参数估计和预测的有偏性和精度降低。

所有子集的数目

- $x_1, x_2, \cdots, x_m$
- 每个自变量都有入选和不入选两种情况，这样 y 关于这些自变量的所有可能的回归方程就有 $2m-1$ 个。这里减一是要求回归模型中至少包含一个自变量。
- 包含常数项

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

关于自变量选择的几个准则

- 从数据与模型拟合优劣的直观考虑出发，认为残差平方和 SSE 最小的回归方程就是最好的。
- 复相关系数 R 来衡量回归拟合的好坏。

然而这两种方法都有明显的不足，这是因为：

$$SSE_{p+1} \leq SSE_p$$

$$R^2_{p+1} \geq R^2_p$$

关于自变量选择的几个准则

准则 自由度调整复决定系数达到最大

- 调整的复决定系数为

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

- $R_a^2 \leq R^2$，R_a^2 随着自变量的增加并不一定增大。因为尽管 $1-R^2$ 随着变量的增加而减小，但由于其前面的系数 $(n-1)/(n-p-1)$ 增大起折扣作用。
- 从拟合优度的角度追求最优，则所有回归子集中 R_a^2 最大者对应的回归方程就是最优方程。

准则 1：自由度调整复决定系数达到最大

从另一个角度考虑回归的拟合效果，

- 回归误差项方差 σ^2 的无偏估计为

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

此无偏估计式中也加入了惩罚因子 $n-p-1$

- 用平均残差平方和 $\hat{\sigma}^2$ 作为自变量选元准则是合理的。
- 残差平方和和复决定系数 R_a^2 有什么关系？

$$R_a^2 = 1 - \frac{n-1}{SST} \hat{\sigma}^2$$

由于 SST 是与回归无关的固定值，因此 R_a^2 与 $\hat{\sigma}^2$ 是等价的。

关于自变量选择的几个准则

AIC 与 BIC 准则

- AIC 准则是日本统计学家赤池 (Akaike)1974 年根据极大似然估计原理提出的一种较为一般的模型选择准则，人们称它为 Akaike 信息量准则 (Akaike Information Criterion, 简记为 AIC)。
- AIC 准则既可用来作回归方程自变量的选择，又可用于时间序列分析中自回归模型的定阶上。
- 由于该方法的广泛应用，使得赤池乃至日本统计学家在世界的声誉大增。

设模型的似然函数为 $L(\boldsymbol{\theta}, \boldsymbol{x})$，$\boldsymbol{x}$ 的维数为 p，为随机样本（在回归分析中随机样本为 $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)'$），则 AIC 定义为：

$$AIC = -2 \ln L(\hat{\boldsymbol{\theta}}_L, \boldsymbol{x}) + 2p \qquad (*)$$

其中 $\hat{\boldsymbol{\theta}}_L$ 为 $\boldsymbol{\theta}$ 的最大似然估计，p 为未知参数的个数。

AIC 准则

假定回归模型的随机误差项 ϵ 服从正态分布，即

$$\epsilon \sim N(0, \sigma^2)$$

对数似然函数为

$$\ln L_{max} = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}_L^2) - \frac{1}{2\hat{\sigma}_L^2} SSE$$

将 $\hat{\sigma}_L^2 = \frac{1}{n}SSE$ 代入得

$$\ln L_{max} = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\frac{SSE}{n}) - \frac{n}{2}$$

将上式代入 $(*)$，这里似然函数中未知参数得个数为 $P+2$，略去与 p 无关得常数，则回归模型得 AIC 公示为

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

对每一个回归子集计算 AIC，其中 AIC 最小者所对应得模型是最优回归模型。

BIC 准则

- 赤池于 1976 年对 AIC 准则给予了改进，而施瓦茨 (Schwartz) 在 1978 年根据 Bayes 理论也得出同样的判别准则，称为 BIC 准则 (Bayesian information criterion)，也称为 SBC(Schwartz's Bayesian criterion) 准则，加大了对自变量数目的惩罚力度，
- BIC 达极小。

$$BIC = n\ln(SSE) + \ln(n)p \tag{5.11}$$

- R 软件可以计算 BIC，计算形式大致为

$$BIC = n\ln(\frac{SSE}{SST}) + 1 + \ln(2\pi) + \ln(n)p \tag{5.12}$$

式 (5.11) 与 (5.12) 是等价的，两者的差值只与 n 和 SST 有关，与 p 无关。

关于自变量选择的几个准则

- 则 C_p 统计量达到最小

1964 年马勒斯 (Mallows) 从预测的角度提出一个可以用来选择自变量的统计量——C_p 统计量。根据性质 5，即使全模型正确，但仍有可能选模型有更小的预测误差。C_p 正是根据这一原理提出来的。

考虑在 n 个样本点上，用选模型式作回报预测时，预测值与期望值的相对偏差平方和为：

$$J_p = \frac{1}{\sigma^2}\sum_{i=1}^{n}(\hat{y}_{ip} - E(y_i))^2$$
$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}(\hat{\beta}_{0p} + \hat{\beta}_{1p}x_{i1} + \cdots + \hat{\beta}_{pp}x_{ip} - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im}))^2$$

J_p 的期望是

$$E(J_p) = \frac{E(SSE_p)}{\sigma^2} - n + 2(P+1)$$

C_p 统计量达到最小

略去无关的常数 2，据此构造出 C_p 统计量为

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

其中 $\hat{\sigma}^2 = \frac{1}{n-m-1} SSE_m$，为全模型中 σ^2 的无偏估计。

这样我们得到一个选择变量的 C_p 准则：选择使 C_p 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。

- 例 5.1 和例 5.2

逐步回归

- 变量的所有可能子集构成 $2m-1$ 个回归方程，
- 当可供选择的自变量不太多时，用前边的方法可以求出一切可能的回归方程，然后用几个选元准则去挑出“最好”的方程，
- 但是当自变量的个数较多时，要求出所有可能的回归方程是非常困难的。

为此，人们提出了一些较为简便、实用、快速的选择“最优”方程的方法。人们所给出的方法各有优缺点，至今还没有绝对最优的方法，

- 目前常用的方法有“前进法”、“后退法”、“逐步回归法”，而逐步回归法最受推崇。
- 在后边的讨论中，无论我们从回归方程中剔除某个自变量，还是给回归方程增加某个自变量都要利用偏 F 检验，这个偏 F 检验 t 检验是等价的，F 检验的定义式的统计意义更为明了，并且容易推广到对多个自变量的显著性检验，因而采用 F 检验。

$$F_j = \frac{\Delta SSR_{(j)}/1}{SSE/(n-p-1)}, \quad t_j = \frac{\hat{\beta}_j}{\sqrt{c_{jj}}\hat{\sigma}}$$

前进法

前进法的思想是变量由少到多，每次增加一个，直至没有可引入的变量为止。具体的做法是首先将全部 m 个自变量分别对因变量 y 建立一元线性回归方程，并分别计算这 m 个一元线性回归方程的 m 个回归系数的 F 检验值，记为 $\{F_1^1, F_2^1, \cdots, F_m^1\}$，选其最大值记为

$$F_j^1 = \max\{F_1^1, F_2^1, \cdots, F_m^1\}$$

给定显著水平 α，若 $F_j^1 \geq F_\alpha(1, n-2)$，则首先将 x_j 引入回归方程，为了方便，设 x_j 就是 x_1。

接下来因变量 y 分别与 $(x_1, x_2), (x_1, x_3), \cdots, (x_1, x_m)$ 建立二元线性回归方程，对这 $m-1$ 个回归方程中 $x_2, \cdots, x_m$ 的回归系数进 F 检验，计算 F 值，记为 $\{F_2^2, F_3^2, \cdots, F_m^2\}$，选其最大值记为

$$F_j^2 = \max\{F_2^2, F_3^2, \cdots, F_m^2\}$$

若 $F_j^2 \geq F_\alpha(1, n-3)$，则将 x_j 引入回归方程。

依上述方法接着做下去。直至所有未被引入方程的自变量的 F 值均小于 $F_\alpha(1, n-p-1)$ 时为止。这时，得到的回归方程就是最终确定的方程。

每步检验中的临界值 $F_\alpha(1, n-p-1)$ 与自变量数目 p 有关，在用软件计算时，我们实际使用的是显著性 P 值（或记为 sig）做检验。例 5.4

后退法

- 后退法与前进法相反，首先用全部 m 个变量建立一个回归方程，然后在这 m 个变量中选择一个最不重要的变量，将它从方程中剔除。
- 设对 m 个回归系数进行 F 检验，记求得的 F 值为 $\{F_1^m, F_2^m, \cdots, F_m^m\}$，选其中最小者记为：

$$F_j^m = \min\{F_1^m, F_2^m, \cdots, F_m^m\}$$

  给定显著水平 α，若 $F_j^m \leq F_\alpha(1, n-m-1)$，则首先将 x_j 从回归方程中剔除，为了方便，设 x_j 就是 x_m。
- 接着对剩下的 $m-1$ 个自变量重新建立回归方程，进行回归系数的显著性检验，像上面那样计算出 F_1^{m-1}，如果又有 $F_j^{m-1} \leq F_\alpha(1, n-(m-1)-1)$，则剔除 x_j，重新建立关于 $m-2$ 个自变量的回归方程，
- 以此类推，直至回归方程中所剩余的 p 个自变量的 F 检验值均大于临界值 $F_\alpha(1, n-p-1)$，没有可以剔除的变量为止。这时，得到的回归方程就是最终确定的方程。
- 续例 5.4

逐步回归法

- 逐步回归的基本思想是“有进有出”。具体做法是将变量一个一个引入，当每引入一个自变量后，对已选入的变量要进行逐个检验，当原引入的变量由于后面变量的引入而变得不再显著时，要将其剔除。
- 这个过程反复进行，直到既无显著的自变量选入回归方程，也无不显著自变量从回归方程中剔除为止。
- **优点**避免了前进法和后退法各自的缺陷，保证了最后所得的回归子集是“最优”回归子集。
- 在逐步回归中需要注意的一个问题是引入自变量和剔除自变量的显著性水平 α 值是不相同的，要求$\alpha_{进} < \alpha_{出}$，否则可能产生“死循环”。
  - 当 $\alpha_{进} \geq \alpha_{出}$ 时，
  - 某个自变量的显著性 P 值在 $\alpha_{进}$ 与 $\alpha_{出}$ 之间，那末这个自变量将被引入、剔除、再引入、再剔除、$\cdots$，循环往复，以至无穷.
- 续例 5.5

作业

- 使用 R 实现书中所有的例子（不需要提交）
- p.151 5.9(提交)
- Consider a random sample $X_1, X_2, \ldots, X_n \sim Unif(0, \theta)$.
  1. Find the estimator for θ through MoM,denoted by $\hat{\theta}_{MM}$.
  2. Find the MLE $\hat{\theta}_{MLM}$.
  3. What are the expetation and variance of $\hat{\theta}_{MM}$ and $\hat{\theta}_{MLM}$? Which estimator is better?
- 扩充阅读： CH11 Regression Analysis By Example 5th