《线性回归》

杨　　瑛

清华大学　数学科学系
Email: yyang@math.tsinghua.edu.cn
Tel: 62796887

2019.05.23

Outline

1. 变量选择问题
   - 变量选择
   - 变量选择
2. 模型与变量的扩展
   - 线性回归模型的扩展

Outline

变量选择方法

变量选择(variable selection) 是线性回归中非常重要的研究课题之一.

- 变量选择的目的之一在于精简模型, 用尽可能少的变量描述模型, 将多余的变量剔除，用简单的方法解释数据，
- 对于响应变量y提供合理而又准确的预报.
- 利用变量选择方法可以消除共线性, 但变量选择的目的不完全是为了处理共线性. 如果变量之间有共线性的话，可能多个变量在做同样的事情！
- 成本：如果我们的模型是用来做预测的话，不用测量多余的变量就可以节省金钱和时间!

变量选择方法

- 好的模型中应包含对响应有重要影响的、贡献大的变量，不包含无效变量。
- 对于模型贡献不大的变量依照一定的标准(例如,最小化均方误差)将其剔除之.

变量选择之前的工作

- 甄别异常值和影响点，至少要暂时排除这些点.
- 对变量做适当的变换是合适的.

变量选择方法(续)

- 向前回归法、向后回归法、逐步回归法(stepwise selection)、完全子集法、交叉核实法(cross validation)、AIC、BIC、LASSO、SCAD、Bayes变量选择、最佳子集选择(Best Subset Selection)、特征选择(Feature Selection)、Group Selection、Penalized Variable Selection in High-Dimensional Data、Robust Model Selection、等
- 这些方法在SAS的PROG REG中有非常详细的论述。
- 在R中也有一些package专门讨论变量选择的方法。可在http://cran.r-project.org/ 找到合适的package.

变量选择方法：向后回归法

基本想法：将所有可能对响应变量产生影响的自变量都纳入模型，然后逐个将最没有价值或者贡献的变量从模型中剔除，直到留在模型中的变量不能剔除或者模型没有变量为止。

1. 拟合完全模型(将所有可能对响应变量产生影响的自变量都纳入模型)；
2. 用F统计量较小者或者p值较大者的变量【可能是无效的变量】剔除之，得到较小的模型；
3. F统计量超过指定的临界值或者p值较小者的变量保留下来；
4. 重新拟合模型；
5. 回到第二步。

变量选择方法: AIC (Akaike Information Criterion)

- AIC 准则是由日本统计学家Akaike (赤池)根据最大似然原理提出的变量选择标准：

$$AIC = n\left\{\log\left(RSS/n\right) + 1 + \log(2\pi)\right\} + 2 \times (p+1), \quad (1)$$

  其中RSS 是拟合残差平方和，p 是选入模型的变量个数，n 为样本量。
- AIC 原则的想法：当选入模型的变量增加时，(1) 中的拟合残差平方和RSS减小，即(1) 中的第一项较小，而第二项随着入选模型变量增加而增加。当由变量增加带来的方差减少的作用大于变量增加带来的惩罚时，AIC 的值逐渐减小；而当变量个数达到一定数目时，由变量增加带来的惩罚大于变量增加带来的方差减少时，AIC的值将会逐渐增加。
- AIC 选择变量的原则：使AIC达到最小的模型是‘最优’的模型。

变量选择方法: BIC (Bayesian Information Criterion)

- BIC 定义为

$$BIC = n\left\{\log\left(RSS/n\right) + 1 + \log(2\pi)\right\} + \log(n) \times (p+1), \quad (2)$$

- BIC 选择变量的原则：使BIC达到最小的模型是‘最优’的模型。
- 与AIC不同是的惩罚项不一样，BIC中变为$\log(n) \times (p+1)$，当$n \geq 8$时，$\log(n) > 2$, 因此，BIC的惩罚项的力度要比AIC来得大，通常BIC选出的变量个数少于AIC 选出的变量个数。
- AIC 较保守，选出模型变量的个数往往多于真实模型参数的个数，即过拟合(overfit);
- BIC 准则确定的最优模型通常与真实模型更为接近！

AIC和BIC在R中的实现：

- AIC 的实现：
  lm.aic=step(lm(y∼ x, data=dat), trace=F);
  summary(lm.aic)
- BIC 的实现:
  lm.bic=step(lm(y∼ x, data=dat), k=log(n), trace=F);
  summary(lm.bic)

R中的其它的变量选择方法

package leaps: regression subset selection
在R的packages目录搜索关键词'selection' 可以找多个变量选择方法(包括传统的方法和现代的方法)。

变量选择方法：交叉核实(cross validation)

基本想法：

1. 将数据适当的分为两部分: $S_c = \{(x_i, y_i), i = 1, \cdots, n_c\}$, $S_v = \{(x_i, y_i), i = n_c, \cdots, n_v + n_c\}$, $n = n_c + n_v$.
2. 用S_c 来建立线性模型$Y_i = X_i\beta + \epsilon_i$, $i = 1, \cdots, n_c$, 得到回归方程$Y = X\widehat{\beta}_c$.
3. 用S_v 来核实模型: 计算在$x = x_i$, $n_c + 1 \leq i \leq n$出的预测值$\hat{y}_i$,
4. 好的模型应该使得平均平方预测误差: $\frac{1}{n_v}\sum_{j=1}^{n_v}(y_{i+n_c} - \hat{y}_{i+n_c})^2$ 达到最小.
5. 如何选择n_c 和n_v? 详见: Shao, J. (1993), Linear model selection by cross-validation. *J. Am. Stat. Assoc.* 88, 486-494.

现代变量选择方法：

LASSO [Least Absolute Shrinkage & Selection Operator]

1. LASSO是斯坦福大学统计系教授Tibshirani 于1996年发表的著名论文“Regression shrinkage and selection via the LASSO”(Journal of Royal Statistical Society, Series B, 58, 267-288) 中所提出的一种选择变量的方法。与传统的基于AIC或者BIC标准的变量选择方法相比，该方法的计算速度非常快，特别适合处理当代的大规模数据。在过去的十几年中，该方法的理论研究和应用取得了长足的进展。
2. 该方法是线性模型估计和变量选择的的新方法，LASSO 就是在回归系数的值不超过某个常数的条件之下极小化残差平方和。基于这个约束条件，LASSO选择的某些系数恰好是0,从而给出可解释的模型。
3. 研究表明，LASSO 具有子集选择和岭估计的性质。LASSO 的概念是非常一般和广泛的，可以应用于各种统计模型，可以推广到广义的回归模型。

现代变量选择方法：LASSO的定义

- 设数据为$(X_i, Y_i), i = 1, \cdots, n$, 其中$X_i = (x_{i1}, \cdots, x_{ip})^T$ 是预测变量，Y_i是响应变量。假定协变量给定的条件之下，响应值是相互独立的。假定x_{ij} 是被标准化的使得$\sum_i x_{ij}/N = 0$, $\sum_i x_{ij}^2 = 1$.
  设$\hat{\beta} = (\hat{\beta}_1, \cdots, \hat{\beta}_p)^T$, 则lasso 估计$(\hat{\alpha}, \hat{\beta})$ 定义为

$$(\hat{\alpha}, \hat{\beta}) = \arg\min \left\{ \sum_{i=1}^{n} \left(Y_i - \alpha - \sum_j \beta_j x_{ij}\right)^2 \right\}$$
$$\text{subject to} \quad \sum_j |\beta_j| \leq t \tag{3}$$

  其中$t \geq 0$是tuning parameter.
- 方程(3) 的求解计算可以转换为一个有线性不等式约束的二次规划问题。

现代变量选择方法：LASSO的应用

- 参见R的Prostate(lasso2)
- 或者参见R的prostate(faraway) page 59 in faraway.pdf
- 一些有用的命令：

  help(package='lasso2')

  data(package='lasso2')

  data(Prostate)

  example(Prostate)

作业:

利用cross-validation 方法，选择上述例题中的自变量。

Outline

- 内在线性模型
- 内在非线性模型
- 给定协变量$X_2, \cdots, X_p$ 和响应变量Y, 假定

$$Y = F(X_2, X_3, \cdots, X_k, \epsilon)$$

该模型称为内在线性的，如果它可以转化为

$$\begin{aligned} f(Y) &= \beta_1 + \beta_2 g_2(X_2, \cdots, X_k) \\ &\quad + \cdots + \beta_k g_k(X_2, \cdots, X_k) + \epsilon \end{aligned}$$

或者

$$Y^* = \beta_1 + \beta_2 X_2^* + \cdots + \beta_k X_k^* + \epsilon \tag{4}$$

- (4) 中的关系之所以称之为是内在线性的是因为模型关于参数$\beta_1,\beta_2,\cdots,\beta_k$ 是线性的。
- 一些在实际中用到的重要的内在线性模型：

- $Y=\beta_1+\beta_2X_2+\beta_3X_2^2+\epsilon$ (Quadratic polynomial model)
- $Y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1^2+\beta_4x_2^2+\beta_5x_1x_2+\epsilon$ (包括二次效应和交互效应)
- $\log Y=\alpha_1+\alpha_2\log(X_2)+\alpha_3\log(X_3)+\log\epsilon$(对数线性模型)

- $Y=\gamma_1X_2^{\gamma_2}X_3^{\gamma_3}\epsilon^*$ (Multiplicative model)

- $Y = \exp\{\beta_1 + \beta_2 X_2 + \beta_3 X_3\} + \epsilon'$ (Exponential model)
  By taking logarithms of both sides, this model can be transformed to

$$\log Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \log \epsilon$$

- $Y = \frac{1}{\beta_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon}$ (reciprocal model)
- $Y = \beta_1 + \beta_2 \log(X_2) + \epsilon$ (semilog model)
- $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 (X_2 X_3) + \epsilon$ (interaction model).
  Note that if $\log \epsilon \sim N(\mu, \sigma^2)$, then ϵ obeys lognormal distribution .

线性回归模型的扩展

- $\log\left(\frac{y}{1-y}\right) = \beta_0 + \beta_1 X_1 + \epsilon$ (logit model)
- $\Phi\left(\frac{y}{1-y}\right) = \beta_0 + \beta_1 X_1 + \epsilon$ (probit model)
  Φ 为标准正态随机变量的分布函数。
- 对于上述内在线性模型，可以利用LS方法得到参数的估计。

Box-Cox 变换

- 线性模型中的Box-Cox 变换（已单独讲授）