模型的选择与评价

在前面的章节中, 我们考虑了如何拟合简单和多元回归模型, 以及如何从这些模型中进行推断. 在本章中, 我们首先概述了模型构建和模型验证过程. 然后, 我们更详细地考虑了探索性观测研究中预测变量选择的一些特殊问题. 我们在本章的最后详细描述了验证回归模型的方法.

建立回归模型的4个阶段

1. 数据收集和准备
2. 减少解释性或预测性变量
3. 模型细化和选择
4. 模型验证

1 总览

1.1 数据收集阶段

对照实验

在对照实验中, 实验者控制解释变量的水平, 并给每个实验单元分配由解释变量水平组合而成的处理, 并观察其反应. 一种治疗方法包括一种特殊的组合, 即呈现的大小和允许的时间长度. 在对照实验中, 解释变量常被称为*因子*或*控制变量*.

有协变量的对照实验

实验的统计设计利用补充信息, 如实验单元的特征, 来设计实验, 以减少回归模型中实验误差项的方差. 然而, 有时不可能将这些补充信息合并到实验设计中. 相反, 试验者有可能将这些信息纳入回归模型, 从而通过在模型中包括*未受控变量*或*协变量*来减少误差方差.

验证性的观察性研究

这些研究基于观察数据, 而非实验数据, 目的是检验(即证实或不证实)源自以往研究或直觉的假设. 在这些研究中, 收集的数据包括先前的研究表明会影响响应变量的解释变量, 以及假说中涉及的新变量或变量. 在这种情况下, 假说中涉及的解释变量有时被称为*主要变量*, 而反映现有知识

的解释变量被称为控制变量(流行病学中被称为**风险因素**). 这里的控制变量不像在实验研究中那样受控, 但它们被用来说明对响应变量的已知影响.

例. 在一项关于维生素E补充剂对某一类型癌症发生影响的观察性研究中, 已知的风险因素, 如年龄、性别和种族, 将被纳入控制变量, 每日维生素E补充剂的摄入量将成为主要解释变量. 反应变量将是在考虑期间发生的特定类型的癌症.

探索性的观察性研究

在社会、行为和健康科学、管理和其他领域, 通常不可能进行对照实验. 或可能缺乏进行验证性观察性研究的充分知识. 因此, 这些领域的许多研究都是探索性观察性研究, 研究人员寻找可能与反应变量相关的解释变量.

一个解释变量

1. 可能不是问题的根本
2. 可能受到较大的测量误差的影响
3. 可能有效地重复列表中的另一个解释变量
4. 无法测量的解释变量可以被删除或被与之高度相关的代理变量所取代

1.2 解释变量的简化

对照实验

- 单纯对照实验: 很少需要或希望减少解释变量的数量
- 带协变量的对照实验: 删除任何不能减少误差方差的协变量

观察性研究

- 验证性观察性研究: 必须保留所有的控制变量与之前的研究进行比较, 也应该保留所有的主要变量.
- 探索性观察性研究: 通常有许多潜在的预测因素(以及多项式和相互作用). 想要拟合一个简洁的模型来解释Y的大部分变化, 同时尽可能保持一个基本的模型

注记. 通常情况下, 粗心的调查人员会通过拟合包含整个潜在X变量集的回归模型来筛选一组解释变量, 然后简单地删除那些绝对值较小的t^*统计量:

$$t_k^* = \frac{b_k}{se\{b_k\}}$$

然而由于多重共线性的效果, 某些系数的抽样标准差可能相较本身而言更大, 因此上述方法可能去掉了重要的变量, 因此一种良好的搜索方法应当使得能够获取重要的变量.

注记. 对p个预测变量, 就有2^p种不同的组合形式, 因此函数空间的增长是指数形式的.

2 外科的例子

一家医院外科部门对预测接受特定类型肝脏手术的患者的生存率很感兴趣. 随机选取108例患者进行分析. 从每个患者记录中, 从术前评估中提取以下信息: 由于存活时间的分布是严重右偏的,

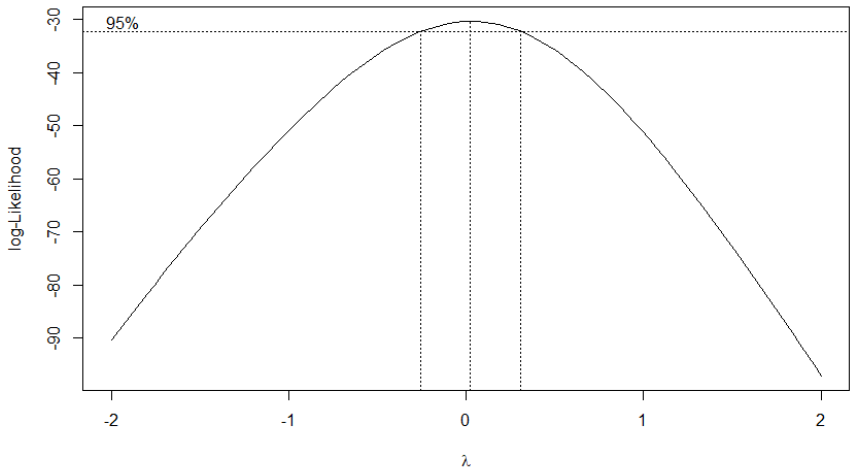
Y	术后存活时间
X_1	凝血值
X_2	预后指数
X_3	酶功能值
X_4	肝功值
X_5	年龄
X_6	性别的指示变量
X_7, X_8	酒精使用的指示变量

表 1:

所以取存活时间的对数$Y' = \log Y$为应变量, 或者我们考虑Box-Cox变换

```r
alldat = read.table('./Data_5e/Data_5e/CH09TA01.txt')
dat0 = alldat[1:54,c(1:4, 9)]
names(dat0) = c('X1','X2','X3','X4','Y')
library(MASS)
fit = lm(Y~X1+X2+X3+X4,data=dat0)
bxcx = boxcox(fit)
```

Listing 1: **boxcox.R**



如果该研究是在一些受试者死亡之前进行的, 那么时间可能会有删失, 则应当使用生存分析的方法, 若仅考虑4个自变量我们有$2^4 = 16$中模型.

3 选择“最佳”回归模型: 模型选择的判别准则

尝试获取一个回归方程时, 实际上面对的是从众多模型中做选择的问题,

- 是不是所有的变量都要包括
- 应不应该去掉那个对预测贡献不显著的变量
- 是否要添加多项式项或交互项提高拟合精度
- 两个具有几乎相同精确度的模型应该选哪一个

使用似然函数的缺点.

考虑线性回归模型:

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

假设 $\boldsymbol{\varepsilon} \mid \boldsymbol{X} \sim N\left(\boldsymbol{\theta}, \sigma^2 \boldsymbol{I}_n\right)$, 则 $\boldsymbol{y} \mid \boldsymbol{X} \sim N\left(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n\right)$, 条件密度函数为:

$$f(\boldsymbol{y} \mid \boldsymbol{X}) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})\right\}$$

那么对于两模型

$$m_1 : y = X\beta + \epsilon$$

$$m_2 : y = X\beta + X^*\beta^* + \epsilon$$

显然对系数的搜索: m_1的搜索空间是m_2搜索空间的子空间, 因此变量增多必然导致似然函数不减, 但是在模型选择时应当对解释变量个数加以限制, 即对似然函数进行一定的惩罚. 详见3.3节.

在基础安装中的anova() 函数可以比较两个嵌套模型(nested models)的拟合优度, 所谓嵌套模型是指, 它的一些项完全包含于另一些模型之中.

```
> rm(list = ls())
> state <- as.data.frame(state.x77[,c("Murder","Population","Illiteracy","Income","Frost")])
> rm(list = ls())
> states <- as.data.frame(state.x77[,c("Murder","Population","Illiteracy","Income","Frost")])
> fit1 <- lm(Murder~Population+Illiteracy+Income+Frost,data = states)
> fit2 <- lm(Murder~Population+Illiteracy,data = states)
> anova(fit1,fit2)
Analysis of Variance Table

Model 1: Murder ~ Population + Illiteracy + Income + Frost
Model 2: Murder ~ Population + Illiteracy
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     45 289.17
2     47 289.25 -2 -0.078505 0.0061 0.9939
```

这里用到的anova()本质上就是一般线性检验.

我们在这里给出6个判别准则

$$R_p^2, R_{a,p}^2, C_p, AIC_p, BIC_p(SBC_p), PRESS_p$$

我们面临2个不同的问题,

- 最小子集的选取: 调整的R^2, $MSE, C_p, PRESS, AIC, SBC$.
- 固定最小子集后最佳模型的选取: R^2

3.1 全子集回归

3.1.1 R_p^2准则

R_p^2准则要求逐个复判定系数$R^2 = \dfrac{SSR}{SST}$以选择一个或者几个X变量的自己, R_p^2表示回归方程中有p个参数或$p-1$个自变量.

R_p^2的散点图$\max(R_p^2)$出现在每个p的顶端.

注记. 在使用R_p^2准则进行比较时, 只有当解释变量X的个数相同时比较才有意义.

3.1.2 MSE_p或R_a^2准则

R_p^2并不会考虑模型中参数的个数, 我们使用调整的复判定系数

$$\begin{aligned} R_a^2 &= 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SST} \\ &= 1 - \frac{MSE}{SST/(n-1)} \end{aligned}$$

这个准则通过自由度考虑了参数个数, 当且仅当MSE减少时R_a^2增加. 使用MSE_p准则, 就是寻找很接近于最小的MSE_p的一个或几个子集.

注记. 其中p为参数个数, $p \geq 1$, 因此$R_a^2 \leq R^2$

3.1.3 Mallow's C_p准则

Mallows建议了一种与拟合值的均方误差相关的准则.

C_p准则是关于不同子集回归模型的n个拟合值的总均方误差.

回顾点估计理论, 对θ的点估计$\hat{\theta}$, 我们如此定义均方误差

$$E(\hat{\theta} - \theta)^2 = MSE(\hat{\theta})$$

以下我们将Y_i的期望值记为μ_i, 则有

$$\begin{aligned} \left(\hat{Y}_i - \mu_i\right)^2 &= \left(\hat{Y}_i - \mathrm{E}\left(\hat{Y}_i\right) + \mathrm{E}\left(\hat{Y}_i\right) - \mu_i\right)^2 \\ &= \left(\mathrm{E}\left(\hat{Y}_i\right) - \mu_i\right)^2 + \left(\hat{Y}_i - \mathrm{E}\left(\hat{Y}_i\right)\right)^2 + \left[E\left(\hat{Y}_i\right) - \mu_i\right]\left[\hat{Y}_i - E\left(\hat{Y}_i\right)\right] \\ &= \mathrm{Bias}^2 + \left(\hat{Y}_i - \mathrm{E}\left(\hat{Y}_i\right)\right)^2 + \left[E\left(\hat{Y}_i\right) - \mu_i\right]\left[\hat{Y}_i - E\left(\hat{Y}_i\right)\right] \end{aligned}$$

因此$\hat{Y}_i$的均方误差是$E\left(\hat{Y}_i - \mu_i\right)^2 = \left(\mathrm{E}\left(\hat{Y}_i\right) - \mu_i\right)^2 + \sigma^2\left(\hat{Y}_i\right)$, 所有$n$个拟合值的总均方误差是

$$\sum \left(E\left(Y_i\right) - \mu_i\right)^2 + \sum \sigma^2\left(\hat{Y}_i\right)$$

度量标准是总均方误差除以真实均方误差σ^2, 用Γ_p表示:

$$\begin{aligned}\Gamma_p &= \frac{\sum\left(E\left(Y_i\right)-\mu_i\right)^2+\sum\sigma^2\left(Y_i\right)}{\sigma^2}\\ &= \frac{\sum \text{ Bias }^2+\sum \text{Var(prediction)}}{\text{Var(error)}}\end{aligned}$$

可见最小化Γ_p和最小化MSE是等价的, 因此Γ_p越小越好.

考虑一个有$p-1$个自变量的(有p个参数的)模型, 则有$E(SSE_p)=\sum(E(\hat{Y}_i-\mu_i)^2+(n-p)\sigma^2$.

证明.

$$Y=\mu+\varepsilon, \varepsilon\sim N(0,\sigma^2 I), E(Y)=\mu$$

回顾二次型的期望

$$E(X^\top AX)=tr(A\cdot Var(X))+(EX)^\top A(EX)$$

$$\begin{aligned}SSE_p &= \sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2=\mathbf{Y}'(\mathbf{I}-\mathbf{H})\mathbf{Y}\\ E\left(SSE_p\right) &= E\left\{\mathbf{Y}'(\mathbf{I}-\mathbf{H})\mathbf{Y}\right\}=\text{tr}[(\mathbf{I}-\mathbf{H})]\sigma^2+\boldsymbol{\mu}'(\mathbf{I}-\mathbf{H})\boldsymbol{\mu}\\ &= (n-p)\sigma^2+[(\mathbf{I}-\mathbf{H})\boldsymbol{\mu}]'[(\mathbf{I}-\mathbf{H})\boldsymbol{\mu}]\\ &= (n-p)\sigma^2+[\boldsymbol{\mu}-\mathbf{E}(\hat{\mathbf{Y}})]'[\boldsymbol{\mu}-\mathbf{E}(\hat{\mathbf{Y}})]\\ &= (n-p)\sigma^2+\sum_{i=1}^{n}\left[E\left(\hat{Y}_i\right)-\mu_i\right]^2\end{aligned}$$

注意: 当模型不正确时, $\hat{\mu}=H\mu$不是μ的无偏估计量, 只有当模型与真实情况一样时, 才有$H\mu=\mu$. □

注记. 若构建的小模型正确, 则大模型也一定正确, 且对$\sum\sigma^2(\hat{Y}_i)$, 我们有

$$\sum\sigma^2(\hat{Y}_i)=tr(\sigma^2(\hat{Y}))$$

而

$$\sigma^2(\hat{Y})=\sigma^2(HY)=$$

可以证明Γ_p的一个估计量$\hat{\Gamma}_p$是C_p:

$$\begin{aligned}C_p &= \frac{\left(\text{SSE}_p-(n-p)\text{MSE}_P\right)+p\text{MSE}_P}{\text{MSE}_P}\\ &= \frac{\text{SSE}_p}{\text{MSE}\left(X_1,X_2,\ldots,X_{P-1}\right)}-(n-2p)\end{aligned}$$

如果$p-1$个自变量的回归模型没有偏差, 那么C_p值落在$C_p=p$附近.

$$\Gamma_p=\frac{0+p\sigma^2}{\sigma^2}=p;\quad E\left(C_p\right)\approx p;$$

偏差很大的模型的C_p值势必远远高于这条直线.

$$\Gamma_p > \frac{0 + p\sigma^2}{\sigma^2} = p; \quad E\left(C_p\right) > p;$$

人们使用C_p准则时, 根据下面两条来识别X变量的子集

1. C_p值小
2. C_p值接近于p

3.2 全子集回归实例

```
rm(list = ls())
states <-
    as.data.frame(state.x77[, c("Murder", "Population", "Illiteracy", "Income", "Frost")])
fit1 <-
    lm(Murder ~ Population + Illiteracy + Income + Frost, data = states)
fit2 <- lm(Murder ~ Population + Illiteracy, data = states)
anova(fit1, fit2)
AIC(fit1, fit2)
library(MASS)
fit <-
    lm(Murder ~ Population + Illiteracy + Income + Frost, data = states)
stepAIC(fit, direction = "backward")
library(leaps)
leaps <-
    regsubsets(Murder ~ Population + Illiteracy + Income + Frost,
               data = states,
               nbest = 4)
plot(leaps, scale = "adjr2")
library(car)
subsets(leaps, statistic = "cp", main = "Cp Plot for All Subsets Regression")
abline(1, 1, lty = 2, col = "red")
```

Listing 2: 全子集回归

初看图1可能令人费解, 从底部开始, 第一行可以看到含有 (intercept)截距项和Income的模型调整的$R_a^2 = -.33$, 含有截距项和Population的模型的调整的$R_a^2 = 0.54$, 而仅含截距项,Population, Illiteracy的模型的$R_a^2 = 0.55$. 此时可以发现含预测变量越少的模型调整的R_a^2越大, 图1表明双预测变量模型(Population和Illiteracy)是最佳模型.

在图2中可以看出对于不同子集的大小, 基于Mallows Cp统计量的四个最佳模型. 越好的模型离截距项和斜率均为1的直线越近, , 图形表明可以选择以下四种模型: 含Population和Illiteracy的双变量模型, 含Population, Illiteracy和Frost的三变量模型, 含Population, Illiteracy和Income的三变量模型(在图形上重叠了, 不易分辨), 含Population, Illiteracy, Frost和Income的四变量模型.

3.3 AIC与SBC(BIC)判别准则

- AIC: Akaike Information Criterion (赤池信息准则)
- BIC: Bayesian Information Criterion

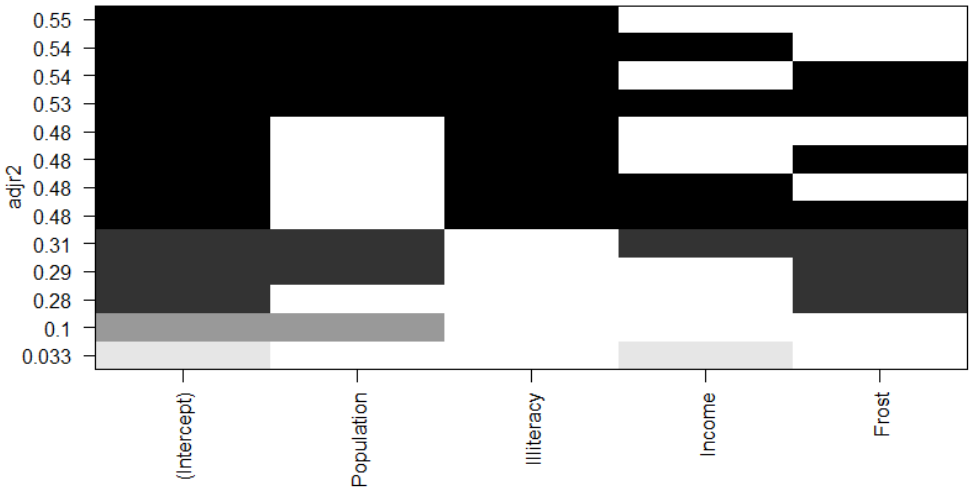


图 1: 基于调整的R^2, 不同子集大小的四个最佳模型

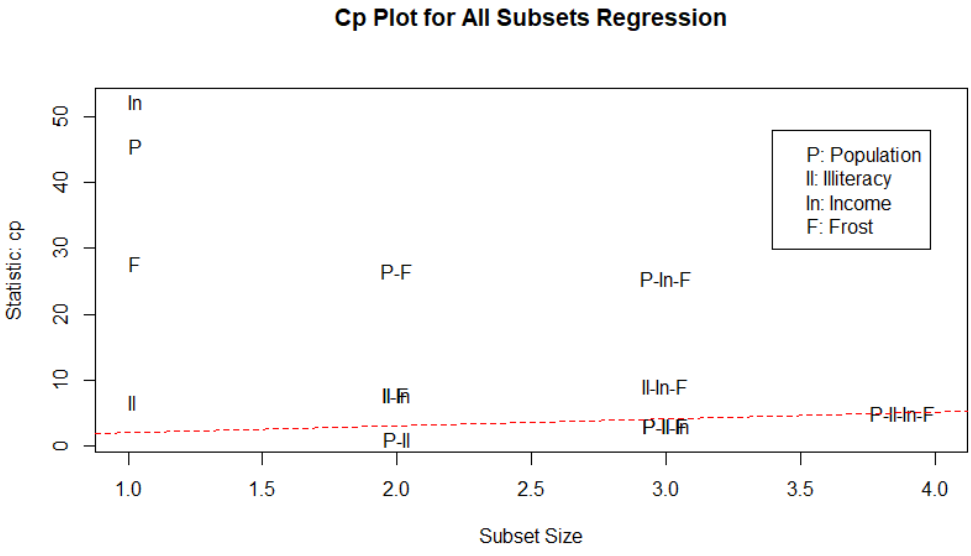


图 2: 基于Mallows Cp统计量, 不同子集大小的四个最佳模型

AIC方法考虑了模型的统计拟合优度以及用来拟合的参数的数目, AIC值较小的模型要优先选择, 该准则可以用`AIC()`函数实现

```
> AIC(fit1, fit2)
     df      AIC
fit1  6 241.6429
fit2  4 237.6565
```

$$\ln L_p\left(\boldsymbol{\beta}, \sigma^2\right) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \mu_i\right)^2$$

$$\text{where } \mu_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_{p-1} X_{p-1,i}$$

$$\ln L_p\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2\right) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2} - \frac{n}{2}\ln\left(\frac{SSE_p}{n}\right)$$

AIC(Akaike信息准则)和SBC(BIC)准则是基于最小化对数似然函数加上一个惩罚项.

$$AIC_p = n\ln\left(\frac{SSE_p}{n}\right) + 2p$$

$$SBC_p = n\ln\left(\frac{SSE_p}{n}\right) + [\ln(n)]p$$

AIC和BIC可以用来比较**非嵌套模型**.

3.4 $PRESS_p$判别准则

所谓PRESS, 即Prediction Residual Error Sum of Squares, 观察预测平方和, 它量化了拟合值预测观测响应的效果. 对于每一种情况i, 使用由其他$n-1$种情况生成的模型预测Y_i 即留一法交叉验证,

$$PRESS(p) = \sum(Y_i - \hat{Y}_{i(-i)})^2$$

4 程序化模型变量选择

1. 最优子集选择算法
2. 向后剔除法: 从包括所有X开始, 找出F^*值最小的那个变量
3. 向前选元法: 逐步回归方法的简化, 略去了检验进入模型的变量是否应该剔除

```
> rm(list = ls())
> alldat = read.table('./Data_5e/Data_5e/CH09TA01.txt')
> dat0 = alldat[1:54,c(1:9)]
> names(dat0) = c('X1','X2','X3','X4','X5','X6','X7','X8','Y')
> full=lm(Y~X1+X2+X3+X4+X5+X6+X7+X8,data = dat0)
> null=lm(Y~1,data = dat0)
> step(null,scope = list(upper=full,lower=null), direction = 'both')
Start:  AIC=647.36
Y ~ 1

       Df Sum of Sq     RSS    AIC
```

```
+ X4     1   3804272 4565248 616.63
+ X3     1   2798310 5571211 627.38
+ X2     1   1479767 6889754 638.85
+ X8     1   1454057 6915463 639.06
+ X1     1   1005152 7364369 642.45
<none>               8369521 647.36
+ X7     1    271062 8098458 647.58
+ X6     1    251809 8117712 647.71
+ X5     1    118863 8250658 648.59

Step:  AIC=616.63
Y ~ X4

       Df Sum of Sq     RSS    AIC
+ X8     1    939077 3626171 606.19
+ X3     1    896004 3669244 606.83
+ X2     1    285592 4279656 615.14
+ X7     1    225396 4339852 615.90
<none>               4565248 616.63
+ X6     1      6162 4559086 618.56
+ X5     1      3726 4561522 618.59
+ X1     1       685 4564563 618.62
- X4     1   3804272 8369521 647.36

Step:  AIC=606.19
Y ~ X4 + X8

       Df Sum of Sq     RSS    AIC
+ X3     1    772164 2854006 595.26
+ X2     1    459359 3166812 600.88
<none>               3626171 606.19
+ X1     1     25196 3600975 607.82
+ X5     1     22048 3604123 607.86
+ X7     1       785 3625386 608.18
+ X6     1       443 3625727 608.19
- X8     1    939077 4565248 616.63
- X4     1   3289293 6915463 639.06

Step:  AIC=595.26
Y ~ X4 + X8 + X3

       Df Sum of Sq     RSS    AIC
+ X2     1    762846 2091160 580.47
<none>               2854006 595.26
+ X1     1     89836 2764170 595.54
+ X7     1      5608 2848399 597.16
+ X5     1      4896 2849110 597.17
+ X6     1         6 2854000 597.26
- X3     1    772164 3626171 606.19
- X8     1    815237 3669244 606.83
- X4     1   1684225 4538232 618.31

Step:  AIC=580.47
Y ~ X4 + X8 + X3 + X2

```

```
       Df Sum of Sq      RSS    AIC
+ X1    1    257444 1833716 575.38
<none>               2091160 580.47
+ X5    1      1326 2089835 582.44
+ X6    1        90 2091070 582.47
+ X7    1        61 2091100 582.47
- X4    1    641710 2732871 592.92
- X2    1    762846 2854006 595.26
- X8    1   1016066 3107226 599.85
- X3    1   1075652 3166812 600.88

Step:  AIC=575.38
Y ~ X4 + X8 + X3 + X2 + X1

       Df Sum of Sq      RSS    AIC
<none>               1833716 575.38
- X4    1     79920 1913636 575.68
+ X5    1      4281 1829436 577.25
+ X6    1      2360 1831356 577.31
+ X7    1       217 1833499 577.37
- X1    1    257444 2091160 580.47
- X8    1    752123 2585839 591.94
- X2    1    930454 2764170 595.54
- X3    1   1324076 3157793 602.73

Call:
lm(formula = Y ~ X4 + X8 + X3 + X2 + X1, data = dat0)

Coefficients:
(Intercept)           X4           X8           X3           X2           X1
  -1178.330       58.064      317.848        9.748        8.924       59.864

> step(null,scope = list(upper=full,lower=null), direction = 'forward')
Start:  AIC=647.36
Y ~ 1

       Df Sum of Sq      RSS    AIC
+ X4    1   3804272 4565248 616.63
+ X3    1   2798310 5571211 627.38
+ X2    1   1479767 6889754 638.85
+ X8    1   1454057 6915463 639.06
+ X1    1   1005152 7364369 642.45
<none>               8369521 647.36
+ X7    1    271062 8098458 647.58
+ X6    1    251809 8117712 647.71
+ X5    1    118863 8250658 648.59

Step:  AIC=616.63
Y ~ X4

       Df Sum of Sq      RSS    AIC
+ X8    1    939077 3626171 606.19
+ X3    1    896004 3669244 606.83
+ X2    1    285592 4279656 615.14
+ X7    1    225396 4339852 615.90
```

```
<none>              4565248 616.63
+ X6    1      6162 4559086 618.56
+ X5    1      3726 4561522 618.59
+ X1    1       685 4564563 618.62

Step:  AIC=606.19
Y ~ X4 + X8

       Df Sum of Sq     RSS    AIC
+ X3    1    772164 2854006 595.26
+ X2    1    459359 3166812 600.88
<none>              3626171 606.19
+ X1    1     25196 3600975 607.82
+ X5    1     22048 3604123 607.86
+ X7    1       785 3625386 608.18
+ X6    1       443 3625727 608.19

Step:  AIC=595.26
Y ~ X4 + X8 + X3

       Df Sum of Sq     RSS    AIC
+ X2    1    762846 2091160 580.47
<none>              2854006 595.26
+ X1    1     89836 2764170 595.54
+ X7    1      5608 2848399 597.16
+ X5    1      4896 2849110 597.17
+ X6    1         6 2854000 597.26

Step:  AIC=580.47
Y ~ X4 + X8 + X3 + X2

       Df Sum of Sq     RSS    AIC
+ X1    1    257444 1833716 575.38
<none>              2091160 580.47
+ X5    1      1326 2089835 582.44
+ X6    1        90 2091070 582.47
+ X7    1        61 2091100 582.47

Step:  AIC=575.38
Y ~ X4 + X8 + X3 + X2 + X1

       Df Sum of Sq     RSS    AIC
<none>              1833716 575.38
+ X5    1    4280.8 1829436 577.25
+ X6    1    2360.3 1831356 577.31
+ X7    1     217.0 1833499 577.37

Call:
lm(formula = Y ~ X4 + X8 + X3 + X2 + X1, data = dat0)

Coefficients:
(Intercept)           X4           X8           X3           X2           X1
  -1178.330       58.064      317.848        9.748        8.924       59.864

> step(full,scope = list(upper=full,lower=null), direction = 'backward')
```

```
Start:  AIC=581.14
Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8

       Df Sum of Sq     RSS    AIC
- X7    1       572 1826478 579.16
- X6    1      2990 1828896 579.23
- X5    1      5231 1831137 579.30
- X4    1     51016 1876922 580.63
<none>              1825906 581.14
- X1    1    263780 2089686 586.43
- X8    1    576636 2402542 593.96
- X2    1    930187 2756093 601.38
- X3    1   1307757 3133663 608.31

Step:  AIC=579.16
Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X8

       Df Sum of Sq     RSS    AIC
- X6    1      2958 1829436 577.25
- X5    1      4878 1831356 577.31
- X4    1     51951 1878429 578.68
<none>              1826478 579.16
- X1    1    263219 2089697 584.43
- X8    1    726329 2552807 595.24
- X2    1    936544 2763022 599.51
- X3    1   1309433 3135911 606.35

Step:  AIC=577.25
Y ~ X1 + X2 + X3 + X4 + X5 + X8

       Df Sum of Sq     RSS    AIC
- X5    1      4281 1833716 575.38
- X4    1     63596 1893032 577.09
<none>              1829436 577.25
- X1    1    260399 2089835 582.44
- X8    1    723371 2552807 593.24
- X2    1    934511 2763947 597.53
- X3    1   1306483 3135918 604.35

Step:  AIC=575.38
Y ~ X1 + X2 + X3 + X4 + X8

       Df Sum of Sq     RSS    AIC
<none>              1833716 575.38
- X4    1     79920 1913636 575.68
- X1    1    257444 2091160 580.47
- X8    1    752123 2585839 591.94
- X2    1    930454 2764170 595.54
- X3    1   1324076 3157793 602.73

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X8, data = dat0)

Coefficients:
(Intercept)           X1           X2           X3           X4           X8
```

```
  -1178.330       59.864        8.924        9.748       58.064      317.848
```

5 模型验证

当我们有大量数据时, 我们希望看到一个模型对一组数据(训练样本)的适合程度与对一组新数据(验证样本)的适合程度如何, 以及训练模型对新数据的适合程度如何.

训练集的观察量应至少是潜在预测量的6-10倍.

训练模型应用于验证样本时的均方预测误差:

$$MSPR = \frac{\sum_{i=1}^{n^*}\left(Y_i^V - \hat{Y}_i^V\right)^2}{n^*} \quad \hat{Y}_i^V = b_0^T + b_1^T X_{i1}^V + \ldots + b_{p-1}^T X_{i,p-1}^V$$

在$k-$折交叉验证中, 样本被分为k个子样本, 轮流将$k-1$个子样本作为训练集, 另外一个子集保留为验证集, 再对k个预测方程的求值取平均, 当$k=n$时, 该方法又被称为刀切法(jackknifing).

```r
library(bootstrap)
shrinkage <- function(fit, k = 10)
{
    require(bootstrap)
    theta.fit <- function(x, y) {
        lsfit(x, y)
    }
    theta.predict <- function(fit, x) {
        cbind(1, x) %*% fit$coef
    }
    x <- fit$model[, 2:ncol(fit$model)]
    y <- fit$model[, 1]
    res <- crossval(x, y, theta.fit, theta.predict, ngroup = k)
    r2 <- cor(y, fit$fitted.values) ^ 2
    r2cv <- cor(y, res$cv.fit) ^ 2
    cat("Original R-square = ", r2, "\n")
    cat(k, "Fold Cross-Validation R-square = ", r2cv, "\n")
    cat("Change = ", r2 - r2cv, "\n")
}

states <-
    as.data.frame(state.x77[, c("Murder", "Population", "Illiteracy", "Income", "Frost")])
fit1 <-
    lm(Murder ~ Population + Illiteracy + Income + Frost, data = states)
fit2 <- lm(Murder ~ Population + Illiteracy, data = states)


shrinkage(fit1)
shrinkage(fit2)
```

Listing 3: R^2k-折交叉验证

```
> shrinkage(fit1)
Original R-square =  0.5669502
10 Fold Cross-Validation R-square =  0.4600774
Change =  0.1068728
> shrinkage(fit2)
```

```
Original R-square =  0.5668327
10 Fold Cross-Validation R-square =  0.5024496
Change =  0.06438311
```

从上述结果我们可以看到, 基于初始样本的$R^2 = 0.567$过于乐观了, 对新数据更好方差解释率估计是交叉验证后$R^2 = 0.448$. 由于观测值每次被随机分配到k组中, 故每次得到的结果略有不同, 通过选择更有泛化能力的模型, 即验证的R^2的减少较小, 说明该模型更有吸引力.