第 六 次 作 业

1907402030 熊雄

1 *QUESTION.*

假设样本 $(x_1, \ldots, x_p) \sim N(0, I_p)$, 建立回归模型$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, 其中回归系数 β_0, β_1, β_2, $\beta_3 \sim Unif(1,2)$为随机数, 误差项 $\varepsilon \sim N(0,1)$, 并且 $p \in \{10, 20\}$

请重复下述过程 n 次,记录正确(即选择 x_1, x_2, x_3), 多选, 少选(即不全是 x_1, x_2, x_3), 错选的次数(即无 x_1, x_2, x_3), 其中 $n \in \{200, 500, 1000\}$.

1. 生成 β_0, β_1, β_2, β_3;
2. 生成 ε 和 X 数据;
3. 生成 Y 数据;
4. 利用 AIC 准则, BIC准则, R_{adj}, C_p 统计量寻找最优子集.

2 *ANSWER.*

2.1 CODE.

输入以下代码:

```
n <- 200
p <- 10
correct = 0 # 记录正确的个数
add = 0 # 记录多选的个数
less = 0 # 少选的个数
error = 0 # 记录错误的个数
for (i in 1:200) # 与n = 200一起进行动态调整
{
  xmean <- matrix(c(runif(p, 0, 0)), ncol = 1)
  xsigma <- diag(p)
  library(MASS)
  x <- mvrnorm(n,xmean,xsigma)
  beta <- matrix(runif(3, 1, 2))
  beta0 <- matrix(rep(runif(1, 1, 2), n))
```

```
  e <- matrix(rnorm(n, 0, 1))
  y <- x[,c(1:3)]%*%beta + beta0 + e
  dx <- data.frame(x)
  library(leaps)
  sub.fit <- regsubsets(y~.,data = dx)
  best.summary <- summary(sub.fit)
  k <- which.max(best.summary$adjr2) # 调整R方
  k <- which.min(best.summary$bic) # BIC
  k <- which.min(best.summary$cp) # C_p
  best.summary$aic = best.summary$bic - p * log(n)+2 * p # AIC
  k <- which.min(best.summary$aic)
  jg <- coef(sub.fit,k)
  xz <- matrix(rownames(data.frame(jg)),ncol = 1)
  if('X1'%in%xz & 'X2'%in%xz & 'X3'%in%xz )
    {
      if(length(xz) == 4)
        correct = correct + 1
      else
        add = add + 1
    }
  else if('X4'%in%xz | 'X5'%in%xz | 'X6'%in%xz | 'X7'%in%xz |
'X8'%in%xz | 'X9'%in%xz |'X10'%in%xz)
    error = error + 1
  else
    less = less + 1
}
```

2.2 RESULT and ANALYSIS.

由上述代码知， n 为200时的运行结果为:

```
add = 24 # 记录多选的个数为70
correct = 176 # 记录正确的个数为430
less = 0 # 少选的个数为0
error = 0 # 记录错误的个数为0
```

将 n 调整为500，运行结果为:

```
add = 34 # 记录多选的个数为34
correct = 466 # 记录正确的个数为466
less = 0 # 少选的个数为0
error = 0 # 记录错误的个数为0
```

将 n 调整为1000，运行结果为：

```
add = 34 # 记录多选的个数为42
correct = 466 # 记录正确的个数为958
less = 0 # 少选的个数为0
error = 0 # 记录错误的个数为0
```

故大致可以看出，当 n 进行改变时，利用各个准则求得的最优子集记录正确的个数比记录多选的个数多得多，并且记录少选或错选的情况几乎不可能发生.