回归分析第六次作业参考解答

第一题

注：这道题计算量偏大，所以答案用 R 语言计算。大家还是需要熟练掌握相关的概念、方法与计算过程。

(a) 按 S_p 序列算得所有子集回归的残差平方和 RSS，并分别按 RMS_q, C_p, AIC, BIC 准则选出最优子集。

```
# 生成 Sp 序列的函数
Sp <- function(n){
 if(n<2){
  S <- c(n)
 }else if(n>20){
  S <- -1
 }else{
  S <- c(Sp(n-1),n,-rev(Sp(n-1)))
 }
 S
}
```

```
# 矩阵消去变换函数
MatrixEli <- function(A,i){
 n <- nrow(A)
 B <- matrix(0,n,n)
 B[i,i] <- 1/A[i,i]
 for(j in 1:n){
  if(j!=i){
    B[i,j] <- A[i,j]/A[i,i]
    B[j,i] <- -A[j,i]/A[i,i]
  }
 }
 for(j in 1:n){
  if(j!=i){
    for(k in 1:n){
      if(k!=i){
        B[j,k] <- A[j,k]-A[j,i]*A[i,k]/A[i,i]
      }
```

```r
    }
  }
 }
 B
}
```

```r
# 输入初始的矩阵与数据
M <- matrix(c(4112.5,-4291.5,38.85,-26.97,70.16,-4291.5,4834.5,-97.35,16.62,-90.98,38.85
,-97.35,27.545,5.73,6.33,-26.97,16.62,5.73,3.66,0.90,70.16,-90.98,6.33,0.90,2.677),nrow=
5,ncol=5)
n <- 18
p <- 5
s <- Sp(p-1)
# 数据框 d 存储系数估计值与 RMSq 等准则
d <- as.data.frame(matrix(0,2^(p-1)-1,p+5))
colnames(d) <- c("beta1","beta2","beta3","beta4","RSS","RMSq","Cp","abs(Cp-q)","AIC","BIC")
```

```r
M
```

```
##          [,1]     [,2]    [,3]   [,4]    [,5]
## [1,]  4112.50 -4291.50  38.850 -26.97  70.160
## [2,] -4291.50  4834.50 -97.350  16.62 -90.980
## [3,]    38.85   -97.35  27.545   5.73   6.330
## [4,]   -26.97    16.62   5.730   3.66   0.900
## [5,]    70.16   -90.98   6.330   0.90   2.677
```

```r
# 创建向量存储回归变量
f <- c()
# 循环，依次计算回归的结果
for(j in 1:(2^(p-1)-1)){
 # 此次计算过程中涉及的变量
 if(s[j] > 0){
  f <- sort(c(f,s[j]))
 }else{
  f <- f[-which(f==abs(s[j]))]
 }
 # 作一次消去变换
 M <- MatrixEli(M,abs(s[j]))
 # 记录计算的结果：回归系数、RSS
 d[j,f] <- M[f,p]
 d[j,5] <- M[p,p]
```

```r
}
# 计算残差均方
sigma2 <- d[2^(p-2)+2,5]/(n-p)
for(j in 1:2^(p-1)-1){
  q <- sum(d[j,1:4]!=0)+1
  # 计算 RMSq
  d[j,6] <- d[j,5]/(n-q)
  # 计算 Cp
  d[j,7] <- d[j,5]/sigma2 - (n-2*q)
  # 计算 Cp-q 绝对值
  d[j,8] <- abs(d[j,7]-q)
  # 计算 AIC
  d[j,9] <- n*log(d[j,5]) + 2*q
  # 计算 BIC
  d[j,10] <- n*log(d[j,5]) + 2*q*log(n)
}
d
```

```
##           beta1       beta2     beta3       beta4       RSS       RMSq
## 1   0.01706018  0.00000000 0.0000000  0.00000000 1.4800576 0.09250360
## 2  -0.03498658 -0.04987587 0.0000000  0.00000000 0.5939516 0.03959677
## 3   0.00000000 -0.01881891 0.0000000  0.00000000 0.9648560 0.06030350
## 4   0.00000000 -0.01527876 0.1758073  0.00000000 0.1740782 0.01160521
## 5  -0.01043393 -0.02493139 0.1564091  0.00000000 0.1507176 0.01076554
## 6   0.01509031  0.00000000 0.2085221  0.00000000 0.2983187 0.01988791
## 7   0.00000000  0.00000000 0.2298058  0.00000000 1.2223295 0.07639559
## 8   0.00000000  0.00000000 0.2649356 -0.16887461 1.1519447 0.07679631
## 9   0.01553776  0.00000000 0.1971159  0.05179756 0.2925094 0.02089353
## 10 -0.01015631 -0.02519405 0.1419279  0.06326854 0.1420667 0.01092821
## 11  0.00000000 -0.01583452 0.1597631  0.06768505 0.1641585 0.01172560
## 12  0.00000000 -0.01997611 0.0000000  0.33661283 0.5566219 0.03710813
## 13 -0.02480643 -0.04170717 0.0000000  0.25249829 0.3956523 0.02826088
## 14  0.01962101  0.00000000 0.0000000  0.39048597 0.9489525 0.06326350
## 15  0.00000000  0.00000000 0.0000000  0.24590164 2.4556885 0.15348053
##             Cp   abs(Cp-q)        AIC         BIC
## 1   121.434592 119.4345917  11.057458  18.6189452
## 2    42.350309  39.3503090  -3.377235   7.9649956
## 3    74.290396  72.2903959   3.356024  10.9175107
## 4     3.929250   0.9292501 -25.468513 -14.1262829
## 5     3.791609   0.2083906 -26.062253 -10.9392793
## 6    15.298036  12.2980364 -15.772674  -4.4304439
```

```
## 7    97.850843  95.8508434   7.613652  15.1751388
## 8    93.410195  90.4101951   8.546128  19.8883587
## 9    16.766448  12.7664480 -14.126655   0.9963186
## 10    5.000000   0.0000000 -25.126252  -6.2225343
## 11    5.021534   1.0215338 -24.524615  -9.4016407
## 12   38.934413  35.9344129  -4.545642   6.7965883
## 13   26.204676  22.2046759  -8.689951   6.4330228
## 14   74.835131  71.8351309   5.056863  16.3990931
## 15  210.710965 208.7109649  20.171329  27.7328163
```

RMS_q, AIC, BIC 准则的评价标准都是愈小愈好，因此 RMS_q, AIC 准则选出的最优子集都是 $\{X_1, X_2, X_3\}$，BIC 准则选出的最优子集是 $\{X_2, X_3\}$。C_p 准则要综合考虑两方面：C_p 愈小愈好、C_p 与 q 愈接近愈好，二者结果不同但比较接近时一般优先考虑 C_p 愈小愈好，因此 C_p 准则选出的最优子集是 $\{X_1, X_2, X_3\}$。

(b) 分别用逐步回归法、向前法、向后法选择最优变量子集，检验水平 $\alpha = 0.05$。

逐步回归法

步骤 1：引入第一个自变量

```r
# 初始增广矩阵
M1 <- matrix(c(4112.5,-4291.5,38.85,-26.97,70.16,-4291.5,4834.5,-97.35,16.62,-90.98,38.85,
             -97.35,27.545,5.73,6.33,-26.97,16.62,5.73,3.66,0.90,70.16,-90.98,6.33,0.90,
             2.677),nrow=5,ncol=5)
M1
```

```
##          [,1]     [,2]    [,3]   [,4]    [,5]
## [1,]  4112.50 -4291.50  38.850 -26.97  70.160
## [2,] -4291.50  4834.50 -97.350  16.62 -90.980
## [3,]    38.85   -97.35  27.545   5.73   6.330
## [4,]   -26.97    16.62   5.730   3.66   0.900
## [5,]    70.16   -90.98   6.330   0.90   2.677
```

```r
c(M1[1,5],M1[2,5],M1[3,5],M1[4,5])^2/c(M1[1,1],M1[2,2],M1[3,3],M1[4,4])
```

```
## [1] 1.1969424 1.7121440 1.4546705 0.2213115
```

最大值为 $p_2^{(1)} = 1.712$

```r
16*1.712/(2.677-1.712)
```

```
## [1] 28.38549
```

```r
qf(0.05,1,16,lower.tail = FALSE)
```

```
## [1] 4.493998
```

引入变量 X_2。

步骤 2：由于回归方程只引入变量 X_2，此步骤不考虑剔除变量，继续引入变量。

```r
M2 <- MatrixEli(M1,2)
M2
```

```
##             [,1]          [,2]         [,3]         [,4]         [,5]
## [1,] 303.0114800  0.8876822836 -47.56587031 -12.216720447 -10.60133416
## [2,]  -0.8876823  0.0002068466  -0.02013652   0.003437791  -0.01881891
## [3,] -47.5658703  0.0201365188  25.58470990   6.064668942   4.49797952
## [4,] -12.2167204 -0.0034377909   6.06466894   3.602863916   1.21277021
## [5,] -10.6013342  0.0188189058   4.49797952   1.212770214   0.96485595
```

```r
c(M2[1,5],M2[3,5],M2[4,5])^2/c(M2[1,1],M2[3,3],M2[4,4])
```

```
## [1] 0.3709044 0.7907778 0.4082340
```

最大值为 $p_3^{(2)} = 0.791$

```r
15*0.791/(0.965-0.791)
```

```
## [1] 68.18966
```

```r
qf(0.05,1,15,lower.tail = FALSE)
```

```
## [1] 4.543077
```

引入变量 X_3。

步骤 3：由本章习题 1 知，此步骤不考虑剔除变量，继续引入变量。

```r
M3 <- MatrixEli(M2,3)
M3
```

```
##             [,1]          [,2]          [,3]         [,4]        [,5]
## [1,] 214.5792864  0.9251191372  1.8591522240 -0.941577695 -2.23890553
## [2,]  -0.9251191  0.0002226951  0.0007870528  0.008211006 -0.01527876
## [3,]  -1.8591522  0.0007870528  0.0390858448  0.237042709  0.17580733
## [4,]  -0.9415777 -0.0082110058 -0.2370427090  2.165278361  0.14655696
## [5,]  -2.2389055  0.0152787582 -0.1758073295  0.146556963  0.17407818
```

```r
c(M3[1,5],M3[4,5])^2/c(M3[1,1],M3[4,4])
```

```
## [1] 0.023360586 0.009919715
```

最大值为 $p_1^{(3)} = 0.0234$

```r
14*0.0234/(0.1741-0.0234)
```

```
## [1] 2.173855
```

```r
qf(0.05,1,14,lower.tail = FALSE)
```

```
## [1] 4.60011
```

此步骤不引入变量。此时回归方程包含 X_2, X_3 两个变量，由本章习题 1 知，接下来无需考虑剔除变量，逐步回归法过程结束，选出的最优子集为 $\{X_2, X_3\}$。

向前法

此问题向前法的过程与逐步回归法一致，因此向前法选出的最优子集为 $\{X_2, X_3\}$。

向后法

步骤 1：剔除第一个变量

```r
# 初始增广矩阵
N1 <- matrix(c(0.004669,0.004303,0.008199,0.002031,0.0102,0.004303,0.004219,0.009242,
               -0.001922,0.0252,0.008199,0.009242,0.079442,-0.105938,-0.1416,0.002031,
               -0.001922,-0.105938,0.462846,-0.0637,-0.0102,-0.0252,0.1416,0.0637,0.142),
             ncol=5,nrow=5)
N1
```

```
##          [,1]      [,2]      [,3]      [,4]    [,5]
## [1,] 0.004669  0.004303  0.008199  0.002031 -0.0102
## [2,] 0.004303  0.004219  0.009242 -0.001922 -0.0252
## [3,] 0.008199  0.009242  0.079442 -0.105938  0.1416
## [4,] 0.002031 -0.001922 -0.105938  0.462846  0.0637
## [5,] 0.010200  0.025200 -0.141600 -0.063700  0.1420
```

```r
c(N1[1,5],N1[2,5],N1[3,5],N1[4,5])^2/c(N1[1,1],N1[2,2],N1[3,3],N1[4,4])
```

```
## [1] 0.022283144 0.150519080 0.252392437 0.008766825
```

最小值为 $p_4^{(1)} = 0.0088$

```
13*0.0088/0.142
```

```
## [1] 0.8056338
```

```
qf(0.05,1,13,lower.tail = FALSE)
```

```
## [1] 4.667193
```

剔除变量 X_4。

步骤 2：剔除下一个变量

```
N2 <- MatrixEli(N1,4)
N2
```

```
##             [,1]         [,2]         [,3]         [,4]        [,5]
## [1,] 0.004660088  0.004311434  0.008663863 -0.004388069 -0.01047952
## [2,] 0.004311434  0.004211019  0.008802085  0.004152569 -0.02493548
## [3,] 0.008663863  0.008802085  0.055194497  0.228883905  0.15617990
## [4,] 0.004388069 -0.004152569 -0.228883905  2.160545840  0.13762677
## [5,] 0.010479520  0.024935481 -0.156179905  0.137626770  0.15076683
```

```
c(N2[1,5],N2[2,5],N2[3,5])^2/c(N2[1,1],N2[2,2],N2[3,3])
```

```
## [1] 0.02356615 0.14765506 0.44193106
```

最小值为 $p_1^{(2)} = 0.0236$

```
14*0.0236/0.1508
```

```
## [1] 2.190981
```

```
qf(0.05,1,14,lower.tail = FALSE)
```

```
## [1] 4.60011
```

剔除变量 X_1。

步骤 3：剔除下一个变量

```
N3 <- MatrixEli(N2,1)
N3
```

```
##              [,1]          [,2]          [,3]         [,4]        [,5]
## [1,] 214.5882301  0.9251829628  1.8591630722 -0.941627875 -2.24878164
## [2,]  -0.9251830  0.0002221536  0.0007864265  0.008212335 -0.01524001
## [3,]  -1.8591631  0.0007864265  0.0390869623  0.237042040  0.17566304
## [4,]  -0.9416279 -0.0082123354 -0.2370420403  2.164677768  0.14749458
## [5,]  -2.2487816  0.0152400080 -0.1756630413  0.147494578  0.17433298
```

```
c(N3[2,5],N3[3,5])^2/c(N3[2,2],N3[3,3])
```

```
## [1] 1.0454831 0.7894577
```

最小值为 $p_3^{(3)} = 0.7895$

```
15*0.7895/0.1743
```

```
## [1] 67.9432
```

```
qf(0.05,1,15,lower.tail = FALSE)
```

```
## [1] 4.543077
```

此步骤无需剔除变量，向后法过程结束，选出的最优子集为 $\{X_2, X_3\}$。

第二题

对 Hald 水泥数据按 S_p 序列作消去变换，求得所有子集回归的系数表和对应的 RSS，并分别按 RMS_q, C_p, AIC, BIC 准则选出最优子集。

```
# 输入数据
n <- 13
p <- 5
x1 <- c(7,1,11,11,7,11,3,1,2,21,1,11,10)
x2 <- c(26,29,56,31,52,55,71,31,54,47,40,66,68)
x3 <- c(6,15,8,8,6,9,17,22,18,4,23,9,8)
x4 <- c(60,52,20,47,33,22,6,44,22,26,34,12,12)
y <- c(78.5,74.3,104.3,87.6,95.9,109.2,102.7,72.5,93.1,115.9,83.8,113.3,109.4)
# 数据中心化
xm1 <- mean(x1)
xm2 <- mean(x2)
xm3 <- mean(x3)
xm4 <- mean(x4)
ym <- mean(y)
```

```
xc1 <- x1-xm1
xc2 <- x2-xm2
xc3 <- x3-xm3
xc4 <- x4-xm4
Xc <- t(rbind(xc1,xc2,xc3,xc4))
s <- Sp(p-1)
# 初始的增广矩阵
M <- cbind(rbind(t(Xc)%*%Xc,y%*%Xc),rbind(t(Xc)%*%y,t(y-ym)%*%(y-ym)))
d <- as.data.frame(matrix(0,2^(p-1)-1,p+5))
colnames(d) <- c("beta1","beta2","beta3","beta4","RSS","RMSq","Cp","abs(Cp-q)","AIC","BIC")
```

```
M
```

```
##           xc1        xc2       xc3     xc4
## xc1  415.2308   251.0769 -372.6154  -290.0   775.9615
## xc2  251.0769  2905.6923 -166.5385 -3041.0  2292.9538
## xc3 -372.6154  -166.5385  492.3077    38.0  -618.2308
## xc4 -290.0000 -3041.0000   38.0000  3362.0 -2481.7000
##      775.9615  2292.9538 -618.2308 -2481.7  2715.7631
```

```
# 创建向量存储回归变量
f <- c()
# 循环，依次计算回归的结果
for(j in 1:(2^(p-1)-1)){
 # 此次计算过程中涉及的变量
 if(s[j] > 0){
  f <- sort(c(f,s[j]))
 }else{
  f <- f[-which(f==abs(s[j]))]
 }
 # 作一次消去变换
 M <- MatrixEli(M,abs(s[j]))
 # 记录计算的结果，回归系数、RSS
 d[j,f] <- M[f,p]
 d[j,5] <- M[p,p]
}
# 计算残差均方
sigma2 <- d[2^(p-2)+2,5]/(n-p)
for(j in 1:2^(p-1)-1){
  q <- sum(d[j,1:4]!=0)+1
  # 计算 RMSq
```

```
  d[j,6] <- d[j,5]/(n-q)
  # 计算 Cp
  d[j,7] <- d[j,5]/sigma2 - (n-2*q)
  # 计算 Cp-q 绝对值
  d[j,8] <- abs(d[j,7]-q)
  # 计算 AIC
  d[j,9] <- n*log(d[j,5]) + 2*q
  # 计算 BIC
  d[j,10] <- n*log(d[j,5]) + 2*q*log(n)
}
d
```

```
##       beta1      beta2      beta3      beta4        RSS       RMSq         Cp
## 1  1.868748  0.0000000  0.0000000  0.0000000 1265.68675 115.062432 202.548769
## 2  1.468306  0.6622505  0.0000000  0.0000000   57.90448   5.790448   2.678242
## 3  0.000000  0.7891248  0.0000000  0.0000000  906.33634  82.394213 142.486407
## 4  0.000000  0.7313296 -1.0083862  0.0000000  415.44273  41.544273  62.437716
## 5  1.695890  0.6569149  0.2500176  0.0000000   48.11061   5.345624   3.041280
## 6  2.312468  0.0000000  0.4944682  0.0000000 1227.07206 122.707206 198.094653
## 7  0.000000  0.0000000 -1.2557813  0.0000000 1939.40047 176.309134 315.154284
## 8  0.000000  0.0000000 -1.1998512 -0.7246001  175.73800  17.573800  22.373112
## 9  1.051854  0.0000000 -0.4100433 -0.6427961   50.83612   5.648458   3.496824
## 10 1.551103  0.5101676  0.1019094 -0.1440610   47.86364   5.982955   5.000000
## 11 0.000000 -0.9234160 -1.4479712 -1.5570449   73.81455   8.201617   7.337474
## 12 0.000000  0.3109047  0.0000000 -0.4569419  868.88013  86.888013 138.225920
## 13 1.451938  0.4161098  0.0000000 -0.2365402   47.97273   5.330303   3.018233
## 14 1.439958  0.0000000  0.0000000 -0.6139536   74.76211   7.476211   5.495851
## 15 0.000000  0.0000000  0.0000000 -0.7381618  883.86692  80.351538 138.730833
##      abs(Cp-q)       AIC       BIC
## 1  200.5487691  96.86381 103.12361
## 2    0.3217584  58.76433  68.15403
## 3  140.4864069  92.52234  98.78213
## 4   59.4377163  84.38148  93.77118
## 5    0.9587203  58.35554  70.87513
## 6  195.0946526  98.46102 107.85072
## 7  313.1542841 102.41174 108.67154
## 8   19.3731120  73.19693  82.58662
## 9    0.5031756  59.07189  71.59149
## 10   0.0000000  60.28863  75.93812
## 11   3.3374740  63.92023  76.43982
## 12 135.2259198  93.97367 103.36336
```

```
## 13   0.9817665  58.31823  70.83782
## 14   2.4958508  62.08605  71.47574
## 15 136.7308335  92.19598  98.45578
```

RMS_q, AIC 准则选出的最优子集都是 $\{X_1, X_2, X_4\}$，C_p, BIC 准则选出的最优子集都是 $\{X_1, X_2\}$。