#输入数据并检查

cement <- data.frame(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))

cement

#进行简单线性回归，查看效果

lm.sol <- lm(Y ~ ., data = cement)

summary(lm.sol)

##   
## Call:  
## lm(formula = Y ~ ., data = cement)  
##   
## Residuals:  
##    Min     1Q Median     3Q    Max   
## -3.175 -1.671  0.251  1.378  3.925   
##   
## Coefficients:  
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   62.405     70.071    0.89    0.399    
## X1             1.551      0.745    2.08    0.071 .  
## X2             0.510      0.724    0.70    0.501    
## X3             0.102      0.755    0.14    0.896    
## X4            -0.144      0.709   -0.20    0.844    
## ---  
## Signif. codes:  0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.45 on 8 degrees of freedom  
## Multiple R-squared:  0.982,  Adjusted R-squared:  0.974   
## F-statistic:  111 on 4 and 8 DF,  p-value: 4.76e-07  
# 从结果看，截距和自变量的相关系数均不显著。  
# 利用car包中的vif（）函数查看各自变量间的共线情况

library(car)  
vif(lm.sol)

##     X1     X2     X3     X4   
##  38.50 254.42  46.87 282.51  
# 从结果看，各自变量的VIF值都超过10，存在多重共线性，其中，X2与X4的VIF值均超过200.  
plot(X2 ~ X4, col = "red", data = cement)

library(MASS)

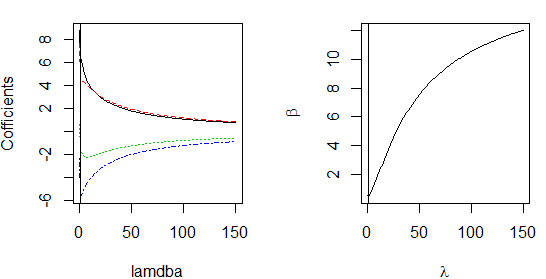
#使用151个lambda值来计算，最后选取使得 广义交叉验证GCV 最小的那个lambda

ridge.sol <- lm.ridge(Y ~ ., lambda = seq(0, 150, length = 151), data = cement,   
    model = TRUE)

names(ridge.sol)  # 变量名字

## [1] "coef"   "scales" "Inter"  "lambda" "ym"     "xm"     "GCV"    "kHKB"    
## [9] "kLW"

ridge.sol$lambda[which.min(ridge.sol$GCV)]  ##找到GCV最小时的lambdaGCV  
## [1] 1  
ridge.sol$coef[which.min(ridge.sol$GCV)]  ##找到GCV最小时对应的系数  
## [1] 7.627  
par(mfrow = c(1, 2))  
# 画出图形，并作出lambdaGCV取最小值时的那条竖直线  
matplot(ridge.sol$lambda, t(ridge.sol$coef), xlab = expression(lamdba), ylab = "Cofficients",   
    type = "l", lty = 1:20)  
abline(v = ridge.sol$lambda[which.min(ridge.sol$GCV)])  
# 下面的语句绘出lambda同GCV之间关系的图形  
plot(ridge.sol$lambda, ridge.sol$GCV, type = "l", xlab = expression(lambda),   
    ylab = expression(beta))  
abline(v = ridge.sol$lambda[which.min(ridge.sol$GCV)])

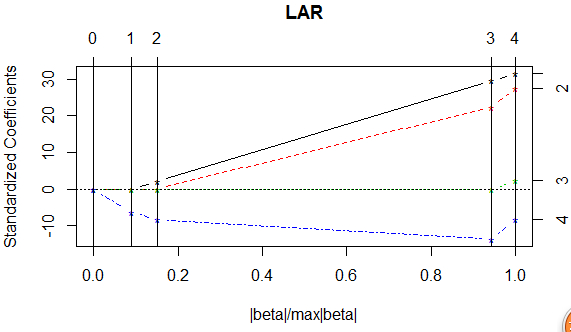


par(mfrow = c(1, 1))  
# 从上图看，lambda的选择并不是那么重要，只要不离lambda=0太近就没有多大差别。  
# 下面利用ridge包中的linearRidge()函数进行自动选择岭回归参数

library(ridge)  
mod <- linearRidge(Y ~ ., data = cement)  
summary(mod)

##   
## Call:  
## linearRidge(formula = Y ~ ., data = cement)  
##   
##   
## Coefficients:  
##             Estimate Scaled estimate Std. Error (scaled) t value (scaled)  
## (Intercept)   83.704              NA                  NA               NA  
## X1             1.292          26.332               3.672             7.17  
## X2             0.298          16.046               3.988             4.02  
## X3            -0.148          -3.279               3.598             0.91  
## X4            -0.351         -20.329               3.996             5.09  
##             Pr(>|t|)      
## (Intercept)       NA      
## X1           7.5e-13 \*\*\*  
## X2           5.7e-05 \*\*\*  
## X3              0.36      
## X4           3.6e-07 \*\*\*  
## ---  
## Signif. codes:  0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Ridge parameter: 0.01473, chosen automatically, computed using 2 PCs  
##   
## Degrees of freedom: model 3.01 , variance 2.84 , residual 3.18  
# 从模型运行结果看，测岭回归参数值为0.0147，各自变量的系数显著想明显提高（除了X3仍不显著）  
# 最后，利用Lasso回归解决共线性问题

library(lars)  
## Loaded lars 1.2  
x = as.matrix(cement[, 1:4])  
y = as.matrix(cement[, 5])  
(laa = lars(x, y, type = "lar"))  #lars函数值用于矩阵型数据  
##   
## Call:  
## lars(x = x, y = y, type = "lar")  
## R-squared: 0.982   
## Sequence of LAR moves:  
##      X4 X1 X2 X3  
## Var   4  1  2  3  
## Step  1  2  3  4  
# 由此可见，LASSO的变量选择依次是X4，X1，X2，X3  
plot(laa)  #绘出图



summary(laa)  #给出Cp值  
## LARS/LAR  
## Call: lars(x = x, y = y, type = "lar")  
##   Df  Rss     Cp  
## 0  1 2716 442.92  
## 1  2 2219 361.95  
## 2  3 1918 313.50  
## 3  4   48   3.02  
## 4  5   48   5.00  
# 根据课上对Cp含义的解释（衡量多重共线性，其值越小越好），我们取到第3步，使得Cp值最小，也就是选择X4，X1，X2这三个变量。