HWp5

Sagar Ganapaneni

November 15, 2016

Problem 5: This question relates to the College data set.

(a)Split the data into a training set and a test set. Using out-of-state tuition as the response and the other variables as the predictors, perform forward stepwise selection on the training set in order to identify a satisfactory model that uses just a subset of the predictors.

library(ISLR)

## Warning: package 'ISLR' was built under R version 3.2.5

library(leaps)

## Warning: package 'leaps' was built under R version 3.2.5

library(gam)

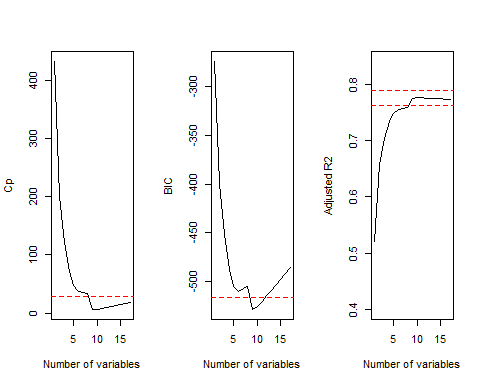
## Warning: package 'gam' was built under R version 3.2.5

## Loading required package: splines

## Loading required package: foreach

## Loaded gam 1.14

data("College")  
attach(College)  
train <- sample(length(Outstate), length(Outstate) / 2)  
test <- -train  
College\_train <- College[train, ]  
College\_test <- College[test, ]  
fit <- regsubsets(Outstate ~ ., data = College\_train, nvmax = 17, method = "forward")  
fit.summary <- summary(fit)  
par(mfrow = c(1, 3))  
plot(fit.summary$cp, xlab = "Number of variables", ylab = "Cp", type = "l")  
min\_cp <- min(fit.summary$cp)  
std\_cp <- sd(fit.summary$cp)  
abline(h = min\_cp + 0.2 \* std\_cp, col = "red", lty = 2)  
abline(h = min\_cp - 0.2 \* std\_cp, col = "red", lty = 2)  
plot(fit.summary$bic, xlab = "Number of variables", ylab = "BIC", type='l')  
min\_bic <- min(fit.summary$bic)  
std\_bic <- sd(fit.summary$bic)  
abline(h = min\_bic + 0.2 \* std\_bic, col = "red", lty = 2)  
abline(h = min\_bic - 0.2 \* std\_bic, col = "red", lty = 2)  
plot(fit.summary$adjr2, xlab = "Number of variables", ylab = "Adjusted R2", type = "l", ylim = c(0.4, 0.84))  
max\_adjR2 <- max(fit.summary$adjr2)  
std\_adjR2 <- sd(fit.summary$adjr2)  
abline(h = max\_adjR2 + 0.2 \* std\_adjR2, col = "red", lty = 2)  
abline(h = max\_adjR2 - 0.2 \* std\_adjR2, col = "red", lty = 2)

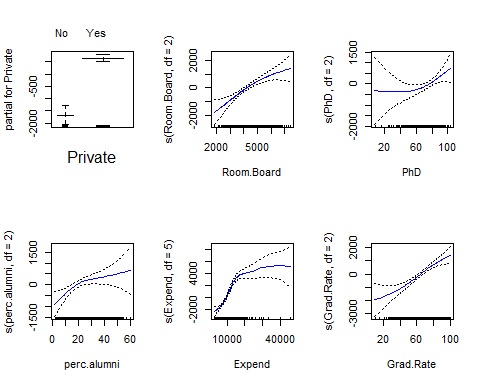
  
Cp, BIC and adjr2 show that size 6 is the minimum size for the subset for which the scores are within 0.2 standard deviations of optimum.

fit <- regsubsets(Outstate ~ ., data = College, method = "forward")  
coeffs <- coef(fit, id = 6)  
names(coeffs)

## [1] "(Intercept)" "PrivateYes" "Room.Board" "PhD" "perc.alumni"  
## [6] "Expend" "Grad.Rate"

1. Fit a GAM on the training data, using out-of-state tuition as the response and the features selected in the previous step as the predictors. Plot the results, and explain your findings.

fit <- gam(Outstate ~ Private + s(Room.Board, df = 2) + s(PhD, df = 2) + s(perc.alumni, df = 2) + s(Expend, df = 5) + s(Grad.Rate, df = 2), data=College\_train)  
par(mfrow = c(2, 3))  
plot(fit, se = T, col = "blue")

  
(c) Evaluate the model obtained on the test set, and explain the results obtained.

preds <- predict(fit, College\_test)  
err <- mean((College\_test$Outstate - preds)^2)  
err

## [1] 3706367

tss <- mean((College\_test$Outstate - mean(College\_test$Outstate))^2)  
rss <- 1 - err / tss  
rss

## [1] 0.7764983

We obtain a test R^2 of 0.77 using GAM with 6 predictors.

summary(fit)

##   
## Call: gam(formula = Outstate ~ Private + s(Room.Board, df = 2) + s(PhD,   
## df = 2) + s(perc.alumni, df = 2) + s(Expend, df = 5) + s(Grad.Rate,   
## df = 2), data = College\_train)  
## Deviance Residuals:  
## Min 1Q Median 3Q Max   
## -7457.88 -1066.58 10.69 1205.80 4269.38   
##   
## (Dispersion Parameter for gaussian family taken to be 3322454)  
##   
## Null Deviance: 6102207011 on 387 degrees of freedom  
## Residual Deviance: 1239276076 on 373.0002 degrees of freedom  
## AIC: 6944.09   
##   
## Number of Local Scoring Iterations: 2   
##   
## Anova for Parametric Effects  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Private 1 1650304201 1650304201 496.712 < 2.2e-16 \*\*\*  
## s(Room.Board, df = 2) 1 1241956325 1241956325 373.807 < 2.2e-16 \*\*\*  
## s(PhD, df = 2) 1 439487340 439487340 132.278 < 2.2e-16 \*\*\*  
## s(perc.alumni, df = 2) 1 220881872 220881872 66.481 5.431e-15 \*\*\*  
## s(Expend, df = 5) 1 617867577 617867577 185.967 < 2.2e-16 \*\*\*  
## s(Grad.Rate, df = 2) 1 145332274 145332274 43.742 1.301e-10 \*\*\*  
## Residuals 373 1239276076 3322454   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Anova for Nonparametric Effects  
## Npar Df Npar F Pr(F)   
## (Intercept)   
## Private   
## s(Room.Board, df = 2) 1 2.6037 0.10745   
## s(PhD, df = 2) 1 2.6635 0.10352   
## s(perc.alumni, df = 2) 1 4.7013 0.03077 \*   
## s(Expend, df = 5) 4 15.5625 8.735e-12 \*\*\*  
## s(Grad.Rate, df = 2) 1 3.1876 0.07501 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

ANOVA shows a strong evidence of non-linear relationship between "Outstate" and "Expend"", and a moderately strong non-linear relationship (using p-value of 0.05) between"Outstate" and "Grad.Rate"" or "PhD".