Stat 897 Fall 2017 Data Analysis Assignment 8

Penn State

Due October 22, 2017

In this assignment we again use the College data found in the ISLR library, with the 600 observation training set (using the rest as the test data).

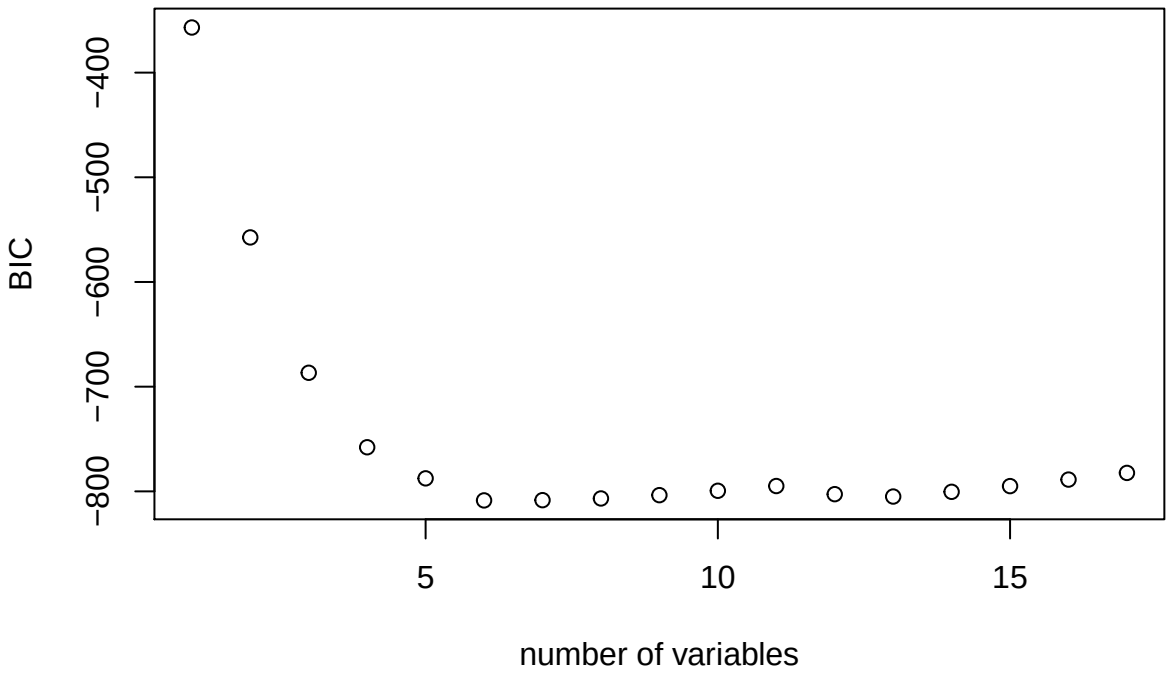
(a) Using out-of-state tuition as the response and the other variables as the predictors, perform forward stepwise selection on the training set in order to identify a satisfactory model that uses just a subset of the predictors. Please set the seed at 801 once at the beginning before choosing the training set.

```
suppressWarnings(library(ISLR))
library(leaps)
suppressWarnings(library(gam, quietly = T))
```

```
## Loaded gam 1.14
```

```
set.seed(801)
train = sample(seq(1:777), 600, replace = FALSE)
College.train = College[ train,]
College.test  = College[-train,]
y.test = College.test$Outstate

### fitting the forward selection model
fit1 = regsubsets(Outstate ~., data = College.train, method = "forward", nvmax = 17)
fit1.summary = summary(fit1)
### plotting the BIC
plot(fit1.summary$bic, xlab = 'number of variables', ylab = 'BIC')
```



```
nvariables = which.min(fit1.summary$bic)
beta1 = coef(fit1, id = nvariables)
varnames = names(beta1)[-1]
varnames[1] = "Private"
### number of variables included in the model by BIC forward selection
nvariables
```

```
## [1] 6
```

```
### coefficients of the chosen variables
beta1
```

```
##   (Intercept)    PrivateYes    Room.Board      Terminal   perc.alumni
## -4123.9790886  2715.3989427     0.9578136    39.0264668    47.6455425
##        Expend     Grad.Rate
##     0.2275206    31.2214554
```

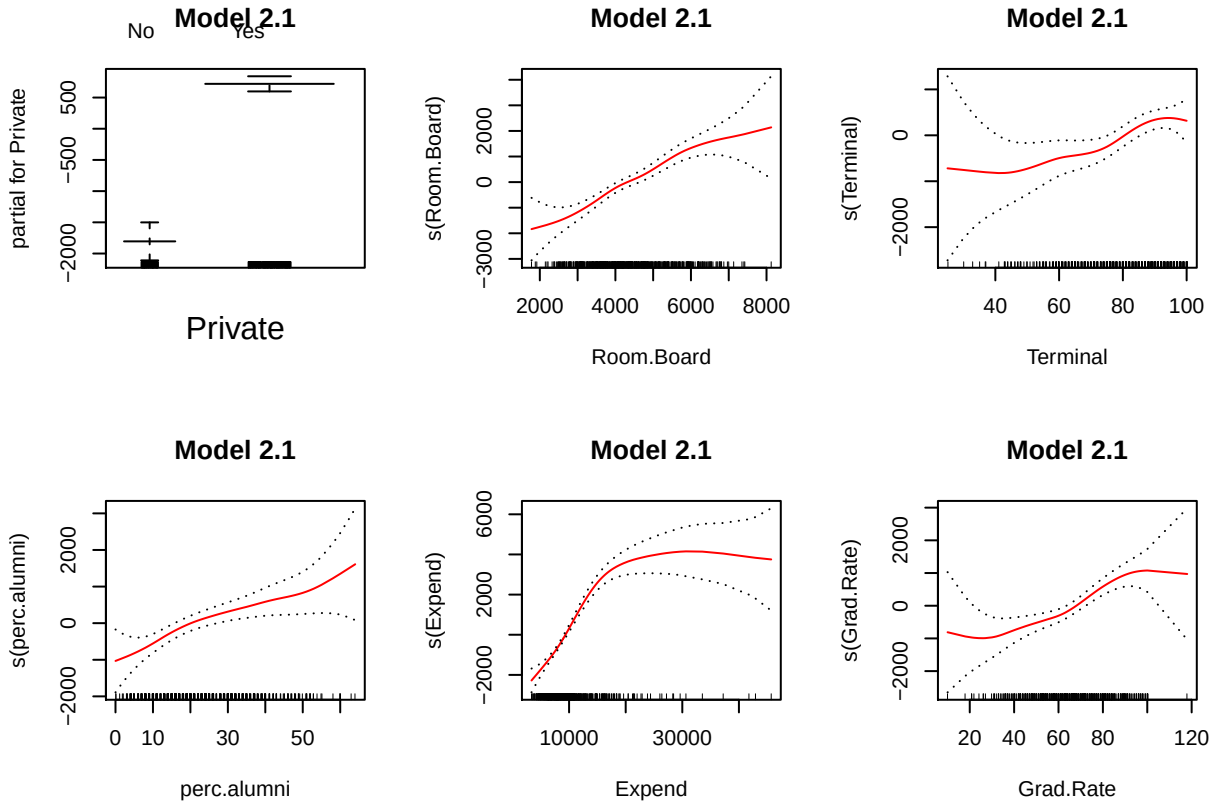
Based on the plot, we can see 11 variables will give the smalleset BIC. However, we will use the model only including 6 variables since (1) it gives a relatively less complex model (2) BIC dose not change a lot even if we include more variables. These 6 variables are

```
### chosen variables
varnames
```

```
## [1] "Private"     "Room.Board"  "Terminal"    "perc.alumni" "Expend"
## [6] "Grad.Rate"
```

(b) 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. Use appropriate nonlinear components (e.g. natural splines, step functions) for the variables that need it as in the salary example from the book. Plot the results, and explain your findings. What nonlinear components did you use?

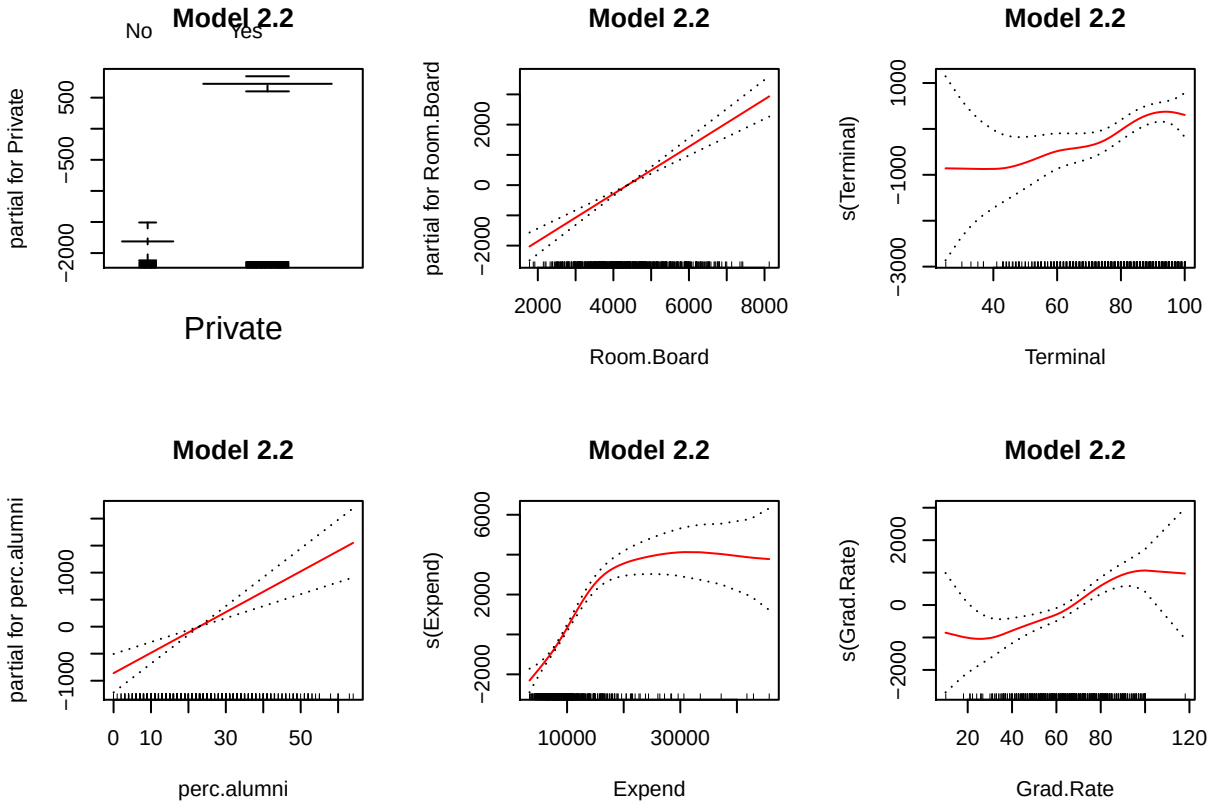
```
fit2.1 = gam(Outstate ~ Private + s(Room.Board) + s(Terminal) + s(perc.alumni) +
                 s(Expend) + s(Grad.Rate), data = College.train)
par(mfrow = c(2,3))
plot(fit2.1, main = "Model 2.1", se = T, col = 'red')
```




```
par(mfrow = c(1,1))
```

We first fit a GAM model with smooth spline without the nonlinear transformation on "Private" since it is a categorical variable. Based on the plot above, we can see the smooth functions for "Room.Board" and "perc.alumni" are quite linear. Therefore, we will not apply nonlinear functions on these two variables.

```
fit2.2 = gam(Outstate ~ Private + Room.Board + s(Terminal) + perc.alumni +
                 s(Expend) + s(Grad.Rate), data = College.train)
par(mfrow = c(2,3))
plot(fit2.2, main = "Model 2.2", se = T, col = 'red')
```




```
par(mfrow = c(1,1))
```

There is a obvious nonliear relationship between "Terminal" and "Outstate" and "Expend" and "Outstate".

(c) Evaluate both models obtained from part (a) and (b) on the test set, and explain the results obtained.

```
### MSE for the forward selection model
fit3 = lm(Outstate ~ Private + Room.Board + Terminal + perc.alumni + Expend +
             Grad.Rate, data = College.train)
ypred1 = predict(fit3, newdata = College.test)
mse1 = mean((ypred1 - y.test)^2)
mse1
```

```
## [1] 4840137
```

```
### MSE for the GAM model
ypred2 = predict(fit2.2, newdata = College.test)

mse2 = mean((ypred2 - y.test)^2)
mse2
```

```
## [1] 3605797
```

We can see the improvement of the GAM model compared with the model chosen by forward selection in terms of the MSE of test data.

(d) For which variables, if any, is there evidence of a non-linear relationship with the response?

There is evidence for variables "Terminal" and "Expend" of a non-linear relationship with the response, but it is hard to tell for the other variables.