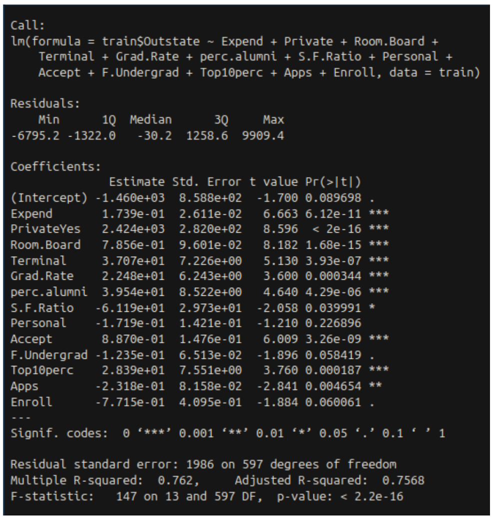
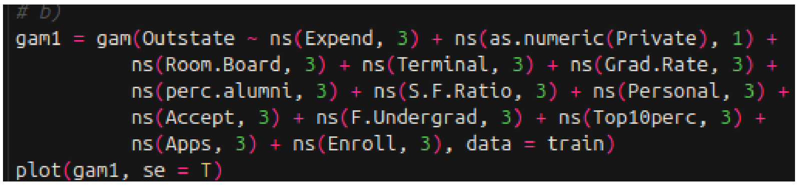
Question 10

a) I split my data into a training and test set at a 4/5 ratio train. I first created an empty linear model, and a full linear model, and then used the step function to perform forward stepwise selection. 14 of the 18 original predictors were used in the Step model.

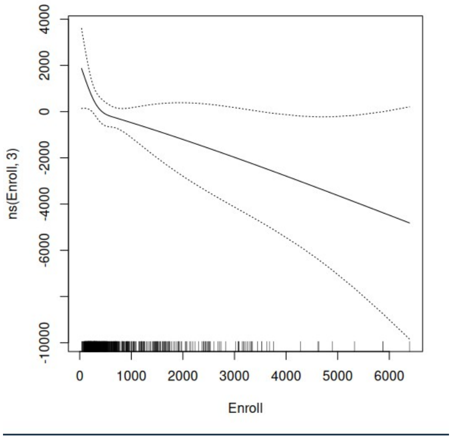


```
Call:
lm(formula = train$Outstate ~ Expend + Private + Room.Board +
    Terminal + Grad.Rate + perc.alumni + S.F.Ratio + Personal +
    Accept + F.Undergrad + Top10perc + Apps + Enroll, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-6795.2 -1322.0   -30.2  1258.6  9909.4

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.460e+03  8.588e+02  -1.700 0.089698 .
Expend       1.739e-01  2.611e-02   6.663 6.12e-11 ***
PrivateYes   2.424e+03  2.820e+02   8.596  < 2e-16 ***
Room.Board   7.856e-01  9.601e-02   8.182 1.68e-15 ***
Terminal     3.707e+01  7.226e+00   5.130 3.93e-07 ***
Grad.Rate    2.248e+01  6.243e+00   3.600 0.000344 ***
perc.alumni  3.954e+01  8.522e+00   4.640 4.29e-06 ***
S.F.Ratio   -6.119e+01  2.973e+01  -2.058 0.039991 *
Personal    -1.719e-01  1.421e-01  -1.210 0.226896
Accept       8.870e-01  1.476e-01   6.009 3.26e-09 ***
F.Undergrad -1.235e-01  6.513e-02  -1.896 0.058419 .
Top10perc    2.839e+01  7.551e+00   3.760 0.000187 ***
Apps        -2.318e-01  8.158e-02  -2.841 0.004654 **
Enroll      -7.715e-01  4.095e-01  -1.884 0.060061 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1986 on 597 degrees of freedom
Multiple R-squared:  0.762,	Adjusted R-squared:  0.7568
F-statistic:   147 on 13 and 597 DF,  p-value: < 2.2e-16
```

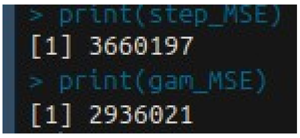
b) I took each of the predictors and created a model using natural splines with 3 degrees of freedom for each. I then plotted the result



```
# b)
gam1 = gam(Outstate ~ ns(Expend, 3) + ns(as.numeric(Private), 1) +
             ns(Room.Board, 3) + ns(Terminal, 3) + ns(Grad.Rate, 3) +
             ns(perc.alumni, 3) + ns(S.F.Ratio, 3) + ns(Personal, 3) +
             ns(Accept, 3) + ns(F.Undergrad, 3) + ns(Top10perc, 3) +
             ns(Apps, 3) + ns(Enroll, 3), data = train)
plot(gam1, se = T)
```

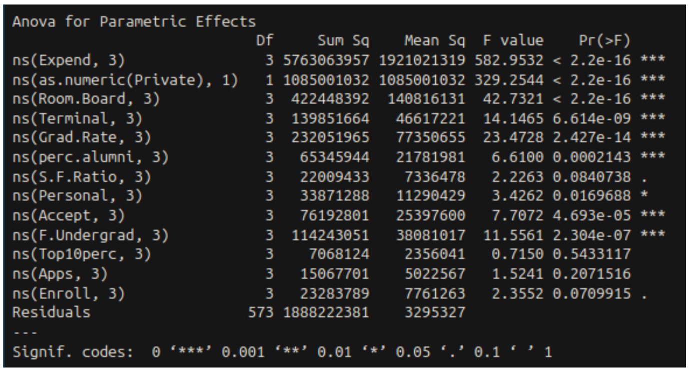


c) I found the test MSE for both the step model and the GAM model. The gam model produced a better result than the stepwise model



```
> print(step_MSE)
[1] 3660197
> print(gam_MSE)
[1] 2936021
```

d) There is evidence that a lot of the variables have a non-linear relationship with the response



```
Anova for Parametric Effects
                            Df     Sum Sq    Mean Sq  F value    Pr(>F)
ns(Expend, 3)                3 5763063957 1921021319 582.9532 < 2.2e-16 ***
ns(as.numeric(Private), 1)   1 1085001032 1085001032 329.2544 < 2.2e-16 ***
ns(Room.Board, 3)            3  422448392  140816131  42.7321 < 2.2e-16 ***
ns(Terminal, 3)              3  139851664   46617221  14.1465 6.614e-09 ***
ns(Grad.Rate, 3)             3  232051965   77350655  23.4728 2.427e-14 ***
ns(perc.alumni, 3)           3   65345944   21781981   6.6100 0.0002143 ***
ns(S.F.Ratio, 3)             3   22009433    7336478   2.2263 0.0840738 .
ns(Personal, 3)              3   33871288   11290429   3.4262 0.0169688 *
ns(Accept, 3)                3   76192801   25397600   7.7072 4.693e-05 ***
ns(F.Undergrad, 3)           3  114243051   38081017  11.5561 2.304e-07 ***
ns(Top10perc, 3)             3    7068124    2356041   0.7150 0.5433117
ns(Apps, 3)                  3   15067701    5022567   1.5241 0.2071516
ns(Enroll, 3)                3   23283789    7761263   2.3552 0.0709915 .
Residuals                  573 1888222381    3295327
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```