Assignment 9
Shaheed Shihan
Modern Applied Statistics

6.
In this exercise, you will further analyze the Wage data set considered throughout this chapter.

(a) Perform polynomial regression to predict wage using age. Use cross-validation to select the optimal degree d for the polynomial. What degree was chosen, and how does this compare to the results of hypothesis testing using ANOVA? Make a plot of the resulting polynomial fit to the data.

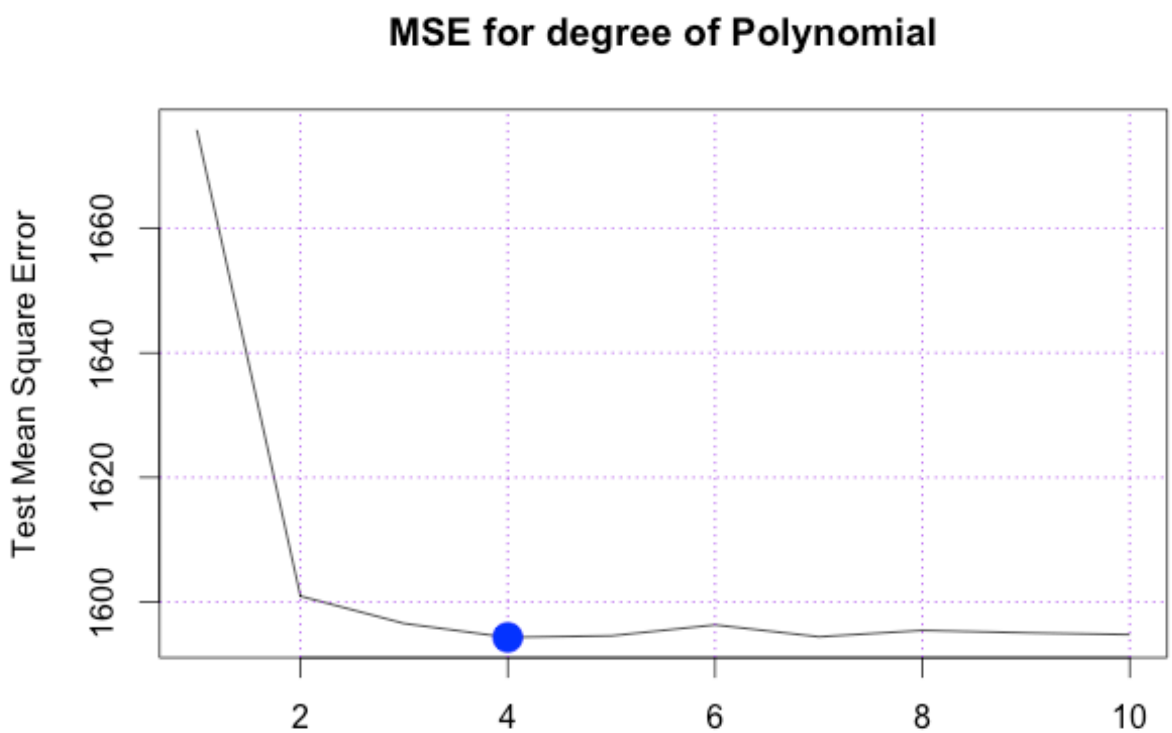

Figure 1

```
> anova(fit1, fit2, fit3, fit4, fit5,fit6,fit7,fit8,fit9,fit10)
Analysis of Variance Table

Model  1: wage ~ age
Model  2: wage ~ poly(age, 2)
Model  3: wage ~ poly(age, 3)
```

```
Model  4: wage ~ poly(age, 4)
Model  5: wage ~ poly(age, 5)
Model  6: wage ~ poly(age, 6)
Model  7: wage ~ poly(age, 7)
Model  8: wage ~ poly(age, 8)
Model  9: wage ~ poly(age, 9)
Model 10: wage ~ poly(age, 10)
   Res.Df     RSS Df Sum of Sq        F    Pr(>F)
1    2998 5022216
2    2997 4793430  1    228786 143.7638 < 2.2e-16 ***
3    2996 4777674  1     15756   9.9005  0.001669 **
4    2995 4771604  1      6070   3.8143  0.050909 .
5    2994 4770322  1      1283   0.8059  0.369398
6    2993 4766389  1      3932   2.4709  0.116074
7    2992 4763834  1      2555   1.6057  0.205199
8    2991 4763707  1       127   0.0796  0.777865
9    2990 4756703  1      7004   4.4014  0.035994 *
10   2989 4756701  1         3   0.0017  0.967529
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

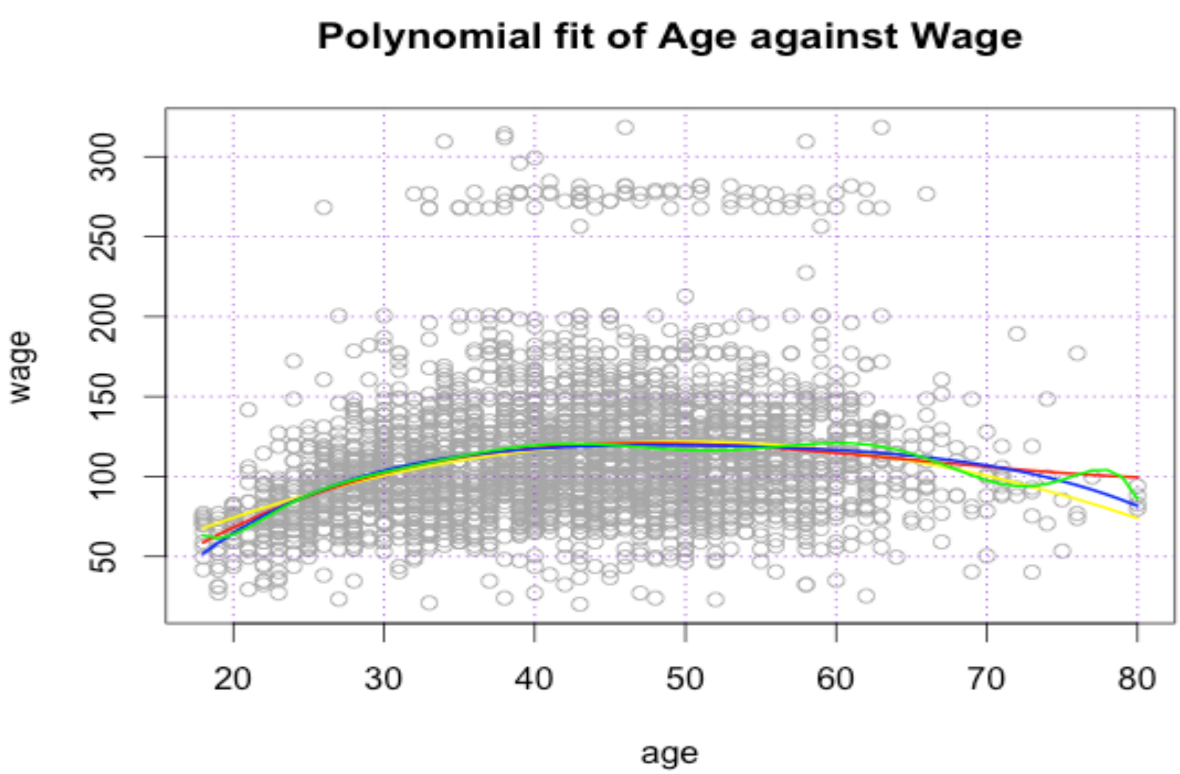

Figure 2

According to the 10 fold cross validation, a degree of 4 had the lowest MSE. The results of hypothesis testing using ANOVA yielded models with degrees 2,3 and 9 as being statistically significant. A plot was fitted using polynomials of degree 2,3,4 and 9.

(b) **Fit a step function to predict wage using age, and perform cross validation to choose the optimal number of cuts. Make a plot of the fit obtained.**

A 10-fold cross validation was run to figure out the optimal number of cuts.

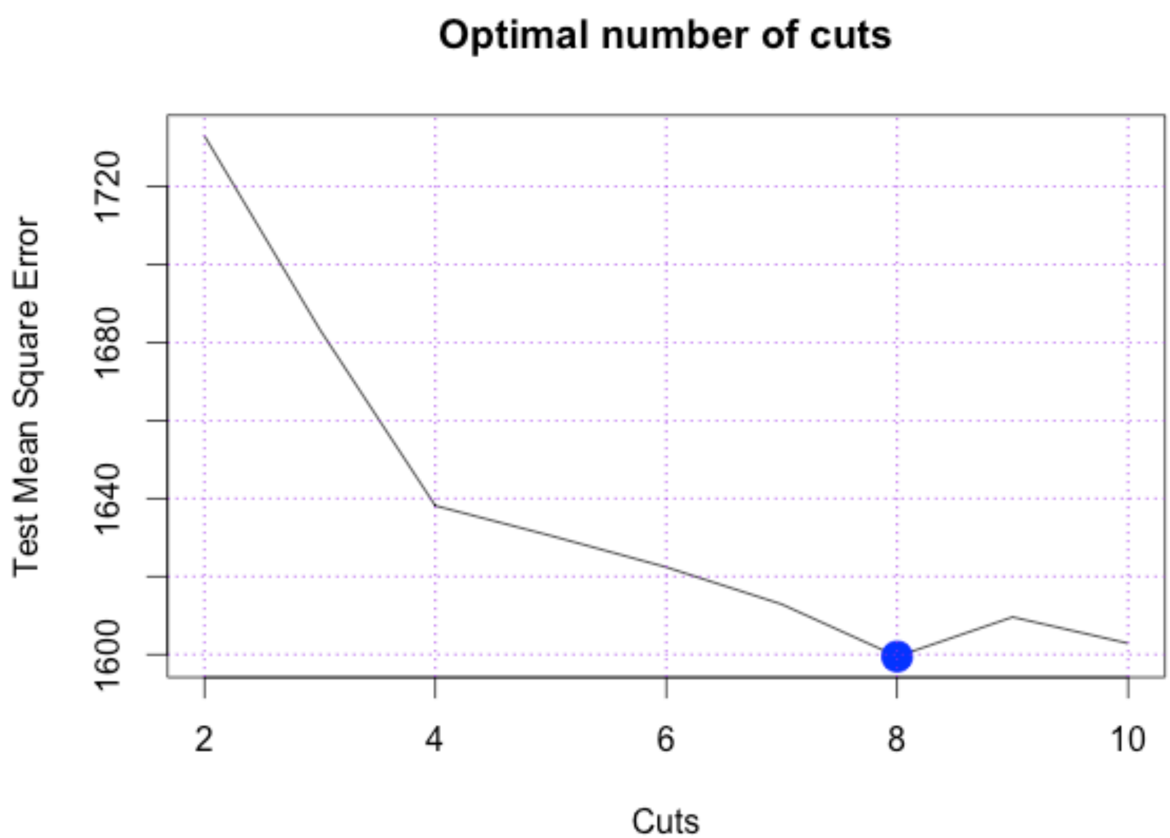

Figure 3

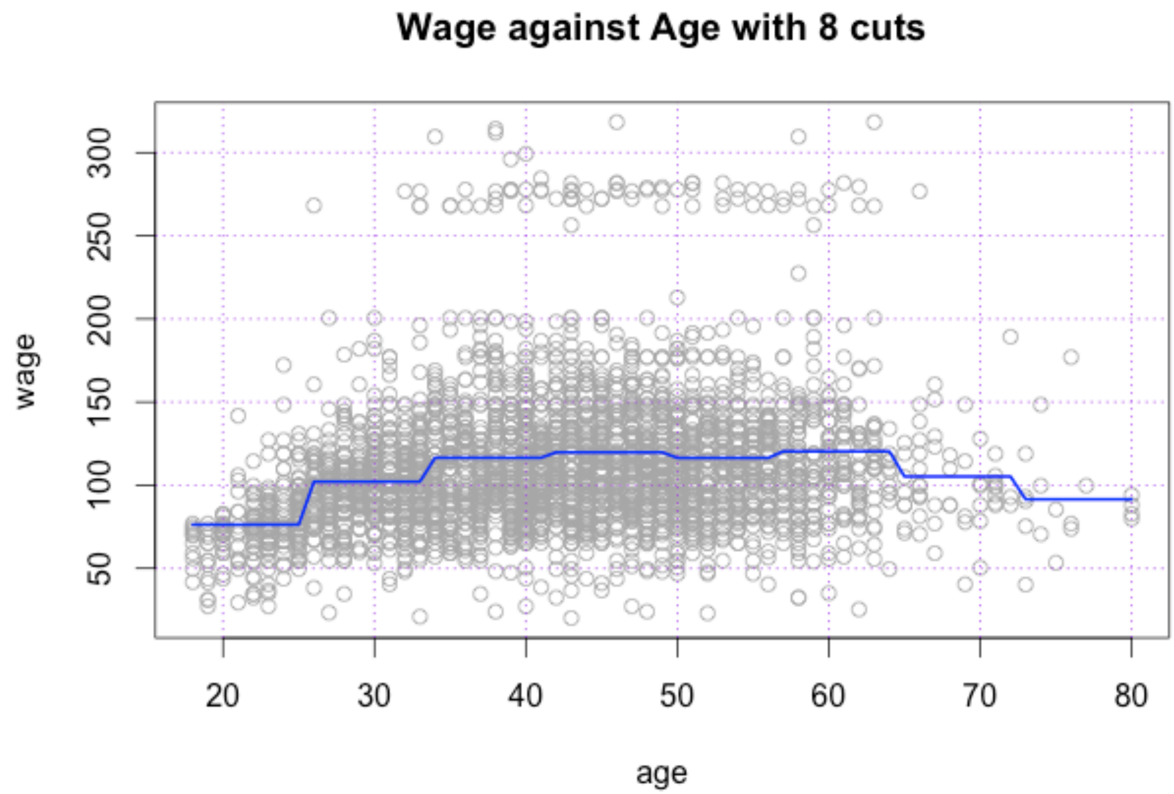

Figure 4

9. This question uses the variables dis (the weighted mean of distances to five Boston employment centers) and nox (nitrogen oxides concentration in parts per 10 million) from the Boston data. We will treat dis as the predictor and nox as the response.

(a) Use the poly() function to fit a cubic polynomial regression to predict nox using dis. Report the regression output, and plot the resulting data and polynomial fits.

```
> summary(fit)

Call:
lm(formula = nox ~ poly(dis, 3), data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max
-0.121130 -0.040619 -0.009738  0.023385  0.194904

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)    0.554695   0.002759 201.021  < 2e-16 ***
poly(dis, 3)1 -2.003096   0.062071 -32.271  < 2e-16 ***
poly(dis, 3)2  0.856330   0.062071  13.796  < 2e-16 ***
poly(dis, 3)3 -0.318049   0.062071  -5.124 4.27e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06207 on 502 degrees of freedom
Multiple R-squared:  0.7148,  Adjusted R-squared:  0.7131
F-statistic: 419.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

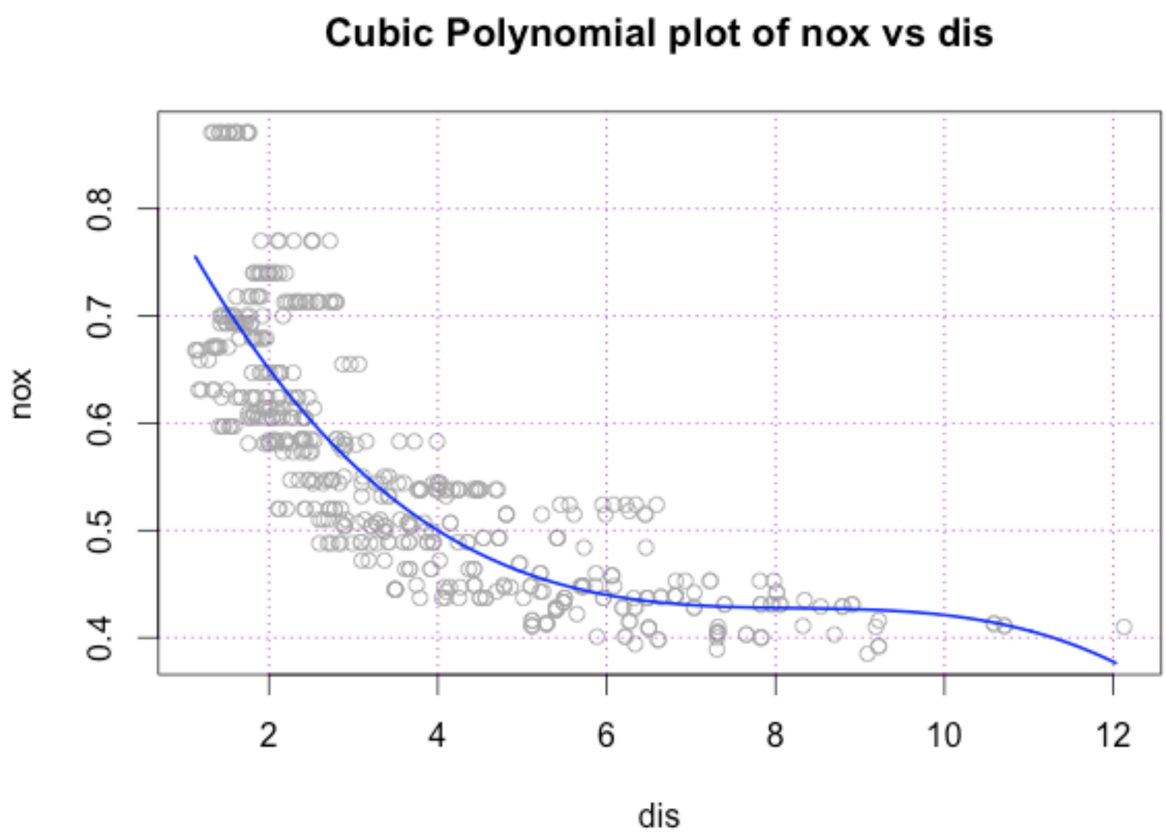

Figure 5

From the results table, we can conclude that all polynomial terms are significant.

(b) Plot the polynomial fits for a range of different polynomial degrees (say, from 1 to 10), and report the associated residual sum of squares.

It can be seen that the lowest residual sum of squares happens at a degree of 15.

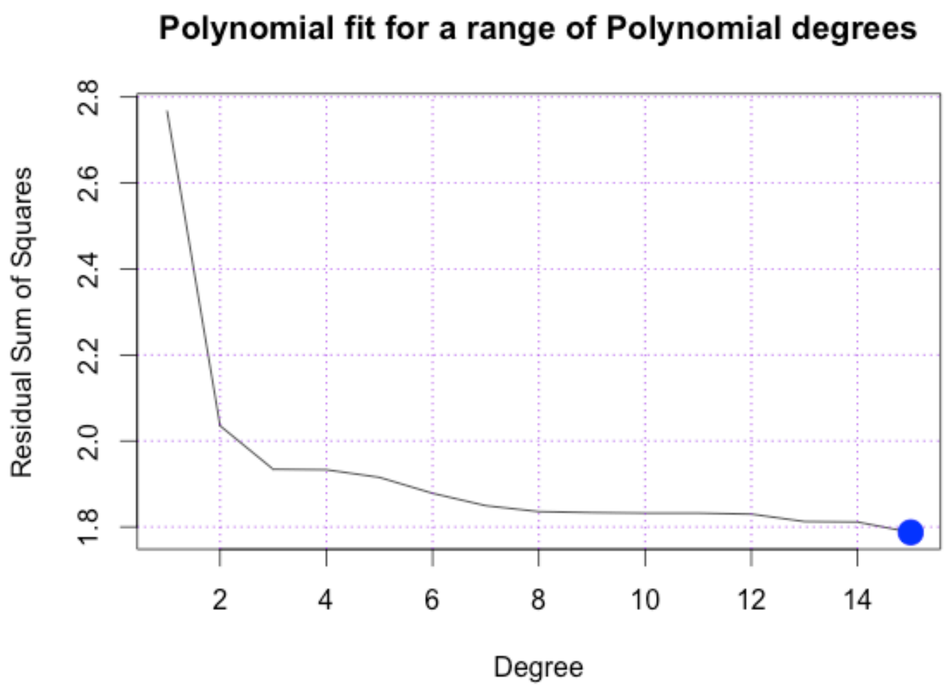

Figure 6

(c) Perform cross-validation or another approach to select the optimal degree for the polynomial, and explain your results.

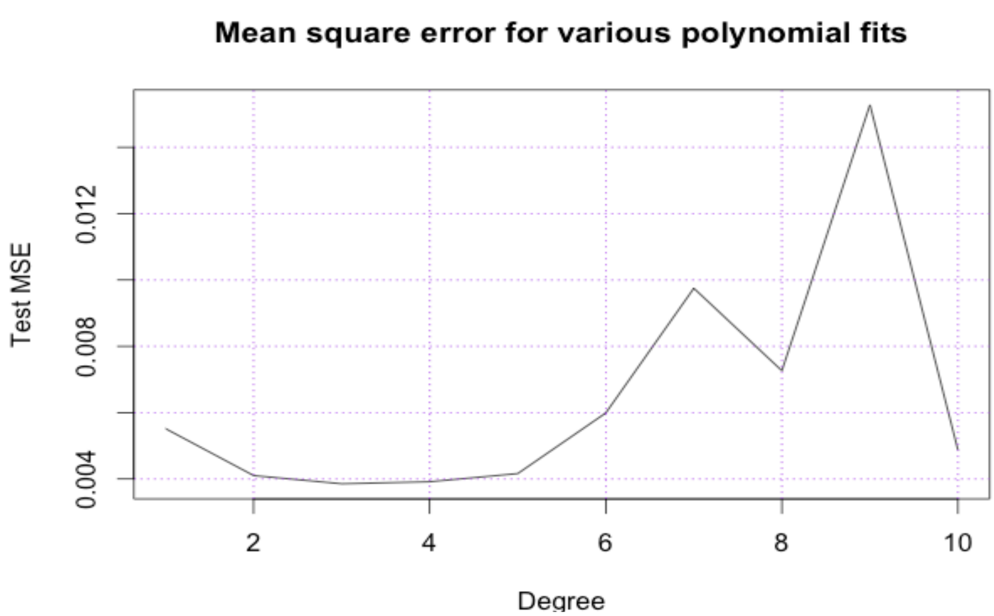

Figure 7

The plot seems to indicate that the lowest mean square error occurs somewhere between the cubic and quartic polynomial model. Higher polynomial models may have less MSE but could result in overfitting.

(d) Use the bs() function to fit a regression spline to predict nox using dis. Report the output for the fit using four degrees of freedom. How did you choose the knots? Plot the resulting fit.

```
Call:
lm(formula = nox ~ bs(dis, df = 4), data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max
-0.124622 -0.039259 -0.008514  0.020850  0.193891

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.73447    0.01460  50.306  < 2e-16 ***
bs(dis, df = 4)1 -0.05810    0.02186  -2.658  0.00812 **
bs(dis, df = 4)2 -0.46356    0.02366 -19.596  < 2e-16 ***
bs(dis, df = 4)3 -0.19979    0.04311  -4.634 4.58e-06 ***
bs(dis, df = 4)4 -0.38881    0.04551  -8.544  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06195 on 501 degrees of freedom
Multiple R-squared:  0.7164,  Adjusted R-squared:  0.7142
F-statistic: 316.5 on 4 and 501 DF,  p-value: < 2.2e-16
```

As asked in the question, I chose 4 degrees of freedom and not knots. In the bs() function, if both knots and df are specified then knots are ignored. Therefore, there was no point in specifying both.

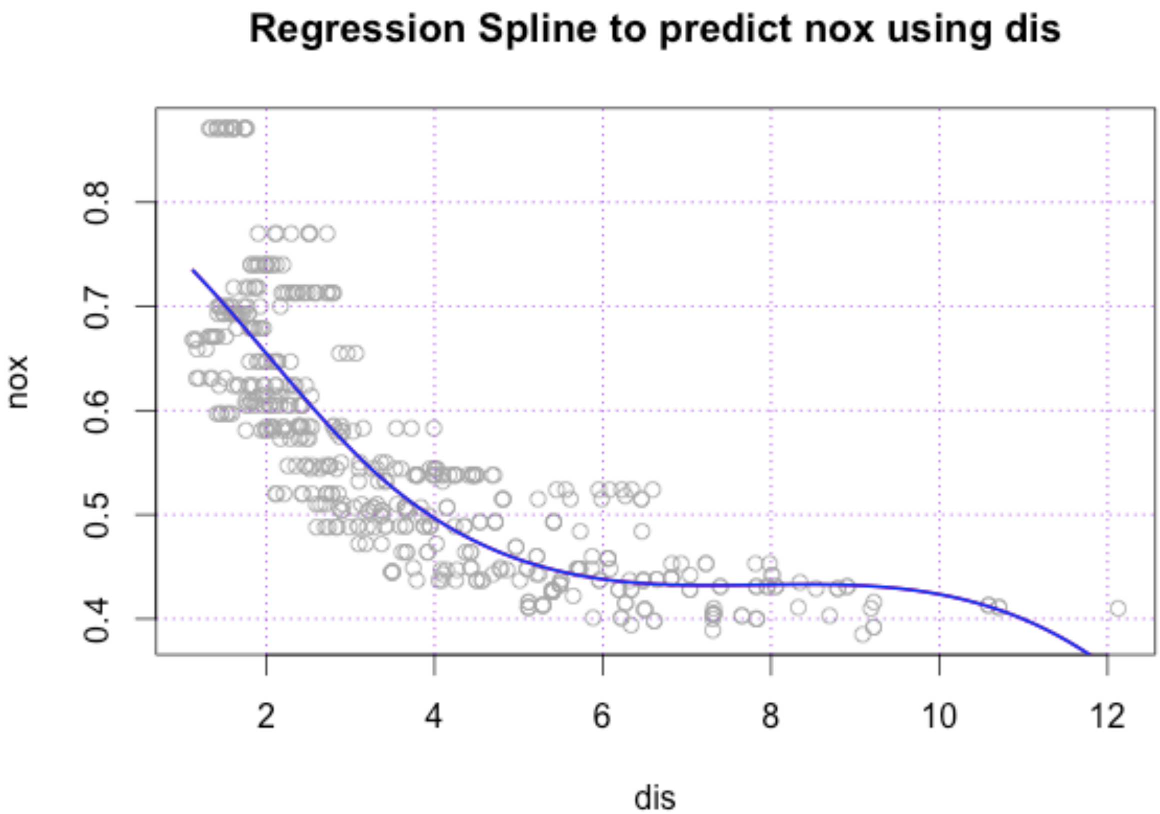

Figure 8

(e) **Now fit a regression spline for a range of degrees of freedom, and plot the resulting fits and report the resulting RSS. Describe the results obtained.**

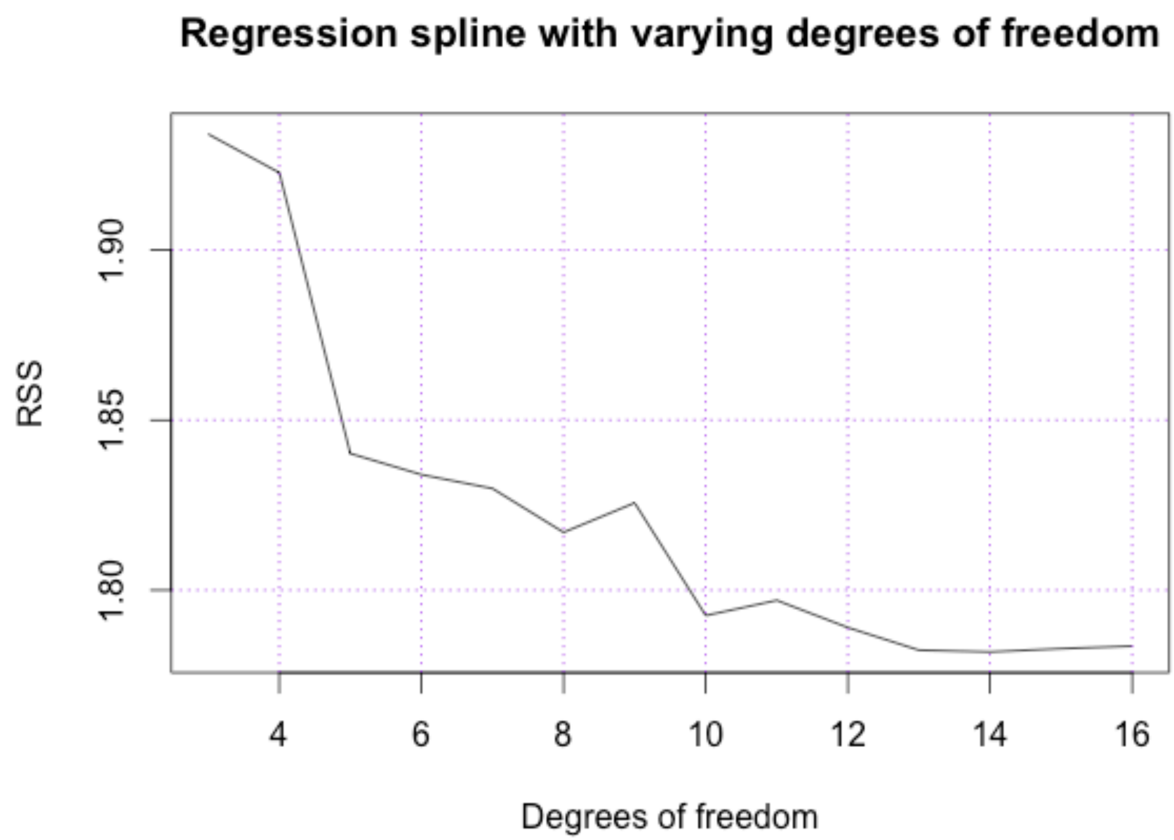

Figure 9

(f) Perform cross-validation or another approach in order to select the best degrees of freedom for a regression spline on this data. Describe your results.

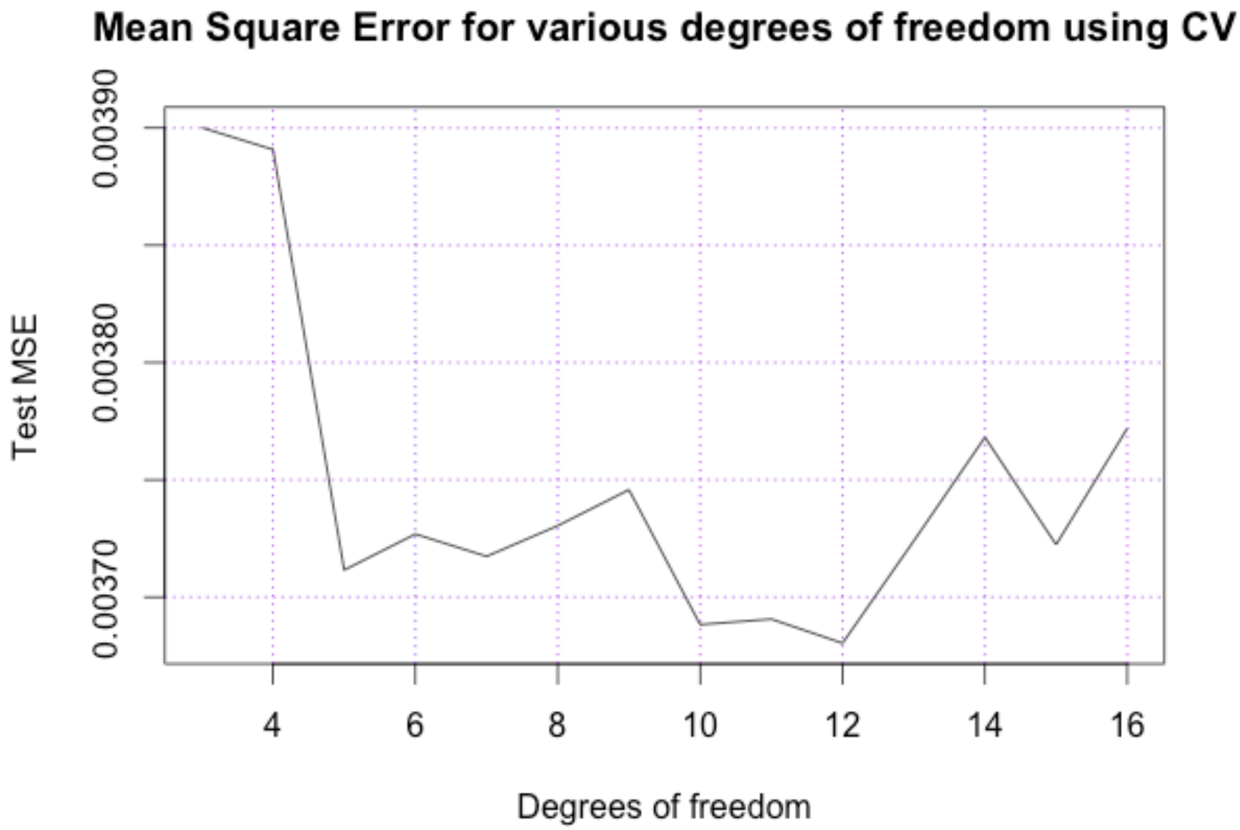

Figure 10

The regression spline model gives us the lowest MSE at 12 degrees of freedom.

10. 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.

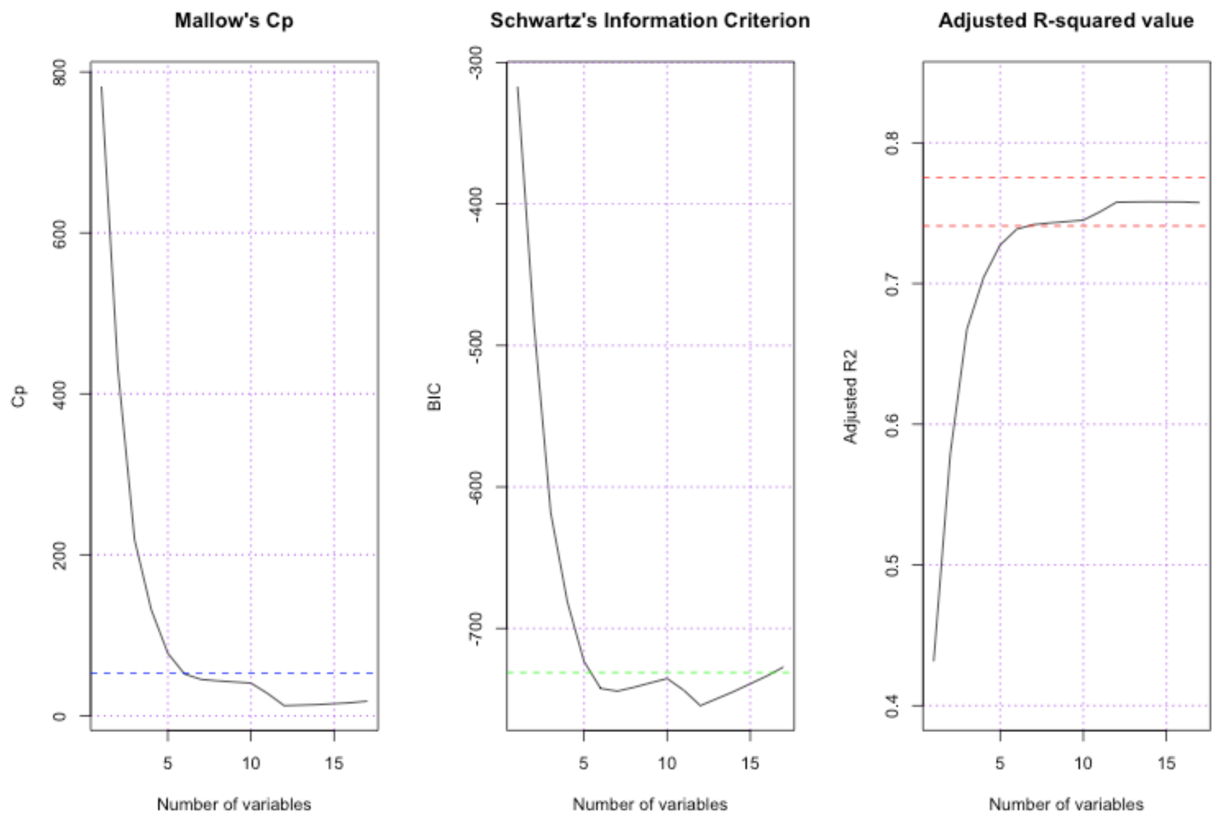

Figure 11

Plotting the results of the forward selection method using regsubsets function in R. Mallow's Cp and Schwartz's BIC seems to be the lowest at 11 or 12 variables while the adjusted R-squared value is the highest at those number of variables.
However, what is important here is the cutoff point. It seems that 6 is the minimum number of variables for which the standard deviation is within 0.2 of the optimum level.

Running a coefficient function the 6 most important variables in order to predict out of state tuition are:

```
> names(coeffs)
[1] "(Intercept)" "PrivateYes"  "Room.Board"  "PhD"        "perc.alumni" "Expend"      "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. Plot the results, and explain your findings.

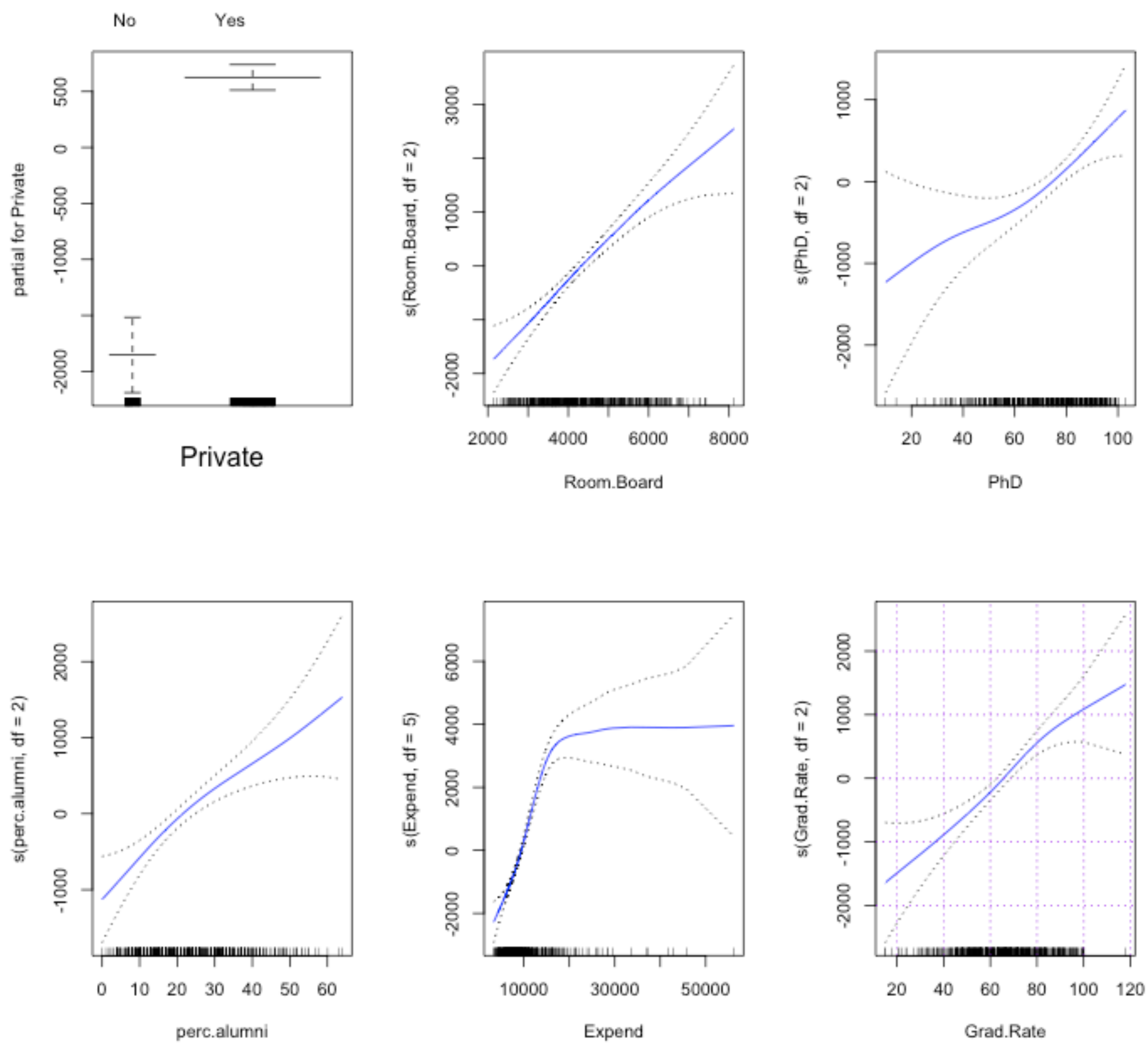

Figure 12

For figure 12, I used 6 variables to predict out-of-state tuition using the College data set. A private university had a way higher out-of-state tuition than a public. Room and board seemed to display a linear relationship with out-of-state tuition. The third plot shows that the higher the percentage of phD faculty holders, the higher the out-of-state tuition. The same argument applies to the percentage of alumni who donate to the school and the graduation rate for the college students. Instructional expenditure has an interesting relationship with out-of-state tuition – out-of-state tuition appears to stay constant for higher levels of instructional expenditure. All of these findings however, are quite intuitive.

It must be mentioned that when conducting effects for each of the predictor variables, other predictor variables were held fixed.

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

```
  error        rss
1 3318035      0.8051305
```

A R-squared value of 80.5% was obtained using the 6 predictors. This means that close to 80.5% of the variability in the response variable can be explained by these 6 predictors.

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

```
> summary(gam.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
-7288.05 -1089.27    23.97  1265.61  8251.17

(Dispersion Parameter for gaussian family taken to be 3550475)

   Null Deviance: 9238455544 on 581 degrees of freedom
Residual Deviance: 2013119109 on 566.9999 degrees of freedom
AIC: 10446.52

Number of Local Scoring Iterations: 2

Anova for Parametric Effects
                        Df     Sum Sq    Mean Sq F value    Pr(>F)
Private                  1 2309362822 2309362822 650.438 < 2.2e-16 ***
s(Room.Board, df = 2)    1 1886269737 1886269737 531.273 < 2.2e-16 ***
s(PhD, df = 2)           1  667519561  667519561 188.008 < 2.2e-16 ***
s(perc.alumni, df = 2)   1  380458329  380458329 107.157 < 2.2e-16 ***
s(Expend, df = 5)        1  676556637  676556637 190.554 < 2.2e-16 ***
s(Grad.Rate, df = 2)     1  131943136  131943136  37.162 2.011e-09 ***
Residuals              567 2013119109    3550475
---
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  0.8618 0.3536
s(PhD, df = 2)              1  2.3598 0.1251
s(perc.alumni, df = 2)      1  1.7232 0.1898
s(Expend, df = 5)           4 23.6837 <2e-16 ***
s(Grad.Rate, df = 2)        1  2.6896 0.1016
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

According to the ANOVA model for nonparametric effects, it does look like there is a strong non-linear relationship between out-of-state tuition and expenditure. This is clear evidence that a non-linear term is required for expenditure. Larger p-values such as that for room and board and percentage of alumni donating reinforce the idea that a linear function is adequate. This was also seen in the gam plots on the previous page.