Homework 9

Prakash Paudyal

Please do the following problems from the text book ISLR. (use set.seed(702) to replicate your results).

1. Question 7.9.6 pg 299

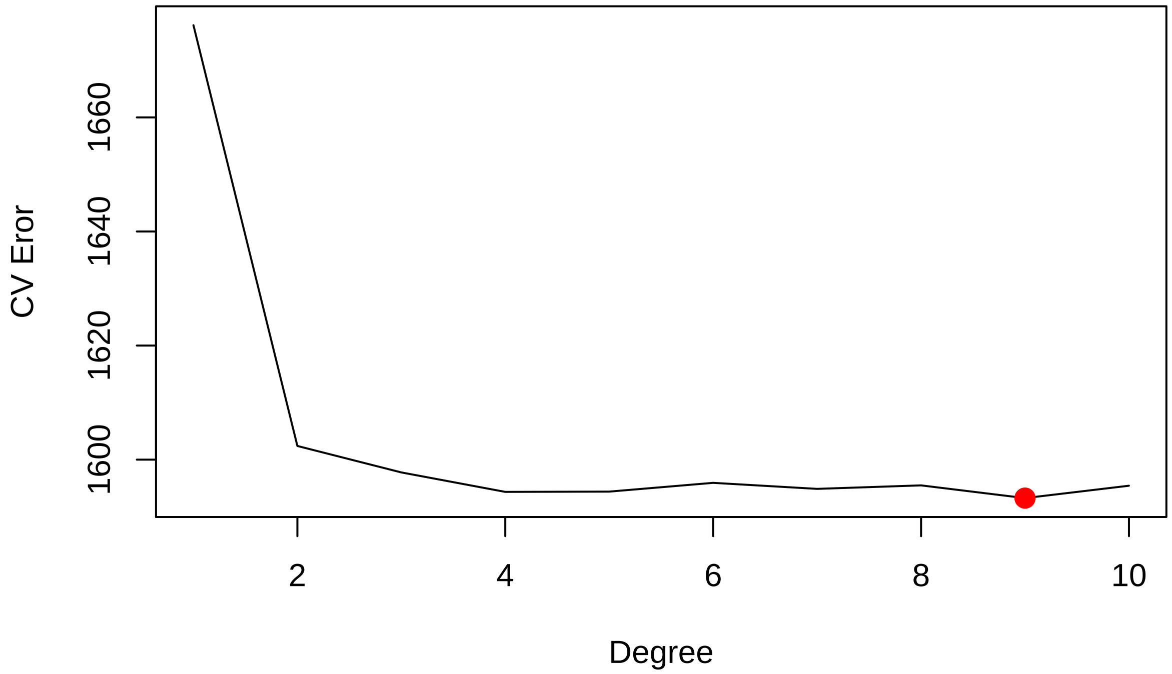
(6.) In this exercise, you will further analyze the Wage data set considered throughout this chapter.

```
## 'data.frame':    3000 obs. of  11 variables:
##  $ year      : int  2006 2004 2003 2003 2005 2008 2009 2008 2006 2004 ...
##  $ age       : int  18 24 45 43 50 54 44 30 41 52 ...
##  $ maritl    : Factor w/ 5 levels "1. Never Married",..: 1 1 2 2 4 2 2 1 1 2 ...
##  $ race      : Factor w/ 4 levels "1. White","2. Black",..: 1 1 1 3 1 1 4 3 2 1 ...
##  $ education : Factor w/ 5 levels "1. < HS Grad",..: 1 4 3 4 2 4 3 3 3 2 ...
##  $ region    : Factor w/ 9 levels "1. New England",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ jobclass  : Factor w/ 2 levels "1. Industrial",..: 1 2 1 2 2 2 1 2 2 2 ...
##  $ health    : Factor w/ 2 levels "1. <=Good","2. >=Very Good": 1 2 1 2 1 2 2 1 2 2 ...
##  $ health_ins: Factor w/ 2 levels "1. Yes","2. No": 2 2 1 1 1 1 1 1 1 1 ...
##  $ logwage   : num  4.32 4.26 4.88 5.04 4.32 ...
##  $ wage      : num  75 70.5 131 154.7 75 ...
```

(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.

```
##  [1] 1676.163 1602.396 1597.762 1594.348 1594.392 1595.934 1594.874
##  [8] 1595.490 1593.259 1595.426
```



Degree,d=9 is optimal degree for polyminal, choosen by cross validation method.

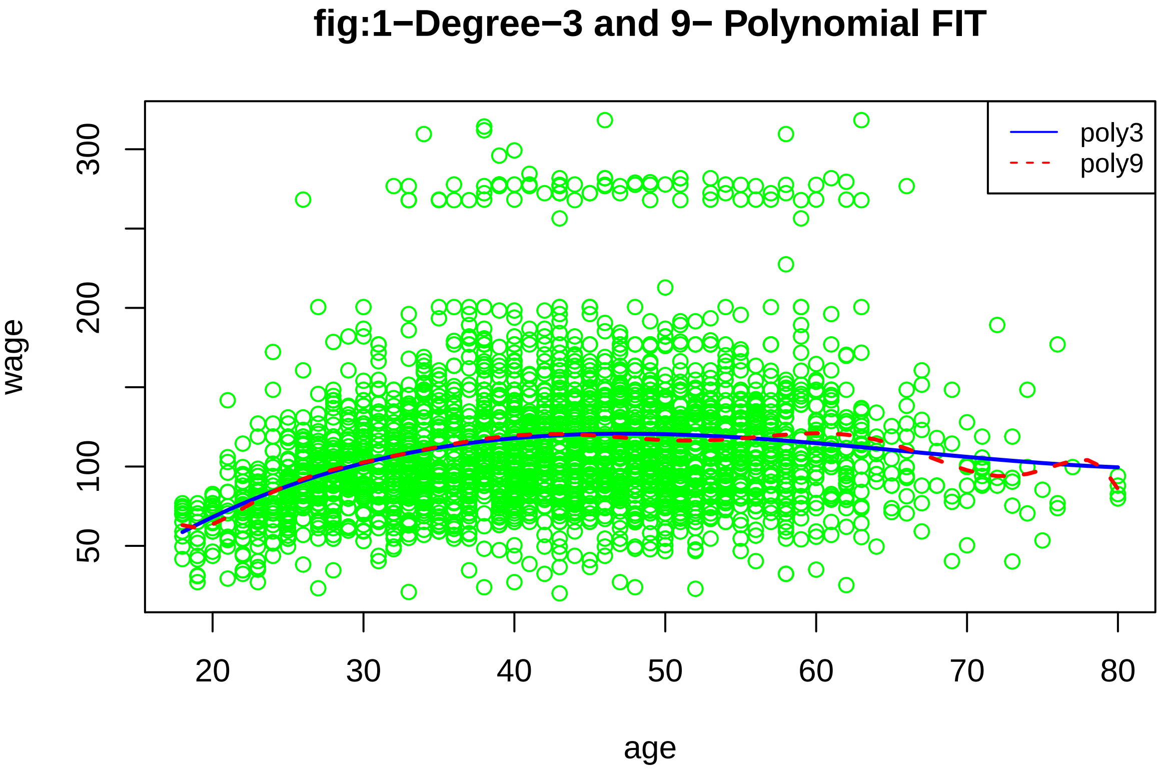
we use the anova() in order to test null hypothesis that a model M_1 is sufficient to explain the data against the altenative hypothesis that a more complex model M_2 is required.

We fit the models upto 10 polynomial degrees of ages.

```
## 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
```

Anova,P-values, shows that 2nd , 3rd and 9th degree of polynomial degree regression provide a reasonable fit to the data. Among them 2nd order polynomial provide best fit to the data. Where as cross validation method provide best fit at polynominal degree of 9.

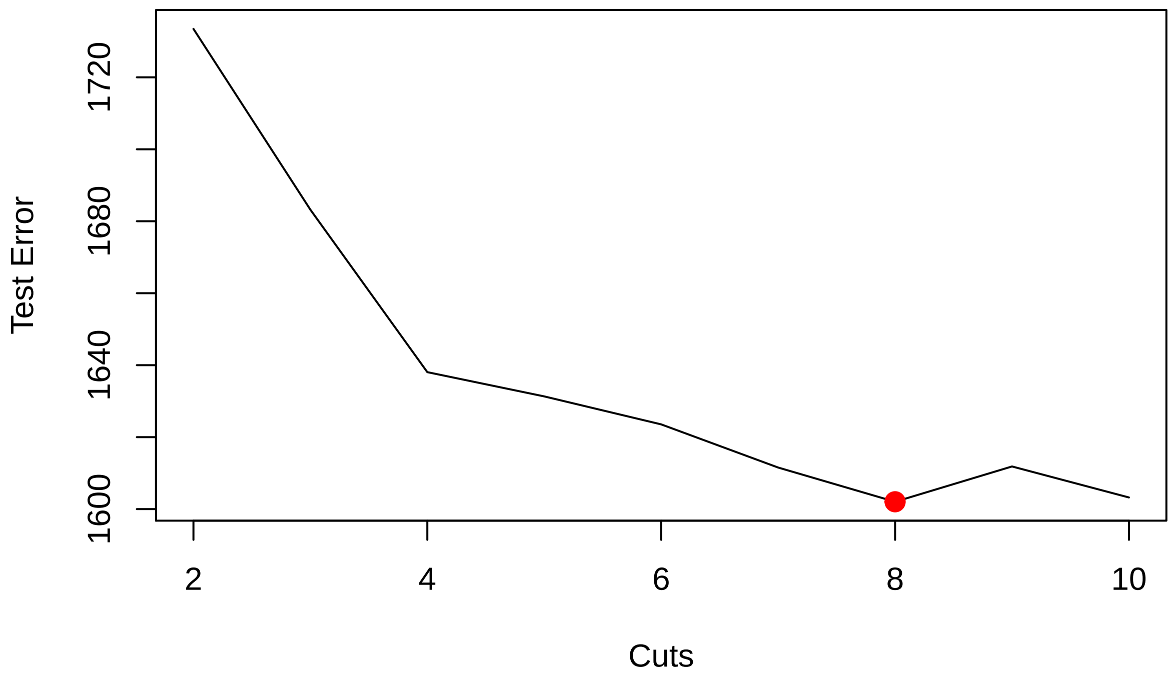
Plotting polynomial fit



(b)

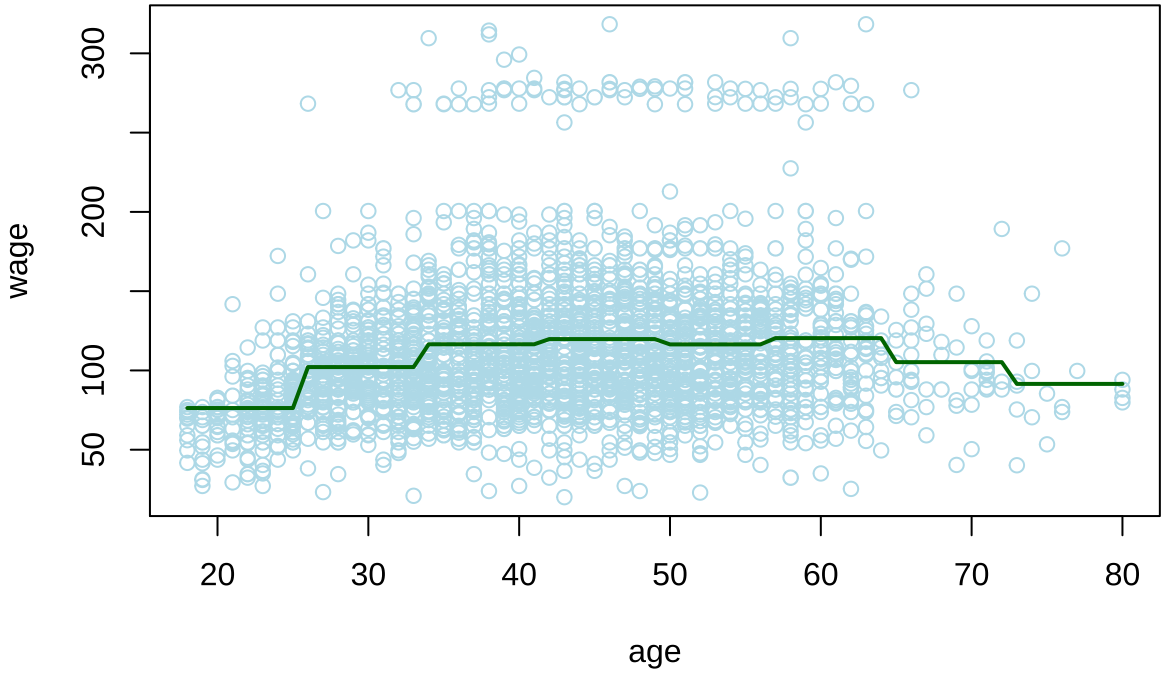
Fit a step function to predict wage using age, and perform crossvalidation to choose the optimal number of cuts. Make a plot of the fit obtained.

Fig:2,10−Fold Cross Validation to find optimal cuts



The cross validation shows that test error is minimum for optimal cuts k=8 cuts. Train the data with cuts, k=8

Fig:3, Scatter plot of data with fitted line at cuts=8



2. Question 7.9.9 pg 299

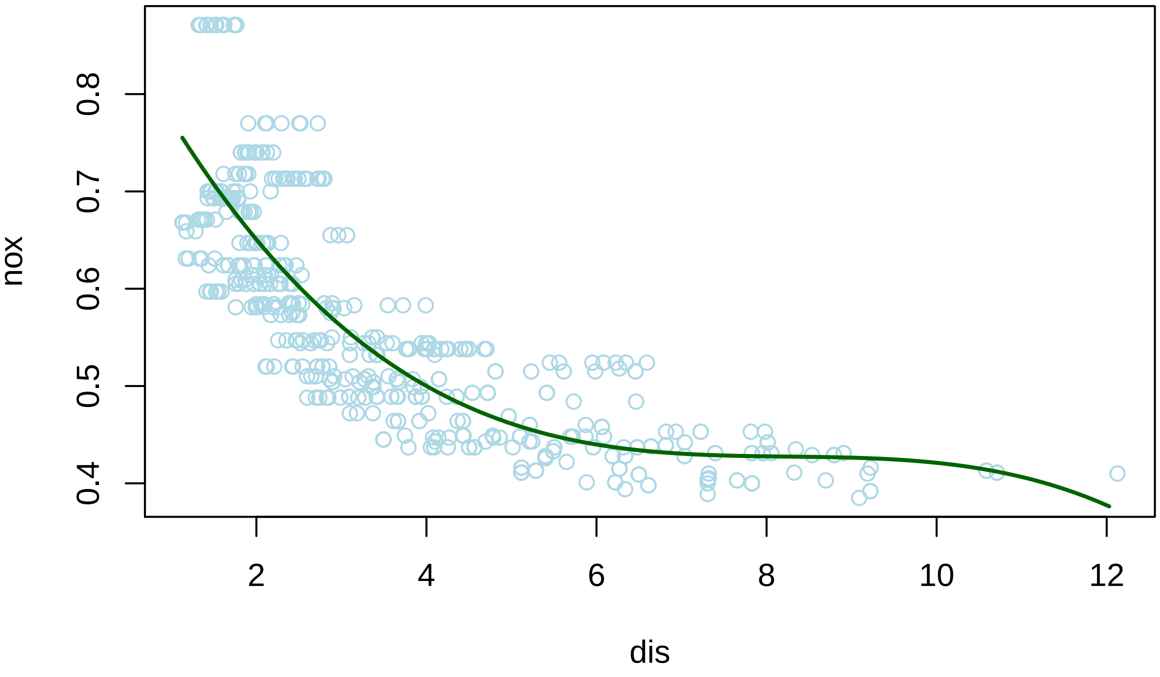
** (9.) This question uses the variables dis (the weighted mean of distances to five Boston employment centers) and nox (nitrogen oxides concentration in parts per 10 million) from the Boston data. We will treat dis as the predictor and nox as the response.**

(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.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.5546951	0.0027594	201.020893	0e+00
poly(dis, 3)1	-2.0030959	0.0620709	-32.271071	0e+00
poly(dis, 3)2	0.8563300	0.0620709	13.795987	0e+00
poly(dis, 3)3	-0.3180490	0.0620709	-5.123959	4e-07

Plot nox vs dis and polynominal fit

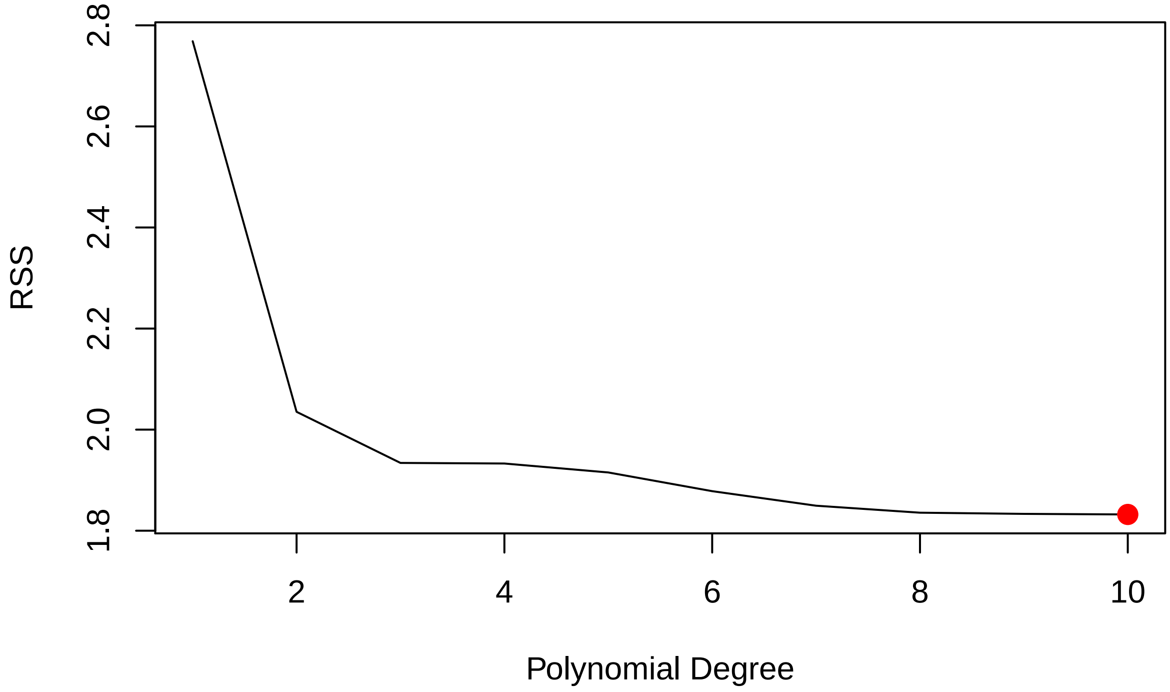


Nox has decrease trend with dis.

(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.

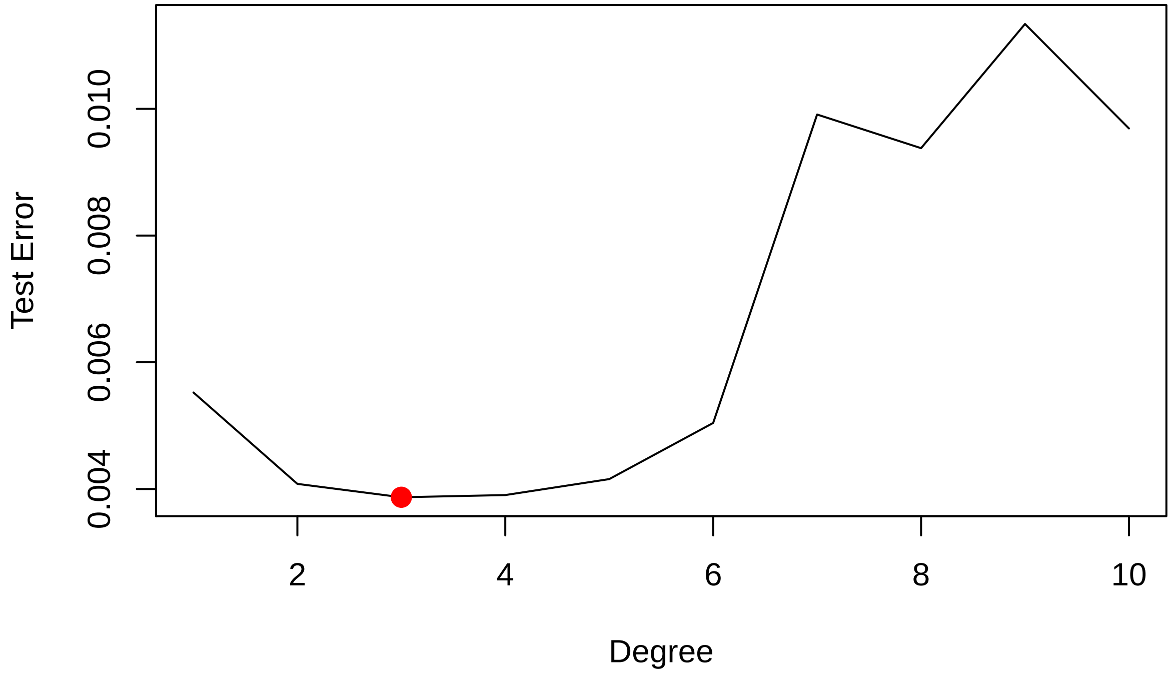
```
##  [1] 2.768563 2.035262 1.934107 1.932981 1.915290 1.878257 1.849484
##  [8] 1.835630 1.833331 1.832171
```



Plot shows that RSS decreases with degree of polynomial and it is minimum at polynomil of degree 10.

(c)

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



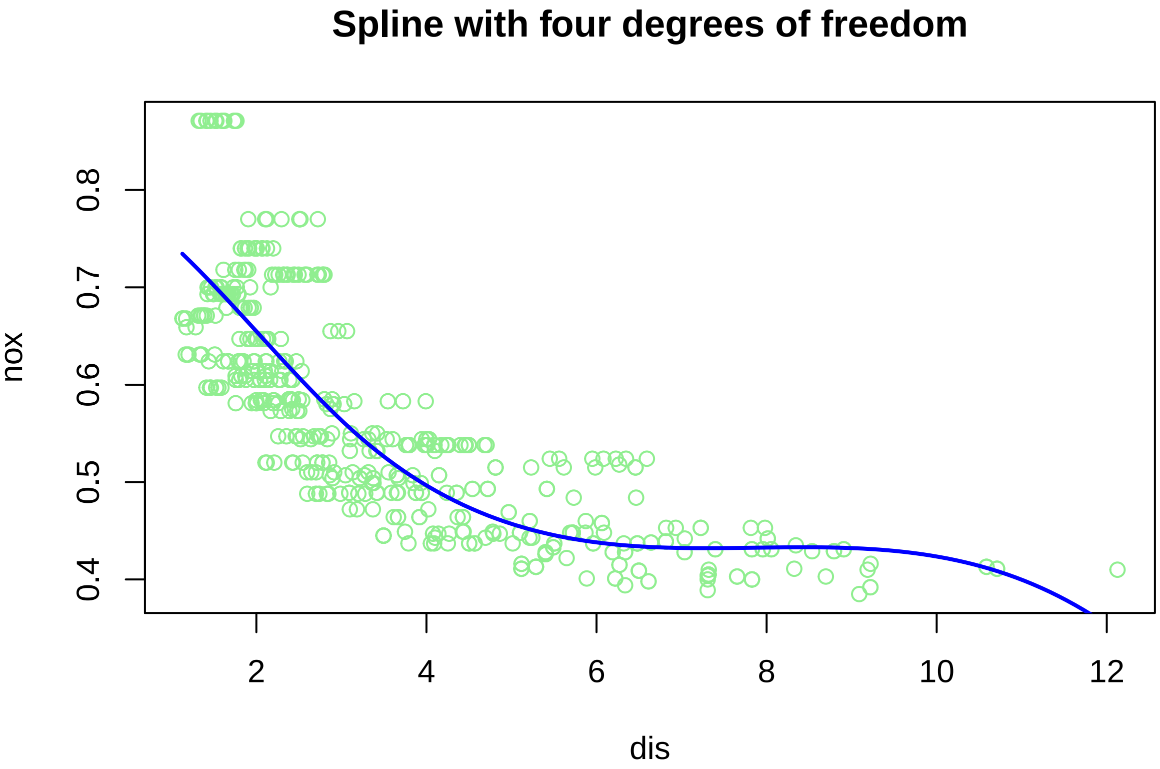
```
## [1] 3
```

Polymonial degree of 3 has minimum error rate in 10 fold cross validation.Above the 3rd degree polymonial test error started to increase and slowly came down at 10th degree.

(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.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.7344739	0.0146003	50.305518	0.0000000
bs(dis, df = 4)1	-0.0580976	0.0218591	-2.657825	0.0081160
bs(dis, df = 4)2	-0.4635635	0.0236559	-19.596141	0.0000000
bs(dis, df = 4)3	-0.1997882	0.0431140	-4.633950	0.0000046
bs(dis, df = 4)4	-0.3888095	0.0455070	-8.543944	0.0000000

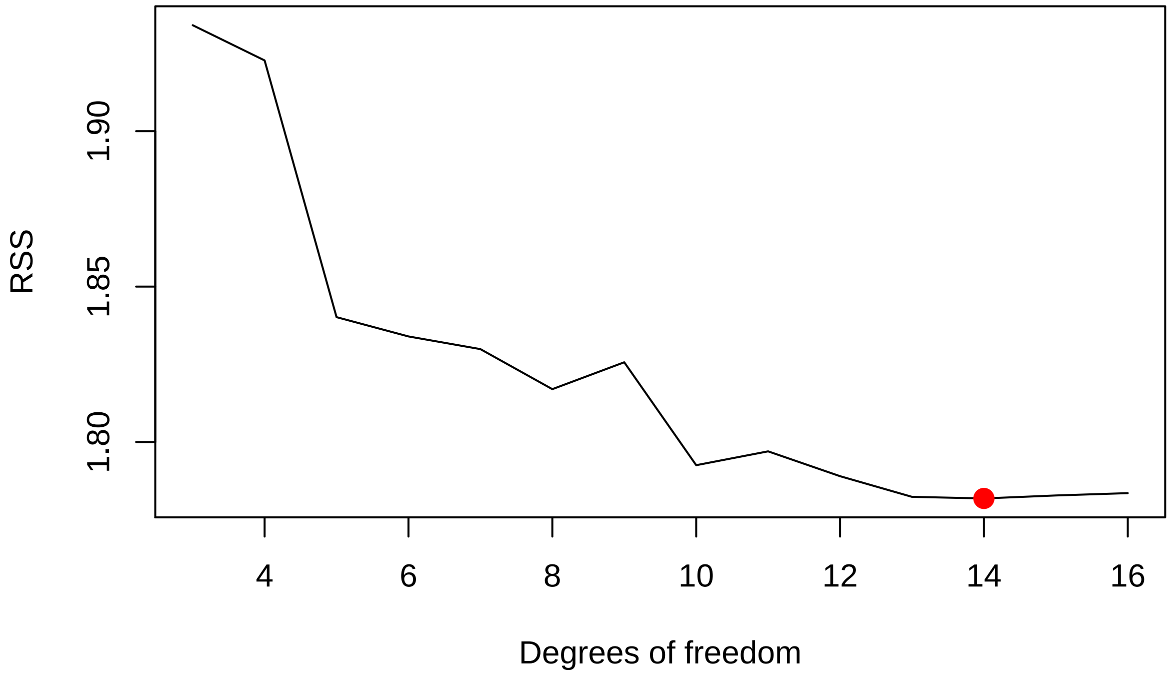


Summary shows that all the spline fit are significant. Plot of fitted line shows that model fited data well. We useed the df option to produce a spline with knots at uniform quantiles of the data

(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.

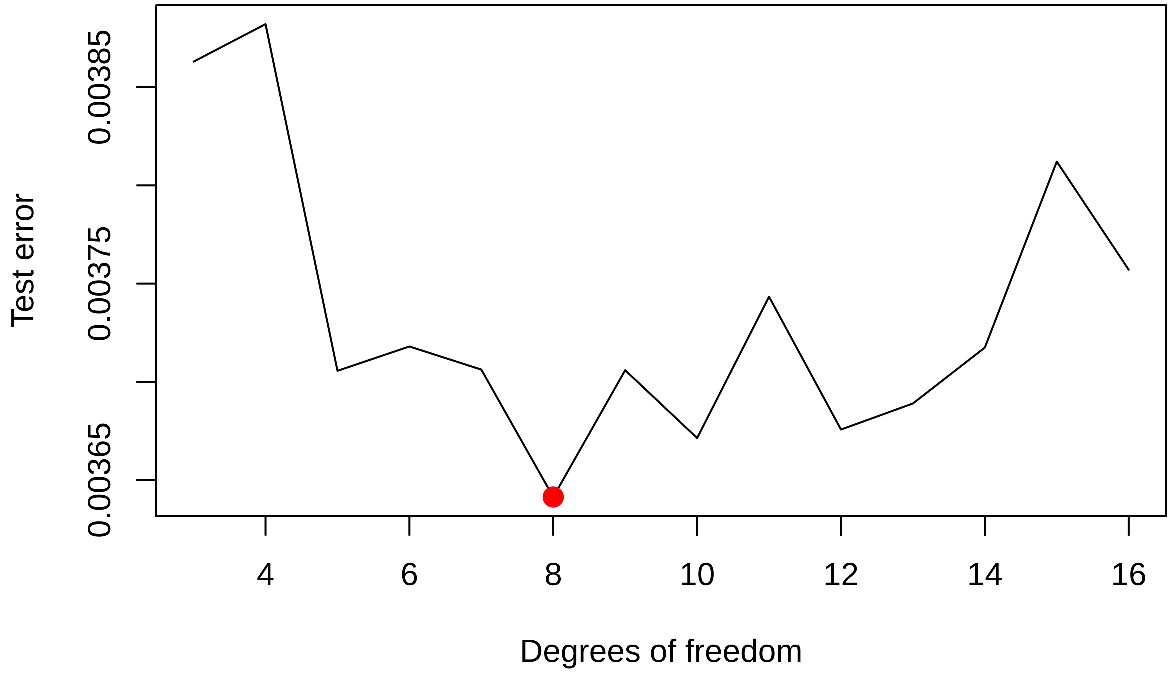
```
##  [1] 1.934107 1.922775 1.840173 1.833966 1.829884 1.816995 1.825653
##  [8] 1.792535 1.796992 1.788999 1.782350 1.781838 1.782798 1.783546
```



The regression splines upto 16th order were fitted and its corresponding RSS were computed. The results showed 14th order with least RSS.

(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.



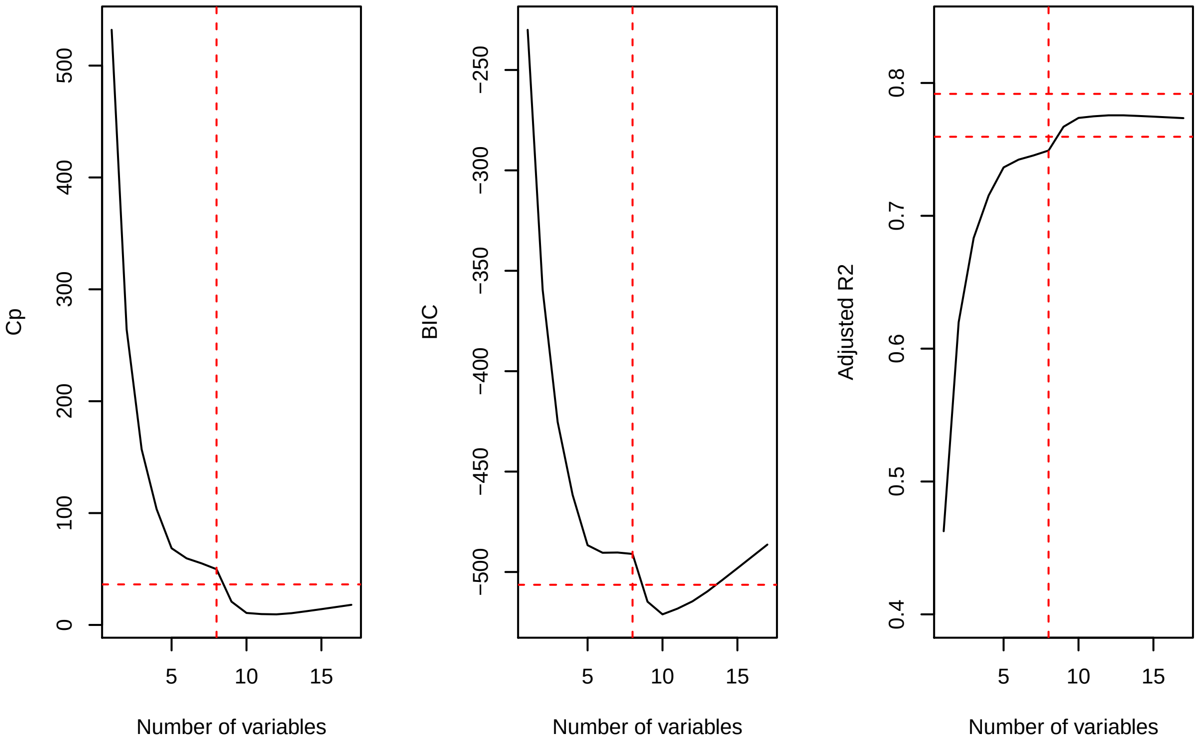
The 10 fold cross validation methods suggest polynomial regression with 8th degree of freedom has least test error.

3. Question 7.9.10 pg 300

(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.



Plot sugessted that subset of 8 variables is with in 0.2 standard deviation of of optimal value of cp,bic and adjuster R-squared. Hence we chhoose subset of 8 variables to fit the model Names of variable of selected subset.

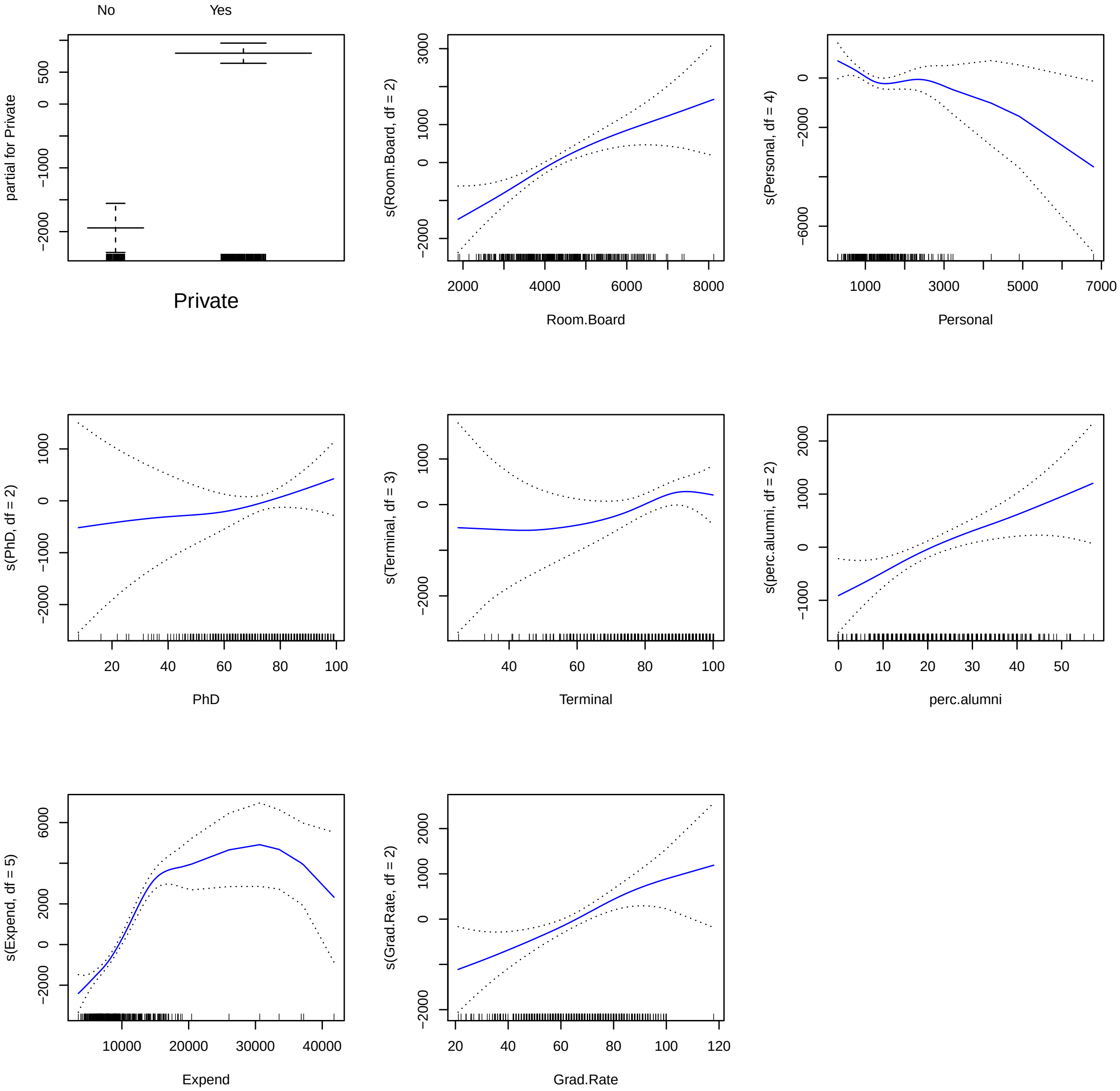
```
## [1] "(Intercept)" "PrivateYes"  "Room.Board"  "Personal"    "PhD"
## [6] "Terminal"    "perc.alumni" "Expend"      "Grad.Rate"
```

(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

```
## Loading required package: foreach
```

```
## Loaded gam 1.15
```



GAM model used subset of 8 predictors selected by permormimg forward stepwise selection and with specified degree of freedom.Since privte is qualitative variable, we did not mention any degree of freedom for it in s().In these plots, the function of Room board, PhD, Terminal, Percent Alumni, Grad Rate looks to have positive slope.While personal variable seems to have decreasing trend and expend variable seems to have increasing and then decreasing trend

(c)

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

```
## [1] 3716430
```

```
## [1] 0.7660947
```

MSE=3716430 R^2 error for test data set is 0.7660947 for GAM model with 8 predictors.

(d)

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

```
##
## Call: gam(formula = Outstate ~ Private + s(Room.Board, df = 2) + s(Personal,
##     df = 4) + s(PhD, df = 2) + s(Terminal, df = 3) + s(perc.alumni,
##     df = 2) + s(Expend, df = 5) + s(Grad.Rate, df = 2), data = College.train)
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -7090.42 -1177.60    56.49  1262.42  4434.01
##
## (Dispersion Parameter for gaussian family taken to be 3411910)
##
##     Null Deviance: 6376456790 on 387 degrees of freedom
## Residual Deviance: 1248759398 on 366.0001 degrees of freedom
## AIC: 6961.048
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##                         Df     Sum Sq    Mean Sq  F value    Pr(>F)
## Private                  1 1891238090 1891238090 554.3048 < 2.2e-16 ***
## s(Room.Board, df = 2)    1 1156607828 1156607828 338.9913 < 2.2e-16 ***
## s(Personal, df = 4)      1   42855502   42855502  12.5606 0.0004450 ***
## s(PhD, df = 2)           1  523556000  523556000 153.4495 < 2.2e-16 ***
## s(Terminal, df = 3)      1   23579040   23579040   6.9108 0.0089287 **
## s(perc.alumni, df = 2)   1  182235965  182235965  53.4117 1.712e-12 ***
## s(Expend, df = 5)        1  506343542  506343542 148.4047 < 2.2e-16 ***
## s(Grad.Rate, df = 2)     1   49448460   49448460  14.4929 0.0001649 ***
## Residuals              366 1248759398    3411910
## ---
## 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  1.5291   0.21705
## s(Personal, df = 4)          3  2.4019   0.06741 .
## s(PhD, df = 2)               1  0.8696   0.35169
## s(Terminal, df = 3)          2  1.4693   0.23145
## s(perc.alumni, df = 2)       1  0.6336   0.42657
## s(Expend, df = 5)            4 19.0495 3.031e-14 ***
## s(Grad.Rate, df = 2)         1  1.0132   0.31480
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The nonparametric component gam.fit shows that there is strong non linear relationship of Expend variables with response variable. And personal variable also has some non-linear repationship with response but not highly significant.

Bonus (1 pt) Question 7.9.11 pg 300 (Not necessary, can do if you want an extra point!)

(11.) In Section 7.7, it was mentioned that GAMs are generally fit using a backfitting approach. The idea behind backfitting is actually quite simple. We will now explore backfitting in the context of multiple linear regression.Suppose that we would like to perform multiple linear regression, but we do not have software to do so. Instead, we only have software to perform simple linear regression. Therefore, we take the following iterative approach: we repeatedly hold all but one coefficient estimate fixed at its current value, and update only that coefficientes timate using a simple linear regression. The process is continued until convergence—that is, until the coefficient estimates stop changing.We now try this out on a toy example.

a)

Generate a response Y and two predictors X1 and X2, with n = 100.

I used this equation to generate response Y

```
set.seed(702)
x1 = rnorm(100)
x2 = rnorm(100)
eps = rnorm(100, sd = 0.1)
y = -.21*x1 + .72*x2 +eps
y
```

```
##   [1]  0.056973823  0.051337084  0.307919643  0.656686218  3.145175176
##   [6] -0.218709812  0.491635180  0.653196752  0.962488375  0.380546592
##  [11]  0.048466272 -1.115580913  1.382164892 -0.212410783  1.082746849
##  [16]  0.103045469 -0.706730539  0.013031592  0.905110183  0.038787755
##  [21] -0.208964650 -0.149184272 -0.377105125  0.855289873 -0.633015837
##  [26] -0.455131294  0.922918397 -1.713910149  0.895820424 -0.693956470
##  [31] -0.106725914  1.033731886 -0.038729811  0.100682191  0.504255675
##  [36]  0.182008156 -0.820074847 -0.348957019  2.116831309 -0.006939019
##  [41]  1.101455971  0.611934280  1.144696179  0.019668254 -0.921270480
##  [46] -0.038327429  0.043927618  0.240552039  0.079864394 -0.106266430
##  [51]  0.300523025  0.188594050  0.136943371 -0.995707463  0.389470334
##  [56] -0.712725492  0.806027790 -1.384794629 -1.570219166  0.371931284
##  [61]  1.278262854 -0.129634178  0.735502777 -0.316151522  0.466116343
##  [66] -0.549902019  0.632975601 -0.628873668  0.279076157  0.603580048
##  [71]  0.224158657 -0.450724191 -0.151748529  0.208893144 -1.255802936
##  [76] -0.633602437 -0.522006621  0.692500615 -0.332566747 -0.072641716
##  [81] -1.999528172  1.348016393 -0.439891137  0.067238801  1.030155533
##  [86]  0.401077933 -1.085408504  1.426160395  0.491455790 -0.568868026
##  [91]  0.459738017  0.503721906 -0.761821781  0.579071967 -0.166435704
##  [96]  0.145629815 -0.314754635  0.075493917 -0.315877403 -0.673657357
```

(b)

Initialize $\hat{\beta}_1$ to take on a value of your choice. It does not matter what value you choose.

Initialize $\hat{\beta}_1 = 8$

(c)

Keeping β_1 fixed, fit the model

y-beta1 x1= beta0 +beta2 x2+esp

You can do this as follows:

```
##       x2
## 1.184977
```

(d)

Keeping β_2 fixed, fit the model y-beta2 x2= beta0 +beta1 x1+esp

You can do this as follows:

```
##         x1
## -0.1787137
```

(e) Write a for loop to repeat (c) and (d) 1,000 times. Report the estimates of beta0,beta1 and beta2 at each iteration of the for loop.Create a plot in which each of these values is displayed, with beta0,beta1 and beta2 each shown in a different color.

Estimates of beta0,

```
##  [1] -0.069114492 -0.003515606 -0.003291867 -0.003291104 -0.003291101
##  [6] -0.003291101 -0.003291101 -0.003291101 -0.003291101 -0.003291101
## [11] -0.003291101 -0.003291101 -0.003291101 -0.003291101 -0.003291101
## [16] -0.003291101 -0.003291101 -0.003291101
```

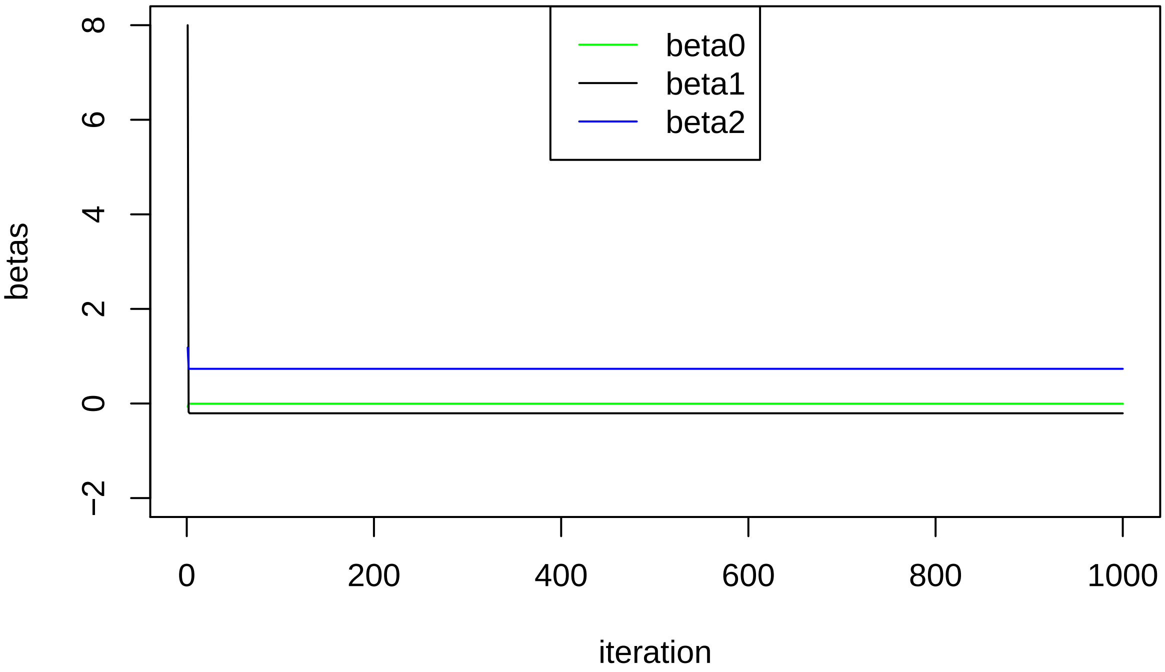
Estimates of beta1

```
##  [1]  8.0000000 -0.1787137 -0.2066090 -0.2067041 -0.2067045 -0.2067045
##  [7] -0.2067045 -0.2067045 -0.2067045 -0.2067045 -0.2067045 -0.2067045
## [13] -0.2067045 -0.2067045 -0.2067045 -0.2067045 -0.2067045 -0.2067045
## [19] -0.2067045 -0.2067045 -0.2067045 -0.2067045
```

Estimates of beta2

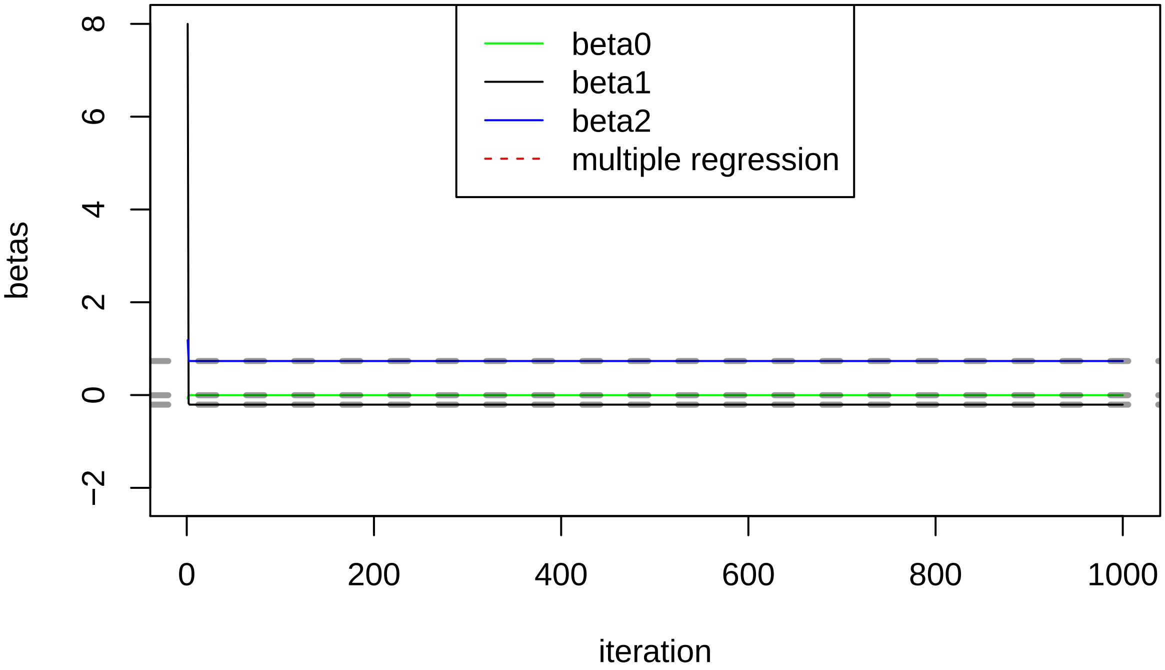
```
##  [1] 1.1849770 0.7348202 0.7332848 0.7332796 0.7332796 0.7332796 0.7332796
##  [8] 0.7332796 0.7332796 0.7332796 0.7332796 0.7332796 0.7332796 0.7332796
## [15] 0.7332796 0.7332796 0.7332796 0.7332796 0.7332796 0.7332796 0.7332796
## [22] 0.7332796 0.7332796
```

Plot



(f)

Compare your answer in (e) to the results of simply performing multiple linear regression to predict Y using X1 and X2. Use the abline() function to overlay those multiple linear regression coefficient estimates on the plot obtained in (e).



Plot shows that coefficents obtaind from backfitting and coefeicents obtained from multiple linear regression are ovelaping to each other.Thats means, both the models predicted exaxtly same coefficents.

(g)

On this data set, how many backfitting iterations were required in order to obtain a “good” approximation to the multiple regression coefficient estimates?

Ans As per the definition of backfitting , Iteration process is continued until convergence—that is, until the coefficient estimates stop changing.Plot of betas shows predected coefficent are almost constant for all iteration but by looking at the estimated value of betas , beta0 ,beta1 and beta2 stop changing after 5 , 5 and 4 iterations respectievly. Hence, we requred to have 5 backfitting iteration to obtain a good appproximation to multiple regression.