Homework 10

Snigdha Peddi

Exercises (ISLR)

Use set.seed(202110) in each exercise to make results reproducible.

Be explicit in citing all of your sources.

Question 1. In this exercise, you will further analyze the rock data set. *You can use Dr. Saunders' toy example from the ridge regression code to help*

a) Perform polynomial regression to predict `area` using `perimeter`. 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.

Answer:

A Quadratic model is fit for the Rock data using *poly()* function and 10 degrees. The coefficients of the model indicate that the linear model with degree 1 and quadratic model with degree 3 are significant at 95% confidence interval with the significant p-values.

```
##
## Dimensions of Rock dataset: 48 4
```

```
## Quadratic Model for Rock Data:
```

```
## glm(formula = area ~ poly(peri, 10), data = rock)
```

```
##
## Coefficients of Quadratic Model for Rock Data:
```

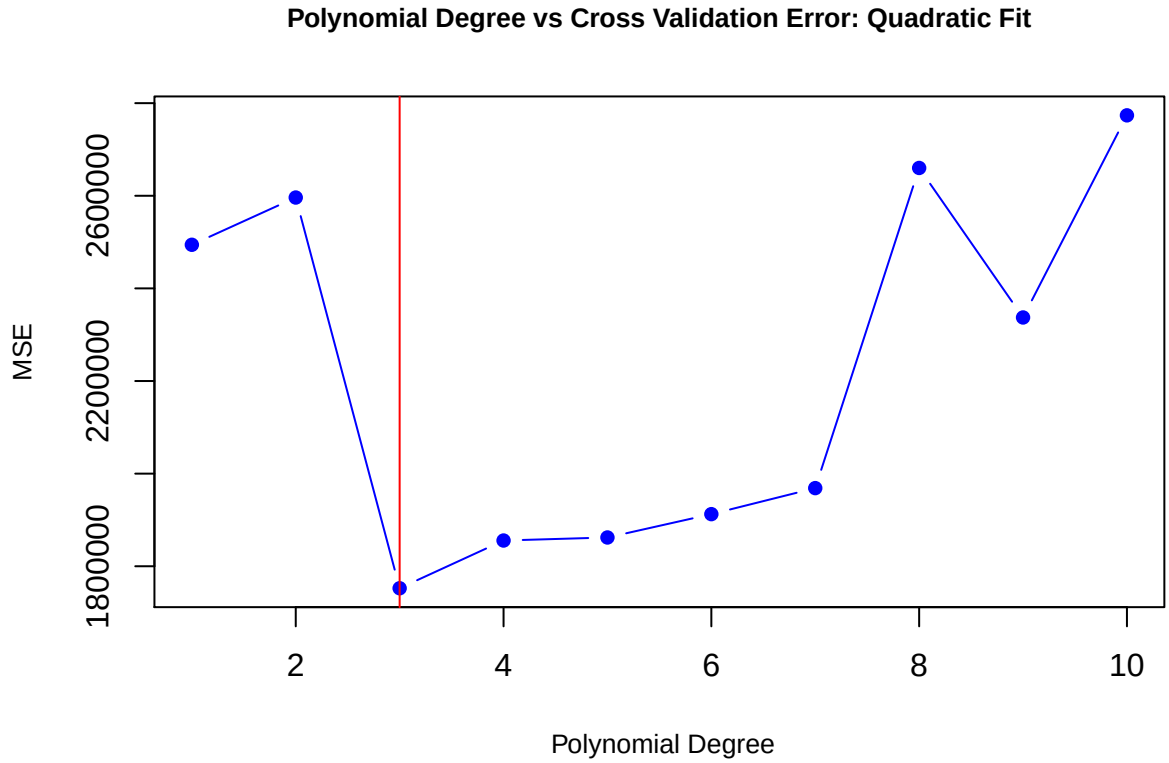
```
##                     Estimate Std. Error     t value     Pr(>|t|)
## (Intercept)       7187.72917   196.0191 36.66850555 1.084609e-30
## poly(peri, 10)1  15133.74008  1358.0605 11.14364216 2.211628e-13
## poly(peri, 10)2  -1440.44225  1358.0605 -1.06066134 2.957187e-01
## poly(peri, 10)3   5819.44550  1358.0605  4.28511511 1.249231e-04
## poly(peri, 10)4   -238.58003  1358.0605 -0.17567703 8.615056e-01
## poly(peri, 10)5   -911.41141  1358.0605 -0.67111253 5.063172e-01
## poly(peri, 10)6    866.53146  1358.0605  0.63806544 5.273607e-01
## poly(peri, 10)7     54.37577  1358.0605  0.04003929 9.682771e-01
## poly(peri, 10)8   -866.34289  1358.0605 -0.63792658 5.274501e-01
## poly(peri, 10)9  -1269.41780  1358.0605 -0.93472847 3.559889e-01
## poly(peri, 10)10  1152.67090  1358.0605  0.84876257 4.014715e-01
```

cv.glm() function is used to perform LOOCV for the Rock Data upto 10 polynomial degrees. From the table below it is clear that a Quadratic fit with a 3rd degree polynomial is a good fit with the lowest MSE of 1752525. This can also be observed in the plot between the polynomial degree and the MSE.The vertical line indicates that the lowest MSE is for a fit using 3rd order polynomial of perimeter variable.

Table 1: LOOCV Errors for Quadratic fit of Rock Data

Degree	MSE
1	2494174
2	2596193
3	1752525
4	1855568
5	1862120
6	1912522
7	1968732
8	2660026
9	2337201
10	2773639

```
##
## Degree at which Lowest MSE observed: 3

##
##
## Lowest MSE observed: 1752525
```

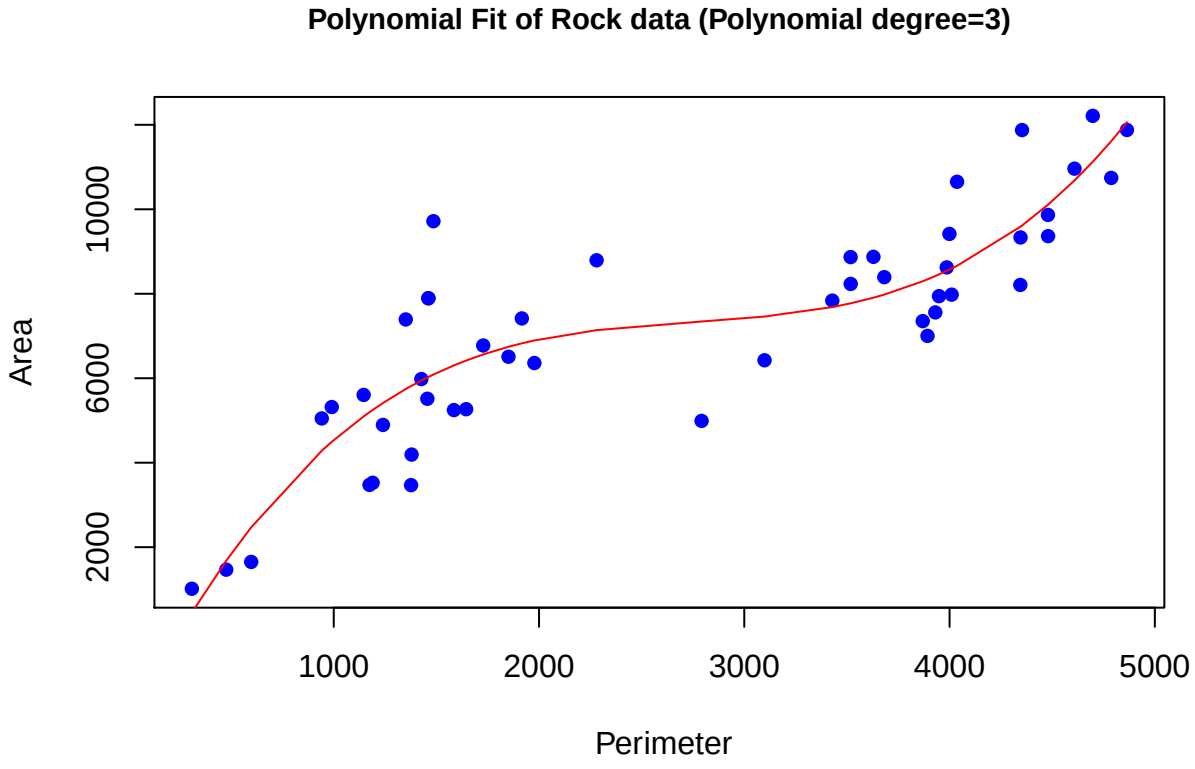


Hypothesis testing is performed using Analysis of Variance,*ANOVA*. To use the anova a series of linear nested models were fit from degree 1 to degree 10 polynomials.The p-value of linear model is zero indicating that this model is not sufficient to explain the data.The significant pvalue of degree 3 polynomial indicates that it failed to reject the null hypothesis and this model is sufficient to explain the data than the higher order complex models. A plot of the resulting polynomial fit is presented below.

Table 2: P-value of different polynomial fits

Degree	Pvalue
1	NA
2	0.29572
3	0.00012
4	0.86151
5	0.50632
6	0.52736
7	0.96828
8	0.52745
9	0.35599
10	0.40147

```
##
## Quadratic model with Optimum Degree,3:

## lm(formula = area ~ poly(peri, 3), data = rok)
```



b) Fit a step function to predict `area` using `perimeter`, and perform cross validation to choose the optimal number of cuts. Make a plot of the fit obtained. Do not print out every single model fit from the step function. If you are having issues, please ask!

Answer:

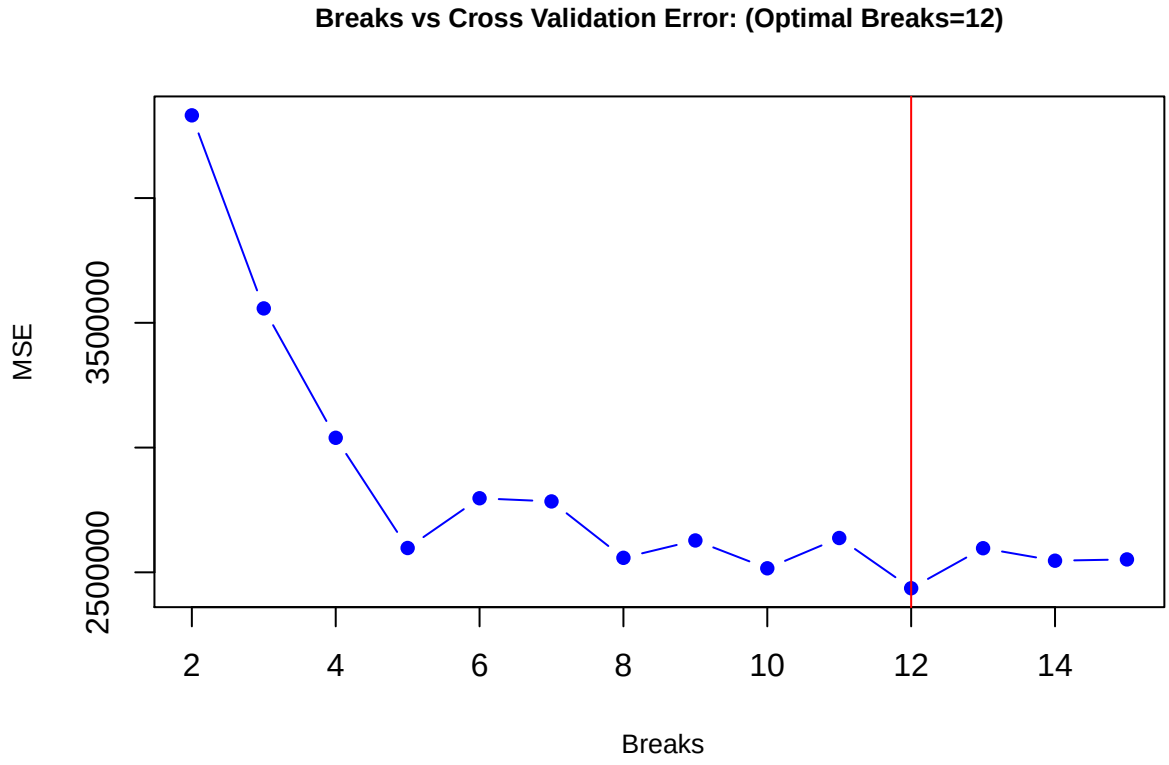
Using a for loop and *cut()* function a number of linear models were fit with cuts ranging from 2 to 15. A LOOCV is simultaneously performed. The table below shows the number of cuts made to the perimeter variable and the MSE of the fits. From the table and the plot between number of cuts and cross validation error it is clear that the lowest MSE is observed when perimeter variable was grouped into 12 bins.

Table 3: Breaks vs MSE of the models

Cut	mse
2	4331433
3	3557706
4	3039155
5	2597468
6	2796897
7	2784226
8	2558119
9	2627751
10	2516150
11	2637565
12	2436256
13	2596596
14	2546616
15	2551737

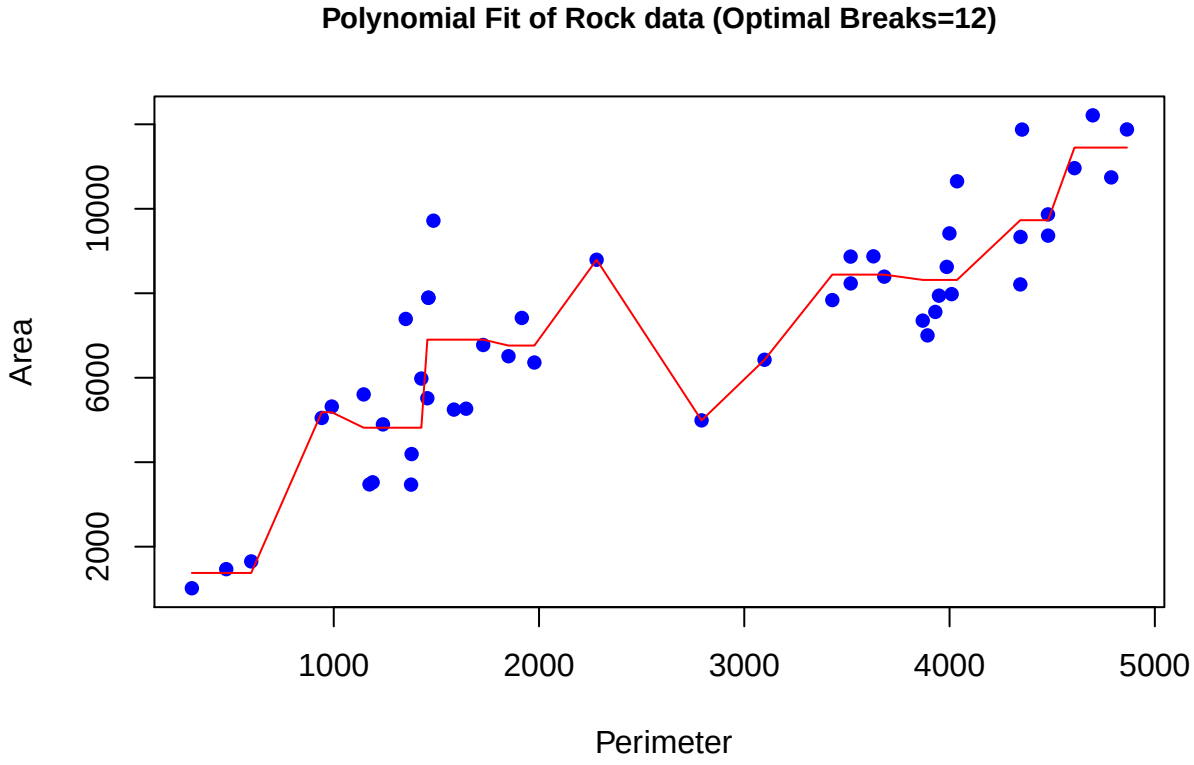
```
##
##
## Lowest MSE observed: 2436256
```

```
##
##
## Cut at which MSE is Lowest: 12
```



A linear model is fit with the optimal number of cuts selected previously(12). The predictions were made and fitted values were plotted. The plot shows some over fitting due to the binning of the perimeter variable.

```
##
## Linear model with Optimal Cuts,12:

## lm(formula = area ~ cut(peri, 12), data = rok)
```



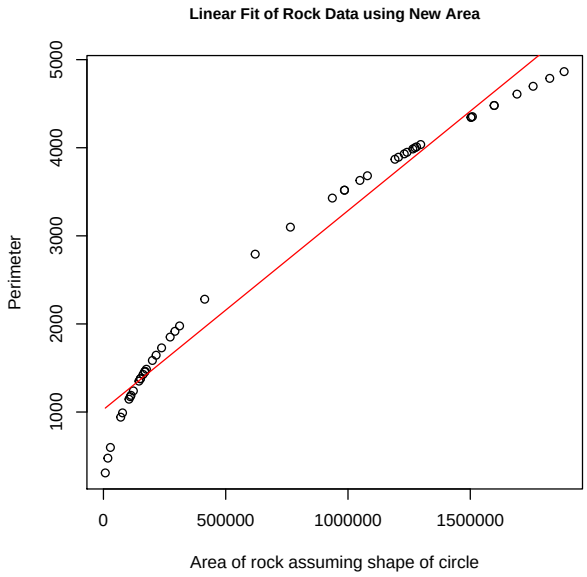
c) If all of the rocks were perfect circles, what would be the relationship between area and perimeter? If it is not linear, what does that tell you about the shape of the rocks?

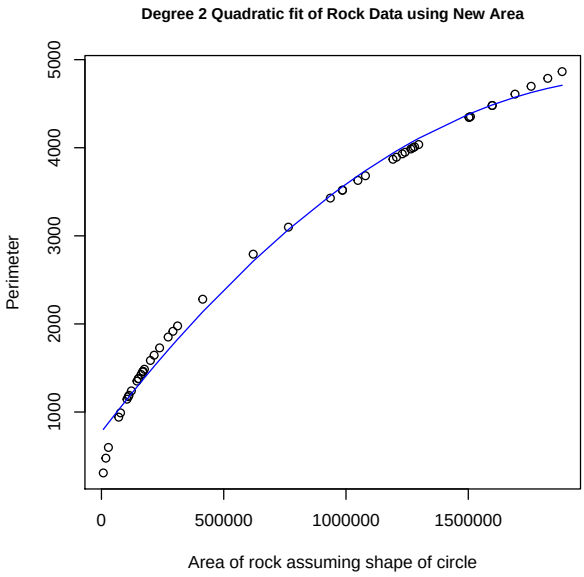
Answer:

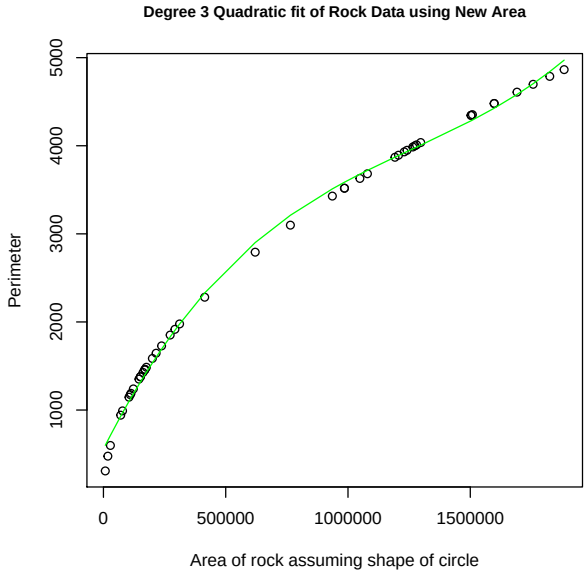
To answer this question I have used perimeter variable to derive the area of the rock (assuming the rocks are perfect circles). I had plotted the new area on x axis and the perimeter on y axis. The scatter plot shows that as the area increase the perimeter of the stone increases. To verify if the relation is linear I had fit a linear model and made predictions. The linear fit shows that though the increase in area and perimeter are directly proportional they do not have a linear fit. Then I fit few quadratic models and from the plots below it is clear that the degree 3 polynomial fit of area perfectly explains the perimeter of the rocks. This shows that the rocks if assumed to be perfect circles they are three dimensional and are spheres.

```
##
## Relationship between Area and Perimeter-Linear model:

## lm(formula = peri ~ area.cir, data = rok)

##
## Relationship between Area and Perimeter-Quadretic model-Degree 2:

## lm(formula = peri ~ poly(area.cir, 2), data = rok)

##
## Relationship between Area and Perimeter-Quadretic model-Degree 3:
```

```
## lm(formula = peri ~ poly(area.cir, 3), data = rok)
```







References

- Lecture by Abbass AI Sharif, *DSO 530: LOOCV and k-fold CV in R*,Oct 4,2013.
- Lecture by thatRnerd, *Analysis of Variance(ANOVA) in R*, May 21,2016.
- Blogpost by stackoverflow, *Plot polynomial regression curve in R.*
- Blogpost by stackoverflow, *Cross Validation step function in R.*
- RDocumentation by DataCamp,*cut:Convert Numeric to factor.*
- Chapter 7,Moving Beyond Linearity,*An Introduction to Statistical Learning with Applications in R* by Gareth James.

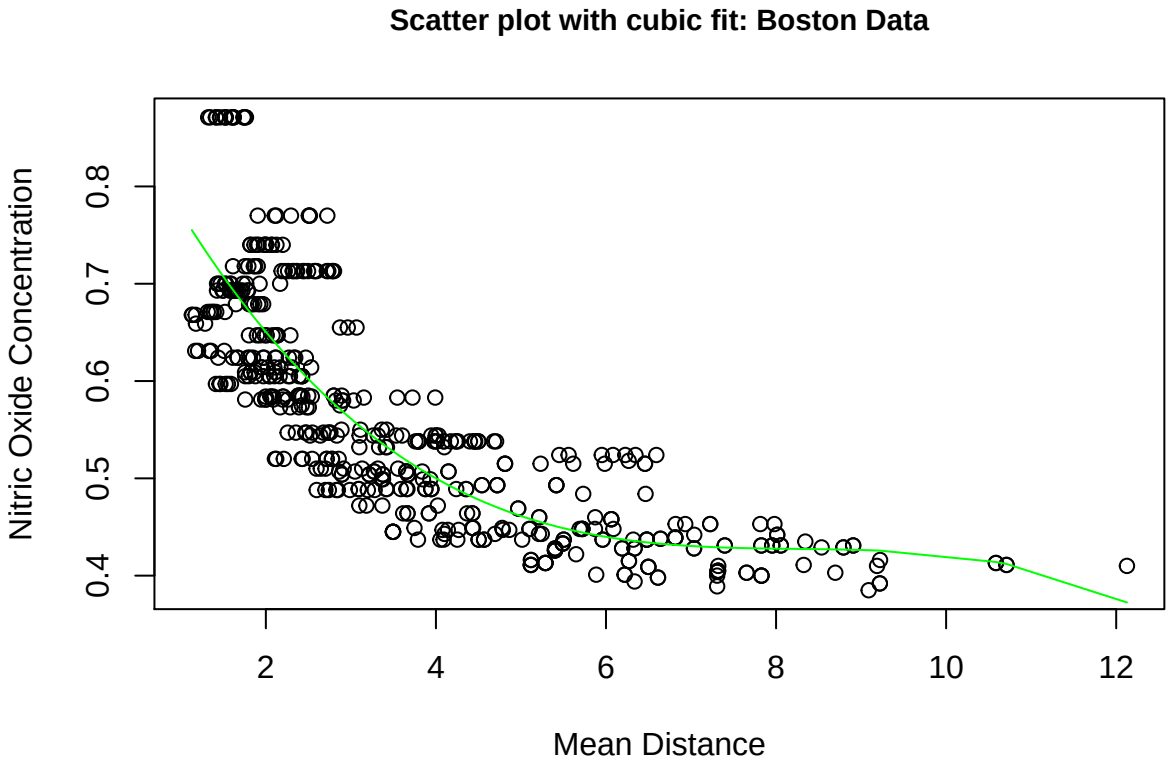
Question 2. Exercise 7.9.9 pg 299 Be explicit in citing all of your sources.This question uses the variables dis (the weighted mean of distances to five Boston employment centers) and nox (nitrogen oxides concentration in parts per 10 million) from the Boston data. We will treat dis as the predictor 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.

Answer:

poly() function is used to fit a quadratic model of 3rd order using *nox* as response variable and *dis* as predictor from Boston Data. The lower p-values from the coefficient estimates of the fit shows that there is a relation between the two variables. A scatter plot of the data and the polynomial fit is plotted. It further confirms that the quadratic fit explains the data very well and no complex model is required.

```
##
## Relationship between Distance and Nitric Oxide Concentration-Quadretic model-Degree 3:

##
## Call:
## glm(formula = nox ~ poly(dis, 3), data = Bos)
##
## Deviance 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
##
## (Dispersion parameter for gaussian family taken to be 0.003852802)
##
##     Null deviance: 6.7810  on 505  degrees of freedom
## Residual deviance: 1.9341  on 502  degrees of freedom
## AIC: -1370.9
##
## Number of Fisher Scoring iterations: 2
```



(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.

Answer:

Multiple quadratic models of order 1 to 10 were fit and associated residual sum of squares were calculated.Observing the RSS values it is clear that as the polynomial order increased the RSS reduced indicating reduced bias with increase in each polynomial order. Plots of the polynomial fits show that with increasing degree of polynomial the fit smoothend to fit few extream values of dis variable.

```
##
## Quadratic model used for Boston data:

## glm(formula = nox ~ poly(dis, i), data = Bos)
```

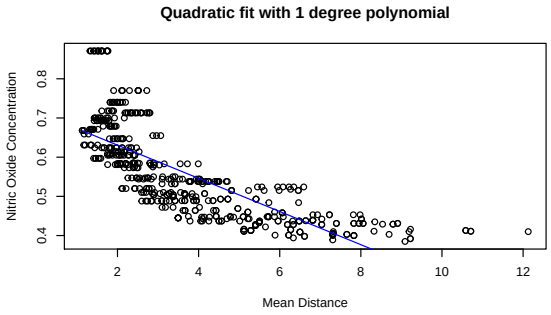
Table 4: Residual Sum of Squares for Quadratic fit of Boston Data

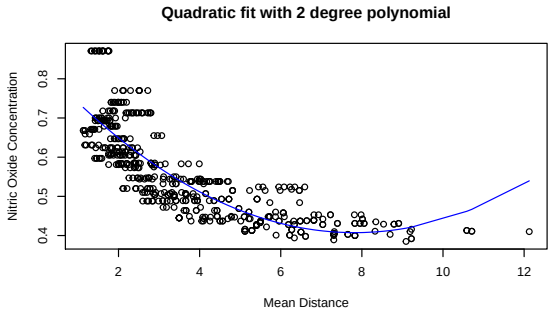
Degree	RSS
1	120.9739
2	120.2406
3	120.1394
4	120.1383
5	120.1206
6	120.0836

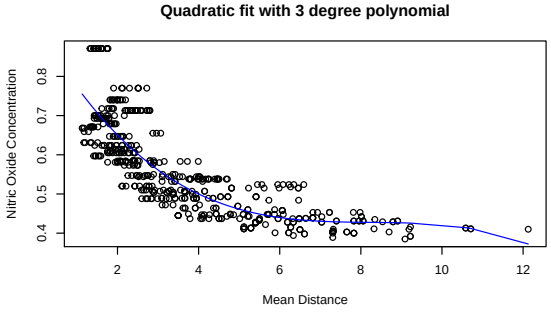
Degree	RSS
7	120.0548
8	120.0409
9	120.0387
10	120.0375

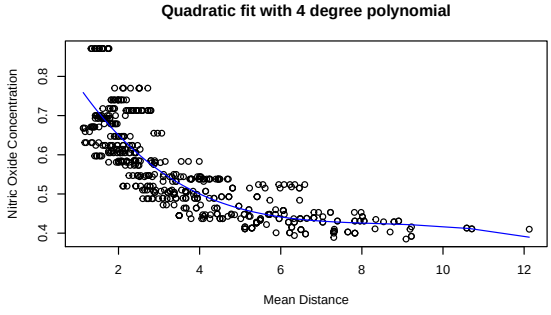
```
##
## Degree at which Lowest RSS observed: 10
```

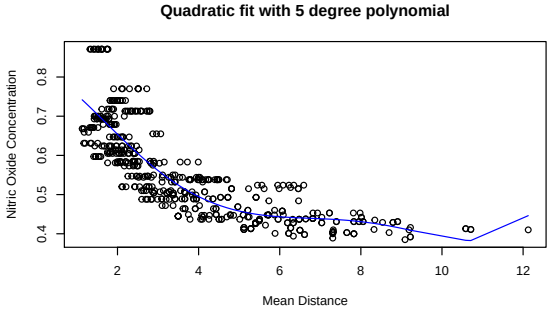
```
##
##
## Lowest RSS observed: 120.0375
```

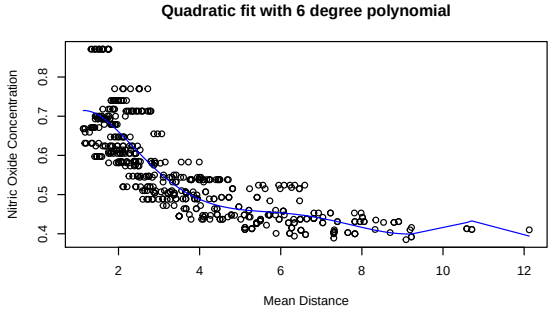

Quadratic fit with 1 degree polynomial
Nitric Oxide Concentration
Mean Distance

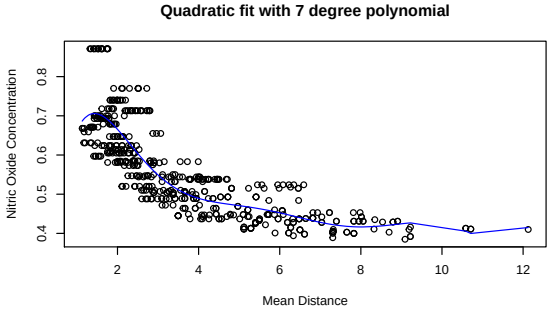

Quadratic fit with 2 degree polynomial
Nitric Oxide Concentration
Mean Distance

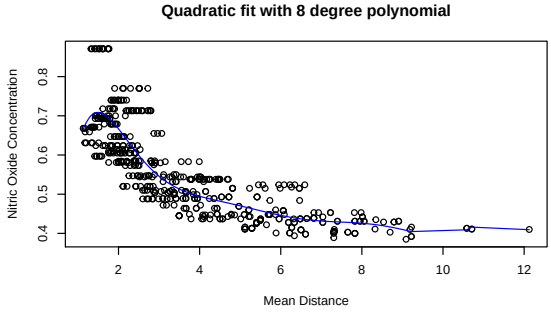

Quadratic fit with 3 degree polynomial
Nitric Oxide Concentration
Mean Distance

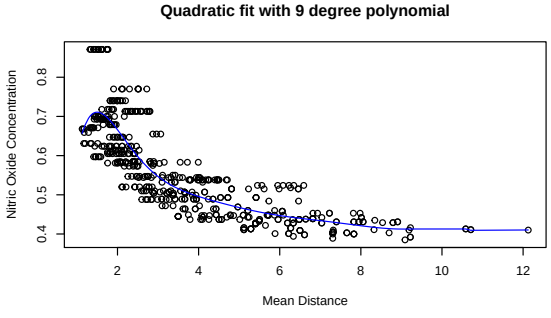

Quadratic fit with 4 degree polynomial
Nitric Oxide Concentration
Mean Distance

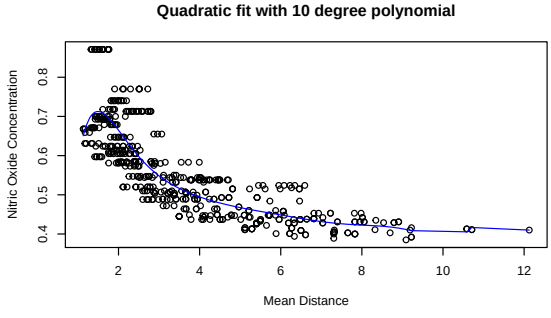

Quadratic fit with 5 degree polynomial
Nitric Oxide Concentration
Mean Distance


Quadratic fit with 6 degree polynomial
Nitric Oxide Concentration
Mean Distance


Quadratic fit with 7 degree polynomial
Nitric Oxide Concentration
Mean Distance


Quadratic fit with 8 degree polynomial
Nitric Oxide Concentration
Mean Distance


Quadratic fit with 9 degree polynomial
Nitric Oxide Concentration
Mean Distance


Quadratic fit with 10 degree polynomial
Nitric Oxide Concentration
Mean Distance

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

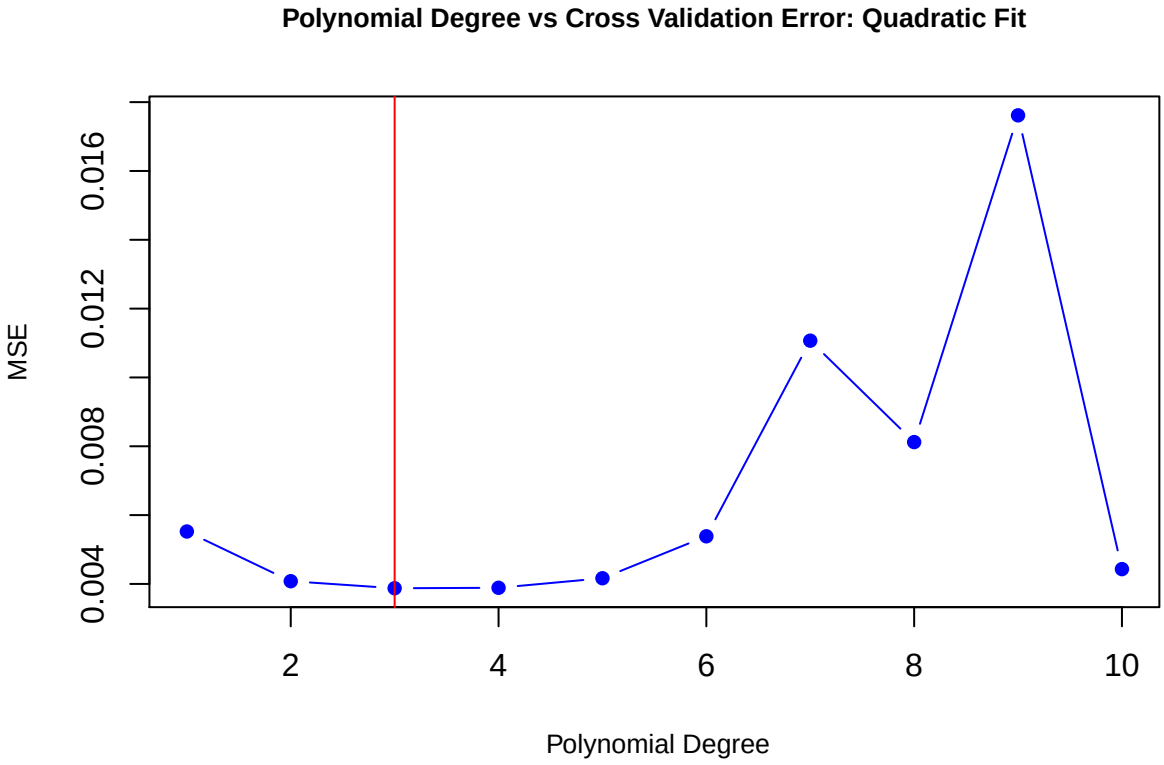
Answer:

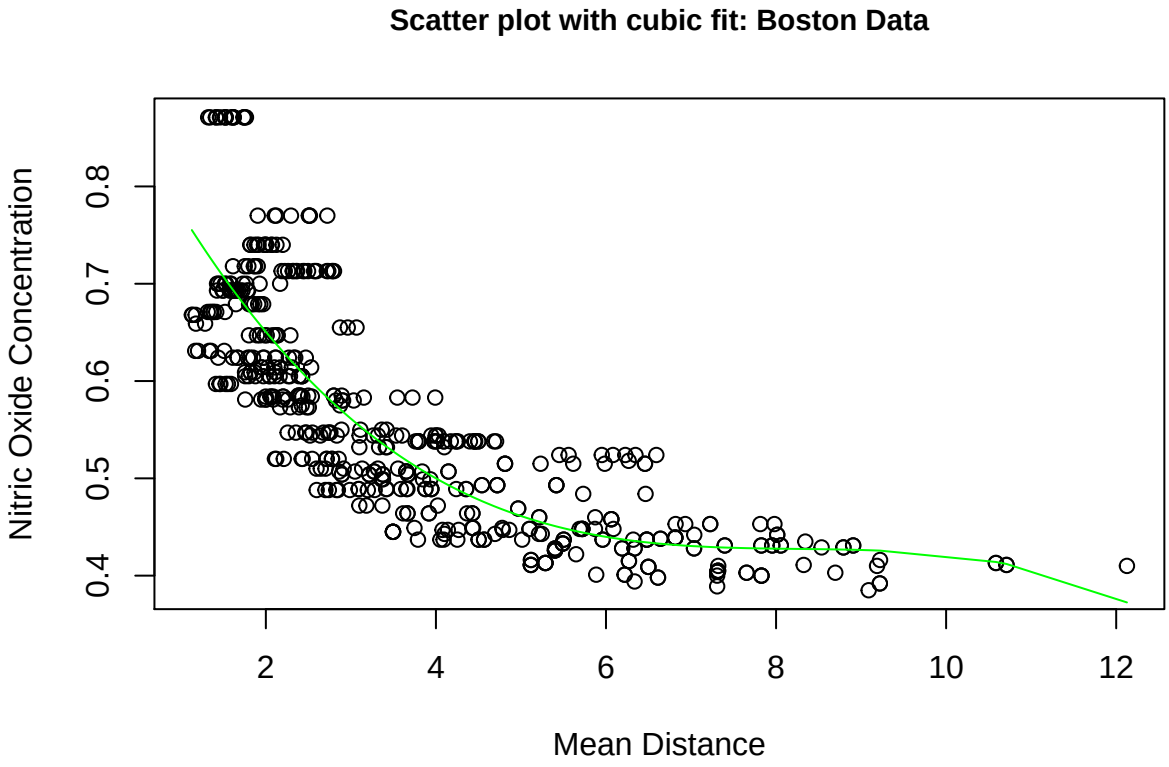
A LOOCV is performed for the quadratic models with degrees 1 to 10. The MSE of each quadratic fit is calculated and presented in the table below. The MSE values and plot between polynomial degree and associated polynomial degree show that quadratic fir of order 3 has the lowest MSE of 0.00387 and best explains the relationship between the mean distance and the nitric oxide concentration. The plot the polynomial fit of order 3 confirms the relationship.

Table 5: LOOCV Errors for Quadratic fit of Boston Data

Degree	MSE
1	0.0055239
2	0.0040794
3	0.0038748
4	0.0038875
5	0.0041649
6	0.0053843
7	0.0110688
8	0.0081214
9	0.0176164
10	0.0044303

```
##
## Degree at which Lowest MSE observed: 3

##
##
## Lowest MSE observed: 0.003874762
```


Polynomial Degree vs Cross Validation Error: Quadratic Fit
MSE
0.004
0.008
0.012
0.016
2
4
6
8
10
Polynomial Degree



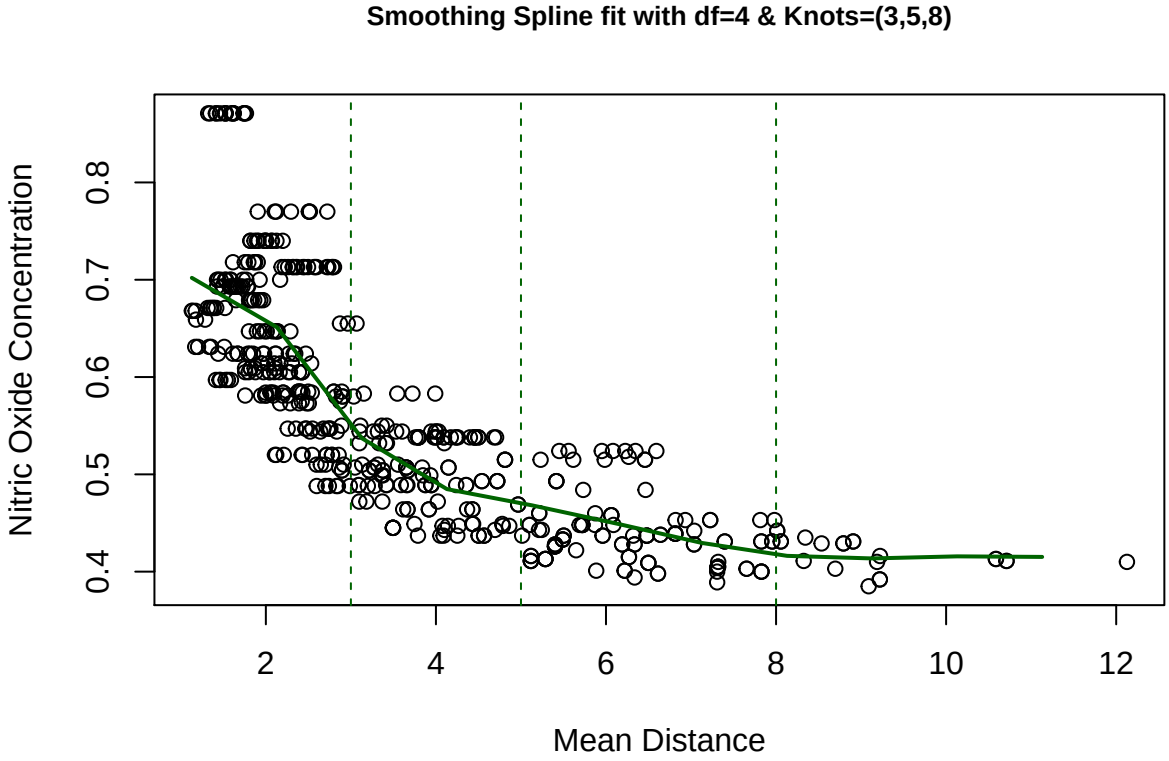
(d) Use the bs() function to fit a regression spline to predict nox using dis. Report the output for the fit using four degrees of freedom. How did you choose the knots? Plot the resulting fit.

Answer:

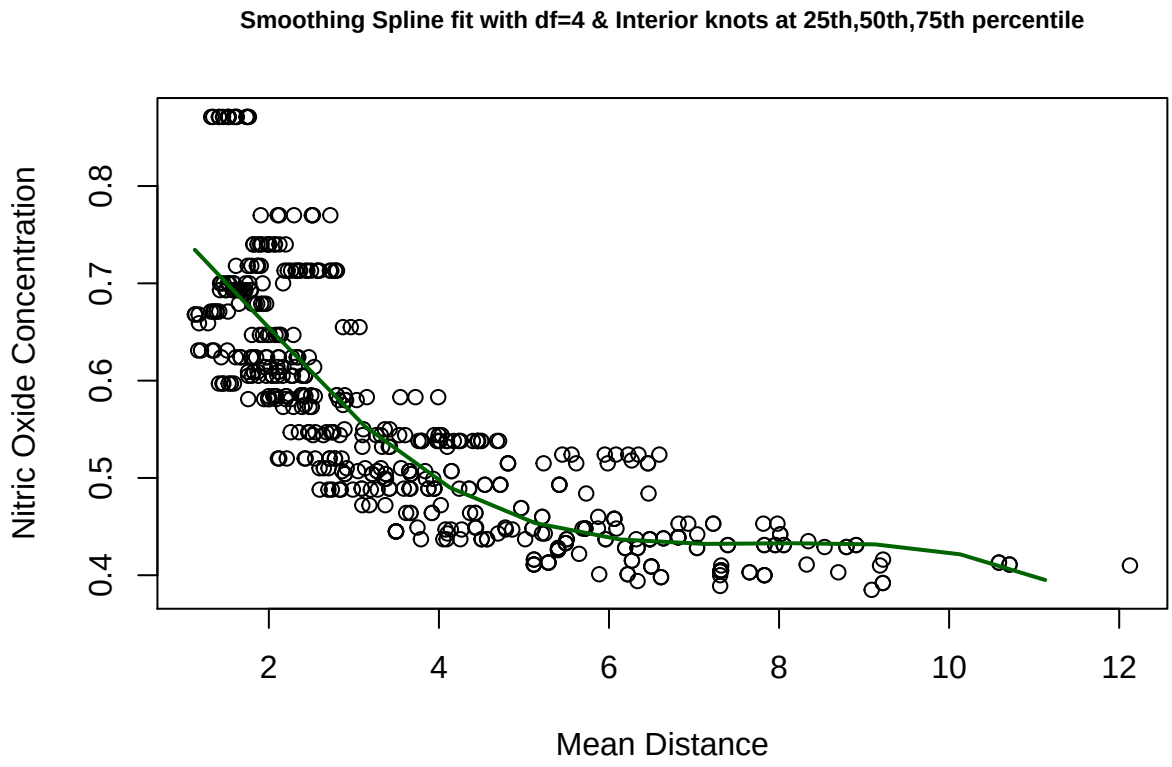
Regression spline was fit for Boston Dataset with Nitrogen Oxide concentration as response variable and Weighted mean distance as predictor variable using *bs()* function from *splines2* package. I picked the position of knots where the function might vary most rapidly and placed fewer knots at the extremes where the function is more stable. I have tried different set of knots like (2,6,9),(3,6,9) etc but I have picked (3,4,8) knots as they give the better fit compared to others.A degree of freedom of 4 and knots at 3,4,8 were used for this fit.The resultant fit is plotted. I also fit a regression spline with just degrees of freedom as 4 (which leads to 3 internal knots) with no specific knots (default knots are at 25th,50th and 75th percentile) and fit the line over the scatter plot. Both the plot with (4 df+K knots) and without (4 df) the knots look vary similar. However I chose to pick the fit with 4 degrees of freedom and knots at (3,5,8) as this fit has lower residual error (0.06102) and Adjusted R square (0.7227) compared to other fit (0.06195 and 0.7142 respectively).

```
##
## Summary of the Regression spline fit with df=4 and Knots=(3,5,8):

##
## Call:
## lm(formula = nox ~ splines::bs(dis, df = 4, knots = c(3, 5, 8)),
##     data = Bos)
##
## Residuals:
```

```
##       Min        1Q    Median        3Q       Max
## -0.133745 -0.037577 -0.009471  0.025106  0.190388
##
## Coefficients:
##                                                 Estimate Std. Error t value
## (Intercept)                                      0.70195    0.01657  42.375
## splines::bs(dis, df = 4, knots = c(3, 5, 8))1  0.02571    0.02757   0.932
## splines::bs(dis, df = 4, knots = c(3, 5, 8))2 -0.21581    0.01761 -12.256
## splines::bs(dis, df = 4, knots = c(3, 5, 8))3 -0.22703    0.02587  -8.775
## splines::bs(dis, df = 4, knots = c(3, 5, 8))4 -0.31543    0.03175  -9.935
## splines::bs(dis, df = 4, knots = c(3, 5, 8))5 -0.27229    0.05504  -4.947
## splines::bs(dis, df = 4, knots = c(3, 5, 8))6 -0.29757    0.06011  -4.950
##                                                 Pr(>|t|)
## (Intercept)                                      < 2e-16 ***
## splines::bs(dis, df = 4, knots = c(3, 5, 8))1    0.352
## splines::bs(dis, df = 4, knots = c(3, 5, 8))2  < 2e-16 ***
## splines::bs(dis, df = 4, knots = c(3, 5, 8))3  < 2e-16 ***
## splines::bs(dis, df = 4, knots = c(3, 5, 8))4  < 2e-16 ***
## splines::bs(dis, df = 4, knots = c(3, 5, 8))5 1.03e-06 ***
## splines::bs(dis, df = 4, knots = c(3, 5, 8))6 1.02e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06102 on 499 degrees of freedom
## Multiple R-squared:  0.726,  Adjusted R-squared:  0.7227
## F-statistic: 220.4 on 6 and 499 DF,  p-value: < 2.2e-16
```



```
##
## Summary of the Regression spline fit with df=4:

##
## Call:
## lm(formula = nox ~ splines::bs(dis, df = 4), data = Bos)
##
## 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 ***
## splines::bs(dis, df = 4)1 -0.05810    0.02186  -2.658  0.00812 **
## splines::bs(dis, df = 4)2 -0.46356    0.02366 -19.596  < 2e-16 ***
## splines::bs(dis, df = 4)3 -0.19979    0.04311  -4.634 4.58e-06 ***
## splines::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
```



(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.

Answer:

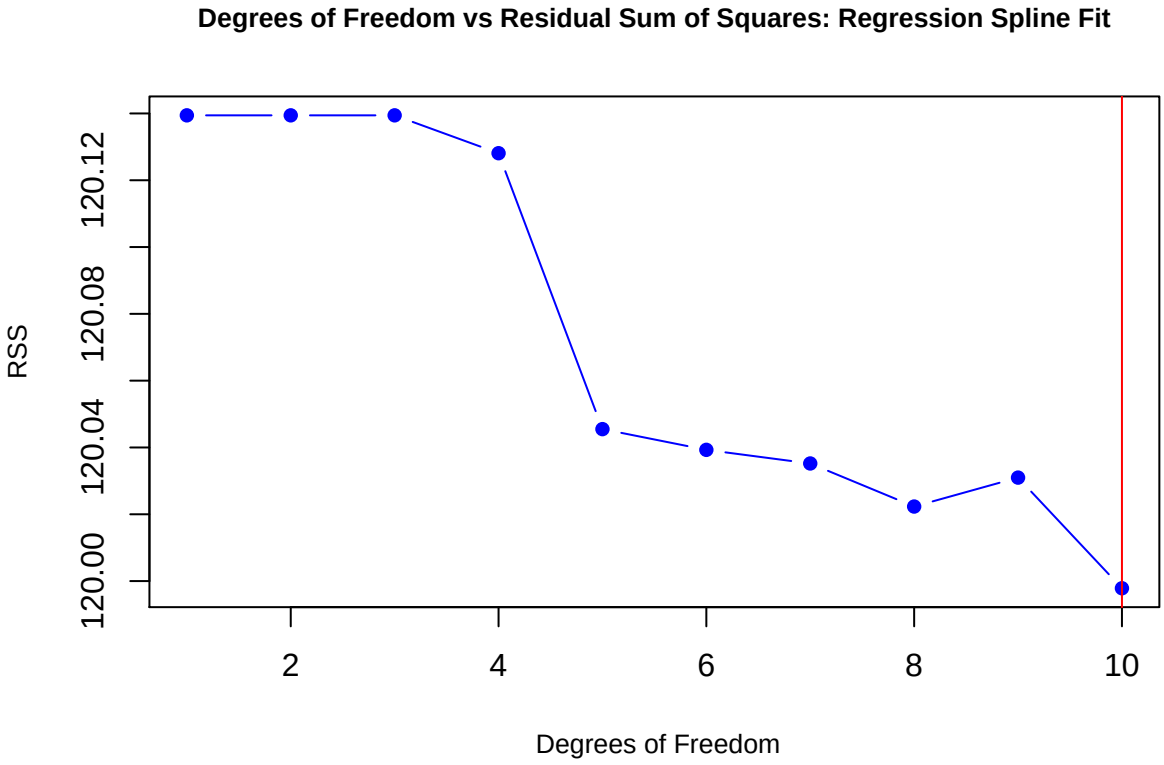
A regression spline is fit for a range of degrees of freedom if 1 to 10 (similar to number of polynomial degrees picked for question 2b). The resulting RSS of all the fits are calculated and reported in the below table. As the number of degrees of freedom increased from 0 to 1 RSS reduced. However, the reduction in RSS not substantial. The fit with 10 degrees of freedom has the lowest RSS of 199.9979. From the plots we can see than from degrees of freedom 1 through 4 the fit looks more smoother towards to smaller values of dis compared to the other fits. However the remaining portion of fit looks similar for all the fits with different degrees of freedom.

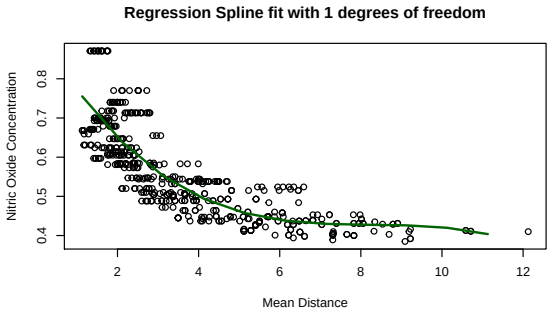
```
##
## Regression Spline model used for Boston data:

## lm(formula = nox ~ splines::bs(dis, df = i), data = Bos)
```

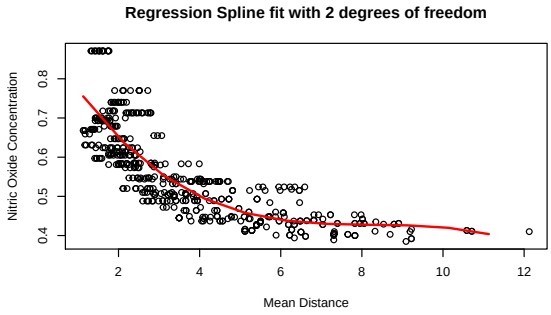
Table 6: Residual Sum of Squares for Regression Spline fit of Boston Data

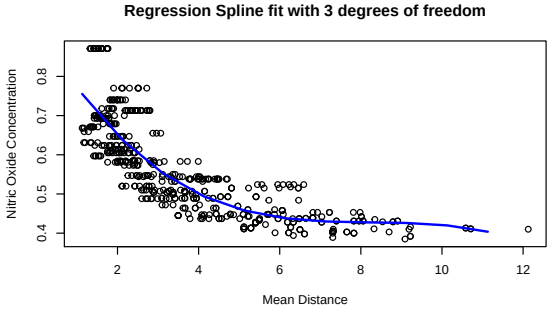
df	RSS
1	120.1394
2	120.1394
3	120.1394
4	120.1281
5	120.0455
6	120.0393
7	120.0352
8	120.0223
9	120.0310
10	119.9979

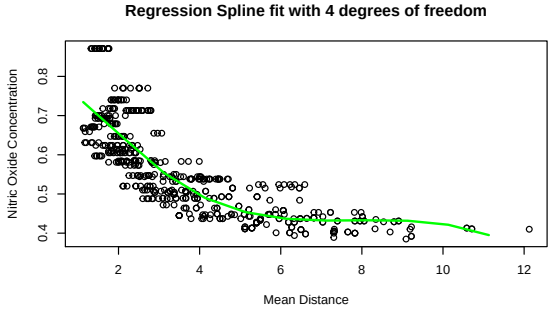
```
##
## Degrees of Freedom at which Lowest RSS is observed: 10

##
##
## Lowest RSS observed: 119.9979
```

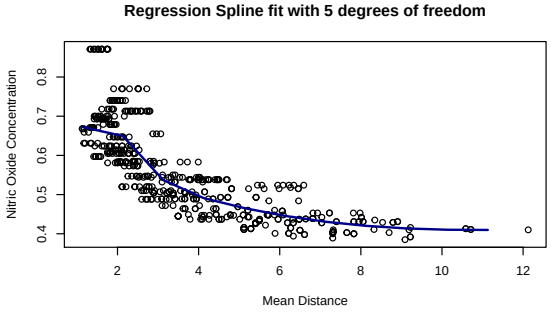

Degrees of Freedom vs Residual Sum of Squares: Regression Spline Fit
RSS
120.00
120.04
120.08
120.12
2
4
6
8
10
Degrees of Freedom

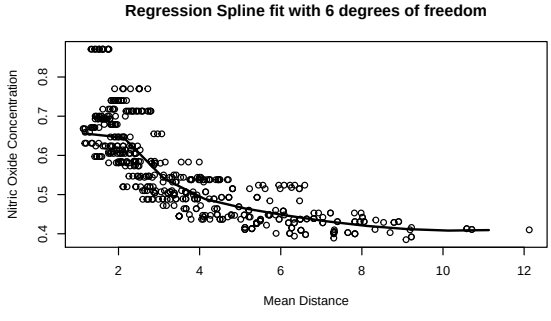

Regression Spline fit with 1 degrees of freedom
Nitric Oxide Concentration
Mean Distance

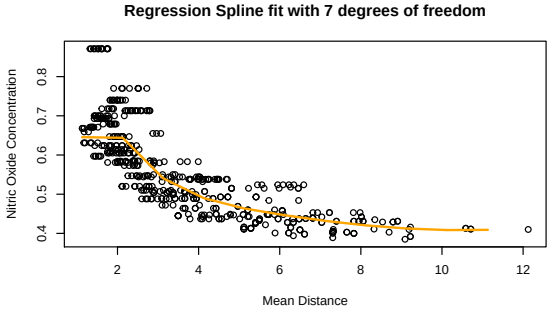

Regression Spline fit with 2 degrees of freedom
Nitric Oxide Concentration
Mean Distance

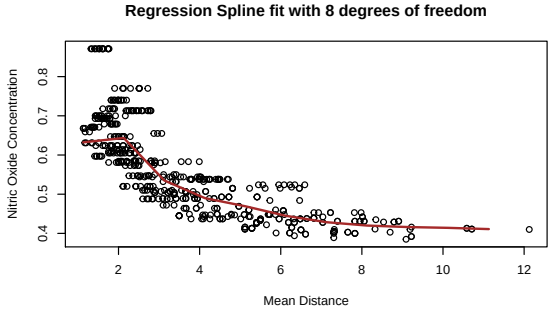

Regression Spline fit with 3 degrees of freedom
Nitric Oxide Concentration
Mean Distance

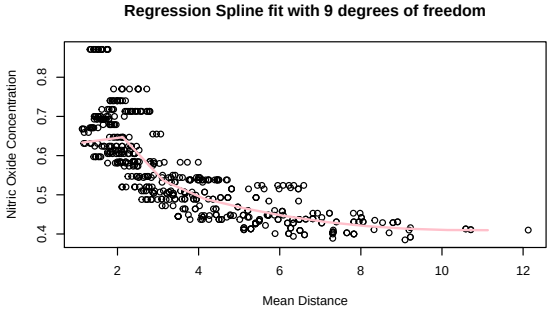

Regression Spline fit with 4 degrees of freedom
Nitric Oxide Concentration
Mean Distance

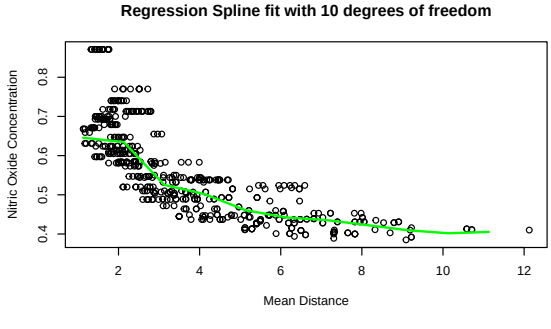

Regression Spline fit with 5 degrees of freedom
Nitric Oxide Concentration
Mean Distance


Regression Spline fit with 6 degrees of freedom
Nitric Oxide Concentration
Mean Distance


Regression Spline fit with 7 degrees of freedom
Nitric Oxide Concentration
Mean Distance


Regression Spline fit with 8 degrees of freedom
Nitric Oxide Concentration
Mean Distance


Regression Spline fit with 9 degrees of freedom
Nitric Oxide Concentration
Mean Distance


Regression Spline fit with 10 degrees of freedom
Nitric Oxide Concentration
Mean Distance

(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.

Answer:

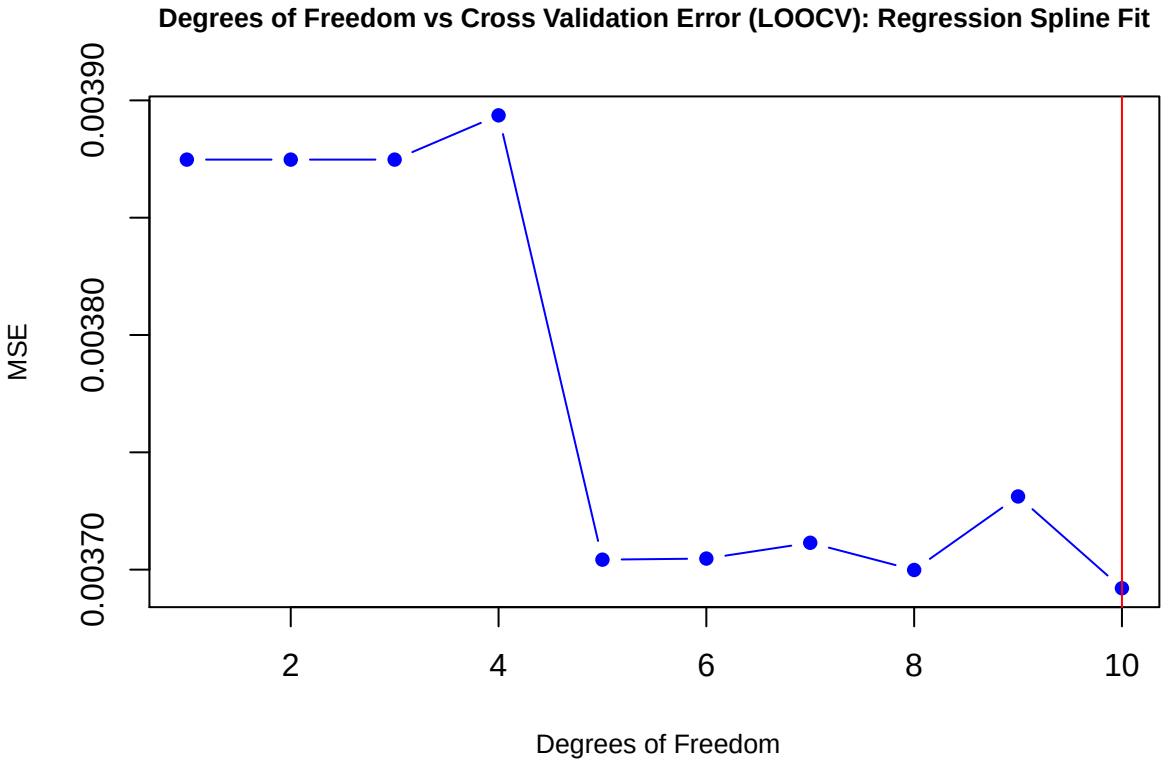
Leave One Out Cross Validation was performed for all the fits with degrees of freedom 1 to 10.The MSE of all the fits was calculated and tabulated below. The MSE of the fits decreases with increasing degrees of freedom that can be clearly observed in the plot. A lowest MSE of 0.00369 is observed with a df of 10.

Table 7: LOOCV Errors for Regression spline fit of Boston Data

df	MSE
1	0.0038748
2	0.0038748
3	0.0038748
4	0.0038936
5	0.0037043
6	0.0037047
7	0.0037114
8	0.0036999
9	0.0037312
10	0.0036921

```
##
## Degrees of freedom at which Lowest MSE is observed: 10
```

```
##
##
## Lowest MSE observed: 0.003692067
```



References

- RDocumentation by DataCamp *bs: B-Spline Basic for polynomial Splines*
- Blogpost by stackoverflow, *bs() fucntion not found.*
- Blogpost by datascience+, *Cubic and Smoothing Splines in R*, April 4,2018.
- Chapter 7,Moving Beyond Linearity,*An Introduction to Statistical Learning with Applications in R* by Gareth James.