Homework 10

Rylie Fleckenstein

Exercises (ISLR)

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

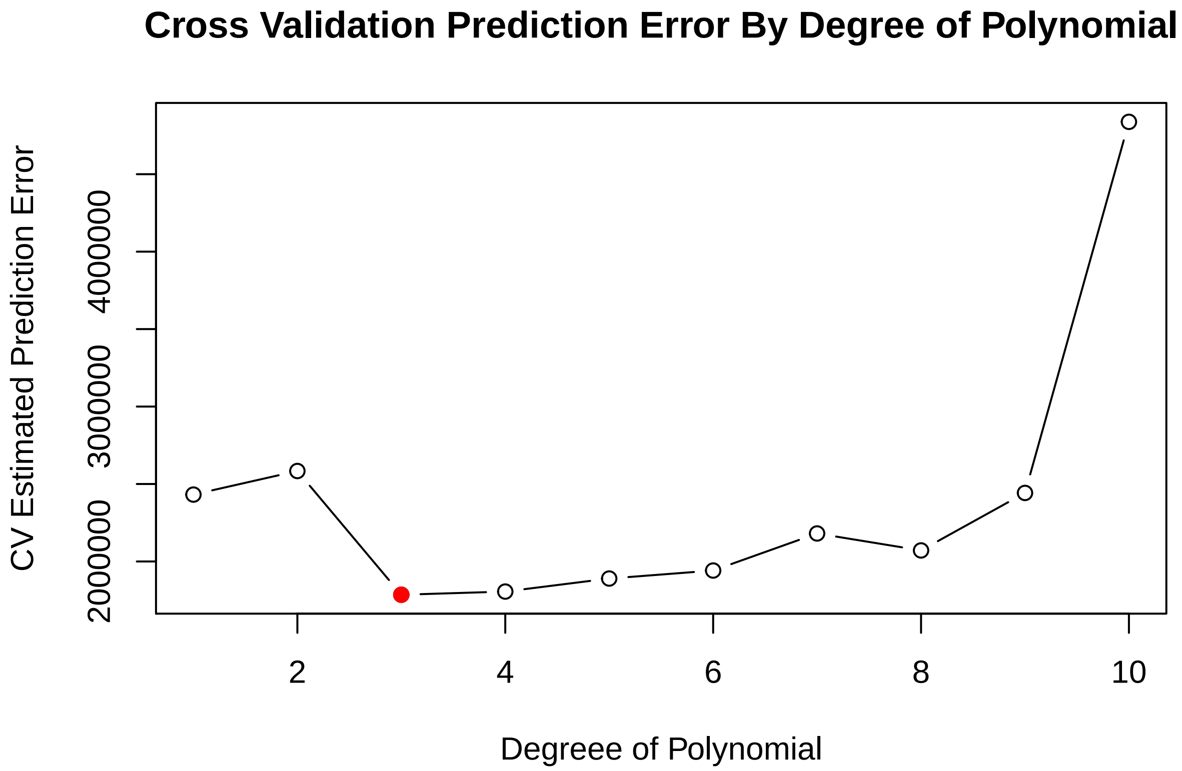
Be explicit in citing all of your sources.

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.

Some code was adapted from the code found on pages 193 and 290 of the textbook "Introduction to Statistical Learning" (ISLR) (Gareth, 2013).

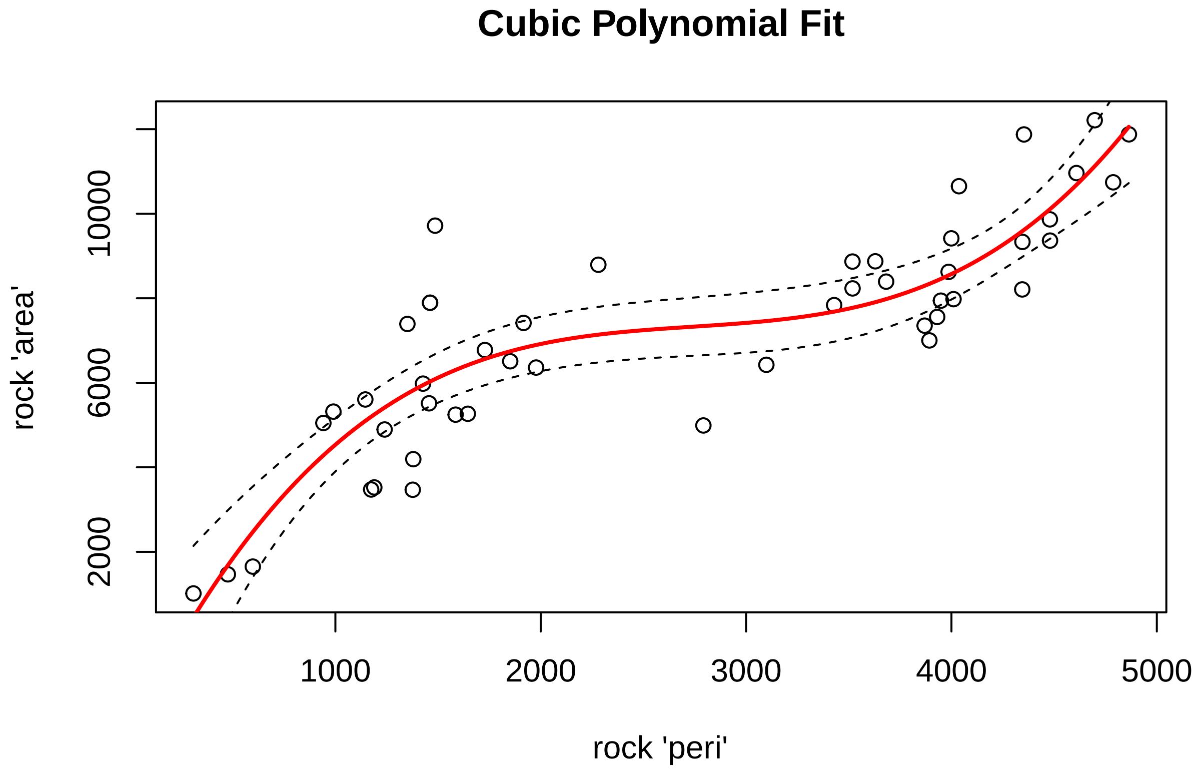
Below I implemented code to perform cross validation for selecting the optimal degree of the polynomial to be used in the linear model $area = \beta_0 + \beta_1 perimeter + \beta_2 perimeter^2 + ... + \beta_n perimeter^n$ In the plot below we can see the lowest prediction error was achieved when using a polynomial of the 3rd degree. In order to validation this result, we then performed an ANOVA test on the each level of polynomial and also found that the model with the polynomial of the 3rd degree is the best fit.

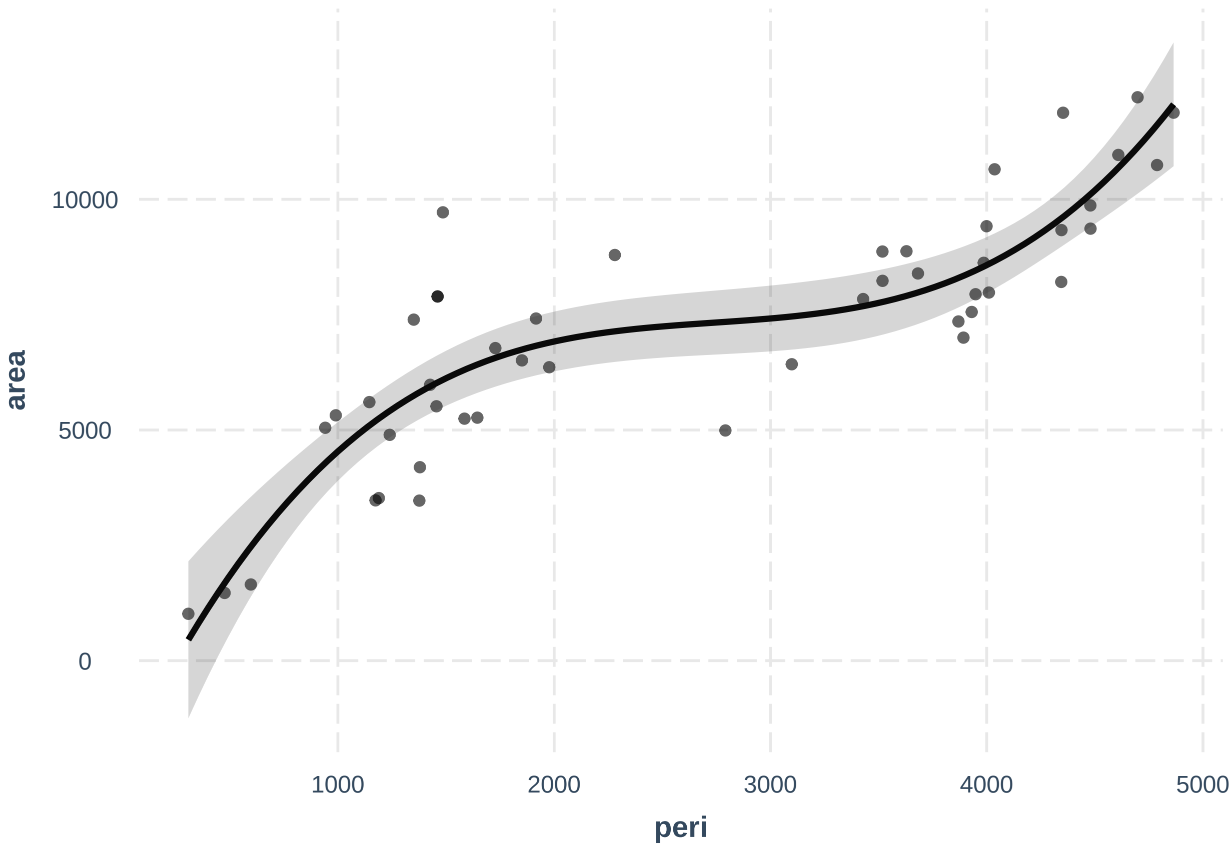


Here we are performing the ANOVA test and determining if each subsequent model better explains the data in comparison to the previous model. We can see that the model with the polynomial in the 3rd degree is significant in improving the fit of the model in comparison to the previous models. Also, the model with the 4th degree polynomial does not improve upon the 3rd degree model.

```
## Analysis of Variance Table
##
## Model 1: area ~ peri
## Model 2: area ~ poly(peri, 2)
## Model 3: area ~ poly(peri, 3)
## Model 4: area ~ poly(peri, 4)
## Model 5: area ~ poly(peri, 5)
## Model 6: area ~ poly(peri, 6)
## Model 7: area ~ poly(peri, 7)
## Model 8: area ~ poly(peri, 8)
##   Res.Df       RSS Df Sum of Sq       F   Pr(>F)
## 1     46 109513013
## 2     45 107438139  1   2074874  1.1368 0.292879
## 3     44  73572193  1  33865946 18.5553 0.000108 ***
## 4     43  73515272  1     56920  0.0312 0.860738
## 5     42  72684602  1    830671  0.4551 0.503889
## 6     41  71933725  1    750877  0.4114 0.525007
## 7     40  71930768  1      2957  0.0016 0.968100
## 8     39  71180218  1    750550  0.4112 0.525097
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here we are showing the fit of the 3rd degree polynomial (the optimal model) to the data.

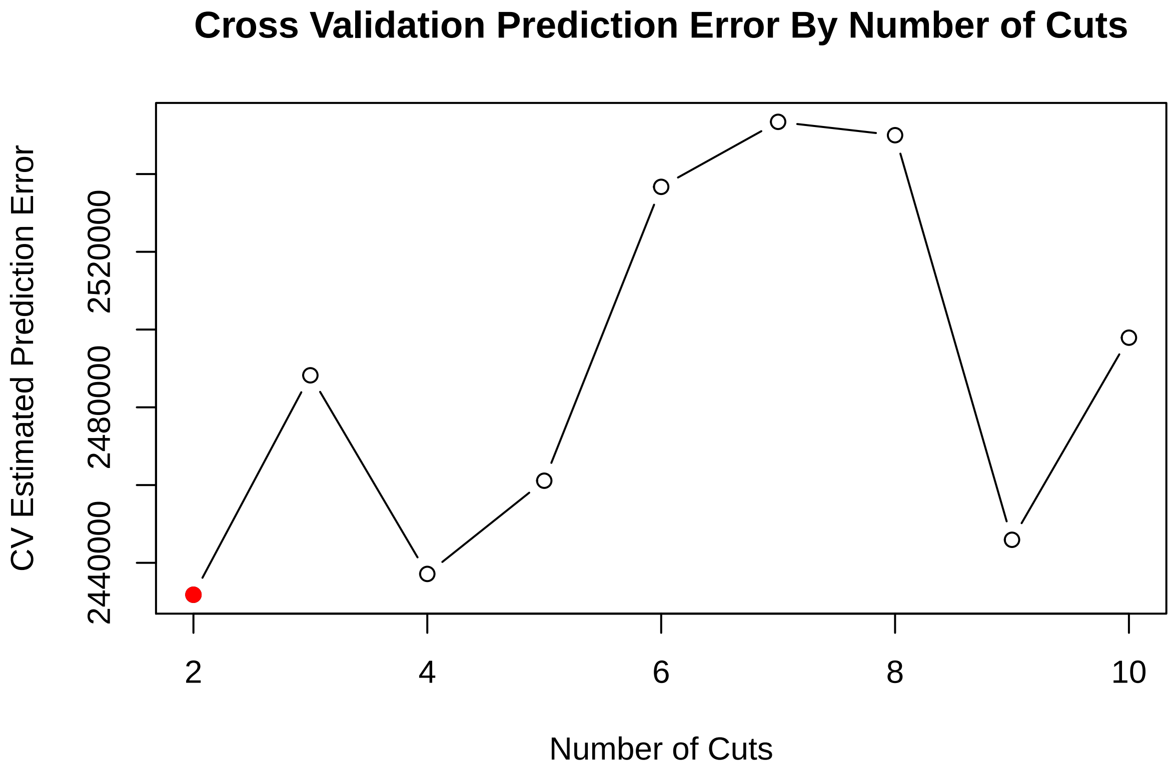




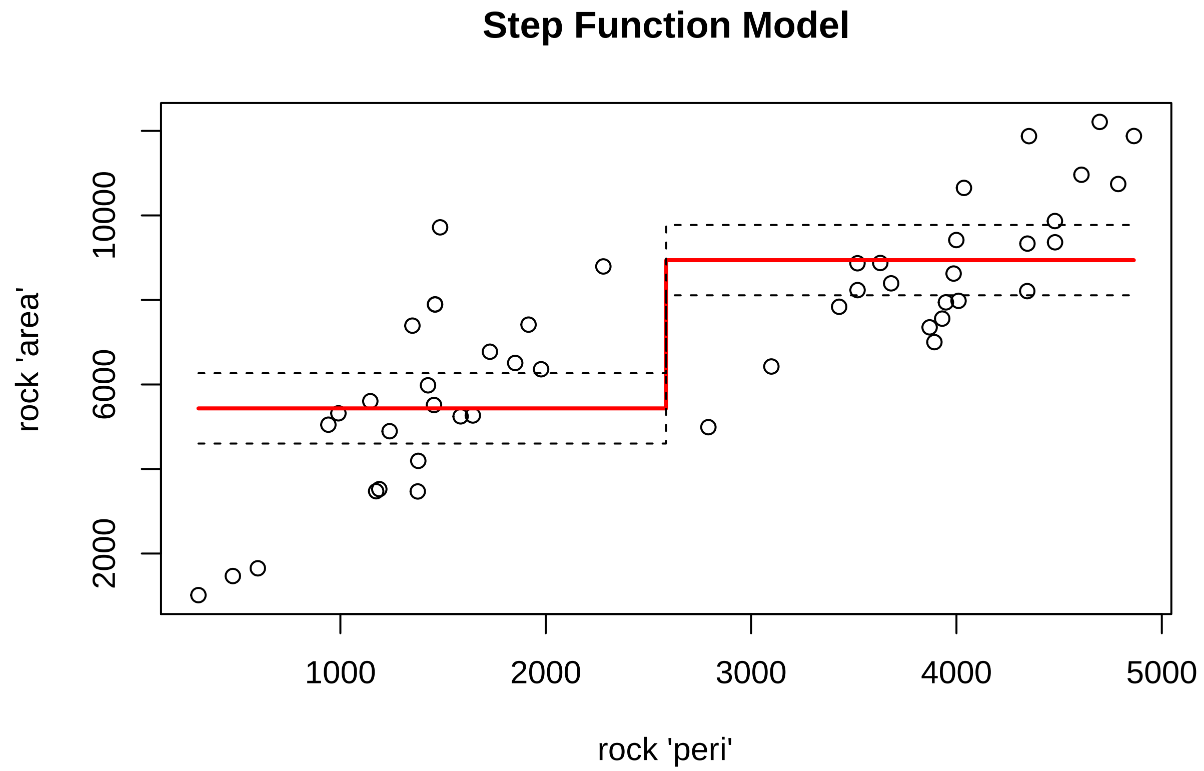
b) Fit a step function to predict `area` using `perimeter`, and perform cross validation to choose the optimal number of cuts. Make a plot of the fit obtained. *Do not print out every single model fit from the step function. If you are having issues, please ask!*

The code was adapted from the example code found on pages 289-291 in the textbook (Gareth, 2013). Here we are fitting a step function to the data through the use of "cuts" in the predicting variable. In order to find the optimal number of steps, we have implemented cross validation and compare the CV error for each number of cuts between 2 and 10.

Here we have a plot of the CV prediction error resulting from the number of steps we implement into the model. We can see that the lowest prediction error resulted from only 2 cuts in the model or 2 steps.



Here we have plotted the resulting fit of the step function model onto the data in order see how this model performs.



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?

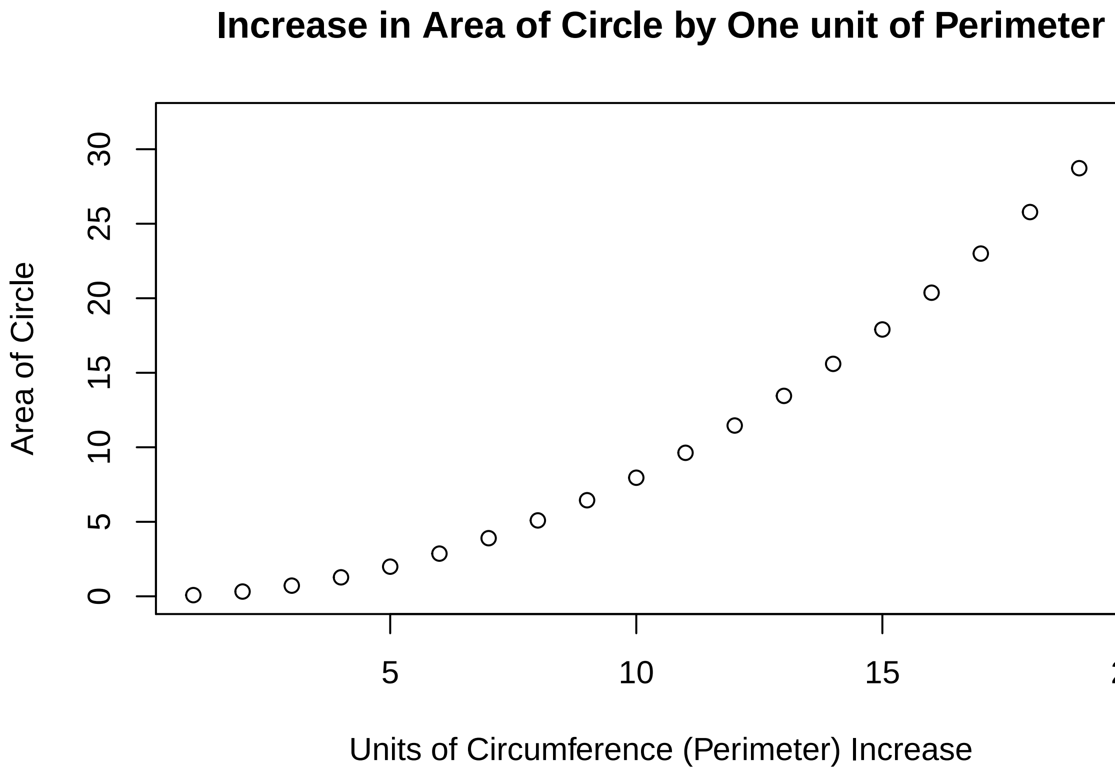
The area of a circle can be explained with the following equation: $A = \pi r^2$ while the circumference or perimeter of circle can be explained with the following equation: $C = 2\pi r$. Therefore, in order to determine the relationship between area and the perimeter of a perfect circle we solve for r in the equation for circumference and substitute it into the equation for area:

$$C = 2\pi r => r = \frac{C}{2\pi} => A = \pi(\frac{C}{2\pi})^2 = \pi(\frac{C^2}{4\pi^2})$$

Therefore, we can model the area of a circle by the increase in units of circumference with the following equation:

$$A = \frac{C^2}{4\pi}$$

If we take a look at the graph, we can see that the relationship is not linear. From this, we can deduce that even though the shape of the rocks is not perfectly circular, it is more circular then not. The model for perfect circles is quadratic in nature and the best fit model for the rocks is cubic in nature.

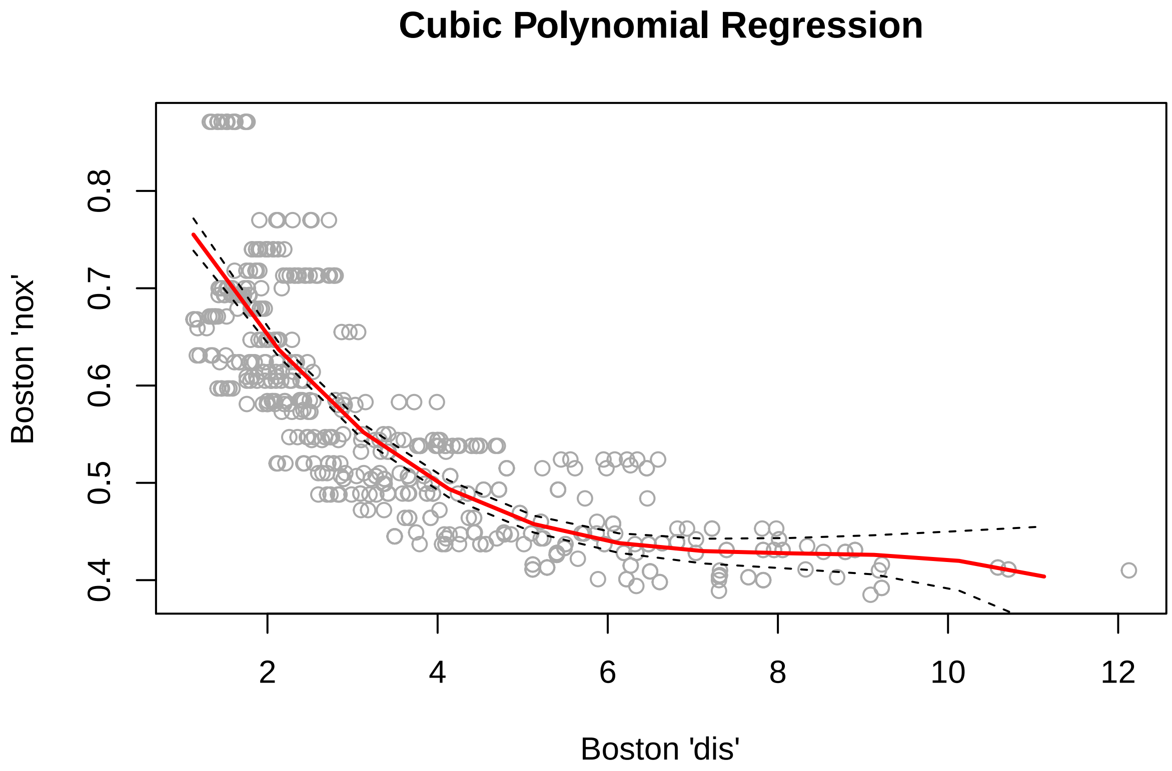


2. Exercise 7.9.9 pg 299 **Be explicit in citing all of your sources.**

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.

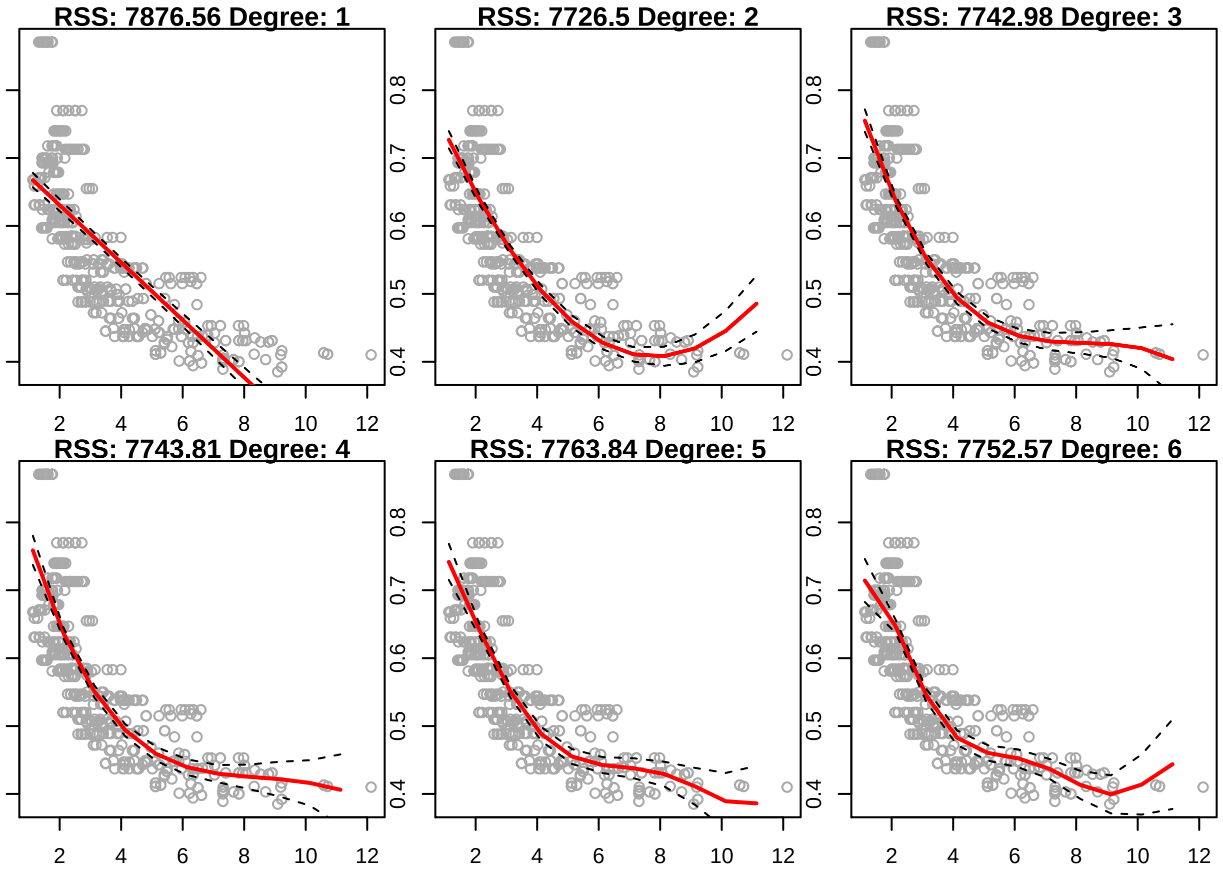
(a) Use the poly() function to fit a cubic polynomial regression to predict nox using dis. Report the regression output, and plot the resulting data and polynomial fits.

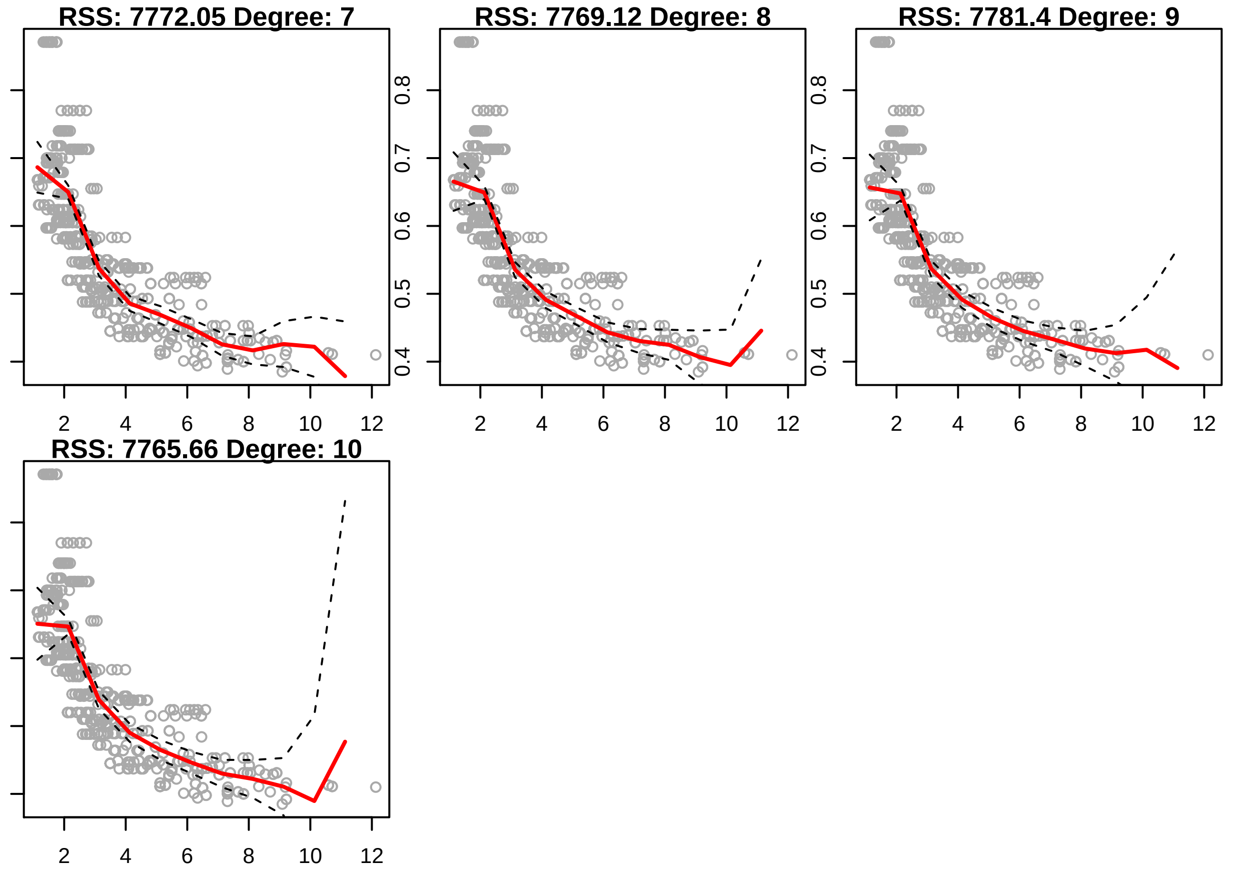
Below I have plotted the cubic polynomial regression fit to the data when predicting ‘nox’ with ‘dis’.



(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.

As we can see, when the degree of the polynomial increases there starts to be a lot of variance at the tails. Also, the RSS doesn't really see much of an increase after the first few polynomial degrees and the optimal degree, with lowest RSS, is 2.


RSS: 7876.56 Degree: 1
RSS: 7726.5 Degree: 2
RSS: 7742.98 Degree: 3
RSS: 7743.81 Degree: 4
RSS: 7763.84 Degree: 5
RSS: 7752.57 Degree: 6
0.4
0.5
0.6
0.7
0.8
2
4
6
8
10
12



RSS: 7772.05 Degree: 7

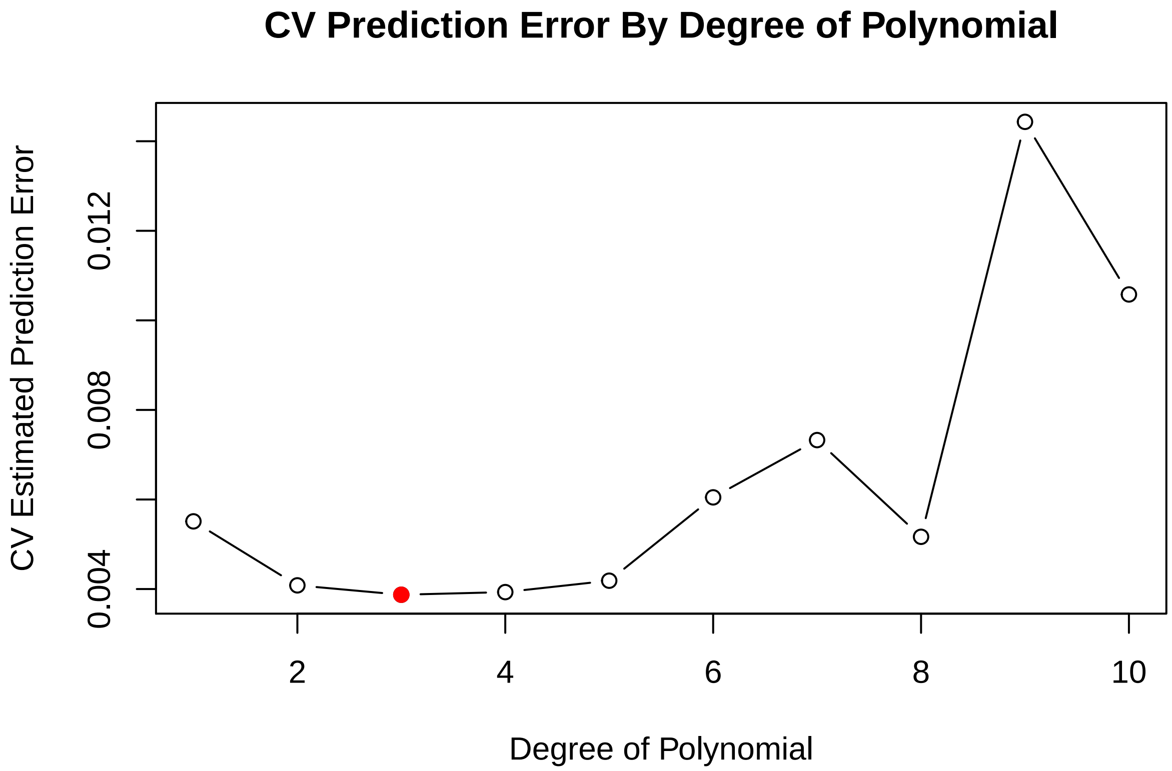
RSS: 7769.12 Degree: 8

RSS: 7781.4 Degree: 9

RSS: 7765.66 Degree: 10

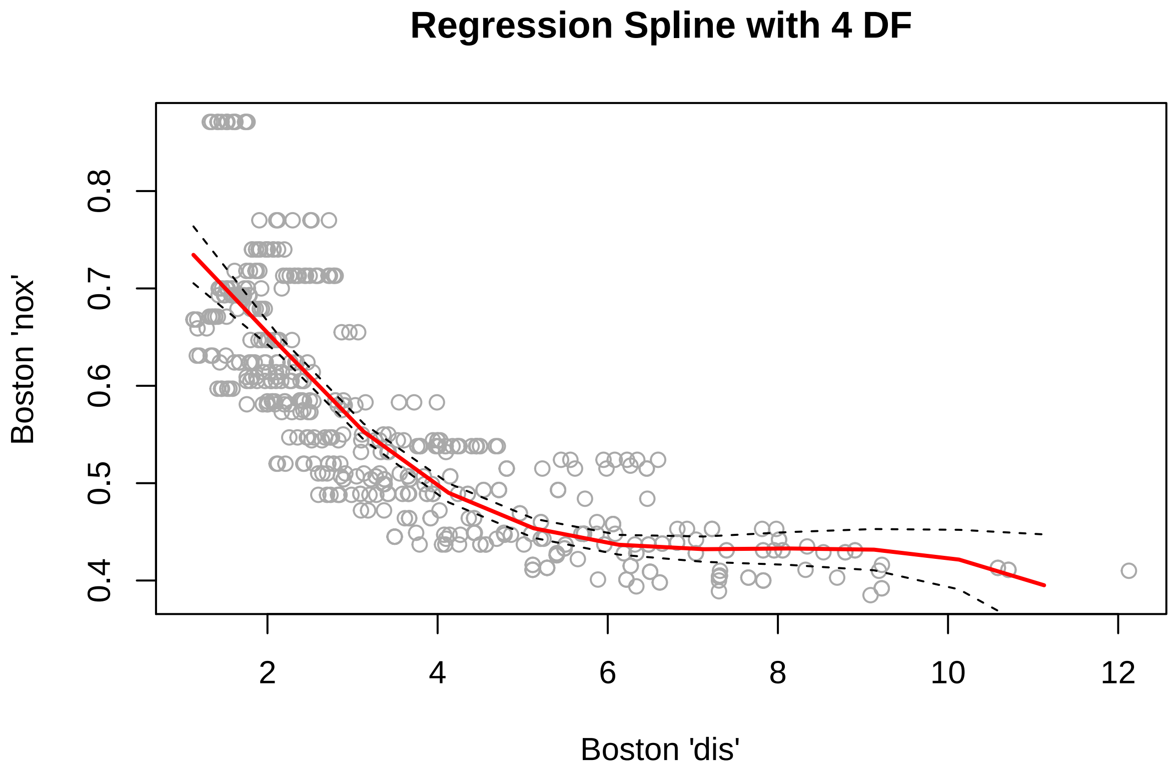
(c) Perform cross-validation or another approach to select the optimal degree for the polynomial, and explain your results.

Here we are going to take a look at the prediction errors through cross validation in order to determine the optimal degree for out polynomial. We can see that the prediction error decreases to a minimum at a 3rd degree polynomial. Then as the degrees increase, we start to see the prediction error increase in value as well as variance. This is due to over fitting the data set and also from the high variance seen in the tails of the models as the polynomial degree gets to be quite high.



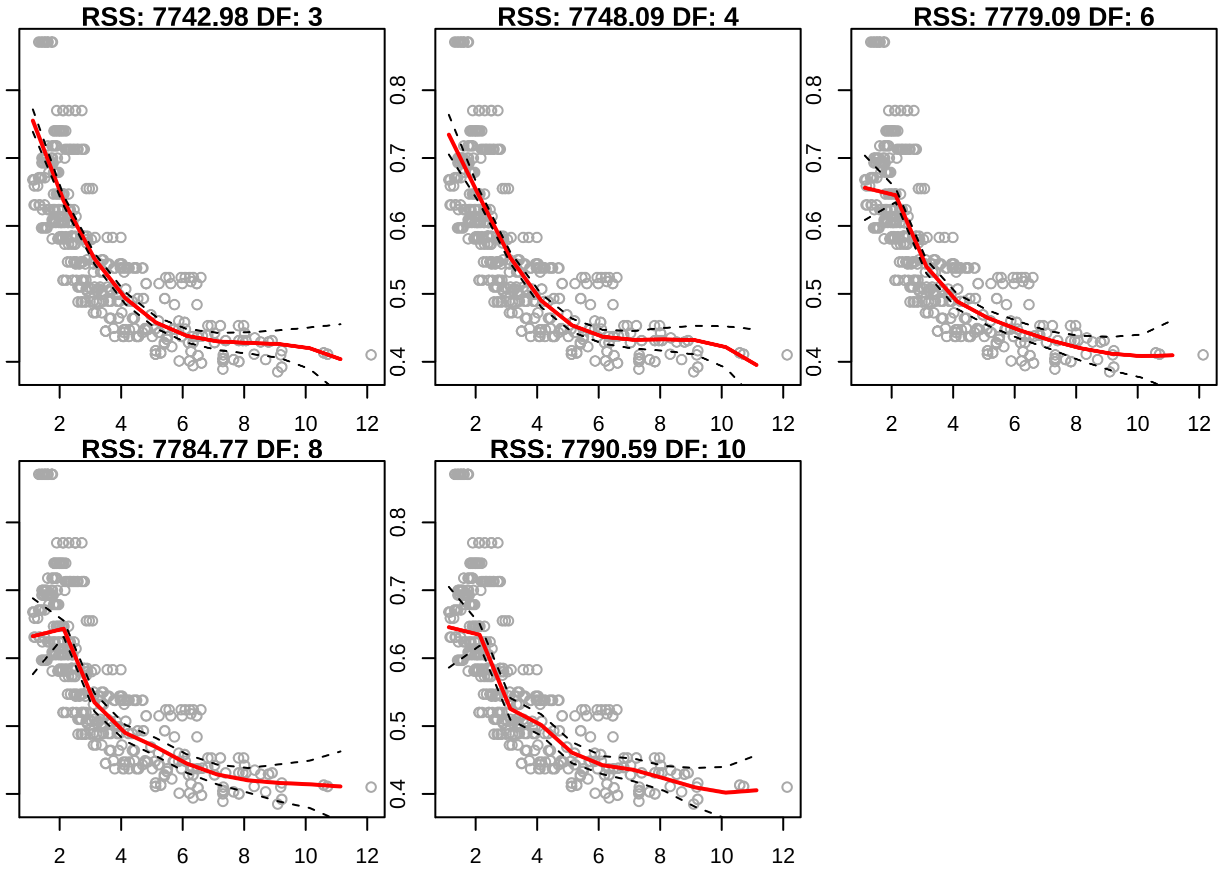
(d) Use the bs() function to fit a regression spline to predict nox using dis. Report the output for the fit using four degrees of freedom. How did you choose the knots? Plot the resulting fit.

I followed the coding ideas found on pages 293-294 of the textbook (Gareth, 2013). The knots are chosen for us by the bs() function since we explicitly stated the degrees of freedom that we want in our model. Below we can see the resulting fit of the regression spline with 4 degrees of freedom.



(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.

The regression splines are an improvement on the higher degree polynomial models we fit above. That is because the regression splines are able to mitigate the variance that is seen in the tails of the model. Also, the optimal model appears to be in the 3-4 degrees of freedom range if we were to solely base that on which model has the lowest 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.

In order to obtain an accurate calculation of the optimal number of degrees of freedom for this model we are going to implement a cross validation technique. By following the coding idea found on page 293 of the textbook (Gareth, 2013), I implemented a smooth.spline() model on the data. The model then determines the optimal smoothness level through cross validation (value for λ in the penalty term in the following smoothing splines regression equation: $\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$). This in turn determines the effective degrees of freedom that will be in the model. As λ approaches infinity from 0, the bias of the model increases and the variance decreases. Henceforth, the effective degrees of freedom decrease from n to 2 (Gareth, 278). Below we can see the model chose 4.13 as the optimal degrees of freedom.

```
## Optimal DF: 4.128915
```

Sources:

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2013). An introduction to statistical learning : with applications in R. New York :Springer

Dr. Saunders class notes