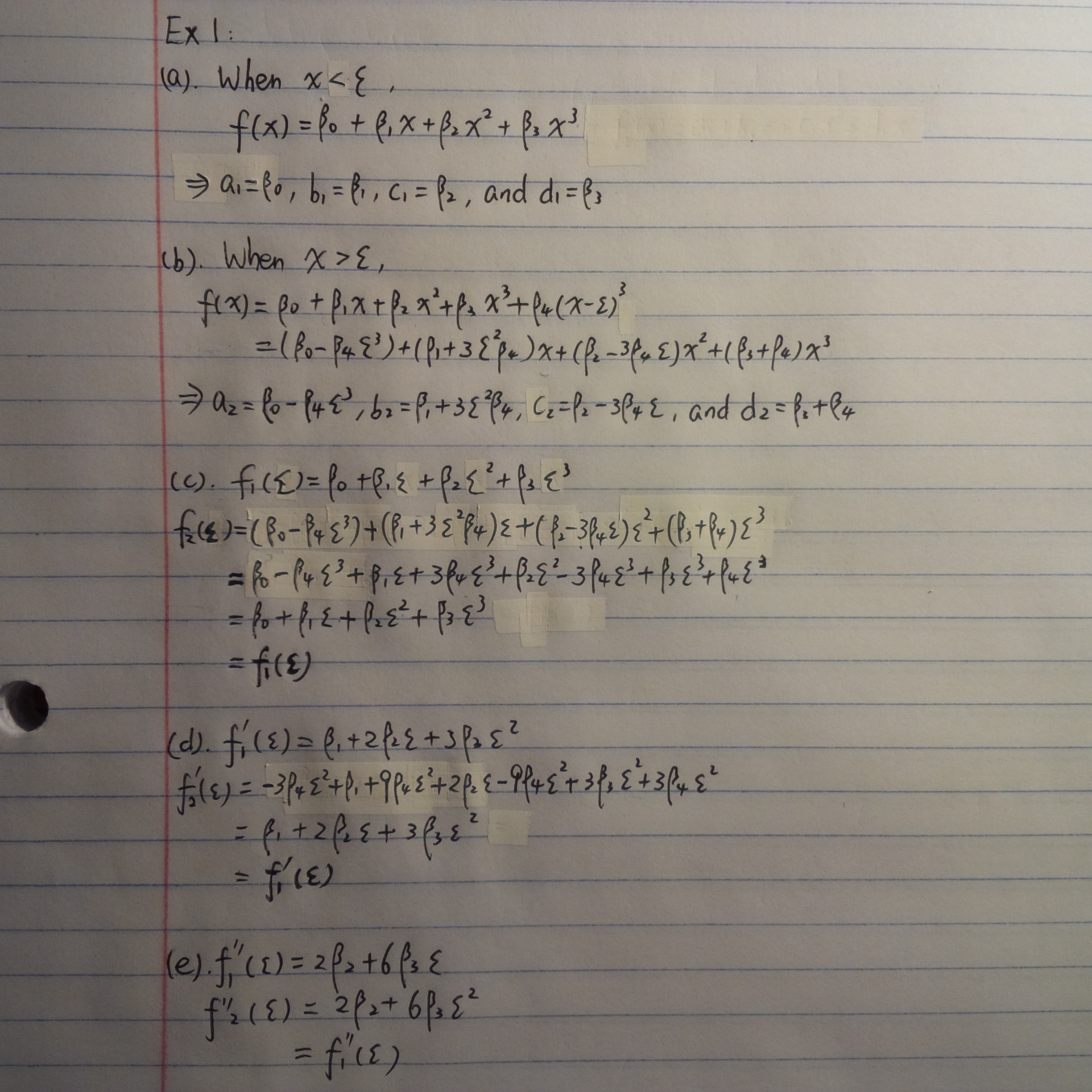
**HW 5 Chapter 7**

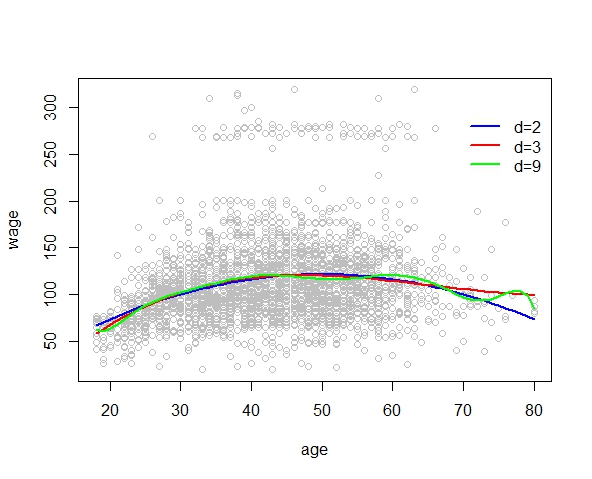
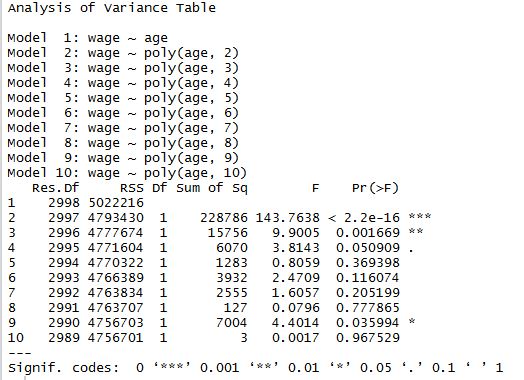
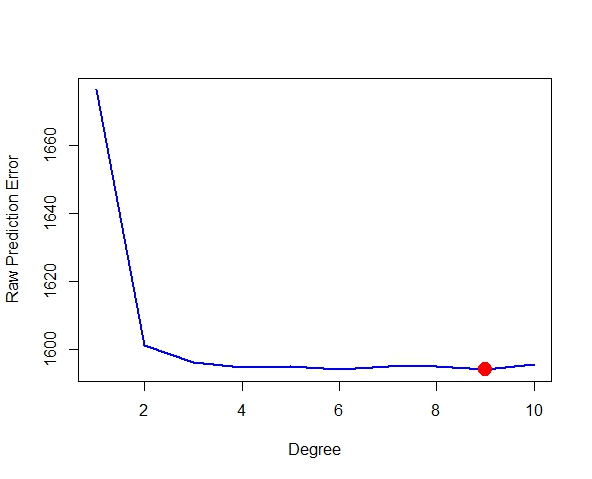
**Luyao Zhang (NetID: lzhang94)**

**Ex 1**

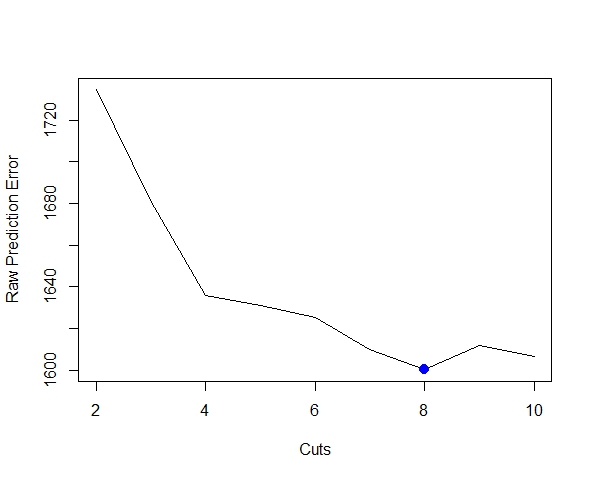


**Ex 6**

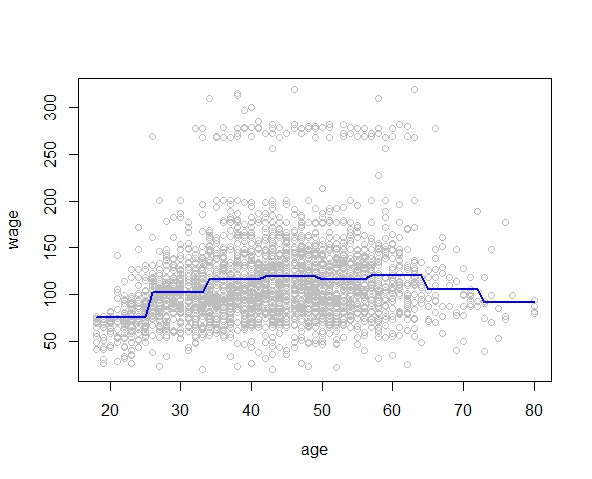
1. I performed a K-fold CV with K = 10. According to the plot below, d = 9 is the optimal. Then, ANOVA was carried out to compare models with different degrees, according to whose results, it seems that a quadratic, cubic or 9-degree polynomial appear to be a reasonable fit. Based on the fit plot, there could be some problem with overfitting when the number of degree is equal to 9.



1. K = 10. It looks like when we there are 8 cuts, the error reaches its minimum.

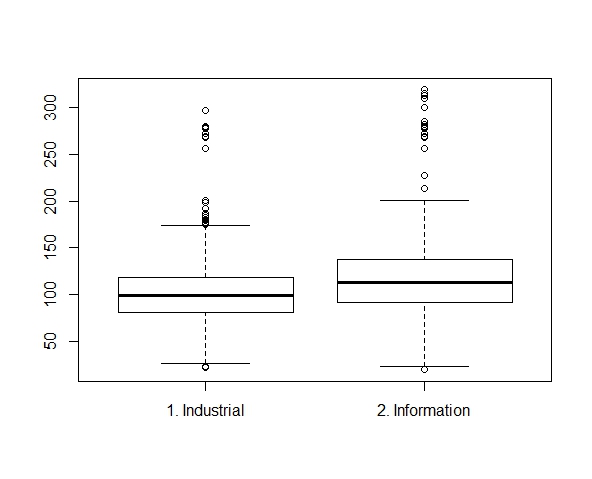
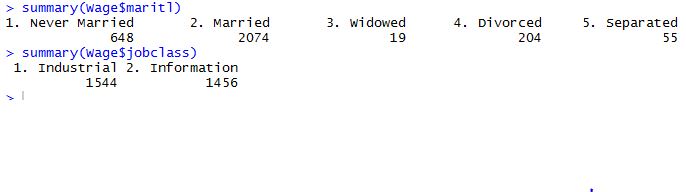


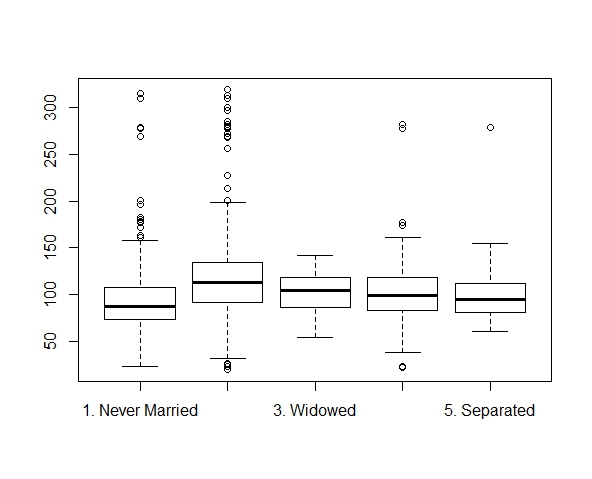
The fit plot is as below.



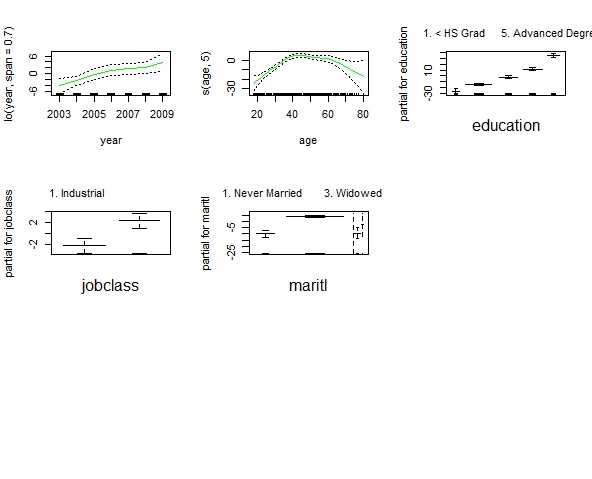
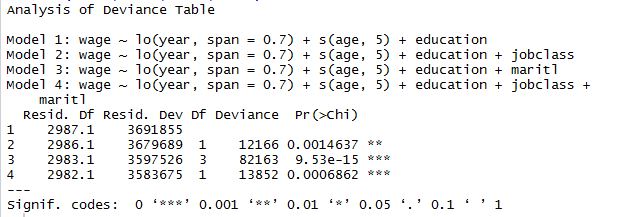
**Ex 7**

1. The summaries of variables “maritl” and “jobclass” are as below. Looks like information jobs have higher pay than industrial, and married couples are paid more compared with people of other marital status.



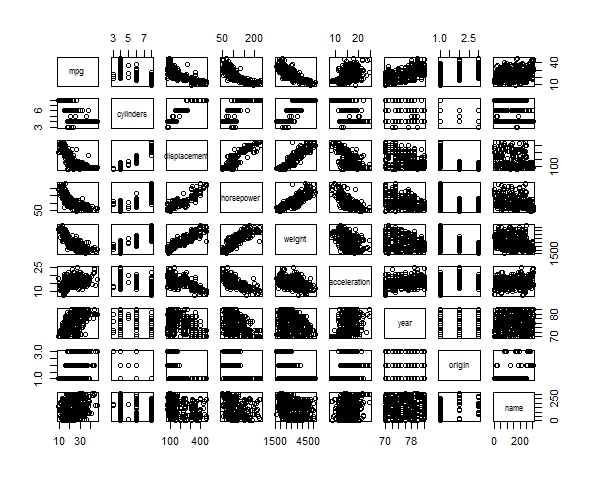


1. GAM was used to predict “wage” using predictors “year”, “age”, “education”, “maritl”, and “jobclass” with natural spline functions. Given the results below, model 4 seems to be the best one.

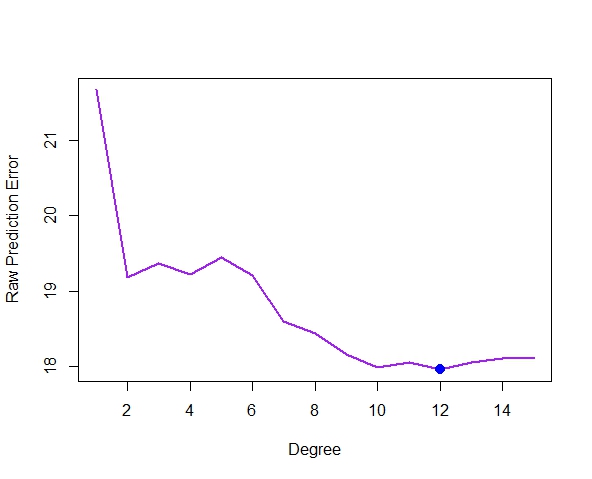


**Ex 8**

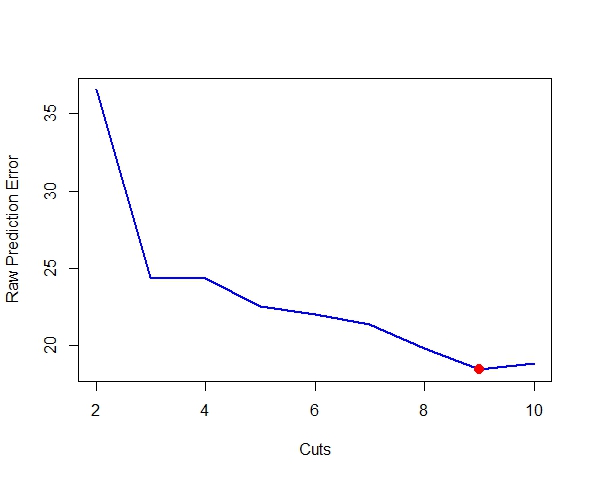
1. First plots by pairs of variables were obtained. According to the plots below, “mpg” is negatively related to “cylinders”, “displacement”, “horsepower”, and “weight”. I chose “displacement” as the predictor for the investigation of non-linear relationship.



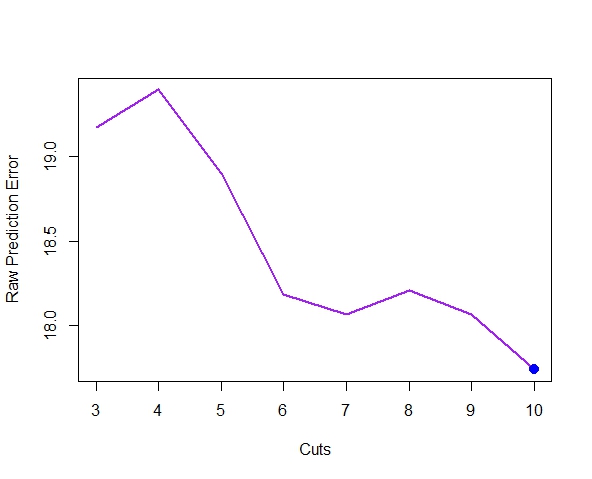
According to the plot below, d = 12 is the optimal with the lowest error.



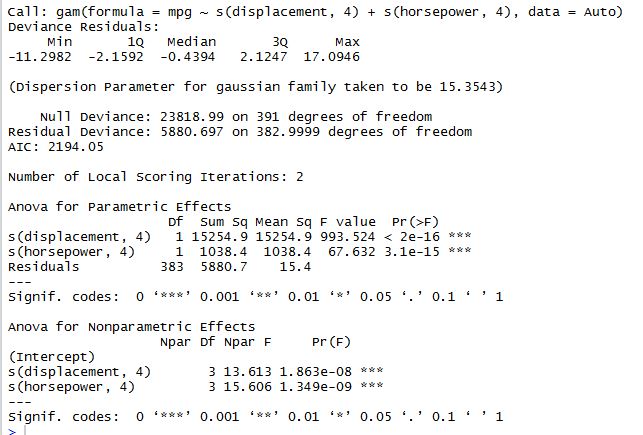
Step functions showed that the error is the smallest when there are 9 cuts.



The spline functions returned degrees of freedom equal to 10 when error is the smallest.

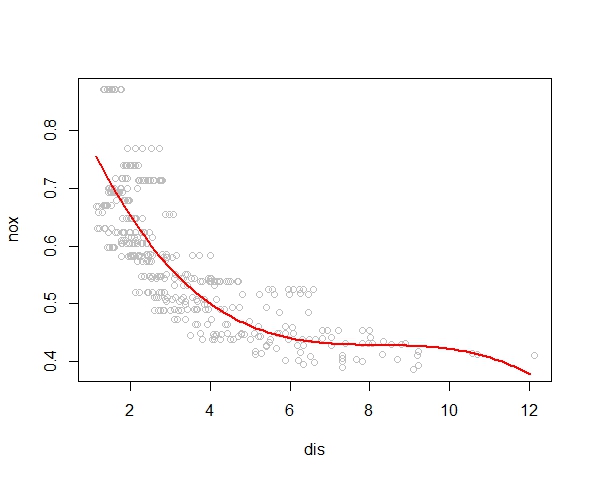
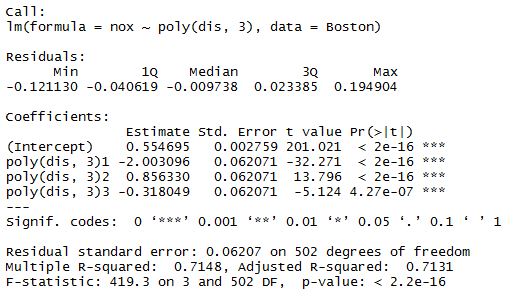


Lastly, GAM was used.

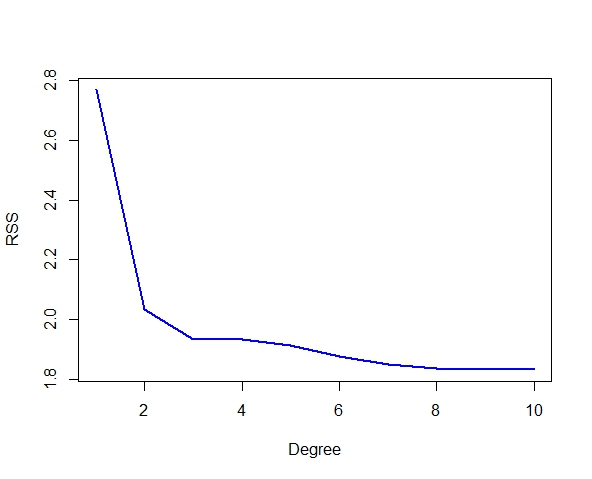


**Ex 9**

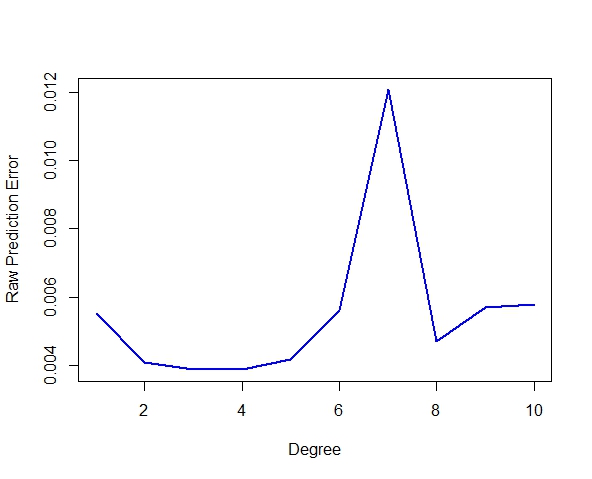
1. The regression output and plot are as below. All polynomial terms are significant at



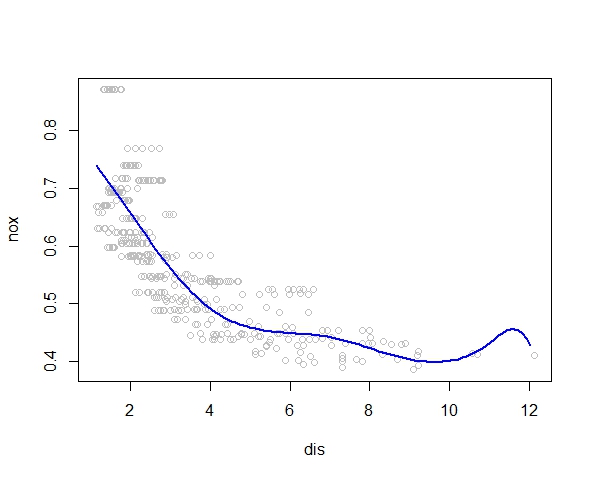
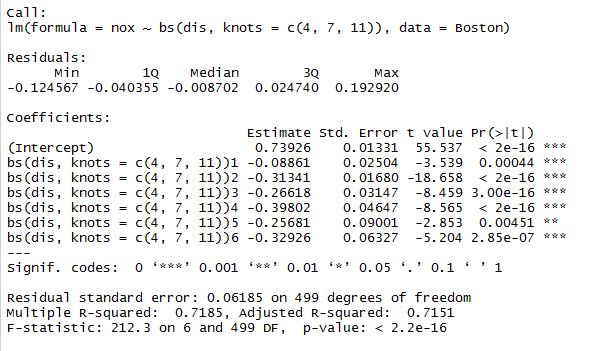
1. According to the plot below, RSS drops as degree goes up, and RSS is the minimum when degree is equal to 10.



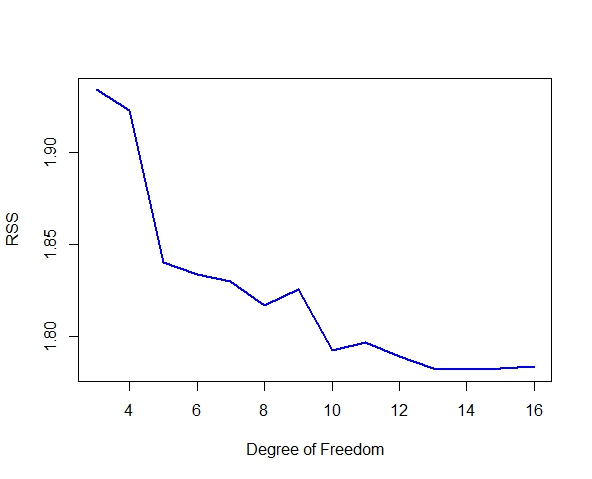
1. Cross-validation is carried out, and the optimal number of degree is either 3 or 4 (probably 4).



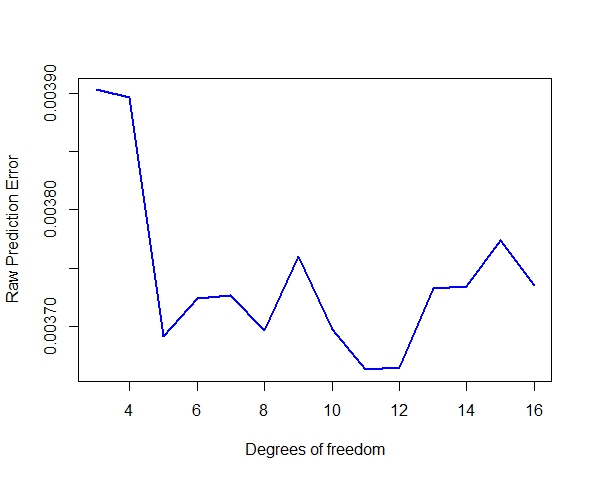
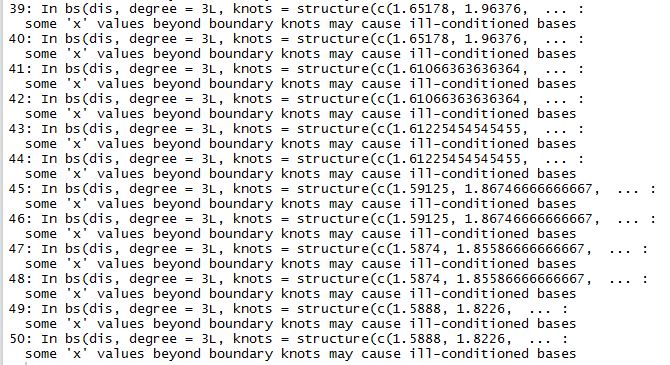
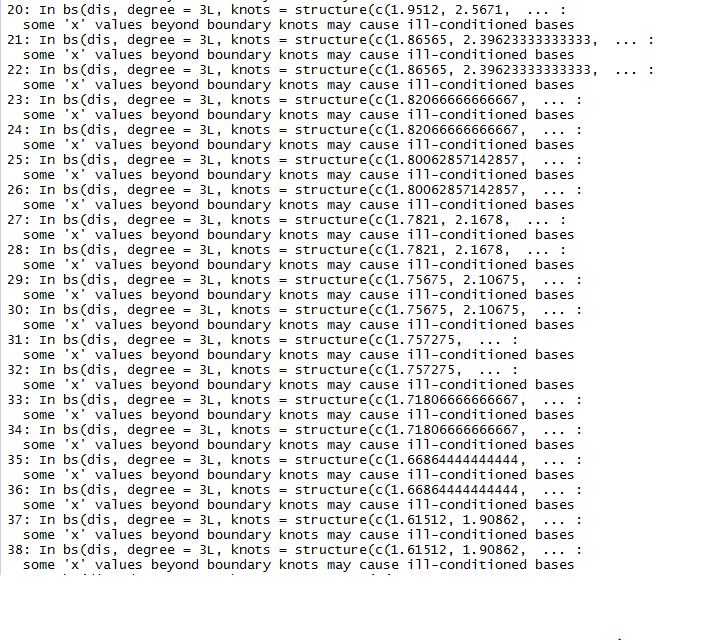
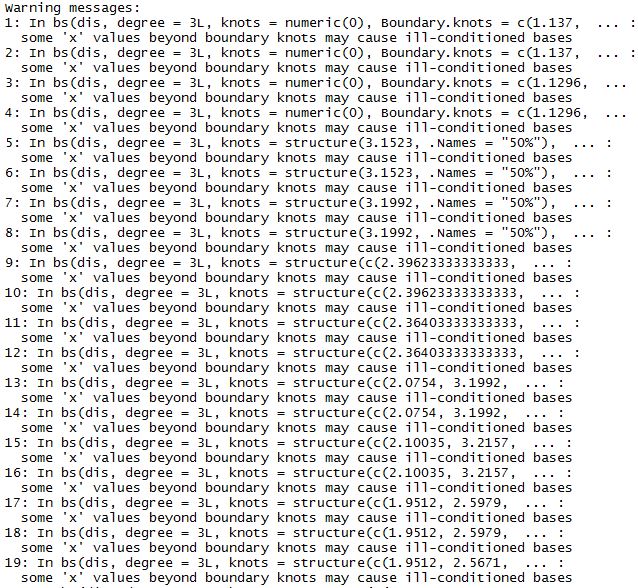
1. Using 4 degrees of freedom, the output and the plot are below. The knots were at 4, 7, and 11. All terms of the spline function are significant. The spline fits well until “dis” values becomes extremely large.



1. The plot is shown below. Looks like the RSS is the minimum with 14 degrees of freedom. After 14, the RSS increases slightly.

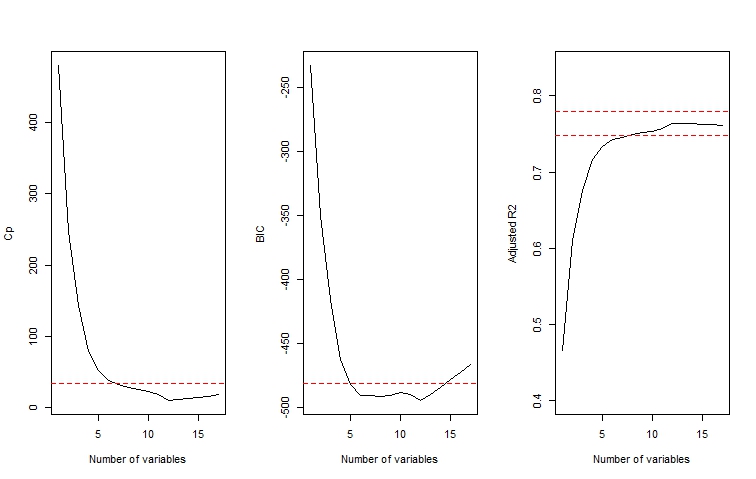


1. A cross-validation was carried out, and 50 warning messages were returned. The error changes more wildly this time. Looks like the error is the minimum at 11 degrees of freedom.

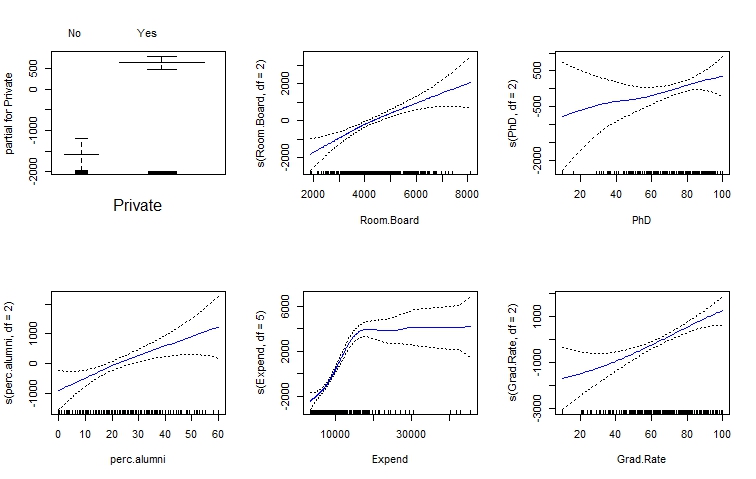


**Ex 10**

1. The model and the plot are as below. According to Cp, BIC, and adjusted R^2, when scores are within 0.2 s.d. of optimum, a subset of 6 is the smallest, and the predictor included are “PrivateYes”, “Room.Board”, “PhD”, “perc.alumni”, “Expend”, and “Grad.Rate”.



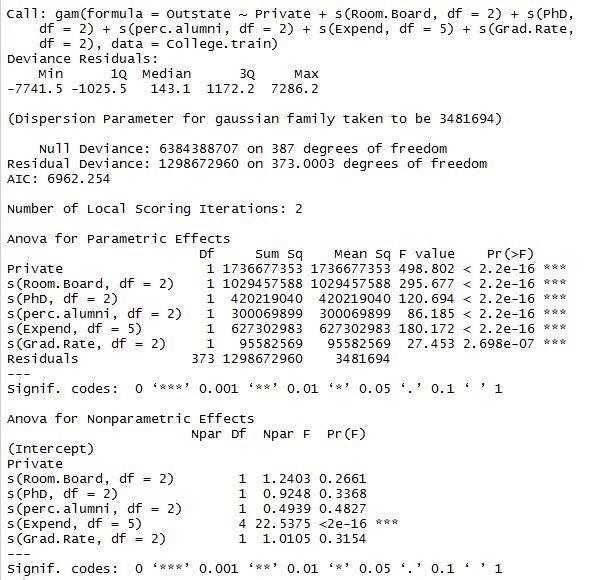
1. The results are as below after fitting the GAM with predictors from (a). Generally speaking, each of the 6 predictors is positively associated with out-of-state tuition when the other 5 are held constant.



1. The test error is 3543901 and the test R^2 is 0.78, which means the 6 predictor altogether explain 78% of the variance in out-of-state tuition, which is quite a lot.

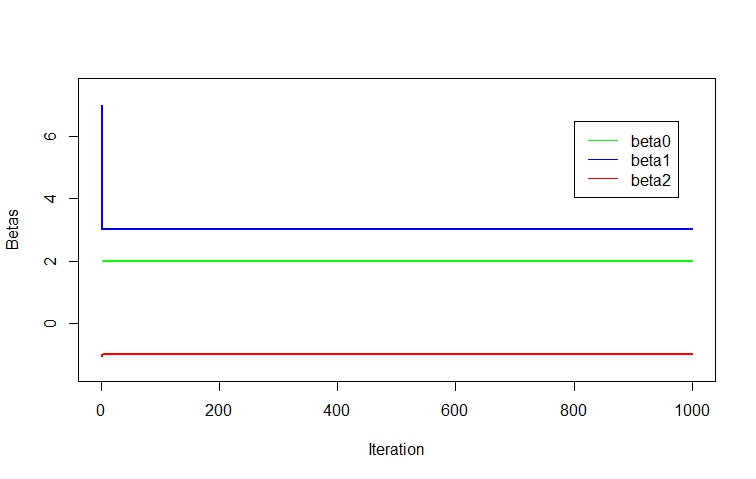


1. According to the *p*-value under “ANOVA for Nonparametric Effects”, “Expend” shows a non-linear relationship with the response (*p*-value < 0.01).



**Ex 11**

1. The equation used to create Xs and Ys is: Y = 2 + 3\*X1 - X2.
2. The initial value of beta1 was set to be equal to 7.
3. **(d) (e)** The plot is as below. It’s very clear according to the plot that the coefficients after very few iterations reached the same values as least square (beta0 = 2, beta1 = 3, and beta2 = -1).



1. The plot below showed that the multiple regression obtained exactly the same results as backfitting.

