Andrew Hillard

Data 3

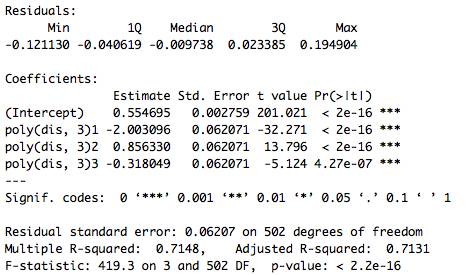
Homework 4

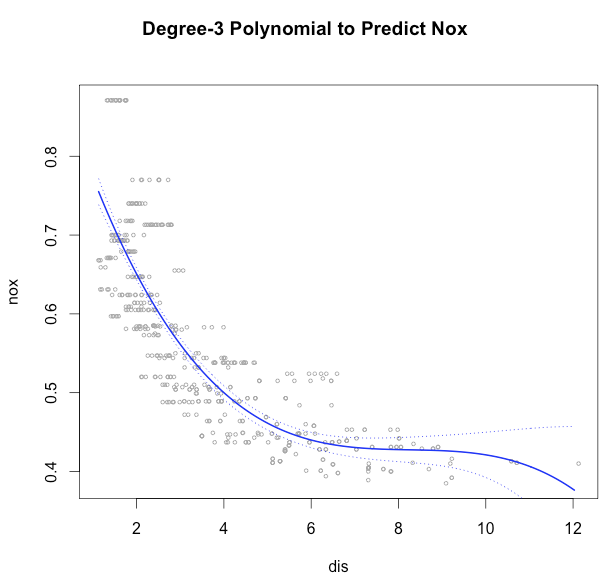
**Problem 1**

See attached notebook paper.

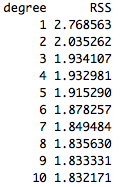
**Problem 2**

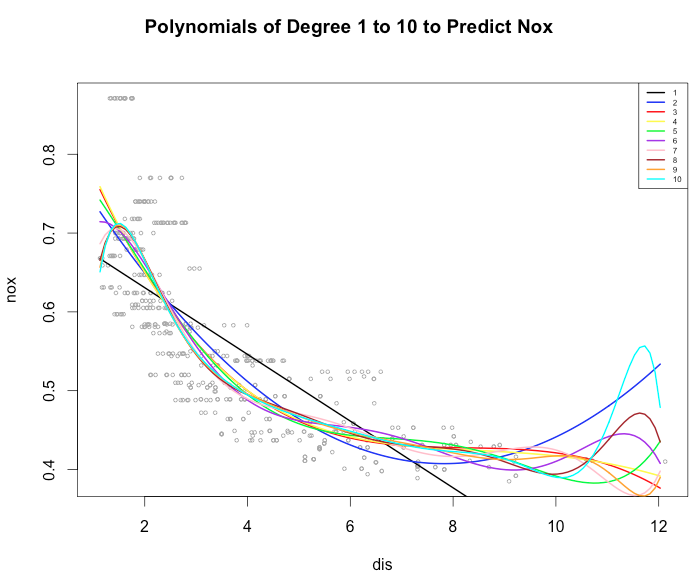
a. The regression output and plot is reported below.



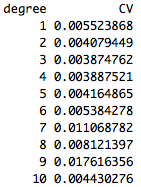


b.

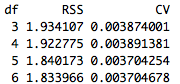


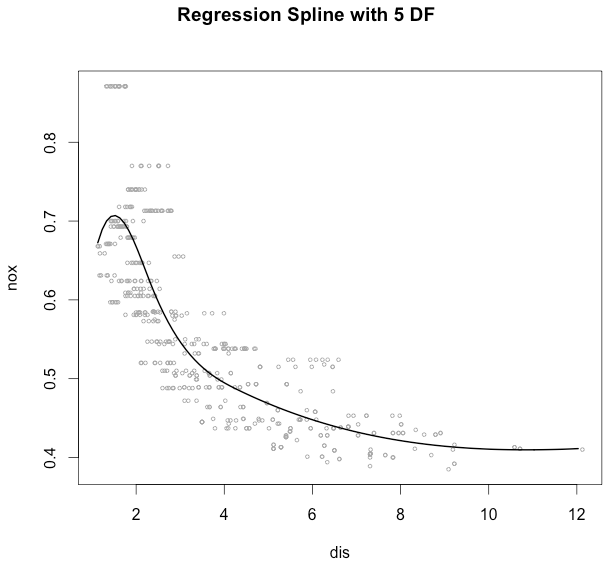


c. The optimal polynomial is the cubic polynomial. The cubic polynomial, degree of 3, had the smallest CV value for 10 fold cross validation.

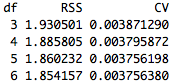


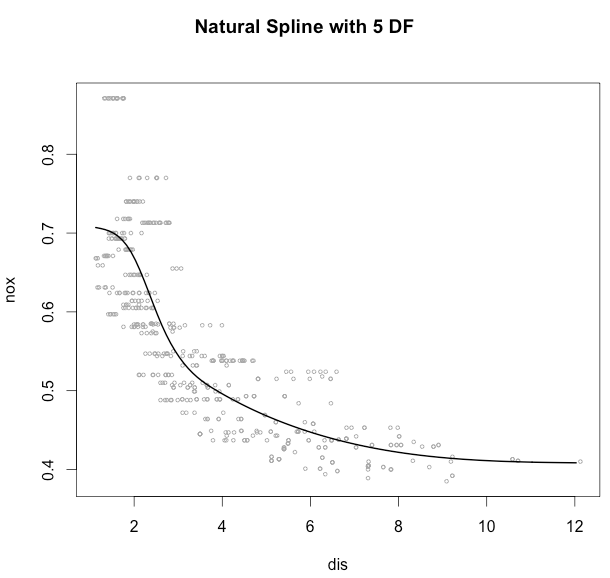
d. The best regression spline had 5 degrees of freedom, an overall RSS of 1.840173, and a 10 fold cross validation value of 0.003704254, which was the smallest CV value for the different tested degrees of freedom. The plot of the best model, degrees of freedom 5, is given below.



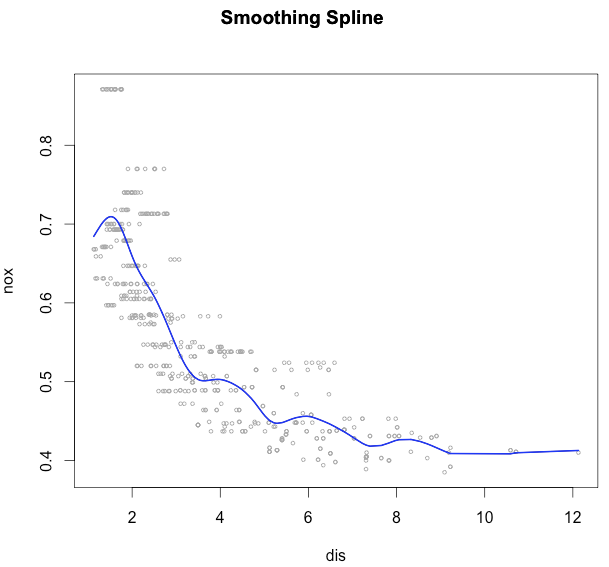


e. The best natural spline also had 5 degrees of freedom, an overall RSS of 1.860232, and a 10 fold cross validation value of 0.003756198. The plot of the best natural splines is given below.

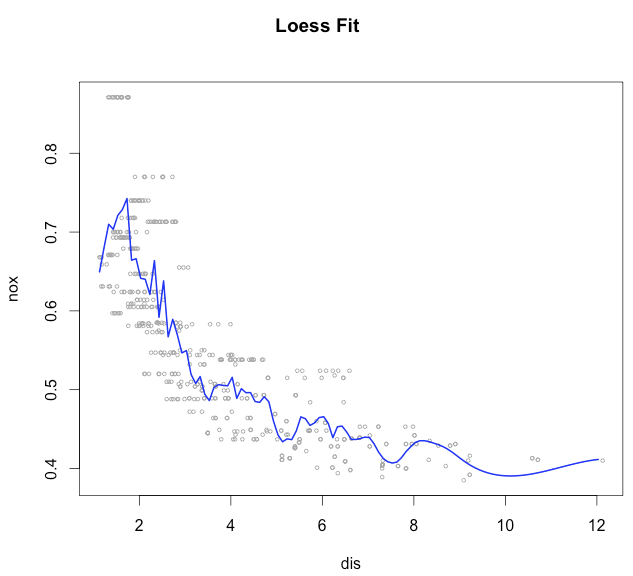




f. The smoothing spline is fit to the data with degrees of freedom 15.42984, chosen through cross validation using the smooth.spline function with cv=TRUE. The plot is given below.



g. I chose a span of 0.0605 using the function loess.wrapper {bisoreg}. The loess.wrapper performs 10 fold cross validation over a range of span values and finds the span that minimizes the cross validation value. I then used the loess function to fit the data with a span of 0.0605. The plot of the loess model is given below.

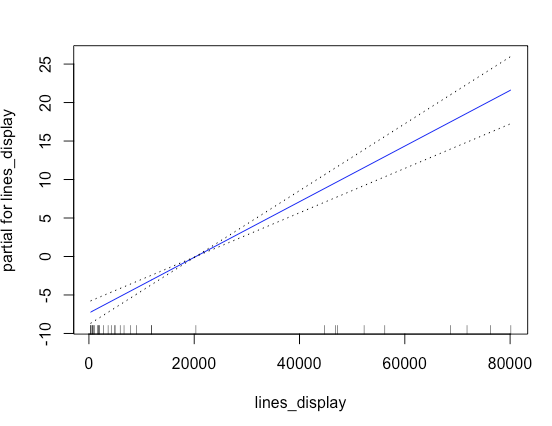


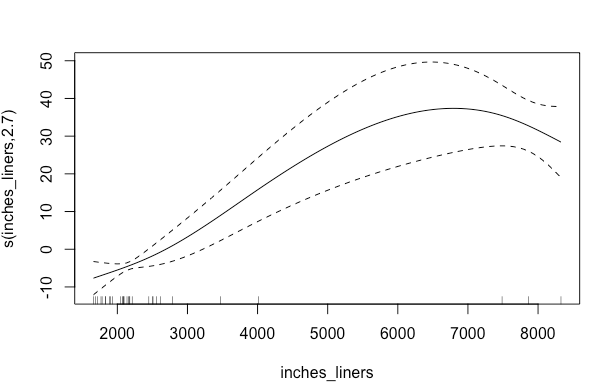
**Problem 3**

I first used the gamclass package to perform 5-fold cross validation on every possible linear regression model using the three predictors and assuming no interactions. The best linear regression model had two predictors: inches liners and lines display. I then built a number of linear and non-linear (smoothing and natural spline) models with each predictor by itself. Each predictor had significantly lower cross-validation values when treated as non-linear as opposed to linear. For inches liner by itself, the regular regression model had a CV value of 143.92, the natural spline model with df=3 had a CV value of 75.1079, and the smoothing spline model with df=3 had a CV value equal to 79.4221. For lines display by itself, the regular regression model had a CV value of 128.63, the natural spline model with df=4 had a CV value of 49.1276, and the smoothing spline model with df=5 had a CV value equal to 36.1977.This is evidence that there is a non-linear relationship between inches liners, lines display, and the response variable.

All combinations of a two variable model, applying a smoothing spline to one or both of the predictors, were then tested. The degrees of freedom for the different models were tweaked until the lowest cross validation value resulted. The GAM model with the lowest cross validation value was when a smoothing spline was applied to inches liners with df=4, and lines display was treated as a normal linear regression predictor. The 5 fold cross validation for this model was 40.1059. The partial additive plots for this model are given below.

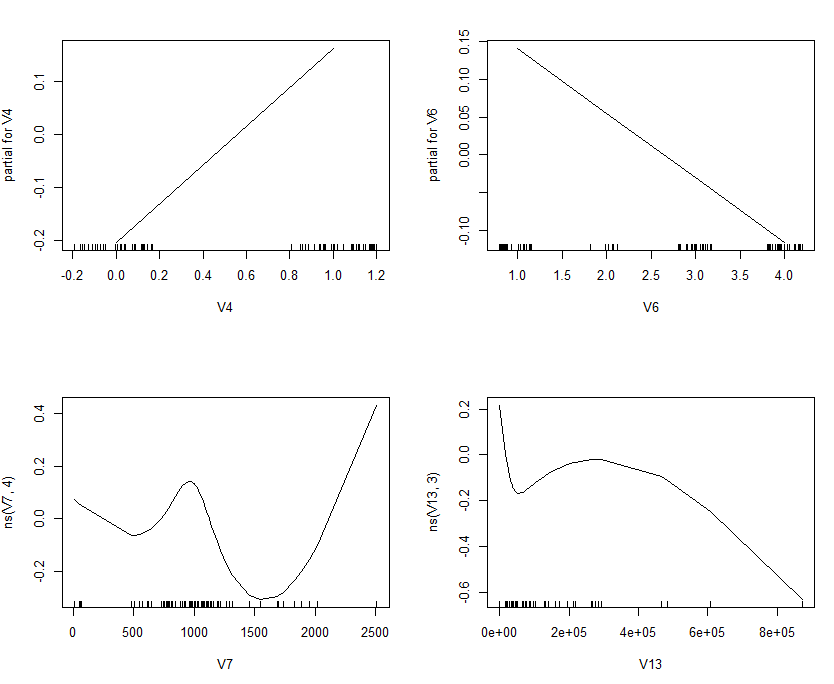
This model does fit better than the linear regression model in Homework 2, which had a cross validation value of 43.4543.





**Problem 4**

I first used lasso regularization on the full logistic regression model to perform variable selection. The only variables that were not set to zero were variables 4, 6, 7, and 13. These numbers correspond to the column number in the dataset. Variables 4 and 6 were categorical and were not modified in the GAM model since regression splines can only be applied to continuous variables. I then ran a base model (levee failure ~ V4+V6+V7+V13) with natural splines and smoothing splines applied to V7, changing the degrees of freedom from 1 to 10 by increments of 1. I also ran the same base model with a loess regression applied to V7, changing the span from 0.2 to 0.8 by increments of 0.1. I applied the same procedure to V13. I then compared the UBRE score (a form of generalized cross validation) for every iteration described above. The best model was levee failure ~ V4+V6+V7+ns(V13, df=3), which had an UBRE score of 0.206493. Then using the model, levee failure ~ V4+V6+V7+ns(V13, df=3), as my new base model, I applied different natural splines and smoothing splines on V7 by changing the degrees of freedom from 1 to 10 by increments of 1. I then checked the UBRE scores on all of these models and identified the following model as the best model: levee failure ~ V4+V6+ns(V7, 4) +ns(V13, df=3). This best model had an UBRE score of 0.1888495. The partial plots for the different variables are given below.



**Problem 5**

Because there were 30 possible predictors, I first used best subsets to perform variable selection. I ran best subsets using the package, regsubsets. This gave me a best model for models of different size. I then performed 5 fold cross validation on all these best models using the procedure from page 250 in JWHT. The model that had the lowest CV error had three variables: sex, Mjob, and failures. Sex and Mjob were categorical so I did not apply a spline or loess regression on these predictors. I did, however, use the base model (G3 ~ sex+Mjob+failures) and apply a natural spline to failures, changing the degrees of freedom from 1 to 10 by increments of 1. I also applied loess regression to failures, changing the span from 0.2 to 0.8 by increments of 0.1. I then compared the generalized cross validation value for each variation on the base model and found that the best model, the model that minimized the CV value, was the GAM model: G3 ~ sex+Mjob+ns(failures, 3). This GAM model had a generalized CV value of 17.80095. The best model from homework 3 was the PCR model with a CV value of 16.60168. Therefore, the GAM model did not perform as well. However, it should be noted that it performed better than the PLS, Lasso, and Ridge Regression model from homework 3.