STA721

Homework 17

The Wage data in `library(ISLR); data(Wage)` includes a set of variables thought to be associated with wages (age, education level, year, etc) for a group of males from the Atlantic region of the US.

- Split the data into a training set and validation set (say 80% training and 20% for validation). (keep your random seed in case you need to reproduce). Using the training data, explore adding additional variables to the generalized additive model for wage to construct a model for the data. (see the documentation for `gam` or `bam` in `library(mgcv)` for options and details on other smoothing splines/priors, i.e. you may wish to fix k to prevent some over-fitting). Narrow down your models to what you might consider your top 5.

- Using several of your "best" models, calculate the likelihood ratio for comparing log(wage) to wage as the best response for normality. (e.g. BoxCox but with just two choices). Which transformation is better? does that agree with residual plots?

- Use your models to predict wages on the validation data. Using boxplots of predicted residuals and predicted MSE, which model(s) are best for predicting Wages? Does this agree with your findings on the training data?

- Using R Markdown or KnitR, provide a write up of up to 3 pages with key figures that introduces the problem, your methods and findings (i.e. key interactions, nonlinearities or other relationships with figures to support key findings) that describe the relationship of wages with the other variables. You may include all of your R code as an appendix but code should not appear in the main document.