1. In R file
2. It would appear as though those who are currently married earn the most out of all marital classifications, while those who are either widowed or separated make the least. Divorced people are also quite unstable in terms of possible wages, and those who have never married forms a sort of “middle ground”. As well, those who work in the information sector appear to be higher earners than those who work in the industrial sector. If we examine race, we may observe that on average, Asians earn the most, followed by whites, then blacks and others. Whites have the most high outliers, however, and the highest black earners are the highest earners in the entire set. According to deviance analysis, marital status and job class are both very significant to the dataset, while race appears to be less significant. However, using race and job class together, or race and marital status together, as predictors seems to be quite effective as well. This suggests a possible synergistic effect between these two pairs of variables.
3. From the analysis done in the R file, it can be concluded that at least the three predictors of displacement, weight, and horsepower have a non-linear relationship with the response mpg (mileage per gallon).
4. Boston
   1. The coefficients of the cubic polynomial fit are 0.934, -0.182, 0.022, and -0.0009, for the intercept, linear, quadratic, and cubic terms, respectively. All of the terms have small p-values, with the linear and intercept having a near 0 p-value, the quadratic term having a p-value of 3.43e^-13, and the cubic term having a p-value of 4.72e^-7. The adjusted R^2 is 0.7131, which is a decent amount of variability explained by this model. Therefore, all of the terms here are significant.
   2. Residual sum of squares (number =degree):
      1. 2.77
      2. 2.04
      3. 1.93
      4. 1.93
      5. 1.92
      6. 1.88
      7. 1.85
      8. 1.84
      9. 1.83
      10. 1.83
   3. According to ANOVA testing, a cubic polynomial is sufficient to fit the data. The quartic and the quintic fit both have very high p-values when compared to the cubic.
   4. The spline with 4 degrees of freedom has an R^2 value of 0.7164, comparable to the cubic polynomial from earlier. The p-values of the intercept, quadratic, cubic, and knot term all appear to be significant. The linear term’s p-value is very high, making it appear to be insignificant. The number of knots was chosen simply by specifying the degrees of freedom to be 4 using the df argument in the bs() function.
   5. In general, as more degrees of freedom are added, the lower the residuals become. This is because the model is allowed to vary more the more degrees of freedom it is allowed, making it more flexible and perform better on the training set. However, this may not apply to other sets used to test it.
   6. According to cross-validation, the best spline to fit the relationship between nox and dis is a cubic with 10 degrees of freedom.
5. College
   1. In R file
   2. The GAM used the 5 variables of Private, Room.Board, PhD, perc.alumni, and Expend. Overall, it seems like the most impactful of the 5 is Private, and the least impactful is Expend, based on the coefficient estimates.
   3. With the 5 variables chosen by forward stepwise selection, the R^2 value resulting from evaluating the 5-variable model on the test set was 0.7415, which explains a decent amount of the variability in the relationship between the out of state tuition of students and its predictors.
   4. Based on ANOVA testing from the summaries of the linear GAM and spline GAM, Room.board, PhD, and perc.alumni have large p-values in the spline GAM and small p-values in the linear GAM, suggesting that these three variables all have a linear relationship with the response. Conversely, the p-value of Expend is nearly zero in both the linear and spline GAM, indicating that a non-linear relationship is highly likely between Expend and the response.
6. Backfitting investigation
   1. In R file
   2. In R file
   3. In R file
   4. In R file
   5. In R file
   6. In R file
   7. On the data set I generated, upon observing the resulting values for the coefficients upon each iteration, it took about 5 backfitting iterations in order to obtain an accurate estimate of the multiple regression coefficient estimates, or after “convergence”.
7. With p=100, the number of backfitting iterations required to achieve a “good” approximation to the multiple regression coefficient estimates varies based on the starting estimates for the coefficients. However, we can see that, based on the R plot of the coefficient estimates provided after 1000 backfitting iterations, the backfitting estimates are still wildly off of the multiple regression estimates, with severely decreasing slope in the estimates per iteration. Therefore, it is safe to conclude that a nearly-infinite amount of backfitting iterations are required to obtain a “good” approximation for the multiple regression coefficients, as the backfitting estimates appear to asymptotically approach the multiple regression coefficients with a slope of perpetually decreasing magnitude.