**Problem 1.**

Use the ***sat*** dataset.

1. Run a linear model with response *total* and predictors *expend*, *ratio*, *salary*, and *takers*. What is the regression coefficient for *ratio*?
2. Create the usual linear model plots. Looking at the normal qq plot, the normality could be an issue. Why? Run a test to verify whether there really is a problem.
3. Looking at the residuals vs. leverage plot, is there a state that we should investigate further?
4. Again looking at the residuals vs. leverage plot, approximately how many high leverage points can you identify? Confirm this using code and identify the states with these high leverage points.
5. Create four plots of the residuals versus each of the predictors. Indicate the points for the state you found in c. Does it seem like the state has unusual values for the predictors?
6. Identify any potential issues with any of the plots produced in e).
7. Give the four states with the highest residuals (possible outliers) in order, highest last.
8. Give the four states with the highest studentized residuals in order, highest last.
9. Why is there a difference between g and h?
10. Create both a partial regression plot with a regression line, and a partial residual plot for *takers*. Which one has more data points in the center (a meaningless question just to establish that you did the exercise)?
11. Perform an eigenvalue decomposition. Do we have collinearity in the predictors *expend*, *ratio*, *salary*, and *takers*? If so, how many eigenvalues point to this?
12. Which two of the predictors have the highest correlation?
13. Which predictors (if any) have variance inflation factors that indicate a problem, and which ones need more investigation?
14. Judging by the ***adjusted R2*** (You will learn more about this in machine learning; it is a better indicator than the regular R2 because it penalizes additional predictors), pick the best model among the original and the two where each one of the two problematic predictors has been removed.
15. Using the best model so far (the one in n) determine if the model can benefit from a transformation of the response. Using the adjusted R2 again, determine which model is best.
16. With the model from n, create orthogonal polynomial predictors for takers. Determine which order polynomial would be best. Is this model better than the two in n and o?
17. For your best model so far, check out the plots and compare to the originals. Do they look better now?
18. Now imagine we settled on a model of total vs. salary and takers. Ratio and expend are also in the data set, but we are not using those in the model. Setting the seed to 123, create a data set sat2 that has 20 random values of salary set to NA. Create a third data set sat3 that is a copy of sat2. Use regression to impute the missing values for the salaries predictor in sat3. At this point, sat will have the original values, sat2 the missing values, and sat3 the imputed by regression values. Compare some of the imputed values versus the originals; does it seem like the process did a good job?
19. Compare the summaries of the three models of total vs. salary and takers for the three data sets. Does the imputation by regression do a better job compared to the default (removing the cases)?
20. Using the data set with missing values (sat2), set the seed to 123 and perform multiple imputation. Compare the regression coefficients and standard errors with those obtained in s.