**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*? -3.6242
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.  
   The qq plot indicates a deviation from normality possibly indicating a long-tailed distribution. Running a shapiro test, we find a p-value equal to .4304, alleviating our worries; the data does appear to be normal.
3. Looking at the residuals vs. leverage plot, is there a state that we should investigate further?  
   Yes, Utah looks to be a potential high influence point.
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.  
   About 4 points have leverage > 2\*p/n (=.2). They are New Jersey, Connecticut, California, and Utah.
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?  
   Yes.
6. Identify any potential issues with any of the plots produced in e).  
   Residuals vs. Expenditure looks somewhat cone-shaped; variance seems to decrease as Expenditure increases.  
   Residuals vs. avg. pupil/teacher ratio seems to have two high leverage points.  
   Residuals versus takers could be quadratic rather than linear.
7. Give the four states with the highest residuals (possible outliers) in order, highest last.  
   Utah, New Hampshire, North Dakota, West Virginia
8. Give the four states with the highest studentized residuals in order, highest last.

New Hampshire, North Dakota, Utah, West Virginia

1. Why is there a difference between g and h?

The studentized residual of a case is based on its residual w.r.t. the linear model without that case.

1. 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)? The partial regression plot.
2. Perform an eigenvalue decomposition. Do we have collinearity in the predictors *expend*, *ratio*, *salary*, and *takers*? If so, how many eigenvalues point to this? Yes we do, we have one point where .
3. Which two of the predictors have the highest correlation?  
   Salary and expenditure.
4. Which predictors (if any) have variance inflation factors that indicate a problem, and which ones need more investigation?  
   We need to investigate expenditure and salary. None are outright problems (although they come close).
5. 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.  
   The model with expenditure removed (leaving ratio, salary, takers) has the highest adjusted R2.
6. 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.  
   The model with the response to the power of -2 has a better adjusted R2 (.8245).
7. 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?  
   The third order polynomial is insignificant, but the second is not. So we create a model with second order polynomials for takers, which has a higher adjusted R2 (.8769) than n and o, so it appears to be the best one so far.
8. For your best model so far, check out the plots and compare to the originals. Do they look better now?  
   Yes they do; residuals vs. fitted no longer look curved (non-linear) and Utah no longer has high influence. The QQ plot also looks more normal.
9. 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?  
   Yes, it certainly seems to have done a better job than a single imputation using the mean would have done. The high collinearity of salary with expenditure helped a lot.
10. 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)?  
    Yes. The betas are close for all three models, but the standard errors are highest for data set sat2 and lowest for sat3. The adjusted R2 for sat3 is more similar to the one for sat.
11. 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.  
    The regression coefficients are comparable to the ones obtained using regression imputation, but the standard errors for multiple imputation are a bit larger, they are closer to the actual standard errors.