**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.  
     
   qqplot looks like having fatter tails problem. But Shapiro test tells it is normal.
3. Looking at the residuals vs. leverage plot, is there a state that we should investigate further?  
     
   None of the points is over 0.5 or 1. Utah is at ~0.5 and could possibly an Influential point. We need to further Investigate the Situation
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.

By looking at graph it looks we have 2 high leverage points.

Utah and California are High leverage points (Observations) 

1. 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?

We can Observe few observations with expend value more then 8.7

We can Observe 2 observations with ratio value more than or equal 24

We can see Few Observations with salary over 44

we can see the two separate behaviors of Takers below 40 and above 40

1. Identify any potential issues with any of the plots produced in e).

i) Residuals vs Expend  
we can we non constant variance.   
ii) Residuals vs Ratio   
Variance Looks pretty much constant   
iii) Residuals vs Salary   
Non constant variance   
iv) Residuals vs Takers   
Non linearity can be Observed. 

1. Give the four states with the highest residuals (possible outliers) in order, highest last.  
   Following are the points with highest residuals and could possibly be outliers.   
   New Hampshire   
   Utah   
   North Dakota   
   West Virginia
2. Give the four states with the highest studentized residuals in order, highest last.  
   Following are the points with highest studentized residuals   
   New Hampshire   
   North Dakota   
   Utah   
   West Virginia
3. Why is there a difference between g and h?
4. 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)?   
   Partial Regression plot has more points in Center.
5. Perform an eigenvalue decomposition. Do we have collinearity in the predictors *expend*, *ratio*, *salary*, and *takers*? If so, how many eigenvalues point to this?   
   # 2 values are high

## [1] 1.000000 3.392475 20.336746 113.750609

1. Which two of the predictors have the highest correlation?  
     
   expend and salary have high corelation of 0.8698015
2. Which predictors (if any) have variance inflation factors that indicate a problem, and which ones need more investigation?  
     
   expend and salary has vif > 5. Needs Invetigation.
3. 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.  
     
   R Squre and adjusted R Squre both decresese in 2nd model. meaning 1st is better. (However Removing them doesnot make much differnce)
4. 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.

Yes we can benefit from transformations/

New Model is better. Best transformation is power of -2 

1. 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?

4rth Polynomial is Best and is better from previous models.

1. For your best model so far, check out the plots and compare to the originals. Do they look better now?  
   Plots Look Much better now.
2. 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?

Almost did a good job.

1. 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)?  
     
   Model with predicted missing salaries performs slightly better than the others
2. 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.