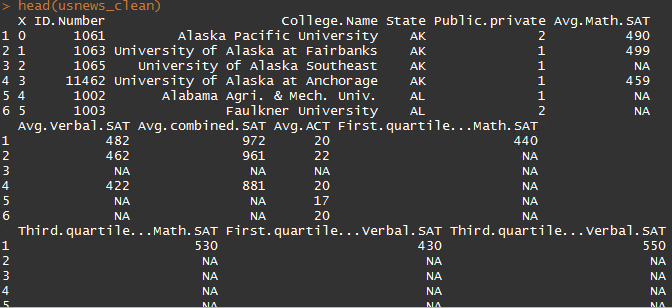
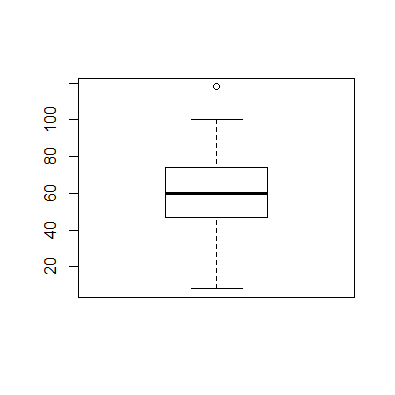
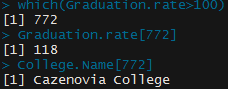
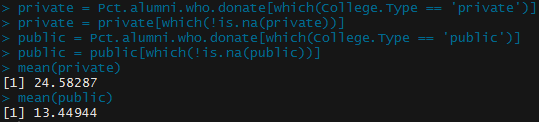
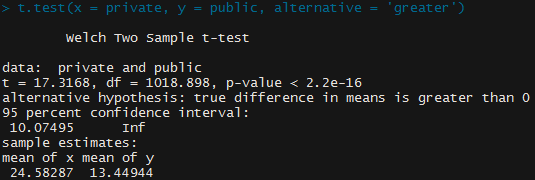
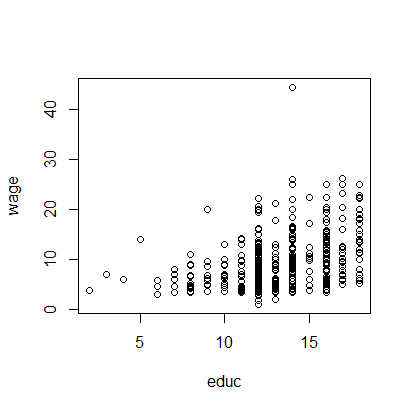
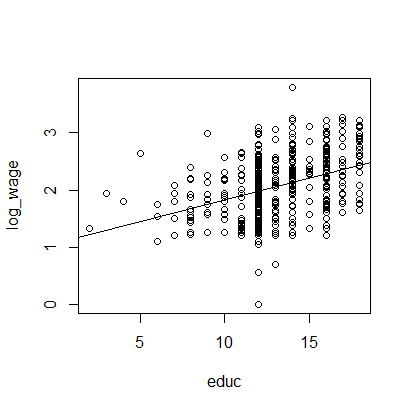
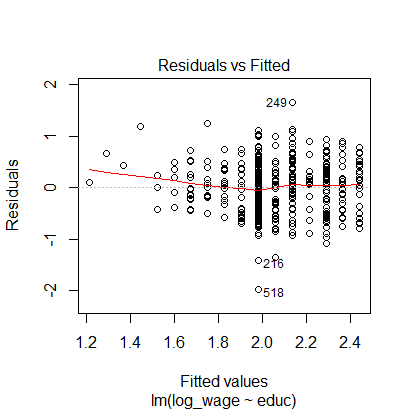
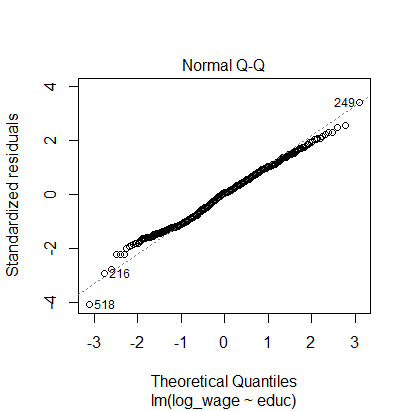
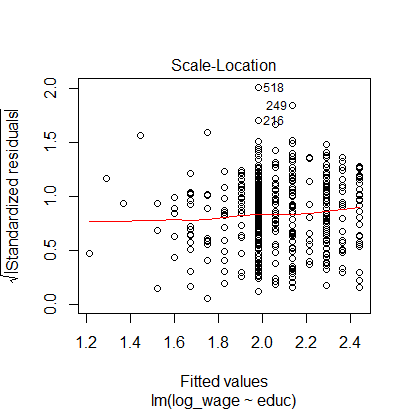
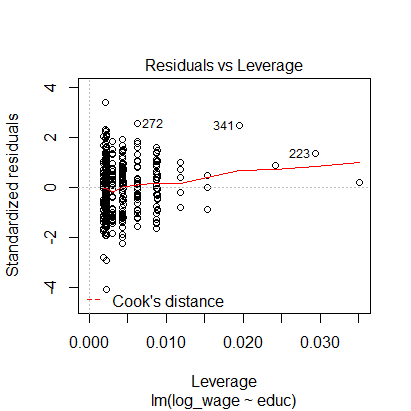
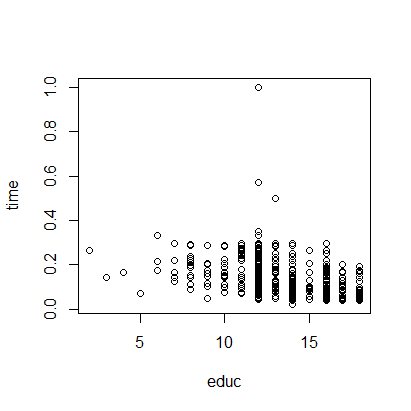
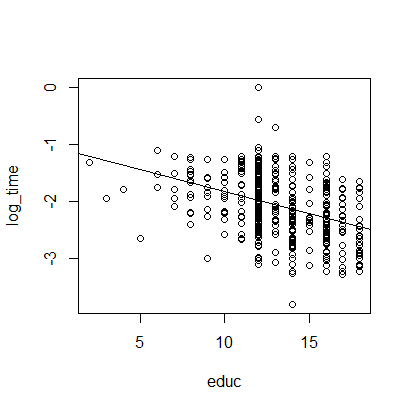
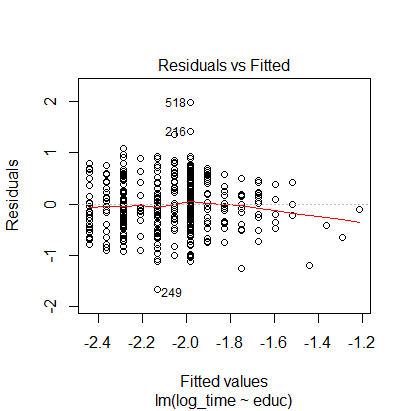
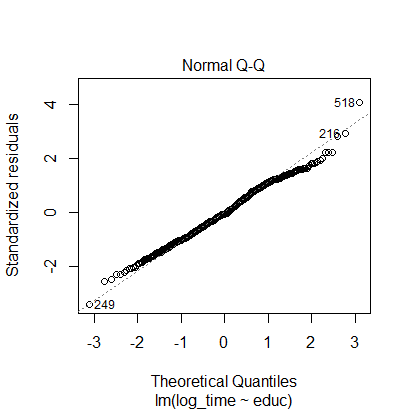
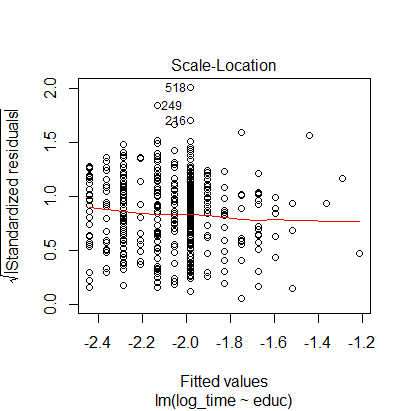
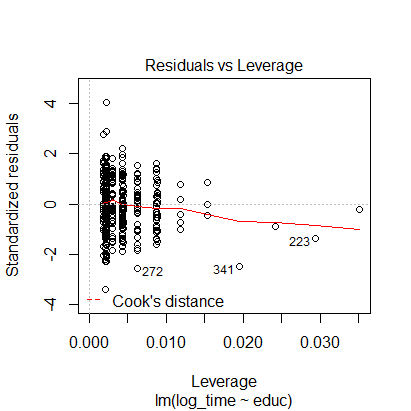
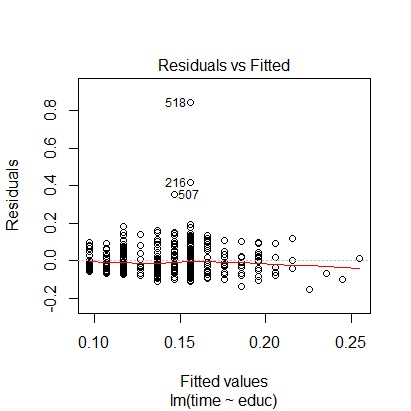
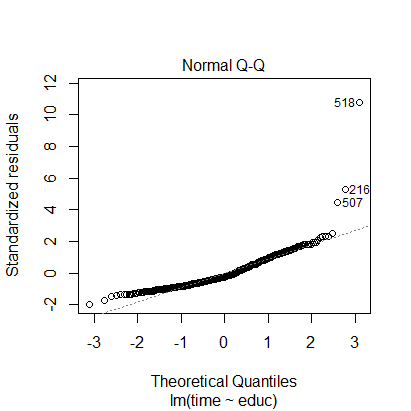
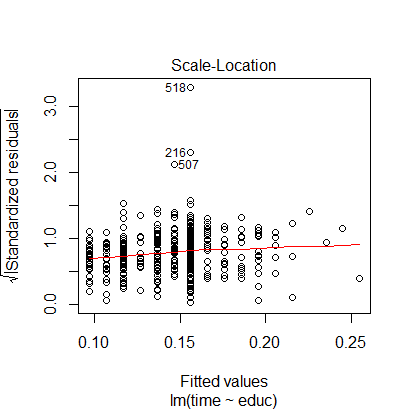
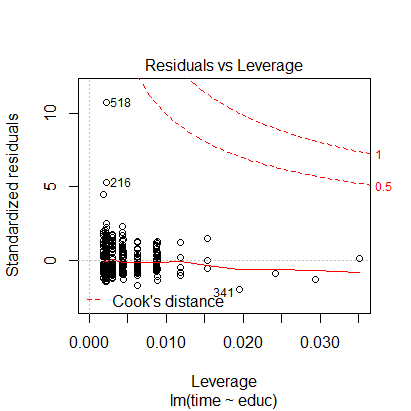
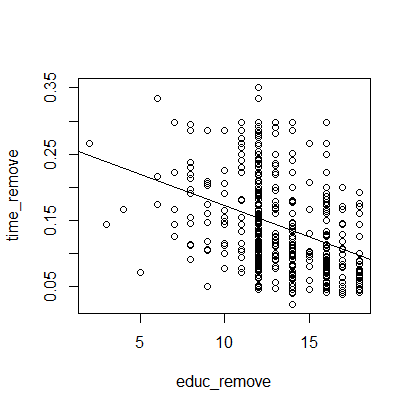
Spencer Swartz

1. In this problem, you will do further cleaning and analysis of the data from the 1995 US News and World Report on colleges and universities in the US.
   1. 
   2. As seen below the max graduation rate is 118% this is imposible so any college that were above 100% were converted to missing data. (Cazenovia College)
      1. 
      2. 
      3. 
      4. 
   3. 
      1. Private mean = 24.58287%
      2. Public mean = 13.44944%
   4. 
      1. With a p-value < 2.2e-16 it is assumed that the mean for private schools is higher than for public schools.
2. The data set cps.csv contains data from the 1985 Current Population Survey.
   1. 
      1. It looks as if a linear model may be used in this plot, it seems that as educ increases so does wage.
   2. We may want to transfor the wage data so that there is not as much varience in the datapoint within each year of education. To do this we will use the log function on wage
      1. Graph after transformation, and plotting of model
         1. 
      2. Diagnostic plots
         1. 
            1. This plot shows that we have pretty good distribution between our residuals and that we likely have a linear relationship between the 2 variables.
         2. 
            1. This plot shows that our residuals are fairly normally distributed as the points follow a straight line fairly well.
         3. 
            1. With this plot showing us a horizontal line and fairly equally spread points, we can assume that the residuals are spread equally along the range of predictors.
         4. 
            1. This plots shows us that there is no strong influence from an outlier. All points are within Cook’s distance.
   3. time <- 1/wage
   4. 
      1. There seems to be the ability to use a linear model on this chart aswell. Expecialy if we transform time using the log of time so that the outliers do not effect the model as much.
   5. 
      1. As the amount of years in education increase the less amount of time it takes to accumulate 1 dollar. Based on these results, yes I am happy with my decision to pursue a master’s.
   6. Diagnostics Plots
      1. 
         1. We likely have a good distribution of our residuals, the data is also likely to be linear.
      2. 
         1. The residuals follow a straight line for the most part showing that the residuals are likely normally distributed.
      3. 
         1. It looks like our residuals are spread normally across the range of predictors.
      4. 
         1. There are no overly strong influences based on outliers in the data. No data point is passed cooks distance.
      5. If I were not to create the model with the log of time we would see int the following diagnostic plots that we would remove records (507,216,518)
         1. 
         2. 
         3. 
         4. 
      6. The model without the outliers would look like this
         1. 
         2. This does not change my conclusion.