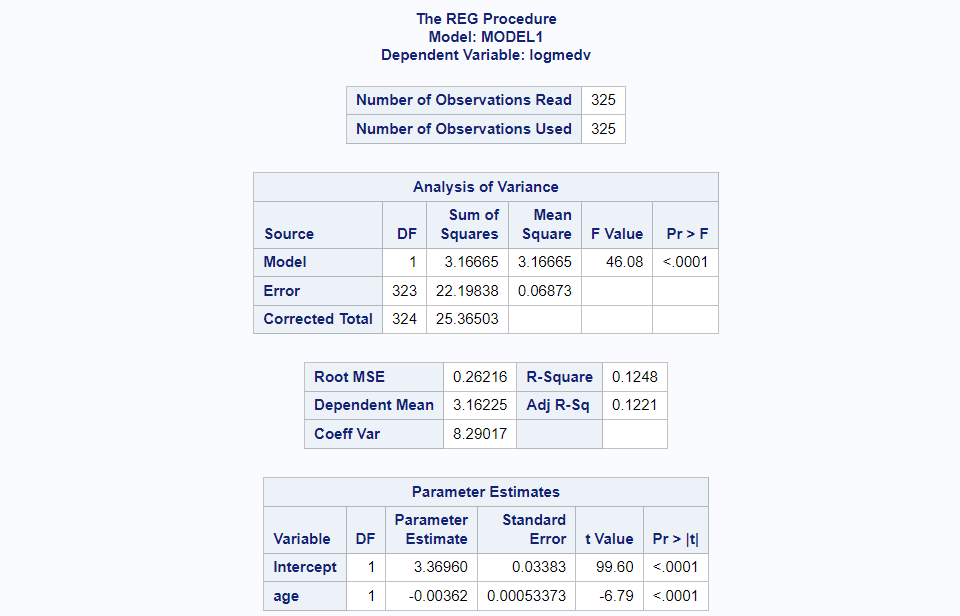
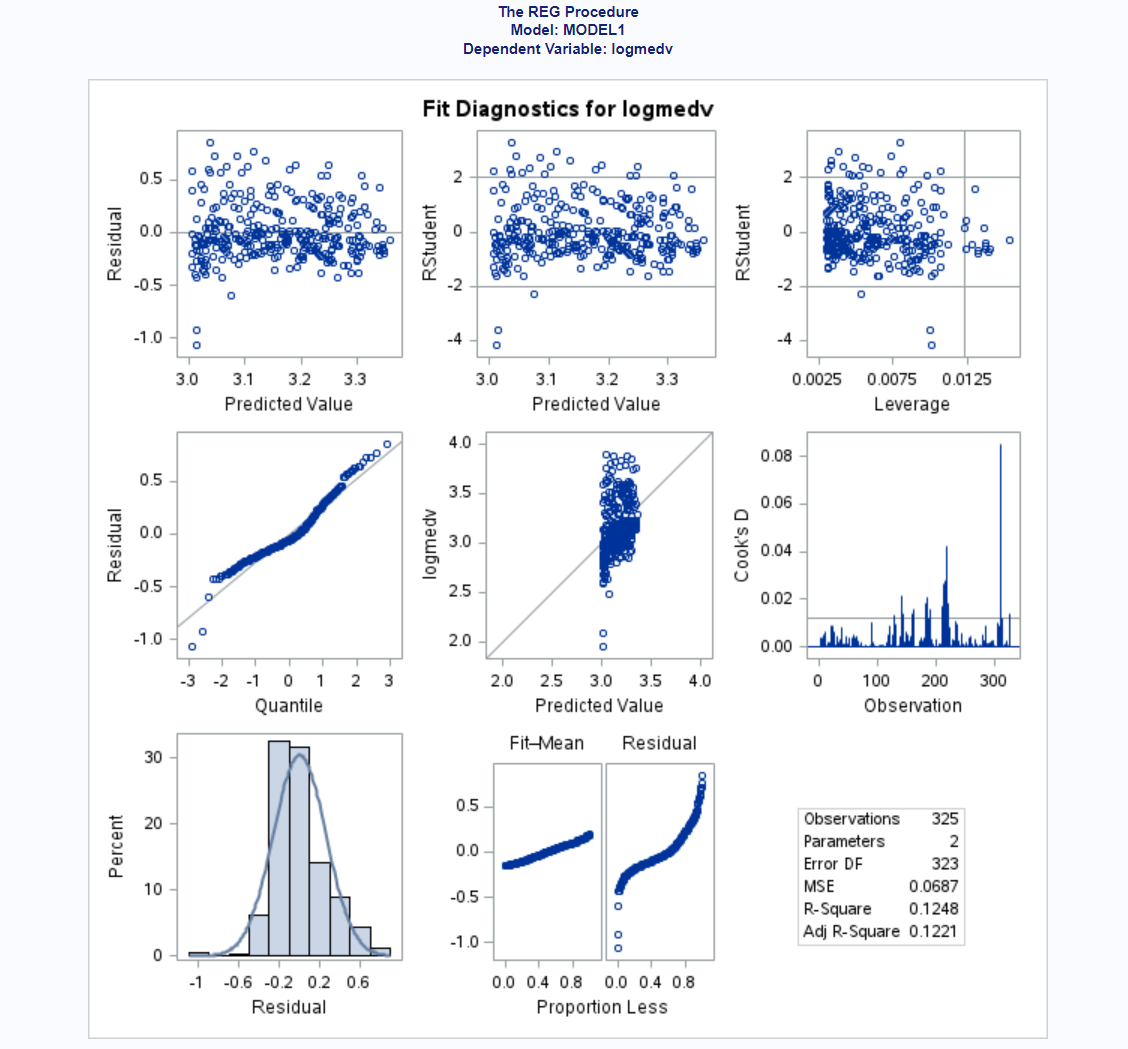
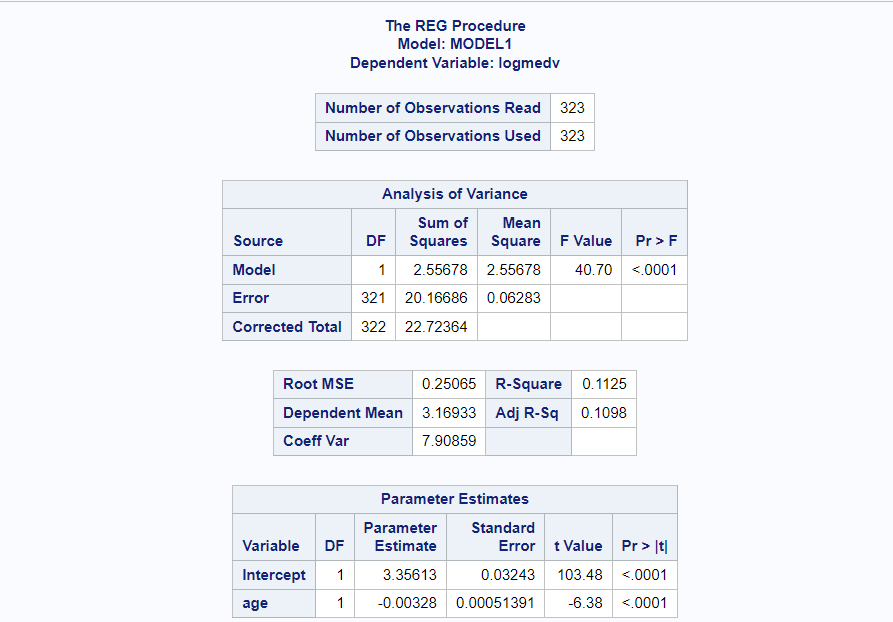
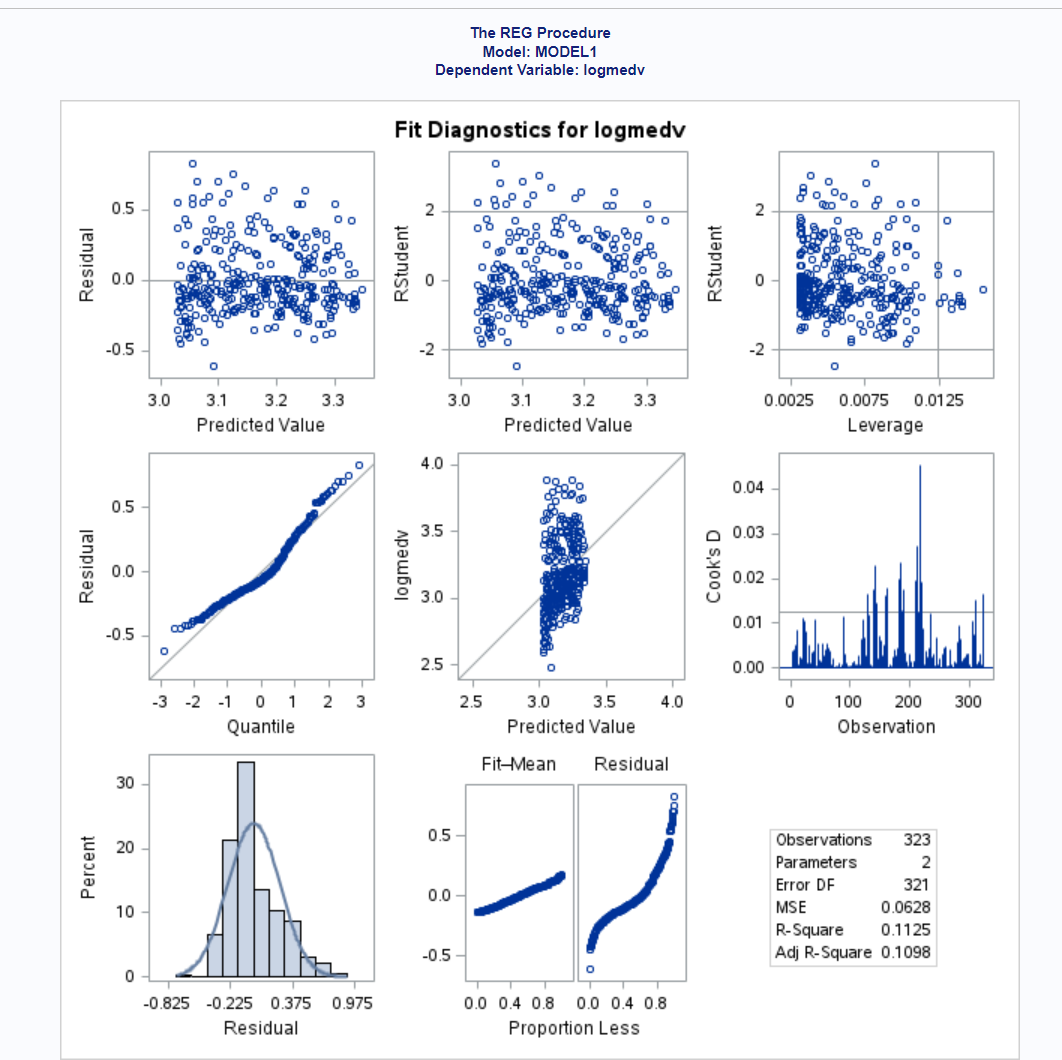
HW 4

## Problem 1

## 1a.)

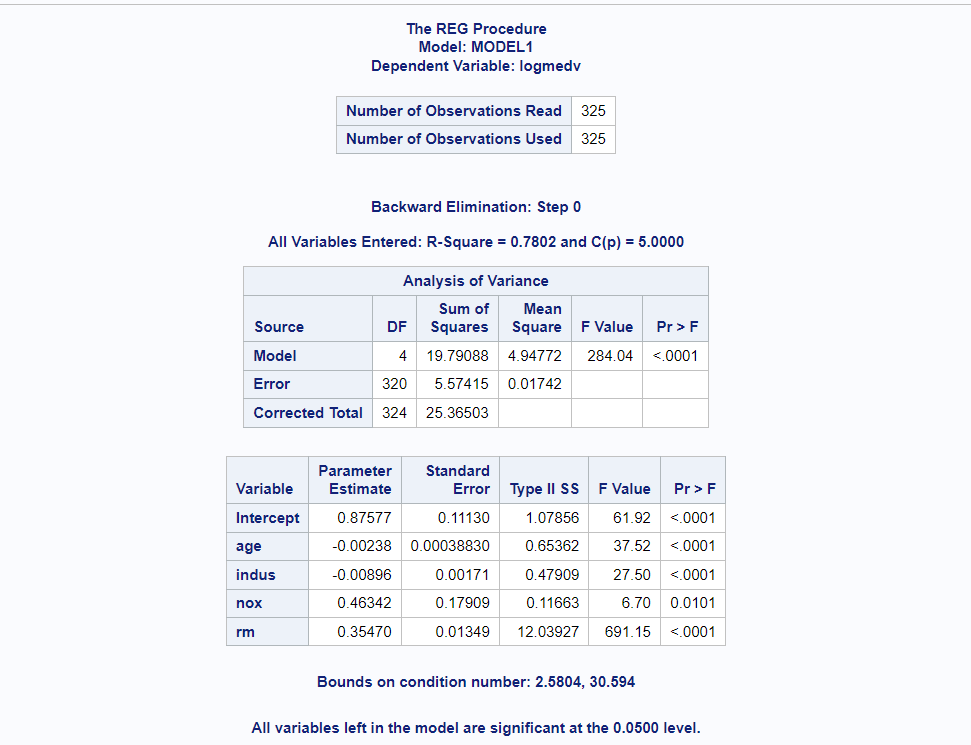
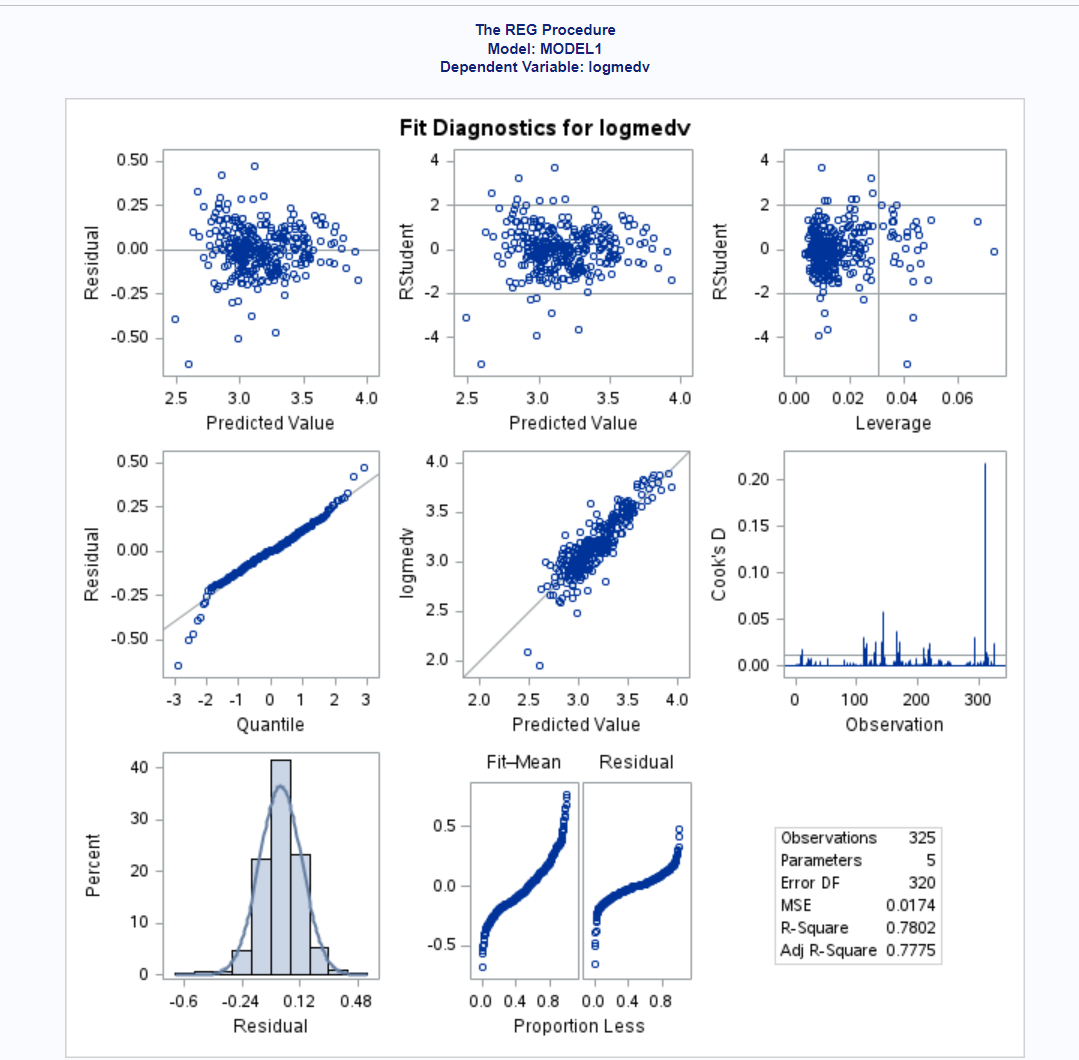
  
  
We see that the log median value vs Predicted value has issues as the points don't roughly fit the diagonal line. Instead they are in an almost vertical line which suggests that the model does not fit the data well. The Residual plots appear generally ok, there could be a slight fan shape with the smaller predicted values having a slightly larger variance. The Cooks Distance & Leverage show a few highly influential points. The Normality Plots also look fine, and the residual fit spread plot shows that the residuals have more spread than the center fitted values.

## 1b.)

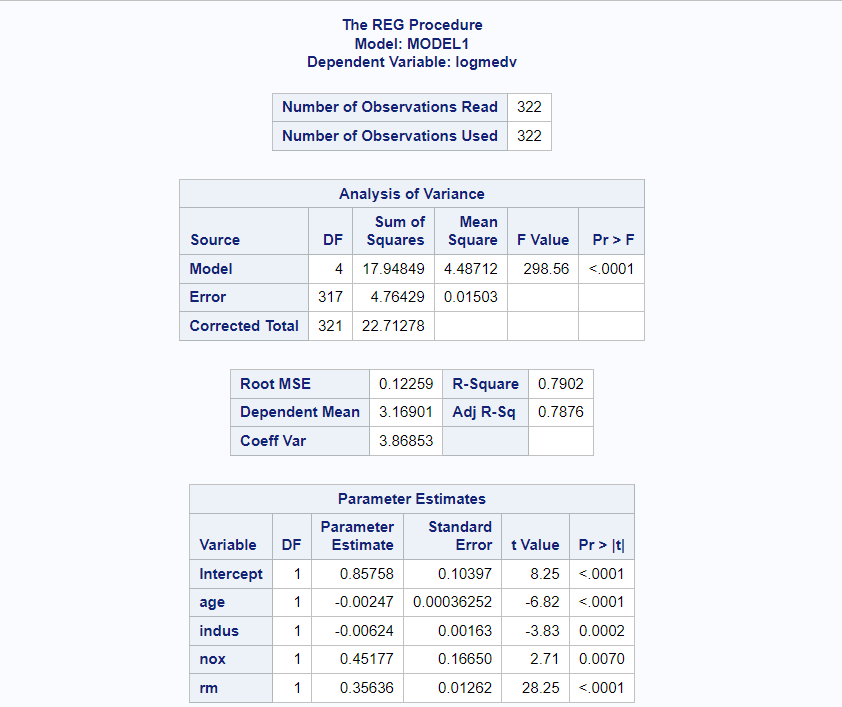
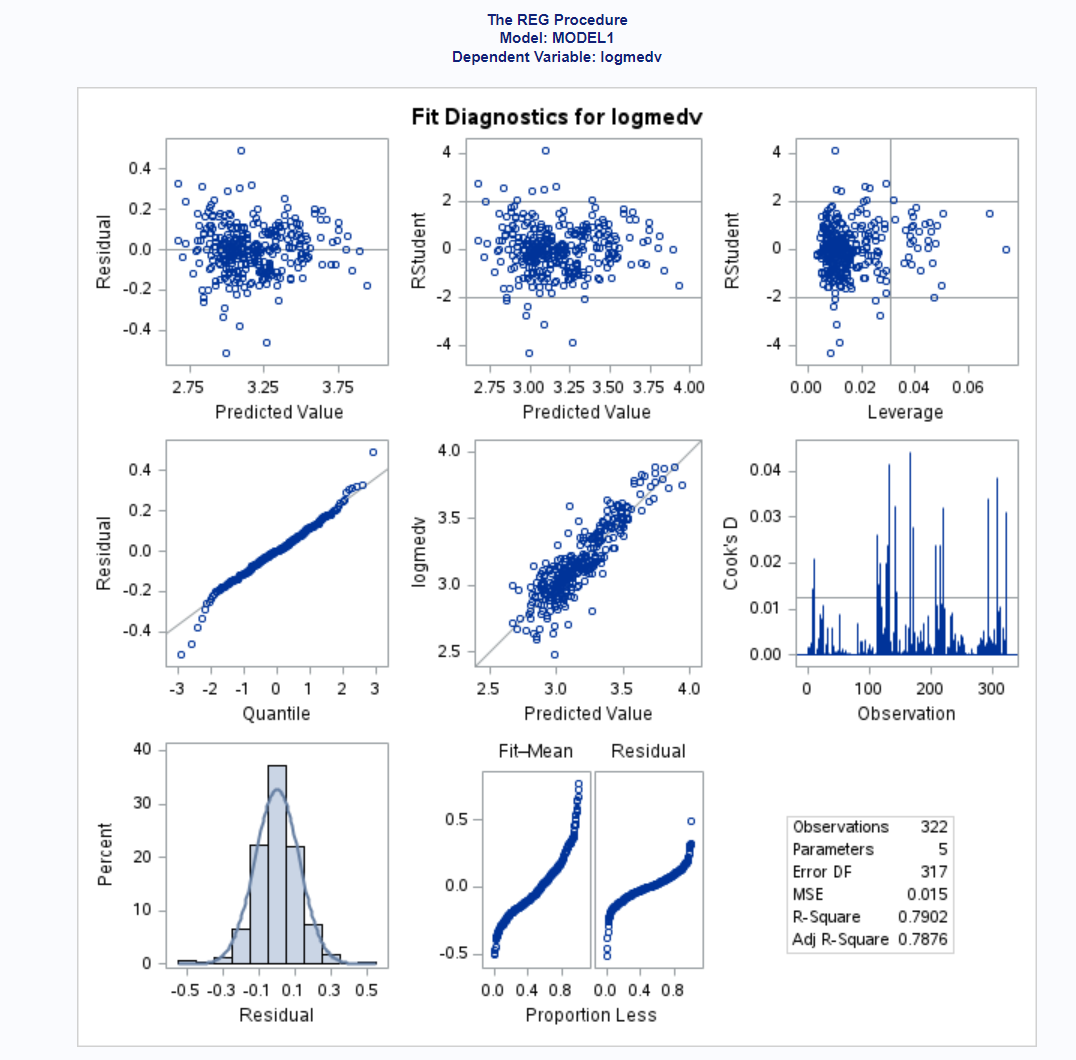
  
After applying the cooks distance cut-off, 2 points were removed, after this no more points exceeded 4 times the cooks distance cutoff. We see that this model is not that good, as the R-Square values are very small, with around 10% of the variation in log median value explained by the age of the home, despite the predictor being significant. We see that as age increases, the median value of a home doesn't appear to change at all. If we take e^(-0.00328) we get approximately 0.99, which is very close to 1 so increase in a unit of age does not really change the expected median value of a home.  
  
The Residual plots now look better with the two influential points removed, there doesn't appear to be any slight fan shape as there was before. The other plots have the same conclusions as the model without the highly influential points removed. To remedy this issue it would be best to either fit a different model or to not use the age variable at all. Based on the diagnostics and results, we see that this model is not very useful when predicting the log median value of a home.

## Problem 2

## 2a.)

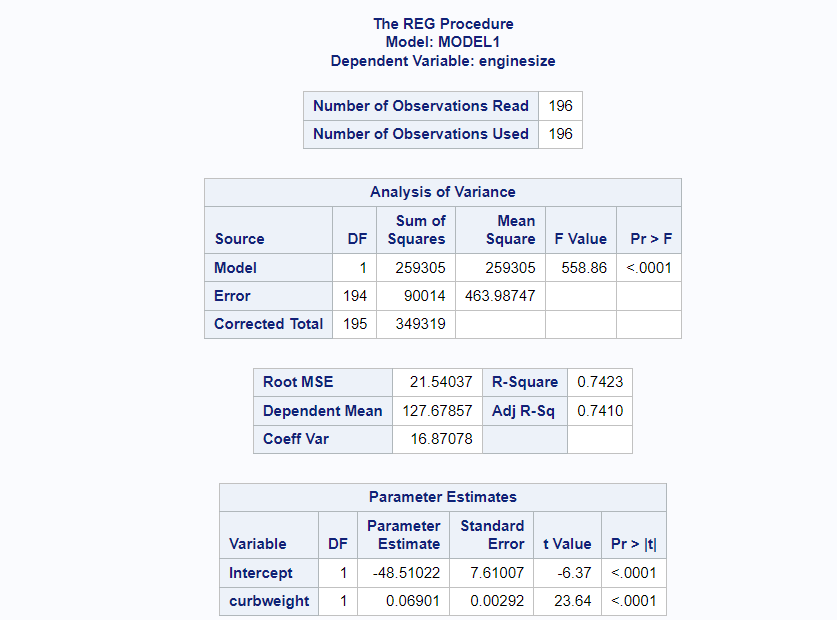
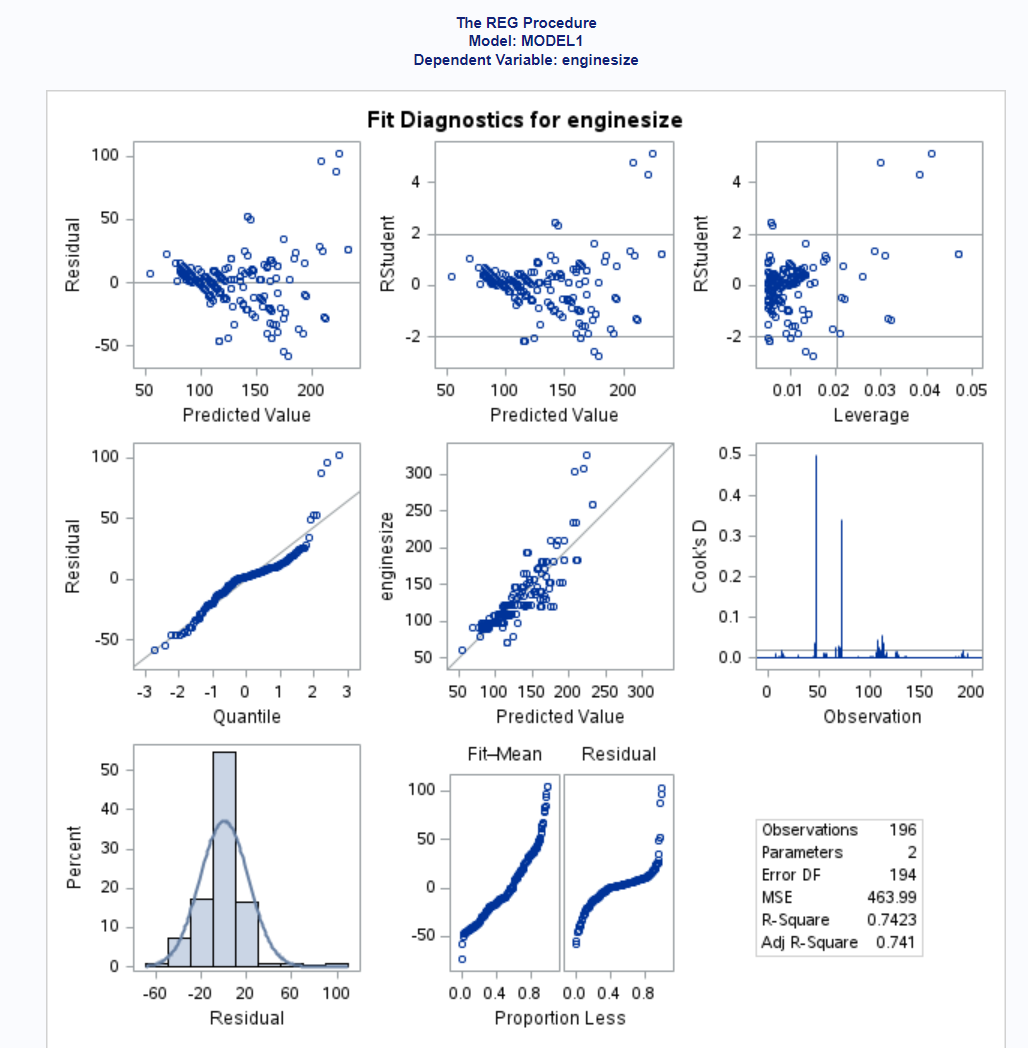
  
Model selection stops on step 0 as all the variable contribute significantly so none are dropped.  
The residual and Normality diagnostics plots are all as expected, they don't have any apparent issues. The Leverage and cooks distance plots show a few highly influential points and the log median value vs predicted value plot has the points following the diagonal line pretty tightly so no issues there. With the Residual fit spread plots we see that the residual values have a smaller spread than the center fitted values.

## 2b.)

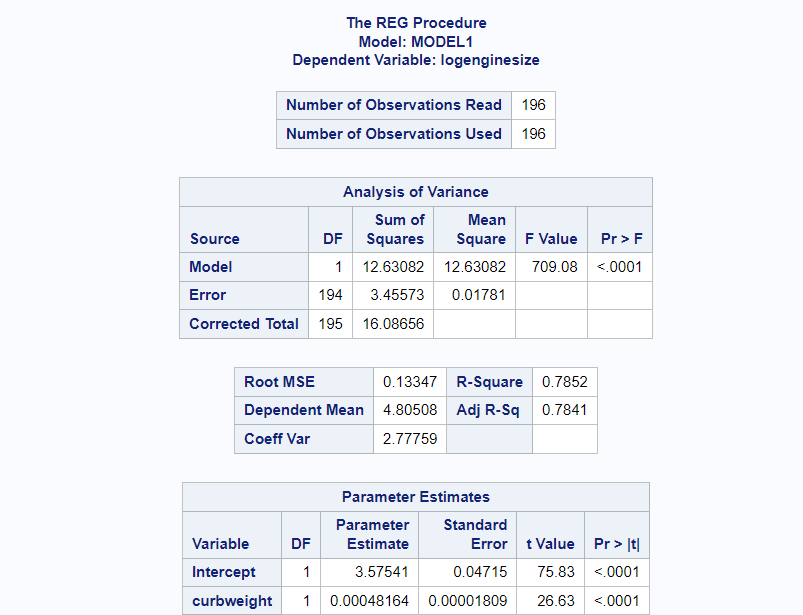
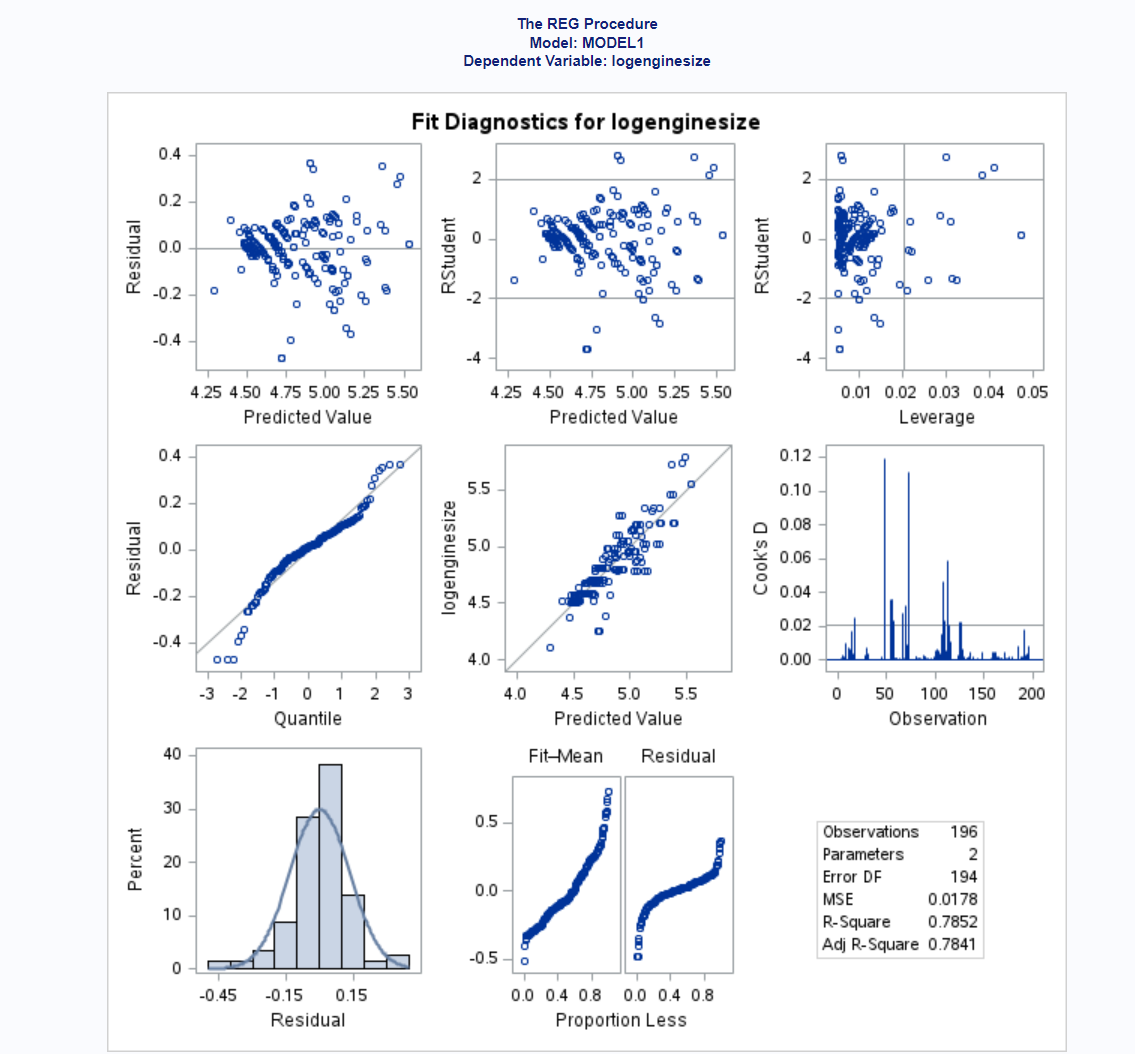
  
We see that this model is pretty good as it has an R-Square value of 0.79 so it explains around 79% of the variation in log median home value in this subset of the data. We dropped 3 highly influential observations that were above 4 times the cutoff. All the terms in the model are still significant. We see that median home value appears to increase as rm & nox increases since taking the exponential of a positive value will be positive. We see that median home value very slightly decreases as age & indus increase.  
  
The only remaining diagnostics issues are the residual fit spread plot that shows the residuals having a smaller spread still and we see some potentially highly influential points but they all have cooks distance less than 4 times the cutoff so they don't get removed.

## Problem 3

## 3a.)

  
  
There appear to be 2 points that look to be highly influential based on the cooks distance plot. Additionally, the residual plots look like they have a slight trend and are not uniformly distributed. The rest of the plots don't appear to have any issues.

## 3b.)

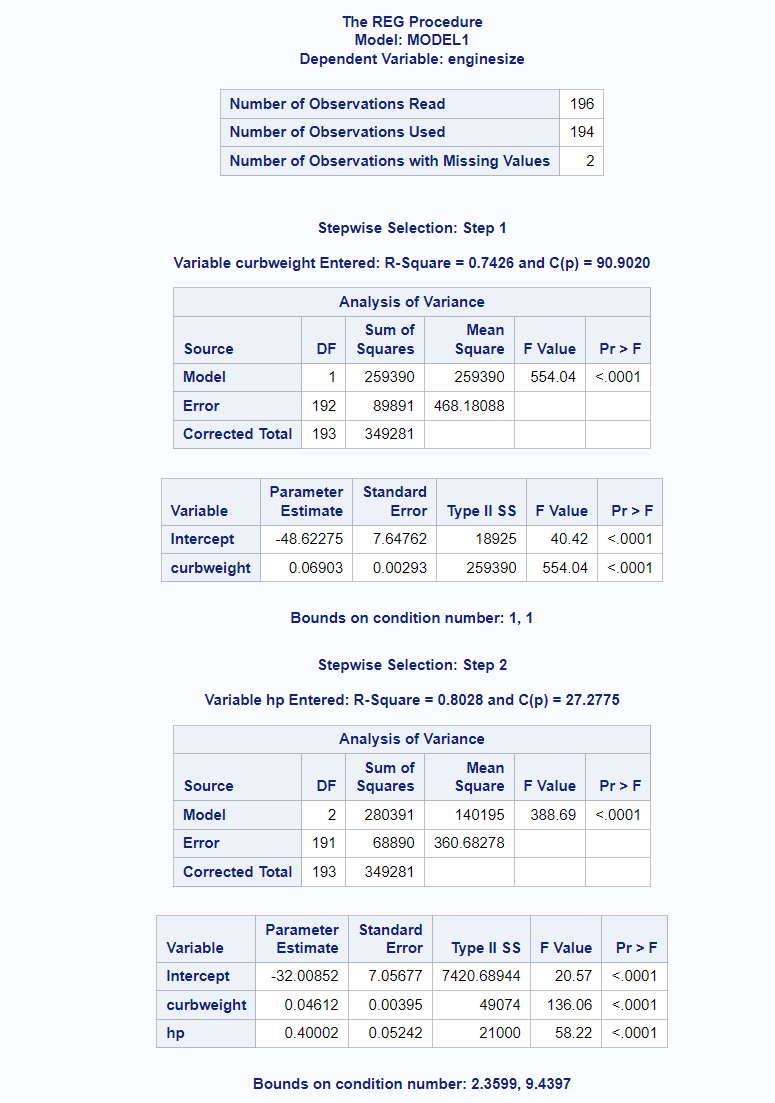
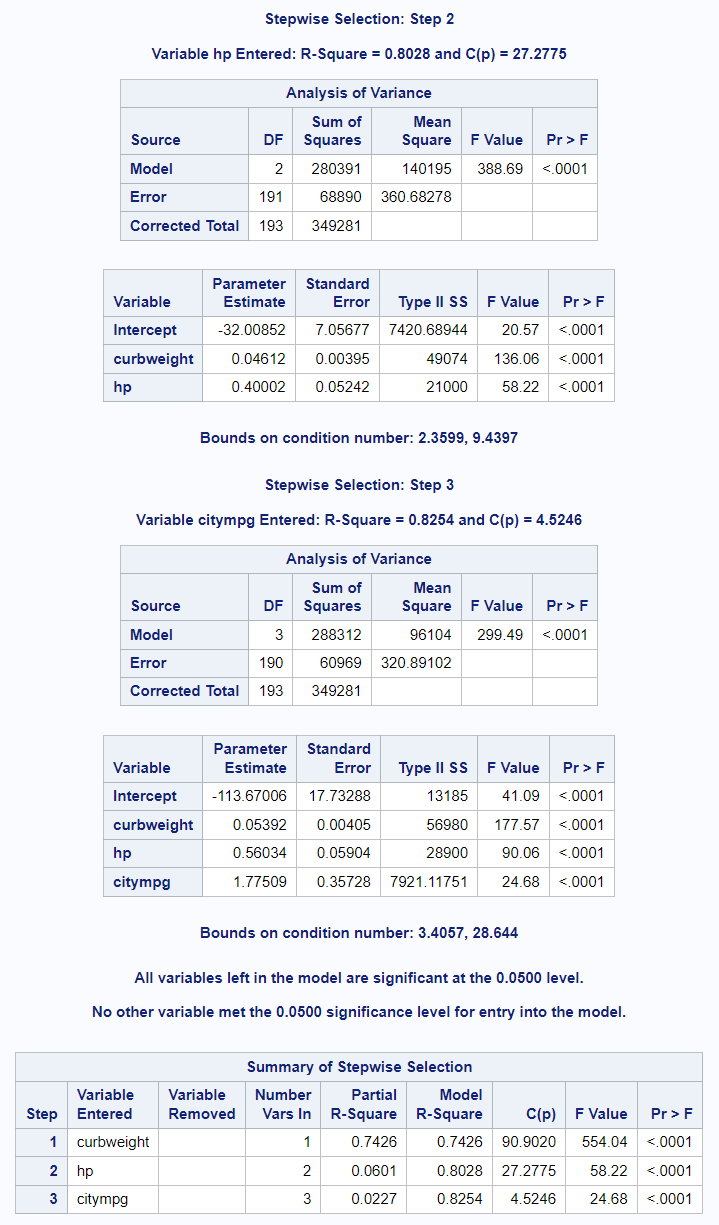
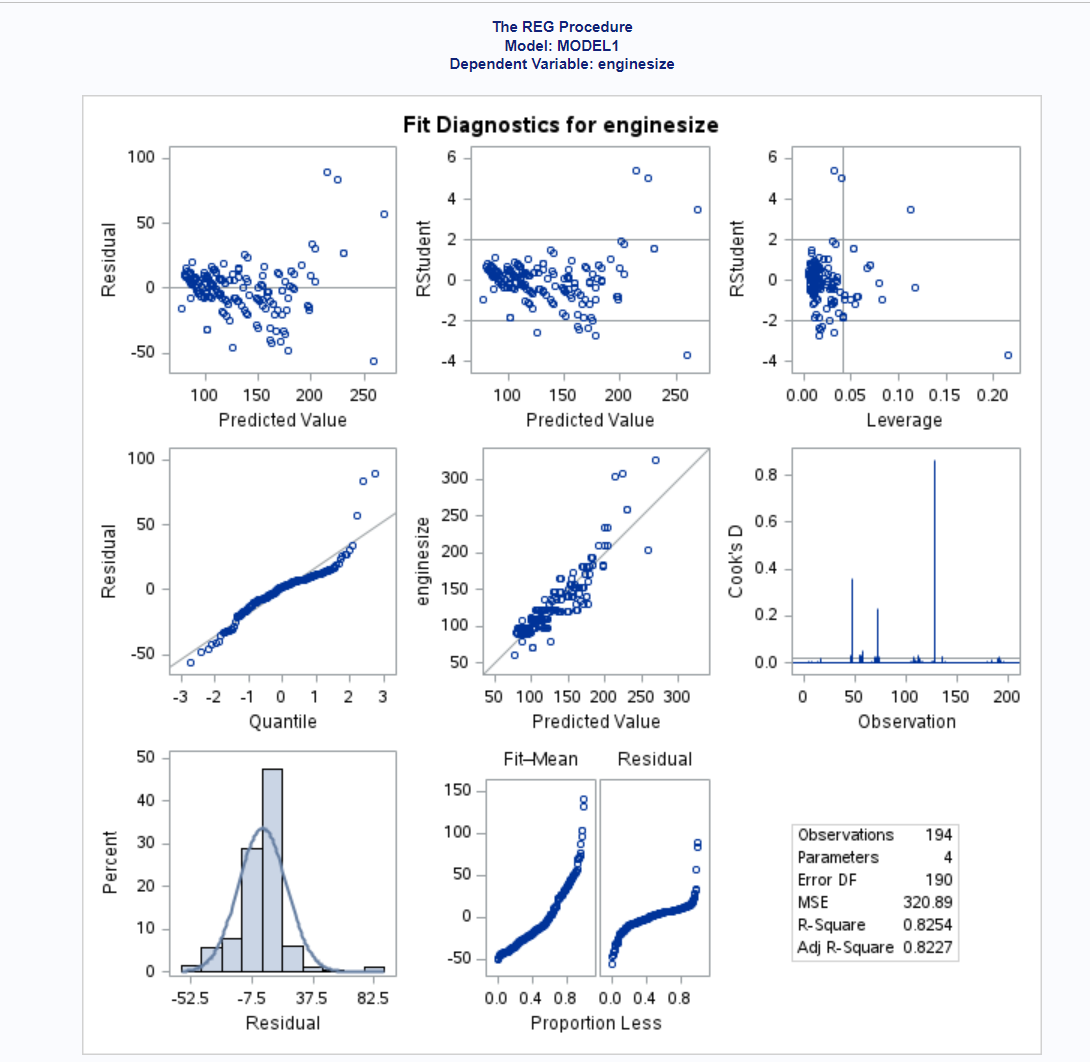
  
  
We see again two points that appear to be above 4 times the cooks distance cutoff. The residual fit spread plot show that the residuals have a slightly smaller spread than the centered fitted values. However the rest of the plots appear to be ok, the residual plots look better than the previous model too as they now don't appear to have any trends.

## 3c.)

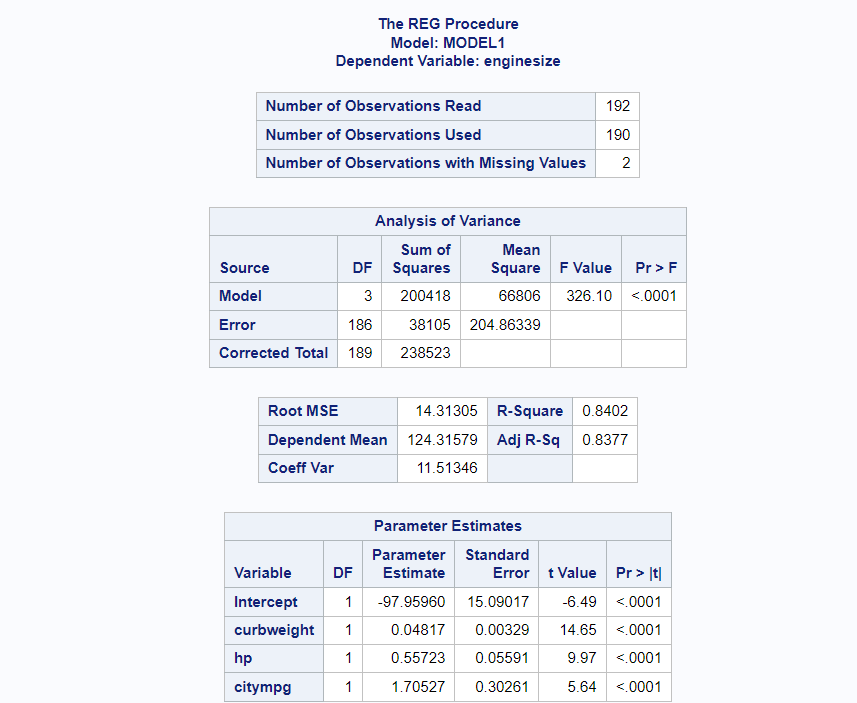
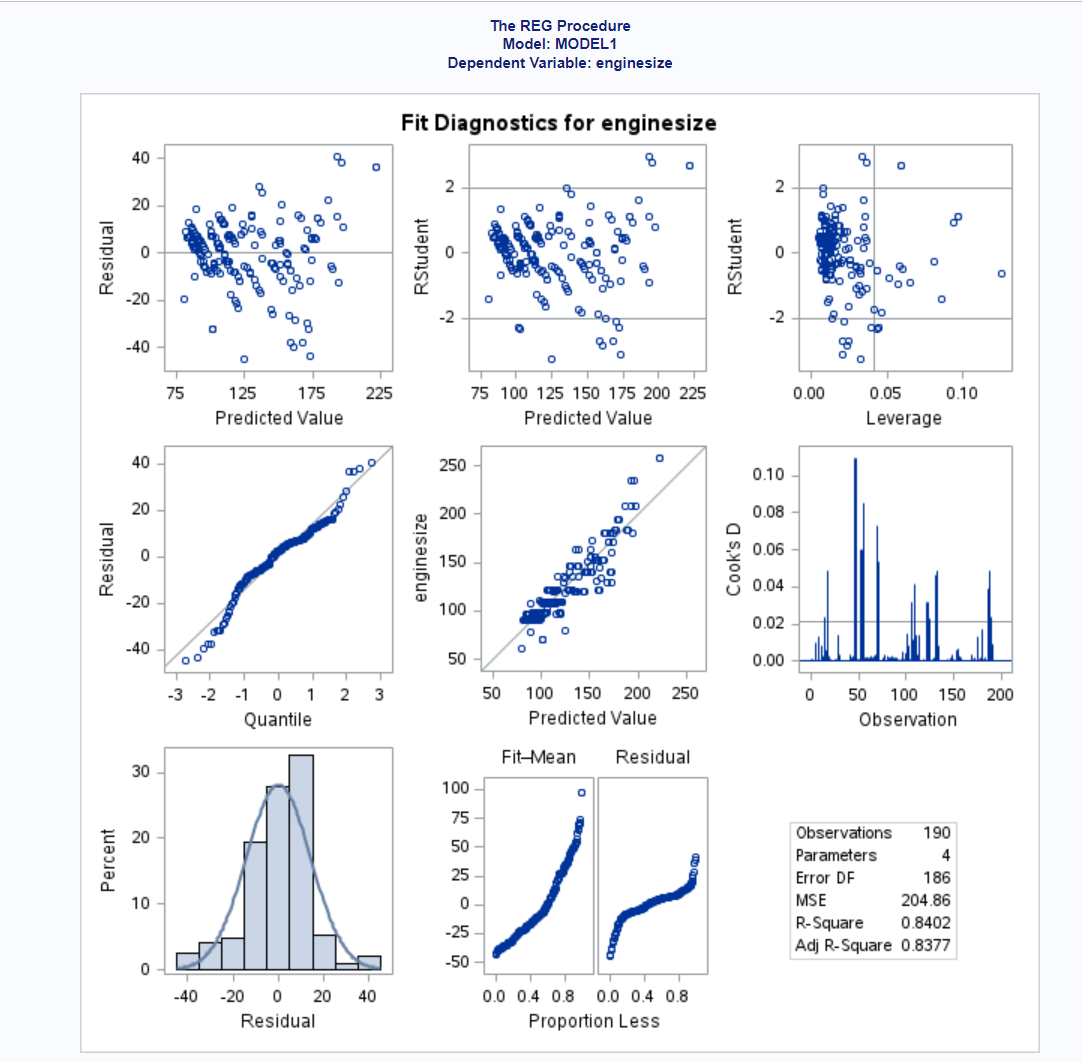
The model from part b would be the better model since it has a higher R-Squared and does not have the residual diagnostic plot issues. So for the model in part b, we explain around 78% of the variation in log engine size using curb weight. We see from the parameter estimates that as curb weight increases by a unit then log engine size increases by 0.0004.

## Problem 4

## 4a.)

  
  
  
I first ran stepwise selection on just the continuous variables and got the model with curb weight, hp & city mpg as the predictor variables. I then removed the highly influential points that had cooks distance over 4 times the cutoff. I ended up removing 2 variables and after this no more points where 4 times above the cooks distance cut off.

## 4b.)

  
The model is significant and it explains around 83 percent of the variation in engine size. As city mpg increases by 1 unit, engine size appears to increase by 1.7 units. While HP increases by 1 unit, engine size increases by 0.55 units, and while curb weight increases 1 unit, engine size increases by 0.04 units. We see that all the variables appear to have a positive correlation with engine size.  
  
The only remaining diagnostics issues are in the residual fit spread plot, where the residuals appear to have a smaller spread than the centered fitted values. The rest of the plots don't have any apparent issues.