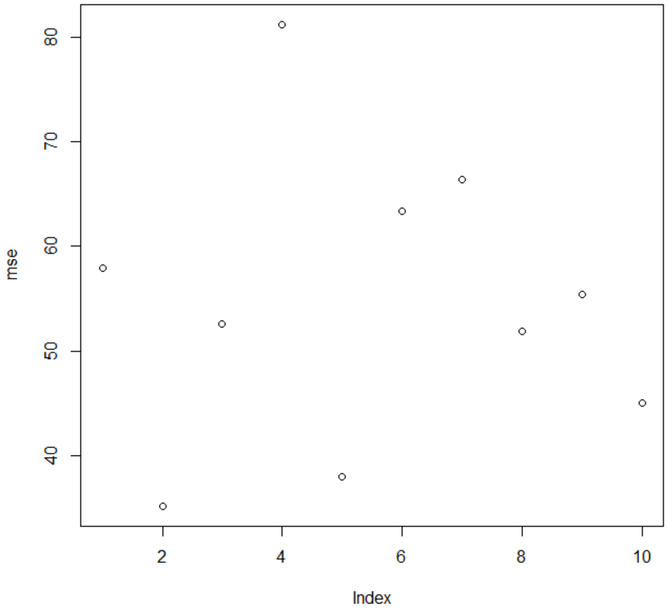
Arun Sangar

Project 1

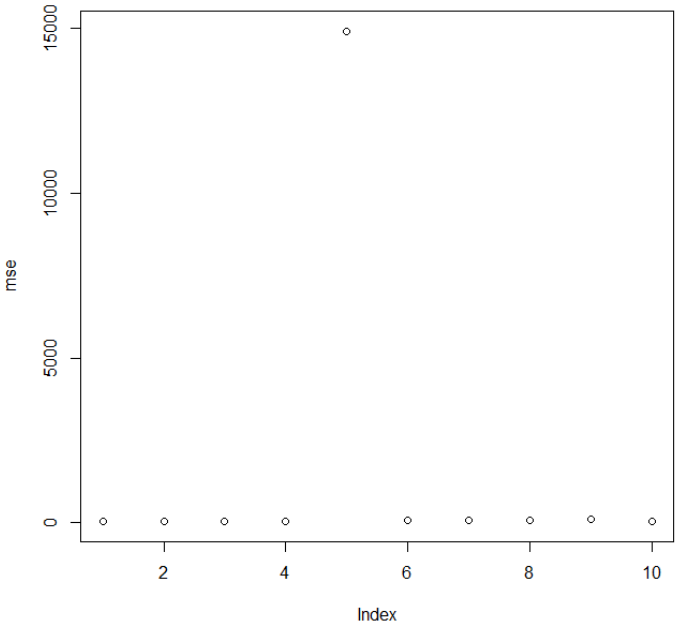
CPSC 375

MSE Analysis

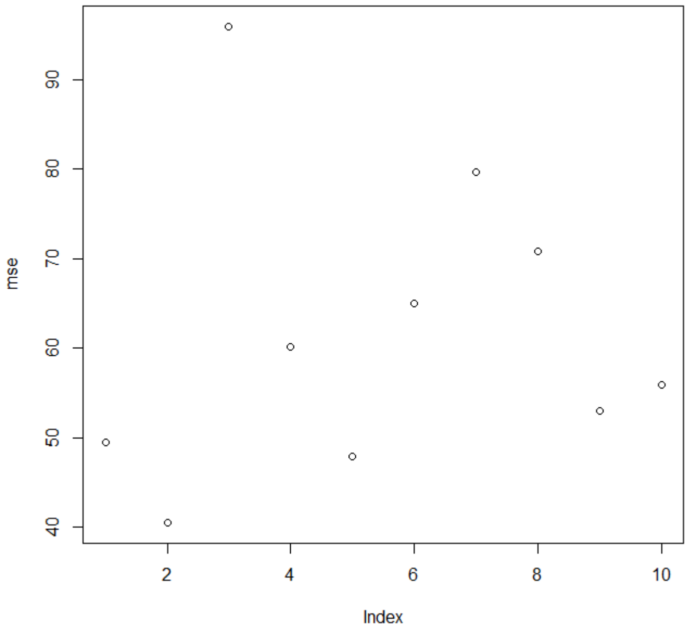
1) The first linear model I tried was simply using all variables with no transformations. The average MSE was around 55.



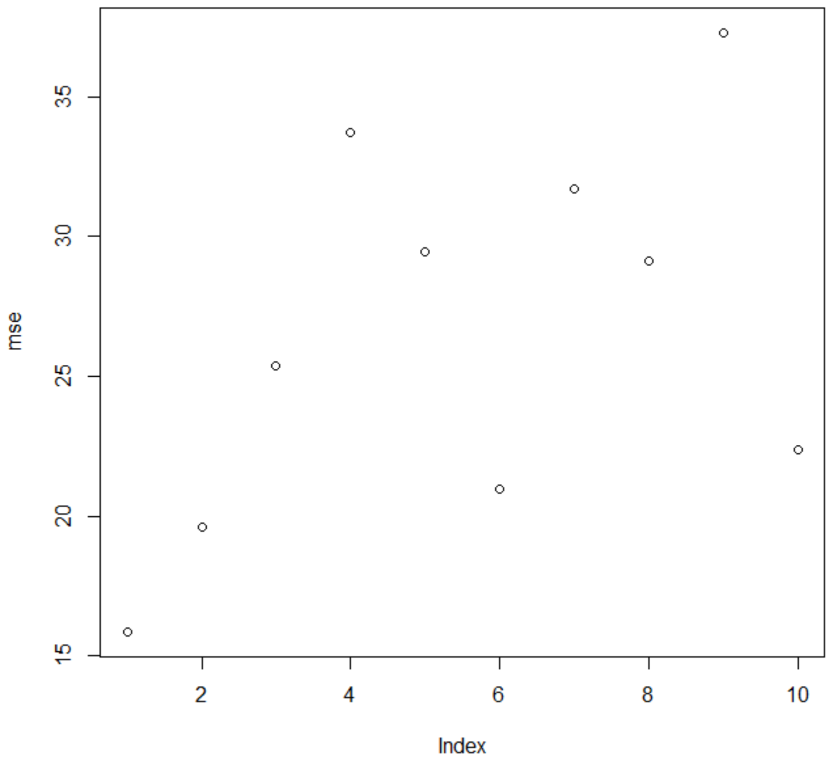
2) The next linear model I tried was to use BMI which involves a transformation of weight and height. The average MSE was always very high (1000+). This was quite interesting because it showed a very strong outlier. From the MSE plot we can see that it is one data point giving the error. Running the linear model multiple times always resulted in this type of plot. This means one data point is responsible. Looking into it further, I found that entry 42 in the body fat data set had a height of 29.5 inches. This is either a very specific case or incorrect data.



After this I removed this entry and retried this method. The average MSE, after removing the outlier, was around 61. This is not a better model than the previous one because BMI gives the outer dimensions of the body and does not give much information on the ratio of lean to fat tissue.



3) The next linear model I attempted was to remove any weak predictors of body fat. These are areas of the body that have less body fat (e.g. neck/knee/ankle/wrists), areas that may be skewed because of muscle mass (e.g. chest/biceps/forearm) and parameters that may not have direct correlation to body fat (e.g. age/height). Area with less body fat or higher muscle mass are likely not good indicators of body fat percentage since they may not change as much relative to body fat. The age may not be good indicator as well because body type varies greatly from person to person. Removing these 9 parameters and running the linear model resulted in an average MSE of around 27. We can see from the average MSE that this is clearly the best model.


mse
Index