**Alex Riles**

**DSCI 425 – Supervised Learning (50 pts.)**

**Assignment 2 – MLR, ACE/AVAS, MARS**

**predicting strength of concrete**

Concrete is the most important material in civil engineering. The concrete compressive strength is thought to be a highly nonlinear function of age and ingredients.

**Variable Information:**  
Given below are the variables contained the file **Concrete.csv** on course website. These data come from a collection of 17 experiments where the compressive strength (MPa) of concrete was determined under different formulations and length of curing (days). These data consist of n = 1030 observations on nine variables (8 predictors and 1 response). There are no cases with missing values!  
  
Name / Data Type / Description/Measurement Units (red denotes variable has zeroes)

* Cement () - continuous – kg of cement per cubic meter of concrete
* Blast Furnace Slag () - continuous – kg of slag per cubic meter of concrete
* Fly Ash () - continuous -- kg of fly ash per cubic meter of concrete
* Water () - continuous -- kg of water per cubic meter of concrete
* Superplasticizer () - continuous -- kg of superplasticizer per cubic meter of concrete
* Coarse Aggregate () - continuous -- kg of course aggregate per cubic meter of concrete
* Fine Aggregate () - continuous -- kg of fine aggregate per cubic meter of concrete
* Age - discrete – age of concrete measured in days (1-365)
* Concrete compressive strength - continuous – compressive strength in *megapascals* ()

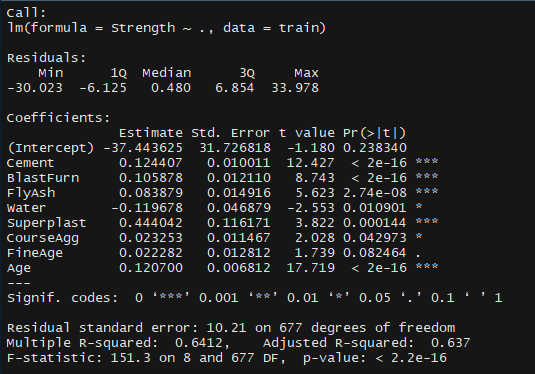
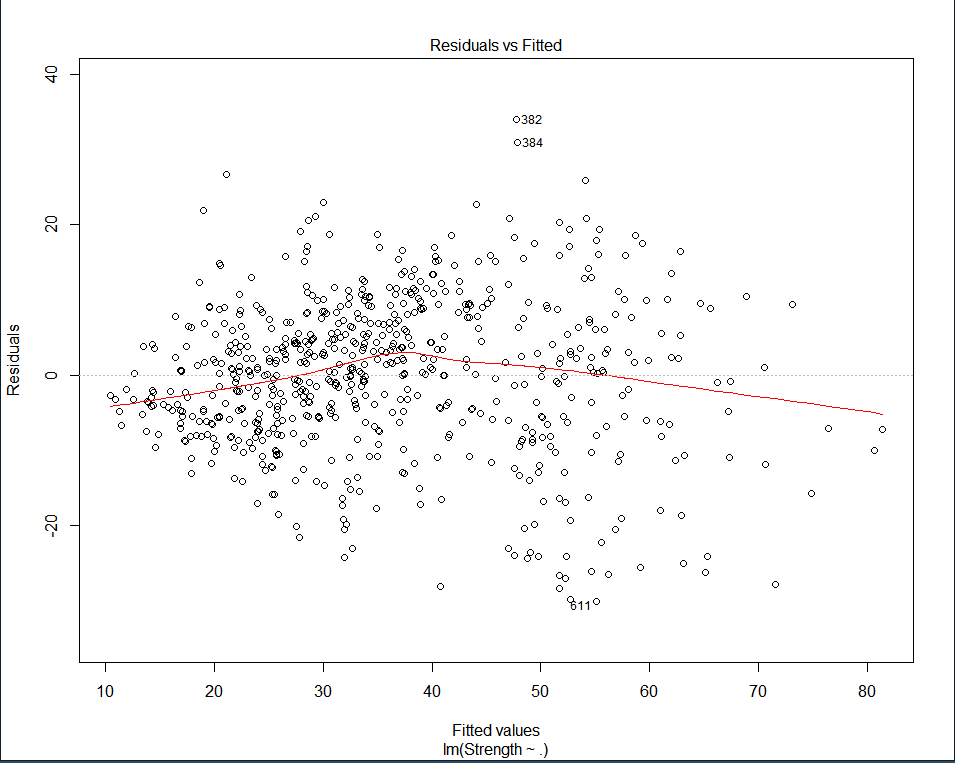
Data source: I-Cheng Yeh, "*Modeling of strength of high performance concrete using artificial neural networks*," Cement and Concrete Research, Vol. 28, No. 12, pp. 1797-1808 (1998).

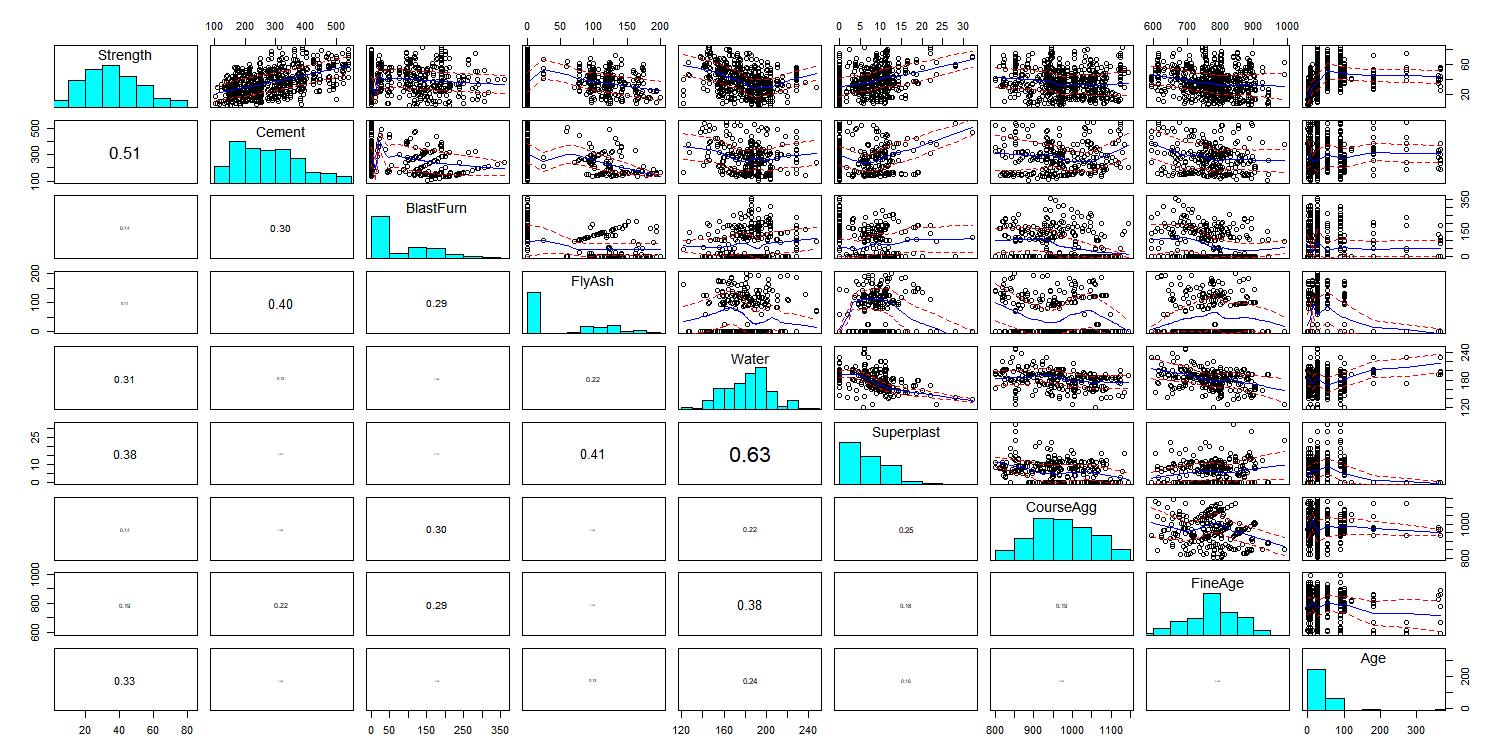
First we will form a training dataset consisting of 66.6% of the original cases. This is done by selecting 66.6% indices (obs. #’s) of the original n = 1030 observations to serve as the training data. The 33.4% the cases not selected will serve as the validation set.

> Concrete = read.table(file.choose(),header=T,sep=”,”) #read in the **Concrete.csv** file.  
> set.seed(1) # You MUST execute this command before creating your training sample.  
> sam = sample(1:1030,size=floor(.6666\*1030),replace=F)  
  
To refer to the training data frame use Concrete[sam,] and to refer to the validation data frame use Concrete[-sam,]. Any transformations you perform can then be applied to the entire dataset without doing it separately for the training and validation sets. For example suppose you want to square the response (which you DON’T) then you could form a new data frame containing the transformed response.

> Concrete.trans = Concrete  
> Concrete.trans$Strength = Concrete$Strength^2

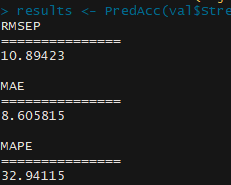
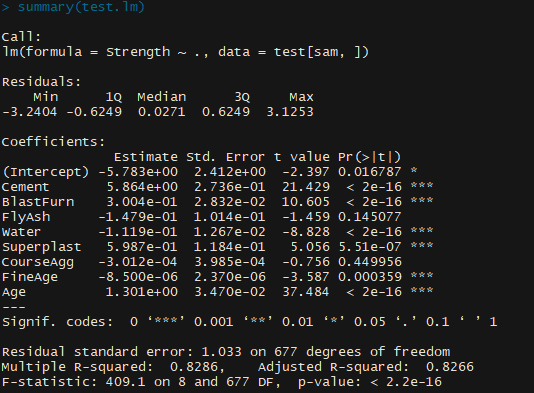
Other transformations could then be performed to create a new data frame Concrete.trans. You can again use Concrete.trans[sam,] and Concrete.trans[-sam,] to refer to the training and validation sets respectively.

1. Fit a MLR model with all variables in their original scales using the training data. Summarize this model and discuss any model deficiencies. (5 pts.)  
     



With using all the predictor in the original scale there is enough evidence with a p-value at 2.2e-16 predict the strength of concrete. There are some deficiencies that can be found in the predictor and response. There deficiencies are that there is some curvature in the residual vs. predicted plot and when looking at the correlation plots there is skewness in the response and in the predictors.

1. Use the model from part (a) to predict the response value using the validation data and compute the prediction accuracy (RMSEP,MAEP,MAPEP) of these predictions by comparing the actual compression strengths of the concrete samples in the validation set. (5 pts.)

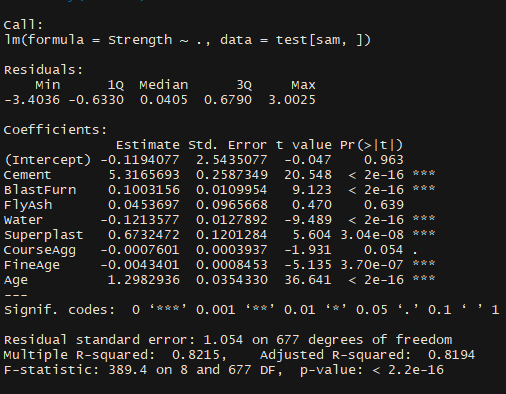
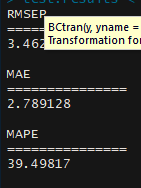
 

RMSEP = 10.89423

MAEP = 8.605815

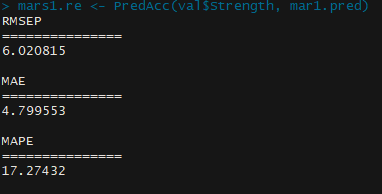
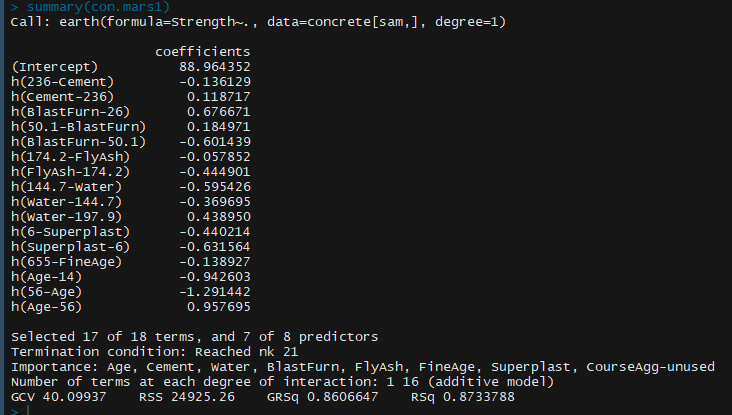
MAPEP = 32.94115

1. Use tools such as Box-Cox transformations, CERES plots, ACE/AVAS, and stepwise model selection to create and choose terms and choose a potential response transformation to address the deficiencies exhibited by the model from part (a). You should explain what tools you used and give a summary of your final MLR model selected. This model should not have the deficiencies identified in part (a). Use this model to predict the compression strength of the validation case in the ORIGINAL SCALE (). Compute the prediction accuracy measures (RMSEP, MAEP, MAPEP) and compare to the results from part (b). Does your more complex MLR model do a better job in terms of these prediction accuracy measures? If it doesn’t you might rethink your model. (10 pts.)



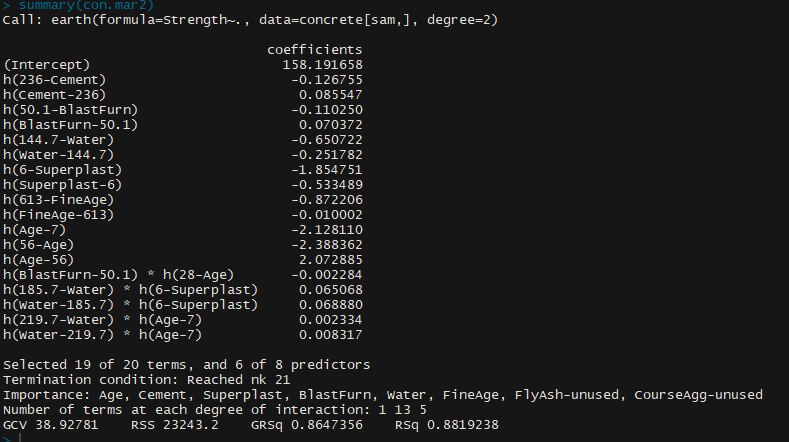
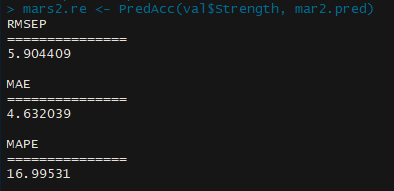
The things that I did is a Box-Cox transformation to find transformation for each variable. Most of the variables used powers to transform but Blast Furnace and Age were the two variables that I found that were transformed using the log function. When it comes to the deficiencies there were still present and I couldn’t find a way to fix it through transforming. The R-squared was . which is an improvement to what was found in part a.

1. Fit a MARS model to the training data with degree = 1 (i.e. no interactions). Use the internal cross-validation features of the earth() function to choose the “best” MARS model with degree = 1. Again predict the compression strength of the concrete samples in the validation set in the original scale and compute RMSEP, MAEP, and MAPEP. How does this compare to the models in part (a) and (c)? (10 pts.)

When looking at R-squared it did better than the first two model. Also the MAPEP was cut in half compared to the other two models which mean there is a less mean percent error when using the MARS model with degrees of freedom of 1.

1. Fit a MARS model to the training with degree = 2 (i.e. including potential interactions). Again use the internal cross-validation capabilities of the earth() function to choose the best degree = 2 MARS model for these data. Again predict the compression strength of the concrete samples in the validation set in the original scale and compute RMSEP, MAEP, and MAPEP. How does this compare to the earlier models? Which predictors seem most important for predicting strength? (10 pts.)

When looking at the MARS Model with degrees of freedom of 2 it increased the r-squared to .8819 and the MAPEP went to 16.995. So, after all the different modeling methods used the MARS with degrees of freedom of 2 was the best model.

1. Fill in the table from fitting all of the models above. (10 pts.)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model Method | (training data) | (predicting validation set) | (MPa) | (MPa) | (%) |
| MLR orig | .6412 | .5768 | 10.89 | 8.605 | 32.94 |
| MLR trans | .8286 | .8338 | 3.390 | 2.734 | 38.560 |
| MARS (deg 1) | .8733 | .8744 | 6.021 | 4.799 | 17.274 |
| MARS (deg 2) | .8819 | .8795 | 5.904 | 4.632 | 16.995 |