Linear Regression Analysis

Final Exam Approach

Nick Morris

12/14/15

Day 1

Variable Evaluation – Correlations and Linear Relationships

My first step in developing a model for the Thermal Load data set was to understand each variable's relationship with the response individually. This was accomplished analytically and graphically. I created a function Transform() to speed up the process of trying different transformations on variables. This function ranks transformations based on the magnitude of correlation between the response and a transformed variable. This correlation magnitude may be misleading by destroying the spread of data in a scatterplot or by improving the correlation by an insignificant amount compared to the untransformed variable. Therefore the top 6 results of the Transform() function for each continuous variable, were plotted to verify that there was still a spread of data that looked significantly more linear. The results of this analysis weren't promising. The work for this can be found in the # CONTINUOUS VARIABLE CORRELATIONS section of the Approach.R file.

Variable Evaluation – Scatterplot Groupings

The prominent characteristic of this dataset shown by the pairs plot, is the groups of data points vertically and horizontally. This led me to use ggplot2 to quickly create scatterplots, capturing multiple variables to better understand the interactions that may be causing these groupings of data points. The results of this analysis were promising, clearly identifying that interactions would accurately pin point the Thermal Load range of possible values due to the groupings. The work for this can be found in the # VARIABLE GRAPHICS section of the Approach.R file.

Regressor Evaluation – Simple Linear Regression Models

The results of transformed variables and identified interactions were used in individual simple linear regression models. These models included untransformed main effects, transformed main effects, two way interactions, three way interactions, and four way interactions. These models allowed me to identify what terms alone have a significant effect on Thermal Load. The work for this was logged into an excel spreadsheet for later use in my approach.

Manual Mixed Stepwise Regression

The next step in my approach was to start with an all main effects model, without any transformations, and manually add in and remove terms. I would add in terms based on the terms identified from the scatterplot groupings, and significant simple linear regression terms. I would remove terms based on a p-value requirement of <= 0.15, and a decreasing RSE requirement. Key aspects of this approach are adding in terms that aren't significant and that make another term which was already in the model, insignificant. This allowed me to switch out terms to see if the overall accuracy of the model improved. This approach allowed me to understand what terms do and don't work together, and what terms tend to over fit the model, particularly the Orientation:WindowDist interaction. This process went through 95 iterations and can be found in the # MANUAL MODEL STEPPING section of the Apprach.R File.

Day 2

DOE Experiments – 6^5 Factorial

This experiment included six levels {No Transform, Squared, Log, Reciprocal, Square Root, Exp} and the 5 continuous regressors. This experiment was indented to see if a particular combination of transformations would reduce the RSE value of the resulting model from the manual mixed stepwise regression. This experiment was set up in excel to quickly develop the R syntax for each combination, and then were copy and pasted into the R console in batches of 100 and evaluated based on RSE values. The results of this analysis weren't promising, the 3033rd iteration proved to decrease the RSE slightly, but failing to drop below 0.5 still.

DOE Experiments – 6^4 Factorial

This experiment included six levels {No Transform, Squared, Log, Reciprocal, Square Root, Exp} and 4 continuous regressors, Surface was excluded due to its tendency to be aliased with Compactness. This experiment was intended to see if the step() function would produce a better model if given a particular combination of transformed variables. The stepwise included a scope up to three way interactions. This experiment was set up in excel to quickly develop the R syntax for each combination, and then were copy and pasted into the R console in batches of 100 and evaluated based on RSE values. The results did not prove to produce a better model than previously developed.

PCA

Principal Components Analysis was performed to see if a linear combination of variables would produce a better model. The calculated loadings using all of the continuous variables in the data set, proved to show Wall and Roof as a linear combination. They were combined to create a Z1 and Z2 variable which were then used in a mixed stepwise regression and drop1 analysis to result in a model that was no more accurate than the previously developed models.

Day 3

Six Major Assumptions

The effort to reduce the RSE value, was no longer adding value to the models created so the next step was to meet the six major assumptions. This was done by using all of the previously developed models, starting from a main effects model, and starting from a three way interactions model. The models were evaluated analytically and graphically. Changes were made to the various models by transforming the response variable with log transformations, suggested box cox transformations, and by replacing terms that had large variance inflation factor values.

The Chosen Model

The final model candidates can be found in the # MODEL CANIDATES section of the Approach.R file. These candidates were all worked on simultaneously and were ranked based on their RSE values from cv.glm() and whether or not the major assumptions were met. Ultimately the preferred model was chosen after a boxcox transformation on the response by square rooting the response. Then the variance inflation factors were decreased by switching out significant interaction terms with other known significant interaction terms. The chosen model can be found in the # CHOSEN MODEL section.