CHL5212H Assignment 2: Variable Selection

This assignment will cover the basics of variable screening, dimensionality reduction and model selection. The requirement is to train several models on a training set and select the best one using a validation set. Then retrain the selected model on the combined training and validation sets and obtain measures of generalizable error from the test set. Use the dataset provided and split off 50% for a test set, then split the remaining data into a training and validation set (2/3 and 1/3). Do the following:

Variable Selection: Evaluate the univariate predictive power of each variable. Generate performance metrics; residual error (RE) and standardized coefficient values, with 5 fold CV applied on the training set. Average across the 5 CV and get the variable ranks for both metrics (RE and coefficient). Plot the univariate performance, ordered by the rank, for both metrics. Using the RE metric, generate a short list (top 100 variables) and a longer list (CV RE <0.925).

Dimensionality Reduction: Apply principal component regression (PCR) to the longer list of your selected variables and select the optimal number of components to include by 5 fold CV on the training set. Use 5, 6, 7, 8, 10, 11, 13 and 16 components. Generate scores from the best PCR model for training and validation sets.

Model Building: Using the training set, apply forward selection using AIC to the short list, long list and PCR scores list. Use k=2 for short and PCR lists and k=3 for the longer list. Extend the forward selection to include first-order interactions for the selected main effects in the three lists. Use k=2 for short and PCR lists and k=4 for the longer list.

Fit a linear regression using each of the lists (short list, longer list, PCR list) x (initial list, AIC selection main effect, AIC selection main effects and first-order interactions) for a total of nine models and compute the MSE error on the validation set.

Selecting a Final Model: Identify the best model using the validation set MSE. Retrain each model on the combined training and validation set and estimate the error on the test set. Plot the results and explain how the selected model compares to other models in terms of performance.

Interpretation:

1. For univariate variable screening, why is RE more reliable (in a model training sense) than coefficient values (what is the extra step)? Do the variable ranks differ substantially? Instead of CV, what would've been an alternative method of splitting the data to reduce bias?
2. Comment on the univariate variable filtering, what are the pros and cons of the selected approach? Were these appropriate cutoffs for short-listing variables?
3. What can you say about the structure of data and the structure of the signal based on the univariate variable rank plots? Comment on the number of PCR components that were optimal as well as the changes in performance.
4. What could be done to further refine the signal from the models you obtained?
5. What were some of the feasibility constraints (e.g. computation time) on the implementation?
6. What are two advantages of step-AIC for variable selection? What are two disadvantages? Did variable ranks differ between non-PCR, step-AIC generated models?

Evaluation: Results 45%, interpretation 35%, coding style 20%. **Due: May 26, 2020**