**Energy efficiency Data Set**

The dataset is accessible from UCI - Machine Learning Repository. Copy of data set and its description are at <https://archive.ics.uci.edu/ml/datasets/Energy+efficiency>

The dataset contains eight attributes (or features, denoted by X1...X8) and two responses (or outcomes, denoted by Y1 and Y2).

The aim of this exercise is to use the eight features to predict each of the two responses (768 samples and 8 features).

**1. Inspecting the data table and providing descriptive statistics and checking for any NA's**

The variables:

X1 Relative Compactness

X2 Surface Area

X3 Wall Area

X4 Roof Area

X5 Overall Height

X6 Orientation

X7 Glazing Area

X8 Glazing Area Distribution

Y1 Heating Load

Y2 Cooling Load

Attribute characteristics: Numeric (double)

Count (sample size): 768

No NA’s identified.

**2. Describing Y1 (Heating Load) and Y2 (Cooling Load)**

Histogram for Y1 and its mean (22.31) which is greater than its median (18.95) indicates its distribution is right skewed. Application of Empirical rule indicates no outliers. Its maximum and minimum values are 6.01 and 43.10 respectively.

Histogram for Y2 and its mean (24.59) which is greater than its median (22.08) indicates its distribution is right skewed. Application of Empirical rule indicates no outliers. Its maximum and minimum values are 10.90 and 48.03 respectively.

**3. Inspecting correlation between the independent variables and dependent variables (Y1 and Y2) and multi-collinearity between the independent variables.**

Correlation > 0.8 = strong relation; Correlation < 0.4 = weak relation; else moderate

**Y1 (Heating Load)**

X1: moderate positive linear relation

X2: moderate negative linear relation

X3: moderate positive linear relation

X4: strong negative linear relation

X5: strong positive linear relation

X6: no linear relation

X7: weak positive linear relation

X8: weak positive linear relation

**Y2 (Cooling Load)**

X1: moderate positive linear relation

X2: moderate negative linear relation

X3: moderate positive linear relation

X4: strong negative linear relation

X5: strong positive linear relation

X6: no linear relation

X7: weak positive linear relation

X8: weak positive linear relation

The distribution for all independent variables is uniform except X3 and X4 which appear to be normal and left skewed respectively.

**4. Inspecting collinearity**

The following variables have collinearity (in descending order). Dealing with pairs with highest collinearity:

X1, X2 (-0.99)

Both variables have strong correlation and have opposing influence on dependent variables. X2 appears to have marginally stronger correlation with Y1 and Y2 than X1. Therefore, dropping X1

X4, X5 (-0.97)

Both variables have strong correlation and appear to be opposite of each other. X5 appears to have marginally stronger correlation with Y1 and Y2 than X4. Therefore, dropping X4.

X1, X4 (-0.87)

X1, X5 (0.83)

X2, X4 (0.88)

X2, X5 (-0.86)

After removing X1 and X4, collinearity between X2 and X5 as inspected. X5 appears to have stronger correlation with Y1 and Y2 than X2. Therefore, dropping X2.

**5. Splitting dataset into training (80%) and testing sets (20%)**

**6. Building model for Y1 - Heating Load**

After two iterations (dropping X6), the significant variables are X3, X5, X7 and X8. The model is overall statistically significant at 5% significance (p-value < 0.05) and the independent variables are individually statistically significant at 5% significance (p-value < 0.05). R-square is 0.92 which means approximately 92% of variation in Y1 can be explained by the variation in all the independent variables.

The model is: Y1 = -24.23 + 0.051 \* X3 + 4.76 \* X5 + 19.25 \* X7 + 0.24 \* X8

**Residual Analysis**

The residual plot indicates constant spread across the horizontal axis with no cone shape formations confirming residual linearity and equal variance. The residual histogram also indicates residuals appear to be normally distributed. The error terms across the horizontal axis do not appear to be related confirming independence of errors assumption is met. The normal probability plot is approximately a straight-line indicating errors are normally distributed.

**Making predictions and checking RMSE, error rate and R-square to check the model's performance**

RMSE is 3.32 representing an error rate of 14.2%. R-square is 0.89, which means the observations and predicted values are highly correlated. Residual standard error of the model is 2.94, which means the actual Y1 values in the model deviate from the true regression line by approximately 2.94 units.

**7. Building model for Y2 – Cooling Load**

After three iterations (dropping X6 and X8), the significant variables are X3, X5 and X7. The model is overall statistically significant at 5% significance (p-value < 0.05) and the independent variables are individually statistically significant at 5% significance (p-value < 0.05). R-square is 0.88 which means approximately 88% of variation in Y2 can be explained by the variation in all the independent variables.

The model is: Y2 = -15.84 + 0.04 \* X3 + 4.6 \* X5 + 14.8 \* X7

**Residual Analysis**

The residual plot indicates constant spread across the horizontal axis with no cone shape formations confirming residual linearity and equal variance. The residual histogram also indicates residuals appear to be normally distributed. The error terms across the horizontal axis do not appear to be related confirming independence of errors assumption is met. The normal probability plot is approximately a straight-line indicating errors are normally distributed.

**Making predictions and checking RMSE, error rate and R-square to check the model's performance**

RMSE is 3.66 representing an error rate of 14.3%. R-square is 0.85, which means the observations and predicted values are highly correlated. Residual standard error of the model is 3.22, which means the actual Y1 values in the model deviate from the true regression line by approximately 3.22 units.

Citation:

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014. An Introduction to Statistical Learning: With Applications in R. Springer Publishing Company, Incorporated.

A. Tsanas, A. Xifara: 'Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools', Energy and Buildings, Vol. 49, pp. 560-567, 2012