Student Number: 20243635

Regression and Cross-Validation

*ABSTRACT*

PURPOSE: Use linear regression to analyze the data, estimate the weight vector of the given "tall and thin" matrix A and vector C, to find the regression variables.

METHODS: Loading the data in code and find it size. Then standardize it so that improve the stability and convergence of algorithms and to prevent numerical instability in some statistical models. Calculate and find the smallest rms error. Read the test data in the file, find it size and standardize the data. Process folds. Create arrays for test set and training set. Compute the rms error for each fold in final.

RESULTS

The purpose of this experiment is to find the variables that are best explained by the variables in the reported data set. Reliability of this variable is obtained by a five-fold cross validation analysis.

CONCLUSIONS

According to the data provided and the evaluation of variables, linear regression model is an effective method to predict the value of these basic commodities. It is worth noting that although the data is scaled when encountering many data with different scales.

INTRODUCTION

PURPOSE

Find the variables that are best explained by other variables in the data set provided and obtain their reliability through 5-fold cross-validation analysis.

METHODS

Load the data into MATLAB, calculate and find the size and content, find the rms errors of linear regression. Get one column as dependent variables and other are independent variables. Use the same method, finds the rms errors of 5-fold cross-validation for the index of the smallest RMS error of the data.

RESULTS

Table 1: RMS Errors in standardized units

|  |  |
| --- | --- |
|  | RMS Error |
| Zinc | 0.2925 |
| WTI Crude | 0.2598 |
| Uranium | 0.2905 |
| Tin | 0.1589 |
| Copper | 0.1238 |
| Hard Logs | 0.4734 |
| Soft Logs | 0.4716 |
| Hides | 0.5478 |
| Lead | 0.2409 |
| Nickel | 0.2988 |
| Rubber | 0.2177 |
| Soft Sawn | 0.5834 |
| Fish Meal | 0.3258 |
| Cotton | 0.3938 |
| Coal | 0.3066 |
| Iron Ore | 0.2251 |
| Hard Sawn | 0.3652 |

Table 2: the index of the variable that is best explained by the other variables is 5 and the corresponding name is copper.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Fold1 | Fold2 | Fold3 | Fold4 | Fold5 |
| Rms train | 0.1164 | 0.1194 | 0.1196 | 0.1238 | 0.1275 |
| Rms test | 0.1649 | 0.2153 | 0.1541 | 0.1376 | 0.1168 |

DISCUSSION

The linear regression analysis of the "goods.csv" dataset found copper to be the best predictor, as determined by a 5-fold cross-validation evaluation method. The data was standardized using the zscore function in MATLAB to improve consistency and address the significant variability in the scale of the data. The analysis did not include an intercept term as it was deemed not logically feasible for a good to have a zero price. The training data showed a consistent trend over time, while the accuracy of the testing data showed some greater fluctuations, with a drop in the final fold.