Linear Data Analysis
Regression and Cross Validation

Cain Susko
20244352

Queen's University
School of Computing

February 10, 2022

Abstract

This report was created with the intent of deriving a procedure for generating a linear model for a data matrix of variables and observations. This was attempted by means of linear regression and k-fold cross validation on data from FRED within MatLab. This procedure yielded a 'best' dependent variable given the data as well as the Root Mean Square errors for training and validating a model for said dependent variable.

Introduction

The scientific purpose of this report was to demonstrate a procedure for modeling data using linear algebra. This report is based upon the math of linear algebra and focuses on linear regression, which is a linear approach to modeling the relation ship between a scalar response and one or more explanatory variables. Typically, it is used in training neural nets or other machine learning systems. The Scientific Question was on how one can use linear regression within MatLab to find the dependent variable c with the lowest root mean square error given a set of data from the U.S. Federal Reserve Bank of St. Louis Economic Data.

Methods

Pretaining to the first question assigned for this report: The data is first stripped of the observations and variables, and then concatenated to the empty matrix A. Once the required data is in A, it is standardized into X using an iterative method over the column vectors of A to find the zero mean, standard variance, and finally the z score of each column–which are stored in Z. Next, each column in X is treated as a dependent variable while the rest are treated as independent and a model is created using this format for each column in X. The root mean square is then computed for each of these models derived from X and the index with the lowest rms value is considered the variable of 'best' regression.
pretaining to the second question assigned for this report: the data is first imported and trimmed exactly the same as above and concatenated to the empty matrix B. This matrix is then used to derive A and c which are the independent and dependent variables for this question respectively. A is

then partitioned into k parts by using the MatLab function `randperm` which chooses from a list of numbers without replacement from 1 to the width of A for all columns in A. This list of random numbers is then used to select the columns that go into each partition of the matrix of independent variables X. Now, for each X in range $1 : k$, the training and validation data is created where each column in X and c are truncated to further partition the data into 2 sets. Finally, the model w is made with the training data and the rms error is calculated and returned for both the training and validation data.

Results

columns 1-8	0.31	0.12	0.40	0.33	0.48	0.37	0.55	0.23	
columns 9-17	0.24	0.30	0.22	0.47	0.59	0.16	0.29	0.26	0.29

Table 1: index = 2, commodity = Copper

rms training data	1047.73	1166.72	977.44	1232.59	791.09
rms validation data	604.44	610.10	942.88	2070.74	439.72

Table 2: index = 2, commodity = Copper

Discussion

From observing this data i can see that either it is incorrect or my models are very unfit for this data as, the rms error is very high (relative to examples we covered in the lectures). A reason for the unfitness of my model is perhaps because the data itself is irratic and perhaps linear regression is not suited for modelling this type of data. However this is also contradicted by table 1, as the rms error is much smaller than in table 2, implying that errors may have been made in the computation of table 2.