Finding best fit dependant vector and model effectiveness testing

Abstract

PURPOSE: Experiment with the method of linear regression to find the best fitting dependent vector, and testing that vector using a 5-fold cross validation, to test for quality.

METHODS: Using Root Mean Square Error as an indicator, test each of the column vectors, treating it as the dependent variable, finding the lowest resulting RMS. Then randomize and section the rows to perform the 5-fold analysis on the lowest RMS dependent variable.

RESULTS: The results indicated that the best dependent variable for the sample was Copper, and the 5-fold cross validation indicated that the sample was sufficient.

CONCLUSIONS: The linear regression was found and allowed for analysis on the elements of the goods. The 5-fold cross validation proved the linear regression was on a fitting data sample, in addition to indicating for this data set, there is no direct correlation between the RMS error and how well the sample creates a model.

Introduction

The purpose of this experiment is to analyze a set of financial data using the method of linear regression. Linear regression is the practice of creating a weight vector which represents a linear combination of the rest of the columns of the matrix. This linear combination creates a good approximation of the dependent vector, while only having information about the other independent vectors. This is useful because we can use this trained weight vector in order to predict the behaviour of the dependent variable as we are presented with new independent data. This can also be used to explore the relationships between one of the elements and the rest, if the RMS is high, it has a low correlation with the rest of the elements, and the converse is also true. If the RMS is low, it has a high correlation with the other elements.

Methods

In order to find the most fitting dependent variable, The data was first standardized in order to reduce biases and allow the data to be analyzed with constant units. The data was z standardized to create a normal relative measure that can be compared to each other. We lost the units however this is not an issue because the desired output is the index of a column, it does not matter the units. After the data was standardized, a model was created for each column, treating the specified column as the dependent variable, then comparing the RMS error of the model's generated value for that column subtracted by the column's actual value. The lowest RMS error is the best fit for the dependent variable and is then marked. After finding the model, It was plotted for visibility, and tested to see if it was able to predict a new piece of data, the mean price of the copper. This data was obtained by creating a new row vector that contains the means of all the columns, then the weight vector was applied to the row vector and the mean of the dependent variable should be outputted.

In order to test if a newly found dependent vector is a good representation of the data, without access to extra materials, we will perform a 5-fold cross validation. Due to the fact that the data is sorted by date, we must take that into account for the choosing of the subsets of data for the cross validation, the data must be shuffled before it can be used. This will be done by taking a random permutation. After randomly shuffling, The data not standardized in order to keep the units, instead an intercept vector was added to account for the lowest number, as the data does not go directly to 0, and thus an intercepted linear regression can more accurately represent the data, representing the y intercept. After the intercept was added, the data was split into 5 groups. One of the groups was selected to be the testing group, the rest was the training. A weight vector was created using the training data, then the RMS was calculated in dollars when tested onto the testing data. This will be looped until all groups have the testing data. This data will be outputted to a table. We will also test not only the best fitting dependent vector, but also the worst fitting dependent vector in order to test if the quality of the dependent vector has a correlation to the quality of the fittiness between the vector and the sample.

Results

Calculated Copper Mean Price = 4610.7
Predicted copper mean price using the weight vector = 4613.1

Table 1: Displays the RMS error for each of the columns when linear regression is applied

Col #	1	2	3	4	5	6	7	8	9
RMS	0.2925	0.2598	0.2905	0.1589	0.1238	0.4734	0.4716	0.5478	0.2409
Col #	10	11	12	13	14	15	16	17	
RMS	0.2988	0.2177	0.5834	0.3258	0.3938	0.3066	0.2251	0.3652	

Table 2: Displays the Testing and Training RMS's for the groupings in the 5-fold cross analysis

Group	Copper Training RMS	Copper Testing RMS	Difference	Hides Training RMS	Hides Testing RMS	Difference
1	296.0262	414.2819	118.2557	9.2324	11.2281	1.9957
2	312.9995	380.6417	67.6422	8.7425	12.8669	4.1244

3	301.8077	400.7359	98.9282	9.7248	8.9953	-0.7295
4	308.1973	359.1201	50.3124	9.6866	9.1755	-0.5111
5	309.8087	385.8537	76.0450	9.0209	12.3889	3.368
Total	NA	NA	411.1835	NA	NA	10.7287
Mean	305.7679	388.1267	82.2366	9.2815	10.9309	NA
Standard Deviation	6.8028	20.8905	15.4802	0.4247	1.7882	NA

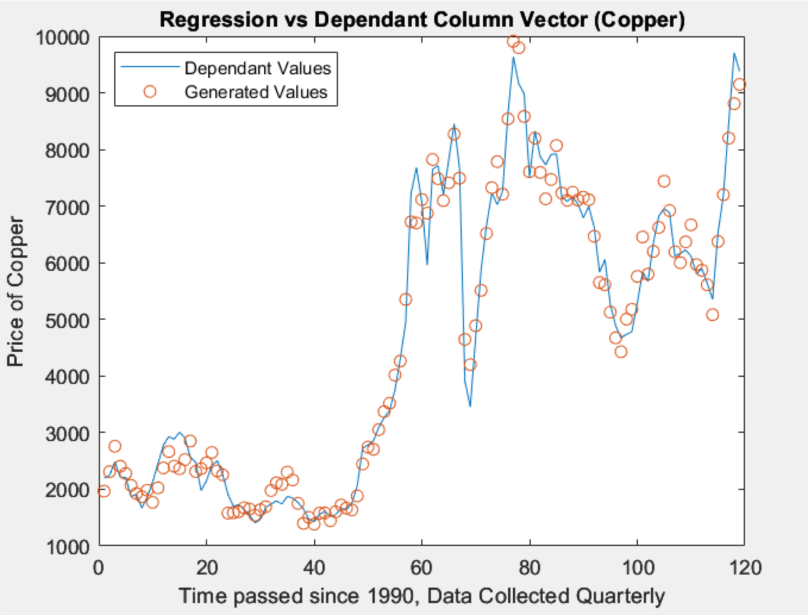

Figure 1:A chart depicting the relationship between the chosen best dependant value (blue) and the values generated by the linear regression (red).

Discussion

The linear regression generated by the method resulted in the best representation for a column that can be created was for the Copper column. This is shown by the fact that copper in the column has the smallest RMS compared to the rest, meaning that overall it had the most accurate predictions when compared to the actual vectors. Thus when the vectors get projected into the vector space of the matrix, the sum of the error vectors is the smallest. Using the data we was able to create a prediction of a “new” piece of data, that was only 2.4$ off of the actual calculated, less than a 1% difference, however due to the fact that both the weight vector and the mean are created from the same data set, it is skewed to be more correct.

Compared to the rest of the RMS's there were only two options less than 0.2 RMS, looking at column 4, it had a 0.158, this shows that both column 5 and 4 have a strong relationship with the other materials on the list. This shows that the price of these materials rely heavily on the price of the other materials showing that these prices fluctuate more than other prices and change rapidly with the market. This information can also be used conversely, the highest RMS's have a low connection to the market, column 8 and 12 have the two highest, this corresponds to them having the most static prices, not shifting with the market.

When the data was tested using the 5-fold analysis, the data was sufficiently shuffled due to the algorithm, and thus the error of the data being sorted by time is not an issue. On average, the RMS was around 300$ off compared to the price of 3-10 thousand, which was always less than 10% off. When comparing the results of the Copper training and testing data, and the Hides(The dependant vector with the greatest RMS) training and testing data, the fittingness of the model was very similar, the percent difference was also less than 10% error, the mean was proportionally similar to the Copper data. This leads to the indication that the size of the RMS error between predicted vector and the actual vector and the fittingness of the data has very little correlation. So with how well the model fits the data, does not affect whether or not the sample has the capability to create a good model. This is interesting as the linear regression is defined by the sample it is derived from. The model was consistently better on the training than it is on the testing, proving a little bit of overfitting, however the difference is not large showing that the model is wide enough to well fit the dependent variable. Gathering the mean and standard distribution between the mean difference between the Testing and Training is 82$, a relatively low number when the copper numbers are 2000$ and above. This low difference between the Testing and Training indicates a good sample for the model. This indicates a good model because the RMS is similar between training and testing, indicating that the sample is indicative of the whole population.