APPLICATION OF LINEAR REGRESSION ON FINANCIAL DATA

ABSTRACT

PURPOSE: To understand the applications of linear regression and how it can be used to describe relationships between data.

METHODS: A linear regression and 5-fold cross-validation are performed on twenty-eight years of quarterly commodity pricing to assess the commodity that is best described by the remaining data set.

RESULTS: Out of the given financial data set, copper was found to have the lowest RMSE score when used as the dependent vector and has an approximate error of 8% in estimating copper's price based on the other commodities as compared to copper's average price.

CONCLUSIONS: Out of the given financial data set, copper is found to be the commodity that is best described by the remaining commodities, which parallels how copper can be used as a proxy for other resources in financial markets.

INTRODUCTION

On stock markets, commodity values can be difficult to predict when analyzing them individually. Due to this, traditionally, sustainable commodities that are rare and have high demand are used as a proxy for other commodities. Copper is an example of one these resources. Due to copper's demand in the industries such as the technological and construction industries, it is priced high and has an inelastic demand [1]. Therefore, copper can be used as a proxy, or a commodity that is described by and describes the trend of other commodities. However, it is unclear that copper can maintain this property regardless of other commodities relation to it.

Using the linear algebra taught in CISC 271, specifically the lectures outlined through Week 5 & 8, implement the core concepts of vector projection, linear regression and k-fold cross-validation. Furthermore, understand how these concepts can be applied to applications outside of linear data analysis, in this case financial data analysis. Linear regression is the process of finding a "line of best fit", or a linear relation that will approximate the data given. The process for finding a linear regression builds off of the concept of vector projection [2], and a detailed explanation be found in the CISC 271 notes [3].

The data within the study does contain missing values in a few of the variables, these missing values are dealt with by repeating the previous recorded value to that quarter. This is done so that the missing data does not impact the regression in a meaningful way. Rather than estimating the trend of the value, the value is assumed to have been stable, which will reduce the skew in favour or against the commodity. Leaving the value at zero will describe an inaccurate fluctuation of the price and taking the average of the next data points may be ineffective if the remaining values for the commodity are missing. Therefore, when the value of a commodity is missing, the value repeats to the latest non-missing data, if it is the first data point then it assumes the value is zero.

Moreover, when validating the findings of the initial linear regression using k-fold cross validation [4], the study uses random fold selection in the cross-validation. Typically, when dealing with time sensitive data, data forwarding is used as a means of validation, in order to predict future values using previous data. However, in this case, although the data is time sensitive, the purpose of the validation is not to validate how well the regression can predict the future, rather to validate how well one commodity can be described by the others. Therefore, data forwarding would not validate the regression in terms of this study's purpose, that is why cross-validation is used. Furthermore, randomly arranging the data in the fold so it is not in chronological order is another tool to eliminate bias. The regression demonstrates that at any point in time, one commodity can be described by the others, therefore chronological order would introduce a bias. For example, the regression might be predicting the commodity price not how well it is characterized by other prices. That is why the validation of the linear regression is done with a random fold selection.

The objective of the report is to understand how linear regression can be applied in understanding the relationship between data set, with a focus on financial data. The hypothesis that this study proposes is given the financial data on a select subset of commonly traded basic commodities with copper as a vairable, copper can be used as a proxy for other commodities.

METHODS

The experiment is conducted using MATLAB R2019b. In MATLAB, the three primary functions used, *a3q1*, *a3q2* and *mykfold*. The data for the study was gathered from the U.S. Federal Reserve Bank of St. Louis Economic Data (FRED) website and placed into a CSV file. The data variables are world prices of a select group of commonly traded basic commodities. The prices are recorded by their quarterly date starting in January 1st, 1992 and ending in October 1st, 2019. The program is run and tested through the MATLAB console at which results, as outlined in **Table 1** and **Table 2**, are printed to the console as well as the plot being displayed in a separate window.

Function *a3q1* finds the RMS errors of linear regression of the data by treating each column as a vector of dependent observations, using the other columns of the data as observations of independent variables. It first reads in the CSV file into a 112 x 16 matrix, *Amat*, and the commodity headers into a one-dimensional array, *labels*. The matrix is then standardized. A for loop is used to iterate through each commodity as the dependent variable and stores its RMSE into a vector, *rmsvars*. On every iteration a different dependent observation is put into a vector, while the remaining observations are gathered into a matrix. Linear regression is performed by multiplying the independent variable matrix by the weight vector, as outlined in the CISC 271 notes [3]. The weight vector accounts for how much data each independent variable can describe the dependent variable at any given time. Once the regression calculates the projected vector, *p*, the root-mean square error (RMSE) is calculated, as outlined in the notes [3]. The RMSE calculates on average, how accurate the regression is to the dependent variable, the closer the RMSE is to zero, the less amount of errors. Once all the commodities have been chosen, the index of commodity with the lowest RMSE is saved as *lowndx*. The same linear regression, as described previously, is performed again on the unstandardized data for a comparison with units. Using *lowndx* as the index of the dependent variable, *yvec*, the remaining independent variables are gathered into a matrix, *Xmat*. The projection of *yvec* onto *Xmat* is plotted to **Figure 1**.

Function *a3q2*, finds the RMS errors of 5-fold cross-validation for the variable *lowndx*. The data for the matrix, *Amat*, is inputted using the same process as in the function *a3q1*. The linear regression on the dependent vector using *lowndx* as the index for the commodity is performed to create the vector *yvec*. The remaining independent vectors are gathered into a matrix, *Xmat*. The vector and matrix, *yvec* and *Xmat*, are standardized and passed through to the final function *mykfold* along with *yvec*'s standard deviation so the RMS score is in USD, and returns the vectors, *rmstrain* and *rmstest*, as outlined in *mykfold*.

The final function, *mykfold*, creates and performs a k-fold validation of the least-squares linear fit of *yvec* to *Xmat* as outlined in the notes [4]. In this case *k* was chosen to be 5. The *yvec* and *Xmat* partitioned into five subsets where in every cycle of cross-validation, four are used to train and the fifth is used to validate the regression. Furthermore, *yvec* and *Xmat* are randomly reorganized such that the rows stay intact, but the order of the rows is not chronological. A for loop is used to perform the cross-validation. On every iteration, a weighting vector¸ *wvec*, contains the weights for the regression's training sets. The RMSE, using *wvec* is calculated for each fold and multiplied by the standard deviation of *a2q3*'s *yvec* vector, to give the RMSE units in USD for both the training and testing sets stored in two vectors at their respective fold indices, *rmstrain* and *rmstest*, respectively. The vectors, *rmstrain* and *rmstest*, are then returned.

RESULTS

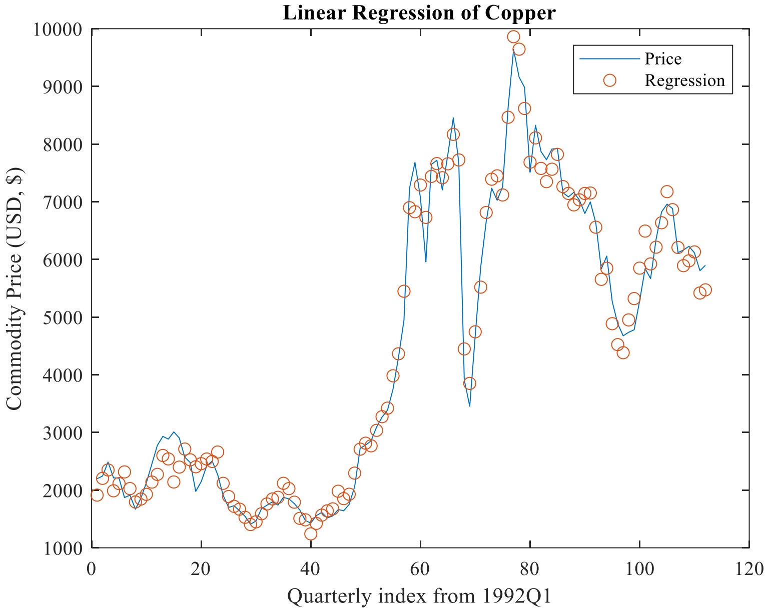

Figure 1: Plot of the second indexed commodity, copper, with the quarterly index as the independent axis and its price as the dependent axis. The actual value of copper is plotted as the blue line while the linear regression of copper are the orange circles.

Table 1: The root-mean squared error of every commodity; every commodity is given their index, their name and RMS score, respectively. All RMS scores are unitless due to the standardization of the data.

Index	Commodity	RMS Score
1	*Coal*	0.2812
2	*Copper*	0.1148
3	*Cotton*	0.3826
4	*Fish Meal*	0.3099
5	*Hard Logs*	0.5592
6	*Hides*	0.6356

7	*Iron Ore*	0.2436
8	*Lead*	0.2291
9	*Nickle*	0.2957
10	*Rubber*	0.2042
11	*Soft Logs*	0.4927
12	*Soft Sawn*	0.6065
13	*Tin*	0.1596
14	*Uranium*	0.2882
15	*WTI Crude*	0.2351
16	*Zinc*	0.2617

Table 2: The root-mean squared error of the training and the testing sets at each fold of the second indexed commodity, copper, in USD $.

RMS Value Set	Fold #1	Fold #2	Fold #3	Fold #4	Fold #5
Training	267.7799	294.0157	295.8872	264.5635	271.0467
Testing	377.0607	277.3516	262.3180	374.8206	389.6855

DISCUSSION

The hypothesis that the study poses, is if that given financial data of commodities, can copper be used as a proxy commodity for the data set. The commodity that is found to be best described by the other commodities when standardized is the second indexed commodity, copper. Standardization was chosen because this section experiment’s goal was to find the commodity that best is described by the other commodities. Many of the commodities have different initial positions at the first quarter of 1992 and have very different variances. Comparing their prices would not accurately describe the impact that one commodity can have on the other commodities. For instance, commodities like Iron Ore are priced in the tens of dollars with a variance in the dollars and commodities like Tin are priced in the thousands with a variance in the hundreds of dollars. A comparison between the two would not yield any meaningful conclusions to this study. Therefore, when the data standardized, the commodities are zero-meaned, the prices become unitless and reduced to a measure of deviation from their average. Therefore, when we compare commodities like Iron Ore and Tin, it produces a more objective comparison; rather than comparing prices, the general trends and variation from the average can be compared. Furthermore, no intercept term is used in any of the matrices due to the standardization of the data, therefore a bias in terms of an intercept column in *Xmat*, would not provide any meaningful addition to the results of the experiment. The results show that copper’s general trend line is best described by most of the other commodities trend lines, when the data is standardized. **Table 1** shows the RMS values for each commodity when they are treated as the dependent variable. A lower RMS score means that there is little error at any given time between the dependent and regression vector, thus a more accurate description of the other commodities. Copper is found to have the lowest RMS score of all the commodities; thus, it is the index that can be used to is best characterized by the other commodities. Note, all RMS scores in **Table 1** are unitless due to the data is standardization. Furthermore, **Figure 1** is a plot of the copper, with the quarterly index as the independent axis and its price as the dependent axis. The blue line represents its price at a given quarter, and the orange circle represent the regression’s estimation of copper’s price. The plot demonstrates that although the regression isn’t completely accurate, it does however give a valid approximation of the price. The regression follows copper's general trend correctly and further supports the hypothesis' claim.

Before reaching any conclusions from the first experiment, the results must be validated with another experiment run through the function *a3q2*. Using the index found in the first experiment, in this case the second index, the data was standardized and ran through 5-fold cross validation. The purpose for standardization is, for reasons similar stated above, because the study compare the data's value curves, not the values itself. The multiplication of the standard deviation of copper is necessary as it reverse the standardization process and ensures that the RMS score has units. The results from the each of the five sets acting as the testing set, while the remaining being the training are outlined in **Table 2**. The RMS scores are given for both the model against the trained data, and test data. The results show that the regression is off by about $278.66 on the training data and $336.25 on the test data, which is an acceptable RMS considering that copper is measured in thousands of dollars and has a range of $1 000 to $10 000. Therefore, the regression has an error of about 7.59% on test data when compared to copper's average price. Thus, copper is described by the trendline of every other given commodity with an accuracy of about 92%. Although it may seem quite accurate, there are other biases that affect the accuracy such as the data being overfitted. If the data was trained, for example, on the years from 1992 to 2014, then validated on the years 2015 to 2019, it would give a better estimation on the error. However, because the data was trained on the entire data set, there is concern that the data is overfit, even with the random partitioning, and thus gives a better RMS score.

Nonetheless, these two experiments show that with the commodities given, copper can be used as a proxy for the other commodities. Copper's price curve best mimics the price curve, on average of other resources such as Iron and Tin. The same can not be said for copper's value. The accuracy at which copper is described by the values in the data set given is about 92%, however, further studies need to be taken to prove effectiveness of using copper as a proxy for commodities outside of this data set. These findings reflect how commodities such as gold and copper are used as a basis for currency and value for other resources. Thus, by understanding and predicting how commodities such as copper react in certain financial markets, by proxy, one can understand how an unrelated resource such as Soft Logs will react as well. It is a more efficient way of understanding the market, and thus why other commodities values trends will reflect that of copper's, even those of weaker relation.

In conclusion, the results show that with the given data set, copper is proven to be the commodity that is best characterized by the values of other commodities at a given time. This reflects how copper is used as a proxy in the stock markets and validates the study's hypothesis.

REFERENCES

[1] Martin Stuermer: Industrialization and the Demand for Mineral Commodities, 2017

[2] Dr Randy E Ellis: CISC 271 Class 14 Graphs - Projection into a Subspace, 2020

[3] Dr Randy E Ellis: CISC 271 Class 15 Graphs – Patterns: Least-Squares, 2020

[4] Dr Randy E Ellis: CISC 271 Class 22 Graphs - Cross-Validation of Linear Regression, 2020