Linear Regression and K-fold Cross Validation

Matthew Sun

1 Abstract

The purpose of this assignment to use linear regression on a given data to find the variable with the lowest RMS error when chosen as the sole dependent variable and perform k-fold cross-validation on the regression model to evaluate its fit to the data set. This is done by finding the weight vector for each column of a the data matrix which provides an approximate solution to the overdetermined system. The RMS error is the root mean square of the residual error between the approximate solution and the actual values. k-fold cross-validation is performed by taking out one partition as the testing set, training on the remaining data, and then comparing the results across each fold to determine the fit. It was found that copper is the least RMS error variable but its regression model performed poorly in a 5-fold cross-validation.

2 Introduction

The objective of the assignment was to find the variable in a given data set containing commodities and their prices with the smallest root mean square (RMS) error when linear regression is applied to said variable as the dependent variable while all other variables in the data set are considered independent variables. In addition, the reliability of the variable with the smallest RMS error as a proxy for all other variables is evaluated using 5-fold cross-validation. To compensate for varying magnitudes in a given data set, calculate standardized matrix $X = [A - [\vec{1}\bar{A}]]D_A^{-1}$, where A is the data matrix, $\bar{A}$ is the zero mean matrix with entries of vector $\bar{a} := \frac{1}{m}\sum_{i=1}^{m} a_i$, D_A is the diagonal standard deviation matrix of A with entries $d_{ii} = \sigma_{a_i} := \frac{||\vec{a_i} - \vec{1}\bar{a_i}||}{m-1}$. This is equivalent to treating every column as a normal distribution and assigning every entry its corresponding Z-score.

Linear regression is the process of projecting a dependent variable vector $\vec{c}$ onto a separate subspace $\mathbb{V}$ whose basis vectors are defined by the columns (variables) of a data matrix $A \in \mathbb{R}^{m\times n}$. This projection is the vector $\vec{p} := A\vec{w}$ where $\vec{w}$ is the weight vector which contains scale each column vector of A, $\vec{a_1}, \vec{a_2}, ..., \vec{a_n}$ such that $\vec{p} \in \mathbb{V}$ is geometrically the closest to $\vec{c}$. As such, $\vec{p} := w_1\vec{a_1} + w_1\vec{a_1} + ... + w_n\vec{a_n}$. The goal of finding $\vec{w}$, and thus $\vec{p}$, is to minimize the residual error between the projection $\vec{p}$ and the vector $\vec{c}$. The distance between the projection $\vec{p}$ and the actual vector $\vec{c}$ is denoted as the error vector $\vec{e}$. This is a measure of the distance between the projection and the actual vector. Each entry of $\vec{e}$ is a residual error $e_i := c_i - p_i$. This error vector $\vec{e}$ will be orthogonal to the basis vectors of A, thus $A^T\vec{e} = \vec{0}$. This property can be used to find the normal equation for the data matrix A which describes how to find the $\vec{w}$ that projects $\vec{c}$ into A.

$$A^T\vec{e} = \vec{0} \tag{1}$$
$$A^T(\vec{c} - \vec{p}) = \vec{0} \tag{2}$$
$$A^T\vec{c} - A^T\vec{p} = \vec{0} \tag{3}$$
$$A^T\vec{p} = A^T\vec{c} \tag{4}$$
$$A^T[A\vec{w}] = A^T\vec{c} \tag{5}$$
$$[A^TA]\vec{w} = A^T\vec{c} \tag{6}$$

Since data matrix A has full column rank then the left transpose product $[A^TA]$ will be a symmetric positive definite matrix. Thus, $[A^TA]$ is invertible and there is a unique solution for equation (6). In many cases, the system is overdetermined (more rows than columns) and this formula will provide an approximate solution $A\vec{w} \approx \vec{c}$. Knowing this, $\vec{w}$ can be written as $\vec{w} = [A^TA]^{-1}A^T\vec{c}$.

A linear regression model can be assessed using k-fold cross-validation. Given a matrix of independent variables A and a dependent variable $\vec{c}$, A is partitioned into k sets which will be denoted as $S_1, S_2, ..., S_k$.

Then cross-validation is performed k times where for each S_j, the regression is trained of all data sets $i \neq j$ and the trained regression is tested on S_j. The effectiveness of regression on each testing set S_j can be evaluated using the RMS error. The weight vector $\vec{w}$ provides an approximate solution to $A\vec{w} \approx \vec{c}$ is assessed using the RMS error. The RMS error is calculated using the data matrix A, the dependent variable $\vec{c}$, and the weight vector $\vec{w}$. The RMS error is defined as $RMS(A, \vec{c}, \vec{w}) := \frac{||\vec{e}||}{\sqrt{m}}$, where $\vec{e}$ is the error vector of $A\vec{w}$ and $\vec{c}$ and m is the number rows (observations) in A.

The scientific question to be explored is: What variable in the given data set is best explained by all the other variables and how does the regression model for said variable perform in 5-fold cross-validation? After each variable in the given data set is used as the dependent variable and a linear regression model is found, the variable with the lowest RMS error value will be chosen as the variable that is best explained by the other variables. 5-fold cross-validation will then be performed with this variable as the dependent variable. The effectiveness of the linear regression model will be evaluated by comparing the RMS error of the training set against the RMS error of the testing set for each fold of data. If the RMS error of the training set is significant lower than the RMS error of the testing set, it could be said the model is over-fitted to the training data and a poor model for the test data. If there is no significant difference between the mean and standard deviation of the training RMS errors and the testing RMS errors, it could be said that the model is a good fit for the data.

3 Methods

Let A be the given data matrix containing the prices of commodities, where the columns are the variables (commodities) and the rows are observations of prices separated by three-month intervals. The choice was made to standardize the data due to the varying degrees of magnitude in the prices amongst the commodities. The intercept form was not used because it is redundant to apply both methods. The "zscore" Matlab function was used to perform the necessary calculations to standardize the data. Going forward, A will be a standardized data matrix.

The RMS error of each variable when used as the dependent variable while all other variables are considered independent variables are found by iterating through every column of the standardized data matrix $a_1, a_2, ..., a_n$; letting column $\vec{a_i}$ be the dependent variables and creating another matrix containing all other columns of A which will be the independent variables. The weight vector was calculated using the using the Matlab function "\" which performs a similar calculation to the normal equation although it should be noted there are subtle differences. The projection vector was calculated by multiplying A with the weight vector. The error vector was calculated by subtracting corresponding values of the dependent variable vector from the projection vector. The RMS error was calculated using the "rms" function in Matlab. The RMS error value is then stored in a row vector. This process is repeated for all columns of A. After all column vectors have been used as the dependent variable, the index of the variable corresponding with the smallest RMS error is saved. The regression of the unstandardized version of the variable with lowest RMS error was then plotted alongside its original data. The understandardized version was used because it provides more clarity over the price compared to using standardized units.

5-fold cross-validation was then performed using the variable with lowest RMS error as the dependent variable while collecting all other variables into a matrix which will be used as the independent variables. The data used for cross-validation are unstandardized to keep the magnitude of the price values relevant and what effects it may have on the RMS error values. The number of observations in each fold was calculated by taking the number observations and dividing it by $k = 5$ and rounding the quotient to a whole number which will be denoted as n. A random generation seed is set for consistent results and for each index j from 1 to $k = 5$, the following process is followed. First, the row of the dependent variable and independent variables are shuffled. The choice was to shuffle the observations since the prices of the commodities tend to rise over time and thus shuffling the rows should produce a better result. The index of the start of the testing set is $(j-1) \cdot n + 1$. The index of the end of the training set is the start index $+ n - 1$. There is also a conditional statement to check if this end index has surpassed the number of rows and if it does the

end index will bet set to the number of rows. Two logical arrays were then created where one contained all false except the indices included in the start and end index, this array keeps track of which observations to use for testing and the other contains all trues except the indices of the testing set are set to false, this will keep track of the indices of the training set. Using the "\" function in Matlab, the weight vector of the training set is calculated. This weight vector is then used to find the projection for training set and testing set. The RMS error for both sets are calculated using the "rms" function in Matlab and stored in separate row vectors. The mean and standard deviation for each set of RMS errors are also calculated.

The plot of the undstandardized regression of the variable with the lowest RMS error is then evaluated from visual observation to see how closely the regression model fits with the original data. The difference between the RMS error values from the training and testing sets of the 5-fold cross-validation are then observed to evaluate the performance of regression model. Judgements of the difference between the training and testing set can be made using the mean and standard deviation of each set of RMS errors.

4 Results

Table 1: Each commodity and the associated RMS error, index 5 (copper) has the lowest RMS error

Index	Commodity	RMS error (standardardized units)
1	Zinc	0.2925
2	WTI Crude	0.2598
3	Uranium	0.2905
4	Tin	0.1589
5	Copper	0.1238
6	Hard Logs	0.4734
7	Soft Logs	0.4716
8	Hides	0.5478
9	Lead	0.2409
10	Nickel	0.2988
11	Rubber	0.2177
12	Soft Sawn	0.5834
13	Fish Meal	0.3258
14	Cotton	0.3938
15	Coal	0.3066
16	Iron Ore	0.2251
17	Hard Sawn	0.3652

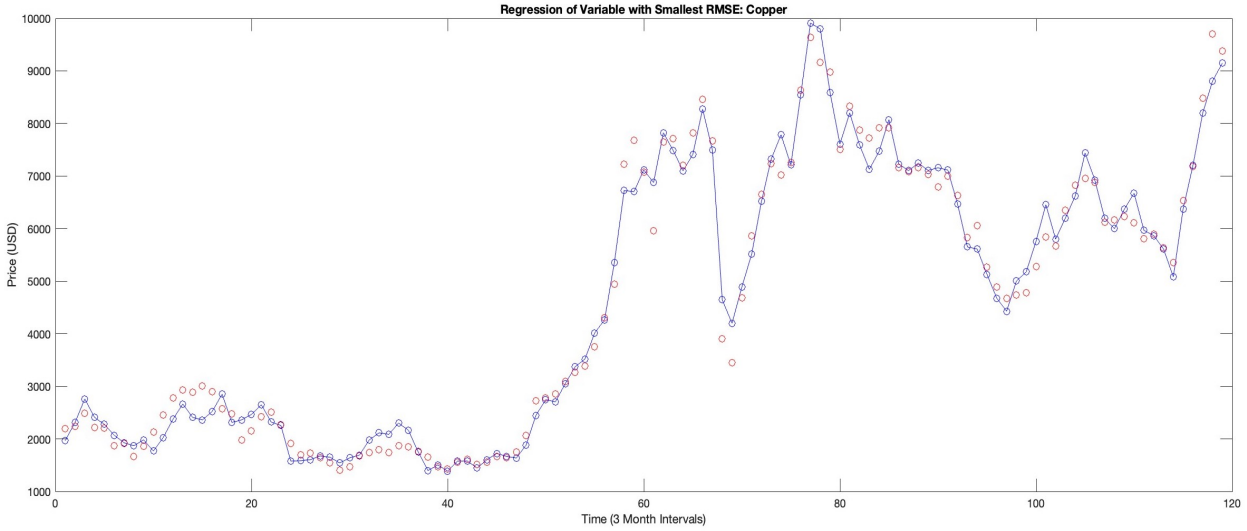

Figure 1: Unstandardized regression of variable with lowest RMS error: index 5 (copper), original data in red, regression model in blue

Table 2: Resulting RMS errors of training and testing set from 5-fold cross-validation of variable in index 5 (copper)

Training RMS errors (USD)	**Testing RMS errors (USD)**
299.0770	421.4926
299.9267	627.9161
304.7132	414.1664
315.7226	352.0761
309.6467	396.9051
Mean = 305.8172	Mean = 442.5113
SD = 6.9631	SD = 107.0965

5 Discussion

For the given data set of commodities and their prices over time, copper was the variable that was best explained by all other variables when chosen as the dependent variable for linear regression with an RMS error of 0.1238 standardized units. From visual observation of the plotted undstandardized data and regression, linear regression seems to provide a good model for copper. After performing 5-fold cross-validation, it was found that the mean of the training RMS errors is 305.8172 USD and the mean of the testing RMS errors is 442.5113 USD. The standard deviation of the training RMS error was found to be 6.9631 USD and the standard deviation of the testing RMS was 107.0965. Based on the evaluation requirement that for the model to be a good fit for data the mean and standard deviation between the training and testing set must be relatively close, this unstandardized model is a poor fit for the data set as there is large difference between the mean and standard deviation of the test set is significantly higher than that of the training set. Notice that the standard deviation of the testing RMS errors are more than 15 times higher than that of the standard deviation of the training RMS errors. It can be said that the model is overfit to the training data as it performs much better on the training data than it does on testing data.

Despite Figure 1 appearing to show that the regression model for copper fits very well with the given data set, it still performed very poorly on the 5-fold cross-validation. This could be due to the nature of the data being an ordered data set. Patterns exist across time but shuffling those observations will disrupt the patterns that exists in the data which could help to explain the discrepancy between how well the regression model seems to fit in Figure 1 and how the model overfits to the training data in the 5-fold cross validation. Further testing using a different method of selecting folds such that it preserves the order of observations to some degree should performed.

Another limitation for the 5-fold cross validation is that the data used was unstandardized and the trial was performed with a preset random generator seed. A trial was run with standardized data and the same random generator seed. The difference between the training and testing RMS errors appear to be much smaller. Standardized training set RMS errors: 0.1170, 0.1185, 0.1202, 0.1217, 0.1221. Standardized testing set RMS errors: 0.1615, 0.2246, 0.1513, 0.1431, 0.1498. The differences are still not insignificant despite not influenced by commodities with price values that orders of magnitudes of higher than others. This suggests that the model may not be as overfitted to the testing data as the unstandardized 5-fold cross-validation suggests but for the most part it is a poor fit. The largest difference between training and testing RMS errors are found between folds one and two. It could be that by chance, much of the data that are closer to being outliers were shuffled into those folds which drastically increased their RMS errors. Observe that even for the 5-fold cross-validation done using standardized data, folds one and two still have the largest difference between the training and testing RMS errors. It can be said that this increase most likely cannot be attributed to the varying magnitudes of price values amongst the commodities since standardizing the data would eliminate these discrepancies. Further testing using k-fold cross validation with different random generator seeds and different number of folds should be performed.

There are also other commodities that have relatively low RMS errors that are explained well when chosen as the dependent variable. For example, the commodity with the second lowest RMS error is tin (0.1589 standardized units). There is not a significant difference between the RMS error of copper and

tin and one would expect that tin would also be a good proxy for all the other commodities in the given data set. This is an interesting observation because both these commodities are metals that are very commonly used, especially the commodity with the lowest RMS error: copper. The results presented in this result suggest that copper is a good indicator for the performance of an economy. Copper is popular as a commodity that economists and those in commodities markets have coined the term "Dr. Copper" do describe the effectiveness of copper as indicator for economic health. In a 2021 report [CDA21] created by the "Copper Development Association", estimates that 46% of global copper goes to building construction, 21% to electrical, 17% to consumer products and machinery, and 16% to transportation. Copper's wide use across a multitude of industries makes it a fitting indicator for economic health. From an economic perspective, this partially explains why copper is a suitable proxy for all other commodities provided in this data set.

For the given data set it was found that copper is the variable with the lowest RMS error when considered the dependent while all other variables were used as independent variables. When 5-fold cross-validation was performed on the undstandardized form of the data with copper as the dependent variable and all other variables as independent, it was found that the model performs poorly as it overfits to the training set. From an economic perspective, it is reasonable that copper would be a the most suitable proxy as it widely used.

References

[CDA21] CDA. *Annual Data 2020 - copper*. 2021. URL: https://www.copper.org/resources/market_data/pdfs/annual-data-book-2020_final.pdf.