**Data Analytics Capstone Topic Approval Form**

**Student Name:** Michelle N.

**Student ID:** Click here to enter text.

**Capstone Project Name:** Ridge Regression and Time Series Analysis on Silver Pricing Dataset

**Project Topic**: Predictive Model for Commodity Market Trends

**X This project does not involve human subjects research and is exempt from WGU IRB review.**

**Research Question:** Can a ridge regression or time series model be constructed on the silver pricing dataset?

**Hypothesis**: **Null hypothesis**- A predictive ridge regression model with a model accuracy greater than 70% cannot be constructed on the market research dataset. **Alternate Hypothesis**- A predictive ridge regression model with a model accuracy greater than 70% can be constructed on the market research dataset.

**Context:** The contribution of this study to the field of Data Analytics and the MSDA program is to create a predictive model, regression or time series, that can estimate the price of silver in USD so that a solar panel manufacturing company can proactively manage their raw material costs. Silver is an important raw material for manufacturing solar panels (*Silver and Solar Technology*, n.d.). This study will utilize two approaches to attempt to forecast silver prices. The first is to create a ridge regression model using LBMA data on silver prices and macroeconomic data from FRED to analyze the significant predictor variables for the price of silver in USD. A previous study from the University of Melbourne, Australia found that using ridge regression is viable as a way to predict commodity prices from macroeconomic data and performs better than other regression methods because it reduces the effects of parameter estimation error by using shrinkage (Gargano & Timmermann, 2014). The second approach will be to use an ARIMA model based solely on LBMA data. A previous study conducted by Kriechbaumer et al. concluded that ARIMA models “perform marginally better and are a useful tool for […] companies to predict metal prices” (Kriechbaumer et al., 2013). Predicting the price of silver is crucial because macroeconomic factors and historical trends in the price of silver can be used to predict the future price of silver, allowing the cost of raw materials to be minimized and the bottom line to be maximized.

**Data:**

The base dataset was sourced from Kaggle, but the data itself comes from the LBMA, also known as the London Bullion Market Association. The Kaggle provides the date, silver USD, silver GBP, and silver euro columns in the table below. The base dataset from Kaggle is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0,) which allows the use of the dataset as long as appropriate credit is given, a link to the license is provided, the data is not used for commercial purposes, and no derivatives of the dataset are shared (*Deed - Attribution-NonCommercial-NoDerivatives 4.0 International - Creative Commons*, n.d.). The rest of the columns come from FRED, also known as Federal Reserve Economic Data, where one dataset provides each of the remaining five columns. The FRED website’s terms of service indicate that datasets can be used so long as proper attribution is included and the purpose for dataset use is non-commercial in nature (*Legal Notices, Information and Disclaimers*, n.d.). The pricing of silver is recommended for study by CAPEX, a firm that performs similar research (Cochintu, 2024).

These datasets were combined into one larger dataset as shown below, which has 19,834 cleaned rows total representing a date range starting from the earliest date in any of the six original datasets and ending with the latest date in any of the 6 original datasets. A link to this dataset can be found here: <https://westerngovernorsuniversity-my.sharepoint.com/:x:/g/personal/mnel522_wgu_edu/EX74iym0Hl5MtdvuhU9EXBcBUs2U164UFvqV8FWEATg2Pw?e=oBfuUs>

The variable types in the combined dataset are broken down as follows:

|  |  |  |
| --- | --- | --- |
| Cleaned Column Name | Data Type | Variable Type |
| Date | Numeric | Independent |
| Silver\_USD | Numeric | Dependent |
| US\_AUD\_Rate | Numeric | Independent |
| Unemployment | Numeric | Independent |
| Inflation | Numeric | Independent |
| US\_Recession | Categorical | Independent |
| China\_Recession | Categorical | Independent |

Limitations: One of the more restrictive limitations on this study is that there is a significant amount of sparsity in the original, uncleaned dataset. However, there are 19,834 cleaned rows in the dataset. Another limitation is that the datasets all cover different time periods and the window where all the datasets have data is limited to January 2nd, 2003 through April 21st, 2022. Thus, the ridge regression model will have to be run on this time period. Likewise, the LBMA data used for the ARIMA model only contains data from January 2nd, 1968 through April 21st, 2022, so the ARIMA model will have to be run on this time period.

In addition, for this study, two approaches will be used: a ridge regression and an ARIMA model. For the ridge regression, all of the columns in the table above will be used. For the ARIMA model, only the Silver\_USD and Date columns will be used. ARIMA models only take a date and one column of time series data whereas ridge regression predicts differently, using a slew of other factors besides patterns in historical prices.

Delimitations: There are no significant delimitations for this study.

The macroeconomic data is important to this study because previous studies have suggested that commodity prices can be predicted from macroeconomic factors such as inflation, unemployment rate, commodity currency exchange rates, and recessions (Gargano & Timmermann, 2014). Thus, the selection of data from FRED reflects what the researchers from this similar study found to be important predictor variables for commodity prices using several kinds of regression models. The LBMA data is important to the study because it provides the datapoints to be predicted, Silver\_USD, which contains the price of silver in USD over time. In this study, this is the dependent variable.

**Data Gathering:** The Treatment of the Data: The LBMA data will be downloaded from the publicly available CSV file on Kaggle.com which shows silver prices in three currencies for the time period specified above. The various FRED datasets will be downloaded from publicly available Excel files on fred.stlouisfed.org which provides data on the various macroeconomic factors listed above. The datasets will be imported into Jupyter Notebook and each contained in individual dataframes. Each dataset will then have the columns renamed to more human-readable names. Each dataset’s “Date” column will be converted to datetime, and any other datatype conversions will be completed during this step as well, such as converting the two categorical variables to integer representations (1 or 0.) Since the dataframe containing unemployment data has monthly datapoints, the data will have to be converted to daily data by inserting rows for all of the missing days and filling those via linear interpolation. To combine the datasets, the individual dataframes must first span the same date range, as merging will be done on the “Date” column in each respective dataset. Rows will be inserted into each dataframe to ensure each dataframe has all the dates between February 1st, 1947, and January 7th, 2025. Then, the dataframes will be merged on the “Date” columns. The data will be checked for missing, null, or duplicate values. Any missing values in any column between January 2nd, 2003, through April 21st, 2022, except the two recession columns will be filled using linear interpolation for the purpose of the ridge regression piece of this study. Since the two recession columns are categorical, the missing values here will instead be filled using forward fill. Any missing values in the Silver\_USD column between January 2nd, 1968, through April 21st, 2022, will be filled using linear interpolation for the purpose of the ARIMA piece of this study. Linear interpolation is a recommended method that can be used with time series data (GeeksforGeeks, 2024). Rows from the combined dataset prior to January 2nd, 1968, will be dropped. Similarly, rows after April 21st, 2022, will be dropped. Since the two categorical variables are already binary, creating dummy variables is not necessary—this would be required for categorical variables with more than two categories. The Silver\_GBP and Silver\_Euro columns will be dropped.

The quality of the data is high as each dataset was sourced from a credible organization. For the silver price data, the organization is the London Bouillon Market Association. For the macroeconomic data, this is the Federal Reserve Bank of St. Louis. The combined dataset contains both quantitative and qualitative variables.

The combined dataset’s final overall sparsity is 14.73%.

**Data Analytics Tools and Techniques**: The Design of the Study: The main data analysis technique that will be run is ridge regression. The ridge regression analysis will use a type of linear regression to describe the contribution of each independent variable in predicting the dependent variable, Silver\_USD. Ridge regression is appropriate for this data because it is useful when multiple linear regression cannot be used because the multivariate data does not follow a normal distribution (NCSS Statistical Software, n.d.). This will be followed by an ARIMA model, which will be compared to the ridge model. The goals and expectations of the study are to predict silver prices using macroeconomic factors and identify the macroeconomic factors that most contribute to the fluctuation of silver prices. The presentation layer will include univariate and bivariate visualizations and a Tableau dashboard explaining this study’s findings.

While ridge regression and ARIMA analysis do not require normality, a Q-Q plot and Shapiro-Wilk test will nevertheless be run to determine normality. Univariate and bivariate graphs will be created to determine if variable relationships are linear and whether or not the residuals have constant variance (are homoscedastic.) Visualizations will be created using the tool “seasonal\_decompose” to check the Silver\_USD and Date columns for factors that influence what type of ARIMA model should be used, such as seasonality. The ACF and PACF plots will also be created to determine the ARIMA model’s order.

The data, using all columns, will be split into a training set and a testing set where 80% of the data is used to train the model and the remaining 20% is used to conduct an assessment of the ridge regression model. For the second portion of this study, the ARIMA model, only the date and Silver\_USD columns will be used. The data for the ARIMA model will also be split into 80% for the training data and 20% for the testing data, which will allow an assessment of the ARIMA model. An ARIMA model will look for patterns in the historical price of silver and use the patterns within the price itself to predict future silver prices.

Both analyses will be performed in a Jupyter Notebook using Anaconda.

**Justification of Tools/Techniques:** Python will be used to clean the data including the mitigation of issues such as missing data, anomalies, and so on. According to a book published in 2022 by Marco Peixeiro called “Time Series Forecasting in Python,” R used to be the gold standard for statistical analysis. However, many robust statistical analysis packages have since been designed in Python. In addition, in recent years, all data scientists have been required to be proficient with Python and it is now “the most widespread language in the field of machine learning.” Python is very versatile, more so than R (Peixeiro, 2022). Regarding SAS, Python is the better choice because it possesses graphing packages that are much more adaptable and robust than SAS, and it is open-source and updated more quickly than SAS so users have the best features sooner (Dutta, 2021).

**Project Outcomes**: The project seeks to create a ridge regression model predictive of the price of silver in USD based on economic factors. Support for the alternative hypothesis that either a ridge regression model that is accurate and predictive can be constructed from macroeconomic and/or commodity price data can be found in Gargano and Timmermann (2014).

**Projected Project End Date**: 4/15/2025

**Sources**:

Cochintu, C. (2024, July 16). *Silver Forecast & Price Predictions Q2 2024: Breakout from a three-and-a-half-year range.* Retrieved January 28, 2025, from <https://capex.com/en/overview/silver-forecast-price-prediction>

*Deed - Attribution-NonCommercial-NoDerivatives 4.0 International - Creative Commons.* (n.d.). Creative Commons. Retrieved January 28, 2025, from <https://creativecommons.org/licenses/by-nc-nd/4.0/>

Dutta, B. (2021, August 15). *8 Differences Between SAS and Python*. Analytics Steps. Retrieved January 28, 2025, from <https://www.analyticssteps.com/blogs/8-differences-between-sas-and-python>

Gargano, A., & Timmermann, A. (2014). Forecasting commodity price indexes using macroeconomic and financial predictors. *International Journal of Forecasting*, *30*(3), 825–843. Retrieved January 28, 2025, from <https://doi.org/10.1016/j.ijforecast.2013.09.003>

GeeksforGeeks. (2024, January 18). *Handling missing values in time series data.* GeeksforGeeks. Retrieved January 28, 2025, from <https://www.geeksforgeeks.org/handling-missing-values-in-time-series-data/>

Kriechbaumer, T., Angus, A., Parsons, D., & Casado, M. R. (2013). An improved wavelet–ARIMA approach for forecasting metal prices. *Resources Policy*, 39, 32–41. Retrieved January 28, 2025, from <https://doi.org/10.1016/j.resourpol.2013.10.005>

*Legal Notices, Information and Disclaimers.* (n.d.). Federal Reserve Bank of St. Louis. Retrieved January 28, 2025, from <https://fred.stlouisfed.org/legal/>

NCSS Statistical Software. (n.d.). *Ridge Regression.* Retrieved January 22, 2025, from Retrieved January 28, 2025, from <https://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Ridge_Regression.pdf>

Peixeiro, M. (2022). Understanding Time Series Forecasting [Google Books]. In *Time Series Forecasting in Python* (pp. 3–4). Manning Publications Co. Retrieved January 28, 2025, from <https://books.google.com/books?hl=en&lr=&id=4H6HEAAAQBAJ&oi=fnd&pg=PA1&dq=time+series+analysis+python+or+R&ots=6-bjU9uB4J&sig=MKLgxNAb9yKTELKf1jd38Hd643s#v=onepage&q=time%20series%20analysis%20python%20or%20R&f=false>

*Silver and solar technology*. (n.d.). The Silver Institute. Retrieved January 28, 2025, from <https://silverinstitute.org/silver-solar-technology-2/>

**Course Instructor Signature/Date:**

The research is exempt from an IRB Review.

An IRB approval is in place (provide proof in appendix B).

Course Instructor’s Approval Status: Approved

Date: Click here to enter a date.

Reviewed by:

Comments: Click here to enter text.