**D214 Task 3: Executive Summary**

Michelle Nelson

Department of Information Technology, Western Governor's University

Dr. William Sewell

March 31, 2025

**Research Question and Hypotheses**

Predicting the price of silver is crucial for certain businesses, in this context, a solar panel manufacturing company, because if a predictive model can be made, the cost of raw materials can be minimized, which maximizes profits. Furthermore, identifying which macroeconomic factors most influence the price of silver can help a business be proactive in ordering raw materials while they are still cheap. Thus, the research question for this analysis is: Can a ridge regression or time series model be constructed on the silver pricing dataset?

To answer this question, two data sources were used. The London Bullion Market Association (LBMA) calculates the daily price of silver via the ICE Benchmark Administration (IBA) using a daily electronic auction that balances buy and sell orders within a certain tolerance level until an equilibrium price is established (LBMA Silver Price FAQs, n.d.). In addition, Federal Reserve Economic Data (FRED) collects and compiles macroeconomic factor data from various agencies onto its website, many of which are available in a daily format.

The goals of this study are to predict daily silver prices using macroeconomic factors and to identify the macroeconomic factors that most contribute to the fluctuation of silver prices. The alternative hypothesis derived from these goals is summarized as follows: A predictive ridge regression model with a model accuracy greater than 70% can be constructed on the market research dataset. Conversely, the null hypothesis is: A predictive ridge regression model with a model accuracy greater than 70% cannot be constructed on the market research dataset.

**Summary of Data Analysis Process**

The datasets from the LBMA and FRED were cleaned and combined into one dataset. The data was cleaned by addressing data frequency issues, missing data, and outliers. One of the several datasets from FRED, unemployment rate, contained data updated on a monthly basis, with each listed as the first of the month for which the observation was noted. Because the rest of the datasets were daily in nature, linear interpolation was used to fill in the “missing data” in these gaps. Missing data outside of the largest common date range among the datasets was dropped, since imputing in these ranges is extrapolation. Outliers were detected in quantitative columns using the interquartile range (IQR) method, though none were found to be concerning. Finally, missing data within the largest common date range was imputed using linear interpolation or forward filling, depending on column type. This resulted in a data frame with 7 columns and 19,834 rows. One column contained the date, which is not counted among the independent or dependent variables.

Exploratory data analysis was conducted to test that the data fit the assumptions of a ridge regression model. A Q-Q plot was generated to visually assess if the data was normal, along with a more objective Shapiro-Wilk Test. Both conclusively showed that the data was non-normal. VIF was used to verify whether or not multicollinearity was present in the data. This test clearly showed that high multicollinearity was present in the data. Ridge regression’s advantage over ordinary least squares regression is that it is robust to both of these conditions, so a ridge regression model was chosen.

To create the initial ridge regression model, the data was split first into independent (X) and dependent (y) variables, then into training and testing datasets. The dependent variable was the silver price in USD, while the independent variables were macroeconomic factors including the US/AUD exchange rate, unemployment rate, inflation rate, a US recession flag, and a China recession flag. The last two independent variables were categorical variables encoded as either 0 or 1. 80% of the data was randomly allocated to the training set and 20% to the testing set.

Because ridge regression is sensitive to scale, the X\_train and X\_test datasets were scaled using StandardScaler. A ridge model with alpha equal to 1 was fitted on the training data, which was then used to predict on the testing data. This produced the performance metrics below.

A close-up of a number

AI-generated content may be incorrect.

With accuracy metrics just above the cutoff for rejecting the null hypothesis, some investigation was required to attempt to create a better predictive model. For this study, MAPE was of particular importance—to reject the null hypothesis, it must be under 30% (representing a greater than 70% accuracy, as stated in the hypothesis.)

The one assumption that was not well-verified prior to running the initial ridge regression model was the existence of a linear relationship between the dependent variable and the quantitative independent variables. As such, the following graphs were generated to illustrate those relationships. None of the three showed much of a linear relationship.

A group of blue dots

AI-generated content may be incorrect.

Despite the lack of a linear relationship, the plots did reveal some sort of structure. Most notably, the USD to AUD exchange rate and the inflation rate followed a curvilinear trend and appeared somewhat parabolic. Because of this, it seemed justified to attempt using polynomial features to try and capture these more complex relationships in the data. This allows for the introduction of new features generated by raising the original variables to different powers and considering interaction terms. Several degrees were tested, but ultimately, a degree of 4 was chosen because higher values tended toward overfitting while lower values didn't achieve the necessary amount of complexity needed to accurately predict.

Polynomial features with degree 4 was applied to the X\_train and X\_test datasets, then they were scaled using StandardScaler, just as before. Finally, a ridge regression model with alpha equal to one was fitted on this new training set, and predictions were made. Below are the performance metrics produced by the polynomial ridge regression model.

A black text with black text

AI-generated content may be incorrect.

These metrics were compared to the metrics produced by running the model on the testing dataset to ensure that the model had not overfitted and generalized well to new data. Those metrics are shown below.

A black text on a white background

AI-generated content may be incorrect.

Because the metrics are so close, it was determined that the model had not overfit.

**Study Findings**

With a MAPE of 21.76% on the testing dataset, well below the threshold set in the hypothesis, the final model was found to be robust, and the null hypothesis was rejected in favor of the alternate hypothesis. The other performance metrics concur, as all are better when smaller.

A graph of the predictions was also created. Below, the red line represents the predictions, and the blue line represents reality. Except for predictions made between 2004 through roughly 2006 and predictions made between 2020 through roughly 2022, the model's predictions align closely with reality. It is important, however, to note that there is a caveat to reading this graph—in reality, it would be better if this graph were a graph of dots, as the predictions don't represent a continuous time period. Rather, they are dates chosen at random during data splitting. Originally, a plot of dots was created, but it was excessively hard to read since the dots overlapped quite a bit. The line graph's shape looks almost identical to the dot plot but is easier to read. Thus, when reading the line graph, keep in mind that it connects actual prediction points with lines, even though predictions may not exist between those points.

A graph with red and blue lines

AI-generated content may be incorrect.

In addition, it was determined that the top ten most influential factors on the price of silver in USD were the following, shown below, ordered by magnitude.

A screen shot of a white and black text

AI-generated content may be incorrect.

**Study Limitations**

The primary limitation of this analysis was the interpretability of the results of the ridge model. Because polynomial features were necessary to create a robust model, including many interaction terms, it was somewhat difficult to discern exactly what the model suggests are the key factors that drive silver prices. None of the top terms are linear—that is, they are not the original set of 5 factors and are instead either interactions or raised to some power. Squared and cubed terms and the like can be interpreted to mean that the magnitude of the impact of the variable (such as unemployment) changes depending on how high unemployment itself is. Interaction terms indicate the price of silver changes based on the relationship between two variables, such as the US/AUD exchange rate and the unemployment rate, and a change in one alters the effect of the other. Given these explanations, it is easy to see how the interpretation of the coefficients and their factors quickly gets cumbersome and complicated.

**Proposed Actions**

A recommended course of action is to study the coefficients and their factors before making any business decisions based on this ridge analysis. Determining what some of these mean might require looking at the various functions involved. Once the analyst understands these factors, including how each one works (specifically the complex interaction terms,) the analyst will know what to look out for as omens of higher silver prices to come. However, some recommendations based on the simpler coefficients and their factors from this preliminary analysis are as follows:

1. The factor US/AUD Exchange Rate³ suggests it would be recommended to purchase silver when the exchange rate is higher since the coefficient is negative. This indicates that higher exchange rates drive down silver prices.
2. The factor Unemployment² suggests that it would be recommended to buy silver when the unemployment rate is low since the coefficient is positive. High unemployment drives the price of silver up.
3. The factor Inflation³ suggests that it would be recommended to buy silver when inflation is low since the coefficient is positive. High inflation drives the price of silver up.

While this study showed that a ridge regression model can be constructed on the silver price dataset, there are still improvements that can be made—or other models that can be tried. Because of the lack of a linear relationship and only a somewhat recognizable curvilinear one, it may be better to run a model that does not rely on the presence of trends that can be defined by the equation of a line. One such model is a random forest regression model, which has very few assumptions at all. Running such a model may even solve the interpretability problem encountered with the ridge regression model, since RFR models use a series of decision trees that should be simpler to interpret than convoluted interactions and squared factors.

**Expected Benefits**

As previously stated, the most important benefit of this analysis is that it provides a way for a company to be proactive about its purchasing of raw materials—in this case, silver. Being able to predict the price of silver means that the cost of raw materials for manufacturing can be minimized, which maximizes profits. Identifying a set of macroeconomic factors that most influence the price of silver helps to warn a business when the price of their raw materials is about to skyrocket so the purchasing agent can make proactive choices.

**References**

*LBMA Silver Price FAQs.* (n.d.). LBMA. Retrieved February 20, 2025, from <https://www.lbma.org.uk/prices-and-data/lbma-silver-price/lbma-silver-price-faq>