**Develop Modeling Report**

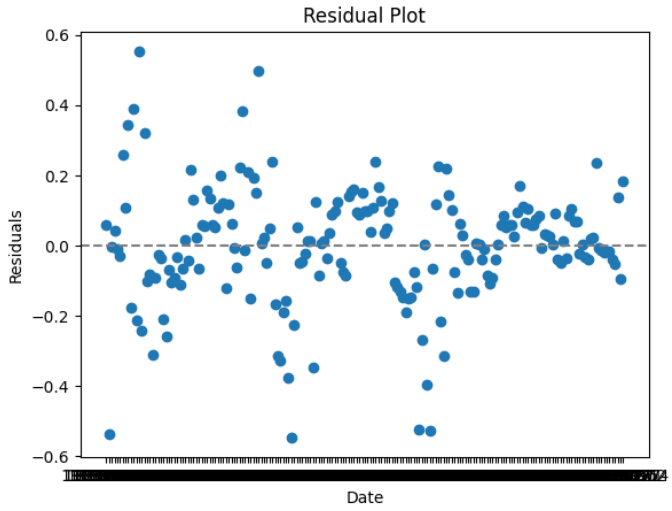
The model approach is to begin with fitting an OLS model. Starting with an OLS model helps in variable selection as well as describes the obstacles for more complex models. As such, these models begin with rudimentary linear models using a selected subset of hyperparameters for the number of variables, overfitting with all of the variables, reverse variable selection, and then ridge regression.

1. Simple OLS model

For this model, all of the variables in the training set were first evaluated for their p-value. This is a simplistic and straightforward test that allows for ranking of specific variables' relevance to prediction of the output variable. It does suffer, however, from the assumption of normality in the sample set, and does not consider interaction effects. Then, the top 15 variables in terms of p-value were selected for use in a simple OLS model. This hyperparameter was chosen somewhat haphazardly: doing a visual test, including the top 10 variables included a mix of macroeconomic variables, sentiment score, and also a vectorized word. This achieved at least one metric from the range of variables intended to be included in the model during variable creation.

The model is not as complex as other models to come, and is not intended to be. This will serve as a baseline model to which further models will be compared. As such, the creation of the model is simple as well, using the "Difference" variable as the dependent variable and including the resulting 10 independent variables as predictors.

Regarding accuracy, this model performs very badly. It achieves an $R^2$ of only 0.324 and an adjusted $R^2$ of .267. This adjusted $R^2$ penalized model overfitting, and tells us that even 10 variables is too many. Not only that, but they are only good at predicting 26.7% of the variability in the "Difference" variable. Further, it looks like only the "Retail Sales" and "Risk" vectors were statistically significant at the 99% level.




Looking at the residual plot, there is no obvious evidence of correlation in errors, which is encouraging. Correlation in errors would suggest a spurious variable that is not being captured.

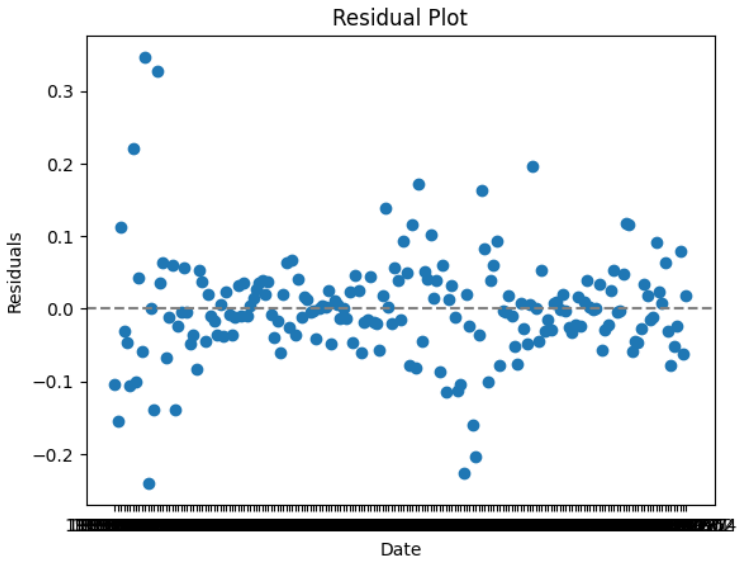
2. All variables subset

Next, the full training set was used, leveraging the swathe of vectorized words that were available as well as the macroeconomic variables that investors are likely to be monitoring as well. During variable

augmentation, the hyperparameter of 100 variables was chosen as a balance between variable selection and simplicity. This resulted in 26 macroeconomic variables or augmented variables, while the rest were vectorized words. While this is a clumsy way to fit a model, the adjusted $R^2$ should penalize for unnecessary variables. This should provide a balance in model selection after the first OLS model.

The only complexity this model adds is dimensionality. There are a few problems with fitting too many variables. First is overfitting, where the variables start fitting to the random element rather than the underlying trend. The second is multicollinearity, where some of the predictor variables are endogenous to each other. This is certainly the case, as the LEI, CEI, and LAG indicators cover many of the other macroeconomic variables, and there are some rolling averages already included. The last is model complexity, though this is not as much a problem as 74% of the variables are just vectorized words. That said, actual utilization of a model this complex would prove difficult.

This model performs much better than the last one, with an $R^2$ of 0.867 and an adjusted $R^2$ of .747. This is surprising, given the vastly greater number of variables. Most of the variables were not statistically significant, though, which does suggest overfitting. Further, a model that has few statistically significant variables and performs relatively well suggests the tradeoff between accuracy and model simplicity.




This residual plot also does not suggest any sort of trend in errors, which would be an issue. That said, the magnitude of the errors is not inspiring, as the degree of most monetary policy actions is only 25 basis points.

3. Reverse Variable Selection

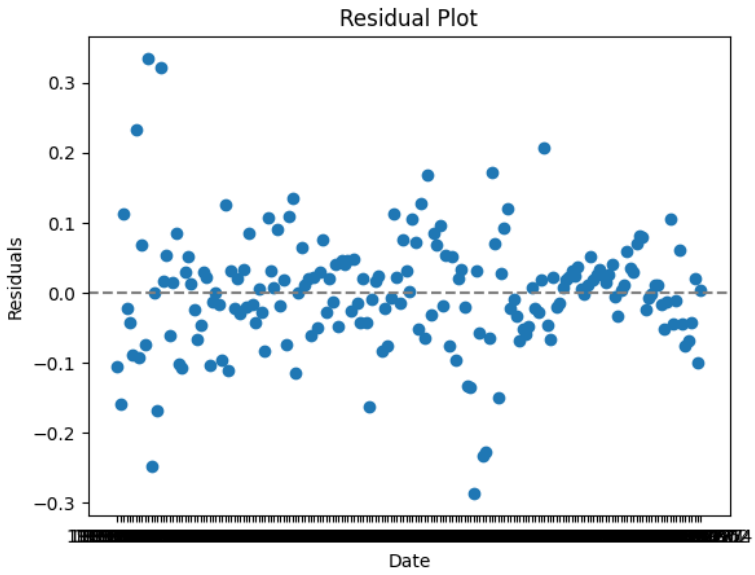
Starting from a model that includes all of the variables, it make sense to perform reverse variable selection to cut down the number of variables and increase ease of use. The benefit of these models is that it iterates the significance of each variable until no further variables are eligible for removal. It results in a smaller set of variables that hopefully all have statistical significance—in this case, the p-value.

It also benefits from retaining model simplicity in that it is still an OLS model, just with better-chosen variables than hand-picking them. To that end, replicability becomes a strength of these models relative to using every single variable and overfitting.

The model performs admirably compared to the previous model in that it has an $R^2$ of 0.823 and an adjusted $R^2$ of 0.796. The higher adjusted $R^2$ relative to the last model suggests some overfitting and that the reduced number of variables contains only variables that are deemed necessary to the model. To that end, it contains 7 macroeconomic variables, 4 Minutes-derives variables (Net Sentiment Score, Positive Frequency, Negative Frequency), and 16 vectorized words. Further, it includes macroeconomic variables that sync with textbook theories of monetary policy action in that it includes the "Level", "Average Hourly Earnings", and "LEI" variables.

The benefit of reverse variable selection is that you can also filter by the hyperparameter AICc. This is a metric that calculates goodness of fit, but also penalizes models with many parameters relative to the sample size. Because the sample size is relatively small for a dataset, the AICc is useful as it provides a more accurate measure of model quality.

This model is even more accurate, with an $R^2$ of 0.831 and an adjusted $R^2$ of 0.804. This process seems to have benefitted the goodness-of-fit both in an absolute manner but also with fewer required variables. This variable has forgone the Minutes-related variables and has elected to use 10 macroeconomic variables and 16 vectorized words.




The residual plot also does not show any signs of trend within the plot. There is some skewness in that earlier dates seem to have more variation, which could be addressed in the validation.

4. Ridge Regression

Finally, ridge regression also provides some benefits in terms of variable selection and accuracy. Further, it allows for selection of the strength of regularization in the form of an alpha variable. Further, the best value for alpha can be determined iteratively using RidgeCV. The benefits of ridge regression are that it prevents overfitting through regularization, handles multicollinearity, and offers model simplicity in coefficient shrinkage.

Ridge regression models are visually similar to OLS models in that it regresses a predicted variable on a string of predictor variables with corresponding weights. The difference is that it utilizes a weighted error function alpha * $|b|^2$, which assists in variable selection.

| Training | $R^2$ | Adjusted $R^2$ | RMSE | AICc |
|---|---|---|---|---|
| Simple OLS | 0.324 | 0.267 | 0.172 | -62.480 |
| All Variables | 0.867 | 0.747 | 0.101 | 21.630 |
| Reverse Variable Selection (AICc) | 0.831 | 0.804 | 0.089 | -265.885 |
| Ridge Regression | | | 0.421 | -104.398 |

| Validation | $R^2$ | Adjusted $R^2$ | RMSE | AICc |
|---|---|---|---|---|
| Simple OLS | -0.201 | | 0.279 | 48.513 |
| All Variables | 0.658 | | 0.149 | -154.683 |
| Reverse Variable Selection (AICc) | 0.518 | | 0.177 | -407.096 |
| Ridge Regression | 0.613 | | 0.159 | -151.537 |

Note: some of the calculations of these metrics returned improbable answers, with $R^2$ and adjusted $R^2$ above |1|. Will require a second look.

The best way to compare these models based on both accuracy and a penalty on an excessive number of variables is the AICc. It provides a trade-off between goodness of fit and complexity. The lower the AICc, the more preferred the model should be. As such, the Reverse Variable Selection model seems to be the strongest out of all of them. This was true in the training data, and is even more accurate in the validation set. This may be by design, as it was optimized to reduce the AICc. That does not take away from the fact that it is a strong model. Unfortunately, it also has the largest RMSe, and lowest $R^2$, which shows that it may benefit from some overfitting. All models have different strengths in the validation set, but the Reverse Variable Selection model outshines the ret by far in AICc for a marginal decrease in RMSE and $R^2$. As such, this model will be chosen as the best model for this week.