Multicollinearity in Financial Data
Final Project Executive Summary

Harold Miao

December 2023

1 Introduction

Multicollinearity is a common problem arising in multi-dimensional linear regression. Collinearity refers to two random variables closely linearly related to each other. Multicollinearity in the multi-dimensional linear regression context refers to the presence of a strong linear relationship among independent variables in the data set.

The study presents three methodologies to systematically identify the presence of multicollinearity in a given data set and further introduces methodologies to address the problem.

2 Materials and Methods

2.1 Data

The data set of the study comes from the Financial Market. The independent variables are the daily log returns of 12 stocks between 2021 and 2022 chosen from 4 major sectors: Microsoft Corp (MSFT), Apple Inc.(AAPL), and Alphabet Inc.(GOOG) from the Tech sector; Procter & Gamble Company (PG), Pepsico Inc.(PEP), and Mondelez International Inc.(MDLZ) from the Consumer Retail sector; JPMorgan Chase & Co.(JPM), Bank of America Corp (BAC), and Citigroup Inc.(C) from the Financial sector; and Exxon Mobil Corp (XOM), Chevron Corp (CVX), and Conocophillips (COP) from the Energy sector. The independent variable is the daily log return of the SP500 between 2021 and 2022. All of the data are directly quoted from Yahoo Finance. Due to the scope of the study, the time-serial structure of the daily log return is not taken into account.

2.2 Impact of Multicollinearity

Multicollinearity may introduce additional variance to the best estimate of the linear coefficient of the model. It also violates the underlying assumption that all independent variables are mutually independent, undermining the model result's explainability. Further, it poses challenges to the stepwise feature selection procedure of the model.

2.3 Quantifying Multicollinearity

2.3.1 Pairwise Scatterplot and Correlation Coefficient

The most straightforward way of detecting the presence of multicollinearity is through pairwise scatterplot and correlation coefficient. A strong pairwise linear relationship will lead to a "better fitted" pairwise scatterplot and a higher correlation coefficient, which may be calculated as

$$r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$$

From multiple studies, a pairwise correlation value greater than 0.8 indicates the presence of multicollinearity issues.

2.3.2 Variance Inflation Factor

While the pairwise correlation coefficient explains the pairwise linear dependency, multicollinearity may arise even when pairwise linear correlation is low. For example, if $X_1 = 2X_2 - 3X_3 + 5X_4 - 7X_5 + 11X_6$, there might be the case that X_1 is not pairwise highly correlated with any of the rest independent variables but remains a pure linear combination of X_2 to X_6 nonetheless. In this case, pairwise correlation fails to identify the presence of multicollinearity correctly. To address this issue, the Variance Inflation Factor (commonly referred to as VIF) is introduced. For an independent variable X_j, it is defined as

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the standard R-squared score obtained by regressing the X_j with respect to the rest $(n-1)$ number of random variables. Since the R-squared score is between 0 and 1, with a larger numerical value indicating stronger linear explainability, we expect $1 - R_j^2$ to drop as the linear relationship strengthens. Hence, VIF_j increases as the linear relationship of X_j and the rest of the independent variables increases.

In practice, a VIF score greater than 4 suggests moderate multicollinearity and a VIF score greater than 10 suggests significant multicollinearity.

2.3.3 Eigenvalue Method

From PCA theory, we know that the eigenvalue of the independent variables' correlation matrix sheds light on the 'variability' of the independent variables in each eigenvector direction. If all variables were perfectly independent and no multicollinearity existed, we would expect the variability to be roughly evenly distributed across all the n dimensions. Suppose the variability of the independent variables is concentrated in a subset of the n dimensions, which is manifested through an uneven distribution of the n eigenvalues. In that case, multicollinearity is potentially present in the model. From existing literature, the minimum eigenvalue being less than 0.05 is a standard threshold for the presence of multicollinearity.

The drawback of this approach compared to VIF is that the model does not identify which independent variable introduces the multicollinearity. Whereas in the VIF approach, we know exactly that the variable with a high VIF_j score causes multicollinearity, in the eigenvalue approach, should we observe an uneven distribution of the eigenvalues, we can only deduce that multicollinearity exists. However, we cannot correctly identify which variable causes the multicollinearity. This method would not be as explainable as the VIF method.

2.4 Addressing Multicollinearity

2.4.1 PCA Regression

Multicollinearity may be addressed via PCA regression. PCA regression regresses the dependent variable on the principal components of the independent variables, which are orthogonal by design. The regression performs a significantly better estimation by optimally choosing the number of components via AIC/BIC/cross-validation. PCA regression addresses the root cause of multicollinearity.

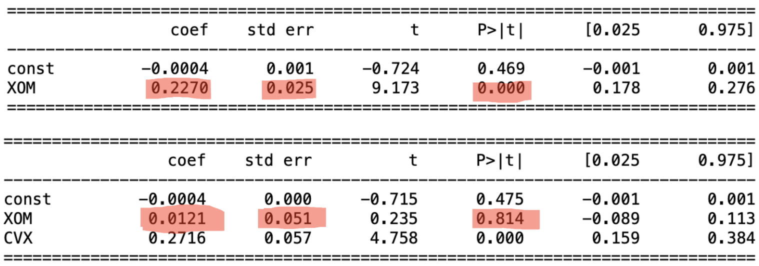
2.4.2 Regularization

Multicollinearity may also be addressed via regularization, such as including a Ridge penalty term. Multicollinearity is often associated with the model incorporating too many independent random variables that are potentially linearly correlated to each other, and the introduction of the penalty term in regularization may address the issue of multicollinearity by forcing the model to be limited in its complexity. However, unlike PCA, regularization does not directly ”De-linearize” the independent variables.

3 Results

3.1 Impact of Multicollinearity

The issue caused by multicollinearity is evident in the data. Let us compare a model with two independent variables, X_1 representing the Exxon Mobil Corp (XOM) and X_2 representing the Chevron Corp (CVX), against its nested model with only one independent variable, X_1 representing the Exxon Mobil Corp (XOM). We can see that the coefficient estimates, standard errors, and P values are significantly different (Figure 1). This poses a challenge in the traditional feature-selection procedure through stepwise forward and backward selection.



```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0004      0.001     -0.724      0.469      -0.001       0.001
XOM            0.2270      0.025      9.173      0.000       0.178       0.276
==============================================================================

==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0004      0.000     -0.715      0.475      -0.001       0.001
XOM            0.0121      0.051      0.235      0.814      -0.089       0.113
CVX            0.2716      0.057      4.758      0.000       0.159       0.384
==============================================================================
```

(a) Figure 1

3.2 Quantifying Multicollinearity

3.2.1 Pairwise Scatterplot and Correlation Coefficient

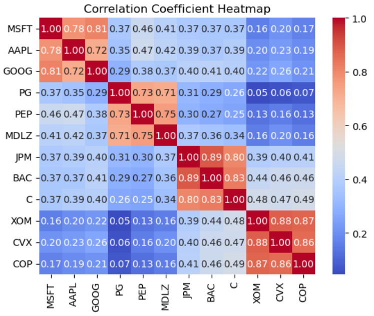
The following heatmap plot (Figure 2) of the pairwise correlation coefficient suggests that the presence of multicollinearity is evident. While inter-sector correlation is generally below 0.4, intra-sector pairwise correlation is significantly higher, with the Financial Sector and Energy Sector exceeding 0.8.

3.2.2 Variance Inflation Factor

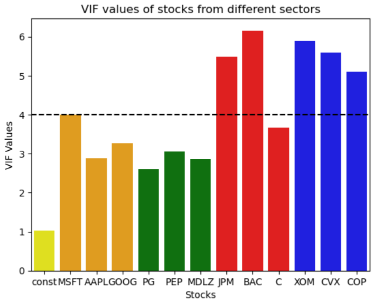
The VIF methodology draws similar results. Based on the VIF plot of each individual stock (Figure 3), we can see that five variables exceed the threshold of 4 in their VIF values. The five stocks are from the Financial and Energy sectors, matching our previous methodology findings.

3.2.3 Eigenvalue Factor

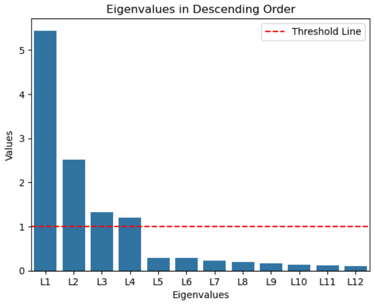
The eigenvalue distribution is uneven in our dataset (Figure 4). 4 principal components account for 87.5% of the total variations present in the data. This matches our domain knowledge since the data set has four different sectors. Even though the least eigenvalue is still above the threshold of 0.05, the two distinct sets of eigenvalues allude to the fact that multicollinearity exists in the financial data. The VIF methodology draws similar results. Based on the VIF plot of each individual stock (Figure 3), we can see that five variables exceed the threshold of 4 in their VIF values. The five stocks are from the Financial and Energy sectors, matching our previous methodology findings.



(a) Figure 2



(b) Figure 3

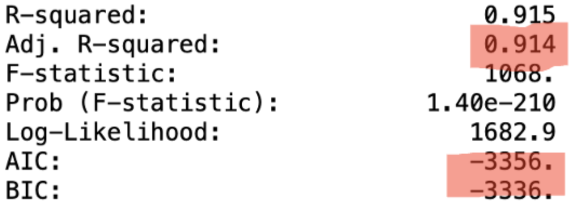


(c) Figure 4

3.3 Addressing Multicollinearity

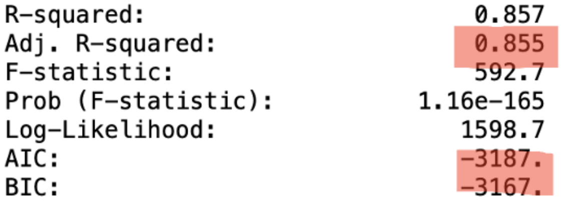
3.3.1 PCA Regression

PCA regression addresses the issue of multicollinearity. BIC chooses the first four principal components, AIC chooses the first nine, and cross-validation R2 chooses all twelve principal components. Suppose we cross-compare the regression on the first four principal components, as suggested by BIC, against the vanilla regression on the four most influential stocks representing each sector (MSFT, P&G, JPM, XOM). In that case, we can see a significantly higher adjusted R-squared score, a better AIC and BIC score (Figure 5 & 6), as well as a lower standard error in the model coefficient estimate. (Figure 7 & 8)



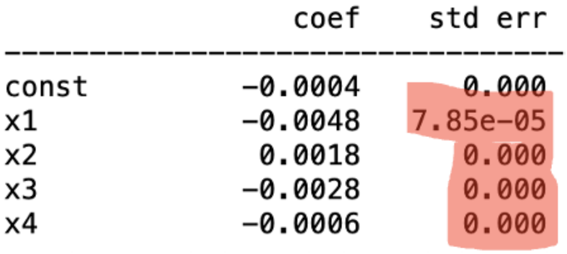
```
R-squared:                       0.915
Adj. R-squared:                  0.914
F-statistic:                     1068.
Prob (F-statistic):          1.40e-210
Log-Likelihood:                 1682.9
AIC:                            -3356.
BIC:                            -3336.
```

(a) Figure 5: Score for PCA Regression



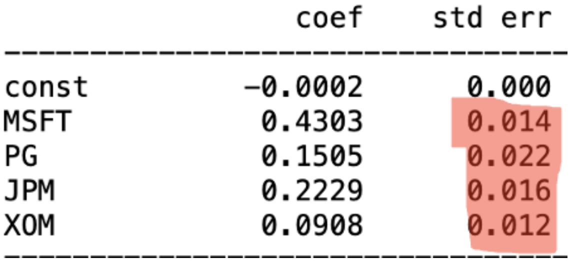
```
R-squared:                       0.857
Adj. R-squared:                  0.855
F-statistic:                     592.7
Prob (F-statistic):          1.16e-165
Log-Likelihood:                 1598.7
AIC:                            -3187.
BIC:                            -3167.
```

(b) Figure 6: Score for Vanilla Regression



```
                 coef    std err
--------------------------------
const         -0.0004      0.000
x1            -0.0048   7.85e-05
x2             0.0018      0.000
x3            -0.0028      0.000
x4            -0.0006      0.000
```

(c) Figure 7: Standard Error for PCA Regression

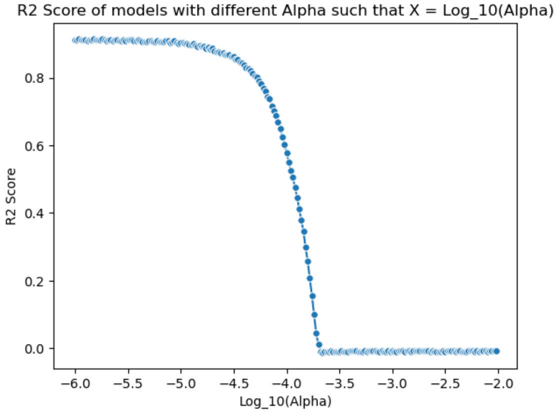


```
                 coef    std err
--------------------------------
const         -0.0002      0.000
MSFT           0.4303      0.014
PG             0.1505      0.022
JPM            0.2229      0.016
XOM            0.0908      0.012
--------------------------------
```

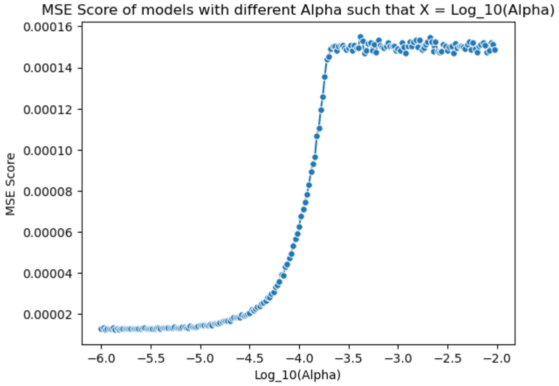
(d) Figure 8: Standard Error for Vanilla Regression

3.3.2 Regularization

According to existing literature, ridge regression should improve the performance of multicollinear regressions. However, while the original model had a cross-validation R2 score of around 0.91, the regularised model does not see an improvement in the cross-validation score. The performance drops significantly (Figure 9) when the alpha penalty parameter exceeds $10^{-}3.7$. A similar drop in performance is evident when we use the MSE metric (Figure 10). Further investigation is needed to assess whether adding a penalty term improves the regression result.



(a) Figure 9



(b) Figure 10

4 Conclusion

Multicollinearity is shown to be present in the data through all three methods. Multicollinearity is most present in the Financial and Energy sector, which aligns with domain knowledge. PCA regression successfully addresses the issue and improves the model in all AIC/BIC/Cross-validation score metrics. Ridge regression does not improve the performance of this dataset. The research may be further extended, incorporating the time-serial structure of the dataset.

References

[1] Shrestha, Noora, *Detecting Multicollinearity in Regression Analysis*, American Journal of Applied Mathematics and Statistics, **8**, 39-42, 2020.

[2] Young, D.S., *Handbook of regression methods*, CRC Press, 2017, 109-136.

[3] Vatcheva, K.P., Lee, M., McCormick, J.B., and Rahbar, M.H., *Multicollinearity in regression analysis conducted in epidemiologic studies*, Epidemiology (Sunnyvale, Calif.), **6 (2)**, 227, 2016.

[4] Belsley, D.A., *Conditioning diagnostics: Collinearity and weak data in regression*, John Wiley & Sons, Inc., 1991.