FACTOR MODELING

AUTHORS:

Yupeng Lu

Chunlu Zhang

Chunxiang Wang

Dejian Zhang

Xuyu Cai

Zhenhang Guo

SUPERVISOR: CHRIS KELLIHER

DECEMBER 2021

Abstract

This experiment tests 14 common factors on the market to see if any of them can explain the S&P 500 return during 2015-2020. Results show that 3 factors passed significance tests, monotonicity test, and multicollinearity test. Then we use multifactor regression to find factor returns for each period. In order to construct our portfolio, we resort to the ARIMA model to obtain the prediction of factor returns. Finally, a mean-variance optimization is performed to find the weights of stocks in S&P500.

Introduction

Our project is to predict the price of certain financial assets in the next day given historical information. This is important in investments because accurate predictions can help traders make investment decisions with reliable sources. A multifactor model is built upon common factors that can be found in the market. The multifactor model is an extension of the single-factor capital asset pricing model (CAPM). After building up the multifactor model, we design a portfolio based on this model to maximize returns. All data are extracted from the Bloomberg terminal, all needed information can be found in the links and references we cited.

Data Extraction

The first step of the project is data extraction, collecting financial data from disparate types. Many of the data are poorly organized or completely unstructured but still able to be used. Data extraction prepares data for later processing, which, therefore, is the most important part of building up the model. Our factors include the Value Factor, Growth Factor, Financial Quality Factor, Leverage Factor, etc. Each of these factor types includes several smaller factors.

Data Cleaning

To clean all these factors and make them the same indices, we made them in two steps.

Transforming data

Firstly, all factors are transformed into monthly data and the same columns names which are dates. Filling missing columns with blank data, then we tried to use linear interpolation to fill missing data but which will affect the distribution of original data. Thus, we finally forward filled missing data with previous data.

Processing data

We first calculate reciprocal for each factor exposure. Then we depolarized, standardized, and forward fill all missing data.

Statistical Analysis and Factor Evaluation

WLS-Significance Test

Significance test can be done by different methods, WLS is chosen because residuals should always satisfy the requirements of linear regression. Thus, heteroskedasticity should be removed if it exists. WLS can well remove heteroskedasticity, WLS becomes OLS(Ordinary Least Squares) after assigning weights to residuals.

$$S\&P\ return_{t_{i+1}} = \beta * factor + \sum_{i=1} \beta_i * Industry\ return * 1_{industry_i} + \epsilon \quad (1)$$

$$weight = market\ values \quad (2)$$

In this experiment, 11 industries are labeled inside S&P 500 stocks, weights are market values of S&P 500 stocks. β is the regression coefficient of each factor, added β_i which is up to i=11 for each industry, the indicator $1_{industry_i}$ is an identity matrix that has 1 for each stock if the stock is in the given industry else 0. Therefore, from the above equations, t-statistics of each factor can be calculated, they are checked to see how much each factor can explain the total market return in the next period. The sum of absolute values of t-statistics is sorted to compare which factor is more significant.

IC(Information Coefficient) Test

Information Coefficient, referred to as IC, represents the capacity factor to predict stock returns. The calculation of IC is to calculate the correlation between the ranking of all stocks at the beginning of the period and the income ranking at the end of the period. Large IC means the stock-picking ability is strong. The maximum value of IC is 1, which means that the stock selection of this factor is 100% accurate, and corresponds to the stock with the highest-ranking score, selected stocks have the largest increase in the next adjustment cycle. Normal IC-Pearson correlation is calculated. The cross-sectional correlation coefficient between the factor exposure in all stocks and its return in the next period at a certain point in time as explained in (3)

$$IC_A = correlation(f_A, r) \quad (3)$$

In addition to the study of single factors, a more realistic situation is to calculate the IC value of the compound factors as multiple factors are combined together. The IC obtained at this time is the IC of this compound factor. For example, EP and BP can be put into the ranking conditions at the same time to form a composite factor.

Monotonicity Test

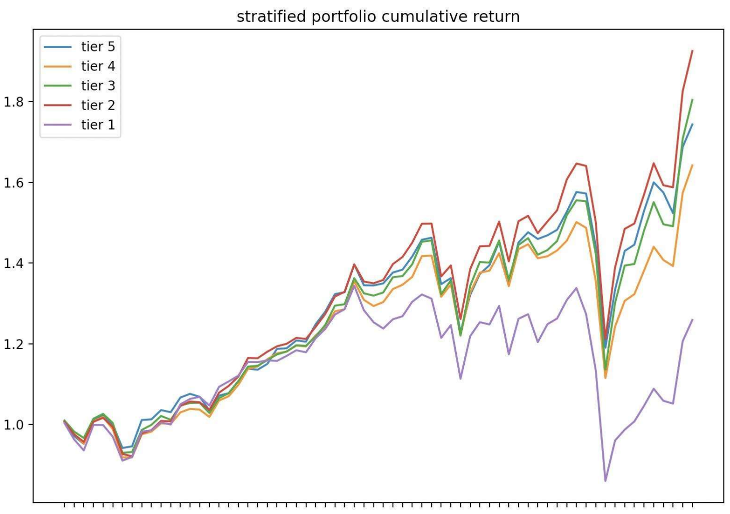


Figure 1. Monotonicity Test for volatility

All stocks are divided into 5 ranks within each industry based on the amplitude of their factor exposures. For every rank, stocks are picked up to make up a portfolio and the return of it is calculated to compare with S&P return. As expected, the higher rankings of stocks can always construct portfolios that have higher returns. Figure 1 is the results of the volatility factor, which shows monotonicity, the other 2 factors are added to Appendix A.

Multicollinearity-Correlation Test

The last step to filter factors is to test the multicollinearity of factors. We evaluate correlations of factors under the same category, like Value Factor, Growth Factor, and etc.

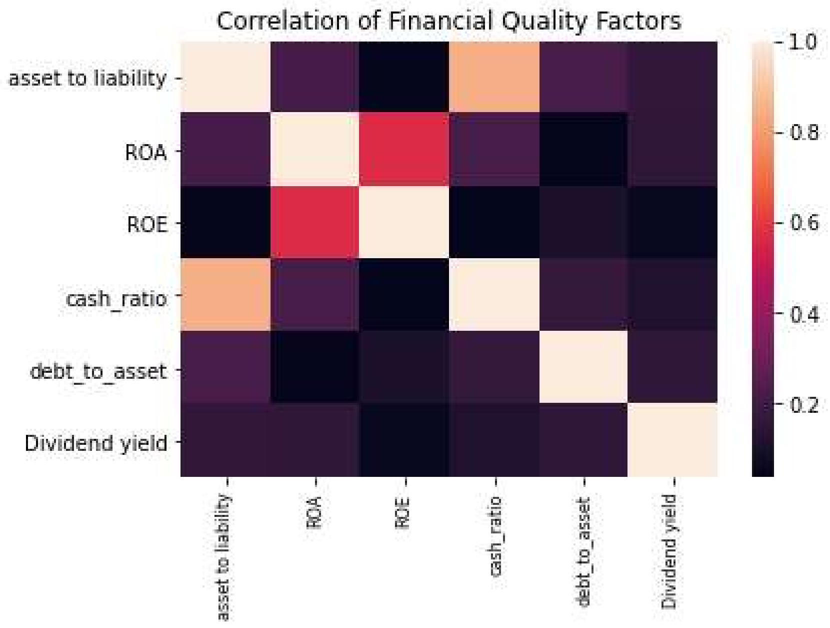


Figure 2. Correlation between Financial Quality Factors

For example, Figure 2 shows correlations between financial quality factors. Because most of them are higher than 0.2, we only keep ROE inside Financial Quality Factors. So do other factors, the results are added to Appendix B. Before that process, factors left are Value Factors, Growth Factors, Turnover Factors and Volatility Factors. After that process, the factors left are ROE, EBITDA/EV and volatility.

Stock Return Calculation by OLS

The purpose of this linear regression is to calculate the historical βs that explain the S&P 500 returns and support the following time series model prediction. The explained variable is the S&P 500 return. The explanatory variables are 3 factors left from the steps above and S&P 500. The S&P 500 is viewed as a measure of how well the stock market is performing overall.

In statistics, ordinary least squares (OLS) is a type of linear least-squares method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters of a linear function of a set of explanatory variables by the principle of least squares: minimizing the sum of the squares of the differences between the observed dependent variable (values of the variable being observed) in the given dataset and those predicted by the linear function of the independent variable.

$$Y_{t+1} = \sum_{i}^{4} \beta_{i,t} X_{i,t} + \beta_{5,t} S\&P500_{t} + \varepsilon_{t} \qquad (4)$$

Where Y is 500 stock returns, X is the return of each stock.

Historical βs are obtained from the OLS model. The βs would be used for establishing a time series model for later ARIMA prediction. T-statistics are kept to present the significance of the βs. Also, white-test p-values are checked because if the p-values are smaller than 0.05, it could be said that it has heteroscedastic problems under the 5% significance level.

Table 1: Average of the absolute t-values

variable	ROE	EBITDA_EV	volatility	S&P 500
Average of absolute t-statistics	0.67326	1.55539	0.84269	9.9183

Next, the average of the absolute t-statistics of three factors is calculated. Large values mean the βs are more significant. Table 1 shows that the S&P 500 and EBITDA_EV are the top two powerful Explanatory variables.

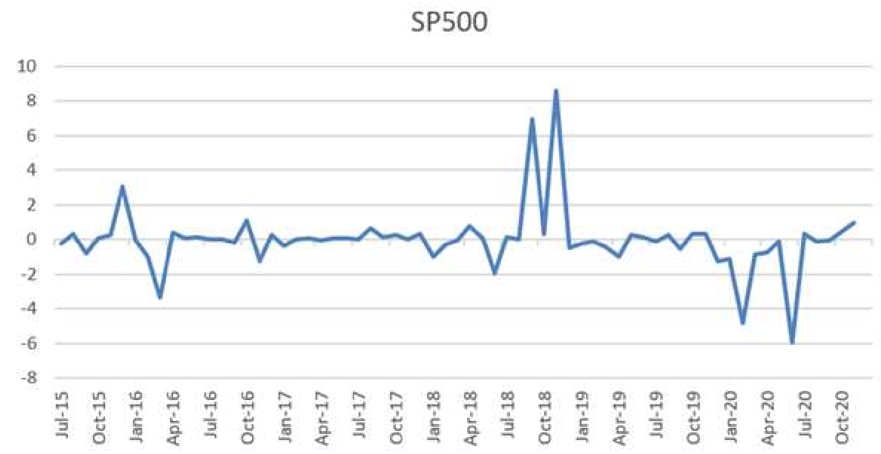


Figure 3 the historical βs of S&P 500

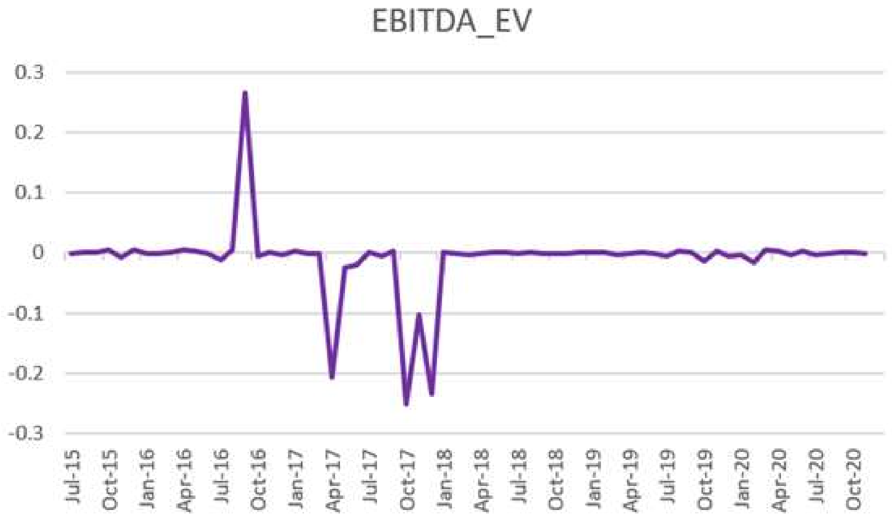


Figure 4 the historical βs of EBITDA_EV

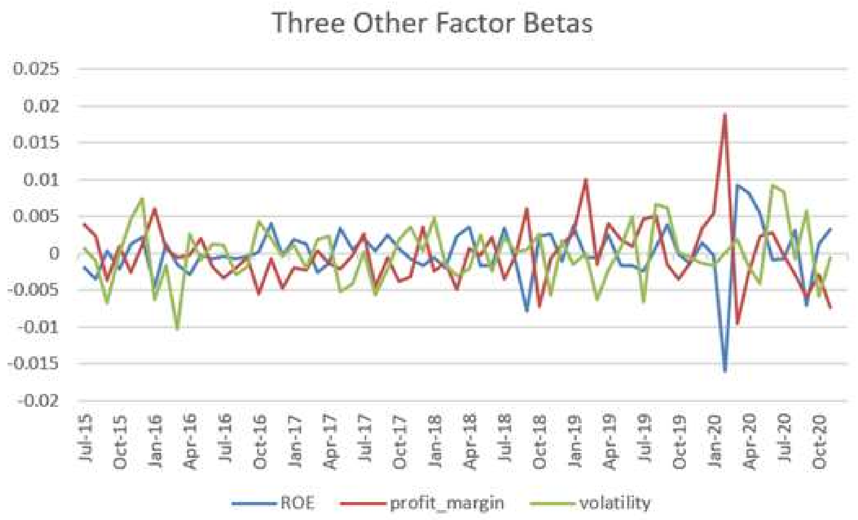


Figure 5 the historical βs of ROE, profit_margin, and volatility

In Figure 3, for S&P 500, the highest two positive points are in November and September 2018. When S&P 500 increases one unit and the variables do not change, the stock return would increase 690.74% and 856.73% respectively. The highest negative influence took place in June 2020. When it increases one unit and the variables do not change, the stock return would decrease 115.95%. In Figure 4, for EBITA_EV, the highest positive point is in September 2016. When it increases one unit and the variables do not change, the stock return would increase 26.61%. The three highest negative points are April, October and December 2017. In Figure 5, for the other three factors(ROE, profit_margin and volatility), they have an influence between -0.5% and 0.5% mostly. At the beginning of 2020, their role of ROE and volatility expanded reaching -1.5% and 2% respectively.

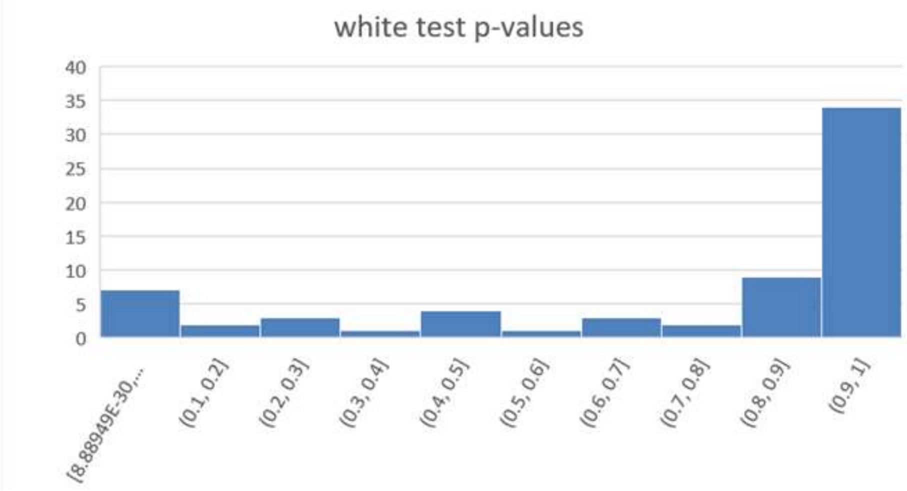


Figure 6 White test p-values

Figure 6 shows that, overall, there are no heteroscedastic problems in this experiment.

β **Prediction by ARIMA**

After getting historical βs, β_t can only be calculated by prediction because β_t is calculated using all information at $t + 1$, which means we cannot know the latest coefficient β_t unless we predict it based on historical data.

$$\beta_t = c + \emptyset_1 \beta_i + \cdots + \emptyset_p \beta_{t-p} + \theta_1 e_{t-1} + \cdots + \theta_p e_{t-p} + e_t (5)$$

Where β_t= the beta we want at time t, c= intercept, ∅= coefficient of each parameter q, θ= coefficient of each parameter q, e= residuals or errors in time t.

(5) is used to get predictions of βs. Methods to predict time series are various, we chose this method because βs are tested and result in good autocorrelations.

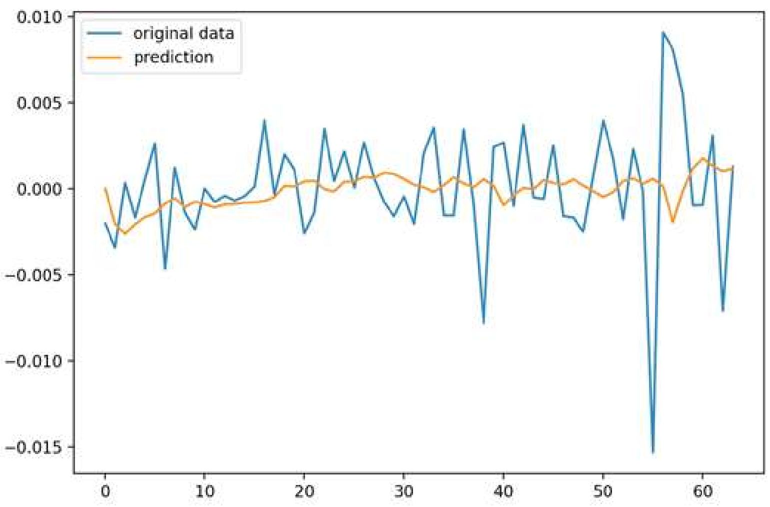


Figure 7 ARIMA β Prediction of Next 5 Periods

Figure 7 is the ARIMA prediction of βs for EBITDA_EV in the next 5 months. With these results, we can measure how accurate the prediction is using a testing set of data. Other 2 factors predictions are added to Appendix C.

Portfolio Optimization

With stock returns calculated, the portfolio can be constructed by assigning weights to each stock. The method used to optimize portfolio return is Lagrange Multipliers, which optimizes the given function along with constrained conditions.

$$\begin{cases} \mathrm{f(w)} = -\mathrm{w} * \mathrm{R} \\ \mathrm{w} * \Sigma * \mathrm{w}^{\mathrm{T}} = \mathrm{a} \\ \mathrm{w} * \mathrm{I} = 1 \end{cases} \qquad (6)$$

Where f(w)= function to minimize, w= weights of stocks, shape= (1,500), R= stock returns, shape= (500,1), I = identity matrix, a = S&P variance, Σ = covariance matrix of S&P 500 returns. By applying (6), the return of the optimal portfolio at a certain period is returned.

Conclusion and Suggestion

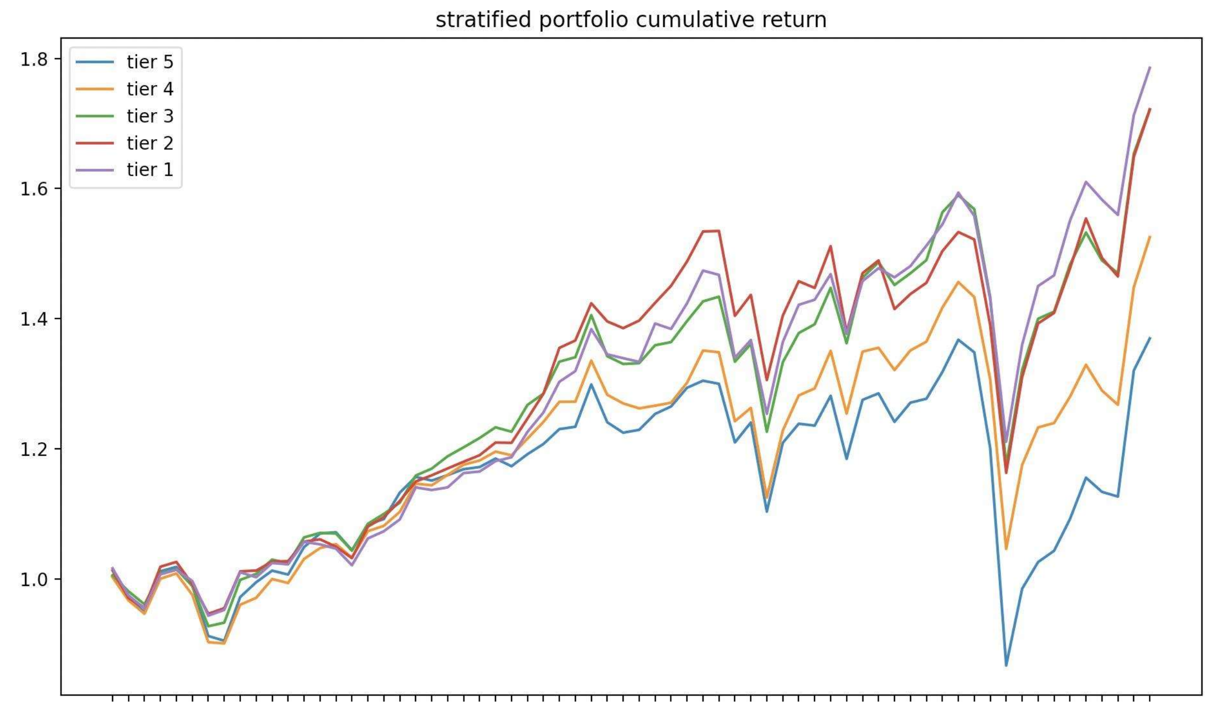
We propose a basic framework for stock portfolio construction by factor picking. This paper examines 14 common factors that can be found on the market easily, which makes the data extraction part easy to understand and work. However, common factors don’t always have good explanation power toward the market return and that’s why our significance tests filtered most factors. In the future study, we might consider adding factors we designed ourselves and test more factors from other researches.

References

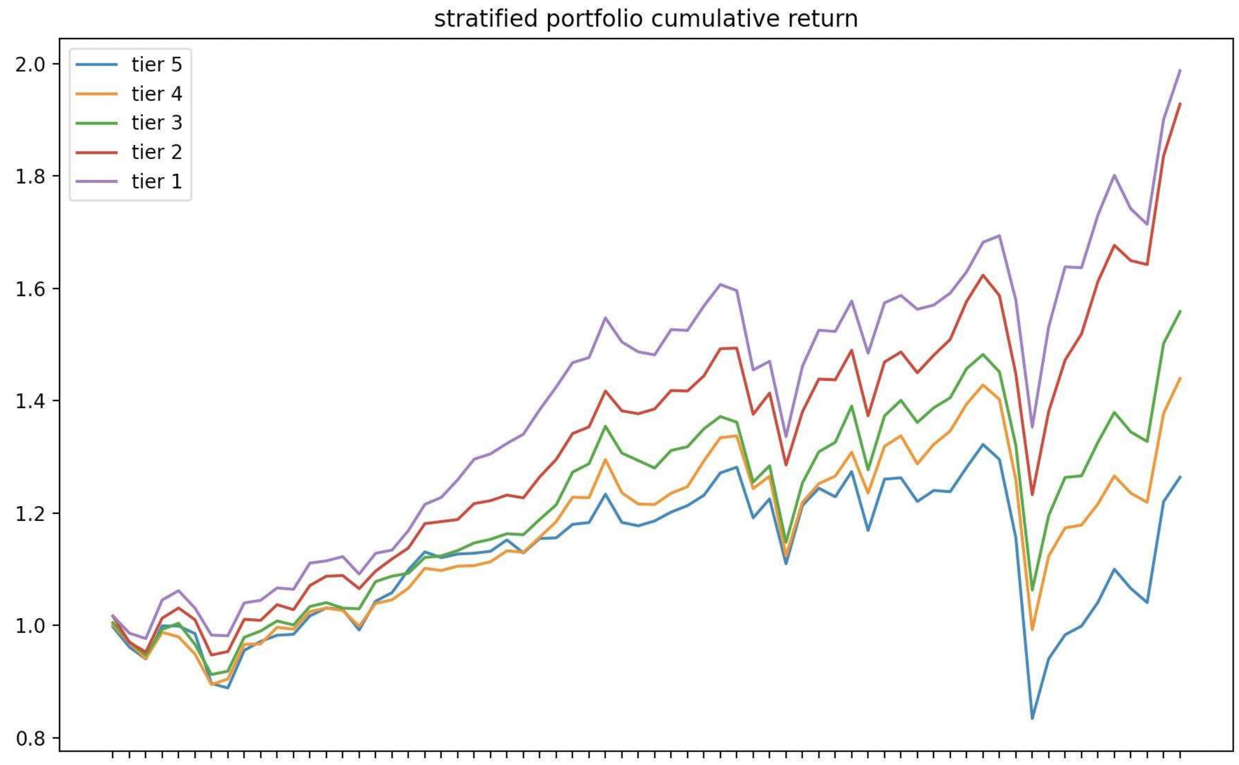
[1] Z. Ding and R. D. Martin, "The Fundamental Law of Active Management: Redux," Journal of Empirical Finance, vol. 43, pp. 91-114, 2017.

[2] T. Oyeniyi and L. Ma, "U.S. Stock Selection Model Performance Review 2016 – The Year of “Risky” Value," 2017.

APPENDIX A

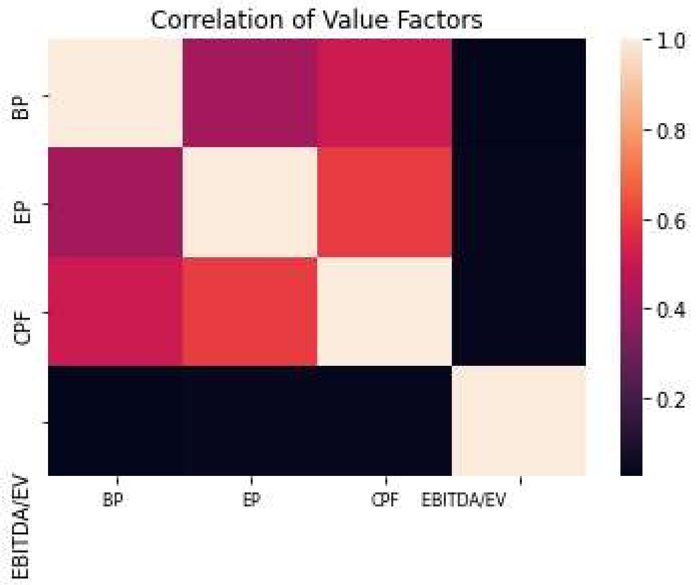


Appendix A.1

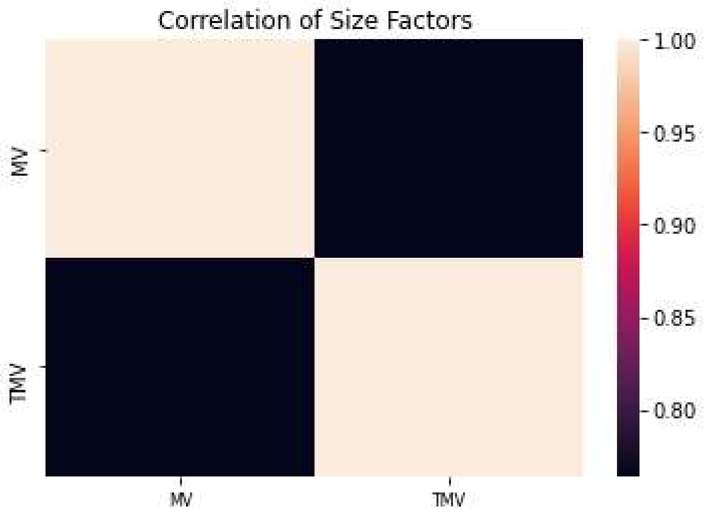


Appendix A.2

APPENDIX B

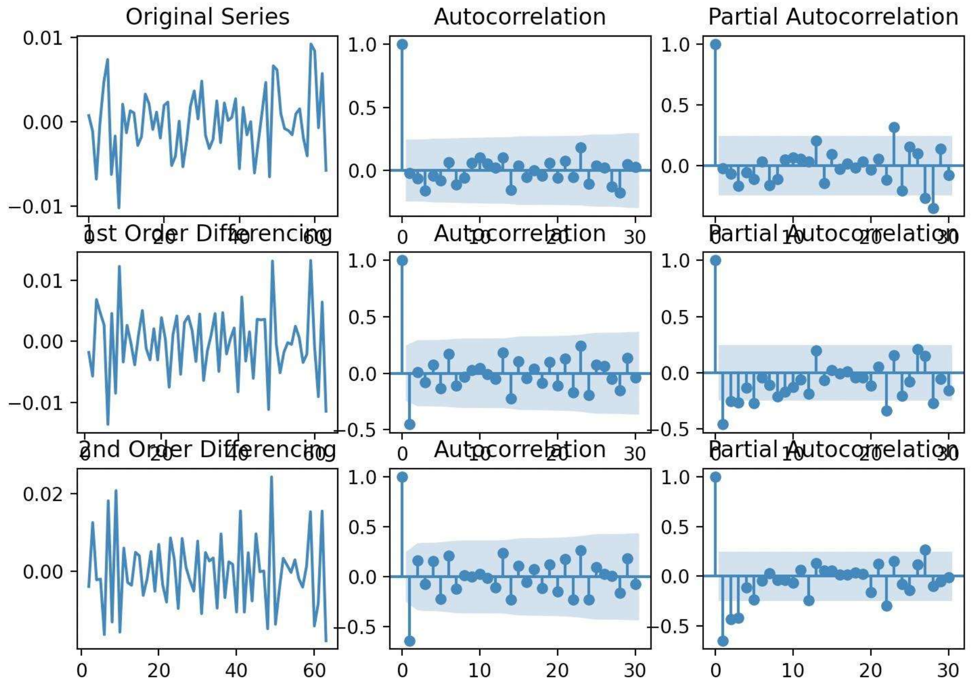


Appendix B.1

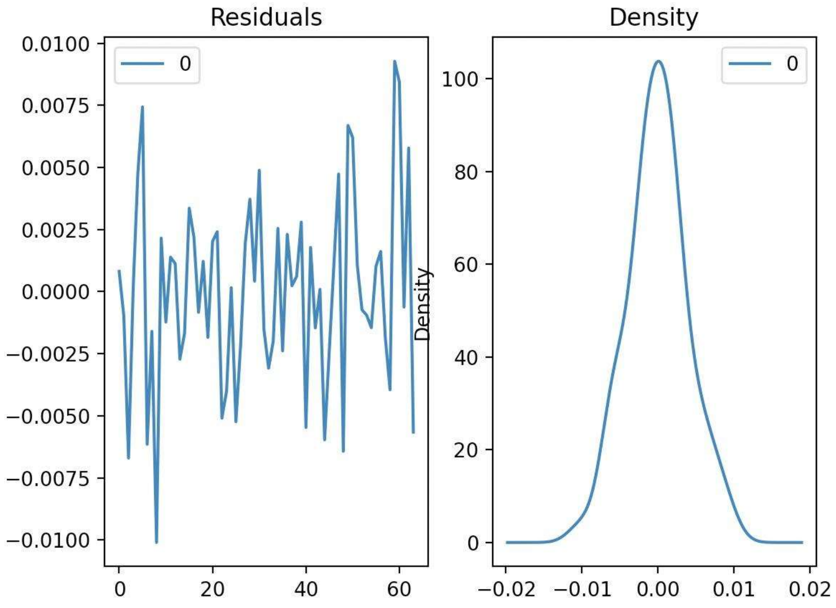


Appendix B.2

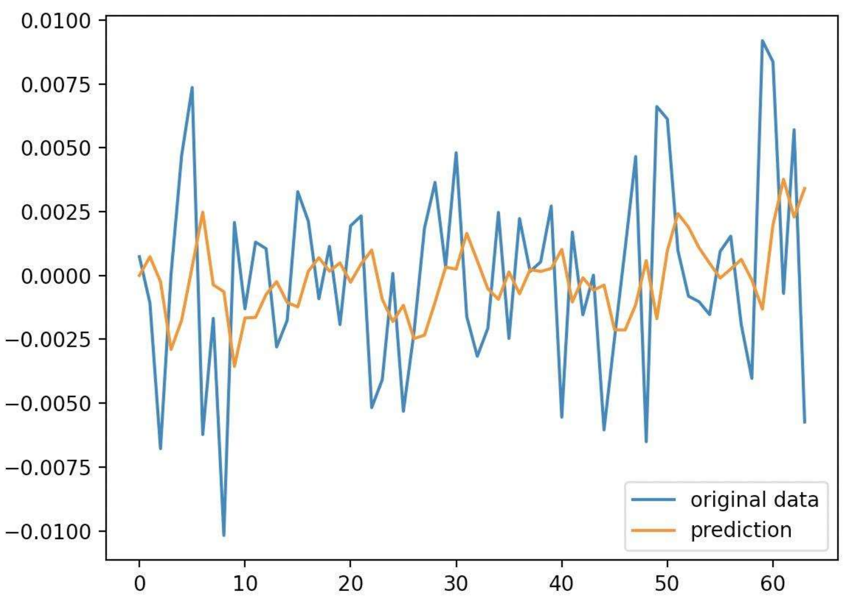
APPENDIX C



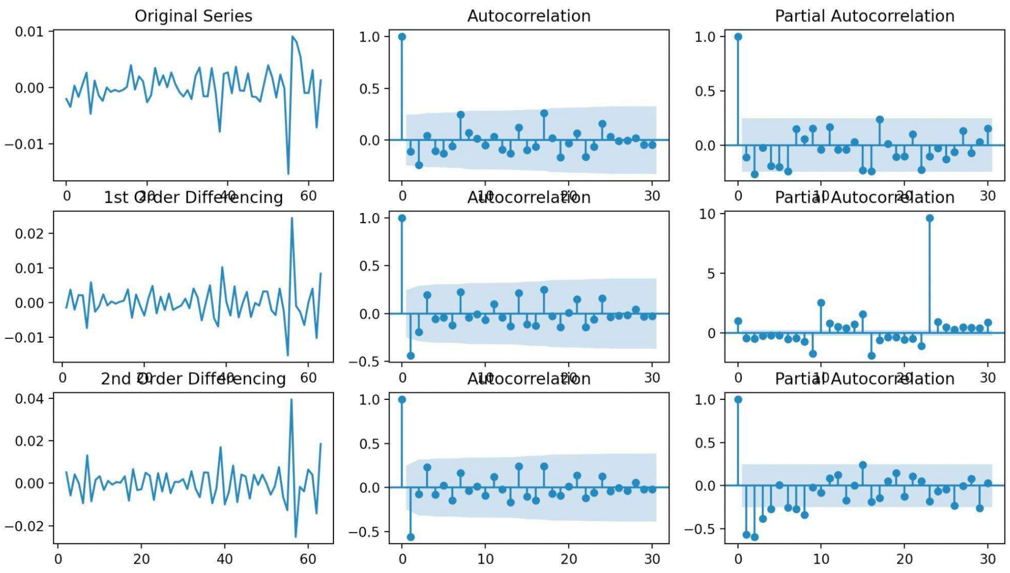
Appendix C.1



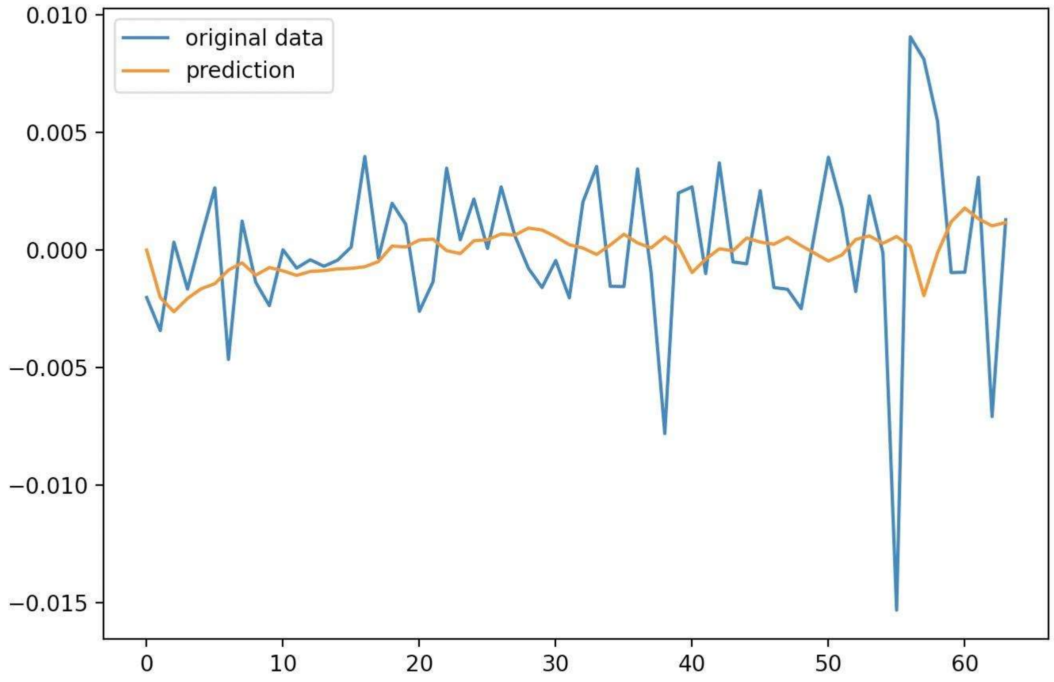
Appendix C.2


0.0100
0.0075
0.0050
0.0025
0.0000
−0.0025
−0.0050
−0.0075
−0.0100
0
10
20
30
40
50
60
original data
prediction

Appendix C.3



Appendix C.4



Appendix C.5