# Empirical Analysis of Two Methods for Forecasting S&P 500 Index Returns

Huiyi Zhang, Shiqi Chen, Yiwei Wang, Jie Zhou

School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, China

**Abstract.** The S&P 500 index is one of the important barometers of the global financial market trends, and therefore predicting its direction is of significant importance to the financial market. This article aims to improve the predictive ability for the S&P 500 index returns (excess returns) using various methods. Economic variable forecasting is introduced, including 15 new economic variables such as inflation rates and treasury bond yields. For this purpose, this article uses data from May 1937 to April 1972, a total of 35 years, as in-sample data to estimate the forecasting model, and data from May 1972 to December 2023 as out-of-sample data to test the model's effectiveness. In the process, we rely on Principal Component Analysis (PCA) to extract a common factor from the predictive variables. Empirical results indicate that the PCA regression model corresponding to these 15 economic variables cannot successfully predict the S&P 500 index returns within the sample and out of sample. At the same time, Python was used to analyze each economic variable and provided a certain explanation for the inability to predict. We also conducted ARMA model forecasting for the S&P 500 index returns.

**Keywords:** Yield forecasting; Principal Component Analysis (PCA); Economic variable forecasting; ARMA model forecasting.

## 1. Introduction

### 1.1. Research Background

The S&P 500 Index, as the flagship index of the U.S. stock market, has been a focal point for global investors since its inception in 1957, thanks to its broad market coverage and precise measurement of market performance. Compiled by Standard & Poor's Corporation, the index aggregates the stocks of the top 500 listed companies in the United States, representing approximately 80% of the total market capitalization of the U.S. stock market, and is one of the key indicators for measuring the performance of the U.S. stock market.

The robust performance of the S&P 500 Index is attributed to the careful selection of its constituent stocks. These companies are leaders in their respective industries, possess strong profitability, and have promising prospects, offering investors a wealth of investment options. The index's calculation method is scientifically rigorous, employing a market capitalization-weighted approach, with dynamic adjustments to its components to ensure it accurately reflects the overall performance of the U.S. stock market. Therefore, the S&P 500 Index is not only a barometer of the U.S. economy but also a microcosm of the global economy.

Over the past few decades, the S&P 500 Index has experienced numerous bull and bear market transitions, witnessing the ups and downs of the U.S. economy. By studying and tracking the trends of the S&P 500 Index, investors can better grasp the pulse of the global economy and make more informed investment decisions.

In the investment field, the S&P 500 index has a wide range of applications. It serves not only as the benchmark for passive investment products such as index funds and exchange-traded funds (ETFs) but also leads the development of investment strategies and asset allocation plans based on it. Moreover, the historical return data of the S&P 500 index provides investors with valuable research resources. Through the analysis of historical data, investors can understand the volatility characteristics, influencing factors, and long-term trends of the S&P 500 index, providing important references for future investment decisions.Among the many factors influencing the return of the S&P

500 index, economic variables play a significant role. Macroeconomic factors such as economic growth, inflation levels, and interest rate environments, as well as changes in government policies and regulatory measures, have a direct impact on the index's fluctuations. Moreover, market sentiment, corporate performance, corporate bond yields, government long-term bond yields, and factors such as natural disasters and public health events also affect the index's returns. Therefore, to more accurately predict the returns of the S&P 500 index, it is necessary to consider a variety of economic variables, including but not limited to inflation rates, consumer price indices, risk-free interest rates, and government long-term bond yields.

Dealing with multiple economic variables, using Principal Component Analysis (PCA) can greatly alleviate the difficulty of the work.Neely et al. (2014) use the PCA approach to combine information from technical indicators and macroeconomic variables and find that new predictors can significantly improve the equity risk premium predictability. By employing the PLS method, Huang et al. (2015) propose a novel investor sentiment index. The new index exhibits stronger predictive power than the sentiment index constructed by Baker and Wurgler (2006, 2007) using the PCA method. Ma et al. (2022) examine the relationship between macroeconomic attention and stock return predictability using the PCA, scaled PCA, and PLS approaches.

David et al.(2024) using the ARMA model in stock market returns is also a relatively common method of forecasting.

### 1.2. Research Purpose

This article aims to explore effective predictive methods for the S&P 500 index through two approaches, namely, a comprehensive analysis of a series of economic variables (15 in total). The focus of the study is on constructing a predictive model that can accurately capture market dynamics and effectively forecast future trends. By integrating economic data such as inflation levels, interest rate changes, policy guidance, long-term government bond yields, and returns, the goal is to reveal the specific impact mechanisms of these economic variables on the S&P 500 index returns. Through analysis of individual economic variables and methods such as Principal Component Analysis (PCA), the study seeks to identify the specific impacts of these economic variables on the S&P 500 index returns. Alternatively, using the autocorrelation properties inherent to the index, the ARMA model is employed to forecast the index's excess returns.

## 2. Establishment of Theoretical Models for Economic Variable Forecasting

### 2.1. Variable Analysis

The dependent variable: the excess return of the S&P 500 index, which is calculated by subtracting the corresponding risk-free rate from the logarithmic return of the S&P 500 index. Excess returns are used for measurement and prediction.

$$\mathrm{r_{t+1}} = \overline{\mathrm{r_{t+1}}} - \mathrm{r_{f,t}}$$

$$\bar{\mathrm{r}} = \ln\left(\frac{\mathrm{p_t}}{\mathrm{p_{t-1}}}\right) = \ln(\mathrm{p_t}) - \ln(\mathrm{p_{t-1}})$$

Where $\mathrm{r_{f,t}}$ represents the risk-free interest rate that is determined by the t-th period, and $\mathrm{p_t}$ is the value of the S&P 500 index for the t-th period.

Explain variables (economic variables): Inflation rate, corporate bond yield, government bond yield, and other indicators, have always been a hot topic of research for economists regarding their impact on market yield (S&P 500 index yield). Many scholars have constructed numerous theories from different perspectives to explain this, and it can be found that there are many factors affecting market yield. Many scholars explain these based on different evidence and perspectives.Such as Zhang Y. et al.(2019 2023 2024), Wen, D.et al.(2022)

This article, based on theories and existing literature, and considering issues such as data selection and interpretation, mainly selects macroeconomic variables such as the inflation rate, as well as Empirical Analysis of Two Methods for Forecasting S&P 500 Index Returns.

## 2.2. Selection of economic variables

### 2.2.1 12-month dividends（d12）

For a stock's yield, whether the company pays dividends and how much it pays affect its fluctuations. The S&P 500 Index aggregates stock data from the top 500 listed companies in the United States. Therefore, this article has collected the dividend payout data of the constituent companies of the S&P 500 Index from May 1937 to December 2023, corresponding to the data with dividends over the past 12 months. The 12-month dividend indicator is typically used to assess the total amount of dividends a company has paid to shareholders over the past year and is one of the important indicators that investors consider for investment returns and the company's operating conditions. A company that can consistently and steadily pay dividends usually means that its business situation is good and has sufficient profitability to support shareholder returns, which helps to enhance investor confidence in the company and increase its market value.

### 2.2.2 12-month earnings（e12）

Similarly, a company's profit level has a significant impact on the returns of its stock, akin to "d12". Data has been collected on the profit levels of companies corresponding to the S&P 500 index from May 1937 to December 2023, matching the profit growth rate of the past 12 months to this data. The profit indicator for December can be used to assess a company's total profit over the past year, and it is also one of the important indicators that investors consider when evaluating investment returns and the company's operational status.

### 2.2.3 Book-to-Market Ratio（b/m）

Book-to-Market Ratio, it is the ratio of a company's market value to its book value, used to assess the company's growth potential and value. It also has a significant impact on stock prices, and the required b/m is obtained by weighting the b/m values of the target companies.

### 2.2.4 Treasury Bills（tbl）

T-bill, or Treasury Bills, commonly referred to as U.S. Treasury bills or T-Bills, are short-term bonds issued and guaranteed by the U.S. government. As a type of U.S. Treasury security, T-Bills are favored by investors for their high creditworthiness, low risk, and short maturities (usually not exceeding one year). Common T-Bill maturities include 3 months, 6 months, and 1 year, offering investors flexible options for capital allocation. T-Bill rates typically reflect the interest rate levels in the short-term Treasury market and investor expectations for the U.S. economy.

### 2.2.5 AAA bond yield（AAA）

AAA bond yield,This refers to the return rate that investors can obtain from holding AAA-rated bonds. This rate of return is usually expressed in annualized form, meaning the proportion of earnings an investor would receive from investing in AAA-rated bonds over the course of a year. AAA-rated bonds are issued by companies with excellent credit ratings and are characterized by low credit risk, stable earnings, and good liquidity. Due to their low credit risk and stable earnings, the yield on AAA corporate bonds is generally lower than that of treasury bonds with the same maturity but higher than the yield on corporate bonds with other credit ratings. This is mainly used to characterize the earnings situation of high-quality bonds.

### 2.2.6 BAA bond yield（BAA）

BAA bond yield, is considered to be of general grade have an average ability to pay. This data represents the yield that investors can expect to receive from an investment in BAA-rated bonds over a year. Here, BAA is mainly used to indicate the yield situation of slightly lower-grade bonds.

### 2.2.7 long govt yield（lty）

Long govt yield,is one of the important indicators for measuring a country's credit risk and also a significant reference interest rate in the financial market. It reflects investors' expected return on long-term government bonds and the market's expectations for future economic trends.

### 2.2.8 net equity issuance（ntis）

Net equity issuance refers to the issuance ratio of the company's ordinary shares that are free from restrictive rights and special terms to raise funds, which can to some extent affect the company's stock price and thereby impact the return rate of the S&P 500 index.

### 2.2.9 riskfree return（Rfree）

Risk-free return refers to the sum of the time value of money (pure interest rate) and the inflation compensation rate, which represents the minimum rate of return an investor can expect without any risk. It has a crucial impact on the yield of any stock. Excess return is based on this.

### 2.2.10 inflation（infl）

Inflation, the inflation rate has a comprehensive impact on a country's overall financial market. Inflation indicators are very important economic indicators in macroeconomics, as they help people understand the current economic situation and provide a basis for decision-making for policymakers. At the same time, they can greatly affect investors' operational behavior in the financial market and have different impacts on various products in the financial market.

### 2.2.11 corporate bond return（corpr）

The corporate bond return refers to the ratio between the income an investor earns from holding corporate bonds and the purchase price of the bonds. It can reflect the credit status of the company and also indicate the state of the financial market. It is one of the important bases for investors to make investment decisions.

### 2.2.12 Other factors（ltr，svar，CRSP_SPvw，CRSP_SPvwx）

In addition to these factors, there are many other factors that can influence the return of the S&P 500 index.

### 2.3. Data Source and Statistics

Due to space constraints, only data from May 1937 to August 1938 is provided (rounded to two decimal places) at the end of the document. The data comes from the S&P 500 related data provided by Professor Gong Xue in class, the website is:https://teacher.njust.edu.cn/_upload/article/files/12/40/b52fd80c4ba58e6bc74a6c3dc0b9/9ce051af-9fd0-4abc-9440-741102f10c60.xlsx

## 3. Estimation of Economic Variables Forecasting Models

### 3.1. Predictive regression models

To verify whether these economic variables can predict the return of the S&P 500 index, we used data from May 1937 to April 1972, a total of 35 years, as in-sample data for model estimation, and the remaining data from May 1972 to December 2023 as out-of-sample data for out-of-sample forecasting.

$$r_{t+1} = \alpha + \beta_i x_{i,t} + \varepsilon_{t+1} \quad (1)$$

Where rt+1 is the log excess return of the S&P 500 index in month t+1, and xt,i is the value of the corresponding economic variable in period t.

Out-of-sample predictions are based on the model above.eq(1):

$$\overline{r_{t+1}} = \overline{\alpha} + \overline{\beta_i}x_{i,t} \tag{2}$$

The parameters $\overline{\alpha}$ and $\overline{\beta_i}$ are both obtained through OLS estimation of model (1).

### 3.2. Principal component analysis

We use principal component analysis (PCA) to construct common factors from financial market economic variables.The PCA model extracts the diffusion indexes via the following equation:

$$x_{i,t} = \varphi_i' F_t^{PCA} + e_{i,t} \tag{3}$$

Where $F_t^{PCA}$ denotes the PCA diffusion indexes extracted from the financial market economic variables and is a K-vector. $\varphi_i'$ is a K-dimensional parameter to be estimated. Then, we run the PCA regression, as follows:

$$r_{t+1} = \beta_0 + \sum_{k=1}^{K} \beta_k F_{k,t}^{PCA} + \varepsilon_{t+1} \tag{4}$$

where $F_{k,t}^{PCA}$ denotes the kth principal component extracted from the financial market economic variables.Table 1 reports the summary statistics of the fifteen financial market economic variables.

**Table 1** summary statistics (economic variables)

| Variables | Mean | Std. Dev. | Maximum | Minimum |
|---|---|---|---|---|
| D12 | 13.2581045192308 | 16.6803542 | 70.3037 | 0.51 |
| E12 | 31.5536331730769 | 43.04861412 | 197.91 | 0.62 |
| b/m | 0.53506954343365 | 0.250143072 | 1.206529805 | 0.120510188 |
| tbl | 0.03515567307692 | 0.031199932 | 0.163 | 0.0001 |
| AAA | 0.05834028846154 | 0.028723456 | 0.1549 | 0.0214 |
| BAA | 0.06836240384615 | 0.030514117 | 0.1718 | 0.0294 |
| lty | 0.05166038461538 | 0.028843526 | 0.1482 | 0.0062 |
| ntis | 0.01225392768558 | 0.018683174 | 0.051160886 | -0.055954167 |
| Rfree | 0.00283663461538 | 0.002569053 | 0.0135 | -0.0006 |
| infl | 0.00297416973552 | 0.004656076 | 0.058823529 | -0.019152895 |
| ltr | 0.00470048076923 | 0.025618769 | 0.1523 | -0.1124 |
| corpr | 0.0049225 | 0.023068464 | 0.156 | -0.0949 |
| svar | 0.00223720324543 | 0.004618703 | 0.073153143 | 0.000107534 |
| CRSP_SPvw | 0.0095118125 | 0.045419764 | 0.253519 | -0.254106 |
| CRSP_SPvwx | 0.00651995 | 0.045511565 | 0.249304 | -0.257951 |

Note: This table reports the summary statistics of economic variables.The sample period is from May 1937 to December 2023.

# 4. Empirical results of the economic variables forecasting model

## 4.1. In-sample analysis

As shown in Table 2, we have summarized the in-sample regression results for various economic variables, including the β values by OLS estimate, t-stat,R2 (%)and P-value.It was found that among the 15 variables given, only 4 were significant at the 10% level, with only one significant at the 5% level, and the majority of the variables were not significant.

**Table 2** In-sample estimation results.

| Variables | β | t-stat | R2 | P-value |
|---|---|---|---|---|
| D12 | 0.008641051 | 1.078538072 | 0.09689819 | 0.670444302 |
| E12 | 0.002681708 | 0.905035977 | 0.062022864 | 0.501529547 |
| b/m | 0.36723164 | 0.595204161 | 0.03973499 | 0.786117586 |
| tbl | -9.337243717* | -1.994170699 | 0.400230398 | 0.082831254 |
| AAA | -7.501354287 | -1.484034899 | 0.218970327 | 0.213007467 |
| BAA | -6.825185569 | -1.448736859 | 0.204578252 | 0.259550258 |
| lty | -7.975672055 | -1.603007279 | 0.249607971 | 0.175456263 |
| ntis | -7.073025837 | -0.649697253 | 0.082176335 | 0.311438358 |
| Rfree | -100.4631106 | -1.764612548 | 0.314135342 | 0.162446768 |
| infl | -88.93121429* | -2.868682806 | 0.806352834 | 0.073388394 |
| ltr | 11.28697937* | 2.308042871 | 0.393938805 | 0.068191447 |
| corpr | 17.37918011** | 2.812350437 | 0.756177194 | 0.021098819 |
| svar | -34.66755665 | -0.513189883 | 0.120840609 | 0.500990294 |
| CRSP_SPvw | -0.668951934 | -0.154025972 | 0.004352242 | 0.942019579 |
| CRSP_SPvwx | -0.921628846 | -0.213095001 | 0.008294032 | 0.91905722 |

Note:***Statistical significance at 1% level.
**Statistical significance at 5% level.
*Statistical significance at 10% level.

## 4.2. Out-of-sample analysis

### 4.2.1 Forecast evaluation method

Following Wei et al. (2021), Liang et al. (2022), Neely et al. (2014), and Rapach et al. (2010), we use out-of-sample $R^2$ ($R^2_{OOS}$) statistics to evaluate the out-of-sample forecasting performance. We compute $R^2_{OOS}$ using the following equation:

$$R^2_{OOS} = 1 - \frac{MSE_M}{MSE_B} \tag{5}$$

Where:

$$MSE_M = \frac{1}{N}\sum(r_t - \overline{r_{M,t}})^2 \tag{6}$$

$$MSE_B = \frac{1}{N}\sum(r_t - \overline{r_{B,t}})^2 \tag{7}$$

Following Zhang, Ma, and Zhu (2019) and Rapach et al. (2010), we use the historical average, $\overline{r_{B,t}} = \frac{1}{t}\sum_{i=1}^{t} r_i$, as the benchmark.A positive R2 OS means that the forecasting models generate more accurate predictions for the index returns than the benchmark model.Furthermore, we use the CW statistic to test the null hypothesis $R^2_{OOS} \leq 0$ against the alternative hypothesis $R^2_{OOS} > 0$. Mathematically, the CW statistic can be computed with the following equation:

$$f_{t+1} = (r_{t+1} - \overline{r_{B,t+1}})^2 - \left[\left(r_{t+1} - \overline{r_{M,t+1}}\right)^2 - \left(\overline{r_{B,t+1}} - \overline{r_{M,t+1}}\right)^2\right] \tag{8}$$

By regressing $f_{t+1}$ on a constant, we can obtain the CW statistic, which is equivalent to the t-statistic of the constant.

### 4.2.2 Out-of-sample results

We divide the entire sample into an in-sample period and an out-of-sample period. The in-sample period spans from May 1937 to April 1972. The out-of-sample period is from May 1972 to December 2023. We use an expanding window for the analysis to take full advantage of the entire sample.

Table 3 reports the out-of-sample forecasting results.Shows the corresponding $R^2_{OOS}$ and CW statistics for each variable after PCA dimensionality reduction.Figure 1 presents the cumulative squared forecast error (CSFE).And Figure 2 presents the real returns(from May 1937 to December 2023)and the historical average returns(from May 1972 to December 2023)

**Table 3** Out-of-sample estimation results.

| Variables | $R^2_{OOS}$(%) | CW-stat |
|---|---|---|
| D12 | -0.938172109 | -0.441140347 |
| E12 | -1.040769329 | -0.003834016 |
| b/m | -2.042631144 | -0.793022306 |
| tbl | 0.1415554 | 1.386276442 |
| AAA | -0.191858981 | 0.796029424 |
| BAA | -0.341701733 | 0.644732555 |
| lty | -0.164683812 | 0.932820749 |
| ntis | -0.54849304 | 0.491777399 |
| Rfree | -0.050107969 | 0.984451564 |
| infl | 0.415172331 | 1.451011033 |
| ltr | 0.01844448 | 1.489396864 |
| corpr | 0.655601456 | 2.031565738 |
| svar | -0.968600774 | -0.002482302 |
| CRSP_SPvw | -0.321945112 | -1.571955627 |
| CRSP_SPvwx | -0.354022261 | -1.39875802 |
| PCA | 0.1584932137 | 1.29656485 |

Note: This table reports the out-of-sample forecasting performance. The outof-sample $R^2$ ($R^2_{OOS}$) and Clark and West (2007) statistic (CW-stat) are reported. The sample period is from May 1937 to April 1972 and the out-of-sample period is from February May 1972 to December 2023.

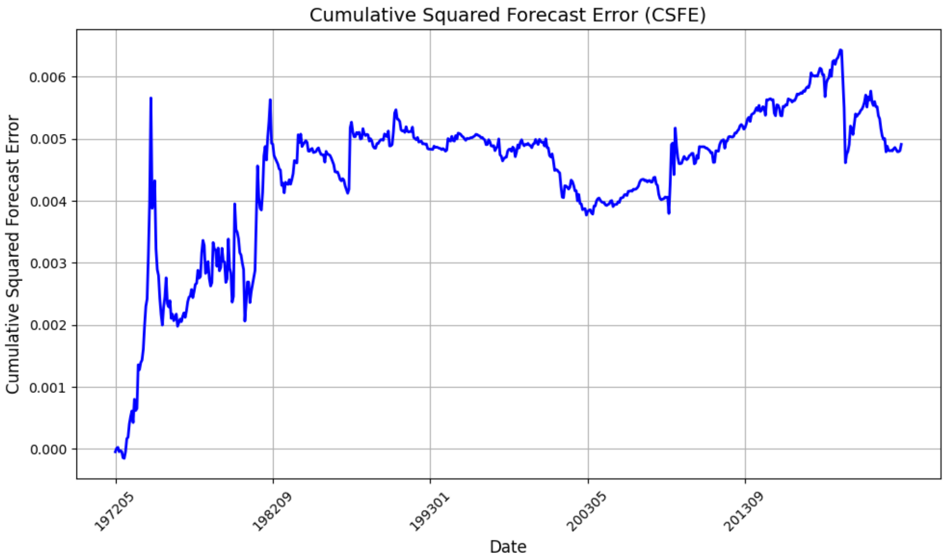


**Figure 1:** cumulative squared forecast error (CSFE),from May 1972 to December 2023.

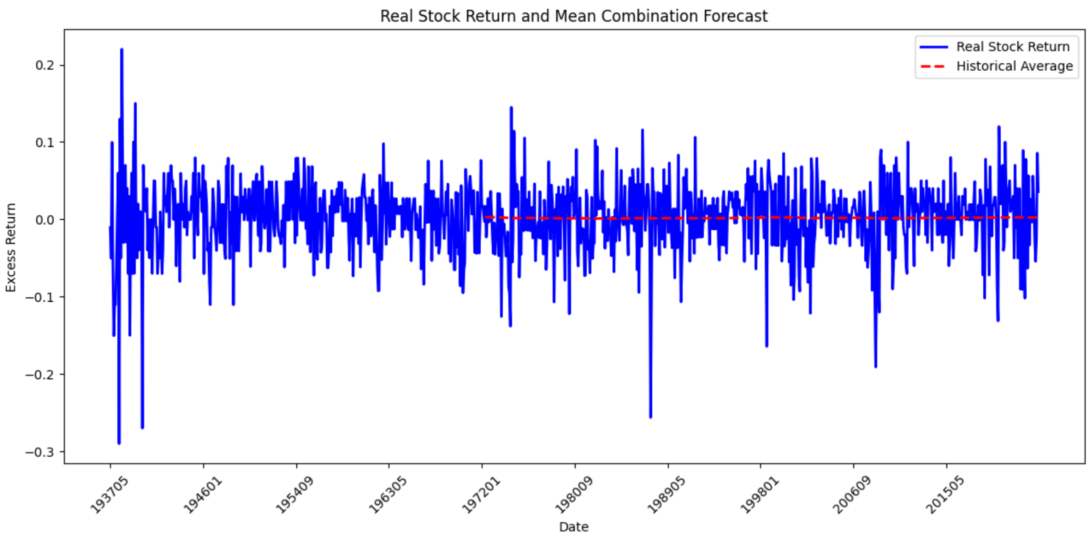


**Figure 2 :** the real returns(from May 1937 to December 2023,blue,Solid line)and the historical average returns(from May 1972 to December 2023,red,Dashed line)

# 5. ARMA model

## 5.1. Analysis of Excess Return Distribution Characteristics

We conduct a distributional characteristic analysis of the excess returns of the S&P 500 index, with Figure 3 showing the time series plot and Figure 4 showing the histogram and summary statistics of the excess returns, respectively(from May 1937 to December 2023).

As can be seen from Figure 3, the excess monthly return fluctuates significantly, often experiencing extremely high or low excess returns, indicating that it is occasionally impacted by significant events.

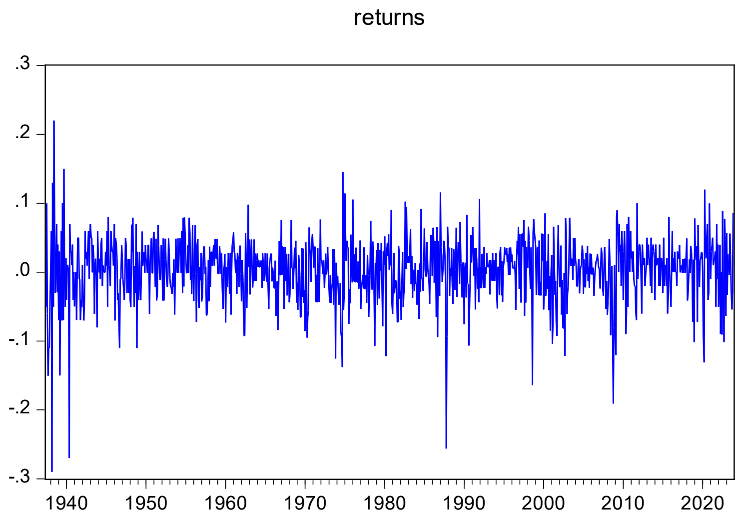


**Figure 3 :** the time series plot of the excess returns.

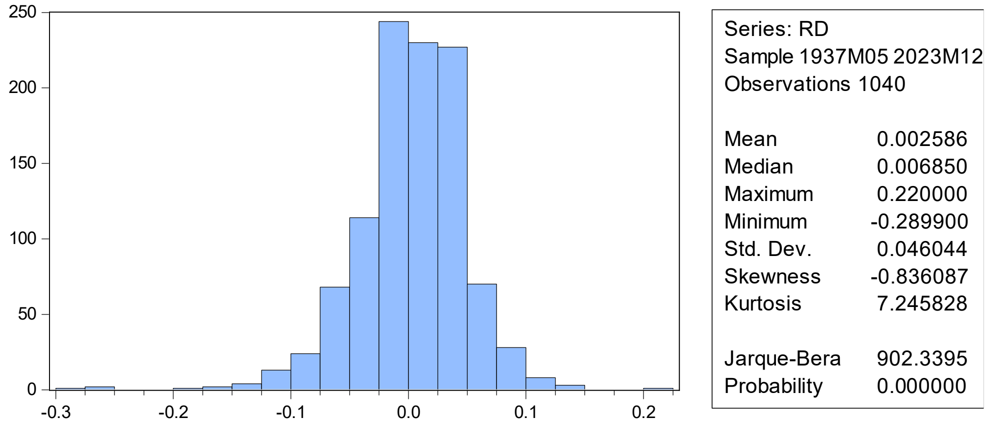


**Figure 4 :** the histogram and summary statistics of the excess returns .

## 5.2. Stationarity Test (ADF Test)

To forecast the excess returns of the S&P 500 index, it is necessary to determine if it is stationary, and therefore an Augmented Dickey-Fuller (ADF) test needs to be conducted.Table 4 reports the t-statistic values and their corresponding p-values for the Augmented Dickey-Fuller (ADF) test, respectively, with intercept, with trend, without intercept, and without trend.

As can be seen from Table 4, the excess return of the S&P 500 index is stationary.

**Table 4** ADF test results.

| | t-Statistic | Prob.* | The t-Statistic of significance at 1% level |
|---|---|---|---|
| Trend and intercept | -32.49008 | 0.0000 | -3.966905 |
| Intercept | -32.48897 | 0.0000 | -3.436425 |
| None | -32.40062 | 0.0000 | -2.567190 |

### 5.3. ARMA(p,q) model

#### 5.3.1 Model recognition

Figure 5 is the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) graph for the S&P 500 index excess returns.

As can be seen from the graph, ACF shows first-order truncation, while PACF exhibits a tailing phenomenon, suggesting an ARMA model might be appropriate.

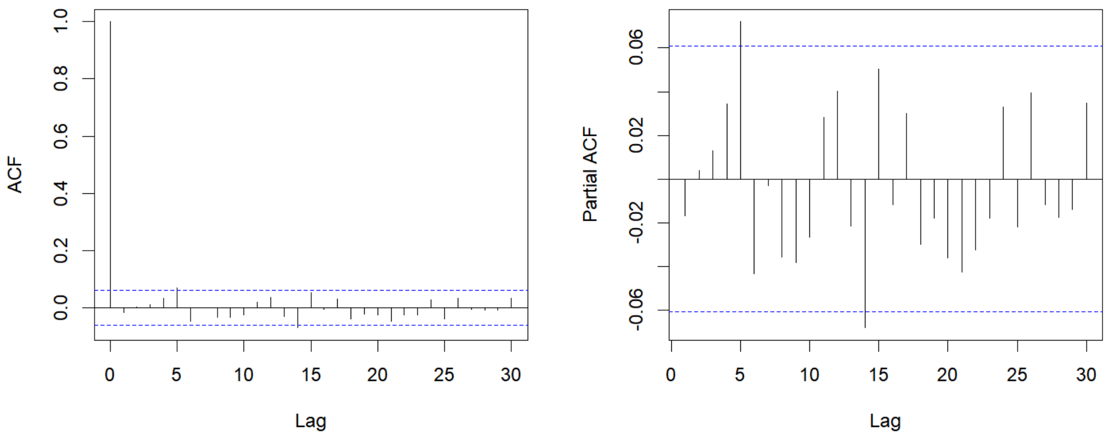


**Figure 5 :** ACF and PACF of excess returns

#### 5.3.2 Determination of lag order

Table 5 records whether the coefficients are significant for ARMA (0,1) to ARMA (5,1), (If the truncation of ACF is very obvious, then the parameter of MA must be 1.)as well as the values of information criteria such as AIC, SC, and HQ.

From the results of Table 5, it can be concluded that the ARMA(5,1) model should be chosen for prediction.

**Table 5** Determination of lag order

| Model | Coefficient significance(10%) | AIC | SC | HQ |
|---|---|---|---|---|
| ARMA(0,1) | 0 | -3.313724 | -3.299454 | -3.308310 |
| ARMA(1,1) | 0 | -3.311818 | -3.292791 | -3.304600 |
| ARMA(2,1) | 0 | -3.313452 | -3.289669 | -3.304430 |
| ARMA(3,1) | 0 | -3.309916 | -3.281376 | -3.299090 |
| ARMA(4,1) | 0 | -3.309460 | -3.276163 | -3.296829 |
| ARMA(5,1) | 1 | -3.312535 | -3.274481 | -3.298099 |

Note: The significance level for determining significance is 10%, with 1 indicating significance and 0 indicating insignificance.

The basic model is set as follows:

$$r_t = c + \beta_1 r_{t-1} + \beta_2 r_{t-2} + \beta_3 r_{t-3} + \beta_4 r_{t-4} + \beta_5 r_{t-5} + \varepsilon_t + \theta_1 \varepsilon_{t-1} \quad (9)$$

### 5.3.3 Model determination and parameter estimation

Table 6 shows the results of the OLS estimation for ARMA(5,1). According to the results, AR(2) and AR(3) are not significant, so AR(2) and AR(3) terms are removed.

Therefore, the final model is :

$$r_t = c + \beta_1 r_{t-1} + \beta_4 r_{t-4} + \beta_5 r_{t-5} + \varepsilon_t + \theta_1 \varepsilon_{t-1} \tag{10}$$

Based on the regression results, the model is:

$$r_t = 0.002581 - 0.49 r_{t-1} + 0.4988 r_{t-4} + 0.08993 r_{t-5} + \varepsilon_t + 0.4804 \varepsilon_{t-1}$$

**Table 6** OLS Estimated result

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 0.002581 | 0.001828 | 1.412413 | 0.1581 |
| AR(1) | -0.489993* | 0.265616 | -1.844741 | 0.0654 |
| AR(2) | 0.001651 | 0.029238 | 0.056467 | 0.9550 |
| AR(3) | 0.021423 | 0.02901 | 0.738483 | 0.4604 |
| AR(4) | 0.04988* | 0.029171 | 1.709959 | 0.0876 |
| AR(5) | 0.089932*** | 0.028533 | 3.151924 | 0.0017 |
| MA(1) | 0.480371* | 0.266492 | 1.802576 | 0.0717 |
| SIGMASQ | 0.0021 | 5.92E-05 | 35.48837 | 0.0000 |

**Note:***Statistical significance at 1% level.**
****Statistical significance at 5% level.**
***Statistical significance at 10% level.**

## 6. Conclusion

This article uses two models to predict the returns of the S&P 500 index. One model uses 15 economic variables to forecast the returns and employs PCA for dimensionality reduction. However, the results were not very good, with a p-value of around 0.1685 and an out-of-sample R-squared of only 0.2%. Although this still exceeds the historical mean method, the overall effect is not satisfactory. The reason for this could be the strong correlation between the 15 economic variables fitted, which indicates the presence of multicollinearity. Additionally, introducing too many "non-effective" variables might also be a reason for the poor fitting results. The other model utilizes the autocorrelation of the index itself, using an ARMA model to predict the index returns, and after adjustments, the fitting results are quite good.

This article innovatively uses these 15 types of economic variables to predict the index, although the results are not very good, it is hoped that it can provide some help for the discoveries of subsequent researchers.

## Acknowledgments

We thank seminar participants at Nanjing University of Science and Technology for helpful comments and suggestions. This work is supported by the National Undergraduate Training Program for Innovation and Entrepreneurship [202410288052Z].

## Reference

[1] Baker, M., & Wurgler, J. (2006). Investor sentiment and the crosssection of stock returns. The Journal of Finance, 61, 1645–1680.

[2] Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. Journal of Economic Perspectives, 21, 208–230.

[3] David,A., M. Belén.,& Antonio P.(2024) Hybrid ARMA-GARCH-Neural Networks for intraday strategy exploration in high-frequency trading.Pattern Recognition,148,110-139

[4] Huang, D., Jiang, F., Tu, J., & Zhou, G. (2015). Investor sentiment aligned: A powerful predictor of stock returns. The Review of Financial Studies, 28, 791–837.

[5] Liang, C., Wei, Y., Lei, L., & Ma, F. (2022). Global equity market volatility forecasting: New evidence. International Journal of Finance and Economics, 27, 594–609.

[6] Neely, C. J., Rapach, D. E., Tu, J., & Zhou, G. (2014). Forecasting the equity risk premium: The role of technical indicators. Management Science, 60, 1772–1791.

[7] Wei, Y., Bai, L., Yang, K., & Wei, G. (2021). Are industry-level indicators more helpful to forecast industrial stock volatility? Evidence from Chinese manufacturing purchasing managers index. Journal of Forecasting, 40, 17–39.

[8] Rapach, D., Strauss, J. K., & Zhou, G. (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. Review of Financial Studies, 23, 821–862.

[9] Wen, D., He, M., Zhang, Y., & Wang, Y. (2022). Forecasting realized volatility of Chinese stock market: A simple but efficient truncated approach. Journal of Forecasting, 41, 230–251.

[10] Zhang, Y., He, M., Liao, C., & Wang, Y. (2023). Climate risk exposure and the cross-section of Chinese stock returns. Finance Research Letters, 55, 103987.

[11] Zhang, Y., Ma, F., & Zhu, B. (2019). Intraday momentum and stock return predictability: Evidence from China. Economic Modelling, 76, 319–329.

[12] Zhang, Y., Wei, Y., Ma, F., & Yi, Y. (2019). Economic constraints and stock return predictability: A new approach. International Review of Financial Analysis, 63, 1–9.

[13] Zhang, Y., Zeng, Q., Ma, F., & Shi, B. (2019). Forecasting stock returns: Do less powerful predictors help? Economic Modelling, 78, 32–39.

[14] Zhang,Y.,Han,Q.&He,M.(2024).Forecasting stock market returns with a lottery index: Evidence from China.Journal of Forecasting,43,1595-1606

**Table 7:** Partial data display

| yyymm | ndex | eturns | 12 | 12 | /m | bl | AA | AA | ty | tis | free | nfl | tr | orpr | var | C RSP_S Pvw | C RSP_S Pvwx |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 93705 | 6.26 | 0.01 | .81 | .15 | .49 | .00 | .03 | .05 | .03 | .03 | .00 | .01 | .01 | .00 | .00 | - 0.01 | - 0.01 |
| 93706 | 5.40 | 0.05 | .84 | .17 | .50 | .00 | .03 | .05 | .03 | .03 | .00 | .00 | .00 | .01 | .00 | - 0.05 | - 0.06 |
| 93707 | 6.98 | .10 | .82 | .19 | .46 | .00 | .03 | .05 | .03 | .03 | .00 | .01 | .01 | .00 | .00 | 0 .10 | 0 .10 |
| 93708 | 6.04 | 0.06 | .79 | .20 | .48 | .00 | .03 | .05 | .03 | .03 | .00 | .00 | 0.01 | .00 | .00 | - 0.05 | - 0.06 |
| 93709 | 3.76 | 0.15 | .77 | .22 | .55 | .00 | .03 | .05 | .03 | .04 | .00 | .01 | .00 | .00 | .02 | - 0.14 | - 0.14 |
| 93710 | 2.36 | 0.11 | .78 | .19 | .62 | .00 | .03 | .06 | .03 | .04 | .00 | .00 | .00 | .01 | .03 | - 0.10 | - 0.10 |
| 93711 | 1.11 | 0.11 | .79 | .16 | .69 | .00 | .03 | .06 | .03 | .04 | .00 | 0.01 | .01 | .01 | .02 | - 0.08 | - 0.10 |
| 93712 | 0.55 | 0.05 | .80 | .13 | .71 | .00 | .03 | .06 | .03 | .03 | .00 | 0.01 | .01 | .01 | .01 | - 0.05 | - 0.05 |
| 93801 | 0.69 | .01 | .79 | .08 | .70 | .00 | .03 | .06 | .03 | .04 | .00 | 0.01 | .01 | .00 | .01 | 0 .01 | 0 .01 |
| 93802 | 1.34 | .06 | .79 | .02 | .66 | .00 | .03 | .06 | .03 | .03 | .00 | 0.01 | .01 | .00 | .01 | 0 .07 | 0 .06 |
| 93803 | .50 | 0.29 | .78 | .97 | .89 | .00 | .03 | .06 | .03 | .03 | .00 | .00 | .00 | 0.01 | .02 | - 0.25 | - 0.26 |

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 9380 4 | .70 | .13 | .77 | .90 | .79 | .00 | .03 | .06 | .03 | .03 | .00 | .01 | .02 | .01 | .02 | 0 .14 | 0 .14 |
| 9380 5 | .27 | 0.05 | .75 | .84 | .82 | .00 | .03 | .06 | .03 | .03 | .00 | 0.0 1 | .00 | .00 | .01 | - 0.04 | - 0.05 |
| 9380 6 | 1.56 | .22 | .74 | .77 | .66 | .00 | .03 | .06 | .03 | .02 | .00 | .00 | .00 | .01 | .01 | 0 .25 | 0 .25 |
| 9380 7 | 2.40 | .07 | .71 | .72 | .63 | .00 | .03 | .06 | .03 | .02 | .00 | .00 | .00 | .01 | .01 | 0 .07 | 0 .07 |

**Data from May 1937 to July 1938 (retaining two decimal places) (partial)**