Predictability of stocks

Dissertation submitted by

Elena Miltiadous

to

Department of Economics,

University of Cyprus

In partial fulfillment of the requirements for the degree of Master in Monetary and Financial Economics

Supervised by: **Philippos Louis**

Cyprus, (04/12/2023)

Abstract

An important question in the financial market is the predictability of stock returns, which has essential and broad economic implications. Stock predictability is highly linked with the efficiency of the capital markets in allocating the available resources combined with their highly valued uses. This thesis analyzes the predictability of stock market returns in the USA. The aim of this research is to identify the factors that affect stocks and make them predictable. As a result, there is a pattern in the stock prices that traders can exploit and make profit from, i.e., beat the market. To find these factors, bivariate models are developed. The period of the analysis started from the 1st of January 1900 and ended up to 31st of December 2021. Regarding the results, there were some factors that could be used in order to predict stock prices. In addition, a homoscedasticity test is shown in order to examine the stability of the variance of the residual. Furthermore, an ARCH test is used to examine whether there is dynamic homoscedasticity. According to the results, the null hypothesis was rejected, that is, there was a strong dynamic heteroskedasticity.

This research provides clear evidence of stock market return predictability in the United States of America using financial variables. Future research can be carried out to compare alternative ways to examine the stock predictability, such as splitting the period into sub periods for a better understanding of the relationship as well as including COVID-19 pandemic variables in order to see the effect of the pandemic on stocks.

Table of Contents

List of Tables

List of Figures

List of Diagram

List of Equations

List of Abbreviations

Definition	Abbreviation
Akaike Informational Criteria	AIC
Autoregressive Conditional Heteroskedatsicity	ARCH
Bayesian Informational Criteria	BIC
Book to Market ratio	bm
Coefficient of Determination	R^2
Corporate issuing activity	ntis
Correlation	CORR
Cumulative Sum	CUSUM
Default yield spread	dfy
Dividend price ratio	dp
Dividend yield	dy
Dow Jones Industrial Average	DJIA
Earnings price ratio	ep
Efficient Market Hypothesis	EMH

Error term	U
ERT	Excess Returns
Generalized Autoregressive Conditional Heteroskedasticity	GARCH
In Sample	IS
Inflation	Infl
Investment to Capital ratio	ik
Lagrange Multiplier	LM
Long Term rate of return	ltr
Mean Absolute Error	MAE
Mean Absolute Percentage Error	MAPE
Model Selection	ms
Neutral Network	NN
Out Of Sample	OOS
Percent Equity issuing	eqis
Price Earnings ratio	pe
Root Mean Square Error	RMSE

Standard and Poor’s 500	S&P 500
Stock Variance	svar
Term Spread	tms
Treasury bill	tbl
United Kingdom	UK
United States of America	USA
Vector Autoregressive Model	VAR

1. Introduction

The two most significant values for stock evaluation are the present value of an investment and the predicted future value. Despite the obvious riskiness of stock transactions, investors always expect a significant return on their investments. Generally, the historical price of stocks has been utilized as the primary determinant in predicting future values, severely affecting investment decisions. The stock markets in the United Kingdom (UK) and the United States of America (USA) are both volatile. However, there are widely established methodologies for stock predictability that are used in worldwide markets. Some investors, for example, avoid stocks that rise while expecting a sharp decline. At the same time, other investors avoid falling stocks for fear of a long-term decline. Regardless of the technique chosen, the predictability of stock prices can play an important role in economic protection by buffering the economy from unexpected and ineffective shocks.

The recent financial crisis of 2007-2008 created major problems for the market. Prior to the shock, people believed that everything was working efficiently, but the markets proved the opposite. Therefore, it is important to conduct relevant research aimed at obsessing over possible solutions and preventive measures. In every economy, asset prices can help with the understanding and prediction of various macroeconomic issues. Decision-making can be influenced by the indexes of prices since the prediction of various behaviors is possible.

The most recent health crisis caused by pandemic COVID-19 generated interest among some scientists to analyze stock return predictability during this shock, as well as the price volatility in the U.S. More specifically, Hui Hong, Zhicun Bian, and Chien-Chiang Lee, through their article, examined the stock market performance, that is, the stocks' predictability and price volatility, during the short period 2019 (1)–2020 (6). Through their research, they conclude that stock predictability as well as price volatility showed an important increase during the crisis. As a result, some investors had the chance to increase their profits, while others suddenly lost their money, and in this way, an unequal distribution of wealth was created in the market. In general, COVID-19 was related to market inefficiency in the United States during that time period (Hong et al., 2021).

Pesaran & Timmermann (1995) investigate whether there is a possibility to predict the USA's stock market. The article focuses on how problematic stock returns can affect the whole performance of the economy. The volatility of the return has been shown to affect the expected return to a significant degree. With publicly available information, it is possible to make significantly accurate estimates of stock returns. Some of the information that can be used contains time-series data on macroeconomics and financial variables. This conclusion, according to the authors, applies to all markets, regardless of time horizon differences. Theoretical efforts to predict the excess returns will face certain uncertainties and inaccuracies. Using methods that have been verified through their use in investment strategies is one of the ways that contradictions such as inefficiencies in the markets can be managed to prevent them from affecting the result. The outcomes from the article are consistent with the fact that it is possible to predict stock performance in Britain and the U.S.A. efficiently since the recommended methods are applied worldwide (PESARAN & TIMMERMANN, 1995).

Over the years, stock performance forecasting has been widely researched, with various authors using different methodologies. Welch & Goyal (2008) base their arguments on the fact that most models from the past are unstable and cannot be used effectively in determining stock performance. Variance, book-to-market (b/m), different interest rates, income ratios, and wealth are some of the variables that have been used by the authors. To determine the level of usability and practicality, the authors re-reviewed empirical data from 2006. According to the authors, the insignificance of most models is the reason for unstable outcomes to the extent that the performance of stock returns concerns us. Constant reference is made to data from the S&P 500. The fact that the article is based on real data implies that the suggested models are applicable to the stock markets of the United States and Great Britain (Welch & Goyal, 2004).

To investigate the economy's performance, we can use the excess returns of stock markets. According to Qi (1999), it is possible to predict excess returns. The economic bibliography is focused on linear predictability and nonlinearity. The relationship between predicting variables and excess returns is investigated using a neural network model. This model is particularly more attractive than the other models because of its ability to perform nonlinear functional approximation with more flexibility. The author of this article was working to improve previous studies of other authors. For instance, the author changed the set of investors' choices to identify the change in output. The simple and popular linear regression model has been widely used to predict stock returns using economic and financial variables. Neural networks have also been used to understand trends in stock performance due to their increased ability to provide better results for both linear and non-linear functions (Qi, 2012).

Marcellino, Stock, and Watson (2006) intend to achieve a more practical approach to forecasting by distinguishing between direct and empirically repeated predictions and between univariate and bivariate models. The main reason that they decided to proceed with something more than a simple theoretical approach was that there were always concerns about the effectiveness of iterated forecasts in contrast with direct predictions. According to their research, where a model is able to choose long-lag specifications, iterated forecasts perform much better than direct forecasts. The performance of iterated forecasts is improving with each change in the prediction's time horizon. Choosing between a one-period iterated model and a multi-period model is frequently necessary. The multiperiod is often projected using a loss function that has been substantially adjusted to match the forecast horizons (Marcellino et al., 2015).

2. Literature Review

The predictability of stock returns pushed the researchers to try to find the factors that determine equity in order to identify whether the stocks can follow a specific path. This section gives an overview in order to conduct a comparative review of some important key variables that affect stock returns. In the literature review, five studies are analyzed. The two of them used the same methodological approach, which was the IS model and the OOS model. In addition, one study used the VAR method. Furthermore, the fourth study used the recursive regression approach. The last study showed the impact of COVID-19 on the predictability of stock market returns, and it used the VAR model and the GARCH model. These studies differ since they used different approaches as well as different variables; however, the aim of these studies is the same.

Marcellino, Stock, and Watson (2004) studied the use of out-of-sample simulation techniques to analyze iterated and direct forecasts in univariate and bivariate autoregressions. In addition, they used 170 monthly US time series, and the period started on 1st of January 1959 and ended up to 31st of December 2002. They also used five categories of series (Marcellino et al., 2015).

Regarding the results for univariate model, in the case of the distributions of the ratios of the MSFE, in general, it was mentioned that if the model is adequately defined, iterated predictions can perform better, while direct forecasts are more robust to the errors that may exist in the model. The general dilemma in the text is which of the two approaches is more trustworthy for the majority of time series and, more specifically, if the MSFE is smaller for the direct forecasts. Models with a fixed lag order as well as models with a choice of lag order that depends on the data are studied through the Akaike Information Criterion (AIC) or the Bayes Information Criterion (BIC). Forecasts are made in terms of one quarter, six months, one year, and two years. The forecast error for direct forecasts appears to be lower than the error for iterated forecasts, according to the author's investigation using the presented models (Marcellino et al., 2015).

According to the results for univariate autoregressions, whether the iterative or direct estimator is superior depends not only on the lag selection method but also on the time. For example, if the lag order is equal to 4 (p = 4), which is selected with the BIC method, and the time horizon is equal to 3 (h = 3), then the direct prediction method seems to be slightly more efficient. This can happen by defining a lag through the BIC method for a period of time up to h = 12. On the other hand, for a longer period of time (up to h = 24), the iterated method seems to be more efficient for all lag selection methods (Marcellino et al., 2015).

The results for the nominal prices, wages, and money series differ from the results for the other series. For this category, the direct method seems to be preferred for forecasts across all time horizons, but not for all values of lag. The larger the value of the lags, the larger the forecast error, so the direct forecast model loses its performance and the iterated forecast model becomes better. On the other hand, for all other categories of series, the iterated method is preferable for each time horizon. More specifically, the best performance is at smaller lags in these categories of series. In general, the iterated method with lag selection through the AIC has, for all variables, the lowest error for all time horizons. According to the results for the bivariate forecasts, they used a stratified random sample of the VARs, as well as, they compared the iterated and direct forecasts, while the lag length selection was the same. The results shown that, they were same as the results from the univariate models. The authors conclude that iterated forecasts are more efficient, which contradicts the theoretical literature, which emphasizes the advantages of direct forecasts (Marcellino et al., 2015).

Compared with the master thesis, this study used bivariate models without using lag order selection and, consequently, a VAR model. Moreover, Marcellino, Stock, and Watson (2004) focused on out-of-sample forecasts by using both univariate and bivariate models, while in this thesis the estimation was about the bivariate models until time t. That is, the outcome didn't focus on what happened after time t. Another difference is that they used AR models for the univariate model as well as VAR models in the case of the bivariate models. However, in this thesis, the methodology follows the bivariate models and a single regression without the use of a VAR model.

Goyal & Welch (2007) examined the performance of variables that are suggested by the academic literature as good predictors of the equity premium. The dependent variable was the equity premium. Regarding the methodology, the authors used OLS in order to estimate the regression coefficients. Moreover, they used the in-sample (IS) and out-of-sample (OOS) procedures. They investigate three time periods. Firstly, the OOS forecasts 20 years after the available data; the second time period started OOS in 1965; and the third period ignores all the data prior to 1927, even in the estimation (Welch & Goyal, 2004).

Generally, in this article, there is a different approach to stock predictability. Significant, insignificant, and time-changing models are used for the analysis. Through insignificant models, the variables that were studied were d/p, d/y, and e/p. Based on the elements of significance, these price indices are statistically significant at a 90% significance level. However, through significant models, they studied the variables b/m, i/k, ntis, and eqis. The b/m variable is statistically significant at the 6% IS level, and the i/k variable is statistically significant at the 5% IS level. B/m performed well only in the first semester of the sample in both IS and OOS. Both variables have superior performance. Eqis maintains its strong performance until 1930, but soon gives back the additional benefits. Moreover, eqis has superior performance during the oil crisis for both IS and OOS, and it is the only variable that is statistically significant for OOS on yearly data. That is the significant difference between those two variables. They investigated the variables caya and ms using time-varying models. Some models can predict long-term returns better than short-term ones. In conclusion, the models do not lead to correct predictions and are unstable. The author explains that the literature generally argues that it is important to be able to forecast such values; consequently, it is important to make predictions in models that are as trustworthy as possible and that use updated variables (Welch & Goyal, 2004).

This thesis relies largely on this research, but with some differences in the methodology. Specifically, most of the variables are the same. However, the methodology differs between the two studies. More specifically, in this study, they used the IS model as well as the OOS model, while this master thesis used simple regression with the bivariate models. Another difference is that this study used the equity premium as the dependent variable, while in the thesis; the excess return of the S&P 500 was used (Welch & Goyal, 2004).

The predictability of excess returns has been classified as linear or nonlinear in the finance literature. Qi (1999), in this article, examined whether the nonlinear side performs well and can be used for such predictions. He examined the nonlinear predictability of the excess returns. The aim is to compare the goodness of fit of linear regression and neutral network models (non-linear), as well as the control of portfolio performance using the alternative forecasts. According to the methodology, it is assumed that stock returns can be predicted by the mean of a set of financial and macroeconomic indicators without knowing the prices of the variables. As a result, the best thing an investor can do in such a situation is to select the most capable model that will help in predicting. The model that can make the best prediction without knowing the exact values of the parameters available is the neutral network (NN). In order to prevent any impact on the data from future periods, predictions are made through the model before the data for that time becomes available. In addition, forecasts are generally adjusted when new information becomes available, allowing investors to adapt to changing economic conditions. This research investigates the performance of linear and nonlinear models in stock predictability, with a focus on nonlinear Neutral Network (NN) models. Relevant studies have found that in daily stock returns, these models can work in a sample but lose their ability to predict outside of the sample. For monthly stock returns, no substantial difference was found between linear and nonlinear models. The author used nine financial and economic variables. The same data is used in predictions with linear and nonlinear models to derive valid results for which model are more accurate in prediction. There are 468 observations in the sample, dating from the 1st of January 1954 until 31st of December 1992. Moreover, the recursive forecasts started from 1st of January 1960 and ended up to 31st of December 1992 with predictions made by both linear and nonlinear models (Qi, 2012).

Regarding the results, a comparison of the fit and forecasting accuracy of the models is made due to the following measures: RMSE, MAE, MAPE, CORR, and sign. According to these measures, neutral network models seem to have a better fit than linear models. In the case of the in-sample (IS), MAE, MAPE, and RMSE are smaller for NN models, while CORR as well as sign are larger for NN. According to the author, the better-fitting nonlinear NN models imply that they capture substantial nonlinearity, which linear models cannot fully capture. In the case of out-of-sample (OOS) models, the nonlinear NN models fit the data better than the linear model. Moreover, the nonlinear NN model provides fairly accurate forecasts. In addition, the nonlinear model showed statistically significant market timing ability in 1960 and 1970. The linear forecast model, on the other hand, displays it only throughout the 1970s. Even though the squared forecast error for NN models is always shown to be smaller, based on the Diebold and Mariano tests, the improvement is not statistically significant. However, at the linear predictions, the coefficient and the intercept are not differentiating from 0, but at the NN predictions, the coefficient is statistically significant at the 5% significance level (Qi, 2012).

Being able to predict something does not necessarily result in our profitability increasing. The profitability of an investor depends on the cost of the transaction and the trading strategy. A strategy based on recursive forecasts does not have such a high profit. When transaction costs are low, even though during the 1970s the portfolio based on recursive linear forecasts appeared to perform better, the performance of the portfolio based on NN forecasts during the periods 1970 and 1980 was better than the market portfolio. When transaction costs are high, the mean return on both forecasts is higher than on the markets in the 1970s (Qi, 2012).

The article concludes that the use of the non-linearity model allows the investor to achieve the highest possible accuracy and the best risk adjustment. In addition, the study shows evidence of nonlinear predictability in US stock market returns using economic and financial variables.
For this study, there are two differences regarding the master thesis methodology. Firstly, this study used the NN model, which is completely different from the model used in the master thesis, which were the bivariate models as mentioned above. Secondly, the author used the IS model as well as the OOP model for forecasts (Qi, 2012).

Pesaran and Timmermann (1995) examined whether the evidence from previous studies about the predictability of stock returns holds, as well as whether it would be possible to use such predictions to increase investor profitability. Some research has found that stock returns can be accurately forecasted using publicly available information across international markets and time periods. The data was in monthly frequency, and the time period started from the 1st of January 1954 and ended up to 31st of December 1992. Moreover, the recursive model selection as well as the estimation was based on monthly observations. All measurements of the variables included in the forecasting process have been recorded from January 1954 until December 1992. Stock prices were measured by the S&P 500 (PESARAN & TIMMERMANN, 1995).

According to empirical results about the robustness of the predictability of stock returns, the authors found that the excess returns using the recursive model, which was based on AIC, BIC, and SIC criteria, have similar patterns. In contrast, the standard error of recursive excess returns has the tendency to increase over time. Furthermore, they found that the predictability of stock returns by the recursive forecasts relative to the actual excess returns increased over time. This implies that the predictability of excess returns in the US is affected by periods of economic instability (PESARAN & TIMMERMANN, 1995).

Examining the hypothesis that excess returns can be predicted will always be dependent on the model used. Two separate evaluation approaches are employed to determine whether such forecasts could benefit the economy. The first approach involves portfolio managers assessing in real time if these portfolios continuously provide excess returns, and this method has the considerable advantage of allowing decisions to be based on past values. On the other hand, disadvantages may arise since the factors that produce the forecast are not evaluated, as well as the dilemma that managers may not use only public information. The second approach that, if used, requires more attention is replicating real-time investor decisions using information that is publicly available for parameters deemed to be relevant to predicting stock returns. However, many studies show this method to be ineffective because, in "real time," no investor could have acquired parameter estimations based on the whole sample. Similar issues might arise with forecasting models since they may ignore uncertainty and its consequences for investors' portfolio strategies. Additionally, transaction costs are taken into account while simulating the decisions of investors' portfolios. This article analyzes cases where transaction costs are both low and high in order to determine whether the information provided may be used to enhance the economy. The existence of several model selection criteria demonstrates the problem of uncertainty in model selection (PESARAN & TIMMERMANN, 1995).

Whether an investor believes that stock returns are predictable despite not knowing the true values of the parameters, the best he can do is choose the most suitable model for his situation.

It is worth noting that, until the 1970s, the AIC and BIC model selection criteria were not recognized. The variables considered crucial in predicting stock returns are as follows: company earnings, liquidity measures, short and long interest rates, dividend yields, the inflation rate, and industrial production. According to the authors, stock returns depend not only on the evolution of the business cycle but also on the magnitude of the shocks. Moreover, based on the recursively selected regression models, they found that during the relative calm market of 1960, there was no excess return that would be gained using a switching strategy (PESARAN & TIMMERMANN, 1995).

Based on the data in this article, the author concludes that an investor who follows such predictions does not definitely increase his earnings. One reason is that the excess returns do not follow a standard distribution, so they may not be predictable enough to generate profit. Additionally, transaction costs might have an impact on an investor's profitability. The author concludes that if any investor might have benefited from these predictions, it would have been during a period of market volatility in the 1970s (PESARAN & TIMMERMANN, 1995).

This research seems to present some differences with this master thesis methodology. In order to examine the predictability of the stock returns, the authors divide the period into subperiods for a better understanding of this relationship. Moreover, they used the recursive selection regression model, while in this thesis, the bivariate models were used (PESARAN & TIMMERMANN, 1995).

Hong, Bian, and Lee (2021) examined how the COVID-19 pandemic affected the stock market's predictability and price volatility in the U.S. The data used in the study is collected daily and employed for more accurate identification of structural breaks in the regression models. Moreover, they used data starting from 1st of January 2019, which was the beginning of the pandemic, until 30th of June 2020. They noticed a break in the forecast models for both price volatility and stock predictability in February 2020 (Hong et al., 2021).

According to the results, when they used the S&P 500 return model and the DJIA return model, all the coefficients were statistically significant at the 1% significance level. In addition, the authors did some robustness tests as well as a heteroskedasticity test and a serial correlation test. The results show that the estimated coefficients remain statistically significant at the 1% significance level. In the case of heteroskedatsicity and serial correlation, the S&P 500 return model and DJIA return model are approximately affected. According to the authors, the regression models are valid to predict US stock returns with no statistically significant changes in the estimated coefficients after possible problems are accounted for. Regarding the results for the structural break, the authors found that there is one break prediction model for both the S&P 500 and DJIA. Also, the predictive regression model is expressed by structural instability. Furthermore, the results show that the estimated coefficients for the S&P 500 and DJIA changed significantly following the break (Hong et al., 2021).

In addition, according to the authors, the stock return predictability was dependent on the structural break, which was attributed to COVID-19 during the investigated period. Moreover, they stated that COVID-19 was responsible for the market inefficiency. This implies that it allowed for stock return predictability; that is, it gave traders profitable opportunities. Furthermore, they test for breaks and their significance. According to the results, they found a structural break in the volatility of the S&P 500 and DJIA. Since both the modified CUSUM test and the LM test reject the null hypothesis for constant variance at the 5% significance level when they use cross-validated bandwidths, the important implication is that they found a single break around 21st of February 2020. To conclude, COVID-19 is responsible for the U.S. market's inefficiency during that period, which produced profitable opportunities. Finally, crises may induce wealth and income inequality (Hong et al., 2021).

Regarding this study, there are some differences compared to the master thesis. In this study, the authors used some COVID-19 pandemic variables to show the effect of the pandemic on the predictability of stock returns, while in the master thesis, these COVID-19 variables were not included. Secondly, the data they used was collected on a daily basis. In contrast, the data used in the master thesis was expressed in terms of monthly frequency. Finally, they used structural break tests, while in this thesis, this test was not used (Hong et al., 2021).

3. Theoretical Background

3.1 Theoretical background and explanation of the expected results

This chapter presents the methodology that will be used to analyze the predictability of stock market returns for the United States of America. Following this, a brief description of the variables that will be used in the model will follow. In addition, a preliminary analysis using descriptive statistics is presented. Furthermore, a comprehensive analysis of the econometric method and the model is being shown. Also, figures and tables regarding the results of the descriptive statistics as well as the results from the econometric analysis will be presented.

Efficient Market Hypothesis (EMH)

In efficient markets, stock prices should be equal to their fundamental value. The stock prices may present an upward or downward trend at any time on a particular day. Maurice Kendall (1953) did not find any specific and predictable pattern in stock price changes. An important statement of the Efficient Market Hypothesis is that stock prices already reflect all the available public information. This implies that there is no other available information that investors can use and exploit in order to make a profit in the market. Another important point in the Efficient Market Hypothesis is that a forecast for a stock that has favorable performance must lead to favorable current performance, as the market participants want to trade the information. As a result, the stock prices change until the expected returns are commensurate with the risk. Finally, the Efficient Market Hypothesis (EMH) assumes that information is available at zero cost, which is a strong assumption (DOWNEY et al., 2023).

According to the Efficient Market Hypothesis, the new information is unpredictable, since if it were predicted, then the prediction would be part of today's information. As the new information changes the stock prices, then the stock prices must be unpredictable. That is, the stock prices changes follow a random walk (DOWNEY et al., 2023).

The Efficient Market Hypothesis includes three types. The first type is the Weak EMH, which contains an information set of past prices. In addition, it assumes that past prices cannot be used to predict the future prices of any financial instrument. The second type is the semi-strong EMH, which contains an information set of all the available public information. Regarding the semi-strong EMH, stock prices can adjust quickly in response to new publicly available information that is disclosed. As a result, there is no way for traders to beat the market by using both fundamental analysis and technical analysis. The last type is the Strong EMH, which contains an information set of all the available public information as well as private information. Finally, most of the time, the tests are based on a semi-strong form of efficiency (MAVERICK et al., 2023).

Furthermore, there are some issues that make markets inefficient. Firstly, there is the magnitude issue. According to this issue, only managers of large portfolios can make large trading profits since they invest a large amount of money. As a result, they can make the exploitation of minor mispricing worth the effort. Secondly, there is the issue of selection bias, where only unsuccessful investment schemes are publicly available. The good schemes remain private. Finally, there is the lucky event issue, which is about whether an investor can beat the market with luck or manage to beat the market because there was a pattern and they exploited it by making a profit (DOWNEY et al., 2023).

To conclude, in this thesis, the result is expected to converge with the theory of the Efficient Market Hypothesis, i.e., the stocks in the United States of America are not predictable. Thus, there is no pattern that can help market participants make profits. In addition, in the case of whether the market is efficient, the empirical results showed that the performance of professional managers is broadly consistent with the Efficient Market Hypothesis.

Random Walk Theory

The Random Walk (RW) theory refers to a mathematical equation for stock market prices. The supporters of this theory believe that the stock prices are following a random walk. It is a statistical phenomenon where a variable cannot predict the path that will follow since there is no discernible pattern or trend. Moreover, it moves seemingly at random. Furthermore, the Random Walk theory was laid out by Burton Malkiel and implies that the price of any security moves randomly, and therefore, any attempt to predict the future movement of that stock, either through fundamental analysis or technical analysis, leads to failure. In addition, the proponents of this theory recommend to the traders the ‘’buy and hold’’ strategy and the selection of stocks that represent the overall market, such as the S&P 500 Index (CFI Team, 2023b).

There are also some fundamental assumptions for the random walk theory. First of all, it assumes that the price of each investment in the stock market moves in a random walk. The second assumption refers to the fact that the price change in one stock is independent of the price movement in another stock. Generally, it is impossible to predict the future movement of the stock price, and as a result, it is impossible for a trader to beat the market in the long run, based on the random walk theory. Following that, an investor's ideal strategy is to invest in a market portfolio, which is a portfolio that reflects the whole stock market and whose prices completely reflect the change of each security's price in the market (CFI Team, 2023b).

In contrast, there are some criticisms of the random walk theory. The main criticism is that the market consists of a huge number of investors, and the amount of time spent by each investor is different. As a result, it is possible in the short run for trends to be shown in the prices of the securities. Thus, investors can beat the market by buying when the price of a security is low and selling when the price has an upward trend. In addition, there is another criticism of the random walk theory. It implies that stock prices can follow patterns or trends even in the long run. They argue that the stock price can be affected by an extremely large number of factors (CFI Team, 2023b).

4. Methodology

4.1 Description of the methodology

This chapter presents the methodology which will be used to analyze the predictability of the stocks in the United States of America. Following this, a brief description of the variables that will be used in the model will follow. In addition, a preliminary analysis using descriptive statistics is presented. Furthermore, a comprehensive analysis of the econometric method and the model is presented. Also, in the case of heteroskedasticity, this chapter includes and presents ways to solve this issue. Also, figures and tables regarding the results of the descriptive statistics will follow.

In order to examine the relationship between the risk and stocks returns in the USA, data were used from the Federal Reserve Bank of St. Louis Goyal & Welch (2008). The data is a monthly time series and the period of the analysis starts from January 1st of 1990 and ends up to December 31st of 2021. The data is time series since it refers a sequence of statistical data for a specific period of time. In this study the following eleven (11) variables were used:

4.1.1 Theoretical explanation of the variables

The dependent variable in the bivariate models is defined by Excess return (ERT), which is calculated as the difference between the stock return and the risk-free rate.

- **Stock returns:** Stock returns are consistently outperformed by stock index returns; in our model, we will use the S&P 500 stock index.

- **Risk-free Rate:** The risk-free rate is defined as an investment that is not subject to financial loss. It is the interest that an investor would receive from an entirely risk-free investment over a certain time period. Also, the risk free rate represents the treasury-bill.

The independent variables are defined as follows:

1. **Dividend price ratio (dp):** Dividends are amounts paid out of the S&P 500 index and last for a year. They are influenced by price indices because as the index rises, the owner's dividend will also increase, and vice versa. Dividend prices are represented as dp. The dividend price ratio is the difference between the logarithm of dividends and the logarithm of prices. It shows how much a company pay dividend relative to its stock price.

2. **Dividend Yield (dy):** Dividend yield is the dividend paid by a company as a percentage of the cash value of the share for a specific period of time. It is the difference between the logarithm of dividends and the logarithm of lagged prices.

3. **Stock Variance (svar):** Stock variance is stock volatility, which is measured by the sum of the squares of the daily returns of the S&P 500 index.

4. **Treasury Bills (tbl):** Treasury bills are short-term bonds issued by the government and have the lowest risk in the money markets.

5. **Inflation (infl):** Inflation is measured by the consumer price index. It has negative effects on the economy since it reduces the purchasing power of incomes and restricts the flow of savings.

6. **Corporate Bond Return-Default Yield Spread (dfy):** It is the corporate bond yield, where dfy represents the difference between the nominal yield and corporate bonds (BAA-AAA).

7. **Book-to-market ratio (bm):** The book-to market ratio is defined as the comparison between the market value of a stock, to its book value. Its value it is obtained by the current stock price over the book value per share for the previous quarter.

8. **Long Term Rate of Returns (ltr):** The long-term rates of returns are the long-term returns that have a positive relationship with stock indices. In addition, is the rate of return, that an investor can get when investing in fund with long term frame.

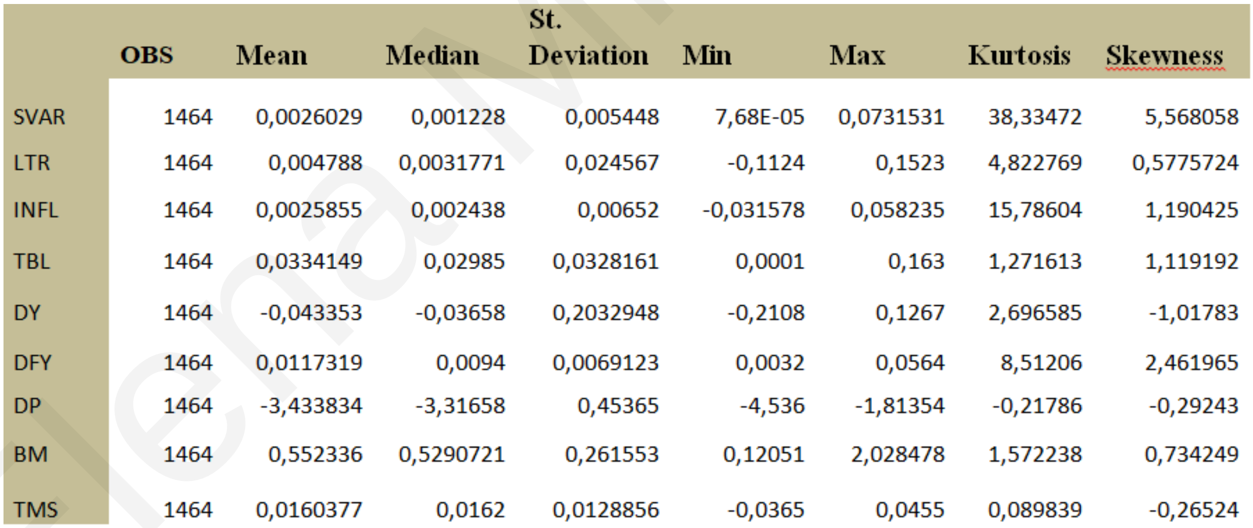
9. **Term Spread (tms):** Term spread is the difference between the long-term yields on government bonds and T-bills. Moreover, investors use the spread as an indication of the relative pricing or valuation of a bond.

4.1.2 Statistical description of the variable

Table 1 below summarizes the main characteristics of the variables that will be used in the study and may give a first picture of the variables. Descriptive data are based on a monthly basis, from January 1900 to December 2021. The maximum number of observations is one thousand four hundred sixty-four (1464). As shown, the BM variable has the highest average return, while the DP variable has the lowest value. Moreover, the DP variable shows the highest standard deviation, while the SVAR variable has the lowest standard deviation value. The DP variable has the minimum value. However, the BM variable has the highest value.

Further elements that are given on the table are skewness and kurtosis. In general, a frequency distribution can be either symmetric or asymmetric. As we can see from Table 1, the SVAR variable has the highest kurtosis, while the DP variable has the lowest kurtosis. Finally, regarding skewness, the SVAR variable has the highest value. In contrast, the DY variable has the lowest value.

Table 1: Descriptive statistics of the variables

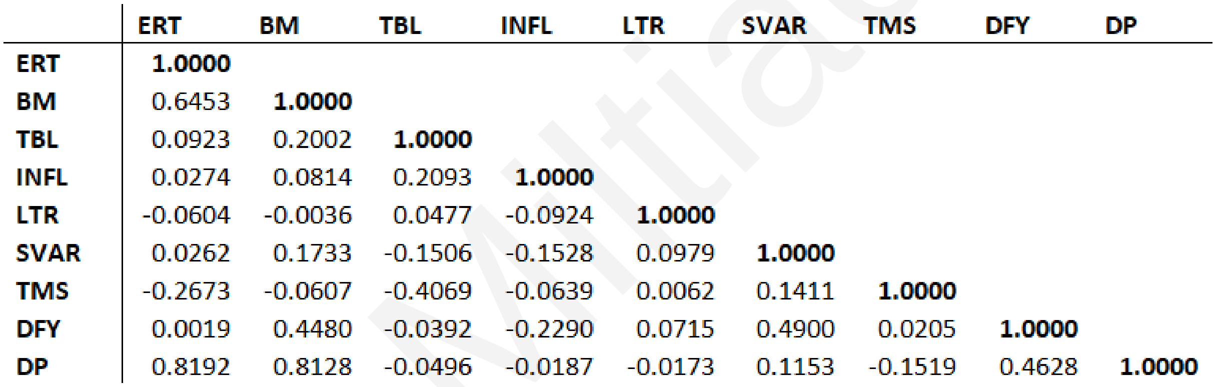


	OBS	Mean	Median	St. Deviation	Min	Max	Kurtosis	Skewness
SVAR	1464	0,0026029	0,001228	0,005448	7,68E-05	0,0731531	38,33472	5,568058
LTR	1464	0,004788	0,0031771	0,024567	-0,1124	0,1523	4,822769	0,5775724
INFL	1464	0,0025855	0,002438	0,00652	-0,031578	0,058235	15,78604	1,190425
TBL	1464	0,0334149	0,02985	0,0328161	0,0001	0,163	1,271613	1,119192
DY	1464	-0,043353	-0,03658	0,2032948	-0,2108	0,1267	2,696585	-1,01783
DFY	1464	0,0117319	0,0094	0,0069123	0,0032	0,0564	8,51206	2,461965
DP	1464	-3,433834	-3,31658	0,45365	-4,536	-1,81354	-0,21786	-0,29243
BM	1464	0,552336	0,5290721	0,261553	0,12051	2,028478	1,572238	0,734249
TMS	1464	0,0160377	0,0162	0,0128856	-0,0365	0,0455	0,089839	-0,26524

4.1.3 Correlation of variables

The correlation among variables can take values between -1 (where there is a negative correlation) and +1 (where there is a positive correlation). The correlation is characterized as negative when the small values of one variable correspond to the large values of the other variable, and vice versa. On the other hand, in positive correlation, small values of one variable correspond to small values of the other, and vice versa. More specifically, with the sign of the correlation, the relationship between the variables becomes observable. The greater their absolute value, the stronger the correlation between the two variables. The perfect correlation is achieved when the correlation is equal to 1.

Table 2: Correlation matrix of the variables



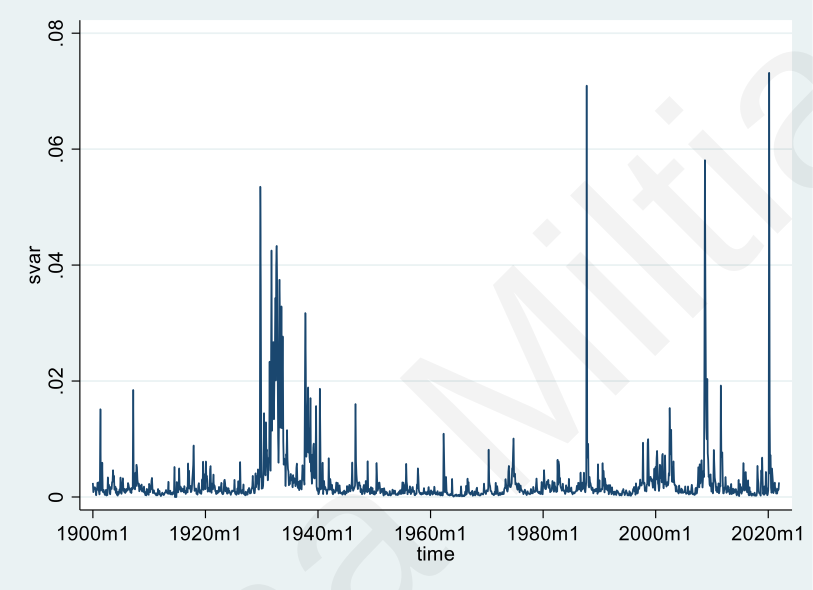
	ERT	BM	TBL	INFL	LTR	SVAR	TMS	DFY	DP
ERT	**1.0000**								
BM	0.6453	**1.0000**							
TBL	0.0923	0.2002	**1.0000**						
INFL	0.0274	0.0814	0.2093	**1.0000**					
LTR	-0.0604	-0.0036	0.0477	-0.0924	**1.0000**				
SVAR	0.0262	0.1733	-0.1506	-0.1528	0.0979	**1.0000**			
TMS	-0.2673	-0.0607	-0.4069	-0.0639	0.0062	0.1411	**1.0000**		
DFY	0.0019	0.4480	-0.0392	-0.2290	0.0715	0.4900	0.0205	**1.0000**	
DP	0.8192	0.8128	-0.0496	-0.0187	-0.0173	0.1153	-0.1519	0.4628	**1.0000**

Table 2 above shows the correlation matrix between the dependent variables and the explanatory variables. The dependent variable ERT has a negative correlation with the variables TMS and LTR and a positive correlation with the variables SVAR, BM, TBL, DFY, INFL and DP. It is obvious that there is almost a full correlation between the dependent variable and DP with 0.8192. The variable BM has a negative correlation with the variables LTR and TMS and a positive correlation with the variables DFY, TBL, INFL, DP, and SVAR. This variable also shows a high positive correlation with the DP with 0.8128 (close to 1). Moreover, the variable TBL shows a negative correlation with the variables TMS, DFY, DP, and SVAR, a positive correlation with the variables, and a positive correlation with LTR, and INFL. According to table 2, the variable TMS shows a negative correlation with DP and a positive correlation with DFY. The variable LTR has a positive correlation with DFY, TMS, and SVAR but a negative correlation with DP. Inflation has a negative correlation with LTR, SVAR, TMS, DFY, and DP, as shown. SVAR has a positive correlation with the variables TMS, DFY, and DP. Finally, the variable DFY is positively correlated with the variable DP.

4.2 Variables over time

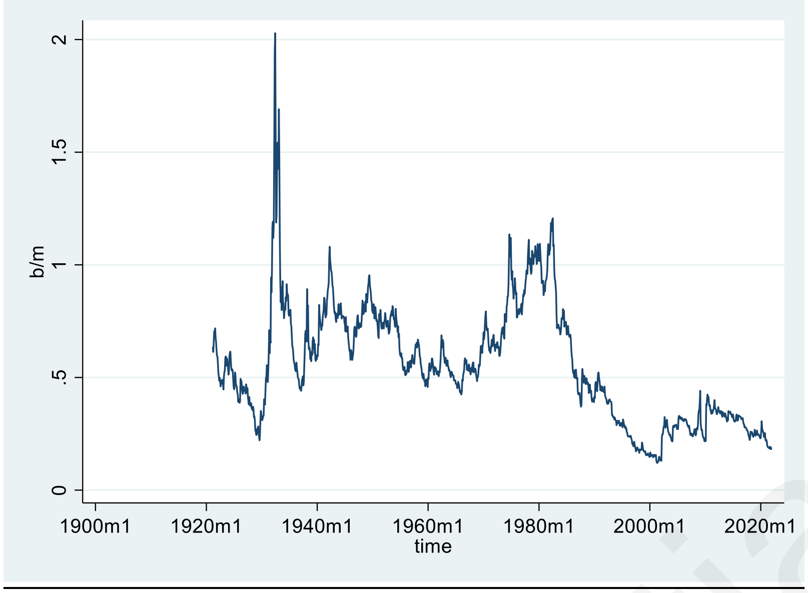
→ The diagrams below show how the variables behave over time:

Diagram 1: Stock Variance (SVAR)



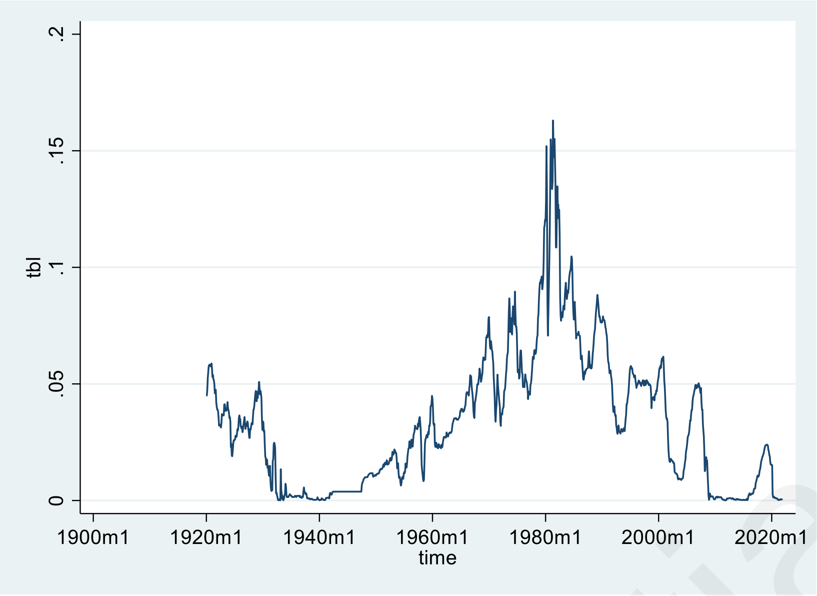
Observations: Diagram 1 above shows the chronological order of the American price index of the SVAR variable. As we can see from diagram 1, the variable is not stationary because the mean and variance are not constant. We notice that there were sharp fluctuations during the time period when the financial crisis and pandemic broke out.

Diagram 2: Book to Market Ratio (B/M)



Observations: In diagram 2, the variable B/M is characterized as a non-stationary time series in the first time frames, but then there is a stationarity with a minimal decrease. This phenomenon is explained on the basis of the crisis that broke out.

Diagram 3: Treasury Bills (TBL)



Observations: The diagram 3 shows the evolution of the time series for the Treasury Bills variable. The control can be based on the graphic display of the series. We notice that both the mean and the variance are not constant over time.

Diagram 4: Long Term Rate of Returns (ltr)

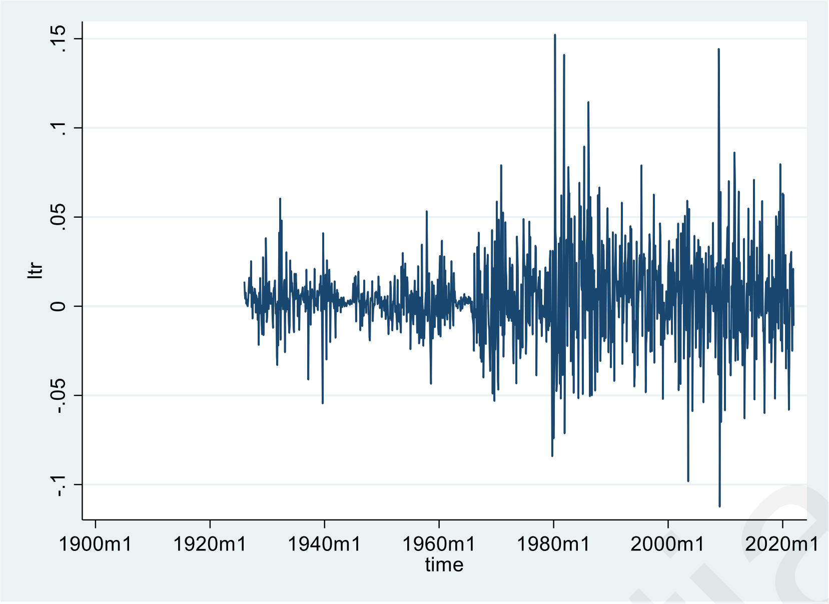
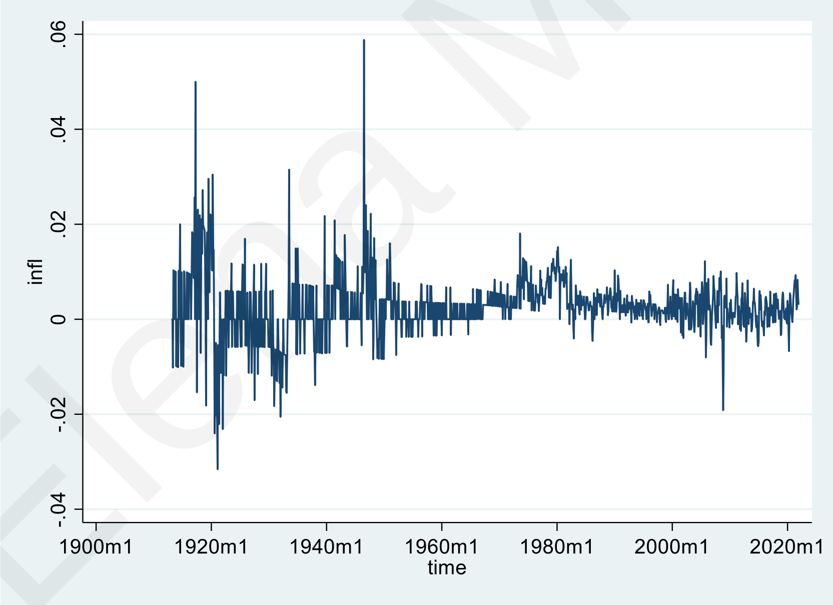
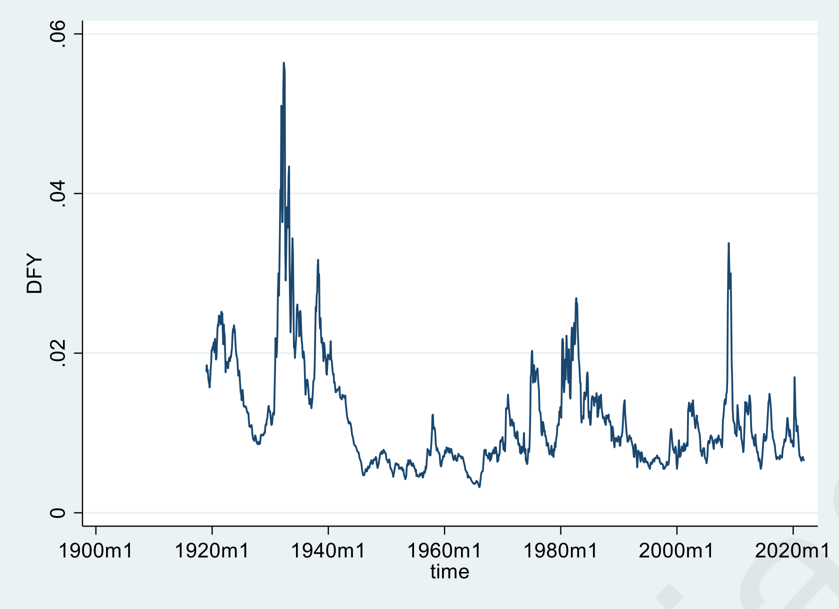


Diagram 5: Inflation (infl)



Observation for diagrams 4 & 5: In diagram 4, we observe that the variable LTR is stationary because the mean and the variance are constant over time. The same happens in diagram 5 with the inflation; stability in mean and variance over time is observed.

Diagram 6: Default Yield Spread (DFY)



Observations: In diagram 6, the series is characterized as non-stationary because the autocorrelations do not decrease geometrically. The mean and variance are not constant, and there is a sharp increase and, immediately after that, a decrease at the same rate.

Diagram 7: Term Spread (tms)

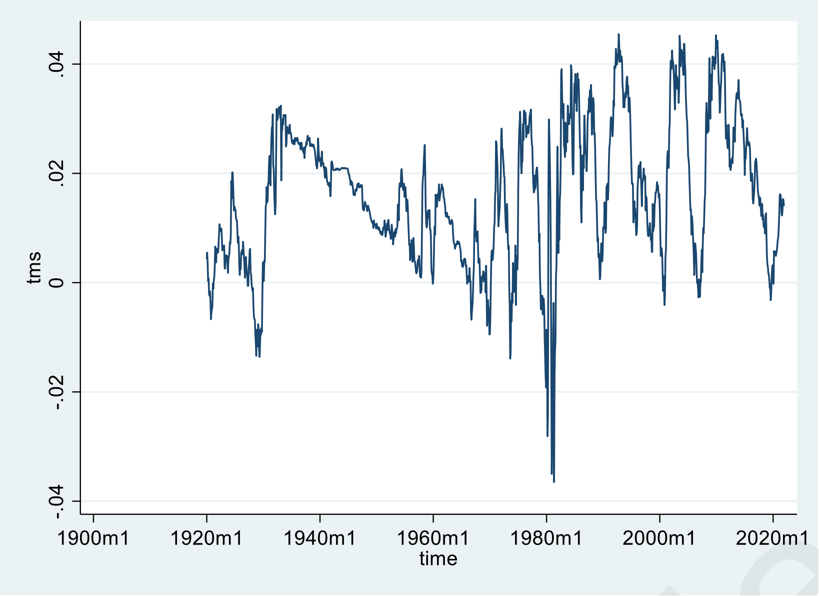
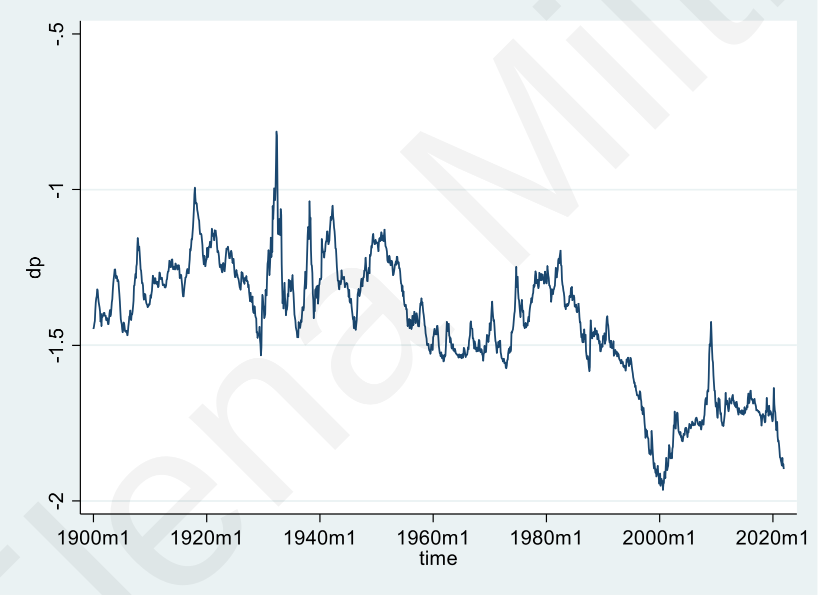


Diagram 8: Dividend Price Ratio (dp)



Observations: In the diagrams 7 & 8 which represent the variables TMS and DP, it is noticed that they do not have a fixed mean and variation; there are fluctuations, and therefore they are characterized as non-stationary.

4.3 Description of the Econometric Method and Model

4.3.1 Description of the Econometric Method

To examine the predictability of stock returns in the United States of America, it is necessary to use a model. Those models are the Bivariate Linear Regression Model. The method that will be used to analyze the model is the bivariate model. The econometric software that will be used to estimate this impact is Stata 17.

4.3.2 Model Description

Econometric Model

The following econometric model will be used in order to estimate the predictability of stock returns in the United States of America. Below will be presented the estimation of the econometric model in bivariate models setting that includes the nine (9) variables dp, dy, tms, infl, svar, dfy, bm, and ltr. The models are defined by the below equations:

Equation 1: Bivariate Models

$$ER_t = \beta 0 + \sum_{i=1}^{P} \beta 1 \; DP_{t-i} + \varepsilon$$
$$ER_t = \beta 0 + \sum_{i=1}^{P} \beta 2 \; DY_{t-i} + \varepsilon$$
$$ER_t = \beta 0 + \sum_{i=1}^{P} \beta 3 \; TMS_{t-i} + \varepsilon$$
$$ER_t = \beta 0 + \sum_{i=1}^{P} \beta 4 \; INFL_{t-i} + \varepsilon$$
$$ER_t = \beta 0 + \sum_{i=1}^{P} \beta 5 \; SVAR_{t-i} + \varepsilon$$
$$ER_t = \beta 0 + \sum_{i=1}^{P} \beta 6 \; TBL_{t-i} + \varepsilon$$
$$ER_t = \beta 0 + \sum_{i=1}^{P} \beta 7 \; BM_{t-i} + \varepsilon$$
$$ER_t = \beta 0 + \sum_{i=1}^{P} \beta 8 \; LTR_{t-i} + \varepsilon$$
$$ER_t = \beta 0 + \sum_{i=1}^{P} \beta 9 \; DFY_{t-i} + \varepsilon$$

Unit root

In a time series, the data must satisfy some assumptions. One important assumption is stationarity. More specifically, when a time series is stationary, it means that the mechanism does not vary significantly over time. In contrast, non-stationarity is a serious problem in time series data and, in this case, in stock predictability since some variables present some trend over time, that is, they change over a time period. As a result, the stock prices cannot be forecast since the results are spurious. (IORDANOVA et al., 2022).

Heteroskedasticity

Heteroskedasticity is another problem that may be present in the data. In such a case, the variance of the error term changes over time. More specifically, the heteroskedasticity in the model causes some problems in the econometric results. Specifically, when heteroskedasticity occurs in the model, then the OLS standard error is not valid. Furthermore, it makes the t-statistic as well as the F-statistic invalid for either accepting or rejecting the null hypothesis. Moreover, heteroskedasticity causes model misspecification. Finally, the estimated coefficients are not efficient. As a result, the variance of the stock is not constant, and the stock predictability is not valid (CFI Team, 2023a).

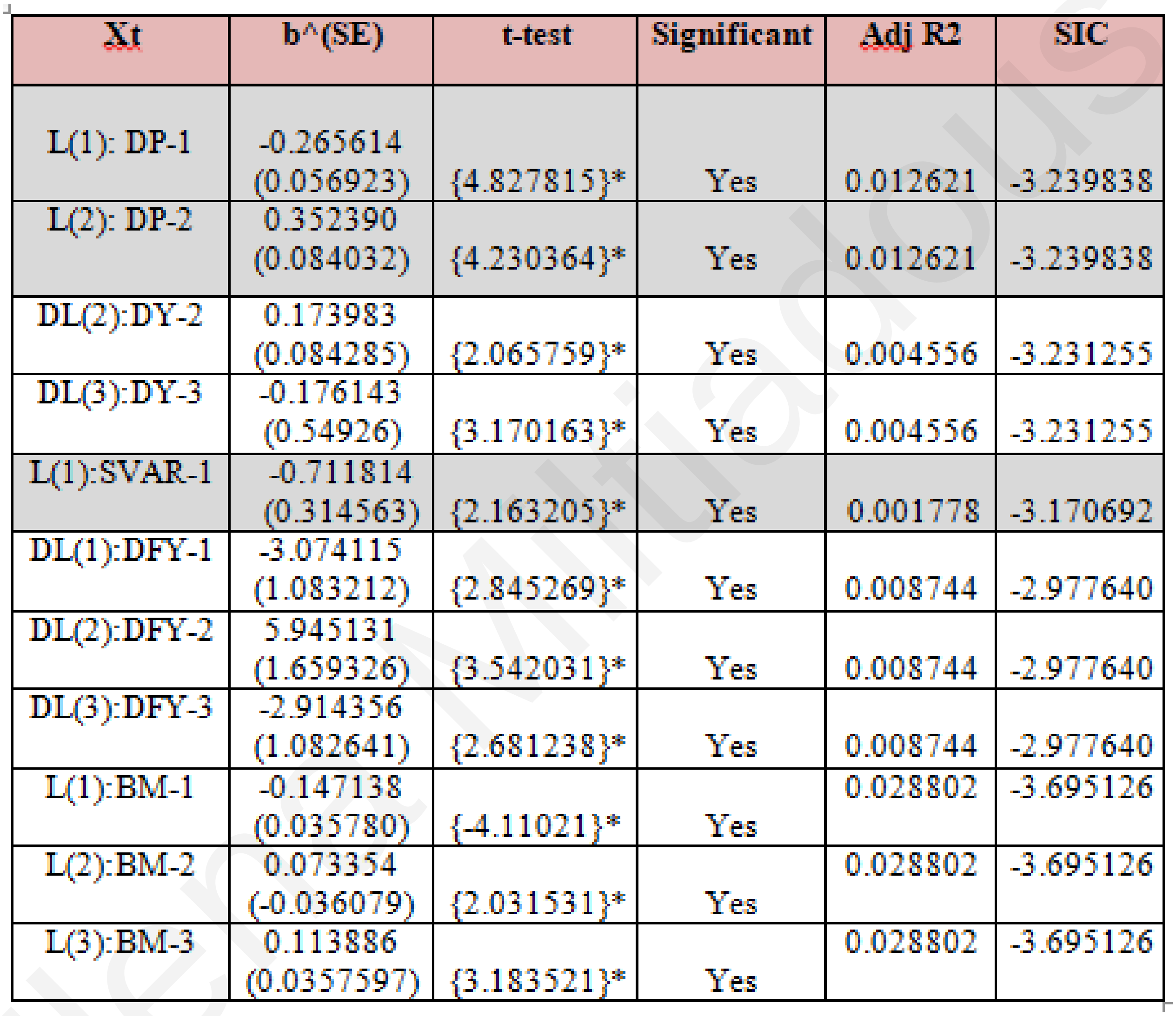
Parameter instability and Structural break

Parameter instability and structural breaks are problems that occur in time series data. Specifically, they show whether a variable changes unexpectedly over time. Since it can lead to huge forecasting errors and unreliability, That is, stock predictability is not valid. This change could involve either a change in the mean or a change in another parameter in the model (StataCorp LLC, 2023).

5. Results

5.1 Table 1: Significant Predictors

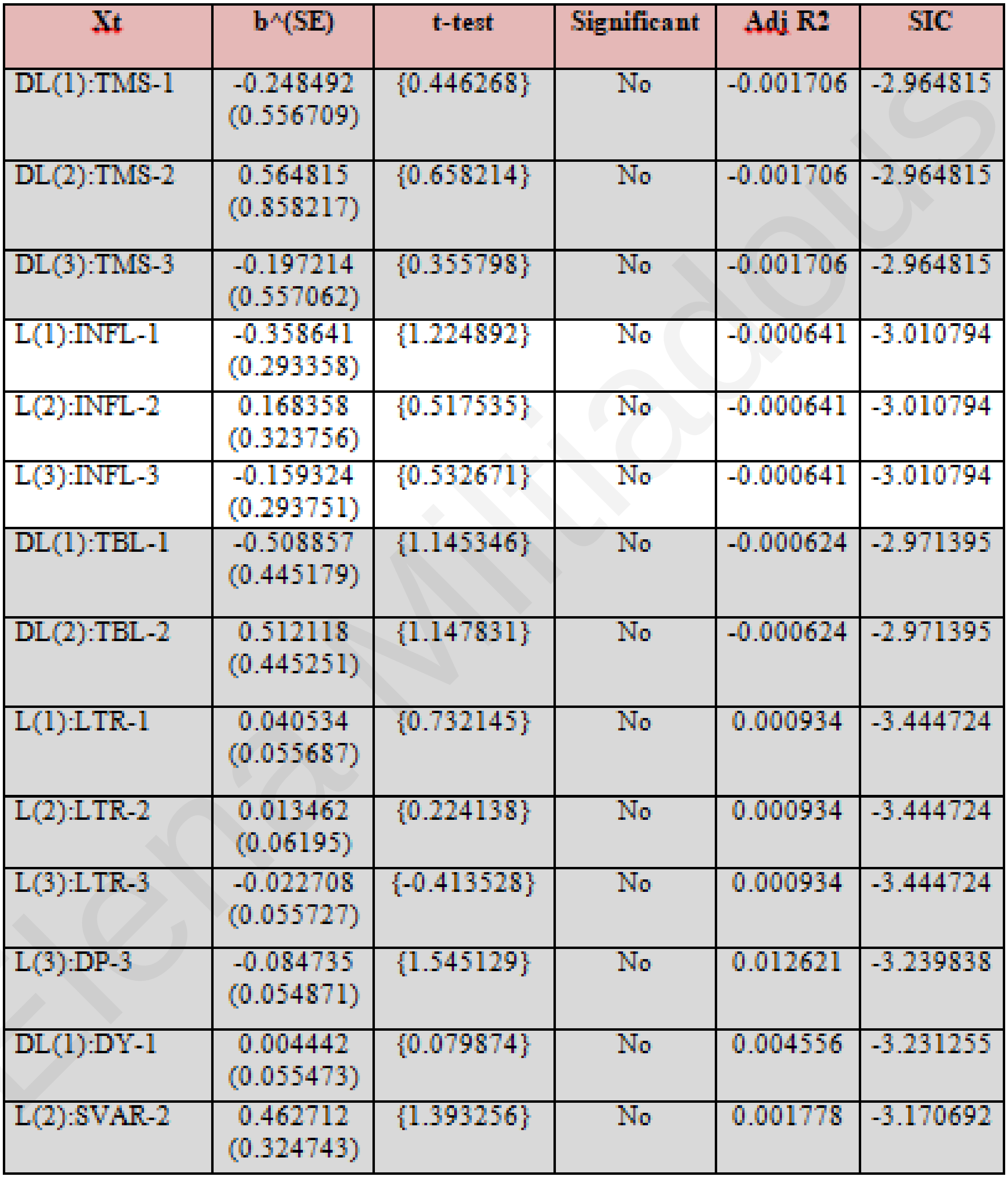
Table 3: Significant predictors



Xt	b^(SE)	t-test	Significant	Adj R2	SIC
L(1): DP-1	-0.265614 (0.056923)	{4.827815}*	Yes	0.012621	-3.239838
L(2): DP-2	0.352390 (0.084032)	{4.230364}*	Yes	0.012621	-3.239838
DL(2):DY-2	0.173983 (0.084285)	{2.065759}*	Yes	0.004556	-3.231255
DL(3):DY-3	-0.176143 (0.54926)	{3.170163}*	Yes	0.004556	-3.231255
L(1):SVAR-1	-0.711814 (0.314563)	{2.163205}*	Yes	0.001778	-3.170692
DL(1):DFY-1	-3.074115 (1.083212)	{2.845269}*	Yes	0.008744	-2.977640
DL(2):DFY-2	5.945131 (1.659326)	{3.542031}*	Yes	0.008744	-2.977640
DL(3):DFY-3	-2.914356 (1.082641)	{2.681238}*	Yes	0.008744	-2.977640
L(1):BM-1	-0.147138 (0.035780)	{-4.11021}*	Yes	0.028802	-3.695126
L(2):BM-2	0.073354 (-0.036079)	{2.031531}*	Yes	0.028802	-3.695126
L(3):BM-3	0.113886 (0.0357597)	{3.183521}*	Yes	0.028802	-3.695126

5.2 Table 2: Insignificant Predictors

Table 4: Insignificant Predictors



Xt	b^(SE)	t-test	Significant	Adj R2	SIC
DL(1):TMS-1	-0.248492 (0.556709)	{0.446268}	No	-0.001706	-2.964815
DL(2):TMS-2	0.564815 (0.858217)	{0.658214}	No	-0.001706	-2.964815
DL(3):TMS-3	-0.197214 (0.557062)	{0.355798}	No	-0.001706	-2.964815
L(1):INFL-1	-0.358641 (0.293358)	{1.224892}	No	-0.000641	-3.010794
L(2):INFL-2	0.168358 (0.323756)	{0.517535}	No	-0.000641	-3.010794
L(3):INFL-3	-0.159324 (0.293751)	{0.532671}	No	-0.000641	-3.010794
DL(1):TBL-1	-0.508857 (0.445179)	{1.145346}	No	-0.000624	-2.971395
DL(2):TBL-2	0.512118 (0.445251)	{1.147831}	No	-0.000624	-2.971395
L(1):LTR-1	0.040534 (0.055687)	{0.732145}	No	0.000934	-3.444724
L(2):LTR-2	0.013462 (0.06195)	{0.224138}	No	0.000934	-3.444724
L(3):LTR-3	-0.022708 (0.055727)	{-0.413528}	No	0.000934	-3.444724
L(3):DP-3	-0.084735 (0.054871)	{1.545129}	No	0.012621	-3.239838
DL(1):DY-1	0.004442 (0.055473)	{0.079874}	No	0.004556	-3.231255
L(2):SVAR-2	0.462712 (0.324743)	{1.393256}	No	0.001778	-3.170692

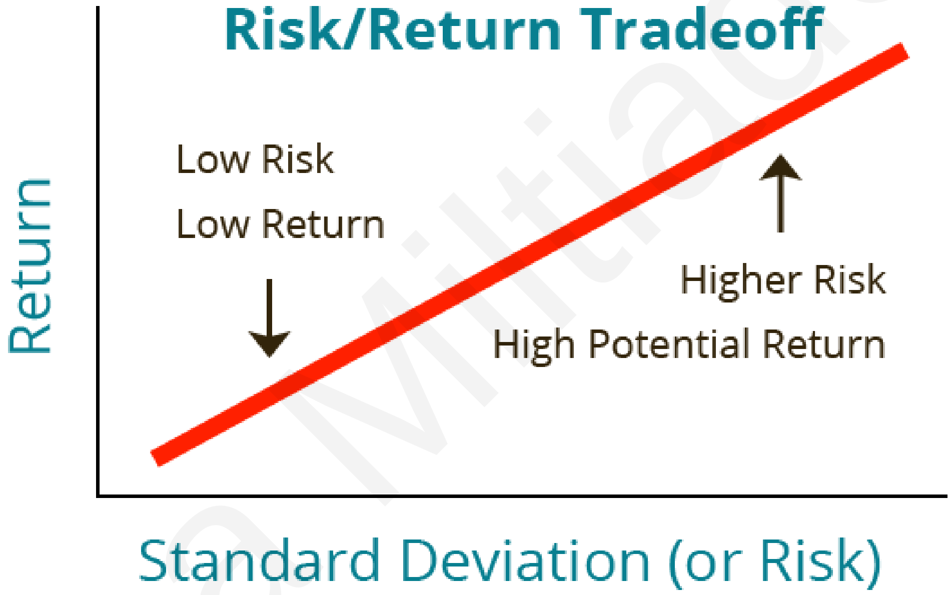
The main purpose of this research is to examine the bivariate models of the production of US stock returns for the period January 1900–December 2021 and identify the variables that affect the stock returns the most. The sample consists of monthly stock returns of the index S&P 500, and the connection of accounting parameters with stock returns and systematic risk is investigated. This association could probably be used by investors to make their portfolios more profitable. That is the reason for the use of parametric methods to find the quarterly data, which probably connects with the return and the danger of stocks.

The following results are based on the empirical data. Table 3 presents the statistics of the variables that are statistically significant at the 10% significance level and are denoted by [*]. Moreover, Table 4 shows the variables that are not statistically significant at the same level of significance.

First of all, the variable Dividend Price Ratio is statistically significant, noting that returns react up to 2 months to the change in the risk of stocks. Dividends are a percentage of annual profits that will be paid to shareholders as income, so that is why returns react directly to changes in a stock's risk. The dividend yield variable reacts in the same way. It is related to the annual profits of the returns, which imply that as the profits of the company increase, the dividend of the owners of a share also increases. Dividend yield is statistically significant, and it reacts to the time lag. Going 1 period back, it is not statistically significant, but at 3 periods back, it starts to react. This was expected because the higher the dividend yield, the more attractive a stock is considered, all other factors being equal. This mainly interests an investor with a long-term perspective who has invested his funds with the aim of collecting a satisfactory income, which he hopes will increase over time with a gradual increase in the stock.

Between stock variance and excess stock returns, there is a negative relationship, as shown in the chart below, but it is noticeable that the returns react positively in 2 months with the change in risk. It is also observed that the variable is statistically significant at the 10% level of significance, in contrast with a macroeconomic variable such as inflation, whose returns do not react at 1, 2, or 3 months. When inflation increases, it will not affect either negatively or positively the excess returns. The same applies to the term spread variable, which is not statistically significant. With a three-month time lag, the results of the t-statistic do not change, so neither does its reaction to the change in risk. Term spreads affect bonds to a greater degree than stocks. Furthermore, the Treasury Bills variable is not statistically significant. It does not respond to a risk change after a two-month lag. Default Yield Spread is a term used to describe a spread in the yield on a bond. It is statistically significant at the 10% significance level and does not react negatively to the quarterly lag. Therefore, this means that the price index is affected by the value of the default yield spread.

Figure 1: Risk and Return Tradeoff



6. Homoscedasticity Test

In statistics, a sequence or vector of random variables is homoscedastic if all the random variables have the same finite variance. The absence of this property is called heteroscedasticity. The existence of homoscedasticity implies stability of variance, i.e., the variance of the dependent variable Excess Returns is equal for different levels of each independent variable.

The non-stationarity of variance implies that risk is not constant over time. Investors, however, are becoming more and more demanding of the quality of information regarding the characteristics concerning the evaluation of their investment options. The risk is the most important of these. The key dimensions for proper evaluation are returns and risk. Risk and danger in finance are the uncertainty that exists in the realization of the outcome. A portfolio or investment becomes more attractive when it has a high return and low risk. That is also observable in the figure above. While the return is increasing on the vertical axis, the risk is increasing on the horizontal axis. There is a positive relationship between them.

The existence of homoscedasticity or heteroskedasticity is studied through the Breusch-Pagan Godfrey test and ARCH test methods. Then the F-statistic test and the p-value are used to check for equality of variances, with the purpose of determining if all the residuals have the same variances.

6.1 Breusch-Bagan Test

Null Hypothesis- H_0: σ^2

Alternative Hypothesis- H_1: σt^2 (xt^2)

Table 5: Breusch – Bagan Test Results

	F-Statistic	P-value
Dividend Price Ratio (dp)	8,19	0,00
Dividend Yield (dy)	8,93	0,00
Stock Variance (svar)	68,03	0,00
Default Yield Spread (dfy)	69,29	0,00
Book to Market (bm)	11,19	0,00

Table 5 above shows the homoscedasticity test for the four statistically significant variables. This test is performed to determine if there is a constant variance.

When the p-value is less than the 5% significance level, then the null hypothesis H_0 is rejected and heteroscedasticity exists.

According to the results shown in table 5, the p-value value is less than 5% for all statistically significant variables, so there is heteroscedasticity and non-constant variation.

P-value = 0,00 < 0,05 ➡ Reject H_0 ➡ Heteroskedasticity

6.2 ARCH Test

Null Hypothesis- H_0: σ^2

Alternative Hypothesis- H_1: σ_t^2 (u^2_{t-1})

Table 6: ARCH Test Results

	F-Statistic	P-value
Dividend Price Ratio (dp)	92,18	0,00
Dividend Yield (dy)	83,21	0,00
Stock Variance (svar)	79,68	0,00
Default Yield Spread (dfy)	60,85	0,00
Book to Market (bm)	75,34	0,00

The aim of the ARCH test is to examine whether there is dynamic homoscedasticity. According to the results shown in table 6, the p-value value is less than 5% for all statistically significant variables; therefore, the null hypothesis is rejected, which implies there is a strong heteroscedasticity. Both tests concluded with the same result, which means that there is strong heteroscedasticity and non-stability of variances.

P-value = 0,00 < 0,05 → Reject H_0 → Dynamic heteroskedasticity

7. Cumulative sum test for parameter stability

This section examines the stability of the parameter of the regression. More specific, it tests whether the coefficients in time series are stable over time. This test is constructed from the cumulative sum of Ordinary Least Squares residuals. That is, it tests for structural breaks in the residuals.

Diagram- Dividend Price Ratio

Table 7: Cumulative sum test for stability of the Dividend Price Ratio

Cumulative sum test for parameter stability				
Variable	t - statistic	1% Critical Value	5% Critical Value	10% Critical Value
Dividend Price Ratio	13.7636	1.6276	1.3581	1.224

Diagram 9: OLS cusum plot of Dividend Price Ratio

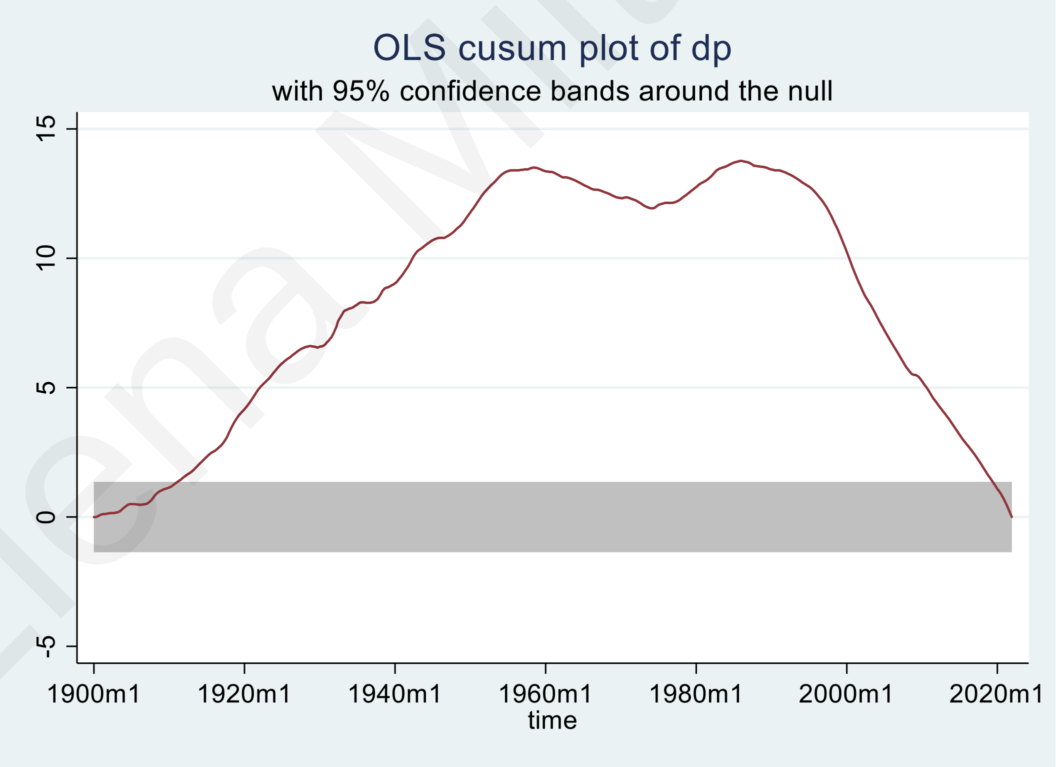
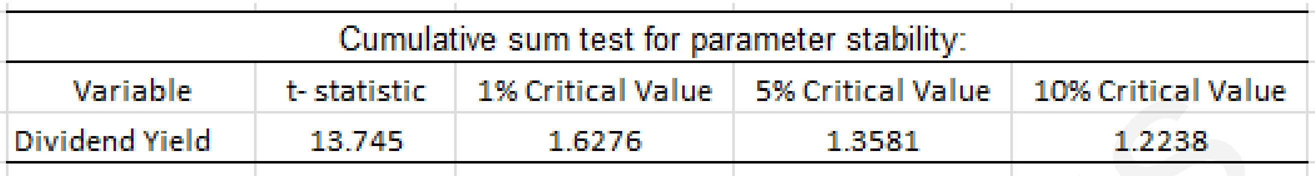


Diagram- Dividend Yield

Table 8: Cumulative sum test for stability of the Dividend Yield



Cumulative sum test for parameter stability:				
Variable	t- statistic	1% Critical Value	5% Critical Value	10% Critical Value
Dividend Yield	13.745	1.6276	1.3581	1.2238

Diagram 10: OLS cusum plot of the Dividend Yield

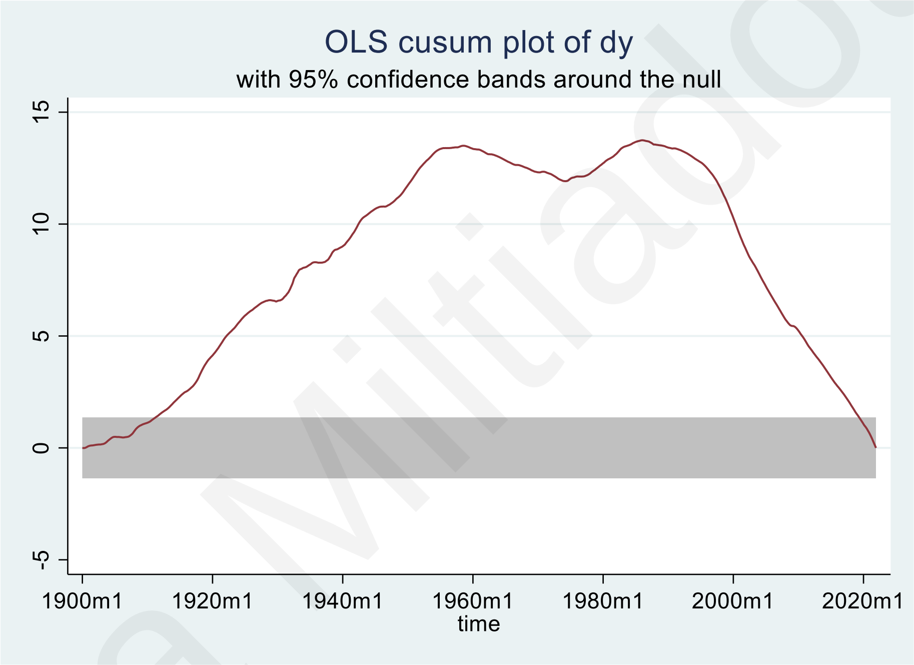


Diagram- Book to Market Ratio

Table 9: Cumulative sum test for stability of the Book to Market Ratio

Cumulative sum test for parameter stability				
Variable	t - statistic	1% Critical Value	5% Critical Value	10% Critical Value
Book to Market Ratio	11.8193	1.6276	1.3581	1.224

Diagram 11: OLS cusum plot of the Book to Market Ratio

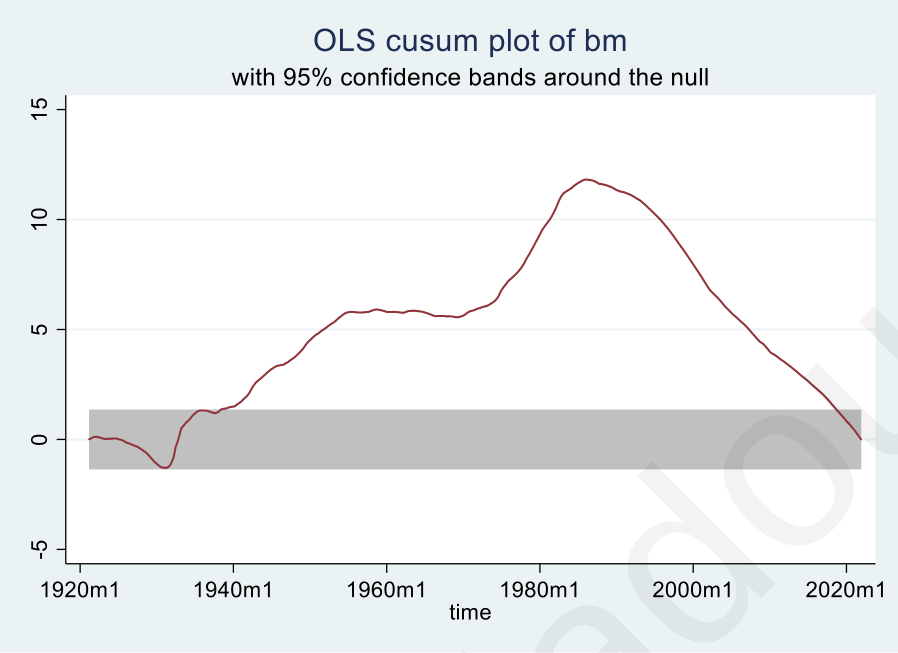


Diagram- Stock Variance

Table 10: Cumulative sum test for stability of the Stock Variance

Cumulative sum test for parameter stability				
Variable	t - statistic	1% Critical Value	5% Critical Value	10% Critical Value
Stock Variance	2.8900	1.6276	1.3581	1.224

Diagram 12: OLS cusum plot of Stock Variance

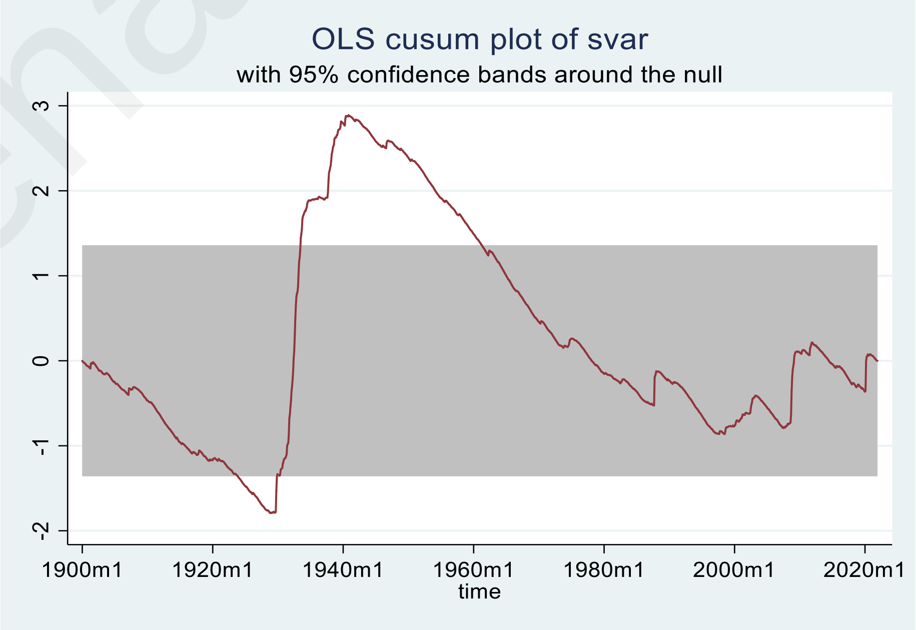
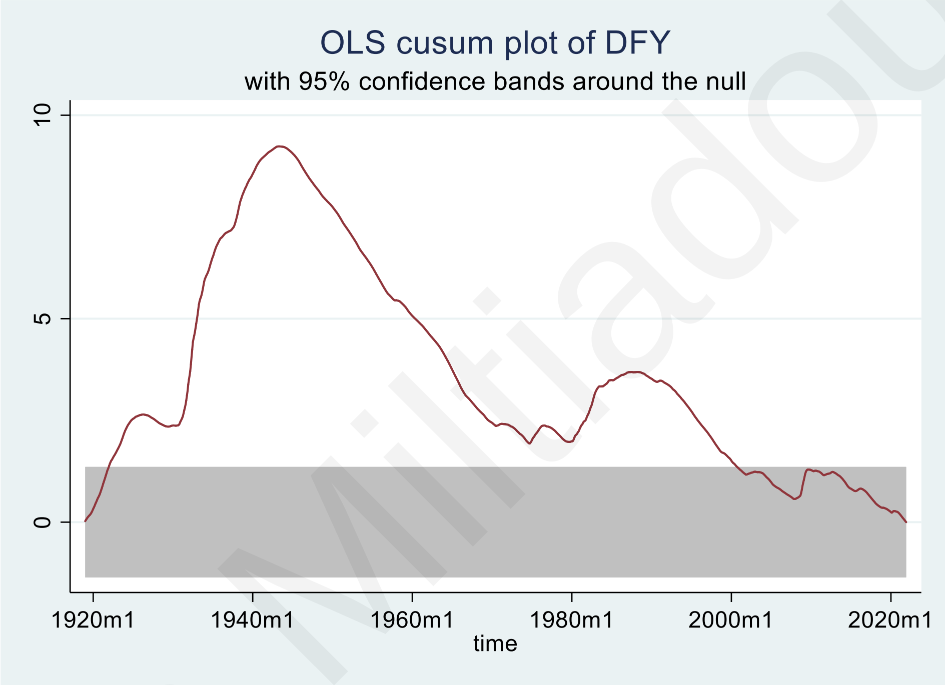


Diagram- Default Yield Spread

Table 11: Cumulative sum test for stability of Default Yield Spread

Cumulative sum test for parameter stability				
Variable	t - statistic	1% Critical Value	5% Critical Value	10% Critical Value
Default Yield Spread	9.2344	1.6276	1.3581	1.224

Diagram 13: OLS cusum plot of Default Yield Spread



Tables 7, 8, 9,10 and 11 show the OLS cumulative sum of OLS residuals. The analysis spans the years 1900 to 2021, beginning on January 1, 1900, and ending on December 31, 2021. The null hypothesis of this test is that the estimated coefficient is stable over time. However, the rejection of the null implies that there is no parameter stability. In addition, the significance level is 5%. According to the results of the aforementioned tables, the null hypothesis is rejected since the t-statistic is greater than the critical value of 1%, 5%, and 10% significance levels in all cases. As a result, the null hypothesis is rejected at all significance levels. That means the parameters in the regression model are unstable for this specific time period.

The above diagrams show graphically the cumulative sum of OLS residuals for dividend price ratio, dividend yield, book-to-market ratio, stock variance, and default yield spread with 95%

confidence. According to the diagrams, the null hypothesis is rejected at the 5% significance level since the cusum statistic goes outside the bound. Moreover, diagrams confirm the results exacted from the tables. As a result, the parameters of the regression model became unstable during that time.

To sum up, from the above diagrams, it is evident that the parameter stability is not constant over time, so the OLS residuals are not homoscedastic. Some of the change in variance, i.e., the volatility of the risk, is due to a financial crisis. Furthermore, all the above variables are outside of the bound, and either an upward or downward trend has developed. This implies that the process has changed, while the process may be affected by special causes.

A change in variance implies a change in risk. This causes uncertainty for any type of investment. Individuals' behavior differs radically depending on whether they face risk or uncertainty. Although the words "risk" and "uncertainty" are sometimes used for the same purpose, technically they are different. The term risk means that the one who makes a decision knows the possible consequences of his decision and the related probabilities at the time he makes the decision. On the other hand, uncertainty is a situation in which the probability of possible consequences is unknown. One of the main problems faced by investors is making rational decisions in the face of uncertainty, incomplete information, and risk.

The factors that create fluctuations in the returns and prices of shares are elements of risk. Some external factors, such as changes in interest rates, the level of inflation, the level of unemployment, and exchange rates, cannot be controlled by the company and affect a large number of shares. Some other internal factors, such as the company's bargaining power with its suppliers, competition within the same industry, and research and development of new products and services, can be largely controlled by the company.

8. Conclusions

My thesis is based on stock predictability. It scrutinizes the present value of an investment and its expected future value. These two values are very critical in stock analysis. Despite the obvious risk and danger associated with stock trading, every investor hopes to receive high returns on their investment. Stock prices are used as the primary basis for investment decisions. Stock prices are used as the primary basis for investment decisions. It is important to be able to predict stock returns because this helps protect the economy by preventing unnecessarily large shocks.

Stock return forecasting has been extensively conducted over the years by various authors using different methods. The main article on which my study is based is Amit Goyal & Ivo Welch(2008). The authors base their arguments on the fact that most past models are unstable and cannot be used effectively in determining stock performance. This article studies stock returns and forecasting over the time period 1872–2005. Constant reference is made to data from the S&P 500. Some of the variables that have been extensively used by the authors include volatility, B/M (book-to-market) ratio, different levels, income ratio, and wealth. Through the study of Goyal & Welch, the variables d/p, d/y, e/p, d/e, dfy, infl, and svar are not statistically significant, and the variables b/m, i/k, ntis, and eqis are statistically significant.

My study is based on the period 1900–2021, analyzing the variables dp, dy, svar, dfy, tms, bm, infl, ltr and tbl. We examined bivariate models of US stock returns and identified the variables that best corresponded to the factors affecting stock returns. Our sample consists of monthly S&P 500 stock returns, and we investigate the association of the parameters with stock returns and stock risk. The investors use this association to build their portfolios. Time lags are used in order to check in how many months each variable reacts to the change in risk. The significance level we use in our test was 10%. The variables that are statistically significant at 10% are the dividend price ratio (dp), dividend yield (dy), stock variance (svar), book-to-market ratio (bm) and default yield spread (dfy).

Therefore, we conclude that DP (dividend price ratio) and DY (dividend yield) react directly to the change in risk. This was expected because both variables are directly related to the annual profits that will be paid to shareholders as income; therefore, they affect stock returns and react to the change in risk. The higher the dividend yield, the more attractive a stock is considered. All investors place their funds with the goal of collecting a satisfactory income, which they hope will increase over time as the stock rises. This is also confirmed by our assessment. We observe that increasing DP will decrease ERT returns by -0.26 in the first month, but then a positive relationship is observed with an increase in ERT of 0.35. In this way, the relationship of each independent variable with the dependent variable is determined. On the other hand, the variables infl, tbl, and tms do not affect stock returns since they are non-statistically significant, so they are not useful for stock predictability. From the estimation and the test for constant variance, we noticed that a homoscedasticity problem arose, which we can solve with heteroscedastic standard errors.

In conclusion, we come to the same conclusion as the authors, Goyal and Welch, in their own study: that there is a problem of misspecification and instability in the models. This needs to be corrected to get accurate and valid results for stock predictability. A financial crisis can cause uncertainty and greater volatility. Therefore, financial relationships may not be valid and may be unstable. For us, this is a challenge to find other relationships and model them.

References

CFI Team. (2023a). *Heteroskedasticity*. Corporate Financial Institute. https://corporatefinanceinstitute.com/resources/data-science/heteroskedasticity/

CFI Team. (2023b). *Random Walk Theory*. Corporate Financial Institute. https://corporatefinanceinstitute.com/resources/capital-markets/what-is-the-random-walk-theory/

DOWNEY, L., SCOTT, G., & VELASQUEZ, V. (2023). *Efficient Market Hypothesis (EMH): Definition and Critique*. Investopedia. https://www.investopedia.com/terms/e/efficientmarkethypothesis.asp

Hong, H., Bian, Z., & Lee, C. (2021). *COVID-19 and instability of stock market performance: evidence from the U.S.* https://jfin-swufe.springeropen.com/articles/10.1186/s40854-021-00229-1

IORDANOVA, T., KHARTIT, K., & LI, T. (2022). *An Introduction to Non-Stationary Processes*. Investopedia. https://www.investopedia.com/articles/trading/07/stationary.asp

Marcellino, M., Stock, J. H., & Watson, M. W. (2015). *A Comparison of Direct and Iterated Multistep AR Methods for Forecasting Macroeconomic Time Series*. https://www.princeton.edu/~mwatson/papers/hstep_3.pdf

MAVERICK, J. B., POTTERS, C., & JASPERSON, H. D. (2023). *The Weak, Strong, and Semi-Strong Efficient Market Hypotheses*. Investopedia. https://www.investopedia.com/ask/answers/032615/what-are-differences-between-weak-strong-and-semistrong-versions-efficient-market-hypothesis.asp

PESARAN, M. H., & TIMMERMANN, A. (1995). Predictability of Stock Returns: Robustness and Economic Significance. *THE JOURNAL OF FINANCE, L*(4), 1–28. https://www.jstor.org/stable/2329349

Qi, M. (2012). Nonlinear Predictability of Stock Returns Using Financial and Economic Variables. *Journal of Business & Economic Statistics*, *17*(4), 419–429. http://dx.doi.org/10.1080/07350015.1999.10524830

StataCorp LLC. (2023). *Tests for structural breaks in time-series data*. StataCorp LLC. https://www.stata.com/features/overview/structural-breaks/#:~:text=What’s this about%3F,process that produce the series.

Welch, I., & Goyal, A. (2004). *A Comprehensive Look at The Empirical Performance of Equity Premium Prediction* (No. 04; 11). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=517667