TU Dortmund

Case Studies

Project I: Forecasting The Equity Premium Using Ordinary Least Squares

Lecturers:
Prof. Dr. Matei Demetrescu
Dr. Paul Navas

Author: Bushra Tariq Kiyani

Group number: 4

Group members: Sarmistha Bhattacharyya, Oliver Fischer

November 17, 2023

Contents

1 Introduction

In the highly dynamic and constantly evolving landscape of financial markets, the ability to predict stock market returns remains a significant challenge for investors, financial analysts, and portfolio managers. Accurately forecasting stock returns holds immense potential. It not only promises to refine investment strategies but also to enhance risk management techniques, ultimately leading to increased overall earnings from investments. Financial markets are characterized by inherent uncertainty and volatility. Precise forecasts have the potential to improve asset allocation decisions, strengthen risk management strategies, and result in more profitable investment portfolios. Understanding the dynamics of stock returns, the influence of various economic indicators, and the effectiveness of forecasting methods can empower investors to navigate financial markets confidently. Accurate and dependable forecasts of excess stock returns stand as the cornerstone of prudent investment decisions (Julio et al., 2022).

This project aims to produce forecasts of excess stock returns using a range of data-driven approaches. The excess return of a stock, defined as the difference between the stock's return and the risk-free rate, is a vital measure in financial analysis. Our task consists of forecasting the excess return using lagged predictors and taking into account different time frequencies. The quality of these forecasts are rigorously evaluated through the Root Mean Squared Forecast Error (RMSFE). This project utilizes a comprehensive dataset that includes monthly, quarterly, and yearly data, encompassing a numerous financial and economic indicators. The analysis is conducted on a diverse set of predictors, including stock market indices, interest rates, corporate bond yields, and various macroeconomic variables. It involves time series analysis, regression modeling, and the application of statistical and econometric techniques. Models are constructed using Ordinary Least Squares (OLS). The choice of the appropriate lag order for the autoregressive model is informed by ACF and PACF.

The second section describes the structure and quality of the dataset in more detail. Additionally, the goals of the project are stated in the second section. The third section explains the Linear regression, Autoregressive (AR) model, Backward Selection, Forward Selection, The Akaike Information Criterion (AIC), Autocorrelation Function (ACF) plots and partial autocorrelation function (PACF) plots. The fourth section focuses on the application of these methods and the interpretation of the plots and results. Finally, the fifth section summarizes the most important findings

2 Problem statement

2.1 Dataset and Data Quality

This project uses the data collected by the authors of the paper "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction" (Welch and Goyal, 2007). Updated data (up to 2022) is downloaded from the personal webpage for Amit Goyal, a researcher in finance and economics (Goyal, 2022). Data were collected through different websites, detailed description of data collection is mentioned in the paper.

The dataset spans from the year 1871 to 2022 and focuses on the equity premium as the dependent variable, which is computed. It is segmented into three time intervals: monthly, quarterly, and yearly. The number of independent variables and observations varies across these time spans. Specifically, the dataset has 18 independent variables and 1824 observations for monthly data, 22 covariates and 608 observations for quarterly data, and 21 covariates and 152 observations for yearly data. Table 4 in the appendix displays the description of all variables. The dataset is gathered for a scientific study, so it is expected to be high quality in terms of accuracy, consistency, relevance, and a trusted source. There are missing values in almost all variables that are simply omitted from the analysis.

2.2 Project Objectives

The goal of this project is to conduct an in-depth analysis and produce accurate forecasts of excess stock returns using Ordinary Least Squares (OLS) regression. The first step will be to calculate the excess returns by subtracting the risk-free rate from the stock returns. Next, plots of the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) for the excess return series will be generated.

Succeeding this, an Autoregressive (AR) model will be fitted using information from the ACF and PACF to determine the AR model's order (p). Subsequently, forecasts will be generated based on the full-sample autoregressive fit, and the Root Mean Squared Forecast Error (RMSFE) will be computed. The project will also consider the challenges of applying this process in real-life situations. Further, linear models with individual lagged predictors, using Ordinary Least Squares (OLS), will be fitted. The RMSFEs will be used again to compare the effectiveness of these models against the autoregressive model.

After that, a linear predictive model using all available predictors will be developed, and its RMSFE will be compared with the RMSFEs from models using one predictor at a time. Explanations for observed differences will be provided, enhancing the interpretability of the findings.

The final step involves selecting the best predictor variables for the multiple prediction regression using backward selection and AIC as information criteria. This selection process will be repeated using forward stepwise model selection, offering insights into the effectiveness of both predictor selection methods.

3 Statistical methods

3.1 Linear Regression

The linear regression model is a statistical modeling approach employed to depict a relationship between a continuous variable of interest, denoted as Y (the response variable), and a collection of explanatory variables $x_1, ..., x_k$, which can take on continuous or categorical values. This association between Y and $x_1, ..., x_k$ is expressed through a function denoted as $f(x_1, ..., x_k)$, incorporating additive errors. This relationship can be represented by the following equation:

$$Y = f(x_1, ..., x_k) + \epsilon$$

Where ϵ is a random variable, and the function f is a linear combination of covariates, given by:

$$f(x_1, ..., x_k) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

The parameters $\beta_0, \beta_1, ..., \beta_k$ are unknown and require estimation. β_0 is the intercept term. By consolidating covariates and unknown parameters into $p = k + 1$ dimensional vectors, denoted as $\mathbf{x_i} = (1, x_{i1}, ..., x_{ik})'$ and $\boldsymbol{\beta} = (\beta_0, ..., \beta_k)'$, then relationship is:

$$\hat{y}_i = \mathbf{x_i}'\boldsymbol{\beta} + \epsilon_i$$

With a total of n observations, the model can be expressed using vector notation. The vectors $\mathbf{y}$ and $\boldsymbol{\epsilon}$, along with the design matrix $\mathbf{X}$, defined as follows:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}$$

The model can be written as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

The following assumptions are made: the errors are assumed to have a mean of zero, $E[\epsilon_i] = 0$. It is assumed that the errors exhibit a constant variance σ^2 across observations, referred to as homoscedastic errors. The covariance between errors for different observations is zero. The design matrix possesses a full column rank, It is assumed that the errors follow a normal distribution. (Fahrmeir et al., 2022, p. 74-75)

3.1.1 Estimation

The estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$ are represented as $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\sigma}}^2$, respectively.

Regression parameters estimation: Estimation of regression parameters usually involves the use of **least squares**. Following this principle, the unknown regression coefficient $\boldsymbol{\beta}$ is estimated by minimizing the sum of squared deviations between the true response value y_i and the predicted value $\hat{y}_i = \mathbf{x}_i'\hat{\beta}$:

$$\text{LS}(\beta) = \sum_{i=1}^{n}(y_i - \mathbf{x}_i'\beta)^2$$

in Matrix notation: $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. The estimation of $\boldsymbol{\beta}$ involves minimizing the LS($\boldsymbol{\beta}$) by setting its first derivative to zero and solving for $\boldsymbol{\beta}$. This estimator is (Fahrmeir et al., 2022, p. 105)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Predicted values and residuals: The estimator for the mean of y_i is given by:

$$\widehat{E[y_i]} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_k x_{ik} = \mathbf{x}_i'\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Usually, $\widehat{E[y_i]}$ is referred to as $\hat{y}_i$. The difference between the true value y_i and the estimated value $\hat{y}_i$, expressed as $\hat{\epsilon}_i$, is called the residual (Fahrmeir et al., 2022, p. 107):

$$\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

Estimation of the error variance: The maximum likelihood estimator of the error variance, as described in (Fahrmeir et al., 2022, p. 108), is given by:

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}'}{n}$$

However, this estimator is biased. As an alternative, the Restricted Maximum Likelihood Estimator (REML) is commonly used for estimating σ^2, which is expressed as:

$$\hat{\sigma}^2 = \frac{1}{n-p}\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}'$$

3.2 Autoregressive (AR) model

An Autoregressive (AR) model is a type of time series model used to describe a sequence of observations where each observation is modeled as a linear combination of past observations. The general form of an AR(p) model of order p is given by:

$$Y_t = \phi_1 Y_{t-1} + ... + \phi_k Y_{t-p} + \epsilon_t$$

where Y_t is the value of time series at time t. The parameters $\phi_1, ..., \phi_t$ are the autoregressive coefficients. $Y_{t-1}...Y_{t-p}$ are the past observations of the time series, which determine the strength and sign of the relationship between the current observation and its past values. They are estimated from the data during the model fitting process. ϵ_t is the white noise error term at time t, representing unobserved factors influencing the series. It is assumed to have a mean of zero and constant variance.

The order of the AR model, denoted as p, specifies how many past observations are included in the model. For example, an AR(1) model includes only the immediate past observation, AR(2) includes the two most recent observations, and so on. (Brockwell and Davis, 1991, p. 239)

3.3 Root Mean Squared Forecasting Error (RMSFE)

The Root Mean Squared Forecasting Error (RMSFE) assesses forecast accuracy in time series analysis, quantifying how well a forecasting model aligns with actual values. Mathematically, it is expressed as:

$$\text{RMSFE}_y = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}$$

where T is the total number of observations. y_t is the actual observed value at time t and $\hat{y}_t$ is the forecasted value at time t produced by the forecasting model. A lower RMSFE signifies improved forecasting, indicating smaller average squared differences between actual and forecasted values. The square root operation enhances interpretability, offering a measure of the typical error in the original data's units. It is particularly useful when comparing forecasting performance across different models or time series. (Brockwell and Davis, 1991, p. 49)

3.4 Model Selection Methods

Backward Selection Method: Backward selection begins with the full model and removes the least important variables until a stopping criterion is reached. This criterion can be a predetermined significance level, a set number of predictors, or a chosen evaluation metric (e.g., AIC). The process continues until the stopping criterion is met, yielding a final model with the most impactful predictors. (Fahrmeir et al., 2022, p. 151)

Forward Selection Method: Forward selection begins with an intercept-only model and adds the most significant predictors one at a time until meeting a stopping criterion, like a predetermined significance level, a set number of predictors, or a chosen evaluation metric (e.g., AIC). This continues until the stopping criterion is met, yielding a final model with the most influential predictors. (Fahrmeir et al., 2022, p. 151)

3.5 The Selection Criteria

The Akaike information criterion (AIC) is commonly employed for model selection in likelihood-based inference. It quantifies the balance between the model's goodness of fit and its complexity, and is defined as:

$$\text{AIC} = -2l(\hat{\beta}_M, \hat{\sigma}^2) + 2(|M| + 1)$$

Here, M represents the number of covariates, $\hat{\beta}_M$ and $\hat{\sigma}^2$ are the maximum likelihood (ML) estimators, and $l(\hat{\beta}_M, \hat{\sigma}^2)$ represents the maximum value of the log-likelihood. A lower AIC value signifies a superior model fit (Fahrmeir et al., 2022, p. 148-149).

3.6 Autocorrelation Function (ACF) Plot

The ACF plot in time series analysis illustrates autocorrelation by showing correlation coefficients on the vertical axis and lagged time points on the horizontal axis. Lag 0 represents the correlation with itself, and subsequent lags indicate correlations with past time points. Significance bands help identify statistically significant autocorrelations, with values outside suggesting significance. Positive peaks above the band indicate positive autocorrelations, while negative dips below indicate negative autocorrelations. The decay pattern of coefficients informs about the time series model's order, with a slow decay suggesting long memory in the data. (Brockwell and Davis, 1991, p. 91-100)

3.7 Partial Autocorrelation Function (PACF) Plot

The PACF plot in time series analysis focuses on direct relationships between a variable and its past values at specific lags by removing intermediate influences. Similar to ACF plots, it uses lag values on the horizontal axis and partial autocorrelation coefficients on the vertical axis. Significance bands indicate statistically significant partial autocorrelations, with positive peaks above the band and negative dips below. The decay pattern of partial autocorrelation coefficients helps identify the order of autoregressive (AR) terms in a time series model. PACF plots are often used alongside ACF plots for comprehensive model specification. (Brockwell and Davis, 1991, p. 91-100)

4 Statistical analysis

This section presents a descriptive analysis of the dataset using statistical measures and plots described in the previous section. For calculation of all statistical measures and graphical representations R software (R Core Team, 2022) Version, 4.2.1 is used with additional packages **ggpubr** (Kassambara, 2022), **dplyr** (Wickham et al., 2022),

ggplot2 (Wickham, 2016), **GGally** (Schloerke et al., 2021), **tidyr** (Wickham et al., 2023b), **readr** (Wickham et al., 2023a), **stats** (R Core Team, 2023), **forecast** (Hyndman et al., 2023).

4.1 Exploration of Time Series Aspects in Excess Returns Series: ACF and PACF Plots

Excess returns series is generated using the formula:

$$\text{excess returns} = \text{stock returns} - \text{risk-free rate}$$

where the stock returns are the growth rates of the series Index, computed as:

$$\text{stock returns} = \frac{\text{Index} - \text{lag(Index)}}{\text{lag(Index)}}$$

To explore serial correlation of excess returns series the ACF and PACF plots are produced. Figure 1 shows ACF and PACF plots of yearly, quarterly and monthly excess returns. In Figure 1(a) the ACF plot of yearly excess returns demonstrates significant dips at lags 2 and 5, indicating autocorrelation. This suggests a negative correlation between current-year excess returns and those from two and five years ago. Similarly, Figure 1(b) shows the PACF plot with significant dips at lag 2 and 5. These spikes indicate a direct negative correlation between the current-year excess returns and those from two and five years ago, excluding the effects of other intermediate lags.

In Figure 1(c), the ACF plot for the quarterly data displays a significant positive spike at lag 3, indicating a positive autocorrelation. This suggests a correlation between the current quarter's data and the data from three quarters ago. Additionally, significant dips at lags 4 and 7 in the ACF plot indicate a negative correlation at these lags, implying a negative relationship between the current quarter's data and that from four and seven quarters ago. In Figure 1(d), the PACF plot reveals a strong positive correlation at lag 3, indicating a direct association between the current quarter's data and that from three quarters ago, removing the effects of other intermediate lags. Additionally, significant fluctuations at lags 4, 7, 10, and 21 suggest varying correlations. Lags 4 and 7 represent negative correlations, while the spike at lag 10 signifies a positive relationship between the current quarter's data and that from ten quarters ago, excluding intermediate lags. Lag 21 also suggests a negative correlation.

Figure 1(e) and (f) display ACF and PACF plots for monthly excess returns. ACF exhibits spikes at lags 1 and 5 and dips at 3, 14, 20, and 21, whereas PACF shows spikes at 1, 4, 5, 7, and dips at 3, 6, 13, 14, 20, 21, and 25.

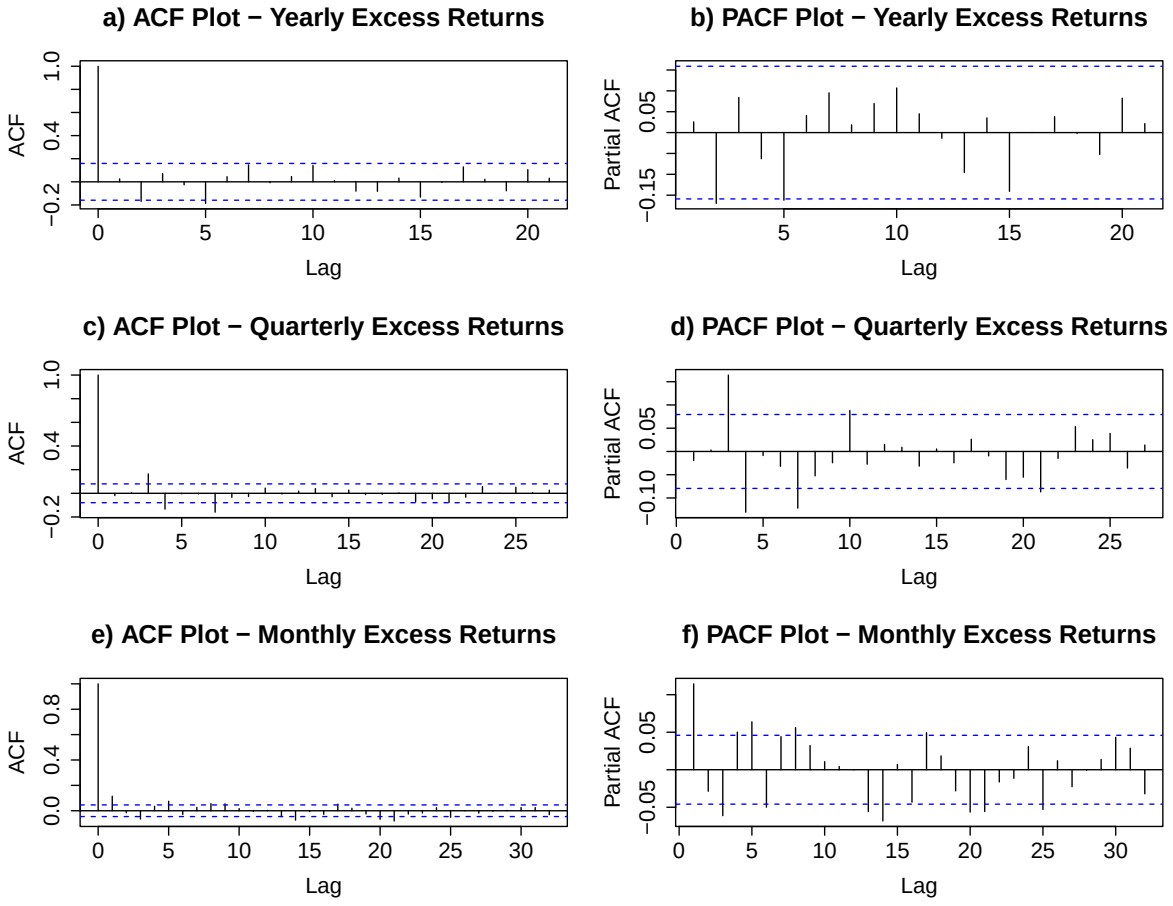

Figure 1: ACF and PACF plots

4.2 AR(p) Model Fitting and Selection Using Information Criterion

After reviewing the ACF and PACF plots, significant peaks were identified, indicating potential values for the parameter 'p' in the autoregressive (AR) model. Initially, a range of potential 'p' values was considered: 1 to 5 for yearly data, 1 to 7 for quarterly data, and 1 for monthly data, acknowledging a significant drop in the PACF plot after the first lag in monthly data. Subsequently, the best 'p' values for each dataset were determined using the AIC, prioritizing predictive accuracy. The chosen 'p' values for yearly, quarterly, and monthly data, obtained through AIC, are 2, 7, and 1, respectively. These selections align with the observations derived from the PACF and ACF plots.

Next, an AR(2) model was applied to the yearly dataset, yielding an RMSFE of 0.183 and an AIC of -77.09. For the quarterly dataset, an AR(7) model was employed, resulting in an RMSFE of 0.093 and an AIC of -1140.39. Lastly, an AR(1) model was fitted to the monthly data, producing an RMSFE of 0.047 and an AIC of -5958.87.

4.3 Forecasting Using Full-Sample Autoregressive Model and Comparison with Actual Returns

In Figure 2, the yearly forecasted values demonstrate a close alignment with the actual returns across the timeline. Notably, there appears to be greater variability in the actual values compared to the forecasted values.

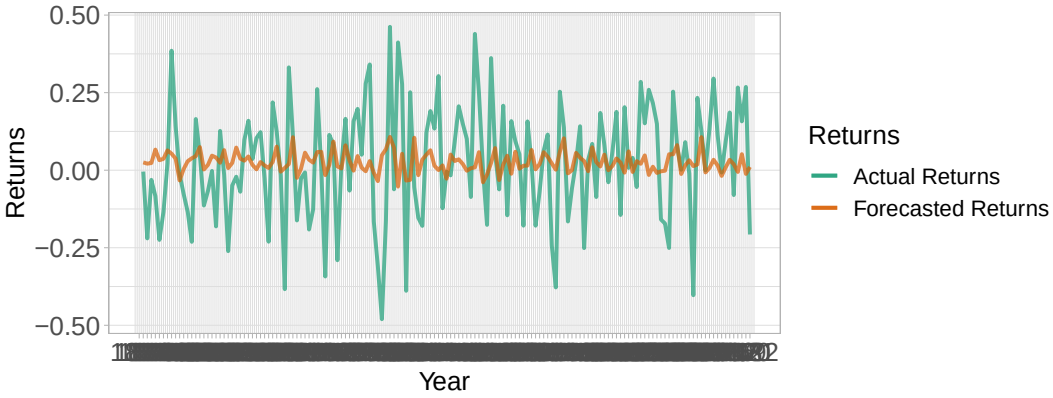

Figure 2: Actual vs Forecasted Excess Returns (Yearly Data)

In Figure 3 and Figure 4, the quarterly and monthly forecasted values exhibit a similar pattern to the actual values, although the actual data shows more variability than the forecasts. An unexpected spike observed in the actual data among the forecasting plots suggests a sudden and unanticipated surge in the observed values, which may not have been accurately predicted by the model. In real-life forecasting scenarios, utilizing the entire dataset for fitting the AR model and then using the same data for out-of-sample forecasting might not accurately replicate real-world conditions. Employing a rolling or expanding window approach is more reflective of real-life forecasting practices.

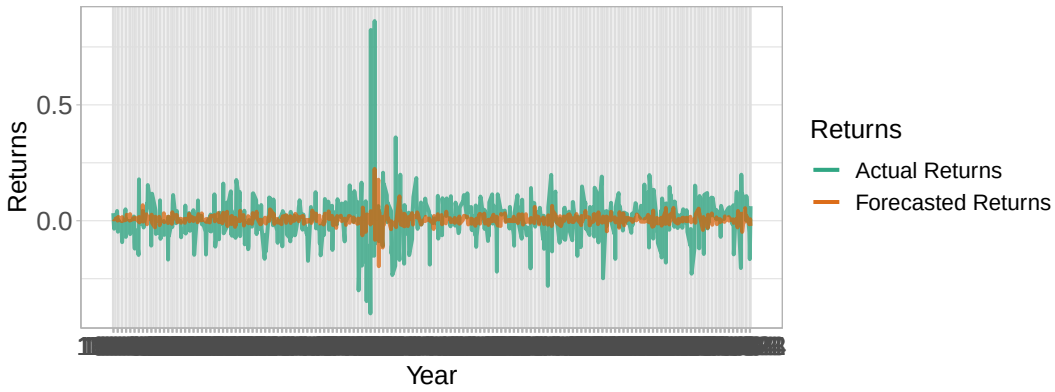

Figure 3: Actual vs Forecasted Excess Returns (Quarterly Data)

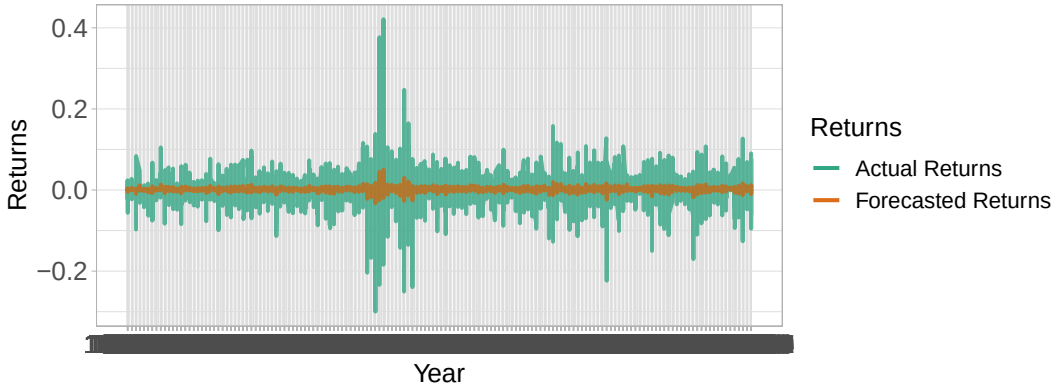

Figure 4: Actual vs Forecasted Excess Returns (Monthly Data)

4.4 Individual Predictor Analysis: Estimation using OLS and Comparison with Autoregressive Model's RMSFEs

Linear predictive models for each predictor, excluding Date, Stock Returns Including Dividends, Stock Returns Excluding Dividends, Index, and Risk-free-rate (as they are part of the dependent variable), are created. Models are estimated via OLS. The predictors are lagged to capture their temporal connection with the target variable.

Table 5 in appendix presents an overview of the RMSFEs of all yearly, quarterly, and monthly data individual predictive models. For yearly data, some predictors exceed the AR(2) model's RMSFE of 0.183, such as Long-term yield and Long-term Corporate Bond Returns (0.192), while others, e.g., Investment to Capital Ratio (0.162) and

Consumption-wealth-income ratio (0.168), perform relatively better in predicting yearly excess returns.

For quarterly data, AR(7) exhibits an RMSFE of 0.093, while certain predictors like Book-to-Market-ratio and Treasury Bills exhibit 0.107, which is higher than AR(7). Conversely, other predictors like Consumption-Wealth-Income-ratio, Net Equity Expansion, and Quarterly Dividends show lower RMSFEs (0.075 - 0.079), suggesting better predictive power. For monthly data, some predictors show higher RMSFEs (0.048 - 0.057) than the AR(1) model's 0.047, while Cross-Sectional beta Premium has the lowest RMSFE (0.046), hinting at better predictive performance.

4.5 Full Predictive Model and Comparative Forecast Analysis with Individual Predictors

A complete predictive model incorporating all predictors is developed, and its forecasting errors (RMSFEs) are compared with those from models using single predictors. This comparison aims to understand differences in forecast accuracy. Table 1 displays the RMSFEs of the full linear predictive models for yearly, quarterly, and monthly data.

	Yearly	Quarterly	Monthly
RMSFE value	0.116	0.056	0.041

Table 1: RMSFEs of Full Predictive Model for Yearly, Quarterly, and Monthly Data.

The full linear predictive model for yearly data, with an RMSFE of 0.116, demonstrates better accuracy than the individual predictors. For quarterly data, the full predictive model yields an RMSFE of 0.056, lower than all individual predictive models. Similarly, for monthly data, the full model's RMSFE (0.041) is notably lower than that of individual predictive models. The comprehensive model tends to perform better due to the combined effect of multiple predictors capturing various aspects of the data. Single predictor models may not account for the multi-dimensional nature of the data, leading to higher errors. Additionally, the full model can benefit from correlations among predictors, thereby enhancing its predictive accuracy

4.6 Optimizing Predictors in Multiple Regression: Backward Selection with AIC, Forecasting, and Comparative Analysis

The backward selection method is employed to identify the optimal subset of independent variables for excess returns. Using backward selection becomes computationally advantageous when the full linear regression model has already been fitted. The AIC-based model selection, favoring more complex models for improved predictive accuracy, leads to distinct outcomes across yearly, quarterly, and monthly datasets.

Table 6 in the appendix shows the selected variables. For the yearly data, it identifies 8 covariates: Earnings, Treasury Bills, Corporate Bond Yields on BAA-rated Bonds, Long-Term Yield, Inflation, Long-term Corporate Bond Returns, Cross-Sectional beta Premium, and Investment to Capital Ratio with an AIC value of -189.13 and an RMSFE of 0.12, closely matching the full model's RMSFE. In the case of quarterly data, 9 covariates are selected: Dividends, Treasury Bills, Corporate Bond Yields on AAA-rated Bonds, Long-Term Yield, Inflation, Long-term Rate of Returns, Long-term Corporate Bond Returns, Stock Variance, and Quarterly Dividends, yielding an AIC of -312.25 and an RMSFE of 0.057, again closely aligning with the full model's performance. Similarly, the monthly dataset showcases the identification of 9 covariates: Book-to-Market-ratio, Corporate Bond Yields on AAA-rated Bonds, Corporate Bond Yields on BAA-rated Bonds, Long-Term Yield, Net Equity Expansion, Long-term Rate of Returns, Long-term Corporate Bond Returns, Stock Variance, and Cross-Sectional beta Premium with an AIC of -4995.88 and an RMSFE of 0.041, matching the performance of the full model. Table 2 provides an overview of these outcomes for all models.

	Yearly	Quarterly	Monthly
Total Covariates	21	22	18
Covariates Selected	8	9	9
RMSFE	0.121	0.057	0.041

Table 2: Overview of Linear Regression Model with Backward Selection.

4.7 Optimizing Predictors in Multiple Regression: Forward Selection with AIC, Forecasting, and Comparative Analysis

Subsequently, the forward stepwise model selection approach was employed to identify the most promising predictors using AIC. Table 6 in the appendix shows the selected

variables. In the case of yearly data, this method selected four predictors: Inflation, Cross-Sectional beta Premium, Treasury Bills, and Investment to Capital Ratio with an AIC of -184.56 and an associated RMSFE of 0.137. Notably, this RMSFE is higher than those obtained from the backward selection method model and the full model. Similarly, for the quarterly data, four covariates were selected: Stock Variance, Inflation, Treasury Bills, and Corporate Bond Yields on AAA-rated Bonds, yielding an AIC of -312.58 and an RMSFE of 0.062, again showcasing a higher value than the backward selection method model and the full model. For the monthly dataset, the forward stepwise model selection process identified six covariates: Stock Variance, Long-term Corporate Bond Returns, Cross-Sectional beta Premium, Book-to-Market-ratio, Net Equity Expansion, and Long-term Rate of Returns with an AIC of -4990.89 and an RMSFE of 0.042, slightly surpassing the values obtained from the backward selection method model and the full model. Table 3 presents an overview of the model fit using the forward selection method with AIC

	Yearly	Quarterly	Monthly
Total Covariates	21	22	18
Covariates Selected	4	4	6
RMSFE	0.137	0.062	0.041

Table 3: Overview of Linear Regression Model with Forward Selection.

4.8 Summary

The project involved applying autoregressive (AR) and linear predictive techniques to forecast excess returns using yearly, quarterly, and monthly datasets. Determining the AR model's order involved analyzing ACF, PACF plots, and AIC. For the yearly dataset, an AR(2) model resulted in an RMSFE of 0.183, whereas the quarterly dataset showed an RMSFE of 0.093 for an AR(7) model, and the monthly data had an RMSFE of 0.047 for an AR(1) model. While the forecasted values generally aligned with actual returns, unanticipated spikes in the actual data highlighted unforeseen surges not captured by the models, underscoring the limitations of using the same data for fitting and forecasting purposes.

Further analysis included scrutinizing individual predictors using OLS, revealing varied performance among predictors. Some predictors in the yearly dataset, such as Long-term yield and Long-term Corporate Bond Returns (0.192), exceeded the AR model's RMSFE

(0.183), suggesting limited predictive power. Conversely, predictors like Investment to Capital Ratio (0.169) and Consumption-wealth-income ratio (0.168) performed relatively better in predicting returns across different time frequencies. For quarterly data, certain predictors like Book-to-Market-ratio and Treasury Bills showed an RMSFE of 0.107, higher than AR(7), while others like Consumption-Wealth-Income-ratio, Net Equity Expansion, and Quarterly Dividends exhibited lower RMSFEs (0.075 - 0.079), indicating better predictive power. In monthly data, Cross-Sectional beta Premium had the lowest RMSFE (0.046).

Developing a comprehensive predictive model that integrated all predictors for yearly (RMSFE of 0.116), quarterly (RMSFE of 0.056), and monthly data (RMSFE of 0.041) demonstrated enhanced accuracy compared to individual predictor models and AR models. This emphasized the advantages of incorporating multiple predictors capturing diverse aspects of the data, resulting in reduced forecasting errors.

Backward selection via AIC identified subsets of covariates for yearly, quarterly, and monthly datasets, showcasing optimal subsets akin to the performance of the full predictive model. Additionally, forward selection with AIC proposed alternative predictor subsets that slightly outperformed or performed similarly to the backward selection method. The forward selection method generally selected fewer predictors compared to the backward selection method. For instance, in yearly data, both methods selected Inflation, Cross-Sectional beta Premium, Treasury Bills, and Investment to Capital Ratio. Overall, the backward selection method chose 8 predictors with an RMSFE of 0.121, while the forward selection method selected 4 predictors with an RMSFE of 0.137. Similarly, for quarterly and monthly data, both methods yielded subsets of predictors with varying numbers and comparable RMSFEs.

Additionally, it is recommended to expand the comparison of various time series models beyond autoregressive models (AR) to encompass other methodologies like moving average (MA), autoregressive integrated moving average (ARIMA), or machine learning models such as random forests and neural networks. This broader comparison would provide a comprehensive understanding of the models' performances under different conditions. Implementing rolling window or expanding window approaches can simulate real-time forecasting conditions. Evaluating the robustness of models trained on different time periods over time would further enrich the analysis

Bibliography

Peter J. Brockwell and Richard A. Davis. *Time Series: Theory and Methods*. Springer Science+Business Media, New York, 1991. ISBN 978-1-4419-0320-4.

Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian D. Marx. *Regression*. Springer-Verlag GmbH, 2022. 2nd Edition.

Amit Goyal. A Comprehensive Look at the Empirical Performance of Equity Premium Prediction, 2022. URL `https://sites.google.com/view/agoyal145`.

Rob Hyndman, George Athanasopoulos, Christoph Bergmeir, Gabriel Caceres, Leanne Chhay, Mitchell O'Hara-Wild, Fotios Petropoulos, Slava Razbash, Earo Wang, and Farah Yasmeen. *forecast: Forecasting functions for time series and linear models*, 2023. URL `https://pkg.robjhyndman.com/forecast/`. R package version 8.21.1.

Raky Julio, Andres Monzon, and Yusak O. Susilo. Identifying key elements for user satisfaction of bike-sharing systems: a combination of direct and indirect evaluations. *Journal Name*, 2022. doi: 10.1007/s11116-022-10335-3. URL `https://link.springer.com/article/10.1007/s11116-022-10335-3`.

Alboukadel Kassambara. *ggpubr: ggplot2' Based Publication Ready Plots*, 2022. URL `https://rpkgs.datanovia.com/ggpubr/`. R package version 3.4.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL `https://www.R-project.org/`.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL `https://www.R-project.org/`.

Barret Schloerke, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Jason Crowley. *GGally: Extension to 'ggplot2'*, 2021. https://ggobi.github.io/ggally/, https://github.com/ggobi/ggally.

Ivo Welch and Amit Goyal. A Comprehensive Look at The Empirical Performance of Equity Premium Prediction. *The Review of Financial Studies*, 21 (4):1455–1508, 03 2007. ISSN 0893-9454. doi: 10.1093/rfs/hhm014. URL `https://doi.org/10.1093/rfs/hhm014`.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*, 2016. URL `https://ggplot2.tidyverse.org`.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2022. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.

Hadley Wickham, Jim Hester, and Jennifer Bryan. *readr: Read Rectangular Text Data*, 2023a. URL `https://readr.tidyverse.org`. R package version 2.1.4, https://github.com/tidyverse/readr.

Hadley Wickham, Davis Vaughan, and Maximilian Girlich. *tidyr: Tidy Messy Data*, 2023b. URL `https://tidyr.tidyverse.org`. R package version 1.3.0, https://github.com/tidyverse/tidyr.

Appendix

A Additional tables

Variable	Description
Date	Date of observation of Timeseries data.
Index	A benchmark value employed to monitor the performance of a collection of assets.
Dividends	Portion of a company's profits distributed to its shareholders.
Earnings	A company's profits or net income.
Book-to-Market-ratio	Ratio of book value to market value for the Dow Jones Industrial Average.
Treasury Bills	Short-term debt securities issued by the Govt.
Corporate Bond Yields on AAA-rated Bonds	Yield on AAA-rated corporate bonds.
Corporate Bond Yields on BAA-rated Bonds	Yield on BAA-rated corporate bonds.
Long Term Yield	Yield on long-term government bonds.
Net Equity Expansion	Change in equity or capital structure of a company over time
Risk-free-rate	Theoretical interest rate of an investment with no risk of financial loss.
Inflation	The rate at which the general level of prices for goods and services increases over time.
Long-term Rate of Returns	Average return on an investment over an extended period.
Long-term Corporate Bond Returns	Historical returns on long-term corporate bonds.
Stock Variance	Degree of variation in the returns of a stock
Cross-Sectional beta Premium	The excess return expected from holding a stock with a beta higher than the market's beta.
Stock Returns Including Dividends	Stock returns inclusive of dividend payouts
Stock Returns Excluding Dividends	Stock returns exclusive of dividend payouts
Consumption-Wealth-Income-ratio	Consumption Wealth Income ratio
Investment-to-Capital-ratio	Investment to Capital ratio
Percent Equity Issuing	Proportion of equity issuance to total issuance.

Table 4: Data Description

Covariates	Yearly	Quarterly	Monthly
Dividends	0.1847	0.0960	0.0481
Earnings	0.1850	0.0962	0.0481
Book-to-Market-ratio	0.1897	0.1058	0.0555
Treasury Bills	0.19	0.1066	0.0556
Corporate Bond Yields on AAA-rated Bonds	0.1911	0.1070	0.0556
Corporate Bond Yields on BAA-rated Bonds	0.1921	0.1072	0.0556
Long Term Yield	0.1921	0.1071	0.0556
Consumption-wealth-income ratio (Y, Q)	0.1680	0.0792	-
Net Equity Expansion	0.1680	0.0792	0.0569
Inflation	0.1654	0.0788	0.0571
Percent Equity Issuing (Y)	0.1669	-	-
Long-term Rate of Returns	0.1700	0.0786	0.0571
Long-term Corporate Bond Returns	0.1704	0.0782	0.0571
Stock Variance	0.1690	0.0793	0.0572
Cross-Sectional beta Premium	0.1750	0.0791	0.0456
Investment to Capital Ratio (Y, Q)	0.1693	0.0782	-
Quarterly Dividends (Q)	-	0.0754	-
Quarterly Earnings (Q)	-	0.0761	-

Table 5: RMSFEs of Individual Predictive Models for Yearly, Quarterly and Monthly Data

Dataset	Selected Variables Backward Method	Selected Variables Forward Method
Yearly	1. Inflation 2. Treasury Bills 3. Cross-Sectional beta Premium 4. Investment to Capital Ratio 5. Earnings 6. Long-term Corporate Bond Returns 7. Corporate Bond Yields on BAA-rated Bonds 8. Long Term Yield	1. Inflation 2. Treasury Bills 3. Cross-Sectional beta Premium 4. Investment to Capital Ratio
Quarterly	1. Stock Variance 2. Treasury Bills 3. Corporate Bond Yields on AAA-rated Bonds 4. Inflation 5. Long Term Yield 6. Long-term Rate of Returns 7. Long-term Corporate Bond Returns 8. Dividends 9. Quarterly Dividends	1. Stock Variance 2. Treasury Bills 3. Corporate Bond Yields on AAA-rated Bonds 4. Inflation
Monthly	1. Book-to-Market-ratio 2. Net Equity Expansion 3. Long-term Rate of Returns 4. Long-term Corporate Bond Returns 5. Stock Variance 6. Cross-Sectional beta Premium 7. Corporate Bond Yields on AAA-rated Bonds 8. Corporate Bond Yields on BAA-rated Bonds 9. Long Term Yield	1. Book-to-Market-ratio 2. Net Equity Expansion 3. Long-term Rate of Returns 4. Long-term Corporate Bond Returns 5. Stock Variance 6. Cross-Sectional beta Premium

Table 6: Selected Variables after Backward and Forward Selection in Yearly, Monthly, and Quarterly Datasets