FAC 413

INTRODUCTION TO MATHEMATICAL FINANCE AND FINANCIAL ENGINEERING

MINI PROJECT 2

PRESENTED BY :

ARUSHI AGARWAL 2210110201
SRISHTI KHANNA 2210110597

CONTENTS

01.
Abstract
02.
Introduction
03.
Data
04.
Methodology
05.
Answer to Project Questions
06.
Conclusion
07.
Limitations
08.
Results Snapshots
09.
References

ABSTRACT

This project investigates the predictability of stock returns from 2014 to 2018 using firm-specific financial metrics known at the end of 2014. The analysis tests the Efficient Market Hypothesis (EMH), which posits that stock returns should not be systematically predictable based on publicly available information. Using a dataset of 142 observations, the study employs linear regression models to examine the relationship between stock returns and key financial indicators, including Return on Equity (ROE), Return on Capital (ROC), Debt to Capital Ratio (DKR), Earnings per Share (EPS), Net Income, CEO Compensation, and Stock Price.

The methodology involves hypothesis testing to determine the joint and individual significance of the independent variables, as well as an exploration of their statistical characteristics. Interaction terms, polynomial transformations, and domain-specific ratios (e.g., net income to market capitalization) are incorporated to capture non-linear relationships and improve model performance. The results reveal whether firm characteristics from 2014 can predict stock returns, providing insights into market efficiency and potential strategies for stock selection. Key findings address the financial and economic implications of the predictors, the validity of the hypotheses, and the overall strength of evidence for stock return predictability. The study concludes with a discussion of limitations and recommendations for future research.

INTRODUCTION

Stock market efficiency is a fundamental concept in financial economics, primarily guided by the Efficient Market Hypothesis (EMH). According to the EMH, stock prices reflect all available information at any given time, making it impossible to consistently achieve abnormal returns through publicly known data. This study aims to empirically test whether firm-specific financial characteristics known at the end of 2014 have any predictive power over stock returns in the subsequent four-year period (2015–2018).

HYPOTHESIS

This report examines the hypothesis that:

Firm-specific characteristics known at the beginning of the period (end of 2014) do not significantly predict stock returns over the subsequent four-year period (2015–2018).

While financial metrics such as Return on Equity (ROE), Return on Capital (ROC), and Earnings Per Share (EPS) provide insights into a company's past performance, stock prices are continuously influenced by macroeconomic conditions, monetary policy adjustments, global economic trends, and geopolitical events. Since markets integrate new information rapidly, historical firm-level data may have limited ability to forecast future stock returns.

To test this hypothesis, a multiple regression model will be employed, using stock return percentage change from 2014 to 2018 as the dependent variable. Independent variables will include financial performance indicators such as ROE, ROC, Debt to Capital Ratio (DKR), EPS, Net Income, and CEO Salary, with necessary transformations applied where required. The findings of this study will help evaluate whether historical firm-specific data contains any exploitable patterns or whether stock prices truly reflect an efficient market.

DATA

This study utilizes a dataset comprising 142 firm-level observations spanning a four-year period from 2014 to 2018. The dataset includes financial and stock performance metrics recorded at the end of 2014, which serve as independent variables, and the percentage change in stock price from 2014 to 2018, which serves as the dependent variable. The objective is to determine whether firm-specific characteristics available at the start of the period can predict future stock returns, thereby assessing the efficiency of financial markets.

Description of Variables

- ### Dependent Variable:
  - **Stock Return (%)** – The percentage change in stock price from the end of 2014 to the end of 2018.
- ### Independent Variables:
  - **ROE (Return on Equity, 2014)** – A measure of financial performance, calculated as net income divided by shareholders' equity.
  - **ROC (Return on Capital, 2014)** – A profitability ratio that considers both debt and equity financing to evaluate capital efficiency.
  - **DKR (Debt to Capital Ratio, 2014)** – A measure of financial leverage, calculated as total debt divided by total capital.
  - **EPS (Earnings Per Share, 2014)** – Represents net income per outstanding share, indicating firm profitability.
  - **Net Income (2014, in millions ₹)** – The firm's total earnings after all expenses and taxes.
  - **CEO Salary (2014, in thousands ₹)** – Total compensation of the CEO, which may serve as a proxy for managerial effectiveness or agency costs.
  - **Stock Price (sp14, sp18)** – Stock price at the end of 2014 (sp14) and stock price at the end of 2018 (sp18). These are used to calculate stock return and are not included as independent variables.

METHODOLOGY

This study employs a multiple regression model to examine the relationship between firm-specific financial characteristics at the end of 2014 and stock returns over the subsequent four-year period (2015–2018). The objective is to determine whether these firm-specific factors can predict stock price movements, thereby testing the implications of the Efficient Market Hypothesis (EMH).

Data Cleaning

To ensure the dataset is well-structured and free from inconsistencies, several preprocessing steps were undertaken:

Importing Data and Loading Required Libraries

- The dataset was imported from an Excel file using the readxl package.
- Essential libraries such as moments, psych, corrplot, car, lmtest, and sandwich were installed and loaded to facilitate statistical analysis, diagnostics, and visualization.

Initial Data Inspection

- The first six rows of the dataset (head(EM_ReturnsData)) were viewed to understand its structure.
- The variable types were checked using str(), ensuring that all relevant variables were numeric and appropriately formatted.

Summary Statistics and Missing Value Check

- The summary() function was used to obtain an overview of the dataset, including key statistics such as mean, median, and range.
- Missing values were identified using colSums(is.na()) to determine whether imputation or removal was necessary.

Duplicate Detection

- The dataset was checked for duplicate observations using sum(duplicated()) to prevent redundant data points from affecting the analysis.

Data Type Validation

- The sapply() function was used to confirm that all variables were numeric, ensuring compatibility with statistical methods.

Outlier Detection and Visualization

- Boxplots were generated for Return on Equity (ROE) and Stock Returns to visually inspect potential outliers.

- The Interquartile Range (IQR) method was applied to detect extreme values in ROE, with thresholds set at 1.5 times the IQR beyond the first and third quartiles.
- Observations outside this range were flagged as potential outliers for further assessment.

These preprocessing steps ensured that the dataset was clean, free from missing values and duplicates, and ready for statistical analysis.

Variable Transformation

To improve model interpretability and address issues related to skewness, non-linearity, and coefficient interpretation, the following transformations were applied:

Log Transformation for sal and netinc

- Since Sales (sal) and Net Income (netinc) exhibited high skewness and kurtosis, a log transformation was applied to normalize their distributions.
- This transformation also allows for percentage-based interpretation of the regression coefficients, which is more meaningful in financial analysis.
- To handle zero or negative values, a small constant was added before taking the logarithm.

Quadratic Term for dkr (Debt-to-Equity Ratio)

- A squared term (dkr_sq = dkr^2) was introduced to account for non-linear effects and diminishing marginal returns of leverage on stock returns.
- This helps capture the possibility that while leverage initially improves returns, excessive debt may have negative consequences.

Calculation of Price-to-Earnings (P/E) Ratio

- The P/E ratio was computed as Stock Price (sp14) divided by Earnings per Share (eps), a standard measure in financial analysis.
- This variable provides insight into market valuation and investor expectations regarding future earnings potential.

These transformations ensure that the model better captures relationships between financial variables and stock returns while addressing statistical concerns such as skewness and non-linearity.

Descriptive Statistics

To summarize the dataset and understand the distribution, relationships, and potential issues with the independent variables, the following statistical measures were computed:

Summary of Original and Transformed Variables

- The summary() function was used to generate key descriptive statistics (mean, median, minimum, maximum) for both original and transformed variables.
- This helped assess the impact of transformations on normalizing distributions and reducing extreme values.

Calculation of Descriptive Statistics

- For each independent variable, the mean, median, mode, variance, standard deviation (SD), minimum, maximum, and range were computed.
- The mode was determined using a custom function, as R does not have a built-in mode function for continuous data.
- These statistics provided insights into the central tendency and dispersion of the variables.

Skewness and Kurtosis Analysis

- Skewness was calculated to check for asymmetry in variable distributions.
- Kurtosis measured the "tailedness" of the distribution, identifying potential outliers.
- High skewness or kurtosis indicated the need for transformations to meet normality assumptions.

Quantile Distribution

- The 0%, 25%, 50% (median), 75%, and 100% quantiles were computed for all independent variables.
- This helped identify variable distributions and detect outliers beyond the interquartile range.

Correlation Matrix and Multicollinearity Check

- A correlation matrix was generated to measure the linear relationships between independent variables.
- A correlation heatmap was created using corrplot() to visualize strong correlations that may indicate multicollinearity.
- This step was essential for identifying variables that might cause estimation issues in the regression model.

These descriptive statistics provided a foundational understanding of the dataset, ensuring informed decision-making for regression analysis and assumption testing.

Running the Regression

To analyze the relationship between stock returns and financial indicators, a Multiple Linear regression model was estimated.

Model Specification

- $\text{return} = \beta_0 + \beta_1(\text{roe}) + \beta_2(\text{dkr}) + \beta_3(\text{dkr}^2) + \beta_4(\text{eps}) + \beta_5(\log \text{netinc}) + \beta_6(\log \text{sal}) + \beta_7(\text{pe\_ratio}) + \varepsilon$
- This model captures the effect of profitability, leverage, firm size, and market valuation on stock returns.

Selection of Independent Variables

- *Return on Equity (ROE):* A key measure of profitability, indicating how efficiently a company generates returns for shareholders.
- *Debt-to-Equity Ratio (DKR) & Squared Term (DKR²):* Captures the effect of leverage on returns, with the squared term accounting for potential diminishing marginal effects.
- *Earnings Per Share (EPS):* Represents a firm's profitability on a per-share basis and is a key determinant of investor confidence.
- *Log of Net Income (log_netinc):* Used to adjust for skewness and interpret changes in percentage terms, reflecting overall firm profitability.
- *Log of Sales (log_sal):* Included as a proxy for firm size, as larger firms often experience different return patterns than smaller firms.
- *Price-to-Earnings Ratio (PE Ratio):* A valuation metric indicating market expectations of future earnings growth.

Variable Exclusions and Justification

- *Stock Prices (2014 & 2018):* Excluded as they were used to calculate stock returns, making them redundant in the regression.
- *Return on Capital (ROC):* Dropped due to high correlation with ROE, which could introduce multicollinearity. ROE was retained as it is a widely used measure of profitability.

Model Estimation

- The lm() function in R was used to estimate the model.
- The summary(model) output provided insights into coefficient significance, R^2, adjusted R^2, and residual diagnostics.

This model serves as the foundation for testing the Efficient Market Hypothesis (EMH) by examining whether firm-specific financial metrics predict stock returns.

Assumption Testing

To validate the reliability of the regression model, the five classical assumptions of Multiple Linear Regression (MLR) were tested.

Linearity in Parameters

- The model was specified in a linear functional form, ensuring that it satisfies the assumption of linearity in parameters.

Random Sampling

- It was assumed that the dataset provided for the project was obtained through random sampling, making the observations independent of each other.

No Perfect Multicollinearity (Variance Inflation Factor - VIF Test)

- Variance Inflation Factor (VIF) values were calculated, and a bar plot was generated with a reference line at VIF = 10 to identify potential multicollinearity issues.
- All VIF values were within an acceptable range, confirming the absence of severe multicollinearity.

Zero Conditional Mean (ZCM) Assumption

- The residuals were regressed on the independent variables, and a t-test on residuals was conducted to check if their mean is significantly different from zero.
- The Ramsey RESET test was performed to detect possible omitted variable bias or incorrect functional form.
- Both tests confirmed that the assumption held, indicating no evidence of model misspecification.

Homoskedasticity (Breusch-Pagan Test)

- The Breusch-Pagan (BP) test was conducted, and a Residuals vs. Fitted plot was visually inspected.
- The BP test did not indicate significant heteroskedasticity, confirming that the assumption of homoskedasticity holds.

Normality of Residuals (Shapiro-Wilk Test & Visual Inspection)

- The Shapiro-Wilk test was performed to formally test for normality.
- A histogram with a density curve and a Residuals vs. Fitted plot were examined for visual confirmation.
- The Shapiro-Wilk test returned a significant p-value, indicating that the residuals deviate from normality. This was the only assumption that was violated.

All other regression assumptions were satisfied, confirming that the model provides unbiased and efficient estimates. The violation of normality was addressed in the next step.

Correcting the Violation of Normality Assumption

Since the Shapiro-Wilk test and visual inspections of residuals indicated a violation of the normality assumption, steps were taken to address this issue.

Identifying Skewed Variables

- Histograms of all independent variables were plotted to visually inspect skewness.
- Skewness and Kurtosis were computed to quantitatively assess deviations from normality.

Detecting and Removing Outliers

- Cook's Distance was computed to identify highly influential observations that could distort regression estimates.
- Observations exceeding the threshold (4/n) were flagged as outliers and removed from the dataset.
- The regression model was then re-estimated using the cleaned dataset.

Re-Evaluating Normality

- The Shapiro-Wilk test was repeated on the residuals of the updated model.
- A histogram of residuals and a Q-Q plot were generated to visually inspect improvements in normality.

Final Model Assumption Checks

- The Breusch-Pagan test was conducted again to confirm that homoskedasticity was still satisfied.
- The Ramsey RESET test was re-run to check for potential misspecification after modifying the dataset.
- The Zero Conditional Mean (ZCM) assumption was reassessed by computing correlations between residuals and independent variables.
- VIF values were recalculated to ensure that removing outliers did not introduce multicollinearity.

After removing outliers, the Shapiro-Wilk test showed an improvement in residual normality, ensuring that all classical MLR assumptions were satisfied, confirming the validity of the model.

ANSWERS TO PROJECT QUESTIONS

MLR Model :

$$\text{return} = \beta_0 + \beta_1(\text{roe}) + \beta_2(\text{dkr}) + \beta_3(\text{dkr}^2) + \beta_4(\text{eps}) + \beta_5(\log \text{netinc}) + \beta_6(\log \text{sal}) + \beta_7(\text{pe\_ratio}) + \varepsilon$$

Running the Model :

```
Call:
lm(formula = return ~ roe + dkr + dkr_sq + eps + log_netinc +
    log_sal + pe_ratio, data = EM_ReturnsData_cleaned)

Residuals:
    Min      1Q  Median      3Q     Max
-64.512 -20.634  -3.469  17.441 109.038

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.24796   40.19448  -0.354  0.72359
roe          -0.77946    0.40176  -1.940  0.05465 .
dkr          -1.21096    0.65782  -1.841  0.06805 .
dkr_sq        0.02388    0.01200   1.990  0.04882 *
eps           0.06325    0.08989   0.704  0.48299
log_netinc   -0.69071    3.25394  -0.212  0.83225
log_sal       5.15940    6.77970   0.761  0.44811
pe_ratio     -0.99116    0.35494  -2.792  0.00607 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.75 on 123 degrees of freedom
Multiple R-squared:  0.1187,    Adjusted R-squared:  0.06856
F-statistic: 2.367 on 7 and 123 DF,  p-value: 0.02647
```

1. Are your IVs jointly significant at the chosen SL?

Joint Hypothesis :

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$$

$$H_1 : \text{At least one } \beta_i \neq 0$$

The F-test for joint significance yields a p-value of 0.02647. Since this p-value is less than the chosen significance level of 5%, we reject the null hypothesis, indicating that the independent variables are jointly significant in explaining stock returns.

2. What’s the financial and economic implication for the above (item 1)?

Since the null hypothesis (that all independent variables have no joint effect on stock returns) is rejected, this implies that the independent variables jointly explain a significant portion of stock returns, even if some individual coefficients are not significant.

Financial Implications:

- Investment Strategies: Fundamental factors (ROE, EPS, P/E ratio) influence stock returns, aiding in valuation and risk assessment.
- Risk Management: Leverage (dkr, dkr^2) affects returns, requiring firms to manage debt carefully.
- Market Efficiency: The significance of these variables suggests potential inefficiencies, creating opportunities for informed investors.

Economic Implications:

- Regulatory Impact: Authorities may need to monitor corporate leverage to ensure financial stability.
- Capital Allocation: Firms may adjust financial strategies based on profitability and investor expectations.
- Growth and Stability: Corporate financial decisions influenced by these factors can impact broader economic conditions.

3. Is any explanatory variable individually significant?

Yes, the ones which are individually significant are:

- Yes, the variables that are individually significant are:
- pe_ratio ($p = 0.00607$) → Significant at the 1% level.
- dkr^2 ($p = 0.04882$) → Significant at the 5% level.
- Other variables such as roe ($p = 0.05465$) and dkr ($p = 0.06805$) are marginally significant at the 10% level, while eps ($p = 0.48299$), log_netinc ($p = 0.83225$), and log_sal ($p = 0.44811$) are not significant.

4. Which of your hypotheses are corroborated?

- The impact of roe on returns is not strongly significant ($p = 0.05465$), meaning the expected relationship between profitability and stock returns is not fully supported.
- The effect of eps on returns is insignificant ($p = 0.48299$), contradicting the hypothesis that higher earnings per share positively impact stock returns.
- log_netinc and log_sal are not significant ($p = 0.83225$ and $p = 0.44811$, respectively), indicating that net income and firm size (proxied by salary expenses) do not have a clear impact on returns in this model.

5. Which of your hypotheses are not corroborated?

- The non-linear impact of debt (dkr^2) on returns is significant, supporting the hypothesis that debt has a quadratic relationship with stock returns ($p = 0.04882$).
- The negative effect of pe_ratio on stock returns is significant ($p = 0.00607$), confirming the hypothesis that overvalued stocks tend to yield lower returns.

6. What are the financial and economic implications for the above (item 3)?

- **ROE's Weak Significance** – The relationship between return on equity (ROE) and stock returns is weaker than expected, suggesting profitability may not be a strong predictor of returns.
- **EPS's Insignificance** – Earnings per share (EPS) does not significantly impact returns, indicating investors may focus on other factors beyond basic earnings figures.
- **Net Income & Salary's Insignificance** – Log-transformed net income and salary expenses are not significant, implying firm size and net profitability do not strongly influence stock returns in this model.
- **Debt's Mixed Impact** – While dkr^2 is significant, dkr itself is only marginally significant, suggesting a nonlinear effect of debt on returns.

7. Examine the statistical characteristics of the IVs netinc and sal. What do you think about their nature and distribution? Is there a better way to use them in your model?

Nature and Distribution of Net Income and Sales

- From the summary statistics provided in the results section:

Net Income (netinc):

- Exhibits high skewness (3.3269) and high kurtosis (15.7263), indicating a heavily right-skewed distribution with the presence of extreme values.
- The mean (512.3) is much higher than the median (229.5), further confirming the skewness.
- The range is quite large, from a minimum of 4.0 to a maximum of 4237.0.

Sales (sal):

- Also shows substantial skewness (6.3439) and extreme kurtosis (49.3984), meaning that a few firms have disproportionately high sales.
- The mean (1325.1) far exceeds the median (1040.5), reinforcing the skewed nature.
- The maximum value (14,336) is drastically higher than the third quartile (1364.2), indicating the presence of outliers.

Justification for Log Transformations

Given these characteristics, using raw net income and sales in the regression model would have introduced severe skewness and outlier influence, violating the normality assumption. To correct this:

- Log transformations were applied to both variables (log_netinc and log_sal) before running the regression.
- This helped normalize their distributions and improve the interpretability of coefficients, allowing us to analyze percentage changes rather than absolute differences.
- This decision was based on the statistical insights drawn from the summary statistics, which confirmed the need for transformation before proceeding with regression analysis.

Instead of using netinc and sal in their original forms, we transformed them into log_netinc and log_sal based on the observed skewness and extreme values. This approach improved the model's robustness and ensured that the assumptions of normality and linearity were better upheld.

8. If so, do any of your conclusions from the earlier execution change at all?

The summary statistics revealed that netinc and sal exhibited high skewness and kurtosis, indicating potential issues with normality and extreme values. Recognizing this, log transformations were applied at the outset to improve the distribution of these variables.

Since these transformations were incorporated before executing the model, the key conclusions remain unchanged. The adjustments ensured:

- Improved normality of residuals, reducing the impact of outliers.
- Better interpretability, as log transformations allow for an analysis of proportional changes rather than absolute values.
- Preservation of relationships, maintaining economic meaning while addressing statistical concerns.

Because these modifications were made before running the analysis, there is no need to revise any prior findings. Instead, the early application of transformations strengthened the robustness of the conclusions.

9. What about dkr and eps? Are they alright in your model?

Examining the statistical characteristics:

- dkr has a skewness of 0.43 and kurtosis of 2.94, indicating that it is approximately normally distributed and does not require transformation.
- eps exhibits higher skewness (1.89) and kurtosis (14.26), suggesting a right-skewed distribution with heavy tails.

The dkr term and its squared counterpart (dkr_sq) were included in the model to improve interpretability and capture potential non-linear effects. This allows for a more flexible representation of the relationship between dkr and stock returns, ensuring that diminishing marginal effects are appropriately modeled.

Despite the skewness in eps, the model did not violate the classical assumptions of multiple linear regression. Residual diagnostics, including normality tests and graphical assessments, confirmed that the residuals behaved appropriately. Given that no severe assumption violations were detected, eps was retained in its original form to preserve financial interpretability.

10. Overall, is the evidence for predictability of stock returns strong or weak?

Please find the answer included under the section of ‘Conclusion’.

RESULTS AND CONCLUSION

The regression analysis provides insights into the relationship between stock returns and key financial variables. The model's explanatory power appears limited, and while some variables show significant effects, the overall ability to predict stock returns remains weak. The key findings are outlined below:

1. Model Fit and Explanatory Power

- The multiple R-squared value of 0.1187 suggests that the model explains only 11.87% of the variation in stock returns, indicating weak explanatory power.
- The adjusted R-squared of 0.06856 further reinforces that the inclusion of independent variables does not significantly improve the model's predictive ability.
- The F-statistic of 2.367 with a p-value of 0.02647 indicates that at least one independent variable is statistically significant in explaining stock returns. However, the model still explains only a small portion of return variations.

2. Statistical Significance of Variables

- **Statistically Significant Predictors (at 5% level):**
  - pe_ratio (p = 0.00607) is statistically significant, indicating a strong negative relationship between the price-to-earnings ratio and stock returns. The negative coefficient (-0.99116) suggests that firms with higher PE ratios tend to experience lower stock returns, which aligns with valuation-based theories.
  - dkr_sq (p = 0.04882) is significant, indicating a nonlinear effect of the debt-to-equity ratio on stock returns.

- **Insignificant Predictors at 5% Level:**
  - roe (p = 0.05465) and dkr (p = 0.06805), while approaching significance at the 10% level, do not meet the 5% threshold, meaning their relationships with stock returns are not statistically strong enough for firm conclusions.
  - eps (p = 0.48299), log_netinc (p = 0.83225), and log_sal (p = 0.44811) are far from statistical significance, indicating they do not meaningfully impact stock returns in this model.

3. Interpretation of Key Variables

- The negative coefficient of dkr (-1.21096) and positive coefficient of dkr_sq (0.02388) suggest a nonlinear relationship between the debt-to-equity ratio and stock returns. Initially, higher leverage negatively impacts returns, but at extreme levels, the effect may reverse. This is counter intuitive.
- The negative coefficient of pe_ratio supports the idea that stocks with high PE ratios tend to underperform, possibly due to overvaluation.
- The insignificance of log_netinc and log_sal suggests that, even after transformation, these variables do not add meaningful explanatory power to the model.
- The lack of statistical significance for eps implies that earnings per share do not have a strong direct impact on stock return movements in this dataset.

4. Conclusion

- Among the variables tested, only pe_ratio and dkr_sq are statistically significant at the 5% level.
- The low R-squared value indicates that the model captures only a small fraction of return variations, suggesting weak overall predictability.
- These findings imply that additional financial and macroeconomic factors, or alternative modeling techniques, may be necessary to improve predictive accuracy.

LIMITATIONS

While the regression analysis provides insights into the relationship between financial indicators and stock returns, there are several limitations that must be considered:

1. Low Explanatory Power

- The adjusted R-squared value (0.06856) indicates that the model explains only a small portion of the variation in stock returns. This suggests that important factors influencing returns are missing from the analysis.

2. Omitted Variables

- The study focuses on a limited set of financial indicators, excluding key macroeconomic variables such as interest rates, inflation, market sentiment, and sector-specific trends.
- Non-financial factors such as company management, geopolitical events, and investor behaviour may also play a crucial role in determining stock returns.

3. Assumption of Linearity

- The model assumes a linear relationship between independent variables and stock returns, except for the squared term of dkr (dkr_sq). However, financial markets are often influenced by complex, nonlinear dynamics that may not be captured by this approach.

5. Sample-Specific Findings

- The results are specific to the dataset used, covering a particular time period and set of firms. Findings may not generalize well to different market conditions or time horizons.

6. Potential Data Quality Issues

- The presence of outliers or extreme values in financial variables (e.g., net income and sales) could affect the model's estimates, even though transformations were applied.
- Measurement errors in financial reporting could also introduce noise into the analysis.
- The sample size is small.

7. Lack of Market Risk Consideration

- The study does not account for systematic risk factors such as beta, volatility, or broader market indices, which are critical in asset pricing models like CAPM or Fama-French.

While the model provides some insights into stock return predictability, its limitations highlight the need for a broader set of explanatory variables, nonlinear modeling approaches, and robustness checks to improve predictive accuracy. Future research could incorporate macroeconomic indicators, investor sentiment, and alternative statistical techniques to enhance the model's effectiveness.

RESULTS SNAPSHOT

1- View first 6 rows

	roe	roc	dkr	eps	netinc	sp14	sp18	sal	return
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	18.7	17.4	4	48.1	1144	59.4	47	1090	-20.8
2	1.6	2.4	27.3	-85.3	35	47.9	43.5	1923	-9.14
3	4.9	4.6	36.8	-44.1	127	39	72.6	1012	86.2
4	11.1	8.6	46.4	192.	367	61.2	142	579	132.
5	5.6	4.5	36.2	-60.4	214	58	53.2	600	-8.19
6	3.5	2.9	18.7	-79.8	118	68.2	50.5	735	-26.0

2- Summary Statistics for Original Variables

roe	roc	dkr	eps	netinc
Min. : 1.10	Min. : 1.400	Min. : 0.00	Min. :-89.30	Min. : 4.0
1st Qu.:12.15	1st Qu.: 8.425	1st Qu.:12.90	1st Qu.:-15.70	1st Qu.: 151.2
Median :15.80	Median :12.050	Median :26.40	Median : 6.70	Median : 229.5
Mean :17.73	Mean :13.711	Mean :25.65	Mean : 2.02	Mean : 512.3
3rd Qu.:21.18	3rd Qu.:16.850	3rd Qu.:35.48	3rd Qu.: 17.07	3rd Qu.: 520.8
Max. :61.10	Max. :43.300	Max. :79.50	Max. :266.60	Max. :4237.0

sp14	sp18	sal	return
Min. : 13.38	Min. : 9.625	Min. : 267.0	Min. :-84.888
1st Qu.: 36.25	1st Qu.: 30.781	1st Qu.: 733.5	1st Qu.:-30.740
Median : 46.25	Median : 44.900	Median : 1040.5	Median : -8.673
Mean : 55.67	Mean : 47.489	Mean : 1325.1	Mean : -4.043
3rd Qu.: 62.19	3rd Qu.: 57.688	3rd Qu.: 1364.2	3rd Qu.: 14.786
Max. :564.12	Max. :242.500	Max. :14336.0	Max. :131.837

3- Skewness and Kurtosis of Original Variables

	Skewness	Kurtosis
roe	1.3703211	6.488881
roc	1.3845499	5.312104
dkr	0.4362501	2.938663
eps	1.8939094	14.261033
netinc	3.3269811	15.726308
sal	6.3438468	49.398426
sp14	7.4286778	71.482595
sp18	3.3003967	23.370059
return	0.9289140	4.095208

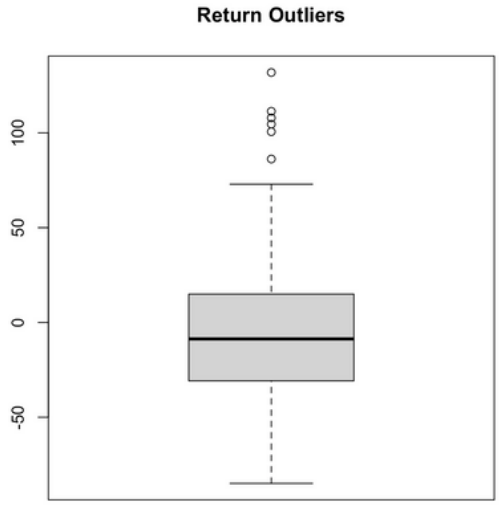
4- Check for missing values

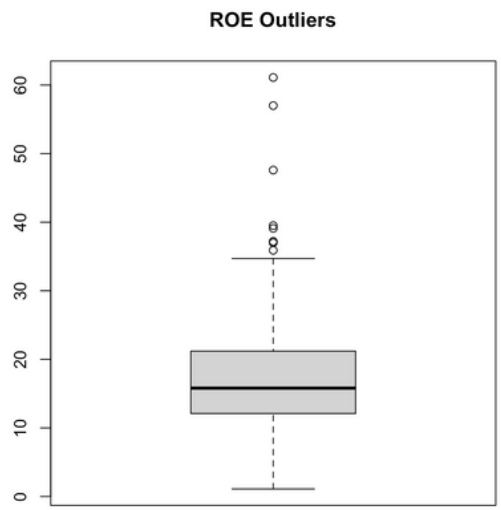
roe	roc	dkr	eps	netinc	sp14	sp18	sal	return
0	0	0	0	0	0	0	0	0

5 - Check if all columns are numeric

roe	roc	dkr	eps	netinc	sp14	sp18	sal	return
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

6- Boxplots to visualize outliers





7- Summary Statistics for Transformed Variables

log_sal	log_netinc	dkr_sq	pe_ratio
Min. :5.587	Min. :1.386	Min. : 0.0	Min. :-108.750
1st Qu.:6.598	1st Qu.:5.019	1st Qu.: 166.5	1st Qu.: -1.014
Median :6.947	Median :5.436	Median : 697.0	Median : 1.500
Mean :6.958	Mean :5.615	Mean : 933.7	Mean : 1.799
3rd Qu.:7.218	3rd Qu.:6.255	3rd Qu.:1258.5	3rd Qu.: 4.318
Max. :9.571	Max. :8.352	Max. :6320.2	Max. : 50.611

8- Descriptive statistics for Independent Variables

	Mean	Median	Mode	Variance	SD	Min	Max	Range
roe	17.730986	15.800000	13.700000	9.770641e+01	9.8846552	1.100000	61.100000	60.000000
roc	13.711268	12.050000	11.800000	6.212257e+01	7.8817871	1.400000	43.300000	41.900000
dkr	25.648592	26.400000	0.000000	2.778361e+02	16.6684172	0.000000	79.500000	79.500000
dkr_sq	933.729789	696.970000	0.000000	1.086237e+06	1042.2269725	0.000000	6320.250000	6320.250000
eps	2.019718	6.700000	12.800000	1.862300e+03	43.1543761	-89.300000	266.600000	355.900000
log_netinc	5.615309	5.435901	4.990433	1.303841e+00	1.1418585	1.386294	8.351611	6.965316
log_sal	6.957958	6.947431	6.835185	3.523154e-01	0.5935617	5.587249	9.570529	3.983280
pe_ratio	1.799192	1.500189	1.234407	1.680396e+02	12.9630077	-108.750000	50.611111	159.361111

9- Skewness and Kurtosis of Variables

	Skewness	Kurtosis
roe	1.3703211	6.488881
roc	1.3845499	5.312104
dkr	0.4362501	2.938663
dkr_sq	2.0807119	9.036644
eps	1.8939094	14.261033
log_netinc	-0.2779684	4.284120
log_sal	0.9756558	6.328570
pe_ratio	-3.5329498	41.206436

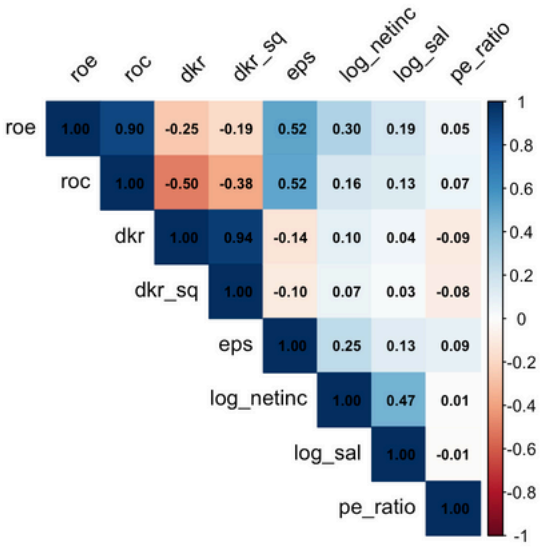
10 - Quantiles

	0%	25%	50%	75%	100%
roe	1.100000	12.150000	15.800000	21.175000	61.100000
roc	1.400000	8.425000	12.050000	16.850000	43.300000
dkr	0.000000	12.900000	26.400000	35.475000	79.500000
dkr_sq	0.000000	166.530000	696.970000	1258.477500	6320.250000
eps	-89.300000	-15.700000	6.700000	17.075000	266.600000
log_netinc	1.386294	5.018930	5.435901	6.255261	8.351611
log_sal	5.587249	6.597827	6.947431	7.218358	9.570529
pe_ratio	-108.750000	-1.013961	1.500189	4.317852	50.611111

11- Correlation Matrix

	roe	roc	dkr	dkr_sq	eps	log_netinc	log_sal	pe_ratio
roe	1.00000000	0.90343467	-0.25394634	-0.18743612	0.52009169	0.295719705	0.18660102	0.054946850
roc	0.90343467	1.00000000	-0.50354485	-0.37988705	0.51871103	0.163095486	0.13449973	0.068029614
dkr	-0.25394634	-0.50354485	1.00000000	0.93628488	-0.13965674	0.100162234	0.04306088	-0.094641470
dkr_sq	-0.18743612	-0.37988705	0.93628488	1.00000000	-0.10069991	0.072159141	0.02527220	-0.081102365
eps	0.52009169	0.51871103	-0.13965674	-0.10069991	1.00000000	0.245853995	0.12519012	0.093672144
log_netinc	0.29571971	0.16309549	0.10016223	0.07215914	0.24585399	1.000000000	0.46762897	0.009264487
log_sal	0.18660102	0.13449973	0.04306088	0.02527220	0.12519012	0.467628967	1.00000000	-0.013683826
pe_ratio	0.05494685	0.06802961	-0.09464147	-0.08110237	0.09367214	0.009264487	-0.01368383	1.000000000

12 - Correlation Heatmap



13 - Running the Base Model

```
Call:
lm(formula = return ~ roe + dkr + dkr_sq + eps + log_netinc +
    log_sal + pe_ratio, data = EM_ReturnsData)

Residuals:
    Min      1Q  Median      3Q     Max
-64.584 -24.867  -4.357  23.267 109.746

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -20.171066  38.117699  -0.529   0.5976
roe          -0.923722   0.403100  -2.292   0.0235 *
dkr          -1.082271   0.568812  -1.903   0.0592 .
dkr_sq        0.021271   0.008867   2.399   0.0178 *
eps           0.154683   0.087281   1.772   0.0786 .
log_netinc   -2.661477   3.333120  -0.798   0.4260
log_sal       8.050631   6.084280   1.323   0.1880
pe_ratio     -0.544345   0.247686  -2.198   0.0297 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 37.81 on 134 degrees of freedom
Multiple R-squared:  0.1255,    Adjusted R-squared:  0.07986
F-statistic: 2.748 on 7 and 134 DF,  p-value: 0.01067
```

Joint Hypothesis :

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$$

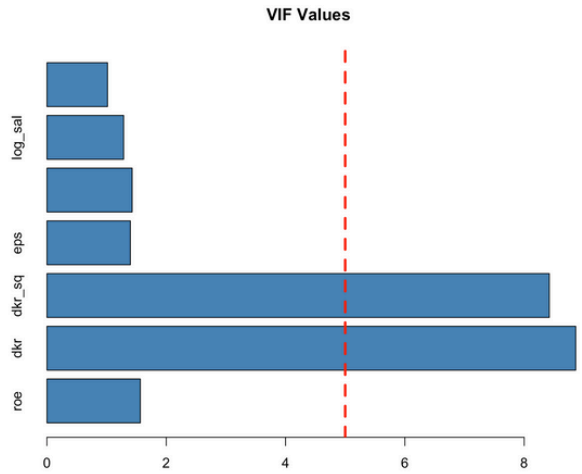
$$H_1 : \text{At least one } \beta_i \neq 0$$

The F-test for joint significance yields a p-value of 0.01067. Since this p-value is less than the chosen significance level of 5%, we reject the null hypothesis, indicating that the independent variables are jointly significant in explaining stock returns.

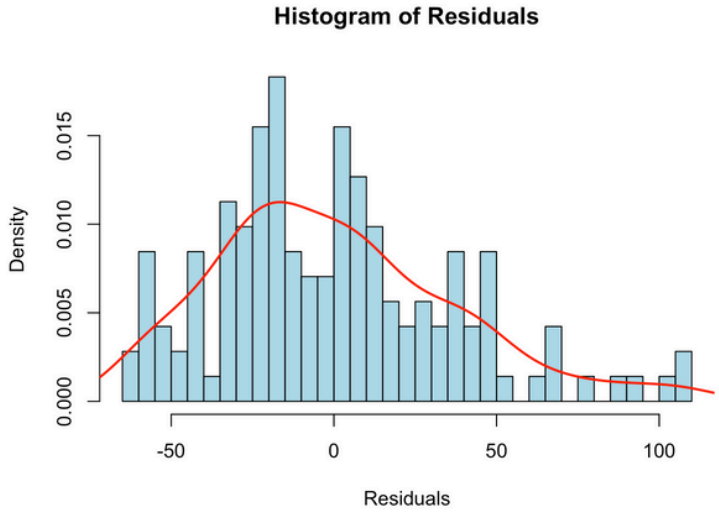
14 - VIF Result

roe	dkr	dkr_sq	eps	log_netinc	log_sal	pe_ratio
1.565655	8.864917	8.421777	1.399069	1.428481	1.286169	1.016625

15 - VIF Bar Plot



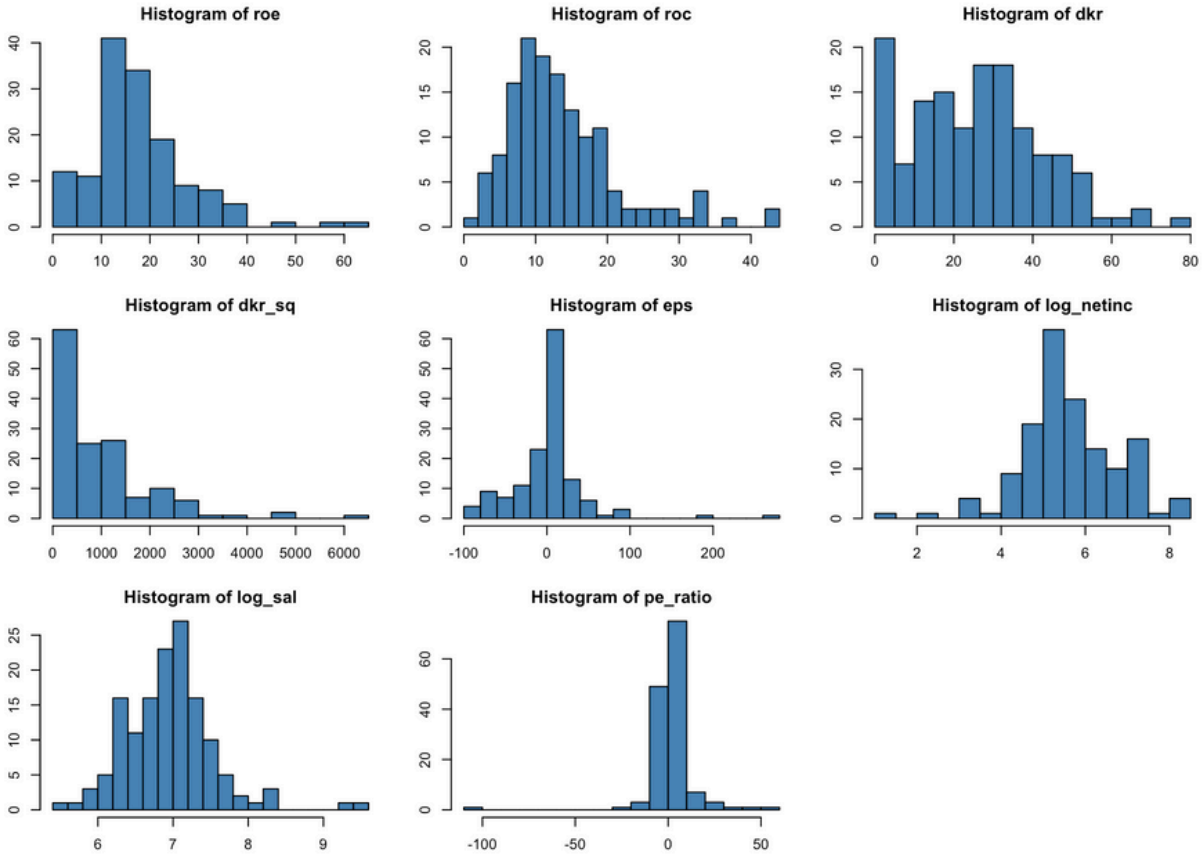
16 - Histogram of Residuals



17 - Shapiro - Wilk Normality test

```
data:  residuals(model)
W = 0.96442, p-value = 0.0009501
```

18 - Histogram of IV's



19 - Correlation matrix of final IV's

	roe	dkr	dkr_sq	eps
roe	" 1.000000"	"-0.253946"	"-0.187436"	" 0.520092"
dkr	"-0.253946"	" 1.000000"	" 0.936285"	"-0.139657"
dkr_sq	"-0.187436"	" 0.936285"	" 1.000000"	"-0.100700"
eps	" 0.520092"	"-0.139657"	"-0.100700"	" 1.000000"
log_netinc	" 0.295720"	" 0.100162"	" 0.072159"	" 0.245854"
log_sal	" 0.186601"	" 0.043061"	" 0.025272"	" 0.125190"
pe_ratio	" 0.054947"	"-0.094641"	"-0.081102"	" 0.093672"

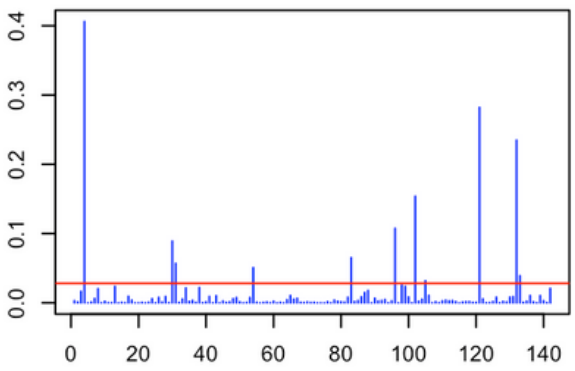
	log_netinc	log_sal	pe_ratio
roe	" 0.295720"	" 0.186601"	" 0.054947"
dkr	" 0.100162"	" 0.043061"	"-0.094641"
dkr_sq	" 0.072159"	" 0.025272"	"-0.081102"
eps	" 0.245854"	" 0.125190"	" 0.093672"
log_netinc	" 1.000000"	" 0.467629"	" 0.009264"
log_sal	" 0.467629"	" 1.000000"	"-0.013684"
pe_ratio	" 0.009264"	"-0.013684"	" 1.000000"

20 - Skewness and Kurtosis of Residuals (Initial vs Cleaned Model)

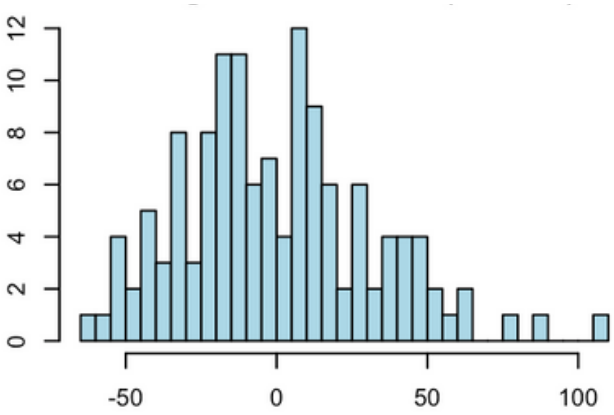
	Statistic	Value
1	Skewness	0.6764612
2	Kurtosis	3.2903221

```
data:  residuals(model_cleaned)
W = 0.9801, p-value = 0.05158
```

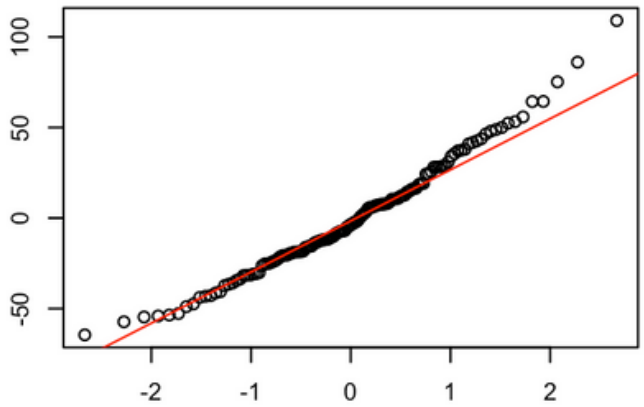
21 - Cook's Distance



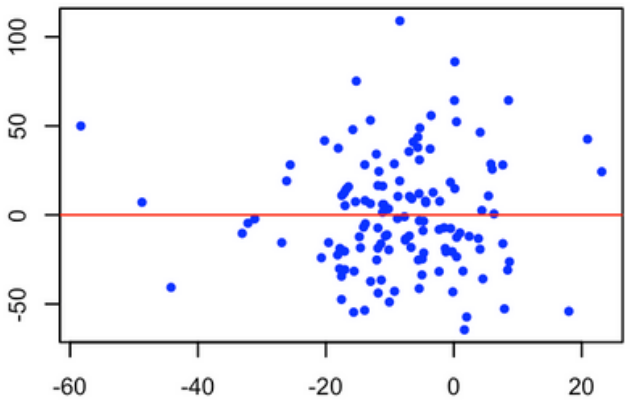
22 - Histogram of Residuals (Cleaned)



23 - Normal Q-Q Plot



24 - Residuals Vs. Fitted



25 - Breusch-Pagan Test

```
data:  model_cleaned
BP = 4.3719, df = 7, p-value = 0.7361
```

26 - Ramsey RESET Test

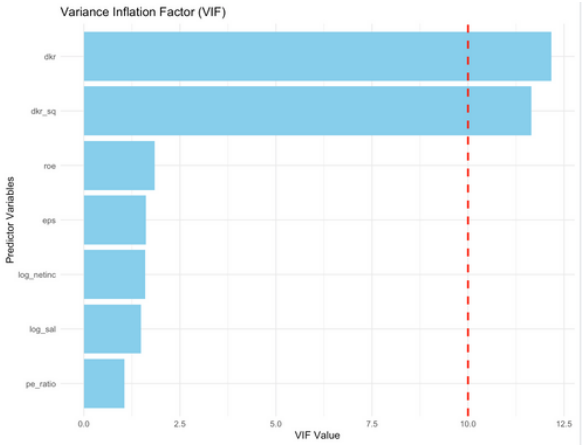
```
data:  model_cleaned
RESET = 0.85844, df1 = 1, df2 = 122, p-value = 0.356
```

27 - Correlation between Residuals and Predictors

```
           roe            dkr         dkr_sq            eps
6.195440e-18   7.623969e-17   5.167186e-17   3.274853e-17
  log_netinc        log_sal       pe_ratio
9.748888e-17   2.253832e-16 -6.815623e-17
```

28 - VIF

```
           roe            dkr         dkr_sq            eps
6.195440e-18   7.623969e-17   5.167186e-17   3.274853e-17
  log_netinc        log_sal       pe_ratio
9.748888e-17   2.253832e-16 -6.815623e-17
```



29 - Final Model Statistics

```
Call:
lm(formula = return ~ roe + dkr + dkr_sq + eps + log_netinc +
    log_sal + pe_ratio, data = EM_ReturnsData_cleaned)

Residuals:
    Min      1Q  Median      3Q     Max
-64.512 -20.634  -3.469  17.441 109.038

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.24796   40.19448  -0.354  0.72359
roe          -0.77946    0.40176  -1.940  0.05465 .
dkr          -1.21096    0.65782  -1.841  0.06805 .
dkr_sq        0.02388    0.01200   1.990  0.04882 *
eps           0.06325    0.08989   0.704  0.48299
log_netinc   -0.69071    3.25394  -0.212  0.83225
log_sal       5.15940    6.77970   0.761  0.44811
pe_ratio     -0.99116    0.35494  -2.792  0.00607 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.75 on 123 degrees of freedom
Multiple R-squared:  0.1187,    Adjusted R-squared:  0.06856
F-statistic: 2.367 on 7 and 123 DF,  p-value: 0.02647
```

REFERENCES

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. The Journal of Finance, 25, 383-417.

Malkiel, B. G. (2003). A Random Walk Down Wall Street. Princeton University Working Paper.

Morningstar Inc. (2024). Active/Passive Barometer Study.

Lo, A. W., & MacKinlay, A. C. (1999). A Non-Random Walk Down Wall Street. Princeton University Press.

Thaler, R. H. (2000). From Homo Economicus to Homo Sapiens. Journal of Economic Perspectives, 14(1), 133-141.

Cootner, P. H. (1964). The Random Character of Stock Market Prices. MIT Press.

Da-Souza, C. A. (2022). EMH vs Behavioral Finance. Saint Peter's University Working Paper.