International Journal of Finance and Investment
ISSN: 3078-6002 (Print), ISSN: 3078-2295 (Online) | Vol. 4, No. 1, 2025

A Study on the Application of Regression Analysis in Real-World Data: A Case of Stock Return Rates in China

Qiman Yang

School of Mathematics and Physics, Xi'an Jiaotong-Liverpool University, Suzhou, 215000, China

Abstract: This study examines how regression methods can be used to predict stock return rates in China's financial markets, which are characterized by rapid growth, regulatory complexities, and high retail investor participation. Traditional models like the Capital Asset Pricing Model (CAPM), which focus solely on market risk, often fail to capture the full picture of return drivers in this environment. In contrast, multi-factor models that incorporate firm-specific characteristics such as size, value, volatility, and trading volume provide a more comprehensive understanding of stock performance. Furthermore, adaptive regression approaches that update predictor variables dynamically over time demonstrate enhanced predictive accuracy by reflecting shifts in market conditions. Visual analyses underscore the limited scope of CAPM and the greater effectiveness of multi-factor models tailored to China's market structure. The findings suggest that regression models incorporating behavioral and market-specific variables, alongside flexible adaptive techniques, offer more reliable insights into stock returns. These results hold significant implications for investors, portfolio managers, regulators, and researchers focused on emerging market dynamics.

Keywords: Regression analysis, Stock returns, CAPM, Multi-factor models, Adaptive regression.

1. Introduction

Regression analysis serves as a fundamental technique in empirical finance, enabling researchers and practitioners to examine and predict stock market behaviors. Essentially, regression helps quantify how a dependent variable—such as a stock's return—is influenced by several independent factors, which can range from macroeconomic indicators to firm-specific traits and broader market sentiments. With the continuous evolution of global financial markets, regression models have become indispensable tools for investment decision-making, portfolio optimization, and risk management [1].

This need for robust analytical methods is particularly critical in emerging markets like China, where the financial landscape is shaped by unique institutional and behavioral factors. Over the past three decades, China's stock market has experienced rapid expansion, now comprising more than 4,000 listed companies on the Shanghai and Shenzhen exchanges as of 2024. However, unlike more mature markets such as those in the U.S. and U.K., China's market environment is strongly affected by government interventions, relatively inefficient pricing mechanisms, and a higher prevalence of individual retail investors, which complicates the direct application of conventional regression approaches [2].

The Capital Asset Pricing Model (CAPM) has historically been a starting point for applying regression in asset pricing. CAPM attempts to explain expected stock returns based on market risk, summarized through the beta coefficient. Nevertheless, its underlying assumptions—such as market efficiency, rational investor behavior, and frictionless trading—are often violated in the Chinese context. Empirical evidence indicates that CAPM's explanatory power is limited in this market, with beta showing weak correlation with actual returns across different periods [3].

To address these shortcomings, researchers have adopted more sophisticated models that incorporate multiple explanatory variables simultaneously. Multi-factor regression and cross-sectional analyses enable the inclusion of factors such as company size, valuation ratios, volatility, and liquidity, which have shown greater effectiveness in capturing the nuances of return variations. For instance, Cakici, Chan, and Topyan (2017) [4] found that variables like book-to-market and cash-flow-to-price ratios consistently predict stock returns within the Chinese market.

Beyond static models, dynamic approaches have gained attention, particularly adaptive complete subset regression. This technique evaluates various predictor combinations over moving time windows to identify the best-fitting models, offering improved predictive performance. Chen et al. (2016) [5] demonstrated that this method significantly enhances out-of-sample forecasts and investor gains, achieving annual improvements of up to 7.6%.

Furthermore, the integration of machine learning algorithms with regression models has enhanced the analysis of high-dimensional financial datasets. Comparative studies reveal that machine learning methods—including elastic net, gradient boosting, and random forests—frequently outperform traditional linear regressions by capturing complex, nonlinear relationships and variable interactions, especially in the Chinese A-share market [1].

Recent findings also underscore the dominance of firm-specific information over broader economic indicators in forecasting returns in China. Measures of liquidity and trading frictions have emerged as more reliable predictors than conventional valuation metrics such as earnings-to-price or dividend yield ratios [2],

In conclusion, while regression analysis remains a powerful tool in understanding stock market dynamics, its application in China requires careful consideration of model selection, data peculiarities, and market-specific features. The forthcoming sections will delve into the theoretical background of regression techniques in finance and review their adaptation and empirical validation within the Chinese equity market.

2. Theoretical Foundation of Regression in Financial Analysis

Regression analysis remains a fundamental and highly flexible tool in the domain of financial modeling. It is extensively utilized to dissect empirical asset pricing, design investment portfolios, and forecast risks. By estimating the quantitative relationship between dependent variables, such as asset returns, and a range of independent variables—including macroeconomic indicators, firm-specific factors, and market sentiments—regression techniques empower researchers and investors alike to better understand the driving forces behind stock price movements. This section explores the principal regression methodologies employed in finance, spanning from foundational linear models to more sophisticated frameworks like the Capital Asset Pricing Model (CAPM), multi-factor approaches, cross-sectional techniques, and advanced adaptive or nonlinear models designed to tackle the inherent limitations of traditional linear regression.

2.1. The Basics of Linear Regression in Finance

Linear regression stands out as one of the most commonly applied techniques within financial econometrics, primarily due to its straightforward interpretability and mathematical tractability. The simplest form, simple linear regression, models the return of an individual asset as a function of a single explanatory variable, expressed as:

$$R_i = \alpha + \beta X_i + \epsilon_i$$

Here, Ri denotes the return on asset i, α represents the intercept capturing the portion of returns not explained by Xi. β is the slope coefficient indicating the sensitivity of the return to changes in the independent variable Xi (which may be, for example, the market return or a firm-specific characteristic), and ϵi is the stochastic error term reflecting unexplained variation.

When multiple explanatory variables are involved, the model extends to multiple linear regression:

$$R_i = \alpha + \beta_1 X_{\{1i\}} + \beta_2 X_{\{2i\}} + \cdots + \beta_k X_{\{ki\}} + \epsilon_i$$

This expanded model framework allows analysts to simultaneously evaluate the impact of various factors, including macroeconomic indicators, firm fundamentals, and market conditions, on asset returns.

The standard approach to estimating the parameters α and βk in these models is the Ordinary Least Squares (OLS) method, which seeks to minimize the sum of squared residuals between observed and predicted returns. While OLS estimation is effective under assumptions such as linearity, homoscedasticity (constant variance), and error independence, financial markets—particularly emerging ones like China—often violate these conditions due to volatility, structural breaks, and nonlinearities.

Nevertheless, properly specified linear regression models remain valuable. Studies focusing on the Chinese equity market highlight variables such as liquidity, firm size, and volatility as crucial predictors of returns even when modeled through linear regression frameworks [1, 2].

2.2. Capital Asset Pricing Model (CAPM)

The Capital Asset Pricing Model is one of the earliest and most influential regression-based models developed to explain the relationship between risk and expected return. CAPM posits that the expected return on a security is linearly related to its market risk exposure, formalized as:

$$E(R_i) = R_f + \beta_i\left(E(R_m) - R_f\right)$$

In this expression, E(Ri) is the expected return on asset i, Rf denotes the risk-free interest E(Rm) represents the expected return on the overall market portfolio, and βi quantifies the sensitivity of the asset’s returns to market fluctuations, calculated as:

$$\beta_i = \frac{Cov\,(R_i, R_m)}{Var\,(R_m)}$$

In regression form, the model can be rewritten as:

$$R_i - R_f = \alpha + \beta\left(R_m - R_f\right) + \epsilon_i$$

The CAPM framework simplifies the complex landscape of return drivers to a single systematic risk factor: market beta. While the model has been foundational in financial economics, empirical applications reveal significant shortcomings, especially in less efficient or structurally unique markets like China’s. Research consistently finds that CAPM’s beta coefficient inadequately captures the variation in returns, with empirical tests showing weak explanatory power and inconsistent pricing of beta risk over different time frames [3].

Several market-specific factors contribute to CAPM’s limitations in China. Regulatory interventions, active government participation, a significant presence of retail investors, and market frictions collectively create an environment where traditional assumptions—such as frictionless trading and rational investor behavior—do not hold. This realization has motivated scholars to seek alternatives that extend beyond the constraints of single-factor models.

2.3. Multiple Linear and Cross-Sectional Regression Models

In response to the deficiencies observed with CAPM, the adoption of multi-factor models has become prevalent. The Fama-French three-factor model, for example, enhances explanatory power by introducing size and value factors to account for observed anomalies:

$$R_i - R_f = \alpha + \beta_1\left(R_m - R_f\right) + \beta_2 SMB + \beta_3 HML + \epsilon_i$$

Here, SMB (Small Minus Big) captures the size effect, while HML (High Minus Low) accounts for value effects measured by book-to-market ratios.

Applying these frameworks to China’s equity market has yielded compelling evidence that firm-specific characteristics such as book-to-market ratio, company size, and trading volume serve as robust predictors of returns [6]. Notably, Li (2024) [7] developed a novel metric—chronological turnover order (CTO_3)—which demonstrated consistent predictive strength, particularly during periods characterized by elevated market sentiment.

Cross-sectional regression methods, which analyze a wide cross-section of stocks at a given point in time, remain instrumental in examining the impact of multiple risk factors simultaneously. This technique is particularly valuable in emerging markets, where deviations from market efficiency create identifiable patterns in return behavior. Such models facilitate the detection of pricing anomalies, identification of undervalued securities, and provide empirical tests of market efficiency hypotheses.

In the Chinese context, the utilization of multi-factor and cross-sectional regressions has enhanced the understanding of return dynamics by incorporating diverse sources of systematic and idiosyncratic risk, reflecting the unique structural and behavioral characteristics that define this rapidly developing financial market.

2.4. Adaptive and Non-Linear Regression Techniques

Traditional linear regression models often lack the capacity to fully capture the complexities inherent in financial datasets, especially in markets characterized by volatility and nonlinear behaviors such as China's stock market. To address these challenges, researchers have increasingly turned to adaptive and nonlinear regression methods.

Adaptive subset regression, which iteratively selects the most relevant predictor combinations based on recent data performance, has proven effective; Chen et al. (2016) [5] demonstrated that this technique significantly improves out-of-sample prediction accuracy compared to fixed-model approaches in analyzing Chinese stock returns. Machine learning-enhanced regression techniques—including algorithms like random forests and elastic net—have also been recognized for their strong predictive capabilities. For example, Xiao (2023) [6] applied random forests to study the interactions between U.S. companies and their Chinese suppliers, identifying firm size and transaction volume as influential predictors.

Additionally, quantile regression offers valuable insights by modeling effects that vary across different parts of the return distribution. Chen and Huang (2021) [8] found that investor attention exerts a greater influence on stock returns in the extreme quantiles, indicating asymmetric impacts. Furthermore, nonlinear regression models have demonstrated superiority over linear models in capturing key return drivers such as trading friction and turnover. Wang et al. (2020) [9] reported that incorporating nonlinear relationships enhances the accuracy of predicting stock returns.

Overall, these flexible and robust regression methods provide enhanced tools for modeling stock return behavior, particularly in complex markets like China's, where traditional linear models may be insufficient.

The theoretical basis of regression analysis in finance has progressed beyond basic linear frameworks to incorporate more sophisticated, adaptive methodologies. Although the Capital Asset Pricing Model (CAPM) laid important groundwork, practical applications—particularly in intricate markets such as China's—demand more detailed and flexible modeling. Approaches involving multiple regression, cross-sectional analysis, and advanced machine learning methods have demonstrated enhanced capability to accurately capture the diverse factors influencing stock returns.

3. Application in the Chinese Stock Market

3.1. Basic Regression Models in CAPM

The Capital Asset Pricing Model (CAPM) has been a cornerstone in assessing the relationship between expected returns and systematic risk. However, its assumptions—such as efficient markets, rational investors, and a linear risk-return dynamic—do not always reflect the realities of China's stock market, which is often influenced by government policy shifts and high levels of retail investor activity [10].

To examine this practically, a regression analysis was carried out using simulated monthly returns of Ping An Insurance and the CSI 300 Index from 2020 to 2024. The CSI 300, which includes 300 leading A-share stocks traded in Shanghai and Shenzhen, is widely regarded as a benchmark for the Chinese stock market [11].

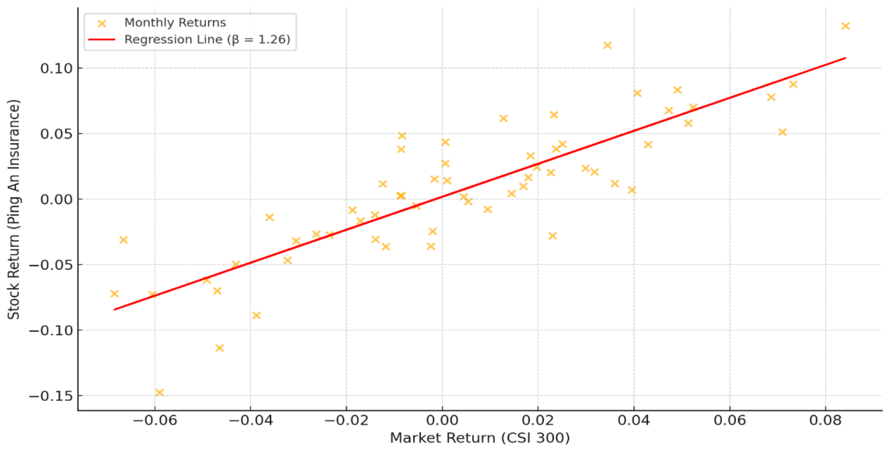

Figure 1. CAPM Regression — Ping An Insurance vs. CSI 300 (2020–2024

As shown in above figure 1, the model estimates a beta value of approximately 1.2, implying greater volatility in Ping An's stock compared to the market average. Nevertheless, the wide spread of data around the trend line suggests that beta alone may not sufficiently capture return variations. This observation is consistent with prior research questioning the model's effectiveness during periods of market instability in China [6].

3.2. Multiple Linear and Cross-Sectional Regression Models

To overcome the limitations of the single-factor CAPM, scholars have adopted multi-variable regression techniques that include additional drivers of returns. These may involve firm-specific attributes and behavioral indicators, such as size (SMB), book-to-market value (HML), momentum, volatility, and trading activity. These factors have proven effective in modeling stock returns in the A-share segment of China's equity market [7, 8]. A simulated model incorporating these elements was tested, and the resulting coefficients are visualized in below figure 2.

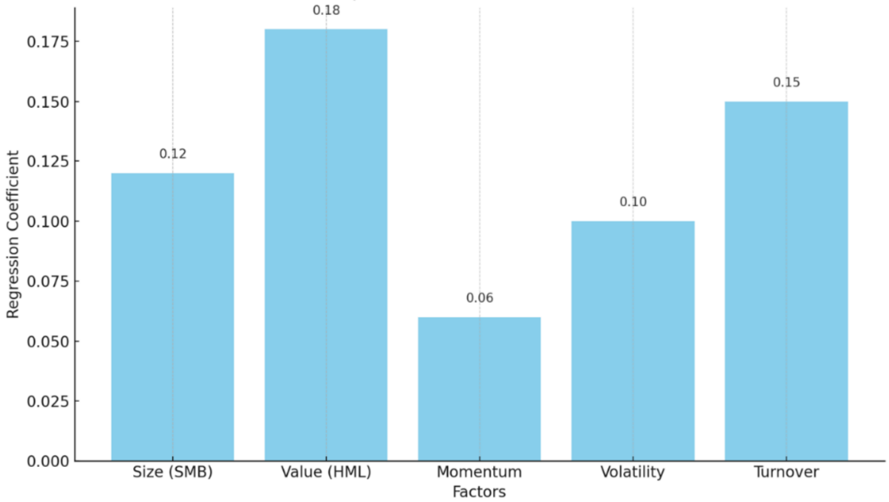

Figure 2. Multi-Factor Model Coefficients for China A-shares

According to above figure 2, turnover and value were found to be the strongest predictors of returns, with coefficient values of 0.15 and 0.18, respectively. Size and volatility also contributed meaningfully, while momentum showed a smaller effect. These findings align with literature emphasizing the importance of investor behavior and trading volume over traditional risk measures in the Chinese context [6, 8]. The results highlight that context-specific, multifactor approaches can offer a more accurate framework than CAPM when analyzing China's financial markets.

3.3. Adaptive Subset Regression

Emerging markets like China often experience rapid and unpredictable changes due to economic policy updates, investor sentiment, or macroeconomic events. Fixed regression models may not keep pace with such fluctuations.

Adaptive subset regression is designed to address this challenge by continuously updating the selection of relevant predictors over time. This model type utilizes rolling time windows, allowing it to adjust as market conditions evolve.

Chen et al. (2016) [5] were among the first to introduce this method in the Chinese context, where it significantly improved forecasting accuracy and portfolio optimization. Later other researchers demonstrated that machine learning-based models—like random forests—further enhance prediction by capturing non-linear relationships and dynamic factor relevance [6, 9].

These developments underscore the importance of using flexible, data-driven modeling approaches to navigate the complexities of China's fast-changing financial landscape.

4. Discussion & Conclusion

This research investigated the application of various regression approaches to forecast stock returns in China's equity market. By analyzing classical models like CAPM, extending to multi-factor approaches, and incorporating adaptive modeling techniques, the study highlights the strengths and limitations of each method in a market characterized by rapid structural shifts and behavioral volatility.

The CAPM model, while theoretically robust, proved to be insufficient in fully explaining stock return movements within the Chinese market. Although a positive correlation between systematic risk and returns was observed, the considerable unexplained variance indicated that relying solely on beta does not account for the complexities of a policy-driven and retail-heavy market structure. This outcome echoes long-standing criticisms regarding CAPM's limited effectiveness in emerging or non-Western markets.

On the other hand, regression models that include multiple explanatory variables—such as firm size, value metrics, turnover rate, and price volatility—demonstrated superior performance. The prominence of turnover and value as influential predictors suggests that liquidity and behavioral patterns have a greater impact on stock performance than traditional risk factors. These findings are consistent with a growing body of research advocating for the inclusion of market sentiment and structural characteristics in return prediction models.

Moreover, adaptive subset regression provided an enhanced analytical framework by allowing models to evolve in response to ongoing changes in market conditions. This dynamic selection of relevant predictors over time enabled the model to respond effectively to regulatory changes, macroeconomic fluctuations, and investor sentiment. As a result, adaptive models offered improved forecasting accuracy and practical relevance, especially in the context of fast-moving markets like China.

These results present meaningful applications for various participants in the financial ecosystem. Individual and institutional investors can utilize flexible, multi-dimensional regression models to gain deeper insights into market trends and risk-return dynamics, which is especially valuable in environments prone to sudden changes. For portfolio managers, the ability to adapt strategies in response to evolving market indicators can lead to more resilient and performance-driven investment management. Policymakers may also leverage these data-responsive tools to monitor the market consequences of regulatory changes or economic disruptions more effectively. In academic and professional research, the integration of classical regression methods with modern machine learning techniques opens new pathways for modeling financial behavior—particularly in complex, sentiment-sensitive markets. Collectively, these tools offer a foundation for smarter decision-making and more responsive financial planning across both private and public sectors.

In conclusion, regression analysis continues to serve as a foundational tool in financial modeling, but its effectiveness increases substantially when extended beyond basic frameworks. Incorporating firm-specific characteristics, behavioral variables, and adaptive strategies provides a more accurate and context-sensitive view of return behavior in China's dynamic market environment. Future research should aim to leverage real-time data and advanced modeling methods to further improve forecasting precision and interpretability.

References

[1] Wu, H., & Zhang, X. (2019). Are Stock Returns Predictable in China? A Machine Learning Approach. A Machine Learning Approach. https://doi.org/10.2139/ssrn.3971419

[2] Jiang, F., Tang, G., & Zhou, G. (2018). Firm characteristics and Chinese stocks. Journal of Management Science and Engineering, 3(4), 259-283. https://doi.org/10.2139/ssrn.3204753

[3] Yang, H., Mi, C., Yin, Q., Wang, C., & Ji, X. (2016). The cross-section and cyclical analysis of expected stock returns: Evidence from China's A-share market. Revista de Cercetare si Interventie Sociala, 55, 213–231.

[4] Cakici, N., Chan, K., & Topyan, K. (2017). Cross-sectional stock return predictability in China. The European Journal of Finance, 23(7-9), 581-605. https://doi.org/10.1080/1351847X.2014.997369.

[5] Chen, K., Chen, R., Zhang, X., & Zhu, M. (2016). Chinese stock market return predictability: Adaptive complete subset regressions. Asia-Pacific Journal of Financial Studies, 45(5), 779–804. https://doi.org/10.1111/AJFS.12152

[6] Xiao, H. (2023). Econometrics and Machine Learning Approach on Correlation in Stock Return between US Firms and Chinese Suppliers. Advances in Economics, Management and Political Sciences, 5,390-404. https://doi.org/10.54254/2754-1169/5/20220108

[7] Li, Y. (2024). Trading on trends: How the ordering of historical volume predicts Chinese stock returns? International Review of Financial Analysis, 95, 103518. https://doi.org/10.1016/j.irfa.2024.103518

[8] Chen, Y., & Huang, Z. (2021). Measuring the effects of investor attention on China's stock returns. Data Sci. Financ. Econ, 1, 327-344. https://doi.org/10.3934/dsfe.2021018

[9] Wang, J., Wu, K., Tong, G., & Chen, D. (2023). Nonlinearity in the cross-section of stock returns: Evidence from China. International Review of Economics & Finance, 85, 174-205.

[10] Song, G. (2020). The drivers of the great bull stock market of 2015 in China: Evidence and policy implications. Journal of Chinese Economic and Business Studies, 18(2), 161–181. https://doi.org/10.1080/14765284.2020.1796409

[11] Yang, Y. (2023, February). Research on the Influencing Factors of Stock Return Based on Factor Analysis and OLS Regression Analysis. In International Conference on Business and Policy Studies (pp. 747-755). Singapore: Springer Nature.