Research on Stock Picking Strategies Based on the Three-Factor Model: The case for the SSE STAR and A-share Markets

Junzhe Gu[1, †], Yingying Huang[2, †], Kang Liu[3, *, †], Sisi Liu[4, †]

[1]School of Finance, Chengdu Jincheng College, Sichuan, China

[2]Beijing Institute of Technology, Zhuhai, China; Bryant University, USA

[3] College of International Education, South China Agricultural University, Guangdong, China

[4] Accounting College, WUHAN TEXTILE UNIVERSITY, Hubei, China

*Corresponding author: kk20375@stu.scau.edu.cn

[†]These authors contributed equally

Abstract. Investors care about how to choose a portfolio that is worth investing in from the huge stock market in China because they want to obtain a high rate of return (RoR) by controlling the investment risk. Then, the three-factor model as a quantitative investment strategy can effectively satisfy investors' requirements by selecting influential factors highly related to stock prices. Moreover, current existing journals on three-factor stock selection strategies in China mostly use constituent stocks of CSI 300 as research data, and there are few studies on the SSE Composite Index, especially the SSE STAR Market which was officially opened in June 2019. Therefore, based on the demand of investors and the innovation of sample data, this paper selects the SSE STAR Market as the stock pool to construct a three-factor model and uses data of the A-share market to construct another model simultaneously for comparison. Finally, the research result of this paper demonstrates that the three-factor model can predict the RoR of the SSE STAR Market and A-share market well, so it can be used to predict the RoR in the actual investment process and then provide an objective suggestion for the investors in the real-world stock market.

Keywords: Quantitative investment, Three-factor model, STAR Market, A-share.

1. Introduction

With the rise of computer technology, quantitative investing, as a new investment model that has emerged, makes investment decisions more objective and efficient compared to traditional value investing approaches. It is a technical investment strategy in which the user uses advanced mathematical methods to construct models to quantitatively analyze historical data and obtain excess returns by means of programmed trading. The Securities Law of the People's Republic of China defines programmed trading as the act of automatically generating or placing trade orders for trading through a computer program (https://gkml.samr.gov.cn/nsjg/bgt/202106/t20210610_330492.html).

According to He, the transactions of financial-oriented quantitative trading have reached about 70% in the U.S. and European institutions. At present, this investment mode is developing rapidly in China. Then, an upsurge of quantitative investment appears in the Chinese financial market, and the scale of the quantitative trading industry exceeded one trillion-yuan which accounts for around 20% of the turnover in the A-share market (http://www.zqrb.cn/). Compared with traditional value investing, quantitative investment can enhance the effectiveness, science, and objectivity of investment, and it is the trend to rise in the domestic financial community.

As a quantitative investment strategy, the three-factor stock selection model is constructed by selecting specific three effective factors that have a significant impact on the direction of the stock market for quantitative analysis in order to explore their impact on stock returns. In terms of academic significance, the three-factor stock selection model covers multiple disciplines such as accounting, finance, and econometrics, which can provide a theoretical basis for quantitative analysis of the Chinese securities market. In terms of practical significance, the quantitative analysis of the three-factor model is conducive to helping investors scientifically understand the actual situation of the

Chinese stock market through objective data analysis and reducing the influence of subjective factors on investment decisions in traditional value investing. At the same time, the three-factor model can be continuously improved in the empirical test, which can better promote the development of the investment industry.

Referring to the characteristics of the Chinese stock market, this paper constructs the three-factor stock selection models using the data of SSE STAR Market and A-shares and excluding the st, st* and delisted stocks during the experiment period. First of all, the samples were analyzed by descriptive statistics to extract the basic sample characteristics such as maximum, minimum, and average value, then ADF smoothness test and correlation analysis were performed, and finally OLS linear model regression was constructed. The results of the test analysis show that the three-factor model of SSE STAR Market and A-share constructed in this paper has significant overall effect and good predictive ability. It is a stock selection strategy composed of three factors and suitable for predicting the situation of Chinese stock market, which can provide reference for investors' stock selection and investment.

The rest of this paper is organized as follows. Section 2 shows the literature review while section 3 depicts the data and introduces the principle of the three-factor model. Then, section 4 analyzes two types of models and discusses the results and section 5 draws a conclusion.

2. Literature Review

The capital asset pricing model (CAPM), which marks the derivation of asset pricing theory, was developed by Sharpe and Lintner based on the modern portfolio theory (MPT) of Markowitz. However, CAPM is a highly theoretical model. Specifically, under this theory all the investors are assumed to be completely rational in the market and the capital market should be perfectly efficient. In other words, Sharpe-Lintner CAPM is based on unrealistic assumptions [1].

In 1976, Ross developed the arbitrage pricing theory (APT) which mentioned the opportunity of risk-free arbitrage exists as long as the stock market is out of balance. Also, the theory believed the excess return of an individual stock or asset portfolio is influenced not only by the systematic factor of market portfolio return, but also by other factors that are highly correlated with market portfolio risk, and namely a common set of factors make an influence on the return on an asset portfolio [2].

With the develop of the theory of quantitative analysis, the CAPM model gradually losses its explanatory power as many financial anomalies cannot be demonstrated by it. Therefore, the validity of the CAPM model were questioned by scholars, for example, the excess return on assets could not be explained by the CAPM model. Fama and French pointed out that firm characteristics such as size, book-to-market (B/M) equity and long and short-term past returns affect average stock returns [3]. However, this theory was not explained by the CAPM model, and they argued that pricing factors in asset pricing models can be found in market anomalies that are not explained by the CAPM model and based on this they proposed the classical three-factor pricing model.

The theoretical system of the three-factor model was developed on the basement of CAPM, APT, and other financial tools. As one of the essential models in quantitative investment, the three-factor model requires the selection of effective factors to predict future returns. In other words, obtaining the excess return is the ultimate purpose. Stock return is the key variable to measure stock risk and is also significant for the stock pricing process. Moreover, the rate of return (RoR) is influenced by many factors and therefore the research on it promoted the development of theoretical system of the multi-factor model [4].

Also, many studies on asset pricing theory in the domestic literature exist. For example, Fan & Yu [5], Yang & Chen [6], and other economic scholars conducted an empirical analysis on the Fama-French three-factor model. Also, Feng & Liu mentioned that the explanator ability of the three-factor model works well when it faces small-cap stocks, but it needs to be enhanced when it comes to large-cap stocks [7].

Sun, Li, and Han proposed that the correlation between variables and factors affecting the stock market is constantly changing [8]. Therefore, the selection of indicators for constructing the model needs to be based on the specific situation of the stock market in different periods and the changes of specific data disclosed. Furthermore, Jiang believed that the three-factor model can reflect cross-sectional variations in the average returns of asset portfolios with different book-to-market (B/M) ratio and the company size [9].

To conclude, the choice of various components will cause the same strategy's impacts to change; other stock selection procedures will generate alternative returns on investment. As a result, the strategy of stock selection is a question worthy of research. Besides, the present research about the application of the three-factor model in China's stock market is still in the stage of development. Especially for the SSE STAR Market, academic journals about it are relatively rare. Therefore, this paper constructs a three-factor model using the data of the SSE STAR Market and analyzes the final result based on current comprehensive theories.

3. Data and research methods

3.1 Selection and Processing of Data

To obtain research results that are more in line with the current status of the stock market in China, this paper chooses the data of the SSE STAR Market from July 1, 2020, to June 24, 2022. In other words, this group of stocks is used to represent the changes in the whole STAR Market. Also, the weekly data of A-shares in the same period is used to construct another set of three-factor models, as a reference, to observe the predictive ability of the model constructed based on STAR Market's data. Besides, all calculations are done using the circulation market value.

Furthermore, two aspects are considered during the data chosen process: first, the A-share market has been dominated by retail investors whose investment decisions are often not fully rational, plus the preliminary supervision system was not sound enough, so the phenomenon such as market manipulation and insider trading show out with high frequency, which distort the stock returns to a certain extent. Then, the threshold for entry into the STAR Market is relatively high, which means the investors are required to have certain financial strength and investment experience. In other words, it exists differences between them and the investors of the A-share market in investment philosophy and investment style. During the chosen period, both the A-shares and STAR Market have significantly changed in the degree of institutionalization of participants and improvement of the regulatory system.

The sample data are reliable because of the same purpose of data study and source of data root. However, the size of the sample is broad, and it has a large fluctuation inside. Also, it exists anomalies such as abnormal collapse on March 16, 2022. Therefore, this paper preprocesses the data of the SSE STAR Market by truncating the tails and outliers and uses the interpolation method to deal with the missing values that the missing values are namely replaced with the average value. It makes the data more stable and convenient for constructing prediction models.

For data sources, the closing indices of STAR Market come from the Tongdaxin Financial Data Terminal while the rate of excess return, market risk-free rate, and the other factor including market risk, B/M, and market capitalization are obtained from China Stock Market & Accounting Research Database (CSMAR). Besides, the index data are first-hand; factors and RoR are second-hand. Moreover, this paper selects the required indicators for both A-share and STAR Market from CSMAR and conducts descriptive statistics based on the data of the two sectors respectively. By screening, outliers, missing values, and extreme values are eliminated. Finally, 103 sample data are retained for analysis; the data processing software is spss27, SPSSPRO, and Excel.

The below tables show the overall result of descriptive statistics on the data of the STAR Market and the A-shares market. Specifically, they analyze the value of the risk premium1, SMB1 (Small Minus Big), HML1 (High Minus Low) and excess return respectively, including their sample size,

mean value, standard deviation (Std Dev), etc. It is used to observe the overall situation of the selected data.

Table 1. Descriptive statistics of samples of STAR Market

Variable	Minimum	Maximum	Average	Standard deviation	Kurtosis	Coefficient of variation
Risk Premium1	-0.054	0.055	-0.003	0.027	-0.752	-10.086
SMB1	-0.036	0.036	-0.001	0.017	-0.539	-23.136
HML1	-0.036	0.041	0	0.019	-0.527	45.199
Excess return	-0.083	0.022	-0.03	0.027	-0.653	-0.894

As above the table 1 confirms, the mean value of risk premium, size factor (SMB1), and book-to-market ratio (HML1) is -0.003, -0.001, and 0 respectively. It means no substantial scale influence and significant value premium exists in the STAR Market from week 25 to week 27 of 2022.

Additionally, the Std Dev of the B/M ratio factor and size factor are 0.017 and 0.019 respectively. Both of them are less than the risk premium of 0.027, proving that the selection of stocks of large-capitalization companies and stocks of low B/M ratio companies in the STAR Market generates low volatility of excess return.

Table 2. Descriptive statistics of samples of A-Share

Variable	Minimum	Maximum	Average	Standard deviation	Kurtosis	Coefficient of variation
Risk Premium1	-0.051	0.078	0.002	0.023	0.871	13.578
SMB1	-0.065	0.039	0.002	0.019	0.953	12.422
HML1	-0.034	0.046	0	0.016	-0.027	39.294
Excess return	-0.079	0.047	-0.027	0.022	0.855	-0.83

Table 2 above demonstrates that the mean values of the A-shares market are similar to that of the STAR Market, which means no significant scale effect or value premium exists in this market. Specifically, the mean values of the risk premium and SMB1 are the same, which is 0.002 respectively. Besides, the average of HML1 is 0. From the perspective of average excess return, the return of large-cap stocks is higher than small-cap stocks.

Moreover, the Std Dev of the HML1 is 0.016, which is smaller than the standard deviation of the risk premium and SMB1. It demonstrates in the A-share market, the excess return of the low B/M ratio stocks is less volatile than that of the large-scale stocks.

3.2 Principle of the Three-factor Model

In 1993, Fama and French formalized the three-factor model which explained the excess return phenomenon from a risk-return perspective. Furthermore, their research in 1996 further described that all the anomalies, except for the medium-term return momentum, can be expounded by the three-factor model [10].

According to Fama and French, for a portfolio of assets, its excess return can be interpreted by three risk factors, namely the market-related risk factor -- market excess return (Rm-Rf), the size-related risk factor (SMB), and the market value-related risk factor (HML).

$$Rit - Rft = \alpha i + bi\,(Rmt - Rft) + si\,SMBt + hi\,HMLt + \varepsilon it \quad (1)$$

Specifically, $Rit - Rft$ stands for the expected excess return which is equal to the total return of portfolio i minus the risk-free rate of return at period t. Also, $Rmt - Rft$ means the total market portfolio return mins the return of risk-free rate when the time period is t. Moreover, SMBt shows the difference between the return on the small firm portfolio and the large firm portfolio; HMLt equals to the return on the high B/M ratio portfolio and the return on the low B/M ratio portfolio.

To sum up, the three-factor model completely explains lots of market anomalies from the aspect of the risk premium and it describes the new risks other than systemic risk. By the empirical test, the model ultimately condenses many influencing factors into the size factor which is the total value of shares issued by a listed company at market prices and the B/M ratio factor. In addition, if the B/M ratio and the size factor are proxy variables of risks, the three-factor model constructs a more comprehensive and closer description of risk than the ideal state of the CAPM.

4. Model analysis and discussion of results

Returns are the core of quantitative investing, and what the model is trying to predict is the movement of the returns brought by various factors. In this paper, past stock market data is used to predict the future trend of the stock index, which is similar to technical indicators in investment. Moreover, the past weekly k-chart excess rate of return is treated as the dependent variable while the factors including risk premium, market size, and B/M ratio are considered as independent variables to explore the relationship between the RoR and the selected factors.

4.1 Data Model of STAR Market

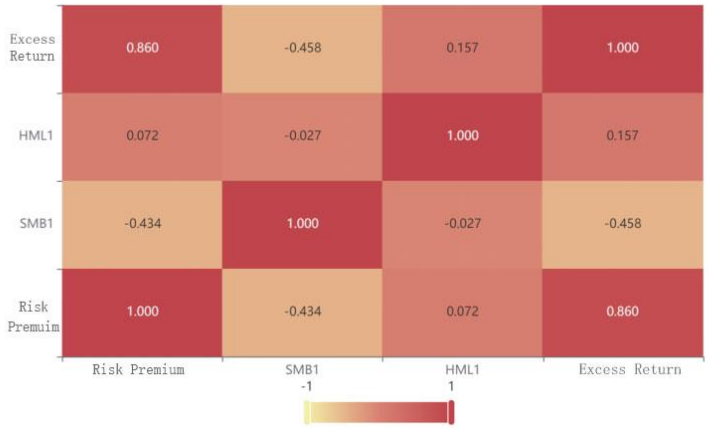

Figure 1 Thermal map of STAR Market

The above figure shows the value of the correlation coefficient in the form of a heat map, mainly by color shades to indicate the magnitude of the value.

Next, correlation analysis with regard to the above indicators is performed to test whether there is a significant correlation between the above independent and dependent variables. According to the heat map of the STAR Market, the absolute value of correlation between the Risk premium1 and HML1 in the STAR Market is 0.072, which shows a weak correlation. Besides, the correlation coefficient between Risk premium1 and SMB1 is slightly larger than the aforementioned one that the absolute value exceeds 0.3. However, it also reflects a weak correlation because the magnitude of the exceedance is small. Moreover, the correlation between SMB1 and HML1 is weak because its absolute value is 0.027. To conclude, a linear substitution relationship comes into no existence; Risk Premium1 is positively correlated with the excess return of STAR Market, while in turn SMB1 and HML1 are negatively correlated with the excess return.

Then, this paper performs ADF smoothness tests for each of the three factors (Risk Premium1, SMB1, and HML1) to determine the smoothness of the time series by whether the p-value is less than the original hypothesis of 0.05.

Table 3. ADF test of STAR Market

	Risk premium1	SMB1	HML1	Excess Return
ADF statistics	-8.693	-9.737	-9.426	-7.984
Smoothness	Smooth	Smooth	Smooth	Smooth

The test results are shown in table 3. Specifically, the three independent variable factors' ADF statistics are completely less than the threshold values at various confidence levels, so the original time series does not contain a unit root. It indicates that the time series of the independent variables from STAR Market's data is smooth. In other words, the linear regression analysis conducted in the following paper does not have the problem of pseudo-regression.

Table 4. STAR Market's linear regression analysis

Linear regression analysis results n=103						
	Non-standardized coefficient		t	VIF	R^2	F
	B	Standard error				
Intercept	-0.028***	0.001	-21.001	-	0.757	F=103.014***
Risk Premium1	0.807***	0.055	14.654	1.237		
SMB1	-0.171*	0.089	-1.924	1.232		
HML1	0.135*	0.07	1.941	1.005		
Dependent variable: Excess return (STAR Market)						

Note: ***, **, * represent 1%, 5%, 10% significance levels, respectively

The above table demonstrates that the standardized coefficients of the three factors, namely Riskpremium1, SMB1, and HML1 are 0.807, -0.106, and 0.096 respectively. Both Risk Premium (RP) and B/M ratio (HML1), have a positive relationship with the excess return, whereas a negative correlation exists between the size factor (SMB1) and the excess return. Besides, all three factors passed the significance test at the confidence level of 0.1. Hence, the information will be considered that the three factors have the ability to predict the excess return of STAR Market. And the goodness of fit R^2 is 0.757, which states clearly that the regression relationship can explain 75.7% of the variance of the dependent variable and the model has a good fitting degree.

In addition, the intercept term delegates that the excess return cannot be interpreted by the model, because the three-factor model can have adequate interpretive force, only when the intercept term equals 0. From the empirical results, it can be seen that the value of the intercept term is close to 0. The estimation of the coefficient of risk premium is in line with our assumption that the higher the risk, the higher the return. Moreover, the coefficients of risk premium all fluctuate around 1 and are highly significant, which illustrates that the market risk has a relatively strong ability to explain the cross-sectional variation of stock returns.

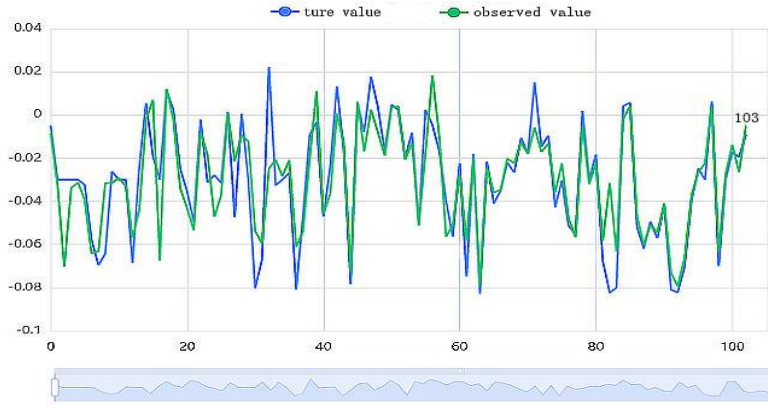

Figure 2 Fitting effect graph of STAR Market

Figure 2 depicts the raw data plot, model fitted values, and model predicted values for this model. The coefficients of the market excess returns all fluctuate around 1 and are greatly significant, which also testifies that the risk of market has a relatively strong capability to explain the cross-sectional variation in stock returns.

4.2 Data Model of A-share

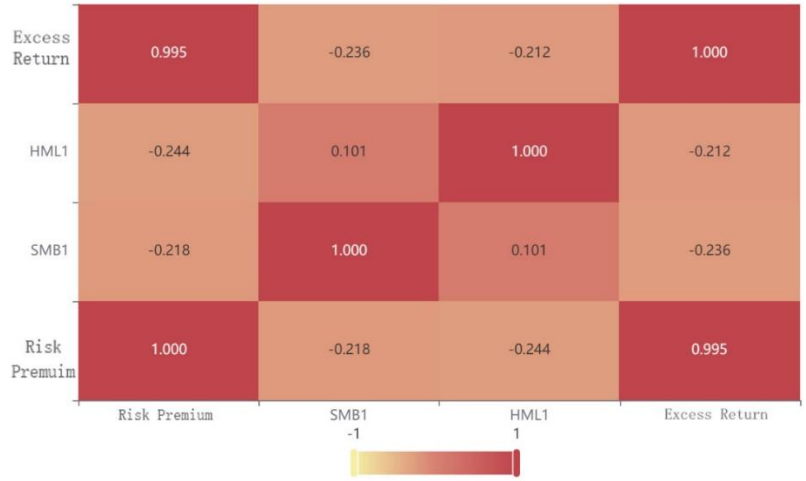

Figure 3 Thermal map of A-share

This paper repeats the steps of the STAR Market of the data study to test whether there is a significant correlation between the above independent and dependent variables. Judging from the heat map of A shares, the absolute value of correlation between Riskpremium1 and SMB1 equals 0.218 smaller than that of correlation between Riskpremium1 and HML1, 0.244. Since the absolute value is less than 0.3, these two groups are weakly correlated by comparison. Also, because of the absolute value of correlation between SMB1 and HML1 which is equal to 0.101, it shows a very weak correlated. Therefore, it is similar to the case of the STAR Market that the linear substitution relationship does not exist. The conclusions are shown in the figure 3 above.

In summary, RiskPremium1 is positively correlated with the excess return of STAR Market, while in turn SMB1 and HML1 are negatively correlated with the excess return. This finding is similar to that of the correlation analysis of the STAR Market data.

Table 5. ADF test of A-share

	Riskpremium1	SMB1	HML1	Excess Return (A shares)
ADF statistics	-10.633	-8.649	-4.175	-10.653
Stability	Stable	Stable	Stable	Stable

As table 5 shows, it is available to get the same conclusion during the analysis of STAR Market, so there is no unit root in the original time series, which indicates that the time series of the data of the independent variables of A shares is smooth, and the study results do not suffer from pseudo-regression.

Table 6. Linear regression analysis results of A-share

Results of linear regression analysis n=103						
	Non-standardized coefficients		t	VIF	R^2	F
	B	Standard error				
Constants	-0.028***	0	-143.199	-	0.992	F=4184.497***
RiskPremium1	0.977***	0.009	106.743	1.108		
SMB1	-0.025**	0.011	-2.408	1.053		
HML1	0.046***	0.012	3.746	1.066		
Dependent variable: Excess return (A shares)						

Note: ***, ** and * represent 1%, 5% and 10% significance levels respectively

Specifically, the analysis of the results from the F-test shows that the model constructed is significant overall and rejects the original hypothesis that the regression coefficient is 0. Therefore, the model basically satisfies the requirements for the performance of variable cointegration, with VIF all less than 10. Together with the analysis results, the goodness of fit R^2 is at 0.992, implying that the variation of 99.2% in the dependent variable can be understood through the regression relationship, and namely the fitting effect is perfect.

From the empirical results, it depicts that the value of the intercept term, -0.028, is close to 0, which means that the three-factor model constructed in this paper and based on the A-share market data can fully explicate the variation of cross-sectional stock returns in the Chinese market.

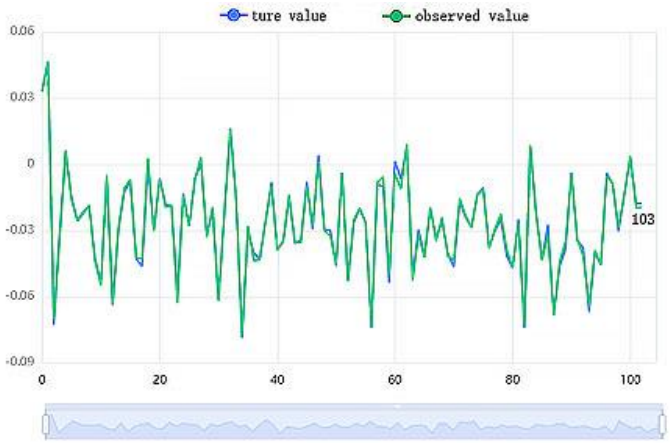

Figure 4 Fitting effect graph of A-share

The judgment method is similar to that of the STAR Market. Specifically, the graph reflects the degree of fit between the predicted value and the true value. Moreover, the predicted value is highly fitted to the true value, and the model is well fitted.

4.3 Comparative Analysis of Models on STAR Market and A-share Market

In this paper, we randomly perform correlation test and find that the risk premium and excess return are positive and significant for both models. The SMB1 of STAR Market is significantly negative, which is consistent with the SMB1 in A-shares. The HML1 in A-shares is significantly negative with excess return (A-shares).

The equation for the linear model of STAR Market:

$$y = -0.028 + 0.807 * \text{RiskPremium1} + (-0.171) * \text{SMB1} + 0.13 * \text{HML1} \quad (2)$$

The equation for linear model of A-share:

$$y = -0.028 + 0.977 * \text{RiskPremium1} + (-0.025) * \text{SMB1} + 0.046 * \text{HML1} \quad (3)$$

Then, this paper compares the OLS regression models of STAR Market with A-share and finds that the intercept term of both sets of equations is the same value, -0.028. It shows both models have a good capability to expound the excess returns that cannot be explained by CAPM.

Moreover, the coefficients of the risk premiums for both sets of equations are similar, i.e., the coefficients of both are very close to 1 and are highly significant, indicating that market risk has a strong explanation for the cross-sectional variation in stock returns. In between two markets, SMB1 is negatively correlated in the STAR Market, but it is negatively correlated in the A-share and through the significance inspection.

In addition, the coefficient of HML1 of the linear model equation for the STAR Market version is 0.135 greater than that of 0.046 for the A-share version, which implies that the return of STAR

Market will be larger than that of A-share on the B/M factor. In addition, the model fits better for both sectors, specifically, the R^2 of 0.757 for the STAR Market and 0.992 for the A-share market, revealing that the interpretive force of the three-factor model is strong in both markets.

5. Conclusion

This paper chooses the market risk factor, market capitalization factor, and B/M factor from the A-share and STAR Market index and excess return data to construct a regression model, referring to the Fama-French's theory of the three-factor model. Then, by comparing the two groups of models, it concludes that the three-factor model can predict the return of the STAR Market well.

As for data, extreme values and outliers in the data are removed and missing values are filled with mean values. As for the model, this paper uses OLS linear model for regression. As for the test, the statistical methods such as ADF smoothness test and correlation analysis are used to compare A-share and STAR Market. To conclude, the market risk and B/M factor are positive factors, while the market capitalization factor is a negative factor.

The significance and fit are very high in both A-shares and STAR Market, and the empirical results of this group of factors are good. The three-factor model meets expectations and fits with China's stock market and can be used for STAR Market return forecasting.

Reference

[1] E.F. Fama, K.R. French, The capital Asset pricing model: theory and evidence, Journal of Economic perspectives, (2004): 25-46.

[2] S.A. Ross. The arbitrage theory of capital asset pricing, Journal of Economic Theory, 13(1976): 341-360.

[3] E.F. Fama, K.R. French, Common risk factors in the returns on stocks and bonds. Journal of Financial Economics, 33, (1993): 3-56.

[4] X. Yang, Z.H. Chen. An empirical study of a three-factor asset pricing model for the Chinese stock market. Quantitative Economic and Technical Economics Research, (2003): 137-141.

[5] L.Z. Fan, S.D. Yu. A three-factor model for the Chinese stock market. Journal of Systems Engineering, 17(2002): 537-546.

[6] K. Chen, H.Y. Yan. The application of three-factor model stock selection strategy in China's securities market, Cooperative Economics and Technology, (2020): 50-51.

[7] S.J. Feng, Z. Liu. An empirical study of a three-factor pricing model for the Chinese stock market. Journal of Economic Research, (2018): 145-148.

[8] Y.D. Sun, H.H. Li, M.X. Han, Application of multi-factor model to stock selection in Chinese stock market. Modern Marketing: Business Edition, (2020): 242-243.

[9] M.H. Jiang, Analysis of SSE 50 constituents based on Fama-French three-factor model. Marketing World (2021): 98-100.

[10] X. Yang, X.Z. Wang. Empirical Study on Performance Evaluation Factor Models of Chinese Mutual Funds. Systems Engineering Theory Practice, 23(10) (2003): 30-35.