Portfolio Establishment Based on Fama-French Five-Factor Model in China Stock Market

Zihui Gong, Qianqian Shi, Guangjie Xu, and Yuzhi Zhou*

***Abstract*—In the field of modern finance, investors like to use the ACPM model to analyze their portfolios to reduce risks and maximize returns. And the main purpose of our investigation is to choose the six stocks and use the R-studio to analyze the data to see whether the five-factor model can be applied well in China stock market. We start our investigation by collecting data and setting up multiple linear regression models. Then we observe the correlations between the five factors and the excess returns of different stocks and test if all the values of the population parameters and some certain parameters are equal to 0. Finally, we test if multicollinearity existed. We can conclude from the analysis that HML is the most significant factor in all of the portfolios. Besides, the factor CMA has the least significance in portfolios 1 and 2 and the factor SMB is the least significant factor in the rest of the regressions of portfolios. Based on the result, we find that the five-factor model is also applicable in China stock market. So Chinese stock investors can use the five-factor model to help them achieve better investment returns.**

***Index Terms*—ACPM, five-factor model, portfolios, SMB**

I. Introduction

To minimize the risk in the investment, investors will be more likely to choose a diversified portfolio that is applied to the CAMP model. In the background of the CAPM model, the capital market is supposed to be the perfect condition in which investors become price-taker, and there are no transaction costs or information costs (Fama and MacBeth, 1973). Harry Markowitz wants to find the logic of the CAPM, so it assumes all investors are risk-averse and try to maximize the utility. The results show all investors will choose the mean variance-efficient portfolios, which can minimize the variance of the portfolio return to give the expected return or maximize the expected return to give the variance (Fama and French, 2004).

The *g* represents the borrowing with some risky portfolios and the risk-free lending. The Rf represents all investments in the risk-free asset. Tobin found that efficient portfolios are always combinate with the risk-free asset. After that, they also assume the market portfolios, as the asset market to clear, the M needs to be on the minimum variance frontier (Fama and French, 2004). According to A (cited in Logue and Merville), the equation can be shown in the $E(Rj)= Rf + [E(Rm) - Rf]$. Where the $E(Rj)$ is the real investment in security, and the Rf is the risk-free interest rate, the $E(Rm)$is the expected return of the market index, *Bj* could measure the ratio of the covariance about the return in the individual security (Weston, 1973). To conclude, the CAPM model showed how the risk-averse construct the best portfolio.

As the Corporate finance institute stated the option pricing model is be classified as a mathematical model which uses some variable to calculate the value of the option. This could help us provide the fair value option (Corporate Finance Institute, 2021). There are two main options that investors can trade on stocks are call options and put options, and they are normally classified by the different types of contracts. While the put option always gives the right to sell the share, the call option is giving the right to buy the share (Brealey *et al*., 2020). According to Copeland Weston Shastri, Black, and Scholes state that the value of the call options in the firms is equal to the equity in the levered firm. As equity is the call option of the firm, we can know this formula $S+P=B+C$ (Copeland *et al*., 2013).

However, European options can only be exercised at the maturity date, and American options can be exercised between the purchase and maturity date (Corporate Finance Institute, 2021). Whereas they have binominal option pricing model and Black-Scholes model both can price the options. Firstly, we talk about the Binominal option pricing model, which is the simplest method to measure the options due to the assumption of the perfect efficient market. Therefore, the model can be used to price at each of the specified times. The Obaidullah Jan also claimed the option pricing model is quite different from the B/S model, it is more suitable for valuing independent options (Obadidullah Jan, ACA, CFA, 2019). Another model is the Black-Scholes Model, the Black-Scholes models are mainly developed in the European options stocks. Regarding the James d. Macbeth and Larry J. Merville's statement, the predictions of the B-S model price were related to the deviation, and this could be shown as the formula below (Macbeth and Merville, 1979):

While the *C* represents the call option about the market value; and *S* is the underlying security's price; the *X* is the stock exercise price; *r* is the time to expiration; whereas *r* is the constant interest rate among this period; As Blake, D. stated that there are also have some factors that affect the options price, one is the premium on a European call option, and the other is the spot price of the underlying security, and the exercise price and time to expiry (Blake, 1989). In conclusion, the Black-Scholes model is trying to determine the value of the option, because they assume it has a riskless strategy. Whereas in the actual market, the option prices are always determining the supply and demand (Figlewski, 1989).

After researching of the basic information about the two models, the main purpose of our investigation is to choose the six stocks and use the R-studio to analyse the data to see whether the five-factor model can be applied in China stock market. In this process, we will regard the five factors in

Manuscript received May 7, 2022; revised July 14, 2022; September 5, 2022.

Zihui Gong, Qianqian Shi, and Yuzhi Zhou are with Lancaster University, UK.

Guangjie Xu is with Hohai University, China.

*Correspondence: y.zhou28@lancaster.ac.uk

doi:10.18178/joebm.2023.11.1.731

independent variables and the expected return of the six stocks independent variables.

II. Data

We use the five factors (MKR, SMB, HML, RMW, CMA) of the five-factor model as the independent variables ($x_{i,}$ i=1,2,3,4,5) of regression models. Five factors are calculated by the traded market value-weighted method that could reflect the influence of changes in the values of different company shares on the entire market.

We selected the risk returns of six different stocks on a share market. The stock code is 600036.sh, 601318.sh, 600519.sh, 600760.sh, 300059.sh, 600276.sh respectively. all the sic stocks have been on the market for a long time (more than 10 years). then we chose Chinese ten-year treasury yields as the risk-free return, and we subtracted the risk-free rate of return from the risk-free rate of returns of the six stocks respectively to obtain the excess returns of the six stocks, which were taken as the dependent variables (j=1,2,3,4,5,6) of the regression model.

III. Methods

A. Set up, Multiple Linear Regression Models

$$R_{\mathrm{i}} = \mathrm{a}_i + b_i + R_M + s_i + E(SMB) + \mathrm{h}_{\mathrm{i}} + E(HML) + r_i + E(RMW) + c_i + E(CMA) + e_i \tag{1}$$

This model represents the excess return of stock I relative to the risk-free return. And shows the excess return of the market return relative to the risk-free return. E(SMB) is the value of the expected excess return of the return of small market value companies relative to the return of large market value companies. E(HML) is the expected excess return of high B / M company stock compared with low B / M company stock is the regression residual term. E(RMW) is the difference between the returns of high / low-profit stock portfolios, while E(CMA) is the difference between the returns of low/high reinvestment ratio stock portfolios.

B. Observe the Correlations between the Five Factors and the Excess Returns of Different Stocks

We used the R^2 and adjusted R^2 to see the fit of the regression models to the sample observations, more precisely, the correlations between all independent variables(factors) and the dependent variables (excess returns).

According to the calculation formula of R^2, a regression model completely fits the sample observations if and only if the R^2 is equal to 1. The closer the R^2 is to 1, the better the regression model fits the sample observations. Since R^2 increases when the number of independent variables increases, we need a statistic that not only simply reflects correlations between all independent variables and the dependent variables but also reflects the effect of the number of independent variables included in a regression model, and that is exactly the adjusted R square.

C. Test if All the Values of the Population Parameters Are Equal to 0. Set up, Multiple Linear Regression Models

Normally we could not observe the value of the population parameters and we wanted to test if they were equal to 0 because the regression makes sense only if the parameters are not all equal to 0.

The principle of the hypothesis test is that if a small probability event occurs, we have reason to reject the null hypothesis, so we can set the null hypothesis from the opposite side of our expectations to verify whether the null hypothesis is a small probability event.

Firstly, we tested if all the values of the parameters were equal to 0. Here is the null hypothesis.

1) set: H_0: β_{1j}=β_{2j}=β_{3j}=β_{4j}=β_{5j}=0(j=1,2,3,4,5,6)

We used the p-value to see if we can reject the null hypothesis. The p-value is the probability of obtaining a result at least as extreme as the one that was observed, given that the null hypothesis is true. The smaller the p-value is, the more we can reject the null hypothesis. Just imagine a case that we have run a lot of tests and the probability of getting our observation under the null hypothesis was very small, and the p-value would only be smaller. But such a small probability event happened that we could no longer believe the null hypothesis, so we rejected the null hypothesis and that meant not all the parameters were equal to 0. To make our regression have statistical significance, we would expect to see a very low p-value.

We also used the quantile of the distribution F (confidence level, k, $n-k-1$) to compare with the F statistics (k is the number of the arguments and n is the sample size). The function qf() gave the quantile of the F distribution of our samples.

If the F statistics are much greater than the quantile of the F distribution, we can reject the null hypothesis and say not all the values of the parameters are equal to 0.

D. Test if the Value of a Certain Parameter is Equal to 0

Both the R^2 test and F test consider all independent variables as a whole to test their correlation with dependent variable Y and the significance of the regression. However, for multiple regression models, the overall significance of the regression model does not mean that each independent variable (the factor) has a significant impact on the dependent variable (the excess return). If a factor is not significant, it should be removed from the regression model. Therefore, significant tests must be performed for each factor.

After testing if all the values of the parameters were equal to 0 and getting the answer “no”, we tested if the value of a certain parameter was equal to 0. Here is the null hypothesis.

2) set: H_0: β_{ij}=0(i=1,2,3,4,5; j=1,2,3,4,5,6)

We used function lm () and summary() which automatically give the t-stat and the p-value of the test.

We still used the p-value to see if we could reject the null hypothesis.

If we get the p-value of a certain factor that is relatively high, then we cannot reject the null hypothesis and that means this factor is not significant.

The above four steps are the main steps of our research. However, except for these main steps, there was still a test that needed to be conducted. We tested if multicollinearity existed.

E. Test If Multicollinearity Existed

When multicollinearity exists, there is at least one independent variable in the regression model, which is a linear combination of other independent variables in the same regression model. The regression model hardly estimates accurately when multicollinearity exists.

We used the function VIF() which gives the variance inflation factors (VIF) of the regression models. If VIF<10, there is no multicollinearity; If VIF>100, the multicollinearity is so severe that some adjustment may be needed.

IV. Result of Regression Analysis

We test the regression model established according to the above method and get the following regression results (Table I).

TABLE I: Results of the Regressions on Six Portfolios and Candidate Factors Dependent Variable

	China Merchants Bank	Ping an Insurance	Kweichow Moutai	SACC	Easy money	Hengrui Medicine
	1	2	3	4	5	6
Rf	1.629**	1.689***	1.345**	1.552**	2.319***	1.075*
	(0.344, 2.914)	(0.426, 2.951)	(0.069, 2.621)	(0.193, 2.910)	(0.943, 3.694)	(−0.184, 2.334)
MKT	0.594	0.398	0.552	0.227	−0.081	0.873
	(−2.736, 3.924)	(−2.873, 3.669)	(−2.753, 3.858)	(−3.292, 3.746)	(−3.645, 3.484)	(−2.389, 4.134)
SMB	−1.93	−2.365	−3.749**	−3.932**	−3.445*	−3.562**
	(−5.278, 1.418)	(−5.654, 0.924)	(−7.073, −0.425)	(−7.471, −0.394)	(−7.029, 0.139)	(−6.842, −0.283)
HML	5.454**	5.405**	6.642**	4.322	3.491	5.894**
	(0.206, 10.702)	(0.249, 10.560)	(1.432, 11.852)	(−1.225, 9.869)	(−2.126, 9.109)	(0.754, 11.034)
RMW	0.915	0.786	1.536	2.92	0.876	0.802
	(−3.810, 5.641)	(−3.857, 5.428)	(−3.155, 6.228)	(−2.075, 7.915)	(−4.183, 5.935)	(−3.827, 5.431)
CMA	−3.462***	−3.467***	−3.455***	−3.444***	−3.434***	−3.458***
	(−3.538, −3.386)	(−3.541, −3.392)	(−3.530, −3.379)	(−3.525, −3.364)	(−3.515, −3.353)	(−3.532, −3.383)
Observations	138	138	138	138	138	138
R^2	0.083	0.097	0.119	0.098	0.13	0.11
Adjusted R^2	0.048	0.063	0.085	0.064	0.097	0.077
Residual Std. Error (df = 132)	0.431	0.423	0.428	0.456	0.461	0.422
F Statistic (df = 5; 132)	2.382**	2.829**	3.556***	2.867**	3.952***	3.277***
Note:						$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

For all the portfolios, HML is the most significant factor because the possible coefficient range of HML is higher than other factors' range. The factor CMA has the least significance in portfolios 1 and 2. The factor SMB is the least significant factor in the rest of the regressions of portfolios. In our model, portfolio 5 has the highest R square value, which is 0.13, and it follows by port3 (0.119). The R square value of portfolio 6 ranked third. Portfolio one has the lowest R square value. Hence portfolio 5 best fits over the regression model compared to other factors. No portfolio performs badly in the regression as there is no negative adjusted R square value.

From the regression table shown above, it can be concluded that we can reject the null hypothesis as the p-value for our F statistic is all smaller than 0.05, the population does not equal to 0 is not a coincidence.

When comparing F distribution and F Statistic, it could be found out that all the F statistics of portfolios are larger than the distribution value at least 0.5, except for portfolio1.

In the regression of portfolios 1 and 2, factor 2(SMB), factor 3(HML) & factor 5(CMA) are not very significant as they are larger than 0.05, the null hypothesis cannot be rejected based on these three factors. Factors 2 and 5 are not significant enough to reject the null hypothesis in the regression of portfolio 3. The t statistic of the portfolio 4 regression model indicates that factors 2,4 and 5 cannot reject the null hypothesis All the factors are not significant in regression of portfolio 5, except for factor one (MKT). Only factors 3 and 4 are relatively significant in the regression of portfolio 4.

The sample size of the five factors is 138, which ranged from March 2010 to August 2021. MKT has the largest range while CMA has the smallest range. All the standard deviations of the factors are very small, not even reaching 0.01, it could be said that the factors do not have many fluctuations (Table II and Fig. 1).

TABLE II: Basic Statistical Indicators

	MKT	SMB	HML	RMW	CMA
n	138	138	138	138	138
mean	0.004	0.007	−0.001	0.0003	0.0003
sd	0.062	0.047	0.037	0.027	0.02
median	0.005	0.005	−0.0001	0.0003	0.001
trimmed	0.003	0.006	−0.001	−0.0002	0.0002
mad	0.047	0.038	0.028	0.023	0.02
min	−0.243	−0.221	−0.144	−0.082	−0.047
max	0.175	0.229	0.16	0.1	0.061
range	0.418	0.45	0.303	0.182	0.109
skew	−0.079	0.035	0.162	0.212	0.154
kurtosis	1.9	5.793	3.337	1.387	0.338
se	0.005	0.004	0.003	0.002	0.002

The distribution of RMW is not as close to a normal distribution shape as other factors do, but it can still be considered symmetric as the skewness is only 0.212 (Fig. 2).

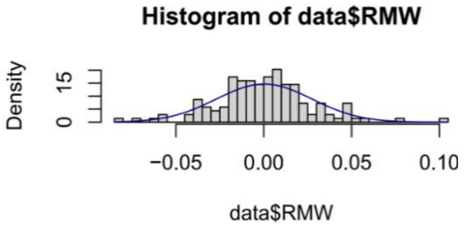

Fig. 1. Histogram of Data$RMW.

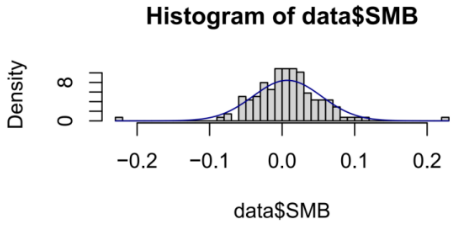

Fig. 2. Histogram of Data$SMB.

Although the shape of the distribution of SMB is the closest to the symmetrical among all the factors. However, its Kurtosis value is up to 5.7, almost 3 times of the Mesokurtic (kurtosis = 3). hence it shows a heavy tail on both sides (see Table III).

TABLE III: VARIANCE INFLATION FACTOR

Variance Inflation factor of the regression model				
MKT	SMB	HML	RMB	CMA
1.214	4.749	2.895	3.921	1.75

The correlation coefficient's values range between −1 and 1. The closer the coefficient to 1 or −1, the stronger the two variables are related.

From the visualization of a correlation matrix:

1) SMB and RMW have the strongest relationship between them (−0.79), they are negatively related.

2) SMB is also negatively related to HML (−0.72)

3) HML and RMW have a positive relationship but are not very strong, the same phenomenon can be found with CMA and RMW, they are negatively related to each other but not considered to be very strong.

From the method of variance inflation factor, we can see that all the value is below 10, we choose 10 as our critical value is because we have a large sample size. Hence, it can be concluded that there is no multicollinearity problem between factors.

V. CONCLUSION

In the research of modern finance, the research on asset pricing is one of the hottest topics. Many research results have been made by scholars such as Black, Scholes, and Harry Markowitz. The purpose of our study is to explore whether the five-factor model can be applied to China stock market. We chose six stocks that are considered as leading stocks in different industries in China stock market and used the R-studio to analyze the data to see whether the five-factor model can be applied in China stock market. Based on the analysis of data, we found that HML is the most significant factor in all the portfolios. Besides, the factor CMA has the least significance in portfolios 1 and 2 and the factor SMB is the least significant factor in the rest of the regressions of portfolios. And all the five factors do not have many fluctuations because all the standard deviations of the factors are very small and there is no multicollinearity problem between the five factors. Through the above analysis, we find that the five-factor model can be well applied to China's stock market. So Chinese stock investors can use the five-factor model to help them achieve better investment returns like minimizing the variance of the portfolio return to give the expected return or maximizing the expected return to give the variance. Although we have achieved good research results, there are still many deficiencies in our research. We only selected six leading stocks in different industries in the study so it may not be enough to represent the applicability of the five-factor model to the whole Chinese stock market. Besides, because the development time of China’s stock market is short, to select as many rises and fall data as possible for a long time, we only select large market capitalization stocks so there is no diversity in the selection of stocks. Our future research will increase the diversity of stock types as much as possible.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Guangjie Xu conducted the background research and wrote the summary, some part of the introduction conclusion. Zihui Gong explained the basics of the ACMP model in the introduction part. Qianqian Shi collected the research data and introduced the method part. Yuzhi Zhou explained the result of the regression analysis and examined the variance inflation factor.

REFERENCES

Blake, D. 1989. Option pricing models. ***Journal of the Institute of Actuaries,*** 116(3): 537﹣58.

Brealey, R., Myers, S., & Allen, F. 2020. ***Principles of Corporate Finance (***11th ed.): 513-515.

Copeland, T., Weston, J., & Shastri, K. 2013. ***Financial Theory and Corporate Policy.*** Pearson New International Edition.

Corporate Finance Institute. 2021. ***Option pricing models.*** Available: https://corporatefinanceinstitute.com/resources/knowledge/valuation/option-pricing-models/

Fama, E. F., & French, K. R. 2004. The capital asset pricing model: Theory and evidence. ***The Journal of Economic Perspectives,*** 18(3): 25-46.

Fama, E. F., & MacBeth, J. D. 1973. Risk, return, and equilibrium: empirical tests. ***Journal of Political Economy,*** 81(3): 607-36.

Figlewski, S. 1989. What does an option pricing model tell us about options prices? ***Financial Analysts Journal,*** 45(5), 12-15.

Macbeth, J. D., & Merville, L. J. 1979. An empirical examination of the black-scholes call option pricing model. ***The Journal of Finance,*** 34(5): 1173-1186.

Obadidullah Jan, ACA, CFA. 2019. Binomial option pricing model. Available: https://xplaind.com/552187/binomial-options-pricing-model

Weston, J. F. 1973. Investment decisions using the capital asset pricing model. ***Financial Management,*** 2(1): 25-33.