Research on Stock Return Prediction Based on Stepwise Regression-incremental Linear Regression Analysis

Sisi Wang

School of Economics and Management, Chongqing University of Posts and Telecommunications, Chongqing 400065, China.

1152758964@qq.com

Abstract

Stock return is an indicator reflecting the level of stock returns. The most important thing for investors to purchase stocks or bonds is how much income they can get. The measure of a securities investment income is expressed in terms of yield. Whether stock market returns can be effectively predicted is one of the hot issues in current financial research. Based on this purpose, this paper selects the factors related to the return rate from 2013 to 2017, and uses stepwise regression method to screen out the strong correlation factors from the many factors that may affect the stock market rate of return as effective factors to predict the stock market rate of return . At the same time, the effective factor is used to establish a mapping relationship between the effective factor and the rate of return through a weighted linear regression model. Through the significance test of the regression coefficient and the statistical test of the model, it is found that the regression results conform to economic laws. Comparing the effect of the predicted value and the real value, the deviation between the two is within a reasonable range, so the purpose of forecasting stock market returns is achieved through the above calculations.

Keywords

Stock market returns, Effective factors, Weighted linear regression, Prediction of return.

1. Introduction

What factors determine the return on stock portfolios have always been a hot issue for researchers in the capital markets. By analyzing a large amount of stock market transaction data, stock investors can find certain rules, which can guide investors to establish effective investment strategies and enable them to obtain long-term excess returns.

The Capital Asset Pricing Model (CAPM) was developed by American scholars William Sharpe, John Lintner, Jack Treynor, and Jan Mossin in 1964. Developed on the basis of asset portfolio theory and capital market theory, which mainly studies the relationship between the expected return of assets in the securities market and risky assets, and how the equilibrium price is formed, which is the backbone of modern financial market price theory, Widely used in investment decision-making and corporate financial management. Arbitrage pricing theory APT (Arbitrage Pricing Theory) is an extension of CAPM. The pricing model given by APT is the same as CAPM. It is a model under equilibrium. The difference is that APT is based on a multifactor model. Arbitrage pricing theory uses multiple factors to explain the return on risk assets, and according to the principle of no arbitrage, it is obtained that there is a linear relationship between the equilibrium return on risk assets and multiple factors.

Fama and French's 1992 research on the factors that determine the difference in return on different stocks in the US stock market found that the beta value of the stock market cannot explain the difference in return on different stocks, and the market value, book-to-market ratio, and price-earnings ratio of listed companies can explain stock Difference in returns. Later, they found that in addition to the above risks, there are profit level risks and investment level risks that can also bring excess returns to individual stocks, and published a five-factor model in 2013. Carhart introduced momentum factors on the basis of Fama and French three-factor models, and constructed a four-factor model. The empirical results show that the four-factor model can better explain the difference in return on investment portfolios. With many "market anomalies" appearing in the capital market after the 1980s, such as the value premium phenomenon, scale effects, momentum reversal effects, calendar effects, etc., it shows that the market is not completely effective.

2. Theory and Model

2.1 stepwise regression

The basic idea of stepwise regression is to model the variables one by one, and then perform an F test after each explanatory variable, and then gradually perform a t test one by one for the explanations that have been selected. When the original published explanation is changed from the date of the explanatory variable to be repeated, One process, until neither significant explanatory variables are selected into the regression equation, nor is it to ensure that the final set of explanatory variables is optimal.

The specific steps for stepwise regression are as follow:

step 1: candidate regression independent variables $\boldsymbol{X_1}$, $\boldsymbol{X_{2,...,}X_m}$, and establish a regression relationship with the dependent variable Y as follows:

$$\mathbf{Y} = \boldsymbol{\alpha_0} + \boldsymbol{\alpha_i X_i} + \boldsymbol{\varepsilon}\,, \boldsymbol{i} = \mathbf{1}, \mathbf{2}, \dots, \boldsymbol{m}$$

Calculate the F-test statistic of the regression coefficient corresponding to the variable $\boldsymbol{X_i}$, $\boldsymbol{F_1^{(1)}}$,..., $\boldsymbol{F_m^{(1)}}$, iterate through the values of each F statistic, iterate through the values of each F statistic and take the largest one: $\boldsymbol{F_{i_1}^{(1)}}$, $\boldsymbol{F_{i_1}^{(1)}} = \mathbf{max}\{\boldsymbol{F_1^{(1)}},\dots, \boldsymbol{F_m^{(1)}}\}$.

For a given significance level $\boldsymbol{\alpha}$, let the critical value be $\boldsymbol{F^{(1)}}$. If the condition $\boldsymbol{F_{i_1}^{(1)}} \geq \boldsymbol{F^{(1)}}$ is satisfied, let $\boldsymbol{X_{i_1}}$ choose included as the model variables, let $\boldsymbol{I_1}$ be the selected variable set.

step 2: Establish dependent variable Y and independent variable subset $\{\boldsymbol{X_{i_1}}, \boldsymbol{X_1}\}$,...,$\{\boldsymbol{X_{i_1}}, \boldsymbol{X_{i_1-1}}\}$,$\{\boldsymbol{X_{i_1}}, \boldsymbol{X_{i_1+1}}\}$,...,$\{\boldsymbol{X_{i_1}}, \boldsymbol{X_m}\}$ binary regression models, a total of *m*-1. Calculate the statistic of the F-test of the regression coefficient of the variable, and record it as $\boldsymbol{F_n^{(1)}}$. Select the one with the largest value, and record it as $\boldsymbol{F_{i_2}^{(2)}}$. The condition $\boldsymbol{F_{i_2}^{(2)}} = \mathbf{max}\{\boldsymbol{F_1^{(2)}}, \dots, \boldsymbol{F_{i_1-1}^{(2)}}, \boldsymbol{F_{i_1+1}^{(2)}}, \boldsymbol{F_m^{(2)}}\}$. Given the significance level $\boldsymbol{\alpha}$, the critical value is $\boldsymbol{F^{(2)}}$. If the condition $\boldsymbol{F_{i_2}^{(1)}} \geq \boldsymbol{F^{(2)}}$ is satisfied, let $\boldsymbol{X_{i_2}}$ be selected is a model variable, otherwise the program terminates.

step 3: Repeat step 2 until no variables are selected into the model.

2.2 Weighted Linear Regression

After finding a set of characteristic factors that are strongly correlated with the response variables, the importance of the characteristic factors is verified by a linear regression model of the response variable R. The formula is as follows:

$$\mathbf{R} = \boldsymbol{X_s^T \alpha} + \boldsymbol{X_I^T \gamma} + \boldsymbol{\alpha_0} + \boldsymbol{\varepsilon_r};\ \ \boldsymbol{\varepsilon_r} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{\sigma_0^2}) \qquad (1)$$

X_s is the vector set of important strong correlation variables, X_I is the vector set of irrelevant feature variables, and ε_r is the residual. Before starting the experiment, the important variable subset S of the variable X_s needs to be determined.

X_I can then be assumed to be distributed in $N(\mu_1, \sigma_1^2)$ and added to the residual ε_r. We then use the important strong correlation feature factors to build a simplified regression model for the response variable R, as shown below:

$$\boldsymbol{y} = \boldsymbol{X^T\alpha} + \boldsymbol{\alpha_0} + \boldsymbol{\varepsilon};\ \ \boldsymbol{\varepsilon} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\sigma^2}) \tag{2}$$

Where α is the parameter vector, α_0 is the absolute term (constant), X and y are important strong correlation special factors and response variables, and ε is the error term, which reflects the random variable and the unconsidered variability in the variable. Errors can be heteroscedastic or correlated.

The parameters of the model are estimated as follows:

$$\boldsymbol{\widehat{\alpha}} = \boldsymbol{a} = (\boldsymbol{X^TWX})^{-1}\boldsymbol{X^TWy} \tag{3}$$

Among them, X is an important strong correlation feature factor matrix included in the model, and W is a diagonal weight matrix estimated according to the regression projection matrix H. The formula is as follows:

$$\mathbf{W} = (\mathbf{I} - \mathbf{diag}[\mathbf{X}(\boldsymbol{X^TX})^{-1}\boldsymbol{X^T}]^{-1}) = (\mathbf{I} - \mathrm{diag}\mathbf{H})^{-\mathbf{1}} \tag{4}$$

3. The datasets

3.1 Sample selection and data source

Because different industries have different operating characteristics, their corresponding resource acquisitions are very different, and the costs, prices, and risks they face are also different, so different industries show different characteristics. In order to enhance the comparability of the indicators and remove the different interference between industries, the samples selected in this paper are listed companies in the manufacturing industry. This article selects companies that were ST and * ST from 2012 to 2017, excluding companies with incomplete sample data, and finally selected 540 manufacturing listed companies that have never been ST and * ST.

3.2 Candidate variables

Candidate factors are selected based on those factors that have an impact on stock returns. Currently, there are many factors for stock analysis in the market. However, not all factors have an effect on stock returns. Based on the review of the introduction and extensive literature analysis, we selected 30 factors for selecting the initial variables of the predictors. These are: operating income growth rate, net asset growth rate, net profit growth rate, operating profit growth rate, cash flow growth rate of operating activities, return on net assets, return on assets, gross sales margin, net sales margin, total asset turnover rate, inventory turnover rate, current asset turnover rate, accounts receivable turnover rate, current ratio, quick-freeze ratio, asset-liability ratio, total market value, market value in circulation, one-month momentum, two-month momentum, three-month momentum, six-month momentum, consensus forecast earnings per share, consensus forecast net profit, consensus forecast return on net assets, consensus forecast operating income, price earnings ratio（P/E）, price-to-book ratio (P/B), price cash flow ratio（PCF）, price-to-sales (PS). As shown in Table 1.

Table 1 Financial parameters and categories.

Categories	Variable code	Constituent candidate variables
growth factor	X1	operating income growth rate
	X2	net asset growth rate
	X3	net profit growth rate
	X4	operating profit growth rate
	X5	cash flow growth rate of operating activities

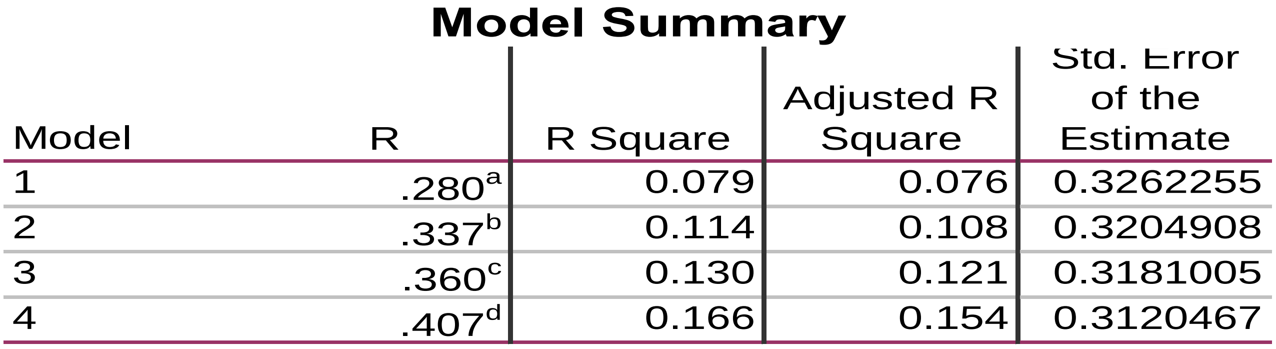
profitability	X6	return on net assets
	X7	return on assets
	X8	gross sales margin
	X9	net sales margin
operational capability	X10	total asset turnover rate
	X11	inventory turnover rate
	X12	current asset turnover rate
	X13	accounts receivable turnover rate
Leverage factor	X14	current ratio
	X15	quick-freeze ratio
	X16	asset-liability ratio
Scale factor	X17	total market value
	X18	market value in circulation
Momentum factor	X19	one-month momentum
	X20	two-month momentum
	X21	three-month momentum
	X22	six-month momentum
Predictive factor	X23	consensus forecast earnings per share
	X24	consensus forecast net profit
	X25	consensus forecast return on net assets
	X26	consensus forecast operating income
Value factor	X27	price earnings ratio（P/E）
	X28	price-to-book ratio (P/B)
	X29	price cash flow ratio（PCF）
	X30	price-to-sales (PS)

4. Empirical findings and discussion

4.1 Factor selection

Collecting 30 factor data of 540 listed companies through Guotai'an database, combining the 30 factors into a data matrix, using MAD (median absolute deviation) method to remove outliers of the data matrix, and normalizing the data of the data matrix To eliminate the impact of different dimensions on the model. Using the principle of stepwise regression above, the specific steps of implementing stepwise regression on the model are set as $Y=(y_1,y_{2,\dots,}y_{540})^T$ represents the company's stock return rate, $X=(X_1,X_{2,\dots,}X_{30})$ is a data matrix of candidate factors composed of 540 companies. The stepwise regression is used to screen single factors, and the process of model establishment is the model solving process. This paper takes the confidence level α = 0.5 and uses SPSS 22 to realize the stepwise regression process.

Table 2 Overview of the model



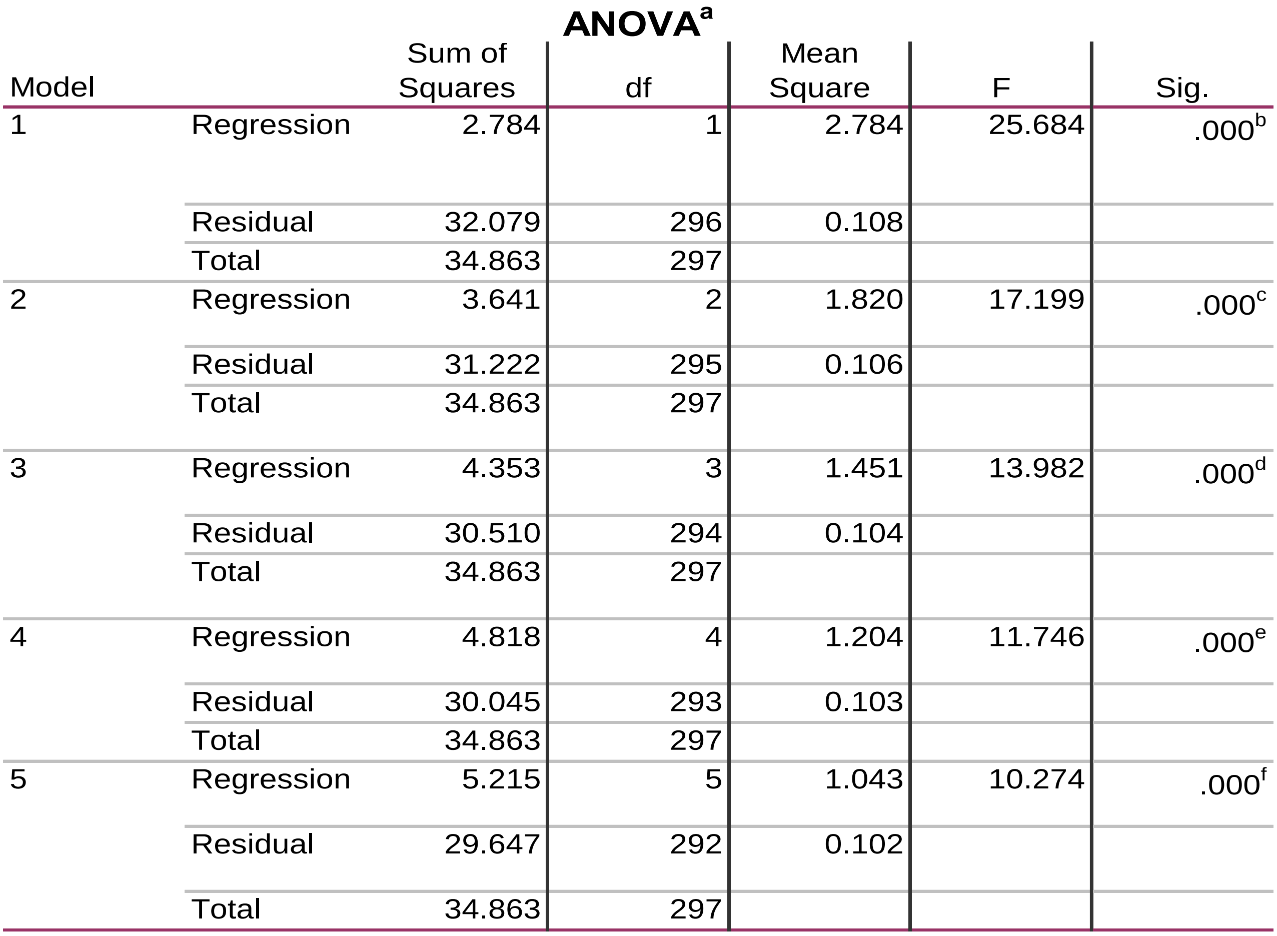
Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.280[a]	0.079	0.076	0.3262255
2	.337[b]	0.114	0.108	0.3204908
3	.360[c]	0.130	0.121	0.3181005
4	.407[d]	0.166	0.154	0.3120467

The above table 2 is an overview of the stepwise regression model. We see the four parameters marked in the above table, which are the negative correlation number, the determination coefficient, the correction determination coefficient, and the estimated value of the random error. These values (except the estimated value of the random error) The larger both are, the better the model is. According to comparison, the fourth model should be the best.

The following table 3 gives the results of the analysis of variance for all five models calculated by the model. This table can test whether all partial regression coefficients are all 0. A sig value less than 0.05 can prove that at least one of the partial regression coefficients of the model is not zero.

Table 3 Variance analysis table

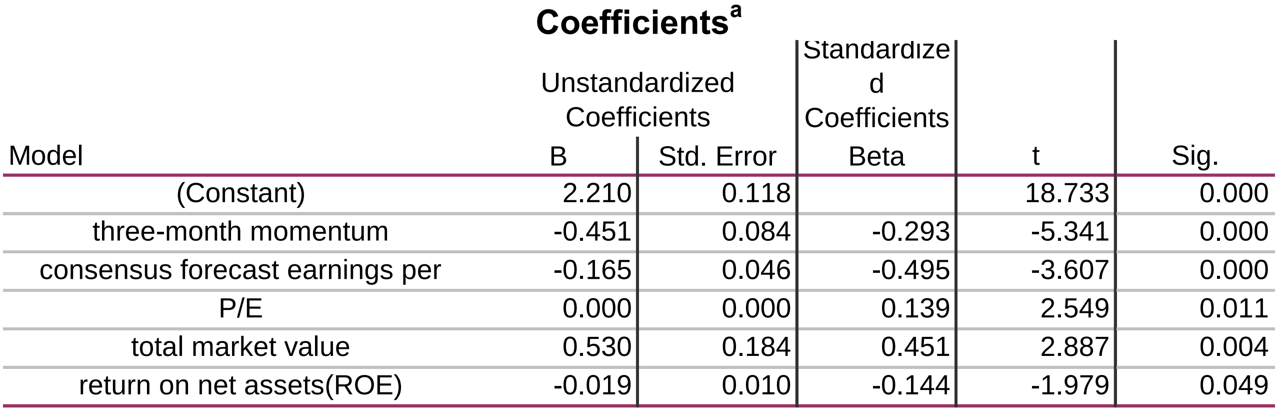


ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2.784	1	2.784	25.684	.000[b]
	Residual	32.079	296	0.108		
	Total	34.863	297			
2	Regression	3.641	2	1.820	17.199	.000[c]
	Residual	31.222	295	0.106		
	Total	34.863	297			
3	Regression	4.353	3	1.451	13.982	.000[d]
	Residual	30.510	294	0.104		
	Total	34.863	297			
4	Regression	4.818	4	1.204	11.746	.000[e]
	Residual	30.045	293	0.103		
	Total	34.863	297			
5	Regression	5.215	5	1.043	10.274	.000[f]
	Residual	29.647	292	0.102		
	Total	34.863	297			

The parameter test of Model 5 is shown in Table 4. The test of parameters. This table gives a test of partial regression coefficients and standard partial regression coefficients. Partial regression coefficients are used for comparison of different models, and standard partial regression coefficients are used for the same The test of different coefficients of the model, the larger the value, the greater the impact on the dependent variable.

Table 4 Parameter check



Coefficients[a]

	Unstandardized Coefficients		Standardized Coefficients		
Model	B	Std. Error	Beta	t	Sig.
(Constant)	2.210	0.118		18.733	0.000
three-month momentum	-0.451	0.084	-0.293	-5.341	0.000
consensus forecast earnings per	-0.165	0.046	-0.495	-3.607	0.000
P/E	0.000	0.000	0.139	2.549	0.011
total market value	0.530	0.184	0.451	2.887	0.004
return on net assets(ROE)	-0.019	0.010	-0.144	-1.979	0.049

In the table above, the three-month momentum, consistent forecast of earnings per share, price-earnings ratio, total market value, and return on net assets were tested. At the significance level, the five factors passed the significance test.

4.2 Weighted linear regression

The above five factors selected through stepwise regression are used to establish a regression relationship between the factors and the return using weighted linear regression. The weight matrix is solved by the above formula (4), and the regression coefficient of the factor is calculated by the formula (3). Compare the calculated predicted value with the real value. The comparison between the actual and predicted values is shown in Figure 1.

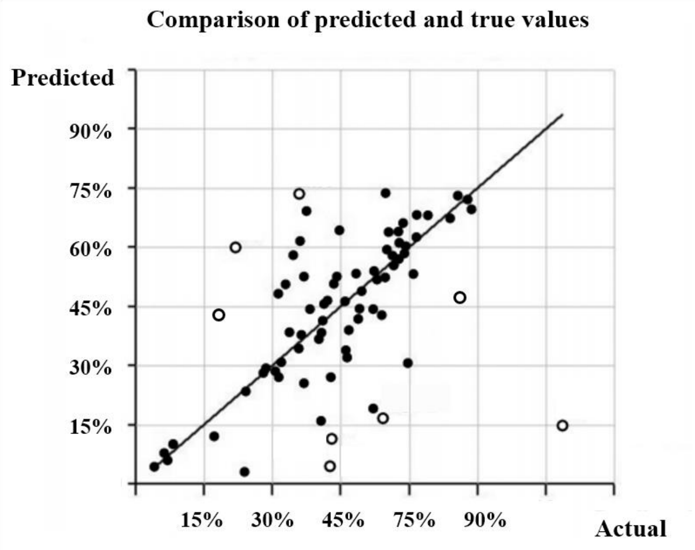


Fig.2. Comparison of predicted and true stock returns

5. Conclusion

The article constructs 30 factors that have an impact on market returns. Using the stepwise regression method, five factors related to the rate of return were screened out of 30 factors, and the confidence level α was set to 0.05. The parameters of the five factors were tested and all passed the significance test. They are: total market value, 3-month momentum, return on net assets, price-earnings ratio, and consensus earnings per share. Through weighted linear regression, using five factors to calculate the predicted rate of return, and comparing the predicted rate of return with the true rate of return, it can be seen that some forecasting effects are still considerable. It shows that the five factors are indeed effective factors that have an impact on the rate of return.

References

[1] Huang S M , Tsai C F , Yen D C , et al. A hybrid financial analysis model for business failure prediction[J]. Expert Systems with Applications, 2008, 35(3):1034-1040.

[2] Gregory, Alan, Tharyan, Rajesh, Christidis, Angela. The Fama-French and Momentum Portfolios and Factors in the UK[J]. Ssrn Electronic Journal.

[3] Fama E F. Efficient capital markets: a review of theory and empirical work[J].Journal of Finance,1970(25): 383-417.

[4] Markowitz H. Portfolio Selection[J].The Journal of Finance,1952,7(1): 77-91.

[5] Fama E F, French K R. Profitability, investment and average returns[J].Journal of Financial Economics,2006,82(3): 491-518.

[6] Guerard J J B, Markowitz H, Xu G. Earnings forecasting in a global stock selection model and efficient portfolio construction and management[J].International Journal of Forecasting,2015,31(2): 550-560.

[7] Lund, Iver A. An Application of Stagewise and Stepwise Regression Procedures to a Problem of Estimating Precipitation in California[J]. Journal of Applied Meteorology, 10(5):892-902.

[8] M Montanaro Gauci, T F Kruger, K Coetzee. Stepwise regression analysis to study male and female factors impacting on pregnancy rate in an intrauterine insemination programme[J]. Andrologia, 2001, 33(3):135-141.

[9] Darren T. Andrews, Liguo Chen, Peter D. Wentzell. Comments on the relationship between principal components analysis and weighted linear regression for bivariate data sets[J]. 34(2):231-244.