The Optimization and Empirical Study of Stock Excess Return

Congtong Qiu*

International College, Xiamen University, Xiamen, Fujian, 361102, China

*Corresponding author. Email: qiucongtong0418@163.com

Abstract. In order to explore the influencing factors of stock excess return and provide a particular reference for predicting the development of stocks in the country, I comprehensively refer to relevant economic theories and academic research and use the 2018 and 2019 CSI 300 index data provided by the wind to establish an econometric model. Moreover, carry out inspection and evaluation. I choose the logarithm (R) of the stock excess return as the explained variable, and the risk coefficient, market capitalization, book-to-market value ratio, turnover rate, etc., as the explanatory variables for regression analysis. According to the support of economic theory and the observation of the scatter plot trend, I removed the two explanatory variables of price-earnings ratio and net profit. In order to reduce the influence of outliers on model parameter estimation, I use the SPSS case diagnosis function to remove outliers. I performed a least-squares estimation of the initial model and performed t-tests and F-tests for the parameters and the overall equation for the remaining data. The variance inflation factor shows that the model has a multicollinearity problem, which I remedy by removing the year*ln_ME term. The model did not pass White's heteroskedasticity test. I used Eviews to perform White's heteroscedasticity correction on the model to correct the model standard error and reduce the degree of heteroskedasticity. After remediation, I used the RESET test on the model again, and although the F value increased slightly, it was still in the receptive field, so the model remediation did not make the model setting severe problems. The Jacques Bella test is shown that the model residuals do not completely obey the normal distribution, but due to the relatively large sample size, the model residuals can be approximated by a normal distribution according to the central limit theorem. In addition, I also obtain numerical prediction results through point prediction, and the scatter plot shows a high degree of interval overlap. Finally, I summarize the form presented by the final model and return to the practical significance, which proposes a reference standard for evaluating excess return from the market environment and company size, and further affirm the economic significance of excess return as a diversified evaluation of stock quality.

Keywords: Excess Rate of Return, Multiple Linear Regression, CSI 300 Constituent Stocks.

1. Introduction

Excess rate of return refers to the rate of return that exceeds the normal (or expected) rate of return, which is equal to the difference between the rate of return on a certain day minus the normal (expected) rate of return required by investors (or the market) on that day, where the normal rate of return is at The expected rate of return if the event does not occur. Under the background of stock market volatility, excess return measures the amount of particular income of listed companies from the perspective of change, which reflects the market environment and the company's profitability[1].

In order to further explore the factors affecting the excess rate of return in recent years, so as to provide a certain reference standard for predicting the correlation between the future situation of my country's stocks and the company's operating conditions, I analyzed the data of some CSI 300 constituent stocks in 2018 and 2019. . The selected indicators include indices related to stock market conditions, such as risk coefficient (beta), market value of listed companies, turnover rate and trading volume of individual stocks, etc.; also cover indices related to the company's operating conditions, such as the ratio of assets to net assets, net profit, etc[2][3].

2. Data collection and processing

The original data used in this model establishment and exploration process are from the data collected by wind, and the data has economic significance. The data has been processed for incomplete values, so the integrity, consistency, and reliability of the data are guaranteed.

3. Notations

See Table 1 for variable classification and interpretation.

The explained variable of this paper is the logarithm of the annual excess return of individual stocks in year t. Excess rate of return refers to the rate of return that exceeds the normal (or expected) rate of return, which is equal to the difference between the rate of return on a certain day minus the normal (expected) rate of return required by investors (or the market) for that day.

Under the theoretical support of the literature and preliminary correlation judgment, I eliminated the dummy variables of price-earnings ratio and a price-earnings ratio, and selected risk coefficient, the market value of listed companies, asset-to-market value ratio, book-to-market value ratio, turnover rate, trading volume, assets/net assets, year A total of 8 indicators and seven dummy variables year interaction terms with other explanatory variables, a total of 15 model explanatory variables.

Table 1. Explanatory Variable Description

Feature performance	Variable	Variable Code	Definition
Rate of Return (explained variable)	Take the logarithm of the annual excess return of individual stocks in year t	R	Total income earned from investing in stocks/Original investment
Risk characteristics	Risk factor calculated by CAPM	Beta	Calculated from the Capital Asset Pricing Model (CAPM)
Scale characteristics of listed companies	Take the logarithm of the market value of listed companies reported at the end of t-1	ln_ME	Ln (total shares * closing price of individual shares)
	Asset to Market Ratio	AM_ratio	Net assets per share/market value per share
	Book-to-market ratio	BM_ratio	Total shareholders' equity/total market value
Market Valuation Features	Last year's end-of-year price-earnings ratio	E_P	EPS/Stock Price
	Dummy variable of inverse price-earnings ratio, equal to 1 when net profit is negative, 0 otherwise	EP_dummy	1, E_P<0
			0, E_P>0
Liquidity Characteristics	The turnover rate of individual stocks in year t	tor	Number of shares traded on the day/number of all outstanding shares
	The trading volume of individual stocks in year t	Vol	The number of deals for a transaction in a unit of time
Financial characteristics	Assets/Net Assets	AB_ratio	Assets/Net Assets

Table 2. First regression coefficient

Variable	Unstandardized coefficients		Standardized coefficient	t	Significance
	B	standard error	Beta		
(Constant)	1.519	0.445		3.410	0.001
year*ln_ME	-0.109	0.026	-3.039	-4.227	0.000
year*tor	-0.008	0.001	-0.452	-8.789	0.000
ln_ME	-0.059	0.018	-0.143	-3.335	0.001
Vol	5.139E-09	0.000	0.166	5.194	0.000
beta	0.065	0.017	0.174	3.927	0.000
year	2.516	0.650	2.829	3.870	0.000
BM_ratio	-0.078	0.020	-0.142	-3.840	0.000
year*AM_ratio	0.076	0.019	0.155	4.065	0.000
year*beta	-0.075	0.033	-0.082	-2.292	0.022

Table 3. Model summary

Model	R	R^2	Adjusted R^2	Standard error
1	0.769a	0.592	0.585	0.286
a. Predictor: (constant), year*beta, year*tor, Vol, BM_ratio, ln_ME, year*AM_ratio, beta, year*ln_ME, year				
b. Dependent variable: R				

4. Model Establishment

4.1 Parameter estimation and hypothesis testing of the initial model

After the theoretical basis was obtained initially, I directly performed stepwise regression on 15 variables through the stepwise command in the SPSS software. At the same time, after using the stepwise method to determine the preliminary regression equation, I used the SPSS case diagnosis function to eliminate eight residuals in the sample data other than three times the standard deviation and then estimated the equation:

$$\begin{aligned} R &= B_1 + B_2 year + B_3 beta + B_4 \ln ME + B_5 BM\_ratio \\ &+ B_6 Vol + *beta + B_8 year*tor + B_9 year*AM\_ratio \\ &+ B_{10} year*\ln\_ME \end{aligned} \tag{1}$$

The above results are shown that the t values of all coefficients and constant terms are very significant. The correction R^2 is 0.585, the model F value is 87.066, and the corresponding P-value is significantly lower than 0.05. Therefore, the equation fits the sample data well, and finally I get the original equation as:

$$\begin{aligned} R &= 1.519 + 2.516 year + 0.065 beta - 0.059 \ln\_ME \\ &- 0.078 BM\_ratio + 5.139*10^{-9} Vol - 0.075 year*beta \\ &- 0.008 year*tor + 0.076 year*AM\_ratio - 0.109 year \\ &*\ln\_ME \end{aligned} \tag{2}$$

4.2 Hypothetical test

In order to test whether the model has the problem of missing variables, I use the python language to perform the Ramsey RESET test.

$$\begin{aligned} H_0 &: F = 0 \\ H_1 &: F > 0 \end{aligned} \tag{3}$$

After calculation, the F value of the RESET test is 0.359, and there is no good reason to reject the null hypothesis, so I believe that the model does not have a serious setting error problem.

4.3 Further optimization of the model

4.3.1. Diagnosis and Remediation of Multicollinearity

4.3.1.1. Diagnosis of multicollinearity - variance expansion factor and correlation coefficient

After obtaining the preliminary regression results stepwise, I found that the VIF values of year*ln_ME and year are 684.579 and 708.125, respectively. Generally speaking, the regression model has multicollinearity when VIF>10, which shows that this model has serious multicollinearity. Collinearity.

However, only the variance inflation factor cannot explain which variables are collinear, so I checked the pairwise correlation coefficients between the variables. As can be seen from the figure, the correlation between year*ln_ME and year with a high VIF value The coefficient is 0.998, and the two variables are collinear.

4.3.1.2. Remedy for Multicollinearity - Removing Variables

From the results of the above VIF and correlation coefficients, it can be preliminarily estimated that year*ln_ME and other variables have collinearity, so I choose to delete year*ln_ME.

After deleting year*ln_ME, the VIF values of the remaining variables are all less than 10, and the adjustment of the model R^2 and the significance of the remaining variables have not changed significantly, so the multicollinearity problem has been well resolved.

4.3.2. Diagnosis and Remediation of Heteroskedasticity

4.3.2.1.Test for heteroscedasticity

According to the Eviews operation, the following results are obtained:

Table 4. White's test results

Heteroskedasticity Test: White

F-statistic	6.920105	Prob. F(36,514)	0.0000
Obs*R-squared	179.8754	Prob. Chi-Square(36)	0.0000
Scaled explained SS	328.4101	Prob. Chi-Square(36)	0.0000

The test results can be obtained: $nR^2 = 179.875$. The null hypothesis is rejected, indicating that the model has heteroscedasticity.

4.3.2.2.Heteroskedasticity Remedy

Since the model has multiple explanatory variables, I performed variable transformations. By selecting $1/R_i^3$ the weight and applying the weighted least squares method, the following results are obtained:

Table 5. WLS regression results

White-Hinkley (HC1) heteroskedasticity consistent standard errors and covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-17.17463	4.752051	-3.614151	0.0003
YEAR*TOR	0.028562	0.021068	1.355747	0.1757
LN_ME	0.667284	0.186657	3.574929	0.0004
VOL	-3.28E-08	9.06E-09	-3.618899	0.0003
YEAR*AM_RATIO	-0.354740	0.167584	-2.116787	0.0347
YEAR	1.658407	0.707017	2.345641	0.0194
BETA	0.125475	0.079719	1.573960	0.1161
YEAR*BETA	-0.095994	0.173199	-0.554239	0.5796
BM_RATIO	0.494747	0.145418	3.402244	0.0007

Weighted Statistics

R-squared	0.999999	Mean dependent var	-8.510572
Adjusted R-squared	0.999999	S.D. dependent var	164.1187
S.E. of regression	0.120432	Akaike info criterion	-1.379267
Sum squared resid	7.861072	Schwarz criterion	-1.308839
Log likelihood	388.9880	Hannan-Quinn criter.	-1.351747
F-statistic	99687752	Durbin-Watson stat	1.932319
Prob(F-statistic)	0.000000	Weighted mean dep.	0.154384

Unweighted Statistics

R-squared	-16.315342	Mean dependent var	-0.007569
Adjusted R-squared	-16.570919	S.D. dependent var	0.444586
S.E. of regression	1.863600	Sum squared resid	1882.369
Durbin-Watson stat	0.419923		

The results of the White test can be obtained: $nR^2 = 8.885927$ accept the null hypothesis, indicating that the model heteroskedasticity has been eliminated. However, due to the large change in the model form and the lack of theoretical economic basis, each variable loses its practical significance. The coefficients of the three variables are not significant, which may affect the accuracy of prediction and the rationality of the model. Therefore, I choose the original model for subsequent predictions and take further heteroskedasticity corrections.

4.3.2.3.White's Heteroskedasticity Correction

We use Eviews' White's correction for heteroskedasticity to obtain robust standard deviations. After weakening the heteroscedasticity, the coefficient values of each variable did not change, the parameter t-test was almost all significant, and the F test was also significant, which effectively improved the accuracy of the model.

5. Retesting of Basic Assumptions of Classical Linear Models

5.1 RESET

After removing the variables, considering that the original equation may have missing variables, I performed the RESET test on the equation again.

The calculated F-value for the RESET test is 1.005, a slight increase but still not statistically significant, so I consider the model still free of significant specification error.

5.2 Random errors follow a normal distribution

5.2.1 Residual Histogram Test

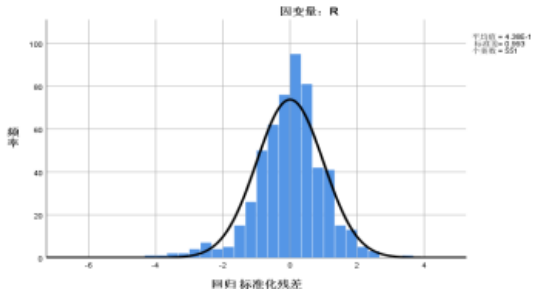

Figure 1 Residual test histogram

Through SPSS software, I can get the normalized residual frequency histogram of regression. It can be seen from the figure that the regression residual approximately obeys the normal distribution, and it is preliminarily determined that the regression residual meets the normality test.

5.2.2. Residual probability plot

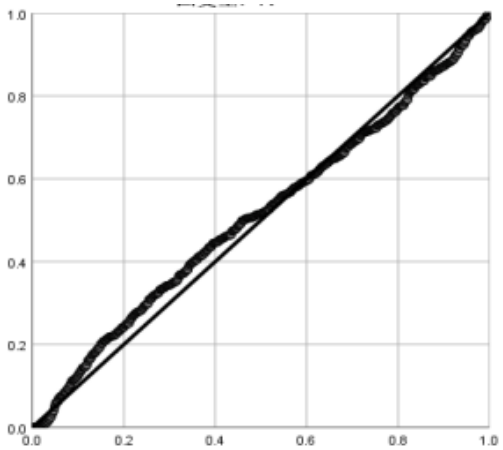

Figure 2 Normal P-P diagram of regression standardized residuals

We can draw the P-P diagram of the normality test. The horizontal axis is the percentage cumulative distribution probability of the residual, and the vertical axis is the expected cumulative probability of the normal distribution. It can be seen that the scattered points of the residual are evenly distributed on the slash, indicating that the residual is better subject to the normality hypothesis.

5.2.3 Jacques-Bella test

We use the Jacobellis test to analyze the normality of residuals quantitatively.

$$H_0 : JB = 0$$
$$H_1 : JB > 0 \tag{4}$$

We use Eviews software to calculate JB statistics and get the skewness s=-0.61. If the skewness k=4.77 is calculated as JB=105, there is sufficient reason to reject the original hypothesis. Therefore, I cannot judge that the residual completely obeys the normal distribution.

5.2.4 Central Limit Theorem

Although the Jacques Bella test did not pass, the sample size n=551>30, using the central limit theorem, the residual can be regarded as approximately obeying the normal distribution.

5.3 Autocorrelation test

We use the runs test method to test the autocorrelation hypothesis [4].

We use the python language to calculate the number of runs of residuals k=254 and the number of positive and negative residuals n_1=296, n_2=255.

Because n_1>10, n_2>10, the residuals are independent. If k can be approximated as a normal distribution, where:

$$E(k) = \frac{2n_1n_2}{n} + 1 = 274.97, \quad \sigma_k^2 = \frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)} = 135.98 \tag{5}$$

Construct the 95% confidence interval:

$$(E(k) - 1.96\sigma_k, E(k) + 1.96\sigma_k) \tag{6}$$

According to the principle of the runs test, when the number of k runs falls within the confidence interval, the variable has no autocorrelation

$k = 254 \in (252.1, 297.8)$, it can be determined that the regression equation has no autocorrelation.

6. Empirical Analysis and Forecasting

6.1 Model final equations and parameter estimates

The final model obtained in this paper is:

$$R = 2.659 - 0.224\ \text{year} + 0.053\ \text{beta} - 0.105 \ln\_ ME - 0.065\ \text{BM\_ratio} + 4.97 * 10^{-9}\text{Vol} - 0.079\ \text{year} * \text{beta} - 0.007\ \text{year} * \text{tor} + 0.061\ \text{year} * \text{AM\_ratio} \quad (7)$$

Table 6. Parameter estimation table of the final model

White-Hinkley (HC1) heteroskedasticity consistent standard errors and covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.659492	0.514076	5.173348	0.0000
YEAR*TOR	-0.007126	0.001328	-5.367701	0.0000
LN_ME	-0.104740	0.020550	-5.096826	0.0000
VOL	4.97E-09	1.07E-09	4.651543	0.0000
YEAR*AM_RATIO	0.060995	0.019240	3.170173	0.0016
YEAR	-0.223718	0.056346	-3.970449	0.0001
BETA	0.052809	0.014006	3.770475	0.0002
YEAR*BETA	-0.078634	0.050920	-1.544266	0.1231
BM_RATIO	-0.065487	0.016804	-3.897048	0.0001

R-squared	0.578084	Mean dependent var	-0.007569
Adjusted R-squared	0.571856	S.D. dependent var	0.444586
S.E. of regression	0.290904	Akaike info criterion	0.384554
Sum squared resid	45.86692	Schwarz criterion	0.454982
Log likelihood	-96.94471	Hannan-Quinn criter.	0.412074
F-statistic	92.82700	Durbin-Watson stat	1.789595
Prob(F-statistic)	0.000000	Wald F-statistic	91.42588
Prob(Wald F-statistic)	0.000000		

6.2 Model empirical analysis and prediction analysis

6.2.1 Temporal characteristics

The influence of different years (times) on the excess rate of return.

The fitting equation shows that taking 2019 as the benchmark class, the excess return in 2018 is 22.4% lower than the excess return in 2019.

According to the assumptions in the first part, I believe that the stock market is divided into a bull market and a bear market based on the t-year situation, and this market nature will affect the excess return. However, due to the lack of a theoretical basis, I cannot make assumptions about the positive and negative coefficients. In this sample, 2018 and 2019 represent bear and bull markets, respectively, and the sample data shows that bear markets have lower excess returns than bull markets.

We analyze that because the excess rate of return is one of the indicators to measure income, its nature is roughly the same as yield and other indicators. Under the condition of a good stock market environment, the rate of return or excess rate of return will increase; when the stock market is not optimistic, yield, or excess return, falls.

6.2.2 Market Risk characteristics

The influence of the risk coefficient (beta) calculated according to the CAPM on the excess rate of return

In this sample, the risk coefficient beta positively correlates with the year. For every 1 unit increase in the risk coefficient, the average excess return increases by 5.3%. The CAPM model shows that the risk coefficient beta is positively correlated with stock returns. Therefore, combined with the analysis, it can be further concluded that the positive correlation also holds for excess returns.

On the other hand, the difference slope coefficient between the risk coefficient beta and the year is negative. When other conditions remain unchanged, the excess rate of return in 2018 is 7.9% lower than the excess rate of return in 2019, and the beta coefficient in 2018 is -0.026. The beta coefficient in 2019 was 0.053. Combined with the first point, this result reflects that the positive correlation

between the risk coefficient beta and the excess return weakens in a bear market and even reverses changes. "A bear market seeks stability, a bull market seeks progress" in the case of an unoptimistic market environment, high risks tend to reduce returns, and excess returns will also decrease as risks rise.

6.2.3 Scale characteristics of listed companies

It can be seen from the model that there is a negative correlation between the company's market value and excess return at the end of the previous year. For every 1% increase in market value, the excess return decreases by 0.105% on average. It is the same as the assumption that small-cap stocks have higher excess returns due to irrational volatility.

Here, the book-to-market ratio increases by 1 unit, and the excess return decreases by 6.5% on average, showing an inverse relationship. This result is consistent with the hypothesis that the lower the book-to-market ratio, the higher the stock excess return.

The results show that the asset-to-market ratio is positively correlated with excess return in 2018. If the asset-to-market ratio increases by 1 unit, the excess return will increase by 6.1% on average; however, there is no significant correlation with excess return in 2019.

Empirical research is shown that the lower the price-to-book ratio, the higher the asset-to-market ratio, and the higher the investment value. In a bear market environment, the impact of a high asset market capitalization ratio on high excess return is more prominent. Since net assets per share are one of the essential reference indicators for judging the intrinsic value of an enterprise in an environment of market volatility, the nature of this indicator increases the stability and accuracy of the measurement, which is better when explanatory variables change.

6.2.4 Liquidity Characteristics

The model shows that the excess return increases by an average of % for every 1 unit increase in the trading volume of individual stocks in the year. According to theoretical assumptions, the impact of individual stock trading volume on excess returns depends on the trend of stock prices, so it usually shows a negative correlation in bear markets.

Our sample data shows that the results in 2018 are not in line with the assumptions. The possible reason is that the operating conditions of the CSI 300 constituent index are less affected than the overall market, so there is a positive change with a small amount of change. However, the limited sample data may also lead to certain errors in this result.

The results show that the turnover rate was negatively correlated with the excess return in 2018. When the asset-to-market ratio increased by 1 unit, the excess return decreased by 0.7% on average; in 2019, there was no obvious correlation with the excess return.

The turnover rate is an important indicator that reflects the activity of the market, but as a separate indicator, it cannot make a clear explanation or prediction for the stock excess return. However, when combined with the market environment analysis, the linear relationship between the turnover rate and the excess return will be more obvious. According to behavioral finance theory, people exhibit irrational behaviors such as overconfidence in a bull market environment, which enhances the correlation between these two variables. Meanwhile, according to the model results, I found that when the market environment is good, high turnover has a negative effect on excess returns, but the effect is limited.

6.3 In-Sample Prediction Assessment

6.3.1 Point prediction

We directly substituted the sample data into the model and used SPSS software to calculate residuals e_i and predicted values $\hat{y}_i$. In order to explore the prediction effect, I comprehensively use the mean square error MSE and the Akaike Information Criterion AIC as the evaluation criteria, among which:

$$MSE = \frac{SSE}{n-s-1} = \frac{\sum (y_i - \hat{y}_i)^2}{n-s-1}$$
$$AIC = 2(s+1) + n\ln(\frac{SSE}{n}) \quad (8)$$

The calculated MSE = 0.083243 and AIC = 0.384554. Compared with other forms of models, the model MSE and AIC are relatively small, and the model prediction effect is good.

7. Conclusion

The arima model is simple and practical. Fewer parameters are required in the Arima model. Compared to the single AR model and the MA model, the Arima model combines the methods of auto-regression and moving average method, which has better accuracy. The model can be considered a five-year investment product consisting of three assets: U.S. dollars, gold, and Bitcoin, adjusting the proportion of the assets randomly according to the fluctuation of the market. I use the Arima model to do linear regression over time series and assess how the value of various products changes over time. Based on this, I use online learning with a passive attack algorithm to do simple optimization for each day and automatic update of the weight vector.

After running the model, it was found that the effect was good, the return was almost always increasing in the long run, and the final asset was $18,269.06. The above is the model of the team. From the perspective of the final assets, the model meets the requirements of the actual situation.

References

[1] Banz R. The relationship between return and market value of common stocks[J]. Journal of Financial Economics, 1981, 9(1):3— 18

[2] Qin Kan. Research on Influencing Factors of Shanghai Stock Index Fluctuation [D]. Shanghai Academy of Social Sciences, 2011(10)

[3] Deng Changrong, Ma Yongkai. "An Empirical Study of the Three-Factor Model in China's Securities Market", Journal of Management, No. 5, 2005

[4] Li Xue et al. Run-length test of China's stock market efficiency [J]. Statistical Research, 2001, 18(12): 43-46Wang Huiying, Hao Yongtao Stock Price Trend Prediction Algorithm Based on Technical Index and Random Forest