Analysis and Prediction about Yields Rate of Security Investment Based on R Software

Lu Wang[1], Yingying Zhang[1]

[1]Department of Statistics and Actuarial Science, College of Mathematics and Statistics, Chongqing University, China

Abstract

With the fantastic spurt of science and modern technology throughout the world, employing statistical techniques as the guidance of investment has become a new trend in finance field. R software is providing us a scientific platform for effective investment. Using R software as a basic tool, this paper discusses in detail about the normality hypothesis of yields rate which plays a significant role in the portfolio selection theory by Harry Markowitz. Firstly, we conduct normal distribution fitting of the hypothesis through line chart, histogram, curve of kernel density estimation, curve of empirical distribution function, and QQ chart. Secondly, we verify the normality hypothesis by the Shapiro-Wilk test and the Kolmogorov-Smirnov test. Finally we use the expectancy method and the exponential smoothing to make corresponding prediction, and arrive at the conclusion that the exponential smoothing is superior to the expectancy method.

Keywords: R Software, Portfolio Selection, Rate of Yield, Normal Distribution Fitting, Test of Normality, Prediction, Exponential Smoothing

1. Introduction

Harry Markowitz won the Nobel Prize in economics in 1990. His pioneering research of modern financial economics provides the theoretical basis for investors and scholars to measure the risks and benefits of different financial investments. His analysis on portfolio selection helps investors to select the most favorable combination and make the investment the highest paid but with the smallest risk[1]. The basic hypothesis of this theory is that the yields rates of securities or portfolios are indicated by the anticipated yields rate, and the risk is measured by the variance of yields rate. The rate of yields, whose properties are indicated by mean and variance, can be regarded as a random variable with normal distribution. This paper mainly uses R software to analyze, verify, and predict this hypothesis. Because of its powerful libraries, various statistical analysis functions, and free programming space, R software is extensively used nowadays with the result of being recognized as an innovative platform for investors as well as researchers.

2. Theoretical Basis

2.1. The Markowitz Investment Portfolio Selection Model[2]

Suppose r_i is the yields rate of security i, and r_i is a random variable. Let $u_i = Er_i$ be the anticipated yields rate of security i. Let σ_{ij} be the covariance of r_i and r_j (σ_{ii} is the variance of r_i). Let ω_i ($\omega_i \geq 0$) be the investment ratio in security i. Then the yields rate of the investment portfolio, which is a random variable, can be expressed as $\sum_{i=1}^{N} \omega_i r_i$. Let $E = \sum_{i=1}^{N} \omega_i u_i$ be the anticipated yields rate. Let $V = \sum_{i=1}^{N} \sum_{j=1}^{N} \sigma_{ij} \omega_i \omega_j$ be the variance of the yields rate. Then the Markowitz investment portfolio selection model can be expressed as

$$min \sum_{i=1}^{N} \sum_{j=1}^{N} \sigma_{ij} \omega_i \omega_j$$

$$s.t. E = \sum_{i=1}^{N} \omega_i u_i,$$
$$\sum_{i=1}^{N} \omega_i = 1,$$
$$\omega_i > 0, i = 1,2,\cdots,N.$$

This paper mainly analyzes one of the

important hypotheses on which the portfolio model is based, that is, the securitie's investment yields rate which can be regarded as a random variable obeys the normal distribution, and its properties are indicated by the mean and the variance[3].

2.2. Basic Knowledge on Descriptive Analysis of Statistical Data

Descriptive statistics is the analysis of data distribution through drawing charts, compiling statistics, and calculating statistics etc.. It is a basic step in data analysis, which is also the foundation of statistical inference[4].

2.2.1. Descriptive Statistics

The mean, median, mode, percentiles can be used to describe the central tendency of the quantitative data; the variance, standard deviation, range, quartile range, variation coefficient, and standard error can be used to indicate the degree of data dispersion; the skewness coefficient and the kurtosis coefficient can be used to measure the shape of the distribution [5].

① Calculating various descriptive statistics: data_outline().

② Five number summary: fivenum(x, na.rm = TRUE).

2.2.2. Data Distributions[5]

① The R software provides us with many functions, which can help us calculate the distribution function, the distribution law, the probability density function, and the inverse function of the distribution function of some typical distributions, e.g., the normal distribution, the Poisson distribution, etc..

② Relevant functions about the test of normal distribution in R software:

Histogram: hist(x);

Curve of kernel density estimation: density(x);

Curve of empirical distribution function: ecdf(x);

QQ chart: qqnorm(x); qqline(x).

2.2.3. Normality Test Functions[5]

① W normality test: shapiro.test(x).

② Kolmogorov-Smirnov test for the empirical distribution of a single population: ks.test(x).

2.3. Predictions of Yields Rates

2.3.1. Expectancy Method

Using expectancy method to obtain the predicted yields rate is to calculate the mean of the yields rates in recent N weeks, and then let the mean to be the predicted yields rate in week $N+1$, namely

$$\tilde{r}_{N+1} = \frac{1}{N}\sum_{i=1}^{N} r_i.$$

This method has been used repeatedly by Markowitz and his followers[3]. The advantages of this method are simple and practicable.

2.3.2. Exponential Smoothing

Exponential smoothing (ES) is put forward by Robert G. Brown. He thought the trend of time series is stable and regular, so time series can be reasonably postponed along with its trend. He believes that the most recent past situation, to a certain extent, will continue into the future, so he put greater weight on the data that is nearer to the target time. This method is used frequently in prediction and in the field of economic development and production[1].

Using exponential smoothing to predict the anticipated yields rate is mainly through calculating the weighted average. According to the continuity of securities investment and its yields, the closer the predicted time to the target time, the closer relationship between the predicted value and the actual value of yields rate. Therefore, when manipulating the known data, we can give different weight to the actual yields rate in different period, with the result that the closer the predicted time to the target time, the greater weight to corresponding rate of yields, that is to say, it has greater influence on the predicted rate of yields. The basic mathematical model is as shown below

$$S_t = \alpha Y_{t-1} + (1-\alpha) S_{t-1},$$

where S_t is the smoothing value of time t; Y_{t-1} is the actual value of time $t-1$; S_{t-1} is the smoothing value of time $t-1$; α ($\alpha \in$

[0,1]) is the smoothing constant, which decides the influence degree of Y_{t-1} and S_{t-1} to S_t. As to S_t, we find through observation and analysis that it has a property of information tracing by period, thus making it possible to search for the origin up to S_1, which can easily find all data. Therefore, the simple formula could be obtained

$$S_t = \alpha Y_{t-1} + \alpha(1-\alpha)Y_{t-2} + \cdots + \alpha(1-\alpha)^{t-2}Y_1 + (1-\alpha)^{t-1}S_1.$$

During the process of prediction, the coefficient of Y_i $(i = t-1, \cdots, 1)$ is exponentially decreasing, because of which named this method exponential smoothing. The crucial point of this method is to select the value of smoothing constant α, which determines the degree of smoothing and the response speed of the difference between the predictive value and the actual value. The closer smoothing constant α is to 1, the greater influence from the actual value Y_{t-1} of time $t-1$ on the smoothing value S_t of time t; the closer smoothing constant α is to 0, the greater influence from the smoothing value S_{t-1} of time $t-1$. Consequently, if the time series is stable enough, the actual value Y_{t-1} of time $t-1$ will approach the actual value Y_t of time t, so we can choose a bigger smoothing constant α when forecasting. On the contrary, if the time series fluctuate obviously, there will be a greater difference between the actual values of two adjacent times. In this condition, it is reasonable to use the smoothing value S_{t-1} of time $t-1$ to conduct the corresponding prediction, so a small value of α would be a better choice.

3. Normality Test of Rate of Yields and Corresponding Prediction

3.1. Normality Test

Searching for the opening price and the closing price of 61 weeks in the Shanghai Composite Index from the DaZhiHui software, 56 weeks of which are used to verify the hypothesis and the rest of which are used to compare and analyze the predicted yields rate. See the annex for the detailed data.

Suppose the actual yields rates of a security in the last N weeks are $r_1, r_2, \cdots, r_N$. Let c_i^0 be the opening price of the first day in week i. Let c_i^1 be the closing price of the last day in week i. Then

$$r_i = \frac{c_i^1 - c_i^0}{c_i^0} \times 100\%.$$

Read data through R software to calculate the weekly yields rate p of this security and draw the line chart. See Figure 1 for the graphical output of p.

```
C = read.csv(file = "data.csv", header = TRUE)
p = as.vector((C$close - C$open) / C$open); p
class(p)
plot(p, type = "l")
```

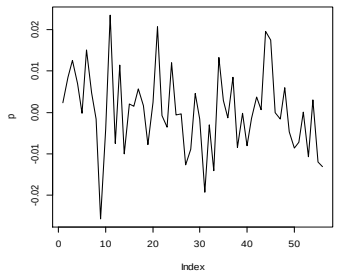

Figure 1. The line chart of the weekly yields rate p.

Then we preliminarily analyze the data of the weekly yields rate and calculate the basic descriptive statistics. The R codes and the results are as follows:

```
> source("data_outline.R")
> data_outline(p)
```

N	Mean	Var
56	0.0002191542	9.677276e-05
std_dev	std_mean	CV
0.009837315	0.001314566	4488.765
R	Skewness	Kurtosis
0.04920133	0.110744	0.3656516

In the above result, N is the number of observations in the sample; $Mean$ is the sample mean; Var is the sample variance; std_dev is the standard deviation; std_mean is the standard error of the sample; CV is the sample

coefficient of variation; R is the sample range; $Skewness$ is the skewness coefficient of the sample; $Kurtosis$ is the kurtosis coefficient of the sample.

Then we calculate the five number summary of the weekly yields rate. The R code and the result are as follows:

> fivenum(p)

[1]-0.0256846026,-0.0073585691,-0.0001639052, 0.0052556464,0.0235167320

The five numbers in the above result which can reflect the important characteristics of the data are the most representative values of data analysis, including the minimum min, the lower quartile Q_1, the median m_e, the upper quartile Q_3, and the maximum max.

In order to analyze the distribution of the weekly yields rate, we can draw a histogram through R software and then draw a kernel density estimation curve and a normal density estimation curve on it to discuss the distribution characteristics of the weekly yields rate clearly. See Figure 2 for the histogram, the kernel density estimation curve, and the normal density curve of p. The R codes are as follows:

w = seq(min(p), max(p), length.out = 51)

Vector = c(density(p)$y, dnorm(w, mean(p), sd(p)))

ylim = c(min(Vector), max(Vector))

dev.new(); hist(p, freq = FALSE, ylim = ylim, main = paste("Histogram of p"))

lines(density(p), col="blue", lty = 1)

lines(w, dnorm(w, mean(p), sd(p)), col="red", lty = 2)

leg.txt = c("Density estimation curve", "Normal density curve")

legend("topleft", legend = leg.txt, lty = 1:2, col = c("blue", "red"))

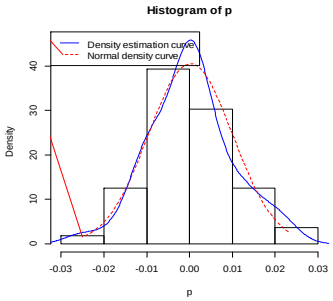

Figure 2. The histogram, the kernel density estimation curve, and the normal density curve of p.

To see whether the weekly yields rate p is from the normal distribution, we draw its empirical cumulative distribution function (cdf) curve and a fitted normal cdf curve on the same graph. See Figure 3. We see from Figure 3 that the empirical cdf curve and the fitted normal cdf curve agree basically. The R codes are as follows:

plot(ecdf(p), verticals = TRUE, do.p = FALSE)

x = seq(-0.03, 0.03, 0.001)

lines(x, pnorm(x, mean = mean(p), sd = sd(p)), col = "red")

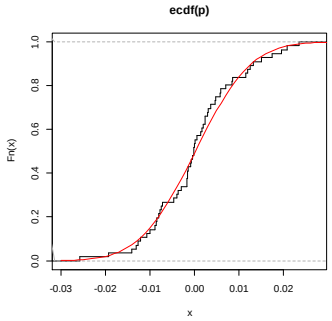

Figure 3. The empirical cdf curve and a fitted normal cdf curve.

Then we draw a QQ chart through R software so that we can discern whether the distribution of our sample is from normal. See Figure 4. We see from Figure 4 that most dots fit the normality very well; only partial dots have slight and non-significant deviations from the normality. The R codes are as follows:

qqnorm(p)

qqline(p)

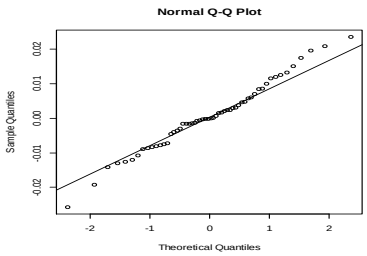

Figure 4. Normal QQ chart.

Then we conduct the Shapiro-Wilk test. The R code and the results are as follows:

shapiro.test(p)

Shapiro-Wilk normality test

data: p

W = 0.9847, p-value = 0.6966

Suppose the significance level is $\alpha = 0.05$, we see from the results that $p - value = 0.6966 > 0.05$. Therefore, the sample can be identified to derive from a normal distribution, that is, the securities yields rate approximately obeys a normal distribution.

For the sake of ensuring authenticity and reliability of the test results, we conduct a Kolmogorov-Smirnov test. The R code and the results are as follows:

ks.test(p, "pnorm", mean(p), sd(p))

One-sample Kolmogorov-Smirnov test

data: p

D = 0.0852, p-value = 0.7796

alternative hypothesis: two-sided

Similarly, we suppose the significance level is $\alpha = 0.05$, we see from the results that $p - value = 0.7796 > 0.05$. So the sample can also be identified to obey a normal distribution.

3.2. Predictions of Yields Rate

3.2.1. Expectancy Method

The R codes and the output result of using the expectancy method to calculate and predict the yields rate are as follows:

> mean(y)

[1] 0.0002191542

So the predicted yields rate through the expectancy method is $\tilde{r}_{N+1} = 0.0002191542$.

3.2.2. Exponential Smoothing

Suppose the yields rates of a security in the last N weeks are $r_1, r_2, \cdots, r_N$, the predicted yields rates are $\hat{r}_1, \hat{r}_2, \cdots, \hat{r}_N$. Then the formula of exponential smoothing is

$$\begin{cases} \hat{r}_1 = \dfrac{r_1 + r_2}{2}, \\ \hat{r}_{i+1} = \alpha r_i + (1-\alpha)\hat{r}_i, \quad i = 1,2,\cdots,N. \end{cases}$$

If the smoothing constant α is given the value of 0.1, we can calculate the predicted yields rate through exponential smoothing. The R codes and the results are as follows:

> source("ExponentialSmoothing.R")

> ExponentialSmoothing(p, alpha = 0.1)

[1] -0.002938993

Therefore, the predicted yields rate through the exponential smoothing is $\hat{r}_{N+1} = -0.002938993$.

3.2.3. Analysis of Forecasting Results

Use the opening price and the closing price of the last 5 weeks to conduct a comparison between the two predicted results. The R codes are as follows:

```
C1 = read.csv(file = "data1.csv", header = TRUE)
p1 = as.vector((C1$close - C1$open) / C1$open)
m = length(p1); m
Predict_ES = rep(0, m)
Predict_mean = rep(0, m)
alpha = 0.1
p_temp = p
for (i in 1:m){
Predict_ES[i]= ExponentialSmoothing(p_temp, alpha = alpha)
Predict_mean[i] = mean(p_temp)
p_temp = c(p_temp, p1[i])
}
p1
Predict_ES
Predict_mean
```

The comparisons of the actual yields rate and the predicted yields rates through two methods are shown in Table 1. In the table, p1 is the actual yields rate; $Predict\_ES$ is the predicted yields

rate through the exponential smoothing; $Predict\_mean$ is the predicted yields rate through the expectancy method.

Table 1. The comparisons of the actual yields rate and two predicted yields rates.

p1	-0.0036 70053	-0.0180 48582	-0.01823 6665	-0.00594 1881	-0.00859 8816
Predict_ ES	-0.0029 38993	-0.0030 12099	-0.00451 5747	-0.00588 7839	-0.00589 3243
Predict_ mean	0.00021 91542	0.00015 09224	-0.00016 28621	-0.00046 91978	-0.00056 04091

We then draw a line chart of the actual yields rate and two predicted yields rates to visualize their relationships. See Figure 5.

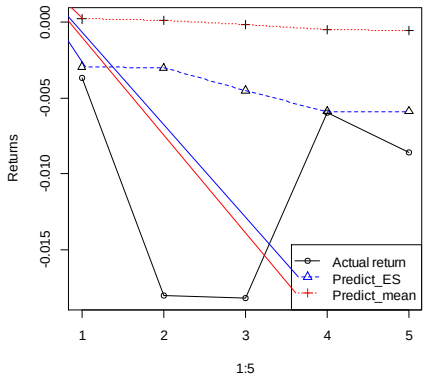

Figure 5. The line chart of the actual yields rate and two predicted yields rates.

We see from Figure 5 that the predicted yields rate through the expectancy method remains close at 0, which is quite far from the actual yields rate and do not reflect the variation trend of the actual yields rate. So the expectancy method is just a rough method of estimation and its sensitivity to the change of yields rate is almost 0, besides, the expectancy method can not reflect the future yields rate of securities which are at the rising or falling stage. In view of this, the expectancy method is impractical. While the exponential smoothing considers the influences on yields rate from different time to the predicted time and the influences are given different weights. In the long run, the predicted value through the exponential smoothing will get increasingly closer to the actual value. Therefore, the exponential smoothing is more practical and accurate. To conclude, the exponential smoothing is obviously superior to the expectancy method.

4. Conclusion

This paper discusses via examples about how to conduct validation, analysis, prediction, and application with a tool of R software in the field of financial investment. First of all, we analyze the normality hypothesis of the famous Markowitz portfolio selection theory and successfully verify that the securities investment rate of yields can be regarded as a random variable and it approximately obeys a normal distribution. Next, to make this theory apply in the investment decision, we predict the perspective yields rate using the expectancy method and the exponential smoothing respectively based on R software. Meanwhile, we compare and analyze the two methods in detail and finally discover that the exponential smoothing is superior to the expectancy method when predicting. This conclusion makes it possible for the prediction of yields rate keeping up with the pace of changes in securities market, which can provide stockjobbers with rational and scientific guidance. In addition, it can also give a new reference method to investment analysts, financial scholars as well as economics lovers.

References

[1] Bodie Z, Kane A, and Marcus AJ. *Investment*. Beijing: Machinery Industry Press. 2005.
[2] Song FM. *Financial Engineering Theory: Arbitrage Analysis with No Equilibrium*. Beijing: Tsinghua University Press. 1999.
[3] Markowitz H. Portfolio selection. *The Journal of Finance*, 7 (1): 77-91, 1952.
[4] Yang H, Liu QS, and Zhong B. *Mathematical Statistics*. Beijing: Higher Education Press. 2004.
[5] Xue Y and Chen LP. *Statistical Modeling and R Software*. Beijing: Tsinghua University Press. 2007.