Stochastic Processes and Advanced Mathematical Finance

Models of Stock Market Prices

Rating

Mathematically Mature: may contain mathematics beyond calculus with proofs.
Section Starter Question

What would be some desirable characteristics for a stochastic process model of a security price?

---

Key Concepts

1. A natural definition of variation of a stock price $s_t$ is the **proportional return** $r_t$ at time $t$

   $$r_t = \frac{(s_t - s_{t-1})}{s_{t-1}}.$$

2. The **log-return**

   $$\rho_t = \log(s_t/s_{t-1})$$

   is another measure of variation on the time scale of the sequence of prices.

3. For small returns, the difference between returns and log-returns is small.

4. The advantage of using log-returns is that they are additive.

5. Using Brownian Motion for modeling stock prices varying over continuous time has two obvious problems:

   (a) Brownian Motion can attain negative values.

   (b) Increments in Brownian Motion have certain variance on a given time interval, so do not reflect proportional changes.

6. Modeling security price changes with a stochastic differential equation leads to a Geometric Brownian Motion model.

7. Deeper statistical investigation of the log-returns shows that while log-returns within 4 standard deviations from the mean are normally distributed, extreme events are more likely to occur than would be predicted by a normal distribution.
Vocabulary

1. A natural definition of variation of a stock price $s_t$ is the proportional return $r_t$ at time $t$

   $$r_t = \frac{s_t - s_{t-1}}{s_{t-1}}.$$

2. The log-return

   $$\rho_i = \log\left(\frac{s_t}{s_{t-1}}\right)$$

   is another measure of variation on the time scale of the sequence of prices.

3. The compounding return at time $t$ over $n$ periods is

   $$\frac{s_t}{s_{t-n}} = (1 + r_t)(1 + r_{t-1}) \cdots (1 + r_{t-n+1})$$

4. The Wilshire 5000 Total Market Index, or more simply the Wilshire 5000, is an index of the market value of all stocks actively traded in the United States.

5. A quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
Mathematical Ideas

Returns, Log-Returns, Compound Returns

Let $s_t$ be a sequence of prices on a stock or a portfolio of stocks, measured at regular intervals, say day to day. A natural question is how best to measure the variation of the prices on the time scale of the regular measurement. A first natural definition of variation is the proportional return $r_t$ at time $t$

$$r_t = \frac{s_t - s_{t-1}}{s_{t-1}}.$$

The proportional return is usually just called the return, and often it is expressed as a percentage. A benefit of using returns versus prices is normalization: measuring all variables in a comparable metric, essentially percentage variation. Using proportional returns allows consistent comparison among two or more securities even though their price sequences may differ by orders of magnitude. Having comparable variations is a requirement for many multidimensional statistical analyses. For example, interpreting a covariance is meaningful when the variables are measured in percentages.

Define the log-return

$$\rho_t = \log(s_t/s_{t-1})$$

as another measure of variation on the time scale of the sequence of prices. For small returns, the difference between returns and log-returns is small. Notice that

$$1 + r_t = \frac{s_t}{s_{t-1}} = \exp \left( \log \left( \frac{s_t}{s_{t-1}} \right) \right) = \exp(\rho_t).$$

Therefore

$$\rho_t = \log(1 + r_t) \approx r_t, \text{ for } r_t \ll 1.$$

More generally, a statistic calculated from a sequence of prices is the compounding return at time $t$ over $n$ periods, defined as

$$\frac{s_t}{s_{t-n}} = (1 + r_t)(1 + r_{t-1}) \cdots (1 + r_{t-n+1}).$$

Taking logarithms, a simplification arises,

$$\log \left( \frac{s_t}{s_{t-n}} \right) = \log(s_t) - \log(s_{t-n})$$
the logarithm of the compounding return over \( n \) periods is just the difference between the logarithm of the price at the final and initial periods. Furthermore,

\[
\log \left( \frac{s_t}{s_{t-n}} \right) = \log \left( (1 + r_t)(1 + r_{t-1}) \cdots (1 + r_{t-n+1}) \right) = \rho_1 + \rho_2 + \cdots + \rho_n.
\]

So, the advantage of using log-returns is that they are additive. Recall that the sum of independent normally-distributed random variables is normal. Therefore, if we assume that log-returns are normally distributed, then the logarithm of the compounding return is normally distributed. However the product of normally-distributed variables has no easy distribution, in particular, it is not normal. So even if we make the simplifying assumption that the returns are normally distributed, there is no corresponding result for the compounded return.

### Modeling from Stochastic Differential Equations

Using Brownian Motion for modeling stock prices varying over continuous time has two obvious problems:

1. Even if started from a positive value \( X_0 > 0 \), at each time there is a positive probability that the process attains negative values, this is unrealistic for stock prices.

2. Stocks selling at small prices tend to have small increments in price over a given time interval, while stocks selling at high prices tend to have much larger increments in price on the same interval. Brownian Motion has a variance which depends on a time interval but not on the process value, so this too is unrealistic for stock prices.

Nobel prize-winning economist Paul Samuelson proposed a solution to both problems in 1965 by modeling stock prices as a Geometric Brownian Motion. Let \( S(t) \) be the continuous-time stock process. The following assumptions about price increments are the foundation for a model of stock prices.

1. Stock price increments have a deterministic component. In a short time, changes in price are proportional to the stock price itself with constant proportionality rate \( r \).
2. The stock price increments have a random component. In a short time, changes in price are jointly proportional to the stock price, a standard normal random variable, and the time increment with constant proportionality rate $\sigma$.

The first assumption is based on the observation that stock prices have a general overall growth (or decay if $r < 0$) rate due to economic conditions. For mathematical simplicity, we take $r$ to be constant, because we know that in the absence of randomness, this leads to the exponential function, a simple mathematical function. The second observation is based on the observation that stock prices vary both up and down on short times. The change is apparently random due to a variety of influences, and that the distribution of changes appears to be normally distributed. The normal distribution has many mathematically desirable features, so for simplicity the randomness is taken to be normal. The proportionality constants are taken to be constant for mathematical simplicity.

These assumptions can be mathematically modeled with a stochastic differential equation

$$dS(t) = rS\,dt + \mu S\,dW(t), \quad S(0) = S_0.$$ 

We already have the solution of this stochastic differential equation as Geometric Brownian Motion:

$$S(t) = S_0 \exp((r - (1/2)\sigma^2)t + \sigma W(t)).$$

At each time the Geometric Brownian Motion has a lognormal distribution with parameters $(\ln(S_0) + rt - (1/2)\sigma^2 t)$ and $\sigma \sqrt{t}$. The mean stock price at any time is $\mathbb{E}[X(t)] = S_0 \exp(rt)$. The variance of the stock price at any time is

$$\text{Var}[X(t)] = S_0^2 \exp(2rt) [\exp(\sigma^2 t) - 1].$$

Note that with this model, the log-return over a period from $t - n$ to $t$ is $(r - \sigma^2/2)n + \sigma[W(t) - W(t - n)]$. The log-return is normally distributed with mean and variance characterized by the parameters associated with the security. This is consistent with the assumptions about the distribution of log-returns of regular sequences of security processes.

The constant $r$ is often called the drift and $\sigma$ is called the volatility. Drifts and volatility are usually reported on an annual basis. Therefore, some care and attention is necessary when converting this annual rate to a daily
rate. An “annual basis” in finance often means 252 days per year because there are that many trading days, not counting weekends and holidays in the 365-day calendar. In order to be consistent in this text with most other applied mathematics we will use 365 days for annual rates. If necessary, the conversion of annual rates to financial-year annual rates is a straightforward matter of dividing annual rates by 365 and then multiplying by 252.

Testing the Assumptions on Data

The Wilshire 5000 Total Market Index, or more simply the Wilshire 5000, is an index of the market value of all stocks actively traded in the United States. The index is intended to measure the performance of publicly traded companies headquartered in the United States. Stocks of extremely small companies are excluded.

In spite of the name, the Wilshire 5000 does not have exactly 5000 stocks. Developed in the summer of 1974, the index had just shy of the 5,000 issues at that time. The membership count has ranged from 3,069 to 7,562. The member count was 3,818 as of September 30, 2014.

The index is computed as

\[ W = \alpha \sum_{i=1}^{M} N_i P_i \]

where \( P_i \) is the price of one share of issue \( i \) included in the index, \( N_i \) is the number of shares of issue \( i \), \( M \) is the number of member companies included in the index, and \( \alpha \) is a fixed scaling factor. The base value for the index was 1404.60 points on base date December 31, 1980, when it had a total market capitalization of $1,404.596 billion. On that date, each one-index-point change in the index was equal to $1 billion. However, index divisor adjustments due to index composition changes have changed the relationship over time, so that by 2005 each index point reflected a change of about $1.2 billion in the total market capitalization of the index.

The index was renamed the “Dow Jones Wilshire 5000” in April 2004, after Dow Jones & Company assumed responsibility for its calculation and maintenance. On March 31, 2009 the partnership with Dow Jones ended and the index returned to Wilshire Associates.

The Wilshire 5000 is the weighted sum of many stock values, each of which we may reasonably assume is a random variable presumably with a
finite variance. If the random variables are independent, then the Central
Limit Theorem would suggest that the index should be normally distributed.
Therefore, a reasonable hypothesis is that the Wilshire 5000 is a normal ran-
dom variable, although we do not know the mean or variance in advance.
(The assumption of independence is probably too strong, since general eco-
nomic conditions affect most stocks similarly.)

Data for the Wilshire 5000 is easy to obtain. For example, the Yahoo
Finance page for W5000 provides a download with the Date, Open, Close,
High, Low, Volume and Adjusted Close values of the index in reverse order
from today to April 1, 2009, the day Wilshire Associates resumed calculation
of the index. (The Adjusted Close is an adjusted price for dividends and
splits that does not affect this analysis.) The data comes in the form of a
comma-separated-value text file. This file format is well-suited as input for
many programs, especially spreadsheets and data analysis programs such as
R. This analysis uses R.

The data from December 31, 2014 back to April 1, 2009 provides 1449
records with seven fields each. This analysis uses the logarithm of each of
the Close prices. Reversing them and then taking the differences gives 1448
daily log-returns. The mean of the 1448 daily log-returns is 0.0006675644
and the variance of 1448 daily log-returns is 0.0001178775. Assume that the
log-returns are normally distributed so that the mean change over a year of
252 trading days is 0.1682262 and the variance over a year of 252 trading
days is: 0.02970512. Then the annual standard deviation is 0.1722922. The
initial value on April 1, 2009 is 8242.38.

Use the values $r = 0.1682262$ and $\sigma = 0.1722922$ with initial value $S_0 =
8242.38$ in the stochastic differential equation

$$dS(t) = rS \, dt + \mu S \, dW(t), \quad S(0) = S_0.$$ 

The resulting Geometric Brownian Motion is a model for the Wilshire 5000
index. Figure 1 is a plot of the actual data for the Wilshire 5000 over the
period April 1, 2009 to December 31, 2014 along with a simulation using a
Geometric Brownian Motion with these parameters.

**Testing the Normality Hypothesis**

Comparison of the graphs of the actual Wilshire 5000 data with the Geo-
metric Brownian Motion suggests that Geometric Brownian Motion gives
Figure 1: The Wilshire 5000 Index from April 1, 2009 to December 31, 2014 plotted in blue along with a Geometric Brownian Motion having the same mean, variance and starting value in red.
plausible results. However, deeper investigation to find if the fundamental modeling hypotheses are actually satisfied is important.

Consider again the 1448 daily log-returns. The log-returns are normalized by subtracting the mean and dividing by the standard deviation of the 1448 changes. The maximum of the 1448 normalized changes is 4.58 and the minimum is −6.81. Already we have a hint that the distribution of the data is not normally distributed, since the likelihood of seeing normally distributed data varying 4 to 6 standard deviations from the mean is negligible.

In R, the `hist` command on the normalized data gives an empirical density histogram. For simplicity, here the histogram is taken over the 14 one-standard-deviation intervals from −7 to 7. For this data, the density histogram of the normalized data has values

\[
0.0006906077 \quad 0.0000000000 \quad 0.0027624309 \quad 0.0062154696 \quad 0.0269337017 \\
0.0877071823 \quad 0.3542817680 \quad 0.4088397790 \quad 0.0877071823 \quad 0.0179558011 \\
0.0048342541 \quad 0.0020718232 \quad 0.0000000000 \quad 0.0000000000
\]

This means, for example, that 0.00069 of the 1448 log-returns, that is 1, occurred in the interval (−7, 6] and a fraction 0.4088397790, or 592 points, fall in the interval (0, 1]. The normal distribution gives the expected density on the same intervals. The ratio between the empirical density and normal density gives an indication of the deviation from normality. For data within 4 standard deviations, the ratios are about what we would expect. This is reassuring that for reasonably small changes, the log-returns are approximately normal. However, the ratio on the interval (−7, 6] is approximately 70,000, and the ratio on the interval (4, 5] is approximately 66. Each of these is much greater than we expect, that is, the extreme tails of the empirical density have much greater probability than expected.

Quantile-Quantile Plots

A quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the data point at which 30% percent of the data fall below and 70% fall above that value. As another example, the median is the 0.5 quantile.
A q-q plot is formed by plotting estimated quantiles from data set 2 on the vertical axis against estimated quantiles from data set 1 on the horizontal axis. Both axes are in units of their respective data sets. For a given point on the q-q plot, we know the quantile level is the same for both points, but not what the quantile level actually is.

If the data sets have the same size, the q-q plot is essentially a plot of sorted data set 1 against sorted data set 2. If the data sets are not of equal size, the quantiles are usually picked to correspond to the sorted values from the smaller data set and then the quantiles for the larger data set are interpolated from the data.

The quantiles from the normal distribution are the values from the inverse cumulative distribution function. These quantiles are tabulated, or they may be obtained from statistical software. For example, in R the function \texttt{qnorm} gives the quantiles, e.g. \texttt{qnorm(0.25)} = -0.6744898, and \texttt{qnorm(0.90)} = 1.281552. Estimating the quantiles for the normalized Wilshire 5000 data is more laborious but is not difficult in principle. Sorting the data and then finding the value for which 100\% of the data is less and 100(1 − \% of the data is greater gives the \% percentile for 0 ≤ \% ≤ 1. For example, the 0.25-quantile for the scaled and normalized daily changes is −0.4342129 and the 0.90-quantile for the scaled and normalized daily changes is 1.072526. Then for the q-q plot, plot the values (−0.6744898, −0.4342129) and (1.281552, 1.072526). Using many quantiles gives the full q-q plot.

In R, to create a q-q plot of the log-changes for the Wilshire 5000 data with a reference line against the normal distribution use the two commands \texttt{qqnorm()} and \texttt{qqline()} on the data set. In the q-q plot in Figure 2 the “twist” of the plot above and below the reference line indicates that the tails of the normalized Wilshire 5000 data are more dispersed than the standard normal data. The low quantiles of the normalized Wilshire 5000 quantiles occur at more negative values than the standard normal distribution. The high quantiles occur at values greater than the standard normal distribution. However, for quantiles near the median, the data does seem to follow the normal distribution. The plot is a graphical representation of the fact that extreme events are more likely to occur than would be predicted by a normal distribution.
Figure 2: The q-q plot for the normalized log-changes from the Wilshire 5000.
Summary

Using log-returns for stock price has the simplifying advantage of additivity. If we assume that log-returns are normally distributed, then the logarithm of the compounding return is normally distributed.

Two problems with using Brownian Motion for modeling stock prices varying over continuous time lead to modeling stock prices with the stochastic differential equation for Geometric Brownian Motion. The log-return over a period from $t - n$ to $t$ is $(r - \sigma^2/2)n + \sigma[W(t) - W(t - n)]$. The log-return is normally distributed with mean and variance characterized by the parameters associated with the security.

The modeling assumptions can be tested against data for the Wilshire 5000 Index. The Geometric Brownian Motion models using the parameters for the Wilshire 5000 Index resemble the actual data. However, deeper statistical investigation of the log-returns shows that while log-returns within 4 standard deviations from the mean are normally distributed, extreme events are more likely to occur than would be predicted by a normal distribution.

Sources

Algorithms, Scripts, Simulations

Algorithm

This script simulates the Wilshire 5000 Index over the period April 1, 2009 to December 31, 2014 using Geometric Brownian Motion with drift and standard deviation parameters calculated from the data over that same period. The resulting simulation is plotted over that 5.75 year time period. Set the drift $\mu$, variance $\sigma$, time interval $T$, and starting value $S_0$. Set the number of time divisions $n$ and compute the time increment $\Delta$. Compute a Wiener process simulation, then apply the exponential to create a Geometric Brownian Motion with the parameters. Plot the simulation on the time interval, or output the data to a file for later plotting or use.

Scripts

R

R script for Stock Market Model.

```r
mu <- 0.1682262
sigma <- 0.1722922
T <- 5.75
# length of the interval [0, T] in time units of years
S0 <- 8242.38

N <- 1448
# number of end-points of the grid including T
Delta <- T/N
# time increment ,

t <- seq(0, T, length = N + 1)
W <- c(0, cumsum(sqrt(Delta) * rnorm(N))) # Wiener process,

GBM <- S0 * exp(mu * t + sigma * W)

plot(t, GBM, type = "l", xaxt = "n", ylab = "Simulated Wilshire 5000 Index")
axis(1, at = c(0.75, 1.75, 2.75, 3.75, 4.75, 5.75), label = c("2010", "2011", "2012", "2013", "2014", "2015"))
```

Octave

Octave script for Stock Market Model.

```octave
mu = 0.1682262
sigma = 0.1722922
T = 5.75
# length of the interval [0, T] in time units of years
S0 = 8242.38

N = 1448
# number of end-points of the grid including T
Delta = T/N
# time increment ,

t = [0:1:N+1]' 
W = [0; cumsum(sqrt(Delta) * randn(N))]

GBM = S0 * exp(mu * t + sigma * W)

plot(t, GBM, type = "l", xaxt = "n", ylab = "Simulated Wilshire 5000 Index")
axis(1, at = [0.75:1:5.75], label = ["2010", "2011", "2012", "2013", "2014", "2015"])
```
mu = 0.168226;
sigma = 0.1722922;
T = 5.75;
# length of the interval [0, T] in time units of years
S0 = 8242.38;

N = 1448;
# number of end-points of the grid including T
Delta = T / N;
# time increment

W = zeros(1, N + 1);
# initialization of the vector W approximating
# Wiener process

t = linspace(0, T, N + 1);
W(2:N + 1) = cumsum(sqrt(Delta) * stdnormal_rnd(1, N));
GBM = S0 * exp(mu * t + sigma * W);

plot(t, GBM)
set(gca, 'xtick', [0.75, 1.75, 2.75, 3.75, 4.75, 5.75])
# initialization of the vector \( W \) approximating Wiener process
\[
\begin{align*}
\text{t} & = \text{ones} ( \text{N} + 1 ) * \text{zeros} ( \text{N} + 1 ) \rightarrow \text{xlinvals} ( 0, \text{T} ) ; \\
\text{Note the use of PDL dim 1 threading rule (PDL name for R recycling)} \\
\text{W} ( 1 : \text{N} ) . = \text{cumusumover} ( \sqrt{\Delta} * \text{grandom} ( \text{N} ) ) ; \\
\text{GBM} & = \text{S0} * \exp ( \text{mu} * \text{t} + \text{sigma} * \text{W} ) ;
\end{align*}
\]

# file output to use with external plotting programming such as gnuplot, R, octave, etc.
# Start gnuplot, then from gnuplot prompt
# plot "stockmarketmodel.dat" with lines;

SciPy [Scientific Python script for Stock Market Model.]

```python
import scipy
mu = 0.1682262
sigma = 0.1722922
T = 5.75

# length of the interval [0, T] in time units
S0 = 8242.38

# number of end-points of the grid including T
Delta = T / N
```

16
Problems to Work for Understanding

1. Show that

\[ \rho_i = \log(1 + r_i) \approx r_i, \text{ for } r \ll 1. \]
2. If standard Brownian Motion is started from a positive value $X_0 > 0$, write the expression for the positive probability that the process can attain negative values at time $t$. Write the expression for the probability that Brownian Motion can ever become negative.

3. Choose a stock index such as the S & P 500, the Dow Jones Industrial Average etc., and obtain closing values of that index for a year-long (or longer) interval of trading days. Find the growth rate and variance of the closing values and create a Geometric Brownian Motion on the same interval with the same initial value, growth rate, and variance. Plot both sets of data on the same axes, as in Figure 1. Discuss the similarities and differences.

4. Choose an individual stock or a stock index such as the S & P 500, the Dow Jones Industrial Average, etc., and obtain values of that index at regular intervals such as daily or hourly for a long interval of trading. Find the log-changes, and normalize by subtracting the mean and dividing by the standard deviation. Create a q-q plot for the data as in Figure 2. Discuss the similarities and differences.

---

**Reading Suggestion:**

**References**


I check all the information on each page for correctness and typographical errors. Nevertheless, some errors may occur and I would be grateful if you would alert me to such errors. I make every reasonable effort to present current and accurate information for public use, however I do not guarantee the accuracy or timeliness of information on this website. Your use of the information from this website is strictly voluntary and at your risk.

I have checked the links to external sites for usefulness. Links to external websites are provided as a convenience. I do not endorse, control, monitor, or guarantee the information contained in any external website. I don’t guarantee that the links are active at all times. Use the links here with the same caution as you would all information on the Internet. This website reflects the thoughts, interests and opinions of its author. They do not explicitly represent official positions or policies of my employer.

Information on this website is subject to change without notice.
Steve Dunbar’s Home Page, http://www.math.unl.edu/~sdunbar1
Email to Steve Dunbar, sdunbar1 at unl dot edu
Last modified: Processed from \TeX source on May 10, 2016