

This lab is about performing some exploratory data analysis for common stocks and stock indexes. First, a few things must be understood. Stocks and many other physical or economic systems are too poorly-understood to be modeled deterministically. Instead we model them in time or space by an appropriate stochastic process.

For stocks, the most common model is the log-normal diffusion model, known as geometric Brownian motion (GBM). Not wishing to confuse, the solution to the equation for time-evolution of this model is $X_i(t) = X_i(0)e^{\mu'_i \Delta t + \sigma_i \sqrt{\Delta t}}$. Although this model has consistently been shown to not truly represent the markets, in many respects it is “close enough.” The nice thing about GBM is that is parameterized just by two population parameters, its growth and variance, or $\theta = (\mu', \sigma)$, where $\mu'_i = \mu_i - \frac{\sigma_i^2}{2}$. These parameters are usually “quoted” as annual values, i.e. $\mu = .10$ is 10% per year.

In our analysis, we usually refer to the time ticks Δt as 1-day increments. The process is modeled in terms of price CHANGES (to get to normality!), i.e. $R_t = \frac{X_{t+1}}{X_t}$, from which we take the natural log to get $r = \ln(R)$. This r is modeled as Gaussian (Normal). For instance, if the price of the stock today is 100 and yesterday it was 95, then $R = 100/95 = 1.053$ (note that the percent increase is $R - 1$, or 5.3%), and its “mathematical” return $r = \ln(1.053) = .0516$, which is CLOSE to the “percent” return but not exactly equal to it. Some factoids are that $r \leq r_{pct}$, $r_{pct} \doteq r$ for small r , and $r_{\%} = R - 1 = e^r - 1$. However, we are mostly concerned with the basic definition $r = \ln(R) \doteq \text{Normal}$.

Once one has estimates for the population parameters, it is easy to simulate stock path progression and other things. The standard estimators for μ and σ^2 are $\hat{\sigma}^2 = \frac{S_r^2}{\Delta t}$ and $\hat{\mu} = \frac{\sigma^2}{2} + \frac{\bar{r}}{\Delta t}$, with $\Delta t = 1/252$ (1 day). The daily variance is especially easy to calculate, it's just the sample variance of the daily returns r . For daily data, $\hat{\sigma}_{annual} = S_{r_daily} \sqrt{252}$, recalling $\sigma = \sqrt{\sigma^2}$. The growth rate $\hat{\mu}_{daily} = \frac{\sigma_{daily}^2}{2} + \frac{\bar{r}}{1}$, with $\hat{\mu}_{annual} = \hat{\mu}_{daily} \cdot 252$.

For a typical stock or market index we might get $GBM(\hat{\mu}_{gbm}, \hat{\sigma}_{gbm}) = GBM(.082, .165)$, showing an annual growth rate of 8.2% and annual “volatility” of 16.5%. So for a stock with annual volatility of 100%, we might expect that in 95% of the cases the price of the stock could double within a year.

It turns out that the big indexes, such as the Dow or the SP500, have annual volatilities around 15.8%. Why is coincidental number so nice? Because the daily volatility $\sigma_{daily} = \sigma_{annual} / \sqrt{252} = .01!$ That means on a given day a 1% change is just 1-sigma!

One can calculate a z-score for a daily return as $z = \frac{r - \hat{\mu}}{\hat{\sigma}}$, which of course is $N(0,1)$.

From this one can quantify the likelihood of a particular return. For example, if a stock is $GBM(\frac{.10}{252}, \frac{.25}{\sqrt{252}})$ and we obtain a return $R=1.03$ (a 3% increase), $r = \ln(1.03) = .02955$

so $z = \frac{.02955 - .0004}{.0157} = 1.85$. This has a *p-value* of .064, which is NOT a significantly high return. We would expect this return once in 15.625 days ($1/p\text{-value}$), or .062 years between occurrences ($\text{days} \cdot \frac{1 \text{ yr}}{252 \text{ d}}$).

Market indexes are “averages” of individual stocks meant to define some reference benchmark. The two types are weighted arithmetic and geometric indexes. By far the most common are the arithmetic indexes. Popular brands are the price-weighted Dow Jones indexes, and the market-value weighted indexes such as the Nasdaq Composite and the Standard & Poor’s indexes. Less well-known are geometric indexes such as the Value Line, but these are gaining in popularity, due to their better mathematical properties.

The arithmetic indexes are variations on \bar{X} . In the case of the price-weighted Dow indexes, $I_{Dow} = \alpha \sum X_i = \alpha n \bar{X}$, or $I_{Dow} = \alpha' \bar{X}$. As of April 2004, the published Dow divisor of 0.144, gives $\alpha' = \frac{30}{.1409} = 212.92$. The geometric weighted index, is

$I_G = \alpha \left(\prod X_i^{w_i} \right)^{1/n}$, just the geometric mean of the components.

We will be constructing Dow-type and Geo-type indexes for a group of 4 stocks, which we will call the ISCO index. For each trading day we take the stocks prices and construct the index value, D-ISCO and G-ISCO. This is simple, as in

| date | A | B | C | D | "Dow" | "Geom" |
|--------|-------|-------|----|-------|---------|---------|
| 9/8/00 | 77.88 | 63.88 | 21 | 19.81 | 45.6425 | 37.9292 |

If we wanted to scale this to a common value A on a certain day, we multiply by our index value by A/I_0 . In the above example, if we wish to scale the indexes to 11,000 on day 1, we multiply each calculation by A/I_0 , or $11111/45.6425$ for the Dow-type, and $11000/37.9292$ for the geometric index.

Questions.

1. Calculate the annual and daily growth and volatility estimates for the DJIA and 4 stocks comprising the ISCO index. Note that the daily growth values will be pretty small.
2. Plot price charts for the DJIA and each component stock.
3. Perform a correlation and regression analysis for each ISCO stock and the DJIA. Try to make statements on what the correlations mean. Plot your regression lines.
4. Calculate and plot the D-ISCO and G-ISCO indexes, scaling to the value of the DJIA on 8/30/2000.
5. Find the correlation between the 4 component stocks and each of the ISCO indexes. Comment on this correlation with respect to those for the DJIA.
6. Would it be better to invest in the ISCO index or the DJIA? Why?
7. Would it be better to invest in the ISCO index or for an individual stock in the ISCO index? Why?
8. Suppose after the Bush re-election each of the ISCO stocks gained 10% return on the day. What are the p-values for these returns how often should such a return be encountered in terms of years (or centuries, or millennia).
9. Do your results in question 8 make you wonder about the Normality assumptions of the log of the daily returns?
10. What other questions does your analysis bring to mind?