Investment Management

Stephen R. Rush, PhD
Chartered Financial Analyst
Companion Website

March 31, 2018

Introduction

A competent financial analyst must be familiar with quantitative and qualitative methods for security selection, portfolio construction, and performance measurement. This booklet summarizes the basic tools needed for managing investments. It is meant for those who have already learned this information from a more complete source such as Investments by Bodie, Kane and Marcus or the CFA Candidate Body of Knowledge. This book is currently a work in progress and is not comprehensive. It complements the material presented in class and I will add more content as time permits.

The Introduction to Finance Analytics guide is written for the non-programmer with some knowledge of finance. It is meant to compliment my teaching and is not a general programming guide. I will focus on the R programming language but I will also include some information on python. In order to reduce the number of revisions to this document, I try to reference outside sources as much as possible. There are plenty of well written books and blogs available to those who want more assistance.

Your first step should be to work through the vignettes included on the CRAN package website and the examples included at the bottom of the documentation for each function. If you are one of my students, please do not hesitate to come to office hours even if you are learning R or python on your own. If you find that some code does not work, please email me.

This material is used to teach undergraduate finance students on Windows computers. For the most part, everything with the exception of Bloomberg exercises can be done on Linux and OSX computers. I aim to provide the best cross-platform method to accomplish a task rather than the most computationally efficient method. I also consider the most natural way for a finance student to address a finance problem using programming rather than the easier way for a programming student to address a finance application. Any suggestions should take these considerations into account.

Contents

Investments

Security Selection

"Two stocks diverged in the market and I-
I took the one with more cash flow,
And that has made all the difference."

-Noted security analyst, Robert Frost

Fundamental Analysis

An equity recommendation starts with buy, sell, or hold. This recommendation is supported by evidence from your macroeconomic forecast, industry analysis, financial statement analysis, and valuation. The top-down approach starts with the economy and works down to the company. The bottom-up approach starts with selecting a company and working up to the economy. A good analyst uses the top to inform where to start looking for a firm to do the bottom-up analysis. With limited time an analysis cannot do a bottom-up analysis on all stocks so the macroeconomic and industry analysis are used to restrict the number of firms that must be analyzed.

Macroeconomic Analysis

The National Bureau of Economic Research (NBER) is responsible for determining recession periods. Information about government debt is available from the (Federal Reserve).

The Yield Curve

Using expectations theory to calculate forward rates:

$f_{1x1} = [\frac{(1+i_2)^2}{(1+i_1)^1}] - 1$

where i_2 is the annualized 2 year rate, i_1 is the annualized 1 year rate, and f_{1x1} is the 1 year rate in 1 year.

The Taylor Rule

The Taylor Rule (1993) may be used to estimate the nominal short term interest rate based on inflation expectations and GDP growth. Bloomberg has a pre-built function.

Inflation Expectations

The yield to maturity on US treasury bonds represents the nominal risk-free rate. Treasury Inflation Protected Securities (TIPS) are US treasury bonds where the coupon increases with inflation. The yield to maturity on TIPS represents the real risk-free rate. The spread between UST and TIPS is the market's expectation of inflation. $\pi = [\frac{(1+UST)}{(1+TIPS)}] - 1$ Information about the money supply can be found at the (Federal Reserve).

Unemployment

The Bureau of Labor Statistics (BLS) is responsible for producing the unemployment rate statistics but the data is more easily accessed through the Federal Reserve database. The U-3 series is official headline unemployment rate. U-6 measures unemployment including people available but not looking for work.

Consumer Sentiment

There are multiple competing indices of consumer sentiment available but the Consumer Sentiment (University of Michigan) is widely used and easily accessible from the Federal Reserve.

Economic Indicators

Leading Economic Indicators

- Average weekly hours of production workers (manufacturing)
- Initial claims for unemployment insurance
- Manufacturers' new orders (consumer goods and materials industries)
- Fraction of companies reporting slower deliveries diffusion index
- New orders for nondefense capital goods
- New private housing units authorized by local building permits
- Yield curve slope: 10 yr UST - FF rate
- Stock prices, broad index
- Money supply (M2) growth rate
- Index of consumer expectations

Coincident Economic Indicators

- Employment on nonagricultural payrolls
- Personal income less transfer payments
- Industrial Production
- Manufacturing and trade sales

Lagging Economic Indicators

- Average duration of unemployment
- Ratio of trade inventories to sales
- Change in index of labor cost per unit of output
- Average prime rate charged by banks
- Commercial and industrial loans outstanding
- Ratio of consumer installment credit outstanding to personal income
- Change in consumer price index for services

Industry Analysis

"I kinda like it, crazy little thing called S.W.O.T."

-Consulting firm, Queen

SWOT Analysis

- Strengths
- Weaknesses
- Opportunities
- Threats

Porter's 5 Forces

- Substitutes
- Competitive Rivalry
- Threats of New Entrants
- Bargaining Power of Buyers
- Bargaining Power of Suppliers

Operating Leverage

$\frac{\%\Delta\text{Operating Income}}{\%\Delta\text{Sales}}$

Does selling 1% more increase profits by more or less than 1%?

Capacity Utilization

Industry Cycles

Demand Elasticity

Financial Statement Analysis

Financial Statement Analysis includes benchmarking the firm against other firms with a similar business model and of similar maturity. Key ratios to benchmark are: Gross Margin, Net Margin, ROE, ROA, ROIC, Debt/Equity ratio, and Current ratio. Ensure that all firms have similar fiscal year ends or use the trailing twelve months (TTM) of quarterly data to line up the ratios.

Common Sizing

In order to compare finacial statements across time and between firms, we must common size the income statement and balance sheet. Divide each line item on the income statement by sales in that period. Divide each line item on the balance sheet by total assets in that period.

DuPont Analysis

In a frictionless and efficient market, the price of a stock should increase with the return on equity. It is important to understand where the value is being created to evaluate whether it will continue.

$ROE = \frac{\text{Net Income}}{\text{Sales}} \text{ x } \frac{\text{Sales}}{\text{Total Assets}} \text{ x } \frac{\text{Total Assets}}{\text{Equity}}$

The first component is profitability defined as net margin. The second component is operational efficiency measured by Total Asset Turnover. The third component is financial leverage measured by the equity multiplier.

The Cash Conversion Cycle

"...better have my money"

-Accounts receivable analyst, Rihanna

The cash conversion cycle measures how effective a firm manages its working capital. Receivables are generated when the firm makes a sale but does not immediately receive cash. Typical B2B transactions include a period of 30 days to pay. Payables are generated when the firm orders raw materials from its suppliers. Working capital is determined by the amount of inventory needed and the difference in collections and payables. The longer the cash conversion cycle, the more working capital is needed to support the business.

CCC = Inventory Conversion + Collection - Deferral

- Inventory Conversion Period: Inverse of Inventory Turnover: $365\text{x}\frac{Inventory}{COGS}$
- Average Collection Period: Days Sales Outstanding (DSO): $365\text{x}\frac{Receivables}{Sales}$
- Payables Deferral Period: $365\text{x}\frac{Payables}{COGS}$

A growing cash conversion cycle indicates a greater need for capital and a reduction in the free cash flow available to the firm.

Credit Analysis

"If you got credit problems, I feel bad for you son. I got 99% leverage but my default probability ain't 1."

-Credit Analyst, Ice-T
(Copied by Junior Credit Analyst, Jay-Z)

Relative Valuation

"Two firms, both alike in valuation
In fair markets, where we lay our scene."

-Valuation expert, Bill Shakespeare

Relative valuation is based on the assumption that ratios are either constant or mean reverting. The less volatile a ratio is, the more useful it is for predicting stock prices. Some industries are valued well based on one or two ratios while others may be valued well using different ratios.
Time Period Conventions

- Trailing Twelve Months (TTM)
- Next Twelve Months (NTM)
- Next Fiscal Year (FY1)
- Last Fiscal Year or Fiscal Year End (FYE)

The exact data needed to forecast a stock price changes with the time convention used. Let's assume that we have TTM EPS and the current price. Calculate the TTM P/E ratio and multiply it by the NTM estimated EPS to find the price in one year.

$$Price_1 = \frac{Price_0}{EPS_{TTM}} * EPS_{NTM} \tag{1}$$

The same approach applies to the other multiples. In general, forward looking ratios are more stable and more reliable for predicting prices than backward looking ratios. There is a trade-off between the accuracy of the denominator: forecasted in a forward ratio versus reported in a trailing ratio. High growth or volatility of the denominator makes trailing ratios even less accurate.

Price to Earnings

Price to Earnings is sensitive to estimated EPS growth and perceived growth opportunities. Using the Dividend Discount Model, we can calculate a justified P/E ratio:

TTM P/E ratio $= \frac{(1+g))\text{Payout Ratio}}{r-g}$ and NTM P/E ratio $= \frac{\text{Payout Ratio}}{r-g}$

For very long term investing, there is Robert Schiller's Cyclically Adjusted P/E Ratio (CAPE).

Price to Sales

Divide four quarters of sales by the number of shares outstanding. Both values found on the income statement. Use estimated sales per share to obtain a one year price target. Price to Sales is sensitive to the business model's margins. This means that an investor is willing to pay more for a single dollar of sales when profit margins are high. It also means that comparing P/S ratios of different industries is not useful.

Price to Book

The book value of equity is the residual value of the firm if management were to liquidate the firm's assets and pay off all liabilities. The book value of equity represents a floor below which the price should not fall without an impairment in the value of a firm's assets. Unfortunately, all assets are not productive and may not have value during liquidation. Intangible items do not add value to the firm so we may need to adjust book value by subtracting items such as goodwill to arrive at tangible book.

Few analysts publish their estimates of the book value of equity. The current book value per share is calculated by dividing common equity found on the balance sheet by the number of share outstanding found on the income statement. Calculating the next book value per share is based on the fact that the book value of equity increases with retained earnings. Thus, the next period's book value per share is the prior period's book value per share plus the earnings per share in the next period minus dividends per share. $BVPS_1 = BVPS_0 + EPS_1 - DPS_1$

Price to Cash Flow

The relevant measure of cash flow is free cash flow to the firm per share. Some analysts only report estimates of Cash Flow from Operations so the ratio may be based on CFO per share. We can take to cash flow from operations and subtract capital expenditures to find the free cash flow to the firm. Divide by the number of shares outstanding to get FCFF per share.

EV to EBITDA

Enterprise Value to EBITDA is often used as a proxy for P/CF since EBITDA approximates cash flow. Using enterprise value rather than equity also normalizes for different capital structures (different proportions of debt and equity). It is a good ratio to use when comparing firms with different business models.

Growth Companies

"Mo' money, mo' problems."

-The Notorious Growth Stock Analyst

Growth companies often have extremely high market multiples. This creates a valuation risk where the company performs well and the market multiple decreases (multiple compression). Growth companies are also risky because their value depends more on assumptions about the future than value companies. To quantify the extent to which a company is valued on growth, we use the concept of present value of growth opportunities (PVGO).

$$Price = \frac{EPS}{r} + PVGO \tag{2}$$

Dividing both sides by the price, we can find the proportion of the price that is derived from continuing operations versus the proportion of the price that is derived from growth assumptions. Since we cannot observe PVGO directly, we can find the proportion of the share price that is based on growth assumptions.

$$1 - \frac{\frac{EPS}{r}}{Price} = 1 - \frac{EPS}{r * Price} \tag{3}$$

The higher this ratio becomes, the more volatile the stock price will be based on changes in growth assumptions.

We can also standardize the P/E ratio with earnings growth and calculate the PEG ratio. $\frac{P/E}{g}$ The PEG ratio is not as stable as other multiples because small changes in the growth rate affect both the numerator and denominator. The result is a ratio that does not distinguish well between two growth stocks but does distinguish well between growth stocks and non-growth stocks.

Forecasting Financial Statements

All discounted cash flow methods require some form of forecasting financial statements. There is no one correct way to forecast financial statements but there are better ways and worse ways. There will always be a trade-off between effort and accuracy. The method described is what I consider to be the most accurate for the amount of effort. The goal of forecasting financials is not necessarily to be the best forecaster. The model that you build requires that you understand how the firm operates and the history of the firm's performance. It is used to evaluate future expectations and to conduct scenario analysis in order to develop a valuation range rather than a single number.

Start with an open-ended forecast and add linkages in the statements as time permits. Retained earnings flow from the income statement to the balance sheet. Increase in cash flows from the statement of cash flows to the balance sheet. Capital expenditures flow from the statement of cash flows to the balance sheet. Changes in debt affect the statement of cash flows and interest expense on the income statement.

Income Statement

Use the percentage of sales method to forecast the income statement. Start with a top line sales forecast derived from historical sales, company guidance found in earnings call transcripts, industry analysis, and sell-side analyst estimates. Next, look for trends in the common sized income statement. Be careful of assuming that margins will expand or contract forever. Learn about the types of costs and the seasonality in industry cycles in order to adjust your expectations for peak margins. For each line item, multiple the percent of sales by the new forecasted sales number. Consolidate items where possible.

Balance Sheet

The key information needed from the balance sheet for valuation purposes is PP&E, Net Work Capital (NWC), and changes in debt. The next focus should be adding assumptions about cash management to ensure there is a positive cash balance that is reasonable but not being supported by excessive short-term debt. Check that short-term debt is not immediately paid off or growing at an unreasonable rate. Estimate a long-term debt increase if necessary. For most firms, it is safe to assume that most of the long-term debt will be refinanced rather than paid off. Short-term debt has a minimum for financing working capital and a seasonal component that fluctuates based on CFO preferences for short-term versus long-term financing.

Statement of Cash Flows

Most of the cash flow from operations can be estimated from the income statement. The next priority is the investment in NWC which is the difference in NWC from the previous period. This is information from the balance sheet. Capital expenditures are reported in cash flow used in investing and are almost always a decrease in cash flow. There is no standard way of using negative and positive signs so analyze the context before putting a number in a formula.

Discounted Cash Flow Valuation

"Free Cash Flow rules everything around me"

-Wu-Tang Financial

The value of any financial asset is the present value of all future cash flows. The problem becomes identifying the relevant cash flows and finding the appropriate discount rate. Quarterly cash flows are discounted using quarterly discount rates and terminal values based on quarterly data need a quarterly growth rate.

Cost of Equity

CAPM: $R_i = \beta(R_{Mkt} - rf) + rf$

- Estimate using a regression in risk-premium format: $R_i - rf = \beta(R_{Mkt} - rf) + \alpha$
- Assume that α will be zero in the long run. Use a 10 year US treasury bond yield as the risk-free rate. Be sure to match the frequency of the yield with the frequency of the data used to estimate β.
- The Market Risk Premium $(R_{Mkt} - rf)$ is mean-reverting so use a long term average
- Be sure to annualize the required return: $EAR = (1 + i)^n - 1$

Cost of Debt

The yield to maturity (YTM) is the cost of debt. This is often similar to but not the same as the coupon rate. To find the yield to maturity, we need market prices of bonds issued by the firm. Bloomberg reports yields and yields are also available on the FINRA website. Use an average of long term bond yields as the cost of debt.

Weighted Average Cost of Capital

Weighted Average Cost of Capital (WACC) $= w_E * r_E + w_D * r_D *$(1-tax rate)
where:
Capital is the market value of debt + market value of equity (price * shares outstanding)
and
$w_E = \frac{Equity}{Capital}$ and $w_D = \frac{Debt}{Capital}$

Free Cash Flow to the Firm

FCFF is cash flow that accrues to all providers of capital. The present value of all future FCFF is the value of the firm. Calculate FCFF for the last 3-5 years and forecast FCFF for the next 3-5 years.

Free Cash Flow to the Firm (FCFF) = EBIT(1-tax rate) + Depreciation - Investment in NWC - CapEx

	Historic Period		Current		Forecast Period	
FY_{-3}	FY_{-2}	FY_{-1}	FY_0	FY_1	FY_2	FY_3
$FCFF_{-3}$	$FCFF_{-2}$	$FCFF_{-1}$	$FCFF_0$	$FCFF_1$	$FCFF_2$	$FCFF_3$
						$Terminal_3$
			PV_0	PV_1	PV_2	PV_3

Calculate the terminal or horizon value using the last forecasted FCFF:

$Terminal_3 = \frac{(1+g)*FCFF_3}{r-g}$

The Weighted Average Cost of Capital (WACC) is the appropriate discount rate for FCFF. Use the historic data to begin to estimate the growth rate in FCFF. Adjust this based on your assumptions of future performance. Calculate the present value of $FCFF_0$, $FCFF_1$, $FCFF_2$, $FCFF_3$, and $Terminal_3$. Adjust the number of compounding periods if the current year is not the end of the fiscal year.

$PV_0 = \frac{FCFF_0}{(1+r)^1}$ $PV_1 = \frac{FCFF_1}{(1+r)^2}$ $PV_2 = \frac{FCFF_2}{(1+r)^3}$ $PV_3 = \frac{FCFF_3+Terminal_3}{(1+r)^4}$

The sum of these present values is the value of the firm. The equity value is Firm Value - Market Value of all Debt. Divide equity by the number of shares outstanding to find share price.

Free Cash Flow to Equity

FCFE is cash flow that accrues to equity holders. The present value of all future FCFE is the value of the firm's equity. Net borrowing is the change in short-term and long-term debt. A decrease in debt decreases the cash flow available to equity holders and vice versa. The interest expense reported on the income statement is often not a good approximation of the interest expense paid on debt. When a firm refinances it's debt, the interest expense will reflect the yield on new debt rather than the past yield on old debt. Interest expense may also include interest received from marketable securities. Instead of using interest expense directly, take the average yield on long term debt (the same as you need to calculate WACC) and multiply that by the sum of long-term and short-term debt.

Free Cash Flow to Equity (FCFE) = FCFF - Interest Expense * (1-tax rate) + Net Borrowing

Historic Period			Current	Forecast Period		
FY_{-3}	FY_{-2}	FY_{-1}	FY_0	FY_1	FY_2	FY_3
$FCFE_{-3}$	$FCFE_{-2}$	$FCFE_{-1}$	$FCFE_0$	$FCFE_1$	$FCFE_2$	$FCFE_3$
						$Terminal_3$
			PV_0	PV_1	PV_2	PV_3

Calculate the terminal or horizon value using the last forecasted FCFE:

$Terminal_3 = \frac{(1+g)*FCFE_3}{r-g}$

The required return on equity is the appropriate discount rate for FCFE. Use the historic data to begin to estimate the growth rate in FCFE. Adjust this based on your assumptions of future performance. Calculate the present value of $FCFE_0$, $FCFE_1$, $FCFE_2$, $FCFE_3$, and $Terminal_3$. Adjust the number of compounding periods if the current year is not the end of the fiscal year.

$PV_0 = \frac{FCFE_0}{(1+r)^1}$ $PV_1 = \frac{FCFE_1}{(1+r)^2}$ $PV_2 = \frac{FCFE_2}{(1+r)^3}$ $PV_3 = \frac{FCFE_3+Terminal_3}{(1+r)^4}$

The sum of these present values is the value the firm's equity. Divide equity by the number of shares outstanding to find share price. This estimate of the intrinsic value of a share occurs as of the last reported quarter. Use the required return on equity to move this price to the time horizon needed. Example: suppose the firm ended their last quarter two months ago and we want a one year price target:

Price Target = Intrinsic Value * $[(1 + r_e)^{1.1667}]$

Dividend Discount Model

$Price_0 = \frac{D_1}{r-g}$ where r is the required return on equity and g is the growth rate in dividends

Residual Income Model

Residual Income is what the firm generates in excess of its required return. To find the dollar value of required return, multiply the book value of equity by the required return on equity.

Residual Income = Net Income - Book Value of Equity*Required Return on Equity

Subtract this “equity charge” from net income to find residual income. The market value of equity is the book value of equity and all future residual income.

Equity Value = Book Value + PV of all future Residual Income

Residual Income is discounted using the required return on equity. Divide equity value by the number of shares to find the share price.

Equity as an Option

"...derivatives, how do they work?"

-Insane Clown Options Trader

Common equity can also be viewed as the residual value of firm assets after all liabilities have been paid. Since equity holders have limited liability, if liabilities exceed assets, then the equity holders walk away with no equity value. They cannot have a negative value. The Black-Scholes Option Pricing model is a continuous time implementation of the Binomial Option model with infinite steps.

$C_0 = S_0 N(d_1) - X e^{r\tau} N(d_2)$ where: $d_1 = \frac{ln(S_0/X)+(r+\sigma^2/2)*\tau}{\sigma\sqrt{\tau}}$ and $d_2 = d_1 - \sigma\sqrt{\tau}$

Variables

- C_0 = Call Option Price
- S_0 = Stock Price
- X = Exercise Price
- r = risk-free rate
- σ = volatility of the stock
- τ = T - t = time to expiration
- N(d_1) = the cumulative density of a standard normal distribution at d_1 (In Excel, use NORMSDIST function)

If common equity is a call option on the assets of the firm, we can rewrite the Black-Scholes variables with the following substitutions:
Corporate Finance Variables

- C_0 = Market value of the firm's equity
- S_0 = Market value of firm assets
- X = Market value of debt
- r = risk-free rate
- σ = volatility of the firm's assets
- τ = T - t = time to debt maturity

Quantitative Security Selection

"When you believe in things that you don't understand,
Then you suffer,
Superstition ain't the way."

-Quant visionary, Stevie Wonder

Arbitrage Pricing Theory

Ross (1976)

$R_i - rf = \alpha_i + \beta_1 RP_1 + \beta_2 RP_2 + ... + \beta_N RP_N$

Each β is called a factor loading. Each RP is a risk premium such as the market risk premium or value premium.

Fama-French (1996)

Adding factors to the original single factor CAPM lowers risk-adjusted returns (α) and unexplained variance of returns (ϵ).

$R_i - rf = \alpha_i + \beta_{Market}(R_{MKT} - rf) + \beta_{HML} RP_{HML} + \beta_{SMB} RP_{SMB} + \epsilon_i$

Three Risk Factors

- Market sensitivity (same as CAPM)
- Value (High B/M - Low B/M)
- Size (Small firms - Big firms)

Other factors: Momentum (trailing 12 month return) and Liquidity
Calculate the required return on equity

- Estimate factor loadings using a regression
- Estimate risk premiums starting with a historic average for each risk factor
- Multiply the factor loading by the risk premium and take the sum plus the risk-free rate to estimate the required return
- Be sure to annualize the required return: $EAR = (1 + i)^n - 1$

Compare t-statistics of alpha between securities. Select the security that produces the most reliably high alpha.

Statistical Arbitrage

Form a zero beta portfolio that generates positive alpha.

Trading

Pairs Trading

Use the correlation between two securities to short one and go long the higher returning security and generate low volatility returns.

Cointegration

Portfolio Management

Portfolio Construction

Investment Policy Statement

The Investment Policy Statement (IPS) states the objectives of the portfolio, risk preferences, tax considerations, asset class restrictions, portfolio benchmark, and any other considerations.

Modern Portfolio Theory

Markowitz Mean-Variance Optimization is the dominant framework that investors use to build portfolios. The idea behind Modern Portfolio Theory (1952) is to maximize the portfolio Sharpe ratio by over-weighting securities with high Sharpe ratios and under-weighting those with low Sharpe ratios taking into account the correlations between securities. The method requires knowing expected returns and forecasting the covariance matrix. Since these are estimated, the resulting portfolio is only as good as the estimates used to build it.

Portfolio Calculations

Define Matrices:

$$r = \begin{bmatrix} return_1 \\ return_2 \\ \dots \\ return_n \end{bmatrix} \qquad w = \begin{bmatrix} weight_1 \\ weight_2 \\ \dots \\ weight_n \end{bmatrix} \qquad w' = \begin{bmatrix} weight_1 & weight_2 & \dots & weight_n \end{bmatrix}$$

Calculate the return and standard deviation of the portfolio:

$$\text{Portfolio Return} = r_p = w'r \qquad \text{Portfolio Variance} = \sigma_p^2 = \text{w'}\Sigma\text{w}$$

Find the covariance matrix Σ in Excel using Data Analysis. One dimensional matrices are arranged vertically. The transpose function changes a vertical array to a horizontal array and vice versa. Use the mmult function in Excel for matrix multiplication and use Control-Shift-Enter and the end of the formula. Calculate the Sharpe ratio of the portfolio using a risk-free rate that matches the frequency of the returns.

Portfolio Optimization

Once the formulas are set up, use the Solver function in Excel to maximize the Sharpe ratio by changing the weights of the portfolio. Constrain the weights to sum to 100%. An optional constraint is limiting weights to positive values to simulate a long-only portfolio. Sometimes the optimal weights result in an unrealistically high weight in one position with a corresponding negative weight in another. This can be fixed by adding an additional constraint that individual weights must fall between arbitrary limits like 200% and -200%. In practice, these limits are set by leverage restrictions.

Enhanced Indexing

Tracking error tends to increase faster than alpha. To maximize the information ratio, take a few small deviations from the benchmark and use the rest of the portfolio to replicate the benchmark. This results in a small concentrated portfolio combined with the benchmark. This structure is operational efficient because it is feasible for a single analyst to cover the active positions rather than requiring multiple analysts covering all the positions in the portfolio.

Treynor-Black

"Treynor-Black is the new black"

Treynor-Black (1973)

Active weight $w_i = \frac{\frac{\alpha_i}{\sigma_{\epsilon_i}}}{\sum \frac{\alpha_i}{\sigma_{\epsilon_i}}}$

Black-Litterman

"Paint It Black-Litterman"

-Rolling Stones Financial Management

Black-Litterman (1992)

Style Analysis

Replicate the performance of a portfolio with unknown holdings.

Rebalancing

Limit Prices

There are nearly no instances where using a market order is appropriate. Use a limit order for every transaction even when you want execution within the next second. Setting the limit price depends on the goal of trading and the time horizon. Assuming that the time horizon is more than a few months and that we are not trading during an earnings release or other important information event, we can afford to wait a few days for execution. The distribution of returns is not a normal distribution. We will use Chebyshev's Inequality to set a limit with a reasonable likelihood of execution. This approach is similar to Bollinger Bands. Ideally, we would want to use intraday prices but usually daily prices are available. Calculate the returns on the last month (22 trading days) of prices. If possible use Volume Weighted Average Prices (VWAP) rather than daily close. Calculate the mean and standard deviation of these returns. We want a probability that the next day's return exceed our limit to be no more than 25%. This means that our trade execution will not only capture a day when the price moves in our direction but also a significant move in our direction. Our expected return, μ is the average return.

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2} \tag{4}$$

If we want to execute a buy order, we want the price to be low. Set a limit where we would expect no more than 50% of the observations to fall outside of our interval $|X-\mu|$ so that no more than 25% of the observations will be below X-μ (one tail). When $k = \sqrt{2}$, then the probability inside the full interval is at least 50% and the probability below our limit is no more than 25%. Solving for our target return, we get:

$X = \mu - \sqrt{2}\sigma$ for a buy limit and $X = \mu + \sqrt{2}\sigma$ for a sell limit

Multiply the most recent price by one plus the return calculated above to find the limit price. Limit Price = (1+X)Current Price

Kelly Criterion

The Kelly Criterion (1956) was published shortly after the work of Markowitz. The first application was to betting systems but it has been applied to portfolio rebalancing.

$$w = \frac{p}{R_D} - \frac{q}{R_U} \tag{5}$$

Variables:

- p = probability of success
- q = 1-p = probability of failure
- R_U = return if successful
- R_D = return if fail
- w = weight in the portfolio

In practice, investors use a variation called "fractional Kelly" in order to reduce changes in weights. The problem with the Kelly criterion is that it assumes that individual security payoffs are independent and that payoffs may be simplified to binary outcomes.

Performance Measurement

Portfolio Risk

Portfolio Beta

Estimate portfolio β in the same way that we estimate the beta of an individual stock. For portfolios with a small number of assets, portfolio β is also the weighted average β of the portfolio holdings. The weighted average is problematic because constituent weights are constantly changing.

Risk-Adjusted Return

"Get alpha or die tryin"

-Portfolio Manager, 50 Cent

Jensen's alpha is the average return per period after accounting for risk. The magnitude and statistical significance of alpha is dependent upon the risk factors that are assumed to be relevant. In general, alpha will decrease as the number of risk factors increases. Estimate portfolio alpha in the same manner as estimating the factor loadings of a single stock. If you already used a regression to estimate the portfolio β the constant term, α, is Jensen's alpha.

Tracking Error

Two types of tracking error: $TE_1 = \sigma_{R_p - R_B}$ and $TE_2 = \epsilon$

The first type of tracking error is the standard deviation of the difference in portfolio returns and benchmark returns. The second type is the error term of the regression: $R_P = \beta R_B + \alpha$

Portfolio Evaluation

Sharpe Ratio

Use the Sharpe Ratio to choose between competing investment opportunities. Divide the portfolio risk premium by the standard deviation of the portfolio return.

$\frac{R_p - rf}{\sigma_p}$

Treynor Measure

The Treynor measure is useful when evaluating performance within a larger portfolio. Divide the portfolio risk premium by the CAPM beta of the portfolio.

$\frac{R_p - R_{rf}}{\beta_p}$

M^2 Measure

$R_p - R_m = \sigma_m(SR_p - SR_m)$

Useful for comparing strategies with very different volatilities

Information Ratio

The Information Ratio is the active return divided by the tracking error.

$\frac{\alpha}{\sigma_{alpha}}$

Performance Attribution

Brinson Attribution

Decompose portfolio return into a security selection component and an allocation component.

Risk Factor Analysis

Identify risk factor loadings generating the portfolio return in order to predict sensitivities to future market events.

Introduction to Financial Analytics

Financial Data

Data Sources

Free Data Sources

In industry there are no data police that ensure standardization of data types. Think before you calculate. Oftentimes a percentage may be shown as a whole number rather than a decimal. Look at your data to determine whether is is reasonable that a risk-free bond returns 50% in one month and adjust accordingly. Make sure the frequency of your data matches before combining it. Do not use a monthly stock return with an annualized risk-free rate.

Federal Reserve

Federal Reserve Economic Data FRED contains a wide range of macroeconomic and market data. Use constant maturity series to find US Treasury bond yields. Yields are always quoted as annualized rates. Be sure to adjust them as necessary.

SEC

The US Securities and Exchange Commission (SEC) maintains the EDGAR database to distribute public company information to the public. Form 10-K has annual financial statements and form 10-Q has quarterly statements. This data is then distributed to premium vendors such as Reuters and Bloomberg. Finally, websites like Yahoo Finance distribute a version of the Reuters data.

Fama-French Factors

Data downloaded from the Ken French Data Library is reported in whole number percentages. Divide values by 100 to convert to decimal form. The risk-free rate here is already in monthly terms.

Bloomberg

Disclaimer: Bloomberg updates their interface every few months without notification. It is virtually impossible to write a guide with every key stroke necessary to get the desired results. Type the <help> button twice to chat with a live representative 24/7.

Login

Create a Bloomberg Login if you have not done so already. You will need a mobile phone. This login will be tied to the Bloomberg terminals at BGSU. When you leave BGSU you will need a new Bloomberg login at your company.

Economic Analysis

Type <ECON> to get to main Economics menu
Taylor Rule <TAYL> for predicting Federal Reserve actions

Industry Analysis

Type <BI> for Bloomberg Intelligence
Select your sector from the left side of the screen
Select an industry within your sector
Select <Industry> to see some reports on industry trends
Use the other sections in the Data Library to develop and support your investment thesis
Type <EQS> for Equity Screening
Use metrics such as ROA, P/E, Cash Flow growth, etc.. to search for possible investment opportunities

Company Information

Type in a company ticker and press the <Equity> button or start typing the company name and select from the dropdown menu
Select the ticker under the "Security" heading
Select "Relative Value" or type <RV> after you are looking at the security
Select the "Comp Sheets" tab
Use the Settings dropdown menu near the top of the screen to adjust the list of comparables
Type <CN> to get company news stories
Type <FA> for Fundamental Analysis
Type <CF> for company SEC filings
If you are trying to export to Excel and you only see formulas or #NAME in the cells:
Settings → Report and Excel → Change setting from "Formulas" to "Values"
Use the <HCP> function to display a list of dates, prices, and percentage changes

Portfolio Management

Display the BGSU Student Managed Fund performance relative to the S&P 500:

- Type <PORT> for the portfolios linked to your account
- Click on the BGSU Student Managed Fund if it is available.
- Select the Performance tab, then Total Return
- Click on the "Value" button above the graph.
- Right-click on the graph and select Table View
- Click on "Actions" at the top of the screen and select "Generate Report (Bloomberg)"

- Choose “Current Tab (Unformatted xls)” to export the portfolio values to Excel

Display Portfolio Holdings

- Select the SPX Index in the 2nd dropdown menu
- Change the dates to the relevant range, hit “Enter”
- Use the Holdings tab to display positions and weights in the portfolio

Exporting Data

Click on “Settings” and “Global Settings” and make sure that Excel Export Mode is set to “Static Data”
Click on “Output” and select “Excel”

Other Vendors

There are many good quality alternatives to Bloomberg. The most used in industry are FactSet and Reuters. In general, good quality data is going to cost more than a retail investor or small advisory business can afford. Non-financial companies often share a single terminal with an entire department.

Capital IQ

Capital IQ is available through the BGSU library and may be accessed from off-campus. It offers reliable, standardized financial statements and price data.

Quandl

Financial Analytics

Setup R

"Start me up"

-Help desk technician, Mick Jagger

Install R

Download the latest version of R for Windows, Mac, or Linux from CRAN. You can accept all the default values here.

Install RStudio

It is much easier to work with a programming language from within an Integrated Development Environment (IDE). My preferred IDE for R is RStudio. Once it is installed, you can use the bottom right pane to search for help on how to use a specific function. The top right pane will show you the objects in memory. The bottom left pane is the output from your current R session. The top left pane is for working with scripts.

Install on a USB drive

While most people will install R on their computer, we will also install the Windows version of R on a USB drive. This will allow us to use our own version of R on a Bloomberg terminal. This will not require formatting the USB drive so you can use any USB drive with at least 1GB of free space.

Download the Windows installer for both R and RStudio. Install R in a directory named R on your USB drive. Windows Defender may mistakenly identify the uninstall program as malware. You may ignore this file and continue with the installation. I installed R version 3.3.2 in a directory named R on my USB drive which is letter E. In order to test the program, I double click E:\R\R-3.3.2\bin\i386\Rgui.exe for the 32 bit version or E:\R\R-3.3.2\bin\x64\Rgui.exe for the 64 bit version.

Next install RStudio in the R directory on your USB drive. Start RStudio by double clicking E:\R\RStudio\bin\rstudio.exe. A dialog box will ask you to select your version of R. Browse to either E:\R\R-3.3.2\bin\i386\or E:\R\R-3.3.2\bin\x64 for either the 32 bit or 64 bit versions of R.

R Session

Objects loaded into memory or created within R are lost when you quit RStudio unless they are explicitly saved. To remove a specific object like a data.frame called "data" without restarting your R session, use the command `rm(data)`. When working with large data sets, be mindful of how much RAM your R session is using. This can slow down calculations without proper planning.

Packages

R comes with many functions built into base R. Additional functions can be added by installing packages. Let's install the package "readr". Type the command `install.packages('readr')`. To use the package in a script, use the command `library(readr)`. This command only needs to be used once per session to access the package.

Setup Python

Even though python is used on almost every system, you should use a separate version for data analysis to avoid problems with versions and dependencies.

Install Python

Anaconda and Enthought Canopy are two distributions of python. Both have versions that are available for free. Choose one to install on your system.

Libraries

Additional functions can be added to python by installing libraries. Use the python distribution software (Anaconda or Canopy) to install libraries. Once the library is installed, it must be loaded each time you start a session in python. Use the `import` command to load a library. The following command loads the Pandas library:

```
import pandas as pd
```

Pandas

The best tool for working with finance data in python is the Pandas Data Analysis Library. This software simplifies many common tasks so I will often refer to the Pandas documentation. Since Pandas is under active development, I cannot provide a direct link but everything is fairly easy to find using the index on the left side of the documentation.

Scripts

A script is a text file with commands that can be interpreted by a programming language. R scripts should be saved with the ".R" extension; python scripts should be saved with the ".py" extension. In RStudio, you can run the entire script by clicking on "Source" or you can run a single line by clicking on the line and "Run". When writing your script, it is important that you leave comments that explain how your code works so when you open it a year from now, you will understand how what it does. Use the hash sign to begin a comment line that will not be run.

Data Frames

A data.frame is similar to a spreadsheet. Each column may contain different types of data. The contents of a csv file are converted to a data.frame when using the readr package. Suppose we have a data.frame called "data" with 5 rows and 3 columns. We want to name the columns: Date, Ticker, and Price.
`colnames(data)=c('Date','Ticker','Price')`
Use `str(data)` to verify that "data" is a data.frame and to verify that the column data types are correct. Use `summary(data)` to get summary statistics on each column of the data.frame.

Subset a Data Frame

If we want to select just the Price column of "data": `data[,'Price']`
If we want the 4th row of the Price column: `data[4,'Price']`
Let's take the 4th and 5th rows of the Price column: `data[4:5,'Price']`
...or the 2nd and 4th rows: `data[c(2,4),'Price']`
Drop the entire Date and Ticker columns and just keep the remaining columns:
`subset(data,select=-c(Date,Ticker))`
Select only the Ticker and Price columns: `subset(data,select=c(Ticker,Price))`

Logical Operations

`==`	equal to	`!=`	not equal
`>`	greater than	`<`	less than
`>=`	greater than or equal to	`<=`	less than or equal to
`is.na()`	gives TRUE is the value is NA	`is.inf()`	gives TRUE is the value is infinite

Conditional Subset

Choose all observations where Price is greater than or equal to 10: `data[data[,'Price']>=10,]`
Maybe we want just Prices greater than or equal to 10 with no other columns: `data[data[,'Price']>=10,'Price']`

Assigning Values to a Data Frame

Changing an existing value or values in a data.frame is as easy as subsetting. Setting the Price of the 3rd row to 20 can be done with `data[3,'Price']=20`

Suppose you want to add a column that is a count of the number of records. First, find the number of rows using `nrow()`. Next, create a sequence from 1 to the number of rows using `seq()`. Then, assign it to a new column called "Count" in the existing data.frame. The whole thing on one line looks like this:
`data[,'Count']=seq(1,nrow(data))`

Operations on a Data Frame

Operations are typically applied to each item in the data frame (dot product). The command `data*2` would multiply every number in the data frame by 2. For matrix multiplication, use the `%*%` operator.

Import/Export

It is a good idea to work with csv files because they can easily be shared between programs and people that don't use R. The downside to csv files is that they are slow to read into R. For faster access, use binary files in .RData format.

Reading and Writing to CSV Files

R

Base R comes with a `read.csv` function that is perfectly fine for small files. I recommend getting familiar with the readr package because it will be necessary to read csv much faster when you work with larger data sets.

First load the readr package with `library(readr)`. Next we need to find the path to the file. If you are working on a Windows computer, it will look something like "C:/username/folder/data.csv". If you are using Mac or Linux, it will look like "/home/user/folder/data.csv". In RStudio, select the Help tab in the bottom right panel and type "read_csv" into the search box. This will bring up the documentation for the function.

When we read the file data.csv into memory, we need to assign it to a variable.

```
data=read_csv(file="C:/username/folder/data.csv",n_max=1000)
```

This will read the first 1,000 rows of data.csv into a data.frame called "data".

To write a data.frame to a csv file, use the `write_csv` command. We need to specify that we are writing the "data" data.frame and location where the file should be written.

```
write_csv(data,"C:/username/folder/Saved.csv")
```

This command will overwrite any existing file in that location unless you set the option `append=TRUE`.

Python

Refer to the Pandas manual for IO Tools. Look for the `read_csv` function and related functions.

RData Files

Binary files are much faster to read and write. To save multiple objects in a binary file, use the `save()` command. Let's say that we have two data.frames "data" and "test" in memory and we want to save them to a file called "MyFile.RData".

```
save(data,test,file='MyFile.RData')
```

To read it back into memory use the `load()` command.

```
load('MyFile.RData')
```

Other Formats

A few packages help import and export data saved in other formats.
readxl read and write directly to Excel files
haven work with files formatted for SAS, Stata, and SPSS
foreign work with files formatted for Minitab, SAS, SPSS, Stata, and more

Loops

> "Ain't no party like a recursive party cuz there ain't no party like a recursive party cuz..."
>
> -Noted mathematician and computer programmer, Coolio

If/Else Statements

Conditional Statement

```
if() {
} else {
}
```

The foreach Package

Foreach loops a set of instructions so that a large problem can be solved in small chunks. The package can also recombine the results of each chunk and assign it to a variable. The foreach package can be used with either single core or multicore processing. When writing a foreach function, it is a good idea to test everything using a single core before utilizing parallel processing in order to make troubleshooting easier.

The first argument in foreach is called the iterator. Foreach allows for many types of iterators but it is usually easier to start with a numerical sequence as your iterator. Let's assume you want to repeat a set of instructions 5 times and save the output in a variable called "results". Your iterator should be `i=1:5`. The commands that are part of the foreach loop should be included in braces like this:

```
results=foreach{i=1:5} %do% {
   do something here
   do some more stuff
}
```

We usually take some chunk of data, apply a set of functions to it, then aggregate the results. The `.combine` option defines how the results will be aggregated. If our results are a data.frame and we want to stack the data.frame results into a single long data.frame, we would set `.combine=rbind`. If multiple results are created in our foreach loop, we need to define which variable has the results to be aggregated. In this example we have created a data.frame called "temp2" that has the results from each loop.

```
results=foreach{i=1:5,.combine=rbind} %do% {
   temp1=do something here
   temp2=do some more stuff
   return(temp2)
}
```

Data Formatting

Pivot Table

R

There are many ways to do a pivot table in R. The best method at the time of this writing is to use the tidyr package.

```
library(tidyr)
spread(data,c,level)
```

Python

```
df.pivot
df2.reset_index()
```

Dealing with Dates

First identify what format the data that you want to convert to dates is in. Oftentimes dates may be read as character strings or integers. You can check with the `str()` command. Many functions use abbreviations for the different date components. Here are some of the most common:

%Y - four digit year
%y - two digit year
%m - two digit month
%b - three letter month
%d - two digit day number
%H - two digit hour
%M - two digit minute
%S - two digit second

Some useful functions are `as.Date` and `as.POSIXct` in base R and `as.yearmon` in the zoo package that comes with xts.

The lubridate Package

More info at the CRAN page

The xts Package

The xts package is very useful when you have all numerical data with one row for each datetime. If you need to do a lot of time subsets or resampling, you should learn this package. More info at the CRAN page.

Time Series Analysis

"Time keeps on slippin', slippin', slippin'...into the future."

-Noted forecaster, Steve Miller

Time Series analysis is a special animal that I may update later. For now, refer to the CRAN Task View. Some notable packages are:

The forecast Package

Use this package for ARIMA modeling. More info at the CRAN page.

The vars Package

Use this package for Vector Auto Regressive models. More info at the CRAN page.

The MTS Package

Use this package for Multivariate Time Series models. This package is by Ruey Tsay who has also written a number of good books for MBA students on financial time series analysis. More info at the CRAN page.

The rugarch Package

Use this package for GARCH models. More info at the CRAN page.

Finance Data

The quantmod Package

The quantmod package is great for pulling in data from the Federal Reserve, Yahoo Finance, and Google Finance. Let's suppose you want to make a chart of the last 36 months of the following interest rates: 1 year U.S. Treasury, 10 year TIPS, 30 year fixed mortgage, Prime, Moody's Aaa Corporate, and Moody's Baa Corporate:

```
library(quantmod)
getSymbols("DGS1",src="FRED") #1 Year US Treasury Daily Constant Maturity rate
getSymbols("DPRIME",src="FRED") #Daily Prime Loan Rate
getSymbols("DAAA",src="FRED") #Daily Moody's Aaa Corporate Bond Yield
getSymbols("DBAA",src="FRED") #Daily Moody's Baa Corporate Bond Yield
getSymbols("MORTG",src="FRED") #30 Year Conventional Mortgage Rate
getSymbols("FII10",src="FRED") #10 Year TIPS Constant Maturity
# Chart Data
layout(matrix(1:6, nrow=3))
chartSeries(last(DGS1,'36 months'),name="US Treasury 1 Year Constant Maturity",layout=NULL)
chartSeries(last(FII10,'36 months'),name="10 Year TIPS Constant Maturity",layout=NULL)
chartSeries(last(MORTG,'36 months'),name="30 Year Fixed Mortgage Rate",layout=NULL)
chartSeries(last(DPRIME,'36 months'),name="Prime Loan Rate",layout=NULL)
chartSeries(last(DAAA,'36 months'),name="Moody's Aaa Corporate Yield",layout=NULL)
chartSeries(last(DBAA,'36 months'),name="Moody's Baa Corporate Yield",layout=NULL)
```

Another example shows the rolling correlation of the S&P 500 ETF and the Dow with price data from Yahoo Finance.

```
library(quantmod)
getSymbols(c("SPY","^DJI"),src="yahoo")
data=data.frame(SPY[,6],DJI[,6])
data=as.xts(data,order.by=as.Date(row.names(data),"%Y-%m-%d"))
c1=rollapply(data,65,cor,by.column=F)
Correlation=c1[,2]; chartSeries(Correlation)
```

Data on precious metals is available from Oanda. This example charts gold, platinum, palladium, and silver.

```
# Charts Gold prices in USD from Oanda
require(quantmod)
#fetch data
getSymbols(c("XAU/USD","XPT/USD","XPD/USD","XAG/USD"),src="oanda")
#setup the plot area, each TA is another graph
layout(matrix(1:4, nrow=2))
#charts to be included
chartSeries(XAUUSD,name="Gold (.oz) in USD", layout=NULL)
chartSeries(XPTUSD,name="Platinum (.oz) in USD", layout=NULL)
chartSeries(XPDUSD,name="Palladium (.oz) in USD", layout=NULL)
chartSeries(XAGUSD,name="Silver (.oz) in USD", layout=NULL)
```

For more information on quantmod, refer to the documentation on CRAN page and the website for quantmod.

Using Bloomberg in R

The blpapi package is the best place to start with Bloomberg. There is a brief introduction here. Bloomberg charges for data using a proprietary algorithm which means we cannot predict with certainty how much data can be used. If you use too much data or make too many data requests, the terminal will have data access suspended. In a university lab, this may result in one of the terminals becoming unusable for several months. Please be careful not to download excessive amounts of data.

For more information on blpapi, refer to the documentation on CRAN page and the GitHub website for blpapi.

Econometrics

CRAN Task View

"Everyone who confuses correlation with causation ends up dead."

Ordinary Least Squares

Let's estimate a stock's beta and alpha using the CAPM. First prepare the data by calculating stock returns and market returns. Subtract the risk-free rate from both and form a data.frame called "urdata" where ExRet is the excess return on the stock and MktRP is the risk premium on the market. Next, specify the formula for the estimated model and call it "m2". Use the `lm()` function to estimate the model and assign the results to a variable called "ols". Finally, use `summary()` to see the regression results.

```
m2=ExRet~MktRP
ols=lm(m2,data=urdata)
summary(ols)
```

If you are using a lot of fixed effects, try using `coef(ols)[1:5,]` to get the first 5 rows of coefficients.

Robust Regression

Robust regression is used to minimize the influence of outliers on the estimates of beta. Load the package with `library(robustbase)`. First get the standard OLS coefficient estimates and sigma estimate then feed those estimates into `lmrob()`.

```
m2=ExRet~MktRP
ols=lm(m2,data=urdata)
model=lmrob(m2,data=urdata,init=list(coefficients=coef(ols),scale=summary(ols)$sigma))
summary(model)
```

More information on the CRAN page and robust regression on Wikipedia.

Panel Regression

R

More information on the CRAN page

Python

Assume that you want to run a panel regression with firm and time fixed effects and double-clustered standard errors. The code below shows the setup assuming that you have all the data in a data frame named "data", the variable "Date" is formatted correctly, and each firm has a unique identifier as the variable "FirmID". More information on the GitHub documentation.

```
data=data.set_index(['FirmID','Date'])
import linearmodels
from linearmodels import PanelOLS
ols = PanelOLS(data[['DependantVar']],data[['X1','X2']],entity_effects=True,time_effects=True)
out=ols.fit(cov_type='clustered', cluster_entity=True,cluster_time=True)
out
```

GMM

The Generalized Method of Moments is a method to estimate the model betas without necessarily using the same assumptions as OLS. More information on the CRAN page and Wikipedia.

Higher Performance in R

"I need more cowbell"

-Noted computer scientist, Christopher Walken

R is not always the best tool for the job. Sometimes old command line tools have better performance. I suggest learning the following tools in this order:

- File Operations: cp, mv, rm (Type cp -help for syntax and options)
- nano: simple text editor (Use Ctrl as the meta key so Ctrl-x exits the program)
- rsync: quickly copy large amounts of files
- wget: fetch files from the Internet
- curl: more advanced Internet functionality
- gawk: manipulate text files
- sed: manipulate text output
- irssi: text based IRC client

tmux: use multiple command line sessions in a single terminal

I love monitors but they suck power and take up too much space. Using a terminal multiplexer like tmux allows you to cram more info into a smaller amount of real estate. Use Ctrl-b to issue commands.

Secure Shell: use the command line of another computer over the network

Be sure to install openssh on Cygwin or Linux. OS X comes with it. Syntax is: ssh user@serverURL Use "ssh-keygen -t rsa" once to generate a key and "ssh-copy-id user@serverURL" to put the public key on the server. This will allow logins without typing your password every time.

mosh: make your sessions robust to network disruptions

Do you like closing your laptop for a while and then resuming exactly where you left off? Mosh does this for secure shell sessions so you don't need to reestablish your connections manually. One of the great tools that "just works".

The data.table Package

A data.table is an improved data.frame. Many operations execute faster and use less RAM. The trade-off is that you need to learn a new syntax to use it. There is a good introductory blog post on the package here. There is a paid DataCamp online course designed by the programmers of the data.table package. The DataCamp notes are available here for free. The CRAN page also has more examples.

The Rcpp Package

When speed really matters and you have the programming expertise, you can embed C++ code with R. Many R packages already incorporate C++ code so you will likely need this package installed but will not use it directly. For more information, see the Rcpp CRAN page.

Parallel Processing or Single Processing?

In many cases, the overhead associated with parallel computing result in slower execution time than single core processing.

Parallel Computing

"MOAR cores!"

I will cover two methods for using multiple processors. The first uses all the cores on your local machine and the second is for using multiple machines together. When running long jobs that use multiple processors, it is a good idea to keep one core in reserve to run the computer's operating system and ensure that your system does not freeze. R has many more capabilities than are mentioned here. For an updated list of packages related to parallel computing, see the CRAN Task View.

Running foreach in Parallel

Recall from the section on loops, that a task may be iterated using the i index and the results compiled.

```
results=foreach{i=1:5,.combine=rbind} %do% {
   temp1=do something here
   temp2=do some more stuff
   return(temp2)
}
```

To run this loop in parallel, register a parallel back-end and change `%do%` to `%dopar%`. If the functions in your loop are part of a package, you must specify the package using the `.package` argument when running the loop in parallel.

For more information, see the CRAN page and the examples included in the documentation.

Parallel Back-ends

"That back-end though"

In order for foreach to run in parallel, you need to register a parallel back-end. These are my preferred parallel back-ends but there are others available. You should start with doParallel and only use doRedis once you are certain that you require more processors on a single foreach loop. Usually, it is less work to simply use more machines running separate R sessions where each session uses doParallel to fully utilize the processing power of that machine.

doParallel

The doParallel package is for using multiple processors on a single machine. I prefer it to other back-ends because it runs on Windows, Mac, and Linux which makes your code fully portable. First load the package `library(doParallel)`. Next use `detectCores()` to find how many processors R can detect in your machine. Finally, use `registerDoParallel` to allow R to use the appropriate number of cores. If you want to keep one processor to run the operating system and use the rest, simply use the command `registerDoParallel(detectCores()-1)`. Verify that doParallel has the correct number of cores with `getDoParWorkers()`. For more information, see the CRAN page.

doRedis

The doRedis package is for using multiple machines over a network. I like this package because it is tolerant of network communication errors and individual machine failures. This is a package where you really need to read the documentation on the CRAN page. Before opening R, make sure you have installed a Redis server on one of your machines. This machine will become the master for all the slave processes. The documentation will explain how to do this.

Big Data

"I like big data and I cannot lie..."

-Noted data scientist, Sir Mix-a-Lot

The term "Big Data" is often used as a marketing term rather than a technical term. For the purpose of this guide, I will define big data as any data that exceeds your computer hardware resources under normal operation. R has many more capabilities than are mentioned here. For an updated list of packages related to high performance computing, see the CRAN Task View.

Here are a few tips that are useful when dealing with data that is constrained by your computing resources. Use `object.size()` to get the RAM usage of an object in your R session. Operations on data.frames usually involve making a copy of a data.frame. This means that you can only use about half of your computer's RAM without using swap space. Any software that uses the hard drive when RAM is limited will be much slower than an operation that is run entirely in RAM because the access speed of RAM is much greater than that of any hard drive.

The SQLite Database

One strategy for dealing with data that doesn't fit into memory is to use a database. Rather than setting up a full MySQL database, you can use a SQLite database for many problems where a single computing process needs read/write access. SQLite is not appropriate when multiple processes need to access the database at the same time.

The RSQLite Package

```
# Read data
dataset=read.csv("/yourdirectory/yourfile.csv",skip=1,header=TRUE)
# Create DB
testdb=dbConnect(SQLite(), dbname = "/yourdirectory/testdb.dat")
# Write the data set into a table called table1
dbWriteTable(testdb,"table1",dataset)
# List the tables available in the database
dbListTables(testdb)
# Remove the original data to free up memory
rm(dataset)
# Retrieve the entire table
dbReadTable(testdb,"table1")
# Retrieve the single column called series from the table
buf<-dbSendQuery(testdb,"select series from table1")
testing<-fetch(buf,n=-1); dbClearResult(buf)
dbDisconnect(testdb); rm(buf,testdb)
```

SQLite in Python

```
import sqlite3 as sql
import pandas as pd
pd.read_sql("select * from table",conn)
```

The bigmemory Package

More information is on the CRAN page and the big memory website.

Resources

Books

Finance Textbooks

Investments by Bodie, Kane, and Marcus

Finance Books

Security Analysis by Benjamin Graham
The Intelligent Investor by Benjamin Graham
Investing Philosophies by Aswath Damodaran
The Misbehavior of Markets: A Fractal View of Financial Turbulence by Benoit Mandelbrot
Mostly Harmless Econometrics by Joshua Angrist

Financial History

When Genius Failed: The Rise and Fall of Long Term Capital Management by Roger Lowenstein
Against the Gods: The Remarkable Story of Risk by Peter Bernstein

Programming Books

R Cookbook by Paul Teetor
Analysis of Financial Time Series by Ruey Tsay

Internet

Personal Website

Finance Library
Market Microstructure
Selected Podcasts and Media Features

Regression Basics

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_N X_N$$

R^2 - Ranges from 0 to 1 and shows the proportion of variation in Y explained by the X variables together. This measure increases as you add more X variables.

Adjusted R^2 - This measure penalizes you for adding X variables so it is a better measure for comparing different models. Interpretation is the same as R^2.

Coefficients (β) - how much Y changes due to a one unit change in X

T-statistic - measures whether β is statistically different from zero. Absolute value greater than about two and a half mean that β is different from zero. This corresponds to p-values.

P-value - measures the certainty that β is different from zero. P-values lower than 0.05 are OK but P-values lower than 0.01 are better.

$\beta = 0$ - This means that the variation in X does not explain the variation in Y

Low R^2 - This means that your collection of X variables does not explain the variation in Y

Things to Remember:

Correlation is not causation! - A high t-stat does not mean that X causes Y

Bad things happen to well-intentioned regressions. - Before implementing the results of a regression, many tests much be done to ensure that the results are valid.

More data is better than less data. - If you have the choice between a short period of time or a long period of time, it is usually better to choose the longer period of time.

Regression in Microsoft Excel
Run a regression by going to Data → Data Analysis → Regression
If it is not there, go to file → options and ensure that the Analysis Toolpack is loaded.

Tricks:

Try using the natural logarithm of Y or X or both if the relationship between the two variables is not a straight line.

Try using X^2 in addition to X if the logarithm trick doesn't work.

Try using the first difference in a time series if you think that the relationship between variables may change over time.

If you do transform a variable, remember to make sure that X and Y are in the same scale. If X is in decimal form but Y is in percentage form, the estimated β is harder to interpret.