FE670 Algorithmic Trading Strategies

Lecture 5. Cross-Sectional Models and Trading Strategies

Steve Yang

Stevens Institute of Technology

steve.yang@stevens.edu

09/30/2021

Overview

Cross-Sectional Methods for Evaluation of Factor Premiums

Performance Evaluation of Factors

Methodologies for a Factor-Based Trading Strategy

Backtesting

Python Sample - Fundamental Factor Long Short Strategy

There are several approaches used for the evaluation of return premiums and risk characteristics to factors. We discuss four most commonly used approaches:

1. *Portfolio Sorts*
2. *Factor Models*.
3. *Factor Portfolios*.
4. *Information Coefficients*.

* In practice, to determine the right approach for a given situation there are several issues to consider 1). the structure of the financial data. 2). the economic intuition underlying the factor. 3). validity of the underlying assumptions of each approach.

Portfolio Sorts

The portfolios are constructed by grouping together securities with similar characteristics (factors). The goal of this process is to determine whether a factor earns a systemic premium.

The return of each portfolio is calculated by equally weighting the individual stock returns. The portfolios provide a representation of how returns vary across the different values of a factor. By studying the return behavior of the factor portfolios, we may assess the return and risk profile of the factor.

Overall, the return behavior of the portfolios will help us conclude whether there is a premium associated with a factor and describe its properties.

The construction of portfolio sorts on a factor is straightforward:

1. Choose an appropriate sorting methodology.
2. Sort the assets according to the factor.
3. Group the sorted assets into N portfolios (usually $N = 5$, or $N = 10$).
4. Compute average returns (and other statistics) of the assets in each portfolio over subsequent periods.

The standard statistical testing procedure for portfolios sorts is to use a Student's $t-$test to evaluate the significance of the mean return differential between the portfolios of stocks with the highest and lowest values of the factor.

Choosing the Sorting Methodology

The sorting methodology should be consistent with the characteristics of the distribution of the factor and the economic motivation underlying its premium. Here six ways to sort factors:

Method 1:

Sort stocks with factor values from the highest to lowest.

Method 2:

Sort stocks with factor values from the lowest to highest.

Method 3: (Example: dividend yield factor)

First allocate stocks with zero factor values into the bottom portfolio.

Sort the remaining stocks with nonzero factor values into the remaining portfolios.

Compute average returns (and other statistics) of the assets in each portfolio over subsequent periods.

Choosing the Sorting Methodology

Method 4:

1). Allocate stocks with zero factor values into the middle portfolio.

2). Sort stocks with positive factor values into the remaining higher portfolios (greater than the middle portfolio).

3). Sort stocks with negative factor values into the remaining lower portfolios (less than the middle portfolio).

Method 5: (Example: rank stocks according to earnings growth on a sector neutral basis)

1). Sort stocks into partitions.

2). Rank assets within each partition.

3). Combine assets with the same ranking from the different partitions into portfolios.

Choosing the Sorting Methodology

Method 6: (Example: share repurchase factor)

1). Separate all the stocks with negative factors values. Split the group of stocks with negative values into two portfolios using the median value as the break point.

2). Allocate stocks with zero factor values into one portfolio.

3). Sort the remaining stocks with nonzero factor values into portfolios based on their factor values.

* The portfolio sort methodology has several advantages. The approach is easy to implement and can easily handle stocks that drop out or enter into the sample. The resulting portfolios diversify away idiosyncratic risk of individual assets and provide a way of assessing how average returns differ across different magnitude of a factor.

Example 1: Portfolio Sorts Based on the EBITDA/EV Factor

Exhibit 7.1 contains the cross-sectional distribution of the EBITDA/EV factor. This distribution is approximately normally distributed around a mean of 0.1, with a slight right skew. We use method 1 to sort the variables into five portfolios. Therefore, a strategy that goes long on portfolio 1 and short 5 appears to produce abnormal returns.

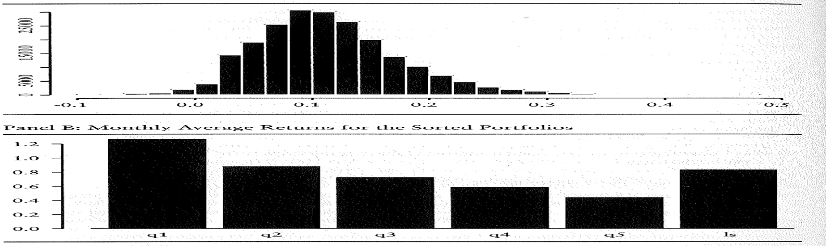


EXHIBIT 7.1 Portfolio Sorts Based on the EBITDA/EV Factor

Panel A: All Factor Values

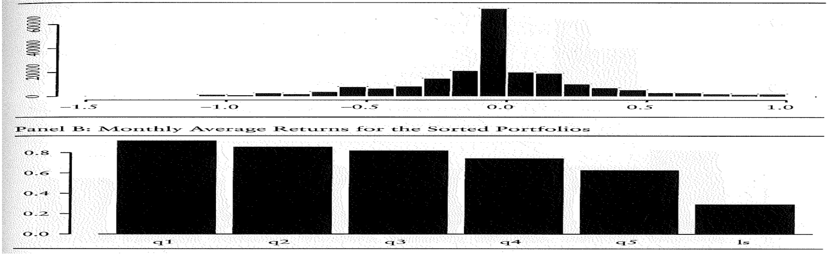
Panel B: Monthly Average Returns for the Sorted Portfolios

Example 2: Portfolio Sorts Based on the Revisions Factor

Exhibit 7.2 shows that the distribution of earnings revisions is leptokurtic around a mean of about zero, with the remaining values symmetrically distributed around the peak. We use method 3 to sort the variables into five portfolios. The stocks with positive revisions are sorted into portfolio 1 and 2- while negative revisions stocks are sorted into 4 and 5. Therefore, a strategy that goes long on portfolio 1 and short 5 appears to produce abnormal returns.

EXHIBIT 7.2 The Revisions Factor

Panel A: All Factor Values



Panel B: Monthly Average Returns for the Sorted Portfolios

Example 3: Portfolio Sorts Based on the Share Repurchase Factor

Exhibit 7.3 shows that the distribution of share repurchase is asymmetric and leptokurtic around a mean of about zero, with the remaining values symmetrically distributed around the peak. We use method 6 to sort the variables into seven portfolios. There is a large difference in return between the extreme portfolios. Therefore, a strategy that goes long on portfolio 1 and short 5 appears to produce abnormal returns.

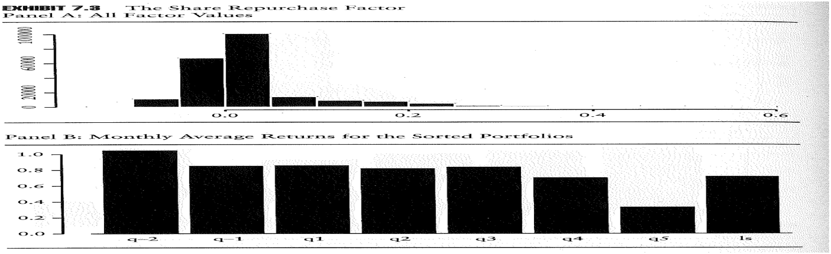


EXHIBIT 7.3 The Share Repurchase Factor

Panel A: All Factor Values

Panel B: Monthly Average Returns for the Sorted Portfolios

Information Ratios for Portfolio Sorts

- The information ratio (IR) is a statistic for summarizing the risk-adjusted performance of an investment strategy. It is defined as the ratio of the average excess return to the standard deviation of return.
- For actively managed equity long portfolios, the IR measures the risk-adjusted value a portfolio manager is adding relative to a benchmark.
- IR can also be used to capture the risk-adjusted performance of long-short portfolios from a portfolio sorts.
- When comparing portfolios built using different factors, the IR is an effective measure for differentiating the performance between the strategies.

New Research on Portfolio Sorts

- The standard statistical testing procedure uses a Student's *t*-test to evaluate the mean return differential between the two portfolios containing stocks with the highest and lowest values of the sorting factor.
- However, this approach ignores important information about the overall pattern of returns among the remaining portfolios.
- The Monotonic Relation(MR) test can reveal whether the null hypothesis of no systemic relationship can be rejected in favor of a monotonic relationship predicted by economic theory.
- By MR it is meant that the expected returns of a factor should rise or decline monotonically in one direction as one goes from one portfolio to another.

Factor Models

In investment management, risk is a primary concern. The majority of trading strategies are not risk free but rather subject to various risks. Here we describe some common risks to factor trading strategies as well as other trading strategies.

Classical financial theory states that the average return of a stock is the payoff to investors for taking on risk. One way of expressing this risk-reward relationship is through a factor model. A factor model can be used to decompose the returns of a security into factor-specific and asset-specific returns:

$$r_{i,t} = \alpha_i + \beta_{i,1} f_{1,t} + ... + \beta_{i,K} f_{K,t} + \epsilon_{i,t}$$

where $\beta_{i,1}, \beta_{i,2}, ..., \beta_{i,K}$ are the factor exposures of stock i, $f_{1,t}, f_{2,t}, ..., f_{K,t}$ are the factor returns, α_i is the average abnormal return of stock i, and $\epsilon_{i,t}$ is the residual.

Factor Models

This factor model specification is *contemporaneous*, that is, both left- and right-hand side variables (returns and factors) have the same time subscript, t.

For trading strategies one generally applies a *forecasting* specification where the time subscript of the return and the factors are $t + h(h \geq 1)$ and t, respectively. In this case, the econometric specification becomes:

$$r_{i,t+h} = \alpha_i + \beta_{i,1} f_{1,t} + ... + \beta_{i,K} f_{K,t} + \epsilon_{i,t+h}$$

How do we interpret a trading strategy based on a factor model? The explanatory variables represent different factors that forecast security returns, each factor with its associated factor premium.

Factor Models

Therefore, future security returns are proportional to the stock's exposure to the factor premium

$$E(r_{i,t+h}|f_{1,t}, f_{2,t}, ..., f_{K,t}) = \boldsymbol{\alpha}_i + \boldsymbol{\beta}_i \mathbf{f_t}$$

and the variance of future stock return is given by

$$Var(r_{i,t+h}|f_{1,t}, f_{2,t}, ..., f_{K,t}) = \boldsymbol{\beta}_i^{'} E(\mathbf{f_t f_t^{'}}) \boldsymbol{\beta}_\mathbf{i}$$

where $\boldsymbol{\beta}_i = (\beta_{i,1}, ..., \beta_{i,K})'$ and $\mathbf{f_t} = (\mathbf{f_{1,t}}, ..., \mathbf{f_{K,t}})'$.

Econometric Considerations for Cross-Sectional Factor Models

In cross-sectional regression where the dependent variable is a stock's return and the independent variables are factors, inference problems may arise that are the result of violations of classical linear regression theory. The most common problems:

Measurement Problems Some factors are not explicitly given, but need to be estimated. These factors are estimated with an error. The estimation errors in the factors can have an impact on the inference from a factor model. This problem is commonly referred to as the "errors in variables problem".

Common Variation in Residuals The residuals from a regression often contain a source of common variation. Sources of common variation in the residuals are heteroskedasticity and serial correlation.

Econometric Considerations for Cross-Sectional Factor Models

- **Common Variation in Residuals**

  Heteroskedasticity occurs when the variance of the residual differs across observations and affects the statistical inference in a linear regression. In particular, the estimated standard errors will be underestimated and the t-statistics will therefore be inflated. Ignoring heteroskedasticity may lead the researcher to find significant relationships where none actually exist. Several procedures have been developed to calculate standard errors that are robust to heteroskedasticity.

- Serial correlation occurs when consecutive residual terms in a linear regression are correlated, violating the assumptions of regression theory. If the serial correlation is positive then the standard errors are underestimated and the t-statistics will be inflated.

Econometric Considerations for Cross-Sectional Factor Models

- **Multicollinearity** Multicollinearity occurs when two or more independent variables are highly correlated. We may encounter several problems when this happens.
  1) First, it is difficult to determine which factors influence the dependent variable.
  2) Second, the individual p values can be misleading – a p value can be high even if the variable is important.
  3) Third, the confidence intervals for the regression coefficients will be wide.
- There is no formal solution based on theory to correct for multicollinearity. The best way is by removing one or more of the correlated independent variables. It can be reduced by increasing the sample size.

Fama-MacBeth Regression

To address the inference problem caused by the correlation of the residuals, Fama and MacBeth proposed the following methodology for estimating cross-sectional regressions of returns on factors. For notational simplicity, we describe the procedure for one factor. The multifactor generalization is straightforward:

First, for each point in time t we perform a cross-sectional regression

$$r_{i,t} = \beta_{i,t} f_t + \epsilon_{i,t}, i = 1, 2, ..., N$$

In the academic literature, the regressions are typically performed using monthly or quarterly data, but the procedure could be used at any frequency.

The mean and standard errors of the time series of slopes and residuals are evaluated to determine the significance of the cross-sectional regression.

Fama-MacBeth Regression

We estimate f and ϵ_i as the average of their cross-sectional estimates:

$$\hat{f} = \frac{1}{T}\sum_{t=1}^{T}\hat{f}_t, \hat{\epsilon}_i = \frac{1}{T}\sum_{t=1}^{T}\hat{\epsilon}_{i,t}$$

The variations in the estimates determine the standard error and capture the effects of residual correlation without actually estimating the correlation.

We used the standard deviations of the cross-sectional regression estimates to calculate the sampling errors for these estimates,

$$\hat{\sigma}_{\hat{f}} = \frac{1}{T^2}\sum_{t=1}^{T}(\hat{f}_t - \hat{f})^2, \sigma^2_{\hat{\epsilon}_i} = \frac{1}{T^2}\sum_{t=1}^{T}(\hat{\epsilon}_{i,t} - \hat{\epsilon}_i)^2$$

Information Coefficients

- To determine the forecast ability of a model, practitioners commonly use the information coefficient (IC). The IC is a linear statistic that measures the cross-sectional correlation between a factor and its subsequent realized return

  $$IC_{t,t+k} = \mathrm{corr}(\mathbf{f}_t, \mathbf{r}_{t,t+k})$$

  where $\mathbf{f}_t$ is a vector of cross sectional factor values at time t and $\mathbf{r}_{t,t+k}$ is a vector of returns over the time period t to $t+k$.
- Just like the standard correlation coefficient, the values of the IC range from -1 to $+1$. A positive IC indicates a positive relation between the factor and return. A negative IC indicates a negative relation. ICs are usually calculated over an interval, for example, daily or monthly. We can evaluate how a factor has performed by examining the time series behavior of the ICs.

- An alternative specification of this measure is to make $\mathbf{f}_t$ the rank of a cross-sectional factor. This calculation is similar to the Spearman rank coefficient. By using the rank of the factor, we focus on the ordering of the factor instead of its value. Ranking the factor value reduces the unduly influence of outliers and reduces the influence of variables with unequal variances.

- The subsequent realized returns to a factor typically vary over different time horizons. For example, the return to a factor based on price reversal is realized over short horizons, while valuation metrics such as EBITDA/EV are realized over longer periods. It therefore makes sense to calculate multiple ICs for a set of factor forecasts whereby each calculation varies the horizon over which the returns are measured.

- Information coefficients can also be used to assess the risk of factors and trading strategies. The standard deviation of the time series (with respect to) of ICs for a particular factor can be interpreted as the strategy risk of a factor. Examining the time series behavior of $std(IC_{t,t+k})$ over different time periods may give a better understanding of how often a particular factor may fail.

- The expected tracking error can be used to understand the active risk of investment portfolios. Qian and Hua defined an expected tracking error as:

  $$\sigma_{TE} = \text{std}(\mathbf{IC}_{t,t+k})\sqrt{N}\sigma_{\text{model}}\text{dis}(\mathbf{R}_t)$$

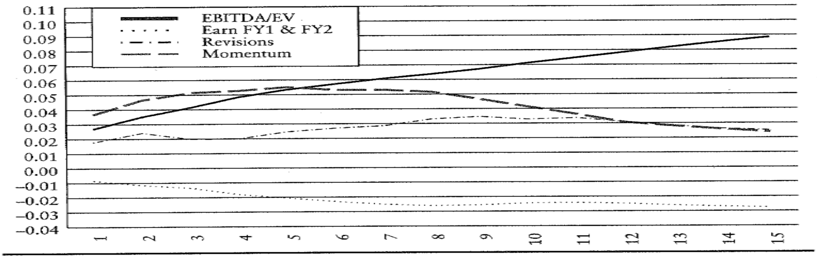
  where N is the number of stocks in the universe (breath), σ_{model} is the risk model tracking error, and $\text{dis}(\mathbf{R}_t)$ is dispersion of returns defined by

  $$\text{dis}(\mathbf{R}_t) = \text{std}(r_{1,t}, r_{2,t}, ..., r_{N,t})$$

Example: Information Coefficients

Exhibit 7.4 displays the time-varying behavior of ICs for each one of the factors EBITDA/EV, growth of fiscal year 1 and 2 earnings estimates, revisions, and momentum. The graph depicts the information horizons for each factor. The EBITDA/EV factor earns higher returns. The overall pattern shows that the return realization pattern to different factors varies.

EXHIBIT 7.4 Information Coefficients over Various Horizons for EBITDA/EV, Growth of Fiscal Year 1 and Fiscal Year 2 Earnings Estimates, Revisions, and Momentum Factors



Factor Portfolios

- Factor portfolios are constructed to measure the information content of a factor. The objective is to mimic the return behavior of a factor and minimize the residual risk. Similar to portfolio sorts, we evaluate the behavior of these factor portfolios to determine whether a factor earns a systematic premium.

- Typically, a factor portfolio has a unit exposure to a factor and zero exposure to other factors. Construction of factor portfolios requires holding both long and short positions. We can also build a factor portfolio that has exposure to multiple attributes, such as beta, sectors, or other characteristics. Portfolios with exposures to multiple factors provide the opportunity to analyze the interaction of different factors.

A Factor Model Approach

- By using a multifactor model, we can build factor portfolios that control for different risks. We decompose return and risk at a point in time into a systematic and specific component using the regression:

  $$\mathbf{r} = \mathbf{X}\mathbf{b} + \mathbf{u}$$

  where **r** is an N vector of excess returns of the stocks considered, **X** is an N by K matrix of factor loadings, **b** is a K vector of factor returns, and **u** is a N vector of firm specific returns (residual returns).
- We assume that factor returns are uncorrelated with the firm specific return. Further assuming that firm specific returns of different companies are uncorrelated, the N by N covariance matrix of stock return **V** is given by:

  $$\mathbf{V} = \mathbf{X}\mathbf{F}\mathbf{X}' + \mathbf{\Delta}$$

where $\mathbf{F}$ is the K by K factor return covariance matrix and $\boldsymbol{\Delta}$ is the N by N diagonal matrix of variances of the specific returns.

- We can use the Fama-MacBeth procedure discussed earlier to estimate the factor returns over time. Each month, we perform Generalize Least Square - GLS regression to obtain

$$\mathbf{b} = (\mathbf{X}'\boldsymbol{\Delta}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Delta}^{-1}\mathbf{r}$$

  OLS would give us an unbiased estimate, but since the residuals are heteroskedastic the GLS methodology is preferred and will deliver a more efficient estimate. The resulting holdings for each factor portfolio are given by the rows of $(X'\Delta^{-1}X)^{-1}X'\Delta^{-1}$.

An Optimization-Based Approach

- A second approach to build factor portfolios uses mean-variance optimization. Using optimization techniques provide a flexible approach for implementing additional objectives and constrains. We would like to construct a portfolio that has maximum exposure to one targeted factor from $\mathbf{X}$ (the alpha factor), zero exposure to all other factors, and minimum portfolio risk. Let us denote the alpha factor by $\mathbf{X}_\alpha$ and all the remaining ones by $\mathbf{X}_\sigma$. Then the resulting optimization problem takes the form:

$$\max_{w}\left\{\mathbf{w}'\mathbf{X}_\alpha - \frac{\mathbf{1}}{\mathbf{2}}\lambda\mathbf{w}'\mathbf{V}\mathbf{w}\right\}$$

$$s.t.\mathbf{w}'\mathbf{X}_\sigma = 0$$

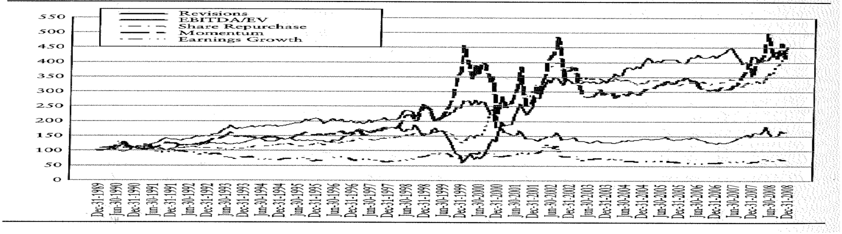
- The analytical solution is given by:

$$h^* = \frac{\mathbf{1}}{\lambda}\mathbf{V}^{-\mathbf{1}}\left[\mathbf{I} - \mathbf{X}_\sigma(\mathbf{X}_\sigma'\mathbf{V}^{-\mathbf{1}}\mathbf{X}_\sigma)^{-\mathbf{1}}\mathbf{X}_\sigma'\mathbf{V}^{-\mathbf{1}}\right]\mathbf{X}_\alpha$$

Performance Evaluation of Factors

- Analyzing the performance of different factors is an important part of the development of a factor-based trading strategy. A researcher may construct and analyze over a hundred different factors, so the means to evaluate and compare these factors is needed. Most often this process starts by trying to understand the time-series properties of each factor in isolation and the study how they interact with each other.

EXHIBIT 7.6 Cumulative Returns of Long-Short Portfolios



Performance Evaluation of Factors

- To better understand the time variation of the performance of these factors, we calculate rolling 24-month mean returns and correlations of the factors.

EXHIBIT 7.9 Rolling 24-Month Mean Returns for the Factors

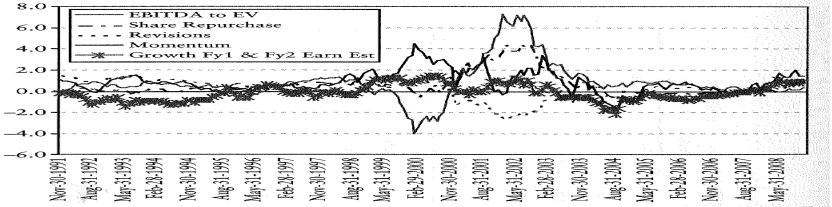
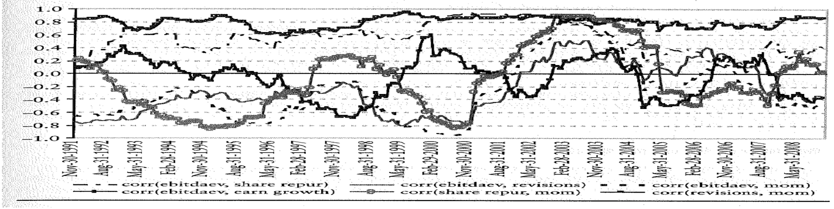


EXHIBIT 7.10 Rolling 24-Month Correlations of Monthly Returns for the Factors



Model Construction Methodologies for a Factor-Based Trading Strategy

- The key aspect of building a model is to (1) determine what factors to use out of the universe of factors that we have, and (2) how to weight them.

- We describe four methodologies to combine and weight factors to build a model for a trading strategy. These methodologies are used to translate the empirical work on factors into a working model.

- It is important to be careful how each methodology is implemented. In particular, it is critical to balance the iterative process of finding a robust model with good forecasting ability versus finding a model that is a result of data mining.

The Data Driven Approach

- A *data driven approach* uses statistical methods to select and weight factors in a forecasting model. This approach uses returns as independent variables and factors as the dependent variables. There are a variety of estimation procedures, such as neural nets, classification trees, and principal components, that can be used to estimate these models.

- Many data driven approaches have no structural assumptions on potential relationships the statistical method finds. Therefore, it is sometimes difficult to understand or even explain the relationship among the dependent variables used in the model.

The Factor Model Approach

- The goal of the factor model is to develop a parsimonious model that forecast returns accurately. One approach is for the researcher to predetermine the variables to be used in the factor model based on economic intuition. The model is estimated and then the estimated coefficients are used to produce the forecasts.
- A second approach is to use statistical tools for model selection. In this approach we construct several models - often by varying the factors and the number of factors used - and have them compete against each other, and then choose the best performing model.
- Factor model performance can be evaluated in three ways. We can evaluate the fit, forecast ability, and economic significance of the model.

The Heuristic Approach

- Heuristics are based on common sense, intuition, and market insight and are not formal statistical or mathematical techniques designed to meet a given set of requirements. The researcher decides the factors to use, creates rules in order to evaluate the factors, and chooses how to combine the factors and implement the model.

- There are different approaches to evaluate a heuristic approach. Statistical analysis can be used to estimate the probability of incorrect outcomes. Another approach is to evaluate economic significance.

- There is no theory that can provide guidance when making modeling choices in the heuristic approach. Consequently, the researcher has to be careful not to fall into the data mining trap.

The Optimization Approach

- In this approach, we use optimization to select and weight factors in a forecasting model. An optimization approach allows us flexibility in calibrating the model and simultaneously optimize an objective function specifying a desirable investment criteria.

- There is substantial overlap between optimization use in forecast modeling and portfolio construction. The factors provide a lower dimensional representation of the complete universe of the stocks considered. Besides the dimensionality reduction, which reduces computational time, the resulting optimization problem is typically more robust to changes in the inputs.

Backtesting

- Model scores are converted into portfolios and then examined to assess how these portfolios perform over time. The backtest should mirror as closely as possible the actual investing environment incorporating both the investment's objectives and the trading environment.
- In-sample backtesting is referred to use the same data sample to specify, calibrate and evaluate a model.
- Out-sample backtesting is where the researcher uses a subset of the sample to specify and calibrate a model, and then evaluates the forecasting ability on a different subset of the data (split-sample method and recursive out-of-sample method).

Python Sample - Fundamental Factor Long Short Strategy

- In this tutorial we implemented a long/short equity strategy based on fundamental factors. The idea comes from AQR white book: A New Core Equity Paradigm.
- The original version is a long only strategy. We developed it into a long/short version. The paper strategy used some fundamental data as measures of value, quality and momentum, and then ranked all the stocks in the universe according to the factors.
- The strategy only long the stocks ranking at the top, but our algorithm would at the same time short the stocks ranking at the bottom. This strategy consistently beats the market and has solid economic intuition.

  See Python Code.