NORTHFIELD | modeling financial markets worldwide

Non-Linear Equity Factor Models?

Dan diBartolomeo
Northfield Webinar
April 2025

Background

Since the 1970s, the key property of equity factor models has been the presumed linear relationship between security returns and factor exposures.

This assumption arises out of the Arbitrage Pricing Theory (Ross, *Journal of Economic Theory*, 1976).

However, the necessary conditions under which the APT would be assumed to strictly hold in the real world do not exist, as the APT assumes that liquidity is essentially infinite, and therefore, transaction costs are non-existent, which allows traded markets to always achieve a clearing price.

While equity factor models have been used successfully to describe the relationship between returns and sources of risk, the recent popularity of "machine learning" has spawned new interest in exploring non-linear relationships.

Today's Discussion

Inspired by our recent research into the non-normality in distribution of the cross-section of stock returns (see Variety is the Spice of Active Management (as well as Life), this webinar will explore a middle ground between traditional linear models and machine learning models.

We will illustrate the estimation of factor models based on simple specifications such as a linear relationship between "heavy tailed" returns and the absolute value of factor exposure, or between such returns and the square of factor exposure.

While the explanatory power of such relationships is lower in traded equities than the linear model, there appears to be material information content in these alternative specifications.

A Simple Example

Let's assume we break stock returns into three groups based on firm size, similar to one of three factors described in the original Fama-French work (*Journal of Finance*, 1992).

In some periods, we might observe that large cap stocks performed well, small cap stocks performed well, but mid-cap stocks performed poorly.

If we assume there must be a linear relationship between size and return, our analysis will conclude that there is no relationship between size and return for the observed period since the relationship cannot be fit to a straight line.

The Northfield performance attribution system has recognized this issue from the early 1990s and provided comparative return analysis (i.e. portfolio and benchmark) by user defined quantiles of all factors.

Let's Start from the CAPM

The traditional Capital Asset Pricing Model (Sharpe, *Journal of Finance*, 1964) posits a *linear* relationship between expected return and risk, where the definition of risk is a rescaled measure of the **covariance** between an asset and a hypothetical "market portfolio" called "beta" as a representation of *relative* economic risk.

- All returns not driven by the market returns are considered residual.

Under certain simplifying assumptions, the CAPM can be derived from the earlier specification of portfolio theory (Markowitz, Journal of Finance, 1952).

- Among these assumptions is that the future is a single period, which eliminates the need for consideration of liquidity or non-zero trading costs.

Building on this infinite liquidity assumption, Ross's Arbitrage Pricing Theory (1976) generalizes the linear relationship between expected returns and risk factors.

- *The APT might better be called the "no-arbitrage pricing theory" since the embedded assumptions presume that a market clearing price can always be established so arbitrage profits should not be available.*
- For illiquid assets, the idea that a clearing price will also exist is obviously a poor assumption.
  - See Belev and diBartolomeo ("Investor Utility and Asset Pricing", Forthcoming in *Journal of Investing*)

The Big Step Forward

Rosenberg and Guy (*Financial Analyst Journal*, 1976) argued that beta values should not be estimated statistically via a set of historical observations *but rather should be **predicted** from the fundamental characteristics (i.e. factor exposures) of stocks.*

The relationship between beta as a measure of risk and many fundamental variables seems intuitive:

- Firms with less volatile earnings will have a lower beta than firms with volatile earnings
- Firms with more balance sheet leverage will have a higher beta than firms with less debt

As a measure of covariance, differences in beta values across firms may arise from differences in either stock return variances or the correlation of stock returns to the market portfolio.

- Some groups of firms (e.g. industries) are less related to the broad world economy (e.g. gold mines) and therefore member stocks will have return lower correlation to the market portfolio, but higher pairwise correlations between any two securities in the group.
- *This effect is known as "extra-market" covariance.*

Common Models of Covariance

Since the seminal work of Rosenberg and Guy, equity factor models have become commonplace as the way in which investment practitioners estimate the risk of their portfolios, either in absolute return terms, or returns relative to some related "benchmark" equity index as representation of the relevant market portfolio (i.e. the set of investment opportunities).

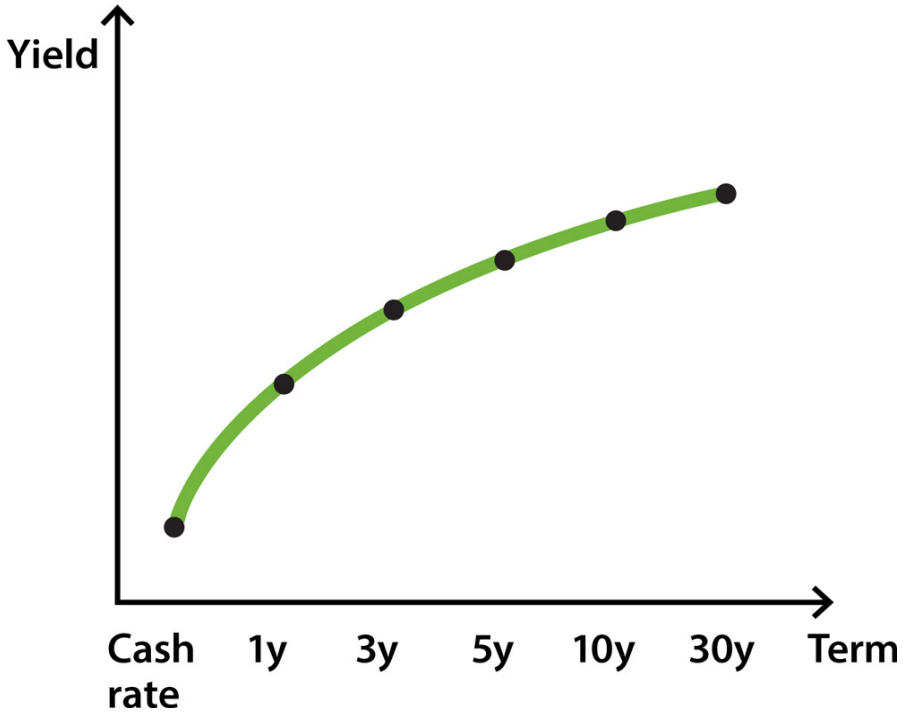
There are three popular ways to specify equity factor models that **all *assume linearity*** between risk factor exposures and returns.

- The model types are endogenous, exogenous, and statistical
- For more discussion of equity model specifications see diBartolomeo (The Oxford Handbook of Quantitative Asset Management, 2012).

In 2003, Northfield introduced *hybrid* models which combined "statistical factors within residuals" with one of the other methods.

- The hybrid method providing for the same underlying factor to be captured twice in an overlapping fashion, allowing for potential non-linear behavior expressed as two linear segments instead of one).

The Armed Weightlifter: Bullets versus Barbells



Fixed income investors have long recognized the non-linear nature of factor representations of interest rates.

Put simply, the "yield curve" is curved.

- Common yield curve (i.e. "term structure") models of interest rates typically involve three factors: the exposure to the average level of interest rates ("duration"), exposure to the difference between rate between long and short maturity bonds (i.e. "slope"), and exposure to *the degree of observable curvature in yields*.

Equal effective duration portfolios are clearly not equivalent, leading to the *"bullet versus barbell"* terminology.

- A portfolio consisting of 50% one-year duration bonds and 50% nine-year duration bonds is not equivalent to a portfolio of five-year duration bonds.

Moving Away from Linearity Assumptions for Equities

In recent years, active equity managers have frequently tried to model the relationship between return and factor exposures using "data driven" techniques such as machine learning.

- Frequent evidence of complex relationships between factor exposures and returns, *but the statistical significance (i.e. stability) of such relationships is difficult to evaluate resulting in severe "overfitting".*
- A rigorous examination of overfitting problem was presented in Bailey, Borwein, dePrado, and Zhu ("Financial Charlatanism", *Notices of the American Mathematical Society*, 2012) ams.org

Clarke, DaSilva, and Thorley (Financial Analyst Journal, 2024) examine five equity return factors in a fashion consistent with Fama and French.

- They conclude that the evidence for the linearity assumption is very weak for the period of 1964 to 2023
- Information ratios for managed portfolios can be increased by controlling risk/return effects of non-linear nature (e.g. factor exposure squared).

Today's empirical analysis starts with the concepts of Rosenberg and Guy, rather than Fama and French.

Starting from the Classic Linear Model

The Northfield US Fundamental factor risk model largely follows the endogenous specification of Rosenberg and Guy.

- All securities traded on US exchanges are covered including ADRs and cross-listings.
- The factor list includes market beta, eleven fundamental characteristics (e.g. log of market cap, dividend yield) and fifty-five industry groups.
- The estimation of the actual commercial model involves many nuances (outlier handling, weighting of observations, Parkinson (*Journal of Business*, 1980) adjustment for non-normal returns)

A sample of summary data on Fundamental Model ***ex-ante risk estimates*** for the post-COVID period from October of 2020 through November of 2024

- The average number of stocks in each monthly analysis is 6700.
- At the individual stock level, the pooled average ex-ante R-squared is 32% for the equal weighted universe (emphasis on small cap), comparable figure for capitalization weight (emphasis on large cap) is 50%.

Let's Look at March 2025

T-Stats		
	Earnings/Price	0.10
	Book/Price	5.62
	Dividend Yield	1.24
	Trading Activity	3.24
	Relative Strength	13.75
	Log of Market Cap	-1.68
	Earnings Variability	-1.86
	EPS Growth Rate	1.10
	Revenue/Price	0.26
	Debt/Equity	-1.15
	Price Volatility	-7.67

Using our Fundamental Model data (factor exposures as of 2/28/25, returns from March 2025) we estimated a cross-sectional regression where the dependent variable are stock returns net of the returns associated with the risk-free rate and the market beta (i.e. consistent with the CAPM).

- Observations are square root of capitalization weighted and stocks with capitalization below $250M are excluded from the estimation process, consistent with the commercial model.
- *If CAPM is a full explanation of returns, the incremental ex-post explanatory power of the model factors should be zero as the factor influences should be embedded in the beta values.*
- The R-squared of the XMC portion is 16% for the linear estimation with five of eleven characteristics having significant T-stats. *T-stats have not been adjusted for weighting of observations and so are slightly upward biased,* but equal weighted regressions were almost identical.

Non-Linear Factor Returns for March 2025

- We then repeat the analysis assuming the relationship is between returns and factor exposure squared, then assuming return and the absolute value of factor exposure.
  - *In both cases the explanatory power of the revised XMC specification is 12% (about 75% of the linear assumption) with multiple significant T-stats (far more than should arise randomly)*
  - R-squared for including both linear and quadratic relationships is 18% and 17% for the combination of linear and absolute value specification.

T-Stats	Quadratic	
	Earnings/Price_squared	0.75
	Book/Price_squared	0.79
	Dividend Yield_squared	0.09
	Trading Activity_squared	0.41
	Relative Strength_squared	6.70
	Log of Market Cap_squared	-0.16
	Earnings Variability_squared	3.55
	EPS Growth Rate_squared	0.85
	Revenue/Price_squared	-0.37
	Debt/Equity_squared	-1.37
	Price Volatility_squared	-0.06

T-stats	Absolute	
	Earnings/Price_absval	0.10
	Book/Price_absval	-0.24
	Dividend Yield_absval	-0.20
	Trading Activity_absval	1.01
	Relative Strength_absval	4.60
	Log of Market Cap_absval	1.06
	Earnings Variability_absval	2.89
	EPS Growth Rate_absval	0.14
	Revenue/Price_absval	-0.23
	Debt/Equity_absval	-1.31
	Price Volatility_absval	2.50

Same Linear Analysis for August 2022

T-Stats		
	Earnings/Price	-2.22
	Book/Price	1.54
	Dividend Yield	2.39
	Trading Activity	-2.33
	Relative Strength	29.15
	Log of Market Cap	2.47
	Earnings Variability	0.18
	EPS Growth Rate	-2.83
	Revenue/Price	1.50
	Debt/Equity	-1.47
	Price Volatility	12.80

The linear model had a slightly better explanatory power for the XMC (factor returns net of beta influence) at 21% versus 18% for March 2025 (exposures from July 31, 2022, returns from August 2022)

Non-Linear Factor Returns for August 2022 Dominate

- We then repeat the analysis assuming the relationship is between returns and factor exposure squared, then assuming return and the absolute value of factor exposure.
  - *For the quadratic specification, the R-squared is **56%,** more than double that of the linear model. The R-squared for the absolute model is 22% (comparable to linear). In both cases the explanatory power of the revised XMC specification is large with multiple significant T-stats, including a massive influence from the relative strength (momentum) factor.*
  - R-squared for including both linear and quadratic relationships is 58% and 26% for the combination of linear and absolute value specification.

T-Stats	Quadratic	
	Earnings/Price	-1.26
	Book/Price	0.39
	Dividend Yield	1.64
	Trading Activity	-2.90
	Relative Strength	67.97
	Log of Market Cap	1.89
	Earnings Variability	0.11
	EPS Growth Rate	-4.71
	Revenue/Price	0.35
	Debt/Equity	-0.18
	Price Volatility	0.00

T-stats	Absolute	
	Earnings/Price	-1.26
	Book/Price	-0.95
	Dividend Yield	2.96
	Trading Activity	-4.29
	Relative Strength	29.86
	Log of Market Cap	4.29
	Earnings Variability	0.06
	EPS Growth Rate	-3.22
	Revenue/Price	-0.78
	Debt/Equity	0.45
	Price Volatility	0.90

A Clue from Merton (1974)

Merton (*Journal of Finance*, 1974) proposed the "contingent claims" model of corporate credit risk wherein a risky corporate debt can be represented as a two-asset portfolio consisting of theoretically risk-free debt (e.g. US Treasury bonds) and a portion of equity in the corporate issuer.

It should be intuitive that risky debt with material likelihood of default will produce an ex-ante return distribution with negative skew and positive excess kurtosis arising from credit default events.

- Since credit risk is represented by equity in this formulation, it must follow that the equity must have the same properties of high moments in some situations of low liquidity such as high yield bonds.

For more detailed discussion, see diBartolomeo (*Journal of Investing*, 2010) and Northfield News-March 2013

Conclusions

The concept of a linear relationship between expected return and risk pervades the literature of formal equity analysis, forming the basis on which almost all quantitative factor models operate.

While the linear model assumption has proven both intuitive and useful in equity management, straightforward empirical analysis shows that this presumption ***is supported only for the most part.***

Machine learning methods may reveal complex relationships between factor exposure and subsequent returns, but are difficult to rely upon as the reliability of such relationships is hard to assess.

A reasonable middle ground between the rigid assumptions of a purely linear specification, and the "data tells all" of machine learning models is that factor returns may arise from simple but non-linear specifications, such as quadratic or absolute value.