Critique of “Factor Premia and Factor Timing: A Century of Evidence”: The Good, The Bad, The Ugly

Venkat Eleswarapu, Ph.D., Managing Director

August 5, 2019

In a recent paper, Ilmanen *et al.* examine the time-series predictability of four factors (Value, Momentum, Volatility, and Carry) in six asset classes over nearly a century and conclude that they cannot reliably find any evidence of successful factor timing. Let us first preface our comments by noting that they have undertaken an impossible task of proving a negative. As basics of logic and statistical inference would indicate, their inability to find a relationship does not prove that it is *not* there. Arguably, their lack of evidence is driven by the data they use and the statistical techniques they deploy. Our own live investment experience and systematic back-tests lead us to an opposite conclusion (see Eleswarapu, 2019). Here is a critique of their paper:

THE GOOD

They painstakingly compile data for comparable definitions of factors for six asset classes over the longest possible history. The good news is that the four factors deliver positive returns consistently over nearly a century. Also, there is no deterioration in the excess returns in the latest sub-period when compared to the historical period prior to the initial period of study in the identification of the factors in the finance literature. Thus, the factors’ performance is not an outcome of data mining, nor are they being arbitraged away in the recent time period.

THE BAD

Their choice of macroeconomic variables to predict factor premia is questionable. They use *realized* GDP growth rates and inflation rates averaged across four countries (U.S., U.K., Germany, and Japan). Of course, these data are reported with a lag in real-time. Also, the actual reporting lag across the countries and through time would vary. They just arbitrarily assume a one-quarter lag for the entire 100-year period and for all the countries. Also, more importantly, the historical GDP and inflation data are notoriously revised much later. Clearly, the data they use are not point-in-time and not known to the average investor in real-time. There are better ex-ante measures of economic output that they do not use. Even their measures of market volatility are slow-moving by construction since they are based on the past 36 months of return data. It surely is a poor proxy of the conditional stock market volatility.

They combine their predictive signals in simple multivariate O.L.S. regressions and do not use the latest statistical techniques available to optimally extract signals from noisy time-series data. To make it worse, they use a brute-force method and try out 11 signals with 19 regression specifications for all 26 assets (11x19x26 = 5434 tests!) There is no attempt at dimension reduction that keeps the number of estimated parameters to a reasonable set to minimize estimation errors and avoid spurious relationships. This is puzzling given the recent work done on machine-learning techniques by their own AQR colleague, Bryan Kelly (Gu, Kelly, and Xui, 2019). Given the number of tests they run, they then argue that their significance level threshold should be 0.09% (instead of 5%). As a result, they are left to conclude that even a factor-timing strategy that delivers an information ratio of 0.60 out-of-sample over 100 years is not *statistically* significant.

The most surprising finding is the claim that their factor-timing strategy increases annual turnover by 200% relative to the turnover of 415% for their base static model. This points to how *unstable* the estimated parameters are in their approach. In our experience and back-tests, we see only additional turnover of 10 to 15%, on the average, from the dynamic nature of our multi-factor models. For example, the Valant EM strategy that relies on dynamic factor weights for stock selection has annual turnover of less than 100% per year, similar to our quant peers.

THE UGLY

Although they use 11 predictive signals, they spend much of their discussion and computing power to show that the value-spread is not useful to time factors. In Table 6, they implement 19 regression specifications to show that this specific signal may not be useful. This is clearly flogging a dead horse. This is a blind alley that misses on more fruitful avenues of inquiry. Surely, real-time indicators of market and macroeconomic conditions that capture investors' time-varying risk appetite will be more useful for capturing the factor dynamics over time.

In conclusion, although Ilmanen and his colleagues can be commended for their valiant attempt to examine factor-timing across many asset classes over a long time, their failure to find a reliable relationship is *not* proof that it does not exist. Obviously, they cannot prove a negative proposition. We argue that their choice of predictive variables is poor, and their statistical approaches are weak. Thus, we disagree with their conclusion that there is no reliable way to predict factor-premia over time.

REFERENCES

Eleswarapu, Venkat, 2019, Averages are not good enough: A case for dynamic stock selection, Chicago Equity Partners White Paper.

Ilmanen, Antti, Ronen Israel, Tobias Moskowitz, Ashwin Thapar, and Franklin Wang, 2019, Factor Premia and Factor Timing: A Century of Evidence.

Gu, Shihao, Bryan Kelly, and Dacheng Xui, 2019, Empirical Asset Pricing via Machine Learning, University of Chicago, Working Paper.