HIGHVISTA STRATEGIES

Home Runs and Strikeouts:
Complexity Bias in Baseball and Investing

November 9, 2018

Jesse Barnes

Christopher Covington

Rahul Gupta

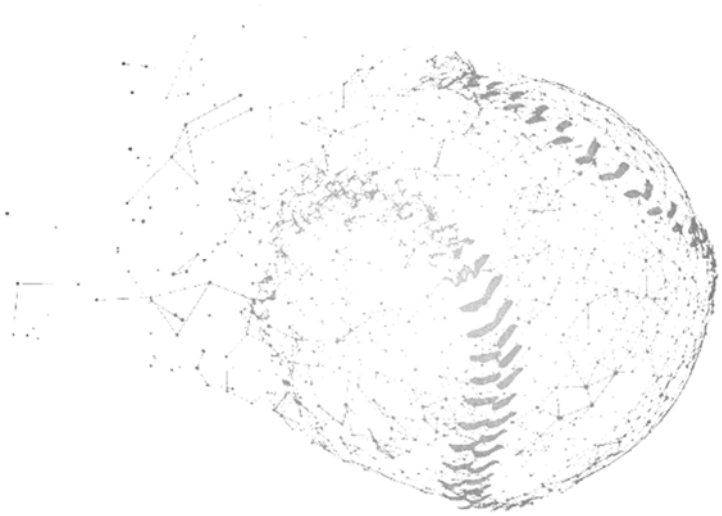
HighVista Strategies LLC
200 Clarendon Street, 50th Floor
Boston, MA 02116
617.406.6500
highvistastrategies.com

HighVista Strategies

HighVista Strategies is a Boston-based investment firm established in 2004. Today, HighVista manages over $3 billion in assets on behalf of institutions and individuals, including systematic investment strategies focused on capturing risk premia to enhance returns. Equity strategies include US, International, and Emerging Markets, which can be constructed relative to desired benchmarks including ACWI, ACWI Ex-US and World.

Executive Summary

Investor interest in quantitative strategies has surged while data availability, computing power, and quantitative talent have never been greater. Both allocators and quantitative investors face a daunting array of choices and investment options. We describe a bias which skews these decisions toward complexity and away from more simple and robust solutions—we call this Complexity Bias. It arises through a combination of three forces: external pressure from clients to innovate, internal pressure from researchers to contribute, and the tempting improvement in backtest performance that complexity inevitably brings. This performance improvement is illusory and failing to recognize and guard against this bias yields opaque investment processes with subpar out-of-sample performance. We utilize a more approachable subject, baseball, to illustrate these principles while demonstrating parallels to quantitative investing themes.



Home Runs and Strikeouts: Complexity Bias in Baseball and Investing

"Every strike brings me closer to the next home run," said baseball legend Babe Ruth, and we can't help but see similarities in the flurry of activity in quantitative investing today.

Teams of Ph.D.'s brimming with complex techniques and unending computing power wildly swing their bats in an effort to explain the unexplained, gain an edge over the rest of the industry and relieve a bit of physics envy. The result is often gratuitous complexity. The recent surge in demand for quantitative investment products and the ubiquity of Ph.D. talent and computing power has created an arms race which has reinforced this complexity as practitioners seek to innovate and expand their offerings (which usually implies more complexity) toward differentiated solutions. The ultimate price is paid by investors who suffer underperformance as the hopes generated by lofty backtests end in strikeouts.

In what follows we use what for many is a more approachable subject—baseball—to explore this phenomenon and its implications for investors. Baseball affords us a neutral ground to illustrate principles that are easily mapped to practical challenges faced by quantitative investors while limiting the impact of our preconceived biases as to what the "right" models should look like. Whether the subject is baseball or finance, forecasting using a quantitative model is ultimately an exercise in disentangling information from an immense amount of noise. Even the most reliable relationships often have low signal to noise ratios that allow many false positives and spurious data to obscure the truth.

We use our simple baseball example to illustrate what we have termed "Complexity Bias" with specific parallels to systematic investing. Complexity Bias arises first from the tendency of more complicated models to show outperformance in backtests—a well-known pitfall in quantitative investing. Unfortunately, this tendency is strongly reinforced by two additional factors. First, external pressure from clients to find new and innovative solutions drives firms to increase research staff, budgets and scope and to reflect less critically on those suspect backtests. This begets a second, internal pressure from these ever-larger teams of researchers who strive to contribute, make a mark, and advance their careers. The result is a bias away from the simple and robust toward the opaque and complex.

Illustrative Example: Predicting Baseball Team Wins

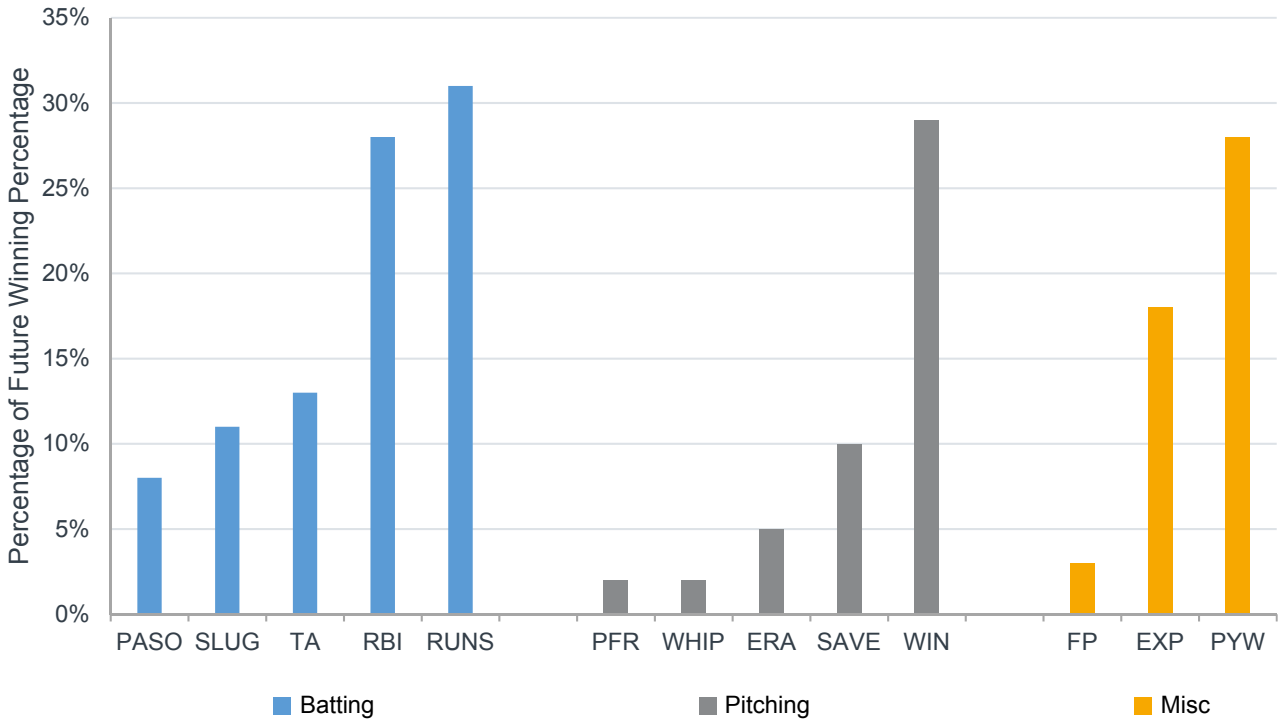
The depth and cleanliness of data for the sport of baseball makes a statistician salivate. It rivals and perhaps surpasses the data available for financial markets and creates a venue to test a variety of techniques. Suppose, for example, you wanted to forecast the performance of baseball teams in terms of winning percentage. There are many variables to choose from: prior year winning percentage, success in hitting (runs scored or batted in), success in pitching (pitcher wins or saves, earned run average), success in fielding (fielding percentage) and other variables such as player experience. A selection of these variables is listed in **Figure 1**. How might one best translate these many data inputs into a forecast for a team's winning percentage?

Figure 1: Selected statistics and descriptions

Factor	Signal	Description
Batting		
	RUNS	Runs Scored: Number of times a player crossed home plate
	RBI	Number of runs attributed to batter action
	TA	Total average: total bases, + walks, + hit by pitch, + steals, - caught stealing / by at bats, - hits, + caught stealing, + grounded into double plays
	SLUG	Number of bases divided by at-bats
	PASO	Number of times at bat for every strikeout
Pitching		
	WIN	Win: number of games where the pitcher was pitching while his team took the lead and went on to win
	SAVE	Save: number of games where the pitcher enters a game led by the pitcher's team, finishes the game without surrendering the lead, is not the winning pitcher
	ERA	Earned run average: total number of earned runs, multiplied by 9, divided by innings pitched
	WHIP	Walks and hits per inning pitched
	PFR	Power finesse ratio: sum of strikeouts and walks divided by innings pitched.
Miscellaneous		
	PYW	Prior Year Win
	EXP	Years Experience
	FP	Fielding percentage: total plays (chances minus errors) divided by the number of total chances

The first step is to identify whether any of these statistics are predictive of winning percentage on their own. To measure this, we first compiled the trailing 2-season average statistic for each player for each year. For each team we then aggregated these statistics using opening day rosters to yield a team score for each variable. Finally, we compared these team scores to winning percentage to measure their usefulness in forecasting. These tests were univariate, meaning we tested each variable on its own, a similar approach to using a single investment factor such as value or momentum. **Figure 2** illustrates the results.

Figure 2: Forecasting percentage of single-variable models



The chart displays the percentage of the future winning percentage that is explained by each variable. For example, the strongest predictor in this sample is RUNS, which on its own explains 31% of the next year's winning percentage. Note that this is highly correlated with RBI (Runs Batted In) and other hitting variables so you can't just add these to get a much higher percentage. For baseball fans, the data suggest some interesting takeaways—for example it appears that both pitching and hitting are roughly equal in importance for predicting team wins. Note also that years of experience has a strong positive relationship with wins, which is perhaps less intuitive in baseball than it is in investing.

To construct a forecast using these data there are at least three key questions to answer:

- Which factor(s) to use?
- How many factors to use?
- How to combine them into a forecast?

Each of these questions alone deserves its own paper and all are susceptible to Complexity Bias. For brevity, here we focus on the latter two, though we present an interesting aside on the first question in the Appendix.

How Many Factors to Use?

While one simple approach might be to build a model using only the "best" factor, this seems clearly suboptimal. If for example a measure of hitting (such as Runs Batted In) is the single most predictive variable, wouldn't including pitching in some way in our model improve our results? It is intuitive that additional factors may improve the model because they may measure different things which are not perfectly correlated to each other. Hitting and pitching in baseball are the perfect examples, but in the finance world Value and Momentum might be similarly unrelated to each other, yet both intuitively linked to future performance. Including an additional factor should not only improve average model performance (making better predictions of future wins) but should also make our model more robust (less prone to underperformance when one factor stops working).

This intuition is proven out in our baseball data. In **Figure 3**, pitcher Wins underperforms Batter Runs by a significant amount for most of the period; however, this relationship reverses for most of the last two decades. Notably, the simple combination of the two factors is the most accurate throughout the entire forecasting period.

Figure 3: 10 Year rolling predictive accuracy (rank correlations)

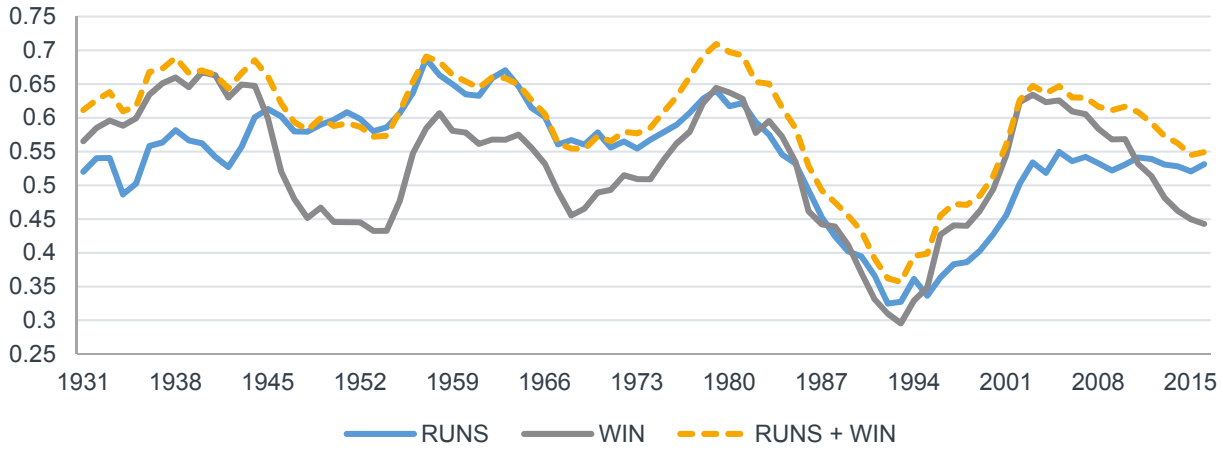
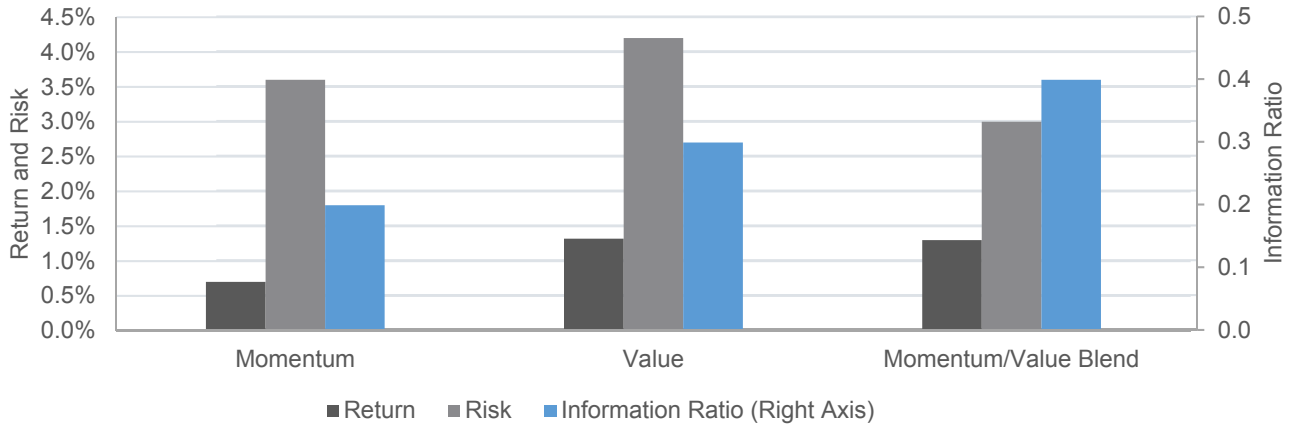


Figure 4 illustrates this same principle in financial markets using Momentum and Value. Note that the two-factor model has a substantially higher Information Ratio (return per unit of risk—in blue) than either factor alone, which indicates a more consistent outperformance.

Figure 4: Combining Momentum and Value

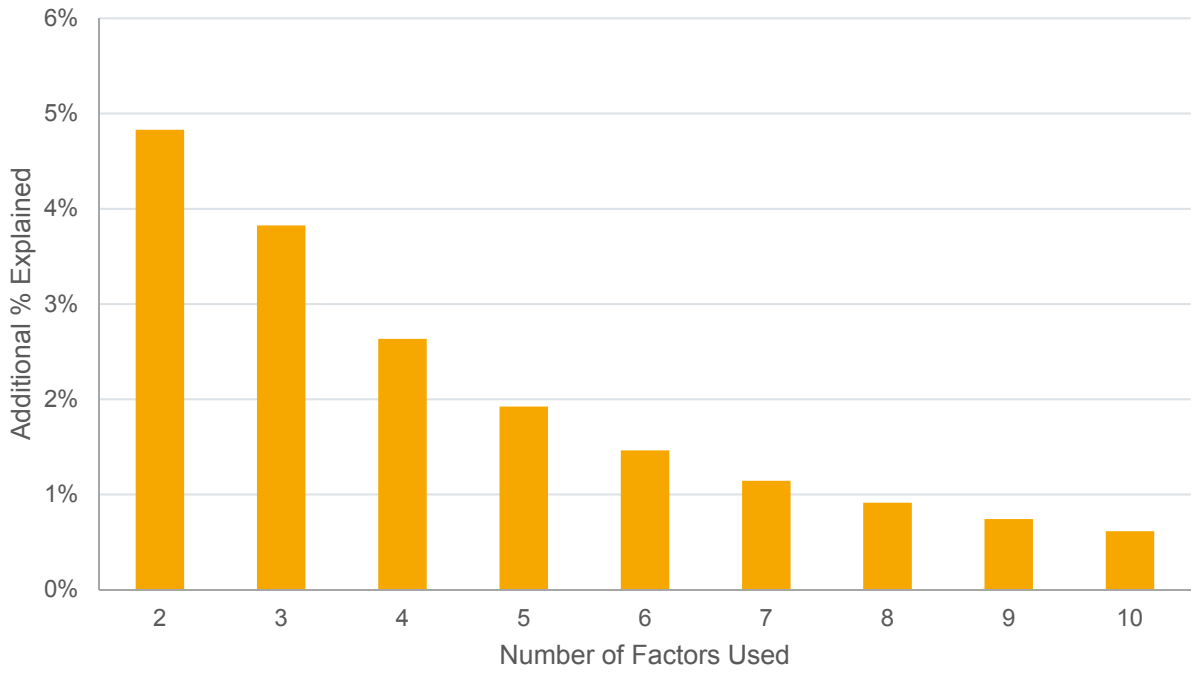


Diminishing Returns from Additional Factors

The same intuition that leads to the conclusion that two factors may be better than one might lead us to conclude that many factors would be even more powerful and robust. In baseball we might include not just Runs but other hitting statistics such as hitting percentage, slugging, and RBIs, and not just Pitching Wins but ERA, strikeouts, etc. And then there are many other factors that might also have some explanatory power: player experience, manager history, strength of schedule, fielding skill, etc. Given an overwhelming amount of data and today's incredible computing power, we might expect to build a comprehensive, powerful, and robust predictive model!

While additional complexity—including additional factors—is additive, there are diminishing returns. The benefit from adding a third factor is likely less than going from one to two, and the benefit from adding the tenth factor will be far lower still. This is illustrated below in **Figure 5** which shows the marginal benefit in predictive percentage from adding additional factors to our baseball model.

Figure 5: Percentage improvement in backtest forecast accuracy from additional factors



The forecast improvements in Figure 5 above are all positive by construction—adding more factors will always be helpful in a backtest. Unfortunately, there is an invisible counter-weight which moves in precisely the opposite direction: the more factors we include the more noise our model will mistake for signal and the larger the gap will be between backtest and live performance. Given that adding many factors adds only marginally even in the backtest, we can be confident that their net effect on live performance will be negative.

This same principle holds in a quantitative investment model, where the number of factors to choose from may be even greater. With clients eager to know that they are invested with a manager that is on the cutting edge, the incentive is to err on the side of adding too many factors rather than too few. As new and alternative data sources or factor methods arise it is inevitable that clients will ask about their potential inclusion, and reticence to incorporate them may risk being viewed as less sophisticated than more aggressive peers.

Combining Factors

Given an appropriate number of factors, the next challenge is how to sensibly combine them in a model. In the two-factor model above we took a simple average, which worked pretty well. As the number and variety of factors increases this method seems inadequate: is strikeouts just as important as pitching wins? Or hitting percentage the same as ERA? Some of these will have higher or lower correlations to other factors and will have higher or lower explanatory power for future wins.

There are many methods for combining factors that range from very simple to very complex. Here is another source of Complexity Bias—the temptation is to attempt to extract every last ounce of information from our factors through complicated analysis. Models that are more complex have strong intuitive appeal as the world (in baseball and investing) is complex! It is very easy to identify shortcomings in simple models and related suggestions to make them "better". More complex models can account for relationships among the factors, how these change over time, their differing magnitudes and correlations to each other, and the precise accounting of each that optimizes future win predictions.

Figure 6 presents four examples of potential approaches of combining factors with varying complexity. The actual menu of approaches is nearly unlimited, but these give a sense of the types of models one might consider.

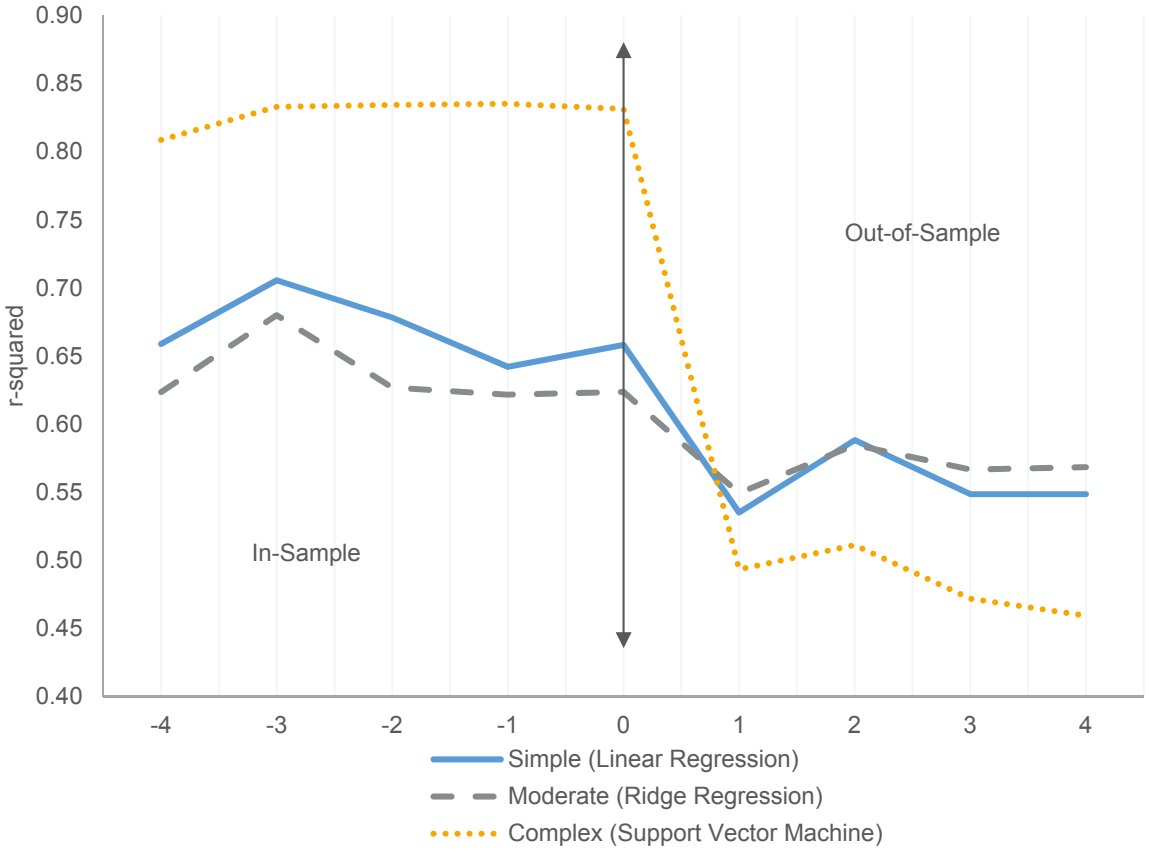
Figure 6: Approaches to combine factors

	Very Simple	Simple	Moderate	Complex
Method	Simple Average	Standard Linear Regression	Ridge Regression	Support Vector Machine
Opacity	Clear	Pretty Clear	Pretty Clear	Fully Opaque
Explanation	Simple blend	Using past relationships to determine how much of each	Using past relationships to determine how much of each, but with skepticism in statistical relationships (shrinking strength towards zero)	Transforming data repeatedly until a relationship emerges

A key point of our thesis is that Complexity Bias arises because complexity always outperforms in a backtest. To illustrate this we used the models in Figure 6 to forecast baseball team wins. For a 10-year period we built each of these models, estimated their parameters, and predicted team wins over that same period. We then averaged the performance over all the 10-year periods to evaluate how well these models did in forecasting. It is critical to note that these forecasts are *in-sample*—their performance is measured in the same data set that was used to construct them. This would be called a backtest in the financial world. We then also calculated the performance of these models over the 5-year period following each 10-year test period. Because we did not use this 5-year data to construct or calibrate the model it is *out-of-sample* or analogous to live performance in the financial world. For each 5-year out-of-sample period we measured forecasting performance, then averaged over all periods to evaluate how each model did in live performance.

Figure 7 below plots the result. The x-axis measures time with the vertical line at 0 representing the transition from in-sample to out-of-sample years. The colored line for each model represents the accuracy for that year on the y-axis. As expected, the most complex model is by far the best at predicting wins for the in-sample period, explaining better than 80% of team wins vs closer to 60% for the simpler models.

Figure 7: Model complexity in-sample and out-of-sample



More interesting however is the out-of-sample 5-year period. Note the strong convergence of the lines in the first observed year (+1 on the x-axis) after the tested period. This is the first "live performance" year for the models and stands in clear contrast to the performance in the preceding years. In the out-of-sample period the most complex model does not outperform, in spite of its promising performance in-sample. This demonstrates a key theme of Complexity Bias. Complexity always yields a better in-sample performance—a better backtest—which draws research interest like a hanging pitch over the middle. But like a well-delivered curveball, it is an illusion. The live performance disappoints, the batter whiffs, and we are left wondering what went wrong.

What went wrong is that there is a strong trade-off between complexity and intuition, which is never evident in a backtest but is sadly made clear in live performance. We have become incredibly adept at extracting every bit of information from a particular data sample to maximize model power in what we have already observed. But this also dramatically increases the likelihood that our models will be built on randomness, noise, or circumstances that are unlikely to persist. Opacity and complexity are natural enemies to long-term robustness.

The Right Tool for the Job

The first key to balancing these tradeoffs in a model is to recognize the inherent but invisible loss that increased complexity brings. One side of the equation (better backtests) is clearly visible while the other side (actual future performance) may not be. Much of quantitative investing rests on the behavioral biases inherent in human nature—we should recognize that those same biases in ourselves may lead to poor outcomes in our model construction or factor selection.

More fundamentally, it is critical to have a firm intuition about the models we deploy and to understand the tools at our disposal so that we may select the correct tool for the task at hand. For example, in the comparison of model performance in Figure 7, the technique that works best out-of-sample (Ridge Regression) is built with an inherent skepticism in the statistical relationships that appear in the data and is well suited for variables that are correlated. As noted earlier, many of our baseball statistics are highly correlated, making this a fair choice in this situation. In another data set this may be overly complex or overly simplistic or simply an inferior fit relative to an alternate method.

This brings us to a popular buzzword in the investment industry today: *Artificial Intelligence ("AI")*. This powerful tool has been deployed in many applications and has tremendous promise in many fields including quantitative investing. It is in some sense the perfect example of non-intuitive model construction: unlimited flexibility combined with unlimited computing power yields a model which perfectly extracts all information from a data set. As our example earlier demonstrates, this does not mean it will perform better out-of-sample. This is not an attack on machine learning or AI—they are tools we know and love as much as any other quant. But for their power to be put to good use it is necessary to be very careful where they are deployed. For example, using them to estimate a particular parameter or factor within a broader model—where their "sandbox" is circumscribed—can harness their power toward a broader and more intuitive end, while limiting their potential to lead to poor out-of-sample outcomes.

Conclusion – Avoiding Complexity Bias

When given a new toy, a child's focus becomes unilateral—life is this toy! Likewise, when a quant learns of a new method or factor, he or she must find a way to deploy it to better forecast returns, improve batting average, etc! As computing power has grown over the past decade, many new toys have become available. Above the behavioral aspect, there is unending external pressure to innovate. Clients want to ensure they are invested with firms that are on the cutting edge, not those becoming stale. Adding to these pressures, complexity looks tantalizingly impressive in backtested results! All of these lead to what we term ***Complexity Bias***. As noted above, some complexity is necessary in modelling and forecasting. In the spirit of Occam's Razor, however, complexity should only be added while recognizing the inherent costs that may not be apparent at the outset. Understanding and utilizing the right tools for the right job, focusing on intuitive relationships and parameters, and requiring rigorous analysis of performance out-of-sample are key protections from this bias. The measure of a good quantitative process is as much what it does not do as what it does, for this is the more difficult force to resist.

Appendix – An Aside on Factor Selection

In our baseball example above we focused on the number and combination of factors, but which factor(s) to use is also a critical question. It is tempting to focus solely on the factor(s) that has performed best historically. For example, from Figure 2 above one might build a model based solely on batter runs given that it outperformed all other factors over time. The fallacy in this approach is revealed by a familiar phrase in the finance industry – "Past performance is not indicative of future returns". Just because a factor has been predictive in the past does not mean it will be most predictive in the future. Has the game changed over this period? Have pitchers become better or worse? Has game strategy evolved? Have the rules or schedule or technology made pitching or hitting more or less valuable than it was in past years?

To illustrate this point consider that the best-performing model by far in 1970 would have been Previous Team Wins (the number of games the team won the prior year). Through 1970 this had been a strong and reliable predictor and substantially outperformed all other potential factors. Historical data would have fully supported using this variable to predict team wins going forward. Unfortunately, its future as a predictor was poorer than one would have expected in 1970 owing to a critical development in Major League Baseball: the advent of free agency. With this change players obtained more freedom to move between teams, which meant that the composition of teams was less stable year-to-year. A team's previous year wins were less reliable as a predictor of future wins because the players that earned those wins might not be on the team anymore! Failing to account for this change would have left the naïve quant with a poor model that would have required years to correct using data alone.

This issue has a clear parallel to those questions that should be asked in factor selection for investing as well. If a factor has outperformed historically, could this be due to market structure, investor behavior, or data availability or technology that were different than they are today? History must be a guide but is dangerous when relied upon exclusively without a more complete understanding of how each factor relates to investor behavior and return.