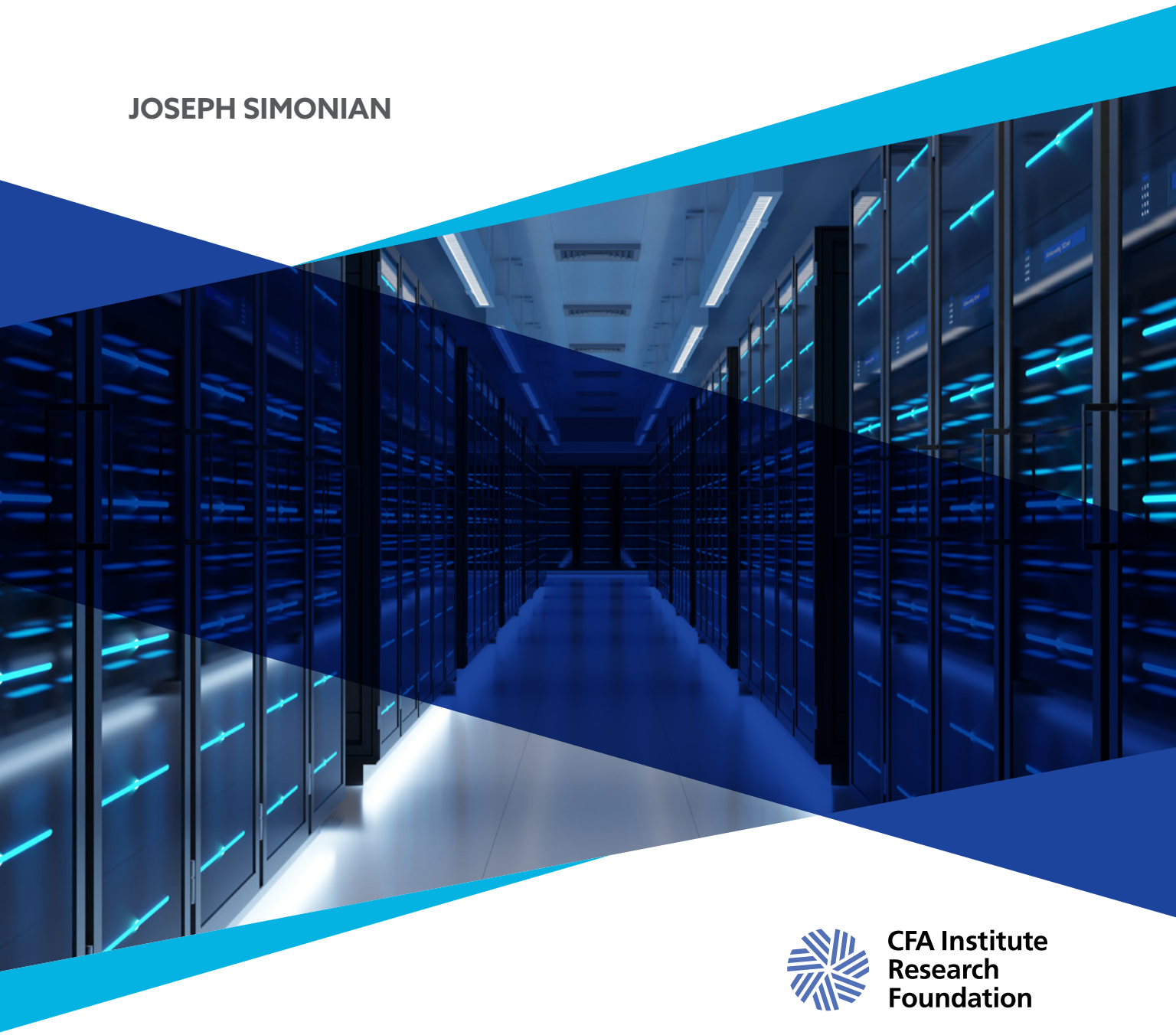
CFA INSTITUTE RESEARCH FOUNDATION / **MONOGRAPH**

INVESTMENT MODEL VALIDATION
A GUIDE FOR PRACTITIONERS

JOSEPH SIMONIAN



CFA Institute Research Foundation

INVESTMENT MODEL VALIDATION

A GUIDE FOR PRACTITIONERS

JOSEPH SIMONIAN

CFA Institute
Research
Foundation

Statement of Purpose

The CFA Institute Research Foundation is a not-for-profit organization established to promote the development and dissemination of relevant research for investment practitioners worldwide.

Photo credit: Jasmin Merdan / Moment RF / Getty Images

ISBN: 978-1-952927-43-0

BIO

Joseph Simonian is a globally renowned investor and researcher who has conducted extensive research in quantitative finance, machine learning, factor investing, and portfolio construction. Over the course of a 20-year career in the investment industry, he has held senior portfolio management and research positions in several prominent asset management firms. He is also the founder and CIO of Autonomous Investment Technologies.

Simonian is a noted contributor to leading finance journals and a prominent speaker at investment events worldwide. He is co-editor of the *Journal of Financial Data Science*, is on the editorial board of the *Journal of Portfolio Management*, chairman of the board of directors of the Financial Data Professional Institute, and a member of the CFA Institute Research and Policy Center Technical Committee. Simonian has authored over 40 publications in leading investment journals. He is co-author of the book *Quantitative Global Bond Portfolio Management* and author of *Computational Global Macro*, slated for release in 2024. In addition to his portfolio management and research activities, Simonian has extensive experience teaching in both academia and industry.

He holds a PhD from the University of California, Santa Barbara; an MA from Columbia University; and a BA from the University of California, Los Angeles.

CONTENTS

CFA Institute

PROFESSIONAL LEARNING QUALIFIED ACTIVITY
This publication qualifies for 1.75 PL credit under the guidelines of the CFA Institute Professional Learning Program.

FOREWORD

In the complex and dynamic world of finance, the importance of rigorous financial model validation cannot be overstated. As financial instruments and markets have become increasingly intricate and intertwined, the need for robust models for alpha generation and risk management has never been more critical. Today, these models play an indispensable role in driving decisions that can have far-reaching ramifications, impacting everything from individual investment choices to the stability of the global financial system. Model validation is thus a cornerstone of contemporary financial management. When implemented properly, model validation entails a meticulous verification of the soundness and reliability of a model's mathematical foundations, assumptions, and outcomes. In an environment where errors can lead to significant financial losses and an erosion of client trust, the value of thorough model validation cannot be underestimated.

Recognizing the industry-wide need for a framework for conducting proper model validation, Joseph Simonian has written a comprehensive guide to the essential practices and principles of financial model validation. Simonian brings a wealth of expertise and experience, drawing on both his extensive record of investment research and years of investment practice. His insights offer a rich and nuanced perspective on all facets of contemporary model validation.

This monograph covers a wide array of topics essential to understanding and implementing effective model validation. From foundational approaches to advanced techniques, each section is designed to equip the reader with the knowledge and tools necessary to effectively validate a wide variety of investment models. The monograph should serve as both a practical guide and a compendium of research sources with which investment practitioners can develop their own specific model validation processes. Whether one is a seasoned professional or a newcomer to the investment world, this monograph should be an indispensable resource for years to come.

Frank J. Fabozzi, CFA

INTRODUCTION: THE IMPORTANCE OF MODEL VALIDATION TO FINANCIAL PRODUCT DEVELOPMENT

Our model of Nature should not be like a building—a handsome structure for the populace to admire, until in the course of time someone takes away a corner stone and the edifice comes toppling down. It should be like an engine with movable parts.

—Sir Arthur Stanley Eddington

Investing is generally not considered one of the "healing arts," but it should be. Just as medical practitioners are concerned with the physical well-being of their fellow humans, investment managers are focused on improving the financial well-being of individuals and organizations. And as is well known, a person's financial well-being is often intimately connected to their psychological and emotional health.

But the analogy between medicine and investing can be carried further. Just as medical professionals use pharmaceutical products to help individuals attain physical health, investors use financial products to help their clients attain financial health. Building financial products is therefore no different from building any other type of product: Great care must be taken to ensure that the product delivered to consumers is robust and reliable. While quantitative managers rely on models more heavily than do fundamental managers, the vast majority of portfolio management teams use models of some type to help them develop their strategies. However, in any application of models, investors must confront *model risk*: the risk that the models they are using are less than robust or are being incorrectly applied in some way.

While there are similarities between pharmaceutical and financial product development, there are limits to the parallels that can be drawn. Both financial and pharmaceutical products go through lengthy product development cycles, but there is a considerable gap between the level of scientific rigor applied to the evaluation of each product type. Pharmaceutical products go through an extensive process of testing and approval, culminating in commercial approval by, for example, the US Food and Drug Administration (FDA). Along the way, numerous trials and studies are conducted, benefiting from multitudes of test subjects that assist pharmaceutical companies in developing the drugs they will ultimately bring to the market. In contrast, the investment industry has neither the benefit of such an approval process nor a regulator such as the FDA that has the ability to pass judgment on whether a given financial product is likely to do more benefit than harm (the basic standard that the FDA uses). Instead, most people simply assume that investment firms are making good faith efforts to thoroughly and carefully develop models and investment products that will ultimately help investors in successfully achieving their financial goals.

When the foregoing assumption proves false and firms use flawed or outdated models, they (and their clients) may experience substantial financial losses. By thoroughly validating models, portfolio managers and researchers can identify potential weaknesses, biases, or inaccuracies in the models' assumptions and outputs and thus reduce model risk. This process helps in refining models to better capture the underlying dynamics of financial markets and adapting them to changing conditions. In essence, model validation serves as a tool for enhancing the overall resilience of investment strategies for client benefit.

Related to model risk is the *reputation risk* that faulty model validation presents. Validating models is imperative to ensure their relevance and performance in real-world scenarios and to establish and maintain investor trust. Clients and stakeholders rely on portfolio managers to make informed and prudent investment decisions. If the models underpinning these decisions are not validated, the integrity of the entire investment process can be called into question. Properly validated models provide a transparent and defensible framework for decision making, instilling confidence among investors and enhancing the credibility of portfolio managers and their strategies.

While clients are undoubtedly concerned with their financial well-being, so are regulators. Financial institutions and asset managers are subject to myriad regulations that, explicitly or implicitly, demand the use of sound and validated models for decision making. Regulatory bodies, such as the US Securities and Exchange Commission (US SEC) and the European Securities and Markets Authority (ESMA), have increasingly emphasized that investment firms must adhere to best practices and maintain a high standard of due diligence. Failure to do so can result in severe legal and financial consequences.

To sum up, firms need to be concerned with developing comprehensive model validation processes in order to ensure the highest-quality products for their clients and to protect their businesses from financial and legal repercussions. This monograph is designed to equip investment professionals with the knowledge and tools that will allow them to implement a rigorous approach to model validation by providing a practical yet detailed overview of the various model validation methodologies that investment practitioners have at their disposal.

Roadmap to the Monograph

- The topics covered in this monograph include empirical methods for testing models, both traditional approaches and more recent, data science–driven approaches. While model validation is often thought to be synonymous with backtesting, it is, in fact, a significantly broader component of investment practice than many think.
- Accordingly, the discussion of model testing is accompanied by an overview of the most important aspects of performance measurement and benchmarking. Because data paucity is a major obstacle to conducting proper model validation, this monograph also provides a comprehensive overview of the various methods of creating synthetic time series, including those based on machine learning techniques.
- Rounding out the monograph are examinations of the role of investment theory in model validation and the importance of proper documentation of the validation process.

By providing a comprehensive discussion of the frameworks and techniques that can be used to assess the accuracy, reliability, and appropriateness of the models that drive investment processes, this monograph seeks to enhance the practice of investment product development. Thus, the ultimate beneficiary of this monograph will be the investing public that uses the services of professional investment managers.

1. A WORKING PHILOSOPHY OF MODEL VALIDATION

Model validation is the process used to verify and validate financial models to ensure that they meet their intended business use and perform within design expectations. There are a number of reasons why model validation in finance is more challenging than in the natural sciences. One of the primary reasons is that finance does not have the benefit of the rich datasets that are available to natural scientists. For example, instead of having hundreds of test subjects (e.g., in Phase 3 drug trials) each with their own history, finance has one S&P 500 Index history, one Russell 2000 Index history, and so on. The relative paucity of data presents a serious challenge to the developers of financial products to test and validate the strategies that they will ultimately offer to their clients. Moreover, unlike research in the natural sciences, research in finance does not have the benefit of closed experiments. This further complicates investment researchers' ability to draw robust and generalizable conclusions from their work. Perhaps the biggest challenge for investment research, however, is that unlike phenomena in the natural world, the primary drivers of financial markets are human psychology and intentionality, which are not mechanistic in the way that many, if not most, natural phenomena are. For these reasons, portfolio managers and researchers are at risk of being easily lulled by initially successful results generated by simple backtests. Most investment and trading strategies developed in this manner will be "false positive" strategies that ultimately end up collapsing under greater scrutiny or real-time testing in an actual portfolio.

As markets have become more saturated with information and data, the challenges to developing robust investment strategies and building financial products that can withstand the multitude of market gyrations, macroeconomic shocks, and political headlines have become even greater. These factors can compromise the accuracy and robustness of investment models. Anyone who has ever been on a portfolio management or research team and been part of the process to develop successful investment strategies has undoubtedly experienced these challenges.

Thus, to ensure that asset owners have access to investment products that possess the requisite level of robustness, investment firms must have in place a comprehensive model validation process. However, as critical as model validation is for the reliability and effectiveness of investment strategies, it is remarkable how decidedly *unscientific* investment strategy development and model validation often are.

This situation is all the more surprising given the "science envy" that economics and finance have had for the last century. How do scientists test models and theories? Their process can be summed up in one word: *falsification*. That is, a scientist will develop a theory or build a model and then proceed to subject it to numerous empirical and logical tests in an attempt to falsify it. If the theory or model cannot be "broken" despite the scientist's falsification attempts, then the theory or model will be accepted as providing some explanatory value.

In contrast, in the asset management industry, strategy and model development often proceeds in the opposite manner. Research teams build various versions of a given model, conduct some simple historical backtests, and scream "eureka!" when they discover one that works well over the specific time period they are using to conduct the analysis. This behavior is known as

data snooping and is likely to result in false positive strategies. To achieve the type of model development encapsulated in Eddington's quote that opens this monograph, a much more exhaustive and rigorous set of tools must be used.

With these considerations in mind, it could be useful to develop a general approach to model validation that is inspired by the concept of *prophylaxis* from the game of chess.[1] This concept, which has become an important element of every top chess player's arsenal, also holds valuable lessons for anyone developing and validating investment models. Prophylaxis is the idea that one's moves in a game of chess should not only advance one's immediate position in the game but also serve to prevent an improvement in an opponent's positioning. Moves that are considered prophylactic accordingly require a player to think ahead and contemplate an opponent's likely responses to one's moves. Investment professionals can likewise benefit by considering counterarguments and counterexamples to their models. For example, in validating any model, it is important to consider potential criticisms relating to the benchmark(s) and performance metrics used, the likely impacts of turnover and transaction costs on model implementation, the number and type of robustness checks used (e.g., cross-validation tests), and the consistency of the model with economic and investment theory.

[1]Notable practitioners of prophylaxis in the history of chess include Aron Nimzowitsch, Tigran Petrosian (world champion from 1963 to 1969), and Anatoly Karpov (world champion from 1975 to 1985). Petrosian in particular was perhaps the most dedicated adherent to prophylaxis in the history of the game. As Bobby Fischer once exclaimed, "He will 'smell' any kind of danger 20 moves before!"

2. BACKTESTING

Backtesting is a procedure that examines and assesses the historical performance of a model by comparing its predictions with actual outcomes. This retrospective analysis helps researchers understand a model's respective strengths and weaknesses in different historical periods. It also enables researchers to assess whether a strategy is more likely to make or lose money and measure the frequency of wins versus losses, among other relevant statistics. In addition, examination of the historical performance of a model helps discern the potential sources of its success or failure. For example, was a model able to detect profitable market trends or changes in market liquidity? And in what market periods or "regimes" was it most successful? A careful study of the historical track record of a trading model can give researchers a good idea of what conditions will be beneficial or harmful to a model's performance in actual market conditions.

At the outset, it is important to note that backtesting is not only used to test a finalized model in the validation phase of model development but is also often used in the earlier stages of product design. For example, based on some economic reasoning or observed regularity in the market, researchers often subject their initial "toy" models to a basic backtest. Moreover, as a model is further refined, it is generally subjected to backtests on a continual basis.

Thus, backtesting is among the basic validation tools used throughout the model's development; rather than a static "one-shot" process, it is dynamic in nature. As market conditions, economic and political variables, and investor behavior evolve over time, regular backtesting allows portfolio managers to adapt their models to changing market developments, ensuring that the models remain relevant and accurate. This iterative process of validation and adaptation is essential for staying ahead of market trends and maintaining a competitive edge over competing firms. Taking a proactive approach to backtesting acknowledges the inherent uncertainties in financial markets and seeks to enhance the robustness and reliability of investment models. As the financial landscape continues to evolve, the importance of rigorous and detailed backtesting will only grow, in lockstep with the need for more informed and resilient investment decisions.

Aside from helping determine the potential effectiveness of a given investment model, backtests can also provide insight into the challenges that investors may encounter when trying to implement their models in live trading. For example, the impact of *transaction costs*, which are the direct or indirect costs of trading a portfolio, can be studied in a backtest. Direct transaction costs include trading fees and commissions. Indirect transaction costs include the costs stemming from bid–ask spreads and market impact, the latter being the difference between the actual price of the security and the price of the security in the absence of the transaction. It is sometimes the case that a particular investment model provides useful information yet shows itself to be unprofitable in practice due to the costs associated with implementing it. If transaction costs erode performance to a significant extent, the model may be deemed too impractical to implement in practice.

Another implementation issue could be related to the types and sizes of trades that the model uses to produce any observed performance. For example, imagine a trading model that produces superior results in a backtest spanning multiple decades, but upon further examination, it is revealed to have generated the majority of its gains from a small number of trades scattered over the life of the backtest. This finding could indicate that the model is not actually identifying any systematic and exploitable pattern in the markets but is merely capitalizing on a few "lucky"

trades. If this is the case, then it is unlikely that it will produce superior performance in the future. Or consider a case where an equity model's performance is based on successfully trading a small subset of its investment universe. If the model is only informative regarding a small handful of stocks, it could indicate that the model is not generalizable to a sufficient degree. Moreover, such a strategy, if implemented in a live investment product, could subject clients to undue concentration risk.

A final implementation issue that backtests can help resolve is related to the impact of various operational time lags on a strategy's real-world effectiveness. When an investment model is developed, it is important to examine its effectiveness in realistic operational conditions where the ideal of "instantaneous trading" is jettisoned. As known from experience, such ideal conditions do not exist during live trading. Rather, delayed execution, for market and firm-specific reasons, is the norm. It is inadvisable to assume that a signal received at time t can also be used to trade at t. It is more realistic to assume that the trade occurs at $t + 1$ or $t + 2$. Delays in execution can also arise due to lagged releases of data that are pertinent inputs to a model. For example, macro data are often released with a lag. If a model's effectiveness in a backtest is materially reduced when one assumes that such execution delays occur, then it would be an indication that the model is not tenable as a commercial product.

Of course, even if a model demonstrates its viability in a backtest, it is not guaranteed that it will perform well in the future. As discussed in later chapters, because of the risk of producing "false positive" strategies based on sole reliance on historical data, backtests cannot be the only tool for model development and validation. Indeed, the validation procedures discussed later are designed to measure models' predictive power and robustness in ways that backtests cannot. However, as a way to filter and weed out models and strategies that have performed poorly over the long term, a backtest is an invaluable tool.

The Building Blocks of Backtests

Backtests are a type of empirical study over a specified historical period. Choosing which historical period to use in a backtest, however, is not as simple as it may seem. Because investment research is data poor relative to the natural sciences, the first instinct of any researcher is to try to obtain the longest time series possible. Accessing the largest data sample is understandable from a purely statistical standpoint. But in an economic or financial study, doing so is not always advisable because economic phenomena change more rapidly than natural phenomena, which may result in very long time series containing information that is not relevant to understanding contemporary markets. For example, consider that as late as the 1930s, the horse and buggy were still in use, albeit rapidly declining in popularity due to the rise of the automobile. Is such an economic world relevant to today's markets? The technological landscape may be less relevant for the modeling of some asset classes, such as commodities, but for most asset classes, including equities, technological considerations are important. In addition, fixed-income markets are continually evolving because of changing fiscal and monetary policies. Thus, no universally appropriate time-series length exists. Proper model validation, however, must account for the economic relevance of the data or risk implementing models that are insufficiently responsive to contemporary market behavior.

In this monograph, the term "backtest" generally refers to *out-of-sample* tests. These are backtests where a model is developed using a subsample that is separate from and chronologically prior to the subsample used to evaluate the efficacy of the model. In contrast, *in-sample* refers

to tests in which the data used to develop a model are the same as those used to evaluate it. Such tests are a form of "cheating" because they do not provide an objective and unbiased method of validating the effectiveness of a model. Further, they do not provide any useful information regarding the likelihood of a model performing successfully on new sets of data.

While it is common for researchers to conduct backtests over one long, contiguous time horizon, it is also often useful to split a time series into a sequence of economically meaningful subsamples. One rigorous way to do this is to use what are known as *regime-switching* models. While the forecasting power of most regime-switching models has been poor,[2] they can be used as effective tools for classifying time-series data in econometrically identifiable ways. For example, consider the regime-switching framework presented by Filardo (1994), which is an extension to Hamilton's (1989) well-known regime model. Filardo followed Hamilton in setting up his model as an autoregressive process:

$$y_t = \mu_{S_t} + \phi_1(y_{t-1} - \mu_{S_{t-1}}) + \phi_2(y_{t-2} - \mu_{S_{t-2}}), \tag{1}$$

where

y_t is the dependent variable

μ_{S_t} is the intercept

s_t is a regime process

ϕ is the autocorrelation with random shocks $\varepsilon_t \sim N(0, \sigma^2)$

Period-to-period transitions from one regime to another are expressed as follows:

$$P(S_t = s_t \,|\, S_{t-1} = s_{t-1}) = \begin{bmatrix} p_{00,t} & p_{10,t} \\ p_{01,t} & p_{11,t} \end{bmatrix}, \tag{2}$$

where $p_{i,j,t}$ is the probability of transitioning from regime i to regime j in period t. However, Filardo extends Hamilton's model by also allowing for exogenous regressors to influence transition probabilities:

$$p_{i,j,t} = \frac{\exp\left\{x'_{t-1}\beta_{i,j}\right\}}{1 + \exp\left\{x'_{t-1}\beta_{i,j}\right\}}. \tag{3}$$

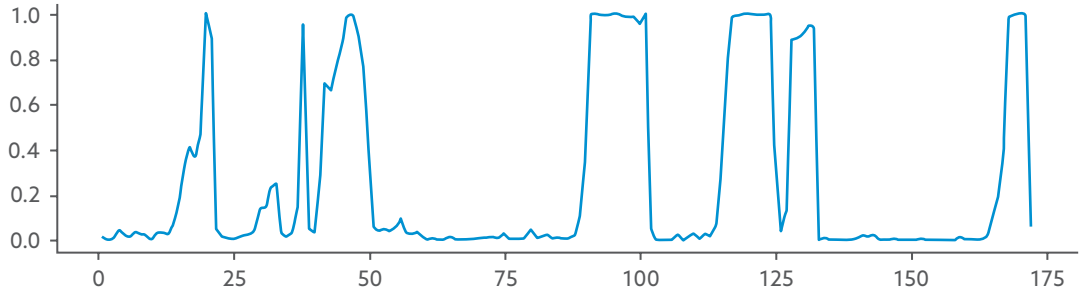
In Equation 3, $\beta_{i,j}$ represents the coefficients that relate the transition probabilities to a vector of exogenous regressors x_{t-1}. Using macro variables to classify regimes is common—for example, using the change in gross domestic product (GDP) as an endogenous regressor with the 10-year yield as the exogenous regressor.[3] **Figure 1** shows sample output from the latter model using quarterly data from 1978 to 2023.

---

[2]See Simonian and Wu (2019) for an in-depth discussion of this point.

[3]According to the Federal Reserve Bank of St. Louis, change in GDP is proxied by "Real Gross Domestic Product, Percent Change, Quarterly, Seasonally Adjusted Annual Rate" and the 10-year Treasury yield is proxied by the "Market Yield on U.S. Treasury Securities at 10-Year Constant Maturity, Quoted on an Investment Basis, Percent, Quarterly, Not Seasonally Adjusted."

Figure 1. Example Regime-Switching Model Output: Probability of Low-Growth Regime (*y*-axis) over 180 Time Periods (*x*-axis)



Source: Data from the Federal Reserve Bank of St. Louis.

In addition to determining the time frame over which data should be drawn and how the historical data being used should be classified, it is also important to ensure that the data being used are free of error. The quality and accuracy of input data are critical because they form the foundation of the backtesting process. While most data provided by vendors are usually relatively error free, errors may be introduced into time series once they are processed by a firm internally by various systems and personnel. Thus, it is important to review each time series for missing, duplicate, and/or erroneous values. One way of doing so could be to compare the same time-series data provided by different vendors (e.g., S&P 500 price levels). If there are discrepancies between them, they should be investigated and resolved. Today, much of this work can be automated.

Running a Backtest

Once the preliminary setup has been accomplished, running a backtest is relatively simple. The basic requirements are as follows:

- One or more *signals*: If a model is driven by more than one trading signal, each can be tested independently or in unison with the other signals. Indeed, it is often useful to test signals individually and in various groupings to see how they perform on their own and in different combinations with other signals. It is often the case that a given signal will show itself to be relatively uninformative on its own yet highly informative when used as part of a set of signals. Discovering such "informational synergies" can be an important part of the model development process.
- The establishment of *entry/exit rules*: Often a signal is broadly directional and must be refined in order to function in practice within an actual commercialized strategy. Part of this refinement involves determining when assets will be bought or sold based on the signal(s) emanating from a model. It is often an iterative process. Through continued testing, the optimal thresholds for opening and closing trades can be found, which will generally involve consideration of various performance and risk metrics, operational feasibility, and transaction costs.

- *Trade sizing* and *rebalancing rules*: Signals and entry/exit rules tell you what and when to trade, but they do not tell you how much. Backtests should also look at how a model performs under various trade sizing protocols. In a backtest, if a model is observed to generate sufficient returns only with large, concentrated trades, a strategy that is driven by the model might be subject to significant downside risk, suffering large losses in the event that a particular trade does not work out. Related to trade sizing is *rebalancing*, which is the process of realigning a portfolio's asset weights to a predetermined benchmark or reference weighting. Rebalancing ensures that a portfolio's positions do not drift too far from what are deemed to be their longer-term or "strategic" portfolio weights, which are generally selected in accordance with the risk and return objectives of the strategy in question. Backtests can be useful in providing insight into the impact of different rebalancing rules on a model's performance. There is no fixed rebalancing frequency to which portfolio managers must adhere. However, a set of acceptable rebalancing frequencies is usually a stated or predetermined part of a portfolio management team's overall investment process.

Possible Pitfalls in Backtests

While backtests are useful tools for model validation, they can also lead researchers astray, namely by imbuing them with more confidence in a given model than is justified. Thus, various pitfalls and challenges are associated with backtesting that need careful consideration. One major risk is the potential for overfitting the strategy to historical data. Overfitting occurs when a strategy is tailored too precisely to past market conditions, performing exceptionally well in historical tests but failing in new, unseen market scenarios.

In statistical terms, a model that is overfitted has low *bias* (error) but high *variance*. There is generally a tradeoff between bias and variance in a model. Bias measures how well a model captures the regularities in the *training data*, the data used to develop a model. Variance measures how well a model responds (generalizes) to new data. Overfitting occurs when the model fits the training data, possibly including noisy data, too well. While ideal models would have both low bias and low variance, this situation is generally not possible. Rather, the better one can fit a model's parameters to past data, the less the model will typically generalize to new data.

Another potential sample-related pitfall is called *inception point risk*. Many readers will be familiar with fund presentations showing the backtested compound returns of a given strategy over a given historical period. These types of exercises must by definition begin compounding at a specific point in time. While there is nothing inherently wrong with doing that, the use of a different inception point may materially impact the backtested performance and reveal that the strategy's success depends on the point at which investing begins. To ensure inception point robustness, it is therefore necessary to measure the compound returns that a strategy produces across multiple inception points. Inception point risk is a type of *selection bias* in which specific time periods or subsets of data that support a model's validity are chosen while neglecting data that would potentially undermine the model as an acceptable forecasting tool.

Other pitfalls include the various biases that can infect the choice of investment universe in designing an investment strategy. For example, if a researcher wishes to use stocks from a particular equity index for its universe of assets, it is important that the researcher use stocks that were in the index at the time that the backtest begins. Because stocks are often included in indexes because of their previous performance, including recent additions to an index would bias the backtest because it will contain "sure winners."

A related bias, known as *survivorship bias*, arises when using only stocks that have stood the test of time—that is, survived. Without consideration of an expansive universe of stocks, including those of companies that have failed, it is impossible to know whether a model is capable of discerning the fortunes of ultimately unsuccessful companies.

Another type of bias is called *look-ahead bias*, which arises when information that would not have been available during the time period being studied is used during the backtesting process. A blatant example of such a bias is if, when testing the efficacy of an optimization model for an equity portfolio, the minimum weight constraints for the historically most successful stocks in a manager's investment universe were set higher than for stocks with lower returns. While that example is a bit forced, there are subtler types of look-ahead bias that may creep into the model development and validation processes. The reason this happens is that every researcher has lived through and observed financial history and has presumably learned some lessons from it. Some of these lessons undoubtedly will make their way into model development despite researchers' best efforts to implement a bias-free process.

For example, after the Global Financial Crisis (GFC) of 2008–2009, the relationship between equity and credit risk was on every investment professional's "risk radar." It is inconceivable that any researcher or portfolio manager today would build a model of credit risk without accounting for the relationship between equity and credit, whereas before the GFC it was relatively common to do so. Thus, even without any kind of explicit data manipulation, it is possible to introduce biases into the model development process. Of course, all biases are not created equal, and some represent more egregious examples of cheating than others. What is most important from the standpoint of "statistical hygiene" is that model builders take appropriate care to eliminate the most significant biases in their backtests so that these tests can help them select viable candidates for further testing and development.

Backtesting is a critical process in investment management because it uses historical data to assess the performance of a strategy or a model. Its primary role is to simulate how a particular investment strategy would have performed in the past. When executed diligently, with the requisite level of statistical and economic detail, it can serve as a valuable tool for researchers and portfolio managers, aiding them in the development and refinement of models that have the potential to perform well in various market conditions.

That said, while backtesting is a valuable tool, it is not without its limitations. It is thus important to approach backtesting with a critical eye and to understand its constraints and potential biases. By acknowledging these pitfalls, investors can make more informed decisions and develop more robust investment models. In the chapters ahead, I present a number of methods that address the various inherent shortcomings in backtesting that stem from its reliance on historical data, because such data might not accurately represent current or future market conditions, rendering backtested models less effective in real time. First, however, I will discuss a model-validation approach that can be considered an extension of standard backtesting: cross-validation.

3. CROSS-VALIDATION

In the previous chapter, I emphasized the importance of out-of-sample backtesting. The relationship between out-of-sample backtesting and time-series cross-validation lies in their shared aim of evaluating a model's performance on unseen future data in the context of time-series analysis. Out-of-sample testing represents a straightforward split between model training and testing data in a time-ordered manner, whereas time-series cross-validation techniques are designed to systematically simulate this process, considering different subsets (known as *folds*) of historical data for training and testing to assess the model's performance over various time periods.

More generally, cross-validation, which originated in the 1930s (Larson 1931) and began to be developed in earnest in the 1960s (Mosteller and Tukey 1968; Lachenbruch and Mickey 1968), is a method used to assess how well a model generalizes to an independent dataset. It consists of the following basic elements:

- *Training set*: This subset of the complete dataset is used to train the model. The model learns the patterns and relationships within these data.
- *Validation set*: This separate subset of the data is not used during the training phase but is used to fine-tune the model's hyperparameters and assess its performance during training.
- *Test set*: The test set is another separate subset of the data that is not used in either the training or validation phases. It is reserved for the final evaluation of the model's performance.

Basic Cross-Validation Methodologies

The following list provides details on several basic methodologies for cross-validation.

- *Holdout method*: This is one of the simplest forms of cross-validation, where a portion of the dataset (often around 20%–30%) is withheld from the model during training and used for evaluation. While straightforward, this method might lead to high variance in the performance estimate due to the randomness in the selection of the training and test sets.
- *K-fold cross-validation*: In this method, the dataset is divided into k subsets (folds). The model is trained on $k - 1$ folds and validated on the remaining fold. This process is repeated k times, with each fold being used once as the validation set. The final performance metric is the average of these k iterations, providing a more stable estimate of the model's performance.
- *Leave-one-out cross-validation (LOOCV)*: This cross-validation framework takes the concept of k-fold cross-validation to the extreme by using each observation as a validation set and the remaining observations as the training set. This process is repeated for each data point, with a different observation being left out for testing in each iteration. While computationally expensive, LOOCV provides an unbiased estimate of model performance.

Time-Series Cross-Validation Methodologies

As is well known, time series contain unique properties that distinguish them from other types of structured data. Perhaps the most fundamental is that the values in a time series at each point possess a memory of past values. *Time-series cross-validation*[4] is intended to address the memory of the past that is embedded in data exhibiting chronological dependencies. It differs from standard types of cross-validation in its requirement that training data temporally precede validation data. The basic approaches to time-series cross-validation are as follows:

- *Rolling window cross-validation*: In this method, the training set size remains constant and the validation set moves forward in time. At each step, the model is trained on the current training set and validated on the subsequent time period.
- *Expanding window cross-validation*: In this method, the training set expands over time, including all data points up to the current validation point. The validation set also moves forward in time.
- *Time-series split cross-validation*: This method splits the data into multiple folds, each containing a training set and a validation set. The validation set always comes after the corresponding training set in time, ensuring that the model is evaluated on future data. For example, the data can be split into training sets from specific beginning and end dates and tested from that date forward on another specified date range.

Although time-series cross-validation solves the memory problem in chronologically dependent data, overfitting can still be a problem because validation and test samples are likely to have some memory of samples that immediately precede them in time. As a remedy, *block time-series cross-validation* has been introduced. This procedure is defined by the omission of some observations (blocks) that lie between training, validation, and test samples during cross-validation to reinforce the memoryless nature of the operation.

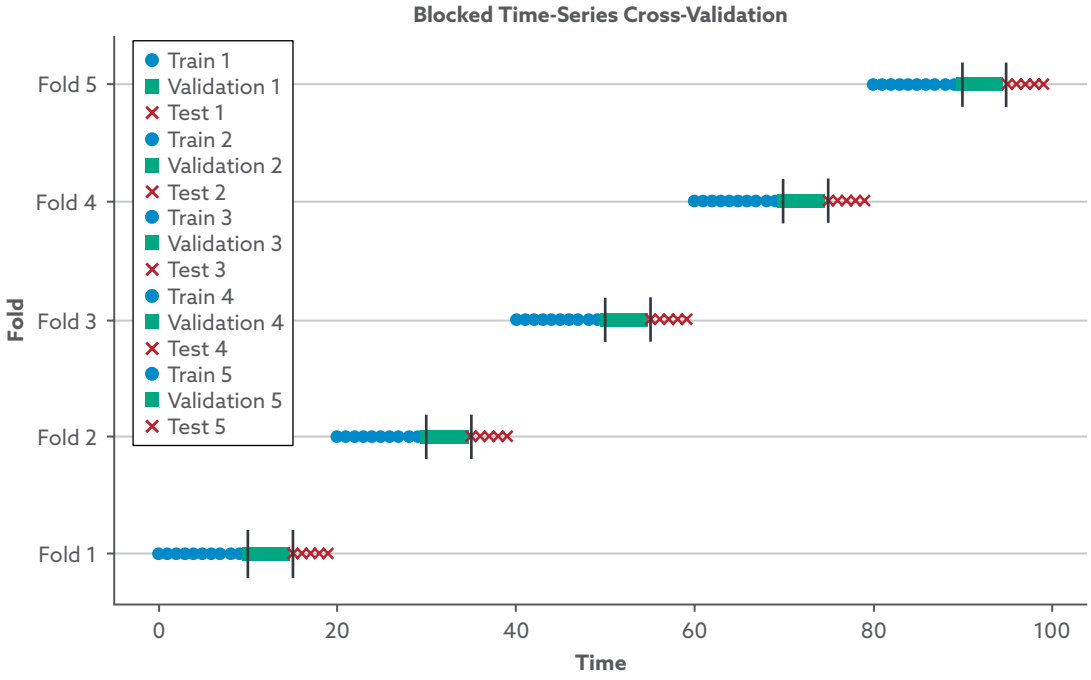
In setting up a time-series cross-validation, the question of the number and length of the folds used for testing, as well as the block size used between folds, naturally arises. Time-series cross-validation by itself, however, does not provide a clear-cut way to choose values for these parameters. The regime-switching models discussed in the previous section can provide some guidance in this respect because they specify the number and length of regimes and the order of the autoregressive process that drives regimes. The number and length of regimes can be used to guide the choice of folds, while the order value seems to be a natural choice for block size; it indicates the length of time beyond which the time series forgets.[5] **Figure 2** shows a stylized example of one way in which blocked time-series split cross-validation can be implemented. As the chart shows, the folds proceed one after another in time. The training, validation, and test samples—colored blue, green, and red, respectively—are divided by the blocked observations, which are represented by black vertical lines. There are also gaps between each fold.

---

[4]See Burman, Chow, and Nolan (1994); Racine (2000); Bergmeir and Benítez (2012); Roberts, Bahn, Ciuti, Boyce, Elith, Guillera-Arroita, Hauenstein, et al. (2017); and Bergmeir, Hyndman, and Koo (2018) for discussions on various aspects of time-series cross-validation.

[5]For an illustration of an application of regime-switching models in this manner, see Simonian (2020).

Figure 2. Example of Blocked Time-Series Cross-Validation



Finally, note that it is possible to extend cross-validation by using synthetic data. That is, it is possible to use a simulated time series to train, validate, and test a model-driven investment strategy. I will discuss methods for synthetic data generation later in the monograph.

4. PERFORMANCE MEASUREMENT AND BENCHMARKING

One of the important aspects of model validation is understanding and quantifying the performance and risks associated with a particular model. This pertains to the metrics that apply both to the model being validated individually and to those in relation to a benchmark. This chapter provides an overview of the major metrics and frameworks for performance and risk measurement, as well as the tenets of proper benchmarking. Performance measurement provides answers to three questions:

- What was the return on a strategy?
- Why has a strategy performed the way it has?
- How can the strategy's performance be improved?

To help answer these questions, researchers have a multitude of performance metrics at their disposal. Performance metrics can be classified into two broad categories: *return metrics* and *risk metrics*. Calculating a strategy's returns, in absolute or relative terms, is of course the fundamental way that the success or failure of an investment strategy is determined. Measures of risk can be viewed as ways of assessing the potential cost, in the form of uncertainty, that investors pay for a strategy's returns.

Return Metrics

During the model validation process, a variety of return metrics are used.[6] These typically include well-known return metrics, such as the *simple rate of return*[7] and *cumulative return*.[8] While these return measures are important, calculating the *time-weighted return*[9] is also important because it reveals the rate of growth (or shrinkage) that, if earned equally in every period, would match the return experience that actually occurred. In addition, there are different money-weighted return calculations, such as the *modified Dietz*[10] method, that are often used by liability-driven and other cash flow–sensitive investors.

---

[6]See Bacon (2004) for a comprehensive inventory of return calculation formulas and methodologies.

[7]The simple rate of return is calculated as $\frac{V_{t+1}}{V_t} - 1$, where V is the value of the portfolio under consideration and t is the time at which the portfolio's value is measured.

[8]Cumulative return is calculated as $(1 + r_1) \times (1 + r_2) \times (1 + r_3) \times \ldots \times (1 + r_{n-1}) \times (1 + r_n) = (1 + r)$, where r is the strategy return and n is the number of periods under consideration. This return measure is often depicted in "growth of a dollar" charts, where the growth or decline of an initial amount of capital is shown over time.

[9]The time-weighted or geometric average return is calculated as $r_G = \left(\prod_{i=1}^{n} (1 + r_i) \right)^{f/n} - 1$, where n is the number of periods under analysis and f is the number of periods within the year (e.g., 12 months).

[10]The modified Dietz return is calculated as $r = \frac{V_E - V_S - C}{V_S + \Sigma C_t \times W_t}$, where V_S signifies the starting value of the portfolio, V_E signifies the ending value of the portfolio, C represents the total external cash flow for the period under consideration, and C_t is the external cash flow received on day t. $W_t = \frac{TD - D_t}{TD}$, where TD is the total number of days contained within the time

Risk Measurement

The counterpart to investment return is of course, investment risk. In validating any investment model, numerous risk metrics must be considered to evaluate the effectiveness and robustness of the model. Investment risks must be properly identified and controlled to harvest the potential rewards that an investment offers. Various types of risk are important for investors, including operational, legal, and counterparty risk, but most investment model validation concerns the measurement and assessment of various forms of *market risk*: the risk due to the price fluctuations of continuously traded financial assets.

Risk metrics can measure the *absolute* level of risk in an asset or portfolio or the *relative* risk of an asset or portfolio in relation to some benchmark or reference value. Most assessment of asset and portfolio risk begins with variance or standard deviation, which are the primary measures of asset return variability. (Standard deviation is simply the square root of variance.) Variability is regarded as undesirable for investors because it reduces the certainty with which assets will produce positive payoffs. In a portfolio context, variance and standard deviation (the latter is also called *volatility*) are important because they play a central role in the most commonly used type of portfolio optimization, *mean–variance optimization*.[11] Variance, signified by σ^2, is the average of the squared deviations of returns from the mean return and is formally defined as

$$\sigma^2 = \frac{\sum_{i=1}^{i=n}(r_i - \bar{r})^2}{n}, \tag{4}$$

where n is the number of observations, r_i is the return in period i, and $\bar{r}$ is the mean return. As just mentioned, standard deviation is simply the square root of variance:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{i=n}(r_i - \bar{r})^2}{n}}. \tag{5}$$

To annualize standard deviation, σ_A, as is often desired in investment analysis, simply multiply the standard deviation at a given observation frequency (e.g., monthly) by the square root of the number of observations in a year (e.g., 12); that is, $\sigma^A = \sigma \times \sqrt{t}$. Portfolio variance is derived by consideration of the variances of the individual assets in a portfolio along with a covariance matrix that describes the directional co-movements of the portfolio's assets (covariance is described in detail later). Formally, this is expressed by $w'\Sigma w$, where w signifies a vector of the portfolio's asset weights and Σ the asset covariance matrix. It considers how a portfolio's assets are weighted and how similar or dissimilar their return behavior is. Portfolio volatility is simply the square root of portfolio variance.

Volatility is important not only as a standalone risk measure but also as an input into the most commonly used measure of portfolio efficiency, the Sharpe ratio (see Sharpe 1966):

$$\text{Sharpe ratio} = \frac{r_P - r_f}{\sigma_P}, \tag{6}$$

horizon being considered and D_t represents the total number of calendar days since the start of the period when the cash flow was received.

[11]Mean–variance optimization was introduced by Markowitz (1952).

where r_p and r_f are the portfolio (or asset) return and risk-free return, respectively, and σ_P is the portfolio (or asset) volatility. The Sharpe ratio is a measure of return per unit of (absolute) risk. Given two investments, the higher Sharpe ratio investment would be considered the more efficient investment.

Mean and variance, while important, are not the only measures that contribute to an understanding of investment risk. Indeed, when examining the return distribution of an investment, investors are also typically concerned with *skewness* and *kurtosis*. Skewness measures the asymmetry in a distribution and is expressed mathematically as

$$\text{Skewness} = \sum \left(\frac{r_i - \bar{r}}{\sigma_P} \right)^3 \times \frac{1}{n}, \tag{7}$$

where the terms are defined as before. A perfectly symmetrical distribution will have a skewness of zero. Such is the case with the normal, or Gaussian, distribution, which is commonly assumed to describe asset returns in investment analysis. As known, however, asset returns are rarely described accurately by the normal distribution and often exhibit significant positive or negative skewness. A positively skewed investment strategy will be expected to experience frequent small losses and a few large gains. The trend-following strategies often used by commodity trading advisers (CTAs) are examples of strategies that have been observed to exhibit positive skewness. In contrast, a negatively skewed investment strategy will be expected to experience frequent small gains and a few large losses. Equity market neutral strategies often exhibit a moderate amount of negative skewness, whereas catastrophe bonds, which pay off only in the event of a rare extreme event (e.g., an earthquake), are examples of investments with a high degree of negative skewness.

Along with the asymmetry of a distribution, investors are also generally concerned with the kurtosis of an investment's return distribution. Kurtosis measures the "tailedness" of a distribution—that is, the incidence of observations that are significantly distant from the mean return. The formula for kurtosis is

$$\text{Kurtosis} = \sum \left(\frac{r_i - \bar{r}}{\sigma_P} \right)^4 \times \frac{1}{n}. \tag{8}$$

The "fatness" of a return distribution's tails, especially the left tail, is naturally important for assessing the risk of any investment. The kurtosis of a normal distribution is 3. Distributions with kurtosis values greater than 3 (*leptokurtic* distributions) exhibit fatter tails and thus contain more extreme observations.

Equity returns have been observed to exhibit behavior that departs from that assumed by a normal distribution. However, many investment models still assume a normal distribution, mainly due to its mathematical tractability. Nevertheless, they do so to the potential detriment of investment performance. Consider, for example, a study conducted by Johansen and Sornette (1998) in which they found that in one million years of simulated trading, the generalized autoregressive conditional heteroscedasticity, or GARCH (1, 1), model,[12] a well-known volatility forecasting model that assumes that asset returns are normally distributed,

[12]The GARCH model was introduced by Bollerslev (1986).

fails to predict drops in the Dow Jones Industrial Average that approach the magnitude of the three largest actually observed drops in the index during the 20th century. Although the GARCH (1, 1) model can be (and has been) improved on, the noted predictive failure of the latter model highlights how difficult forecasting can be, even with a relatively sophisticated quantitative tool. Given the unexpected frequency of large market drops, the importance of understanding and properly measuring kurtosis to both risk management and model development cannot be overstated.

Rounding out the fundamental metrics of performance measurement are covariance and correlation, measures that describe the linear co-movement of assets with one another. As explained previously, covariances play an important role in mean–variance optimization, namely in determining a portfolio's variance. All things being equal, the lower a portfolio's asset covariances, the lower a portfolio's variance—hence, the emphasis on diversification in contemporary investment theory. The covariance between two assets *a* and *b* is derived in the following manner:

$$\text{Covariance} = \frac{\sum_{i=1}^{i=n}(a_i - \bar{a}) \times (b_i - \bar{b})}{n}, \tag{9}$$

where a_i is the return of asset *a* in period *i*, $\bar{a}$ is the mean return of asset *a* at the specified frequency, b_i is the return of asset *b* in period *i*, and $\bar{b}$ is the mean return for asset *b* at the specified frequency. Correlation, in turn, is derived from covariance as follows:

$$\text{Correlation} = \frac{\text{Covariance}}{\sigma_a \times \sigma_b}, \tag{10}$$

where σ_a and σ_b are the respective volatilities for assets *a* and *b*.

Related to correlation is *autocorrelation*, which measures the degree to which a time series is correlated with its own past values. Among the most important practical risks that autocorrelation poses to model development and validation is that autocorrelation can lead to faulty hypothesis testing. Standard errors may be underestimated if autocorrelation is not appropriately addressed, leading to an increased likelihood of Type I errors (false positives) or Type II errors (false negatives). Autocorrelation can also affect the accuracy of model forecasts. Specifically, autocorrelation can lead investors to underestimate or overestimate the forecasted values for model variables, particularly if the model does not consider the influence of past observations. Finally, autocorrelation may create spurious relationships in the data. This situation can lead to the identification of false patterns in the data, which may not accurately reflect the underlying processes being modeled.

Additional risk measures, both absolute and relative, are often used during the model validation process. Some of these metrics are focused on downside risk. For example, *semi-standard deviation* measures the volatility of an investment's downside variability, considering only the negative returns below a specified threshold, such as the risk-free rate or the benchmark return, among others. Downside risk is defined mathematically as

$$\text{Downside risk} = \sqrt{\sum_{i=1}^{n} \frac{\min\left[(r_i - r_T), 0\right]^2}{n}}, \tag{11}$$

where r_T is the threshold return (which could be a minimum target or benchmark or risk-free return). With the downside risk measure, the downside analog of the Sharpe ratio, the *Sortino ratio*, can be derived:

$$\text{Sortino ratio} = \frac{(r_P - r_T)}{\sigma_D}. \tag{12}$$

Another popular risk measure is *value at risk* (VaR), a metric used to estimate the potential loss, at a given confidence level, that an investment or portfolio might face over a specified time horizon. For instance, a 5% one-day VaR of $1 million means that there is a 5% chance that the portfolio may lose more than $1 million over a single day. The formula for VaR is

$$\text{VaR} = P \times \text{Volatility} \times Z\text{-score},$$

where

P is the portfolio value of the investment being analyzed

Volatility is the standard deviation of the portfolio's returns over the given time horizon

Z-score is the number of standard deviations corresponding to the chosen confidence level (for instance, for a 95% confidence level, the *Z*-score is 1.645, assuming a normal distribution)

Aside from the basic parametric VaR framework, a *historical VaR* can be implemented. This method of calculating VaR relies on historical data to estimate potential future losses, without any assumptions about the underlying distribution. To implement historical VaR, one simply chooses a historical time frame over which to calculate VaR, orders historical returns or price changes from the selected period in ascending order, and estimates VaR by selecting the return that corresponds to the desired confidence level and time horizon. For example, with weekly return data for the past year, one could calculate a one-week 95% VaR and would find the return value at the 5th percentile of the ordered historical one-week returns. This value represents the potential loss that would be exceeded only 5% of the time based on historical data.

While historical VaR is convenient, its use does carry some risk because it assumes that historical patterns will repeat. It thus does not account for potential changes in market conditions or for tail risks that are not present in the historical data. As a result, it is often used to complement other VaR methodologies in a more holistic assessment of investment risk. Beyond the foregoing two approaches to the calculation of VaR, it is also possible to use simulation techniques, such as the Monte Carlo method. Simulation methods will be discussed in Chapter 5.

Aside from absolute measures of risk, several *relative risk measures* typically play an important role in model validation. Among the most important is what is known as the *information ratio* (IR), which is the relative return analog of the Sharpe ratio and is often used by portfolio managers to demonstrate the value of their active management process. The formula for the IR is

$$\text{IR} = \frac{\overline{ER}}{\hat{\sigma}_{ER}}, \tag{13}$$

where $\overline{ER}$ is the arithmetic average of excess returns over a given time horizon and $\hat{\sigma}_{ER}$ is the standard deviation of excess returns from the benchmark, otherwise known as *tracking error*.[13]

An additional important relative risk metric is the *Treynor ratio*,[14] another analog to the Sharpe ratio, which measures the excess return on an investment in relation to its systematic risk. It measures the return earned in excess of that which could have been earned on an investment that has no diversifiable risk. It is derived as follows:

$$\text{Treynor ratio} = \frac{r_P - r_f}{\beta_P}, \tag{14}$$

where β_P is the portfolio's beta (linear sensitivity) to a benchmark or market proxy. Related to the Treynor ratio is what is known as *Fama decomposition* (see Fama 1972), which tries to provide more insight into the relationship between excess returns and systematic risk. The Fama decomposition formula has the following components:

$$\underbrace{r_P - r_f}_{\text{Excess return}} = \underbrace{r_P - r_f - \beta_P(r_b - r_f)}_{\text{Selection return}} + \underbrace{\beta_P(r_b - r_f)}_{\text{Systematic risk}}. \tag{15}$$

As discussed previously, performance and risk measurement are usually conducted against a reference portfolio or other performance standard, otherwise known as a *benchmark*. We turn to the topic of benchmarking next, with a special emphasis on its role in the model validation process.

Benchmarking

The discussion of relative risk measures in the previous section attests to the fact that models are not evaluated in isolation: Investment managers typically measure their performance against some type of benchmark. In model validation, a benchmark is used in a manner that is complementary to its use in performance measurement. Thus, benchmarking a model's performance provides context for evaluating whether a model's performance satisfies whatever criteria a portfolio management team deems important. While portfolio managers with live strategies use benchmarks to understand how their investment decisions impacted their performance *ex post*, benchmarks can also be used to assess the robustness of a strategy *ex ante* before it is made available commercially. If a proposed investment strategy demonstrates that it can beat an appropriate benchmark in both backtests and simulations, then that may provide convincing evidence that it is likely to outperform when the strategy goes "live" as a commercially available investment product.

Several characteristics are considered important for any valid benchmark.[15] In the context of model validation, the most important characteristics of benchmarks include the following:

- *Investable*: Portfolio managers must be able to hold the assets of a benchmark and have the ability to hold and trade these assets without limitations.
- *Measurable*: The benchmark returns are readily calculable on a reasonably frequent basis.

---

[13]Tracking error can be calculated using $\sigma_P \times \sqrt{(1-\rho_{P,B})^2}$, where $\rho_{P,B}$ is the correlation between the portfolio and the benchmark.

[14]See Treynor (1965). Manager alpha can also be used in the numerator.

[15]See Sharpe (1992), Siegel (2003), and Lo (2016) for further discussion of the roles and characteristics of investment benchmarks.

- *Appropriate*: The benchmark is consistent with the manager's investment style or area of expertise; it is the "home portfolio" to which the manager would return if he or she had no views on any of the securities in the benchmark.
- *Unambiguous and known*: The identities and the weights of the securities or factor exposures that comprise the benchmark are clearly defined and identified.
- *Accessible*: Portfolio managers and researchers need to have access to all relevant statistics pertaining to a benchmark, as well as benchmark constituents and weights.

While the foregoing criteria are concisely stated and can be easily satisfied for some asset classes, such as liquid public equities, for other asset classes, benchmark selection is not always a simple matter. For example, in the case of fixed income, it is often the case that benchmarks are very complex. The multifaceted risk exposures in fixed-income instruments (e.g., interest rate, credit, prepayment) make bond index construction a challenging task. The mechanics of some bonds (e.g., those with embedded options) also add to the complexity of benchmark construction and selection.

Also, bond benchmarks need to be constantly refreshed with new bonds as the old ones mature (or age out of the maturity range of the benchmark). New issues of bonds in the market sector represented by the benchmark must be included in the benchmark. Including a bond in a benchmark index has an immediate impact on the liquidity and turnover of that bond. These are all important considerations when using or considering a given bond benchmark as a standard for performance measurement.[16] Whatever the asset class, in selecting a benchmark for model validation, it is important to consider any challenges to building or selecting a stable and understandable standard by which a model's viability can be evaluated.

Sometimes, more than one benchmark will be used during model validation. This situation contrasts with the use of benchmarks for performance measurement and reporting, where a single benchmark is typically used by portfolio managers, as well as their clients and employers, to assess the effectiveness of a strategy. In model validation, it is certainly the case that the benchmark that will eventually be used for performance measurement and reporting will also be used for model validation. This single benchmark, however, is not the only possible benchmark that could be used to assess the effectiveness of the investment manager's model or the signals being used to make investment decisions.

For example, one useful benchmark that can be used alongside any "official" benchmark is an equal-weighted portfolio of securities in the manager's investable universe. This benchmark is useful because it represents the "no information" portfolio—the portfolio one would presumably hold in the absence of either investment views on or knowledge of the market caps of any of the securities in the portfolio under consideration. The equal-weighted portfolio therefore represents a valid way to measure the informativeness of the signals being produced by an investment model. In addition, an equal-weighted portfolio is also often very difficult to beat and thus represents a high standard with which to evaluate a model's effectiveness.

With some investment models, of course, it is more appropriate to use so-called absolute return benchmarks, which are not portfolios of assets but a reference return of some type,

[16]For a detailed discussion of bond benchmarks, see Konstantinov, Fabozzi, and Simonian (2023).

such as the 30-day SOFR[17] rate + 5% or CPI[18] + 3%. Even when an investment strategy has an investable benchmark as described, however, it is often useful to test its ability to outperform a fixed reference rate, such as 0%.

Finally, one can consider an "informational benchmark." In developing a given model, one may observe that it indeed performs well along a number of metrics but nevertheless investigate whether it outperforms a much simpler model. If it does not, then that would be an indication that the model under consideration does not provide as much information as believed.

Luckily, there is a convenient way to test the informative power of a model by using the concept of *Granger causality* (see Granger 1969), which states that for a given predictor x and target variable y, x causes y if

(1) x temporally precedes y and

(2) x provides more predictively useful information than a rival "naive" predictor.

One way of formally expressing Condition 2 of Granger causality is $P(Y_{t+1} \mid X_t \cdot Y_t) > P(Y_{t+1} \mid Y_t)$, which can be interpreted as saying that given a target variable Y, the probability of a model predicting the one-step-ahead value Y_{t+1} is higher given prior information about variable X in addition to prior information about Y alone. Granger causality may also be articulated in a more general form:

C causes E if there is information transmission from C to E.

For active managers, one of the benefits of using benchmarks is that they play a prominent role in the most widely accepted return attribution frameworks. These frameworks can be applied to such studies as backtests to determine the sources of a model's returns. The goal of any active investment strategy is to generate excess return, also known as *alpha*, over a specific benchmark. Using the popular attribution framework set forth by Brinson and his coauthors in the 1980s, there are four sources of a portfolio's return:[19]

- Benchmark return: $W_i \times b_i$
- Asset allocation alpha: $(w_i - W_i) \times b_i$
- Security selection alpha: $W_i \times (r_i - b_i)$
- Interaction effects: $(w_i - W_i) \times (r_i \cdot b_i)$,

where

W_i is the weight of the benchmark in the ith asset class

b_i is the return of the benchmark in the ith asset class

r_i is the return of the portfolio assets in the ith asset class

w_i is the weight of the portfolio in the ith asset class

---

[17]SOFR stands for the Secured Overnight Financing Rate, a benchmark interest rate for dollar-denominated derivatives and loans. It is based on transactions in the Treasury repurchase market. SOFR replaced LIBOR (the London Interbank Offered Rate), which was based on estimated future borrowing rates rather than observable market data.

[18]CPI stands for the Consumer Price Index, a precise definition of which can be found at https://fred.stlouisfed.org/series/CPIAUCSL.

[19]Adapted from Brinson and Fachler (1985) and Brinson, Hood, and Beebower (1986).

This framework can be used to analyze any kind of portfolio with a benchmark conforming to the criteria cited earlier. Of course, additional, more granular types of analysis can be conducted alongside the basic attribution described by the previous formulas. For example, it is possible to analyze how much alpha has been generated by decisions pertaining to currency.[20]

Regarding the term *alpha*, note that while a very general definition was used earlier, it is important to remember that within the context of a regression or a traditional factor model of the form $Y = \beta_0 + \beta_n X_n + \varepsilon$, alpha is the intercept term β_0, often interpreted as the baseline value of the dependent variable (Y) when the independent variables (X_n) are zero.[21] I will provide a more detailed discussion of factor models in Chapter 7.

[20]See Karnosky and Singer (1994) for a discussion of currency decisions as they pertain to return attribution.

[21]ε is the error term, representing unobserved factors or random noise.

5. SIMULATING ALTERNATIVE HISTORIES WITH SYNTHETIC DATASETS

In a well-known episode[22] of the original *Star Trek* television show, the crew of the Enterprise encounters an Earth-like planet where the Roman empire is still in existence, albeit with 20th century technology. The episode thus presents an alternative history, one that resembles the Earth's actual history in many ways but differs in others. These differences create challenges that the Enterprise crew must confront and adapt to. Simulation techniques likewise aim to create alternative economic and market histories so that the robustness and adaptability of investment models can be tested.

As emphasized in this monograph, the lack of sufficient data is a challenge to any kind of investment activity, including model validation. Consequently, the construction of synthetic time series should be a central concern when assessing the robustness of investment models. After all, it is highly desirable—mandatory, in fact, if performance in future "live action" is to be good—for our models to be able to perform well in market scenarios that we have not yet experienced. To that end, in this chapter, I provide a detailed overview of various simulation techniques that can be used to supplement historical data during the model validation process.

Monte Carlo Simulation

Monte Carlo simulation is perhaps the most familiar simulation technique to investment professionals.[23] Monte Carlo methods encompass various related computational approaches to modeling the probability of different outcomes in systems that contain random elements. It is named after the Monte Carlo Casino in Monaco, known for its games of chance and randomness. In a basic *parametric* Monte Carlo simulation, the underlying model or system being simulated is described by a precise set of parameters assumed to follow a known distribution. Monte Carlo simulations rely on random sampling, which, in turn, requires generating random numbers that follow specific probability distributions (e.g., uniform, normal, exponential) using pseudo-random number generators. For example, when using Monte Carlo simulation to test an investment strategy, asset returns might be assumed to follow a normal distribution. In that case, a standard way of setting up the model would be to assume a mean return and volatility for each asset used in the strategy being evaluated, as well as a correlation matrix for the assets. Then, given these inputs, a sufficiently high number of simulations would be run (e.g., 10,000) so that the final distribution of the entire set of simulations has the shape of the stipulated distribution. So, for a normal distribution, the final shape of the distribution should resemble a bell curve.[24]

---

[22]"Bread and Circuses," Season 2, Episode 25, 1968.

[23]For an in-depth treatment of Monte Carlo methods in finance, see Glasserman (2003).

[24]In terms of statistical theory, the effectiveness of Monte Carlo simulations relies on the law of large numbers, which states that as the number of random samples increases, the sample mean converges to the true mean. In simulations, as the number of iterations increases, the simulated results tend to converge toward the expected value or the true solution.

It is also possible to implement *nonparametric* Monte Carlo simulations in which the distribution of the variables might not be explicitly stipulated. Instead, historical data or empirical distributions might be used without the assumption that the distribution takes a specific form. The choice between parametric and nonparametric Monte Carlo simulations often depends on the availability of data, investors' confidence in their capital market assumptions, and the level of detail or complexity required in the simulation.

Worked Example

Consider a portfolio with three assets, with an initial amount of capital allocated to each asset. Next, posit an investment horizon over which to test the strategy. Once input variable values are acquired, set up a process that describes the assets' behavior. For example, changes in asset values over time can be represented by a stochastic differential equation, $dV_t = V_t \times \left(\sum_{i=1}^{3} \mu_i dt + \sum_{i=1}^{3} \sigma_i dZ_i \right)$, where $V_t = [V_{t,1}, V_{t,2}, V_{t,3}]$ represents the portfolio values for three assets, σ_i is the volatility for asset i, and dZ_i is a random increment for asset i from a multivariate normal distribution with mean 0 and asset covariance matrix Σ.

Given initial values $[V_{0,1}, V_{0,2}, V_{0,3}]$, the simulated portfolio values at time T for N simulations are obtained by iterating

$$V_T^{(j)} = V_0 \times \exp\left(\left(\mu - \frac{1}{2}\sigma^2 \right) T + \sigma \times \sqrt{T} \times Z^{(j)} \right),$$

where

$V_T^{(j)}$ represents the portfolio values at time T for the jth simulation

$\boldsymbol{\mu} = [[\mu_{1L}, \mu_{1U}], [\mu_{2L}, \mu_{2U}], [\mu_{3L}, \mu_{3U}]]$ is the vector of mean return ranges with upper and lower bound values for each asset return (while it is possible to use point estimates, using ranges allows for more definition regarding the shape of the distribution; ranges can also be used for volatilities and correlations, although I have not done so here for the sake of expository simplicity).

$\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \sigma_3]$ is the vector of volatilities

$Z^{(j)}$ is a vector of random increments from a multivariate normal distribution with mean 0 and covariance matrix Σ (the simulated values of the portfolio for each asset at time T from the simulation are represented as $V_T^{(1)}, V_T^{(2)}, V_T^{(N)}$, where N is the number of simulations).

To look at some concrete input values, consider the following:

$\boldsymbol{\mu} = [[2\%, 7\%], [4\%, 6\%], [-2\%, 10\%]]$ is the vector of asset mean returns ranges.

s = [15%, 12%, 20%] is the vector of asset volatilities

Then stipulate a correlation matrix for the assets

$$\text{Corr} = \begin{matrix} 1 & -0.4 & 0.3 \\ -0.4 & 1 & 0.2 \\ 0.3 & 0.2 & 1 \end{matrix}.$$

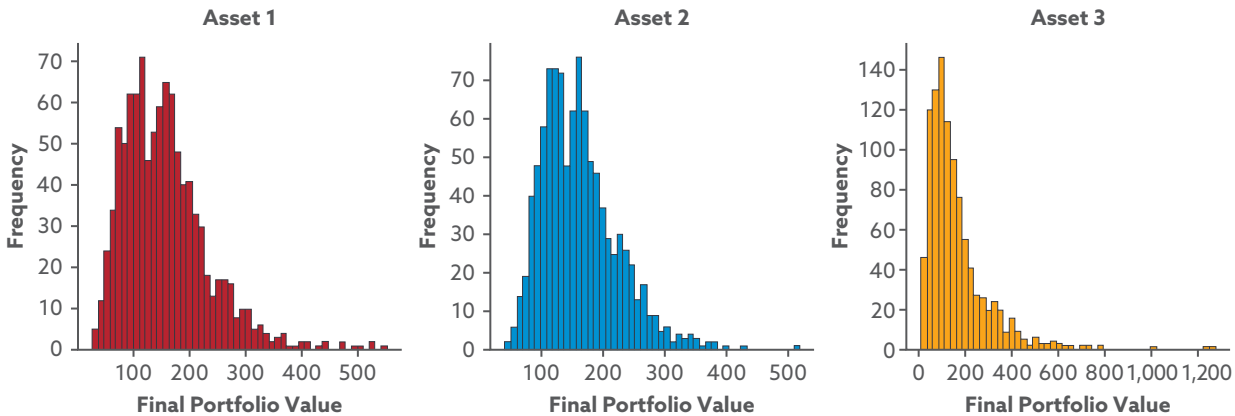
The simulation consists of 1,000 runs, or trials, of returns over a 10-year investment horizon for each asset. Assume that the starting value for each asset is 100. **Figure 3** shows the results of this simulation.

As shown in the three panels, each asset has a distribution with a different shape. Asset 3 in particular exhibits a long right tail; however, this potential upside is tempered by the fact that in several simulations, it has an ending asset value at or close to zero. In contrast, Asset 2 has the most range-bound outcomes, and its minimum values outcomes are well above zero. Asset 1 occupies somewhat of an intermediate position between Assets 2 and 3 but is closer to Asset 2 in terms of the shape of its distribution. Like Asset 2, Asset 1 has a considerably shorter right tail compared with Asset 3. Asset 1's left tail also resembles Asset 2's more than it does Asset 3's. Thus, from a statistical standpoint, Asset 1 and Asset 2 are somewhat similar. Nevertheless, even without knowing any other characteristics about the assets, it may still make sense to include both assets in a portfolio given that they are negatively correlated ($\rho = -0.4$) and hence should be diversifiers against one another.

Once the time series have been generated, they can be tested in a model validation context. For example, with an optimization model, one could dynamically generate portfolio weights through time and observe the range of outcomes for the optimized portfolios. With a trading model, one could see, through the set of simulated histories, how profitable the model-generated buy and sell signals are.

While Monte Carlo simulations can be helpful in providing high-quality simulated data, the technique does have its limitations. Perhaps most critical is that in the standard approach to Monte Carlo simulation, one is forced to simulate data based on a specified process or distribution. This is the weakness of the Monte Carlo technique: Although it offers some insight into market processes, important aspects of observed asset dynamics are, as in the example, often left out or insufficiently specified. Examples include such phenomena as momentum and mean reversion (i.e., positive and negative autocorrelation). Incomplete model specification is generally due to researchers' own limitations: It is impossible for anyone to have complete insight into every aspect of the drivers of asset behavior. For this reason, over time, researchers have gravitated toward more empirically based approaches to synthetic data creation. *Bootstrapping*,

Figure 3. Example of Monte Carlo Simulation Output



discussed next, is one of the more prominent data generation techniques that is grounded in observed data. It provides some remedy to the limitations encountered in standard Monte Carlo simulations.

Bootstrapping

Bootstrapping was introduced by Efron (1979) and could be considered a type of "historical combinatorics" because it literally is a method for recombining pieces of actual history in different ways to form new histories. More formally, it is a resampling technique used to estimate the sampling distribution of a statistic by generating multiple samples from the observed data.

Standard Bootstrapping (Nonparametric Bootstrapping)

The basic approach to bootstrapping involves randomly sampling, with replacement, from an original dataset to create multiple bootstrap samples. In sampling with replacement, each element is selected from the population, and after selection, it is put back into the population before the next draw. In the context of financial applications, an example of sampling with replacement would be to allow for the possibility of selecting the same historical return more than once in a single simulation. This action could reflect a belief that the market events observed in the past could occur again in the future.

Each bootstrap sample is the same size as the original dataset, and the statistic of interest (such as the sample mean) is computed for each resampled dataset. The distribution of these statistics across the bootstrap samples provides an approximation of the sampling distribution of the statistic. An important feature of bootstrapping from the standpoint of model validation is that it can be used to perform hypothesis tests by resampling under the null hypothesis. This process involves generating a null distribution of the test statistic by resampling from the entire original dataset under the assumption that the null hypothesis is true. The *p*-value can then be estimated by comparing the observed test statistic with the null distribution. Formally, the basic bootstrapping procedure can be described as follows:

(1) Given a dataset $X = \{x_1, x_2, \ldots, x_n\}$ of size n, generate B bootstrap samples $X_1^*, X_2^*, \ldots, X_B^*$ by sampling with replacement from X. For each bootstrap sample X_B^*, compute statistic θ.

As can be seen in (1), the standard bootstrap approach is nonparametric. However, it is possible to implement parametric versions of the procedure as well:

(2) Given a dataset $X = \{x_1, x_2, \ldots, x_n\}$ of size n, fit a parametric model $f(\theta|\hat{\theta})$ to the observed data. The term $\hat{\theta}$ represents the estimated parameters. Generate bootstrap samples $X_1^*, X_2^*, \ldots, X_B^*$ from the fitted model, where the parameters $\hat{\theta}$ are used to generate simulated data from $f(\theta|\hat{\theta})$.

To estimate a $(1 - \alpha)$ confidence interval for a statistic θ, the $\alpha/2$ and $1 - \alpha/2$ percentiles of the bootstrap distribution are taken as the lower and upper bounds of the confidence interval. For example, to find a 95% confidence interval, the 2.5th and 97.5th percentiles of the bootstrap distribution are taken as the lower and upper bounds, respectively.

To improve the accuracy of confidence intervals generated by standard bootstrapping methods, one can use *bias-corrected and accelerated bootstrapping* (BCAB). This approach adjusts for potential bias and skewness in the bootstrap distribution, providing more accurate confidence

intervals, especially for smaller sample sizes or skewed distributions. It uses three components: bias $\hat{z}$, acceleration $\hat{a}$, and correction factor $\hat{c}$. The bias-corrected confidence interval is calculated as $\left(\theta^* - \frac{\hat{z}}{\hat{a}}, \theta^* - \frac{\hat{z}}{\hat{c}}\right)$, where θ^* is the observed statistic and $\hat{z}$, $\hat{a}$, and $\hat{c}$ are estimated from the bootstrap distribution.

Block Bootstrap

Over the years, several extensions to the basic bootstrapping framework have been developed. Perhaps the most prominent is *block bootstrapping*. Time-series data often exhibit strong chronological dependencies, especially for financial time series, where such phenomena as momentum and mean reversion are commonly observed. Thus, a standard bootstrap that draws observations one by one will be unsuitable for most investment applications. Block bootstraps are designed to remedy this gap in the standard bootstrapping procedure. As implied by its name, block bootstrapping works by selecting blocks of temporally connected data rather than individual observations. Given time-series data X_t, where $t = 1, 2, \ldots, T$ data are divided into B blocks $X_1^*, X_2^*, \ldots, X_B^*$ of size m with $m < T$, blocks are sampled with replacement, and the resampled series is formed by concatenating the selected blocks, preserving the dependence structure within each block. Block bootstraps can be run in a number of ways. The most common approach is to select a single block size—for example, six months—that is used throughout the simulation. Other approaches, however, vary the block size in different ways. For example, the *stationary bootstrap* approach described in Politis and Romano (1994) uses random exponentially distributed block sizes, with the stipulation that the average block size match some predetermined number (e.g., three months).

Finally, it is possible to use bootstrapping to create more genuinely synthetic yet still empirically grounded data. We do this by means of what is known as a *wild bootstrap* (Wu 1986; Mammen 1993). The wild bootstrap method is specifically designed to account for the heteroscedasticity of the residuals found in a dataset. Instead of directly resampling observed residuals, however, it involves resampling the residuals with a random scale factor applied to each residual. This scale factor can be generated from a distribution, thus preserving certain properties of the data's structure. For example, assume that the observed data points are $y_1, y_2, \ldots, y_n$, and then proceed to fit the data to a model, such as a regression of the form $Y_i = \beta_0 + \beta_1 X_1 + \epsilon_i$. Compute the residuals $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ by subtracting the model predictions from the observed data and then apply a scaling factor $\gamma_1, \gamma_2, \ldots, \gamma_n$ to the residuals. Then, follow these steps:

(1) Generate random scale factors from a distribution centered at 1: $\gamma_1^*, \gamma_2^*, \ldots, \gamma_n^*$.

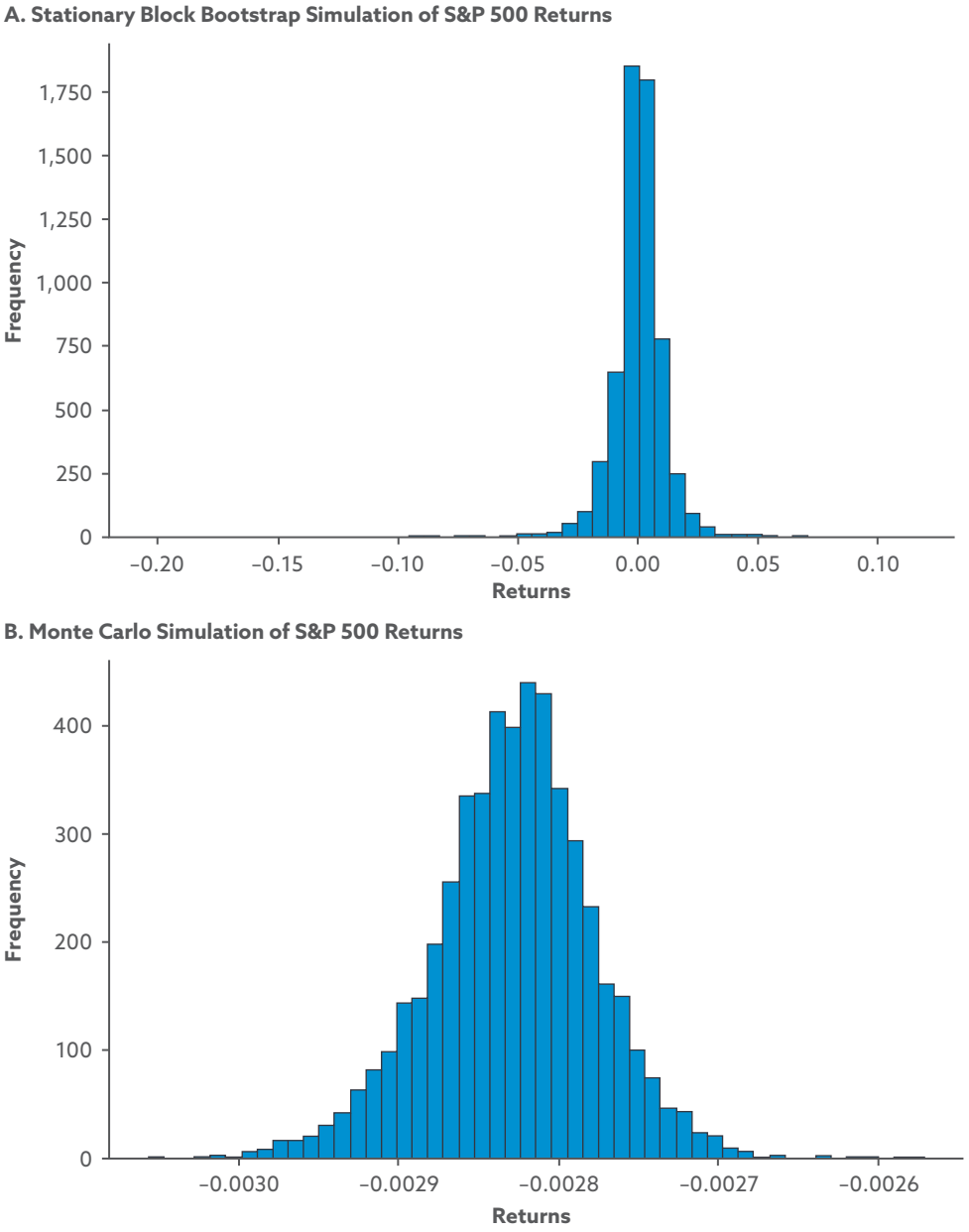
(2) Generate new "wild" residuals by multiplying the original residuals by the scale factors: $\epsilon_1^*, \epsilon_2^*, \ldots, \epsilon_n^* = \gamma_1^* \cdot \epsilon_1^*, \gamma_2^* \cdot \epsilon_2^*, \ldots, \gamma_n^* \cdot \epsilon_n^*$.

(3) Reconstruct the resampled dataset using the modified residuals: $\gamma_1^*, \gamma_2^*, \ldots, \gamma_n^* = \text{Predicted values} + \epsilon_1^*, \text{Predicted values} + \epsilon_2^*, \text{Predicted values} + \epsilon_n^*$.

(4) Compute the statistic of interest $\hat{\theta}$ (e.g., mean) from the resampled datasets.

As alluded to earlier, bootstrap simulations often provide more empirically realistic distributions compared with Monte Carlo simulations. I demonstrate this by means of an example. Panel A of **Figure 4** shows a bootstrap simulation of the S&P 500 using daily data from 1 January 1964 to 1 January 2024. We can see that the distribution is quite unlike a normal distribution.

Figure 4. Bootstrapping vs. Monte Carlo Simulation



It is considerably peaked and most notably contains long tails, especially on the left-hand side. Indeed, consistent with history, several days appear to have a daily return approaching –10%. Panel B shows a distribution that is generated via Monte Carlo simulation, assuming a normal distribution and using the S&P 500's mean return and variance over the same date range used to run the bootstrap simulation.

It is no surprise that the Monte Carlo distribution, shown in Panel B, resembles a normal distribution to a far greater extent, including far shorter and thinner tails, than the bootstrap simulation shown in Panel A. Thus, in its basic form, bootstrap simulation produces significantly more realistic synthetic datasets when compared with Monte Carlo simulations. Although the Monte Carlo simulation possesses the advantage of being able to generate completely novel datasets, it generally takes a considerable amount of work to achieve a level of data realism suitable for rigorous model validation. Moreover, as discussed here, such techniques as wild bootstrapping may be invoked in order to introduce data novelties into synthetically generated time series.

Generative Adversarial Networks

Another possible approach to generating synthetic time series for use in model validation is to use generative adversarial networks (GANs; Goodfellow Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville, and Bengio 2014). GANs are a type of machine learning architecture that consists of two neural networks, the *Generator* and the *Discriminator*, which are engaged in a minimax game. The Generator network, on the one hand, takes random noise as input and transforms it into data, such as images or time series. At the beginning of the game, its output is random and does not resemble the actual data of interest. Over multiple rounds (called *epochs*), the goal of the Generator is to learn how to produce data that are indistinguishable from real data. The Discriminator network, on the other hand, takes in both real data and data created by the Generator. It tries to correctly classify whether the given data are real or fake. In the initial epochs, the synthetic data created by the Generator are quite different from the data it is trying to reproduce, while the data generated in later epochs are much closer to the original data. Thus, when the algorithm terminates, what remains is a set of synthetic data, with some elements of the set more closely resembling the original data compared to other elements in the set.

For example, in the case of time series, in each epoch, a new time series will be created. When the algorithm terminates, the set of time series created will range from those that strongly parallel the original time series to those that scarcely resemble it. The different time series produced thus allow investors to determine the robustness of their strategies by testing them using a variety of alternate market histories.

The key idea behind GANs is the adversarial process. The Generator and Discriminator are in a competition. For example, imagine a case where a counterfeiter (Generator) wants to create fake artwork that looks like real masterpieces. On the other side is a detective (Discriminator) whose job is to tell whether a piece of art is genuine or fake. The Generator wants to create data that are so realistic that the Discriminator cannot distinguish the data from real data. The Discriminator, in turn, wants to become better at distinguishing real data from fake data. The training process involves alternating between training the Generator and training the Discriminator.[25] The Generator starts by creating data from random ideas. Over time, however, the Generator tries to improve its skills by looking at the Discriminator's feedback and refining its synthetic data in each epoch. The Discriminator, in turn, becomes better at catching fakes. This back-and-forth continues through successive epochs as the Generator gets better at creating realistic data and the Discriminator gets better at detecting fakes. If the algorithm is successful, the synthetic data created by the Generator eventually become indistinguishable from real data, and the Generator can be used on its own to generate new data that resemble the training data. **Exhibit 1** formally describes the GAN algorithm.

[25]The neural networks in the GAN in the example are trained using the Adam optimizer. See Kingma and Ba (2015) for more detail.

Exhibit 1. GAN Algorithm

Consider a Generator G and a Discriminator D engaged in a minimax game (a game where the Generator tries to minimize the difference between real and generated data while the Discriminator simultaneously tries to maximize its ability to distinguish between real and generated data). Let z be a latent variable (a random input to the Generator network that represents unobservable features of the data) sampled from $p_z(z)$ and x be a data sample from an unknown distribution $p_{\text{data}}(x)$. The Generator G maps z to a generated sample x, parameterized by θ_g. The Discriminator D computes $D(x;\theta_d)$, indicating the probability that x is genuine. Both networks have parameters θ_g and θ_d.

Train the algorithm by iteratively updating G and D until convergence as follows:

Inputs: Real data samples x from $p_{\text{data}}(x)$, latent samples z from $p_z(z)$, learning rates α_g and α_d, and hyperparameters k and n

Step 1. Randomly initialize θ_g and θ_d.

Step 2. Update Discriminator. For k steps:

Create a sample of real data x from $p_{\text{data}}(x)$.

Create a sample of latent variables from $p_z(z)$.

Compute Discriminator loss:

$$\mathcal{L}_d = \frac{1}{k}\sum_{i=1}^{k}\left[\log D(x_i;\theta_d) + \log\left(1 - D(G(z_i;\theta_g);\theta_d)\right)\right].$$

Update θ_d using gradient ascent (the optimization method): $\theta_d \leftarrow \theta_d + \alpha_d \cdot \nabla_{\theta_d}\mathcal{L}_d$.

Step 3. Update Generator. For n steps:

Create a sample of latent variables from $p_z(z)$.

Compute Generator loss:

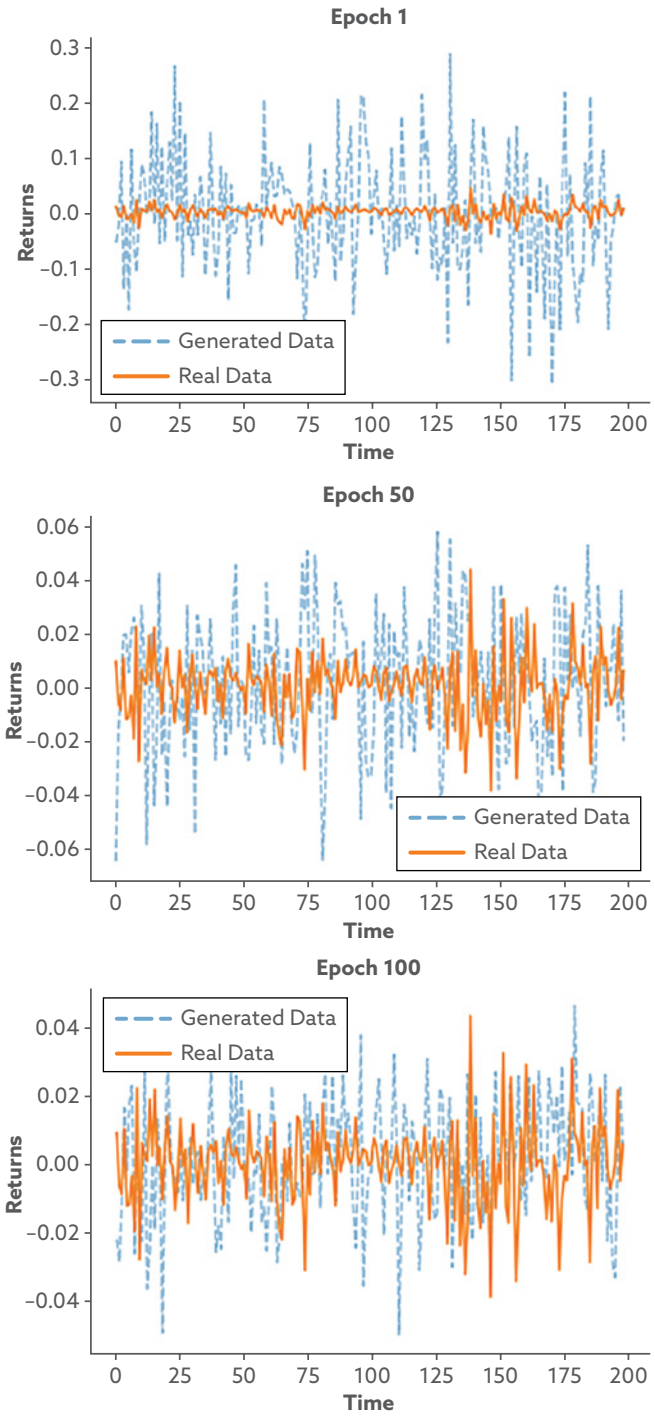
$$\mathcal{L}_g = -\frac{1}{n}\sum_{i=1}^{n}\log D(G(z_i;\theta_g);\theta_d).$$

Update θ_g using gradient ascent: $\theta_g \leftarrow \theta_g + \alpha_g \cdot \nabla_{\theta_g}\mathcal{L}_g$.

It is also possible to combine GANs with other techniques, such as bootstrapping. Specifically, we can use bootstrapping to create synthetic training data for a GAN. For example, suppose a researcher wants to test an investment model against the S&P 500. The first step in a *bootstrap GAN* would be to generate a number of bootstrapped time series. Next, the researcher would apply the GAN algorithm to try to replicate it. While moving through a number of epochs, the researcher will be generating a set of synthetic time series, each bearing various degrees of similarity to the original bootstrapped time series. **Figure 5** shows this process with a bootstrapped GAN over 100 epochs. As the figure shows, the generated data increasingly resemble

the original bootstrapped data as the epochs progress. Of course, the generated time series in Epoch 100 is not identical to the original bootstrapped time series. But that is the point; it is a realistic time series that differs in relevant ways from the source data.

Figure 5. Bootstrap GAN Example Output



6. MODEL COMPARISON

Often, researchers want to compare the performance of different models to identify the most accurate and reliable one. The models being compared could be models with similar performance statistics yet developed with completely different methodologies, or they could be variations of the same model. To compare models with an acceptable level of rigor requires using some formal criteria to evaluate them with.

The Akaike Information Criterion and Schwarz Criterion

Two of the most commonly used methodologies for model selection are the Akaike information criterion (AIC) and the Schwarz criterion (SC):

$$\text{AIC} = 2k + n\ln\left(\frac{\text{RSS}}{n}\right); \tag{16}$$

$$\text{SC} = n\ln\left(\frac{\text{RSS}}{n}\right) + k\ln(n), \tag{17}$$

where

n is the number of data points in the time series

RSS is the residual sum of squares

k is the number of factors in a model

The intuition behind model selection when using the AIC and SC is that when researchers identify a model, there is always a certain amount of information loss. For example, a factor model applied to portfolio returns cannot be better than the true model that describes the portfolio returns. One can think of a true model as a hypothetical or imaginary model that is "perfect." It generates the data one is trying to explain. Therefore, portfolio managers should select the model with the smallest information loss, which is what the AIC and SC seek to do. The smaller the number, the smaller the information loss—thus the closer the model is to the true model. Therefore, for both criteria, the model with the lowest value is preferred. Both the AIC and SC consider the number of explanatory variables. There is an advantage in using the SC when model simplicity is prioritized, however, because the SC imposes a higher penalty on models with many explanatory variables, thus favoring parsimonious models that adequately explain the data without unnecessarily adding complexity.

Worked Example

What follows is an example of how the AIC and SC work in practice. Consider a case in which a researcher wants to evaluate three predictive models for stock returns: linear regression,

random forest,[26] and an LSTM network.[27] Assume the following input values for each of the models:

$n = 100$.

Linear regression: $k_{LR} = 1$; $RSS_{Basic} = 200$.

Random forest: $k_{RF} = 2$; $RSS_{Alpha} = 180$.

LSTM neural network: $k_{LSTM} = 3$; $RSS_{Active\ value} = 160$.

Using these inputs, the AIC values are as follows:

$AIC_{LR} = 2k_{LR} - \ln(RSS_{LR}/n) = 2(1) - \ln(200/100) = 71.31$.

$AIC_{RF} = 2k_{RF} - \ln(RSS_{RF}/n) = 2(2) - \ln(180/100) = 62.78$.

$AIC_{LSTM} = 2k_{LSTM} - \ln(RSS_{LSTM}/n) = 2(3) - \ln(160/100) = 53.00$.

And these are the SC values:

$SC_{LR} = n \times \ln(RSS_{LR}/n) + k_{LR} \times \ln(n) = 100 \times \ln(200/100) + 1 \times \ln(100) = 73.91$

$SC_{RF} = n \times \ln(RSS_{RF}/n) + k_{RF} \times \ln(n) = 100 \times \ln(180/100) + 2 \times \ln(100) = 67.99$.

$SC_{LSTM} = n \times \ln(RSS_{LSTM}/n) + k_{LSTM} \times \ln(n) = 100 \times \ln(160/100) + 3 \times \ln(100) = 60.82$.

As the results show, the LSTM model has the lowest AIC and SC values. This result is interesting given the relative complexity of the model (3 factors) and the manner in which the criteria evaluate models. The AIC emphasizes goodness of fit while penalizing model complexity, favoring models that explain the data well without being overly complex. Lower AIC values suggest better models with a balance between fit and simplicity. The SC prioritizes both fit and model complexity equally but penalizes more severely for additional parameters. It tends to favor simpler models more strongly than the AIC. Because it includes a term that scales with the logarithm of the sample size, the SC makes the penalty for additional parameters larger as the sample size increases. By penalizing complex models more harshly, the SC often leads to the selection of simpler models that have a better chance of generalizing well to new, unseen data.

While the choice of the "best" model might depend on whether one prioritizes models with strong goodness of fit or simpler models that might generalize better to new data, in investment contexts, the latter is often more important. Because investing, especially active

[26]A random forest is a type of machine learning algorithm that uses decision trees. For regression-type problems, decision trees start from a topmost or root node and proceed to generate branches—with each branch containing a condition—and a prediction in the form of a real-valued number, given the condition in question. Trees are composed of a series of conditions attached to decision nodes, which ultimately arrive at a leaf or terminal node whose value is a real number. The latter value represents a predicted value for a target variable given a set of predictor values. For technical details on the random forest algorithm, see Breiman (2001). For an application of the random forest algorithm to predictive modeling, see Simonian, Wu, Itano, and Narayanam (2019).

[27]An LSTM (long short-term memory) network is a type of recurrent neural network architecture designed to effectively model long-term dependencies in sequential data. As the name implies, neural networks are (partially) modeled on the functioning of the human brain. Their basic design consists of a collection of data processors organized in layers, called neurons (or nodes). Information is processed via the responses of neurons to external inputs. These responses are then passed on to the next layer and so on until the final output. The interconnectedness of neurons and their ability to pass information back and forth to each other facilitates the efficient solution of problems. See Hochreiter and Schmidhuber (1997) for more details on LSTM networks.

management, is essentially an exercise in forecasting, models that are likely to exhibit stronger predictive power present an advantage to portfolio managers. By contrast, simpler models are often more attractive with regard to their explainability to clients and colleagues.

The McNemar Test

Given the importance of predictive power as a criterion for model selection, it is useful to employ a test that provides detailed insight into the predictive capabilities of the models being compared. The McNemar test, introduced in McNemar (1947), is one such model. It is a nonparametric statistical test for paired comparisons that can be applied to compare the performance of two predictive models. The McNemar test is also referred to as the "within-subjects chi-squared test," and it is applied to paired nominal data based on a version of a 2×2 contingency table that compares the predictions of two models. In the McNemar test, one formulates the null hypothesis that neither of the two models performs better than the other. Thus, the alternative hypothesis is that the predictive efficacy of the two models is not equal.

The McNemar chi-squared test statistic is computed as $\chi^2 = \frac{(|b-c|-1)^2}{(b+c)}$, where b is the number of observations that Model 1 correctly predicted and Model 2 incorrectly predicted and c is the number of observations that Model 2 correctly predicted and Model 1 incorrectly predicted.[28] The test statistic measures how much the observed counts in the contingency table deviate from what would be expected if there were no difference between the models. A higher chi-squared value suggests a larger difference between the models' predictions.

Worked Example

Consider a case where we are testing two investment models for their effectiveness at correctly predicting at time t whether the S&P 500 will have a positive or negative return at time $t+1$. The hypothetical test results are displayed in **Table 1**.

Looking at the matrix shows that Model 1 got five predictions correct that Model 2 got incorrect and Model 2 got 11 predictions correct that Model 1 got incorrect. So, the ratio is 11: 5, giving Model 2 somewhat better performance than Model 1. When calculating the McNemar chi-squared test statistic, its value is 1.56, suggesting that there is not a large difference between the models. This conclusion is supported by a p-value of 0.2113, indicating that the probability of observing differences between the models is approximately 21.13%, even if there is no true

Table 1. Example McNemar Matrix

	Model 2 Correct	Model 2 Incorrect
Model 1 Correct	15	5
Model 1 Incorrect	11	9

[28]The version of the statistic used here was first presented in Edwards (1948).

difference between them. Moreover, since the *p*-value is greater than 0.05, one might fail to reject the null hypothesis. This result suggests that, based on the observed data, there is an absence of sufficient evidence to claim a significant difference between the two models according to the McNemar test. Luckily, in such cases as this, one can conduct additional predictive tests to aid in drawing more definitive conclusions regarding the choice between Model 1 and Model 2.

Measures of Predictive Accuracy

Next, I describe some of the major measures of predictive accuracy and show how they can give a more holistic picture of model strength. The measures are precision, true positive rate (recall), accuracy, and F1. I also define a measure called receiver operating characteristic (ROC).

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}.$$

A higher precision score indicates fewer false positives, meaning the model is more precise in its positive predictions.

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}.$$

Higher recall implies that the model captures a larger proportion of actual positives.

$$\text{Accuracy} = \frac{\text{True positives} + \text{True negatives}}{\text{Total samples}}.$$

A higher accuracy score (closer to 1) suggests that the model makes fewer mistakes, correctly predicting more observations.

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

F1 scores range from 0 to 1. Higher F1 scores indicate a balance between precision and recall.

ROC Curve

ROC is a graphical metric with 1 as its maximum. It is based on a curve determined by two input metrics, the previously described true positive rate and the false positive rate:

$$\text{False positive rate} = \frac{\text{False positives}}{\text{False positives} + \text{True negatives}}.$$

The ROC curve shows how well a model can distinguish between the two classes by plotting the tradeoff between its true positive rate and false positive rate across different decision thresholds. A good ROC curve will tend toward the upper left corner of the plot, indicating high sensitivity and specificity. Thus, a larger area under the curve (AUC) value indicates a model that is better at distinguishing between positive and negative classes. In contrast, a random or ineffective model would produce a curve close to a diagonal line.

Applying these metrics to our hypothetical Models 1 and 2, **Figure 6** and **Figure 7** show that Model 2 exhibits significantly superior predictive power. This result is interesting given that the McNemar test revealed that there are no significant differences between the models. Nevertheless, the statistical tests considered here show a clear difference between the two models.

Figure 6. Predictive Tests for Hypothetical Models 1 and 2

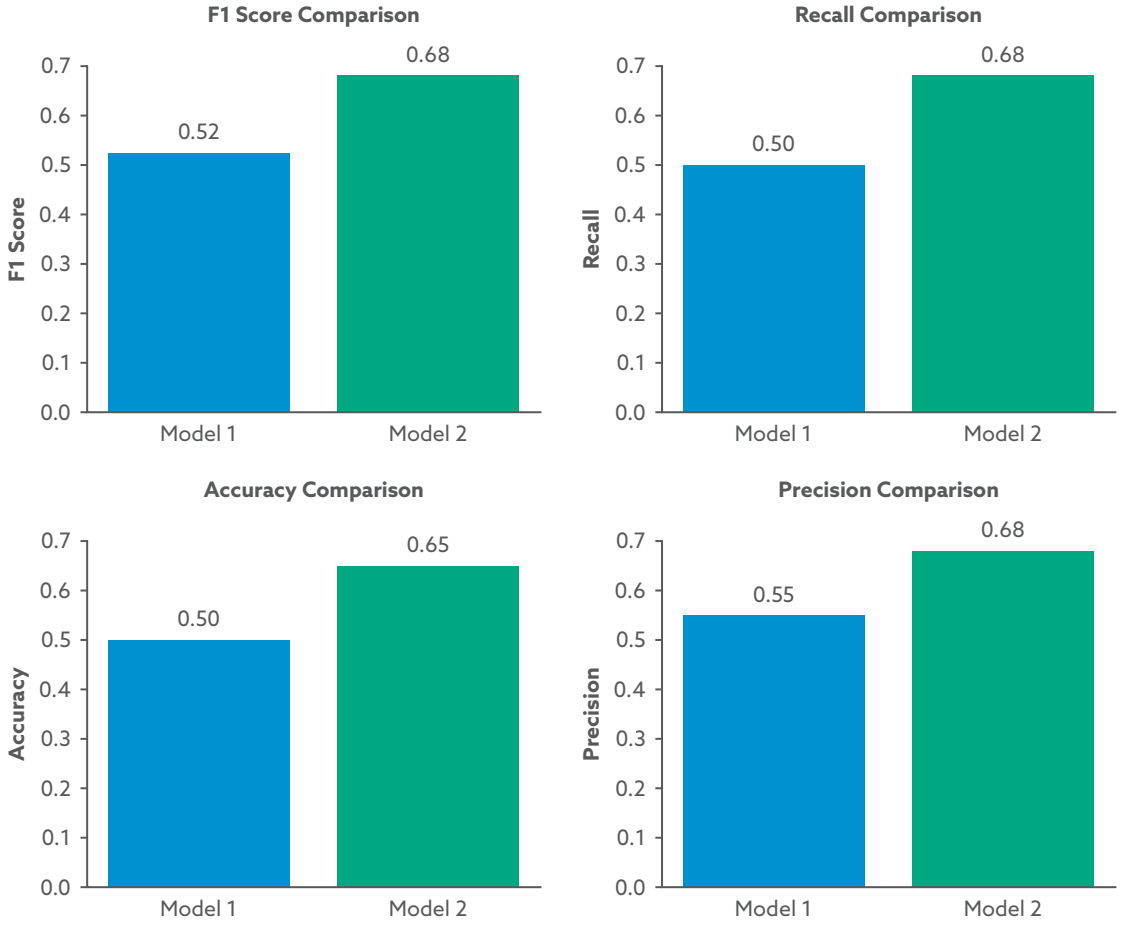
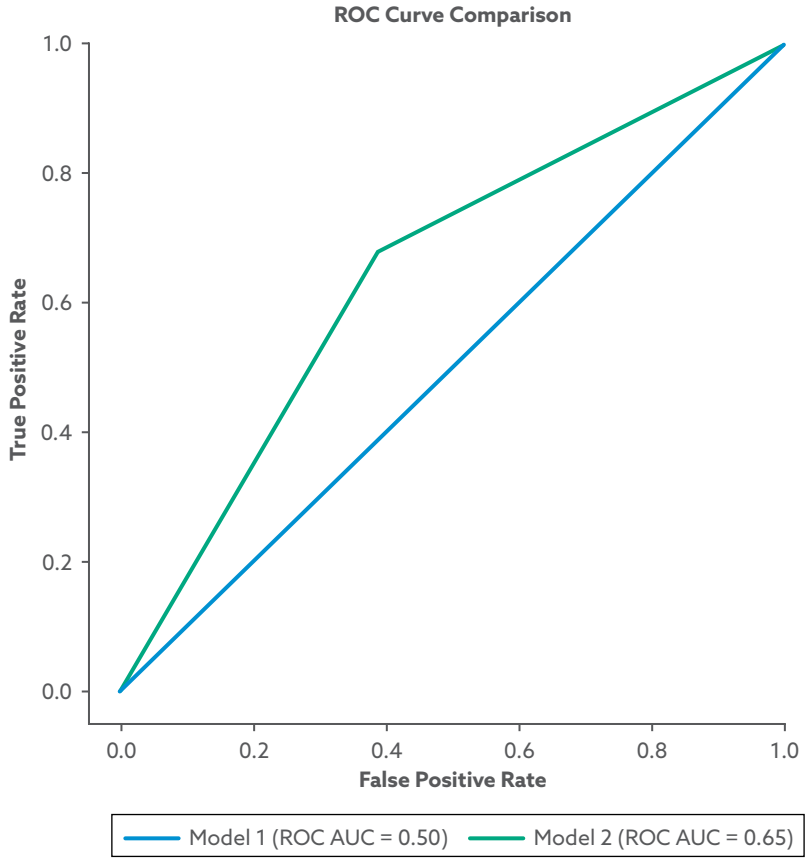


Figure 7. ROC Curves for Hypothetical Models 1 and 2



7. STRESS TESTING AND SCENARIO ANALYSIS

The most fundamental way to test the robustness of a machine, structure, or product is to try to break it. In engineering, stress testing and scenario analysis are fundamental aspects of project development. Their aim is to assess the robustness of physical objects to various rare and/or extreme natural and manmade phenomena that could damage the object or impair its functioning. Financial products are likewise vulnerable to unforeseen negative shocks. Thus, stress testing and scenario analysis are typically an integral part of the model validation process.

Stress Testing

Stress testing seeks to quantify the impact of low-probability yet high-impact market events, with the intent of allowing investors to form some reasonable expectation of the magnitude of portfolio losses if such events occur. Part of the stress-testing process involves the calculation of statistics that allow for a quantitative assessment of extreme investment losses. The VaR measure described previously is one such metric.

The primary focus of stress testing, however, is calibrating the primary risk factors in a model in disadvantageous ways in order to observe the reactions of a portfolio and to understand the most vulnerable dimensions of a particular investment strategy. For example, in a fixed-income model, various changes in credit spreads and the level, slope, and curvature of the yield curve may be posited to isolate the points of fragility. Or consider a multiasset portfolio for which one might assume various values for the correlations that exist between the commodity, currency, stock, and bond positions in a portfolio to assess how truly diversified it is.

Stress testing that encompasses this type of sensitivity analysis is also known as *factor push*. As is implied by its name, factor push assumes a factor model that describes a portfolio or strategy. Many, if not most, quantitative models today are constructed on the basis of risk factors.

Multifactor models are by now commonplace in investment management.[29] The primary attraction of factor models is that they help explain asset behavior using a parsimonious set of drivers, hence simplifying the analysis of portfolios consisting of many assets.

Because factor models provide a transparent view of the systemic risks that a portfolio is exposed to, they can be used for both risk management and alpha generation. One of the challenges in building factor models is that they must explain asset behavior adequately under the condition that the set of explanatory variables must be parsimonious. Given this constraint, the primary challenge for anyone building a factor model is to settle on a set of factors that, on the one hand, can adequately explain portfolio behavior over time and, on the other, is simple enough to remain computationally tractable. In this way, the challenge faced in building a factor model is the same as that faced by scientists when building theories to explain natural phenomena, in which the tradeoff between informative power and simplicity is also a fundamental consideration.

[29]See Connor (1995) for an overview of the various types of factor models.

The type of stress testing exemplified by factor push presumably relies on the factors that have already been selected to build a model. Thus, the choice of which factors to use to build a model is inextricably connected to the factors used in stress testing. This choice can be made in a number of ways. Given the fact that investors need to be concerned with both prediction and explanatory power, however, the ideal choice for a factor selection tool would be one that balances these two priorities. Luckily, the analytical framework known as *LASSO* (least absolute shrinkage and selection operator) is precisely such a tool (Santosa and Symes 1986; Tibshirani 1996). LASSO is a regression-based methodology that can help in both selecting variables for and mitigating against overfitting in models to enhance their predictive accuracy and interpretability. LASSO is described formally in the following manner:

$$\text{Minimize}\left(\frac{1}{2n}\sum_{i=1}^{n}(y_i-\alpha_0-\sum_{j=1}^{P}\beta_j x_{ij})^2+\lambda\sum_{j=1}^{P}\left|\beta_j\right|\right). \tag{18}$$

Equation 18 is an optimization in which the objective is to minimize the combination of the sum of squared errors and a penalization term. In it, y_i is the variable to be predicted, α_0 is an intercept, x_{ij} signifies the value of the jth predictor (factor) for the ith observation, and β_j is the coefficient for the jth predictor. The regularization term $\lambda\sum_{j=0}^{P}\left|\beta_j\right|$ penalizes the absolute values of the coefficients. The strength of the penalty is controlled by the regularization parameter λ. When λ takes a value of zero, it results in the ordinary least-squares model. A sufficiently large value of λ, however, will force some of the coefficients β_j to become zero, thus excluding them from the model. If one wanted to select a subset of factors (say, 5) from a larger set of candidate factors (say, 40), one would add the constraint $\sum_{j=1}^{P}\mathbb{I}(\beta_j\neq 0)\leq k$ to Equation 18, where k is the desired number of factors, P is the number of candidate factors, and $\mathbb{I}$ is an indicator function.

There are many convenient and practical ways to use LASSO in the context of model development. For example, when building a multiasset model, typically dozens of potential factors could provide some informational value. However, in keeping with the goal of parsimony for factor models, it would be impractical to use a multitude of factors, only some of which would be the primary drivers of portfolio behavior. It would be much more efficient to select a subset of a more expansive array of risk factors. To do this via LASSO, one would simply have to select a large number of factors, specify the final number of factors in our model, and run the procedure described here. Then, LASSO will provide the subset of risk factors that have the most explanatory and predictive power while simultaneously having the lowest correlation with one another. With a LASSO-derived factor model in hand, one can proceed with sensitivity analyses, such as factor push.

One of the most important aspects of sensitivity analysis is the recalibration of assumed correlations between risk factors and/or assets. The reason why correlations play such a prominent role in stress testing is that diversification is the cornerstone of portfolio construction. As described earlier in this monograph, lower correlations between assets and the factors that drive a portfolio's behavior will, all things being equal, result in lower portfolio volatility. Thus, within the context of stress testing, it is often useful to assume that factors and assets have significantly higher but still realistic correlations with one another compared to their historical average correlations.

Reverse Stress Testing

Stress testing evaluates how a model would perform under adverse market conditions, such as a severe market crash or a sudden spike in interest rates. Reverse stress testing, in contrast, is an approach to risk analysis where the traditional stress-testing process is reversed. Instead of testing the impact of predefined stress scenarios on a portfolio, reverse stress testing begins with determining the extreme adverse outcomes that would lead to the failure of an institution, portfolio, or financial system. The researcher then figures out what combination of market events would produce those outcomes.

For example, suppose a portfolio management team expects that in the event of a 25% drawdown, they will lose 75% of their invested capital due to client withdrawals. The latter is the team's definition of "catastrophic loss." Once defined, they can proceed to determine what turn of events could lead to such a scenario.

Scenario Analysis

Scenario analysis is a process in which portfolio outcomes are evaluated under different market scenarios. The scenarios may be based on actual events, but they do not have to be. The use of data from recessionary or otherwise adverse market events (such as inflation shocks), however, is common.

The market events used as templates in scenario analysis could be regularly occurring dislocations or more idiosyncratic events. For example, Packham and Woebbeking (2019) present a correlation stress-testing case study based on the "London Whale" episode, which resulted in a $6.2 billion loss on a credit derivative portfolio at JPMorgan due to rogue trading by a single individual. A primary driver of the portfolio losses was the breakdown of the correlations between positions that were assumed to be hedges for each other. The authors show that correlation stress testing would have revealed this risk early on and allowed the appropriate risk management response to be implemented. The risk factors used in the case study were chosen to match the characteristics of a credit derivative portfolio, such as credit quality and maturity.

Stylized scenarios do not necessarily have to be *ex post*, drawn from actual history, but may be based on *ex ante* expected events. For example, a concrete risk event may be clearly seen on the horizon even though there are no actual cases to draw specific data from. An example of such a scenario was the risk of a "fiscal cliff" in 2012, when a number of tax increases and spending cuts were scheduled to take effect simultaneously at the beginning of 2013. The key elements of the fiscal cliff included the expiration of the Bush-era tax cuts, the end of a temporary payroll tax cut, automatic spending cuts (also known as "sequestration") that were mandated by the Budget Control Act of 2011, and the expiration of extended unemployment benefits. The risk to the markets was that Congress would be unable to reach a bipartisan compromise regarding how to address the impending contraction in fiscal support to the economy. The fear was that, due to the forthcoming tax increases and spending cuts, negative repercussions would be transmitted to financial markets.

The most extreme scenario was ultimately avoided due to the passage of the American Taxpayer Relief Act, which made permanent the tax cuts for most Americans but allowed some tax increases for higher-income earners. The automatic spending cuts were also delayed for a brief period. From the standpoint of scenario analysis, the specific combination of tax increases

and spending cuts that created the fiscal cliff scenario in 2012 was unusual. It is thus an example of a type of scenario that can prove challenging during model validation. Rather than being able to draw directly on historical data, one would have to construct the scenario from different historical occurrences. For example, one could assume that in the worst fiscal cliff scenario, the US economy would contract by 3%. The next step would be to study the various episodes when the US economy did experience that level of contraction. Then, possible ranges for movement in the stock market, credit spreads, commodities, and so on, could be derived. Next, a given portfolio could be run through various combinations of values falling in the ranges posited for each asset in the portfolio, and one could thereby derive some understanding of possible losses to the investment strategy under consideration.

Of course, it is possible to create entirely fabricated scenarios. The basic way that this is accomplished is by assuming extreme risk and return assumptions for the risk factors and assets of interest. As mentioned previously, among the most important metrics is the set of correlation coefficients between assets. Thus, creating "shock" correlation matrices is typically a central aspect of creating stylized scenarios. One way of generating appropriate correlation matrices for stress testing is to choose a real-valued reference matrix R—for example, one representing a more tranquil or commonplace market environment—and then "pull" it toward a real-valued target matrix T representing a more turbulent or low-probability state of the world.

Doing this mathematically generally involves using a methodology called *shrinkage* (Ledoit and Wolf 2004) to produce a convex linear combination of the two matrices. The precise values of the shrinkage matrix $S_\lambda = (1 - \lambda)R + \lambda T$ are determined by accounting for a shrinkage factor $\lambda \in [0,1]$, which expresses the degree to which we wish S_λ to resemble a given market scenario. One heuristic that can help us calibrate λ is to assume that it represents the probability of the stress scenario materializing. The stipulated probability can be objective, based on historical frequency, or subjective, based on forward-looking views.

8. VALIDATING MODELS AGAINST ECONOMIC THEORY

In building any investment model, it is important to assess whether a model aligns with basic economic intuition and theory. There are many theories of market behavior encompassing a wide array of ideas regarding market efficiency, risk and return tradeoffs, utility maximization, behavioral biases, and macroeconomic themes. And while these theories are not always consistent with one another, there are nevertheless some fundamental drivers of economic behavior that have been accepted by the majority of economists and investors. The understanding of these principles should inform the conceptual framework on which investment models are built. While it is beyond the scope of a monograph on the topic of model validation to judge the merits of any investment theory, I can provide some insight into how investment theories can be used to assess the consistency of a model during the validation process.

First consider one of the cornerstones of contemporary investment theory, the *capital asset pricing model* (CAPM; Treynor 1961, 1962; Sharpe 1964; Lintner 1965; Mossin 1966). The CAPM provides a framework for estimating the expected return of an asset given its risk (specifically *systematic risk*, that part of the asset's total risk that is correlated with the cap-weighted portfolio of all stocks or all risky assets). The model says that asset expected returns, in excess of the riskless rate, are proportional to the systematic risk of the asset. The systematic risk inherent in a given asset is called the asset's *beta*, measured on a scale where the beta of the overall market is 1. In the CAPM, a stock's expected return is the sum of the risk-free rate and a risk premium, which is the product of the stock's beta and the expected excess return of the market as a whole above the risk-free rate (the market risk premium).

The CAPM can be used in model validation in a number of ways. The CAPM assumes that in an efficient market, the only way to achieve returns above the market is by taking on additional risk (beta). If a trading model consistently generates excess returns (alpha) that cannot be explained by CAPM, it suggests the model is adding value beyond what would be expected based on systematic risk. A related application is in cross-validation. One can compare the predictions of a trading model, evaluated against the CAPM, by observing the alphas it generated to determine whether the trading model produces material improvements over some other model that is already in use or being considered. If it does, that could be an indication that the model is additive to a portfolio manager's investment process.

Of course, the CAPM involves a number of assumptions that undermine its status as a complete and accurate descriptor of market behavior.[30] One of these assumptions is that the market beta is the sole source of all explainable variation in asset returns. Over the years, this assumption has been brought into question and multifactor models—models that posit multiple sources of systematic risk—have been developed in its stead. In the realm of equities, perhaps the most well-known of these models is the *Fama–French–Carhart model* (Fama and French 1992, 1993; Carhart 1997), which extends the CAPM framework by introducing three new factors in addition to the market factor: the size factor (small-cap stock returns minus large-cap stock returns),

[30]For a complete list of these assumptions, see Elton, Gruber, Brown, and Goetzmann (2003).

value (high-book-to-price stock returns minus low-book-to-price stock returns), and momentum (high-returning stocks over some previous period, usually a year, minus low-returning stocks). As in the case of the CAPM, multifactor models can also be used in model validation to assess the effectiveness of trading models. For example, it is often important for a trading model to generate alpha via security selection (idiosyncratic risk) rather than by taking on systematic exposures to risk factors. Using a multifactor model during the model validation process can help determine the sources of a model's performance relative to a benchmark. If a model maintains the same factor exposures as its designated benchmark yet still produces alpha, that provides evidence that the model is adding value. Multifactor models can also be used in the same manner as the CAPM to test the predictive efficacy of an investment model.

Another mainstay of investment theory is what is known as the *efficient market hypothesis* (EMH),[31] which asserts that asset prices reflect all available information. Most investors accept that the strongest version of the EMH is not true—that some inefficiencies are characteristic of all markets to varying degrees. This belief is not hard to understand. Given that security prices are a function of information, it is implausible that every piece of relevant information is reflected in asset prices. The "information loss" is especially pronounced in less efficient markets, such as those found in emerging economies or in the high-yield bond sector. Exploiting such informational inefficiencies is what many trading models are designed to do. Thus, when evaluating an investment model's performance, it may be important to ask the following questions: What inefficiencies is the model exploiting, if any? Are the inefficiencies that are being exploited the ones intended by the model? How reliable is the model in terms of its ability to exploit its targeted inefficiencies? Is the model better at exploiting some inefficiencies over others?

One of the ways that investment models often capitalize on market inefficiencies is by effectively detecting and trading around the various aspects of investor behavior that cause market inefficiencies. The body of investment theory that studies such behavioral drivers of market dynamics is known as *behavioral finance*.[32] This field of study combines insights from psychology and economics to understand and explain how individuals make financial decisions. Unlike traditional financial theories that assume rationality and efficiency in financial markets, behavioral finance recognizes that human behavior and emotions play a significant role in shaping market anomalies, such as bubbles, crashes, momentum, and the value effect (the outperformance of undervalued stocks over time).

Behavioral finance often focuses on the related concepts of *biases* and *heuristics*. Various cognitive biases impact decision making, including *confirmation bias* (seeking or giving extra weight to information that confirms preexisting beliefs) and *availability bias* (relying on the most readily available information). *Familiarity bias* refers to the propensity of individuals to invest in familiar assets, sectors, or regions rather than considering unexplored investment opportunities. *Overconfidence*, the tendency of individuals to overestimate their own abilities, knowledge, or judgments, is another common cognitive bias. *Anchoring* is a cognitive bias where individuals rely heavily on the initial information they receive. *Recency bias* is exhibited when individuals give more weight to recent information or experiences when making decisions. The foregoing list of cognitive biases, while not exhaustive, inventories some of the most important cognitive biases in finance. Heuristics are the mental "shortcuts" or "rules of thumb" that help

[31]See Fama (1970) for a detailed discussion of the EMH.

[32]For a review of behavioral finance, see Shleifer (2000).

us make decisions quickly. While heuristics can assist in accelerating the decision-making process, they can also lead to systematic errors if used constantly without giving any scrutiny to the appropriateness of their application. Examples of heuristics include the following of *expert opinion* and abiding by *majority rule*, basing decisions on team or firmwide popularity rather than thoughtful deliberation.

Often, investment models are built on behavioral foundations; perhaps the most obvious are momentum strategies. Indeed, the recognition of behavioral concepts and their application to investment products is now widespread. Thus, when evaluating a model during the validation process, it is important to answer three questions: Is the model, in fact, successfully exploiting one or more aspects of investor behavior? Is the model able to exploit said behaviors in a systematic fashion? And how much of a model's performance can be attributed to the exploitation of the behaviors(s) in question versus other factors?

Finally, as mentioned earlier, diversification as a means to control risk in portfolios is almost axiomatic in modern investment practice. However, investment models often call for portfolio weights that deviate, at times significantly, from those of a mean–variance-optimal portfolio. Comparing an investment model's performance to a mean–variance-optimal or otherwise diversified portfolio (e.g., a risk parity portfolio)[33] in backtests and simulations can often provide insight into the risk and return tradeoffs inherent in using a given model. This is especially true when an investment model uses risk measures other than variance or standard deviation as inputs.

[33]For a discussion of the development of risk parity strategies, see Fabozzi, Simonian, and Fabozzi (2021).

9. PREPARING MODEL DOCUMENTATION

Model documentation must be comprehensive, accurate, and well maintained so that the validation process is transparent and facilitates understanding of the model by stakeholders. The documentation should describe both the model and the various tests that have been used throughout the validation process. Preparing two types of model documentation is useful. First, a comprehensive set of documents should be prepared for internal use and for the purposes of any external audits. This set of documents should be considered the *canonical model documentation*. This type of documentation is important because it facilitates the following aspects of model acceptance and implementation:

Reproducibility: The canonical set of documents should be detailed enough that the model and any tests involved in validation could be reproduced by any member of the firm possessing sufficient technical capabilities. Such stakeholders should be able to understand the steps taken, the data used, the metrics evaluated, and the decisions made during validation. This reproducibility helps in verifying and validating the model's performance and results.

Transparency and accountability: Documentation provides transparency about the methods, assumptions, and choices made during model validation. This transparency is essential for accountability and allows stakeholders to understand how conclusions and assessments were reached during the validation process.

Traceability: Detailed documentation helps trace the lineage of the model, including data sources, preprocessing steps, factor selection, and the reasoning behind specific model choices. This traceability is important for understanding the model's life cycle and for identifying potential biases or errors.

Problem diagnosis and improvement: In case of issues or unexpected results, documentation acts as a reference for diagnosing problems in a model. Understanding the validation process also helps identify where improvements or adjustments to a model could be made.

Regulatory and firm compliance: In regulated industries, such as banking, documentation is often a regulatory requirement. Comprehensive documentation ensures that the model validation process complies with industry standards and regulations. Even in the absence of industry-wide regulations, firms may have "best practices" with regard to model development and testing. Proper documentation provides a tangible record that firm-wide best practices have been adhered to.

Communication: Well-documented validation processes facilitate communication among team members, allowing them to understand each other's work, collaborate effectively, and build on previous validation efforts. Communication is also important for the second type of model documentation, called *expository model documentation*. This type of documentation is intended for those stakeholders who, while potentially important for the approval, dissemination, and/or use of the model within a firm, may not be technically proficient enough to follow every technical and mathematical detail pertaining to it or the model validation process. Expository documentation should thus transmit the relevant specifications of the model and validation process in conceptual terms that could be understood by any investment professional.

There is an additional, residual benefit to preparing thorough and detailed model documentation—the potential to publish schematic or otherwise more "skeletal" versions of the model documentation as research in peer-reviewed journals. While care must be taken to withhold any proprietary information (the "secret sauce") in any published paper, transmitting the main ideas and results of model development and validation in a journal can serve a useful purpose for firms: It exposes their work to subject matter experts who can provide valuable feedback on the reasonableness of model assumptions and outputs. It can also serve to publicize the firm's research efforts to the wider investment community.

10. CONCLUSION

The goal of this monograph is to provide an in-depth overview of the most important dimensions of model validation. The general assumption is that when validating financial models, investment professionals should take a scientific approach. What this means in practice is that the purpose of model validation should be to try to *falsify* an investment model. If attempts at falsification fail despite the use of different approaches and techniques to falsify it, the model can be considered valid and suitable for implementation.

Various methods can be used in an attempt to validate (falsify) an investment model. Backtesting—evaluating a model's performance on historical data—is fundamental and often the first step in a model validation process. Because there is no guarantee that the future will resemble the past, however, backtesting has its limitations. For that reason, methods beyond basic backtesting have been developed to further gauge the robustness of investment models. This monograph attempts to summarize and explain in accessible language such additional techniques.

First and foremost is cross-validation, which can be considered a cousin of backtesting. Although cross-validation in its basic form is also based on historical data, it nevertheless provides a way to clearly separate model development and calibration from model testing. It thus serves as a formal way of mitigating overfitting. In both backtesting and cross-validation, it is important to use a range of return and risk statistics and carefully select appropriate benchmarks in order to gain a clear picture of the performance characteristics of a given investment model.

Relative to natural science, finance is data limited. This fact makes the development and application of tools to generate synthetic time series an important component of robust model validation. The monograph discusses two traditional synthetic data generation techniques, Monte Carlo simulation and bootstrapping, as well as a newer machine-learning-driven approach based on generative adversarial networks.

Often, more than one candidate model is being evaluated during the model validation process. It is thus necessary to have the formal means to compare models. The monograph discusses two of the most widely used model comparison methodologies—the Akaike information criterion and the Schwarz criterion—as well as the most widely accepted methods for evaluating the predictive efficacy of models.

A major component of model validation is evaluating a model's response to different extreme or stress scenarios to assess its robustness. These scenarios can be hypothetical or based on actual market events. Stress testing and scenario analysis are useful for understanding how a model performs under various market conditions and how it responds to isolated or multiple economic shocks.

Aside from empirical tests, additional insight regarding a model may be gained by evaluating its consistency with prevailing investment theory. Although investment models do not necessarily have to completely conform to any given investment theory to be considered valid, it would be unusual for the mechanics of an effective investment model to be wholly inconsistent with basic economic intuition and the fundamentals of market behavior.

Finally, it is important to document the model validation process in detail for the benefit of model users and stakeholders, including any external auditors. Model validators may benefit from developing two sets of documentation, a technically detailed canonical set of documents and a more accessible expository set of documents. The latter would serve those stakeholders who do not possess the requisite technical background that would allow them to thoroughly evaluate the quantitative aspects of a model validation process.

Model validation is a critical aspect of any quantitative investment process. This guide provides investment professionals with a blueprint with which they can fashion their own model validation frameworks. By using the various techniques and methodologies described in the monograph, investors will be able to conduct model validation with scientific and statistical rigor and enhance the quality of their investment products for the ultimate benefit of their clients and beneficiaries.

BIBLIOGRAPHY

Bacon, Carl. 2004. *Practical Portfolio Performance Measurement and Attribution*. Hoboken, NJ: John Wiley & Sons.

Bergmeir, Christoph, and José M. Benítez. 2012. "On the Use of Cross-Validation for Time Series Predictor Evaluation." *Information Sciences* 191: 192–213. doi:10.1016/j.ins.2011.12.028.

Bergmeir, Christoph, Rob J. Hyndman, and Bonsoo Koo. 2018. "A Note on the Validity of Cross-Validation for Evaluating Autoregressive Time Series Prediction." *Computational Statistics & Data Analysis* 120: 70–83. doi:10.1016/j.csda.2017.11.003.

Bollerslev, Tim. 1986. "Generalized Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics* 31 (3): 307–27. doi:10.1016/0304-4076(86)90063-1.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. doi:10.1023/A:1010933404324.

Brinson, Gary P., and Nimrod Fachler. 1985. "Measuring Non-U.S. Equity Portfolio Performance." *Journal of Portfolio Management* 11 (3): 73–76. doi:10.3905/jpm.1985.409005.

Brinson, Gary P., L. Randolph Hood, and Gilbert L. Beebower. 1986. "Determinants of Portfolio Performance." *Financial Analysts Journal* 42 (4): 39–44. doi:10.2469/faj.v42.n4.39.

Burman, Prabir, Edmond Chow, and Deborah Nolan. 1994. "A Cross-Validatory Method for Dependent Data." *Biometrika* 81 (2): 351–58. doi:10.1093/biomet/81.2.351.

Carhart, Mark M. 1997. "On Persistence in Mutual Fund Performance." *Journal of Finance* 52 (1): 57–82. doi:10.1111/j.1540-6261.1997.tb03808.x.

Connor, Gregory. 1995. "The Three Types of Factor Models: A Comparison of Their Explanatory Power." *Financial Analysts Journal* 51 (3): 42–46. doi:10.2469/faj.v51.n3.1904.

Edwards, A. L. 1948. "Note on the 'Correction for Continuity' in Testing the Significance of the Difference between Correlated Proportions." *Psychometrika* 13 (3): 185–87. doi:10.1007/BF02289261.

Efron, Bradley. 1979. "Bootstrap Methods: Another Look at the Jackknife." *Annals of Statistics* 7 (1): 1–26. doi:10.1214/aos/1176344552.

Elton, Edwin J., Martin J. Gruber, Stephen J. Brown, and William N. Goetzmann. 2003. *Modern Portfolio Theory and Investment Analysis*, 6th ed. Hoboken, NJ: John Wiley & Sons.

Fabozzi, Frank, Joseph Simonian, and Francisco J. Fabozzi. 2021. "Risk Parity: The Democratization of Risk in Asset Allocation." *Journal of Portfolio Management* 47 (5): 41–50. doi:10.3905/jpm.2021.1.228.

Fama, Eugene F. 1970. "Efficient Capital Markets: A Review of Theory and Empirical Work." *Journal of Finance* 25 (2): 383–417. doi:10.2307/2325486.

Fama, Eugene F. 1972. "Components of Investment Performance." *Journal of Finance* 27 (3): 551–67.

Fama, Eugene F., and Kenneth R. French. 1992. "The Cross-Section of Expected Stock Returns." *Journal of Finance* 47 (2): 427–65.

Fama, Eugene F., and Kenneth R. French. 1993. "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics* 33 (1): 3–56. doi:10.1016/0304-405X(93)90023-5.

Filardo, Andrew J. 1994. "Business-Cycle Phases and Their Transitional Dynamics." *Journal of Business & Economic Statistics* 12 (3): 299–308. doi:10.1080/07350015.1994.10524545.

Glasserman, Paul. 2003. *Monte Carlo Methods in Financial Engineering*. New York: Springer-Verlag. doi:10.1007/978-0-387-21617-1.

Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. "Generative Adversarial Nets." *Proceedings of the 27th International Conference on Neural Information Processing Systems* 2: 2672–80.

Granger, Clive W. J. 1969. "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods." *Econometrica* 37 (3): 424–38. doi:10.2307/1912791.

Hamilton, James D. 1989. "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle." *Econometrica* 57 (2): 357–84. doi:10.2307/1912559.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9 (8): 1735–80. doi:10.1162/neco.1997.9.8.1735.

Johansen, Anders, and Didier Sornette. 1998. "Stock Market Crashes Are Outliers." *European Physical Journal B* 1: 141–43. doi:10.1007/s100510050163.

Karnosky, Denis S., and Brian D. Singer. 1994. *Global Asset Management and Performance Attribution*. Charlottesville, VA: CFA Institute Research Foundation.

Kingma, Diederik P., and Jimmy Ba. 2015. "Adam: A Method for Stochastic Optimization." *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*.

Konstantinov, Gueorgui S., Frank F. Fabozzi, and Joseph Simonian. 2023. *Quantitative Global Bond Portfolio Management*. World Scientific. doi:10.1142/13313.

Lachenbruch, Peter A., and M. Ray Mickey. 1968. "Estimation of Error Rates in Discriminant Analysis." *Technometrics* 10 (1): 1–11. doi:10.1080/00401706.1968.10490530.

Larson, Selmer C. 1931. "The Shrinkage of the Coefficient of Multiple Correlation." *Journal of Educational Psychology* 22 (1): 45–55. doi:10.1037/h0072400.

Ledoit, Olivier, and Michael Wolf. 2004. "Honey, I Shrunk the Sample Covariance Matrix." *Journal of Portfolio Management* 30 (4): 110–19. doi:10.3905/jpm.2004.110.

Lintner, J. 1965. "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets." *Review of Economics and Statistics* 47 (1): 13–37.

Lo, Andrew W. 2016. "What Is an Index?" *Journal of Portfolio Management* 42 (2): 21–36. doi:10.3905/jpm.2016.42.2.021.

Mammen, Enno. 1993. "Bootstrap and Wild Bootstrap for High Dimensional Linear Models." *Annals of Statistics* 21 (1): 255–85. doi:10.1214/aos/1176349025.

Markowitz, Harry M. 1952. "Portfolio Selection." *Journal of Finance* 7 (1): 77–91.

McNemar, Quinn. 1947. "Note on the Sampling Error of the Difference between Correlated Proportions or Percentages." *Psychometrika* 12 (2): 153–57. doi:10.1007/BF02295996.

Mossin, Jan. 1966. "Equilibrium in a Capital Asset Market." *Econometrica* 34 (4): 768–83. doi:10.2307/1910098.

Mosteller, Frederick, and John W. Tukey. 1968. "Data Analysis, Including Statistics." In *Handbook of Social Psychology*. Reading, MA: Addison-Wesley.

Packham, Natalie, and Fabian Woebbeking. 2019. "A Factor-Model Approach for Correlation Scenarios and Correlation Stress Testing." *Journal of Banking & Finance* 101: 92–103. doi:10.1016/j.jbankfin.2019.01.020.

Politis, Dimitris N., and Joseph P. Romano. 1994. "The Stationary Bootstrap." *Journal of the American Statistical Association* 89 (428): 1303–13. doi:10.1080/01621459.1994.10476870.

Racine, Jeff. 2000. "Consistent Cross-Validatory Model-Selection for Dependent Data: *hv*-Block Cross-Validation." *Journal of Econometrics* 99 (1): 39–61. doi:10.1016/S0304-4076(00)00030-0.

Roberts, David R., Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, et al. 2017. "Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure." *Ecography* 40 (8): 913–29. doi:10.1111/ecog.02881.

Santosa, Fadil, and William W. Symes. 1986. "Linear Inversion of Band-Limited Reflection Seismograms." *SIAM Journal on Scientific and Statistical Computing* 7 (4): 1307–30. doi:10.1137/0907087.

Sharpe, William. F. 1964. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *Journal of Finance* 19 (3): 425–42.

Sharpe, William. F. 1966. "Mutual Fund Performance." *Journal of Business* 39 (1): 119–38. doi:10.1086/294846.

Sharpe, William. F. 1992. "Asset Allocation: Management Style and Performance Measurement." *Journal of Portfolio Management* 18 (2): 7–19. doi:10.3905/jpm.1992.409394.

Shleifer, Andrei. 2000. *Inefficient Markets: An Introduction to Behavioral Finance*. Oxford, UK: Oxford University Press. doi:10.1093/0198292279.001.0001.

Siegel, Laurence B. 2003. *Benchmarks and Investment Management*. Charlottesville, VA: CFA Institute Research Foundation.

Simonian, Joseph. 2020. "Modular Machine Learning for Model Validation: An Application to the Fundamental Law of Active Management." *Journal of Financial Data Science* 2 (2): 41–50. doi:10.3905/jfds.2020.1.027.

Simonian, Joseph, and Chenwei Wu. 2019. "Minsky vs. Machine: New Foundations for Quant-Macro Investing." *Journal of Financial Data Science* 1 (2): 94–110. doi:10.3905/jfds.2019.1.004.

Simonian, Joseph, Chenwei Wu, Daniel Itano, and Vyshaal Narayanam. 2019. "A Machine Learning Approach to Risk Factors: A Case Study Using the Fama–French–Carhart Model." *Journal of Financial Data Science* 1 (1): 32–44. doi:10.3905/jfds.2019.1.032.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society, Series B (Methodological)* 58 (1): 267–88. doi:10.1111/j.2517-6161.1996.tb02080.x.

Treynor, Jack L. 1961. "Market Value, Time, and Risk." Unpublished manuscript (8 August). doi:10.2139/ssrn.2600356.

Treynor, Jack L. 1962. "Toward a Theory of Market Value of Risky Assets." Mimeo, final version published in *Asset Pricing and Portfolio Performance: Models, Strategy and Performance Metrics*, edited by R. A. Korajczyk. 1999. London: Risk Books.

Treynor, Jack L. 1965. "How to Rate Management of Investment Funds." *Harvard Business Review* 43 (1): 63–75.

Wu, Chien-Fu J. 1986. "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis." *Annals of Statistics* 14 (4): 1261–95. doi:10.1214/aos/1176350142.

Named Endowments

CFA Institute Research Foundation acknowledges with sincere gratitude the generous contributions of the Named Endowment participants listed below.

Gifts of at least US$100,000 qualify donors for membership in the Named Endowment category, which recognizes in perpetuity the commitment toward unbiased, practitioner-oriented, relevant research that these firms and individuals have expressed through their generous support of CFA Institute Research Foundation.

Ameritech
Anonymous
Robert D. Arnott
Theodore R. Aronson, CFA
Asahi Mutual Life Insurance Company
Batterymarch Financial Management
Boston Company
Boston Partners Asset Management, L.P.
Gary P. Brinson, CFA
Brinson Partners, Inc.
Capital Group International, Inc.
Concord Capital Management
Dai-Ichi Life Insurance Company
Daiwa Securities
Mr. and Mrs. Jeffrey Diermeier
Gifford Fong Associates
John A. Gunn, CFA
Investment Counsel Association of America, Inc.
Jacobs Levy Equity Management
Jon L. Hagler Foundation
Long-Term Credit Bank of Japan, Ltd.
Lynch, Jones & Ryan, LLC
Meiji Mutual Life Insurance Company
Miller Anderson & Sherrerd, LLP
John B. Neff, CFA
Nikko Securities Co., Ltd.
Nippon Life Insurance Company of Japan
Nomura Securities Co., Ltd.
Payden & Rygel
Provident National Bank
Frank K. Reilly, CFA
Salomon Brothers
Sassoon Holdings Pte. Ltd.
Scudder Stevens & Clark
Security Analysts Association of Japan
Shaw Data Securities, Inc.
Sit Investment Associates, Inc.
Standish, Ayer & Wood, Inc.
State Farm Insurance Company
Sumitomo Life America, Inc.
T. Rowe Price Associates, Inc.
Templeton Investment Counsel Inc.
Frank Trainer, CFA
Travelers Insurance Co.
USF&G Companies
Yamaichi Securities Co., Ltd.

Senior Research Fellows

Financial Services Analyst Association

For more on upcoming CFA Institute Research Foundation publications and webcasts, please visit www.cfainstitute.org/research/foundation.

CFA Institute Research Foundation Board of Trustees 2023–2024

Chair
Aaron Low, PhD, CFA
LUMIQ

Vice Chair
Jeff Bailey, CFA
Groveley Associates

Margaret Franklin, CFA
CFA Institute

Giuseppe Ballocchi, PhD, CFA
Alpha Governance Partners
University of Lausanne

Aaron Brown, CFA
City of Calgary

Frank Fabozzi, PhD, CFA*
The Johns Hopkins University
Carey Business School

Bill Fung, PhD
Aventura, FL

Philip Graham, CFA
Consultant – AustralianSuper

Joanne Hill, PhD
Cboe Vest LLC

Roger Ibbotson, PhD*
Yale School of Management

Lotta Moberg, PhD, CFA
ViviFi Ventures

Punita Kumar-Sinha, PhD, CFA
Infosys

Susan Spinner, CFA
CFA Society Germany

Dave Uduanu, CFA
Sigma Pensions Ltd

Kurt Winkelmann, PhD
Navega Strategies

*Emeritus

Officers and Directors

Gary P. Brinson Director of Research
Laurence B. Siegel
Blue Moon Communications

Research Director
Luis Garcia-Feijóo, CFA, CIPM
Coral Gables, Florida

Director of Data Science
Francesco Fabozzi

Treasurer
Kim Maynard
CFA Institute

Director of Operations
Bud Haslett, CFA
Windrift Consulting LLC

Research Foundation Review Board

William J. Bernstein, PhD
Efficient Frontier Advisors

Elroy Dimson, PhD
Cambridge Judge Business School

Stephen Figlewski, PhD
New York University

William N. Goetzmann, PhD
Yale School of Management

Elizabeth R. Hilpman
Barlow Partners, Inc.

Paul D. Kaplan, PhD, CFA
Retired – Morningstar, Inc.

Robert E. Kiernan III
Advanced Portfolio Management

Andrew W. Lo, PhD
Massachusetts Institute of Technology

Alan Marcus, PhD
Boston College

Paul O'Connell, PhD
WaterEquity

Krishna Ramaswamy, PhD
University of Pennsylvania

Stephen Sexauer
CIO – San Diego County Employees Retirement Association

Lee R. Thomas, PhD
ReefPoint LLC

ISBN 978-1-952927-43-0

CFA Institute
Research
Foundation

Available online at rpc.cfainstitute.org