1 *Common Elements in Validation of Risk Models Used in Financial Institutions**

DAVID LYNCH, IFTEKHAR HASAN AND AKHTAR SIDDIQUE

Financial institutions use of models has grown dramatically over the past few decades. The use of these models is both absolutely necessary for financial decision making and widely criticized as adding to the complexity of the financial markets. The financial crisis of 2007–2009 brought to light some of the poor modeling choices that had been made by financial institutions. Often they found themselves in terra incognita, where the models acting as their GPS for the financial markets left them stranded. A great distrust in once-trusted financial models developed. Some of the uses of poor models have been documented in news stories. Models were viewed by some as significant contributors to the problems that banks experienced during the Great Recession. Others, notably the Financial Crisis Inquiry Commission, noted the use of models, but laid the blame squarely on the people using them and the decisions they made whether guided by models or not.

Cases of models being used without adequate validation are well documented.[1] In many cases, the decision to use a model with known failings or to use a model without examining its shortcomings have led to disastrous outcomes. The human biases and failings that led to using flawed models called attention to the need for a more systematic approach to model validation.

* The views expressed in this chapter (and other chapters in the book) are those of the authors alone and do not establish supervisory policy, requirements or expectations. They also do not represent those of the Comptroller of the Currency, Board of Governors of the Federal Reserve or Bank of Finland.

[1] Examples include "Recipe for Disaster: The Formula That Killed Wall Street" Wired, Felix Salmon 2009, "Risk Management Lessons from Long Term Capital Management" European Financial Management, Phillippe Jorion, 2006.

https://doi.org/10.1017/9781108608602.002 Published online by Cambridge University Press

Model risk management generally and model validation particularly have gained special attention in its aftermath. The model risk management guidance issued by US regulators (SR 11-7 from the Federal Reserve and Bulletin 2011–12 from the OCC) and European regulators (the EBA issued GL/2014/13 on Supervisory Review and Evaluation Process SREP) outline expectations from the regulators. The guidance contains perspectives on both governance and modeling. Given the regulatory requirements, validation is mandatory for almost all models used at regulated financial institutions. Even for financial institutions outside the regulated banking sector, validation of models is considered quite important.

For example, in 2018, the Securities and Exchange Commission focused on portfolio allocation models used by AEGON USA Investment Management, an asset manager with $106 billion in assets under management. The SEC issued a cease and desist order and imposed nearly $100 million in penalties against AEGON USA Investment Management, Transamerica Asset Management, Transamerica Capital and Transamerica Financial Advisors. SEC complained that the representations regarding the models were misleading because the advisers and broker-dealers "launched the products and strategies without first confirming that the models worked as intended and/or without disclosing any recognized risks associated with using the models."

To the uninitiated, the model validation process often appears to be a disorganized assortment of statistical tests and individual judgements where some aspects of a risk model are challenged and some are not. Many people within a financial organization will have a stake in the model; many of the techniques used by validators are informal and can lead to disputes between the validators and users without any clear criteria for decision making. A few attempts have been made to organize and formalize the processes within financial institutions.

Model validation has been around since the development of models. However, model validation as a formal discipline became more important with the development and increased use of more formal models. In an influential article, Naylor and Finger (1967) identify three schools of thought in approaching validation of models used in business and economics setting. These are rationalism, empiricism and positive economics. The first two schools mainly address the validity of a model's assumptions; the last addresses the validity of a model's predictions.

Rationalism, ultimately beginning with Kant's beliefs in synthetic a priori, views models as logical deductions from synthetic premises that, on their own, are not verifiable. In such an approach, validation mostly consists of examining the validity of those premises and the logical reasoning. The assumptions of a model must be clearly stated, and those assumptions must be readily accepted. Correct logical deductions from those assumptions are acceptable.

This approach seemingly makes model validation a simple process of examining the assumptions and the internal logic of the model; if these are sound the model is valid. However, this approach can take many possible sets of assumptions as valid, with several possible competing models being acceptable based on an examination of their assumptions and logic.

Under a rationalist philosophical approach, then validation is a semiformal, conversational process. A valid model is assumed to be only one of many possible ways of describing a real situation. No particular representation is superior to all others in any absolute sense, although one could prove to be more effective. No model can claim absolute objectivity, for every model carries in it the modeler's world view. In this approach, model validation is a gradual process of building confidence in the usefulness of a model; validity cannot reveal itself mechanically as a result of some formal algorithms. Validation is a matter of social conversation, because establishing model usefulness is a conversational matter.

An alternative approach to the rationalist approach is logical empiricism. In Naylor and Finger's (1967) succinct summation, validation begins with facts rather than assumptions. Observations are viewed as the primary source of truth. Empiricists regard empirical science, and not mathematics, as the ideal form of knowledge. In the words of Reichenbach (1951): "They insist that sense observation is the primary source and the ultimate judge of knowledge, and that it is self-deception to believe the human mind to have direct access to any kind of truth other than that of empty logical relations." In this view every assumption should be validated by empirical observation. More broadly, outcomes of a model are also to be validated by empirical observation.

With a logical empiricist approach then, validation is seen as a formal and "confrontational" process, in the sense that the model is confronted with empirical data. Since the model is assumed to be an

objective and absolute representation of the real system, it can be either true or false. And given that the validator uses the proper validation algorithms, once the model confronts the empirical facts, its truth (or falsehood) is automatically revealed. Validity becomes a matter of formal accuracy rather than practical use.

A third school of thought on validation is based on the ability of the model to predict the behavior of the dependent variables that are treated by the model. The "Positive Economics" view is most widely associated with Milton Friedman (1953) where a model cannot be tested by comparing its assumptions with reality. A model is judged by whether its predictions are good enough for use or better than predictions from alternative models. The realism of the model's assumptions does not matter, only the accuracy of the predictions.

Finger and Naylor synthesize these three schools of thought into a single multi-stage validation process. The first stage is largely in the rationalist tradition, it involves clearly spelling out the model's assumptions and comparing them to theory, casual observations, and general knowledge. The second stage is in the empiricist tradition and involves empirically testing the model's assumptions where possible. The third stage, in the positive economics tradition, involves comparing the output of the model to the real system.

In the United States, model validation for the very complex and large models deployed in the energy domain was an important concern and several conferences organized in the late 1970s and early 1980s under the auspices of the Department of Energy and national labs were important in how model validation developed as a discipline.[2] Model validation since that time has involved the development of techniques to demonstrate the model's validity, a relaxation of the three-step approach, and an emphasis on data validity. Sargent (2011) provides an overview.

Following the three-step approach, the validation of assumptions of a model presents some difficulties. Some assumptions are to be validated on general knowledge and theory, while others are to be empirically validated. How should the model validator decide when to validate based on theory and general knowledge and when to validate based on empirical observation? Naylor and Finger provide one approach where the assumptions should be empirically tested whenever possible. This seemingly straightforward advice may prove difficult when the number

[2] See National bureau of standards (1980) for example.

of assumptions used in a complex simulation is very large or there are many models to be validated. Empirical testing is time-consuming and it may not be feasible to empirically test all assumptions.

Jarrow (2011) provides an approach to selecting which assumptions to test. He divides the assumptions into critical assumptions and robust assumptions. Conclusions of the model are not sensitive to the robust assumptions, where the implications of the model only change slightly if the assumption changes slightly. The critical assumptions are those where the implications of the model change dramatically if one of the critical assumptions change slightly. Jarrow provides the basis for a model validation process. In order to validate a model, a financial institution should (1) test all implications, (2) test all critical assumptions, (3) test all observable robust assumptions and (4) believe all non-observable robust assumptions. The model should be rejected unless all these four conditions are met.

For financial institutions, the regulatory model guidance builds on the validation approaches described above. It states "model validation is the set of processes and activities intended to verify that models are performing as expected, in line with their design objectives and business uses. Effective validation helps ensure that models are sound. It also identifies potential limitations and assumptions, and assesses their possible impact. As with other aspects of effective challenge, model validation must be performed by staff with appropriate incentives, competence, and influence." The guidance identifies three core elements;

1. evaluation of conceptual soundness including developmental evidence;
2. ongoing monitoring, including process verification and benchmarking;
3. outcomes analysis including backtesting.

The first element blends the first two steps of Naylor and Finger's three-step approach and avoids being definitive about which assumptions should be confronted with empirical facts and which assumptions are to be tested more conversationally based on theory and general knowledge. However, Jarrow's approach can provide some guidance here.

The second element, ongoing monitoring, can be viewed either in a rationalist sense or in an empirical sense. A rationalist view of ongoing monitoring is that it should be considered as a search for models that

are more useful over time. Since no model is objectively the truth one should consider it a search for better assumptions and premises that provide more useful models over time. The empiricist view is that ongoing monitoring provides a set of comparisons of the model to measured variables. These comparisons provide tests of the model that allow it to be accepted or rejected.

The third element blends in the view of positive economics. How well the model describes reality as measured by its predictions takes center stage. A model that does not predict reality is not easily accepted.

It is important to note that how validation of risk management models proceeds across the various parts of a large financial institution also depends on how risk is viewed and measured in the disparate parts of the organization. Risk model validation is also affected by the development of the discipline over time. The distinct elements of how risk is perceived in different parts of the financial institution drive how risk is measured.

In general, the approaches to model validation depend on how risk is perceived. For certain risk types such as credit risk, the riskier outcome of a default is directly observable. In contrast, volatility of returns or other measures of dispersion are not easily observable. As such there are a variety of models that attempt to convert the observable outcomes into the risk measures. Models that measure portfolio risk, such as volatility and Value-at-Risk, are of the second variety.

In spite of disparate approaches taken for the risk measured for a given type of model, the underlying approaches tend to share similarities as well as suffering from similar shortcomings. Nearly all risk model validation approaches contain the three elements, albeit the relative importance of each of the elements may vary.

The model validation function is tasked with applying these core elements to each model used by a financial institution. Practitioners seem stuck using the same validation techniques that they have been using since the 1990s. At the same time, the academic literature on forecast evaluation has developed greatly over the last decades.[3] This literature provides insight on the appropriate methods to performing

[3] See the forecast evaluation chapters in the Handbook of Economic Forecasting Vol 1. (2006) Graham Elliott, Clive W. J. Granger and Alan Timmerman Eds. Elsevier.

https://doi.org/10.1017/9781108608602.002 Published online by Cambridge University Press

backtesting, the comparison of model predictions with actual outcomes, and benchmarking the comparison of two models. Many of these techniques have begun to be applied to risk models by academics, but not necessarily by practitioners. In particular, forecast evaluation techniques and forecast encompassing tests developed primarily for point forecasts in the late 1990s have been adapted to cover the models used in risk management. These models include, quantile estimates like Value-at-Risk models, volatility forecasts, probability density forecasts and probability estimates of discrete events such as corporate defaults. This chapter provides a brief description of forecast evaluation and forecast encompassing, using Mincer-Zarnowitz regressions as a point of departure, showing how these statistical techniques and variants can be applied to the types of models used in risk management in large financial institutions to validate models.

1.1 Mincer-Zarnowitz Regressions

The evaluation of forecasts through regression-based techniques generally starts with Mincer and Zarnowitz (1969). The original Mincer-Zarnowitz approach is derived from the properties of optimal forecasts. The literature has established the following important properties of optimal forecasts:

(1) they are unbiased;
(2) optimal one-step-ahead forecast error is white noise and unforecastable;
(3) optimal h-step-ahead forecast errors are correlated, but at most an MA(h-1);
(4) variance of optimal forecasts increases with the forecast horizon;
(5) forecast errors should be unforecastable from all information available at the time of the forecast.

Based on these properties, the regression of the actual value on the *ex ante* forecast should have a zero intercept and a coefficient of one if the *ex ante* forecast is optimal. If the coefficients are different, it indicates systematic bias in the historical forecasts.

The procedure is fairly simple. Estimate the simple regression

$$y_{n+h} = \alpha + \beta y_{n+h|n} + e_{n+h|n}$$

https://doi.org/10.1017/9781108608602.002 Published online by Cambridge University Press

That is regress, at time $n + h$, the forecasted value of y, h time periods ahead at time n using the information available at time n on the actual value of y. Then test the joint hypothesis $\alpha = 0, \text{and } \beta = 1$. If the hypothesis is rejected the forecast can be improved by adjusting the forecast using the linear equation just estimated by the Mincer-Zarnowitz regression itself to get a better forecast.

Running a Mincer-Zarnowitz regression is one way to meet the third element of the model validation guidance for US banking regulators described above. It is a form of outcomes analysis that can be applied whenever a financial institution's model generates a forecast. If one took the positive economics philosophy to an extreme, this type of testing is all that would be necessary; only the forecast matters and the realism of the model's assumptions do not.

However, there are two other elements to the validation of a model contained in US regulatory guidance. It would seem that one would have to look elsewhere to address the other elements of this guidance since at its core the Mincer-Zarnowitz regression is really a test of the models forecast. Modifications of the Mincer-Zarnowitz regression can provide insight into the other two elements contained in the regulatory guidance. While these modified regressions can help address the other two elements of the model validation guidance, it should be noted that fully addressing these elements would usually include more than just these tests.

The first element is an evaluation of the conceptual soundness of the model. As described above, this typically involves an evaluation of the assumptions of the model. A good forecast should incorporate all useful information. If there is additional information available at the time the forecast is made that could improve the forecast, then that information should be used in making the forecast. Implicitly the assumption is that this other information does not affect the forecast. One can test whether other variables are useful in making a forecast by augmenting the regression model with additional auxiliary exogeneous variables.

$$y_{n+h} = \alpha + \beta_1 y_{n+h|n} + \beta_2 x_{2|n} + \cdots + \beta_k x_{k|n} + e_{n+h|n}$$

Each of these additional variables should have no effect on the model. In this case the hypothesis to be tested is that $\alpha = 0$, $\beta_1 = 1$, $\beta_2 = 0$, $\beta_3 = 0$, $\cdots \beta_k = 0$. If some of the auxiliary

https://doi.org/10.1017/9781108608602.002 Published online by Cambridge University Press

exogenous variables are important, then the model can be improved by using those variables in the model producing the forecast. Furthermore, lagged variables and transformations of variables can be used as auxiliary regressors to detect whether persistence or nonlinearities are important elements of the model.

The second element of the regulatory guidance includes benchmarking of the model. This generally includes a comparison of the forecast to an alternative forecast. The basic Mincer-Zarnowitz regression only includes one forecast on the right-hand side, but it would be fairly easy to expand this to include a second forecast or even more,

$$y_{n+h} = \alpha + \beta_1 y^1_{n+h|n} + \beta_2 y^2_{n+h|n} + e_{n+h|n}.$$

The superscripts 1 and 2 now refer to the two different forecasts that one is comparing in the regression. Most typically, the forecasts are unbiased so that $\alpha = 0$, and the coefficients on the forecasts are constrained to sum to 1. If we test the hypothesis that $\beta_1 = 1$ and $\beta_2 = 0$ and this hypothesis is not rejected, then we say that forecast 1 encompasses forecast 2. The second forecast adds nothing to the first forecast in its ability to make the prediction. Similarly, we can test whether forecast 2 encompasses forecast 1. Alternatively, neither forecast may encompass the other and both forecasts have important information not included in the other forecast. The encompassing regression can be seen as part of the conversational or rationalist approach to model validation where all models have validity and one is searching for more useful models over time.

The use of all these variants of the Mincer-Zarnowitz regressions can be part of a process of model improvement over time. The regressions and tests are run frequently, and the tests are used as indicators that the models need to be updated or changed. In our experience, when these types of tests are run, they are most frequently done when models are first implemented, and rarely done thereafter. As part of meeting the second element of regulators guidance on model validation, the tests should be run on a regular basis to perform ongoing monitoring of the model.

It is also important to note that the implied loss function in the original Mincer-Zarnowitz regression is a mean-squared error function. Subsequent literature has pointed out that the loss function in many forecasting contexts may not be a mean squared error function and may be asymmetric. For example, Patton and Timmerman (2006)

show that the Federal Reserve's loss function on forecasts of GDP may be asymmetric. In such cases the Mincer-Zarnowitz framework has to be adjusted to accommodate the loss function that is actually used.

It is important to evaluate a model based on the loss function of the user. In many cases the loss from making an error is not well represented by a mean squared error model. In many contexts an underprediction can be more costly than an overprediction or vice versa. In these cases, the evaluation should be changed to take into account the actual form of the loss function. Elliott and Timmermann (2016) provide a thorough discussion of evaluating forecasts based on different loss functions.

Importantly, many models are estimated by minimizing something other than mean squared error in contradiction to what is assumed when using the Mincer-Zarnowitz regression to evaluate a forecast. Notably, quantile estimates, widely used for risk management, use a "check" loss function.

$$L = \tau \cdot \max(e, 0) + (1 - \tau) \max(-e, 0),$$

where τ is the quantile of interest and e is the error. Gaglione et al. (2011) provide a method for evaluating quantile estimates using quantile regression analogous to a Mincer-Zarnowitz regression. Giacomini and Komunjer (2005) provide a method for comparing two quantile forecasts and performing encompassing tests. Lopez (1999) suggests that regulators do not have a loss function for VaR models, which is a quantile estimate, based on the check function. He proposes that regulators are more concerned about losses that exceed the regulatory VaR than they are about losses within the regulatory VaR and proposes a regulatory loss function that reflects this. This provides some basis to the claim by Perignon, Deng and Wang (2008) that banks overstate their VaR models, at least during normal market times. Gordy and McNeil (2020) provide an approach to backtesting that can reflect these differences in loss functions.

In the cases of probability forecasts, often used to estimate probabilities of default, the mean squared error loss function is modified to the quadratic probability score but other loss functions surface in the literature, the logarithmic probability score and spherical probability score are examples. Clements and Harvey (2010) provide an overview of forecast encompassing tests for probability forecasts using quadratic and logarithmic probability scores.

Volatility estimates play a large role in risk management and portfolio allocation applications. In this case, the use of Mincer-Zarnowitz regressions is complicated because volatility is not directly observable. The use of proxies, such as squared returns, for volatility introduces additional noise. This means that fits to a Mincer-Zarnowitz regression are very low, which makes evaluation and encompassing tests very difficult to perform without large amounts of frequently observed data.

There are many other aspects to evaluating forecasts that are not covered here. This overview provides some insight how this econometric tool can be adapted to assist in the validation of models not just to meet regulatory requirements, but to actively seek to improve models. Even as these approaches to validation are coming into use, machine learning and artificial intelligence have brought to the fore some other aspects of model validation that are important to consider.

The advent of machine learning models and big data has raised several new validation issues related to model use in the financial industry. An important topic is the ability to explain how these large scale models arrive at their outcome. Traditional linear regression models provide a whole host of model diagnostics that allow what Efron (2020) calls attribution or significance testing. These significance tests allow an explanation of how the model arrives at a decision. Machine learning algorithms have generally sacrificed attribution to provide better predictions.

Breiman (2001) echoing a positive economics view, has taken the view that statistical modeling should start with prediction as the goal rather than a theoretic description of the data generating process. Many of the new algorithmic methods do not lend themselves to easy explanation, nonetheless good explanations need to start with good predictions. The debate seems to mirror much of the philosophical debate that opened this chapter.

Machine learning is beginning to deal with the need for explanation of the results of their prediction algorithms. The National Institute of Standards and Technology (2020) has issued a draft setting the principles for explainable Artificial Intelligence for public comment. They note that Artificial intelligence is becoming part of high stakes decision processes, and that laws and regulations in areas where these models are used require information be provided about the logic of how the decision was made and that explanations would also make artificial intelligence applications more trustworthy.

The NIST draft sets forth four principles for explainable AI. The first, characterized as the explanation principle, requires that a model delivers accompanying evidence or reasons for all outputs of the model. The second requires that the explanations be meaningful, that the explanation is understandable to individual users. The third principle requires that the explanation correctly reflects the systems process for generating the output. The last is to recognize the limits of the system so that the system only operates when there is sufficient confidence in its output. The draft notes that explanations may vary depending on the consumer. It is too much to expect a loan applicant to find the explanation that is useful to a model developer satisfying.

It is typical to consider linear regression, logistic regression and decision trees to be self-explainable. The attribution process is well understood and provides suitable explanations for why decisions are made. On the other hand, bagging, boosting, forests and neural networks are examples of modeling techniques that need further explanation and are referred to as black-box algorithms.

Certainly, there are approaches being developed to provide explanations for the black-box models that are comparable to what is provided for the self-explainable models. These generally come in two categories, global explanations and local explanations.

Global explanations are themselves models that explain the algorithm by using the black-box model to build the explanation. One global explainable algorithm is SHAP (Lundberg and Lee, 2017) based on the Shapley value from cooperative game theory. The importance of a feature is determined by its Shapley value. Additionally, partial dependence plots are often used to describe feature importance. See Friedman (2001) or Hastie, Tibshirani and Friedman (2010) for a description of partial dependence plots.

Local explanations are explanations of each decision made by a black-box model. The explanation does not need to generalize to other models. LIME, or Local Interpretable Model Agnostic Explainer, uses nearby points to build a self-explainable model. The self-explainable model is then used to provide the explanation for the black-box model.

Alternatively, and probably more applicable to explaining decisions to the user, is the use of counterfactuals (Wachter et al., 2017). In this approach the explanation provided is what inputs would have to change, and by how much, to change the outcome of the black-box model. In general, the approach is to provide the minimum change

(according to some distance metric) that would change the output of the model. Thus, a loan applicant may be told that if their income was $10,000 more they would have received the loan.

While the notion of explainable models is important, the widespread use of machine learning models has raised questions of the ethical consequences of using these models to make automated decisions. In other words, if machine learning models are to be used to make consequential decisions, those decisions should be fair for different sociological groups. US courts have described two types of bias. Intended or direct discrimination would be a case of disparate treatment. A protected group is treated differently than other groups in the model. Unintentional bias occurs when a decision process has disparate impact for different groups regardless of intent. Both disparate treatment and disparate impact are illegal in the USA. See Feldman et al. (2015) for a more thorough discussion of disparate impact and the distinction between its legal and statistical description. A model could also be validated as being fair to different groups.

In the context of machine learning models, fairness has become an increasingly important issue. Kusner et al. (2018) provide a short list of definitions of fairness with respect to the treatment of protected attributes such as race or gender. Importantly we can think of several ways of making a machine learning model outcome fair. One approach is to ignore the protected attribute and not use it within the model. While this avoids disparate treatment, unintentional bias and disparate impact can still affect the treatment of individuals with different sociological attributes through variables correlated with a protected attribute.

To test disparate impact is more difficult. A naive approach may link this to the counterfactual explanation. What would happen if the protected attribute of a subject changed? Would the outcome of the model change? If not, one could view this as evidence of a lack of disparate impact. Once again, explanatory variables used in the model may serve as a proxy for a protected attribute, and changing the protected attribute does not reflect the lack of change in these underlying proxies.

Increasingly, fairness is becoming closely related to causality. Zhao and Hastie (2021) describe how causality in machine learning models is closely related to the partial dependence plots. Kusner et al. (2018) and Kilbertus et al. (2018) describe how fairness can be determined in

machine learning models by a protected attribute not causing a decision in the sense of Pearl (2009). Examples are provided in both papers to describe the causal reasoning that would test for fairness in a model. In particular, they provide methods to test fairness that would allow for consideration that other variables besides the protected variable itself may be correlated with the protected variable. These have great promise in advancing the validation of the fairness of a model.

Models are being deployed in more contexts by financial institutions, and model outcomes are becoming more consequential and subject to increase scrutiny. For these reasons model validation continues to be of great importance. With the introduction of models into new areas, the task of model validation has been expanded to include more topics under the validation umbrella.

The rest of this book provides examples of validation procedures in different contexts. Since validation is difficult to separate from model development, of necessity the chapters also describe the types of models in use. As such the chapters in the book can serve as primers on the types of models that are used in the disparate risk areas such as market risk, retail credit risk, wholesale credit risk, operational risk, etc.

The book begins with several chapters on the validation of market risk models. Since the Basel Committee on Banking Supervision (1996) included backtesting as a requirement for the use of internal VaR models, the procedures have received considerable scrutiny and the models have been tested frequently. This provides a lot of experience in model validation over the years. These chapters also describe and make use of the regulatory backtesting data that has been collected by US regulators. The second part of the book consists of chapters that cover validation of lending models. The last part of the book considers difficult-to-validate models, those with long time horizons and infrequent or proxy observations, and non-traditional models.

In Chapter 2 David Lynch provides a description of what applying the full scope of tests implied by the regulatory guidance would entail for VaR modeling of trading activities. The chapter also shows how banks models fare under some tests that are implied by the regulatory guidance. Different loss functions are considered explicitly when VaR models are evaluated and compared to alternative models. Using data from 2013 to 2016, the authors find that the average exceedance rate is 0.4 percent for twenty holding companies, though there are individual

banks that are as high as 2.1 percent. Thus, the results document the conservative nature of regulatory VaR models used by banks.

Chapter 3 provides a framework for adapting VaR backtesting to provide insight into backtesting failures. Conditioning backtests on circumstances of interest provides greater insight into VaR model performance. Examples of how this technique can be used are provided. It provides the techniques for getting continuous feedback on model performance under specific conditioning variables such as historical price variation, specific risk, concentration, etc. The framework introduced here can be used for other types of backtesting as well.

Chapter 4 provides an overview of testing of VaR models using probability integral transforms following Berkowitz (2001) at the trading desk level. Instead of testing just the ninety-ninth percentile, this approach tests the fit of the whole distribution of P&L. Several statistical tests for the fit of the distribution are used and compared. This provides new insight into the modeling practices at large bank holding companies. The authors find that pure exceedance-based tests can often fail to find the more nuanced model misspecifications that are uncovered via the use of probability integral transforms.

Chapter 5 provides a new test of the distribution of P&L based on empirical likelihood methods and it applies it to desk level results. The results are compared to the results of more conventional tests as well. This is an alternative to both the traditional exceedance-based tests as well as the probability integral transform-based tests used in Chapter 4. This chapter finds that relative entropy-based tests are often more discerning compared to the Anderson–Darling tests and probability integral transform (PIT) tests. Thus, empirical likelihood methods can ameliorate some of the conservatism inherent in the other metrics.

Chapter 6 reviews the performance of Bank holding company market risk models during the COVID crisis beginning March 2020. The authors first document that backtesting exceptions can predict future backtesting exceptions. Additionally, the predictability of backtesting exceptions increased during the COVID crisis. It notes that the VaR models did not capture the increase in market volatility over the crisis period particularly well; these results were previously documented in Berkowitz and O'Brien (2001) and Szerszen and O'Brien (2017). The chapter augments the typical backtesting tests with market

risk factors to provide insight into the sources of backtesting failures. They find that no single market factor appeared to drive backtesting exceptions, but rather several different factors mattered. Taken together, these results show that backtesting exceptions are predictable from both recent exceptions and from recent market volatility, indicating that banks' VaR models did not react quickly enough to changes in market conditions.

Chapter 7 is on model validation in the context of stress testing, an area that is hampered because of the paucity of observations of true stress results. Klagge and Lopez provide techniques for monitoring the performance of models despite the lack of concrete stress outcomes and they provide methods to show that the models are working as intended.

Chapter 8, from Eduardo Canabarro, provides an overview of the validation of counterparty credit risk models along with how to manage them. His chapter provides the reader with the understanding of both default risk and credit valuation adjustment (CVA) risk in the context of counterparty credit risk. Canabarro provides clear explanations of the technical concepts used in this domain. He then traces the historical evolution of counterparty credit risk from its beginnings as well as how both the industry and regulators have reacted to the historical loss events such as the 1998 Russia/LTCM Crises and the 2008 Great Recession stemming from counterparty credit risk. He discusses the complex models that are deployed by the major institutions along with the shortcomings in these approaches.

Chapter 9, from Feng Li and SangSub Lee, provides a review of retail credit models and model components. The wide variety of models provide challenges for validating retail models. The chapter highlights issues regarding data techniques and data sampling issues in retail credit validation in particular. They begin with a primer on the various ways that retail credit risk models can be categorized. These include static credit and behavioral scoring models and various multi-period loss forecasting models. Then they provide details on aggregate or segmented pool-level modeling approaches such as roll-rate models versus loan-level models. Then they provide thorough descriptions of the components of the loan-level models such as the probability of default (PD) models and the link to survival models, loss given default (LGD) models and exposures. They describe modeling and validation issues arising in the various types of models in great detail as well as

how to ameliorate them. As an example, they discuss in detail the use of the landmarking approach to estimate the effect of time-varying covariates.

Chapter 10 discusses the validation of the wholesale model and provides the comparative advantages and disadvantages for the various methods involved. The authors rely primarily on the expected loss approach that are commonly used at large financial institutions for both regulatory and internal risk management purposes. Given that the largest banking institutions use their obligor and facility internal risk ratings in PD and LGD quantification for Basel and also for quantification of stressed PD and LGD for CCAR and DFAST, they also address issues that arise in the validation of internal ratings systems used by banking institutions for grading wholesale loans.

Chapter 11 presents some case studies in the context of validation of wholesale models. These are of value in not only wholesale credit risk models but in other risk "stripes" as well. These describe validation for use, that is ensuring that models are validated for the use they are put to. They explain issues that can arise if models that have been developed and validated for a different purpose are repurposed. Then the case studies describe how to conduct validation of data and the distinction between internal and external data that arises in this context. The next step is validation of assumptions and methodologies. Validation of model performance is covered next with techniques such as backtesting, outcomes analysis and the use of benchmark models. The next step describes the model validation report and how to structure it.

Chapter 12 provides an overview of the issues that are encountered in the validation of models for allowance for credit losses, i.e. what is referred to as Allowance for Loan and Lease Losses. What makes them distinct is that these models typically rely on more fundamental credit loss models covered in Chapters 9 through 11. At the same time, there can be significant adaptations of the more primitive credit loss models to meet the requirements to estimate the reserves (allowances) and validation needs to take these adaptations into account. The author also compares and contrasts with CECL models. The author provides observations that would likely be quite valuable for those involved in these validations.

Chapter 13 considers the validation of operational risk models, in particular the loss distribution approach used for capital models and

the regression models that are often used for stress testing. They explore the challenges associated with these models, such as the small historical length of operational risk datasets, the fat-tailed nature of operational losses, and the difficulties assigning dates to operational loss events. They propose possible approaches for robust regression modeling, such as the use of external data. They then discuss back-testing and benchmarking of operational risk models and present practical examples of benchmarks used by US regulators to benchmark operational risk models.

Chapter 14 presents a framework grounded in statistical decision theory to assess model adequacy through utility functions, offering an additional approach to supplement common model assessment criteria. Model users rely on standard measures of statistical goodness-of-fit such as AIC to evaluate models, but selecting (and subsequently validating) a model may not be straightforward if, for example, the differences in comparative metrics are marginal as this can amplify uncertainty in a model choice.

Chapter 15 provides an overview of the validation of enterprise level economic capital model. It provides an evaluation of the model against actual outcomes, despite the lack of data for models of this type. The model is evaluated under different loss functions. The chapter relies on all three methods of aggregation that are commonly practiced, namely 1) variance–covariance approach, 2) copula approach and 3) scenario-based aggregation. The authors propose an empirical statistical framework to test the performance of alternative benchmark models for economic capital estimation. They compare different copula functions and find that the T4 copula performs better than other copulas with the hypothetical bank holding company data used.

Chapter 16 provides a view on validation of interest rate risk models that is entirely in the rationalist tradition. The outcomes are largely unobservable and there is little in the way to choose among ways to model the sensitivity of deposits and other products to interest rates. The approach is then to confirm that the products are modeled correctly in the chosen framework.

Chapter 17 provides a discussion on the validation of asset management models, where the output from a model interacts quite heavily with expert judgment. These include portfolio allocation models. This is primarily descriptive rather than an empirical illustration.

References

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management, *Journal of Business and Economic Statistics*, 19(4), 465–474.

Berkowitz, J. and O'Brien, J. (2002). How accurate Are Value-at-Risk models at commercial Banks? *The Journal of Finance*, 57(3), 1093–1111.

Breiman, L. (2001). Statistical modeling: The two cultures, *Statistical Science*, 16, 199–231.

Carson, E.R. and Flood, R.L. (1990). Model validation: Philosophy, methodology and examples, *Transactions of the Institute of Measurement and Control*, 12(4), 178–185. DOI:10.1177/014233129001200404.

Efron B. (2020). Prediction, estimation, and attribution, *Journal of the American Statistical Association*, 115(530), 636–655, DOI: 10.1080/01621459.2020.1762613.

Elliot, G., Granger Clive W. J. and Timmerman Alan. Eds. (2006). *Handbook of Economic Forecasting*, Vol. 1. Elsevier.

Elliott, G. and Timmermann, A. (2016). *Economic Forecasting*. Princeton University Press.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine, *Annals of Statistics*, 29, 1189–1232.

Gaglianone,W. P., Lima, L. R., Linton, O. and Smith, D. R. (2011). Evaluating Value-at-Risk models via quantile regression. *Journal of Business and Economic Statistics*, 29(1), 150–160.

Giacomini, R., Komunjer, I., (2005). Evaluation and combination of conditional quantile forecasts. *Journal of Business and Economic Statistics*, 23, 416–431.

Gordy, M. and McNeil, A. (2020). Spectral backtests of forecast distributions with application to risk management, *Journal of Banking and Finance*, July, Volume 116. Available at: https://doi.org/10.1016/j.jbankfin.2020.105817.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *Elements of Statistical Learning*, New York: Springer.

Jacobs, Michael and Inanoglu, Hulusi. (2009). Models for risk aggregation and sensitivity analysis: An application to bank economic capital, *Journal of Risk and Financial Management*, 2(1), 118–189.

Jarrow R. (2011). Risk management Models: Construction, *Testing, Usage Journal of Derivatives*, 18(4), 89–98.

Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D. and Scholkopf, B. (2017). Avoiding discrimination through causal reasoning, in *Advances in Neural Information Processing Systems*, 656–666.

Kusner, M. J., Loftus, J., Russell, C. and Silva, R. (2017). Counterfactual fairness, in *Advances in Neural Information Processing Systems*, 4066–4076.

Lopez, A. J. (1999). Regulatory evaluation of Value-at-Risk models. *Journal of Risk*, 1, 37–64.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., eds., *Advances in Neural Information Processing Systems*, 30, 4765–4774. Curran Associates, Inc.

Mincer, J. and Zarnowitz V. (1969). The evaluation of economic forecasts. In J. Mincer (ed.) *Economic Forecasts and Expectations*. National Bureau of Economic Research, New York.

Naylor, Thomas H. and Finger, J. M. (1967). Verification of computer simulation models, *Management Science* , Oct., Vol. 14, B92–B106.

O'Brien J. and Szerszen P. (2017). An evaluation of bank measures for market risk before during and after the financial crisis, *Journal of Banking & Finance*, 80, July, 215–234.

Patton, A. and Timmermann, A. *Testing forecast optimality under unknown loss, working paper*, University of California San Diego, 2006.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press.

Pérignon, C., Deng, Z. Y. and Wang, Z. Y. (2008). Do banks overstate their Value-at-Risk? *Journal of Banking and Finance*, 32, 783–794.

Phillips, Jonathon P., Hahn, Carina A., Fontana, Peter C., Broniatowski, David A. and Przybocki, Mark A. (2020). *Four Principles of Explainable Artificial Intelligence* National Institute of Standards and Technology, Draft.

Popper, Karl, *The Logic of Scientific Discovery*, Julius Springer, Hutchinson & Co, 1959.

Reichenbach, Hans, *The Rise of Scientific Philosophy*, University of California Press, 1951.

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016). Why should I trust you?" Explaining the predictions of any classifier. In *KDD 2016: Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA. ACM.

Sargent, Robert G. (2011). Verification and validation of simulation models, Proceedings of the 2011 Winter Simulation Conference.

Validation and assessment of energy models: Proceedings of a symposium held at the National Bureau of Standards, Gaithersburg, MD, May 19–21, 1980.

Wachter, S., Mittelstadt, B. and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law and Technology*, 31, 841.

Zhao, Qingyuan and Hastie, Trevor. (2021). Causal interpretations of black-box models, *Journal of Business & Economic Statistics*, 39(1), 272–281, DOI:10.1080/07350015.2019.1624293.