Performance Monitoring: Goals and Techniques

Nick Klagge (Federal Reserve Board)
Jose A. Lopez (Federal Reserve Bank of SF)

Fourth Annual Stress Test Modeling Symposium
June 25-26, 2015
Federal Reserve Bank of Boston

Note that the views expressed here are those of the authors and not necessarily of the Board of Governors of the Federal Reserve System or the Federal Reserve Bank of San Francisco.

What is performance monitoring?

SR 11-7 talks about two venues for assessing model performance:

- **Developmental testing**: "…in which the various components of a model and its overall functioning are evaluated to determine whether the model is performing as intended." (6)
- **Ongoing monitoring**: "…confirms that the model is appropriately implemented and is being used and is performing as intended." (12)

In both cases, among other things, we want to see if the model is **performing as intended**.

We will refer to this activity as "**performance monitoring**"

- May encompass hypothesis "testing" but also other kinds of analysis.

Objectives

Develop a structured framework for thinking about performance monitoring for stress testing models:

- Provide a common language to facilitate communication between developers and validators
- Assist in developing performance monitoring plans
- Organize presentation of results
- Consider how performance monitoring may differ for stress testing vs. unconditional models

Give indicative examples of some classes of performance monitoring

- How do existing analyses fit into this framework?
- What types of analysis might fill in gaps?

Black-box perspective

From software development, two classes of testing:

- Black-box: Look at what happens to outputs given different inputs, without looking at the inner workings. For example, sensitivity analysis.
- White-box: Drill down to the inner workings to see what is going on. For example, looking at statistical significance of estimated coefficients.

Both are useful and can complement one another. However, we think performance monitoring lends itself to a black-box perspective.

- Can be applied to a wide variety of models
- Relatively easy to automate
- If a problem appears, can drill down into the white box to diagnose

Simple notation

Consider the following notation for a model:

$f_m(P,S,t;\theta_m)=Y$

- f_m is the model in question
- the vector Y is the model output (ex., 9-quarter projection)
- the vector P represents information on the portfolio or item being stress-tested
- the vector S represents information on the stress-test scenario in question
- the scalar t represents the valuation date or "starting point" of the stress test
- the vector θ_m represents the numerical parameters of f_m

Scenario Sensitivity Analysis

This class of performance monitoring tools measures the response of a model's output to a change in the scenario:

$$f_m(P,S_k,t;\theta_m) \text{ vs. } f_m(P,S_l,t;\theta_m)$$

This diagnostic is important since it tracks changes in model outcome as the macroeconomic scenarios change.

Examples:

- Scatterplots of the outputs across scenarios.
- Regression of more stressed outputs against less stressed or baseline outputs
  - a formalization of the scatterplot diagram
  - expect a slope greater than one
- Spearman rank correlation across two different scenarios.
- Look at change in output relative to change in variable
  - ex., change in HPI for a mortgage loss model

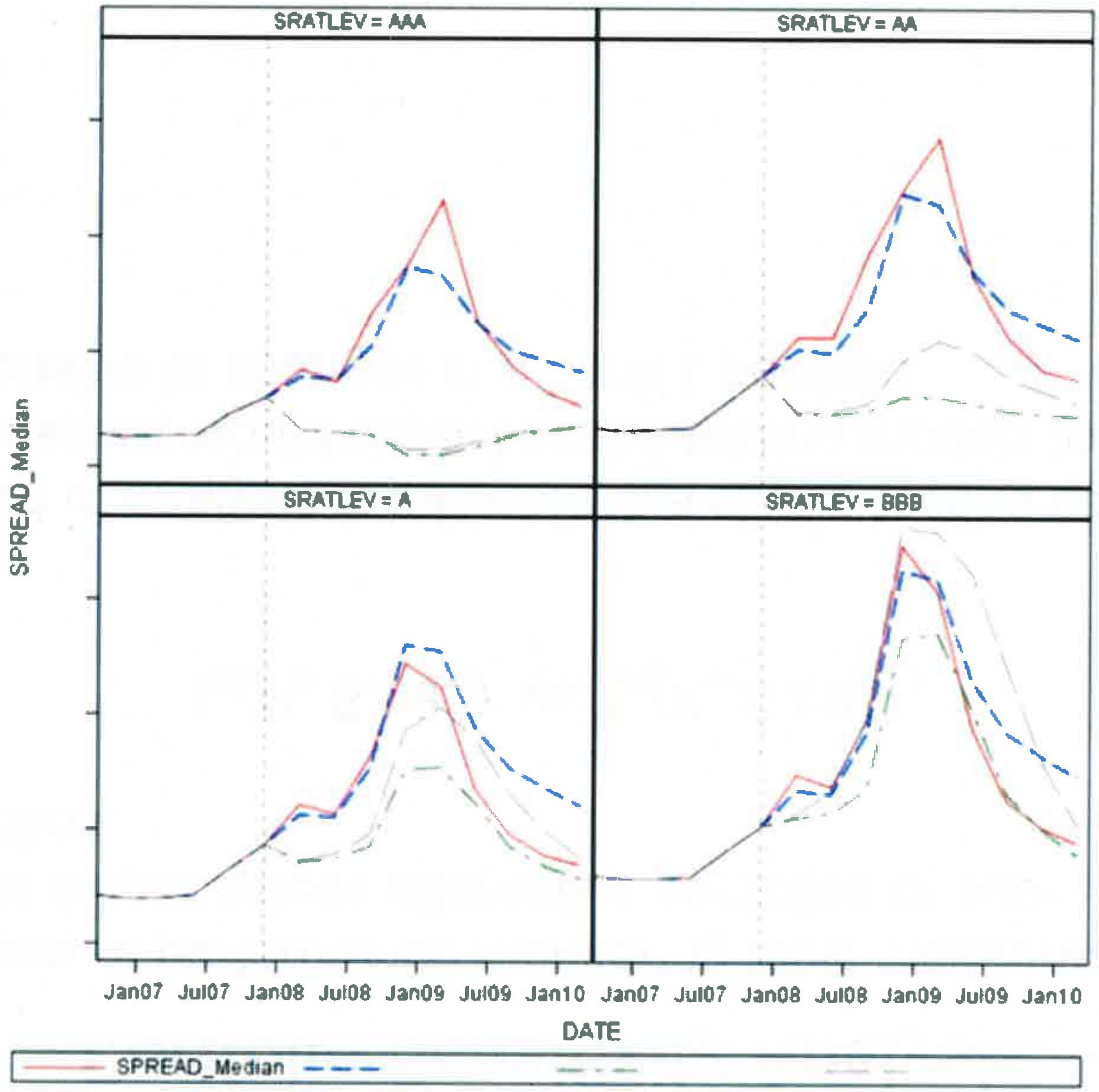
Portfolio Sensitivity Analysis

This class of performance monitoring tools measures the model output across alternative portfolios or sub-portfolios:

$$f_m(P_p, S, t; \theta_m) \text{ vs. } f_m(P_q, S, t; \theta_m)$$

This sort of analysis is valuable in a stress-testing context because we would expect stress-testing models to be responsive to changes in a bank's portfolio.

Portfolio Sensitivity Analysis - Example



Parameter Sensitivity Analysis

This class of performance monitoring tools measures the model output across alternative parameter values:

$$f_m(P,S,t;\theta_{m1}) \text{ vs. } f_m(P,S,t;\theta_{m2})$$

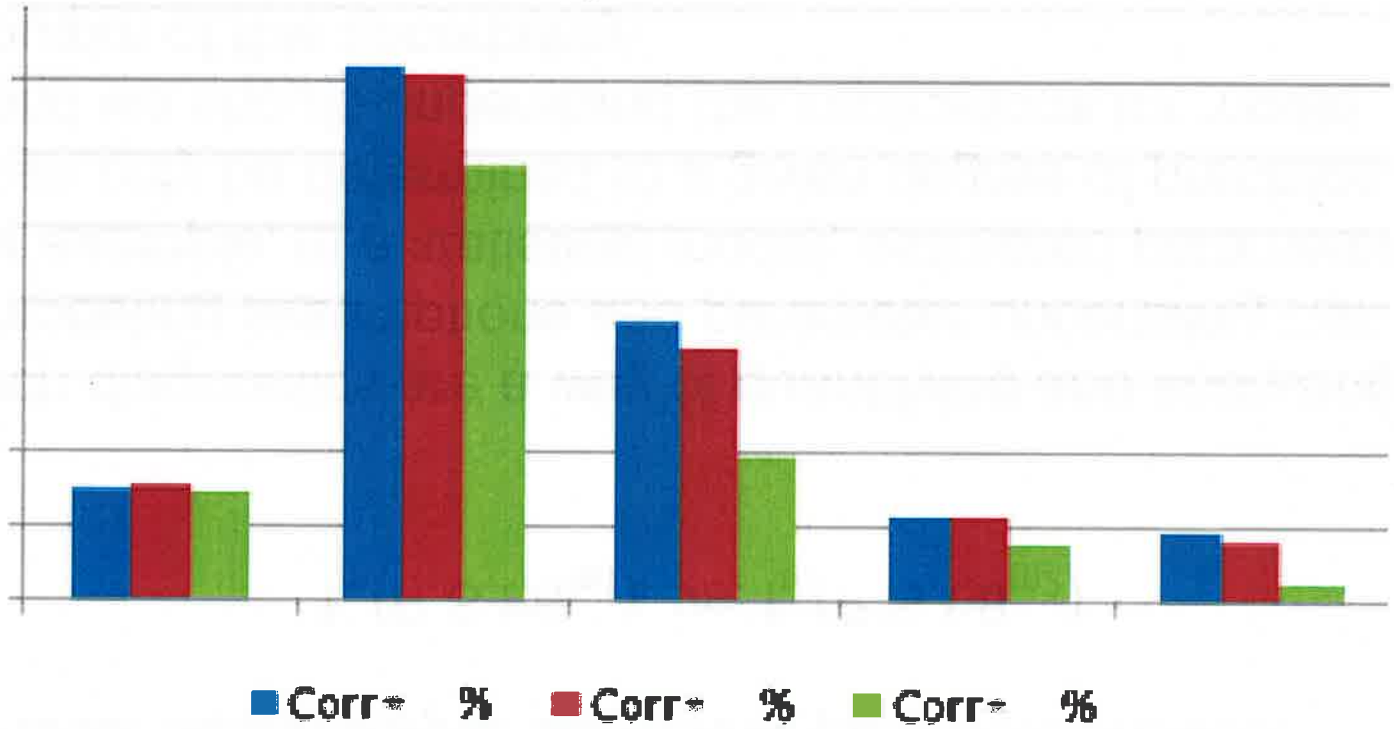
Such diagnostics are a way of quantifying and analyzing modeling assumptions and parameter uncertainty risk.

For example, in a statistical model, estimated parameters can only be determined to a given degree of precision, and we should understand the implications for model output of this uncertainty.

In other cases, models may rely on parameters that are calibrated or even set using expert judgment, and in these cases, we should also understand how the model responds to changes within a reasonable range.

Parameter Sensitivity Analysis

Loss Sensitivity to Asset Correlation Assumption



Output Benchmarking

Output benchmarking compares the outputs of two models given the same inputs.

A benchmark comparison seeks to understand the role of modeling choices while holding other inputs constant.

$$f_i(P,S^*,t;\theta_i) \text{ vs. } f_j(P,S^*,t;\theta_j)$$

where S^* denotes the union of the scenario variables used by the two models.

Possible ways of analyzing the data include:

- Graphs to provide visual comparisons.
- Scatterplots of outputs from models i and j
- Spearman rank correlation coefficients
- Use of statistical tests to determine whether the differences in output across the different models have a zero mean, suggesting agreement in their projections.

Output Backtesting

Output back-testing is probably the most commonly discussed method of performance monitoring for forecasting models; i.e., how similar was the model's prediction to what actually happened?

Particular challenge for tail scenario projections.

Output back-testing in the stress-testing context can be done in two ways.

The first approach is comparing model projections for the supervisory baseline scenario with the actually observed outcomes; however, this approach is a joint analysis of the model and the proximity of the baseline scenario to the actual outcomes.

To focus just on the model, we suggest comparing

$$f_m(P, S_{ACTUAL}, t; \theta_m) \text{ vs. } f_{ACTUAL}(P, S_{ACTUAL}, t)$$

Output Backtesting

Model testing often includes both "in-sample" and "out-of-sample" back-testing.

One type of out-of-sample back-testing, called out-of-time back-testing, evaluates the model's ability to fit outcomes outside of (most commonly after) the historical period from which the estimation sample was drawn.

Out-of-time back-testing helps to assess a model's stability over time.

Since stress test models are intended to be used for out-of-sample forecasts, out-of-sample back-testing, and in particular out-of-time back-testing, is an important diagnostic.

At the same time, stress test models are intended to be used for stressed conditions, but out-of-time periods may not contain stressed conditions. For this reason, in-sample back-testing may also play an important diagnostic role.

Conclusions

- Performance monitoring is a key element of developing and maintaining confidence in a model
- Structured framework for talking about performance monitoring can help in planning analyses and discussing results
- Black-box testing can be used across many model types, and can be complemented by more labor-intensive white-box testing as needed
- Stress testing models require us to draw on many types of analysis, since back-testing can only give us limited information