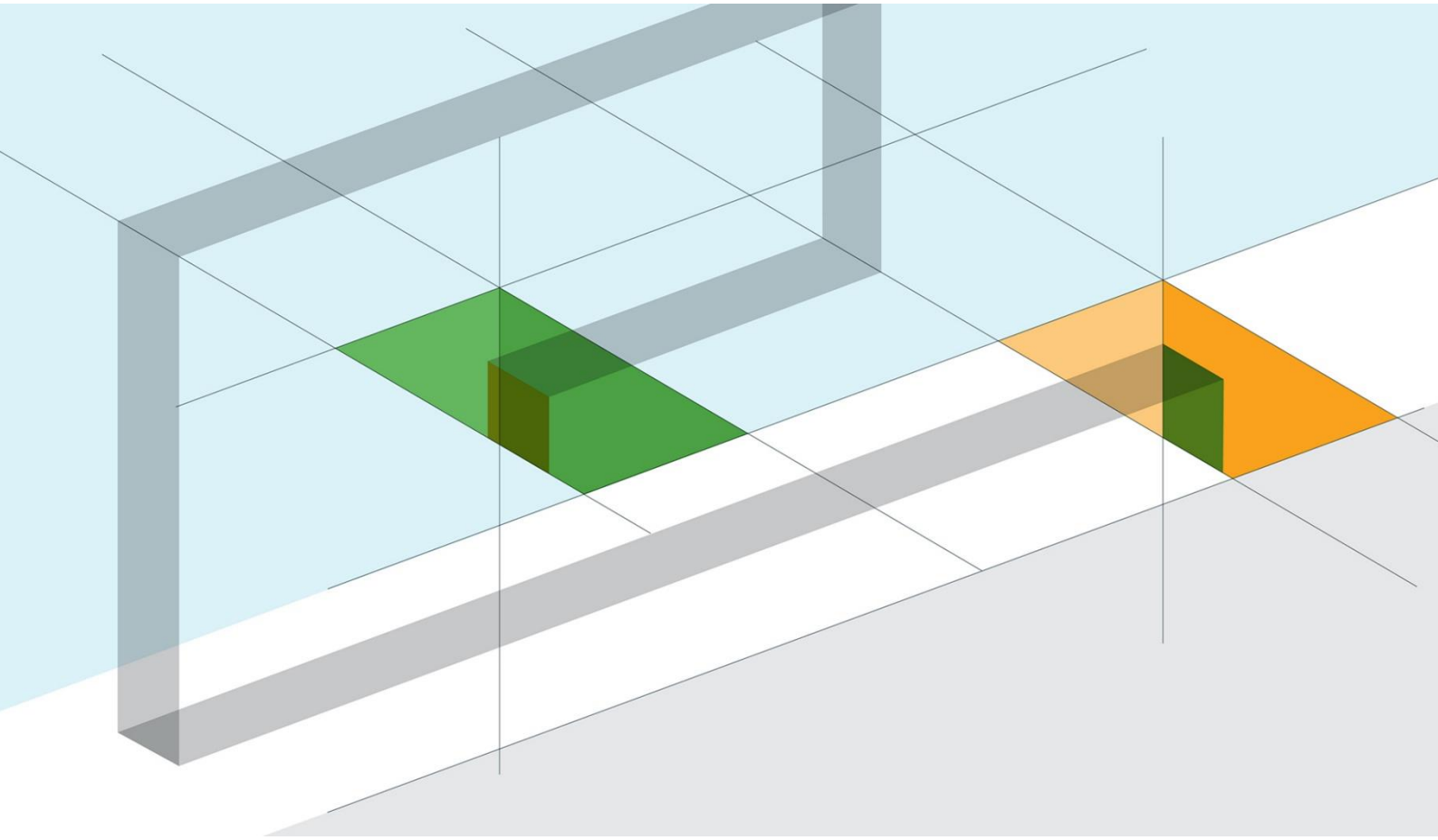
MILLIMAN REPORT

Proxy model validation

April 2020

Michael Leitschkis, DAV
Russel Ward, FIA



Milliman


Milliman

Table of Contents

ACKNOWLEDGEMENTS
The authors would like to thank Matthijs Otten for his contribution to this paper.

1. Introduction

Proxy modelling has now been in use across the UK and Europe for several years, principally in support of Solvency Capital Requirement (SCR) calculations within Solvency II (SII) internal models.

A key feature of a proxy model is the ability to perform capital calculations very quickly and efficiently relative to evaluations from the full cash flow or "heavy" model. This makes it feasible to develop loss distributions based upon hundreds of thousands of risk scenarios even for complex business with embedded guarantees.

The quid pro quo, as the name "proxy model" suggests, is that these models approximate the results that would be generated from the "heavy" model. As such, a key element in the proxy modelling process is to validate the model and thus understand just how good the approximation really is.

Market practice in the area of validation can vary quite markedly, as a recent survey of UK internal model firms by the Prudential Regulation Authority (PRA)1 demonstrated. We are aware this has generated some discussion in the industry and over time we envisage a degree of convergence and also a general raising of the bar.

The aim of this paper is firstly to provide a brief introduction to proxy modelling, and then to contribute to the current debate around validation via a succinct discussion of a number of the tests that might be contemplated and finally an illustration of a potential "validation dashboard."

[1] PRA (14 June 2019). Proxy Modelling Survey: Best Observed Practice. Retrieved 20 March 2020 from https://www.bankofengland.co.uk/-/media/boe/files/prudential-regulation/letter/2019/proxy-modelling-survey-best-observed-practice.pdf.

2. Proxy Modelling

Concisely, the aim of a proxy model is to capture the essential dynamics of the variable of interest (the dependent variable) via a relatively simple function including the key drivers of the variable's behaviour (the independent variables). In this paper, we focus on proxy models developed by fitting curves, typically polynomial functions, using regression techniques which represent the dominant industry approach at present.

In practice, different techniques are in use to achieve the objective, but the core process is common and proceeds as shown in Figure 1.

FIGURE 1: STEPS IN THE PROXY MODEL CALIBRATION PROCESS

Having set the scene, we will now explore the above process in more detail in the context of our favoured approach—Least-Squares Monte Carlo (LSMC). This is, of course, not the only approach available and so we will compare and contrast LSMC with the widely used approach of manual curve fitting (MCF).

2.1 STEP 1: SELECT THE TARGET VARIABLES

Objective: To select the variables whose behaviour is to be approximated. In an insurance setting, the principal use case of a proxy model is often to assess required capital and so the Own Funds / Net Asset Value (NAV) will be a key target variable. In support of these calculations though we may need to estimate other quantities. Thus, models may be required to estimate items such as the value of liabilities (e.g., the BEL under Solvency II) and the present value of shareholder transfers associated with profit sharing. For the enthusiastic proxy modeller, there may be a temptation to go still further and fit models to quite granular breakdowns of balance sheet items. Such an approach may well yield additional information about the behaviour of the balance sheet under stress conditions but the counterweight is a potentially significant increase in the complexity, time and cost of production and the risk of accumulating errors. If the NAV results from the outcome of, say, six proxy model results, then, while the accuracy of each may be acceptable in isolation, the accumulated errors across all six may not.

The other aspect of granularity relates to the structure of the firm and the associated need to fit and validate proxy models across different entities, funds and lines of business.

2.2 STEP 2: RISK DRIVER SELECTION

Objective: Determine the explanatory variables to be included in the model. These are the risk drivers collectively responsible for determining changes in the target variables being modelled, e.g., the NAV or BEL.

Examples of typical risk drivers are:

- Interest rates, often modelled as three principal components (PCs)
- Credit spreads
- Index returns on a range of asset classes such as domestic and international equities
- Inflation, sometimes modelled as two PCs
- Foreign exchange (FX) rates
- Volatilities for interest rates and index returns
- Lapse rates
- Mortality rates, mortality improvements
- Morbidity rates
- Expenses
- Management actions

2.2.1 Factors to consider in selecting risk drivers:

1. Risk drivers should have a material impact on the target variables—including minor risk drivers adds to the complexity of the proxy model but may yield little benefit in terms of overall explanatory power. Furthermore, adding extraneous risk drivers will tend to increase the volume of data required, which pushes up production times and costs.
2. Risk driver behaviour—for a curve to be appropriate to describe a risk driver's behaviour, the relationship between the target variable and risk driver should be continuously differentiable.
3. Risk driver relationships—avoid including risk drivers which exhibit a high degree of co-dependence with other risk drivers. Including such risk drivers adds little to the explanatory power of the data and can obscure key relationships when fitting the model.

LSMC vs. MCF

The considerations vary little between the methods, but risk driver proliferation can be a particular challenge for MCF due to the manual nature of the process and consequent increase in time and effort required.

2.3 STEP 3: SELECTION OF FITTING SCENARIOS

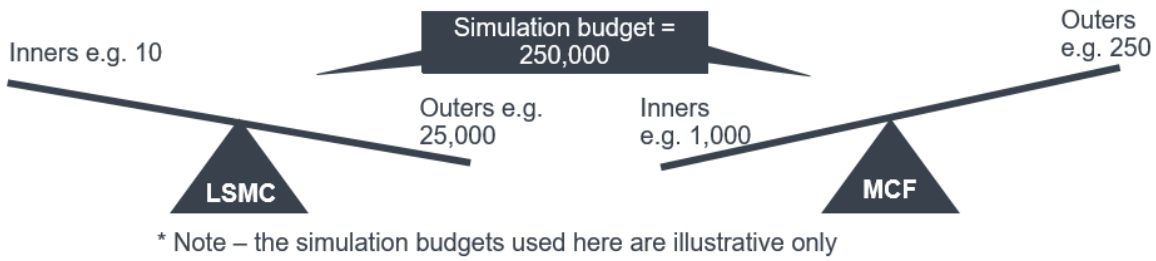
Objective: Determine the values of the various risk drivers that will be used in the fitting process. For example, if equity total return is a risk driver, then what realised values of that do we use to fit our model?

Terminology:

***Outer scenario*:** Represents a realisation of risk driver outcomes (univariate, e.g., equity return +10%, or multivariate, e.g., equity return + 10% and nominal yields + 100 bps and mortality up 5%).

***Inner scenario*:** A risk-neutral scenario path, conditioned upon the outer scenario characteristics, used for Monte Carlo valuation of business with embedded options and guarantees.

FIGURE 2: HOW LSMC AND MCF DIFFER IN USE OF THE AVAILABLE CALIBRATION BUDGET



* Note – the simulation budgets used here are illustrative only

This step exhibits a marked difference between LSMC and MCF:

LSMC vs. MCF

LSMC

- Outer scenarios used are multivariate and derived automatically to provide an even coverage of the multidimensional risk space. To achieve this, low-discrepancy sampling techniques are used, e.g., Sobol numbers.
- Scenario selection is defined by the risk drivers used and simulation budget—no expert judgement is needed in the selection process.
- Use of few inner (risk-neutral) scenarios allows the simulation budget to be more focused on the outer scenarios, enhancing the coverage of the risk space.

MCF

- Outer scenarios are manually selected with the focus on key individual sensitivities and cross-terms—this provides flexibility to target specific scenarios but requires considerable expert judgement.

2.4 STEP 4: GENERATION OF FITTING DATA

Objective: Produce observations of the target variables as the risk driver values are changed. This is another part of the process where LSMC and MCF differ significantly. In short:

LSMC vs. MCF

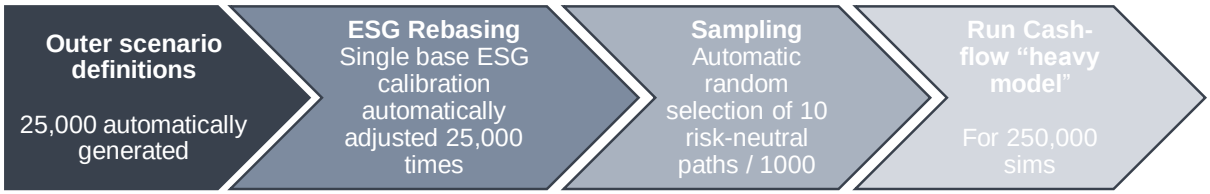
LSMC	MCF
Employs a rich set of risk driver observations BUT Uses inaccurate estimates* of the results for each individual observation	Uses a much more limited set of risk driver observations BUT Each observation is of similar accuracy to a normal valuation result

* Note that there is no loss of accuracy for deterministically valued business.

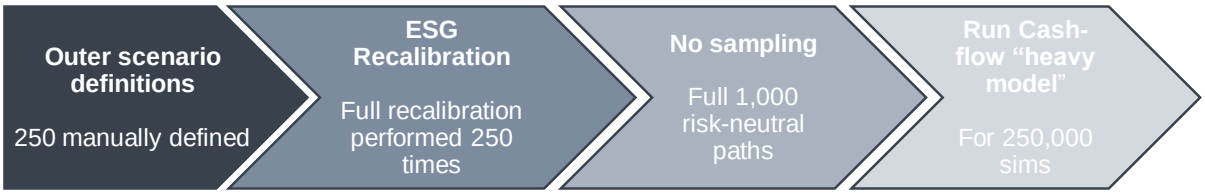
The different approaches adopted by LSMC and MCF result in process variations as described in Figure 3.

FIGURE 3: FITTING DATA GENERATION WITH LSMC AND MCF

LSMC



MCF



* Note – the simulation budgets used here are illustrative only

2.5 STEPS 5 AND 6: FITTING A MODEL TO THE DATA

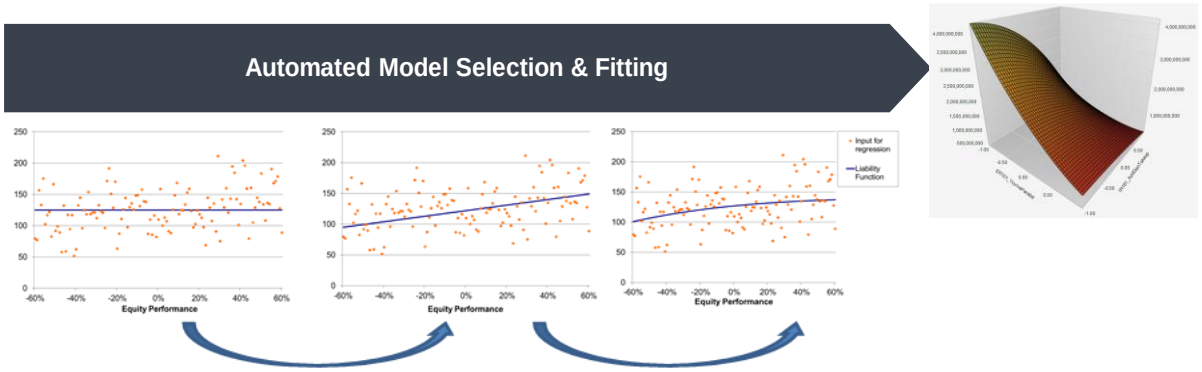
Objective: Select the type of function, e.g., polynomial, select basis functions and calibrate coefficients.

Type of function—polynomials are often employed as they offer considerable flexibility to match the fitted function to the behaviour of the target variable(s). However, it is important that the appropriateness of this choice is checked. For example, by running full heavy model stresses for different levels of the key risk drivers and plotting the results, results showing discontinuities or other materially non-smooth behaviour should be investigated and may require an alternative approach.

In terms of fitting a model to the calibration data, both LSMC and MCF typically use regression techniques to derive basis functions and coefficients that deliver a good fit to the input data. Here again, though, the approaches differ quite markedly in the mechanics of how this is achieved.

FIGURE 4: LSMC MODEL SELECTION

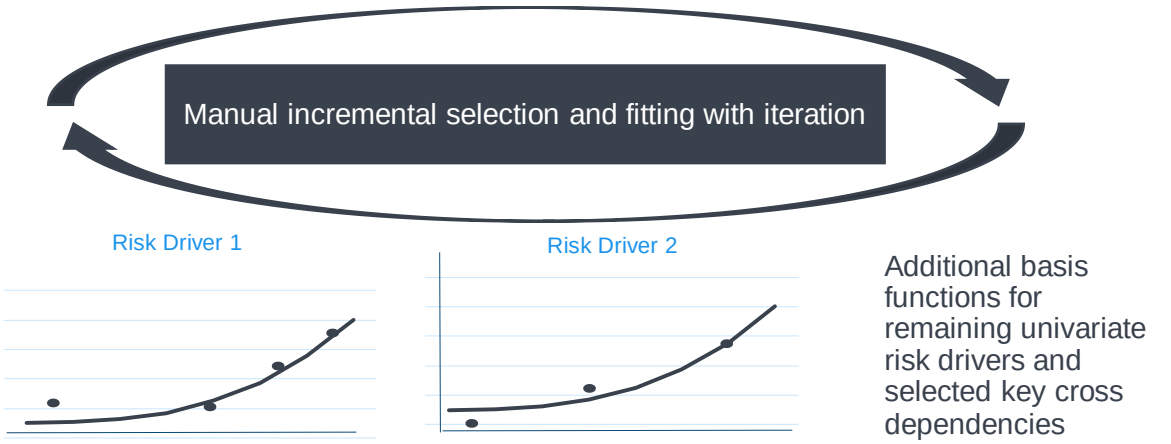
LSMC



Under LSMC, an automated model selection procedure is used to fit the curve. The starting point is the simplest relationship possible—a constant one. Thereafter, the procedure determines a feasible set of additional candidate terms. As indicated in Figure 4, the feasible set considers the simplest terms first before permitting more complex ones (higher powers, cross-terms) and then selects the extra term that provides the greatest improvement in the fit as evidenced by a measure such as the Akaike information criterion (AIC). The risk of including too many terms (overfitting) is mitigated using techniques such as AIC which apply a penalty for each extra term added. The curve is then refitted with the new term included and the process repeated. This algorithmically driven approach avoids the need for expert judgement as to the "best" model structure.

FIGURE 5: MCF MODEL SELECTION

MCF



Under MCF, judgement is required in selecting the fitting points to use for each risk driver included. The number and positioning of these points can be critical to the quality of the fitted curve. In particular, we note that no information is available to guide the fit outside the range of selected points, e.g., in Figure 5, to the right of the final point for risk driver 2. LSMC also requires a calibration boundary to be set, but under this approach scenarios can be populated up to the boundary for all risks automatically. Under MCF a deliberate decision is required to allocate fitting points at the extremes for all risks and there may be insufficient scope in the scenario budget to permit that which can then result in a greater level of extrapolation risk.

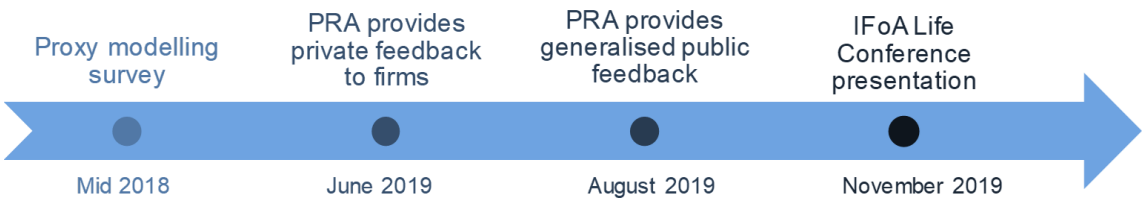
Furthermore, judgement is required in selecting the basis functions to be used in each step of the fitting process and in selecting the cross-term dependencies to include in the model.

The remainder of the paper is now devoted to Step 7—validating that the proxy model derived is fit-for-purpose.

3. Validating the proxy model

3.1 RECENT REGULATORY DEVELOPMENTS

FIGURE 6: PRA PROXY MODEL SURVEY AND FEEDBACK



In May 2018, the PRA's Proxy Modelling specialism issued a survey to a sample of 11 life insurers with a proxy model. They designed the survey questions to capture each firm's modelling process along with each firm's views of best practice. The PRA then compared approaches across firms, and considered the input provided by consultancies and other National Supervisory Authorities to identify the "best observed practice" among the survey responses.[2]

With regard to proxy model validation, the feedback noted that firms should consider all uses of the proxy model when undertaking validation. As different applications may draw on different parts of the overall loss distribution, it is important to ensure the model fit is adequate across all areas of use.

The survey, not unreasonably, placed a good deal of weight on validation activities, which were divided into three areas:

- In-sample testing, restricted to the data used to calibrate the proxy model and performed during the fitting process. The PRA noted a number of tests in use by firms most related to verifying the assumptions which underpin ordinary least squares (OLS) regression models.
- Out-of-sample testing, addressing the comparison of proxy model against "heavy" model results for stresses and scenarios that have explicitly not been used in the calibration process. The PRA focused on the need for firms to demonstrate the adequacy of all of the following:
  - The overall number of scenarios used in light of business and model complexity, in particular the number of risks included.
  - Their coverage in terms of spanning all areas of the loss distribution relevant to use cases of the model.
  - Their coverage in terms of the nature of the scenarios tested, e.g., univariate tests of key risks, bivariate tests of key risk interactions and full multivariate combinations including many or all risks.

  The PRA felt that best observed practice was to select some scenarios using automated judgement-free techniques, but supplementing them with additional points informed by an understanding of the business.
- Other testing, noted a wider use of out-of-sample data in some areas repeating in-sample tests but also considering areas such as the ability of the proxy model to rank risk scenarios consistently with the "heavy" model.

  As noted above, the PRA tended to align the timing of in-sample testing activity with the calibration process itself but that is really an artefact of the approach taken. In a manual curve fitting world, the timing distinction makes sense but in a world of automated data-driven calibration, as under LSMC, such testing can simply be included within the suite of validation information delivered by the production process.

  Having captured the approaches being taken by a number of firms, the full list of possible tests is quite long. This provides a really useful starting point for firms to consider their own validation processes and assess whether they have any significant gaps. However, some of the tests have similar aims and some will be more relevant than others to particular firms. Judgement is therefore required in selecting a validation framework and range of tests that provides the required comfort over the proxy model's fitness for purpose while enabling validation outcomes to be clearly articulated to stakeholders.

[2] PRA, Proxy Modelling Survey: Best Observed Practice, op cit.

3.2 VALIDATION OVERIEW

Similarly to the PRA, we divide the validation of the proxy model into three areas, though our segmentation is slightly different. In the remaining sections of this paper we illustrate a validation for a single fund based on a combination of realistic results from an LSMC calibration alongside some simplified examples for brevity.

Reasonableness tests: The features of the proxy model should be consistent with the expected behaviour of the underlying book of business. For example, if our knowledge tells us to expect a particular set of interactions between the risks then we should check that our proxy model includes the cross-terms reflecting these relationships. At a more basic level, we should also check that our model delivers plausible relationships between the target variable outcome and varying stresses of the explanatory risks—it is a simple matter to verify our proxy model exhibit's reasonable behaviour.

We place this category of tests first within the overall validation framework on the basis that, unless the proxy model's behaviour is demonstrably reasonable, it is hard to place much confidence in the output.

Application tests: In calculating the SCR under a Solvency II internal model via LSMC, we develop a full loss distribution via the evaluation of the change in NAV across literally hundreds of thousands of scenarios, each representing a realisation of the risk drivers included in the proxy model. How comfortable we are with the reasonableness of the results comes down to addressing the question, "how accurately does the proxy model estimate results for data points that have played no role in its calibration?" The application type tests are critical in framing the answer and include:

- So-called out-of-sample tests which compare results from the heavy model with those from the proxy model.
- Tests to verify that the proxy model has not been overfitted.[3]

In-sample tests: These tests consider the relationship between the fitted model and the data used to perform the fit. In common with many statistical models and techniques, OLS regression comes with a set of assumptions forming the ideal conditions for its application. In real-world applications, our experience is that these idealised conditions will not be perfectly met. Nevertheless, we find that curves can still be fitted using LSMC which provide an acceptably accurate approximation to the "heavy" model results. Having said this, we do not dismiss in-sample tests as, in our view, they can provide helpful diagnostic information that may highlight issues for further investigation and ultimately lead to improved calibrations in the future.

Parameter calibration: For completeness, we also note that investigations are needed to establish a suitable set of parameters for the calibration. Such parameters will include:

- The maximum order of univariate and cross-terms permitted
- The maximum number of terms to be included in the fitted function
- The overall calibration budget to be used and its division between outer and inner scenarios

Investigations are required to optimise these parameters upon initial implementation. Thereafter, it should only be necessary to review them in the event of a marked change affecting the regression problem, for example a significant change in the "heavy" model such as the inclusion of extra risk drivers or a material shift in financial conditions expected to impact the significance of the various risks and their interrelationship. Given the different nature of these investigations, we do not cover them further in this paper.

In the remainder of this section we explore a number of tests that can be applied to address our three validation areas. To be clear though the tests we cover are in no way an exhaustive list and we recognise that the selection of tests adopted represents another area of expert judgement.

[3] Overfitting occurs when a model's parameters include not only the genuine relationships in the underlying data but also spurious ones arising solely from noise. The noise element will not be replicated in data outside the fitting set and thus its inclusion reduces the ability of the model to accurately estimate results for scenarios outside the fitting set.

3.3 REASONABLENESS TESTS

3.3.1 Initial analysis of the fitted model ("sniff test")

Even before a proxy model is fitted, there will be a body of knowledge within any firm regarding the hierarchy of risks to which any book of business is exposed and also the key interrelationships between those risks.

When a proxy model is fitted, a first step in checking its reasonableness is to consider the list of basis functions selected and their significance and check this against prior expectations. For example, we may know that a particular book of business has a significant exposure to the level of interest rates, with rising rates expected to reduce the NAV. The same book may also be exposed to the take-up rate of guaranteed annuity options (GAOs), where increasing take-up pushes up the liabilities and reduces the NAV. We may know that there is an interplay between the two risks such that if interest rates fall the GAOs become more valuable and the impact of changes in take-up rates increases. In such a case, it would be reasonable to immediately look for the following elements in our fitted model:

- Univariate terms in the interest rate and GAO take-up rate risk factors are present in the polynomial.
- The coefficients and order of selection of those terms denote a significance consistent with expectations. Note that, for an automated model selection algorithm as used in LSMC, the most significant terms are selected first. In the example being considered, if we found the univariate linear term in the risk "interest rate PC1"[4] had been selected 133rd from 200 terms, then that would immediately feel wrong.
- The signs of coefficients should be consistent with the definition of the risk factors and the direction of their expected relationship with the NAV. Assuming we have defined increases to interest rates to be positive and falls to be negative then we expect to see a negative coefficient on the linear term in "interest rate PC1."
- The presence of a cross-term(s) capturing the expected interaction between the level of interest rate and GAO take-up rate risk factors.

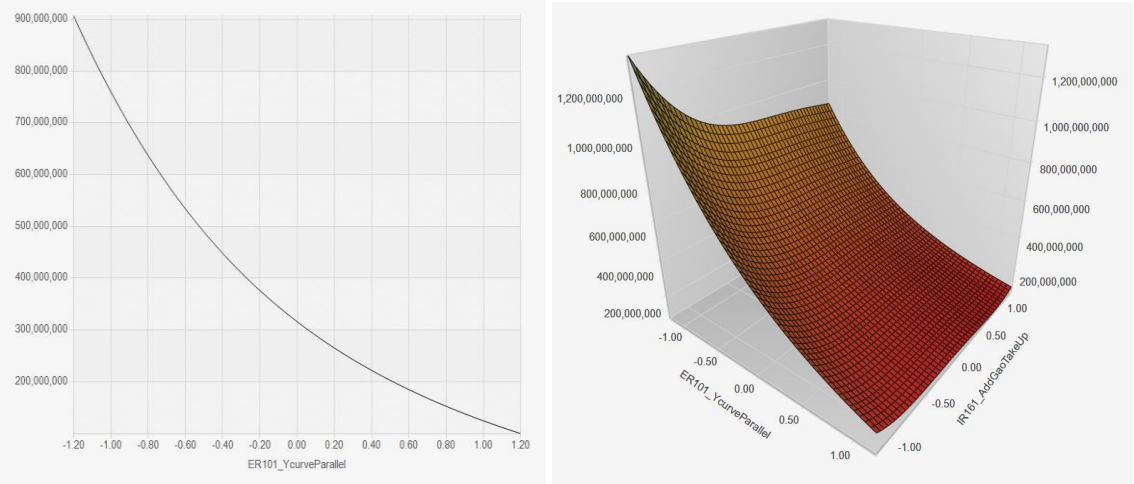
The data required to undertake the analysis above should be readily available as an output from the fitting process.

3.3.2 Graphical analysis

Taking things a step further, we can chart the modelled relationships between our NAV and the key risk factors which influence its variation.

Analysis of univariate risks can be undertaken using standard 2D charts while analysis of key bivariate relationships can be undertaken using more complex 3D charts.

FIGURE 7: GRAPHICAL EXPLORATION OF MODELLED RELATIONSHIPS



[4] To simplify modelling, interest rate risk is often modelled as a set of principal components (PCs), of which PC1 is typically taken to represent the overall level of interest rates.

The left chart in Figure 7 shows the univariate relationship between NAV and the level of interest rates, with NAV falling as rates increase. The right-hand chart then illustrates how the bivariate combination of interest rate level and GAO take-up rate impacts the NAV. Again we can see that NAV falls as rates rise but also that the impact of GAO take-up is far more marked at low rates than at high rates. In this simple example we can readily conclude that the behaviours exhibited by the fitted model are consistent with our ex ante expectations.

3.3.3 Analysis of the change in model form

A final test in the reasonableness section is to consider how the form of the fitted model has evolved over time. By the form of the model we mean the selection of basis functions that make it up such as that shown in Figure 8.

EXHIBIT 8: DECOMPOSITION OF FITTED CURVE INTO BASIS FUNCTIONS

Function	Risk Factor1		Risk Factor2		Selection Order
	Risk	Power	Risk	Power	
1	R1	1			1
2	R4	1			2
15	R6	2			15
16	R1	1	R9	1	16

The information provided in Figure 8 enables us to compare the form of models fitted at different times such as year-end (YE) 2019 versus YE2018. In particular, we can consider:

- How many functions are common to both models?
- How consistent is the selection order for the common functions?
- How many new functions have been added and what are their characteristics?
- How many functions have been dropped and what are their characteristics?

Armed with this information, the changes can now be considered for consistency with what we know about the influences on the book of business during the last year. If at YE2018 this book had little equity exposure but during 2019 that had been significantly increased, then it would not be surprising to find terms added involving the equity return risk factor. If during the year a large block of protection business had been sold, then we might expect to see the importance of terms involving mortality risk factors decline and remaining terms selected later, with some possibly removed altogether. On the other hand, if there had been little change to the business or to financial conditions over the past year, then it would be surprising to see significant changes to the model form, and this would warrant further investigation.

Clearly, this is not a completely scientific exercise and sound knowledge of the business is required along with a good slug of judgement. Nevertheless, the exercise can provide comfort over the changes observed and highlight any that simply don't make sense and so require further investigation, which might start by considering any changes in the "heavy" model behaviour for relevant stresses.

3.4 APPLICATION TESTS

3.4.1 Out-of-sample tests

The premise of the main application test is a straight comparison of the proxy model estimate versus the "heavy" model estimate—the difference being calculated for every test scenario executed. In tabular form the results might look like Figure 9.

FIGURE 9: ILLUSTRATIVE OUT-OF-SAMPLE TEST RESULTS

Scenario	NAV - Heavy Model (£m)	NAV - Proxy Model (£m)	Loss Distribution Percentile	Error (£m)	Abs Error (£m)	Tolerance (£m)	Test Outcome
Multi7	376	362	99.80	14	14	15	PASS
Multi61	398	408	99.55	-10	10	15	
Multi133	405	410	99.50	-5	5	15	PASS
Multi222	430	413	99.45	17	17	15	
Univ12	470	477	87.60	-7	7	15	PASS
Bivar4	567	566	66.30	1	1	15	PASS

Abs Error within tolerance **Abs Error outside tolerance**

Inclusion of the loss distribution percentile implied by the "heavy" model NAV of each scenario can be helpful in demonstrating the coverage of the test set, in particular highlighting any gaps.

The “Error” is included to enable the investigation of the degree of randomness in the direction of the errors. A strong tendency to observe runs of positive or negative results indicates some systematic bias in the proxy model which should then be investigated. To provide a quantitative assessment, a statistical test (such as the runs test) could be undertaken to test the hypothesis that the sequence of positive or negative errors is random.

Assuming our sequence of positive or negative errors is random, it does not concern us which direction any particular error is in. Hence, the absolute value of the error “Abs Error” is used to compare with the tolerance level set to determine the test outcome. This is straightforward enough but the setting of an appropriate tolerance threshold is a nontrivial task.

Under an LSMC approach, calibration scenarios are devised to cover the entire risk space evenly and all such scenarios carry equal weight in the fitting process. Given this, ex ante, there is nothing to suggest the fitted curve will be any more or less accurate in any particular area of the risk space—this is consistent with the adoption of a constant error tolerance across the set of validation tests.

Beyond that, some additional sophistication could be introduced by examining the residual variability of the validation scenarios (for stochastically valued blocks of business, they are typically calculated using say 1,000 to 2,000 risk-neutral simulations—they are themselves estimates). As an example, we note that the level of interest rates often plays a significant role, as the scale and interest rate sensitivity of liabilities with embedded long-term guarantees will typically increase as interest rates fall. In line with this, there is a tendency for validation points reflecting extreme downward stresses to rates to exhibit greater levels of standard error—in our experience, this can be over 50% higher than points reflecting rates around the base valuation level. Thus, recognising differences in the inherent variability of estimates in different areas of the risk space might enable the construction of a slightly more nuanced tolerance, allowing more latitude where there is greater uncertainty in the underlying data.

Clearly, there can be many realisations of multivariate risk scenarios which yield very similar results in terms of NAV but may exhibit different levels of error. In Figure 9 this is the case for scenarios “Multi61,” “Multi133” and “Multi222.” Indeed, for key areas of the loss distribution such as that around the 99.5th percentile we would expect concentrations of validation scenarios to have been considered. In these areas therefore the outcome of the validation can be considered in the round by looking at the average level of error across the validation points within a certain range—this would lead to a pass (average absolute error 10.67) despite the error in “Multi222” being a little outside our tolerance.

The number of validation points to be tested depends on the risk profile and complexity of the business. Understandably, the PRA appears keen for firms to test more points as, provided the points are sensibly chosen, it’s hard to argue that greater comfort will not be provided by the use of more validation points. Of course, we don’t live in an ideal world and there is a price to pay for any such expansion in terms of the increased computational load which will drive up processing times and costs. The PRA clearly recognises this though we expect them to continue to challenge firms to push the boundaries of what is feasible, particularly in light of the move towards “cloud-based” solutions which support targeted high-volume processing.

3.4.2 Testing for overfitting

Overfitting occurs when a regression model adheres too closely to the data used in its calibration. The problem is that the “signal” is not effectively separated from the “noise” in the fitting data and both are baked into the fitted function.

When we come to use the model to predict values outside those used in fitting, the noise element obscures the true relationships and undermines the predictive power of the proxy model, resulting in elevated estimation errors.

A relatively simple check involves calculating the R-squared[5] regression statistic for both the fitting data and the validation data and comparing them. If the R-squared value from the fitting data is markedly higher than that on the validation data, then this may indicate overfitting and potential benefit in generating a simpler model.

With a MCF approach, the comparison is easier as the fitting and validation data are broadly comparable in terms of accuracy and volume. This is not the case for LSMC but a technique called k-fold cross-validation can be employed as follows:

- Divide the calibration data into a number (k)[6] of equal segments and select one as the "test segment." Then the remaining data forms the "calibration segment." For example, if running 40,000 outer scenarios, we might split the data into 10 x 4,000 segments.
- Fit the model to the calibration segment data only, 36,000 scenarios in our example here.
- Evaluate the R-squared and distribution of residuals on both the "test" and "calibration" data segments.
- Repeat the exercise k times, selecting a different segment as the "test segment" each time.

The relative performance of the k-fitted models can then be compared to identify any tendency towards overfitting. A simple exercise with k = 2 is illustrated in Figure 10.

FIGURE 10: ILLUSTRATIVE OVERFITTING TEST RESULTS

Test Statistic	Calibration Segment (20,000 outer scenarios)	Test Segment (20,000 outer scenarios)
Mean residual	0	£100k (≈0)
SD residuals	£38m	£39m
R-squared	97.9%	97.8%

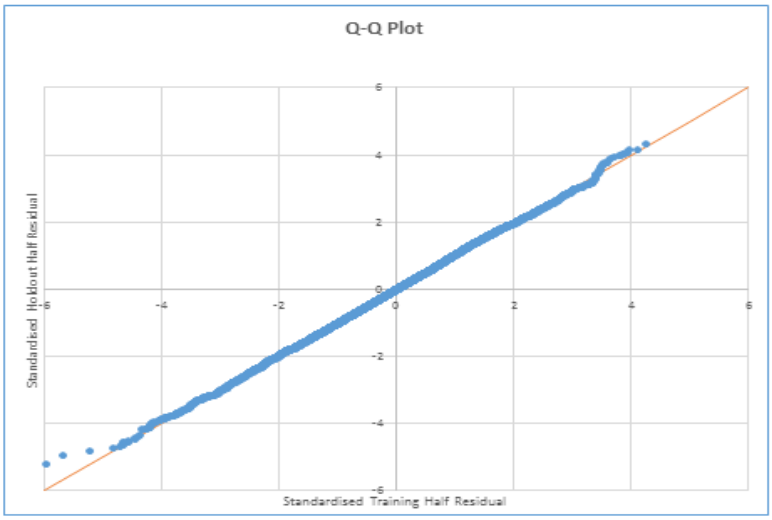


Figure 10 illustrates the results for a with-profit fund and indicates very consistent though slightly poorer-fitting performance on the test segment. These results provide comfort that the model has not been overfitted.

3.4.3 Ranking tests

One of the attributes required of the internal model is to facilitate the ranking of risks to which the insurer is exposed (see article 121(4) of the SII Directive). This was also an area of interest for the PRA, which noted that ranking differences "could be important even if it does not materially affect the monetary amount of the SCR at a given point in time since it can impair the model's ability to rank risks and its role in the capital allocation process

[5] R-squared of a regression shows the proportion of the variation in the data which is explained by the fitted model.

[6] The value of k used is a matter of user judgement.

and other Use Test requirements." Consequently, the PRA encouraged firms to include analysis of the ranking performance of the proxy model within the scope of validation activity.

FIGURE 11: ILLUSTRATIVE RANKING TEST RESULTS

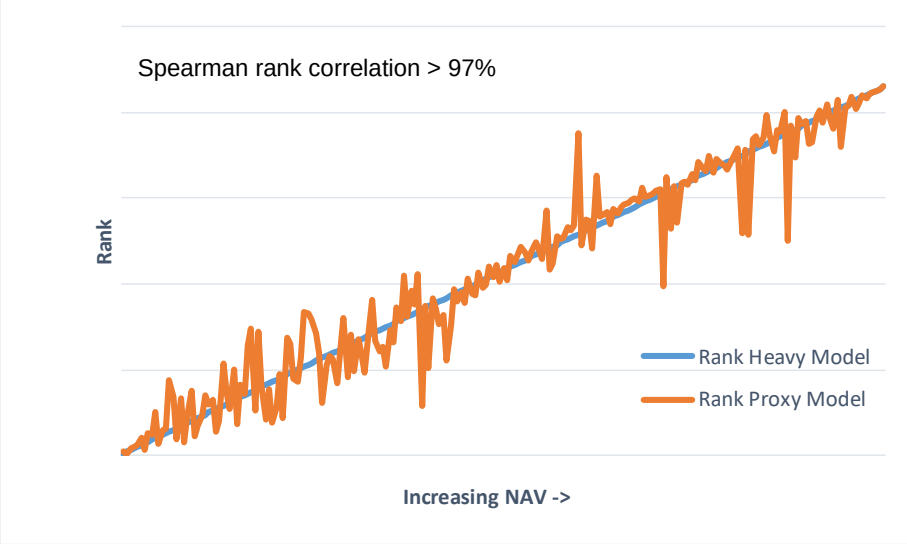


Figure 11 illustrates how a simple ranking comparison might be presented with results ordered by increasing NAV from the "heavy" model scenarios. In the illustration, the proxy model ranking has a clear and strong positive association with that of the "heavy" model but, in reality, the association will not be perfect. Ordering the rank comparison by NAV provides insight on the alignment of ranks across the loss distribution. Finally, computing a rank correlation statistic such as Spearman's rho or Kendall's tau provides a concise measure of the strength of the association that can be compared between fitted models and also tracked over time.

3.5 IN-SAMPLE TESTS

Finally, we come to the in-sample tests. In common with most statistical models, the OLS regression used by our proxy model is underpinned by a number of idealised conditions. Where these conditions fully hold, the parameters (coefficients) fitted are optimal in the sense that they are a best linear unbiased estimator (BLUE) of the underlying population values. The aim of many in-sample tests is to establish how well these OLS conditions hold in the setting of our real-world application. Where departures occur, it is good to be aware of them as they can indicate areas in which improvements can be made in the future. Nevertheless, in our experience, the OLS approach is quite robust to modest departures from the idealised conditions and, in a validation context, we would not typically countenance rejecting a calibration that had performed very well in the application and reasonableness tests solely due to modest departures from the ideal OLS conditions.

A further point to keep in mind is the nature of the regression exercise being undertaken. In many applications of OLS, the whole range of influences on an observed value is unknown and there may be subtle but important drivers of the results hidden to the observer. In an insurance context, our use case is to employ the proxy model to more simply represent results from another model, the "heavy" model. The influences on the "heavy" model are known from the specification of that model—a closed system. The result is that some of the issues that in-sample tests are designed to flag up simply should be avoided at source by careful model design.

The OLS assumptions are:

1. Linear model: The relationship between our dependent variable (NAV, BEL) and the explanatory variables (risk factors) must be capable of being expressed by a "linear in coefficients" model. Such a model has the form:

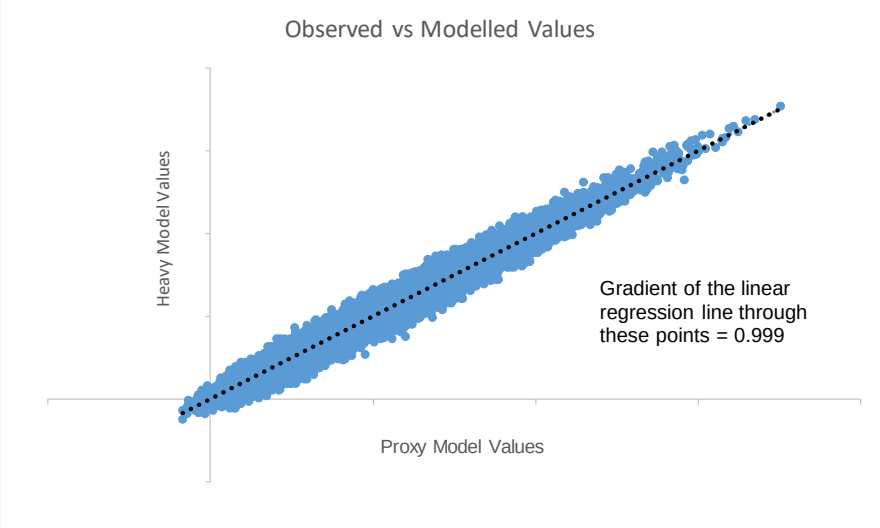
   $NAV = a_0 + a_1R_1 + a_2R_2 + a_3R_3 + \ldots\ldots\ldots\ldots + a_nR_n + error$

2. Where the a_x are the coefficients and the R_x are the basis functions defining the risk factor relationships.
3. Homoscedasticity: The variance of the residuals (fitting errors) should be invariant across observations.
4. Exogeneity: The explanatory variables should be independent of the residuals.
5. No autocorrelation: The residuals should be independent of one another.
6. No bias: The mean of the residuals should = 0.
7. No multicollinearity: None of the explanatory variables is a perfect linear function of the others
8. Normally distributed error terms: Not a strict requirement but helpful when computing certain results such as significance values and confidence intervals.

3.5.1 Linear model

To test for a linear in coefficients model, we plot the "heavy" model results against those produced by the proxy model. In Figure 12, we note that the results are broadly symmetrically distributed around the linear line of best fit and that the gradient of that line is very close to 1. These results indicate that our linear model is perfectly reasonable in this case.

FIGURE 12: CHECKING FOR A LINEAR MODEL



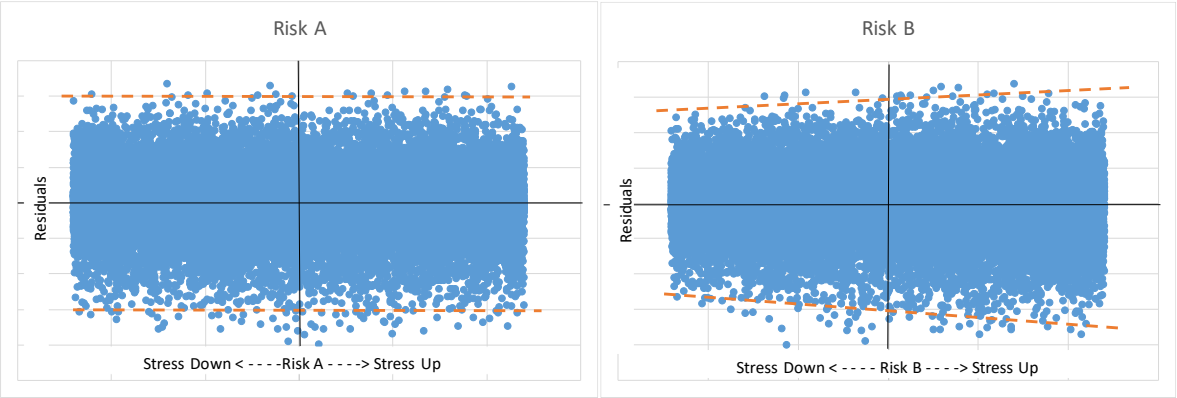
Unless there are some very material changes to the regression problem being considered, this is not a test that it should be necessary to complete as part of a standard "business as usual" (BAU) validation.

3.5.2 Homoscedasticity

The second of the OLS assumptions is that the residuals exhibit constant variance across the range of observations. If this assumption does not hold, it may result in undue weight being placed on part of the data and impacting the accuracy of the parameter estimates derived.

There are statistical tests such as the White test which can be used to indicate whether heteroscedasticity is present in the calibration data. However, such tests provide a binary answer and do not indicate where any issue might lie so they have a limited diagnostic function. A more informative approach can be to undertake a graphical analysis plotting the regression residuals against the values of the key risks. A couple of example risks are shown in Figure 13.

FIGURE 13: CHECKING FOR HOMOSCEDASTICITY



The charts in Figure 13 indicate that Risk A behaves very well while Risk B displays a modest tendency for the variance of the residuals to be positively related to the direction of the stress, though this is more marked in the downward direction.

3.5.3 Exogeneity

Exogeneity requires that there should be no linear relationship between our risk factors and the residuals of the regression—the residuals should be random. If this is not the case, then it implies there is something systematic driving the behaviour of the residuals which is not included in the specification of our model, i.e., a missing risk factor.

Examination can be undertaken using the same approach as for homoscedasticity, though in this case we are looking for a distinct shape or trend in the results; Figure 13 indicates no such issues.

Given that our proxy model is approximating the behaviour of another model with known risk factors, careful model design should address exogeneity. Testing for this makes sense upon initial implementation and if there are subsequent changes to the "heavy" model which may introduce new risk factors. However, absent such changes, testing as part of ongoing BAU validation is unlikely to add much value.

3.5.4 Autocorrelation

Regression residuals should be independent of one another—there should be no autocorrelation. Again, if this is not the case it can indicate a missing explanatory variable, often “time,” as this issue typically arises in the context of a time series model. However, a calibration for SII SCR purposes should not involve time series data and so this particular issue is unlikely to trouble us. A statistical test, for example the Durbin-Watson test, can be undertaken to check for any autocorrelation. Similarly to tests for exogeneity, though, post-implementation checks are unlikely to add a lot of value unless the "heavy" model has seen significant change and the additional comfort provided by a retest is deemed worthwhile.

3.5.5 Bias

The mean of the regression residuals should be zero—no bias. Where an intercept term is included within the regression model this requirement should be met by construction.

3.5.6 Multicollinearity

Strictly speaking, no risk factor should be included whose behaviour is a linear combination of other risks. More broadly though, the regression results can still be adversely impacted if any pair of risk factors has a significant correlation between them. If significant correlation is present between variables then the accuracy of the fitted coefficients for the impacted risks will be reduced, making it more likely that we will observe material errors in out-of-sample tests. Furthermore, the proxy model may display significant instability in the modelled relationships for the correlated risks whereby small changes in the calibration data lead to disproportionately large changes in the derived coefficients, making it difficult to draw robust conclusions about the impact of those individual risks on the NAV or BEL.

To test the strength of the relationships between risk factors, their correlation matrix can be derived and inspected as in Figure 14.

FIGURE 14: CHECKING CORRELATIONS BETWEEN RISK FACTORS

	Risk A	Risk B	Risk C	Risk D	Risk E	Risk F	Risk G	Risk H	Risk I	Risk J
Risk A	1.000	0.017	-0.005	-0.008	0.002	0.038	0.003	0.001	0.001	0.002
Risk B	0.017	1.000	-0.001	0.002	0.000	-0.007	-0.002	0.001	-0.002	-0.001
Risk C	-0.005	-0.001	1.000	-0.001	0.003	0.002	0.001	0.000	0.001	-0.002
Risk D	-0.008	0.002	-0.001	1.000	-0.001	0.003	-0.001	0.003	0.000	0.002
Risk E	0.002	0.000	0.003	-0.001	1.000	0.000	-0.001	-0.001	0.001	0.001
Risk F	0.038	-0.007	0.002	0.003	0.000	1.000	-0.001	-0.002	0.000	-0.001
Risk G	0.003	-0.002	0.001	-0.001	-0.001	-0.001	1.000	0.004	-0.002	-0.001
Risk H	0.001	0.001	0.000	0.003	-0.001	-0.002	0.004	1.000	0.002	0.002
Risk I	0.001	-0.002	0.001	0.000	0.001	0.000	-0.002	0.002	1.000	-0.002
Risk J	0.002	-0.001	-0.002	0.002	0.001	-0.001	-0.001	0.002	-0.002	1.000

In Figure 14 we have derived the correlation matrix between Risks A to J implied by the calibration data. Inspecting the results we note that the highest absolute off-diagonal result is 0.038 for the (Risk A, Risk F) pairing. This very low value provides comfort that the risks included in our model lack any significant linear relationships with each other.

3.5.7 Normally distributed residuals

This is not a requirement for the OLS method to be applicable, but is a desirable property in that it facilitates the more accurate testing of the significance of the fitted model's coefficients via the calculation of their probability values (p-values). This testing can identify whether the fitted model contains parameters which are statistically insignificant. The presence of a number of insignificant parameters may indicate an overfitted model or a model that can be simplified to some degree without material loss of predictive power.

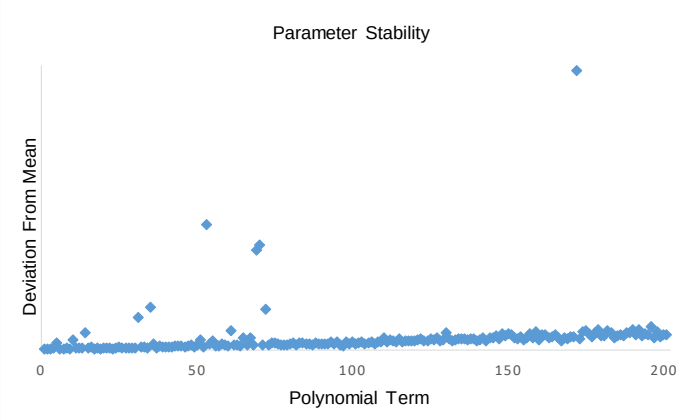
Various approaches can be taken to testing for normality:

1. Create a histogram of the residuals and compare to a normal probability density.
2. Create a quantile-quantile plot (see Figure 10 above) which compares the ranked normalized values of the residuals with the quantile points of a standard normal distribution. This approach generally makes it easier than #1 to discern departures from normality.
3. Calculate the third standard moment (skewness) and fourth standard moment (kurtosis) of the distribution of residuals and compare them to the expected values for a standard normal distribution (0 and 3, respectively). A statistical test can be applied, the Jarque-Bera test, to determine whether the observed values are statistically different from the normal distribution. However, this test is quite sensitive and it does not take too much of a departure from the normal values to generate a significant test result. Using our test data, we found a slightly negative skew around -0.3 and a kurtosis close to 2. Based on this, the Jarque-Bera test produced a highly significant result dismissing the null hypothesis that the residuals follow a normal distribution. However, the question remains as to how much impact this has on any tests of parameter significance.

To explore this a little further, we undertook a standard t-test calculating the p-values of the 200 terms in our fitted model. From this, we determined there were seven terms whose significance might be questioned. If terms are of low significance, then we might expect them to display higher variation in the coefficient estimates from changes to the calibration data set. Returning to the k-fold cross-validation approach we first encountered in section 3.4.2 above, we amended the approach slightly as follows:

1. Fit the model on the full data set.
2. Partition the data set into k segments.
3. Retaining the model structure (selection of basis functions) from step 1, refit the model coefficients k times, excluding a different segment each time.
4. The result for step 4 is k estimates of the coefficient for each model term.
5. Consider the extent of variation observed in the estimates—we used the average absolute deviation as a proportion of the mean.

FIGURE 15: CHECKING PARAMETER STABILITY



Our analysis found very low variation in the vast majority of parameters, though we noted seven where the level of variation was relatively higher. Checking back to our t-test work, we found the same seven terms had been called out by that procedure as having questionable significance. This provides some support that the modest departures from normality of our observed residuals has not significantly disturbed the calculation of the p-values, at least in a relative sense.

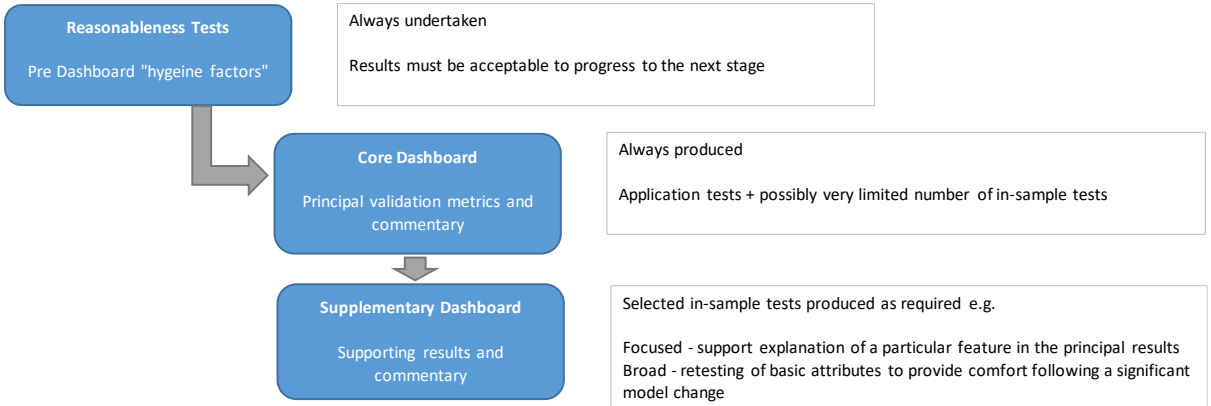
As a final step, we eliminated the seven terms from the fitted model and tested the impact this had on the out-of-sample testing. The outcome was a negligible change in results and in fact a very slight deterioration in the R-squared was noted, indicating there is nothing to be gained by removing these terms. These results provided comfort that the overall model selection and calibration were robust.

4. Organising and presenting the results

Having explored a range of potential tests in section 3 above, we now consider how all that information can be brought together to articulate the validation results and facilitate a decision about whether the current calibration is fit-for-purpose or not.

An obvious point to make is that validation activity has the capacity to generate an awful lot of data and there is real risk of information overload. Hence, we believe it is important to divide the tests into a hierarchy of three different types, as shown in Figure 16.

FIGURE 16: HIERARCHY OF RESULTS



Focusing now on the Core Dashboard, we contemplate the following content:

- Out-of-sample test results in both table and chart form
- Overfitting test results from k-fold cross-validation
- Ranking test results (analysis of top n differences—scenario type, risk factor levels)
- Parameter stability results from modified k-fold cross-validation—useful monitoring test as adverse changes may support the presence of overfitting but also possible introduction of significant correlation between risk factors. The latter possibility can be confirmed or eliminated using the correlation test already described.
- Homoscedasticity—may be worth monitoring for the specific risks prone to this (e.g., interest rate risks) as the extent may vary exogenously with financial conditions.

Figures 17 to 21 illustrate possible content for the Core Dashboard in the context of a single fund using MS Power BI.

FIGURE 17: ILLUSTRATIVE DASHBOARD: OUT-OF-SAMPLE VALIDATION

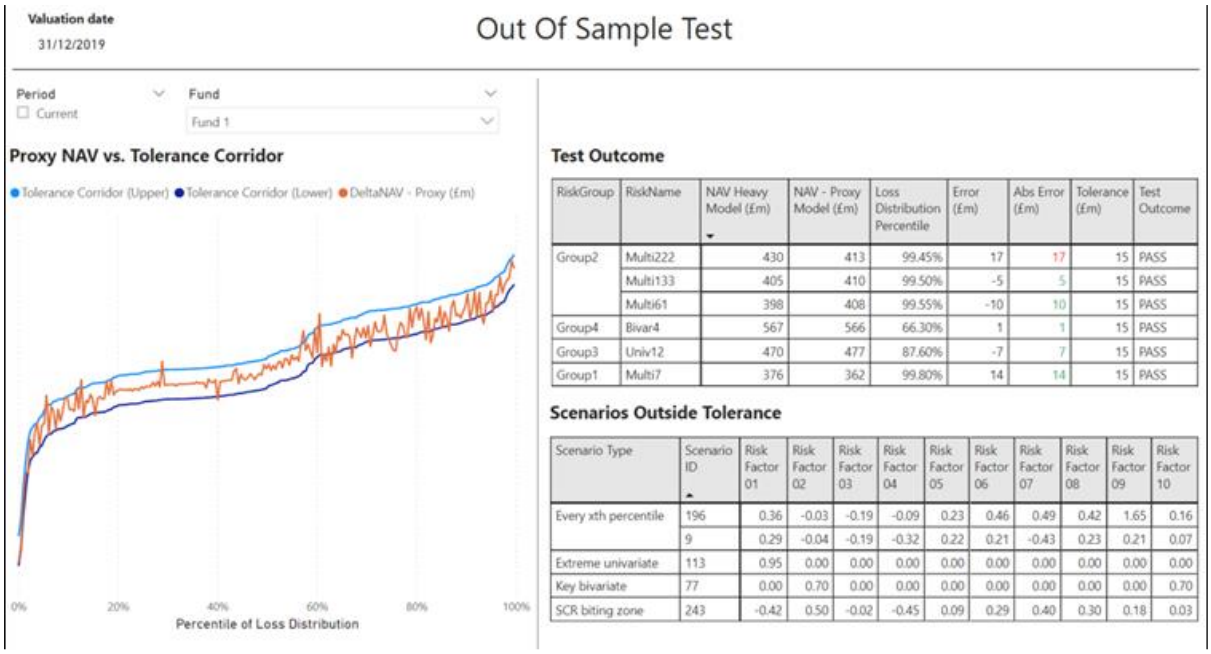
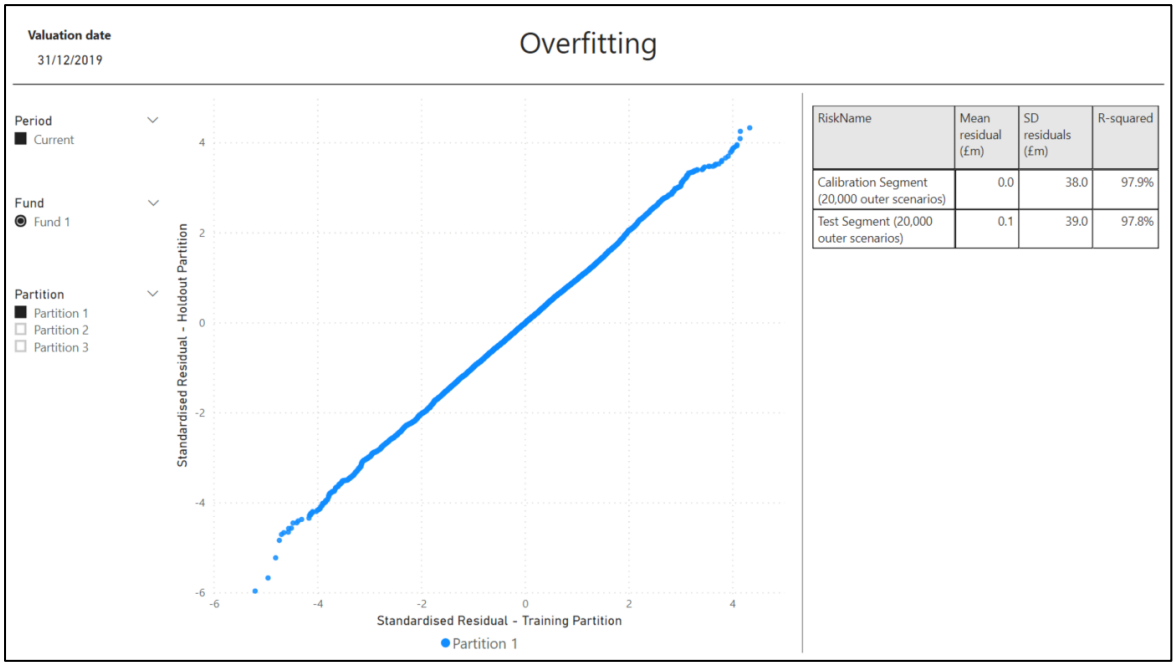
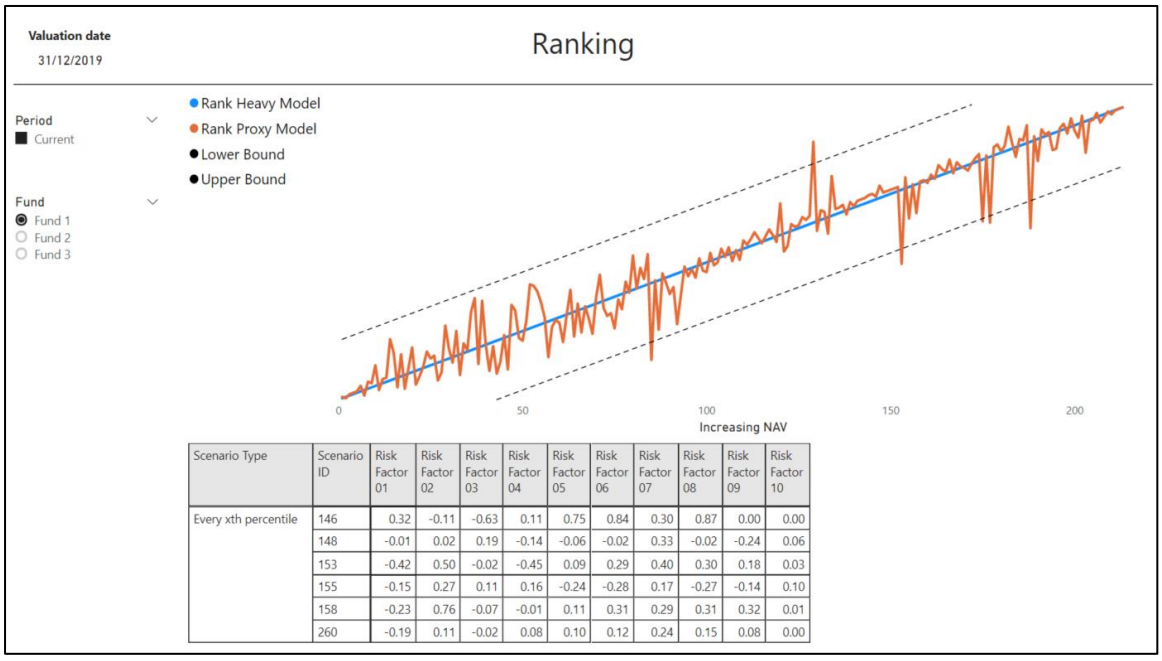


FIGURE 18: ILLUSTRATIVE DASHBOARD: OVERFITTING VALIDATION



RiskName	Mean residual (£m)	SD residuals (£m)	R-squared
Calibration Segment (20,000 outer scenarios)	0.0	38.0	97.9%
Test Segment (20,000 outer scenarios)	0.1	39.0	97.8%

FIGURE 19: ILLUSTRATIVE DASHBOARD: RANKING VALIDATION



Scenario Type	Scenario ID	Risk Factor 01	Risk Factor 02	Risk Factor 03	Risk Factor 04	Risk Factor 05	Risk Factor 06	Risk Factor 07	Risk Factor 08	Risk Factor 09	Risk Factor 10
Every xth percentile	146	0.32	-0.11	-0.63	0.11	0.75	0.84	0.30	0.87	0.00	0.00
	148	-0.01	0.02	0.19	-0.14	-0.06	-0.02	0.33	-0.02	-0.24	0.06
	153	-0.42	0.50	-0.02	-0.45	0.09	0.29	0.40	0.30	0.18	0.03
	155	-0.15	0.27	0.11	0.16	-0.24	-0.28	0.17	-0.27	-0.14	0.10
	158	-0.23	0.76	-0.07	-0.01	0.11	0.31	0.29	0.31	0.32	0.01
	260	-0.19	0.11	-0.02	0.08	0.10	0.12	0.24	0.15	0.08	0.00

FIGURE 20: ILLUSTRATIVE DASHBOARD: HOMOSCEDASTICITY MONITORING

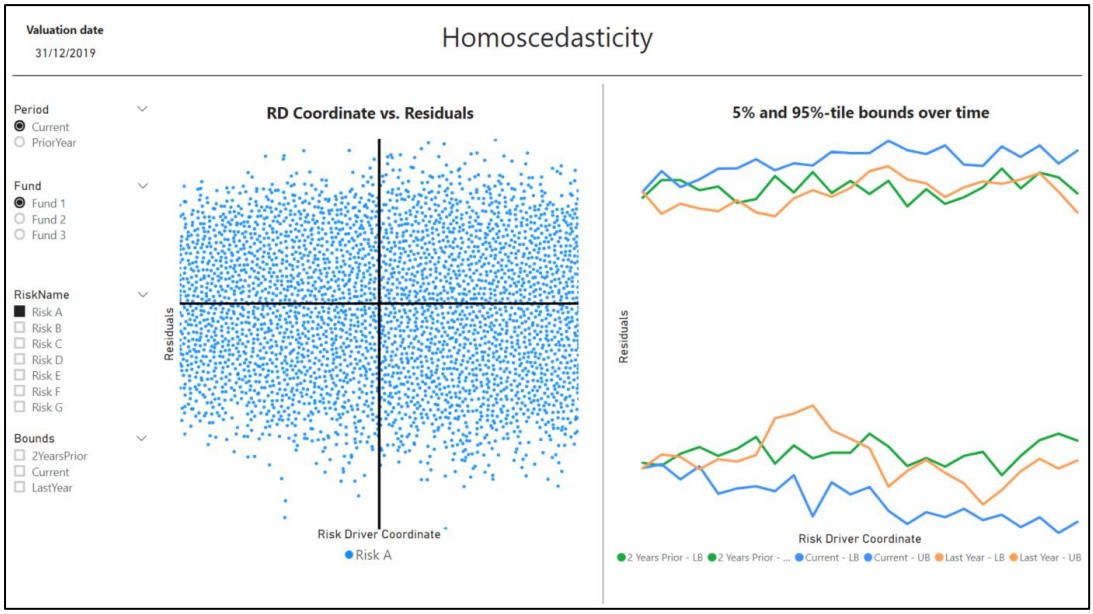
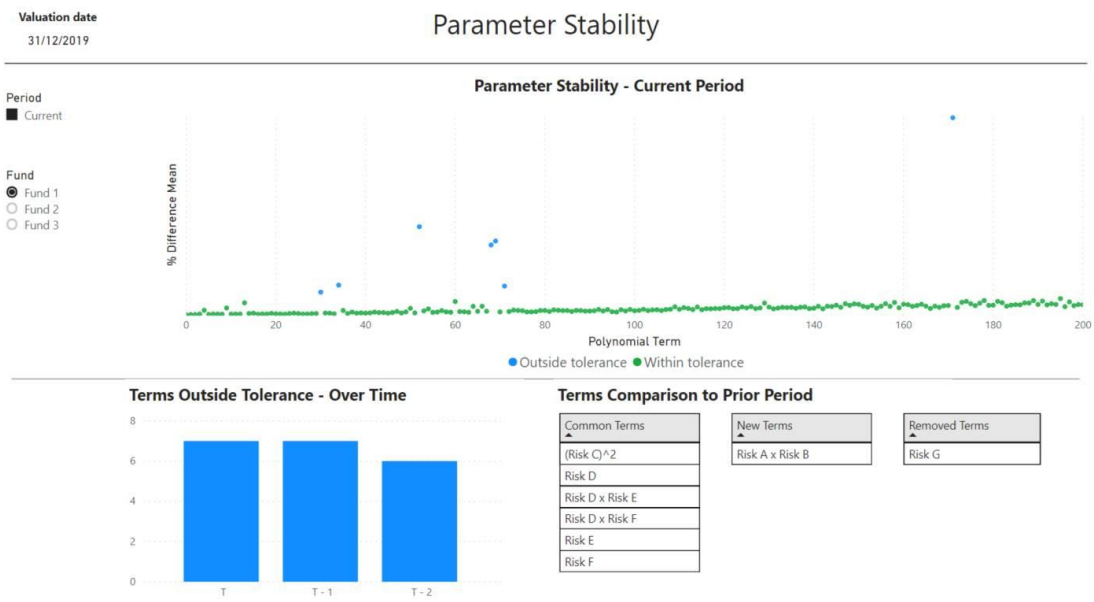


FIGURE 21: ILLUSTRATIVE DASHBOARD: PARAMETER STABILITY MONITORING



The final task for the validator, having been presented with the results, is to conclude whether the calibration is acceptable or not. This is a nontrivial task in itself which requires considerable experience and understanding not only of the proxy model calibration but also of the underlying business. Visualisation of results in the form of a dashboard provides valuable support, in particular:

- The ability to easily switch between funds and to compare results across different funds.
- Being able to compare results over time to examine any developing trends.
- Drilling down from the Core Dashboard into the lower-level more diagnostic tests to investigate particular features observed.

The final decision requires expert judgement based upon the synthesis of the multiple dimensions of the validation results available and full knowledge of the context in which the proxy model is being used.

Milliman is among the world's largest providers of actuarial and related products and services. The firm has consulting practices in life insurance and financial services, property & casualty insurance, healthcare, and employee benefits. Founded in 1947, Milliman is an independent firm with offices in major cities around the globe.

milliman.com

CONTACT
Michael Leitschkis
michael.leitschkis@milliman.com

Russell Ward
russell.ward@milliman.com