British Actuarial Journal (2024), Vol. **29**, e2, pp. 1–49
doi:10.1017/S1357321723000193

SESSIONAL PAPER

Consideration of the proxy modelling validation framework

[Presented at the Institute and Faculty of Actuaries, 1 August 2023, Staple Inn Hall, London]

The IFoA Proxy Model Working Group: Dominic Wollam, Maynard Kuona, Matthew Thomson, Rachael Liu and Stuart Paton

Corresponding author: Dominic Wollam; Email: Professional.communities@actuaries.org.uk

Abstract

Solvency II requires that firms with Internal Models derive the Solvency Capital Requirement directly from the probability distribution forecast generated by the Internal Model. A number of UK insurance undertakings do this via an aggregation model consisting of proxy models and a copula. Since 2016 there have been a number of industry surveys on the application of these models, with the 2019 Prudential Regulation Authority ("PRA") led industry wide thematic review identifying a number of areas of enhancement. This concluded that there was currently no uniform best practice. While there have been many competing priorities for insurers since 2019, the Working Party expects that firms will have either already made changes to their proxy modelling approach in light of the PRA survey, or will have plans to do so in the coming years. This paper takes the PRA feedback into account and explores potential approaches to calibration and validation, taking into consideration the different heavy models used within the industry and relative materiality of business lines.

Keywords: Proxy model; calibration; validation; roll forward

1. Introduction

Since the implementation of Solvency II (European Parliament, 2009) on 1 January 2016, UK insurers are required to calculate their Solvency Capital Requirement (SCR) using either the Standard Formula or an Internal Model (subject to regulatory approval). In order to use an Internal Model in the calculation of the SCR, there are certain minimum standards that must be met: the Use test; Statistical Quality standards; Calibration standards; Profit and Loss attribution; Validation standards; and Documentation standards (all described in Articles 120 to 126 of the SII Directive (2009/138/EC)).

The Calibration standards state that "*Where practicable, insurance and reinsurance undertakings shall derive the Solvency Capital Requirement directly from the probability distribution forecast generated by the internal model of those undertakings, using the Value-at-Risk measure set out in Article 101(3).*" Article 13(38) of the Directive defines the Probability Distribution Forecast as "*a mathematical function which assigns to a set of mutually exclusive future events a probability of realisation*" and clarified further in Article 228(1) of the Solvency II Delegated Regulations which states that "*the exhaustive set of mutually exclusive events . . . shall contain a sufficient number of events to reflect the risk profile of the undertaking*". Guidelines 24 to 27 of the European Insurance and Occupational Pensions Authority ("EIOPA") Guidelines on the use of Internal Models (EIOPA, 2022) provide further guidance on interpretation of "richness of the probability distribution forecast" stressing (inter alia) that "*the probability distribution forecast*

should be rich enough to capture all the relevant characteristics of [an undertaking's] risk profile" and ensure the reliability of the estimate of adverse quantiles is not impaired, whilst "*taking care not to introduce … unfounded richness*".

A number of UK insurance undertakings that use an Internal Model have interpreted this as requiring an aggregation approach which models the different asset and liabilities within the business under a range of scenarios to derive the SCR. The most common approach is the "copula plus proxy model" approach previously discussed in the 2016 IFoA paper "Simulation based capital models testing, justifying, and communicating choices" (IFoA Aggregation and Simulation Working Party, 2016).

Whilst there have been a number of surveys and guidance issued since the 2016 paper, the Prudential Regulation Authority ("PRA") proxy modelling survey in 2019 concluded "*no firm had adopted best observed practice in all areas of proxy modelling*" (PRA, 2019, p. 1). The PRA letter went on to share details of best observed practice. A firm's proxy modelling approach may be dependent on its existing valuation model capabilities, and the nature of the business held may make some of the PRA's best practice areas more relevant to some firms than they are to others. Consequently, firms should focus on developing a proxy modelling approach that it can demonstrate is appropriate for its specific risk profile and robustly justify that its approach is reasonable (noting the PRA feedback within their survey).

The objective of the Working Party ("we", "us", "our", etc.) was therefore to consider the observations raised by the PRA and how businesses can apply this feedback to provide assurance that proxy models are appropriate for use. This is not intended to be an in-depth analysis of the topics discussed but instead provides a framework for implementing this feedback. This framework is intended to provide UK insurance actuaries (and other relevant practitioners) with further guidance on adopting the PRA feedback, enabling them to consider additional steps in providing assurance that the proxy model fit is appropriate. The Working Party experience is primarily from UK Life Insurance firms holding annuities, with-profits and Unit-Linked (operating in both first- and second-line teams) as well as consultancies. This experience spans around eight Internal Model Firms.

We understand that the "copula plus proxy model" method continues to be the most common approach used within Internal Models in the UK life insurance industry. As the main valuation models, commonly referred to as "heavy models", become more efficient, the reliance on proxy models may reduce. However, the Working Party expects that they will remain a key part of insurers' risk management toolkits for some years yet, and advances in technology and the exploration of new techniques could increase the sophistication of proxy models or allow their range of uses to be expanded.

2. Background

The PRA letter provided feedback on proxy modelling but did not contain a definition of a proxy model. Hence, before outlining a framework for calibrating and validating the model, it is important to first set out what is meant by a proxy model (sometimes termed a "lite" model).

This paper defines a proxy model as a model developed to replicate or approximate the output of a more complex model for a given set of input parameters and assumptions for the purpose of Solvency II reporting or other uses, such as solvency monitoring.

Many UK life insurers hold complex assets and liabilities with long durations which are revalued under a large number of simulated scenarios to derive the SCR. As certain liabilities, such as with-profits business with guarantees, are valued using stochastic techniques, it can be resource-intensive to value these under the simulated scenarios using heavy models. Instead, a number of scenarios can be modelled with a proxy model in order to approximate the profits and loss under different scenarios. The proxy model typically is not a single model and instead consists of a

number of proxy functions which describe the changes in assets and/or liabilities under changes in different risk factors. The impact of a single scenario under each of the proxy functions is then aggregated up to approximate the impact on the whole business under the given scenario.

For the proxy model to be appropriate and meet the use test requirements, it needs to be sufficiently representative of the heavy model that it is intending to approximate. It should therefore be subject to a tests to ensure that it is a good fit. Further, given the range of uses of proxy models across insurers and the materiality of model output to areas such as regulatory reporting and risk management, it is vital that the fit itself, and all key underlying assumptions and judgements, are appropriate.

2.1. Annuity Case Study

Throughout this paper, illustrative examples of the different options for various aspects of proxy modelling are set out. These are based on a relatively simple portfolio of fixed (i.e. non-inflation linked) annuity business. While this case study will not reflect the complex nature of the business held by many insurers, its simplicity should allow for the impacts of the different methods and techniques to be more clearly demonstrated.

In particular, the key risks modelled are:

1. Longevity risk (base and trend)
2. Interest rate risk
3. Expense risk (unit cost and expense inflation).

Further details on the case study (including modelling approach, model points and assumptions) are included in the appendices.

2.2. Structure of Paper

In this report, we step through the calibration, validation and considerations for roll forward using the annuity model as an example. The case study example is based on a cash flow projection model[1] with examples differentiated via blue boxes following the relevant section. The structure of the main body of the report is as follows:

Section 4 – Calibration of the model
Section 5 – Validating the fit
Section 6 – Roll forward considerations
Section 7 – Conclusions
Appendices

2.3. Key Definitions

Below we provide definitions for key terms that are used extensively throughout this paper:

- **Scenario:** a scenario consists of a number of stresses to risk drivers which are then used within the proxy model. For example a scenario may just contain an equity stress (and therefore is a univariate stress). Conversely, it may contain multiple stresses. For example, an equity and property fall, lapses increasing and longevity falling. The derivation of these risk calibrations and stresses is outside of the scope of this paper.
- **Proxy model:** a model developed to replicate or approximate the output of a more complex model for a given set of input parameters and assumptions. The general purpose of

[1]Please see Section 2 for more details

developing a proxy model is to be able to produce results that are acceptably close to those that would be produced by the more complex model, in a more efficient way.

- **Full/heavy model:** the more complex model that the proxy model has been designed to replicate/approximate. For many insurers, these models can be costly to run (both in terms of run time and resource required).
- **Calibration:** the process through which the proxy model parameters are determined such that the proxy model output is acceptably close to that of the heavy model, for a given set of inputs.
- **Recalibration:** a further calibration exercise to ensure that the proxy model parameterisation continues to provide results that appropriately replicate the heavy model. This may be carried out as part of a regular process (e.g. quarterly calibrations) or in response to specific trigger events (e.g. a significant movement in financial markets).
- **Validation:** the process of testing a calibrated proxy model against the output of the heavy model. Validation should be carried out using a different set of scenarios than those used to calibrate the proxy model.
- **Risk domain**: the calibration range represented by an n-dimensional space reflecting all possible combinations of the n-risk drivers used within the proxy model. For example, if the proxy model has been calibrated over a range of equity value changes of (40)% to 40% and property value changes from (40)% to 40%, with allowing for interactions, the risk domain can be expressed as the square represented in Figure 1.
- **Roll forward**: the process of updating a calibrated proxy model for certain changes over time without going through a full recalibration (e.g. for movements in certain economic conditions).

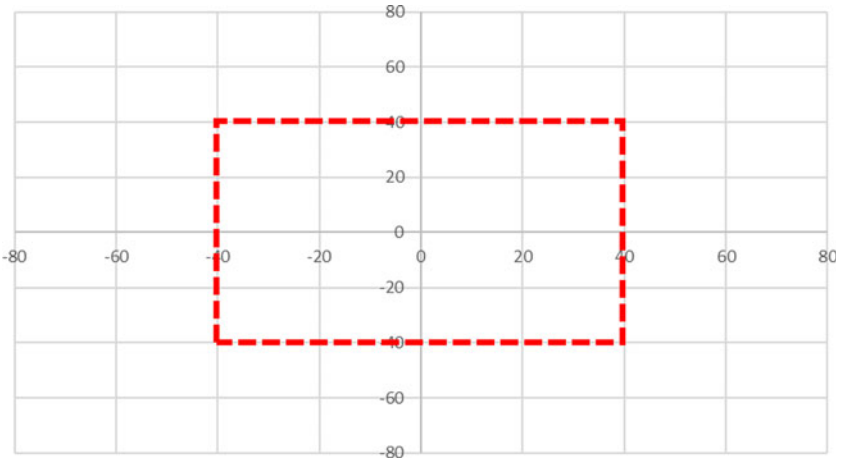

Figure 1. Illustration of risk domain.

3. Methodology

Within this report, an annuity case study is presented which is documented in the Appendix. As the Working Party did not have regular access to a heavy model, a cash flow projection model was developed to project the future benefit and expense cash flows for the portfolio under different scenarios. However, as heavy models can often require significant time to run, a "pseudo-heavy" model was also produced. This used the output of a small number of heavy model runs to allow the user to produce the large number of different scenarios required for calibration and validation of proxy models. The pseudo heavy model uses risk driver coefficients calibrated to the full heavy

model results to model cash flows under different scenarios. Risk drivers of up to the sixth order are used within the model.

The pseudo heavy model allows a set of adjusted cash flows and present values to be produced instantaneously for a scenario with a specified set of risk factors, removing the need for the full model to be run. Whilst this enables analysis to be produced efficiently, it should be noted that the examples are therefore artificial, and we do not expect proxy models used in practice to necessarily show as good a fit.

4. Literature Review

As implied by the results of the PRA's industry survey into best practice, there is currently no literature which provides a comprehensive analysis of proxy modelling, nor the key topics raised within the PRA's letter. This report is intended to, at least in part, address this, and builds on both "Simulation-based Capital Models Testing, Justifying, and Communicating Choices" (IFoA Aggregation and Simulation Working Party, 2016) and "Heavy Models, Lite models and Proxy Models paper" (IFoA Proxy Model Working Party, 2014). These papers represent the foundation of the work outlined within this report and have been supplemented by industry surveys and other IFoA presentations. The focus of this paper is therefore on methodology currently in use within the industry. It is not intended to be a comprehensive review of developments within proxy models. Where literature has been used to inform the discussion, these are included as footnotes enabling the reader to cross-reference the underlying detail.

In combination, these papers outline the development of proxy models prior to and following the implementation of Solvency II and in the run-up to the PRA survey.

5. Calibration

5.1. Background

The initial calibration and subsequent regular recalibrations of proxy models are vital in ensuring that the models are an appropriate proxy for the heavy models that they are intended to replicate. In developing a proxy model, the approach taken for scenario selection (i.e. identifying the specific scenarios to be used for calibrating the proxy model) and the method of fitting the proxy model to the calibration scenarios, including the choice of the form of the proxy model, are potentially the most significant decisions.

This section provides illustrative examples of some of the more common proxy modelling approaches used in the UK, which will be assessed using the various validation tests in Section 5. In this section, we explore different approaches to scenario selection, and to fitting proxy models to the same case study. Their performance is compared using the same out of sample scenarios to illustrate the strengths and weaknesses of each approach.

There are a variety of options available to insurers, and the most appropriate choices depend on several factors, many of which are specific to the insurer and to the exercise being undertaken. The PRA paper noted that there are a variety of approaches taken to proxy modelling fitting across the UK insurance industry and that the PRA has no preferred method, provided insurers can demonstrate the appropriateness of their chosen approach for the relevant uses. One key factor that will impact both scenario selection and fitting method decisions is the expected usage of the proxy model.

Common uses of proxy models, as identified by the PRA, include:

- Economic capital modelling, including calculating regulatory capital requirements (e.g. the Solvency II SCR)
- Business management and decision making

- Sensitivity/stress/scenario testing
- Setting of risk appetite and capital allocation
- New business pricing
- Forecasting
- Liquidity management.

5.2. Fitting Models

Usually, the problem of UK Solvency II firms when fitting a proxy model is to solve a system of linear equations. In particular, the equation to be solved is of the form:

$$y_i = x_i^T \beta_i + \varepsilon_i$$

such that we minimise the sum of squares of the residuals (ε_i), where β_i is unknown for each variable and:

1. y_i represents a single response variable
2. x_i represents a vector of explanatory variables (terms in the model)
3. ε_i represents the error (difference between the predicted and the actual response).

This can also be represented in matrix form as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

For this, we need at least as many calibration points as we have terms in our model (including the intercept term) to have a unique solution to the system of linear equations. The objective of the least squares fitting can be written as follows:

$$\min_{\beta} \left\{ \sum_{i=0}^{N} \left(y_i - x_i^T \beta_i \right)^2 \right\}$$

While linear systems can be solved directly if the form of the polynomial is known (e.g. if the model is quadratic), one of the additional objectives at the model fitting stage is also typically to determine the appropriate form of the model (i.e. to choose between different possible proxy models). We have investigated the following primary methods of calibrating polynomial proxy models:

- Ordinary Least Squares
- Automated model selection (stepwise methods and others).

5.3. Penalised Regression

The outcome of each of these model selection methods is a set of coefficients for the model terms (i.e. the β in the above equations), which in turn specifies the models completely. Examples of methods that can be used for fitting the model are provided in Section 5.2 and the method of fitting the model will in turn influence the selection of scenarios for the models.

5.3.1. Selecting calibration scenarios

The shape of the proxy model, and therefore the quality of the fit of a calibrated proxy model for a given purpose, will be dependent on the selection of scenarios used to calibrate the model. The paper "Efficient Curve Fitting Techniques" by Hursey and Scott (2012) discussed at length the methods of choosing calibration scenarios that achieved a given quality of fit for the minimum

number of calibration scenarios. The paper demonstrated that any required quality of fit can be achieved with the minimum number of scenarios by selecting the scenarios in a way that maximises the quality based on the chosen metric. In particular, the calibration points implied by Legendre nodes produce a fit that (approximately) minimises the average approximation error while Chebyshev nodes produce a fit that (approximately) minimizes the maximum approximation error.

Key considerations in scenario selection include:

- The risk domain of the proxy model
  - This is the range of variables over which the proxy model will be evaluated.
  - The proxy model will be used to evaluate the underlying assets and/or liabilities over a range of potential inputs. All other things being equal, a smaller risk domain should produce smaller (absolute) approximation errors, while a larger risk domain produces larger (absolute) approximation errors.
  - The PRA paper noted that firms demonstrating best practice in model fitting selected a large number of fitting points across the entire domain and supplemented these with additional judgement-based scenarios at specific points of interest. The Working Group believes that this is vital in ensuring the fit is appropriate and considers different uses of the model. The domain should therefore be justified and reconsidered prior to each recalibration of the proxy models.
- Model purpose
  - Proxy models have multiple uses, and different calibrations of the proxy models may lead to models that are optimised for specific purposes at the expense of their wider use (e.g. the SCR at a single point in time). There is typically a trade-off between the different uses of a model. As such, calibration scenarios may be chosen that focus the fit on specific areas of the overall probability density function at the expense of others. For example, choosing a higher proportion of calibration scenarios around the 99.5^{th} percentile for material risks in a model intended to calculate the SCR. In particular, increasing the domain will help to ensure that a model produces smaller residuals over a wider range of possible stresses and may therefore result in a better model for stress and scenario testing. Conversely, this may also make the approximation of the reported SCR less accurate (given the wider domain fit considerations).
  - The PRA paper noted that firms demonstrating best practice in model fitting selected their fitting points by considering all uses to which the model is put, not just focussing on those required to get a good fit around the SCR scenario. The Working Group would note the increased focus on specific stresses, climate-related disclosures and solvency monitoring.
- Complexity
  - More complex proxy models will require a higher number of calibration scenarios (i.e. the more terms in the proxy model, the more calibration scenarios required). Constraints on the number of scenarios available may limit the complexity of the proxy model.
  - In addition, the composition of the calibration scenarios themselves will reflect the terms in the model (i.e. for a given risk driver, the higher the order of terms in the model, the more stresses required).
  - The complexity of the business being modelled is also an important consideration. The PRA paper noted that allowing for the impacts of factors such as options and guarantees, or the behaviour of the Matching Adjustment under stress, will introduce interactions within the proxy model. Firms demonstrating best practice assessed such interactions from first principles when thinking about fitting models. We believe this analysis is particularly important for writers of with-profits business which are typically more complex and include management actions.

- Efficiency
  - Careful selection of the calibration scenarios will enable practitioners to produce a better fit for a specific number of calibration scenarios or to achieve a given standard of fit with fewer calibration scenarios.
  - Different scenario selection methods will be more or less efficient than other selection methods. For example, random scenario selection is likely to be less efficient than some of the interpolation methods suggested by Hursey and Scott (2012).

Examples of the methods that can be used to select calibration scenarios are set out in Section 5.4.

5.3.2. Automation of model and scenario selection

A number of companies are currently developing methods to automate the process of model and scenario selection. The advantages of automating model and scenario selection include:

- Making the process repeatable
- Increasing the efficiency of the process, by removing unneeded complexity
- Reducing the use of expert judgement in the process, or at least codifying the selected rules. The expert judgements can then be applied at a more macro level, e.g., choosing the number of terms in the model, deciding which risk drivers require more complex polynomials and determining the domain of the function
- Making the process more auditable, with clear criteria to demonstrate the appropriateness of the scenarios selected.

While automated selection can be useful for certain selection methods, approaches that rely heavily on expert judgement are harder to automate. The model error introduced through automation processes should also be managed (particularly if AI is used). Whilst PS6/23 (Model Risk Management Principles for Banks) does not currently apply to insurers, this provides an overview of how model risk should be managed by firms.

5.4. Scenario Selection

This section explores three potential methods for selecting the scenarios used to calibrate the proxy model. These methods are:

- Expert judgement based
- Precise interpolation
- Random (or quasi-random) sampling.

5.4.1. Expert judgement based

In this method, the selection of scenarios is informed by the expert's knowledge of the statistical distribution of the risk drivers, their views on the interactions expected between risks, and the complexity of the underlying assets or liabilities being modelled. Typically, the fitting points are chosen from the risk distributions, for example, by choosing the 1-in-200 percentile of the distribution.

The expert will typically employ a heuristic approach to choose the scenarios, with the following being an example of a possible approach that could be taken:

- Identify the domain of the proxy model, taking into consideration the above points.
- Choose the appropriate order of the polynomial terms for each individual risk driver. The order of the polynomial will likely reflect the materiality of the underlying risk drivers, with more material risk drivers requiring higher-order polynomials.
- For each individual risk driver, choose points in the domain that are appropriately spaced apart, with possible choices (among others) being:
  - Points spaced equally across the domain for that risk driver.
  - Points that represent chosen percentiles on the risk distribution, which may also represent points deemed important for specific business uses.

This approach can be extended to chosen interaction terms. The advantages and disadvantages of the expert judgement-based approach are shown in Figure 2.

Advantages of the approach	Disadvantages of the approach
• It is relatively easy to understand and explain to senior management. • It is relatively easy to implement and does not require development of special processes to perform. • It leverages expert judgement to avoid selecting scenarios not material to the calibration, e.g., combinations of risk drivers that are not expected to produce interaction effects.	• The reliance on expert judgement can introduce bias, and decisions made by the expert can be difficult for practitioners to justify. • The process may not be replicable. In particular, different practitioners will make different judgements and may select different looking calibration scenarios. • Scenarios chosen using judgement may not be efficient and can produce unnecessarily large approximation errors. As an example, when using equidistant calibration points, it has been shown by Runge (1901) that the use of higher order polynomials is not guaranteed to improve accuracy. • Needs to be reviewed, and potentially updated, regularly.

Figure 2. Advantages and disadvantages of the expert judgement-based approach.

5.4.2. Precise interpolation

With precise interpolation, the proxy models are designed to pass through specifically chosen calibration points by using as many distinct calibration points as there are terms in the polynomial. For example, when fitting a quadratic polynomial, three points would be required, and the polynomial would pass through all three points. When the calibration points are selected based on Legendre or Chebyshev nodes[2], then the fitted proxy models will be optimal in the sense of approximately minimizing the average or maximum approximation error. The advantages and disadvantages of this approach are discussed in Figure 3.

[2]The Chebyshev nodes minimise the maximum error, while the Legendre nodes attempt to minimise the average error (although this assumes the risks are uniformly distributed across the range). The average (or maximum) error is only approximately minimised – the result would be exact if we were estimating one polynomial with a simpler polynomial (Hursey and Scott, 2012).

Advantages of the approach	Disadvantages of the approach
• The method only requires the user to decide the appropriate domain of the loss function and the appropriate order of the loss function. • The determination of the calibration points, which are the roots of the Legendre or Chebyshev polynomials (depending on the chosen method), is straightforward and can be automated. • The approach can be efficient and produce the smallest number of distinct calibration scenarios required to fit a specific polynomial while producing the smallest error possible over the selected domain. • It is possible to select validation scenarios that are complementary to the calibration scenarios. The method of selecting the validation scenarios will choose points where the model error is expected to be maximised. This can provide further evidence of the appropriateness of the selected calibration scenarios.	• The method relies on choosing an order for the loss function and therefore relies on a key expert judgement. If the order of the polynomial chosen is too low to produce an acceptable fit, the calibration points determined may not be optimal for the loss function being fitted. • The scenarios chosen are unlikely to be an appropriate fit for a different polynomial. Therefore, if the fit is demonstrated to be inadequate, and a different polynomial is required, the previously chosen calibration points will no longer be optimal and new calibration points will need to be selected. This reduces the flexibility of the approach and could also result in significant additional effort being required from practitioners. • This method of scenario selection does not work with automated model selection methods such as stepwise regression or regularised regression.

Figure 3. Advantages and disadvantages of the precise interpolation approach.

5.4.3. Random (or quasi-random) scenario selection

Calibration scenarios can be produced by generating scenarios randomly rather than picking precise interpolation points or using expert judgement. In general, when scenarios are generated randomly, their number is usually significantly in excess of the number of terms expected in the models. We would expect to have in the order of 10 times or more (Steyerberg *et al.*, 2001) calibration scenarios compared to the number of terms in the models.

When random scenario generation is used, the variables can be sampled from any chosen distribution. In practice, the sampling tends to be based on a uniform distribution rather than the expected sampling distribution of the underlying random variables. The use of uniformly distributed random variables produces better coverage of the domain of the proxy model.

The use of quasi-random numbers, such as Sobol sequences, rather than fully random scenarios can ensure a more uniform coverage of the domain of the proxy model. In comparison to random number generation, which will result in clusters of random numbers, the use of quasi-random numbers does not produce such clustering, which may remove some of those artefacts from the modelling process. The advantages and disadvantages of this approach are discussed in Figure 4.

This method of scenario selection works with ordinary least squares model (OLS) fitting, stepwise and regularised regression, which are covered in more detail in the Section 5.5.

5.5. Methods of Fitting Models

5.5.1. Ordinary Least Squares (OLS)

This method directly solves the linear regression model to produce coefficients for the selected polynomial. There are many algorithms for performing this fitting, for example using a factorization of the matrix form of the least squares problem or using a numerical algorithm such as gradient descent. However, given a specific form of the model, and a specific set of calibration points, the model chosen will be unique, provided that an adequate number of distinct calibration points have been selected. The advantages and disadvantages of this approach are discussed in Figure 5.

Advantages of the approach	Disadvantages of the approach
• Using random scenario selection avoids bias in selecting scenarios (unlike other scenario selection methods where the biases of the expert may influence the scenarios chosen). • The scenarios generated will generally be appropriate to use for any model and any method of model-fitting. Therefore, generating calibration scenarios in this way provides the maximum amount of flexibility to define the form of the proxy model as well as the method of calibrating it, provided the overall number of calibration scenarios is adequate. • The large number of calibration scenarios allows the efficient investigation of different polynomial models and the use of automated model selection. Randomly selected scenarios are not optimised for specific models but can be considered as more optimal for exploring different candidate models. • When using randomly generated scenarios, calibration strategies can be devised that ensure that the chosen models and calibrations produce good out-of-sample model fits. An example of such a strategy is using cross-validation to pick the best form of the model, where models are selected based on the cross-validation performance.	• The approach of using (quasi-) randomly generated calibration points is generally not efficient and requires more calibration points to achieve a comparable fit to using interpolation points. As discussed, where precise interpolation will require only as many points as there are terms in the model, when using randomly generated points, a large number is required to ensure adequate coverage of the domain of the loss function.

Figure 4. Advantages and disadvantages of the random scenario selection approach.

Advantages of the approach	Disadvantages of the approach
• This approach to fitting proxy models is simple to explain and implement and is generally familiar to most actuaries (and other practitioners). • This approach is well suited when the proxy models are simple, i.e., have a small number of terms and require a relatively small number of calibration points to achieve an acceptable fit. • If the required or appropriate form of the proxy model is known, then only as many fitting points as there are terms in the model need to be specified. Therefore, this method can be very efficient from a heavy model run time perspective. As discussed by Hursey and Scott (2012), if the required form of the model is known, then scenarios can be chosen that are shown to achieve the best quality fit, subject to the form of the model being a priori appropriate.	• The use of the ordinary least squares model fitting requires both the appropriate form of the model to be known in advance, and to have scenarios that are appropriate for that form of the proxy model to achieve the appropriate fit of the proxy model. If either the form of the model is not appropriate, or the scenarios available are not appropriate, then the model may not achieve an appropriate fit. • Once a specific set of scenarios are chosen which are appropriate to a specific form of proxy model, then those scenarios are generally not as suitable to use to calibrate other proxy model forms.

Figure 5. Advantages and disadvantages of the OLS approach.

5.5.2. Automated model selection (including stepwise methods)

Automated model selection attempts to address two issues simultaneously: (1) choosing the appropriate form of the model and (2) producing the best possible fit for that model.

One of the key optimisation challenges with model selection is ensuring the models chosen generalise well and have good predictive performance. Generally, adding terms to a model allows it to achieve an improved "in-sample" fit (compared to a simpler model). However, this could also lead to poorer out of sample performance. Since, by design, proxy models are intended to be used with out of sample scenarios, the models should not be overly complex and therefore not generalise well. Therefore, the criteria for an improved fit need to be adjusted to ensure that more complex models are only chosen where they are demonstrated to generalise better than the simpler models.

One way to do this is using information criteria, usually the Akaike Information Criterion (AIC) or the Bayes Information Criterion (BIC). The AIC and BIC are defined as follows:

$$AIC = 2k - 2\log\left(\hat{L}\right)$$

and

$$BIC = k\ ln(n) - 2\log\left(\hat{L}\right)$$

where the aim is to minimise these results and:

- n is the number of data points
- k is the number of terms in the model
- $\hat{L}$ is the likelihood.

The AIC and BIC penalise the addition of new terms to the model, and therefore, all other things equal, prefer simpler models. The key difference between AIC and BIC is that BIC has a larger penalty for more complex models where the number of points (n) used to fit the model is larger than 7.

An automated model selection process would calculate the AIC or BIC for the current model and for any adjusted models. If the AIC or BIC is lower for any of the adjusted models, this would demonstrate that the model has improved because of the adjustment.

With automated model selection, it can be useful to consider the set of possible terms, and the possible models that can be built. For example, a liability with 5 risk drivers who are all being modelled with up to quadratic terms, and allowing all possible interactions between any two pairs of univariate terms, would have 51 possible terms within the model (including the intercept term). In this example, it follows that the number of possible different models that can be built from the different combinations of including/excluding each of the 51 terms is 2^{51}.

Given the extreme number of possible models, finding the most optimal model may not be feasible. Instead, practitioners may consider there to be a set of acceptable models and seek to identify a model within this set. The result of this would be the selection of a model that meets the needs of the user but might not be the "best" model out of the 2^{51} possible options. Automated model selection generally requires many calibration scenarios and employs a regression rather than an interpolation strategy.

With a regression strategy, many calibration scenarios are required compared to the number of explanatory variables. Harrell (2022) suggested at least 15 observations (or fitting points) per term in the model as a rule of thumb. However, other authors have suggested even larger ratios when using automated selection procedures to avoid the issues with using stepwise procedures.

Alternatively, in an interpolation strategy the points chosen are those that the model "surface" should pass through. Therefore, the number of points required is limited to the number of terms within the model.

5.5.3. Exhaustive search

An exhaustive model search considers every possible model in a candidate set and chooses the model that achieves the best fit. However, this approach is only useful when the number of possible model terms is relatively small. As noted above, even a relatively small number of terms can produce a large number of candidate models, and an exhaustive search becomes infeasible. Therefore, alternative ways of automatically selecting models are often required to make the process feasible.

5.5.4. Stepwise model selection

One approach for performing this automated model selection is the stepwise algorithm. With this approach, an initial model is repeatedly adjusted in a step-by-step manner with the new models created then compared against the initial model. There are generally two ways of implementing a stepwise algorithm: forward or backward. The forward approach starts with a relatively simple model and adds terms to improve the model (e.g. based on AIC or BIC). The backward model starts with a relatively complex model and removes terms to improve the model. A typical implementation of a (forward) stepwise model would work as follows:

- A starting model is chosen and the AIC or BIC is calculated
- A set of new models is generated by adding one term to the current model. All possible additional terms (that are not already in the model) are tested in this step
- The model with the lowest AIC or BIC is then chosen and this will be either:
  - the original model, at which point the algorithm terminates
  - one of the new candidate models, which then triggers a further step to test an additional term.

The backward algorithm works in a similar way except that it removes terms. A further modification of the algorithm will either add or remove terms, i.e., is bidirectional, but otherwise follows the same overall principles as either the forward or backward algorithms.

Compared with the exhaustive search, the stepwise algorithm potentially investigates a much smaller set of models. For example, in the example discussed above with 5 risk drivers and 51 possible terms, a stepwise algorithm may search less than 2,500 models (based on a bi-directional stepwise search considering 50 models in each step until it considers the full model). This compares with the circa 2^{51} models that one would have to search through for an exhaustive search.

We note that there is literature, including Smith (2018), that discourages the use of stepwise procedures for model selection. The key concern noted is that a model calibrated using a stepwise procedure will perform poorly out of sample. We note that for proxy modelling applications, this risk can be mitigated by using out of sample testing to detect the lack of a good fit, and therefore that this concern can usually be mitigated in practice.

5.5.5. Other automated model selection algorithms

There are other automated approaches to selecting models such as genetic algorithms (also referred to as evolutionary algorithms). One advantage of genetic algorithms is that they can more easily identify combinations of terms that are associated with good models. However, the key challenge is the selection of the "hyperparameters" that control how the algorithm performs its search. Examples of hyperparameters used with genetic algorithms include:

- Mutation rate
- Cross-over points
- Elitism.

Advantages of the approach	Disadvantages of the approach
• Can allow practitioners to explore relationships in the fitting data that they may not have considered based on expert knowledge alone. This may allow the models to capture risks that had not previously been identified or known. • Automated model selection also reduces the dependence of the quality of the fit on the knowledge and expertise of the practitioner. This may be very important where the most appropriate form of the proxy model is not stable from one period to the next, and it would otherwise be a resource intensive process to manually explore, test, and choose a new form of the proxy model for each calibration exercise. • Automated model selection can, by its nature, be automated in the calibration process.	• The key disadvantage of automated model selection is the number of scenarios (or fitting points) relative to the number of terms in the model. As discussed above, the use of regression strategies requires the number of calibration scenarios to be an order of magnitude higher than the number of terms in the model, and the use of automated model selection only increases that ratio. • Other disadvantages include the possibility that automated model selection may produce biased parameter estimates (and therefore biased models). In particular, a stepwise selection method may produce a model that appears to fit well by chance, with the chosen model not generalising well (i.e., being predictively poor). • It can be very difficult to find an appropriate set of hyperparameters that work well for a specific modelling problem. The most appropriate hyperparameters will be different for each modelling problem, and therefore requiring additional expertise e.g., in the specific genetic algorithm as opposed to more general actuarial knowledge.

Figure 6. Advantages and disadvantages of automated model selection approaches.

Broadly speaking, the genetic algorithm is as follows:

1. Start with an initial population of n models. The initial set of models can just be a set of identical models with only the intercept term.
2. Generate a new population of models using one or more genetic model operations. For example:
   a. Randomly select two models from the initial population of models. The models are selected based on their fitness, with the best fitting (lowest AIC/BIC models) more likely to be chosen.
   b. Apply mutation to add or remove terms from the models. This is equivalent to flipping a bit from 0 to 1 (adding a term) or 1 to 0 (removing a term) in the binary representation of the model. Typically, only a small proportion of the terms (e.g. 5%) are changed (added or removed).
   c. Apply cross-over (via an interaction term or a mixture of terms within the model), where two models are mixed, with terms picked from either of the two models that are being mixed.
3. Where "elitism" is also in use, a number of the best models are carried forward from the previous step. This ensures that at each subsequent stage, there is no regression in the best models, i.e., the best models identified so far are always kept.

Each of the models in the new population is then assessed for fitness, the new models chosen and are then subjected to the same set of procedures to generate new models. The advantages and disadvantages of this approach are shown in Figure 6.

5.5.6. Penalised regression

Another method of calibrating proxy models is to use penalised regression: least absolute shrinkage and selection operator (LASSO), ridge regression or elastic nets (a combination of LASSO and ridge regression). While the objective of penalised regression is similar to that of ordinary least squares or automated (stepwise) model selection, there is one key difference in how this is achieved. Penalised regression starts with the same modelling objective as before and attempts to minimise the sum of the square differences, subject to a constraint. In this case, the constraint is that the total sum of the coefficients must not exceed a set number, t (see below).

$$\min_{\beta}\left\{\sum_{i=0}^{N}\left(y_i - x_i^T\beta_i\right)^2\right\} \text{ subject to } ||\beta||_p \leq t$$

This can be relaxed into penalisation optimisation as follows:

$$\min_{\beta}\left\{\sum_{i=0}^{N}\left(y_i - x_i^T\beta_i\right)^2 + \lambda \times ||\beta||_p\right\}$$

where

- $||\beta||_p$ is the appropriate norm
- λ is the regularisation/penalty parameter
- N is the number of fitting points/calibration scenarios.

The other variables are as before.

The key idea behind penalised regression is to introduce a "tension" between minimising the sum of squared error (the left part of the expression) and the penalty term. In particular, whilst the sum of squared error will reduce as terms are added to the model and/or as the coefficients increase, the right part of the equation will increase as the new terms give rise to non-zero coefficients or as those coefficients increase. The explanatory variables in regularised regression tend to be normalised in practice to ensure that the size of the coefficients reflects the importance of the explanatory variables rather than their scale.

The use of the L1 norm ($||\beta||_1$) (sum of the absolute value of the coefficients) gives rise to LASSO regression, while the L2 norm ($||\beta||_2$) (square root of the sum of squares) gives rise to ridge regression. Both LASSO and ridge regression "shrink" the coefficients, with the amount of shrinkage being controlled by the "lambda" parameter. The key difference between LASSO and ridge regression is how they shrink the coefficients as the lambda parameter is increased:

- LASSO regression shrinks all the coefficients by the same amount in absolute terms (until they reach zero). Therefore, LASSO regression also performs variable selection, with larger values of the lambda parameters eventually resulting in many or all of the coefficients being reduced to zero, i.e., the terms are removed from the model.
- Ridge regression shrinks all coefficients by the same proportion. This means that ridge regression does not reduce any coefficients to zero and therefore does not perform variable selection.

Regularised regression trades off between bias and variance (see Figure 7). Bias refers to the tendency of the model to produce estimated results that differ from the inputs (fitting error), while variance refers to the tendency for the model to not generalise well due to the values produced by the model changing too quickly relative to the change in the input/explanatory variables.

In Figure 7, the bias increases as the value of lambda increases (and the model becomes simpler). The optimal model is one where the total error (bias plus variance) is minimised (green shaded area). The advantages and disadvantages of this approach are shown in Figure 8.

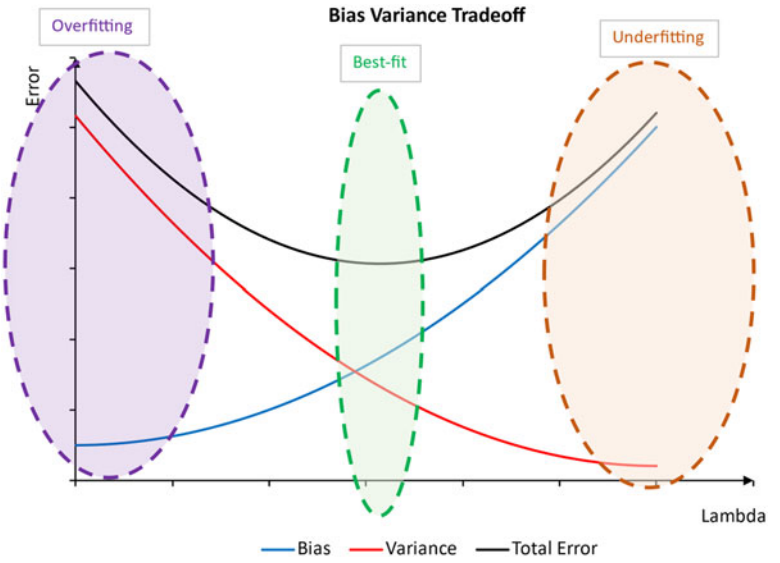

Figure 7. Illustrative bias and variance trade-off.

Advantages of the approach	Disadvantages of the approach
• Regularised regression is an alternative to automated model selection, and also allows both variable selection and the model fitting to happen simultaneously. • Regularised regression explicitly trades off between bias and variance. • Using cross-validation, one can choose a value of the regularisation parameter (λ) that maximises the predictive (out-of-sample) performance of the model by testing a number of different values of lambda and choosing the value that minimises the cross-validation error.	• Shrinking the coefficients through the use of regularisation introduces bias into the model. However, strategies such as the relaxed LASSO approach reduce/remove the bias by using regularisation to perform feature selection, and then using the non-regularised fit to perform predictions.

Figure 8. Advantages and disadvantages of regularised regression approaches.

5.5.7. Least Squares Monte Carlo (LSMC)

This method of calibration is almost exclusively used for the proxy modelling of stochastic liabilities. The key innovation of LSMC is that where a typical valuation of a single stochastic quantity (e.g. BEL) uses a large number of "inner" scenarios (over 1000) to ensure convergence in the asset or liability valuation, the LSMC approach uses a large number of stressed "outer" scenarios with each outer scenario having a smaller number of inner scenarios (e.g. 2–20 scenarios). This essentially means that instead of a small number of accurate scenarios you have a large number of approximate scenarios. The valuation in each outer scenario will be inaccurate due to the lack of convergence given the small number of scenarios. However, this is compensated

Advantages of the approach	Disadvantages of the approach
• The method is efficient and allows the generation of a large number of calibration scenarios (stresses) to fit the model. In particular, when attempting to fit proxy models to liabilities with complex behaviour LSMC allows good coverage of all such scenarios and ensures that the behaviour in those scenarios can be captured. • The large number of calibration scenarios that is produced for LSMC allows the efficient investigation of different polynomial models and the use of automated model selection or regularised regression methods. This reduces reliance on expert judgement as the primary way for a practitioner to determine an appropriate model structure, as alternative model structures can be explored as part of the fitting process.	• The primary disadvantage of this method is that it is not particularly applicable or useful for deterministic liabilities where there is no sampling error. • This method may also require special tools (e.g., an economic scenario generator-type tool) for the generation of the LSMC scenarios as well as potential adaptations to liability models to ensure the generation of appropriate outputs for the LSMC model fitting.

Figure 9. Advantages and disadvantages of the LSMC approach.

for by having a large number of the approximate outer scenarios. There are therefore two primary sources of error when fitting a proxy model using LSMC:

- The sampling error (or noise) due to the lack of convergence of the individual calibration points. This error can be minimised by increasing the number of inner scenarios and achieving greater convergence in the calibration points. The LSMC fitting algorithm "aims" to ignore the noise. Overfitting will occur if/when the fitted proxy function starts to fit the noise or sampling error.
- The approximation error refers to the difference between the estimated value and the true converged value of the quantity being estimated. The approximation error can be minimised by choosing a more complex proxy model, which would, in turn, require a larger number of outer scenarios. There is no limit to the extent to which the approximation error can be reduced, although practical considerations will usually dictate that a significant amount of approximation error will remain.

The LSMC approach aims to reduce the approximation error as much as is reasonable while avoiding fitting to the "sampling error". In practice, we cannot distinguish between the approximation error and the sampling error in the fitting data. However, the LSMC fitting algorithm can mostly ignore the sampling error to provide a reasonable fit to the underlying quantity being estimated, therefore minimising the approximation error. The advantages and disadvantages of this approach are shown in Figure 9.

5.6. Calibration Examples

This section shows examples of the use of the calibration and scenario selection methods set out above for the annuity case study (except for the LSMC approach which we illustrate using a different example in the appendix).

Note that for the stepwise, LASSO and genetic algorithm approaches, we have calibrated the models using the same "universe" of potential model terms, and therefore the differences will mostly reflect the strengths and weaknesses of the different approaches rather than different

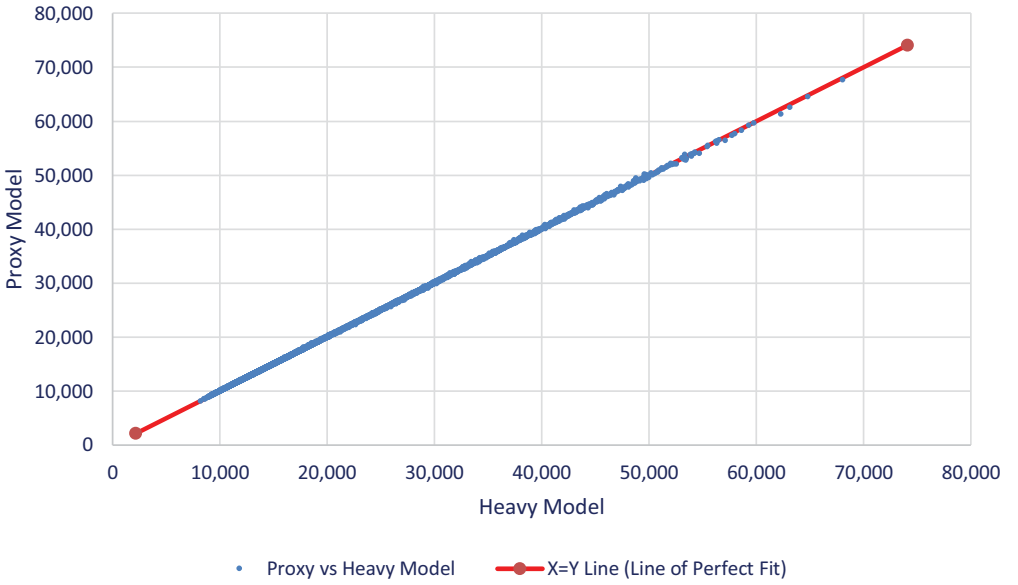

Figure 10. Comparison of validation scenarios under precise interpolation approach.

choices of potential terms in the models. In particular, the stepwise, LASSO and genetic algorithm approaches could choose from 5,233 terms to include in the chosen models.

5.6.1. Ordinary least squares – Precise interpolation

In this example, we have chosen fourth order (quartic) terms for the individual (univariate) and bivariate terms, while we have allowed for three-way interactions between all risk drivers, albeit limiting the less material risk drivers to quadratic terms for the three-way interactions. In total, 1,921 terms were used to calibrate this proxy model.

Figure 10 compares the results from the proxy model calibrated using the "precise interpolation" technique and the heavy model results across the circa 10,000 validation scenarios.

All of the points lie on or very close to the X = Y line. In addition, we note that:

- The root mean square error (RMSE) is circa £60 m
- The largest error is circa £960 m
- 99% of the errors are in the interval (−£254 m, +£278 m).

5.6.2. Stepwise model fitting

In this example, we have chosen to only consider the following terms in the proxy model:

- All risk drivers up to the third order were permitted
- All combinations of risk drivers with combined total order of 8 were allowed in the model.

In total, approximately 5,233 potential terms were allowed to be considered in the model. In addition, we used 20,000 scenarios to calibrate the proxy models with the scenarios randomly chosen from the "domain" of the proxy model.

Figure 11 shows the results of the proxy model fitting. We note that the final version of the model was limited to 150 polynomial terms due to runtime considerations.

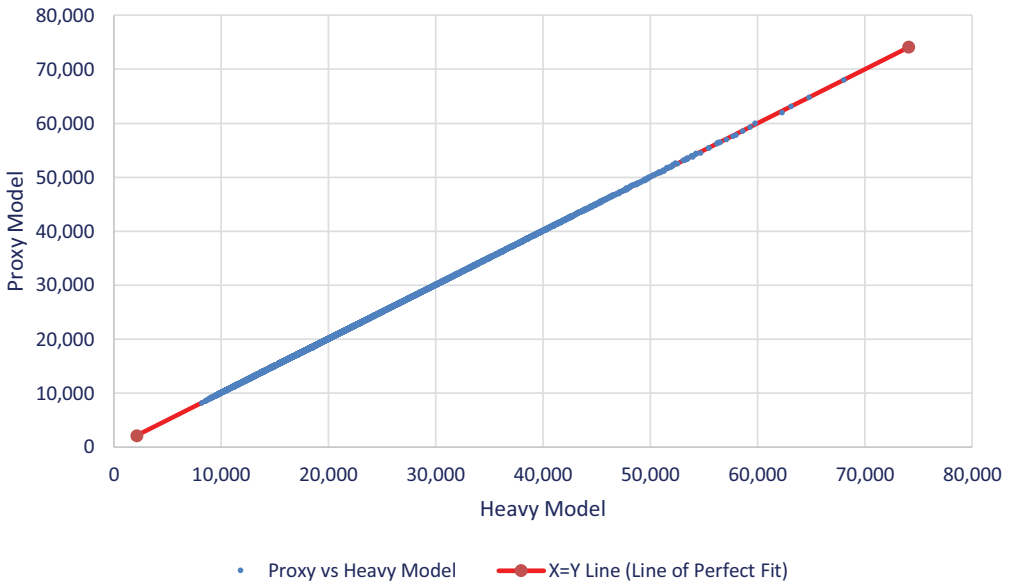

Figure 11. Comparison of validation scenarios under stepwise model fitting approach.

The fit appears visually indistinguishable from the "precise interpolation" fit. All points lie on or very close to the X = Y line.

In addition, we note that:

- The root mean square error (RMSE) is circa £29 m
- The largest error is approximately £342 m
- 99% of the errors are in the interval (–£112 m, +£72 m).

Therefore, in this example, the stepwise approach is "better" because:

- It produces smaller errors than the "precise interpolation" approach
- It requires less application of expert judgement to choose the terms
- It achieves a superior fit compared to the "precise interpolation" approach with significantly fewer terms.

As noted, the stepwise algorithm is relatively slow, largely due to the number of models it checks in each step. In particular, given approximately 5,233 terms to choose from, the model tests 5,233 models in each step, i.e., circa 785,000 model evaluations. A large part of the reason for the (lack of) speed is the 20,000 calibration scenarios being used to perform the evaluation.

One adjustment that we have tested to the stepwise approach is to draw a (different) random sample from the calibration scenarios in each step. The random sample is allowed to grow in proportion with the number of terms in the model being tested. In particular, we have tested a stepwise approach as follows:

- Randomly select a sample of scenarios from the full set of 20,000 scenarios in each step
- Use 10 times the number of calibration scenarios as there are terms in the models being tested, subject to a minimum of 100 scenarios for the fitting
- In each step, the same scenarios are used to test all of the "candidate" models
- The final calibration is then based on the full set of 20,000 calibration scenarios.

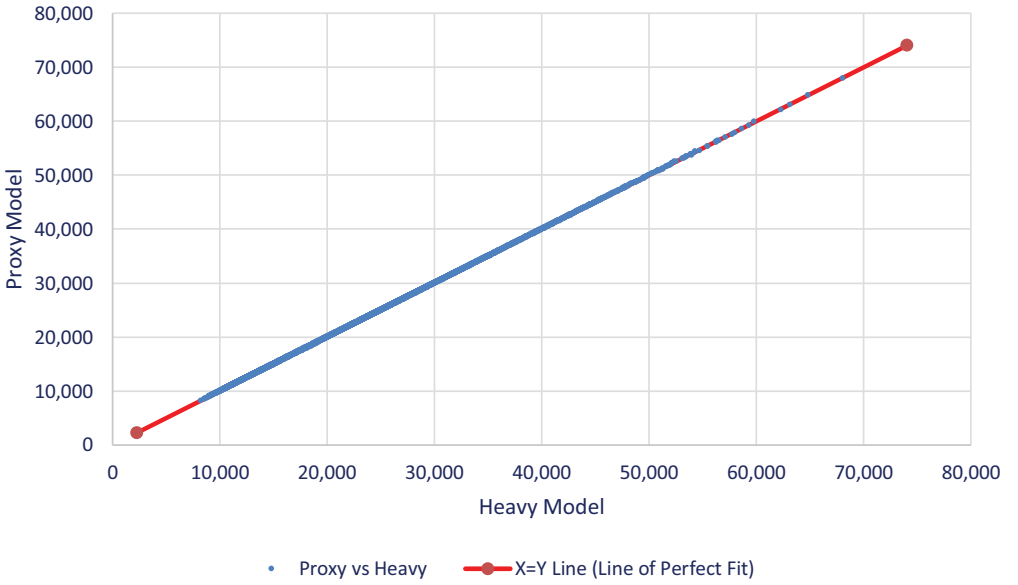

Figure 12. Comparison of validation scenarios under stepwise model fitting with resampling.

The key advantage of this approach is that it allows the stepwise model selection to run a lot faster than the stepwise approach using the full 20,000 scenarios. In addition, randomly varying the scenario in each step builds in some validation into the fitting process. Terms are only kept in subsequent steps in the model if they remain appropriate given the new set of randomly chosen scenarios.

Figure 12 shows the results using this adjusted approach.

We note that:

- The root mean square error (RMSE) is circa £28 m
- The largest error is approximately £269 m
- 99% of the errors are in the interval (–£106 m, +£72 m).

We note that the results remain comparable to the stepwise approach using the full 20,000 scenarios. However, this model requires 301 terms rather than the 151 in the "full" stepwise approach to achieve a comparable fit. This does result in a slightly improved fit.

5.6.3. Genetic algorithm model fitting

For the genetic algorithm approach, we have used the same data and tested the same potential terms as in the stepwise approaches.

Figure 13 shows the results achieved using the genetic algorithm approach.

We note that:

- The root mean square error (RMSE) is circa £35 m
- The largest error is approximately £370 m
- 99% of the errors are in the interval (–£135 m, +£95 m).

We can observe that the genetic algorithm was not able to produce a fit as good as the stepwise approaches. The genetic algorithm, in this case, was significantly quicker to run than the stepwise approaches to achieve this goodness of fit, having taken 50 generations with a population size of 100, i.e., 5,000 models tested. More scenarios may be required, however, to achieve an equivalent fit.

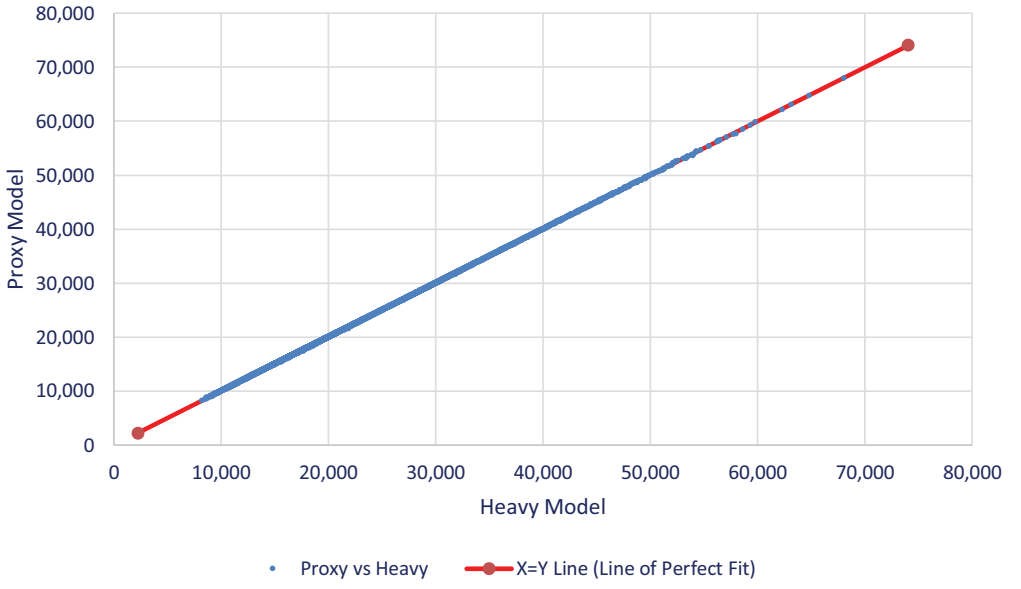

Figure 13. Comparison of validation scenarios under genetic algorithm approach.

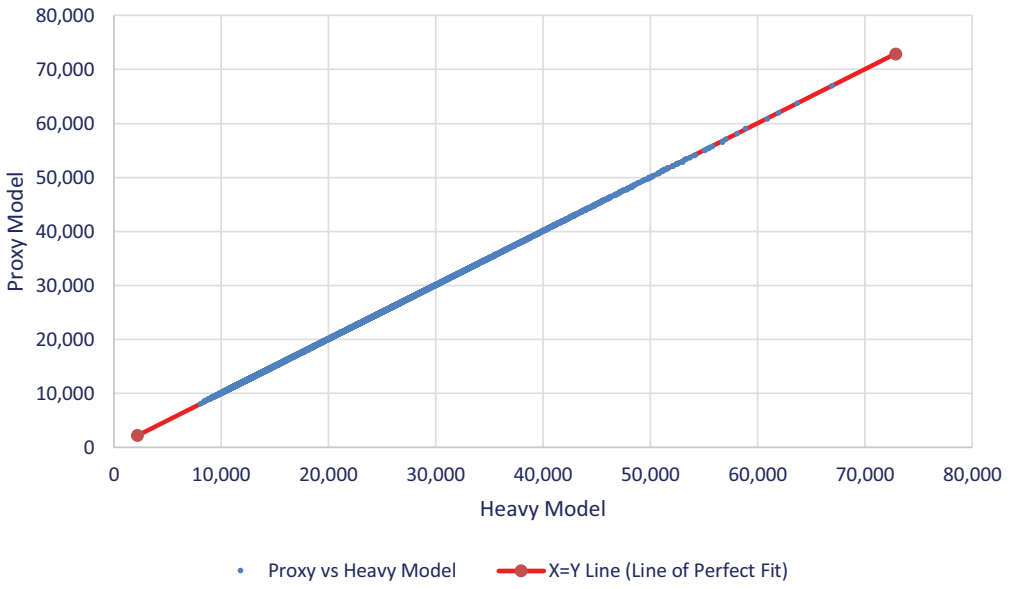

Figure 14. Comparison of validation scenarios under LASSO approach.

5.6.4. LASSO model fitting

For the LASSO model fitting, we have used the same data and tested the same potential terms as in the stepwise approaches. Figure 14 shows the results achieved using the LASSO approach.

- The root mean square error (RMSE) is circa £25 m
- The largest error is approximately £248 m
- 99% of the errors are in the interval (–£92 m, +£62 m).

We note that the results remain comparable to the stepwise approaches.

Table 1. Comparison of Calibration Approaches

Calibration Method	RMSE	Largest (Absolute) Error	99% Confidence Interval (Errors)	No of Terms	Relative Runtime (Approx.) (Lower is Better)
Precise interpolation	60	957	(−254, 278)	1,921	1
Stepwise	29	342	(−112, 72)	151	3,000
Stepwise (with resampling)	28	269	(−106, 72)	301	2,500
Genetic algorithm	35	370	(−53, 210)	2,599	1,200
LASSO	25	248	(−93, 63)	1,655	18

The key advantage of using the LASSO approach is the speed. The LASSO model fitting takes a few minutes compared to the hours that are required with the stepwise approaches to achieve a similar fit.

5.6.5. Summary and conclusion

Table 1 compares the outcomes of the different calibration approaches.

A few things to note about this comparison:

- For computational efficiency, the "stepwise" and the "stepwise (with resampling)" approaches were limited to 150 and 300 steps, respectively, allowing the algorithms to run for longer would have improved the outcomes for those methods (but the relative run time would have increased).
- The precise interpolation approach only admitted three-way cross-terms, while the other approaches allowed 4-way cross-terms.

The precise interpolation approach has the quickest run-time. However, this approach had the weakest goodness of fit, when evaluated on out of sample scenarios. The primary reason for this is likely to be that no four-way interactions were allowed in the modelling. This was at odds with our original intuition that cross-terms involving more than 3 variables were unlikely to be required. However, 10% of the terms in the stepwise fitted models were four-way cross-terms, suggesting that these terms are genuinely important (at least for this specific proxy modelling problem).

The LASSO has the next quickest run-time and was significantly more efficient than the other automated model selection approaches. Of these approaches, the LASSO produces the smallest errors in the metrics that we have used in the table above. The speed of the LASSO approach also confers additional advantages, chiefly that it allows more exploration of different model forms. However, the stepwise approach produces a very comparable fit with a much smaller number of terms which is attractive and provides more confidence that the model is unlikely to be overfitted. We note that the out-of-sample testing showed similar performance for the stepwise and the LASSO approaches.

The genetic algorithm did not perform as well as the LASSO or the stepwise approaches while allowing for the largest number of parameters. We note that genetic algorithms are very difficult to "tune" for specific modelling problems, and one set of hyperparameters will not necessarily work well for a different problem.

6. Validation

6.1. Background

Once a proxy model has been calibrated, it is vital that its fit is validated to provide assurance that it is suitably reflective of the "heavy model" and therefore that its results are reliable. This also

supports the firm in meeting the model validation requirements under the Solvency II Directive (Article 124) (European Parliament, 2009). The validation should consider the results of the proxy model relative to the results of the heavy model using scenarios that are distinct from those used in the calibration. As part of its thematic review into proxy modelling (PRA, 2019), the PRA outlined 11 commonly applied tests used to inform the goodness-of-fit assessment of the loss function. These tests are carried out either on-cycle or off-cycle and used to inform future fits.

The 11 tests are considered in the below sections and have been grouped by their use for:

- Testing the fit of the calibration
  - **Independence of errors**. Tests the implicit assumption that the errors are independent and therefore there is no systemic issue in the fit.
  - **Homoscedasticity of errors**. Homoscedasticity (same variance) is central to linear regression and describes the situation where the error term is consistent across the distribution (i.e. the error does not increase/decrease dependent on the size of the stress).
  - **Normality of errors**. Test to identify whether the errors are normally distributed (a key assumption to the error being "noise").
  - **Over/Understatement of errors**. Proxy models should be reflective of the heavy model and should not systemically over- or under-state capital. This test validates this assumption by reviewing whether there is any statistical significance in the sign of the residual.
  - **Overfitting**. Proxy models are fitted to heavy models at a point in time; if it is overfitted (i.e. it too closely corresponds to the data at calibration) then it would be expected to introduce larger errors at subsequent reporting dates (and following roll forwards).
- Providing validation and a feedback loop
  - **Out-of-sample test (relative error).** An out-of-sample test is one that is carried out on scenarios that did not form part of the proxy calibration set (i.e. these scenarios were not used to inform the structure or coefficients of the proxy models). This tests the proxy model error relative to the heavy model stresses.
  - **Ranking tests of the loss distribution**. The ability of the Internal Model to appropriately rank risk is also a requirement of the Solvency II Directive. The test passes if the proxy and heavy model results are ranked consistently.
  - **Quantification of mis-estimation of the SCR**. As well as testing the level of potential proxy model error, this provides a mechanism to adjust the SCR for the difference between the proxy model and heavy model results around the 99.5th percentile. This can be used if the results of another test indicate this action should be taken and can be extended to other percentiles and uses as necessary.
- Supporting the sign-off process
  - **Analysis of change of the form of the loss function**. A simple table that lists the type of function for each risk factor (e.g. risk factor, power, selection order). The test passes if the above criteria are met, which confirms that the proxy and heavy model results are ranked in a consistent manner.
  - **Graphical analysis of bivariate fit**. A visual chart which shows the relationship between key risk factors, e.g., net asset value. Univariate risks can be shown using two-dimensional charts and bivariate relationships can be shown using three-dimensional charts. This test provides reasonableness checks, examination of turning points, discontinuity points, behaviour at the extremes, etc.
  - **Sensitivity**. Sensitivities are commonly used to assess the impact of changes in key parameters/expert judgements.

6.2. Testing the Fit of the Calibration

The following tests can be used as part of the calibration or as part of the feedback loop and focus on the error (the difference between the heavy and lite models). By definition, as the proxy model is a proxy for the heavy model, there should be no systemic difference between the two. Therefore, a good proxy model would have independent errors (i.e. not vary in size dependent on the stress). Equally it is beneficial for the errors to be identically distributed (i.e. the mean of the errors should be zero and the variance should be constant). That is, the errors (ε_s) can be modelled as independent identically distributed normal random variables with zero mean:

$$\varepsilon_s \sim N(0, \sigma^2)$$

where

- $\varepsilon_s = HM_s - PM$
- HM_s is the heavy model value under scenario s
- PM_s is the proxy model value under scenario s.

The first four tests of this section validate the key assumptions around the error and therefore proxy model.

6.2.1. Independence of errors

The aim is to test whether the errors are statistically independent, and therefore whether there are any systemic issues in the fit. The test can either be performed visually using a scatter plot of the errors (and checking whether there is any visible pattern) or by splitting the errors into groups and performing correlation tests on the subsets (see below). As referenced in previous IFoA papers[3], statistical tests can be less meaningful in this context due to the relatively small number of fitting points and the residuals not being independent. Notwithstanding this, given the ease in which this test can be performed, we would expect this to form a standard validation test within the proxy model fitting process.

Whilst it may be expected that the test will fail, dependent on the number of scenarios, the results may still be informative. In particular, any relationships between the residuals may be illuminated. If a pattern exists, this may indicate the proxy model is failing to capture a behaviour of the underlying model. It also enables comparison between different proxy models.

6.2.2. Homoscedasticity of errors

An implicit assumption within the fitting methodology may be that the errors are independent, supporting the assertion that there is no systemic issue in the fit. This can be tested by calculating the correlation coefficient of the residuals with the test failing if it breaches a pre-specified tolerance.

$$\rho_{\varepsilon R} = \frac{Cov(\varepsilon, R)}{\sigma_\varepsilon \sigma_R}$$

where:

- R is the stress for the risk R
- ρ, ε and σ are as defined above.

[3]https://www.actuaries.org.uk/system/files/field/document/A1_Andrew%20Smith_Gabi%20Baumgartner.pdf

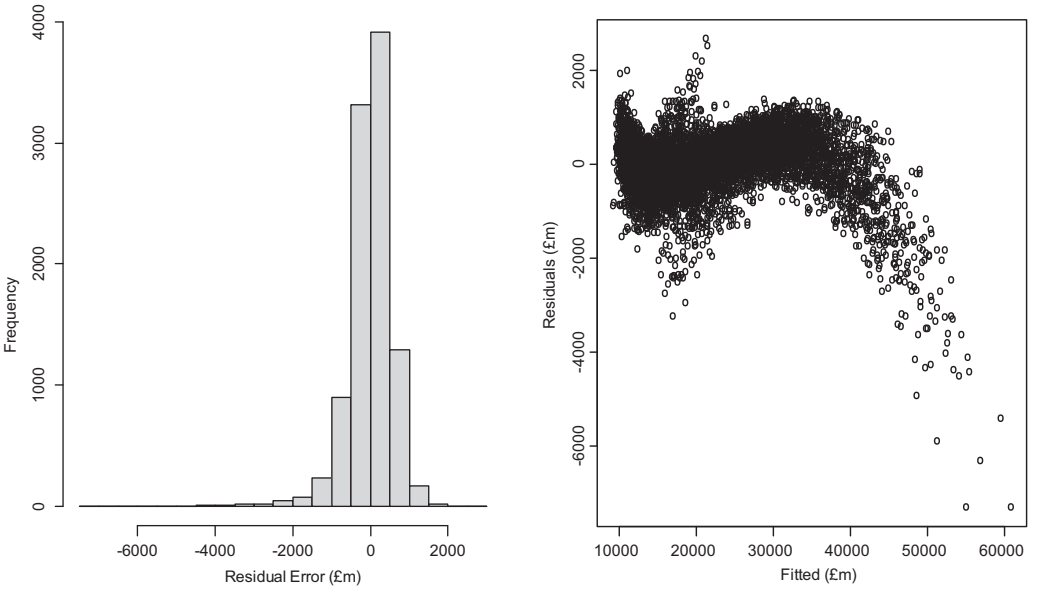

Figure 15. Plotted residuals.

A key judgement in this test is the pass criteria, i.e., the value of ρ that should be considered as a failure and therefore lead to the conclusion that a relationship exists between the size of the stress and the error observed. This judgement should consider the materiality of the risk considered and the number and range of validation scenarios. Specifically:

- For a material risk, a small correlation could result in a large error in the 1-in-200 scenario (as it would be expected to be a relatively large stress in this scenario). Hence, a higher threshold may be set for less material risks to allow resource to focus on more material risks.
- As with all tests, for it to be statistically significant, a sufficient number of validation scenarios must be applied. Additionally, the test is unlikely to be meaningful if the scenarios are not spread over a suitably wide range (as there is unlikely to be much variation in the results).

A graphical plot of the residuals against the proxy model (or inspection of the values) should indicate what scenarios/risk space is causing the test to fail. This can be seen in Figure 15 where the size of the error can be seen to increase as the stress increases, indicating that there is a feature of the heavy model that the proxy model has not captured.

Once a proxy model has been calibrated, the residuals can be used to test for both independence and homoscedasticity. In the case study, the first iteration of fitting has resulted in the distribution of errors shown in Figure 15. We can see that the residuals are not normally distributed, with the size increasing the larger the capital requirement. Analysis of the results shows that the univariate mortality risk stresses are highly correlated with the residual error. This may be contributing to the non-independence observed in the residuals. To address this, higher terms for both mortality and PC1 have been included within the universe of available terms and ultimately the proxy model. The tests have then been re-performed with the results presented in Figure 16.

We can see from these figures that the residuals now better represent a normal distribution, with the covariances of mortality and interest rates (PC1) are 0.83. Introducing further terms therefore appears to have reduced the limitations in the calibration.

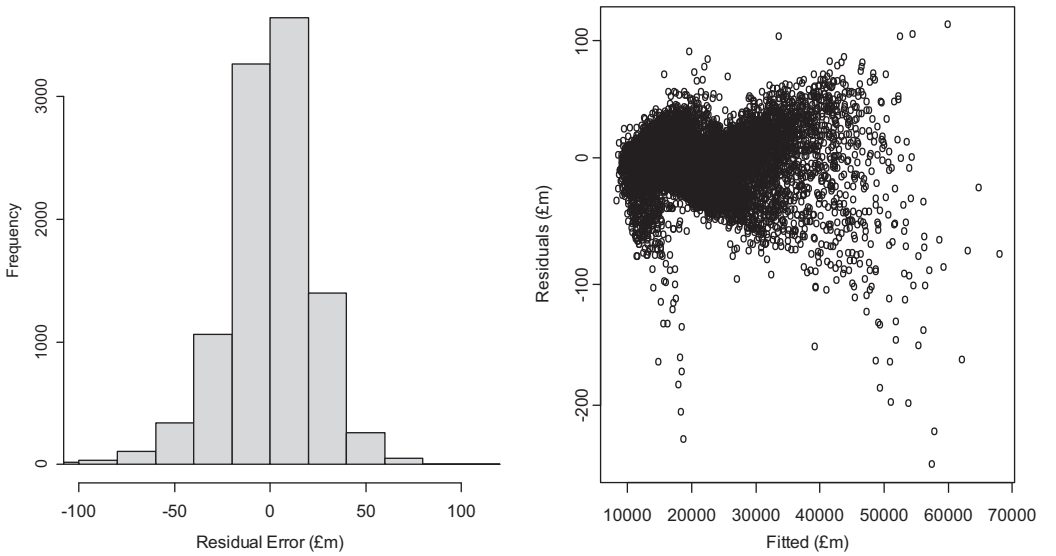

Figure 16. Plotted residuals.

Actions if Fail

If the test fails, a number of actions can be taken:

- Re-fit the proxy model
  The proxy model could be re-fitted to capture the feature of the model currently driving the non-constant variance. This may include adding additional terms for a particular stress type or interaction terms.
- Removal of scenario
  If the test failure is caused by a particular scenario (e.g. an extreme stress having a large residual) then the appropriateness of this scenario should be considered including:
  (a) whether the scenario is realistic. If not, it may be appropriate to remove this scenario from the set as the changes to capture such a feature in an approximate model may not be proportionate (e.g. an asset share increase and estate decrease may be extremely improbable)
  (b) whether there is an error or known limitation. This may be either in the heavy model, proxy model or the scenario run.
- Accept the limitation
  The limitation in the fit of the proxy models may be accepted if, for example, the scenarios causing the failure are far from the region of use for the proxy model (e.g. downside stresses when biting direction is up); the impact of the error is small (e.g. non-independence for immaterial risks); or it would be disproportionate to correct for this error (e.g. additional scenario runs). In these cases, this should be justified to senior stakeholders and documented appropriately.

6.2.3. Normality of errors

Dependent on the proxy model, the residuals may be expected to be normally distributed. There is a suite of literature on testing for normality and so this is not repeated here. We would, however, expect these to be part of the validation process, including: Pearson's Chi-squared test; histogram plots; Q-Q plots; and other statistical tests (such as the Jarque–Bera test). We would note, however, that this assumption may not always hold, e.g., the residuals for burn-through may not be normally distributed (dependent on the arrangement).

6.2.4. Over/Understatement of errors

Proxy models should be reflective of the heavy model and should not systematically over or understate capital. This test aims to validate this assumption by reviewing whether there is any statistical significance in the sign of the residual. That is, if there is a statistically significant number of overstatements (i.e. the error is less than 0) then there is evidence that the calibration has resulted in a biased proxy model (and vice-versa). Whilst this test can be done graphically, it can also be performed via a binomial test with p = 0.5 and n = number of scenarios. This test can be performed at individual polynomial form (e.g. net cost of options, guarantees and smoothing) or at the overall reporting entity level (e.g. with-profits fund). It can also be done taking into consideration the direction of the stress and, as with all tests, its usefulness is dependent on the number of scenarios.

That is, if X is the number of errors (ε_s) greater than 0 then it is assumed that $X \sim \text{Bin}(n, p)$. The test is therefore a hypothesis test where H_0: $p = 0.5$. If k is the number of validation scenarios with an error greater than 0 then:

$$P(X = k) = \sum_{i=0}^{K} \binom{n}{i} p^i (1-p)^{n-i}$$

As the test is interested in both over or understatements, this should be a two-tailed test with the pass criteria taking into consideration the number of scenarios and materiality of the risks (as outlined above).

It should be noted that even if the proxy model is shown to not have bias (e.g. the residuals are positive for 50% and negative for 50%), this can still result in capital being overstated. This is because it does not consider the size of the residual, just the sign. Further, as the capital requirements are derived from a specific point, it is the relative over/understatement at this point that is of particular interest for reporting. This test does, however, provide further insight into the fit of the model and potential areas of focus. For example, if the proxy model always materially understates capital for expense stresses, then this suggests the coefficients or terms for expense may not be suitable (Murphy & Raduin, 2021).

Actions if Fail

If the test fails, a number of actions can be taken:

- Re-fit the proxy model.
  The proxy model could be refitted to capture the feature of the model currently driving the non-constant variance. This may include adding additional terms for a particular stress type or interaction terms.

- Shift the proxy model.
  Simplistically, bias indicates that there is at least some area of the risk space where the proxy model is either "above" or "below" the heavy model. A shift could therefore be applied to the proxy model to adjust for the bias by effectively applying a constant to the polynomial, which can either be fixed across the risk distribution, or stress dependent (for when bias is only present in one direction). It is important that, if a shift is applied, the implications of other validation tests are considered and how the shift has been applied (e.g. if done at underlying liability level then this may introduce bias at entity level and the validation tests at the entity level should be reperformed).

- Accept the limitation.
  The limitation in the fit of the proxy models may be accepted, for example, if the impact of the error is small (e.g. small constant overstatement); or it would be disproportionate to

correct for this error (e.g. additional scenario runs). In these cases, this should be justified to senior stakeholders and documented appropriately.

Performing a bias test on the fit of the interest rate scenarios shows that the proxy model overstates the heavy model in all cases (shown in Figure 17). If uncorrected, the proxy model will be known to give higher capital requirements for interest rate stresses.

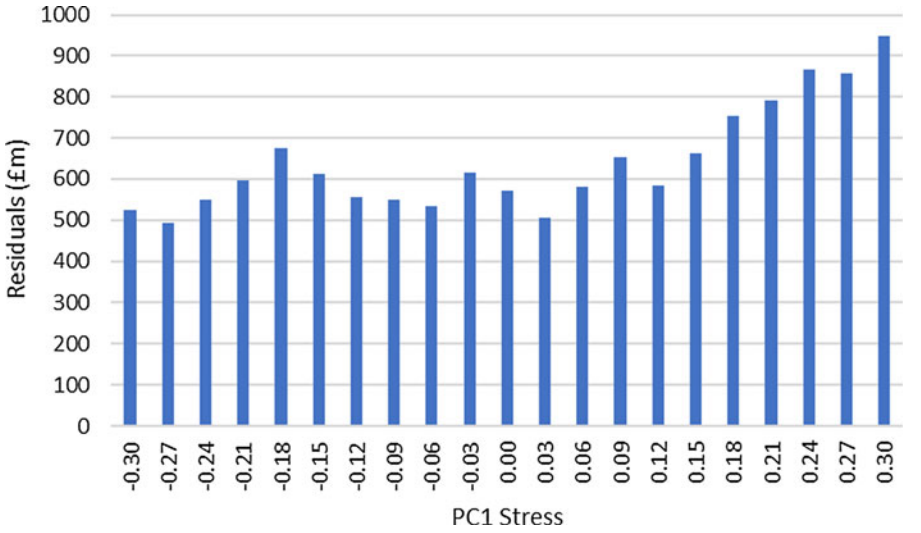

Figure 17. Plotted residuals of bias test.

In this case, shifting the proxy model interest rate stresses by circa £640m (the average of the residuals) would remove the bias, however, clearly the residual would remain large (Figure 18). Hence, it may be more appropriate to include a higher term to remove the bias (noting that this can take longer so simple shifts may be more appropriate if the underlying fit is already acceptable). Increasing the terms included within the model has corrected the bias in this case. This test also indicates heteroscedasticity (a good example of where tests offer multiple validation benefits for minimal effort).

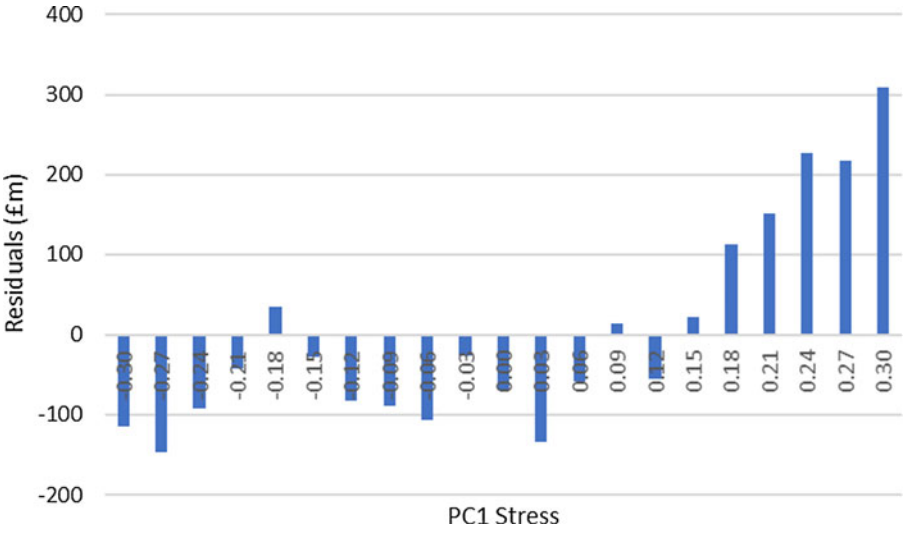

Figure 18. Plotted residuals of bias test following increase in terms.

6.2.5. Overfitting

Proxy models are fitted to heavy models at a point in time. However, if models are overfitted (i.e. too closely corresponds to the data at the calibration date) then there is a risk of larger errors emerging at subsequent reporting dates. There are a number of ways to test for overfitting including:

- Using a calibration process which takes overfitting into consideration explicitly (e.g. through nested models or AIC tests)
- Performing a bootstrapping calibration (i.e. using a sample of the calibration scenarios to fit the proxy models then using the remaining scenarios to test the fit)
- Running the same validation scenarios at calibration and reporting dates to compare the residuals (noting some of this will be caused by the deterioration in fit between dates)
- Applying the other validation tests outlined in this section at both the calibration and validation dates.

If overfitting is identified during calibration, investigations should be performed to identify which proxy model/function is causing this and removing the relevant terms. As overfitting identified as part of the reporting process can be challenging to address appropriately in-cycle, this should be investigated as part of the regular feedback loop.

Akaike Information Criterion

The AIC can be used to compare two models with the one with the lowest AIC being the better' model. The test was discussed in the calibration section (Section 5.5.2).

Bootstrapping calibration

A random sample of calibration scenarios is selected and used for model fitting; the remaining scenarios are then used for validation to test the fit (and gives an indication of overfitting). This test can be repeated using different samples.

6.3. Tests to Support Validation and a Feedback Loop

These validation tests can be used post-calibration to assess whether the proxy fit is satisfactory, and all focus on an out-of-sample scenario set (i.e. scenarios that were not used to inform the structure or coefficients of the proxy models). The outcome can be used as part of a feedback loop to inform the current or future calibration cycles.

We consider the following:

- Out-of-sample relative error test
- Ranking tests of the loss distribution
- Quantification of mis-estimation of the SCR.

6.3.1. Out-of-sample relative error test

The aim of the test is to check proxy model accuracy across the distribution, by considering scenarios that were not used in fitting the proxy models. Here, we focus on testing the proxy model error relative to the heavy model stresses, with overlays relating to the absolute error and heavy model movement to improve the efficacy of the test.

The relative error is defined as:

$$Relative\ error = \frac{(LM_{stress} - LM_{base}) - (HM_{Stressed} - HM_{base})}{HM_{Stressed} - HM_{Base}}$$

Using a number of out-of-sample scenarios (with the size of this set depending on run budget and other priorities), the following is performed:

- For each scenario, compare the change in proxy model and heavy model result from base
- The residual is then divided by the heavy model movement
- If this error is below a pre-determined threshold (say 5%) then the scenario passes
- Analysis over all scenarios is performed (e.g. assessing against a minimum pass proportion).

There are a number of areas where test parameters need to be established:

- The pass threshold for each scenario
- The proportion of scenarios where a pass is required (to determine whether the test passes at an overall level)
- Whether a minimum absolute movement in the heavy model is required to include a scenario in the test, to avoid the large relative errors that tend to occur in these cases
- The maximum absolute error that is considered acceptable, regardless of the size of the relative error, otherwise the test is deemed to have failed.

Judgement will be required when setting these parameters. These could be set using a top-down approach where stakeholders agree on the level of accuracy required in, for example, SCR or risk appetite calculations. The level of accuracy required can then be converted into test constraints but should be back tested for stability. If applying a minimum absolute heavy model movement constraint, this could be set as a fixed value or based on the variance of the movement in the heavy model being considered (so that the constraint is set with some sense of the variability of the model output). The maximum absolute error accepted can be set using the firm's risk appetite for model error, with consideration of the region in which the scenario occurs and whether this is likely to impact the primary uses of the model.

In addition to establishing these test parameters, thought is needed regarding the granularity at which the test is applied (i.e. the level of aggregation). It could be performed at multiple levels to see whether there are offsetting accumulating errors that build up, or whether a particular business line is contributing a disproportionately large amount to the overall error.

Using a relative error test that is dependent on the size of the heavy model movement helps to avoid some issues commonly encountered with other tests:

- Setting a fixed out-of-sample error tolerance is insensitive to the size of the stress being applied, and may introduce circularity in setting the value (e.g. if the constraint is set as a proportion of the SCR, which also does not allow for the level of aggregation at which testing is applied).
- Measuring the error relative to the heavy model value (rather than heavy model movement) can be distorted by items with either base values very close to zero (e.g. derivatives) or with very large base values (e.g. liability reserves on a significant block of business).

The relative error tends to be larger for small heavy model movements. The use of a minimum heavy model movement (set as a fixed value or linked to the variability of the particular heavy model(s) being considered) filters out scenarios that might cause the test to fail when the impact on the accuracy of the proxy models is immaterial. Equally, applying a maximum acceptable absolute error ensures the test does not pass if the model error is beyond appetite. This may occur for instance in a scenario with a large heavy model movement where the relative error is relatively low.

As with all tests, the usefulness will be dependent on selecting a sufficient number of relevant scenarios. It may not provide a full picture if a small number of scenarios are used that do not sufficiently cover the risk domain.

If the test fails, this indicates that the proxy model is not reproducing the heavy model results with a high level of accuracy across a wide range of scenarios, or has a particularly poor fit in a certain area of the risk space. In this case, a number of actions can be taken:

- Note the limitation
  Including the materiality of the limitation where possible.

- SCR adjustment
  Apply a short-term adjustment to the SCR (or relevant result) to allow for the limitation, which encourages the feedback loop below. We would expect any adjustment to be strictly positive (i.e. to increase the capital requirement).

- Feedback loop
  In the long term, a feedback loop should be used to investigate the scenarios failing the test to establish whether there is a common theme emerging. This can then feed into future calibration cycles.

6.3.2. Ranking tests of the loss distribution

The internal capital model is used, in part, to support the management of risks within companies. Hence, a key requirement of the proxy model is to allow firms to rank risks, and it is therefore vital that the proxy model and heavy model rank scenarios in a consistent manner. Ranking tests provide a mechanism to confirm that this is the case. The ability of the Internal Model to appropriately rank risk is also a requirement of the SII Directive (see bullet 4 under Article 121 on Statistical Quality standards) (European Parliament, 2009). These tests are therefore also important from a regulatory perspective.

Using a number of scenarios from across the distribution, evaluated by both the proxy model and heavy model, we can assess whether:

- The allocations of capital are sufficiently similar
- The rank correlation between the results is sufficiently high.

The test passes if the above criteria are met, which confirms that the proxy and heavy model results are ranked in a consistent manner. If SCR accuracy is the key consideration, a higher concentration of scenarios around the 99.5^{th} percentile may be required.

Judgement is required to set the maximum tolerance for capital error and minimum tolerance for rank correlation between the heavy model and proxy model results. Out-of-sample scenarios utilised in other tests should be re-used where possible to assess the rank correlation. A larger number of scenarios may be required to assess capital allocation if the firm's SCR calculation is stochastic in nature, and it is here that run budget may become a limiting factor.

We expect that tests of rank correlation should be high priority and carried out frequently, and this was a key area of best practice that was identified in the PRA Proxy Modelling review (PRA, 2019, p. 12). Tests to assess capital allocation may need to be performed off-cycle with lower frequency due to run budget. Any results falling below the acceptance criteria should be investigated to identify the scenarios that are causing the failure. From this the root cause can be identified and a correction made, or the model developed as appropriate.

6.3.3. Quantification of misestimation of the SCR

As well as testing the level of potential proxy model error, this provides a mechanism to adjust the SCR for the difference between the proxy model and heavy model results around the 99.5th percentile. This can be used if the results of another test indicate this action should be taken and can be extended to other percentiles and uses as necessary.

There are a number of approaches that could be used to test for mis-estimation and apply an adjustment, and we have seen the following examples:

- Analyse the average error from multiple smoothed scenarios around the 99.5th percentile loss
- Analyse the average error from multiple unsmoothed scenarios around the 99.5th percentile loss
- Analyse the error in the 99.5th percentile smoothed scenario
- Analyse the distribution of errors and test for independence from the heavy model results.

For background, stochastic scenarios are ranked according to their impact on the firm's Own Funds. Unsmoothed scenarios are those sampled directly from this ranked set and this can lead to large differences in the prevalent risks when comparing neighbouring scenarios (in terms of Own Funds impacts). For example, an annuity writer with significant longevity risk exposure will typically see scenarios around the 99.5th percentile that exhibit large, unexpected increases in longevity improvement rates. However, there will also be scenarios in the region that are dominated by other risks (such as those arising from credit-related exposures). Comparing the unsmoothed scenario at the 99.5th percentile at different valuation dates can be difficult if the prevalent risks change. Smoothing over a number of scenarios, by averaging the stresses in respect of each risk, provides a single scenario which is representative of that region of the Own Funds distribution. This is more stable over time and provides a more useful means of comparison and monitoring of the magnitude of mis-estimation. The term "biting scenario" refers to the 99.5th scenario from the ranked set. The smoothed 1-in-200 refers to the "average" scenario when X scenarios are taken around the 99.5th. For example, the smoothed 1-in-200 can be derived by averaging the 500 scenarios around the 99.5th (i.e. the biting scenario).

We would expect a larger number of smoothed scenarios to be analysed (as opposed to unsmoothed scenarios) to improve the stability of the result from period to period. A large bias in the errors will tend to result in an adjustment being required. A large variance will have a smaller impact, provided a sufficient number of scenarios have been run to allow these errors to stabilise.

Judgement is required to determine the number of scenarios to test, the size of the window to test, the threshold to apply (if any), and whether adjustments will be made for understatements only (or also overstatements).

If an adjustment is applied, judgement is required to determine how frequently this should be assessed and updated. The adjustment will be determined using a limited number of scenarios, so there is a risk of volatility in the results. The additional use of smoothed scenarios should help to reduce this risk.

There may not be sufficient time, resource or computing power available to perform this on-cycle. Therefore, resource may be required after the reporting cycle to obtain an adjustment that is then applied to the next SCR.

The test fails if the misestimation exceeds the selected threshold, and in that case an adjustment to the SCR is applied. A number of possible approaches are set out below:

- Adjustment equal to calculated error: the adjustment is applied regardless of size, though would typically only be applied if the SCR is understated.
- Adjustment equal to calculated error, if exceeds threshold: the adjustment is applied if a pre-determined threshold is exceeded (e.g. 2.5% of SCR).

- Adjustment equal to trailing average error: to stabilise the adjustment, an average based on the misestimation over previous SCR calculation periods may be used. Though this improves stability, the appropriateness of such an adjustment will depend on the stability of the business itself (e.g. a trailing average used for a rapidly growing firm may not be suitable).
- Adjustment equal to impact of simulated error terms: if the distribution of errors is independent of the heavy model results, a large number of Internal Model scenarios with an independent error term can be simulated, and the impact on the SCR can be calculated.

If the error is excessive, the proxy models should be investigated to determine if there is a root cause. A capital overlay will not fix an underlying issue, so in this case it is important to identify the cause of the problem.

Ranking Test

Performing a ranking test on the lite models shows that the lite models and heavy models do not consistently rank the risk distributions (e.g. the 99.5th percentile scenario for the lite model is broadly equivalent to a 99.25th percentile scenario for the heavy model). This indicates that the two models are not consistently ranking scenarios and further work should be performed to understand the cause of the mismatch (see Figure 19).

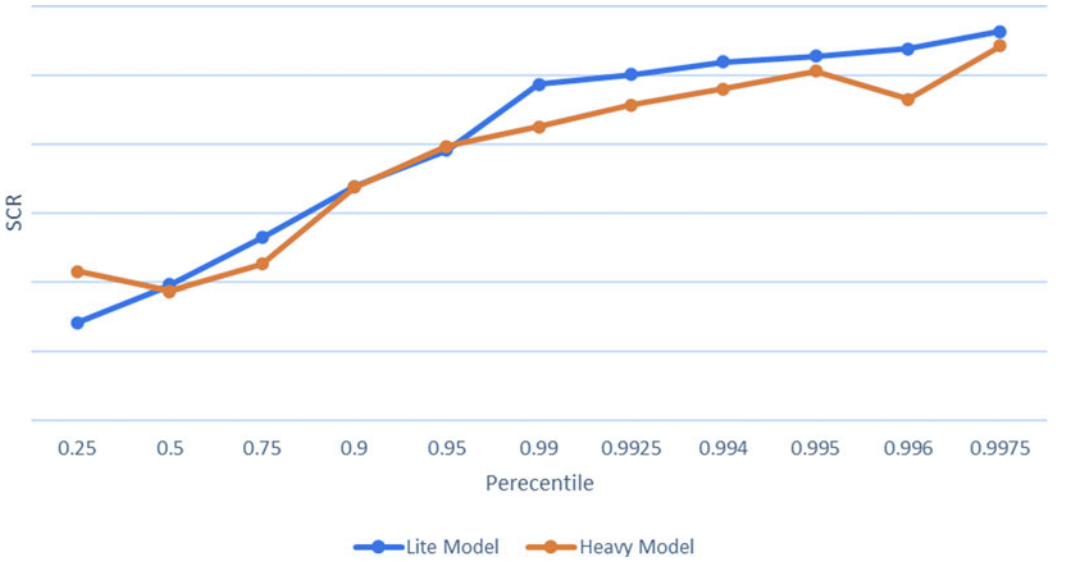

Figure 19. Results of ranking test.

Quantification of Mis-estimation

There are multiple ways to quantify the mis-estimation of the proxy models with Table 2 showing the potential discrepancies within each method.

The suitability of the approach is dependent on the modelling capabilities and the use of the quantification (e.g. using the biting scenario may introduce volatility into any true-up whilst the smoothed scenario may not be suitable for the true-up under any stress and scenario testing).

Table 2. Comparison of Methods for Quantifying Mis-Estimation

Method	*Error*
Error at biting scenario	*1.24*
Error at smoothed 99.5th scenario	*1.36*
Average error from biting 99.25th, 99.5th and 99.75th	*1.26*
Average error from smoothed 99.25th, 99.5th and 99.75th	*1.31*

6.4. Tests to Support Sign-off Process

The following tests can be used to help support the sign-off process. These tests tend to be more visual in nature to help communicate a summary of results and key judgements to senior stakeholders.

6.4.1. Analysis of change of form of loss function

This test is a simple table that lists the type of function for each risk factor (e.g. risk factor, term, selection order). The purpose of the test is to compare the fitted models against previous periods and check how the form of the fitted model has evolved over time (e.g. YE22 versus YE21). This is good for stakeholder management as the table can be accompanied by commentary to make it easy to understand and challenge, i.e., changes explained – what caused the change in form/new terms.

This test is mechanical, but judgement will be applied to comment on whether or not the change in form of loss function is appropriate. Businesses may have the tendency to use the same form of loss functions over time, particularly when the changes are small, and avoid having to explain changes. Good practice will be to challenge and justify the appropriateness of the form of the loss function, regardless of whether the form is changed or retained. The analysis of change is a simple summary of the form of loss functions and does not take much time or effort.

Actions if Fail

A failure will be when the form of loss function is significantly different from previous periods and is unexplained. This will trigger a review of the form of loss function to understand the reason for the change.

6.4.2. Graphical analysis of bivariate fit

The purpose of this test is to:

- Sense check impacts and behaviours for reasonableness
- Examine turning points/discontinuity points
- Visualise behaviours at extreme points.

Visual aids are useful for stakeholder management and enable stakeholders to visualise tail-risk. Graphical analyses are easy to create but can require an elevated level of expertise to interpret the charts and apply expert judgement on the reasonableness of bivariate fit. There is a lack of quantitative pass/fail criteria – the assessment relies heavily on judgement and it can be difficult to interpret/understand without commentary.

Actions if Fail
A failure can occur when, for example, the tail risks are higher than expected, which would trigger further investigation of the behaviours at the extremes.

6.4.3. Sensitivities of key parameters and/or judgements

This is a common test to assess key judgements. The test could include checks on:

- SCR impacts
- Proxy fit (and associated statistics regarding errors versus heavy model)
- Re-ranking

This test assesses the impact of such judgements and supports sign-off of judgements and adjustments. Judgement may be applied to methodology (before calibration) as well as during calibration. This is a quantitative test that can be checked against materiality criteria. The pass/fail criteria may not be clear and will be considered on a case-by-case basis. The actions taken will also depend on the test (and may change how calibration/ validation scenarios are defined). Medium effort is required to assess impacts from changes in parameters. This can also be carried out off-cycle to avoid impacting the critical path.

Actions if Fail
If the impact of changing a key parameter is material, then this test will highlight that the parameter or expert judgement is sensitive (for example a key risk), and further investigation could be triggered to assess whether the parameter or expert judgement is appropriate. For example, the parameter or expert judgement may be updated in response to the sensitivity (new information).

Analysis of Change of Form
Comparing the form of each risk factor is a quick validation check for changes to the loss functions (done in Figure 20). There are no new risks. Forms are consistent with prior year's loss functions except for mortality.

Risk Factor	Decision YE22	Prior Year
• Lapse	• Quadratic	• Quadratic
• Mass Lapse	• Quadratic	• Quadratic
• Longevity Trend	• Quadratic	• Quadratic
• Mortality	• Quadratic	• Linear
• Expense	• Linear	• Linear

Figure 20. Form of loss functions.

Bivariate fit
Graphical analysis can be used to show behaviours between risk factors. The graph to the right in Figure 21 shows there is a change in behaviour at extreme point.

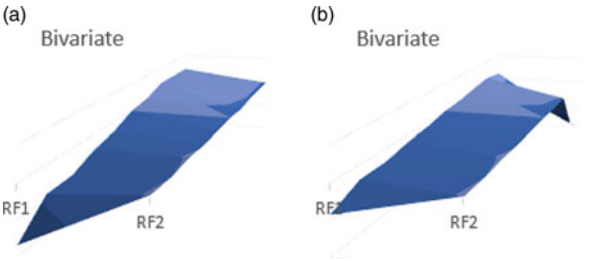

Figure 21. Graphical analysis of risk factors.

Sensitivities

Example 1

Sensitivity analysis can be performed to check the proxy fit. The below results compare the loss function form for lapse risk.

Run 1 – Linear form residual 0.6%

Run 2 – Quadratic residual 0.3%

Example 2

Sensitivities can also be performed to test the materiality of inclusion of a new cross-term X*Y

Running the model including the cross-term X*Y shows an impact on SCR of circa £8 m. This is above the materiality threshold of circa £5 m. Therefore, it is concluded that the cross-term is material and should be included.

6.5. Validation Framework

A firm's approach to validating proxy models should provide assurance to all users of the proxy model that it is fit for purpose. This should take into account the different users of the model; the materiality of the business being modelled and the heavy model it is designed to replicate (in particular any limitations inherent in the model). We believe a defined framework for validation should be outlined within the firm's suite of Solvency II documentation, including justification on the testing performed. This can then be referred back to as part of the sign-off process. Figure 22 outlines a sample high-level framework for informing what validation tests are appropriate. It should be noted that, in our opinion, given their operational ease, the following five validation tests should be included as part of all validation frameworks. The first two provide stakeholders with assurance that the fit is reasonable, and can easily be compared to previous years. The other three provide a goodness of fit assessment quickly and are well established:

- Results of out-of-sample tests
- Bias tests
- Independence of errors
- Homoscedasticity of errors
- Normality of errors.

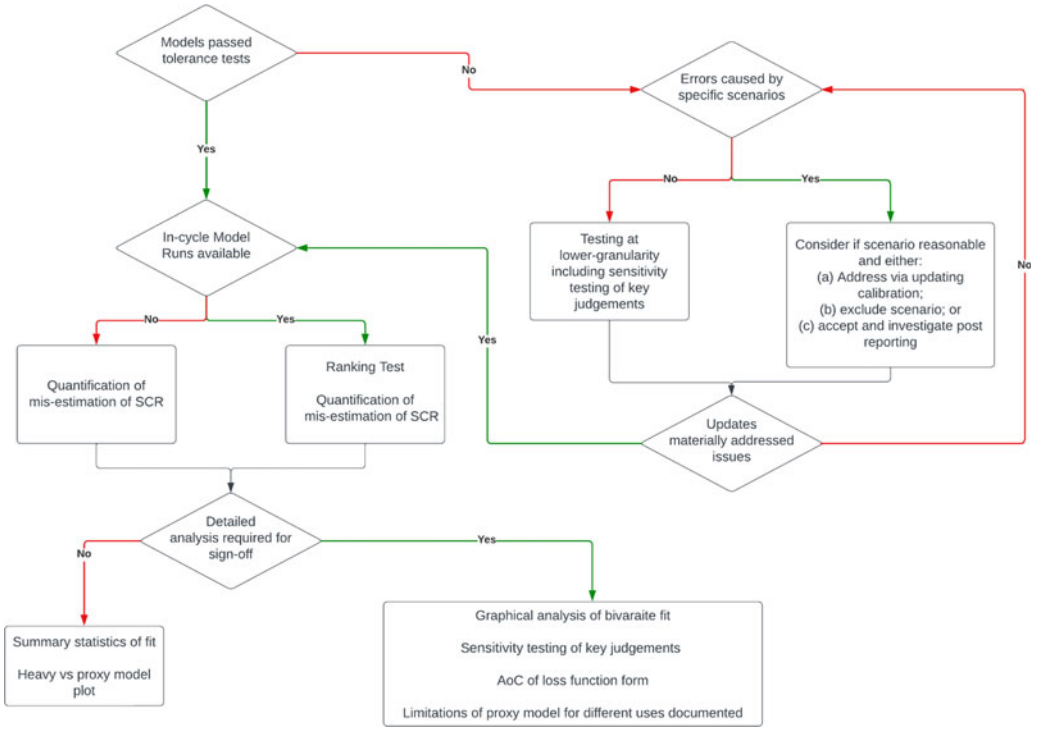

Figure 22. Illustrative decision-making process.

7. Roll Forward

Proxy models are, by definition, approximations to "heavy models" and are often calibrated outside of the reporting period. The proxy model calibration approaches can broadly be grouped into one of three groups:

(1) Proxy model form and parameters calibrated in-cycle
(2) Proxy model form fit out of cycle and parameters calibrated in-cycle
(3) Proxy model form and parameters calibrated out-of-cycle.

From an accuracy and relevance perspective, approach 1 is optimal as the proxy model will capture all material elements of the heavy model. However, this is dependent on the modelling infrastructure of the business and may not be achievable (in particular for stochastic modelling of with-profits business).

Approach 2 offers a more proportionate approach and relies on a smaller number of scenarios used to calculate the coefficients of the polynomial. In particular, for standard curve fitting approaches, Hursey and Scott (2012) outline that the optimal number of fitting points can be derived based on the form of the polynomial. We consider the above two approaches to be consistent with the PRA feedback (PRA, 2019, p. 11), which state "*Firms operating in line with best observed practice have invested in the efficiency of their modelling processes, allowing them to re-calibrate and validate the proxy model in full each quarter (i.e. on-cycle)*".

Approach 3 should only be used where either approach 1 or 2 is not proportionate or feasible given the modelling infrastructure of the business. In these instances, a "roll forward" or "true-up" approach should be applied. This refers to adjusting proxy models for the changes between the

calibration and calculation dates. Where roll forward is used, it is important that sufficient validation is performed to provide assurance the fit remains sufficient. This should ensure:

- Validation is performed in-cycle and prior to reporting
- Additional validation is performed to support the continual development of the roll forward
- Triggers are established to identify when the roll forward is not suitable.

Validation tests are discussed in Section 5 with the below sections discussing the roll forward adjustments and when they may not be suitable.

7.1. Roll Forward Adjustments

Under approach 3, adjustments are required to the proxy models to allow for changes both within the business and externally since they were calibrated. The ease of allowing for these is dependent on the specific change. When allowing for these directly within the model is either impractical or disproportionate, an out of model adjustment may be required. The derivation of these is not discussed further within this report.

A mature proxy model framework should include processes for identifying changes between the calibration and valuation date. These processes should inform the adjustments over a roll forward and include market movements, new business and run-off (versus expectations) and changes within the business (e.g. investment strategy, new product launches). The following sections consider how to allow for this within the proxy model framework.

7.1.1. Market risks

We believe that the risk distributions used within the proxy models should be updated to allow for market movements between the calibration and valuation date. Allowance for market movements is dependent on the risk being modelled and should be applied either via scalars or shifts. This is discussed in Table 3 using two examples prior to the theory being set out.

Table 3. Comparison of Risk Factors

Risk	Stress Type	RF Type	Calibration Date	Reporting Date	1-in-200
Credit spread	Absolute	Shift	75 bps	100 bps	140 bps
Equity	Relative	Scale	6,000	6,200	−40%

Market risks are modelled either as absolute movements, relative movements or a combination of the two (usually dependent on the point of the base value) in the risk. The proxy model therefore shows the impact of either an absolute or relative movement in the risk based on the starting value and the stress applied. Therefore, the proxy model is calibrated to ensure that it provides a materially equivalent impact to the heavy model given a specific stress. The challenge arises when the proxy model is calibrated to different base conditions to the heavy model and so a stress of 5% (say) is different for the proxy and heavy model. The roll forward must therefore correct for this.

Absolute Stresses

Where the risk is modelled as an absolute stress, the roll forward should "shift" the risk distribution by the market movement observed. This ensures that the stress applied in both models is consistent. Using the above example, the risk distribution should be shifted 25 bps "to the right" so that the 140bps stress applied is the same percentile in both the proxy and heavy model (see Figure 23).

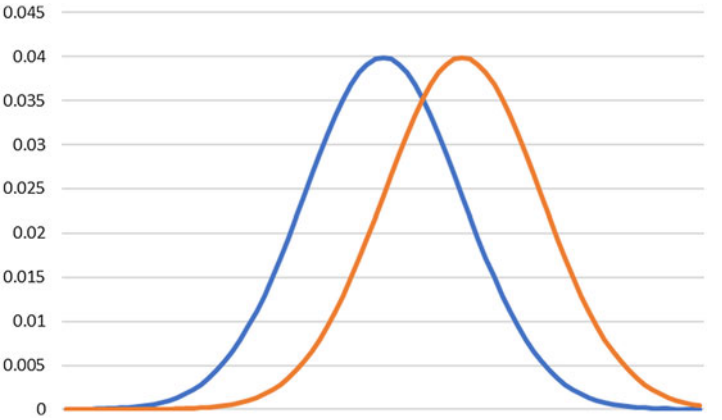

Figure 23. Graphical analysis showing impact of shift for absolute stresses.

Relative Stresses

Where the risk is modelled as a relative stress, the roll forward should "scale" the risk distribution by the market movement observed. Using the above example, the risk distribution should be scaled by 1.03 so that the 40% stress results in an equivalent index score (see Figure 24).

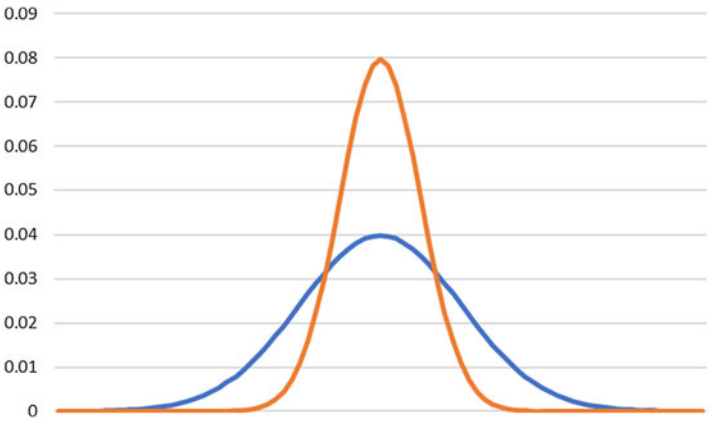

Figure 24. Graphical analysis showing impact of scaling for relative stresses.

7.1.2. Non-market risks

By their nature, we would not expect proxy models to have to be adjusted for changes in non-market risks on a frequent basis. This is because non-market risks are usually updated as part of an annual calibration process which can be timed around the calibration of proxy models. However, where changes to non-market assumptions do occur, these should be allowed for either in a consistent manner to market movements, via an overlay or a "true-up".

The true-up acts by scaling all terms in the polynomial by the change in a particular proxy between the calibration and valuation date. For example, using the change in BEL between calibration and valuation to scale up the coefficients. This is discussed below as part of the new business granularity discussion.

7.1.3. New business

New business written between calibration and valuation dates should be allowed for within the proxy model (otherwise this will contribute to a deterioration in fit). How this is allowed for is dependent on the granularity of the proxy modelling and the materiality of the new business.

Granularity

Proxy models should be calibrated to homogenous groups of products to enable them to accurately reflect the risks inherent. This may therefore allow different volumes of new business to be allowed for within the proxy models. This would enable different adjustments to be applied to different proxy models with the adjustment derived from the volume of new business.

The scalar can be derived by:

- Analysing the volume of new business at a level commensurate to the proxy modelling
- Using an appropriate indicator for the risk (e.g. sum assured, policy counts, premiums), scale the entire proxy model to reflect the increase in business (e.g. multiplier of 1.01).

Materiality

Where certain product lines may write little or no new business, the effort taken to allow for new business may be disproportionate given its materiality. The company should therefore consider whether making no allowance for new business is an acceptable limitation. In this scenario, justification for this should be documented. A more proportionate approach may be to derive a scalar annually/tri-annually (for example) and apply this as an approximation.

7.1.4. Run-off

Similarly to the new business considerations, the run-off of existing business should be allowed for within the proxy model (otherwise this will contribute to a deterioration in fit). The considerations are similar to that of new business (in respect of granularity and materiality) with the run-off indicator being consistent with that used within the heavy model. In practice, the two may broadly offset and therefore be an acceptable limitation.

7.1.5. Developments to the heavy model

Any developments to the heavy model between calibration and valuation can contribute to observed discrepancies between the proxy and heavy models. We are unaware of any standard methods to allow for these, and it is recommended that developments of the heavy model are planned taking into consideration calibration dates. Recognising this is not always possible, the impact of changes to the heavy model should be analysed to understand the impact on proxy models. This may then be adjusted for by:

- Applying scalars to the proxy models
- Holding an overlay.

In practice, we believe that an overlay may be more appropriate given the implicit and explicit assumptions that would be applied in deriving a scalar. If an overlay is used, the impact of the heavy models on the risk profile of the business should be considered. In particular, whether the change increases the exposure to any particular risk and should be therefore considered further as part of analysis of the SCR, sensitivities, and the Use Test perspective.

7.2. Allowing for Interactions and Diversification

Interactions are commonly included as a limitation and whilst it is relatively simple to roll forward individual risk factors through the use of shifts and scalars (as explained above), a roll forward of an interaction between two risks requires shifts and scalars along a plane (i.e. a 3-dimensional axis). The roll forward of individual risk factors will partially allow for interactions but will not capture all of the effects (as can be seen by considering Figure 25 where the roll forward allows for movement across two axes only).

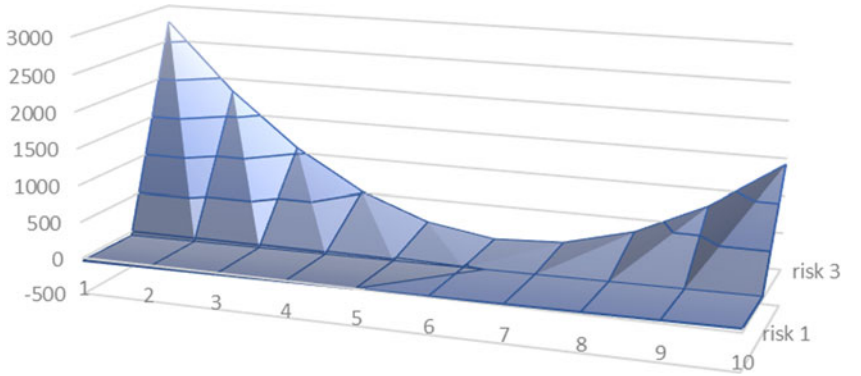

Figure 25. Graphical analysis showing impact of interactions.

Interaction between risks can be considered at different points of the plane, which highly depends on the use of the model. For example, considering interactions at the 1-in-200 smoothed scenario would be very different to the interactions of risk at more extreme scenarios (potentially relevant for Stress and Scenario Testing ("SST")).

Materiality is often used to justify limitations to the roll forward process. It is also equally important to consider the balance between complexity and accuracy.

7.2.1. Validations–Material interaction

Sensitivities can be performed to test the materiality of the cross-term X*Y from one period to the next.

As outlined in the previous section, the roll forward process can be expressed as a combination of additive risks (shift) and multiplicative risks (scalars). Individual risks X and Y are both subject to roll forward. The individual risk factors can be shifted or scaled and the result of roll forward can be compared against the cross-term X*Y for the next period.

If the SCR of the cross-term X*Y is comparable to the roll forward of individual risk factors, and the difference is less than the materiality threshold, then the interaction between the two risks is immaterial and can be included as a limitation.

7.2.2. Rolling forward interactions

Relationships can be modelled using a variance–covariance approach (such as the Solvency II standard formula approach) or copula simulation approach (commonly used within Internal Models). Assuming a copula simulation model is used, allowing for interactions would consider shifts to the copula. It is common to use the Gaussian or Student-t copula to model interactions.

$$f_{XY}(x, y) = c(u, v) f_X(x) f_Y(y)$$

The roll forward process may be used to adjust the proxy functions under each stress but rather than changing individual loss functions, the roll forward process can be applied to estimate impacts based on pairwise loss functions between two risks $f_{xy}(x, y)$ for risks X and risk Y. This will require the application of shifts and scales along a plane to reflect the roll forward of interactions (see Figure 26), but we note that this is not common in practice.

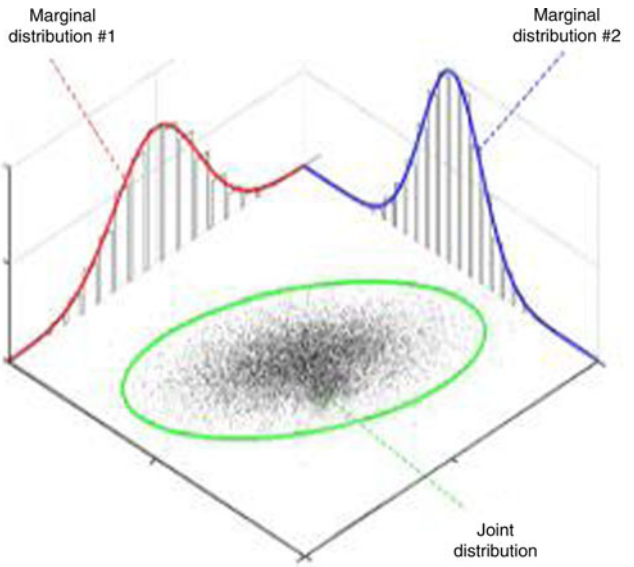

Figure 26. Graphical analysis showing joint and marginal distributions.

7.3. Other Uses of Roll Forward

We understand that the use of roll forward is in decline due to improvements in modelling capabilities, higher processing power and more efficient processes, reducing the need for the use of shifts and scalars. There has been scrutiny from PRA and auditors on use of roll forward and companies using roll forward models are required to perform back-testing to demonstrate reasonableness of the proxy model. This back-testing exercise can be a lengthy process and is therefore performed off-cycle.

Common uses of roll forward models include:

- Solvency monitoring for periods between reporting dates
- Setting risk appetites
- Scenario and sensitivity testing (including for Own Risk and Solvency Assessment (ORSA))
- Business planning

- Responding to ad-hoc requests (e.g. from management or regulators) that require modelled results to be provided. For example, providing estimates on risk capital to inform quick management decisions (although if risk profile is expected to change, it is best not to use roll forward).

7.4. Roll Forward Triggers

The roll forward process can be effective in stable conditions, for example, for business that is in run off, and account for risk movements (e.g. economic changes) and new business. However, the roll forward methodology may fail due to unpredictable events, such as a market crash, or management actions that change the nature of the risk, invalidating the risk distributions that were set at the time of calibration.

7.4.1. Uses of triggers

Triggers can be used to flag when the roll forward methodology is not appropriate and a true-up or re-calibration of the risk model is needed. It provides an indication of the continued validity of the stress test assumptions.

The process may include:

- Defining a trigger event (where realised losses exceed a pre-determined limit)
- Identifying the portfolio where the event was triggered
- Analysing root causes of this event (e.g. market fall)
- Examining how this root cause is reflected in the Internal Model
- Analysing the root causes of large movements in profit or losses.

Figure 27 considers triggers which can be used to identify events that would invalidate the roll forward methodology.

Event	Triggers	Action if trigger breached
Unforeseen regulatory changes e.g., caps/floors, pricing, capital requirements	These changes would invalidate the use of roll-forward	Unable to use roll forward method. May require change in roll forward methodology going forwards.
Large market volatilities	Change to key parameters used in calibration of risks. In particular, where market movements result in the 1-in-200 expected to be outside of the calibration range discussed in Section 4.	Recalibration
Changes in business strategy e.g., M&A/new business	Significant changes in business profile/exposure.	Revisit roll forward methodology, scalars/shifts and drifts.
Crystallisation of a non-modelled risk event e.g., mis-selling event	Sudden drop in share price, mass lapse event and/or significant impacts in NB	Consider materiality, scenarios and key risks. Revisit methodology.
Management actions e.g., hedging activity, de-risking	Significant changes in the risk profile/risk exposure	Changing scalars/shifts and drifts

Figure 27. Events, triggers and actions.

7.4.2. Principles for defining a trigger event

Trigger events can take many forms. The triggers can be defined as change in market movements, change in percentage of asset split, percentage of SCR, or as monetary values (e.g. £10 m). Ideally, the trigger framework should be derived from (or link back to) risk appetite and risk tolerance. For example, if SCR tolerance for a fund is 2.5% of diversified SCR then the trigger framework should be designed such that no single risk or combination of risks causes roll forward error of more than 2.5% SCR. This section sets out some principles for defining a trigger framework.

- The trigger framework should be reviewed when risk appetite is reviewed and at least annually.
- The trigger framework should define pass/fail criteria for each test.
- The trigger framework should specify tolerances for all material individual risk and interactions that are in-scope of the roll forward.
- The trigger framework should be designed such that it is appropriate for the risk. Below are examples of typical indicators that could be used to set triggers:
  - Interest rate: movements in 10-year swap rate
  - Equity: movements in FTSE 100 index
  - Equity volatility: movements in FTSE 100 implied volatility
  - Spreads: movement in credit spreads (bps) for corporate and sovereigns
- Trading activity: asset data/asset holdings changes during the roll forward period (maybe be difficult to obtain live data, data may be one month in arrears)
- How to true-up.

A proxy model is an approximation of the heavy model and at any given valuation date may over or understate the "true" model. Whilst the PRA does not permit "true downs" (i.e. reductions in capital requirements due to the proxy model overstating the heavy model), "true-ups" would be expected to ensure the capital held is appropriate. There is no defined methodology for how a true up is calculated but we believe that it should consider the following principles:

1. Reflect the purpose of the proxy model
   The true up should take into consideration the point on the distribution that is the primary interest of the proxy model. That is, if the proxy model is being used to calculate the SCR, it should take into account the fit of the model at the 1-in-200 point.

2. Timeliness
   As the proxy model will be recalibrated frequently, the true-up should be based on the current proxy model and, where possible, the valuation date. If it is not possible to use the current proxy model as at the valuation date, historical performance should be taken into consideration to avoid excessive fluctuations in any true up with a lag.

3. Use a number of scenarios
   Using a single point on the distribution to derive a true-up is unlikely to be representative of the entire curve and, if a 1-in-200 point is taken, may not be representative of the company's risk profile. Equally, a smoothed 1-in-200 point may not be sufficient to appropriately capture tail risks. We would therefore expect the true-up to consider the overall fit of the model across the distribution, including points stronger than a 1-in-200, and therefore be more representative for alternative uses of the model.

7.5. Back-Testing and Feedback Loops

The above mechanisms allow for adjustments to the proxy model for the valuation period but they are ultimately approximations and will not capture all the variations between the calibration and

valuation date. It is important that the fit following any adjustments is validated and results used to inform future valuation periods (either through an enhanced adjustment methodology or via future adjustments). Many of the tests outlined in Section 5 can be applied using scenarios as of the valuation date and a wide range of scenarios should be selected. This will ensure the adjustment's impacts across the entire distribution can be considered and provide assurance that the roll forwards are appropriate for other uses of the proxy model.

8. Conclusions

The UK life insurance market consists of complex products sold over decades with long run off periods. While the HM Treasury has indicated that, following Brexit, certain requirements of Solvency II are likely to change (under Solvency II UK reform, or SUK), fundamentally the use of models and scenario analysis is expected to continue for the foreseeable future and potentially expand into new areas (e.g. climate change scenarios). This, and the significant increase in interest rates over 2022, emphasises the importance of proxy models and the ability to demonstrate the ongoing appropriateness of the fit of the model. Whilst we would not expect widespread changes to a firm's existing proxy model approach, developments (and improvements in fit) are likely to be iterative.

From a calibration perspective, we believe the use of both judgement and data analysis in deriving the calibration scenarios and proxy models is important. History is no guarantee for the future and as 2022 demonstrated, purely relying on historical data can result in significant limitations. For example, many firms' Internal Model 1-in-200 interest rate stresses were weaker than what actually occurred over 2022. By applying expert judgement, proxy models can allow for known features (or limitations) of the underlying heavy model, or products. Equally, the choice of fitting methodology is dependent on the heavy model and business sold. It is, however, important that proxy models provide a robust fit in different economic conditions.

Whilst the PRA feedback sets out a number of tests to be performed, it is apparent that there is no one size fits all approach. It is therefore vital that firms perform sufficient testing to provide assurance that the proxy model is appropriate for all uses of the model. Further, the significant market volatility since 2021 has demonstrated the importance of roll forwards and in-cycle validation. The in-cycle validation is an area where we believe further development is required (with this dependent on modelling infrastructure which may prove more challenging following the introduction of IFRS17). This is a key control in ensuring that the proxy models continue to rank risks appropriately (this ranking may have varied significantly as interest rates change).

Whilst we would encourage firms to continue to develop their calibration and validation framework, the focus should be on ensuring they can robustly provide assurance that proxy models are representative of the heavy model. Any limitations in the validation approach should be outlined, with triggers in place for review on a periodic basis, or where materiality may increase. Stakeholders should be made aware of these limitations, including in the context of different model uses, with expert judgement applied where necessary.

Glossary.

PRA	Prudential Regulation Authority
PM	Proxy model
HM	Heavy model
AIC	Akaike Information Criterion
BIC	Bayes Information Criterion
LASSO	Least Absolute Shrinkage and Selection Operator
OLS	Ordinary Least Squares
LSMC	Least Squares Monte Carlo
RMSE	Root Mean Square Error
NAV	Net Asset Value
NCOGS	Net Cost of Options, Guarantees and Smoothing
SST	Stress and Scenario Testing
SII	Solvency II
IM	Internal Model
SCR	Solvency Capital Requirement

References

EIOPA. (2022). EIOPA Guidelines 24 to 27, available at https://www.eiopa.europa.eu/document-library/guidelines_en (accessed 23 March 2023).

European Parliament (2009). Solvency II Directive, available at https://www.eiopa.europa.eu/browse/regulation-and-policy/solvency-ii_en (accessed 23 March 2023).

Harrell, F. E. (2022). *Regression Modelling Strategies* (2nd ed.). London: Springer.

Hursey, C., & Scott, R. (2012). Replicating Formulae Efficient Calibration Techniques, available at https://www.noca.uk/wp-content/uploads/2020/12/Replicating-Formulae-Efficient-Calibration-Techniques-Final-1.pdf (accessed 23 March 2023).

IFoA Aggregation and Simulation Working Party. (2016). Simulation Based Capital Models Testing, Justifying, and Communicating Choices, available at https://www.actuaries.org.uk/system/files/field/document/simagg_2016_06_06_v2a_no_tc%20plus%20front%20back%20pages.pdf (accessed 23 March 2023).

IFoA Proxy Model Working Party. (2014). Heavy Models, Lite Models and Proxy Models Paper [pdf], available at https://www.actuaries.org.uk/system/files/documents/pdf/proxy-models-working-party-paper-240214.pdf (accessed 23 March 2023).

Murphy, P., & Radun, M. (2021). Proxy Model: Uncertain Times, The Actuary [e-journal], available at https://www.theactuary.com/features/2021/03/03/proxy-models-uncertain-terms (accessed 23 March 2023).

PRA. (2019). Proxy Modelling Survey: Best Observed Practice, available at https://www.bankofengland.co.uk/-/media/boe/files/prudential-regulation/letter/2019/proxy-modelling-survey-best-observed-practice (accessed 23 March 2023).

Runge. (1901). Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten. *Zeitschrift für Mathematik und Physik*, 46, 224–243 [On empirical functions and interpolation between equidistant ordinates. *Journal of Mathematics and Physics*, 46, 224–243].

Smith, G. (2018). Step away from stepwise. *Journal of Big Data*, **5**, 32. https://doi.org/10.1186/s40537-018-0143-6 (accessed 23 March 2023).

Steyerberg, E.W., Eijkemans, M.J., Harrell Jr, F.E., & Habbema, J.D. (2001). Prognostic modelling with logistic regression analysis: In search of a sensible strategy in small data sets. *Medical Decision Making*, **21**(1), 45–56. https://doi.org/10.1177/0272989X0102100106 (accessed 23 March 2023).

Appendices

A.1. Model Points and Assumptions

Figure A.1 summarises the model points used in the case study. Figure A.2 shows the base economic assumptions used in the case study. Each model point represents 10 individual annuities.

Modelling

A cash flow projection model (the "heavy model") was developed to project the future benefit and expense cash flows for the portfolio under different scenarios. Figure A.3 shows the projected future cash flows (benefits and expenses) in the base scenario and in two stress scenarios (an improvement in baseline mortality rates and an improvement in assumed mortality improvements).

Number of model points	257
Joint / single lives	141 single lives, 97 joint lives, 19 reversion
Single life annuities	72 females, 69 males
Joint life annuities	43 females life 1 and males life 2 54 males life 1 and females life 2
Reversionary annuities	10 females life 1 and males life 2 9 males life 1 and females life 2
Age (life 1)	Between 56 and 90, average 71.3
Age (life 2)	Between 49 and 97, average 70.9
Annuity guarantee period	No guarantee: 21 5 year guarantee: 175 10 year guarantee: 58 20 year guarantee: 3
Duration in force	Between 0 and 20 years, average 8 years
Annuity amount (annual)	Between 12,159 and 59,285, average 42,683
Escalation type	All fixed escalation
Escalation rate	Fixed between 1.52 and 3.52, average 1.96
Reversion rates	Joint life: 50% (31), 67% (40), 75% (21), 100% (24)
Payment frequency	Monthly (114), Quarterly (59), Semi-annually (23), Annually (61)
Payment type	Advance (192), Arrears (65)
Months to escalation	Between 1 and 12, average 6.77

Figure A.1. Model points.

Valuation date	31 December 2020
Yield curve	Bank of England curve as at valuation date
Inflation	Derived from Bank of England RPI spot rates at valuation date
Expense inflation	RPI + 0.50%
Matching adjustment	0.80%
Mortality rates	Male: NLT15-17(E&W) (Male), CMI_2018_M [1.75%] Female: NLT15-17(E&W) (Female), CMI_2018_F [1.75%]

Figure A.2. Model assumptions.

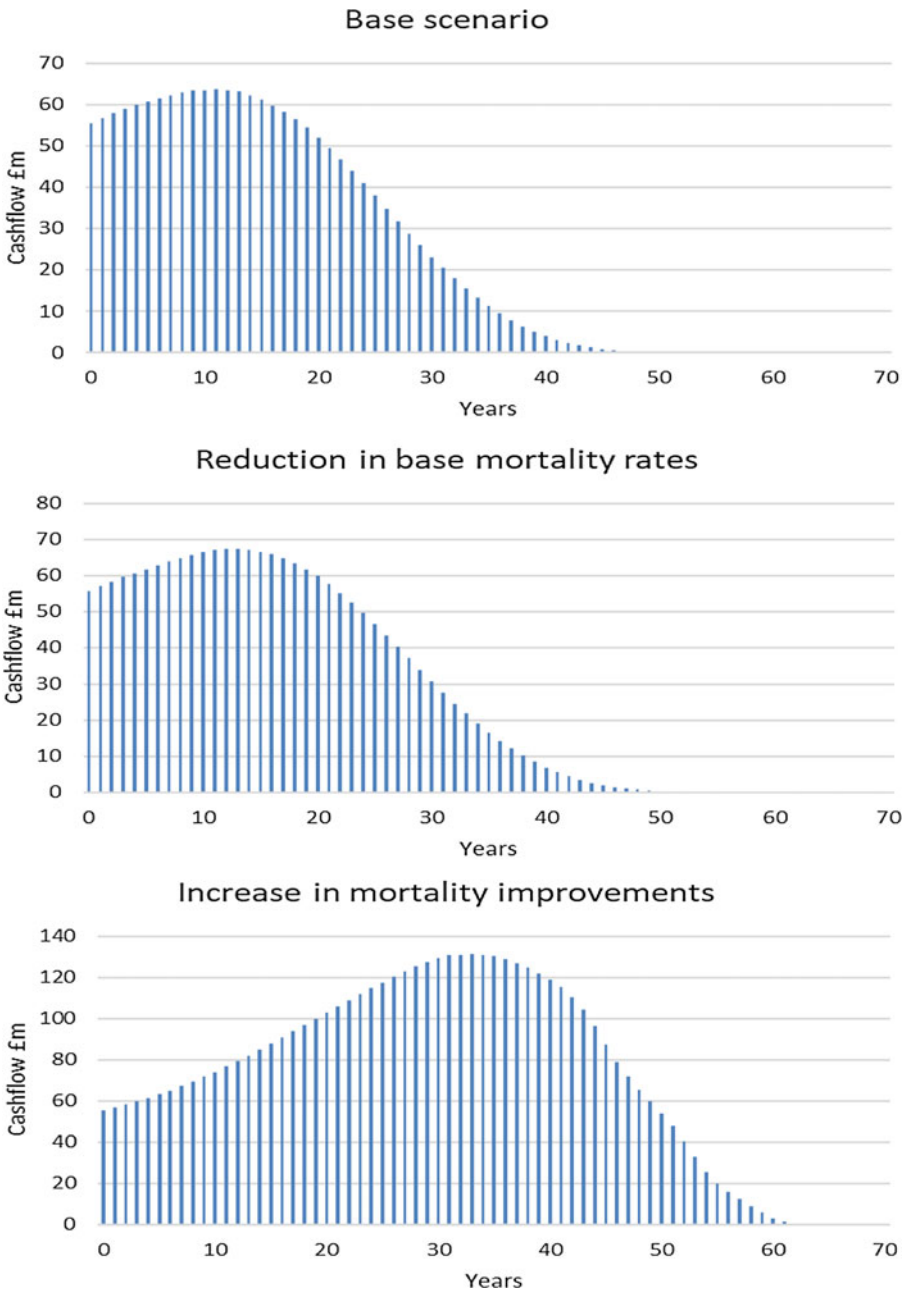

Figure A.3. Cashflow projections.

Pseudo-heavy Model

As heavy models can often require significant time to run, a "pseudo-heavy" model was also produced, using the output of a small number of heavy model runs to allow the user to produce the large number of different scenarios required for calibration and validation of proxy models.

The pseudo heavy model uses risk driver coefficients for longevity risks (base and improvement) and is calibrated to the full heavy model results to model cash flows under different scenarios. These risk types are not easily applied analytically. All other risk drivers – expense, interest rate (3 principal components plus matching adjustment), expense inflation – are applied analytically.

Risk drivers up to the sixth order are used within the proxy models investigated.

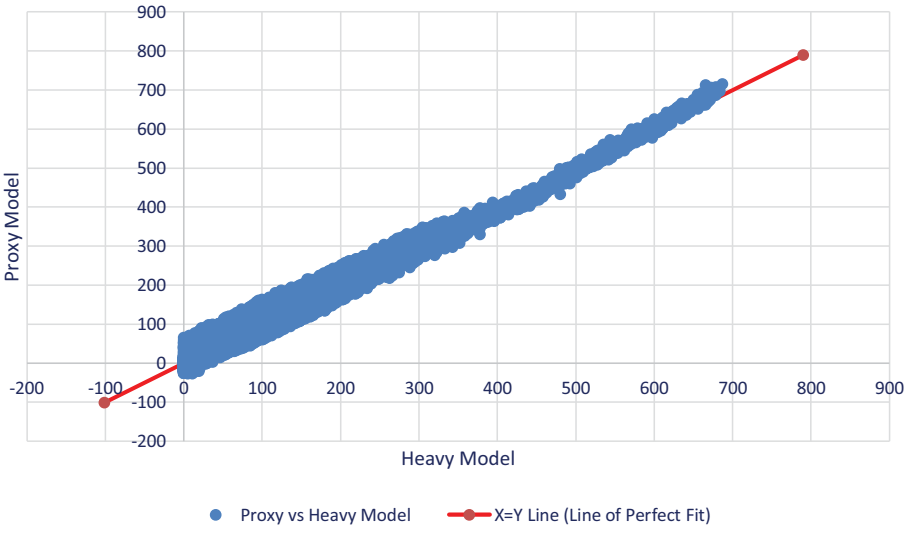

Figure A.4. LSMC validation scenarios.

Base scenario cash flows are adjusted by multiplying the risk drivers by the risk factors specified in any given scenario. The result is an adjusted set of future cash flows (i.e. benefits and expenses for annuities).

The following risk factors are incorporated into the pseudo-heavy model by adjusting the discount rate being used to calculate the present value of the cash flow:

- Interest rate risk
- Inflation risk
- Expense inflation
- Matching adjustment

The adjusted discount rate can then be applied to the adjusted cash flows to calculate a present value of future cash flows. The result of the model is a set of adjusted cash flows and present values that can be produced instantaneously for a scenario with a specified set of risk factors (removing the need for the full model to be run). It is important to note that the order of the coefficients used within the pseudo-heavy model is required to be greater than those used in any proxy model that is being calibrated/validated.

The key benefit of the pseudo-heavy model is that it allows a significant number of additional scenarios to be produced almost instantaneously. In particular, it can be used to produce calibration and validation scenarios and has been used to explore the calibration and validation approaches.

Least Squares Monte Carlo (LSMC)

The LSMC approach was tested with a with-profit fund example, and allowing for four risk drivers – interest rate, equity level, equity volatility and mortality rates. The example allowed a single management action to vary the equity backing of the with-profits liability to maintain solvency.

The "true" cost of guarantee was calculated using Black Scholes, while simulated valuations (the inner scenarios) were performed using a geometric Brownian motion for the equity value.

Figure A.4 shows the fit achieved with this modelling strategy.

- The root mean square error (RMSE) is circa £18 m
- The largest error is approximately £68 m.
- 99% of the errors are in the interval (−£39 m, +£55 m)

In this example, the errors are significant in comparison to the underlying liability being approximated. This is due to the nature of the cost of guarantee which has properties that make it difficult to approximate with a polynomial. To give a concrete example of this difficulty, the cost of guarantee, by definition, does not go below zero. However, a polynomial model is generally not constrained in this way and, as can be seen in the figure above, will produce negative values.

Cite this article: The IFoA Proxy Model Working Group: Wollam D, Kuona M, Thomson M, Liu R, and Paton S. (2024). Consideration of the proxy modelling validation framework. *British Actuarial Journal*. https://doi.org/10.1017/S1357321723000193