



Institute
and Faculty
of Actuaries

Real World Scenario Testing

Andrew Smith, Deloitte UK

Jason Doughty, Deloitte UK

Eamon Howlin, Deloitte Ireland



Disclaimer

The views expressed in this presentation are those of the presenters and not necessarily those of their employers or professional bodies of which they are members



Institute
and Faculty
of Actuaries

Agenda

- Introduction
- In sample testing
- Out of sample testing
- Use of expert judgement
- Tests on generated data



Introduction

- Use of stochastic models now widespread
 - Reserve uncertainty / premium risk models
 - CAT models
 - ESGs
- Important to remember that the true underlying distribution is unknown (possibly unknowable)
 - Use a substitute or “Ersatz” model in place of the real world process
- However we need to understand choices made in building a model
 - Models are becoming more widely relied on
 - Used for more business critical purposes
- A model doesn't need to be perfect – just fit for its purpose

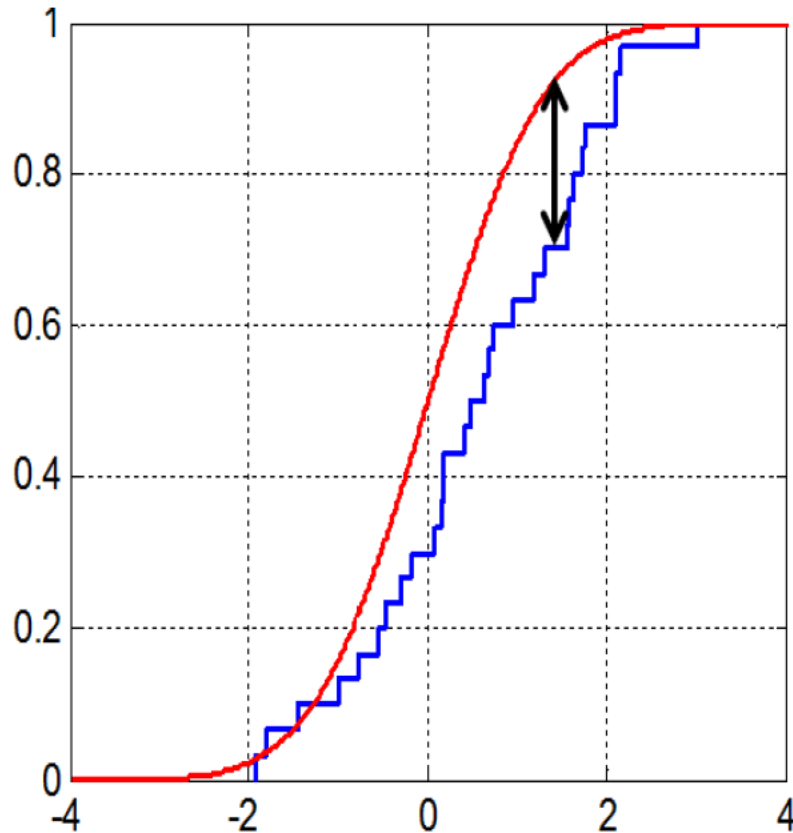


Introduction

- Traditionally model tests have concentrated on the performance of a particular instance of a model
 - Parameterise a model from observed historic data
 - Use “Goodness of fit” tests
- Model builders often augment observed data with judgement from business experts
 - Observed sample is incomplete (and possibly not representative of the population of potential scenarios)
 - Events not in data can lead to understatement of some measures of risk (e.g. tail dependency)



In-Sample tests



- Statistical models are usually calibrated (at least in part) using historic data
 - Motor claims frequency
 - Bootstraps
 - Dependencies
 - Investment returns

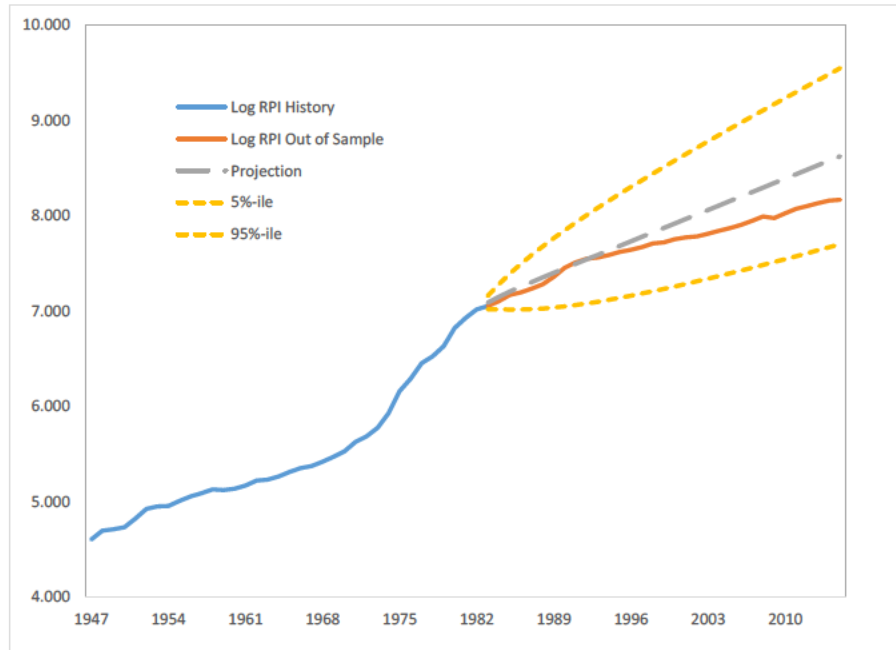


In-Sample Tests

- In-sample tests assess if the fitted model could have generated the data used to calibrate it?
- Satisfactory in-sample test does not imply a good model!
 - e.g. as a result of overfitting
 - e.g. fitting a stock market model to a bull or bear market
 - May not possess reasonable predictive power despite very favourable in sample results

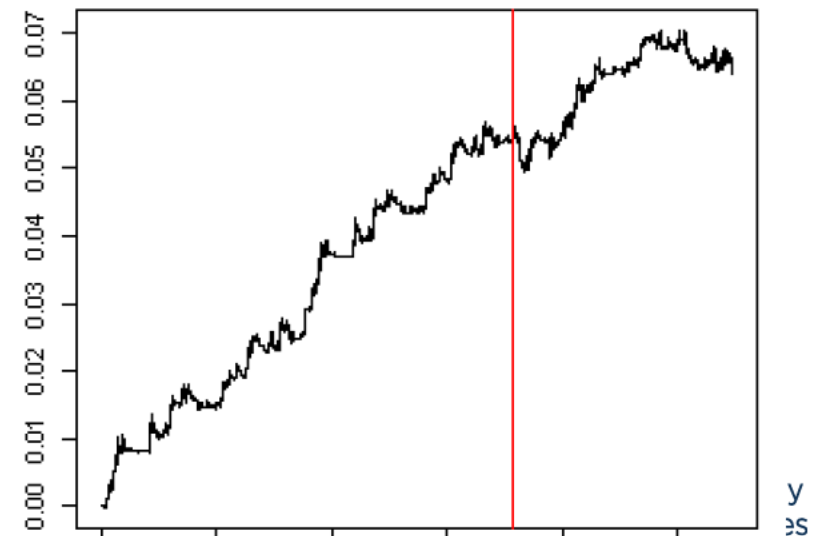


Out of Sample Tests



- Hold out samples
- “Wait and see” approach
- Build model in one sector / region & test on another

Out of sample testing examines whether model predictions are consistent with subsequently emerging data

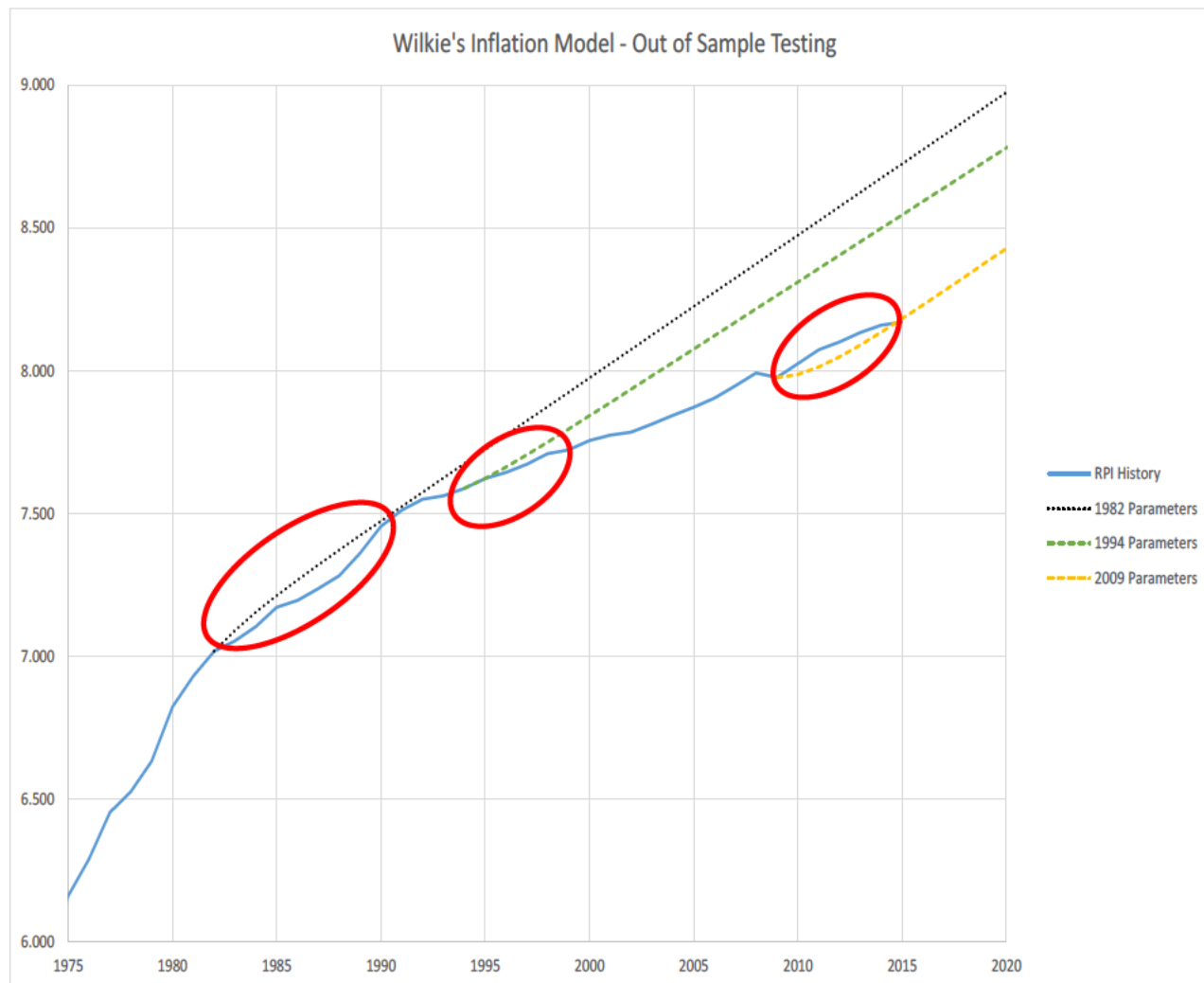


Out of Sample Tests

- Generally more powerful than in-sample tests
 - Can illustrate / overcome over fitting issues
 - Models are generally used for making predictions!
- Often leads to the most high profile model criticisms
 - Recent financial crisis when actual losses often many multiples of the stated VaR
- Use a model to construct a distribution of forecasts
 - Use p-values / confidence intervals
 - Visualised via a “funnel of doubt”
- Some limitations
 - Test result maybe unknown when calibrated (can involve an element of “waiting”)
 - Modeller bias if out of sample testing is performed on a hold out sample



Example – Wilkie Model for inflation



Note: all values are on a logarithmic scale

Model appears to predict well for a number of years after each set of parameters is derived before actual experience diverges from expected

Data Period	QA	QMU	QSD
1919-1982	0.60	0.050	0.0500
1923-1994	0.58	0.047	0.0425
1923-2009	0.58	0.043	0.0400



Institute
and Faculty
of Actuaries

Use of expert judgement

- Incompleteness in data is prevalent in non-life insurance
 - Limited relevant history causing events to not be in the data
 - Changes in social, judicial or legislative environment may cause some data to become irrelevant
 - Climate Change
- Expert Judgement used to supplement data should be based on a foundation
 - Developed from facts which are auditable (Policy Limits, Historic Exposure measures)
 - Based on a process which can be reviewed by another expert
 - The level of uncertainty should be understood
 - A standard and criteria for rejecting expert judgement should exist



Validation of expert judgement

- Expert Judgement is notoriously difficult to validate
 - Often focussed on extreme events (return periods greater than 1-in-50)
 - Based on expert's own experiences (how different would a casualty underwriter today see extreme claims than a casualty underwriter in the 90s?)
 - Inherently forming assumptions where empirical information is insufficient (could introduce bias – which may not be understood)
- Validating Expert Judgement improves the understanding of the Model's Limitations
 - Validation is a process to identify model weaknesses
 - Solvency II requires that judgement is 'falsifiable' (e.g. Lloyd's minimum standard SCU 4.1)
 - Failure in a validation test does not necessarily require a resolution (i.e. It could be the identification of a Model Limitation)
 - The acceptance of judgement (at times) may be an article of faith which outsiders to the process might legitimately question

Why use generated data tests?

- Traditional model tests can have weaknesses:
 - Limited data (especially for extreme events) can lead to low power
 - Risk of cherry picking
 - Risk of over-fitting
 - Philosophical issues (we are actively seeking out a Type II error)
- *“Instead of asking whether our model is correct we should ask whether our objective in building the model has been achieved” – Mark Davis¹*

1. Verification of internal risk measure estimates



Why use generated data tests?

- Tests based on historic data can be a poor test of a model's predictive power
- Your model is only likely to be proved wrong when it fails and costs you money!
- “Lab testing” a model using generated data allows you to evaluate its strengths and weaknesses in a safe environment



Setting the scene²

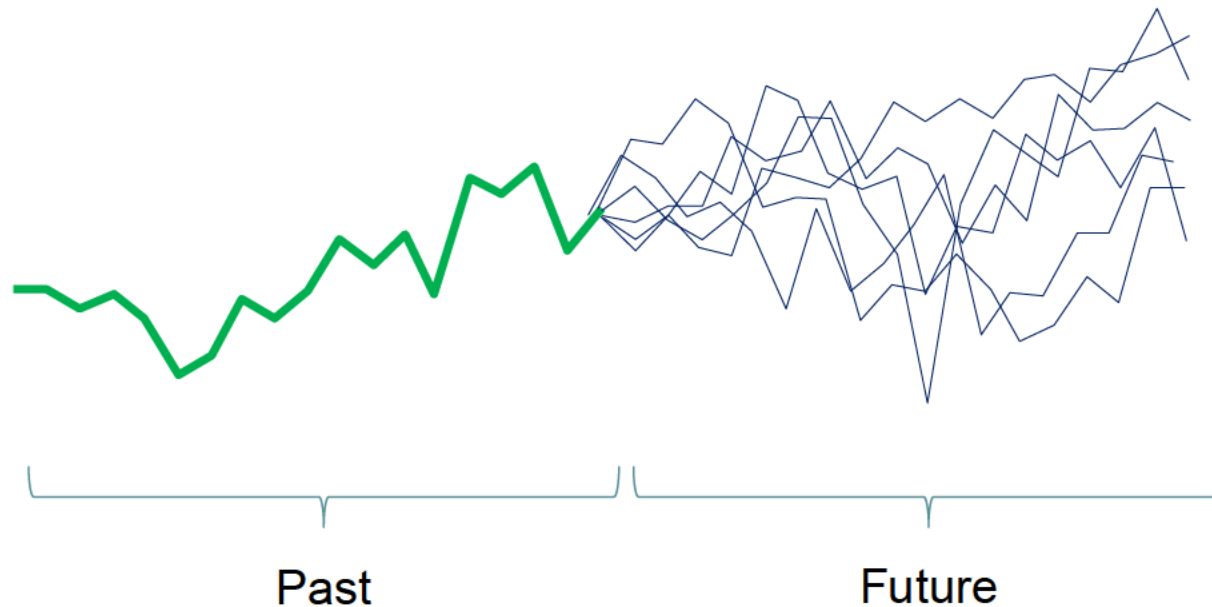
- “All models are wrong but some are useful” (George Box)
- Ersatz models:
 - True probability law is unknown
 - Build a model based on past experience – an imperfect substitute for the true model
- We are testing the way in which models are built rather than a particular instance of the model
 - Allows us to use computer generated data
- Reference models:
 - The models that generate the test data

2. See Ersatz Model Tests – Jarvis, Sharpe and Smith



Institute
and Faculty
of Actuaries

Traditional Approach

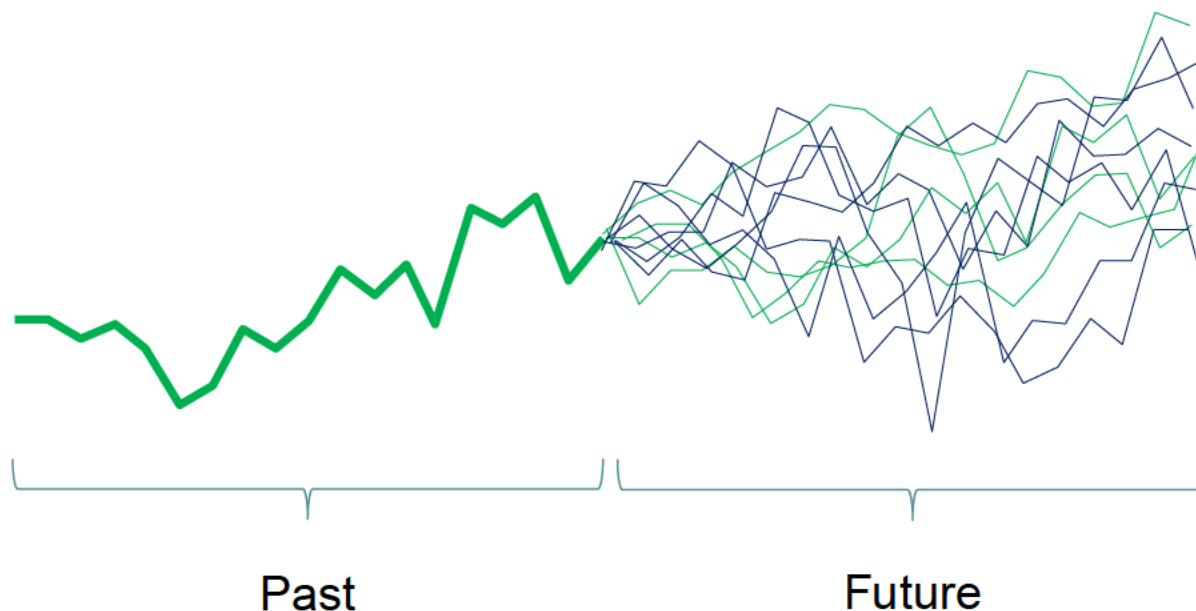


- We want the statistical properties of the ersatz blue (scenarios) to resemble those of the green process
- We see only one version history so we are condemned to only base our forecast on this evidence



Institute
and Faculty
of Actuaries

Generated Data Approach



- If we can observe parallel histories and multiple associated futures then we can assess how close our ersatz model is to the “true” process



Generated Data Test Process

1. Generate a long test data series, split into a past portion and a future portion using the reference model
2. Use the past portion to fit an ersatz model without reference to the original generating process
3. Run the ersatz model based on the past data portion to give forecast future scenarios
4. Compare the future scenarios from the ersatz model to the future scenarios from the original data
5. Repeat for many other generated data series



The model passes if the scenarios from the ersatz model are sufficiently representative of the scenarios from the original generating process.



Institute
and Faculty
of Actuaries

Generated Data Test Results

- Multiple types of statistical tests can be performed:
 - Tests of parameter bias
 - Percentile based tests
 - Monte-Carlo back tests on generated data
- Theoretically there is no limit to the power of a generated data test!
- What are we testing?
 - Consistency – how effective is the model fitting process when the basic assumptions are satisfied?
 - Robustness – how effective is the model fitting process when the assumptions are not satisfied?



Disadvantages of generated data tests

Generated data tests can address questions that are unanswerable from tests based on real data **BUT:**

- Criteria for the choice of reference models is unclear
- Requires the ability to recreate what the fitted model would have looked like under alternative histories
- Only tests the model building process and not a specific instance of the model
- Difficult to subject expert judgement to generated data tests
- Ersatz model calibration can be a time consuming process
- Test results can / will conflict given the unlimited power of the test
- The more precisely we can formulate the expert process, the better we can test it



Institute
and Faculty
of Actuaries

Are generated data tests the solution?

- Key Question: Does using generated data tests mean you prevent model failures?
- Model risk is not just a quantitative issue, social and cultural factors also play a role
- Generated data tests can:
 - Reflect events too rare to feature in historical data sets
 - Provide objective evidence of where the strengths and weaknesses of models lie
 - Help to improve the corporate risk culture towards model risk



Questions

Comments

The views expressed in this [publication/presentation] are those of invited contributors and not necessarily those of the IFoA. The IFoA do not endorse any of the views stated, nor any claims or representations made in this [publication/presentation] and accept no responsibility or liability to any person for loss or damage suffered as a consequence of their placing reliance upon any view, claim or representation made in this [publication/presentation].

The information and expressions of opinion contained in this publication are not intended to be a comprehensive study, nor to provide actuarial advice or advice of any nature and should not be treated as a substitute for specific advice concerning individual situations. On no account may any part of this [publication/presentation] be reproduced without the written permission of the IFoA [*or authors, in the case of non-IFoA research*].



Institute
and Faculty
of Actuaries