Setting Benchmarks for Asset Model Percentiles

This note describes some possible methods for setting numerical standards for stochastic asset models used in capital assessments for life, health and general insurers.

The Purpose of Setting Standards

The purpose for setting standards is to enable meaningful comparison between capital disclosures for different insurers.

Consider the following example of two insurers:

Insurer "A" discloses
- economic capital requirements of £100m
- based on a 0.5% ruin probability
- measured over 1 year
- using a market-calibrated Hull-White model

Insurer "B" discloses
- economic capital requirements of £100m
- based on 10% ruin probability
- with a 30 year run-off period, testing future solvency at 5 year intervals
- using a historically calibrated Wilkie model

Users of this information can reasonably ask which insurer has applied the more stringent capital test. Ideally, analysts would wish to re-run insurer "B" using the model belonging to model "A", and vice versa. This is unlikely to be possible, give that the data and the models may be proprietary.

Setting standards for capital models is inherently more difficult than setting standards for market consistent valuations. Differences in market consistent valuation models arise chiefly from different choices of calibration assets. Standardisation of calibration assets goes a long way towards achieving comparable results. On the other hand, two insurers with exactly the same business and identical stochastic models could still disclose different capital requirements by choosing different percentiles, different time horizons or different solvency testing frequencies.

The definition of default is also contentious. The simplest model is that default occurs when liabilities exceed assets. In company law, this is the point at which control theoretically passes from shareholders to bondholders. However, in the insurance industry the regulator plays a greater role. The degree of regulator intervention is likely to increase well before technical insolvency. If shareholders are averse to losing control of a business, then the modelled ruin even could arguably relate to a failure to meet regulatory solvency margins, rather than technical insolvency.

There is another argument which works in the opposite direction. Many insurers' share price contain a franchise value element which is not captured in their accounting net assets. Shareholders can realise this franchise value by financial restructuring, by injecting more equity or by selling the business to a third party, rather than by exercising the option to default. This is why the financial services industry has seen few bond defaults, but many ailing companies are refinanced or taken over by stronger peers. There is an argument that if an insurer targets a default probability of (for example) 0.5%, they should be permitted to disregard potential ruin scenarios in which a private third party comes to the rescue, provided policyholders do not lose out in this event.

Standardising Model Parameters or Model Output

A framework for comparing outputs from different models can use three elements
(i) restriction on the range of permitted models, for example by standardised parameters or standard tests on output reasonableness. The tests can be two-sided or one-sided. For example, a two sided test could require equity volatilities lie between 20% and 25%, while a one-sided test would require only that the volatility was not less than 20%.
(ii) disclosures of sample model output, so that models can be compared. For example, insurers might be expected to tabulate sample means and standard deviations of equity market levels and interest rates over various horizons. Alternatively, combinations of model and capital definitions might be disclosed. For example, offices might be required to compute the required capital under their model and capital definition for an office who invests 100% in equities to deliver a fixed policyholder cash flow in 10 years' time.
(iii) algorithms for converting one form of model output for another. For example, we might seek algorithms to estimate the 99.5th percentile of a distribution given the 90th percentile and some other properties of the distribution. We might also seek algorithms to estimate a run probability over 5 years given the result over 1 year. These algorithms will inevitably be empirical and valid only for limited purposes. For example, there is no mathematical formula for deducing a 99.5th percentile from a 90th percentile for arbitrary distributions. All we know is that the 99.5th percentile must be the larger figure. However, if we had knowledge of the sort of distributions that arise in insurance work then we may be able to build crude comparisons that are valid for a limited range of businesses.

There is a trade-off between the emphasis given to the different elements. At one extreme, a uniform approach could impose a single model and capital definition on all participants, in which case the model disclosures, (ii), would be the same for everybody and there would be little need for the algorithms in (iii). At the opposite extreme, companies could be given unfettered discretion in model choice, in which case the disclosures would need to be extensive and analysts would need to rely on a large number of conversion algorithms to make sense of the output. It may be costly to develop conversion algorithms which work with sufficient generality. Businesses may also resent the cost of producing figures which are required only for comparison purposes with peers rather than for the running of the business. Every additional restriction on the choice of available models has a benefit of placing fewer demands on the conversion algorithms and reducing the number of required outputs for comparison purposes.

Random Walk Capital Models

The simplest conversion rules arise with capital models based on random walks. We assume that solvency is measured in continuous time, and we define a process X_t by

$X_t = \log\{\text{ assets}(t) / \text{liabilities}(t)\ \}$

The insurer is solvent so long as $X_t \geq 0$. We can distinguish two definitions of ruin:

The "giant leap" method: **Prob**{giant leap ruin} = **Prob**{$X_t < 0$}
The "continuous sampling" method: **Prob**{continuously sampled ruin} = **Prob**{X_s<0 for some $0 \leq s \leq t$}

We can calculate these probabilities analytically under the special condition that $\{X_t\}$ is a geometric random walk. In these cases, we have the following results:

Great leap probability of ruin at date $t = \Phi\left(-\frac{X_0 + \mu t}{\sigma\sqrt{t}}\right)$

Cumulative probability of ruin on or before date t =

$$\Phi\left(-\frac{X_0 + \mu t}{\sigma\sqrt{t}}\right) + \exp\left(-\frac{2\mu X_0}{\sigma^2}\right)\Phi\left(\frac{-X_0 + \mu t}{\sigma\sqrt{t}}\right)$$

Fitting to Default Statistics

The cumulative probabilities are useful because we can compare these to cumulative default frequencies on rated bonds. The S&P European default study tabulates the following default probabilities over horizons from 1 to 5 years:

EU Cumulative Average Default Rates (1981-2003)

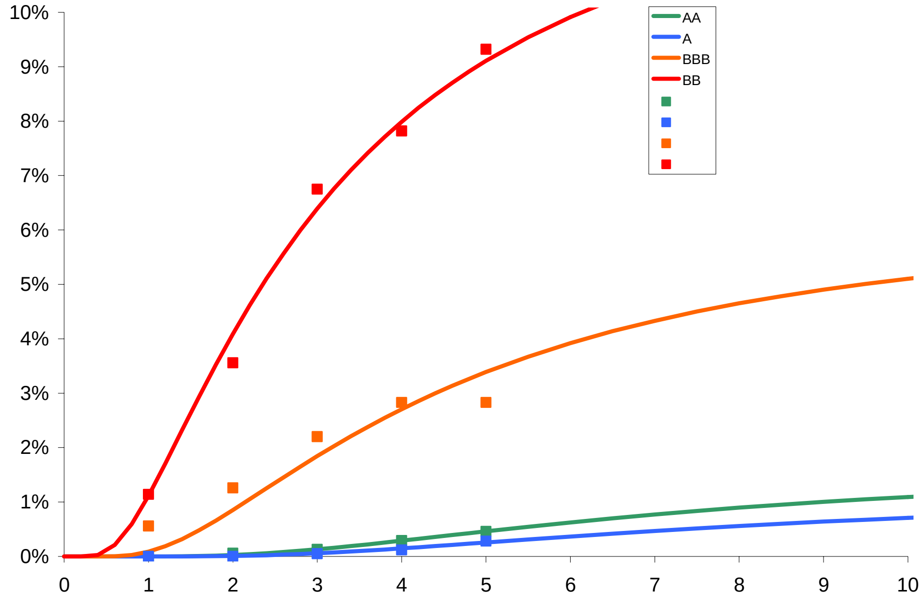
	0	1	2	3	4	5
AAA	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
AA	0.00%	0.00%	0.06%	0.13%	0.29%	0.46%
A	0.00%	0.00%	0.00%	0.05%	0.12%	0.28%
BBB	0.00%	0.56%	1.26%	2.20%	2.83%	2.83%
BB	0.00%	1.14%	3.56%	6.75%	7.82%	9.32%
B	0.00%	7.17%	19.22%	27.58%	31.23%	33.60%
CCC	0.00%	60.71%	60.71%	60.71%	60.71%	60.71%

Assuming a geometric random walk, we can identify parameters that best fit this history. Although these default statistics are real data, they are subject to sampling error. For example, the historical statistics show higher defaults for "AA" bonds than for "A" bonds. This is probably a statistical fluke; most reasonable models should show the reverse.

Let us make the heroic assumption that the bonds of different grades are modelled by the same process, that is, with common parameters μ and σ, but represent different starting points X_0. Possible parameters, fitted by least squares, are tabulated below:

	X0	mu	Sigma
AA	4.2534	0.4896	1
A	4.6231	0.4896	1
BBB	2.8786	0.4896	1
BB	2.1153	0.4896	1

The goodness of fit is illustrated in the chart below. Unfortunately, the fit is worst for the BBB default statistics, on which the FSA has placed greatest emphasis.



Caveats for Comparing Models to Default Statistics

There are a number of reasons why we would not expect model output to replicate historic default statistics for similarly rated entities. These include:

- Differences in default probability for different entities with the same rating. Credit rating models are typically based on financial ratios, combined with a subjective element. Stochastic models have not played any part in the historic ratings, although some agencies are considering the use of such information in future. Therefore, it is wrong to suppose that the default probability is a function of the rating grade alone. The historic default rate is an estimate of the average default probability for all issued bonds with a given grade. There is no reason why a specific modelled firm or industry should not have a higher or lower default probability than the average.
- The default event as modelled is not the default event captured in the statistics. The modelled default event in most stochastic models is assets falling below liabilities, but policyholders do not necessarily lose out in this event, for example if a third party acquirer comes to the rescue.
- Bottom up estimates of the parameters μ and σ may not reflect all the risks that actually cause companies to fail. For example, companies may fail due to fraudulent or incompetent management, but that same management is unlikely to allow for those contingencies in their stochastic model.
- As previously noted, historic default statistics are subject to sampling error, and do not necessarily represent the true default probability. For example, in the last 20 years, no European AAA bond has defaulted, but few analysts would deduce that AAA bonds can never default.

Disclosing Parameters or Model Output

There are at least three levels of possible disclosures around asset model assumptions.

The first level disclosure is the chosen model parameters. If an analyst had access to the model details, then knowledge of the parameters used would (in principle) enable the analyst to replicate the model output. To the extent that model structures are proprietary, or mathematically advanced, parameter disclosures are of more limited use. A modeller can disclose that "$\xi = 25$" but this is only helpful if the user of the disclosures can understand how this parameter is used within a model.

The second level disclosure is sets of asset model output statistics. Common disclosures include means and standard deviations. These statistics are tabulated, for a range of models, by Lee and Wilkie (2002). In the context of solvency tests, adverse percentiles may be of more relevance. See for example [can anyone find chapter and verse on the Canadian regs?]

A third level disclosure is output of required capital for a number of standard situations. This would include the firms' own definition of required capital, and could allow some trade-off between model choice and capital definition. For example, some firms might choose mean reverting equity models while others use a random walk approach. These results could still be consistent provided that the mean reverting firms used a more extreme percentile in their capital assessment.

To illustrate this, we consider a simple case of equity mismatch risk. An equity return index, measured in excess of the risk free rate, is modelled as a geometric random walk with geometric risk premium of 4% per annum and annual volatility of 20%. This equates to an arithmetic risk premium of 6% per annum.

The first level disclosure is a description of the model and parameter tabulation
$\mu = 0.02$
$\sigma = 0.20$

The second level disclosure looks at percentiles of an equity total return index, in this case measured relative to a cash total return index. These are as follows:

horizon	1	2	5	10	20	50
percentile						
0.1%	0.5610	0.4520	0.3067	0.2113	0.1403	0.0935
0.5%	0.6218	0.5228	0.3860	0.2926	0.2223	0.1934
1.0%	0.6536	0.5610	0.4315	0.3426	0.2778	0.2753
2.0%	0.6902	0.6060	0.4875	0.4070	0.3545	0.4048
5.0%	0.7490	0.6803	0.5853	0.5271	0.5111	0.7217
10.0%	0.8055	0.7539	0.6886	0.6633	0.7073	1.2064

The third level disclosure looks at the amount of capital required to support an equity investment over a fixed horizon, to meet a fixed cash flow. The required capital is measured in excess of the present value of the liability as calculated at the risk-free rate – in other words, $\exp(X_0)-1$ in our above notation. Using great leap probabilities, the required capital would be:

horizon	1	2	5	10	20	50
percentile	capital required					
99.9%	78%	121%	226%	373%	613%	970%
99.5%	61%	91%	159%	242%	350%	417%
99.0%	53%	78%	132%	192%	260%	263%
98.0%	45%	65%	105%	146%	182%	147%
95.0%	34%	47%	71%	90%	96%	39%
90.0%	24%	33%	45%	51%	41%	-17%

The negative capital required on the bottom right is not a misprint. It reflects the fact that, if you hold equities for long enough, the probability of underperforming bonds gets smaller and smaller.

Using cumulative probabilities, the capital requirements are higher:

horizon	1	2	5	10	20	50
percentile	capital required					
0.1%	86%	136%	263%	458%	827%	1755%
0.5%	69%	106%	194%	316%	518%	925%
1.0%	62%	93%	166%	262%	412%	681%
2.0%	54%	80%	139%	212%	319%	490%
5.0%	43%	63%	105%	152%	215%	300%
10.0%	35%	50%	80%	111%	149%	195%

We note that the cumulative ruin probabilities are higher than the great leap ruin probabilities. The differences are most noticeable at longer horizons.

Under the cumulative probability methodology, more capital is required the longer a firm is supposed to stay solvent. On the other hand, the great leap method produces capital requirements which first increase with horizon and then fall again.

It is possible offices might use different methods, not ruin probabilities, to set capital. Some may use conditional tail expectations. Others may use methods which trade the benefit of higher capital (security for policyholders and, so some degree, shareholders) against the cost of holding capital. Offices using dynamic programming methods may avoid the need to pick a percentile or a single time horizon. In these cases it is difficult to see how tables comparable to the above could be produced.

The statistics above, or figures like them could be used in two ways

(i) they could form a required disclosure by offices using stochastic models

(ii) they could be specified by a standards board as minimum standards, so that offices could only use models requiring at least the stated amount of capital for their capital assessment.

If the main purpose of this exercise is to achieve comparability then the same form of output should be required for each office. There is an outstanding question of whether an office who has chosen a 5 year horizon should be forced to disclose results on a 1-year basis merely to facilitate comparisons with other offices. Likewise, should an office with a chosen 1-year horizon be expected to comply with standards relating to percentiles of 5-year equity returns?

Fat Tailed Distributions

The models described so far are based on normal distributions. Observed distributions are clearly not normal. For example, the 2004 equity gilt studies tabulates annual log equity returns (capital index, in real turns) with (normalised) skewness of -1 and kurtosis of 5.5.

There are two popular techniques for capturing fat tailed distributions. One method is to use conditionally normal distributions with a stochastic volatility process. The Hardy regime-switching model, used in Canada, is an example oft his technique.

The second method is to make direct use of fat tailed distributions in the residuals, resulting in Lévy processes. These are also well discussed in the literature (see Carr, Geman, Madan & Yor).

Both of these tools have their advocates. The improvement in model fit has to be balanced against the need to calibrate additional parameters. There is little doubt that in the real world, both stochastic volatility and fat tailed residuals are the norm, but a joint calibration of both effects is widely believed to lie beyond what is achievable given the data.

[Tabulations of ruin probabilities, on a great leap and a cumulative basis, using these two models, to be inserted]