Help from Weather Forecasters
From Verification to Validation

Joseph Lo[1]

Abstract

Weather forecasting professionals have been developing and deploying probabilistic forecasts for well over 50 years. In face of the demand made on the general insurance actuaries to estimate, model and communicate uncertainty in all areas of their work, this paper argues that actuaries could gain insights by referencing techniques and debates developed by the weather forecasting profession. The paper demonstrates this with the Brier Score for verifying probabilistic forecasts on binary events, showing possibilities of how such a technique could be applied to validating stochastic models and methods.

Keywords Probabilistic forecasts; subjective probability; forecast validation; Brier Score; binary events

Forecast verification in weather forecasting helps to develop and improve forecasting models. Professionals working in this area developed a wide range of tools to help verify forecasts. A recent book that explores these concepts is (Jolliffe & Stephenson, 2012).

A particular case of weather forecasting is that of binary probabilistic forecasts. Instead of suggesting it will rain tomorrow, the forecast would communicate the uncertainty by saying something like "there will be a 80% chance of rain tomorrow". It is argued that such forecasts convey more truthfully the idea that forecasting is an uncertain exercise, although one needs to be careful with regards to communication and understanding of probabilities. In the U.K., weather forecasting organisations (e.g. the Met Office) are exploring probabilistic forecasts for public uses. The reader may want to reference a BBC report (BBC, 2007) and a Met Office article (Met Office) for more detail.

The present author suspects that ideas from this area could help actuaries in general insurance estimating, modelling and communicating uncertainty. Indeed, the paper (Lahiri & Wang, 2013) that originally interested him in this area was one written by economists, validating GPD decline predictions using these techniques. Our paper explores a case study using the Brier score (defined later) for binary forecasting and considers how this may help us validate exposure curves for pricing. It will also suggest two other general insurance examples where such scores could potentially be helpful.

A health warning is required. Through examples, the paper does not attempt to draw general conclusions on topics other than the demonstration of the testing techniques on models and assumptions. It should not be relied upon as a general market study of specific modelling techniques and assumptions themselves, nor should the reader rely on this paper as an indication as to how the author's own company performs modelling!

1 The Brier Score

The Brier score is an "actual vs expected" measure between a set of probabilistic forecasts of binary events and a corresponding set of future observed outcomes. It takes values between 0 and 1, with smaller values giving indications of better fit. Our key reference will not be the original paper by Brier (Brier, 1950), but (Lahiri & Wang, 2013) mentioned above.

Binary events for (Lahiri & Wang, 2013) were negative growths of GDP. The probabilistic forecasts came from surveys of forecasting professionals that started from 1968 Q4. The professionals were asked to give the probability of a GDP decline in the current quarter, next quarter, next two quarters, etc., and the survey

[1] Head of Actuarial Research & Development, Aspen. Contact email: jo.lo@aspen.co

then averaged the probabilities to give a probabilistic forecast for GDP declines. With around 140 data point pairs (forecasts and subsequent observed values), the Brier scores were used to explore effectiveness of the survey and possible areas of improvements.

To put it more concretely, following (Lahiri & Wang, 2013), the Brier score is defined as

$$\text{QPS} = \frac{1}{T}\sum_{t=1}^{T}(f_t - x_t)^2,$$

where

- QPS is acronym for Quadratic Probability Score (another name for the Brier Score)[2]
- f_t is the forecast probability, made at time *t*, of the event (e.g. GDP decline in the next quarter)
- x_t is the realisation of the event, taking on value 1 if it occurs, and 0 otherwise (e.g. continuing with the above example, if GDP declines in the next quarter after time *t*, then $x_t = 1$, otherwise $x_t = 0$)
- T is the total number of observations. The original intention was for it to denote time: the paper shall keep this notation even when the data point pairs are not made or observed serially in time series.

The Brier Score is proper: it cannot be "played" by the forecasters. The score performs worse if the forecaster does not forecast with what they believe to be the best forecast probability.

A downside of the Brier Score is its focus on binary events with relatively many observations. It would be useful for testing specific details of the forecasts: but for more comprehensive testing, one would use the Brier Score in conjunction with other measures.

2 Exposure Curve Validation

Our example is that of validating property exposure curves (sometimes known as "first-loss curves"). Such curves are popular in pricing property business and are well-known in the industry: we refer the reader to (Ludwig, 1991). The exact mechanics of how exposure curves should be used are not important in our current discussion. The idea to be tested here is that the use of these curves, with a few other assumptions such as loss ratios and limit profiles, would be able to tell us the frequencies at which different severities of losses would come through an excess-of-loss layer. To summarise: we have a mechanism (which depends on exposure curve selection) to provide probabilistic forecasts of whether a layer would be impacted; and our mission is to validate this mechanism against observed hits to layers.

We have used actual industry data and limit profiles, skewed so as to avoid identification and confidentiality issues. The example considers a selection of 220 layers (from different programmes), so that *T* = 220. From exposure curves and pricing assumptions, we have 220 probabilistic forecasts, f_t, of whether a layer would give rise to a loss. From hindsight, we also have 220 observations, x_t, as to whether there is a loss to the layer or not: it has value 1 if there was a loss to the layer *t*, and 0 if there was not. In total, there were 34 layers that had losses (i.e. $\sum_{t=1}^{220} x_t = 34$).

For popular exposure curves, we obtained Brier Score values of around 11% to 12%. It is possible to test whether a set of forecasts, $\{f_{1,t}\}$, has a significantly better (i.e. lower) score than another set, $\{f_{2,t}\}$. Let us define the following quantity for each layer:

$$d_t = 2\left(x_t - f_{1,t}\right)^2 - 2\left(x_t - f_{2,t}\right)^2$$

Then the test statistic $S = \frac{\bar{d}}{\sqrt{2\pi\sigma_d^2/T}}$ is distributed asymptotically with a standard normal (please also refer to (Diebold & Mariano, 1995) for more details).

For a pair of our examples, an exposure curve gives QPS 11.0% and another gives 11.8%. The mean of the $\{d_t\}$ is -1.5% and the SD is 8.5%. So the test statistic *S* is -1.0, giving an asymptotic two-tailed *p*-value of 30%, suggesting that the first exposure curve is not performing significantly better versus the second one.

[2] The original 1950 definition defines the Brier Score as a multiple-category measure. When we have a binary set up (i.e. two categories) some authors define the Brier Score to be twice the QPS.

3 Calibration

The above discussion considers the probability forecasts globally. We continue to follow (Lahiri & Wang, 2013) to consider how well probability forecasts of different levels perform. Intuitively, if a set of layers were predicted to have *f*% chance of a loss, then we should see around *f*% of these layers in the set being impacted by a loss. The unit interval of possible forecast probability values is partitioned into subintervals. A chi-squared test could then be devised to test for significance. The steps are as follows:

1. Partition the unit interval into *J* subintervals.
2. Let T_j be the number of forecasts that are forecast within the *j* th subinterval.
3. Note that $T = \sum_{j=1}^{J} T_j$.
4. Let $x_{j,t}$ be the observed (taking values 0 and 1) of the *t* th forecast in the *j* th subinterval.
5. Let $r_j = \sum_{t=1}^{T_j} x_{j,t}$ be the total number of actual events from those forecast within the *j* th subinterval.
6. Let m_j be a probability representative of subinterval *j*. In the first instance, this can be taken to be the mid-point of the subinterval.
7. Then $e_j = m_j T_j$ is the expected number of events from predictions in the *j* th subinterval.
8. Now calculate the weights of the observations in each subinterval: $w_j = T_j m_j (1 - m_j)$. This helps us to scale the results from each subinterval, taking into account the number of predictions and the variance of event occurrence.
9. For each subinterval, calculate the scaled actual vs expected statistic: $Z_j = \left(r_j - e_j\right) \Big/ \sqrt{w_j}$.
10. Summing the squareds of the Z_j's, we get a statistic that is asymptotically distributed with the chi-squared distribution with *J*-1 degrees of freedom.
11. This statistic allows us to evaluate the overall performance of the forecasts for all *j*.

The following table demonstrates the above steps with an example.

Prob. Intervals							
Min	0.0%	11.0%	19.8%	23.0%	52.5%		
Max	11.0%	19.8%	23.0%	52.5%	100.0%		
Mid	5.5%	15.4%	21.4%	37.8%	76.3%		
Range	11.0%	8.8%	3.2%	29.5%	47.5%		
							Total
Num Progs	153	22	11	25	9		220
Actual Freq	7	7	7	7	6		34
Relative Freq	5%	32%	64%	28%	67%		15.5%
Expected Freq	8.4	3.4	2.4	9.4	6.9		30.5
Weight	7.95	2.87	1.85	5.87	1.63		
							Test Statistic
Test Stat	- 0.50	2.13	3.42	- 1.01	- 0.68		17.94
							p-value
							0.1%

Table 3-1 Example demonstrating the chi-squared test

This example fails the chi-squared test. It also suggests that those layers predicted with probabilities between 11% and 23% tend to get claims more often than expected. In this example, the actuary would focus on improving the exposure curve in this range of probabilities.

4 Resolution

A binary probability forecast is useful if it is well-calibrated, in the sense of Section 3. However, there are strategies of forecasting where unskilled forecasts could actually give better calibrated results. For instance, the constant forecast of the long-run probability of event would be very well-calibrated indeed. Naturally, such a forecast would be useless for pricing, opening a firm to severe anti-selections.

Forecasting methods need to have good resolution. It should be able to differentiate between whether an event would happen or not. The more difference there are between the forecasts that would end up having an event and those that would end up not having an event, the higher the resolution.

The following chart helps to demonstrate this idea. The chart compares the empirical CDF of predictions conditional on whether there is an event. For example, it is saying that (with hindsight!), for those layers that eventually had no loss, 75% of the layers had forecast probabilities of 10% or lower (the blue line). For those layers that eventually had a loss, only 18% of the layers had such forecast probabilities (the red line). By observation, this has much higher resolution than the unskilled forecast, where the plots would both be the same step functions!

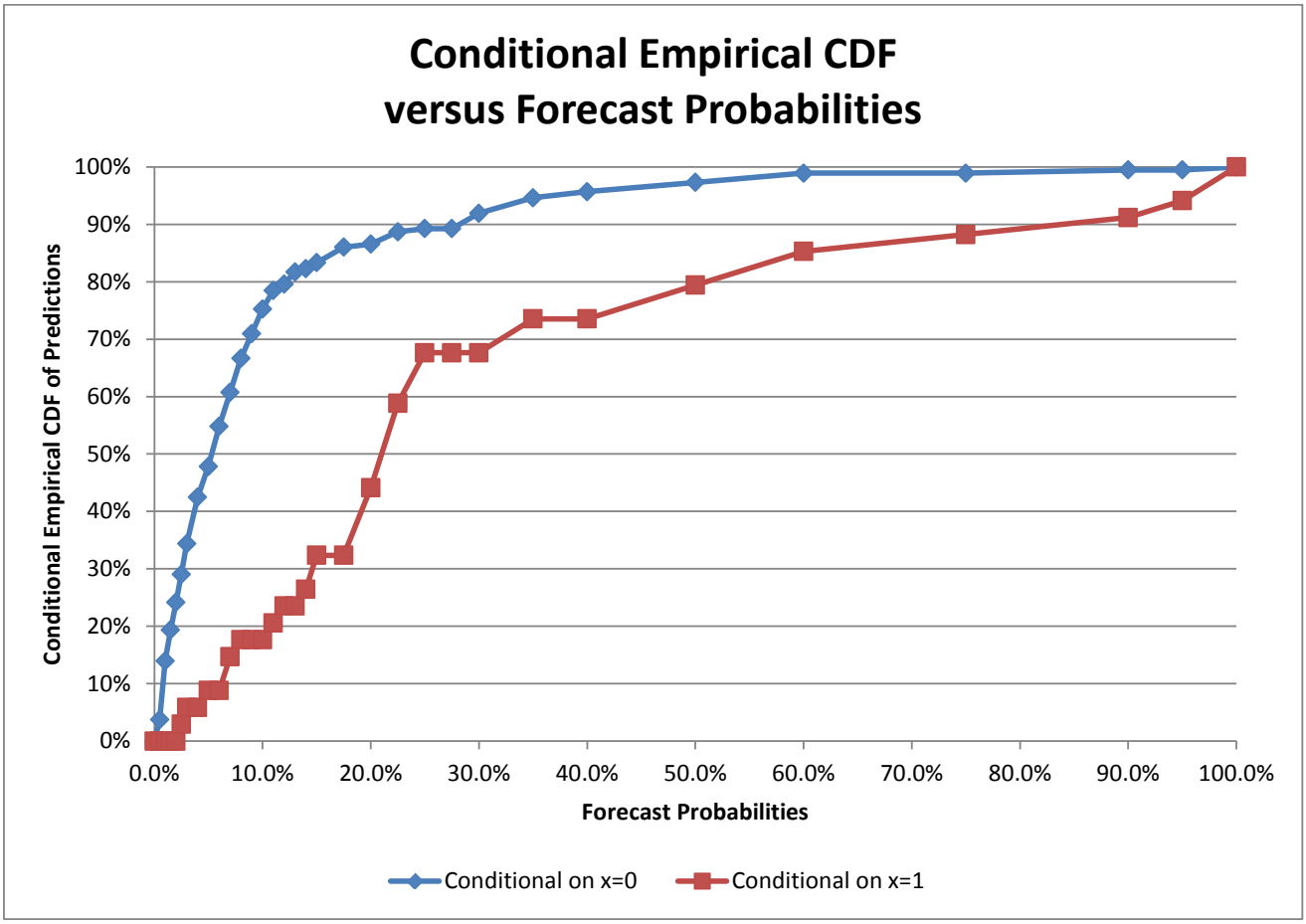

Figure 4-1 Conditional Empirical CDF versus Forecast Probabilities, demonstrating resolution

The Brier Score takes both calibration and resolution into account, as suggested by the so-called Murphy decomposition:

Briar Score = Uncertainty + Reliability – Resolution

Or in symbols:

$$\mathrm{QPS} = \bar{x}(1-\bar{x}) + \frac{1}{T}\sum_{j=1}^{J} T_j\left(\bar{x_j} - m_j\right)^2 - \frac{1}{T}\sum_{j=1}^{J} T_j\left(\bar{x_j} - \bar{x}\right)^2$$

The first term considers the uncertainty of the event to be forecasted. It gives us a baseline to measure the skill of the methodology, as it is akin to considering the (unskilled) constant forecasts with the long-run average. The second term increases the score: it is only zero if the methodology is perfectly calibrated. The third term measures resolutions.

While goodness in calibration is important, it is actually through the resolution term that the skills of a method are channelled. Continuing the exposure curve validation example, Table 4-1 shows the result of a Murphy decomposition. The calibration part increases the score by about 10%, while the resolution brings the score down (i.e. improvement) by around 25% of the uncertainty. In absolute terms, since the resolution term is larger than the calibration term, there is absolute positive skill to be observed with the method.

Murphy Decomposition	
	Absolute
QPS	11.0%
vs Const.	13.1%
Calibration	1.3%
Resolution	3.5%
QPS	10.9%
Check	*-0.1%*

Table 4-1 Example Murphy decomposition

We end the example with a brief discussion of the Yates decomposition, also mentioned in (Lahiri & Wang, 2013). An interesting point with the Yates is that it allows us to consider the excess volatilities. We have seen that a certain amount of volatility is required in the forecast probabilities to give good resolutions. However, too much volatility would actually blur the distinctions between cases with or without events. In our example, large amounts of excess volatility would suggest difficulties of the forecast method to distinguish between layers which would give a loss and layers which would not. As actuaries and underwriters improve their forecasting frameworks, the excess volatility metric points out more clearly what features a better framework would have.

5 Other Applications

Actuarial practice demands assessing uncertainty of estimates and validation of methods and models. The kind of forecast verification technique discussed above is likely useful for stochastic models and methods.

An example of model validation would be stochastic reserving models with fixed finite time horizons – e.g. probabilistic forecasts of claim development results such as (Merz & Wuethrich, 2008) and (Dahms, 2012). A large number of triangles would be passed through with one diagonal removed. Projections would be made, giving probabilistic forecast of, say, a 10% higher claim development result than a baseline model (e.g. chain ladder) predicts. Such a forecast could be tested by considering the actual development result with the held-out diagonal added back in.

An example of method validation would be where the actuaries make or facilitate expert judgments. Such judgments can be extensive in the actuary's work. The claims handlers are putting in their judgments in incurred claims reporting. Reserving for very large claims can involve various professionals such as lawyers. If such expert judgments could be given in terms of probabilistic forecasts (e.g. the claims handler stating there is x% chance that a specific claim would deteriorate by 100%), then these forecasts could also be subject to verification ideas discussed in this paper.

It is interesting that we are brought back to subjective probabilistic forecasts. Subjective probabilities – even if informed by modelling outputs – are now a fact of the actuary's life, especially with respect to new regulatory regimes such as Solvency II. When the U.S. weather forecasting profession was discussing probabilistic forecasts back in the 1960's and earlier (see, for example, (Sanders, 1963) and (Hughes, 1980) for a summary), there were doubts about probability forecasts – not dissimilar to some of our own doubts about being able to put, and communicate, probabilities to insurance events or reserve deteriorations. We could do well referencing and examining their work and debates in our times.

6 Works Cited

BBC. (2007, October 29). *Predictable forecast ... probably?* Retrieved February 26, 2013, from BBC News | Programmes | More or Less: http://news.bbc.co.uk/1/hi/programmes/more_or_less/7067003.stm

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 1-3.

Dahms, R. (2012). Linear stochastic reserving methods. *ASTIN Bulletin*, 1-34.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *J. Bus. Econ. Stat.*, 253-265.

Hughes, L. A. (1980). *Probability forecasting - reasons, procedures, problems.* NOAA.

Jolliffe, I. T., & Stephenson, D. T. (2012). *Forecast Verification: a practitioner's guide in atmospheric science.* Wiley.

Lahiri, K., & Wang, J. G. (2013). Evaluating probability forecasts for GDP declines using alternative methodologies. *International Journal of Forecasting*, 175-190.

Ludwig, S. J. (1991). An exposure rating approach to pricing property excess-of-loss reinsurance. *Proc. Casualty Actuarial Soc.*, 110-145.

Merz, M., & Wuethrich, M. V. (2008). Modelling the claims development result for solvency purposes. *ASTIN Colloquium.* Manchester.

Met Office. (n.d.). *Probabilistic weather forecasts and their interpretation*. Retrieved February 26, 2013, from Met Office: http://www.metoffice.gov.uk/weather/showcase/olympics_moguk_details.html

Sanders, F. (1963). On subjective probability forecasting. *J. App. Meteorology*, 191-201.