On Sequential Probability Forecasting

David A. Bessler
Texas A&M University

March 2010
Prequential Analysis

Let \( x_t' = (x_{1t}, \ldots, x_{mt}) \), \( t = 1, \ldots, T \) be observed values of the \( mx1 \) vector time series \( x_t \). At time \( t \), given known values \( x_t, t = 1, \ldots, T \), a set of probability distributions \( P_{Tk} = (P_{T+j}; j=1,\ldots, k) \) for unknown quantities \( x_{T+j}, j=1,\ldots, k \) are issued.

Phil Dawid *JRSS* 1984 and *Annals of Statistics* 1985 gives formal mathematics on prequential systems.

Prequential analysis refers to a system that predicts sequentially (pre-quential). Not unlike a human (system) making it through each day or an econometric model used for sequential decision-making.

\[ \text{Phil Dawid} \]

---

1. All photos in this presentation are in the public domain and are obtained from Google Image.
How to Judge Goodness?

A prequential forecasting system is judged as "good" or "bad" through the sequence of probabilities it actually issues and subsequent outcomes. It is not judged through a priori considerations, such as agreement with theory or goodness of fit with previous data. ¹

Prequential Analysis finds agreement in spirit with Friedman’s “instrumentalism” or Popper’s “falsification.” “Good” or “bad” are not determined by agreement with priors, but by studying probability forecasts and realizations of “new” data. ²

Actually “prequential analysis” might be viewed as an exercise in external validity (see Campbell and Stanley³).  

Milton Friedman

---

1. Actually the only a priori consideration would be coherence: “do the probabilities meet the “No Dutch Book” condition? Even the “No Dutch Book” might be questioned (not that we are so doing here), as it assumes rationality (and known utility) on the part of the probability assessor (deFinetti, *Theory of Probability* Wiley 1974).
Subjective Probability

Prequential analysis applies to probabilities emanating from either frequency-based probabilities or subjective probabilities (from one’s head).

The assessment of “goodness” of prequential probabilites is based on early work in subjective probability. All of which harkens back to Bruno deFinetti’s work, see *Theory of Probability*, Wiley, 1974 or “La Prevision” *Annales de l’Institut Henri Poincare*, 1937.
... The rational concept of probability, which is the only basis of probability calculus, applies only to problems in which the same event repeats itself again and again, of a great number of uniform elements are involved at the same time. [Probability, Statistics and Truth, MacMillan 1957. page 11]

... the collective “denotes a sequence of uniform events or processes which differ by certain observable attributes, ..., all the throws of a dice made in the course of a game form a collective wherein the attribute of the single event is the number of points thrown. [Probability, Statistics and Truth, MacMillan 1957. page 12]
**Von Mises Collective**

Consider a sequence of similar events. We put these together as a set of uniform events and look at the relative frequency of particular outcomes. Consider tossing a coin. The tosses in aggregate form a *von Mises Collective*. By increasing the number of flips we obtain a more precise estimate of the probability of a head, so as the number of tosses goes to infinity, the ratio of number of heads to number of tosses approaches the probability of a head.

<table>
<thead>
<tr>
<th>Number of Tosses</th>
<th>Number of Heads</th>
<th>Number of Tails</th>
<th>Ratio of Heads to Tosses</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>4</td>
<td>6</td>
<td>.4000</td>
</tr>
<tr>
<td>100</td>
<td>56</td>
<td>44</td>
<td>.5600</td>
</tr>
<tr>
<td>1000</td>
<td>518</td>
<td>482</td>
<td>.5180</td>
</tr>
<tr>
<td>10000</td>
<td>5036</td>
<td>4964</td>
<td>.5036</td>
</tr>
<tr>
<td>100000</td>
<td>50120</td>
<td>49880</td>
<td>.5012</td>
</tr>
</tbody>
</table>
Dawid Collective

While von Mises looked (or apparently looked) at physical characteristics on the sequence of events to form his collective (toss $i$ is no different from toss $j$ a priori), the Dawid Collective is such that events which have the same probability form a collective.

I call this a “Dawid collective,” not because he invented it, but because, apparently, he was the first to apply the Bayesian idea to the degrees of belief emanating from a statistical model. Clearly, Fischhoff, Lichtenstein and Phillips and other psychologists had this idea before Dawid; only they applied it to probabilities emanating from one’s mind – this may be a trivial distinction.

---

1. Hume may have arrived at a similar idea on characteristics as von Mises (Philosophical Essay Concerning Human Understanding), recall the quote of Hume by J.M Keynes, A Treatise on Probability page 217: “Nothing so like as eggs; yet no one, on account of this apparent similarity expects the same taste and relish in all of them. ’Tis only after a long course of uniform experiments in any kind, that we attain a firm reliance and security with regard to a particular event.”

One might consider the Dawid collective as the dual of the von Mises collective. In the former the collective is defined on the probabilities (e.g. the .7 collective) and one checks the data (outcomes) to see if the relative frequencies match the assigned probabilities; in the latter the data (events considered) define the collective and one derives the probabilities as relative frequencies.
Calibration and the Dawid Collective

In my mind:

- the probability that the Eiffel Tower is taller than Big Ben is 0.7;
- the probability that the Suez Canal is longer than the Panama Canal is 0.7;
- the probability that it rains tomorrow in Bryan/College Station, Texas is 0.7;
- the probability that the Aggies win their next football game is 0.7;
- the probability that Boston wins next year’s NBA Championship is 0.7.

All of these events form the “0.7 collective for me” (David Bessler), based on information I have today (right now).

If I wait until each of these event comes to pass (I travel to London and Paris and measure Big Ben and the Eiffel Tower, I wait until tomorrow and see if it rains, etc.) and if 70% of these “0.7 collective” events should actually be true (occur), I am well calibrated at .70.

The same holds for statistical models. These too should show well calibration on issued probabilities.
Calibration pre-dates Dawid’s Work

Early researchers on calibration of subjective probabilities is given in Lichtenstein, Fischhoff and Phillips, reprinted in Kahneman, Slovic and Tversky, *Judgement under Uncertainty: Heuristics and Biases*, Cambridge University press, 1982. This work documents well that people tend to be over confident in their probability assessments.

Weather Forecasting – back to the early 1950’s

Much of the early work on probability forecasting and assessment was done in the field of weather forecasting. These assessors have been shown to be quite good in terms of probability calibration.

Allan H. Murphy

Robert Winkler

Not G.W. Brier but St. Benedict in Brier Patch. Anyone have George’s picture?
Weather Forecasters versus Medical Diagnoses: No Contest

Here Scott Plous plots probability forecasts of weather people and medical diagnosticians on the horizontal axis and the (after the fact) relative frequency. To be well calibrated these should map into the 45° line. The weather forecasters appear to be better calibrated. Note, however, when the weather people forecast events as the sure thing (P=100) the, after the fact, relative frequency is only 90/100. This is evidence of over confidence. But this over confidence of the weather forecasters appears to be trivial relative to that of the medical diagnosticians.

Probability Integral Transform

• A graphical representation of calibration is merely the plot of the realized fractile versus the relative frequency. The fractile (also known as quantile) is the realization under the cumulative distribution function.

• Dawid relies on the application of a beautiful theorem from Mathematical Statistics, the *probability integral transform*.

• Theorem: If X is a random variable with a continuous cumulative distribution function $F_X(x)$, then $U = F_X(X)$ is uniformly distributed over the interval $(0,1)$.

• Realized fractiles will be uniformly distributed.
Apparently the First Use in Econometrics

John Kling has two *Journal of Business* pieces in the late 1980’s where prequential analysis is introduced to econometrics.

Here John and I study calibration and re-calibration on money supply (MS), interest rates (TBR), industrial production (IP) and prices (CPI). Bunn 1984 gets credit for introducing us to re-calibration.

Note, no evidence of over confidence, but read on.
Realization of zero fractiles

- We should never issue probabilities that have realized fractiles of 0 or 1. What a zero or one realized fractile implies is that our probabilities missed the actual realization.

- In fact, our models (and our brains) oftentimes give such realizations. The model that Kling and I built in the 1980’s (previous slide from Jo. Business 1989), apparently issued probabilities that gave zero fractiles on interest rates during the last few months of 2009 – not a good result! Fischhoff, Lichtenstein and Phillips, cited above, found similar “tightness” in probabilities issued by human subjects.

- Over our lifetime (1949 – 2009) we have never seen interest rates as low as we did in December 2009. So how do we get our models to issue probabilities on data that have no historical precedence? I’m not prepared to answer this in this set of notes – perhaps a formal Bayesian modeling approach that assigns a non-zero prior probability (no matter how small) to any deductively possible outcome will work here.
Here the bivariate model's probabilistic forecasts of cash prices (which includes knowledge of past future prices) shows better calibration than the univariate model's probabilistic forecasts of cash prices (the latter based on just past cash prices). Notice no evidence of over confidence.

* with Bessler
Calibration Plots on a Probabilistic version of Campbell’s Stock Market VAR

Here Campbell’s VAR (Economic Journal 1992) gives probability forecasts that are better calibrated than the random walk forecasts. See Ruffley (with Bessler) Applied Economics 2004.
Chi-Squared Test on the Null Hypothesis that Probability Forecasts are Well Calibrated.

<table>
<thead>
<tr>
<th></th>
<th>1 step ahead</th>
<th>2 steps ahead</th>
<th>3 steps ahead</th>
</tr>
</thead>
<tbody>
<tr>
<td>Campbell’s Model</td>
<td>11.00</td>
<td>15.71</td>
<td>13.38</td>
</tr>
<tr>
<td>Random Walk</td>
<td>6.42</td>
<td>42.49</td>
<td>92.16</td>
</tr>
</tbody>
</table>

The critical chi-squared statistic is 30.11.

We fail to reject well-calibration for Campbell’s model at all forecasting horizons. We reject well calibration for two of the random walk forecasts. So by the calibration criterion we say the Campbell-model’s forecasts “beat” the Random Walk forecast of stock market prices.

\[
\chi^2 = \sum_{j=1}^{J} \left[ \frac{(a_j - L_j L_n)^2}{L_j L_n} \right] \sim \chi^2 (J - 1).
\]

Here \( a_j \) is the actual number of observed fractiles in the interval \( j \) and \( L_j \) is the length of interval \( j \). Under weak conditions, not requiring independence for the distributions underlying the forecasts and under the null hypothesis of calibration, the test statistic will be distributed as chi-squared with \( J-1 \) degrees of freedom (Dawid (1984)).
Is Calibration Enough? No

• One can be a perfectly well-calibrated probability forecaster and be near unhelpful.

• A forecaster who gives the long-run relative frequency of rain as her daily forecast of rain in College Station, Texas will be well calibrated. But if I want to play golf tomorrow, her forecast doesn’t give me much help on whether to call in for a tee time.

• She will pale relative to a forecaster who gives probabilities of 1.0, on days that it turns out to rain, and 0.00, on days it turns out to not rain. This last forecaster is also well calibrated, but is able to sort – distinguish days of rain from days of no rain.
Average probabilities assigned to stock market events that ultimately obtain versus (minus) stock market events that ultimately do not obtain

<table>
<thead>
<tr>
<th>Horizon (steps ahead)</th>
<th>t+1</th>
<th>t+2</th>
<th>t+3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Campbell’s VAR</td>
<td>.0168</td>
<td>.0101</td>
<td>.0077</td>
</tr>
<tr>
<td>Random Walk</td>
<td>.0147</td>
<td>.0078</td>
<td>.0004</td>
</tr>
</tbody>
</table>

We want this difference in probabilities to be a large number. That is we want our models to distinguish between events that obtain (assign these a high probability) and events that do not obtain (assign these a low probability). Here Campbell’s VAR again beats a random walk, but does so with very small differences in assigned probabilities.

In the weather forecasting example given above, the person who gives the same long run relative frequency of rain every day as her forecast of rain in College Station has an average probability difference of 0.00. She does no sorting!

And the person who gives a forecast of 1.0 on days when it actually does rain and 0.0 on days when it does not rain has an average probability difference of 1.0. She does perfect sorting!
Trade-off Between Calibration and Resolution

• Elsewhere I have described the problem as a trade-off between calibration and resolution (*Int. Jo Forecasting* 1993 discussion).

• We want both Calibration and Resolution, but oftentimes can’t have both.

• Unfortunately our econometric students are sometimes encouraged to build models that give (perhaps) over-confident predictions; just as some cultures encourage over-confidence in life’s work in general.

• We want well-calibrated confidence. Not just confidence.
Quadratic Rule

The Mean Probability Score:

\[ \overline{PS}(p, d) = \left( \frac{1}{N} \right) \sum_{t=1}^{N} (p_t - d_t)^2 \]

is an average of the single-forecast version of the Probability Score over \( N \) occasions, indexed by \( t = 1, 2, \ldots, N \), where \( p \) is the issued probability forecast of event \( i \), \( d=0,1 \) indicator of the event occurred (1) or did not occur (0). We want to minimize \( PS \) for a good score.

This rule has a multiple event generalization studied in Winkler.

---

Quadratic Scoring Rule

• The quadratic scoring rule, following Brier, allows us to consider the trade-off between resolution and calibration.

• As users of probability forecasts we may be willing to give up some measure of calibration to increase resolution – reflecting movement along one of the curves in the above figure.

• Scoring rules have histories to both encourage honest probability assessment (elicitation) and evaluation of goodness in probability assessment (we use them here for evaluation of goodness).
Yates partitions the quadratic scoring rule (probability score PS) to reflect more deeply on how a set of probability forecasts perform.

\[
\overline{PS}(p,d) = B^2 + S + \sigma_{p,\text{min}}^2 + \sigma_d^2 - 2\sigma_{p,d}
\]

Here the probability score is decomposed into five components: bias, scatter, minimum variance, the variance of the random variable (d) (which we can’t control) and covariance between the random variable and our probability forecast (the latter is key).
Other Scoring Rules

Log Scoring Rule – apparently first introduced by I.J. Good (*JRSS* 1952). Studied by Shuford, et al *Psychometrika* 1966. They show the only proper scoring that has its payoff just in terms of the event which obtains is the log rule – seems to be a big deal for elicitation with untrained subjects (see Moore (with Bessler) *AER* 1979 and Nelson (with Bessler) *AJAE* 1989).

Macro Forecasting

- The Bank of England issues probability forecasts of Inflation and GDP. Gabriel Casillas* has studied these forecasts, using both Calibration and the Yates partition (Journal of Policy Modeling 2006).

- The next two slides give covariance graphs for the Bank’s forecasts (MPC) and those of a group of “other forecasters” (OF).

* With Bessler
Covariance Graphs (Inflation)

MPC

Other Forecasters
• Covariance Graphs (GDP Growth)
## Difference in Average Probabilities on the UK Macro Forecasts

<table>
<thead>
<tr>
<th></th>
<th>Inflation</th>
<th>GDP</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPC</td>
<td>.27 - .18 = .09</td>
<td>.26 - .15 = .11</td>
</tr>
<tr>
<td>Others</td>
<td>.33 - .17 = .16</td>
<td>.28 - .16 = .12</td>
</tr>
</tbody>
</table>

Here, both the inflation forecasts and the GDP forecasts from the set of “other forecasters” show better sorting ability than does the Monetary Policy Committee (MPC).

It is probably not surprising that both the “other forecasters” and the “MPC” show better sorting (larger differences) than we found in the model of stock market Prices (slide 19).

---

Average probability assigned to events that occur minus average probabilities assigned to events that do not occur. Again, we want this difference to be large, closer to one than to zero.
So What Have We Learned?

• Models can give well calibrated probability forecasts.
• However, they too can be over confident when used over new data points.
• The evidence we have worked on shows that models and people don’t show a lot of ability in sorting between events that occur and events that do not occur.
And What Does Prequential Analysis have to do with Agriculture?

- Almost all agricultural decisions are set in risky environments.
- We generally take our probabilities from historical data.
- Almost no one ever checks for well-calibration or sorting.
- Yet our minds and our models have generally shown themselves to be badly calibrated (over-confidence is seemingly ubiquitous).
- Applications in policy analysis, crop insurance, futures markets are all areas where checking for well-calibration and sorting may have non-trivial implications.
Some References


David

Glenda and David at Mt. Lemmon, Arizona July 2009