PREQUENTIAL DATA ANALYSIS

A.P. Dawid, University College London, Department
of Statistical Science, England

Abstract

The basic theory of the prequential approach to data analysis is described, and illustrated by means of both simulation experiments and applications to real data-sets.

Introduction

The prequential approach to the problems of theoretical statistics was introduced by Dawid (1984). It is based on the idea that statistical methods should be assessed by means of the validity of the predictions that flow from them, and that such assessments can usefully be extracted from a sequence of realized data-values, by forming, at each intermediate time-point, a forecast for the next value, based on an analysis of earlier values. The main emphasis is on probability forecasting, requiring that one describe current uncertainty about the predictand by means of a fully specified probability distribution. However, point forecasts, or other forms of prediction, can also be accommodated.

The purpose of the above paper was to indicate the fertility of the prequential point of view for furthering understanding of traditional concerns of theoretical statistics, such as consistency and efficiency. However, the prequential approach is essentially data-analytic. As such, it is particularly well suited to empirical investigation of the structure and properties of real-world observations, and their sources. In this paper, we shall discuss some of the ways in which prequential assessment may be applied in practical problems, including goodness-of-fit, model choice and density estimation. These methods are illustrated, by means of simulation experiments and applications to real data.

Prequential Assessment

Let $\underset{\sim}{Y} = (Y_1, Y_2, \ldots)$ be a potentially infinite sequence of observables, and $\underset{\sim}{Y}^{(k)} = (Y_1, Y_2, \ldots, Y_k)$. We consider methods of forming, for each $k = 1, 2, \ldots$, a *prediction*, $\hat{y}_k$, for Y_k, based on past data $\underset{\sim}{Y}^{(k-1)} = \underset{\sim}{y}^{(k-1)}$; or, more generally, of deciding on an *action* a_k on the basis of $\underset{\sim}{y}^{(k-1)}$, when subject to a loss $L_k(y, a)$ if $Y_k = y$ and $a_k = a$. Such a method M having been applied for $k = 1$ to n, and resulting in actions $(a_1, a_2, \ldots, a_n)$, its performance might be assessed by means of its total *prequential loss*

$$L_n^*(M) = \sum_{k=1}^{n} L_k(y_k, a_k),$$

which measures the success of its earlier forecasts; and comparison amongst methods on this basis provides a guide (albeit imperfect) as to their likely relative future performance.

Starting from a parametric family of such methods, $\mathcal{M} = \{M_\theta\colon \theta \in \mathcal{T}\}$, with M_θ specifying $a_k = a_k(\underset{\sim}{y}^{(k)}; \theta)$, each θ-value is thus assessed by

$$L_n^*(\theta) = \sum_{k=1}^{n} L_k\Big(y_k,\ a_k(\underset{\sim}{y}^{(k-1)}; \theta)\Big).$$

The *optimizing strategy* $\hat{\mathcal{M}}$ based on $\mathcal{M}$ then uses, for selection of a_{n+1}, $M_{\hat{\theta}_n}$, where $\hat{\theta}_n$ minimizes $L_n^*(\theta)$ ($n = 0,\ 1,\ldots$; modification for small n may be required). This itself needs to be assessed by its prequential loss

$$L_n^*(\hat{\mathcal{M}}) = \sum_{k=1}^{n} L_k\Big(y_k,\ a_k(\underset{\sim}{y}^{(k-1)}; \hat{\theta}_{k-1})\Big),$$

which will typically exceed $L_n^*(\hat{\theta}_n)$.

Prequential assessment of past predictive performance is very close in spirit to the method of *cross-validation* (Stone, 1974) but bases its prediction for Y_k on all *previous* outcomes, rather than on all outcomes *distinct* from Y_k. In both methods, the intention is to avoid the bias involved in letting Y_k contribute to its own prediction, and so to produce an honest assessment of uncertainty.

Probability Forecasting

One way to choosing the action a_k, after observing $\underset{\sim}{Y}^{(k-1)} = \underset{\sim}{y}^{(k-1)}$, is to specify a *predictive distribution* P_k for Y_k, and to choose a_k to minimize the *predictive expected loss*

$$\int L_k(y_k,\ a)\,dP_k(y_k).$$

Specification of such a sequence of predictive distributions (P_k), for any data, constitutes a *probability forecasting system* (PFS), and is equivalent to choosing a joint distribution P for the sequence $\underset{\sim}{Y}$. Under broad regularity conditions, it then follows that, with P-probability 1, $\underset{n\to\infty}{\underline{lim}}\ sup(L_n^*(M) - L_n^*(M')) < \infty$, where M is given by the above method, and M' is an arbitrary method. Thus if *Nature* is regarded as generating $\underset{\sim}{Y}$ from P, then using P as a PFS to construct an action sequence will be optimal, for any loss function.

A PFS P for $\underset{\sim}{Y}$, or its associated sequence (P_k) of predictive distributions of Y_k given $\underset{\sim}{Y}^{(k-1)} = \underset{\sim}{y}^{(k-1)}$, can be assessed directly if we take the *action* a_k to be the choice of a distribution Q_k for Y_k, and use a *proper scoring rule* $S_k(y,\ Q_k)$, i.e. such that, for any distribution P_k for Y_k, $E_{P_k}[S_k(Y_k,\ Q_k)]$ is minimized in Q_k when $Q_k = P_k$ (Dawid, 1986). Then the optimal sequence of *actions* is just the sequence (P_k). The assessment becomes particularly simple if we use the *logarithmic scoring rule* $S_k(y,\ P_k) = -log\, f_k(y)$, f_k being the density of P_k. We then obtain $L_n^*(P) = -log\, f(\underset{\sim}{y}^{(n)})$, f being the implied joint density for $\underset{\sim}{Y}^{(n)}$ under P. That is, we can, and henceforth shall, assess and compare PFS's by means of their *prequential log-likelihoods.*

It is interesting to note that, if the distributions P and Q for $\underset{\sim}{Y}$ are mutually absolutely continuous, then $L_n^*(P) - L_n^*(Q)$ will (with probability 1 under either P or Q) remain bounded, and may oscillate between positive and negative values. In this case we shall never achieve an ultimate preference for either PFS, and it seems that we remain forever in a quandary as to which to use for further forecasts. However, a result of Blackwell and Dubins (1962) shows that, in this case, the forecasts produced by P and Q will be asymptotically indistinguishable, so that the choice is unimportant. This is an instance of *Jeffreys's Law* (Dawid, 1984): observationally indistinguishable statistical approaches must be in essential agreement on their assertions about observables.

If $\mathcal{P} = \{P_\theta\colon \theta \in \mathcal{T}\}$ is a parametric family of PFS's, with predictive densities $f_i(y_i;\, \theta)$, the optimizing strategy $\hat{\mathcal{P}}$ based on $\mathcal{P}$ describes Y_{n+1} as having density $f_{n+1}(y_{n+1};\, \hat{\theta}_n)$; $\hat{\theta}_n$ being the maximum likelihood estimator based on data $\underset{\sim}{y}^{(n)}$. The success of this *plug-in MLE* strategy must itself, however, be judged by means of its own prequential log-likelihood, viz.

$$log \prod_{i=1}^{n} f_i(y_i;\, \hat{\theta}_{i-1}),$$

rather than

$$log \prod_{i=1}^{n} f_i(y_i;\, \hat{\theta}_n).$$

Similarly we can judge any other such *statistical forecasting system* (SFS), based on the same model or on another. A SFS might involve plugging-in some estimate of θ from past data, as above; Bayesian or fiducial elimination of θ; or any other suitable (standard or *ad hoc*) procedure. However, any such strategy will itself always be describable as a PFS, and hence as a joint distribution for $\underset{\sim}{Y}$. This allows standard probability theory to be applied in theoretical studies of the performance of a SFS for data generated from $P_\theta \in \mathcal{P}$, and opens up a fresh approach to the traditional problems of statistical theory (Dawid, 1984). In general, (efficient-estimate) plug-in and Bayesian SFS's are asymptotically optimal. The latter yield prequential likelihoods expressible in the form $\int f(\underset{\sim}{y}^{(n)};\, \theta)\pi(\theta)d\theta$, which has computational advantages, as well as being insensitive to reordering of the data.

Empirical Assessment

Sometimes an absolute assessment is required as to whether a PFS P adequately describes data $\underset{\sim}{y}$. If the Y_i are continuous real variables, and F_i denotes the distribution function of Y_i under P_i, then $\underset{\sim}{U} = (U_1, U_2, \ldots)$, where $U_i = F_i(Y_i)$, should be independently uniform on [0,1] if $\underset{\sim}{Y}$ arises from P, and so a variety of tests can be based on the observed values $\underset{\sim}{u}$. To assess uniformity, we might examine the u-plot, i.e. the empirical c.d.f. of the u's, which should be close to the line of unit slope. This could be tested formally using, say, the Kolmogorov-Smirnov statistic. One should also inspect the (u_i) for any sign of non-independence, trend, or dependence on omitted variables. A simple indicator of trend is provided by the *uniform conditional test* (Cox and Lewis, 1966) or y-plot, which forms the empirical c.d.f. of (y_j), where $y_j = \sum_{i=1}^{j} x_i \Big/ \sum_{i=1}^{n} x_i$, with $x_i = -log(1 - u_i)$. These y's are uniform order-statistics under P, and this can again be tested formally.

If the Y_i are 0–1 variables, we can form *calibration plots* in which, for various $\pi \in [0, 1]$, the observed relative frequency of $Y_i = 1$ over the set of occasions having $\Pi_i = \pi$ (where $\Pi_i = P_i(Y_i = 1)$) is plotted against π. This should give an approximate diagonal line. More formally, we can construct test-statistics such as $Z = \Sigma(Y_i - \Pi_i)/[\Sigma\Pi_i(1 - \Pi_i)]^{1/2}$, the sum possibly being restricted to a suitable subset of the data. Under very weak conditions, *not* requiring independence, Z and similar standardized statistics will be asymptotically standard normal under P (Seillier and Dawid, 1987) and independent of statistics based on disjoint subsets. An observed value z can thus be referred to standard normal tables, or a sum of squares of z's based on k disjoint subsets to chi-square tables with k degrees of freedom.

It is noteworthy that all the methods described above are applicable given only the two sequences, of outcomes and of their probability forecasts, and make no reference to the structure of P over outcomes not observed. This is in accord with the *Prequential Principle* (Dawid, 1984).

If P is itself constructed as a SFS based on a parametric model $\mathcal{P} = \{P_\theta\}$, it turns out, again under mild conditions, that the asymptotic distributions of the test-statistics considered above continue to hold under any $P_\theta \in \mathcal{P}$ (Seillier et al., 1988). Consequently, these methods can be used to test the overall goodness-of-fit of a parametric model.

If the distribution or model being used fails to describe the data, it may be possible to *massage* it to provide a better fit. Thus suppose that the (u_i) above look like a random sample, but from a non-uniform distribution. This distribution could itself be estimated, either parametrically or nonparametrically (as in *Density estimation* below). If the estimate based on $\underset{\sim}{u}^{(n)}$ is G_n, then Y_{n+1} could be forecast by requiring that $F_{n+1}(Y_{n+1})$ has distribution G_n, rather than uniform. Alternatively, serial correlation, or other suspected structure, in the (u_i) could be estimated and allowed for. In the $(0 - 1)$ case, if previous occasions on which the same probability forecast as p_{n+1} was issued had resulted in a proportion q of 1's, then p_{n+1} might be replaced by q. Such adaptive recalibration methods can improve the performance of a badly chosen initial model, although there can be no guarantee that they will, since the recalibration is based on the past but applied to the future.

Model Choice

Given a choice between two competing models, say $\mathcal{P} = \{P_\theta\}$ and $\mathcal{Q} = \{Q_\theta\}$, we can first replace each of these by an appropriate SFS, say $\hat{P}$ and $\hat{Q}$, respectively. We might then optimize the choice between these at each time-point. Thus if it were $\hat{P}$, say, rather than $\hat{Q}$, that gave the larger prequential likelihood (or smaller total prequential loss) to the data $\underset{\sim}{y}^{(k)}$ at time k, the probability forecast for Y_{k+1} would be that based on $\hat{P}$. Of course, such a two-stage optimization strategy needs assessing afresh in its own right. The method extends to more stages, and to an arbitrary collection of models at each stage, but clearly less trust can be placed in prequential analyses iterated to more stages: even though the prequential approach avoids obvious bias at each stage, no finite set of data can support more than a certain amount of investigation without throwing up misleading messages.

In place of repeated optimization, one can take a Bayesian approach, assigning prior weights α and $1 - \alpha$ to $\mathcal{P}$ and $\mathcal{Q}$. After observing $\underset{\sim}{y}^{(k)}$, with prequential joint density $f(\underset{\sim}{y}^{(k)})$ under $\hat{P}$ and $g(\underset{\sim}{y}^{(k)})$ under $\hat{Q}$, α is replaced by $\alpha_k = \alpha f(\underset{\sim}{y}^{(k)})/[\alpha f(\underset{\sim}{y}^{(k)}) + (1 - \alpha)g(\underset{\sim}{y}^{(k)})]$, and the forecast density for Y_{k+1} is then the mixture $\alpha_k f_{k+1} + (1 - \alpha_k)g_{k+1}$. The overall prequential likelihood for this strategy is simply $\alpha f(\underset{\sim}{y}^{(n)}) + (1 - \alpha)g(\underset{\sim}{y}^{(n)})$. Again the method extends simply to more models and more stages.

If one has a finite or countable collection of alternative models, and the data arise from some distribution in one of these, either of the above methods will be consistent and asymptotically optimal, in the sense that their forecasts will tend to those given by the true distribution, and at the fastest possible rate. However, for finite data-sets, the forecasts under the two methods may look rather different. In either case, if the true distribution is contained in a model of high-dimensionality, early analysis will generally tend to favor incorrect models of low dimensionality. This is intuitively sensible, since, early on, the mis-modelling bias may well be less of a problem than the imprecision involved in trying to estimate many parameters.

As an alternative to allowing such transient behavior to be entirely data-driven, as above, one might build it in directly, by setting out with a strategy for choosing, at each stage, the complexity of the model to be fitted and how it is to be used for prediction. Different strategies, all yielding consistent estimates of the true model (and which use each fixed model efficiently) will all be *asymptotically* equally good. However, their *transient* behaviors, which may be long-lasting, can be very different, with some yielding much larger prequential log-likelihoods (or, more generally, much smaller prequential losses) than others even though these discrepancies will be bounded as the sample size goes to infinity. More empirical and theoretical work is needed to indicate good forms for such strategies. A sensible super-strategy could be built up from a low-dimensional parametrized family of such strategies, using optimizing or Bayesian methods. This could combine good transient behavior with sensitivity to the data and avoidance of data-mining.

Non-parametric Approximation

Many non-parametric problems, such as density estimation or fitting a stationary time-series, can be approached through a sequence of finitely parametrized methods, such as fitting histogram or kernel density estimates with adjustable bin width, or autoregressive models of various finite orders. One can then apply the techniques of the previous section, even though none of the models used is now expected to contain the distribution generating the data. The component models will generally each be characterized by some quantity, such as kernel width (w) or autoregressive order (p), which controls the balance between over-fitting (tracking noise in the data) and over-smoothing (not picking up the signal). Prequential choice of such a quantity will start out with a preference for smoothing (large w, small p), and then, as the data-sequence grows longer and can support more detailed modelling, gradually move towards fitting the past data more and more closely ($w \to 0$, $p \to \infty$). Such a method will often be prequentially consistent for a wide range of generating distributions, and can provide sensible answers based on finite data-sets, by making the predictively optimal compromise between fitting and smoothing.

Investigation of the structure of good *strategies*, for choosing the model to fit at each stage, is still more vital in this context, since the behavior described as *transient* in the previous section now extends to infinity! Again, much further empirical and theoretical work is required to illuminate this problem area.

Simulations

1. Time-series modelling.

Autoregressive models of varying order k ($0 \leq k \leq 8$) were fitted to several simulated time-series of 500 observations, and their prequential likelihoods calculated using both optimization (plugging-in current least-squares estimates) and Bayesian methods (using a *non-informative* prior), always excluding the first 15 observations. Results were as follows.

(i) Independent standard normal variates: $Y_t = \epsilon_t$; Prequential Log-Likelihoods

k	:	0	1	2	3	4	5	6	7	8
Optimization	:	-715.5	-717.1	-721.0	-724.0	-728.0	-728.7	-735.1	-739.1	-740.5
Bayes	:	-712.8	-713.9	-717.6	-719.4	-722.0	-722.4	-726.0	-728.4	-730.1

The strategy of optimizing over k chose $k = 0$ at all points, except one, beyond the 57th observation, and chose $k = 1$ at all the exceptional points. This strategy itself had a prequential log-likelihood of -714, better than that for any fixed k.

The Bayes strategy (using equal prior probabilities) finished by assigning probability 0.75 to $k = 0$ and 0.25 to $k = 1$. Its prequential log-likelihood too was -714.

(ii) Autoregression: $Y_t = 0.1Y_{t-1} - 0.3Y_{t-2} + 0.2Y_{t-3} + \epsilon_t$; Prequential Log-Likelihoods

k :	0	1	2	3	4	5	6	7	8
Optimization :	-723.9	-725.3	-705.5	-700.1	-701.2	-701.3	-704.4	-708.6	-709.3
Bayes :	-722.8	-724.1	-703.5	-698.3	-699.1	-700.6	-702.7	-705.9	-707.6

There is a clear preference for the true order, with under-fitting being more heaving penalized than overfitting. Optimizing over k chose $k = 2$ up to observation 40, $k = 3$ thereafter. This strategy had a prequential log-likelihood of -700, indistinguishable from that of $k = 3$. The Bayes strategy ended by assigning probability 0.63 to $k = 3$, 0.29 to $k = 4$ and 0.07 to $k = 5$, and itself had a prequential log-likelihood of -700.

(iii) Moving average: $Y_t = 0.5\epsilon_t - 0.2\epsilon_{t-1}$; Prequential Log-Likelihoods

k :	0	1	2	3	4	5	6	7	8
Optimization :	-370.1	-354.8	-355.3	-357.7	-355.5	-356.9	-358.7	-360.3	-364.5
Bayes :	-368.9	-353.4	-353.2	-355.7	-353.2	-355.3	-357.1	-359.1	-362.2

The true process can be expressed as an infinite-order autoregression: $Y_t = -0.4Y_{t-1} - 0.16Y_{t-2} - 0.064Y_{t-3} - \dots + 0.5\epsilon_t$. The optimal autoregressive fit to 500 observations, however, gave $k = 1$ (optimization) or $k = 2$ (Bayes), closely followed by $k = 4$ (for which the estimated coefficient of lag 4 was -0.139, compared with the true value of -0.026). Optimizing over k gave $k = 1$ at all points, except for observations 16 to 33 (for which k was 0) and most points between observations 460 and 486 (with $k = 4$). This strategy had prequential log-likelihood of -355.5. The Bayes strategy assigned probabilities 0.27 to $k = 1$, 0.33 to $k = 2$, 0.03 to $k = 3$, 0.32 to $k = 4$, and 0.04 to $k = 5$, and itself had a prequential log-likelihood of –354.3.

2. Density estimation.

Simple histogram-type density estimators were constructed from data-values in [0,1], based on a division of the unit interval into k equal sub-intervals. For each initial sub-sequence of data, the current density estimate was used to forecast the next observation. This was repeated for $1 \leq k \leq K$.

(i) A random sample of size 1000 from the uniform distribution on [0,1] yielded the following overall prequential log-likelihoods (up to $K = 10$);

k	:	1	2	3	4	5	6	7	8	9	10
log-likelihood	:	0	-3.7	-6.4	-8.8	-12.5	-16.0	-18.8	-22.3	-25.1	-32.4

The deterioration in performance when fitting more intervals than needed (viz. 1) is clear.

The optimizing strategy, formed by selecting, at each point, that value for k yielding the highest prequential likelihood to date, always chose $k = 1$, except at a number of points up to the 52nd observation, for which $k = 2$ was chosen.

(ii) A random sample of size 3000 was generated from the symmetric unimodel density

$$f_1(x) = \frac{1}{2}\pi sin(\pi x) \qquad (0 \leq x \leq 1).$$

With $K = 20$, the optimal k based on all the data was 14, the prequential log-likelihoods for $k = 10$ to 15 being, respectively, 399.3, 389.2, 401.3, 396.9, 402.2 and 399.0. When optimizing over k at all points, the first and last appearances of various values, and their frequencies, were:

k	:	1	2	3	4	5	6	7
First used	:	1	6	21	77	229	148	319
Last used	:	20	43	154	138	388	395	842
Frequency	:	18	6	109	16	46	182	39

k	:	8	9	10	11	12	13	14	>14
First used	:	1423	400	1457	-	1435	-	2856	-
Last used	:	1423	2133	2335	-	2915	-	3000	-
Frequency	:	1	1118	434	0	892	0	139	0

The general message of the above simulations would seem to be that, even for large data sets, it is generally far more effective to fit a very simple model that is approximately true, rather than one which contains the true distribution (or comes close to doing so), but is of highish dimension.

Applications

1. Weather forecasting.

Jain (1983) analyzed a 53-year sequence of daily precipitation records from Morogoro, Tanzania, as discussed in Stern and Coe (1984). The model $\mathcal{P}$

for the conditional probability p_t of rain on day t (coded as $Y_t = 1$), given past outcomes, was a non-stationary two-state second-order generalized linear Markov Chain:

$$logit\ p_t(\theta) = a_{ij0} + \sum_{k=1}^{4}\left[a_{ijk}\ sin(kt') + b_{ijk}\ cos(kt')\right]$$

where $t' = (2\pi t/366)$, i and j are the outcomes of days $t - 2$ and $t - 1$, and θ consists of the a's and b's. The parameters were estimated recursively, with initial estimates fitted, using maximum likelihood, to the first 700 data-points, and the probability forecasts $\hat{p}_t$ of the resulting *plug-in* strategy compared with the actual outcomes ($t > 700$). Calibration plots and test-statistics were constructed for various subsets of the data, corresponding to the months of the year, and to specified outcomes of the three previous days. Table I gives, for each month, the overall proportion $\bar{y}$ of rainy days, and the average forecast probability $\bar{\hat{p}}$. The final line gives values of the test statistic

$$z_B = \frac{\Sigma(y_i - \hat{p}_i)^2 - \Sigma\hat{p}_i(1 - \hat{p}_i)}{\left[\Sigma\hat{p}_i(1 - \hat{p}_i)(1 - 2\hat{p}_i)^2\right]^{\frac{1}{2}}}$$

for assessing departure from expectation of the within-month *Brier Score* $\Sigma(y_i - \hat{p}_i)^2$. These should be approximately independent standard normal variables under the model $\mathcal{P}$. The combined chi-square of 78 on 12 degrees of freedom clearly indicates poor model fit, and closer scrutiny reveals that the model is noticeably under-forecasting rain in April, and when the third previous day was wet.

TABLE I

Month	:	J	F	M	A	M	J	J	A	S	O	N	D
$\bar{y}$	:	.21	.22	.33	.54	.31	.10	.06	.04	.09	.10	.17	.22
$\bar{\hat{p}}$	:	.20	.20	.36	.47	.31	.09	.05	.04	.09	.09	.16	.20
z_B	:	1.57	4.68	2.15	5.00	1.80	2.19	2.26	0.40	0.42	0.66	1.26	2.93

2. Medical diagnosis.

Seillier (1982) analyzed 58 cases of jaundice, caused either by hepatitis ($Y = 1$) or by cirrhosis ($Y = 0$). Various logistic models to discriminate between the two diagnoses were considered, using regressor variables chosen from a set of ten symptoms (A, B, C, D, E, F, $X1$, $X2$, $X3$, $X4$) and a location indicator Q.

Each model was fitted by maximum likelihood to the first k cases (k = 30, 31,...,57), and used to provide a probability forecast $\hat{p}_{k+1}$ for Y_{k+1}, based on its associated regressor variables. The assessment of each model was then based on its overall Brier score $\sum_k (y_k - \hat{p}_k)^2$. The results are shown in Table II, which also gives $\bar{\hat{p}}$, for comparison with $\bar{y} = 0.29$.

TABLE II

Variables	Brier Score	$\bar{\hat{p}}$
A + B + C + D + E + F + X1 + X2 + X3 + X4 + Q	4.7	0.25
A + B + C + D + E + F + X1 + X2 + X3 + X4	3.8	0.36
A + C + D + E + X1 + X2 + X3 + X4 + Q	4.6	0.24
A + C + D + E + X1 + X2 + X3 + X4	3.8	0.36
A + D + E + X1 + X2 + X3 + X4 + Q	4.0	0.22
A + D + E + X1 + X2 + X3 + X4	3.4	0.39
A + D + X1 + X2 + X3 + X4 + Q	4.6	0.32
A + D + X1 + X2 + X3 + X4	3.6	0.38
A + D + X1 + X3 + X4 + Q	3.0	0.23
A + D + X1 + X3 + X4	2.3	0.31
A + D + X1 + X3 + Q	3.4	0.24
A + D + X1 + X3	3.0	0.33
D + X1 + X3 + Q	3.0	0.25
D + X1 + X3	2.9	0.39
D + X1 + Q	4.8	0.24
D + X1	4.3	0.37
D + Q	5.3	0.22
D	5.1	0.36

Fitting all variables leads to poor predictions on this size data-set, as does fitting only two or three. The most successful model, as measured by its Brier score, is $A + D + X1 + X3 + X4$, which also has $\bar{\hat{p}}$ closest to $\bar{y}$. It is of interest that, for any collection of symptom variables, adding in the location indicator Q leads to *worse* predictions. This offers some empirical support for the arguments of Dawid (1976) that suitable diagnostic models should be robust over a range of locations.

3. Educational scaling.

Opie (1983) conducted an analysis to see whether items in an educational testing item-bank fitted the Rasch model, under which P(student i gets item j correct) $= e^{\alpha_i+\beta_j}/(1 + e^{\alpha_i+\beta_j})$. The data-set contained responses to 60 test items from 150 students. At an intermediate stage, a number of items, 1 to $k - 1$ say, have been accepted, and item k is under test. For $m = 75$ to 150, the parameters are estimated (by maximum likelihood) from the responses of students 1 to m on items 1 to k, omitting that of student m on item k. The fitted probability for this omitted response can then be calculated, and the process repeated with m increased by 1. Comparison of these forecast probabilities with the actual responses (where these were not missing) then allows assessment of the fit of item k to the model.

For testing item 60, with all other items included, the probabilities were grouped into 8 intervals, with counts, average probability and relative frequency of a right answer as given in Table III.

TABLE III

Group (g)	Count (n_g)	Average probability (π_g)	Relative frequency ($\bar{y}_g$)
0.0 - 0.1	14	0.07	0.07
0.1 - 0.15	16	0.12	0
0.15 - 0.2	11	0.17	0.09
0.2 - 0.3	12	0.25	0.25
0.3 - 0.4	8	0.33	0
0.4 - 0.5	6	0.44	0
0.5 - 0.6	4	0.55	0.5
0.6 - 1.0	4	0.86	0.5

If the item fits the model, then $\sum_g n_g(\bar{y}_g-\pi_g)^2 \div \pi_g(1-\pi_g)$ should be approximately distributed as chi-square with 8 degrees of freedom. The observed value of 15.8 is significant at 5%, suggesting a failure of calibration on this item, and thus its non-conformity with the Rasch model.

4. Software reliability.

Littlewood et al. (1986) have made a thorough comparison of a number of model-based prediction systems for prequential probability forecasting of the successive inter-failure times of complex software systems. The data comprised 136 inter-failure times ranging between 0 and 6150 seconds, and the models used all incorporated reliability growth (improved performance after each bug-fix). Some forecasting systems used optimization, some were Bayesian, others combined the two methods. The results are summarized in Table IV.

TABLE IV

System	u-plot K-S distance (sig. level)	y-plot K-S distance (sig. level)
1. JM	.190 (1%)	.120 (NS)
2. BJM	.170 (1%)	.116 (NS)
3. GO	.153 (2%)	.125 (10%)
4. L	.109 (NS)	.069 (NS)
5. BL	.119 (NS)	.075 (NS)
6. LNHPP	.081 (NS)	.064 (NS)
7. LV	.144 (5%)	.110 (NS)
8. KL	.138 (5%)	.109 (NS)
9. W	.075 (NS)	.075 (NS)
10. D	.159 (2%)	.093 (NS)

Systems 1, 2 and 3 are all based on essentially the same model, as are 4, 5 and 6. It appears that the method of data analysis is less important here than choosing a good model. Measured by prequential likelihood, the optimal system was 6. The authors also considered adaptive recalibration of the above systems, as well as Bayesian and optimizing strategies for combining them, leading in all cases to improvements in performance.

Conclusion

The prequential method is broad in range, simple in concept, and based on a firm theoretical foundation. However its implementation leaves plenty of scope for variations, and is currently more art than science. Further work should lead to an improved understanding, and give guidance on good strategies of applying the method. Efficient computational methods or approximations will also be essential for routine application.

References

Blackwell, D. and Dubins, L. E. (1962): Merging of opinions with increasing information, *Ann. Math. Statist.* 33, 882-886.

Cox, D. R. and Lewis, P. A. W. (1966): *Statistical Analysis of Series of Events*, Methuen, London.

Dawid, A. P. (1976): Properties of diagnostic data distributions, *Biometrics* 32, 647-658.

Dawid, A. P. (1984): Statistical theory: the prequential approach (with discussion), *J. Roy. Statist. Soc.* A 147, 278-292.

Dawid, A. P. (1986): Probability forecasting, *Encyclopedia of Statistical Sciences*, eds. S. Kotz, N. L. Johnson and C. B. Reid, Wiley-Interscience, Vol. 7, 210-218.

Jain, R. (1983): *Probabilistic Weather Forecasting*, M.Sc. dissertation, Department of Statistical Science, University College London.

Littlewood, B., Abdel Ghaly, A. A., and Chan, P. Y. (1986): Evaluation of competing software reliability predictions, *IEEE Trans. Software Eng.* SE-12, 950-967.

Opie, G. A. (1983): *Educational Scaling*, B.Sc. dissertation, Department of Statistical Science, University College London.

Seillier, F. (1982): *Selection Aspects in Medical Diagnosis*, M.Sc. dissertation, Department of Statistical Science, University College London.

Seillier, F. and Dawid, A. P. (1987): On testing the validity of probability forecasts, *Research Report 57*, Department of Statistical Science, University College London.

Seillier, F., Sweeting, T. J., and Dawid, A. P. (1988): Prequential tests of model fit, *Research Report 61*, Department of Statistical Science, University College London.

Stern, R. D. and Coe, R. (1984): A model-fitting analysis of daily rainfall data (with discussion), *J. Roy. Statist. Soc.* A 147, 1-34.

Stone, M. (1974): Cross-validatory choice and assessment of statistical predictions (with discussion). *J. Roy. Statist. Soc.* B 36, 111-147.