A general method of incorporating forecast cost and loss in value scores

William Briggs

General Internal Medicine, Weill Cornell Medical College
525 E. 68th, Box 46, New York, NY 10021
email: wib2004@med.cornell.edu

April 28, 2005

Submitted to Monthly Weather Review

Abstract: We show how to incorporate cost for correct forecasts in the skill score statistical test developed by Briggs and Ruppert (2005) and how to extend this result to another value score developed in Wilks (2001). We then show that both of these methods are special cases of a more general complete value score, where the loss and costs are allowed to be completely general.

Key words: Skill testing; Skill score; Forecast value; Value Score; Expected loss.

1. Introduction

It is important to have formal statistical tests of both forecast skill and value. This allows us to gauge exactly when reported good forecast performance with respect to some metric is due to chance or it is statistically significant. Many performance (skill and value) metrics exist, though few are accompanied with formal statistical tests. This is usually because the sampling distribution of the metric is unknown, making testing difficult or requiring the use of non-parametric methods such as bootstrapping (see Hamill 1998 for an example of such a test).

Here, we show how to modify the skill/value statistical test method of Briggs and Ruppert (2005) and Briggs et al. (2005) to include cost for correct forecasts. Their original metric, K_θ, only included losses for incorrect forecasts with no loss for correct forecasts: the parameter θ is a measure of this loss. Adding cost is usually done for the cost-loss problem. We show how a cost-loss score, the Value Score (VS) given by Wilks (2001; a similar score is in Richardson, 2000), fits into the Briggs-Ruppert statistical testing framework and can have formal statistical tests of value made with it. We then show that both K_θ and VS are special cases of what we call the Complete Value Score, or CVS. This score can include general costs of any kind. One general cost might be a fee that a user must pay to use a particular forecast.

Lastly, we present a hypothetical forecast example and how each score can be interpreted for this example.

2. Cost for correct forecasts

a. **Skill test.** We first develop some notation. Let a dichotomous forecast be $X \in \{0, 1\}$, and its corresponding dichotomous observation be $Y \in \{0, 1\}$. Also let the loss (and costs) for forecasts be denoted as k_{YX}. Details on all notation can be found

in Briggs and Ruppert (2004) and Briggs and Ruppert (2005, hereafter BR). BR developed a skill score and statistical test for significance for situations in which k_{01}, $k_{10} > 0$, but $k_{11} = k_{00} = 0$, that is, a loss for incorrect forecasts but no loss for correct ones, as is usual in statistical models of "goodness". Cost-Loss skill scores, such as that developed in Wilks (2001) and Richardson (2000), require this last condition be modified to $k_{11} = k_{01} > 0$, but with k_{00} still equal to 0. It is desirable to have the formal statistical test developed in BR apply to these situations and to situations where even $k_{00} > 0$.

It is possible to add loss (or cost) for making correct forecasts to the testing framework given in BR. Let $k_{11} \geq 0$ and $k_{00} \geq 0$ be the losses (or costs) for making correct forecasts. In BR, $\theta = k_{01}/(k_{10} + k_{01})$, which is the relative loss for saying $X = 1$ and having $Y = 0$. We also have that $1 - \theta = k_{10}/(k_{10} + k_{01})$ for saying $X = 1$ and having $Y = 0$. Cost-loss problems usually have $k_{00} = 0$, though this need not be the case: $k_{00} = c \geq 0$ can be thought of the cost or fee one might pay for having access to an expert forecast, while k_{11} might be the cost paid to protect against $Y = 1$ (plus a fee: $k_{11} = k_{01} + c$ where c is the forecast fee, if any). We only have the minimal requirement that $k_{00} < k_{01}$ and $k_{11} < k_{10}$, which says that any fee (or protection) cost must be lower than the loss suffered for using an incorrect forecast. This makes sense because nobody would pay more to use the forecast than it would cost to protect against the maximum possible loss. An example will be given below.

Let $p = P(Y = 1)$, $p_{Y|X} = P(Y|X)$, and $p_{+1} = P(X = 1)$. We define the optimal naive forecast, X^N (where the superscript N denotes the optimally naive forecast), as the forecast with the least expected loss that can be given knowing only p. We assume, for the purposes of development, that observations are rare with respect to the loss, that is, that the optimal naive climate forecast is always 0 (if this is not

the case, the observations and forecasts can always be transformed by $\tilde{Y} = 1 - Y$, $\tilde{X} = 1 - X$; the losses k_{YX} are also suitably transformed; see BR). We also define X^E as the expert forecast, represented by the superscript E, which is any forecast but the optimal naive forecast.

Using the definition of value that a collection of expert forecasts has less expected loss than a collection of optimal naive forecasts leads to the statistical null hypothesis of no value originally given in BR as $H_0 : \ E(k^E) \geq E(k^N)$. We can write these expected losses

$$\begin{aligned}
E(k^N) &= P(Y=1, X=0)k_{10} + P(Y=0, X=0)k_{00} \\
&= P(Y=1|X=0)P(X=0)k_{10} + P(Y=0|X=0)P(X=0)k_{00} \\
&= (p_{1|1}p_{+1} + (1-p_{0|0})(1-p_{+1}))k_{10} + ((1-p_{1|1})p_{+1} + p_{0|0}(1-p_{+1}))k_{00} \\
E(k^E) &= \sum_{i=0}^{1}\sum_{j=0}^{1} P(Y=i, X=j)k_{ij} \\
&= p_{1|1}p_{+1}k_{11} + (1-p_{1|1})p_{+1}k_{01} + (1-p_{0|0})(1-p_{+1})k_{10} + p_{0|0}(1-p_{+1})k_{00}
\end{aligned}$$

For any losses k_{YX} (subject to the above-mentioned restrictions), it is then easy to show that this null hypothesis from BR in the general case becomes

$$H_0 : \ p_{1|1} \leq \frac{k_{01} - k_{00}}{k_{01} - k_{00} + k_{10} - k_{11}} = \theta'. \tag{1}$$

The restrictions that $k_{00} < k_{01}$ and $k_{11} < k_{10}$ follows immediately from this equation. In BR, $\theta' = \theta = k_{01}/(k_{10} + k_{01})$.

Calculation of the likelihood ratio test statistic and other quantities, such as p-values, are as stated in BR as the value of θ' is fixed (nonrandom) in these calculations; that is, no statistical result in BR need be changed and only the value of θ' need be (calculated by the relevant decision maker and) replaced in the equations. We briefly show the development of the test statistic.

Let n_{YX} be the observed counts of the forecast and observations. BR showed that under the null hypothesis of $p_{1|1} \leq \theta$, the MLE of $p_{1|1}$ is $\widetilde{p}_{1|1} = \min\{\frac{n_{11}}{n_{11}+n_{01}}, \theta'\}$. We use a likelihood ratio test, which is formed by taking the ratio of the likelihood maximized under the null and the likelihood maximized in general. It turns out that the only parameter that is different in these two cases is $p_{1|1}$; the other parameters then drop out of the calculation. Calculation of the likelihood ratio statistic (G) is then given by:

$$G = -2\log\left[\left(\frac{\widetilde{p}_{1|1}}{\widehat{p}_{1|1}}\right)^{n_{11}}\left(\frac{1-\widetilde{p}_{1|1}}{1-\widehat{p}_{1|1}}\right)^{n_{01}}\right] = 2n_{11}\log\left[\frac{\widehat{p}_{1|1}}{\widetilde{p}_{1|1}}\right] + 2n_{01}\log\left[\frac{1-\widehat{p}_{1|1}}{1-\widetilde{p}_{1|1}}\right].$$

where $\widehat{p}_{1|1} = n_{11}/(n_{11}+n_{01})$. When $\frac{n_{11}}{n_{11}+n_{01}} > \theta'$ then $\widetilde{p}_{1|1} = \theta'$ and $G > 0$, and when $\frac{n_{11}}{n_{11}+n_{01}} \leq \theta'$ then $\widetilde{p}_{1|1} = \widehat{p}_{1|1}$ and $G = 0$. This allows us to rewrite G as

$$G = \left(2n_{11}\log\left[\frac{n_{11}}{n_{+1}\theta'}\right] + 2n_{01}\log\left[\frac{n_{01}}{n_{+1}(1-\theta')}\right]\right) I\left(\frac{n_{11}}{n_{+1}} > \theta'\right) \tag{2}$$

where $n_{+1} = n_{11} + n_{01}$ and $0\log(0) = 0$.

G has asymptotic distribution $1/2\chi_0^2 + 1/2\chi_1^2$ (Self and Liang, 1987; their case 5: this is a mixture distribution of two equally-weighted χ^2 variables, with 0 and 1 degrees of freedom). Tests are carried out similarly to a standard χ_1^2 test with 1 degree of freedom, except one compares G to the critical χ_1^2 value and divides the p-value by 2.

A common definition of a value score is (see Wilks 1995):

$$\text{Value Score} = \frac{E(k^E) - E(k^N)}{E(k^P) - E(k^N)} \tag{3}$$

Here $E(k^P)$ is defined as the expected loss for perfect forecasts, that is, those X in which $X = Y$. A perfect forecast will have a value score equal to 1. Less than perfect forecasts will have a value score less than 1, with negative value being when the score is less than 0. It is easy to see that the same statistical test as above can be directly

applied to the value score, with the null hypothesis that the value score is less than or equal to 0 (i.e., $H_0:$ Value Score ≤ 0).

In the context of BR (when $k_{11} = k_{00} = 0$), we can calculate

$$\begin{aligned} E(k^N) &= (p_{1|1}p_{+1} + (1 - p_{0|0})(1 - p_{+1}))k_{10} \\ E(k^E) &= (1 - p_{1|1})p_{+1}k_{01} + (1 - p_{0|0})(1 - p_{+1})k_{10} \\ E(k^P) &= 0 \end{aligned}$$

where $E(k^P) = 0$ as there is no loss or cost for perfect forecasts.

Solving for (3) and substituting for the estimated parameter values gives an estimate of the value score. We have the estimate of the value score from BR:

$$\widehat{K}_\theta = \frac{(\widehat{p}_{1|1} - \theta)\widehat{p}_{+1}}{\widehat{p}_{1+}(1-\theta)} = \frac{n_{11}(1-\theta) - n_{01}\theta}{(n_{11} + n_{10})(1-\theta)}.$$

For there to be value, we must have that $\widehat{K}_\theta > 0$; for skill we must have $\widehat{K}_{1/2} > 0$. $\theta = 1/2$ when $k_{01} = k_{10}$ (and $k_{11} = k_{00} = 0$).

b. **Wilks's Value Score.** An example of a Cost-Loss value score is the Value Score (VS) proposed in Wilks (2001). Wilks has that a correct forecast has either a loss (or cost) $k_{11} = k_{01} > 0$ or a loss $k_{00} = 0$. Loss for incorrect forecasts is the same as before ($k_{10}, k_{01} > 0$). In these situations it only makes sense to talk of losses where $k_{01} < k_{10}$: this fits in with the restrictions already noted $k_{00} < k_{01} = k_{11} < k_{10}$; see Wilks (2001) for a complete explanation. We can develop a statistical test of Wilks's VS in the BR framework but just calculating the relevant value of θ' and then by calculating G (see below).

We now estimate the actual score. We have that

$$\begin{aligned} E(k^N) &= (p_{1|1}p_{+1} + (1-p_{0|0})(1-p_{+1}))k_{10} \\ E(k^E) &= p_{1|1}p_{+1}k_{11} + (1-p_{1|1})p_{+1}k_{01} + (1-p_{0|0})(1-p_{+1})k_{10} \\ E(k^P) &= p_{1|1}p_{+1}k_{11} \end{aligned}$$

where $E(k^P) > 0$ is again the expected loss for a perfect forecast and is not equal to 0 as it was in BR. This gives an estimate of

$$\widehat{VS}_\theta = \frac{n_{11}(1-2\theta) - n_{01}\theta}{n_{11}(1-2\theta) + n_{10}(1-\theta)}.$$

which is identical to that given in Wilks (again for "rare" observations), and is nearly the same as the value score K_θ in BR. For there to be value in the forecast, $\widehat{VS}_\theta > 0$.

We can now use the results of BR to create a test of significance for VS. The null hypothesis is still $H_0 :\ E(k^E) \geq E(k^N)$: which for VS translates to

$$H_0 : \quad p_{1|1} \leq \frac{k_{01} - k_{00}}{k_{01} - k_{00} + k_{10} - k_{11}} = \frac{k_{01}}{k_{10}} = \frac{\theta}{1-\theta} = \theta'$$

because $k_{00} = 0$ and $k_{11} = k_{01}$. Again, nothing changes from the results given in BR except the value of θ is different. For small θ, which is likely in the cost-loss problem, $\theta \approx \theta/(1-\theta)$ so the two tests are approximately the same; for larger θ the skill test based on the Value Score is more conservative than the one in BR because $\theta/(1-\theta) > \theta$.

c. **Complete Value Score.** Using equation (3), we can calculate a general loss version of the value score with

$$\begin{aligned} E(k^N) &= (p_{1|1}p_{+1} + (1-p_{0|0})(1-p_{+1}))k_{10} + ((1-p_{1|1})p_{+1} + p_{0|0}(1-p_{+1}))k_{00} \\ E(k^E) &= p_{1|1}p_{+1}k_{11} + (1-p_{1|1})p_{+1}k_{01} + (1-p_{0|0})(1-p_{+1})k_{10} + p_{0|0}(1-p_{+1})k_{00} \\ E(k^P) &= p_{1|1}p_{+1}k_{11} + p_{0|0}(1-p_{+1})k_{00}. \end{aligned}$$

This leads to the estimate of the complete value score of

$$\widehat{CVS} = \frac{n_{11}(k_{10} - k_{11}) - n_{01}(k_{01} - k_{00})}{(n_{11} + n_{10})k_{10} - n_{11}k_{11} + n_{01}k_{00}}. \tag{4}$$

The statistical test for value using the CVS follows from the results in Section 2.1. This implies that, for there to be value in the forecast, that $\widehat{CVS} > 0$. It is easily seen that setting $k_{11} = k_{00} = 0$ in (4) gives $\widehat{K}_\theta$; and that setting $k_{11} = k_{01}$, and $k_{00} = 0$ gives $\widehat{VS}_\theta$.

We have calculated all formulas with the idea that the observations are "rare" with respect to the loss, that is, the the optimal naive forecast was to always say no event ($X^N \equiv 0$). This condition holds when

$$(p_{1|1}p_{+1} + (1 - p_{0|0})(1 - p_{+1}))k_{10} \leq ((1 - p_{1|1})p_{+1} + p_{0|0}(1 - p_{+1}))k_{01}$$

which is when the expected loss of always saying 0 is less than or equal to the loss of always saying 1. Of course, this is not true for all forecasts and observations. As such, we can always write the complete value score as

$$CVS = CVS_0 I(X^N = 0) + CVS_1 I(X^N = 1)$$

where, for example, $I(X^N = 0)$ is 1 when the optimal naive forecast is 0 and is 0 otherwise, and the estimate for CVS_0 is given in (4). So, for completeness we present the estimate of CVS_1

$$\widehat{CVS}_1 = \frac{n_{00}(k_{01} - k_{00}) - n_{10}(k_{10} - k_{11})}{(n_{00} + n_{01})k_{01} - n_{00}k_{00} + n_{10}k_{11}}.$$

Similar decompositions can be given for K_θ and VS_θ: see the original papers for the results.

3. Example and Discussion

A small example will help distinguish the differences between cases of the CVS. Imagine a forecast/observation table such as that in Table 1. We consider three

plausible cost scenarios. The first represents a standard statistical test of climate skill with $k_{11} = 0$, $k_{01} = 10$, $k_{10} = 10$, and $k_{00} = 0$. BR proved that in this case the statistical test of value is identical with that of climate skill; that is, the test of whether the expert forecast has a higher probability of being correct than the optimal naive forecast. The exact values of k_{01} and k_{10} are not important since they are equal. The second scenario is a cost-loss situation with $k_{11} = 10$, $k_{01} = 10$, $k_{10} = 30$, and $k_{00} = 0$. The last scenario is the same cost-loss situation plus a fee of 5 units for the forecast: $k_{11} = 15$, $k_{01} = 15$, $k_{10} = 35$, and $k_{00} = 5$. For Table 1 we estimate $\widehat{p}_{1|1} = 5/(5+5) = 0.50$ which we know must be larger than $\theta' = \frac{k_{01}-k_{00}}{k_{01}-k_{00}+k_{10}-k_{11}}$ for there to be value or skill. Table 2 lists θ' and the relevant value score for the each of the three situations.

For scenario 1, $\widehat{K}_{1/2} = 0.5$ and says no skill, which is not surprising since $\widehat{p}_{1|1} = \theta'$: we have done no better than simply guessing 0 each time. Here $G = 0$ and the corresponding p-value $= 0.5$ (an ordinary χ^2_1 test at this value would give a p-value of 1, which we have divided by 2). Scenario 2 has $\widehat{VS} = 0.31$ and $\theta' = 0.33$, which indicates that value is present. Scenario 3 has $\widehat{CVS} = 0.27$ and again $\theta' = 0.33$, which indicates that value is present. By construction, both scenarios 2 and 3 have $G = 1.2$ and p-value of 0.14, indicating that the shown value is not significant by the usual criterion; not surprising given the small number of observations in the Table. It is also not surprising that CVS and VS say nearly the same thing for scenarios 2 and 3 because, for small k_{00} relative to k_{10} and with $k_{11} = k_{01}$, $\widehat{CVS} \approx \widehat{VS}$.

This generalization should prove useful to forecast verification and value studies across a wide range of situations. Regardless of whether the user is interested in skill, cost-loss, or general value, the statistical method, test statistics, and p-value calculations are the same.

A reviewer to an earlier draft of this paper pointed out that, for this example, the value of the Heidke skill score is 0.23 and that of the Kuipers skill score is 0.36, both of which indicate skill with respect to random chance. Our score showed no skill with respect to climate: the expert forecast had 78% correct, the same as if 0 was forecast each time. Skill with respect to climate is a stronger requirement than that with respect to random chance. In this sense, the Heidke and Kuipers skill scores are closer to traditional tests of independence (between Y and X). This is explored in more depth in BR

Acknowledgements

The development of the Complete Value Score was inspired after comments from three reviewers, without whom this paper would not exist.

References

1. Briggs, W.M., M. Pocernich, and D. Ruppert, 2005: Incorporating misclassification error in skill assessment. Submitted to *Monthly Weather Review*.
2. Briggs, W.M., and D. Ruppert, 2004: Assessing the skill of yes/no forecasts for Markov observations. *17th Conf. on Probability and Statistics in the Atmospheric Sciences*, Seattle, WA, Amer. Meteor. Soc.
3. Briggs, W.M., and D. Ruppert, 2005: Assessing the skill of yes/no forecasts. *Biometrics*, in press.
4. Hamill, T. M., 1998. Hypothesis tests for evaluating numerical precipitation forecasts: *Weather and Forecasting*, **14**, 155-167.
5. Richardson, D.S., 2000. Skill and relative economic value of the ECMWF ensemble prediction system. *Q.J.R. Meteorol. Soc.*, **126**, 649-667.
6. Wilks, D.S., 2001: A skill score based on economic value for probability forecasts. *Meteorological Applications*, **8**, 209-219.
7. Wilks, D.S. (1995). *Statistical Methods in the Atmospheric Sciences*. New York: Academic Press.

List of Tables

TABLE 1. Hypothetical forecast/observation data.

	$Y = 1$	$Y = 0$
$X = 1$	5	5
$X = 0$	2	20

.

TABLE 2. Calculations of θ' for the three scenarios presented in the text. Recall that $\widehat{p}_{1|1} = 0.50$ and must be larger than θ' for value to exist. The relevant value statistic is also calculated.

Losses	Scenario	θ'	Score
$k_{11} = 0$ $k_{01} = 10$ $k_{10} = 10$ $k_{00} = 0$	Skill	0.50	$\widehat{K}_\theta = 0$
$k_{11} = 10$ $k_{01} = 10$ $k_{10} = 30$ $k_{00} = 0$	Cost-loss	0.33	$\widehat{VS}_\theta = 0.31$
$k_{11} = 15$ $k_{01} = 15$ $k_{10} = 35$ $k_{00} = 5$	Fee	0.33	$\widehat{CVS} = 0.26$