Predicating Predictivity

Plus predicaments of error modeling

Sir David Spiegelhalter is a British statistician. He is a strong voice for the public understanding of statistics. His work extends to all walks of life, including risk, coincidences, murder, and sex.

Today we talk about extending one of his inventions.

His invention has to do with grading the performance of people and models that make predictions. A **scoring rule** grades how often predictions are right. But it may not tell how difficult the situations are. It is easy to look good with predictions when they start with a high chance of success. A weather forecaster predicting sunny-versus-rainy will be right more often in Las Vegas than in Boston. Quoting this FiveThirtyEight item:

> If you want to have an easy life as a weather forecaster, you should get a job in Las Vegas, Phoenix or Los Angeles. Predict that it won't rain in one of those cities, and you'll be right about 90 percent of the time.

In a 1986 paper, for a particular scoring rule defined by Glenn Brier in 1950, Spiegelhalter worked out how to equalize the forecaster grading. He applied his **Z-test** not to weather like Brier did but to medical prognoses and clinical trials.

What I am doing with a small group of graduate students in Buffalo is trying to turn Spiegelhalter's kind of Z-test around once more. If a forecaster fares poorly, we will try to flag not the model but the behavior of the subjects being modeled. In weather we would want to tell when Mother Nature, not the models, has gone off the rails. Well, we are actually looking for ways to tell when a human being has left the bounds of human predictability for reasons that are inhuman—such as cheating with a computer at chess. And maybe it can shed more light on whether our computers can possibly cheat with quantum mechanics.

Prediction Scores

Let's consider situations t in which the number $\ell = \ell_t$ is usually more than 2, that is, usually more than "rain" or "no rain." The forecaster lays down projections $\vec{q} = \vec{q}_t = (q_1, \ldots, q_\ell)$ for the chance of each outcome. If outcome r happens, then the *Brier score* for

that forecast is

$$B^{\vec{q}}(r) = (1 - q_r)^2 + \sum_{j \neq r} q_j^2. \tag{1}$$

If the forecaster was certain that r would happen and so put $q_r = 1$, all other $q_j = 0$, then the score would be zero. Thus lower is better for the Brier score.

If you put probability $q_r < 1$ on the outcome that happened, then you get penalized both for the difference and for the remaining probability which you put on outcomes that did not happen. It is possible to *decompose* the score in another way that changes the emphasis:

$$B^{\vec{q}}(r) = 1 + Q - 2q_r \qquad \text{where} \qquad Q = \sum_{j=1}^{\ell} q_j^2.$$

Then Q is a fixed measure of how you spread your forecasts around, while all the variability in your score comes from how much stock you placed in the outcome that happened. The worst case is having put $q_r = 0$, whereupon your Brier penalty is 2.

We would like our forecasts always to be perfect, but reality gives us situations that are inherently nondeterministic—with unknown "true probabilities" $\vec{p}_t = (p_1, \ldots, p_\ell)$. The vital point is that the forecaster should not try to hit $r = r_t$ on the nose at every time t but rather to match the true probabilities. Once we postulate $\vec{p}$, the *expected Brier score* is

$$\begin{aligned} \mathsf{E}_{\vec{p}}[B^{\vec{q}}] &= \sum_{i=1}^{\ell} p_i B^{\vec{q}}(i) \\ &= \sum_{i=1}^{\ell} p_i (1 - 2q_i + Q) \\ &= 1 + Q - 2\sum_{i=1}^{\ell} p_i q_i. \end{aligned}$$

This is uniquely minimized by setting $q_i = p_i$ for each i, which defines B as a **strictly proper** scoring rule. Without the second term $\sum_{j \neq r} q_j^2$ in (1) the rule would not be proper for $\ell > 2$. When $\vec{q} = \vec{p}$, Q becomes equal to $P = \sum_{j=1}^{\ell} p_j^2$. Thus P represents an unavoidable prediction penalty from the intrinsic variance. If all p_i are equal, $p_i = \frac{1}{\ell}$, then the expected score cannot be less than $1 - \frac{1}{\ell}$.

Log-Likelihood Scoring

Another popular scoring rule is the *log-likelihood* rule:

$$L^{\vec{q}}(r) = \log(\frac{1}{q_r}).$$

Now if you put $q_r = 0$ and outcome r happens, your penalty is infinite. Your expected score is:

$$\mathsf{E}_{\vec{p}}[L^{\vec{q}}] = \sum_{i=1}^{\ell} p_i \log(\frac{1}{q_i}).$$

This is the *cross entropy* from the projected probabilities to the true ones, which we discussed at the end of our recent post on quantum supremacy. It hides a term analogous to P above. To wit,

$$\mathsf{E}_{\vec{p}}[L^{\vec{q}}] - \sum_{i=1}^{\ell} p_i \log(\frac{1}{p_i}) = \sum_{i=1}^{\ell} p_i \ln(\frac{p_i}{q_i})$$

is the *Kullback-Leibler divergence* from $\vec{q}$ to $\vec{p}$. This is nonnegative and zero only when $\vec{q} = \vec{p}$, which also implies that the log-likelihood rule is strictly proper. The subtracted-off *entropy* term thus represents an unavoidable penalty in $L^{\vec{q}}(r)$.

One way to generalize these scores is to put a cost function $f(i)$ on the outcomes. Then

$$B_f^{\vec{q}}(r) = f(r)(1 - q_r)^2 + \sum_{j \neq r} f(j) q_j^2$$

remains a proper scoring rule. However,

$$L_f^{\vec{q}}(r) = f(r) \log(\frac{1}{q_r})$$

is generally no longer proper. To see why, suppose $\{f(i)p_i\}$ is again a probability distribution. Then the expectation

$$\mathsf{E}_{\vec{p}}[L^{\vec{q}}] = \sum_{i=1}^{\ell} p_i f(i) \log(\frac{1}{q_i}),$$

is minimized by setting $q_i = p_i f(i)$ rather than $q_i = p_i$ for each i. Nevertheless, the "multinomial" log-likelihood rule

$$\begin{aligned} L'_f(r) &= f(r) \log(\frac{1}{q_r}) + \sum_{j \neq r} f(j) \log(\frac{1}{1 - q_j}) \\ &= f(r) \ln(\frac{1 - q_r}{q_r}) + \sum_{j=1}^{\ell} f(j) \log(\frac{1}{1 - q_j}) \end{aligned}$$

is once again strictly proper. The second term is again an unavoidable penalty, while $\ln(\frac{1-q_r}{q_r})$ is the (negative) logit of the true outcome. Note, both ln and log mean natural logarithm but we use the latter to guarantee that the value is non-negative.

Spiegelhalter's Z

Spiegelhalter's z-score neatly drops out the unavoidable penalty term by taking the difference of the score with the expectation. Schematically it is defined as

$$\mathsf{Z}[B] = \frac{B - \mathsf{E}[B]}{\sqrt{\mathsf{Var}[B]}},$$

where $\mathsf{Var}[B]$ means the projected variance $\mathsf{E}_{\vec{p}}[B^2] - (\mathsf{E}_{\vec{p}}[B])^2$. However, here is where it is important to notate the whole series of forecasting situations $t = 1, \ldots, T$ with outcomes r_t for each t. The actual statistic is

$$\mathsf{Z}_{\vec{p}}[B^{\vec{q}}] = \frac{\sum_{t=1}^{T} B^{\vec{q}_t}(r_t) - \mathsf{E}_{\vec{p}_t}[B^{\vec{q}_t}]}{\sqrt{\sum_{t=1}^{T} \mathsf{Var}_{\vec{p}_t}[B^{\vec{q}_t}]}}. \tag{2}$$

The denominator presumes that the forecast situations are independent so that the variances add. The numerator expands to be

$$\sum_{t=1}^{T} \left(2 \sum_{i=1}^{\ell_t} p_{i,t} q_{i,t} \right) - 2q_{r,t}.$$

The original application is a confidence test of the "null hypothesis" that the projections $\vec{q}$ are good. Thus we plug in $p_{i,t} = q_{i,t}$ for all t and i so that we test

$$\mathsf{Z}_{\vec{q}}[B^{\vec{q}}] = 2\frac{\sum_{t=1}^{T} \left(\sum_{i=1}^{\ell_t} q_{i,t}^2 \right) - q_{r,t}}{\sqrt{\sum_{t=1}^{T} \mathsf{Var}_{\vec{q}_t}[B^{\vec{q}_t}]}}.$$

To illustrate, suppose we do ten independent trials of an event with four outcomes whose true probabilities are $(0.1, 0.2, 0.3, 0.4)$. The sum in parentheses is $10(0.01 + 0.04 + 0.09 + 0.16) = 3$. If the outcomes conform exactly to these probabilities then $q_{r,t}$ equals 0.1 once, 0.2 twice, 0.3 three times, and 0.4 four times. This exactly cancels the 3, so $\vec{q} = \vec{p}$ makes $\mathsf{Z}_{\vec{q}}[B^{\vec{q}}] = 0$, as expected. Most trials will give a nonzero numerator, but in the long run, the numerator divided by T tends toward zero and the denominator scales to match it, thus keeping the Z-statistic normally distributed.

A high Z, on the other hand—highly positive or highly negative—indicates that the forecasting is way off. That (2) is an aggregate statistic over independent trials justifies treating the Z-values as standard scores. This applies also to Z-tests made similarly from other scoring rules besides the Brier score. The test thus becomes a verdict on the model. High Z-values on certain subsets of the data may reveal biases.

Our idea is the opposite. Suppose we know that the forecasts are true, or suppose they have biases that are known and correctable over moderately large data sets. We may then be able to fit $\mathsf{Z}[B]$ as an unbiased estimator (of zero) over large training sets. Then it can become a judgment of whether the data has become unnatural.

Why This Z?

As I have detailed in numerous posts on this blog, my system for detecting cheating with computers at chess already provides several statistical z-scores. Why would I want another one?

The motive involves the presence of multiple strong chess-playing programs, each with its own quirks and distribution of values for moves. They are used in two different ways:

1. As inputs telling the relative values v_i of moves m_i, which my model converts into its probability projections q_i.
2. As output predicates telling how often the player chose the move recommended by a specific program and/or quantifying the magnitude of error for different played moves.

Having multiple engines helps point 1. My intent to blend the *values* v_i from different engines has been blunted by issues I discussed here. Thus I now have to train my model separately (and expensively) for each (new version of each) program. I can then blend the q_i, but point 2 still remains at issue: My tests measure concordance with a specific program. Originally the program Rybka 3 was primary and Houdini 4B secondary. Now Stockfish 7 is primary and Komodo 10.0 secondary—until I update to their latest versions. The second engine is supposed to confirm a positive result from the first one. This already means that my model is not trying to detect exactly which program was used.

Nevertheless, my results often vary between testing engines. The engines compete against each other and may be crafted to disagree on certain kinds of moves. They agree with each other barely 75–80% in my tests. I would like to factor these differences out.

The Spiegelhalter Z-test appeals because its reference is not to a particular chess program, but to the prediction quality of my model itself—which per point 1 can be informed by many programs in concert. It gives a way to *predicate predictivity*. A high value will attest that the sequence of played moves falls outside the range of predictability for human players of the same rated skill level.

The Method

To harness $\mathsf{Z}[F]$ for some scoring rule F, we need to quantify the nature of my model's q_i projections. In fact, my model has a clear bias toward conservatism in judging the frequency of particular non-optimal moves. This is discussed in my August post on my model upgrade and shown graphically in an appended note on why the conservative setting of a "gradient" parameter is needed to preserve dynamical stability. The fitting offsets this in a way that creates an opposite bias elsewhere. I hope to correct both biases at the same stroke by a specific means of modeling how the q_i err with respect to the postulated true probabilities p_i.

We postulate an original source of error terms ϵ_i all i.i.d. as $\mathcal{N}(0, \delta^2)$, where δ governs the magnitude of Gaussian noise. This noise can be *transformed* and related in various ways, e.g.:

1. $q_i = p_i \pm \epsilon_i$,
2. $q_i = p_i(1 \pm \epsilon_i)$,
3. $\frac{1}{q_i} = \frac{1}{p_i} \pm \epsilon_i$,
4. $\log(\frac{1}{q_i}) = \log(\frac{1}{p_i}) \pm \epsilon_i$,
5. $\log(\frac{1}{q_i}) = \log(\frac{1}{p_i})(1 \pm \epsilon_i)$,
6. $\ln(\frac{q_i}{1-q_i}) = \ln(\frac{p_i}{1-p_i}) \pm \epsilon_i$.

There are further forms to consider and it is not yet clear from data within my model which one most applies. We would be interested in examples where these representations have been employed and in observations about their natures.

Given any error terms, we can write each p_i as a function of q_i and ϵ_i. One issue is having at most $\ell - 1$ degrees of freedom among $\epsilon_1, \ldots, \epsilon_\ell$, owing to the constraint that the q_i as well as p_i sum to 1. We handle this by choosing some fixed k as the "pivot" and using the constraints to eliminate p_k and q_k, leaving the other error terms free. In all cases, the proposed method of defining what we notate as $\mathsf{Z}_{\vec{q},\vec{\epsilon}}[F]$ is:

- Substitute the terms with q_i, ϵ_i for each free p_i into $\mathsf{Z}_{\vec{p}}[F^{\vec{q}}]$.
- Compute the expectation over $\epsilon_i \sim \mathcal{N}(0, \delta^2)$ for the numerator and denominator of (2), separately.
- Holding the other previously-fitted model parameters in place, fit δ so that $\mathsf{Z}_{\vec{q},\vec{\epsilon}}[F]$ is zero over the training set (or sets, for each level of Elo rating R, so δ becomes a function of R).

If the resulting Z-scores parameterized by δ_R make sense, the last step will be adjusting them to conform to normal distribution, via the resampling process mentioned recently here and earlier here. We are not there yet. But observations from Spiegelhalter tests with $\vec{q} = \vec{p}$ (equivalently, with δ_R fixed to zero) suggest that the resulting single, authoritative, "pure" predictivity test may rival the sharpness of my current tests involving specific chess porgrams.

Error Quirks and Queries

To see a key wrinkle, consider the first error form. It is symmetrical: $p_i = q_i \pm \epsilon_i$. When we substitute $q_i + \epsilon_i$ for p_i and take $\mathsf{E}_{\epsilon_i \sim \mathcal{N}(0,\delta^2)}[\cdots]$, the symmetry of ϵ_i around 0 makes it drop out of the numerator of (2), and out of everything in the denominator except one place where p_i^2 becomes $(q_i^2 + 2\epsilon q_i + \epsilon_i^2)$. There is hence nothing for δ to fit and we are basically left with the original Spiegelhalter Z.

In the second form, however, we get $p_i = q_i \cdot \frac{1}{1+\epsilon_i}$. If we presume δ small enough to make the distribution of $\mathcal{N}(0, \delta^2)$ outside $(-1, 1)$ negligible, then we can use the series expansion to approximate

$$p_i \approx q_i(1 - \epsilon_i + \epsilon_i^2 - \epsilon_i^3 + \epsilon_i^4).$$

Under normal expectation, the odd-power terms drop out (so their signs don't matter) and we get

$$\mathsf{E}_{\epsilon_i \sim \mathcal{N}(0,\delta^2)}[q_i(1 - \epsilon_i + \epsilon_i^2 - \epsilon_i^3 + \epsilon_i^4)] = q_i(1 + \delta^2 + 3\delta^4).$$

This credits p_i as being greater than q_i. Provided the projections for the substituted indices i were generally slightly conservative, this has hope of correcting them.

Already, however, we have traipsed over some pitfalls of methodology. One is that the normal expectation

$$\mathsf{E}_{\epsilon \sim \mathcal{N}(0,\delta^2)}[\frac{1}{1+\epsilon}] = +\infty,$$

regardless of how small δ is. For any δ, regions around the pole $\epsilon = -1$ get some fixed finite probability. Another is the simple paradox of our second form saying:

> q_i is an unbiased estimator of p_i, but p_i is not an unbiased (or even finite) estimator of q_i.

A third curiosity comes from the fourth error form. It gives $q_i = p_i e^{\epsilon_i}$, so $p_i = q_i e^{-\epsilon_i}$. We have

$$\mathsf{E}_{\epsilon \sim \mathcal{N}(0,\delta^2)}[e^{b\epsilon}] = e^{0.5b^2\delta^2}$$

exactly, without approximation. Again the sign of ϵ_i does not matter. So we get

$$\mathsf{E}_{\vec{\epsilon}}[p_i] = q_i e^{0.5\delta^2} > q_i.$$

But by the original fourth equation we get

$$\mathsf{E}_{\vec{\epsilon}}[q_i] = p_i e^{0.5\delta^2} > p_i.$$

So we have $\mathsf{E}[q_i] > p_i$ and $\mathsf{E}[p_i] > q_i$, with both expectations being over the same noise terms. This is like the famous Lake Wobegon syndrome. What it indicates is the need for care in where and how to apply these error representations.

Open Problems

Have you seen this idea of directly testing (un)predictability in the literature? Might it improve the currently much-debated statistical tests for quantum supremacy?

Which error model seems most likely to apply? Where have the paradoxes in our last section been noted?