Probabilistic Scoring Rules for Multiple Choice Tests and Other Applications in Finance

Andrew B. Carver, The College of New Jersey

ABSTRACT

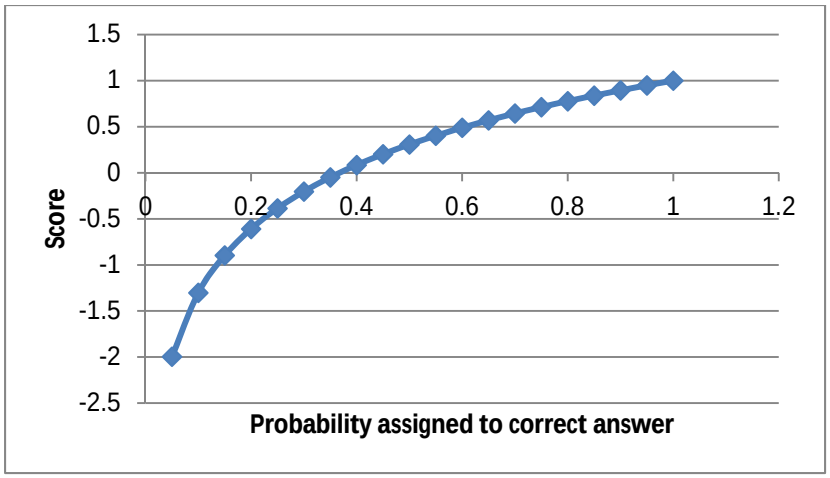
INTRODUCTION

A scoring rule is simply a way to transform one's probability assessment for an uncertain event into a numerical score based on the outcome of that uncertainty. It is an interesting way to evaluate the performance of forecasters. To illustrate, we start with a familiar non-finance example. Each day a weatherman is asked to provide the chance of rain for the following day. Despite the best equipment and most advanced meteorological models, the weatherman cannot be 100% certain of whether or not it will rain tomorrow. Rather, he provides his forecast as a probability--20% chance of rain, for instance.

THE LOGARITHMIC SCORING RULE

How would we evaluate a weather forecaster who provides this probabilistic information? Scoring rules provide a compelling answer. The evaluator can choose from a range of scoring rules that encourage truthful revelation by the weather forecaster. Our preferred scoring rule takes the following logarithmic functional form:

So if the weather forecaster places a 20% probability on the occurrence of rain (and thus an 80% chance of no rain), the forecaster will be awarded -0.61 points if it rains and 0.78 points if it does not rain. The following graph shows the score as a function of the probability placed on the event that actually occurs:



It can be shown that a "strictly proper scoring rule", such as the logarithmic one shown above, encourages the weather forecaster to truthfully reveal his subjective probability assessment if he wants to maximize his expected score. To my knowledge, no weather station has ever based performance appraisals on probabilistic scoring rules but in theory they could.

To see that the proper scoring rule encourages truthful revelation while most other rules do not, consider the alternative linear scoring rule below:

where p is once again is the probability the forecaster places on the event that comes to be. Under this scoring rule a thoughtful forecaster trying to maximize his expected score would choose to place a p=1 when he believes the probability of rain is greater than 50% and will choose p=0 when the probability is less than 50%. This is because placing an extreme probability on the most likely outcome (even if it has a probability of 0.51) will earn the forecaster the greatest score (one), most often. So, this scoring rule is less desirable since it encourages the weatherman to game the system by providing extreme probability forecasts.

OTHER SCORING RULES

In addition to the logarithmic scoring rule demonstrated above, there exist other formulas that encourage truthful revelation (Winkler, 1968). Other so-called proper scoring rules are the quadratic and spherical forms. The spherical scoring rule for an uncertain event with n possible outcomes takes the form:

Where $p_{correct}$ is the probability the forecaster placed on the event that occurred and p_1 through p_n represent the forecaster's probability assessment for outcomes 1 through n. Note that unlike the logarithmic scoring rule, the score assigned to the forecaster depends not only on the probability placed on the event that actually occurred, but also on the probabilities that were placed on events that did not occur. Similarly, a quadratic scoring rule takes the form:

Each of the three scoring rules has as the property that a forecaster who is interested in maximizing their expected (probability weighted) score should assign their true beliefs about the likelihood of each outcome. The differences in the functional form of each formula and how they impact evaluations was described by Bickel (2007). A critical difference between the logarithmic formula and the other two, is that the logarithmic rule gives a score of negative infinity when the forecaster places a probability of zero on an outcome that actually occurs. This may seem intimidating to forecasters, but it encourages them to consider carefully the likelihood of rare, low probability events.

APPLICATIONS IN FINANCE

Like weathermen, analysts and investors are faced with situations in which they need to assess uncertainty, such as a company's earnings next quarter or a firm's stock price in one year. In academic literature, a number of papers have used scoring rules to evaluate the quality of financial forecasts (Yates and McDaniel, 1991; Stael Von Holstein, 1972; Wilkie and Pollock, 1996). In one experiment, Stael von Holstein (1972) asked participants, some of whom were financial professionals, to place probability percentiles on the future prices of stocks. Based on the use of a scoring rule, he found that although participants were able to make probability assessments, most of them did not score well in the exercise, highlighting the difficulty in predicting future stock prices.

SCORING RULES FOR MULTIPLE CHOICE TESTS

In the classroom, scoring rules can enhance a student's ability to recognize and deal with uncertainty. Perhaps the most significant application of scoring rules is as an interesting alternative to standard multiple choice tests (Echternacht, 1972; Bickel, 2010). Under the scoring rule, students are not asked to select a single correct answer; rather, they must place probabilities on whether each of the possible answers is correct.

This may present several advantages over traditional multiple choice tests. First, a student who has limited information about a particular question is not in a position to randomly pick an answer and be rewarded only on the basis of luck. Rather, it is in that student's best interest to truthfully reveal his lack of information by nearly equally distributing probabilities among the candidate answers. In addition, a student who has partial knowledge is awarded with partial credit. Let's take the example of student who must select one of four possible answers. If the student has eliminated two of the four answers based on previous hours of studying, then he or she might answer as follows:

Answer	Student s Subjective Assessment of Answer	Student s Probability
A	Possibly correct answer	0.45
B	Not likely to be correct	0.05
C	Possibly correct answer	0.45
D	Not likely to be correct	0.05

In this case, the student will get 0.2 points [calculation: 1+ln(.45)] if either A or C turns out to be the correct answer. This point total is not as high as it would be if the student had studied harder and was more certain of the correct answer, but it mitigates some of the randomness that comes from guessing either A or C, as would be the case with traditional multiple choice tests.

In many cases, the instructor will need to make adjustments if the student fails to properly give answers in probabilistic form. For example, if the student's probabilities add to greater than 1, the instructor should normalize the scores so that they add to one.

It should be noted that on the familiar, typical multiple choice test, the student is required to select only one answer—similar to placing a probability of 100% on the single most likely answer. As pointed out by Bickel (2010), a probability assignment equal to 100% would be the best answer on a probabilistically answered exam if the scoring rule were Score=p. Thus, by measuring the degree of conviction in a student's answer through a strictly proper scoring rule, the instructor gathers more information about the student's beliefs for each question.

Another advantage of using a scoring rule for multiple choice tests is that it encourages students to think about uncertainty and gives them practice at working with probabilities. Uncertainty surrounds finance professionals in many aspects of their work, whether it be evaluating an investment portfolio in the face of market uncertainty or determining the value of a business expansion project that entails uncertainty about future demand for a product. Having the ability to recognize and deal with uncertainty may prepare students to deal with any number of situations in the real world.

ADDITIONAL FINANCE LEARNING USING SCORING RULES

Scoring rules have been used elsewhere in classroom activities or games. Carver (2008) discusses a classroom game where students are asked to give assessments on whether the stock market will move higher in the following days. In the game, students give probability forecasts, such as "I believe there is a 0.6 probability that the Dow Jones Industrial Average will move higher tomorrow". The quality of the forecasts is scored based on a scoring rule. The exercise is intended to give the students a working knowledge of probability assessment and market efficiency.

To facilitate the game, the instructor first chooses a number of trading days for the assignment. For each day, the student is asked to assign a probability to the likelihood that the Dow Jones Industrial Average will close higher than on the prior trading day. Before the market opens each day, the students' probability forecasts are collected through the course website.

The instructor can choose how much advance information about market efficiency to give to students. In my course, I tell students to do as much or as little investigation as desired to come up with their probability forecasts. Scores are determined by the logarithmic scoring rule.

This exercise allows students to experience how difficult it is to forecast the market, providing a platform to discuss the concept of market efficiency. Since markets are fairly

efficient, students will be hard pressed to consistently perform well by placing extreme probability forecasts. If the efficient market hypothesis holds, a student's highest score in an expected value sense will be achieved by submitting a probability very close to 0.5 each day. When the game is played over a sufficient number of trading days, the highest performing students will be those who place probabilities in the neighborhood of 50% each day.

The exercise may also provide insight into ideas from behavioral finance. Behavioral decision making research suggests people exhibit systematic biases in judging the likelihood of events. One common bias, sometimes called the gambler's fallacy, leads forecasters to believe that a market that has recently moved significantly higher will tend to mean revert by moving lower. On the other hand, some investors, as well as some students who engage in the market forecasting game, tend to believe that markets that recently moved higher will continue to trend in the same direction, which is consistent with Kahneman and Tversky's (1973) finding that humans tend to overweigh the most recent observations. In my experience with this exercise, students often base a given day's probability forecast on the outcome from the previous day. Given what we know about behavioral finance, this finding is not surprising.

CONCLUSION

Scoring rules offer an interesting way to score student probability forecasts. Perhaps the most likely application of scoring rules is for multiple choice tests. The use of such rules can offer a better gauge of students' preparation by allowing the instructor to measure partial knowledge, something not possible with traditional multiple choice exams.

Scoring rules can also be used as part of in-class activities. This paper discussed a classroom game in which students were asked to act as market forecasters. The challenges the students face as forecasters allow the instructor to naturally introduce market efficiency and behavioral finance concepts. Whether used for multiple choice tests or in-class games, students using scoring rules gain practice at working with probabilities to quantify uncertainty. By placing students in a position to think carefully about the uncertainty, probabilistic scoring rules provide a platform for novel pedagogical experiences.

REFERENCES

Bickel, Eric J. 2007. "Some Comparisons Among Quadratic, Spherical, and Logarithmic Scoring Rules." [illegible] 4 (2, June), 49-65.

Bickel, Eric J. 2010. "Scoring Rules and Decision Analysis Education." [illegible] 7 (4, December), 346-357.

Carver, Andrew B. 2008. "Teaching Market Efficiency with the Stock Market Assessment Game." [illegible], 6(1), (Winter 2008), 79-88.

Echternacht, Gary J. 1972. "The Use of Confidence testing in Objective Tests." [illegible] 42 (2, Spring), 217-236.

Kahneman, D. and Tversky, A. 1973. “On the Psychology of Prediction.” [illegible] 80 (1), 237-251.

Stael von Holstein, C.A.S. 1972. “Probabilistic Forecasting: An Experiment Related to the Stock Market.” [illegible] 8, 139-158.

Wilkie, M.E. and Pollock, A.C. 1996. “An Application of Probability Judgment Measures to Currency Forecasting.” [illegible], 12, 25-40.

Winkler, R.L. and Murphy, A.H. 1968. “‘Good’” Probability Assessors.” [illegible], 7 (5), 751-758.

Yates, J.F. and McDaniel, L.S. 1991. “Probabilistic Forecasts of Stock Prices and Earnings: The Hazards of Nascent Expertise.” [illegible] 49, 60-79.