Article from

Predictive Analytics and Futurism

December 2015
Issue 12

A Comparison of Risk Scoring Recalibration Methods

By Geof Hileman

In the July 2015 issue of this newsletter, Shea Parkes and Brad Armstrong published an article titled "Calibrating Risk Score Models with Partial Credibility." In this article, they presented an application of the "ridge regression" technique to the calibration of health-based risk scoring models. This calibration process is often undertaken to tailor a risk scoring model to a specific population on which it is being applied.

This article's publication was timely, as we are currently engaged in updating the SOA's periodic study that compares the predictive accuracy of various risk scoring models. This study has been published three times previously (in 1996, 2002 and 2007), with the 2007 study available at *https://www.soa.org/research/research-projects/health/hlth-risk-assement.aspx*.

This new study, while currently still underway, will include a comparison of the accuracy of prospective and concurrent models out-of-the-box, using the weights provided by each vendor with the models. In addition to this out-of-the-box comparison, which is consistent with the way models are frequently implemented, we have also considered various approaches to recalibrating each of the models included in the comparison. This step is an important element of model comparison, as the comparison of recalibrated models gives insight into the potential predictive ability of each model and normalizes for any differences in the populations on which the offered weights are based.

We have considered three approaches to recalibrating each of the models in the study. The first approach under consideration is full recalibration. Full recalibration is the approach used in the 1996 and 2002 studies and is the approach that would be considered to be the most conventional, given an adequately large data source. To perform a full recalibration, the actual scaled cost level is regressed against the complete set of independent variables to determine new model weights. A critical disadvantage of the full re-specification approach is that full transparency into the workings of the model is required. In order to implement full re-specification without losing any of the clinical logic, one would need to know all of the inputs to the model, including any hierarchical logic or combination variables. Not all vendors provide this degree of transparency along with their models.

The second approach we considered is that which was used in the development of the 2007 study. The 2007 approach differs from a full recalibration in that the dependent variable in the regression equation is the residual, rather than the scaled cost variable. Thus, the equation resulting from the regression gives the expected error for each individual and can be added to the originally-predicted risk score. Without any consideration for statistical significance, the estimated coefficients from the residual approach are by definition equal to the difference in the original model weights and the weights that result from the full recalibration. The authors of the 2007 study introduced a credibility weighting where each coefficient is weighted by $(1-p)^{5.95}$, where p is equal to the p-value associated with that particular coefficient. Accounting for the credibility weighting, the adjustment to each individual's estimated risk score is given by the dot product of three vectors: the estimated coefficients, the credibility weights, and the specific values of each of the independent variables.

The third approach is the ridge regression approach discussed by Parkes and Armstrong. Like the p-value approach used in the 2007 study, this method regresses on the residuals rather than the original dependent variable. However, the blending of the original and the re-estimated coefficients is handled in a less blunt fashion. In an ordinary least squares regression, coefficients are determined to minimize the sum of the squared errors across all observations. In ridge regression, the objective function is modified to incorporate a penalty corresponding to the size of the sum of the estimated coefficients. Thus, the optimal weights strike an appropriate balance between fitting the data and minimizing changes to the original coefficients.

One significant advantage offered by both of the residual approaches is that the details behind the original risk model can remain somewhat obscured. Since both approaches produce an estimate of the expected difference from the original risk score, all of the independent variables that contribute to that risk score do not need to be known. Any variables that are omitted from the re-specification would essentially retain their original weight along with any error that their coefficients contribute toward being absorbed by other variables.

In order to determine the most appropriate approach for our application—comparing commercial risk scoring models—we tested each of the methods in a recalibration of the Clinical Illness & Disability Payment System (CDPS) model. CDPS is an ideal mod-

el for testing the recalibration approaches, because it is entirely transparent and, by virtue of the offered weights being based on a Medicaid population, recalibration on a commercial population should lead to different weights. We selected two samples of just under 700,000 adults from Truven Health's MarketScan databases and used one for recalibrating the weights and the second for computing statistics on the recalibrated models.

To evaluate the effects of the three approaches to recalibration, we first compared the coefficients produced by each of the three methods. The coefficients, while all somewhat different from the original CDPS weights, were very consistent across the three methods. As expected, the coefficients resulting from the 2007 approach were identical to the full recalibration approach in cases where the p-value was 0.0 (and the credibility was thus 100 percent). Larger differences were present across the approaches for the higher-severity lower-frequency conditions.

We also compared the degree to which the recalibrated models explained the variation in the cost data. Using the original weights, we calculated an R-squared of 11.24 percent. Both the full recalibration and the 2007 residual approach resulted in an identical R-squared of 13.70 percent, while the ridge regression returned a slightly higher value of 13.72 percent. Additionally, we computed the correlation coefficient among the four sets of predicted values, shown below in Table 1.

TABLE 1: CORRELATION COEFFICIENTS AMONG PAIRS OF PREDICTED VALUES

	Original Weights	Full Recalibration	2007 Residual Approach	Ridge Regression
Original Weights	1.00000	0.90455	0.90455	0.91483
Full Recalibration	-	1.00000	0.99995	0.99652
2007 Residual Approach	-	-	1.00000	0.99648
Ridge Regression	-	-	-	1.00000

Based on this comparison, we concluded that the selection of a recalibration method for large populations does not need to be guided by statistical fit, but rather by the constraints imposed by the particular models that are being worked with. The method described in the July 2015 newsletter was specifically recommended as being worthwhile when "trying to recalibrate a model for a population that is of moderate size, but not fully credible." Our analysis supports this conclusion, in that the approach provides an incrementally better fit, but is not meaningfully different from the more simplistic approaches when applied to a very large population. ■

Geof Hileman, FSA, MAAA, is VP at Kennell and Associates Inc., in Raleigh, N.C. He can be reached at *ghileman@kennellinc.com*