Comparison of Strategies for Validating Binary Logistic Regression Models

Frank E Harrell Jr
Division of Biostatistics and Epidemiology
Department of Health Evaluation Sciences
University of Virginia School of Medicine
12 March, 1998

fharrell@virginia.edu

Simulation Method

For each of 200 simulations generate a training sample of 200 observations with p predictors (p=15 or 30) and a binary reponse. The predictors are independently U(-0.5,0.5). The response is sampled so as to follow a logistic model where the intercept is zero and all regression coefficients equal 0.5 (which is admittedly not very realistic). The "gold standard" is the predictive ability of the fitted model on a test sample containing 10,000 observations generated from the same population model.

Validation Methods

For each of the 200 training and validation samples several validation methods were employed to estimate how the training sample model predicts responses in the 10,000 observations. These validation methods involving fitting 40 or 200 models per training sample.

g-fold Cross-validation was done using the command validate(f, method='cross', B=4 or B=10) in the Design library for S-Plus. This was repeated 4, 10, 20, or 50 times and averaged using e.g.

```
repeated.vals <- function(fit, method, B, r) {
        z <- 0
        r <- max(r, 1)
        for(i in 1:r) z <- z + validate(fit,method=method,B=B)[c("Dyx",
                "Intercept","Slope","D","U","Q"),]
        z/r
}
```

For bootstrap methods validate(f, method='boot' or '632', B=40 or B=200) was used. method='632' does Efron's ".632" method, labeled "632a" in the output. An ad-hoc modification of the .632 method, "632b" was also done. Here a "bias-corrected" index of accuracy is simply the index evaluated in the observation omitted from the bootstrap re-sample.

The "gold standard" external validations were done using the val.prob function in the Design library. The one-page simulation program is available upon request.

I didn't run all combinations of validation methods and number of predictors, and I only used one sample size (200).

Indexes of Predictive Accuracy

Dxy: Somers' rank correlation between predicted probability that Y=1 vs. the binary Y values. This equals 2(C-0.5) where C is the "ROC Area" or concordance probability.

Q: Logarithmic accuracy score - a scaled version of the log-likelihood achieved by the predictive model

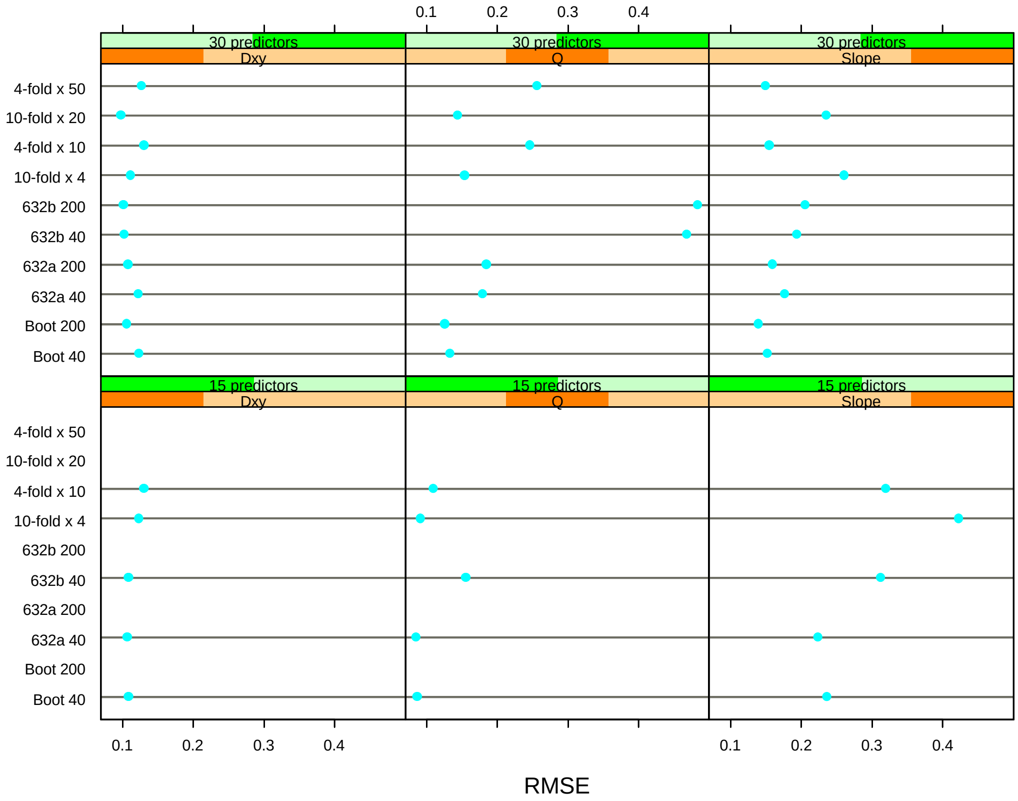
Slope: calibration slope (slope of predicted log odds vs. true log odds)

Measure of Accuracy of Validation Estimates

Root mean squared error of estimates (e.g., of Dxy from the bootstrap on the 200 observations) against the "gold standard" (e.g., Dxy for the fitted 200-observation model achieved in the 10,000 observations).

Summary of Results

This is best seen from the multi-way dot chart produced by S-Plus below.



Some general conclusions from this limited simulation experiment:

1. For p=30 predictors, 40 bootstrap samples was almost as good as 200.
2. The ordinary bootstrap is as good as the 0.632 bootstrap (better for Q when p=30)
3. The 632b method not surprisingly didn't work very well.
4. Ordinary bootstrap is better than or equal to all cross-validation strategies tried
5. 10-fold cross-val repeated 4 times was better then 4-fold cv 10 times except for stimating calibration slope
   10-fold x 20 times was better than 4-fold x 50 except for slope
6. 10 fold x 20 is not much better than 10 fold x 4
7. 4-fold x 50 is not much better than 4-fold x 10
8. Except for slope, 10-fold is better than 4-fold cv
9. The relative comparison of validation methods depends on general on what index you are validating

Apparently, to estimate the calibration slope (shrinkage factor) larger validation (hold-out) samples are needed.

It would be useful to try Efron and Tibshirani's new 0.632+ estimator.

Jim Patrie in our group is running simulations similar to these for R^2 from ordinary regression models, and so far the bootstrap is performing well.

This project was funded by National Center for Health Statistics Order No. 000930153.
Work completed 10Oct91.