Note on Shrinkage: Is the Squared Validity Coefficient always smaller than the R-Squared from the Original Sample?

According to Osborne, 2000 (http://pareonline.net/getvn.asp?v=7&n=2),

"To perform cross-validation, a researcher will either gather two large samples, or one very large sample which will be split into two samples via random selection procedures. The prediction equation is created in the first sample. That equation is then used to create predicted scores for the members of the second sample. The predicted scores are then correlated with the observed scores on the dependent variable **($\mathbf{r_{yy'}}$)**. This is called the *cross-validity coefficient*. The difference between the original R-squared and $\mathbf{r_{yy'}^2}$ is the shrinkage. The smaller the shrinkage, the more confidence we can have in the generalizability of the equation."

So, it is usually the case that the original R-squared is larger than the squared cross-validity coefficient, and the interest is how much shrinkage has occurred. However, the point of this note is that it is possible that the squared cross-validity coefficient is actually larger than the original R-squared.

Suppose you only have one predictor (X). Then, the regression equation from the derivation sample is $Y_1^{'} = a_1 + b_1 X_1$. Here, the subscript 1 indicates Sample 1 (derivation sample). Using those a_1 and b_1, one can obtain $Y_{cv}^{'} = a_1 + b_1 X_2$, where $Y_{cv}^{'}$ is the predicated score for the cross validation sample (Sample 2) based on the regression coefficients from the derivation sample. Now if you correlate Y_2 (observed Y) from the cross validation sample with $Y_{cv}^{'}$, you will get the cross validity coefficient. However, actually, $r_{Y_2,Y_{cv}^{'}}$ (cross validation coefficient) = r_{Y_2,X_2} (Square root of R-squared from the validation sample). It is because the correlation will not change when one of the variables is linearly transformed. $Y_{cv}^{'}$ is a linear transformation of X_2 ($Y_{cv}^{'} = a_1 + b_1 X_2$). Therefore, you are comparing the R-squared from the derivation sample with the R-squared from the validation sample. So it is just as likely to get a larger R-squared from the validation sample.

Of course, this is less likely when you move on to multiple regression. You will start to see that the squared cross validation coefficient is smaller (shrunk) than the original R-squared. In the example above by Osborne, he reports the original R-squared of .55 and the squared cross validation coefficient of .53, thus the shrinkage of 2%. He had 4 predictors in his model.

So, the point of this note is that you should not get surprised if your squared cross validation coefficient is actually bigger (so "expansion" instead of "shrinkage) than the original R-squared especially in the case of simple regression.

For those who are interested in shrinkage, the following classic references are useful:

Crueton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), Educational Measurement (pp. 621 – 692). Washington, DC: American Council on Education. (See p. 692)

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley. (See pp. 284 – 293).

By Chris Oshima 2011.