Predictive R-Squared

Dean P. Foster[*] Robert A. Stine[*]

Department of Statistics

The Wharton School of the University of Pennsylvania

Philadelphia, PA 19104-6340

March 7, 2014

Abstract

Fitting large regression models has become common. The R^2 statistic is perhaps the most popular summary of a regression, and its faults are well-known to all but the most casual users. Adjusted R^2, or $\overline{R}^2$, remedies the omission of degrees of freedom from R^2, but estimates a quantity most users would probably find unnatural. A further, small adjustment produces the predictive R^2 that we define in this short note. Written $\overline{\overline{R}}^2$, predictive R^2 estimates the variation explained when predicting new data rather than variation in the data used to estimate the fit. As the similarity of the name implies, predictive R^2 is nearly the same as the predicted R^2 reported by Minitab, only differing in the simplicity of its calculation. Whereas predicted R^2 performs implicit leave-one-out cross-validation, we use a simple approximation that avoids both cross-validation and the computation of leverages.

Key Phrases: cross-validation, leverage

[*] Research supported by NSF grant 1106743

Table 1: *Comparison of three versions of the r-squared statistic for two large regression models.*

Context		R-Squared			10-fold CV
Features	p	R^2	$\overline{R}^2$	$\overline{\overline{R}}^2$	PMSE
Word frequencies	1,000	0.671	0.620	0.550	6.55
Principal components	500	0.639	0.613	0.582	6.17

1 Introduction

Fitting large regression models has become common in many applications of statistics. As an example, the two, large regression models summarized in Table 1 predict the prices of 7,384 real estate properties listed in Chicago. The regressors are features constructed directly from the text of these listings as described in Foster, Liberman and Stine (2014). If we compare the models using adjusted R^2, we would expect them to be similarly predictive. The first model uses 1,000 features that count the frequency of the most common words in the text of the listings. This large model explains about two-thirds of the variation in prices, with $\overline{R}^2 = 0.62$. The second model uses 500 features that are essentially principal components of the frequency counts used in the larger model. With half as many predictors, this model achieves almost the same adjusted R-squared ($\overline{R}^2 = 0.61$). Though inferior by these measures, the smaller second model is actually more predictive. The predictive R^2 denoted $\overline{\overline{R}}^2$ in Table 1 suggests we should not be surprised. It indicates that the smaller model ought to be slightly more predictive than the larger model — and it is.

To compare the predictive accuracy of these models, we performed 10-fold cross-validation. Leave-one-out cross-validation has well-known limitations (*e.g.*with nonlinearity and model selection Efron, 1987; Shao, 1993), so we use 10-fold cross-validation to estimate the actual out-of-sample predictive error. Rather than rely on one split into ten folds, we repeated the splitting process 10 times. Table 1 shows the prediction mean squared error (PMSE) over the 100 estimated models.

2 Predictive R-squared

To see the problem with adjusted R-squared, it helps to think about what happens for very large samples. Consider the usual linear regression model for which

$$Y = X\,\beta + \epsilon \; , \tag{1}$$

where Y is an n vector, X is a $n \times p$ matrix with a leading column of 1s followed by $p-1$ explanatory variables (*a.k.a.*, regressors, features), and ϵ is independent, mean zero, homoscedastic noise with variance σ^2. Let x_i denote the rows of X. When estimated, the expression for adjusted R-squared is

$$\overline{R}^2 = 1 - \frac{\sum(Y_i - x_i'\hat{\beta})^2/(n-p)}{\sum(Y_i - \overline{Y})^2/(n-1)} = 1 - \frac{\text{RSS}/(n-p)}{\text{TSS}/(n-1)} \tag{2}$$

The numerator in the fraction converges to σ^2.

To appreciate the flaw in $\overline{R}^2$, it is helpful to think of σ^2 as a measure of a prediction error rather than error variance. In particular, we can write σ^2 as

$$\sigma^2 = \mathbb{E}\,(Y_\nu - x_\nu \beta)^2 \; , \tag{3}$$

where the expectation is over pairs (Y_ν, x_ν) that form an independent observation that is consistent with the model (1). In this sense, σ^2 is the expected squared error of predicting a new observation using a model built with infinite data (so that β is known). Consequently, $\overline{R}^2$ ignores the effects of estimation error in $\hat{\beta}$. $\overline{R}^2$ adjusts for the degrees of freedom in the residual sum-of-squares, but that leaves an estimate of σ^2 rather than the error when predicting a new observation with a fitted model.

A variety of methods do adjust for the error when predicting a new observation. These statistics estimate $\mathbb{E}\,(Y_\nu - x_\nu'\hat{\beta})^2$, the error when predicting a new observation with a fitted model. Such statistics have a long history and include Mallows C_p (Mallows, 1973), Akaike's Final Prediction Error (Akaike, 1969), and the prediction sum-of-squares PRESS (Allen, 1974). The most relevant of these for revising the adjusted R-squared is PRESS, which is defined as

$$\text{PRESS} = \sum(Y_i - x_i'\hat{\beta}_{(-i)})^2 = \sum e^2_{(-i)} \; , \tag{4}$$

where $\hat{\beta}_{(-i)}$ is the estimate of β obtained when the ith observation is omitted from the model. PRESS is easily computed without explicitly refitting n regressions by using the well-known expression (*e.g.* Belsley, Kuh and Welsch, 1980; Tarpey, 2000)

$$\hat\beta - \hat\beta_{(-i)} = (X'X)^{-1} x_i' e_i/(1-h_i) \ , \tag{5}$$

where $e_i = Y_i - x_i'\hat\beta$ is the ith residual and $h_i = x_i'(X'X)^{-1}x_i$ is the leverage. Minitab uses PRESS to define a different adjustment to R^2 called the predicted R-squared statistic,

$$\text{Predicted } R^2 = 1 - \frac{\text{PRESS}}{\text{TSS}} \ . \tag{6}$$

Our predictive R-squared is similar but replaces PRESS with an approximation that avoids computing leverages or manipulating large matrices. Using the expression (5), we can write the difference between the ith residual and the leave-one-out error $e_{(-i)}$ defined in (4):

$$e_{(-i)} = Y_i - x_i'\hat\beta_{(-i)} = e_i + x_i'(\hat\beta - \hat\beta_{(-i)}) = e_i/(1-h_i) \ . \tag{7}$$

The leverages are the diagonal elements of the hat matrix $X(X'X)^{-1}X'$ and sum to its rank, p. Predictive R-squared replaces PRESS by setting h_i to the average p/n, obtaining

$$\sum e^2_{(-i)} = \sum \left(\frac{1}{1-h_i}\right)^2 e_i^2 \approx \sum \left(\frac{1}{1-p/n}\right)^2 e_i^2 = \left(\frac{n}{n-p}\right)^2 \text{RSS} \tag{8}$$

This approximation will be accurate so long as no observation has excessively large leverage. A further approximation produces $\overline{\overline{R}}^2$ which more closely resembles $\overline{R}^2$. Simply approximate $((n-p)/n)^2$ by $(n-2p)/n$ (so long as $p \ll n$), and we arrive at the definition of predictive R-squared

$$\overline{\overline{R}}^2 = 1 - \frac{\text{RSS}/(n-2p)}{\text{TSS}/(n-1)} \ . \tag{9}$$

This expression closely resembles equation (2) for $\overline{R}^2$, but replaces the divisor in the numerator $n-p$ by $n-2p$.

3 Discussion

Statistics such as C_p and PRESS that capture the effects of estimation on the predictions from regression have been around for years, but R^2 remains the most common assessment of a model. Adjusted R^2 is an improvement and is routinely computed by software. Our hope is that by building a very similar statistic that more completely adjusts for estimation, users will be less surprised when the predictions from a model fail to meet the expectations from a fitted equation.

Of course, $\overline{\overline{R}}^2$ is only a partial adjustment. It is appropriate when one fits a model to data that was not used to pick that model. Like the other statistics discussed here, $\overline{\overline{R}}^2$ does not adjust for the influences of model selection. Such adjustments would depend on the nature of the selection process and would likely be far more dramatic than the adjustment that produces $\overline{\overline{R}}^2$.

References

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institude of Statistical Mathematics* **21** 243–247.

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **16** 125–127.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* Wiley, New York.

Efron, B. (1987). *The Jackknife, the Bootstrap, and Other Resampling Plans.* CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadephia, PA.

Foster, D. P., Liberman, M. and Stine, R. A. (2014). Featurizing text: Converting text into predictors for regression analysis. Dept of Statistics, University of Pennsylvania.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the Amer. Statist. Assoc.* **88** 486–494.

Tarpey, T. (2000). A note on the prediction sum of squares statistic for restricted least squares. *The American Statistician* **54** 116–118.