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 6, 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, written $\overline{R}^2$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 $\overline{\overline{R}}^2$, the predictive R^2 estimates the variation explained when predicting new data. As the similarity of the name implies, predictive R^2 is nearly the same as 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 approximate 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.*

Model	Features	R^2	$\overline{R}^2$	$\overline{\overline{R}}^2$
Word frequencies	1,000	0.671	0.620	0.550
Principal components	500	0.655	0.630	0.600

1 Introduction

Fitting large regression models has become common in many applications of statistics. As an example, the regression models summarized in Table 1 predict the prices of 7,384 real estate properties listed in Chicago. The regressors are features constructed from only the text of these listings as described in(**?**). The 1,000 features in the first model count the frequency of the most common 1,000 words in the text of all the listings. This large model explains R^2 = percent of the variation in prices, with adjusted R^2 =. The second model uses 500 features that are essentially principal components of the frequency counts used in the larger model. The second model uses half as many predictors but explains almost as much variation: R^2 = and $\overline{R}^2$ =.