

Michael S. Lewis-Beck  
Department of Government, 224 Littauer  
Harvard University  
Cambridge, MA 02138  
617-495-8280 (phone)  
mlewis-beck@latte.harvard.edu  
Word Count: 1,373

**R-squared.** The  $R$ -squared ( $R^2$ ) measures the explanatory or predictive power of a REGRESSION MODEL. It is a goodness-of-fit measure, indicating how well the linear regression equation fits the data.

In regression analysis, it is important to evaluate the performance of the estimated regression equation. To what extent does it account for the phenomenon under study? The  $R^2$  is the leading performance measure for a SIMPLE or MULTIPLE REGRESSION model. Suppose a policy analyst is studying public school expenditures in the 50 American states. The analyst posits a simple regression model,  $Y = a + bX + e$ , where  $Y$  is the DEPENDENT VARIABLE of per pupil public school expenditures (in thousands of dollars) in each state in the year 2000,  $X$  is the INDEPENDENT VARIABLE of urbanization (percentage of population in cities larger than 25,000, as of the 2000 census), and  $e$  is the error term. The model argues that state public school outlays are accounted for, in part, by urbanization. ORDINARY LEAST SQUARES (OLS) estimation yields, hypothetically, the following results:  $Y = 2.11 + .60X + e$ , suggesting that for every additional percentage of urban population, a state expects to spend \$600 more. The  $R^2$ , or coefficient of determination, for the equation is .42, which says that 42% of the variation in state public school expenditures is explained, or at least predicted, by urbanization.

The  $R^2$  shows the gain in predicting  $Y$  knowing  $X$ , as opposed to not knowing  $X$ . Suppose that the analyst knows the scores on  $Y$ , but not the states they are attached to,

and tries to predict school expenditures state by state. The best guess, the one that minimizes the error, would always be the average of  $Y$ , that is,  $Y_m$ . This guess will be way off for most states, with a great distance from the observed expenditure score to the average expenditure score, that is,  $(Y - Y_m)$ . Adding all these distances together (after squaring to overcome the canceling out of the plus and minus signs), they represent the total prediction error not knowing  $X$ ; that is, total sum of squared deviations (TSS) =  $\Sigma (Y - Y_m)^2$ .

Regression analysis promises reduction of this prediction error, through knowledge of  $X$  and its linear relationship to  $Y$ . For example, knowing a state's score on  $X$  is 30% urban, the prediction,  $\hat{Y}$ , is 20.11 ( $\hat{Y} = 2.11 + 18.10 = 20.11$ ), not  $Y_m$ . The distance from the predicted  $Y$  to the mean  $Y$ ,  $(\hat{Y} - Y_m)$ , is the improvement over the baseline prediction made possible by knowing  $X$ . This distance, squared and summed for all observations, is the reduction in prediction error attributable to application of the regression line; that is, regression sum of squared deviations (RSS) =  $\Sigma (\hat{Y} - Y_m)^2$ .

Unless the regression line predicts each case perfectly, some error will still remain, the distance from the observed  $Y$  and the predicted  $Y$ . These distances, squared then summed, represent the variation in  $Y$  that remains unexplained; that is, error sum of squared deviations (ESS) =  $\Sigma (Y - \hat{Y})^2$ .

Total variation in the dependent variable, TSS, thus has two unique parts, RSS accounted for by the regression and ESS not accounted for by the regression. The  $R^2$  reflects the regression portion as a share of the total,  $RSS/TSS = R^2$ . The statistic ranges from 1.0, when all variation is accounted for, to .00, when no variation is accounted for. With real-world data, the  $R^2$  seldom reaches these extreme values. In the above example,

we may say that 42% of the variation in public school expenditures is explained by urbanization. It is important to remember that this explanation may be more “statistical” than “causal.” It could be that urbanization helps predict public school expenditure but does not really explain it in a theoretical sense. In such cases, it is more cautious, and perhaps more correct, to say that  $X$  merely “accounts for” so much variation in  $Y$ . In the bivariate regression case, the  $R^2 =$  the CORRELATION COEFFICIENT squared (the  $r^2$ ).

With a multiple regression equation, the  $R^2$  is referred to as the coefficient of multiple determination. Again, it measures the linear goodness of fit of the model, ranging from 1.0 to .00. For example, when the multiple regression model,  $Y = a + bX + cZ + e$ , is estimated with OLS, the  $R^2$  indicates the fit of a plane to the points in a three-dimensional space. For more than two independent variables, a hyper-plane is fitted. The MULTIPLE CORRELATION COEFFICIENT, symbolized by  $R$ , is the square root of the  $R^2$  and indicates the correlation of the ensemble of independent variables with the dependent variable.

When an independent variable is added to a regression model, the  $R^2$  always increases. Some of this increase is due to chance, for adding independent variables uses DEGREES OF FREEDOM. Especially as the number of independent variables approaches the sample size, the  $R^2$  estimate will exaggerate the real fit of the model to the data. Most analysts report the ADJUSTED  $R$ -SQUARED for a multiple regression model, along with, or instead of, the  $R^2$  itself. The adjusted  $R^2$  will be smaller in magnitude than the  $R^2$ , but seldom much smaller. Unlike the  $R^2$ , the adjusted  $R^2$  can show decreases as independent variables are added to a model. (Rarely, the adjusted  $R^2$  can be negative, when the

unadjusted fit is extremely low in the first place. The unadjusted  $R^2$  has the possibility of being negative only if the constant term is forced to be zero.)

Analysts generally prefer a high  $R^2$ , sensing that a better fit means a better explanation. However, that may or may not be true.  $R^2$  maximization should not be an end in itself. Blind inclusion into a regression model of the variables that have a high correlation with the dependent variable may yield a high  $R^2$ . But when that model is estimated on a subsequent sample, that high fit will likely plummet, since much of it came from chance variation in the first sample. A moderate, even low,  $R^2$  is not necessarily bad. Certain kinds of data, such as public opinion surveys, do not generally admit of high fits. Regardless of data type, a low fit may simply mean that only a small portion of  $Y$  can be explained, the rest being random. There is a case where a low  $R^2$  may be a bad sign, the case of nonlinearity. If the relationship follows a curve rather than a straight line, then linear specification and the  $R^2$  assessment of model performance is inappropriate. Finally, a low  $R^2$  for any model suggests the model will predict poorly.

The  $R^2$  has been subjected to various criticisms. Some consider the baseline for prediction comparison –  $Y_m$  – to be naive. Usually, however, it is difficult to come up with a better alternative baseline. Similarly, some object that it only assesses the linearity of model fit. However, it is often hard to improve on the linear specification in practice. A further issue is whether the  $R^2$  of different models can be tested for STATISTICALLY SIGNIFICANT differences, to assess rival model specifications. This can be done if one model is possibly a nested subset of another larger model. Legitimate use of such significance tests depends partly on whether the  $R^2$  is held to be a descriptive or inferential statistic. A few analysts reject the  $R^2$  altogether, preferring the STANDARD

ERROR OF ESTIMATE OF  $Y$  (the SEE) as a measure of goodness of fit. The difficulty with using the SEE measure alone is that it has no conventional theoretical boundary at the upper end, for example, 1.0. As SEE gets larger, one knows there is more prediction error, but one does not know whether it has reached its maximum. In general, the two measures of fit should both be reported, since they say somewhat different things.

MICHAEL S. LEWIS-BECK

*References*

Lewis-Beck, Michael S., and Skalaban, Andrew. (1990). The  $R$ -squared: Some straight talk. *Political Analysis*. Vol. 2, pp. 153-170.

Pindyck, Robert S., and Rubinfeld, Daniel L. (1991). *Econometric Models and Economic Forecasts* (3rd ed.). New York: McGraw-Hill.

Weisberg, Sanford. (1985). *Applied Linear Regression*. New York: John Wiley.