WHAT DOES R² MEASURE?

Steve Thomson, University of Kentucky Computing Center

The coefficient of determination, squared multiple correlation coefficient, that is, $R^2$, is perhaps the most extensively used measure of goodness of fit for regression models. In my consulting, many people have asked questions like: "My $R^2$ was .6. That's good, isn't it?" or "My $R^2$ was only .05. That's bad, isn't it?". Thus many people seem to use high values of $R^2$ as an index of fit, concluding that if $R^2$ is above some value, they have "good" fit, but that small values indicate "bad" fit. In either case above, probably the best response is "Compared to what?".

One interpretation of $R^2$ is as the limit of the regression sum of squares divided by the limit of the total sum of squares. Or equivalently, it is an estimate of the amount of variation in observation means compared to the variation in means plus the variation in residual error. This can be small even for models with small error, or large for models with large error.

I. Review:

The general linear model can be interpreted as investigating how well an observed nx1 vector, $y$, can be represented in some lower dimensional subspace spanned by the columns of the matrix of regressors. (Recall that the span of a set of vectors is the set of all linear combinations of the vectors, i.e. a subspace) The coefficients of the vectors determining the points in the subspace are the parameters. The least squares solution for the parameters are the parameter values corresponding to the point in the subspace $\hat{y}$ closest to $y$. The various sums of squares in ANOVA can be represented as (squared) Euclidian distances between subspaces spanned by different groups of columns in the regressor matrix.

![Figure 1. Geometry of ANOVA](image)

(Note that distances are square roots of indicated ssq's.)

The total sum of squares, SST, is the (squared) distance from $y$ to the origin. The error sum of squares, SSE, is the (squared) distance from $y$ to $\hat{y}$, in the span of A and B. The model sum of squares, SS(A,B), is the (squared) distance from $\hat{y}$ to the origin. By the Pythagorean theorem,

$$SST = SS(A,B) + SSE.$$  

So heuristically, a good model has a large SS(A,B) compared to SST. This seems to justify the use of their ratio as an index of fit, namely:

$$R^2 = \frac{SS(A,B)}{SST}.$$  

The model sum of squares, SS(A,B), could also be partitioned into sequential sums of squares. One possible partition would be:

$$SS(A,B) = SS(A) + SS(B|A).$$

where both terms on the right are squared distances, and thus positive. The SS(B|A) is the sum of squares for B "adjusted for", or "orthogonalized to" A. So comparing the $R^2$ for the submodel with only A's as regressors to the $R^2$ for the complete model with A's and B's:

$$R^2_A = SS(A) \leq SS(A) + SS(B|A) = R^2$$

Clearly $R^2$ has to increase as variables are added to the model (or at least not decrease). Also, note from the geometry that $0 \leq R^2 \leq 1$.

If the model includes an intercept or constant term, it is particularly easy to adjust, or orthogonalize, to this constant, merely subtract the column mean from each column. Most analyses are performed on such "corrected" models. One last simple observation from the geometry is that there are an uncountably infinite number of subspaces with $R^2 = 1$, namely, any subspace that includes $y$.

Other results on $R^2$:

1. If true replicates are added to the data, so that more than one $y$ value occurs at each $x$, regressors value, then $R^2$ will often decrease (e.g. Healy, 1984, Draper, 1984). In fact, when there are true replicates its achievable upper bound will be less than 1.0. In practice, the results for close-replicates are often similar.

2. For $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i$, with the $\epsilon_i$ independent, homogeneous normal random variables with mean 0, when $\beta_i = \beta_2 = \ldots = \beta_p = 0$, it is easy to show that then the expected value of $R^2$, using corrected sums of squares, is $\frac{p-1}{n-1}$ (e.g. Crocker, 1972, Seber, 1977, pg 115). A similar result holds for $R^2$ using uncorrected sums of squares ($\frac{p}{n}$).

In fact, under these assumptions, $R^2$ follows a beta distribution.

It would seem that a good fit index for comparing completely different regression equations should be optimized by the least squares equation, should be free of the measurement units, and should be large for models with small residual error and small for models with a large residual error. Finally, the regressor variables are usually considered as fixed. So a fit index should be independent of their actual values. $R^2$ satisfies the first two criteria, but fails on the last two.

II. An Alternative Definition:

An alternative definition of $R^2$ that has been recommended (e.g. Draper & Smith, 1981) is

$$R^2 = \frac{SST - SSE}{SST}.$$  

where $SST$ is the total sum of squares, $SSE$ is the error sum of squares, and $SS(A,B)$ is the model sum of squares for the model with only A's and B's. This definition is particularly useful when comparing models with different numbers of parameters. It adjusts for the number of parameters in the model, so that it is not biased towards models with more parameters. This is particularly important when comparing models with very different numbers of parameters. The $R^2$ value will be lower for models with more parameters, reflecting the fact that the error sum of squares will be smaller due to the additional parameters.

The $R^2$ value can also be interpreted as the proportion of variance in the response variable $y$ that is explained by the model. This is particularly useful in regression analysis, where the goal is to find a model that explains as much of the variance in the response variable as possible.

In summary, $R^2$ is a useful measure of the goodness of fit of a regression model, and its interpretation depends on the context in which it is used. It can be used to compare different models, to identify important predictors, and to evaluate the overall fit of a model.
\[ R^2 = \text{squared correlation between regressand (response \( y \)) and predicted value (\( \hat{y} \)).} \]

It is well known that for a linear model "corrected" for the intercept or constant term, estimated by ordinary least squares (so corrected sums of squares are used in the first definition), the two definitions are equivalent. For uncorrected models, \( R^2 \) is easily computed in SAS PROC REG or GLM by OUTPUTing the predicted values and computing correlations in PROC CORR, and finally squaring the result.

To illustrate a possible problem with the second definition, suppose below that \( o \)'s denote observed values, and \( p \)'s denote predicted values from the least squares equation, from a simple regression with no intercept. Then using the second definition, \( R^2 = 1 \).

\[
\begin{array}{ccc}
& o & p \\
0 & 0 & 0 \\
p & p & 0 \\
p & p & 0 \\
\end{array}
\]

Figure 2.

Yet the predicted values, \( p \)'s, are most certainly not doing a good job of predicting the observed values. Of course, in this case, any linear function of the observed \( x \)'s will give an \( R^2 = 1 \). It is slightly more complicated for multiple regression problems, but generally, even if \( R^2 = 1 \), there is no requirement that the least squares hyperplane must be "close" to \( y \). That is, for any \( b \), and nonzero \( a \),

\[
(Corr(y, y))^2 = (Corr(a + b, y))^2.
\]

Since we could generate a hyperplane passing through zero and any specified value, that suggests that unless the possible choices of coefficients is restricted in some way, \( R^2 \) is not explicitly a good measure of "fit".

III. Limiting Behavior of \( R^2 \):

The definition of \( R^2 \) used in SAS procedures seems to make more sense, \( R^2 = \frac{SSR}{SST} \), by default, using corrected sums of squares if the model has an intercept, uncorrected if the model has no intercept. But that particular ratio is exactly what \( R^2 \) is measuring, not necessarily the quality of fit.

Suppose a general linear model is appropriate, i.e. the \( n \times 1 \) regressand vector, \( y \), is appropriately modeled as \( y = X \beta + \epsilon \), where the \( \epsilon \)'s are stochastically independent with mean 0 and variance \( \sigma^2 \). \( X \) is the \( n \times p \) matrix of values of the regressor variables. The usual (strong) assumption for asymptotic regression results is that \( X'X \sim n \rightarrow B \), a positive definite matrix. This implies, with \( \text{plim} \) denoting limit in probability, that \( \text{plim} \frac{X'X}{n} = B \), (since

\[
\text{Var} \left( \frac{X'X}{n} \right) = \frac{\sigma^2}{n} X'X
\]

1. Uncorrected models:

Since it is slightly simpler, first consider a model not adjusted for the intercept. Then, since sums of squares are uncorrected, writing \( P_X \) as the projection onto the subspace spanned by the columns of \( X \), \( P_X = X(X'X)^{-1} X' \).

\[
R^2 = \frac{SSR}{SST} = \frac{\text{y'P}_x \text{y}}{\text{y'y}} = \frac{\text{y'X(X'X)^{-1}X'y}}{\text{y'y}}
\]

\[
= \frac{(p'X'X + 2p'X'e + e'e(X'X)^{-1}X'e)/n}{(p'X'X + e'e)/n}
\]

So as \( n \) gets large, \( \text{plim} R^2 = \frac{\beta' \beta}{\beta' \beta + \sigma^2} \).

If the \( X \)'s were stochastic and regular, \( \beta \) would be the second moment matrix about zero of the regressors. So in a sense, \( R^2 \) is a measure of variation in the \( X \)'s, \( \beta \), weighted by the true coefficients, \( \beta \). Another way to look at this, is if \( \mu_i \) is the mean of the \( i \)-th observation, expressed as a function of the regressors, then

\[
\text{plim} \frac{\sum_i \mu_i^2}{n} = \frac{\beta' \beta}{\beta' \beta + \sigma^2},
\]

i.e. the limiting average value of the individual squared means. The expression on the right hand side has to be interpreted carefully however. The \( \mu_i \)'s are those predicted from the regressors. This means, for example, that for a model with no intercept, if a constant is added to the response, so that observations are moved further from zero, \( R^2 \) will actually tend to decrease, because the model is not adjusted for this new parameter.

Further, note that, asymptotically at least, \( \text{plim} R^2 \) is close to 1 if \( \sigma^2 \) is small (i.e. good fit), or if \( \beta' \beta \) is large (i.e. large mean values). Also the contribution of an individual regressor variable to \( R^2 \) will be small when both the long run mean value of the squared regressor variable is small and when its associated "beta-weight" is small.

2. Corrected models:

One can show that if \( \frac{X'X}{n} \rightarrow B \), a positive definite matrix, then for any orthogonal projection matrix \( P \), there is a \( B' \) so that \( \frac{X'PX}{n} \rightarrow B' \).

For convenience, write the orthogonal projection matrices \( P_X = X(X'X)^{-1}X' \) and for \( j \) denoting the \( n \times 1 \) vector of \( 1 \)'s, \( P_1 = \frac{1_n}{n} \), a \( n \times n \) matrix with all elements \( 1/n \). Since \( j \) is a column of \( X \), there is an \( a \) so that \( j' = Xa \). Also then \( P_1 P_1 = P_1 = P_1 P_X \).

Thus,

\[
R^2 = \frac{SSR}{SST} = \frac{(P_X'y - P_1'y)(P_X'y - P_1'y)}{y'y - P_1'y}
\]

By the law of large numbers, \( \text{plim} \frac{P_1'y}{n} = 0 \), so

\[
\text{plim} \frac{\beta' \beta}{\beta' \beta + \sigma^2} = \frac{\beta' \beta}{\beta' \beta + \sigma^2}
\]

1247
As $n$ gets large, with $X^T X / n \to B_0$, say, then,

$$\text{plim} \frac{X^T Y}{n} = \text{plim} \left( \frac{X^T X \beta_0 + X^T Y \epsilon_0}{n} \right) = B_0 \beta_0.$$

Thus, $\text{plim} R^2 = \frac{\beta_0^T B_0 - B_0 \beta_0}{\beta_0^T B_0 + \sigma^2} = \frac{\beta_0^T B_0}{\beta_0^T B_0 + \sigma^2}$.

where, now, $B'$ is the limiting, centered matrix of regression coefficients, corresponding to a moment matrix of the regressors. Thus $R^2$ will be "large" if $\sigma^2$ is small (i.e., good fit), or if $\beta_0^T B_0$ is large (i.e., large variation among mean values).

Or again, with correct specification, $R^2$ can also be interpreted as

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}(n))^2,$$

basically, the corrected regression sum of squares.

IV. So what does $R^2$ measure?

For a correctly specified model, $R^2$ tends toward the inverse of $1 - \sigma^2$, where, again, $B$ is analogous to a sample covariance matrix of regressors for the intercept case, and a moment matrix for the no intercept case. In either case, the $\beta_0^T B_0$ is a reasonable "regression" sums of squares, but is not much of an indicator of fit. If one has perfect fit, $\sigma^2 = 0$, and $R^2$ will be 1. Of course, $R^2$ will also tend to 1 as the variation in observation means gets large.

In practice, this means that if two experimenters choose equal sample sizes from the same process, the one who chooses his regressor variables to have more variation will tend to have a higher $R^2$. For example, in an economic context, suppose some process remains true over the years and satisfies the standard regression assumptions. Then an economist who uses monthly data, since generally the yearly data will have more variation. Similarly, suppose two nutritionists are modeling weight gain as a function of calories and prior weight. All else being equal the nutritionist who chooses his sample to have more variation in calories and prior weight will tend to have a higher $R^2$. In an educational context, with a true model, a researcher who samples across schools in a state could be expected to have a larger $R^2$ than a researcher who samples schools in a certain county.

Actually this is not surprising. Under the usual assumptions, the variance of the ordinary least squares estimates of the $\beta$'s, is $\sigma^2 (X^T X)^{-1}$. So large variation in the $X$'s tends to reduce the variance of the estimates of the $\beta$'s. Thus more variability in the $X$'s increases efficiency. But $R^2$ is apparently not usually interpreted as a measure of efficiency, but rather fit. However, $R^2$ is not quite a measure of efficiency either. For any set of regressors, $R^2$ would also tend to increase as the $\beta$'s get larger, a change which would have no effect on the precision of the estimates.

The fact that $R^2$ is not quite a measure of fit is not surprising. It is just the squared correlation between the response variable $y$ and a linear combination of the regressor variables. Correlation measures how much variables vary (linearly) together. So $R^2$ has to depend on the regressors as much as on the response variable.

V. So again, what does $R^2$ measure?

Even in a correctly specified model, $R^2$ is primarily an estimate of the relative size of the variation in observation means and the population variance. Within a single set of possible regressors, it could be a useful index of fit to compare various submodels. But, models with small error, and thus good fit, can be a high $R^2$ or a small $R^2$ depending on the observation means, i.e., on the values of the regression variables. So if $R^2$ is large, it does not necessarily indicate "good" fit, while $R^2$ small does not necessarily indicate "bad" fit.

SAS is the registered trademark of SAS Institute, Inc., Cary, NC, USA.

For further information contact:

Steve Thomson
University of Kentucky Computing Center
128 McVey Hall
Lexington, KY 40506-0045
REFERENCES


