What's a good value for R-squared?

A low R2 might well indicate that variables are poorly measured, that important variables have been excluded, or that the model has been miss-specified in other ways (e.g. effects are non-linear or non-additive). But, this does suggest that R2 should generally be of only secondary interest to us. *If a correctly specified model with well-measured variables produces a small R2, then so be it. We should be much more interested in the determinants of R2 than in R2 itself. And, if we are going to make comparisons of R2, we should make sure we are doing so correctly*. Rather than just saying R2 differs across groups, times, or variables, we should try to explain why it differs (and we should definitely avoid misleading statements, such as those which erroneously imply that a larger R2 is the result of larger structural effects.)

Despite the fact that R-squared is a unitless statistic, there is no absolute standard for what is a "good" value. A regression model fitted to non-stationary time series data can have an R-squared of 99% and yet be inferior to a simple random walk model. On the other hand, a regression model fitted to stationarized time series data might have an R-squared of 10%-20% and be considered quite good. When working with stationary stock return data, R-squared values as low as 5% might even be considered significant--if they hold up out-of-sample!
(www.nd.edu/~rwilliam/stats2/l72.pdf, Date Access: 11 July 2007)

The question is often asked: "what's a good value for R-squared?" Sometimes the claim is even made: "a model is not useful unless its R-squared is at least x", where x may be some fraction greater than 50%. By this standard, the model we fitted to the differenced, deflated, and seasonally adjusted auto sales series is disappointing: its R-squared is less than 25%. So what IS a good value for R-squared? The correct answer to this question is polite laughter followed by: "That depends!"

The term R-squared refers to the *fraction of variance explained* by a model, but--what is the relevant variance that demands explanation? We have seen by now that there are many *transformations* that may be applied to a variable before it is used as a dependent variable in a regression model: deflation, logging, seasonal adjustment, differencing. All of these transformations will change the variance and may also change the *units* in which variance is measured. Deflation and logging may dramatically change the units of measurement, while seasonal adjustment and differencing generally reduce the variance significantly when properly applied. *Therefore, if the dependent variable in the regression model has already been transformed in some way, it is possible that much of the variance has already been "explained" merely by the choice of an appropriate transformation*. Seasonal adjustment obviously tries to explain the seasonal component of the original variance, while differencing tries to explain changes in the local mean of the series over time. With respect to which variance should R-squared be measured--that of the original series, the deflated series, the seasonally adjusted series, and/or the differenced series? This question does not always have a clear-cut answer, and as we will see below, there are usually several reference points that may be of interest in any particular case.

So, what is a good value for R-squared? It depends on how you measure it! If you measure it as a percentage of the variance of the "original" (e.g., deflated but otherwise untransformed) series, then a simple time series model may achieve an R-squared above 90%. On the other hand, if you measure R-squared as a percentage of a properly *stationarized* series, then an R-squared of 25% may be quite respectable. (In fact, an R-squared of 10% or even as little as 5% may be statistically significant in some applications, such as predicting stock returns.) If you calculate R-squared as a percentage of the variance in the errors of the *best time series model* that can be explained by adding exogenous regressors, you may be disillusioned at how small this percentage is! Here it was less than 4%, although this was technically a "statistically significant" reduction, since the coefficients of the additional regressors were significantly different from zero.

What value of R-squared should you report to your boss or client? If you used regression analysis, then to be perfectly candid you should of course include the R-squared for the regression model that was actually fitted--i.e., the fraction of the variance of the dependent variable that was explained--along with the other details of your regression analysis, somewhere in your report. However, if the original series is nonstationary, and if the main goal is to predict the *level* (rather than the change or the percent change) of the series, then it is perfectly appropriate to also report an "effective" R-squared calculated relative to the variance of the original series (deflated if appropriate), and this number may be the more important number for purposes of characterizing the predictive power of your model. In such cases, it will often be the case that most of the predictive power is derived from the history of the dependent variable (through lags, differences, and/or seasonal adjustment) rather than from exogenous variables. This is the reason why we spent some time studying the properties of time series models before tackling regression models.

What should never happen to you: Don't ever let yourself fall into the trap of fitting a regression model that has a respectable-looking R-squared but is actually very much inferior to a simple time series model. If the dependent variable in your model is a nonstationary time series, be sure that you do a comparison of error measures against an appropriate time series model.

(More details see http://www.duke.edu/~rnau/rsquared.htm Date Access: 11 July 2007, Associate Professor Robert F. Nau, Ph.D. University of California, Berkeley, 1981 (Expertise: Mathematical Modeling of Decision-Making)