Don’t Let R-Squared Rule You

Mark J. Anderson*

Introduction

R-squared (R^2), also known as the “coefficient of determination,” measures the proportion of the response difference (Δy) caused by the input factors (x’s). Experimenters seeking my advice often express alarm at R^2 results not at levels above 0.9 on par with published studies. One of my clients—a biochemist doing graduate research—deleted perfectly good data just to increase her R^2! After restoring these results, the model produced an R^2 above 0.5—very good by my experience as a process developer and reviewing results from hundreds of industrial experiments.

Much lower R-squareds can be statistically significant. For example, the Framingham Heart Study produced very useful insights on causes of high blood pressure from a model that generated an R^2 of 0.159. The study—still ongoing—encompasses 1000s of subjects, thus powering through the inherent variability in human health and leading to reductions in heart disease.

Happily, my specialty is design of experiments (DOE) for industrial R&D, thus making predictive modeling far easier than doing so for retrospective studies of people. Even so, R-squareds often fall below 0.5 due to a variety of reasons—poorly controlled processes, highly variable measurements and, most commonly, narrow factor ranges that generate a low signal-to-noise ratio, as I will now discuss.

How R^2 depends greatly on the factor range

As illustrated by Figure 1, assume you study the effect of one factor on a process. As shown by the bell-shaped curve, the amount of process variation (σ^2) is relatively high (but not atypical, unfortunately).

*Stat-Ease, Inc. Minneapolis, MN, USA
Corresponding author:
Mark J. Anderson, Stat-Ease, Inc, 6 Pine Tree Dr., Suite 245, Minneapolis, MN 55112, USA
Email: mark@statease.com

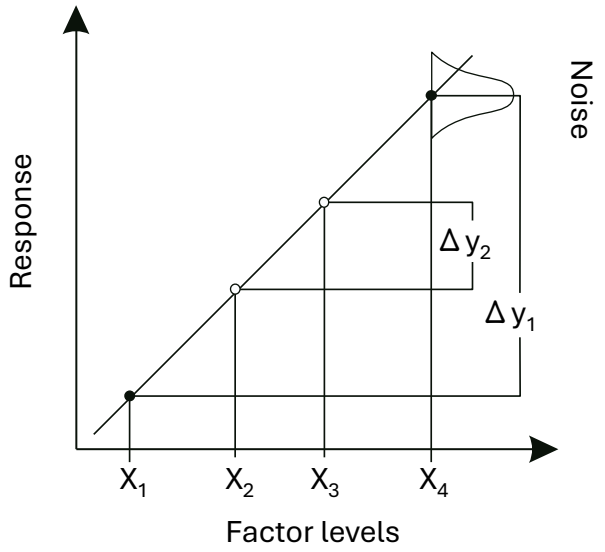

Figure 1: Effect of varying factor levels (x) on response (y)

The experiment is done to quantify the slope of the line, that is, the factor's impact on the response. Choosing levels at extremes of x_1 versus x_4 generates a large signal to noise ratio, thus making it easy to estimate the slope. In this case, the R^2 will be relatively high – perhaps approaching its theoretical maximum of one.

On the other hand, what if levels are tightened up to x_2 and x_3? This generates a smaller signal, thus making it more difficult to estimate the slope. You can overcome this weakness by running replicates. With enough replicates, the power will increase to a level that provides estimates of the slope at the same precision as in the first experiment, despite the narrower range of the factor. However, because the signal (Δy) is smaller relative to the noise (σ), R^2 will be smaller, no matter how many replicates are run!

A numerical example

To put these assertions about R^2 to the test, I set up a simulation that created a 38-run dataset, including several replicates, based on this model: $y = 10 + 2x$, $s = 2$ (standard deviation). See Figure 2 for the linear fit.

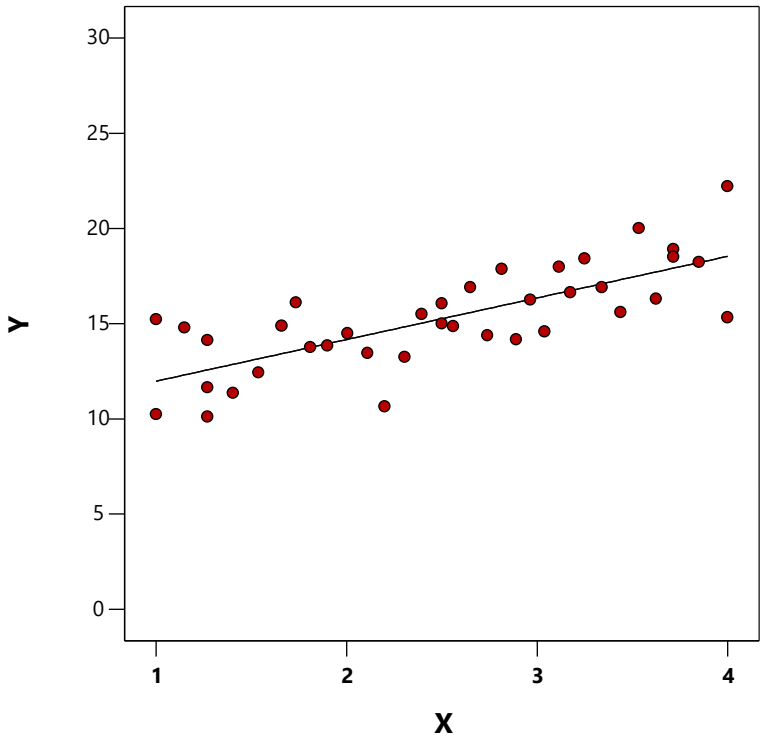

Figure 2: Fitting a line to data with a range from 1 to 4 for x

Keeping in mind that this ends up being a relatively noisy scenario, the statistics for the least-square-regression model are excellent—a highly significant linear fit ($p<0.0001$), no significant lack of fit ($p=0.9844$) with an R^2 of 0.56. Most importantly, the empirically derived slope of the line provides a reasonable approximation of the true surface.

Next, I deleted all runs with x less than 2 or greater than 3. This narrower range includes only 13 runs out of the original 38. See the resulting fit in Figure 3.

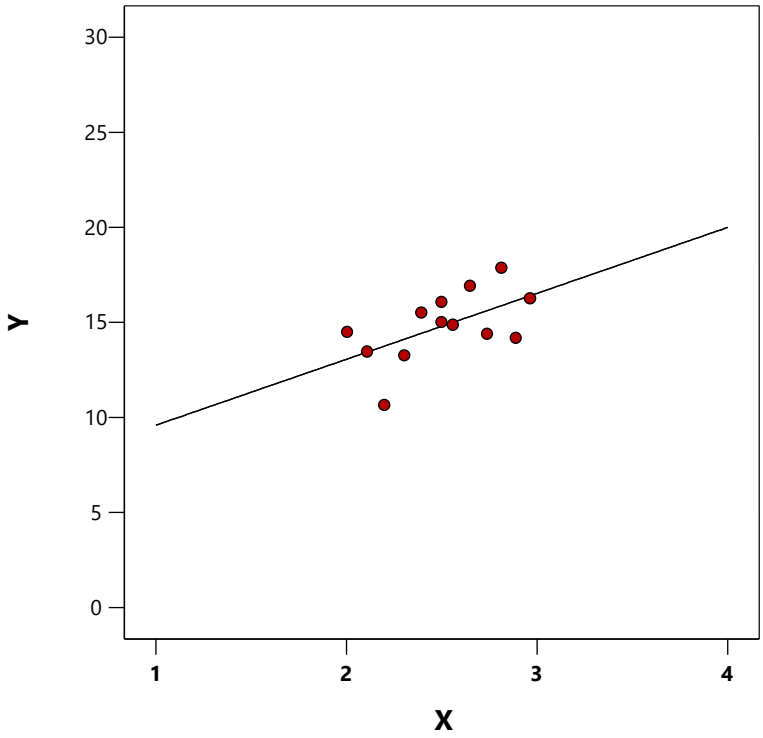

Figure 3: Fitting a line to data with a narrower range of 2 to 3 for x

Notice that the slope remains helpful by predicting that x creates a positive effect on y. However, not surprisingly given the reduction in power due to fewer runs and the smaller effect produced by a narrower range, the statistics fall off from those in my first scenario with the full range of x—the model remaining significant, but barely so (p=0.0422), lack of fit insignificant (p=0.3441) and an R^2 of 0.32.

This is not really a fair comparison because of the reduction in the number of runs. So, I did one more simulation to beef up the one with the narrower range from 13 to 38 runs in an optimal design geared to fit a linear model. See the fit in Figure 4.

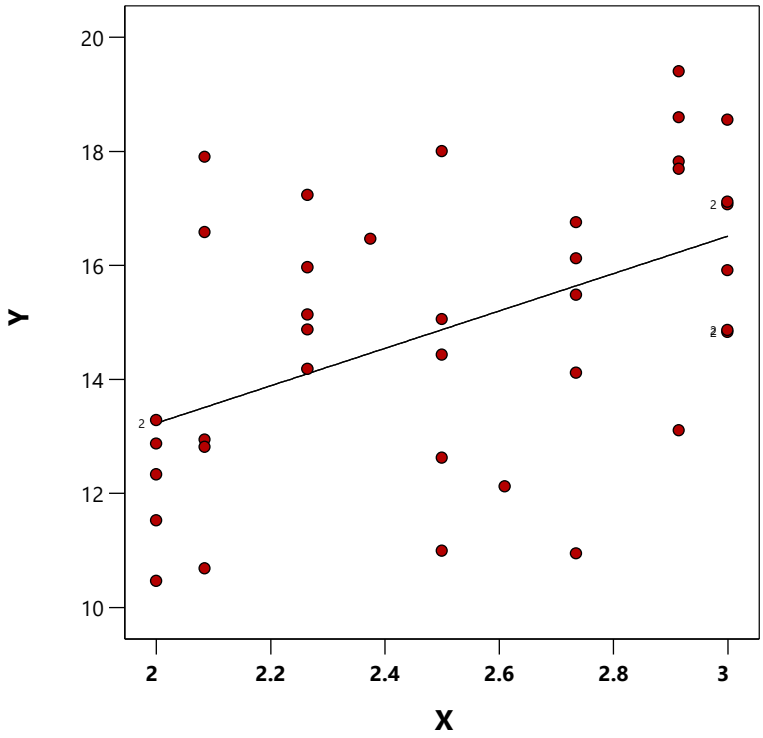

Figure 4: Increasing the number of runs within the narrower range of 2 to 3 for x

The upward slope is consistent with prior results, thus useful from an engineering perspective. With the greater power provided by more runs and an optimal design, the statistics become far better—a highly significant model ($p < 0.0019$) with lack of fit insignificant (0.2342) despite an R^2 of only 0.2373. So far as I'm concerned, this proves my point: Low R-squareds do not invalidate a model.

Conclusion

Don't be ruled by R^2! However, if you cannot help yourself, follow the advice of my colleague Shari Kraber[2] and focus on the adjusted and predicted forms of this statistic, not on the raw R^2. But, most importantly, check your model for statistical significance, consider its sensibility based on subject matter expertise and confirm the findings via follow-up runs.

The goal of DOE is to identify the active factors and measure their effects. However, factor levels often must be restricted, even for experimental purposes. This is nearly always the case for studies done at the manufacturing stage, where upsets to the process cannot be risked. In such cases, make more runs for increased power. Then measure success via ANOVA and (hopefully) its significant p-value. When an experimenter succeeds on these measures, despite an accompanying low R^2, they should be congratulated for doing a proper job of DOE, not shot down for getting poor results! It boils down to this: Although R^2 is a very popular and simple statistic, it is not very well-suited to assessing outcomes from planned experiments.

Footnotes

1. "Cross-Sectional Relations of Digital Vascular Function to Cardiovascular Risk Factors in the Framingham Heart Study," Hamburg, et al, *Circulation*, Volume 117, Number 19, 2008.

2. “R-squared mysteries solved”, Shari L. Kraber, *Journal of Plastic Film & Sheeting*, Volume 37, Issue 4, 2021.

Biography

Mark J. Anderson, Engineering Consultant at Stat-Ease, Inc., is a professional chemical engineer (PE, State of Minnesota) and certified quality engineer (CQE, American Society of Quality). He is the prime author of a trilogy of books on design of experiments (DOE) and response surface methods (RSM): *DOE Simplified*, *RSM Simplified* and *Formulation Simplified*.