STATISTICS ROUNDTABLE

BY CHRISTINE M. ANDERSON-COOK, YONGTAO CAO AND LU LU

Maximize, Minimize or Target

Optimization for a fitted response from a designed experiment

ONE COMMON GOAL in running and analyzing a designed experiment is to find a location in the design space that optimizes the response of interest. Depending on the experiment's goal, we may seek to maximize or minimize the response, or set the process to hit a particular target value. After the designed experiment, a response model is fitted, and the optimal settings of the input factors are obtained based on the estimated response model. The suggested optimal settings of the input factors are then used in the production environment.

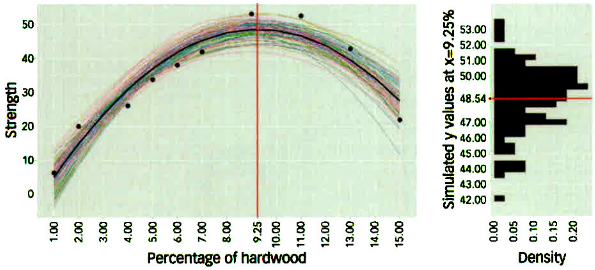
A potential difficulty that often has been ignored in the above procedure is not accounting for uncertainty in the parameter estimation. Uncertainty in the estimated model parameters can lead to choosing a less-than-optimal setting of the inputs for the true underlying process.

Consider a simple example based on an adaptation of a data set that seeks to optimize the strength (y) of kraft paper by altering the percentage of hardwood (x) in the pulp batch.[1] The main plot of Figure 1 shows the observed data, as well as the maximum likelihood (ML) fitted curve to the data assuming a quadratic model, which is given by $\hat{y} = -6.23 + 11.84x - 0.64x^2$.

If a primary goal is to maximize the paper's strength, we can see from Figure 1 that the maximum of the estimated curve with a strength value of 48.54 occurs at 9.25% of hardwood concentration. The exact location of the maximum is calculated analytically by solving the equation of setting the first derivative of the fitted model to equal 0 (here, $11.84 - 1.28x = 0$).

But how confident should we feel about being able to reproduce these results if we implement our findings in the production environment, or even if we just duplicated this experiment and repeated the analysis? To answer the first part of the question, we must know more about how similar the environment where we ran the designed experiment is to the actual production environment. So let's focus on the second part of the question, which considers only the role of the variability of the observed data and how parameter estimation uncertainty impacts our optimization.

Observed data for the strength of the kraft paper example / FIGURE 1



Note: The solid black line represents the best fitting line to the data from the maximum likelihood estimates. The multicolored curves represent simulated curves consistent with the observed data. The right-hand histogram shows the distribution of estimated strength values when the percentage of hardwood in the pulp is set at 9.25%.

Focus on variability

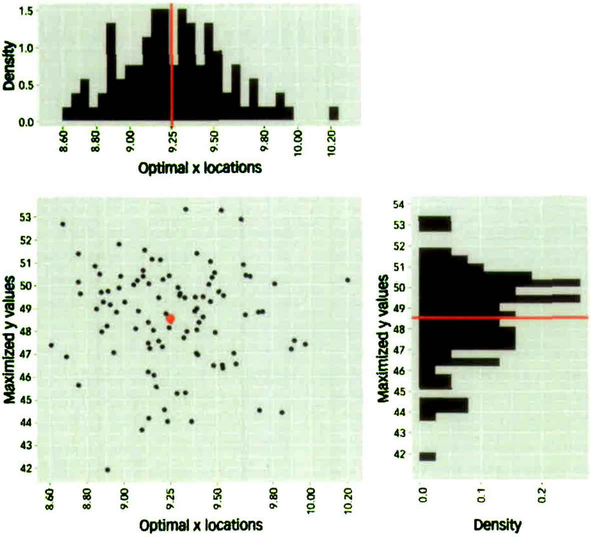
There are two ways in which our optimal choice might turn out to be incorrect:

1. The estimated value of the response—the strength of the kraft paper in our example—will likely not turn out to be exactly 48.54 when we set the input setting at 9.25% if the experiment is repeated with different data observed. Because there is uncertainty associated with the estimated response curve, we might miss above or below the estimated response.
2. The location in the x-space, the percentage of hardwood in the pulp batch associated with the maximum might not be the actual optimal location. In some sense, this might be more important than the first category of the miss because it means we are setting our process to run at a suboptimal set of values.

To better understand what range of values we might experience for the response and the x-setting, consider a simulation in which we explore what other curves are consistent with the observed data by generating M model parameters using:

$$\beta_r \sim MVN\ (\hat{\beta}, \hat{\sigma}^2\ (X'X)^{-1}$$

Best locations to maximize strength / FIGURE 2



Note: The scatter plot shows the best locations (percentage of hardwood in the pulp) from simulated curves for maximizing strength with estimated strength values. The top and right plots show histograms of the estimated optimal *x*-locations and maximized strength values across simulated values. The red point and lines shows the results from the fitted response based on the observed data.

in which $r = 1, \dots, M$ and β and $\hat{\sigma}^2$ are the ML estimates of the model coefficients and variance parameters,[2] and X summarizes the design matrix of what combinations of hardwood percentages were considered in the experiment.

We generate $M = 100$ curves that are in the estimated uncertainty for the model parameters, which are shown as the multicolored lines in the main plot of Figure 1. For each of these curves, we can explore what range of response values might be obtained from the chosen optimal setting based on the point estimate of the curve.

The histogram on the right of Figure 1 shows how much the estimated values for the strength would vary when the hardwood percentage is set at 9.25%. Specifically, the curve values here can reasonably be expected to range from 42 to 53, although many of the values are clustered quite closely to 48.5. The impact of this variability would be reflected in the performance of the process after it is set up at the chosen optimum setting. The average performance might be better or worse than anticipated.

To examine the second type of miss, we found the x-setting (percentage of hardwood in pulp) that maximizes each of the 100 simulated curves. The results of this are shown in the top portion of Figure 2.

While many optimal settings for the simulated curves are close to the identified best location of 9.25%, there is some variation to the values (between 8.6 and 10.2) in which the process should be run. This should be a cautionary note to remind us that the choice of 9.25% for the input for our process might not be strictly optimal. There is some reassurance that we should be relatively close to the best location.

In addition to the percentage of hardwood selected to obtain the maximum strength shifting, the value of the estimated strength also changes. The right-hand portion of Figure 2 shows the range of estimated maxima for the various curves, while the main portion of Figure 2 shows a scatter plot of how the x-setting and y_{max} values are related.

The precision with which we are able to predict the value of the curve at a given setting and where the optimal setting lies are a function of the original data that we have collected. If we double the amount of data collected (using the whole set of the original data),[3] we find that the range of strength values observed at the optimal settings shrinks from [42, 53] to [44, 51], and the range of anticipated optimal percentage locations shrinks from [8.6, 10.2] to [8.8, 9.9]. Therefore, additional data or an improved designed experiment can reduce the uncertainty for both of these potential misses and help guide an improved choice of optimal x-location, as well as provide better calibrated expectations of future performance.

If we think about finding a minimum based on an estimated curve, we might imagine that in many instances, the process would look quite similar to what we have outlined for the maximization case in our example. The curve might be inverted in the region of interest, and we would be looking for the low point on the curve. So the characteristics of the optimization would be quite similar.

Because hardwood is typically one of the **more expensive ingredients in the pulp,** we will likely **choose the 6% setting** because it is cheaper.

If we are looking to get close to a target, however, the characteristics of the optimization differ. We now consider the same example, where instead of maximizing strength, we want to achieve a strength value close to 42. For some uses of the paper, achieving a particular target strength is important given how the paper will be used.

The main portion of Figure 3 shows the same estimated curves as shown in Figure 1, but now with the horizontal target strength line added. Clearly for the estimated curve, there are two hardwood percentage values (at approximately 6% and 12.45%) that are predicted to hit this target. We can observe where the ML curve crosses the target line, or use an analytic strategy to find this optimum by solving the equation:

$$\hat{y} = -6.23 + 11.84x - 0.64x^2 = 42.$$

We can again use the 100 simulations to show the range of values for average strength we can expect to observe if we were to implement one of these settings into production. The left and right plots in Figure 3 show how much variability around the target strength we might expect at these settings. Clearly, from our two potential optimal settings, we might obtain values that differ considerably from the target. It is worth noting that the variation of the fitted response values is different for the two possible optimal settings.

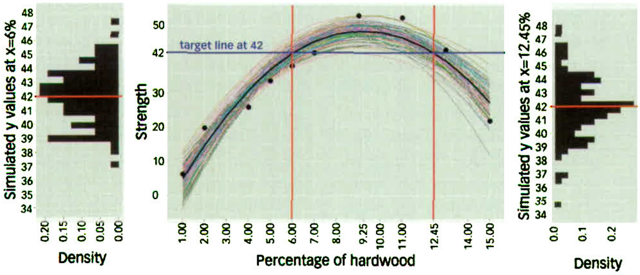
At x = 6%, for instance, the estimated strength ranges from 37.28 to 47.33. But when x = 12.45%, the estimated strength ranges from 34.55 to 47.65. Because hardwood is typically one of the more expensive ingredients in the pulp, we will likely choose the 6% setting because it is cheaper and leads to solutions that have less variability around the target value.

When we examine the range of locations at which values of 42 are obtained from the 100 simulated curves, we find that in Figure 4, there is even greater variability in the results compared to when we were seeing to maximize the response. This is a result that vertical shifts of the curve and changes in the shape of the curve where the curve crosses the threshold.

In addition, because the curves are not near a natural stationary point, the rate of change of the response around these optima can be much larger. It is also interesting to note the single location between the two modes, where x = 8.9%, results in a maximum strength of 41.94.

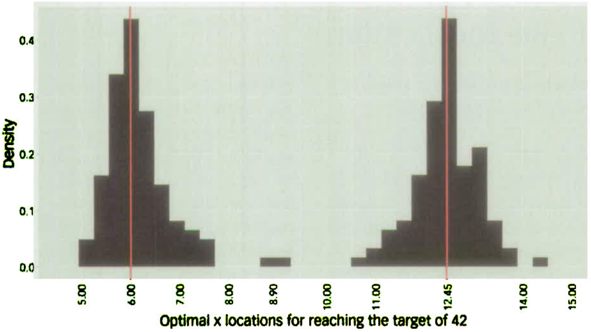
So for this particular simulated curve, we are never able to reach the target of 42. In this case, instead of solving the equation $\beta_0 + \beta_1 x = \beta_2 x^2 = 42$ for x, we find the location on the curve that gets as close to 42 as possible (here, the maximum).

Data and best fitting line to the data relative to optimal target value of 42 / FIGURE 3



Note: The multicolored curves represent simulated curves. Left and right-hand plots show the range of estimated strength values when the percentage of hardwood in the pulp is set at 6% and 12.45% (optimal settings based on the ML curve estimated from the observed data), respectively.

Best locations to target strength of 42 / FIGURE 4



Note: The red lines show the two locations from the fitted response based on the observed data. The bin at 8.9 corresponds to the single simulation that was not able to the hit the target, but can achieve a strength value of 41.94 in closest vicinity of the target value at 8.9% hardwood concentration.

Among all 100 simulations, 99 curves were able to hit the target, and one curve missed by a very small distance. Hence, the corresponding scatter plot and right-hand histogram from Figure 2 are less interesting in this case and are omitted because they would show all values at or close to 42.

Similar to what we saw with the maximization, if we increase the sample size for the experiment, we are able to reduce the ranges for the response at the selected hardwood percentage and the range of plausible x-locations where we might choose to optimize.

Maximum, minimum or target

Based on these two illustrations, there are several take-away messages. First, if we choose to optimize our response based on the ML point estimates from a designed experiment, we should expect that the production implementation of our analysis might vary around the suggested optimum response value.

Second, the choice of what x-setting to use should not be taken to be absolute, but rather that there are multiple potential locations close to the optimal location identified by the ML curve. Selecting based on the ML estimate is still our best choice based on available information, but we should not be overly confident about this being the true ideal location.

Simulation is an effective way to show what range of solution is consistent with the observed data. If other supplementary information about differences in costs or other factors suggest slight deviations from the estimated optimum, these should be included in the decision-making process.

Third, optimizing with a maximum or minimum leads to some differences compared to aiming for a target value. How we find optimal solutions, as well as the spread around those solutions, differs.

Fourth, the choices that are made about what and how much data to collect in the designed experiment to optimize the process are important because they can have a substantial impact on the precision of the results for x and y. The natural variability of the process also has an impact.

Finally, if there is more than one input factor over which we are seeking to optimize, the ranges of values become regions in the input space. QP

READ AND RATE

If you would like to comment on this article, please post your remarks on this column's webpage at www.qualityprogress.com, or e-mail them to editor@asq.org.

REFERENCES

1. Douglas C. Montgomery, Elizabeth A. Peck and G. Geoffrey Vining, *Introduction to Linear Regression Analysis*, Wiley, fourth edition, 2006, p. 205.
2. Michael Kutner, Christopher Nachtsheim and John Neter, *Applied Linear Regression Models*, McGraw-Hill Education, fourth edition, 2004, p. 227.
3. Montgomery, *Introduction to Linear Regression Analysis*, see reference 1.

CHRISTINE M. ANDERSON-COOK is a research scientist in the Statistical Sciences Group at Los Alamos National Laboratory in Los Alamos, NM. She earned a doctorate in statistics from the University of Waterloo in Ontario. Anderson-Cook is a fellow of ASQ and the American Statistical Association.

YONGTAO CAO is an assistant professor in the department of mathematics at Indiana University of Pennsylvania in Indiana, PA. He earned a doctorate in statistics from the University of Wyoming.

LU LU is a visiting assistant professor in the department of mathematics and statistics at the University of South Florida in Tampa. She was a postdoctoral research associate in the statistical sciences group at Los Alamos National Laboratory. She earned a doctorate in statistics from Iowa State University in Ames, IA.