# Data Snooping Part Three

## What happens when we cannot write models for the data?

Donald J. Wheeler

Parts One and Two illustrated four problems with using a model-building approach when data snooping. This column will present an alternative approach for data snooping that is of proven utility. This approach is completely empirical and works with all types of data.

### THE MODEL-BUILDING APPROACH

When carrying out an experimental study we want to know how a set of input variables affect some response variable. In this scenario it makes sense to express the relationships with some appropriate model (such as a regression equation). The experimental runs will focus our attention on these specific input and response variables, and the model will answer our questions about the relationships.

However, experiments always divide the set of all possible input variables into two groups. Some variables will be studied while the remaining variables will be excluded from the experiment. Occasionally one or two of these other variables may be held constant during the experiment, but most will simply be ignored. In an attempt to keep these extraneous variables from influencing the experimental study we commonly use some form of randomization with the experiment. (Randomization is simply a piece of insurance that attempts to average out the effects of any and all of these extraneous variables.)

With experimental studies the model-building approach has proven very satisfactory at examining the relationships between the factors studied. Yet, as illustrated in Parts One and Two of this series, the model-building approach can run into problems when it is used with data snooping. The four caveats illustrated there describe how we can miss important relationships between variables while also finding spurious relationships. These problems arise from two major differences between performing an experiment and data snooping. Unlike an experimental study, when we go data snooping we do not narrow the focus to a specific set of variables. Neither do we attempt to control the levels for the input variables. When we include all possible variables at whatever levels they happened to have taken on in the past we open the door to the problems identified by the caveats of data snooping. So, while mathematical techniques exist that allow us to extract information from messy databases, the quality of the relationships found will always be somewhat uncertain.

### AN ALTERNATIVE TO MODELING RELATIONSHIPS

Rather than looking for a model to "explain" the response variable we could simply observe the behavior of the response variable over time. When we do this we no longer need to estimate parameters for a regression equation, and we no longer have to specify a regression model for

our analysis. Instead, we simply characterize the past process behavior as belonging to one of two classes. Either the process response variable has behaved predictably or it has behaved unpredictably. So, instead of trying to fit a specific model to our data, this approach consists of characterizing the overall behavior of our data as fitting into one of two broad classifications. This shift from estimation to characterization requires a different way of thinking about our analysis; a different way of organizing our data; and a different set of computations. Until these fundamental differences are clearly understood confusion is inevitable.

## CHARACTERIZING PROCESS BEHAVIOR

Using the same data as in Parts One and Two, our response variable is the amount of steam used each week in the plant, $Y$. (Both the data and the regression models used here may be found in the earlier articles.) We once again begin with the baseline of the first eight weeks of production. But rather than seeking to identify a relationship between $Y$ and some independent variable as before, we place the eight values for $Y$ on a process behavior chart for individual values.

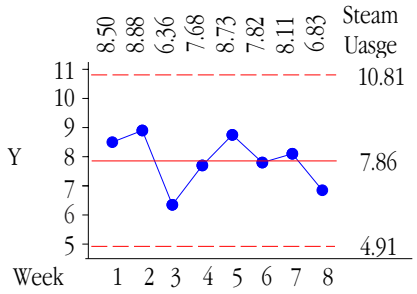


**Figure 1: Baseline Individuals Chart for Weekly Steam Usage $Y$**

While we have not "explained" any of the variation in $Y$, the limits of 4.91 to 10.81 do define the routine amount of variation present in the baseline period. These limits give us something to use in evaluating future values. If future values go outside these limits then we will have strong evidence that the process has changed from how it operated in the past. If new values remain within these limits we can assume that any changes between the past and the present have had a minimal impact and the past may still serve as a guide for the present and future.

Notice how the emphasis has shifted from one of estimating parameters for an assumed model to one of characterizing the behavior of the process over time. This temporal aspect is almost completely missing from traditional approaches to analysis.

Now the limits in Figure 1 are not terribly tight, but they are the limits that "fit" these data in the following sense: They approximate the region where 99 percent to 100 percent of the future values should be found if the process continues to operate in the same way that it operated in the baseline period. They are the voice of the process.

Once we have computed the limits for our baseline period we should pause to see if any of the baseline points fall outside the limits. If no baseline points fall outside the limits, then we can extend the limits and continue to plot sequentially a new value for each time period as these new values become available.

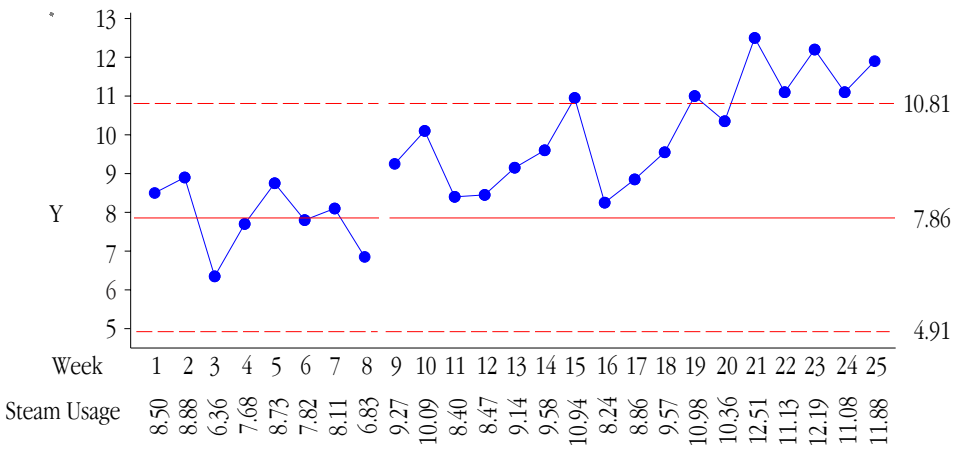


**Figure 2: Individuals Chart for Weekly Steam Usage *Y* for Weeks 1 to 25**

By week 15 we can be sure that the plant's consumption of steam has changed from what it was during the baseline period. This change in behavior is evidence that some process inputs, either known or unknown, have changed. *And we do not need to know the specific relationships to know that this change has happened.*

## ASSIGNABLE CAUSES

Now that we have evidence that the process has changed, we know that there is gold in the mine—one or more assignable causes exist. Assignable causes are input variables that are not being controlled even though their impact is so dominant that when they change levels they take the whole process along for the ride. And as Aristotle taught us, the best time to identify an assignable cause is when the process changes.

Once we have evidence of the effect of an assignable cause we will need to discover what it is. The dominant effect of an assignable cause is what makes our investigation worthwhile. When we identify an assignable cause we can either make it part of the set of control factors or we can seek to compensate for it in the future. Until we identify it we can do neither.

Caveats One and Four explain how we might fail to identify a dominant assignable cause when using a model-building approach with data snooping. Examining the process behavior allows us to wait until a dominant assignable cause changes levels, and this makes it easier to detect the dominant but uncontrolled inputs that we call assignable causes. In addition, characterizing the process behavior also saves us from working to model cause-and-effect relationships that have no real impact upon the overall process behavior.

Thus, points outside the limits not only tell us that the process has changed, but they also represent opportunities to gain new process knowledge. And this knowledge is what allows us to improve the process for the future. In effect, a process behavior chart simultaneously considers all possible input variables. It prioritizes the impact that each of these inputs has on the response variable, and it tells us when a dominant input makes its presence known.

This is how a generic approach to data snooping can work in spite of the caveats. Rather than trying to fit a specific model to our data, we seek to discover if our process can be

characterized by the rather broad model known as a predictable process, or as an alternative, if it shows evidence of changing over time.

## EVALUATING SPECIFIC RELATIONSHIPS

Of course, as soon as we detect the change in Figure 2, we know that any and all relationships detected and modeled during the baseline period are now suspect. However, if we wish to do so, specific relationships can also be investigated by means of a process behavior chart.

To illustrate this point we use the model $Y = f(X2)$ found using the baseline data in Part Two. The regression equation was:

$$\textit{Predicted } Y \;=\; 4.77 + \; 0.045\; X2$$

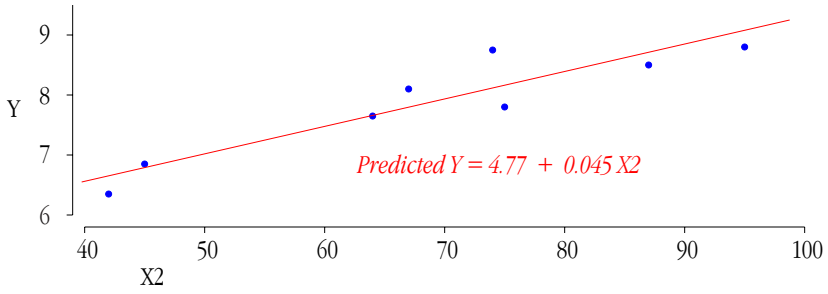


**Figure 3: Regression of Y upon X2 for the Baseline Period**

Remember, one of the paradigm shifts involved with using process behavior charts is looking at the data in time-order sequence. This order is obscured in Figure 3, so we redraw Figure 3 as a time series of actual values with the accompanying predicted values in Figure 4. (In both the predicted values are shown in red and the actual values in blue.) Here we see how the predicted values track the actual values during the baseline period.

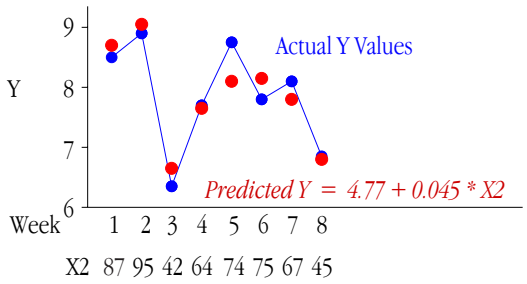


**Figure 4: Actual and Predicted Y Values for Baseline Period**

However, this tracking breaks down following the baseline period. But can we say exactly when this breakdown occurred?

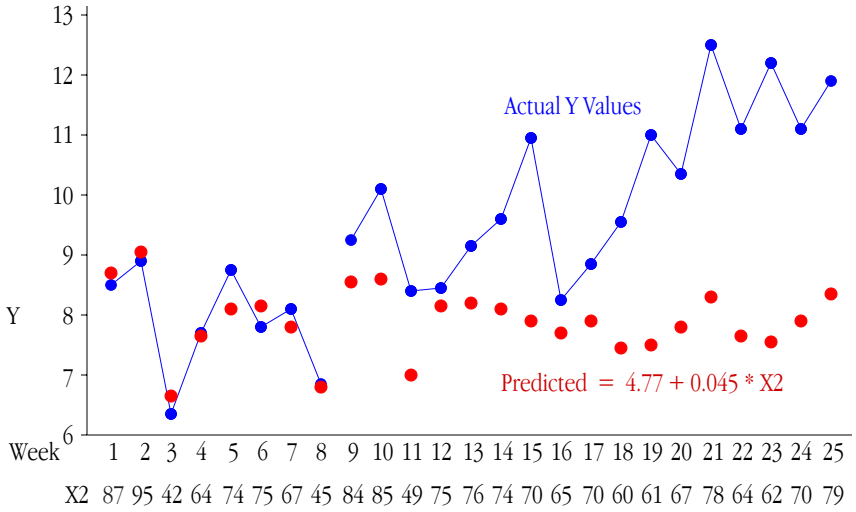


**Figure 5: Predicted and Observed Values for Y**

If we subtract the predicted $Y$ values from the actual $Y$ values we get the residual values. These are shown for the baseline period in Figure 6.

| *Observed* | | *Predicted* | |
|---|---|---|---|
| ***Y*** | ***X2*** | ***Y*** | ***Residuals*** |
| 8.50 | 87 | 8.69 | –0.18 |
| 8.88 | 95 | 9.05 | –0.16 |
| 6.36 | 42 | 6.66 | –0.30 |
| 7.68 | 64 | 7.65 | 0.03 |
| 8.73 | 74 | 8.10 | 0.63 |
| 7.82 | 75 | 8.15 | –0.32 |
| 8.11 | 67 | 7.79 | 0.33 |
| 6.83 | 45 | 6.80 | 0.04 |

**Figure 6: Baseline Residuals for *Y = f(X2)***

These residuals may be placed on their own *XmR* chart. We shall use the baseline period to compute the limits: The baseline average residual is 0.006, the average moving range is 0.426, and the natural process limits are –1.13 to 1.14. The individuals chart for the residuals from the *Y = f(X2)* model is shown in Figure 7.

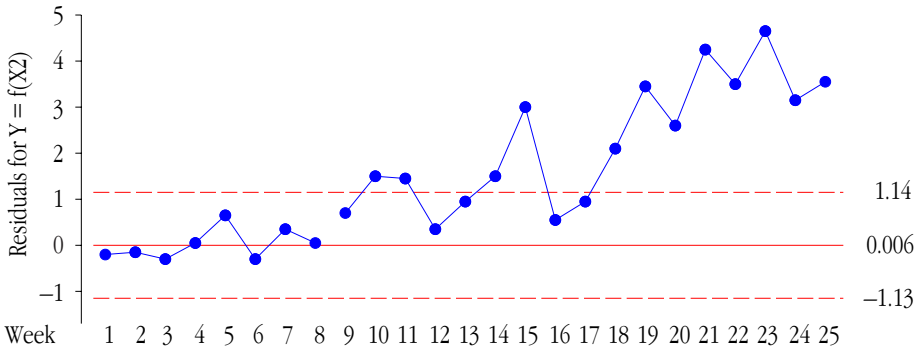


**Figure 7: Individuals Chart for Residuals for *Y = f(X2)***

The chart in Figure 7 shows that the baseline regression equation for *Y = f(X2)* breaks down by week 10. A residual value outside the limits is signaling the end of the usefulness of the regression equation for making predictions. The breakdown of a regression equation does not necessarily mean that the process itself is changing. (Remember the second caveat.) Figure 2 tells us that the process itself was changing by week 15. It is important to note the difference in how we interpret Figures 2 and 7.

If the original relationship makes sense in the context of the data, then a residual value outside the limits may be telling you that additional independent variables are needed. Otherwise, a residual value outside the limits may be telling you that the relationship is spurious.

However, it is still Figure 2, not Figure 7, that tells us when the process has changed. So, while you can use a process behavior chart to examine the idea that a given relationship continues to work over time, this is not the same as detecting a change in the overall process behavior. Even when the relationship is real and continues over time, it may not be a dominant relationship.

In our example, the steam used for process heat will vary with the amount of glycerin produced, but the heating load for the buildings dominates the relationship between *Y* and *X2*. So while our simple regression of *Y = f(X2)* breaks down over time, the relationship for *Y = f(X4, X2)* does a reasonable job as we found in Part Two. It also makes sense in the context. The bivariate regression equation found using all 25 records was:

$$Predicted\ Y = 10.46 + 0.047\ X2 - 0.082\ X4$$

Figure 8 shows these predicted values along with the actual values for steam usage. Figure 9 shows the individuals chart for the residuals for the regression model above.

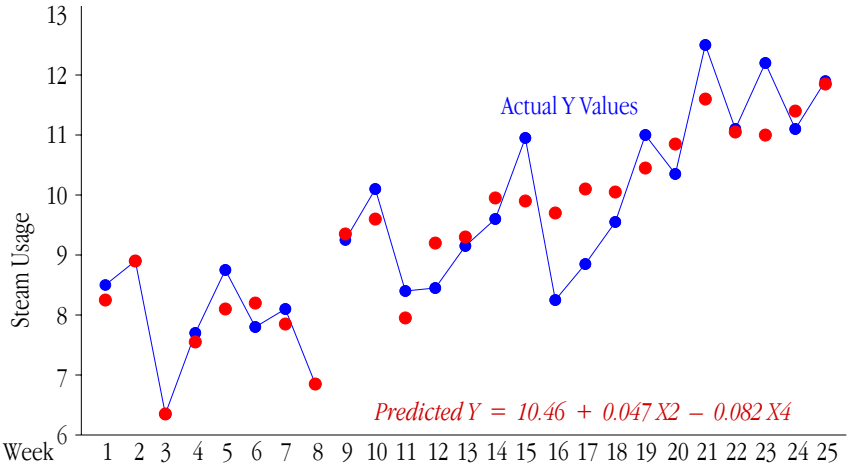


**Figure 8: Actual and Predicted Values for Steam Usage**

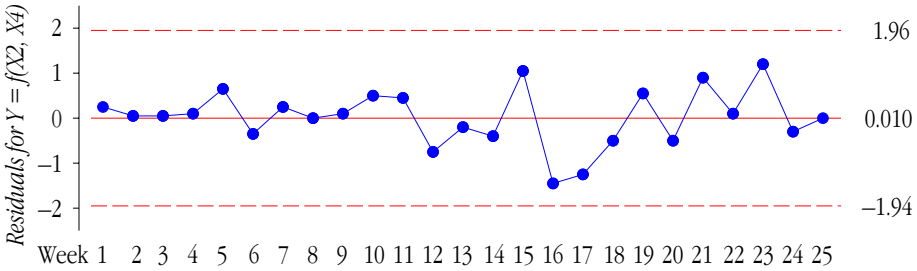


**Figure 9: Individuals Chart for Residuals for *Y = f(X2,X4)***

The limits in Figure 9 also define the amount of uncertainty that may be attached to a predicted value. As long as this regression proves useful, it will err by less than two units. If and when we get a prediction that errs by more than two units we should question the usefulness of this regression model.

## OBSERVATIONAL STUDIES AND DATA MINING

The caveats of Parts One and Two explain why statisticians dislike observational studies. They represent failure modes that can contaminate models fitted to existing data. Yet today anyone with a piece of software can torture the data and fit the noise until they obtain models of the data such as regression equations, classification trees, and cluster analyses. The apparent precision of such models is seductive, but the results may not replicate. This is how data mining can become just one more example of how our ability to carry out computations may outstrip our ability to comprehend the results.

Yet there is an alternative approach to observational studies. It is not sexy. It is not complex. It can be done by hand or with software. Even your boss can understand it. It allows us to take advantage of naturally occurring opportunities to actually improve our processes for the future. This type of observational study is not subject to the caveats of data snooping. It consists of using

process behavior charts to listen to the voice of the process.

When this voice tells us a change has occurred, we look for the cause of the change. When this voice tells us that no change has occurred we do not waste our time looking for causes that do not have a major impact. The simplicity of this approach has been known to cause statisticians to hyperventilate.

How much easier it is to listen to the voice of the process than to get lost in the complexities of data snooping.