Data Snooping Part Two

What pitfalls lurk outside your database?

Donald J. Wheeler

In "Data Snooping Part One" we discovered the basis for the first caveat of data snooping. Here we discover three additional caveats of data snooping.

Last month we discovered:

The First Caveat of Data Snooping:
Important relationships may be missed
when the data set does not contain
the full range of routine values for some variables.

Here we will use the data set from Part One to illustrate three additional caveats. The response variable *Y* represents the weekly steam usage for a chemical plant. *X1* represents the amount of fatty acid in storage. *X2* represents the amount of glycerin produced. *X3* is the weekly number of hours of operation for the plant. (Last month an additional variable was included in the data set, but here we leave it out to illustrate what its absence does to our analysis.) As before, we use the first eight weeks of production as our baseline.

Y	*X1*	*X2*	*X3*
8.50	6.57	87	115
8.88	6.60	95	115
6.36	3.45	42	55
7.68	5.01	64	100
8.73	5.76	74	110
7.82	5.69	75	105
8.11	4.95	67	110
6.83	4.62	45	55

Figure 1: Baseline: Four Variables for Eight Weeks of Production

In Figure 2 we see that each of the three simple regressions has a p-value that is less than 0.05. Furthermore, each of these regression models can explain over 80 percent of the variation in *Y*.

Regression Fitted	Coefficient of Determination	p-value for Slope
$Y = f(X1)$	0.817	0.0020
$Y = f(X2)$	0.864	0.0008
$Y = f(X3)$	0.875	0.0006

Figure 2: Three Simple Linear Regressions

In Part One, using these baseline data, we found that regressions usingtwo independent variables could not really do better than using either $Y = f(X2)$ or $Y = f(X3)$. Since Y represents the amount of steam used, and $X2$ represents the amount of glycerin produced each week, let us use $Y = f(X2)$ for our predictions. The specific equation for this model is:

$$Predicted\ Y\ =\ 4.77 +\ 0.045\ X2$$

Figure 3 shows this regression equation and the scatterplot for the baseline period. As expected, this regression model does a reasonable job of fitting these data.

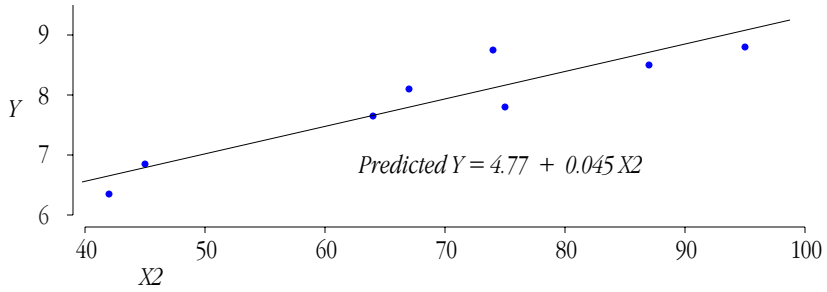

Figure 3: Regression of *Y* upon *X2* for Baseline Period

The data for weeks 9 through 25 are shown in Figure 4.

Y	*X1*	*X2*	*X3*
9.27	6.28	84	105
10.09	6.72	85	110
8.40	3.89	49	100
8.47	5.68	75	105
9.14	6.14	76	100
9.58	5.67	74	100
10.94	5.71	70	115
8.24	4.84	65	100
8.86	5.28	70	100
9.57	4.55	60	95
10.98	5.20	61	100
10.36	5.36	67	100
12.51	6.19	78	115
11.13	5.12	64	100
12.19	4.88	62	105
11.08	5.87	70	110
11.88	6.03	79	105

Figure 4: Data for Weeks 9 through 25

When we pair up the *X2* and *Y* values from Figure 4 we get the 17 points shown in red in Figure 5. Clearly our regression equation from the baseline period does not fit these data. Perhaps it needs tweaking.

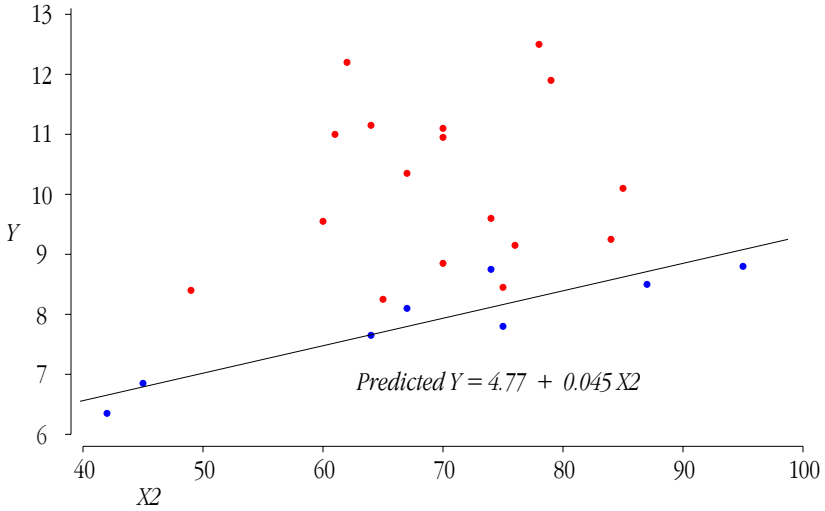

Figure 5: Baseline Regression *Y = f(X2)* with Additional Data Plotted

When we use all 25 weeks of data we find the simple regressions shown in Figure 6. With a p-value of 0.138 the "relationship" between *Y* and *X2* is found to be statistically indistinguishable from a horizontal line.

Regression Fitted	Coefficient of Determination	p-value for Slope
Y = f(X1)	0.147	0.059
Y = f(X2)	0.093	0.138
Y = f(X3)	0.287	0.006

Figure 6: Three Simple Linear Regressions for All 25 Weeks

So while we found a strong relationship between *Y* and *X2* in the baseline period, this relationship evaporates over time. This serves to illustrate the second caveat:

The Second Caveat of Data Snooping:
Apparent relationships between variables
may disappear when additional data are considered.

This caveat effectively pulls the rug out from under using the results of data snooping for making predictions. Even if we split our database into separate portions, and use one portion to "confirm" what we found in the other portion, all of our data will still be historical, and any relationships we confirm will still only describe the past. Since all of the questions of interest will pertain to the future, our models of the past may not be useful for making predictions.

WHAT IF WE USE WHAT WE FIND ?

In Figure 6 only $Y = f(X3)$ shows a detectably non-zero slope. This simple regression explains 28.7 percent of the variation in Y. Can we do better with a bivariate regression? Figure 7 shows the results for adding a second variable to the model $Y = f(X3)$ (using all 25 records).

The bivariate regression model $Y = f(X3, X1)$ explains 28.8% of the variation in the response variable Y, but the conditional p-value for using $X1$ in addition to $X3$ is 0.90, which means that this bivariate regression model is not detectably better than $Y = f(X3)$.

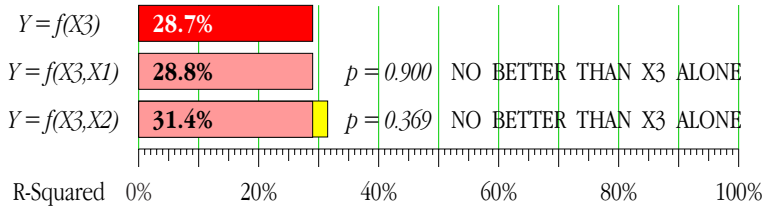

Figure 7: Adding Variables to $Y = f(X3)$

The bivariate regression model $Y = f(X3, X2)$ explains 31.4% of the variation in the response variable Y, but the conditional p-value for using $X2$ in addition to $X3$ is 0.369, which means that this bivariate regression model is not detectably better than $Y = f(X3)$.

Since neither $Y = f(X3, X1)$ nor $Y = f(X3, X2)$ does any better than $Y = f(X3)$ we might decide to use $Y = f(X3)$. This regression equation is:

$$\textit{Predicted } Y = 3.56 + 0.058\ X3$$

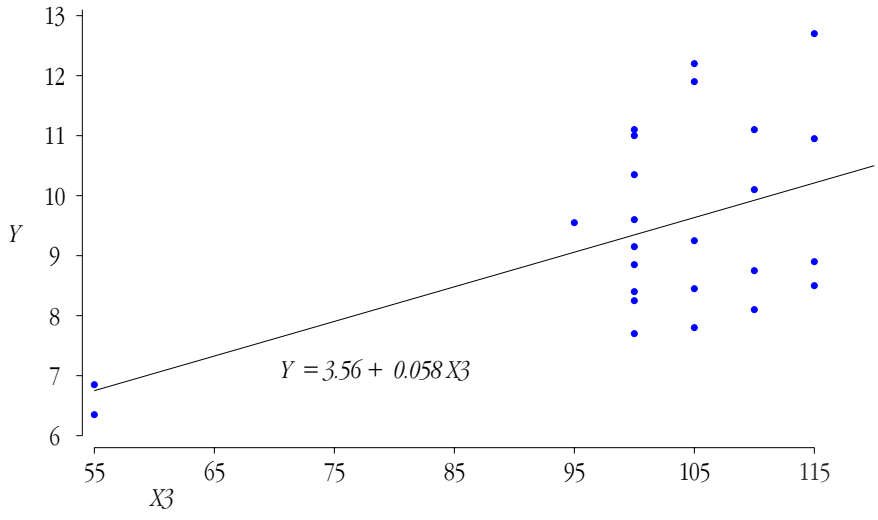

Figure 8: Regression of Y upon $X3$

When we consider the scatterplot in Figure 8 we immediately see that the regression of Y upon $X3$ is dominated by the two extreme points on the left. Remove these two points representing short production runs and the relationship between $X3$ and Y will vanish. While

short production runs clearly result in less steam usage, there is no useful relationship here apart from these two abnormal weeks. This is a common problem of fitting regression models in any situation. Outliers can corrupt your model, and the only antidote is taking the time to look at the scatterplots. This illustrates the third caveat.

The Third Caveat of Data Snooping:
Outliers can create spurious relationships.

At this point our data snooping has led us down two dead ends. First the strong relationship that we found in the baseline data vanished as additional data became available. Then as we analyzed our combined data set we found nothing but a relationship that was dependent upon outliers. In other words, while our data snooping resulted in a handful of regression equations, none of them had any hope of ever being useful in practice.

So even though the data will surrender if you torture them long enough, there is no guarantee that you will find anything useful when you go data snooping in messy data sets.

WHAT HAPPENS WHEN WE ADD *X4*?

In Part One we had an additional variable in our data set that represented the weekly average ambient atmospheric temperature. There, using all the data, we found that the simple regression of steam usage upon temperature, $Y = f(X4)$ explained 71 percent of the variation in steam usage. When we added the amount of product produced, $X2$, we got a bivariate regression that explained 85 percent of the variation in Y. This bivariate regression equation is:

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

A plot of these predicted values versus the observed Y values is shown in Figure 9.

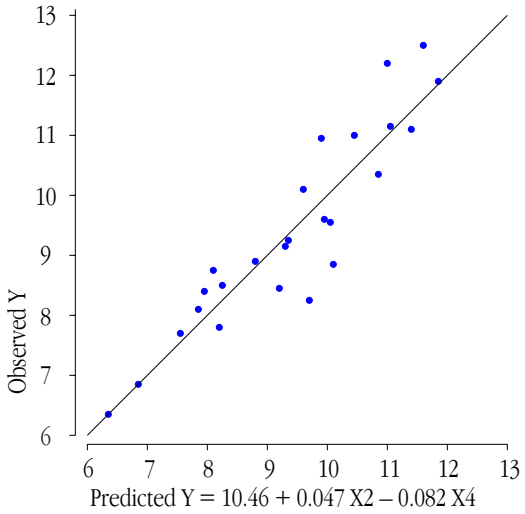

Figure 9: Predicted *Y* versus Observed *Y* for *Y = f(X2, X4)*

So while there is a strong relationship between the steam usage and the temperature, and while the temperature combines with the amount of product produced to give a good

approximation to the observed steam usages, *these relationships cannot be discovered when the temperature data are not included in the data set*. And this illustrates the fourth caveat:

The Fourth Caveat of Data Snooping:
Even dominant relationships will remain hidden
when the independent variables for those relationships
are not included in the database.

THE CAVEATS OF DATA SNOOPING

Parts One and Two have illustrated four caveats for data snooping. The first and fourth caveats pertain to what we may miss. The second and third have to do with what we may find in error. These caveats are sufficiently well known to have names.

Data Snooping Caveat One: Restricted Ranges:
Important relationships may be missed
when the data set does not contain
the full range of routine levels for some variables.

With an existing data set, your variables will only take on those levels that have occurred in the past. When these past levels are restricted in some way you may well overlook some important relationships while modeling relationships of lesser import.

Data Snooping Caveat Two: Colinearity:
Apparent relationships between variables
may disappear when additional data are considered.

Some apparent relationships may be nothing more than serendipity, but there is more to this caveat than a warning about accidental alignments. Messy data sets will generally have what mathematicians call a non-orthogonal data structure. These structures can cause the variation attributable to one variable to appear to be due to another variable. When variables exhibit colinearity (also known as confounding), or when we have an accidental alignment between variables, the apparent relationships we have found may morph, shift, and change as additional data are added.

Data Snooping Caveat Three: Outliers:
Outliers can create spurious relationships.

This is a classic problem where one or two extreme points may create apparent relationships where none really exists. Many different types of regression routines have been developed in attempts to make regression more robust to outliers. But the simple scatterplot still remains the best way to avoid using a model that is highly dependent upon outliers.

Data Snooping Caveat Four: Missing Variables:
Even dominant relationships will remain hidden
when the independent variables for those relationships
are not included in the database.

With all of our mathematical theory, and all of our software, we still do not know how to incorporate *unknown* independent variables into our regression models.

SUMMARY

Existing data sets are always messy. As we include more independent variables in our data set, and as the number of levels for each variable increases, the number of possible combinations of variable levels will increase geometrically. As a result, as a database includes more variables it will typically have an increasing number of missing combinations of variable levels. These missing combinations will create non-orthogonal data structures which will challenge our abilities to extract information about relationships from the data. So, with all these caveats, how can we ever analyze existing data sets?

First, we should not attempt to use data snooping unless we have some *idea* that needs to be examined in the light of the existing data. If we do not know what we are looking for, we are going to have a hard time finding anything in a messy data set.

Second, we cannot ever *establish* or *prove* that a given relationship exists using an existing data set. We can only identify possible relationships to be considered for experimental studies or to be validated by additional data sets.

As with everything in science, when different lines of evidence converge on a given result, that result gains credibility. This applies to the results of data snooping as well as the results of experimental studies.

Nevertheless, the caveats listed here mean that no single bit of data snooping can ever be conclusive. We simply cannot use data snooping to establish or prove that a specific relationship exists. And this shortcoming of data snooping is why my fellow statisticians do not like "observational studies."

Yet, there is a way to utilize observational studies in spite of these caveats. This approach will be the topic of Part Three.