Lurking Variables: Some Examples

Brian L. Joiner

The American Statistician, Vol. 35, No. 4. (Nov., 1981), pp. 227-233.

Stable URL:
http://links.jstor.org/sici?sici=0003-1305%28198111%2935%3A4%3C227%3ALVSE%3E2.0.CO%3B2-8

The American Statistician is currently published by American Statistical Association.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/astata.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org/
Sat Apr 8 01:06:37 2006

Lurking Variables: Some Examples

BRIAN L. JOINER*

Lurking variables are important explanatory variables that might well escape attention in a routine statistical analysis. In this report several examples of lurking variables are given. Important points illustrated include the following. Careful checking, plotting, and thinking are very important. Whenever possible, data and residuals should be examined with respect to time order and spatial arrangement. A variety of plots of the data and the residuals is virtually indispensable. In designing experiments, time order should be considered and, when practical, randomized. Such randomization is not a panacea, however, since lurking variables can still be present.

KEY WORDS: Time-order effects; Plotting; Residuals; Randomization; Lurking variables.

1. INTRODUCTION

Errors, like straws, upon the surface flow;
He who would search for pearls must dive below.
John Dryden (1678), *All for Love*.

A *lurking* variable is, by definition, a variable that has an important effect and yet is not included among the predictor variables under consideration (Box 1966). It may be omitted from the analysis "because its existence is unknown or, if its existence is known, its influence is thought to be negligible or data on it are unavailable" (Hunter and Crowley 1979).

A key question is, "How does one even identify the existence of a lurking variable when, by definition, it is not among the list of factors contemplated by the analyst?" The answer often is to examine the data and residuals with respect to time order, day of the week, spatial arrangement, or some other "diagnoser" variable. It is axiomatic that all data must be collected either serially or cross-sectionally. In fact some data have both temporal and spatial arrangements. For example, crops in various "plots" have a spatial structure but may be chemically analyzed sequentially.

In many respects the problem turns out to be similar to that encountered by physicist Walter A. Shewhart, the father of statistical quality control. Shewhart sought methods for finding "assignable causes" for important variations in industrial production processes. Shewhart found that the data sets he considered, even those from very good laboratory scientists, almost invariably contained peculiarities when examined with respect to time order. He found shifts in level and other patterns and used these patterns to help focus searches for assignable causes of these disturbances.

In other cases, searches may be focused by observing spatial patterns. In still other cases, efforts may be guided by a careful examination of possible causes for one or two "outliers."

In this report we give several examples of lurking variables and tell how their existence was detected. These examples help illustrate the fact that lurking variables can be found—or overlooked!—even in such "clean" situations as carefully designed experiments.

2. VITAMIN B_2 IN TURNIP GREENS

Our first example of a lurking variable is based on data that have been previously analyzed by both Anderson and Bancroft (1959, p. 192) and Draper and Smith (1966, p. 229; 1981, p. 406). These data resulted from an experiment conducted to evaluate the effect of three variables on the amount of vitamin B_2 in turnips. The three variables are x_1 = radiation; x_2 = moisture in soil; and x_3 = temperature. Anderson and Bancroft report a model linear in x_1, x_2, and x_3. Their model gives an (unadjusted) R^2 of .75. Draper and Smith arrive at a model containing x_2, x_3 and x_2^2 with an R^2 of .90.

Even relatively careful analysis of the residuals from the Draper and Smith model, such as the residual plots in Figure 1, reveals no serious problem with the fitted model. Draper and Smith did say (1966, p. 339; 1981, p. 598), "A plot of the residuals reveals runs of + and − signs indicating the presence of unconsidered x-variables."

However, if one plots the original data in the order they were presented in the textbook, a striking pattern can be seen (Figure 2). The data drop off nearly linearly. In fact, a simple straight line fitted to the plot in Figure 2 gives an R^2 of .90! What is the explanation? The answer is not clear, and attempts to get more details about the original experiment have not been successful. One conclusion is that some other factor not recorded—that is, a lurking variable—was the primary responsible party. It may have been that the values are reported in the order measured and that reagents or the turnips themselves decayed over

* Brian L. Joiner is Professor of Statistics and Director of the Statistical Laboratory, Department of Statistics, University of Wisconsin, 1210 W. Dayton St., Madison, WI 53706. The author especially thanks John Crowley, who wrote Section 5, Red Dye 40. He is also indebted to Ellis R. Ott who gave him an early, healthy respect for real data; to John W. Tukey who made a number of useful suggestions including the use of the term "diagnoser"; to Frederick Mosteller for clarification on the Red Dye 40 example; and to Alison Pollack for her helpful advice. This research was supported in part by the United States Army under Contract No. DAAG29-80-C-0041.

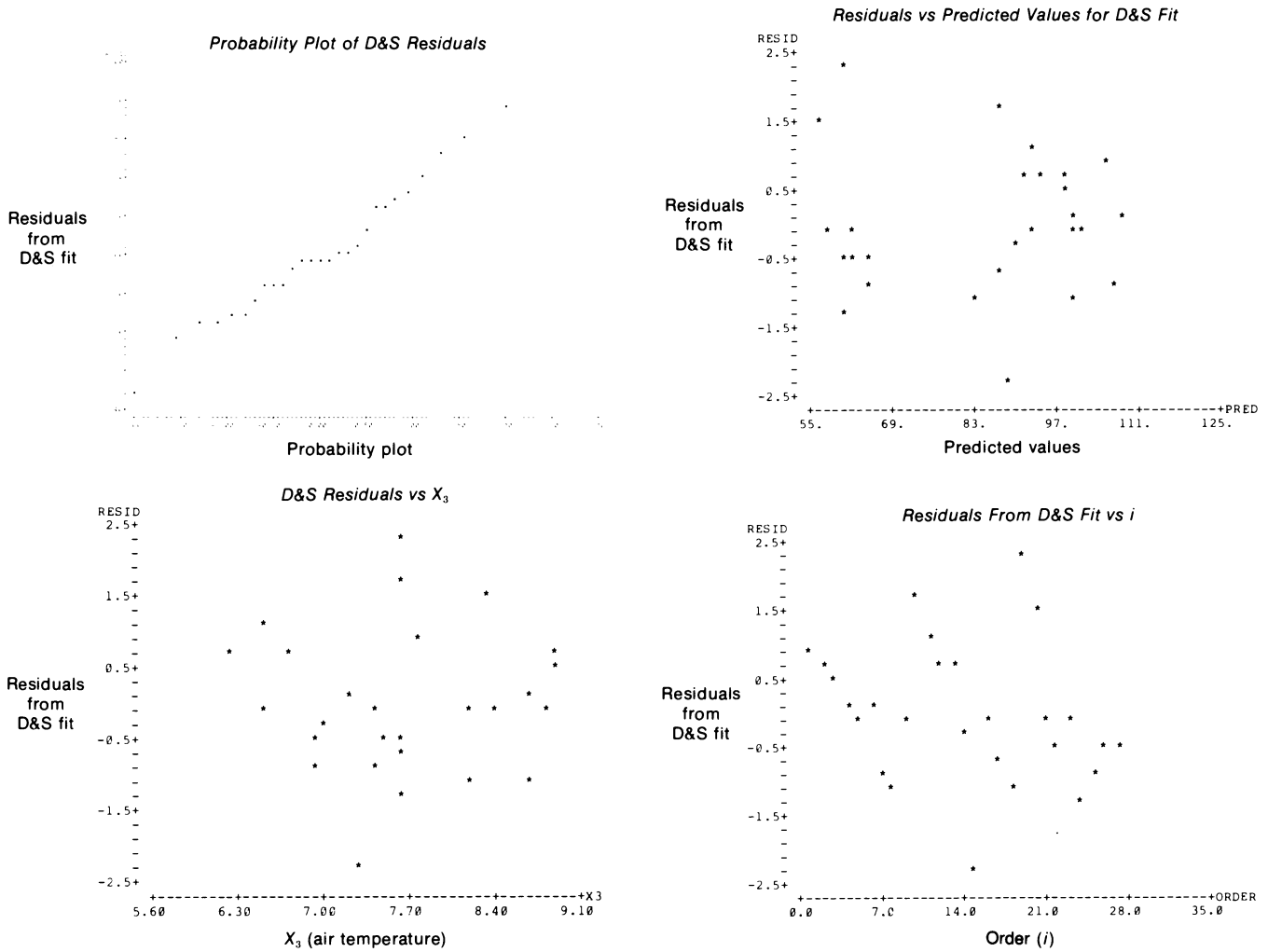

Figure 1. Standardized Residuals From Draper & Smith Fit for Vitamin B_2 in Turnips

time. There is no other ready explanation. The Y values are not merely listed in decreasing order, for a number of inversions are apparent in Figure 2. The argument that x_2 is the most important factor loses credibility on at least two accounts. A quadratic is needed to fit the three levels of x_2, and the Y values in Figure 2 seem to continually drop off with order unaffected by changes in x_2. Careful data checking has opened serious questions about the quality of the data but, in this case, has not identified the culprit.

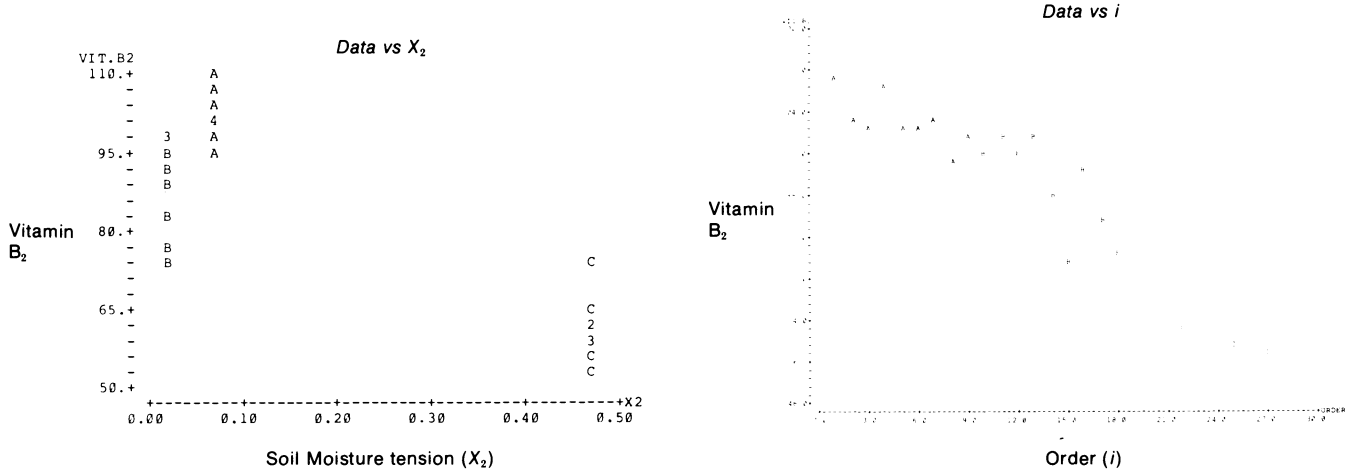

Figure 2. Plots of Data for Vitamin B_2 in Turnips

NOTE: Letters denote values of X_2: A = 0.070, B = 0.020, C = 0.474.

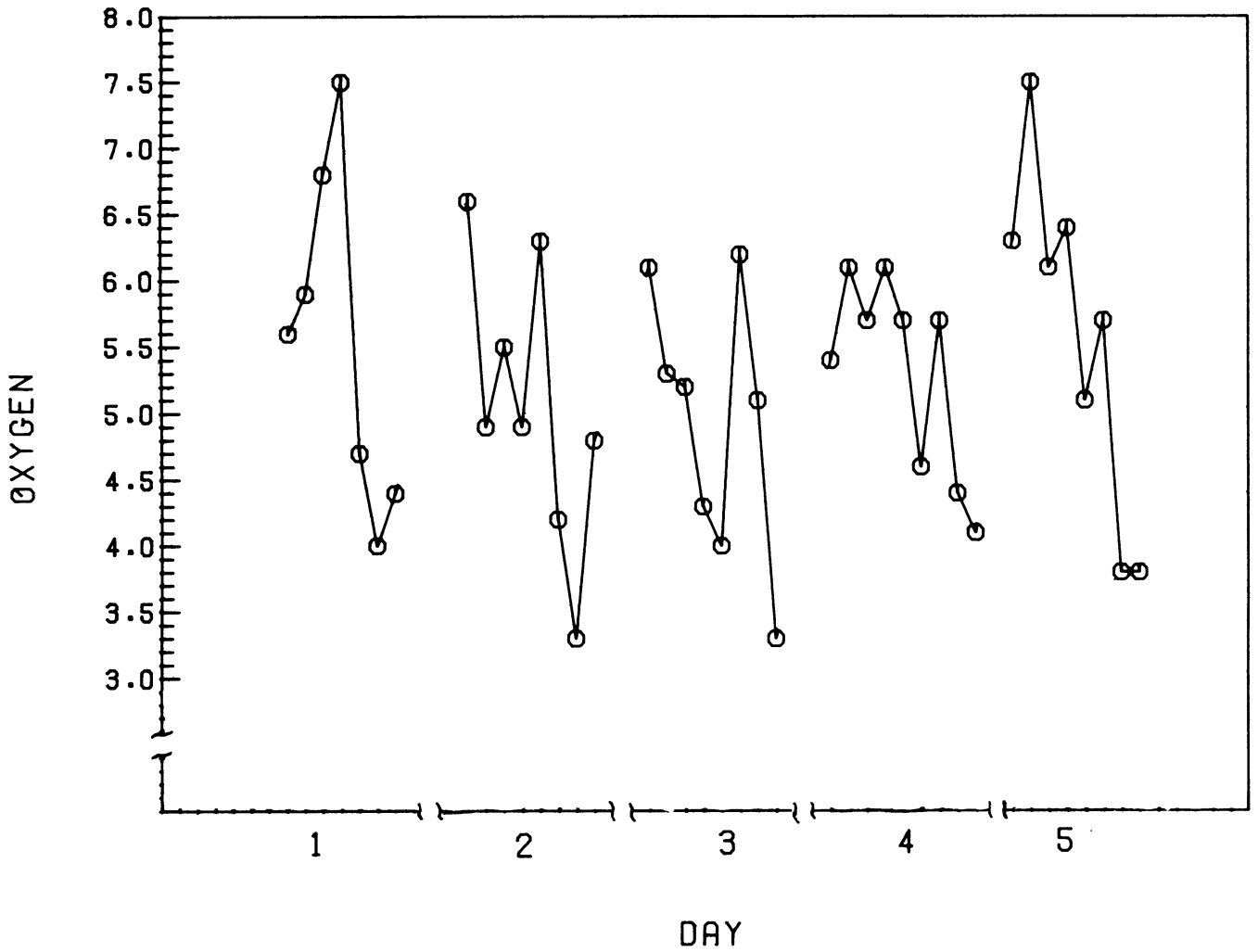

Figure 3. Time-Order Plot of Raw Oxygen-in-Steel Data

Lesson 1: One needs to plot the data themselves, not merely the residuals from some model. Analyzing the residuals does not always make obvious the existence of the lurking variables. (John Tukey has suggested we call the data the "null residuals," i.e., the residuals from a null model, thus avoiding the need to teach people to do something "new.")

3. OXYGEN IN STEEL

Another example of data having an important but not easily detectable time order decay of measurements is provided by the oxygen in steel data reported in Ryan, Joiner, and Ryan (1976, p. 205) and Joiner and Campbell (1976).* These data come from a National Bureau of Standards experiment conducted to evaluate the homogeneity of oxygen in steel rods. Twenty rods were haphazardly selected from a large batch and two measurements were made on each. A standard one-way analysis of variance revealed no significant inhomogeneity from rod to rod, but more variation than expected was present.

Fortunately, the measurements had been made in random order. A study of the residuals was uninformative, but a careful study of a time order plot of the raw data (Figure 3) led to the construction of still another time order plot (Figure 4). The latter plot made it clear that the readings had decreased dramatically within each day. In this case, a careful timely search for causes was made, but unfortunately no explanations were found. Some consolation was taken in the fact that a serious problem has been brought to light by careful statistical analysis.

The original randomization had been critical; without the randomization of measurement order, there would have been little chance of even finding out that there had been a problem. Note, however, that randomization is not a panacea. It enabled us to find the existence of a problem but it did not make the problem go away. Analysts should not be advised to ignore external factors merely because randomization has been practiced.

Lesson 2: Randomization of time order is useful. However, randomization does not obviate the need for

* Oxygen is inadvertently misspecified as nitrogen in this reference.

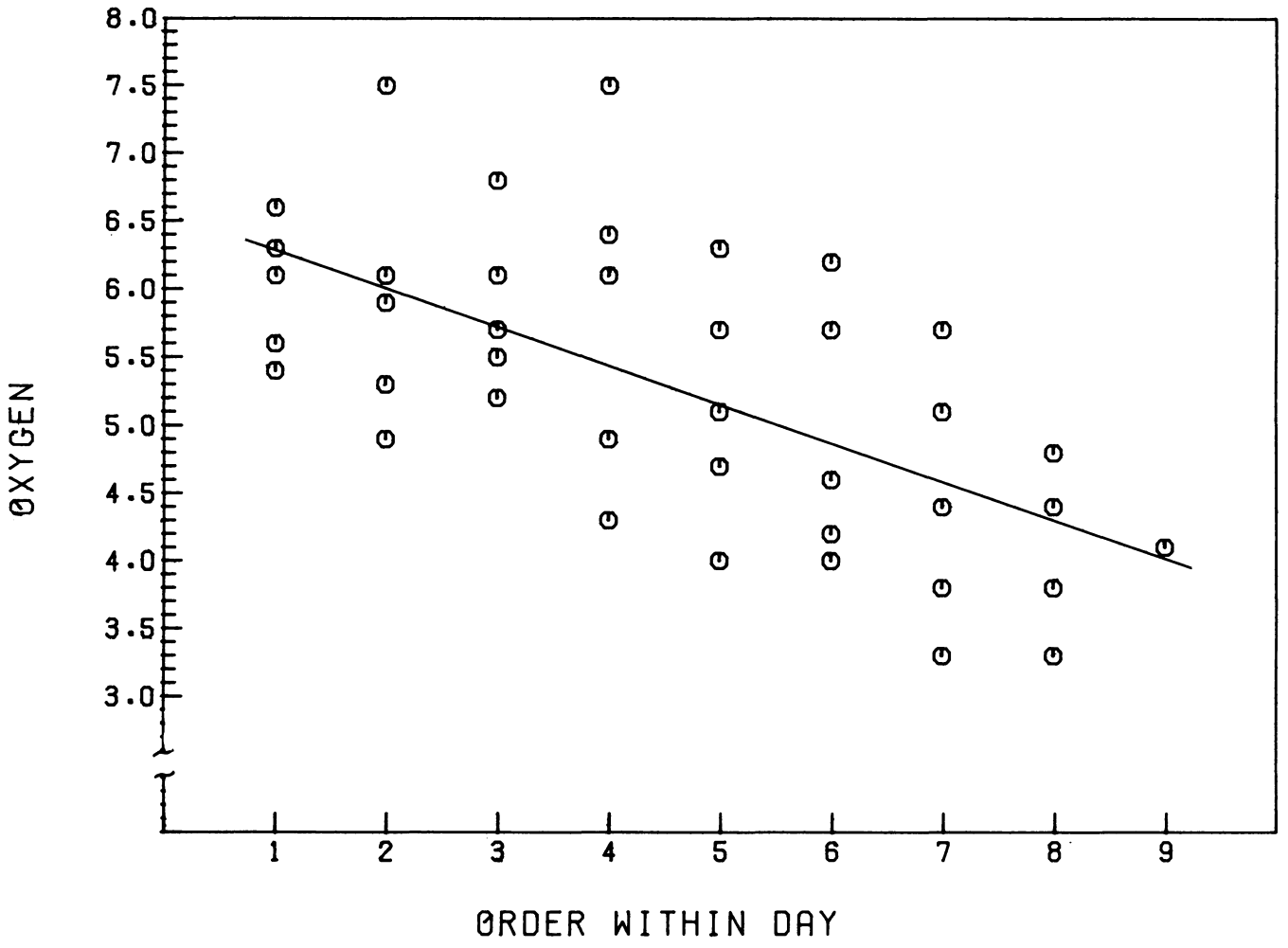

Figure 4. Revised Time-Order Plot of Oxygen-in-Steel Data

careful analysis of possible time-order-dependent effects. In addition, such randomization is not always practical (see, e.g., Joiner and Campbell 1976).

Lesson 3: Even very similar plots often show quite different things. There is no substitute for making and looking carefully at a wide variety of plots, especially given the ease and speed of computer-generated plots.

It is important to note that time itself is seldom the culprit; rather, the problem usually lies with some other factor that has varied with time in a systematic way.

4. FREEZING MEAT LOAF

These data were obtained in an experiment in which the primary interest was in comparing several different methods of freezing meat loaf. The meat loaves were to be baked, then frozen for a time, and finally compared by expert tasters. The data in question here were obtained in a pretest of oven uniformity (see Ryan, Joiner, and Ryan 1976, pp. 207–214).

Eight loaves could be baked in the oven at once and a "uniformity trial" was run to see if the drip loss (weight lost due to drippage during cooking) was different among the eight positions. Three batches were cooked and the results analyzed as a randomized block design. The results indicated that there was a statistically significant difference among the oven positions.

Further checking revealed that the three positions with the lowest drip loss were the three positions in which thermometers had not been inserted in the loaves. Possible remedies in future studies would be to insert at least a dummy thermometer in all loaves or use thermometers in no loaves.

As in many cases, once one knew where to look the "assignable cause" was obvious. An important contribution of good statistical design and analysis is to help researchers identify good places to look for problems or unexpected benefits.

Lesson 4: Look for commonalities among the best and worst (highest and lowest) data points or estimated effects. Ott (1975, e.g., pp. 107–110) and Deming (1981) give a number of interesting industrial examples.

5. RED DYE 40

Red dye number 40 is currently the second most used food coloring in the United States. Does it cause cancer

in mice? Some experts say yes and some say no. Here is a brief outline of the story, and of how a lurking variable played an important role. More details are given in Lagakos and Mosteller (1981).

Red dye 40 was approved for use by the U.S. Food and Drug Administration in 1971 after being tested for carcinogenesis in rats. In that study the rats were fed large doses of the dye throughout their lives, and no significantly greater rate of tumors was found in the red dye group than in the control group. In 1975 another study was initiated, this time using mice. Preliminary results showed a surprisingly early incidence of reticuloendothelial (RE) tumors in the exposed groups. This led to a second, larger, mouse study.

This new study, undertaken in 1976, involved 100 male and 100 female mice in each of two control groups and three dose levels for a total of 1,000 mice. Considerable controversy arose over many aspects of both mice studies; this controversy is still not resolved despite the efforts of several well-known statisticians called in as consultants. One particularly telling set of results is given in Table 1. Sex seemed to be a factor, as expected, but not red dye 40. The difference in rates for the control groups was particularly puzzling.

Statisticians Frederick Mosteller and Stephen Lagakos requested cage information and discovered that the vertical level of the cage and front or back placement were lurking variables (Table 2).

Further investigation revealed that the mice had been placed in the cages based on litters, five mice per cage. All mice in a given cage were of the same sex. The first cage had the three males from litter 1 plus two males from litter 2. The next cage had the remaining male from litter 2 plus three males from litter 3 and one male from litter 4. The next cage had two males from litter 4 plus three from litter 5. This pattern was repeated until all the male cages were filled. The female cages were filled in a similar systematic fashion. Then the cages were placed on the racks in the following systematic order: first came all the male control 1 cages, then all the female control 1 cages, then male control 2, female control 2, male dose 1, female dose 1, male dose 2, female dose 2, male dose 3, then finally female dose 3.

This placement led to complete confounding of the male control 1 versus male control 2 with front versus back of racks and to partial confounding of the other factors of interest with cage placement differentials.

The statistician Bernard Greenberg, who was also called in as a consultant, found evidence for a litter effect. This was particularly troublesome since the mice had been assigned to treatment groups by litters.

Table 1. Incidence Rate of RE Tumors (in %)

Sex	*Control 1*	*Control 2*	*Dose 1*	*Dose 2*	*Dose 3*
Male	25	10	20	9	17
Female	33	25	32	26	22

Table 2. Incident Rate of RE Tumors (in %)

	Top				*Bottom*		
Row	1	2	3	4	5	*Front Cages*	*Back Cages*
Rate	32	24	18	18	17	25	19

Final conclusions on this study are still not available but considerable progress in understanding the results has been made possible by the discovery of cage placement factors and the litter effect.

Lesson 5: Spatial and familial relationships may be important factors, as may sex.

We now give a brief sketch of several other published examples of lurking variables.

6. STACK LOSS

Daniel and Wood (1971, 1980) did a careful analysis of a data set that had already been analyzed by several previous authors. The data represent 21 successive days of operation of a plant oxidizing ammonia to nitric acid. The three explanatory variables are the flow of air to the plant, the temperature of the cooling water, and the concentration of nitric acid in the absorbing liquid. The response variable is the percentage of ammonia that is lost.

In their now classic analysis, Daniel and Wood identified two important time-order effects. The first is a start-up effect (Chapter 5) in which it seems apparent that the first day's operations under a new set of conditions produced values that were quite different from those experienced after the plant has had a chance to "line out."

The second effect is an autocorrelation effect (Ch. 7, p. 126 of 1st edition and p. 138 of 2nd edition); there are really only six *clumps* of data; there is strong correlation among the measurements in each clump.

Lesson 6: Start-up effects and strong correlation among measurements made close together in time are common in data made in time order. Plotting predictor variables versus time as well as responses versus time can help identify potential problems.

7. DANIEL AND WOOD 10-VARIABLE EXAMPLE

Even Daniel and Wood missed the time-order effect in their "10-variable" example in their first edition (1971). The data are weekly figures relating to the operation of a petroleum refining unit. In their second edition (1980, pp. 146–148) Daniel and Wood find that their 10-variable example has plant start-up problems similar to those they found earlier in the stack loss data. They found two points (19 and 20) that had abnormally large influence on the fitted equations. These two points were "observed to have been taken after a

three-week shutdown of the unit (or at least omission of data)."

Lesson 7: Even analysts who are known to do careful work must be on guard lest they miss important lurking variables.

8. CRYOGENIC FLOW METERS

The case study reported in Joiner (1977) provides still another example of time-order dependence. In that study a new facility for calibrating cryogenic flow meters was evaluated. The facility was quite complicated and involved such components as a submerged pump, a heat exchanger, temperature control valves, a weigh tank, a load cell, a set of calibrated weights, and a revolution counter. Each of these factors was subject to some uncertainty as to how it would work at cryogenic temperatures, about 85° Kelvin. A series of tests was done on the system while several meters were being calibrated.

Some lurking variables identified in that study included the following:

1. In one set of data the first three points were "outliers"; the apparent cause was insufficient "exercise" of the weight system (p. 358). This was corrected in further runs by adding more exercise cycles.
2. Another outlier may have been caused by the revolution counter's picking up a stray pulse (p. 359).
3. All the readings on one day were lower than expected, perhaps owing to the fact that a delivery of liquid nitrogen had been received that day that may have abnormally increased the pressure in the system, thereby forcing an extra amount of fluid into the catch tank, which in turn increased the buoyancy force on the weigh system (p. 360).
4. Points taken later in the day sometimes tended to be higher than those taken earlier in the day. This may have been caused by thermally generated voltages in the measuring system (p. 361).
5. A number of other abnormalities were found, but time did not permit the identification of all assignable causes.

In this example the benefit of close collaboration between statisticians and scientists is quite obvious. The analyst identified several suspicious sets of values, which enabled the scientists to focus their search for trouble spots. Further experiments helped confirm the validity of some of the findings. An even more iterative scheme in which preliminary findings suggested further data collection efforts that led to new findings, and so on, would have been even better. In this complex experiment, lurking variables seemed to be everywhere.

Lesson 8: Good continuous communication between statistician and experimenter is very valuable.

9. STRENGTH OF PLASTIC

Wilson (1952, pp. 55–56) reports the results of an experiment which was

> performed to determine the effect of the length of time of pressing in the mold on the strength of a plastic part. Hot plastic was introduced in the mold, pressed for 10 seconds, and removed. Another batch was then introduced into the same mold, pressed for 20 seconds, and so on, the time increasing with each batch. Afterward the strength of each piece was measured and plotted against the duration of the pressure. [The plot seemed] . . . to indicate a strong dependence of strength on duration. However, the research supervisor criticized the experiment because the order of the experiments had not been randomized, and so it was repeated. The results . . . [were replotted with notes as to] the order in which the measurements were taken . . . it was the *order* and not the *duration* which was the controlling variable; the first conclusion was quite erroneous. The origin of the trouble was easily traced after its presence was made known; the mold got warmer and warmer as successive batches of hot plastic were pressed in it.

Lesson 9: In many cases it is a good idea to do a repeat experiment using carefully controlled randomization.

10. CONCLUSIONS

Good statistical analysis often involves careful detective work. Part of this work is based on using some specified set of predictor variables to develop a sound model. Making proper use of these specified variables is often hard enough, but lurking variables can make the task even more difficult. The existence of such variables can usually be detected only by careful study of patterns in the data or the residuals. The failure to detect the presence of lurking variables can sometimes lead to grossly incorrect conclusions.

Lurking variables are to be found in every field. The examples given here have been predominantly from the physical sciences since that is where the author has the most experience. Many of these examples have exhibited time-order dependencies. Careful investigation of time-order dependence is one of the most useful approaches for finding lurking variables since many factors normally vary over time, and not all of them will be recorded for use among the list of predictor variables.

Spatial arrangements also often provide useful clues since some factors change "geographically"—across fields, from floor to ceiling, or in some other way. In other cases, lurking variables are found only by a careful study of what some subset of points have in common. Clusters of observations that may appear only slightly discordant must be examined for common factors.

Advice to search for lurking variables is not new. For example, Anscombe and Tukey (1963) advised an examination of "the relationship of the residuals to external variables, such as time of observation or

geographical position." Such advice is surely much, much older.

Plotting seems to be the best way to identify the presence of lurking variables. Some plots should be done routinely. For example, if data have a time or space ordering, they should be plotted versus that order. Residuals should also be plotted versus order. If the time or space differences between observations are not equal (e.g., if there are occasional big gaps) then a real time or space plot should be made, too. When appropriate, daily, weekly, monthly, or seasonal plots should also be made (see, e.g., Figure 4).

The best general rules in analysis seem to be *think* and *plot*. What could have gone wrong? How could you give patterns a chance to expose themselves? Look carefully at the plots. What hints of possible trouble are there? What else could you do to expose any problems? Ask to see the apparatus or any other source of potential problems. Ask the researcher, Precisely how did you get these measurements? What did you do first? How many measurements were made from each batch of solution? Were any repeat measurements made? What aspects of the process were repeated?

In design, the standard statistical advice is to *block* to eliminate sources of variation suspected to be important and *randomize* as much of the rest as is practical. Note that these are not standard principles of experimental practice in most fields. Generally, the statistician must specify explicitly which measurements are to be made in which order, by whom, on which device, and so on. A mere instruction to "run these randomly" will seldom suffice.

In summary, the statistician who brings to light an important lurking variable makes a very real contribution to the research, far beyond that available from the routine calculation of estimates, tests and p values. Such good work often has a major impact on the research and may result in important new findings. At the very least it reduces the chances that misleading or erroneous results will be reported without cautioning remarks.

[*Received September 1979. Revised June 1981.*]

REFERENCES

ANSCOMBE, F.J., and TUKEY, J.W. (1963), "The Examination and Analysis of Residuals," *Technometrics*, 5, 141–160.

BOX, G.E.P. (1966), "Use and Abuse of Regression," *Technometrics*, 8, 625–629.

DANIEL, C., and WOOD, F.W. (1971; 2nd ed. 1980), *Fitting Equations to Data*, New York: John Wiley and Sons.

DEMING, W. EDWARDS (1981), *On the Management of Statistical Techniques for Quality and Productivity*, unpublished manuscript.

DRAPER, N.R., and SMITH, H. (1966; 2nd ed. 1981) *Applied Regression Analysis*, New York: John Wiley and Sons.

HUNTER, WILLIAM G., and CROWLEY, JOHN J. (1979), "Hazardous Substances, the Environment and Public Health: A Statistical Overview," *Environmental Health Perspectives*, 32, 241–254.

JOINER, BRIAN L. (1977), "Evaluation of Cryogenic Flow Meters: An Example in Non-Standard Experimental Design and Analysis," *Technometrics*, 19, 353–380.

JOINER, BRIAN L., and CAMPBELL, CATHY (1976), "Designing Experiments When Run Order Is Important," *Technometrics*, 18, 249–260.

LAGAKOS, STEPHEN, and MOSTELLER, FREDERICK (1981), "A Case Study of Statistics in the Regulatory Process: The FD & C Red Dye No. 40 Experiments," *Journal of the National Cancer Institute*, 66, 197–212.

OTT, ELLIS R. (1975), *Process Quality Control*, New York: McGraw-Hill.

RYAN, THOMAS A., JR., JOINER, BRIAN L., and RYAN, BARBARA F. (1976), *MINITAB Student Handbook*, North Scituate, Mass.: Duxbury Press.

WILSON, E. BRIGHT, JR. (1952), *An Introduction to Scientific Research*, McGraw-Hill: New York.