Center for Quality and Productivity Improvement
UNIVERSITY OF WISCONSIN
610 Walnut Street
Madison, Wisconsin 53705

(608) 263–2520
(608) 263–1425 FAX

Report No. 149

Quality Quandaries*

Regression Analysis Applied to Happenstance Data

George Box

October 1996

*Edited by George Box and Søren Bisgaard.

C Q P I
UNIVERSITY OF WISCONSIN Report No. 149

Quality Quandaries*

Regression Analysis Applied to Happenstance Data

George Box
Center for Quality and Productivity Improvement
University of Wisconsin
MADISON, WISCONSIN

ABSTRACT

Care is needed in interpreting results when regression analysis is applied to happenstance data. Although a fitted regression model may be useful for prediction of future values of a series, it may be totally misleading for explaining causation relationships between the variables. Furthermore such analysis can be disastrously affected by lack of independence between residual errors.

KEYWORDS: Regression Analysis, Least Squares, Lurking variables, Effect of serial correlation

*Edited by George Box and Søren Bisgaard.

This work was sponsored by grants from the Alfred P. Sloan Foundation, Procter and Gamble and NSF grant #DMI-9414765.

Quality Quandaries*

Regression Analysis Applied to Happenstance Data

George Box

Care is needed in interpreting results when regression analysis is applied to happenstance data. Although a fitted regression model may be useful for prediction of future values of a series, it may be totally misleading for explaining causation relationships between the variables. Furthermore such analysis can be disastrously affected by lack of independence between residual errors.

The method of least squares was invented by Gauss about 1796 and is perhaps the most important single tool of statistics. Although not always evident, it lies behind the analysis of most designed experiments and many other things that we do. Particularly when least squares is used in the analysis of unplanned data (for example, past plant data or economic records), it is often referred to as *regression analysis*. In such an analysis an output or response variable y, such as the yield of a batch process, is supposed to be linked to a number of inputs such as the temperature x_1 and the pressure x_2 used in the preparation of that batch by a model usually of the form

$$y = b_0 + b_1 x_1 + b_2 x_2 + e .$$

In most real examples there are more than two x's but I can illustrate what I need to say in terms of just two. In this equation b_0, b_1, and b_2 are constants (regression coefficients) which are to be estimated and e is an overall error. Regression analysis uses the method of least squares to find those values of the constants b_0, b_1, and b_2 which give the best fit in the sense that they make the sum of squares of the errors as small as possible. Whether the data gathering is planned or unplanned the standard analysis assumes that the errors e each have mean zero, the same standard deviation σ, and are *totally uncorrelated* with each other. To justify the associated t -tests and F-tests that are often applied in the statistical analysis, we also need the assumption that the errors are normally distributed. But contrary to common belief this assumption is less important than those mentioned earlier.

Now, in fact, the error e really represents the effect of what I like to call the *lurking variables* (1). Those are variables - x_3, x_4, and so on - which we have *not* measured often because we don't even know they exist. In practice there could be hundreds of *lurking variables* but for simplicity I will suppose there are only two and I'll call them x_3 and x_4 so that

$$e = c_3 x_3 + c_4 x_4 .$$

The situation is illustrated in Figure 1 where the lurking variables x_3 and x_4 are shown "hidden behind" the wall – frequently what goes on behind the wall nobody really knows.

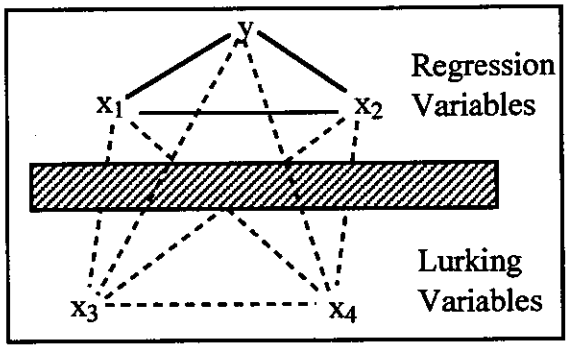

***Figure 1.** The apparent relationship among y, x_1, and x_2, can be affected by unknown links to lurking variables.*

In this diagram the links that the regression equation is specifically looking for are shown as bold lines. Those behind and across the wall about which you have no knowledge, are shown as dotted lines.

The linkages could indicate causative relations. For example, an increase in temperature might *produce* an increase in pressure, or they might just indicate relationships due to correlation. For example, the flow of one of the reactants might be reduced whenever a certain temperature was observed to be high because the operator had been told to run the plant that way.

Now, a very important question to ask is "What do we wish *to do* with the fitted regression equation?" We might:

a) Desire to predict y from passive observation of x_1 and x_2 in the future, assuming the causative and correlative system which operated when the original data was taken has not changed.
b) Wish to discover how to make deliberate *changes* in x_1 and x_2 to *modify* the system and get an improved value for y.

The position is quite different depending on whether prediction from passive observation, or improvement from active intervention, is in mind. This is made clear by an example. Suppose that in a chemical process it has been found that undesirable frothing can be reduced by increasing the pressure x_1. The standard operating procedure is, therefore, to increase pressure whenever frothing appears. Suppose that the amount of frothing actually depends on the impurity x_3 in the feed stock (which is, of course not measured because it is unknown). Finally, suppose that this impurity x_3 not only produces frothing y_2, but also lowers yield. Moreover the yield y_1 is not affected directly by changing pressure x_1. The situation is diagrammed in figure 2.

Now a regression analysis of yield y_1 on pressure x_1 is likely to show a highly significant relationship. The two variables y_1 and x_1 are not *causally* linked. However, provided the system *continues to be run in the same fashion as when the data were recorded* we could use such a relationship to *forecast* the level of the yield y_1 from the pressure x_1 in use on that day. It is, after all, true that on those days when the pressure is high, we usually *will* get a lower yield. On the other hand, as a guide to *action* this regression equation will be utterly misleading. If we hope to increase yield by reducing pressure clearly we will be disappointed.

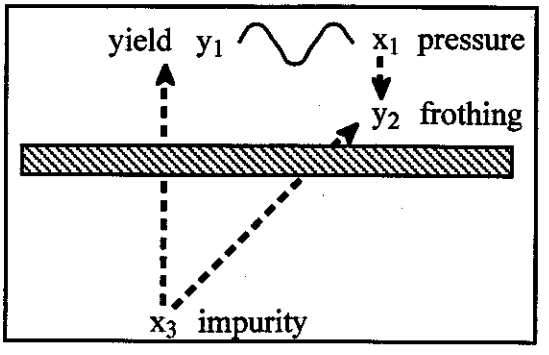

Figure 2. *The lurking variable (x_3) indicates the concentration of an (unknown) impurity which independently causes frothing (y_2) and low yield (y_1). Pressure (x_1) is used to control frothing but has no causative relationship to yield. Such a system can produce a significant regression relationship between x_1 and y_1 which can be used to forecast yield but not to change it. Arrows indicate causative relations and the wiggle line indicates the regression relationship.*

But the bad news does not stop here. Even if we use standard regression methods to fit a model and we appear to obtain highly significant estimates of the coefficients $b_1, b_2, \ldots$, we cannot be sure that we can safely use the model even for prediction. This depends on whether the least squares assumptions have been violated. From elementary statistics courses, we have become so accustomed to modeling data as independent that we often do not stop to consider that this assumption *is not sensible* for serial data. If I look at this piece of paper I'm writing on, and I close my eyes and then look again one second later, it looks pretty much the same as before. I have made a second observation, but have not obtained further information. Observations taken in sequence tend not to be independent but to be positively serially correlated. If data of this kind are treated as if they were independent, then we will greatly overestimate the information they contain and statistical tests can hugely overestimate levels of significance. Unfortunately ordinary least squares makes precisely this assumption of independence.

To give an example, in a paper appearing some time ago (2) it was argued that the stock price y

could be forecast from a regression equation containing x_1 (car production six quarters previously) and x_2 (the consumer price index seven quarters previously).* The regression equation based on 51 successive observations gave t values for the two coefficients b_1 and b_2 of respectively 11.8 and 9.9 indicating enormously significant results and which could lead one to believe that the equation could accurately forecast future levels of stock price. Unfortunately, the data are highly serially correlated and when this was allowed for the t values became 1.8 and 1.0 respectively (3). This indicates, as one would (reluctantly) expect, that there is no evidence of any such relationship (or of the existence of Santa Claus).

It was to overcome such difficulties as those described that Fisher introduced the idea of designed experiments and, in particular of randomization. Here levels of the regression variables are deliberately set *at certain fixed levels given by the design* and run in *random order*. The only *cause* of the particular values which the regression variables take is the randomization process itself. For instance, in the example illustrated in Figure 2, if we had been able to run a designed experiment in which we temporarily dealt with frothing in some other manner and varied randomly the order in which pressure was run at the high and low levels, the "nonsense" correlation between pressure and yield would no longer be produced. Randomization makes it possible to analyze the data *as if* the standard assumptions were true.

These difficulties are by no means the only ones that face us in the analysis of unplanned data. In the operation of an industrial process past experience often shows that certain variables are of major importance. In the routine operation of the process, therefore, care is taken to hold these variables very close to fixed values. As the statistical significance of any variable is greatly affected by the range it covers, it is likely therefore that these most important variables will be dubbed "not significant" by a standard regression analysis. A further difficulty is that with unplanned data, the *regression* variables will frequently be highly correlated with each other, and it may then be almost impossible to discover whether changes in y are associated with x_1 or with x_2, or with both. In designed experiments, of course, one normally arranges that x_1 and x_2 are uncorrelated by using an orthogonal design. See also BH^2 (4).

In the days when computing power consisted of a hand calculator, regression analysis was hard to do and so there was a certain healthy restraint on its use. These days every computer has a regression program and can produce sense (or nonsense) almost instantaneously.

In summary the regression analysis of unplanned data is a technique which must be used with care.

i) it may provide a useful *prediction* of y for a fixed system being passively observed even when lurking variables of some importance exist. However, relationships obtained from serial data are very liable to be spurious unless serial correlation is properly allowed for.

ii) it is one of a number of tools sometimes useful in indicating variables which should be included in some planned experiment to be performed later (in which, if possible randomization will be included as an integral part of the design). However, for reasons obvious from the above, such analysis ought never be used as the sole basis to decide which variables should be *excluded* from further investigation.

To find out what happens to a system when you interfere with it you have to interfere with it (not just passively observe it). If possible you should randomize your interference.

ACKNOWLEDGMENT

This work was sponsored by a grant from the Alfred P. Sloan Foundation.

REFERENCES

Box, G.E.P. (1965), "Use and Abuse of Regression," *Technometrics*, **8**, 625-629.

Coen, P.G. Gomme, E. D., and Kendall, M.G. (1969). "Lagged relationships in economic forecasting." *Journal of the Royal Statistical Society A* , **132**, 133-163.

* The analysis was performed after a linear trend in the data had been removed.

Box, G.E.P. and Newbold, P. (1970). "Some comments on a paper by Coen, Gomme, and Kendall." *Journal of the Royal Statistical Society A*, **134**, 229-240.

Box, G.E.P., Hunter, W.G. and Hunter, J.S. (1978). *Statistics for Experimenters*, New York: Wiley.