TECHNOMETRICS VOL. 8, NO. 4 NOVEMBER 1966

Use and Abuse of Regression[†]

GEORGE E. P. BOX
University of Wisconsin

Let us first restate the usual assumptions and conclusions for linear least squares. Gauss showed that if we have n observations $y_1, y_2, \cdots, y_n$ and *if* an appropriate model for the uth observation is

$$y_u = \beta_0 + \beta_1 x_{1u} + \beta_2 x_{2u} + \cdots + \beta_k x_{ku} + \epsilon_u \tag{1}$$

where the β's are unknown parameters, the x's known constants, and the ϵ's random variables uncorrelated and having the same variance and zero expectation, then estimates $b_0, b_1, \cdots, b_k$ of the β's obtained by minimizing $\sum (y - \hat{y})^2$ with $\hat{y} = b_0 x_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$ are unbiassed and have smallest variance among all linear unbiassed estimates.

The method of least squares is used in the analysis of data from planned experiments and also in the analysis of data from unplanned happenings. The word "regression" is most often used to describe analysis of unplanned data. It is the tacit assumption that the requirements for the validity of least squares analysis are satisfied for unplanned data that produces a great deal of trouble. Whether the data are planned or unplanned the quantity ϵ, which is usually quickly dismissed as a random variable having the very specific properties mentioned above, really describes the effect of a large number of "latent" variables $x_{k+1}, x_{k+2}, \cdots, x_m$ which we know nothing about. If we suppose that it is enough to consider the linear effects of these latent variables (which would often be realistic for small variations in $x_{k+1}, \cdots, x_m$) we should have

$$\epsilon = \beta_{k+1} x_{k+1} + \beta_{k+2} x_{k+2} + \cdots + \beta_m x_m \tag{2}$$

Thus in matrix notation we can write for the column of n observations $\mathbf{y}$

$$\mathbf{y} = \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{X}_2 \boldsymbol{\beta}_2 \tag{3}$$

where $\mathbf{X}_1$ has for elements the n values of the k regression variables and $\mathbf{X}_2$ has for elements the n unknown values of the $m - k$ latent* variables. The situation is illustrated in Figure 1 in which the variables $x_{k+1} \cdots, x_m$ are "hidden behind the wall." In practice various kinds of linkages would occur between the variables indicated by lines. These linkages might indicate causative relations; for instance, an increase in temperature might necessarily produce an increase

Received March 1966.

[†] A talk prepared for the tenth conference on the Design of Experiments in Army Research, Development and Testing. Washington, D. C. November 1964.

* More dramatically described as "lurking" variables.

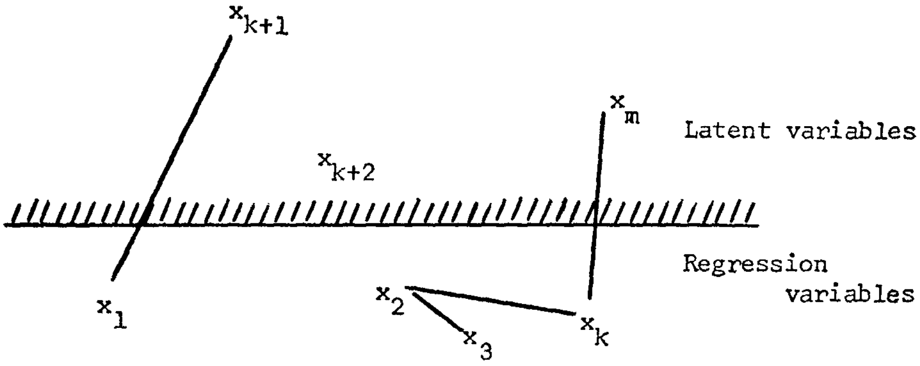

FIGURE 1 Latent variables and regression variables.

in pressure; or merely relationships due to correlation. Thus, an operator in charge of a process might *as a standard operating procedure* always reduce the flow of one of the reactants if a certain temperature was observed to be high.

We must now ask the question, "What do we wish to do with the fitted regression equation?" We might

(i) desire to predict y in the future from passive observation of $x_1 \cdots x_k$. We assume that the causal and correlative system which operated during the data taking has not been interferred with and also operates during the period when predictions are being made.

(ii) to discover how deliberate *changes* in $x_1 \cdots x_k$ will effect y with the intention of actually *modifying* the system to get a better value for y.

The position is quite different depending upon whether prediction from passive observation or improvement from active interference is in mind. This is made clear by the following example.

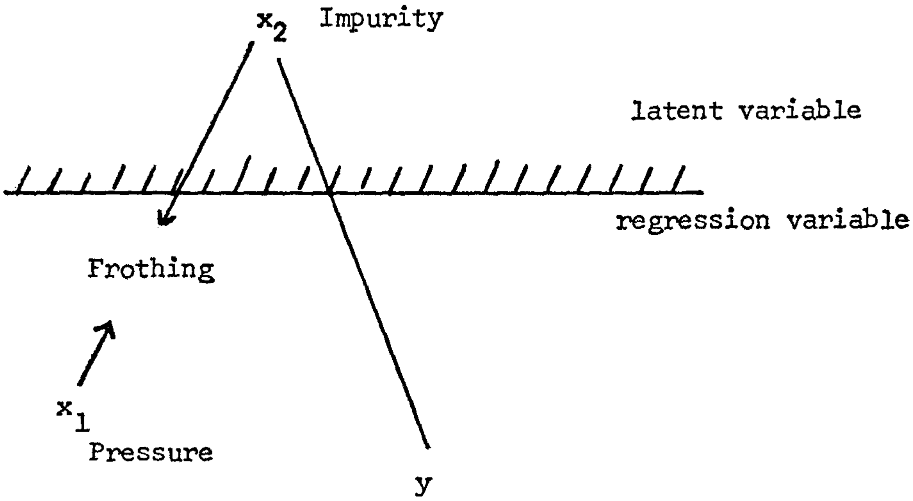

FIGURE 2 Relations between yield, impurity, and pressure.

Suppose that in a chemical process it has been found that undesirable frothing can be reduced by increasing pressure. The standard operating procedure is, therefore, to increase pressure whenever frothing appears. Suppose that the frothing in fact occurs because of an unsuspected impurity x_2 (which is, of course, not measured because it is unknown). Suppose finally that a high value of impurity x_2 not only produces frothing but also lowers yield but that yield is unaffected directly by a change in pressure.

If (with now x and y representing deviations from respective averages) a "regression of yield on pressure" $\hat{y} = b_1x_1$ is fitted by the usual least squares procedure we may well find a highly significant coefficient b_1 .

This well known phenomenon of "nonsense" correlation exhibited in this example is worth studying further. Suppose there is a relationship $y = \beta_1x_1 + \beta_2x_2$ connecting y *exactly* with the two variables x_1 and x_2 (with, in the present instance, $\beta_1 = 0$). Now, of course, the actual levels of x_2 are unknown but suppose that $\hat{x}_2 = ax_1$, is the formal regression of x_2 on x_1 which would be obtained if values of x_2 *were* available. Then it is readily shown that

$$b_1 = \beta_1 + a\beta_2 \tag{4}$$

In this expression β_1 is zero and we appear to obtain a real effect only because of the influence of the bias term ($a\beta_2$). On the other hand using (4) we see that our fitted equation $\hat{y} = b_1x_1$ which ignored x_2 can be written $\hat{y} = \beta_1x_1 + \beta_2ax_1$ or as

$$\hat{y} = \beta_1x_1 + \beta_2\hat{x}_2 \tag{5}$$

This equation which replaces x_2 by $\hat{x}_2$ is the best estimate of y we can expect to get from observing x_1 only. Provided the system *continues to be run in the same fashion as when the data were recorded* we can use pressure to indicate the level of y. Of course, if we had measured x_2 a more accurate (indeed in the present instance an exact) value of y would be deducible, but, lacking knowledge of the importance of x_2 , we might nevertheless appropriately use the simple regression equation $\hat{y} = b_1x_1$.

On the other hand the value of b_1 will be utterly misleading if interpreted as the effect on the variable y of a unit *change* in x_1 . If we hope to increase yield by reducing pressure we will be disappointed.

A similar argument applies for any number of variables. The true model is

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 \tag{6}$$

By including only the variables X_1 in the regression equation our prediction equation for y becomes

$$\hat{\mathbf{y}} = \mathbf{X}_1\mathbf{b}_1 = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} \tag{7}$$

$$= \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2) \tag{8}$$

i.e.

$$\hat{\mathbf{y}} = \mathbf{X}_1\boldsymbol{\beta}_1 + \hat{\mathbf{X}}_2\boldsymbol{\beta}_2 \tag{9}$$

where $\hat{\mathbf{X}}_2 = \mathbf{X}_1\mathbf{A}$ and $\mathbf{A} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2$ is the $k + 1 \times m - k$ matrix of regression coefficients of the latent variables on the regression variables.

Again we see that so far as the passive prediction of $\hat{\mathbf{y}}$ is concerned our simple regression onto the known variables $\mathbf{X}_1$ in effect replaces the unknown $\mathbf{X}_2$ by $\hat{\mathbf{X}}_2$.

On the other hand the regression coefficients $\mathbf{b}_1 = \boldsymbol{\beta}_1 + \mathbf{A}\boldsymbol{\beta}_2$ represent combinations of effects due to regression variables and latent variables and as before it is impossible to draw any valid conclusions as to how interference with the levels of the regression variables will affect the system.

In a designed experiment, we are in quite a different case. It was, of course, to overcome such difficulties as those described above that Fisher introduced the idea of designed experiments and in particular of randomization. When the levels of the regression variables are chosen in some deliberately random manner it is impossible for the levels of a regression variable to be affected by the level of a latent variable. The only cause of the particular values which the regression variables have within the design framework is the throw of an unbiassed die or other random process. Fisher makes it possible to analyze the data *as if* Gaussian assumptions were true by making $\mathbf{X}_1$ a random variable. The regression variables can, of course, still affect the latent variables and these may in turn effect y. Provided, however, we apply our results to the same system for which we obtained our data this will cause no problem. It will be genuinely true that apart from experimental error *manipulation* of regression variables will produce the predicted change in y even though it does it via some latent variable.

The basic difficulty mentioned above is by no means the only one that faces 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 order to control fluctuations in the process, therefore, care is taken to hold precisely these variables very close to fixed values. As the "statistical significance" of any variable is greatly affected by the range it covers there is a strong probability, therefore, that the most important variables will be dubbed "not significant" by a standard regression analysis. A further difficulty is that with unplanned data regression variables will frequently be highly correlated only because of operating policy. The operator is told to reduce x_2 whenever x_1 becomes high. With such data even if difficulties from latent variables could be ignored it may be almost impossible to discover whether changes in y are associated with x_1, 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.

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

(i) It may provide a useful prediction of y in a fixed system being passively observed even when latent variables of some importance exist. For this application computer programs which progressively add or drop variables make some sense.

(ii) It is one of a number of tools sometimes useful in indicating variables which ought to be included in some latter planned experiment (in which randomization will, of course, be included as an integral part of the design). It ought never to be used to decide which variable should be excluded from further investigation for reasons which are obvious from the above.

To find out what happens to a system when you interfere with it you have to interfere with it (not just passively observe it).

Reference

1. Fisher, R. A., 1937. *Design of Experiments,* published by Oliver and Boyd.