DEPARTMENT OF STATISTICS

University of Wisconsin
Madison, Wisconsin

Technical Report No. 46
January 1965

USE AND ABUSE OF REGRESSION

by George E. P. Box

USE AND ABUSE OF REGRESSION

Let us first restate the usual assumptions and conclusions for linear least squares. Gauss showed that if we have \( n \) observations \( y_1, y_2, \ldots, y_n \) and if an appropriate model for the \( u \)th observation is

\[
y_u = \beta_0 + \beta_1x_{1u} + \beta_2x_{2u} + \ldots + \beta_kx_{ku} + \epsilon_u
\]

where the \( \beta \)'s are unknown parameters, the \( x \)'s known constants, and the \( \epsilon \)'s random variables uncorrelated and having the same variance and zero expectation, then estimates \( \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k \) of the \( \beta \)'s obtained by minimizing

\[\Sigma(y - \hat{y})^2\]

with \( \hat{y} = \hat{\beta}_0x_0 + \hat{\beta}_1x_1 + \hat{\beta}_2x_2 + \ldots + \hat{\beta}_kx_k \) are unbiased and have smallest variance among all linear unbiased 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 \( \epsilon \), 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 \( x_{k+1}, x_{k+2}, \ldots, x_m \) 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 \( x_{k+1}, \ldots, x_m \))
we should have

$$
\epsilon = \beta_{k+1}x_{k+1} + \beta_{k+2}x_{k+2} + \ldots + \beta_{m}x_{m}
$$

(2)

Thus in matrix notation we can write for the column of \(n\) observations \(y\)

$$
y = X_1\beta_1 + X_2\beta_2
$$

(3)

where \(X_1\) has for elements the \(n\) values of the \(k\) regression variables and \(X_2\) has for elements the \(n\) unknown values of the \(m - k\) latent variables. The situation is illustrated in Figure 1 in which the variables \(x_{k+1}, \ldots, x_{m}\) 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 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

---

![Diagram](image)

**Figure 1:** Latent variables and regression variables.
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 \( x_1 \ldots x_k \). We assume that the causal and correlative system which operated during the data taking has not been interfered with and also operates during the period when predictions are being made.

(ii) to discover how deliberate changes in \( x_1 \ldots x_k \) 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.

![Diagram]

Figure 2: Relations between yield, impurity, and pressure.
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

\[ y = X_1 \beta_1 + X_2 \beta_2 \]  \hspace{1cm} (6)

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

\[ \hat{y} = \hat{X}_1 \hat{\beta}_1 = X_1 (X_1'X_1)^{-1} X_1' y \]  \hspace{1cm} (7)

\[ = X_1 (X_1'X_1)^{-1} X_1' (X_1 \beta_1 + X_2 \beta_2) \]  \hspace{1cm} (8)

i.e.

\[ \hat{y} = \hat{X}_1 \hat{\beta}_1 + \hat{X}_2 \hat{\beta}_2 \]  \hspace{1cm} (9)

where \( \hat{X}_2 = X_1 A \) and \( A = (X_1'X_1)^{-1} X_1'X_2 \) 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 \( \hat{y} \) is concerned our simple regression onto the known variables \( X_1 \) in effect replaces the unknown \( X_2 \) by \( \hat{X}_2 \).
On the other hand the regression coefficients $b_1 = \beta_1 + A \beta_2$
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 unbiased die or other
random process. Fisher makes it possible to analyze the data as if
Gaussian assumptions were true by making $X_1$ 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
takes 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 later 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).

REFERENCES