Present Position and Potential Developments: Some Personal Views: Design of Experiments and Regression

Author(s): D. R. Cox


Published by: Wiley for the Royal Statistical Society

Stable URL: http://www.jstor.org/stable/2981685


JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at http://about.jstor.org/terms

Royal Statistical Society, Wiley are collaborating with JSTOR to digitize, preserve and extend access to Journal of the Royal Statistical Society. Series A (General)
Present Position and Potential Developments: Some Personal Views
Design of Experiments and Regression

By D. R. COX

Fisher Memorial Lecture

[The Chairman of the Fisher Memorial Committee, Dr W. F. Bodmer, in the Chair]

SUMMARY

The first part of the paper gives a brief historical comment on the development of experimental design and some rather more extensive remarks on randomization. The second and longer part of the paper discusses various difficulties primarily of interpretation connected with regression analysis (in the broad sense). These include the selection of variables, the effect of complex error structure and the effect of errors in the explanatory variables. As an Appendix a few apparently open problems are outlined.

Keywords: BLOCKING; COMPLEX ERROR STRUCTURE; CORRELATION; ERRORS IN VARIABLES; GENERALIZED LINEAR MODEL; LINEAR MODEL; MODEL CHOICE; OBSERVATIONAL STUDY; OPTIMAL DESIGN; RANDOMIZATION; REGRESSION; STRATIFICATION

1. INTRODUCTION

I deeply appreciate being invited to give a Fisher Memorial Lecture.

The comprehensive discussion of design of experiments and of regression half-promised by the title is scarcely feasible. The first and shorter part of the paper makes some comments on experimental design and this is followed by a rather longer, although still highly selective, discussion of regression.

As an Appendix, some apparently unsolved problems are listed.

2. DESIGN OF EXPERIMENTS

For the applied statistician, design of experiments is one of the most interesting and satisfying aspects of our subject. In part this is because involvement in the planning phases of an investigation is usually a symptom of genuine "collaboration", as contrasted with "consultation" over difficulties of analysis at the late stages of a study.

One can distinguish three broad periods of innovation directly associated with applications. Firstly, there is the period up to say 1950 with a primary motivation from agricultural and biometrical experimentation, the books of Fisher (1935) and Yates (1937) being outstanding accounts from that time. Recent work, much of it unpublished, by R. A. Bailey and T. P. Speed, sets many of the classical ideas in a rather "high-powered" mathematical framework. A key distinguishing feature of the "classical" approach is the notion that the analysis of variance table is a consequence of the treatment and unit structure and the randomization, rather than a deduction from an externally specified linear model. More broadly, while the very close links between analysis of variance and the theory of the linear model are of key importance, it would considerably undervalue analysis of variance to regard it as merely an appendage to the theory of the linear model.

Secondly, there is a period between about 1950 and 1970 with a primary motivation from...
industrial experimentation. Here the emphasis has tended to be on treatment structure (response surface form and fractionation) and rather less on the detailed character of the error.

Finally, in recent years there has been much emphasis on clinical trials. Characterizing features are that, superficially at least, treatment structure is simple, experimental units (patients) enter serially, measurement of response may be delayed and that very often a large amount of prognostic information is initially available on each unit.

Of course this division into three phases does not imply lack of continuing interest, theoretical and applied, in the traditional areas. For an agricultural example, see Bartlett (1978) and Wilkinson et al. (1983) on the relation of field trials with spatial stochastic models. Here there are major unresolved issues as to both the practical fruitfulness of the approach, and the merits of rival superficially very similar methods of analysis. The role of cross-over designs attracts continuing attention and controversy in several applied fields. In industrial experimentation, in the UK at least, no doubt the major task is to try to get careful techniques of experimentation more widely adopted; so far as one can judge, basic ideas of experimental design are employed on an enormous scale in Japanese industrial research and development.

From a mathematical point of view, by far the major development in the past 20 years has been the flowering of a systematic theory of optimal experimental design; see Silvey (1980) for a concise account. This theory serves partly to consolidate and codify knowledge about traditional designs and, perhaps more importantly from the viewpoint of immediate application, offers a systematic approach to non-standard problems. For example, suppose that in an experiment with quantitative factors the design region is modified after an initial design has been partly implemented: how should the new design points be chosen? Packaged algorithms, hopefully merely mildly menacing to the user, are needed if such problems are to be tackled routinely under normal working conditions; see Welch (1982).

Limitations on the practical usefulness of designs derived via a formal optimality criterion arise mainly from the provisional character of aspects that have to be precisely specified in order to implement an algorithm. The situation is not essentially different from that with optimality considerations in other practical contexts; for the controversy in operational research between optimizing and so-called satisficing, see Eilon (1972).

On the whole, the theory of optimal design makes rather strong a priori assumptions about the nature of error variability. It has not, for example, yet thrown special light on the role of randomization. That role has always been to some extent controversial; see, for example, the discussion between Fisher and Student in the 1930s.

The least controversial aspect of randomization is probably its role in eliminating systematic error, in particular by concealment. The empirical evidence for biases arising from the entry of personal choice in treatment assignment is so strong that failure to use some impersonally defined procedure can be damning; some element of randomization is often, although not always, the natural solution. The second-order theory of randomization, in conjunction with an assumption of unit-treatment additivity, yields, as noted above, a satisfying unity to the analysis of the classical designs. The relevance of the randomization distribution to the analysis of the design actually employed depends, however, on the absence of a "recognizable subset" in the randomization distribution. Here "recognizable" means with reference to some feature which there is a good scientific reason to judge relevant to the responses.

Put slightly differently, if there are "background ancillaries" the randomization distribution should be conditional on their observed values. Background ancillaries are features of the design for which there is reason, whether based on a formal model or not, to think that their value materially affects the precision of the treatment contrasts of interest. If such an ancillary is recognized at the design stage, then optimum or near-optimum values should be employed. If it is not recognized until the analysis stage, then any randomization analysis should be conditional on values equal or close to those realized in the design employed.

To the first order of asymptotic theory this is achieved via the asymptotic multivariate normal randomization distribution of the background ancillaries and the standard functions of the
response variable. The approximate conditional distribution of the latter can thus be found, leading to a randomization version of the standard analysis of covariance adjustments.

For traditional medium-sized experiments with a limited number of prognostic features per experimental unit, blocking and randomization can be effective in controlling error. In clinical trials, however, quite apart from the special issue of serial entry, there may be difficulties of design associated with the availability of a large number of prognostic features, leading to a conflict between the desire for randomization and the need for balancing or conditioning. Peto, et al. (1976) recommended randomization with blocking on only a few key features, correcting by analysis of covariance for any serious imbalances that might arise. There has been some controversy (Simon, 1979) over this, especially for medium-sized trials where any need for substantial adjustments would both lower efficiency and reduce the intuitive appeal of the analysis.

Cox (1982a) suggested that the appropriate conditional randomization could be formally achieved by repeated rerandomization until a suitably defined measure of imbalance was sufficiently small; an approximate conditional randomization analysis would then in principle be possible.

An explicit notion of conditional randomization clarifies also the limited usefulness of randomization in small experiments, where any design in a potential randomization set is likely to be meaningfully different. It bears also on the desirability of rejecting "bad" designs thrown up by the randomization.

The penetration of statistical ideas on experimental design is very uneven as between different fields of application; particularly patchy areas are the physical sciences and biochemistry. It is possible that new developments or at least changes of emphasis are needed for these subjects.

A final general comment is that, while the distinction between experimental and observational studies is crucial and not to be blurred, there is much scope for parallel development and for improved techniques in observational studies imported partly or wholly from experimental design (Cochran, 1983). See, for example, Rosenbaum and Rubin's (1983) discussion of the possible effect of unobserved covariates on an observational study with binary response. Also, many of the issues connected with randomization in experimental design have parallels in sampling theory although the styles of discussion in the two literatures are surprisingly different.

For a thorough and lucid review of recent theoretical work on design of experiments, see Atkinson (1982).

3. REGRESSION ANALYSIS

3.1. Introductory Comments

Discussions of terminology and of refined distinctions between neighbouring subject areas are sometimes necessary and occasionally important, although possibly more often counter-productive and tedious. A distinction fairly widely accepted and surely important is that between regression and correlation. Suppose that on each individual there are two kinds of observation $y$ and $x$, both possibly vectors:

(a) regression analysis is concerned with how the conditional distribution of $y$ given $x$ varies with $x$, in particular with how the "centre" of that distributions varies with $x$, and, in particular, with how the expected value varies with $x$;

(b) correlation analysis is concerned with departures from independence in the joint distribution of $y$ and $x$, the two being treated on an equal footing.

Of course the product moment correlation coefficient as normally defined has a fairly direct interpretation in regression analysis. On the whole regression analysis is much simpler and more incisive than a comparable correlation analysis, essentially because the dimensionality of the distribution under analysis is reduced. Thus log linear analysis of contingency tables is a form of correlation analysis, whereas the singling out of one variable as a basis say for logistic regression analysis is an instance of regression analysis; it is inviting confusion to use the first when the second is appropriate.
Has the time come to abandon the term "regression" in favour of "dependence", i.e. to talk about dependence coefficients rather than regression coefficients, and so on? The older term "independent variable" for x has obvious drawbacks and seems to be passing out of use. To be attracted by regressand and regressor might be taken as indirect evidence of a biased education; response and explanatory variables seem the most evocative terms.

Linear regression analysis based on the method of least squares remains the most important and widely used and studied form of regression analysis. Many of the ideas, however, apply much more widely; see McCullagh and Nelder (1983) for an excellent account emphasizing so-called generalized linear models. Note, however, that many of the more complicated aspects such as hierarchical error and errors in the explanatory variables have not yet been studied in any generality.

It is trite that specification of the purpose of a statistical analysis is important. Several such purposes can be distinguished for regression analysis including:

(a) the derivation of a concise description of a body of data useful, for instance, in comparing different similar sets of data, and for characterizing individuals as high or low responses relatively to what might have been expected in the light of their explanatory variables;

(b) as a basis for prediction (or calibration) for new individuals;

(c) the estimation of regression coefficients, including the identification of those explanatory variables with "important" effects.

Aspect (c) is closely allied to the prediction for a new individual of the consequences for the response of that individual to changing values of certain components of x, i.e. of intervention. It is the formulation implicit in comparative experiments and therefore also for observational studies performed in lieu of experiments. For that reason (c) will be given primary emphasis in this paper. We may call applications of that type analytical studies.

General issues connected with regression have been discussed by many authors. The following comments draw partly on Cox (1968) and Cox and Snell (1974).

3.2. Choice of Representation

Quite often, especially in observational studies, the dimension of x is fairly large and use of the "full" regression equation is for one reason or another inadvisable. How to deal with this situation is the aspect of regression that causes most concern in applications. To a large, although not entire, extent this problem is a product of the computer era, which has made large-scale fitting by least squares or maximum likelihood a fairly routine matter. Fisher (1938), however, fitted models containing seven explanatory variables (sea-level characteristics of members of an expedition) to explain as response, acclimatization at high altitude for nine individuals; a plausible estimate of error variance was available from the variation within individuals. See Draper and Joiner (1984) for a reanalysis.

The following miscellaneous comments are phrased in terms of least squares regression but are quite broadly applicable to analytical studies, i.e. to applications in which prediction or calibration is not an immediate aim.

(i) It will frequently be sensible to begin by dividing the explanatory variables into treatment variables, intrinsic variables (characterizing the individuals or exogenous, characterizing their environment) and non-specific variables identifying blocks, strata, litters and so on. In observational studies treatment variables identify aspects that would be introduced as treatments in a comparable experiment. Usually interest will focus on the effect of the treatment variables, and their interaction with intrinsic variables.

For example, in an observational study of the effects of smoking and alcohol in pregnancy on the infant's birth weight, measures of smoking and alcohol consumption would be treatment variables, age and sex intrinsic variables; social class would presumably best be considered as a surrogate treatment variable, hopefully incorporating some of the unmeasured variables that in an experimental study would have been controlled by randomization. In the teaching styles study (Aitkin et al., 1981), variables characterizing teachers are treatment variables, initial test
scores on children are intrinsic variables and schools or classes are nonspecific.

(ii) A key issue concerns the introduction of external information about the explanatory variables and their effect. While it would be a mistake to make sweeping statements about what is wise in all practical contexts, on the whole it seems to me best to use qualitative external information about the problem, of the kind set out below, especially in high-dimensional problems where the data alone are incapable of adequate resolution. Where assumptions are made, such as that a battery of related variables can be replaced by one of them, they should, where feasible, be tested for conformity with the data.

The development of a systematic approach to these matters of qualitative judgement is a suitable topic for a computerized "expert system" (IKBS).

A consequence of the inclusion of qualitative external information is that automatic variable selection rules have a role largely limited to the initial reduction of systems with so many explanatory variables that careful examination of all adequately fitting models is not feasible.

There is a broad distinction between:

(a) methods in which some explanatory variables are omitted and some included in the final representation, those included being given "full weight" in a least squares or maximum likelihood procedure;

(b) methods in which all (or most) of the explanatory variables are included, but some have effects "shrunk" towards zero or some convenient reference level.

The occurrence in (b) of a continuous range of possibilities between complete inclusion and complete exclusion, rather than an all or nothing choice, is attractive from a commonsense point of view. It has, of course, an extensive literature under the name ridge regression and a fairly immediate and qualitatively appealing Bayesian interpretation is available. The barriers to practical use are, however, formidable unless the explanatory variables are all measured in comparable units.

We now discuss some of the qualitative considerations that can be useful.

(iii) Where there are several somewhat similar response variables, it will usually be wise to use the same explanatory variables for all. Thus one might take the union of the sets that seem appropriate for the response variables taken individually. For instance, in a study of the size of newborn infants, log weight, log length and log head circumference might be three response variables of interest. While somewhat different sets of explanatory variables might be included by separate analysis of the three variables, interpretation will probably be aided and the chance of overinterpretation reduced by using a common set of explanatory variables for the three analyses. (This is quite a separate issue from that of whether techniques of multivariate analysis are called for.)

(iv) If it appears that one or more suitable subsets of explanatory variables have been isolated as a basis for interpretation it will often be wise to add the omitted variables back one at a time and to examine the resulting estimates and their standard errors.

(v) When "main effects" have been isolated, it will be sensible to look for interactions, especially for treatment \times treatment and treatment \times specific variable interactions, where the corresponding main effects are appreciable (Cox, 1984).

(vi) When there are intermediate response variables included among the explanatory variables, a path model (involving conditional independencies) should if possible be formulated and tested. For contingency tables, see Edwards and Kreiner (1983).

(vii) Where an intrinsic explanatory variable with qualitative levels (e.g. sex) has an appreciable effect, it will be for consideration whether to split the data into sections for separate analysis.

It seems to me that there is some urgency for the incorporation into regression analysis packages of simple-minded guidelines, such as the above or improved versions thereof.

3.3 Effect of Intervention

A fairly common use of multiple regression is to predict not the outcome for a new individual, but rather the effect of an intervention, i.e. the imposition of a change in the system. Suppose
that the explanatory variable $x_1$ is to be changed from $x'_1$ to $x'_1 + d_1$. A crucial question concerns the consequence for the other explanatory variables:

(a) any variables expected to change with $x_1$ in accordance with the same stochastic relations as hold in the baseline data are to be omitted from the regression equation;

(b) any variables to be held fixed should be included in the regression equation or be shown likely to have negligible effect;

(c) there remain variables that may change on intervention but not in the way holding in the data under analysis.

For (c) it is of course necessary to have some data about, or to make some assumptions concerning, the changes induced. It is important also, and often a severe limitation on empirical studies, that major changes in unmeasured explanatory variables may take place on intervention, other than those implicitly represented in the regression equation.

A much oversimplified example is where for a number of individuals a response $y$ is measured and explanatory variables $x_1$ (a binary variable smoker, non-smoker); $x_2$, height; $x_3$, some aspect of diet. What is the effect for an individual of giving up smoking? Variable $x_2$ is of type (b), whereas diet is likely to change but not necessarily in the way indicated by the "between individual" variation. Plausible estimates must be made of the consequent change in $x_3$. Finally suppose that it is suspected that the unmeasured variable weight is important and that height appears in a direct regression equation solely as a surrogate for weight. Then, because weight may change after intervention, some attempt to estimate its effect is desirable.

Formally, in general, if $\beta$ is the vector of parameters in the linear model, estimated by $\hat{\beta}$ with covariance matrix $\Omega_\beta$, we are interested in $\alpha^T \beta$, where $\alpha$ is estimated by $\hat{\alpha}$ with covariance matrix $\Omega_\alpha$. If random errors in $\alpha$ and $\beta$ are uncorrelated

$$E(\hat{\alpha}^T \hat{\beta}) = \alpha^T \beta,$$

$$\text{var}(\hat{\alpha}^T \hat{\beta}) = \alpha^T \Omega_\alpha \alpha + \beta^T \Omega_\beta \beta,$$

so that approximate assessment of precision is possible. No doubt these formulae could be refined.

3.4. Structure of Error

In the context of second-order multiple regression there is a fairly well-established set of procedures for dealing with error variation in which the covariance matrix departs from the simple form $\sigma^2 I$ and in which it is possible to estimate the covariance matrix reasonably accurately. It is important to distinguish between the effect of the covariance matrix on the efficiency of the ordinary least squares estimate and its effect on estimated precision.

The extension of these results to generalized linear models is largely open. Sometimes, however, weighted least squares methods can be adapted after grouping of individuals into sets with approximately the same values of the explanatory variables. Empirical linearizing transformations, e.g. empirical logit transforms, can be found and analysed approximately, provided, of course, that the grouping adopted lends itself to a simple representation also of covariance structure. The widely appreciated problem of overdispersion, or less commonly underdispersion, relative to the binomial or Poisson distribution is a special case (Cox, 1982b).

If the grouping is not just a device employed for convenience but corresponds to grouping of the individuals with direct physical significance, there is the possibility that regression relations between and within groups are different; see, for instance, Cox and McCullagh (1982). Failure to recognize this can lead to major errors of interpretation.

3.5. Concluding Remarks

There are, of course, many further aspects of regression analysis. For instance, regression analysis of time series, "random effects" regression models and kernel regression methods have not even been mentioned. The usefulness of robust regression methods seems to me largely confined to situations where large bodies of data have to be summarized in a somewhat mechanical...
fashion. The current emphasis on regression diagnostics, the isolation of sensitive observations for individual study, is a formalization of more traditional attitudes to the analysis of data and much to be welcomed.

The topics discussed in this paper are very much a personal selection. It is a mark of the vigour and importance of the subject that, despite the very extensive literature developed over a long period, so much remains to be done.

ACKNOWLEDGEMENT

I am grateful to the Science and Engineering Research Council for their support.

APPENDIX

There follow outline statements of one or two open problems. The literature on both design of experiments and regression is so large that it is entirely possible that I have overlooked relevant work. Matters connected with very special designs have been excluded.

1. Is there a simple and useful non-null randomization theory of designs with binary (and other such) responses?
2. Is there a useful Bayesian theory of experimental design recognizing that the prior distribution for the interpreter of the results may be quite different from that of the planner of the investigation? (Equivalently obtaining the data may, of course quite legitimately, itself lead to a change in the prior distribution.)
3. While interesting work has been done on the design of experiments for estimating components of variance, there appears to be no systematic theory. Develop one.
4. What is the role in practice of cross-over designs with more periods than treatments?
5. Classify the kinds of residual effect that might arise, discuss the designs appropriate for each and examine the extent to which it is feasible to distinguish empirically between the different kinds of residual effect.
6. What is the theoretical asymptotic relative efficiency of the optimal randomized block design versus the optimal systematic design when spatial variation follows some (convenient) spatial stochastic model?
7. What is the asymptotic relative efficiency of both the types of design when spatial variation is a self-similar process?
8. Develop flexible families of designs for use when there are a number of factors with quantitative levels (response surface) combined with some factors with qualitative levels.
9. Develop sequential procedures for $D$- and $D_S$-optimal design based on maximum likelihood estimation in a fairly general model. What is the loss of efficiency if the design must be fixed in just two stages?
10. Examine, not solely in a clinical trial context, the so-called principle of analysis by intention to treat (ever randomized, always analysed), as opposed to analysis by treatment actually encountered.
11. In sequential experimentation, explicit changes in the objectives and protocol of the experiment are possible during the course of the investigation. Yet frequent major changes (a different "bright idea" each day) will usually be harmful. Can this usefully be formalized?
12. Develop a simple procedure for detecting whether an apparent dependence on say an explanatory variable $x_q$ could be an artefact arising from measurement errors in one or more of the other explanatory variables.
13. Develop a procedure for binary logistic or probit regression taking account of regressions
“between” and “within” groups of individuals and more refined than the application of normal theory ideas to empirical logit or probit transforms.

14. Extend the discussion of 13 to other generalized linear models and to time series analysis.

15. Suppose that the regression of a response variable on a single explanatory variable is found to be non-linear. Show how to examine whether the non-linearity can be explained by errors of measurement in the explanatory variable. Generalize to multiple regression.

16. Examine in detail the effect in generalized linear models of errors in the explanatory variables.

17. Two (or more) sets of data are available in which the regression relations appear to be different. Suggest how to examine whether the sets are consistent with a single underlying relation, the errors of measurement of the explanatory variables being different in the different sets.

18. Obtain prediction limits, more accurate than those in the text, for the change in response induced by specified intervention on some of the explanatory variables. Extend to the generalized linear model.

19. Discuss carefully the effect on the choice of a multiple regression equation of specific objectives, in particular of the need to extrapolate to a particular point in the space of explanatory variables.

20. Develop versions of ridge regression for the generalized linear model; assume a subset of regression coefficients specified to which the “shrinkage” is not to be applied.

21. Develop a version of multiple regression in which the dependence on one or two explanatory variables of primary interest is assessed by “kernel” methods whereas adjustment for other explanatory variables is carried out by least squares.

22. Develop a systematic theory for the prediction (via intervals) of future values from generalized linear models.

REFERENCES


S. C. Pearce (University of Kent at Canterbury): Professor Cox began by expressing appreciation that he had been invited to deliver this Fisher Memorial Lecture. Let me begin by expressing appreciation that he has done so. He has exhibited two great qualities, both of which were prominent in R. A. Fisher himself. One is width of interest and the other the careful specification of each problem before proceeding to techniques for solving it.

There are two points on which I would like to comment. The first concerns his historical perspective at the start. It is quite true that a period of interest in biological experiments was followed by another in which the needs of industry were paramount, but the two sorts of experiment are so different that I sometimes think that the use of the same word to cover both is little better than a semantic muddle. To start with, in biology data are not reproducible, so if we conduct an experiment and then decide that we ought to have had some additional treatments, we cannot introduce them retrospectively. In a context of physics or chemistry, on the other hand, deciding which determinations to add is a recognized part of the skill. In short, in the one sort everything is comparative within the data set and we can study only contrasts; in the other we are after absolute values. That has an effect on studies of optimality, where a lot of useful work has been done for industrial investigations but much less and rather less successfully for the comparative experiments of the biologists. With a response surface we are after a single quantity, a maximum or something like that; with a comparative experiment there can be several contrasts and we have to weight them to show the ones we want. (For example, two people may design two factorial experiments with identical treatments. One knows that the main effects exist and is curious to know if they interact; the other is fairly sure that they do not, but needs more precise information about the main effects themselves. They will not agree as to which design is optimal.) That single-mindedness of approach in the industrial experiment has facilitated work on optimality. It has also led to misdirection in optimality studies in comparative experiments. People start by assuming that the solution must lie along the path of D-optimality with its use of geometric means, but if one takes the set of possible contrasts and assigns each one a weight and then multiplies that by the variance of estimation or something like it for the design under consideration, it is no good taking a geometric mean of the products. That leads to a product of two quantities, the first being the geometric mean of the weights, which is a constant for all designs, and the second the geometric mean of the variances, which will indeed be a function of the particular design under consideration at the moment but it will not distinguish the contrasts, whether important or not. If we are to proceed, it must be along some other path.

My other point concerns Professor Cox’s suggestion that in regression some explanatory variables can play the role of treatments. I should first remark that the regression coefficient of $y$ on $x$ can depend upon the reason for the $x$-values being different. For example, we may suppose that big boys are stronger than small ones of the same age. We think that a larger weight ($x$) must indicate a larger $y$ (strength). So long as we think of differences of physique and nutrition we are probably right, but we must not forget the boy who spends his time in front of a television set eating sweets compared with one who is active. It is a neglected glory of traditional regression methods that they integrate a range of causes. Given $x$ they will indicate $y$ for the mix of causes operating in that body of data — they will even assign a standard error to $y$ estimated in those conditions — but experience shows that the regression coefficients found do not always transport well to other bodies of data if the mix of causes of variation is different. To take an extreme case in Professor Cox’s example of the smoking and drinking habits of pregnant women, if a woman is alcoholic that could well result from characteristics of mind or body that could also affect her pregnancy. A social scientist might welcome the association and see nothing wrong in the
regression approach, the need being to find out how alcoholic women react to pregnancy, so all concomitants should be included. A physiologist, however, might object that his need is to discover the effect of alcohol in a pregnant woman and her foetus. The best way of finding out would be to adopt the traditional approach by experiment in which one intervenes at random but since we cannot require a random sample of women to take to the bottle for the duration of their pregnancies—the very idea is repulsive—some less satisfactory approach is needed. (As a matter of interest we may recall that it was considerations of that sort that were prominent in Fisher's famous controversy over smoking and lung cancer.) I suggest that no one should speak of treatments unless there has been random intervention or something that is formally identical with it. Further, where intervention has occurred, the appropriate techniques are those we associate with the design of experiments rather than regression methods, though admittedly there is some resemblance between them, a resemblance that I sometimes feel to be misleading.

It gives me great pleasure to propose a vote of thanks to Professor Cox, our thanks being so well deserved.

The author replied later, in writing, as follows:

Professor Pearce's generous comments focus on two important issues that merit longer discussion than is feasible here. There are indeed a number of distinctions between the kind of industrial experiments that motivated much of the work on response surface designs and more traditional biometrical experiments, although I would myself prefer not to overemphasize these differences. Both kinds of experiment seem to be ultimately comparative and one main distinction may be that in many industrial experiments "error" is relatively small, encouraging careful investigation and representation of the form of treatment effect. I agree, of course, that if ideas of optimal design theory are to be used appropriate criteria must be chosen, but there is now a fairly rich variety of such criteria amenable to theoretical and numerical discussion.

Professor Pearce is surely right to stress the major difficulties of interpretation involved with explanatory variables not under the investigator's control. Nevertheless there are very important areas of study where intervention by the investigator is intellectually feasible and desirable but in practice is impossible. My suggestions on terminology were partly designed to clarify interpretation of regression analysis in such situations but were most certainly not intended to blur the absolutely key distinction between experimental and observational studies. Perhaps quasi-treatment variable would be a better term in non-experimental contexts.