A Biometrics Invited Paper with Discussion. Some Aspects of Analysis of Covariance

D. R. Cox; P. McCullagh

Biometrics, Volume 38, Issue 3, Special Issue: Analysis of Covariance (Sep., 1982), 541-561.

Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28198209%2938%3A3%3C541%3AABIPWD%3E2.0.CO%3B2-I

BIOMETRICS 38, 541–561
September 1982

A BIOMETRICS INVITED PAPER WITH DISCUSSION

Some Aspects of Analysis of Covariance

D. R. Cox and P. McCullagh

Department of Mathematics, Imperial College, 180 Queen's Gate, London SW7 2BZ, England

SUMMARY

An account from first principles is given of a number of aspects of analysis of covariance. Six different meanings of analysis of covariance are outlined and the history of this technique is sketched briefly. The development of the key formulae from the method of least squares is described, and generalizations to other distributions in the exponential family are mentioned. Special problems of application in randomized experiments and in observational studies are discussed. Finally, the decomposition of regression relations is considered along with components of covariance.

1. Introduction

The words 'analysis of variance' are used in a number of rather different senses and the same is the case for 'analysis of covariance'. Here are six different meanings.

First, corresponding to any analysis of variance in which the sum of squares of a variable y is a partitioned into components, there is a decomposition of the sum of the products of two variables y and z, and hence the formal possibility of setting out an analysis of covariance table for two or more variables. We assume that the reader is familiar with such tables. In the first place the decomposition can be regarded as a descriptive representation of covariation, broken into components.

Secondly, analysis of covariance is the basis of a numerical technique for improving the precision of a comparative experiment by adjustment for concomitant variables measured before the assignment of treatments to experimental units.

Thirdly, the same formal procedures as employed in precision improvement can be used to provide estimates, confidence limits and significance tests in the normal-theory linear model when the parameters are partitioned into two or more parts. The procedure is especially convenient when the primary analysis of variance has simple structure. In effect this is just a compact way of setting out the standard linear model calculations and as such has been made less important numerically by the wide availability of flexible computer programs for nonorthogonal least squares analyses. It remains, however, an important theoretical technique for handling least squares work. A particular application is the calculation of regression adjustments in unbalanced observational studies.

This paper was read at the Joint Statistical Meetings of the American Statistical Association and the Eastern and Western North American Regions of the Biometric Society held in Detroit on 12 August 1981.

Key words: Adjustment; Component of covariance; Component of regression; Concomitant variable; Fitting by stages; History; Linear model; Precision improvement.

Fourthly, we may decompose covariances of random variables into components. For a single random variable Y, which is the sum of two uncorrelated components $Y^{(1)}$ and $Y^{(2)}$, we have that

$$Y = Y^{(1)} + Y^{(2)}, \qquad \operatorname{var}(Y) = \operatorname{var}(Y^{(1)}) + \operatorname{var}(Y^{(2)}); \tag{1}$$

this is analysis of variance in the most literal sense. Similarly, if the vector $\mathbf{Y} = (Y_1, Y_2)$ is the sum of uncorrelated components,

$$(Y_1, Y_2) = (Y_1^{(1)}, Y_2^{(1)}) + (Y_1^{(2)}, Y_2^{(2)}),$$

then

$$\operatorname{cov}(Y_1, Y_2) = \operatorname{cov}(Y_1^{(1)}, Y_2^{(1)}) + \operatorname{cov}(Y_1^{(2)}, Y_2^{(2)}). \tag{2}$$

This and its obvious generalizations lead to the notion of components of covariance, generalizing the notion of components of variance for univariate random variables. Further, just as in the univariate case we can use synthesis of variance to reconstruct from components of variance new variances applying to modified random systems, so too we can have synthesis of covariance. Note that in principle (1) and (2) generalize to higher order cumulants, when the components are independent to the requisite order.

Fifthly, the formal tests of normal-theory multivariate analysis of variance, and the calculations of canonical regression analysis, are based on the elements of an analysis of covariance table. We regard this aspect as outside the scope of the present paper. Finally, analysis of covariance might be taken as the study of special representations of covariance matrices, for example via linear covariance structures; see Jöreskog (1981) for an excellent review.

Analysis of covariance is due to R. A. Fisher. Eden and Fisher (1927) gave the decomposition of a sum of products and used the corresponding correlation coefficients descriptively. Sanders (1930), at Fisher's suggestion, was the first to use analysis of covariance for precision improvement. The procedure was described in *Statistical Methods for Research Workers* (Fisher, 1932, §49.1) in a form in which standard analysis of variance was applied to $y - \hat{\gamma}z$, where y is the response, z is the concomitant variable and $\hat{\gamma}$ is the regression coefficient estimated from the residual line of the analysis of covariance table. This was recognized to lead only to an approximate F test of treatments. Two papers describing applications of analysis of covariance were read to the Royal Statistical Society in 1934 (Wishart, 1934; Wilsdon, 1934). Fisher contributed to the discussion of both and it is likely that by that point he had appreciated how to set out an 'exact' F test: certainly this is very explicitly described in *Design of Experiments* (Fisher, 1935, Chapter 9). In Wilsdon's paper about statistical methods in the building industry, the emphasis is on the description and decomposition of regression relations rather than on precision improvement. E. S. Pearson contributed an Appendix setting out the calculations in detail. The notion of components of regression was developed by Tukey (1951).

The special 1957 *Biometrics* issue on analysis of covariance is an important source for later work; see especially the lucid review of technique by Cochran (1957) and the discussion by Fairfield Smith (1957) of interpretation when the variable used for adjustment may be affected by treatments.

To the extent that analysis of covariance represents an idea for approaching specific applied issues there is no restriction to the normal-theory linear model. Precision improvement by introducing concomitant variables is, for example, possible in principle very broadly and in particular in generalized linear models (Nelder and Wedderburn, 1972). Also, it would be possible to replace least squares or maximum likelihood estimation by some 'robust' alternative method.

2. Review of Linear Model Results

The main emphasis in the present paper is on interpretation, but in the present section we outline the derivation of the main formal results of analysis of covariance. The results follow immediately from the geometry of least squares. Here, although the treatment is algebraic rather than geometric, the algebra has been set out to exploit the simplicity of the underlying geometry.

Suppose that the $n \times 1$ vector $\mathbf{Y}$ of observed random variables is such that

$$\mathrm{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}, \tag{3}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are $n \times d_X$ and $n \times d_Z$ matrices of constants with $(\mathbf{X}, \mathbf{Z})$ assumed for simplicity to be of rank $d_X + d_Z < n$, and where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are $d_X \times 1$ and $d_Z \times 1$ vectors of unknown parameters. We suppose that the covariance matrix of $\mathbf{Y}$ is $\mathrm{cov}(\mathbf{Y}) = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the $n \times n$ identity matrix and σ^2 is unknown; that is, we suppose that the components of $\mathbf{Y}$ are uncorrelated and that they all have the same variance σ^2.

Typically, in (3), $\mathbf{X}\boldsymbol{\beta}$ represents a model from some standard configuration and $\mathbf{Z}\boldsymbol{\gamma}$ represents the extra terms, corresponding, for instance, to the concomitant variables. When $\boldsymbol{\gamma} = 0$ we speak of the 'model' $\mathbf{X}$. Under this model, let $\mathbf{R}_X$ be the idempotent matrix

$$\mathbf{R}_X = \mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T. \tag{4}$$

Thus for an arbitrary $n \times 1$ vector $\mathbf{u}$, $\mathbf{R}_X\mathbf{u}$ is the vector of residuals from the model $\mathbf{X}$, and $\mathbf{u}^T\mathbf{R}_X\mathbf{u}$ is the residual sum of squares. Further $\mathbf{u}_1^T\mathbf{R}_X\mathbf{u}_2$ is the residual sum of products of two arbitrary vectors $\mathbf{u}_1$ and $\mathbf{u}_2$ and $\mathbf{Z}^T\mathbf{R}_X\mathbf{Z}$ is the $d_Z \times d_Z$ matrix of residual sums of squares and products formed by treating the columns of $\mathbf{Z}$ as observation vectors.

The key step in the derivation of the analysis of covariance formulae is to rewrite (3) with the columns of $\mathbf{Z}$ replaced by their formal residuals under model $\mathbf{X}$. Thus

$$\mathrm{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}^{(0)} + \mathbf{R}_X\mathbf{Z}\boldsymbol{\gamma} \tag{5}$$

$$= \mathbf{X}\{\boldsymbol{\beta}^{(0)} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}\boldsymbol{\gamma}\} + \mathbf{Z}\boldsymbol{\gamma}, \tag{6}$$

so that

$$\boldsymbol{\beta} = \boldsymbol{\beta}^{(0)} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}\boldsymbol{\gamma}. \tag{7}$$

We now form the least squares equations from (5), the matrix defining the model having the particular form $(\mathbf{X}, \mathbf{R}_X\mathbf{Z})$. Thus least squares estimates satisfy

$$\begin{bmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{R}_X\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}_X\mathbf{X} & \mathbf{Z}^T\mathbf{R}_X\mathbf{Z} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}^{(0)} \\ \hat{\boldsymbol{\gamma}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{Y} \\ \mathbf{Z}^T\mathbf{R}_X\mathbf{Y} \end{bmatrix}.$$

Because of the special form of $\mathbf{R}_X$, this reduces to

$$\begin{bmatrix} \mathbf{X}^T\mathbf{X} & 0 \\ 0 & \mathbf{Z}^T\mathbf{R}_X\mathbf{Z} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}^{(0)} \\ \hat{\boldsymbol{\gamma}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{Y} \\ \mathbf{Z}^T\mathbf{R}_X\mathbf{Y} \end{bmatrix}. \tag{8}$$

There now follow immediately five key results which form the essence of the analysis of covariance, viewed as a special algorithm attached to the method of least squares.

(i) The estimates $\hat{\boldsymbol{\gamma}}$ satisfy

$$\mathbf{Z}^T\mathbf{R}_X\mathbf{Z}\hat{\boldsymbol{\gamma}} = \mathbf{Z}^T\mathbf{R}_X\mathbf{Y}, \tag{9}$$

a formal set of least squares equations derived from sums of squares and products residual to the model $\mathbf{X}$. In the rather different context of incomplete block designs, with $\mathbf{X}$

specifying block effects and $\mathbf{Z}$ treatment effects, (9) is called the set of reduced least squares equations.

(ii) The covariance matrix of $\hat{\boldsymbol{\gamma}}$ is $(\mathbf{Z}^T\mathbf{R}_X\mathbf{Z})^{-1}\sigma^2$.

(iii) The least squares estimates $\hat{\boldsymbol{\beta}}^{(0)}$ are those for the model $\mathbf{X}$. They have zero covariance with $\hat{\boldsymbol{\gamma}}$. The least squares estimates of $\boldsymbol{\beta}$ are formed by (7) as

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(0)} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}\hat{\boldsymbol{\gamma}}, \tag{10}$$

the covariance matrix of $\hat{\boldsymbol{\beta}}$ being the sum of those of the two parts. That is, we adjust $\hat{\boldsymbol{\beta}}^{(0)}$ for 'regression' on the columns of $\mathbf{Z}$.

(iv) The residual sum of squares is found by applying to (8) the usual formula 'total sum of squares' minus 'sum of squares for fitting', and is the residual sum of squares from model $\mathbf{X}$ minus the sum of squares for fitting $\mathbf{Z}$, using sums of squares and products residual to model $\mathbf{X}$. The unbiased estimate of σ^2 follows on dividing by $n - d_X - d_Z$.

(v) To test the hypothesis that $\boldsymbol{\beta}_1 = 0$, where $\boldsymbol{\beta}_1$ is a subvector of $\boldsymbol{\beta}$, often representing 'treatments' or some component thereof, we apply the above procedure twice, once under the general model and once under the null hypothesis. This leads to the standard procedure of adjusting for regression the residual from model $\mathbf{X}$ and 'residual plus treatments', the latter being the residual sum of squares from model $\mathbf{X}$ constrained by the null hypothesis.

One immediate generalization applies when $\text{cov}(\mathbf{Y}) = \sigma^2\mathbf{V}$, where $\mathbf{V}$ is a known matrix and least squares estimation is replaced by generalized least squares estimation. For this, $\mathbf{R}_X$ is generalized to

$$\mathbf{R}_X = \mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1},$$

the sums of squares and products $\mathbf{u}^T\mathbf{u}$ and $\mathbf{u}_1^T\mathbf{u}_2$ to $\mathbf{u}^T\mathbf{V}^{-1}\mathbf{u}$ and $\mathbf{u}_1^T\mathbf{V}^{-1}\mathbf{u}_2$, and the residual sums of squares and products to $\mathbf{u}^T\mathbf{R}_X\mathbf{V}^{-1}\mathbf{R}_X\mathbf{u}$ and $\mathbf{u}_1^T\mathbf{R}_X\mathbf{V}^{-1}\mathbf{R}_X\mathbf{u}_2$, where

$$\mathbf{R}_X\mathbf{V}^{-1}\mathbf{R}_X = \mathbf{V}^{-1}\mathbf{R}_X = \mathbf{R}_X^T\mathbf{V}^{-1}.$$

3. Review of Linear Exponential Family Results

Suppose that the $n \times 1$ vector of observed random variables $\mathbf{Y}$ has independent elements Y_i whose distribution, for each i, is in the one-parameter exponential family

$$f_{Y_i}(y_i; \theta_i) = \exp\{y_i\theta_i - g(\theta_i) + d(y_i)\},$$

for suitable functions $g(\cdot)$ and $d(\cdot)$. The model is linear in the canonical parameter θ_i if the $n \times 1$ vector $\boldsymbol{\theta}$ satisfies the linear model

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}, \tag{11}$$

which is the analogue of (3).

In some applications the parameter $\boldsymbol{\gamma}$ is known and the term $\mathbf{Z}\boldsymbol{\gamma}$ is then known as an 'offset'. An example of this arises in the sampling of a Poisson process over intervals of known but unequal length $\exp(z_i)$, where Y_i is the number of events and θ_i is the log of the incidence rate. Here $\gamma = 1$ because the average number of events is directly proportional to the length of the sampling interval.

When $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are both unknown the minimal sufficient statistic is $(\mathbf{X}^T\mathbf{Y}, \mathbf{Z}^T\mathbf{Y})$, of total dimension $d_X + d_Z$. If exact significance tests or confidence intervals are required for $\boldsymbol{\beta}$, these are constructed by examination of the distribution of $\mathbf{T}_1 = \mathbf{X}^T\mathbf{Y}$ given $\mathbf{T}_2 = \mathbf{Z}^T\mathbf{Y}$, this conditional distribution being independent of the nuisance parameter $\boldsymbol{\gamma}$.

When the distribution of Y_i is continuous no particular difficulty arises, at least in principle, although, in practice, computation of the conditional density may prove a formidable task. For discrete distributions the method of conditioning often breaks down because the conditional distribution may be degenerate or nearly so. Such degeneracy differs from that illustrated by Cox (1967) in that the parameter of interest, $\boldsymbol{\beta}$, is here a canonical parameter in the exponential family.

In the absence of a satisfactory small-sample theory, approximate intervals and tests can be constructed via maximum likelihood estimates, likelihood ratio statistics, or score statistics. The assumptions required here are often weaker than those required for an exact theory. We consider first maximum likelihood estimation. It can be shown, on differentiation of (10), that $\mu_i = \mathrm{E}(Y_i) = g'(\theta_i)$ and $v_i = \mathrm{var}(Y_i) = g''(\theta_i) = d\mu_i/d\theta_i$. Let $\mathbf{V} = \mathrm{diag}(v_i)$ be the variance matrix of $\mathbf{Y}$. For the model $\mathbf{X}$ the Newton–Raphson method applied to the maximum likelihood equations leads to the iterative scheme

$$\mathbf{X}^T\hat{\mathbf{V}}_n\mathbf{X}(\hat{\boldsymbol{\beta}}_{n+1} - \hat{\boldsymbol{\beta}}_n) = \mathbf{X}^T(\mathbf{Y} - \hat{\boldsymbol{\mu}}_n), \tag{12}$$

where $\hat{\boldsymbol{\beta}}_n$, $\hat{\mathbf{V}}_n$ and $\hat{\boldsymbol{\mu}}_n$ are current estimates of $\boldsymbol{\beta}$, $\mathbf{V}$ and the mean vector $\boldsymbol{\mu}$, and $\hat{\boldsymbol{\beta}}_{n+1}$ is the updated estimate of $\boldsymbol{\beta}$. A similar iterative scheme with $\mathbf{X}$ replaced by the augmented matrix $(\mathbf{X}, \mathbf{Z})$ and $\boldsymbol{\beta}^T$ replaced by $(\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)$ applies to the extended model (11). A similar but slightly more complicated iterative scheme can be used for multiparameter linear exponential family models.

The essence of the simplification in §2 is that the transformation (5) to orthogonal parameters $(\boldsymbol{\beta}^{(0)}, \boldsymbol{\gamma})$ should not depend on the value of the parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma})$. Now the Fisher information matrix for $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ in (11) is

$$\mathbf{I} = \begin{pmatrix} \mathbf{X}^T\mathbf{V}\mathbf{X} & \mathbf{X}^T\mathbf{V}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{V}\mathbf{X} & \mathbf{Z}^T\mathbf{V}\mathbf{Z} \end{pmatrix},$$

which depends on $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ through $\mathbf{V}$. Hence, for linear exponential families, the transformation to orthogonal parameters is independent of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ only if $v = g''(\theta) = \text{constant}$, and this seems to exclude all but the normal distribution.

If, however, the model is linear not in θ but on the scale in which Fisher information is constant, some of the results of §2 apply. In particular, the covariance matrix of $\hat{\boldsymbol{\gamma}}$ is proportional to $(\mathbf{Z}^T\mathbf{R}_X\mathbf{Z})^{-1}$ with $\mathbf{R}_X$ given by (4). Furthermore, (10) applies and the covariance matrix of $\hat{\boldsymbol{\beta}}$ is the sum of the two uncorrelated components in (10). Despite this slight simplification, no closed-form maximum likelihood estimates exist in general. In particular (8) and (9) do not hold except for the normal theory linear model. Computational details concerning maximum likelihood estimation for such generalized linear models are given by Nelder and Wedderburn (1972).

Approximate confidence intervals for individual treatment contrasts in the presence of concomitant variables can be obtained by a number of asymptotically equivalent methods. We discuss three of these methods here. The first and simplest is to fit by maximum likelihood the full model including the covariate and to use a normal approximation for $\hat{\beta}_s$, where β_s corresponds to the contrast of interest. The approximate $(1-\alpha)$-level confidence interval is $\{\hat{\beta}_s - k^*_{\frac{1}{2}\alpha}\sqrt{I^{ss}}, \hat{\beta}_s + k^*_{\frac{1}{2}\alpha}\sqrt{I^{ss}}\}$, where I^{ss} is the (s, s) element of $\mathbf{I}^{-1}$ and $k^*_{\frac{1}{2}\alpha}$ is the upper $100(1-\frac{1}{2}\alpha)$ percentage point of the standard normal distribution.

The second method is to compute (i) the maximized log likelihood function $l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ and (ii) the maximized log likelihood function holding β_s fixed at some arbitrary value $\beta_s^{(0)}$, i.e. $l(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\gamma}}^*; \beta_s = \beta_s^{(0)})$. Then an approximate $(1-\alpha)$-confidence set for β_s is

$$\{\beta_s^{(0)}: 2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) - 2l(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\gamma}}^*; \beta_s = \beta_s^{(0)}) \leqslant \chi^2_{1,\alpha}\},$$

where $\chi^2_{1,\alpha}$ is the upper $(1-\alpha)$-point of the χ^2_1 distribution.

Finally, using the efficient score $u(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\gamma}}^*; \beta_s = \beta_s^{(0)})$, which is the first derivative with respect to β_s at $\beta_s^{(0)}$ of $l(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\gamma}}^*; \beta_s = \beta_s^{(0)})$, an approximate confidence set for β_s is given by

$$\{\beta_s^{(0)}: -k^*_{\frac{1}{2}\alpha} < u(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\gamma}}^*; \beta_s = \beta_s^{(0)})\sqrt{I^{ss}} < k^*_{\frac{1}{2}\alpha}\}.$$

This third method allows of some variations depending on whether $\mathbf{I}$ is evaluated at $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ or at $(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\gamma}}^*)$ and, for models that are not linear in the canonical parameter, there is also a choice between observed and expected Fisher information. Recent work suggests that the observed information is preferable.

From a purely computational viewpoint the first method, requiring only one maximization of the log likelihood function is best. Its accuracy is greatly improved if the scale is such that $\hat{\beta}_s$ is symmetrically distributed, or nearly so. In particular, the range for β_s should, if possible, be unrestricted. The second and third methods have the strong conceptual advantage of being invariant under 1–1 parameter transformations, but they are computationally cumbersome, at least for construction of confidence interval or sets.

Significance tests concerning a subvector $\boldsymbol{\beta}_1$ of $\boldsymbol{\beta}$ are constructed by comparing $l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ with the maximized log likelihood holding $\boldsymbol{\beta}_1$ fixed at its hypothesized value. This is the direct analogue of (v) in §2. Alternatively, score tests are often simple to construct using a quadratic form in the vector-valued efficient score $\mathbf{u}(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\gamma}}^*; \boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^{(0)})$.

4. Randomization Theory

One approach to the analysis of 'standard' randomized designs appeals only to an assumption of unit–treatment additivity and to the symmetry of the particular randomization scheme employed. This is best regarded as an approach complementary to the assumption of a special linear model: it has the conceptual advantage of leading to a unified second-order theory without *ad hoc* assumptions for each design.

It is known (Yates 1951; Kempthorne, 1955) that 'exact' second-order properties of analysis of covariance for precision improvement do not follow from randomization theory, although some properties can be recovered by a special scheme of weighted randomization in which designs are chosen from a randomization set with unequal weights. There is, however, a sense in which there is an asymptotic randomization theory (Robinson, 1973) and in the light of this Cox (1982) has suggested randomization conditional on 'lack of balance with respect to the covariates being small': this is achieved by repeated rerandomization until a suitable criterion measuring lack of balance in the covariates is sufficiently small.

Further work is needed both on the theory and on the practical implications for design and analysis.

5. Applications

We discuss briefly four broad areas of application. These are distinguished not by differences of numerical technique but by the purpose for which covariance adjustment is used.

5.1 *Adjustment for Precision in Designed Experiments*

In most randomized experiments we have a fairly small number of concomitant variables associated with each experimental unit. We might, for example, use the previous year's yield as a covariate in field trials to reduce variation due to unequal fertility in the experimental plots. The purpose of covariance adjustment in this case is solely to increase precision of treatment contrasts: randomization ensures that, in the absence of information about the covariate, unbiased estimates of treatment contrasts and of their precision are obtained from the usual analysis of the standard configuration $\mathbf{X}$. Adjustment for the covariate reduces the residual variance from $\text{var}(Y)$ to $\text{var}(Y \mid z)$, the conditional variance of Y after linear regression on the covariate z. We write $\text{var}(Y \mid z) = (1-\rho^2)\text{var}(Y)$, where ρ is the correlation between Y and z. The principal effect of covariance adjustment is to increase the precision of treatment contrasts by multiplying the residual variance by the factor $1-\rho^2$. If an allowance is made for errors of estimation in the additional d_Z parameters, where d_Z is the number of concomitant variables, this factor becomes, on average, approximately $(1-\rho^2)(d_R-1)/(d_R-d_Z-1)$, where d_R is the residual degrees of freedom from the standard configuration, $\mathbf{X}$. This is an average precision factor and as such may be misleading in small experiments where designs with substantially greater or smaller precision can occur with nonnegligible probability. Cochran (1957) gives a similar average precision factor with $d_Z = 1$. The question of how to choose a design in the light of a concomitant variable has been investigated by Finney (1957) and by Cox (1957).

If increased precision is the sole criterion, the above analysis suggests the following:

(i) if ρ is less than about .3, covariance adjustment offers little gain in precision;
(ii) if $\rho > .9$ there is substantial gain in precision;
(iii) if the residual degrees of freedom are small it is undesirable to use more than one or two concomitant variables.

The usual analysis assumes parallelism, i.e. the absence of treatment × concomitant variable interaction. Absence of interaction can be tested, for instance, by introducing a further set of variables formed by multiplying $\mathbf{Z}$ by indicator functions for treatment or treatment contrasts, and comparing residual sums of squares. Such interaction, if present, can often be removed by a nonlinear transformation of y, thus simplifying interpretation. Care must, however, be taken to present the conclusions on a readily understood scale. This may well be the original scale of measurement, y, and not the transformed scale.

5.2 *Adjustment for Bias in Observational Studies*

In many investigations, especially those involving human subjects, random allocation to treatments is not possible for ethical, financial or other reasons. For general discussion of comparative nonrandomized studies, see the recent book by Anderson *et al.* (1980). Examples of nonrandomized investigations include studies of the effect of fluoridation of the water supply on the incidence of (i) dental caries and (ii) cancer, and studies of the effect of nuclear fallout on the incidence of various forms of cancer.

The purposes of analysis of covariance in observational studies are again twofold. The first concerns the transition from the observed sample to the whole population. To make a strong case for generalizing the results to the population we need to demonstrate that the effect of interest is constant under a wide variety of conditions. In other words, there should, if possible, be no interaction between the effect of interest and incidental covariates. This may be achievable only after a change in the scale of measurement. Such interactions as cannot be removed by transformation should be readily explainable, preferably by a dichotomy over some classifying variable such as sex.

The second purpose of covariance analysis is to ensure that the groups being examined are comparable and that observed differences cannot be attributed to demographic or other incidental variables. In general the crude estimate of the treatment effect or contrast of interest can be partitioned into three components:

(i) the real group or treatment effect;
(ii) an effect attributable to incidental variables, z;
(iii) unexplained or random variation.

The central idea is to make comparisons at fixed levels of the incidental variables. The question whether or not such comparisons are sensible is discussed briefly in §6. If the data are grouped into sets having the same z values, treatment comparisons would be made within each set, preferably on a scale such that the treatment effect does not depend on z. Information is then pooled to form an overall estimate of the treatment effect adjusted for z.

In general, such a partition of the data does not arise naturally and, to make progress, it is usually necessary to postulate some form of dependence such as (3) or (11). The results of §2 and §3 then apply, assuming of course the adequacy of the model. In particular it is assumed that the effect on the response of any further covariate, observed or unobserved, is absorbed in z. In contrast to the position in randomized studies, where the effect of unobserved variables should, at least in large experiments, be balanced out by randomization, in observational studies such unobserved variables can be strongly associated with the contrast of interest. If such an unobserved or discarded concomitant variable, not orthogonal to the contrast of interest, has an effect on the response additional to that of z, then the conclusions are liable to be misleading. Simpson's paradox (Simpson, 1951) provides a standard example in this context. An example of Simpson's paradox in the context of alleged sexual bias in college admissions policy is given by Bickel, Hammel and O'Connell (1977). See also the ensuing correspondence with W. H. Kruskal in the same reference.

5.3 *Adjustment for Missing Values in Balanced Designs*

This application of covariance analysis is purely a numerical device for dealing with missing values in what would otherwise be a standard balanced design. The general idea is to replace the missing values by any convenient numbers and at the same time to include as covariates an indicator vector for each missing value. Covariance adjustment in the now balanced design gives the full exact analysis for the unbalanced design. For details, see Coons (1957).

With the advent of general-purpose computer packages, missing values can now be treated in a number of different ways, some of which are more convenient than covariance adjustment. There are

(i) to analyse the unbalanced design via the general linear model;
(ii) to complete the design but give zero weight to the missing values;
(iii) to use the EM algorithm (Dempster, Laird and Rubin, 1977) which involves iterative estimation of the missing values. Whatever method is used, if it is required to test the hypothesis that $\boldsymbol{\beta}_1 = \mathbf{0}$, where $\boldsymbol{\beta}_1$ is a subvector of $\boldsymbol{\beta}$, the method described in §2(v) can be used.

5.4 *Adjustment for Historical Controls in Clinical Trials*

In most clinical trials patients are divided at random into two or more groups, one group being given a standard treatment or placebo. These are known as 'concurrent controls' and may comprise about one half of the total number of patients enrolled in the study. Often data are available from similar controlled trials conducted some years previously. There is clearly the possibility, at least in theory, of using the control group from such a previous study as the controls for the new study. These are then known as 'historical controls'. One would naturally expect the historical controls to differ in some respects from the new treatment group, and an adjustment, either by analysis of covariance or by some form of standardization, would be required for such observed differences.

The main potential advantages of using historical as opposed to concurrent controls are the following:

(i) the number of patients required for the new study could be reduced possibly by a factor of one half;
(ii) administrative costs would be reduced because of the smaller number of patients and because there would be no need to randomize patients to one of two groups.

The main disadvantages are fairly clear but we list five:

(i) lack of blindness;
(ii) changes in medical definitions and in methods of measurement;
(iii) changing standards of general medical care;
(iv) the validity of any assumed relationship such as (3) or (10) may be open to question and difficult to check, so the parallelism of relations within new and historical groups can and should be checked;
(v) there may be further unobserved influential concomitant variables.

Some of these disadvantages could be reduced or eliminated by using a combination of concurrent and historical controls, in which case it would be necessary to check that the two sets of controls were mutually consistent. If they were not it is far from clear how the historical controls could be used, or indeed whether they should be used, in formal comparisons. For an empirical discussion of the use of historical controls, see Farewell and D'Angio (1981).

6. Analysis and Interpretation

For linear exponential family models a method for iterative computation of maximum likelihood estimate and likelihood ratio tests was given in §3. For ordinary least squares, iterative computation is unnecessary. The relevant equations are given in §2 and the various sums of squares and products can be set out as an analysis of covariance table. We assume that the reader is familiar with such tables and we concentrate instead on difficulties of interpretation.

One minor difficulty arises in orthogonal designs where unadjusted treatment contrasts and the corresponding sums of squares are independent. This property of independence simplifies interpretation because the various sums of squares are uniquely ascribable to the corresponding treatment contrasts. However, the adjusted treatment contrasts are not independent. In randomized experiments we would expect the degree of dependence to be small and no great difficulty would arise. In nonrandomized studies, however, we could find that the joint effect of two treatments was highly significant but that neither of the adjusted treatment effects was significant.

A major difficulty of interpretation arises when variations in the concomitant variable are caused directly or indirectly by the treatment applied. This cannot occur if the concomitant variable is measured before the assignment of treatments to units. In other situations, where the covariate is affected by treatment, an adjustment by covariance analysis amounts to adjusting the treatment effect for differences caused by the treatments themselves. There is a strong element of circularity here, unless a clear causal path is appropriate, and the adjustment may be inappropriate or even misleading. For an excellent and thoughtful discussion of this problem in the context of agricultural trials, see Fairfield Smith (1957, §§3, 4).

One general consideration, which applies to randomized and nonrandomized studies alike, concerns the propriety of insisting that treatment comparisons be made at fixed values of the covariate. In the context of agricultural trials for cereal yields, different varieties have different germination rates and hence, if the sowing densities are equal, the numbers of plants per plot will be unequal. Should yields for the different varieties be compared unadjusted at unequal plant densities or should an attempt be made to adjust for unequal germination rates? The analysis would clearly be very different in general: both approaches might be relevant, but in answer to quite different questions. In either case an analysis of germination rates would be essential.

In the context of observational studies, Lord (1967, 1969) gives an example where two methods of analysis, one using covariance adjustment, the other not, lead to correct but apparently contradictory conclusions. We use the same example but our discussion differs

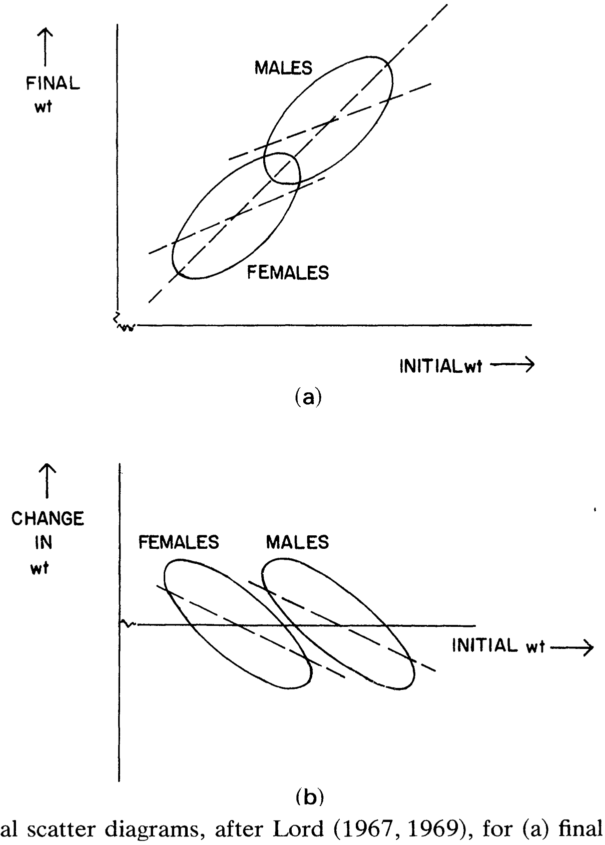

Figure 1. Hypothetical scatter diagrams, after Lord (1967, 1969), for (a) final and initial weight of males (M) and females (F), and (b) change in weight and initial weight of males and females.

slightly from Lord's. Suppose that it is required to investigate the effect of diet on the weight of first-year university students. A large number of students classified only by sex are weighed at the beginning and at the end of term. The response of interest is the gain or loss in weight. It is found that, for males, the weight distribution is unchanged and that the same is true for females. Individual weights have changed, of course, but the overall weight distribution is unaltered. Clearly the average weight gain in each group is zero and therefore the effect of diet on weight gain is the same, namely zero, for both sexes. This is immediately clear from Fig. 1.

Suppose instead that it was decided to analyse the individual changes in weight, this quantity being the response of primary interest, and that initial weight is treated as a covariate. Since initial and final weight are positively correlated for each sex it follows that weight gain is negatively correlated with initial weight. In other words, initially overweight individuals tend to lose weight, and conversely for initially underweight individuals. The problem is illustrated in Fig. 1(b) where it is supposed for the sake of argument that the marginal variances are the same for each sex and that the regression lines are parallel. It is clear from Fig. 1(b) that for each sex the average change in weight is zero, but that after adjustment for initial weight the men gained more weight than the women. This phenomenon is known as Lord's paradox.

Despite the apparent contradiction, both conclusions are clearly correct, as can be seen from inspection of Fig. 1. The paradox is resolved by noting that it is inappropriate to compare males and females at fixed initial weights because this amounts to comparing overweight females with underweight males. Bock (1975, pp. 490–496) gives a lengthier discussion of Lord's paradox.

7. Partitioning of Regression and Covariance

We now turn to problems in which the various lines of the analysis of covariance table are used to describe different aspects of the regression relations present. For simplicity, suppose that there are m groups, and no other structure, that there is a single covariate and that the data are thus (z_{ij}, Y_{ij}), $i = 1, \ldots, m$; $j = 1, \ldots, r_i$.

We assume in the following discussion that the simplest forms of parallelism, constancy of variance, etc. hold; clearly in applications these assumptions need checking and, where appropriate, modifications have to be introduced.

It is convenient to distinguish three broad situations.

Case A. Suppose that the z_{ij} are fixed, possibly controlled by the investigator. Suppose further that for the ith group there is a reference level $z_i^{(0)}$, possibly although not necessarily equal to $\bar{z}_{i.}$, the observed mean in the ith group. Now we can conveniently write

$$E(Y_{ij}) = \mu_i + \gamma(z_{ij} - \bar{z}_{i.}). \tag{14}$$

Thus, for each of the m groups, the reference level and the corresponding expected value define a point P_i,

$$\{z_i^{(0)}, \mu_i + \gamma(z_i^{(0)} - \bar{z}_{i.})\}. \tag{15}$$

We may then consider whether the P_i lie on a smooth curve, e.g. on a straight line. If they do lie on a straight line, say of slope γ^*, we may look for confidence intervals for γ^* and in particular may examine the hypothesis $\gamma^* = \gamma$. This last hypothesis is equivalent to one of total homogeneity,

$$E(Y_{ij}) = \mu + \gamma z_{ij}, \tag{16}$$

and does not depend on the particular choice of reference levels $z_i^{(0)}$.

Now these problems are standard linear ones; the hypothesis that the points P_i lie on a line of slope γ^* is equivalent to

$$E(Y_{ij}) = \mu + \gamma(z_{ij} - \bar{z}_{..}) + (\gamma^* - \gamma)(z_i^{(0)} - \bar{z}_.^{(0)}), \tag{17}$$

where

$$\bar{z}_.^{(0)} = \sum r_i z_i^{(0)} / \sum r_i.$$

Thus, it is routine to find sums of squares for testing relevant hypotheses, and to get estimates and their standard errors. Of course a plot of the estimated points $\hat{P}_i$ is essential.

In the special case $z_i^{(0)} = \bar{z}_{i.}$, the model (17) becomes

$$E(Y_{ij}) = \mu + \gamma(z_{ij} - \bar{z}_{i.}) + \gamma^*(\bar{z}_{i.} - \bar{z}_{..}), \tag{18}$$

γ^* is the regression coefficient between groups in the standard analysis of covariance table, and sums of squares for departure from (18) are readily computed from the analysis of covariance table.

Case B. Suppose now that the m groups are m bivariate populations with means (ν_i, μ_i) and, for simplicity, with the same covariance matrix $\boldsymbol{\Omega}$ within each group. Again such homogeneity must not be taken for granted in applications. Elements of $\boldsymbol{\Omega}$ are estimated from the within-groups line of the analysis of covariance table in the usual way.

If attention concentrates on the points (ν_i, μ_i), a first step will be inspection of the points $(\bar{Z}_{i.}, \bar{Y}_{i.})$; note that we have replaced z by Z because a random variable is now involved. Simple models that may describe (ν_i, μ_i) are

$$\mu_i = \mu_0 + \gamma^*(\nu_i - \nu_0), \tag{19}$$

$$\mu_i = \mu_0 + \gamma^*(\nu_i - \nu_0) + \eta_i, \tag{20}$$

where $\eta_1, \ldots, \eta_m$ are independent random variables of zero mean and variance σ_η^2, independent of the within-group variability, and (ν_0, μ_0) is some notational population mean. Such models were in effect introduced in the present context by Tukey (1951), although (19) has strong connexions with models for so-called functional relationships. Representation (20) is more realistic than (19), which demands exact collinearity; note, however, than (20) in general does induce some asymmetry between the two variables.

A natural way to analyse (20) is by equating the six sums of squares and products in the analysis of covariance table to their expectations in terms of the three elements of $\boldsymbol{\Omega}$, γ^*, $\sum \nu_i^2/m$ and σ_η^2. This is closely related to the components of covariance model to be discussed next.

The connexions between functional models, structural models, canonical correlation analysis and analysis of covariance deserve a separate, more detailed discussion.

Case C. Now suppose that the m groups can be regarded as a random sample from a real or hypothetical population of groups. Make the same working assumptions of homogeneity as before and suppose that the group means have covariance matrix $\boldsymbol{\Omega}^*$. We deal for simplicity with the balanced case in which $r_1 = \cdots = r_m = r$. Then, under assumptions of normality, sufficiency indicates reduction of the data to the analysis of covariance table. If $\mathbf{MS}_w$ and $\mathbf{MS}_b$ denote the matrices of mean squares and products within and between groups, then

$$E(\mathbf{MS}_w) = \mathbf{\Omega}, \qquad E(\mathbf{MS}_b) = \mathbf{\Omega} + r\mathbf{\Omega}^*, \tag{21}$$

leading to the estimate

$$\tilde{\mathbf{\Omega}}^* = (\mathbf{MS}_b - \mathbf{MS}_w)/r. \tag{22}$$

Just as in the one-dimensional case where we may have negative estimates of the 'between' variance, so (22) may not be positive semidefinite. Various modifications are possible, the simplest being to replace negative eigenvalues in the spectral resolution by zero.

There are two main possibilities for further analysis. One is that attention is focussed on a particular aspect of $\mathbf{\Omega}^*$, for example on the regression coefficient of Y on Z, and possibly on its relation with the corresponding feature of $\mathbf{\Omega}$. The second possibility is that one should examine estimates of $\mathbf{\Omega}$ and $\mathbf{\Omega}^*$, or preferably several such pairs of estimates, and look for common features and compact representations. There are very many possibilities. To compare regression coefficients between and within groups we take, in an obvious notation,

$$\tilde{\beta}^* = \frac{MS_{b,yz} - MS_{w,yz}/r}{MS_{b,zz} - MS_{w,zz}/r}, \qquad \tilde{\beta} = \frac{MS_{w,yz}}{MS_{w,zz}}, \tag{23}$$

provided that the denominator of $\tilde{\beta}^*$ is positive.

Two more complex issues, somewhat related to the points above, concern analysis of covariance in designs with hierarchical error structure and analysis of covariance in time series. In the former, for example in split-plot designs, the possibility of differing whole-plot and subplot regressions has to be allowed for. Interpretation of differences in the regressions could be difficult. Two possibilities are first that errors in measuring the concomitant variables are relatively more important at subplot level, and secondly that there is local competition at subplot level. The first would imply that the regression coefficient for subplots is numerically less than that for whole plots, and the second would lead to steeper regression at subplot level. In studying the relation between two or more time series, arranged, say, as years cross-classified with months, analysis of covariance could be used as a convenient way of basing the relationship on the 'high-frequency' component eliminating years and months, using, that is, the 'residual' line of the analysis of covariance table.

Résumé

A partir des principes essentiels nous faisons un exposé sur quelques aspects de l'analyse de covariance. Nous dégageons six différentes significations de l'analyse de covariance et nous résumons rapidement l'histoire de cette technique. Nous décrivons le développement des formules clés à partir de la méthode des moindres carrés, et nous mentionnons des généralisations à d'autres distributions de la famille exponentielle. Nous discutons de quelques problèmes particuliers d'application aux expériences randomisées et aux données d'observation. Enfin, nous considérons la décomposition des relations de régression de concert avec les composantes de la covariance.

References

Anderson, S., Auquier, A., Hauck, W. W., Oakes, D., Vandaele, W. and Weisberg, H. I. (1980). *Statistical Methods for Comparative Studies.* Wiley: New York.

Bickel, P. J., Hammel, E. A. and O'Connell, J. W. (1972). Sex bias in graduate admissions. Data from Berkeley. In *Statistics and Public Policy*, W. B. Fairley and F. Mosteller (eds), 113–130. Reading, Massachusetts: Addison Wesley.

Bock, R. D. (1975). *Multivariate Statistical Methods in Behavioral Research.* New York: McGraw Hill.

Cochran, W. G. (1957). Analysis of covariance: its nature and uses. *Biometrics* **13,** 261–281.

Coons, I. (1957). The analysis of variance as a missing plot technique. *Biometrics* **13,** 387–405.

Cox, D. R. (1957). The use of a concomitant variable in selecting an experimental design. *Biometrika* **44,** 150–158.

Cox, D. R. (1967). Fieller's theorem and a generalisation. *Biometrika* **54,** 567–572.

Cox, D. R. (1982). Randomization and concomitant variables in the design of experiments. In *Essays in Honor of C. R. Rao,* G. Kallianpur, P. R. Krishnaiah and J. K. Ghosh (eds), 197–202. Amsterdam: North-Holland.

Dempster, A. P., Laird, N. and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39,** 1–38.

Fairfield Smith, H. (1957). Interpretation of adjusted treatment means and regressions in analysis of covariance. *Biometrics* **13,** 282–308.

Farewell, V. T. and D'Angio, G. J. (1981). A simulated study of historical controls using real data. *Biometrics* **37,** 169–176.

Finney, D. J. (1957). Stratification, balance and covariance. *Biometrics* **13,** 373–386.

Fisher, R. A. (1932), *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd (13th ed. 1958).

Fisher, R. A. (1935). *The Design of Experiments.* Edinburgh: Oliver and Boyd (7th ed. 1960).

Jöreskog, K. G. (1981). Analysis of covariance structures (with discussion). *Scandinavian Journal of Statistics* **8,** 65–92.

Kempthorne, O. (1952). *The Design and Analysis of Experiments.* New York: Wiley.

Kempthorne, O. (1955). The randomisation theory of experimental inference. *Journal of the American Statistical Association* **50,** 946–967.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin* **68,** 304–305.

Lord, F. M. (1969). Statistical adjustments when comparing pre-existing groups. *Psychological Bulletin* **72,** 336–337.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalised linear models. *Journal of the Royal Statistical Society, Series A* **135,** 370–384.

Robinson, J. (1973). The analysis of covariance under a randomization model. *Journal of the Royal Statistical Society, Series B* **35,** 368–376.

Sanders, H. G. (1930). A note on the value of uniformity trials for subsequent experiments. *Journal of Agricultural Science* **20,** 63–73.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B* **13**, 238–241.

Tukey, J. W. (1951). Components in regression. *Biometrics* **7,** 33–69.

Wilsdon, B. H. (1934). Discrimination by specification statistically considered and illustrated by the standard specification for Portland cement (with discussion). *Journal of the Royal Statistical Society, Supplement* **1,** 152–206.

Wishart, J. (1934). Statistics in agricultural research (with discussion). *Journal of the Royal Statistical Society, Supplement* **1,** 26–61.

Yates, F. (1951). Bases logiques de la planification des experiences. *Annales de l'Institut Henri Poincaré* **12,** 97–112.

Received August 1981

DISCUSSION ON THE PAPER BY PROFESSOR COX AND DR McCULLAGH

Professor Vasant P. Bhapkar (Department of Statistics, University of Kentucky, Lexington, Kentucky 40506, U.S.A.)

Professor Cox and Dr McCullagh have presented a lucid exposition of covariance analysis. They have clearly pointed out how the covariance analysis with linear models can be generalized from the normal case to the general exponential family. I shall elaborate a little on categorical-data applications of such linear-model methodology in the Poisson or the multinomial cases.

I agree with the suggestion in §7 that a distinction should be made between the case in which the covariate z is fixed and that in which the covariate Z is essentially random. For the sake of discussion, consider a three-dimensional contingency table $\{n_{ijk}\}$ with $i=1,\ldots,I$ 'treatments', $j=1,\ldots,J$ levels of the covariate and $k=1,\ldots,K$ levels of the primary response of interest.

If the covariate is regarded as fixed, consider the product–multinomial model with probabilities $\{\pi_{ijk}\}$, where $\sum_k \pi_{ijk}$, $\sum_k n_{ijk}$ are regarded as fixed equal to 1 and n_{ij+}, respectively. In the binary-response case, i.e. $K=2$, let $\lambda_{ij}=\ln(\pi_{ij1}/\pi_{ij2})$ be the logits. Consider models of the type

$$\mathrm{M}_1\colon \lambda_{ij}=\mu+\beta_i+\gamma_j, \qquad \sum\beta_i=\sum\gamma_j=0,$$
$$\mathrm{M}_2\colon \lambda_{ij}=\mu+\beta_i+\gamma z_j, \qquad \sum z_j=0.$$

Here M_2 is a special case of the model M_1 of no interaction between covariates and treatments and is appropriate when z_j represents the score measured from the mean. On the other hand, if there is interaction between the treatments and the covariate, we consider models of the type

$$\mathrm{M}_3\colon \lambda_{ij}=\mu+\beta_i+\gamma_j+\gamma^{(i)}z_j, \qquad \sum\gamma^{(i)}=0,$$
$$\mathrm{M}_4\colon \lambda_{ij}=\mu+\beta_i+\gamma_*^{(i)}z_j.$$

With more than two response levels we consider similar models in terms of multiple logits $\lambda_{ijk}=\ln(\pi_{ijk}/\pi_{ijK})$, $k=1,\ldots,K-1$.

Writing $\theta_{ijk}=\lambda_{ijk}$, these models are all of the type

$$\mathrm{M}\colon \boldsymbol{\theta}=\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{\gamma}, \tag{1}$$

where $\boldsymbol{\theta}_{n\times 1}$ is the 'natural parameter' vector, with $n=IJ(K-1)$, in the pdf

$$p(\mathbf{y};\boldsymbol{\theta})=\exp\{\mathbf{y}'\boldsymbol{\theta}-g(\boldsymbol{\theta})+d(\mathbf{y})\}, \tag{2}$$

where the components of $\mathbf{y}_{n\times 1}$ are not necessarily independent.

With a random covariate, we postulate the product–multinomial model with probabilities $\{\pi_{ijk}\}$, but now $\sum_j\sum_k \pi_{ijk}$, $\sum_j\sum_k n_{ijk}$ are regarded as fixed equal to 1 and n_{i++}, respectively.

In the simple case of binary response ($K=2$) and binary covariate ($J=2$), let

$$\Delta_i=\ln\frac{\pi_{i11}\pi_{i22}}{\pi_{i21}\pi_{i12}}, \qquad \lambda_i^{(R)}=\ln\frac{\pi_{i+1}}{\pi_{i+2}}, \qquad \lambda_i^{(C)}=\ln\frac{\pi_{i1+}}{\pi_{i2+}}.$$

Then we could consider (linear) models in terms of the Δ_i and/or $\lambda_i^{(R)}$ and/or $\lambda_i^{(C)}$ to explore the nature of treatment effect on the measure of association Δ_i or the marginal logits $\lambda_i^{(R)}$, $\lambda_i^{(C)}$. For example, the model

$$\mathrm{M}_5\colon \begin{cases} \lambda_i^{(R)}=\alpha^{(R)}+\beta^{(R)}x_i \\ \lambda_i^{(C)}=\alpha^{(C)}+\beta^{(C)}x_i \\ \Delta_i=\alpha+\beta x_i \end{cases}$$

would be of interest in the case where x_i represents the score (e.g. log dose) for the ith treatment; such relations with the β terms set equal to zero would represent the null hypothesis of no treatment effects (in which case no scores need be postulated).

With arbitrary J and K, we would consider

$$\Delta_{ijk} = \ln \frac{\pi_{ijk}\pi_{iJK}}{\pi_{ijK}\pi_{iJk}}, \qquad \lambda^{(R)}_{ik} = \ln \frac{\pi_{i+k}}{\pi_{i+K}}, \qquad \lambda^{(C)}_{ij} = \ln \frac{\pi_{ij+}}{\pi_{iJ+}},$$

$j = 1, \ldots, j-1$ and $k = 1, \ldots, K-1$. However, we note here that the model M_5 is *not* of type (1), but is of the type

$$\mathrm{M}^*: \mathbf{h}(\boldsymbol{\theta}) = \mathbf{X}\boldsymbol{\beta}, \tag{3}$$

where $\mathbf{h}$ is a 1–1 vector function of the vector of $n = I(JK-1)$ independent natural parameters $\theta_{ijk} = \ln(\pi_{ijk}/\pi_{iJK})$, $j, k \neq J, K$, $i = 1, \ldots, I$.

Alternatively, especially if treatments have an effect on $\lambda^{(R)}_i$ but not on $\lambda^{(C)}_i$, we may focus our attention only on the conditional response, given the covariate level j, and thus consider only the conditional response logits $\lambda^*_{ijk} = \ln(\pi_{ijk}/\pi_{ijK})$. The linear models for the λ^* are essentially of type M_1–M_4, where the covariate z is considered fixed, and thus can be handled as in that case.

With a random covariate another interesting model, is

$$\mathrm{M}_6: \lambda^{(R)}_i = \alpha + \gamma\lambda^{(C)}_i, \qquad J = K = 2;$$

this model is in the same spirit as (19) of Cox and McCullagh. The model M_6 can be expressed in an equivalent form,

$$\mathrm{M}'_6: \frac{\lambda^{(R)}_1 - \lambda^{(R)}_I}{\lambda^{(C)}_1 - \lambda^{(C)}_I} = \cdots = \frac{\lambda^{(R)}_{I-1} - \lambda^{(R)}_I}{\lambda^{(R)}_{I-1} - \lambda^{(R)}_I} = \alpha, \text{ say:}$$

thus it is of the general form

$$\mathrm{M}^{**}: \mathbf{f}_{u\times 1}(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}, \qquad u \leqslant n, \tag{4}$$

where $\boldsymbol{\mu} = \mathrm{E}(\mathbf{Y})$ is a 1–1 vector function of $\boldsymbol{\theta}$. In the model M'_6, $u = I-1$ which is strictly less than $n = 3I$.

It is not quite clear how the model M_6 should be generalized to the case of arbitrary J and/or K. One possibility is to consider models of the type

$$\mathrm{M}_7: \lambda^{(R)}_{ik} = \alpha_k + \sum_{j=1}^{J-1} \gamma_j \lambda^{(C)}_{ij}, \qquad k = 1, \ldots, K-1, \quad i = 1, \ldots, I.$$

A much simpler compromise would be in terms of suitable summary measures such as mean scores [see (19) of Cox and McCullagh], e.g.

$$\mathrm{M}_8: \phi_i = \alpha + \gamma v_i,$$

with $\phi_i = \sum_k w_k \pi_{i+k}$, $v_i = \sum_j z_j \pi_{ij+}$, where $\{w_k\}$, $\{z_j\}$ are the scores for the levels of the response and covariate, respectively.

For the model M^{**}, given by (4), the weighted least squares (WLS) estimates $\tilde{\boldsymbol{\beta}}$ are obtained by minimizing

$$S^2(\boldsymbol{\beta}) = [\mathbf{f}(\mathbf{y}) - \mathbf{X}\boldsymbol{\beta}]'\mathbf{H}_y^{-1}[\mathbf{f}(\mathbf{y}) - \mathbf{X}\boldsymbol{\beta}]$$

and, thus, we have the explicit solution

$$\tilde{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{H}_y^{-1}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{H}_y^{-1}\mathbf{f}(\mathbf{y}); \tag{5}$$

here $\mathbf{H}_\mu \equiv \mathbf{F}_\mu \mathbf{V}_\mu \mathbf{F}'_\mu$ is the asymptotic covariance matrix of $\mathbf{f}(\mathbf{Y})$, with $\mathbf{F}_\mu = [\partial\mathbf{f}(\boldsymbol{\mu})/\partial\boldsymbol{\mu}]$ and $\mathbf{V}_\mu$ is $\mathbf{V} = \mathrm{cov}(\mathbf{Y})$, for the pdf given by (2), when expressed as a function of $\boldsymbol{\mu}$.

For the special case $\boldsymbol{\theta}=\mathbf{f}(\boldsymbol{\mu})$, $\mathbf{H}_y^{-1}=\mathbf{V}_y$, $u=n$ and the WLS estimate is then

$$\tilde{\boldsymbol{\beta}}=[\mathbf{X}'\mathbf{V}_y\mathbf{X}]^{-1}\mathbf{X}'\mathbf{V}_y\mathbf{f}(\mathbf{y}). \tag{6}$$

This form may be compared to the form of the ML (and WLS) estimate in the normal case (with known $\mathrm{cov}(\mathbf{Y})=\mathbf{V}$), namely

$$\tilde{\boldsymbol{\beta}}=[\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}. \tag{7}$$

The apparent discrepancy in the forms (6) and (7) is seen to arise because the estimate in (6) refers to the exponential family model $\boldsymbol{\theta}=\mathbf{X}\boldsymbol{\beta}$ which may be written as $\mathrm{M_E}$: $\mathbf{f}_{n\times1}(\boldsymbol{\mu})=\mathbf{X}\boldsymbol{\beta}$, while the estimate in (7) refers to the normal model $\mathrm{M_N}$: $\boldsymbol{\mu}=\mathbf{X}\boldsymbol{\beta}$. The goodness of fit of the general model M** can be tested using $S^2(\tilde{\boldsymbol{\beta}})$ as a large sample χ^2-criterion on $u-d_x$ df, d_x being the rank of $\mathbf{X}$.

Occasional singularities of $\mathbf{H}_y$, arising from singularity of $\mathbf{V}_y$, and infinite values of $\mathbf{f}(\mathbf{y})$ can be handled as in Berkson (1953) or Cox (1970). The WLS estimates, test criteria etc. have the same first-order large-sample efficiency as the likelihood procedures. However, they are computationally simpler and are especially convenient for dealing with models of the type M** in the presence of nuisance parameters, i.e. with $u<n$.

References

Berkson, J. (1953). A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function. *Journal of the American Statistical Association* **48,** 565–599.

Cox, D. R. (1970). *The Analysis of Binary Data.* London: Methuen.

Professor D. J. Finney (Department of Statistics, University of Edinburgh, The King's Buildings, Edinburgh EH9 3JZ, Scotland).

To open discussion on a paper by Professor Cox and Dr McCullagh is difficult, especially for anyone brought up in the British tradition of emphasizing the errors and shortcomings of the paper discussed. We have had today a masterly account of the many applications of covariance analysis and their relation to general statistical theory. I shall comment briefly on a few points of practice.

Professor Cox and Dr McCullagh placed less emphasis than I expected on the distinction between the narrow literal sense of covariance analysis, partitioning a total sum of products in the same way as analysis of variance partitions a sum of squares, and the broader sense that comprises also the use of an error regression function for adjusting treatment or group means of one variable so as to take account of inequalities in the means of others. In the first sense, the analysis forms an integral part of most procedures for multivariate analysis that are based on linear parameter formulations. Here it is inescapable. As for the broader sense, if it did not exist would we now find any reason to invent it? I am inclined to think that we might prefer to achieve the same results in other ways; because it *does* exist and is well established in our literature, I see much to be said for continuing to use covariance analysis.

The second sense is of course that in which the term was first used, for a method that can now be seen as bridging the gap between analysis of variance and multiple linear regression. Today we realize that all analysis of variance can be formally regarded as multiple regression; though to execute computations in this way is apt to obscure rather than aid interpretation, it can be useful for the occasional data structure not covered by locally-available programs. Equally, general multiple linear regression can be studied in terms of nonorthogonal analysis of variance. The practical merit of covariance analysis is that it permits easy extension of computations for a symmetric design to various complications that seriously affect the orthogonality. These include (i) the 'classical' adjustment for an independent variate unaffected by treatment; (ii) missing plots; (iii) treatments interchanged between blocks; (iv) mixed-up yields; (v) fertility trends; and (vi) additional design constraints of a less symmetric character. The list is not exhaustive, and the items listed are not completely distinct; Outhwaite and Rutherford (1955) gave a good example of (v) and (vi). Cox and McCullagh have emphasized that missing values can now be handled in more convenient ways, and their comment would presumably apply to interchanges and mixing, but I doubt whether any other method is equally flexible in its applicability.

One development that I thought would be mentioned is the use of general nonlinear regression functions. Of course, if the concomitant variates are used in polynomials or in other functions linear in the parameters no new considerations enter. However, one major benefit that modern computing power brings to statistics is the practicability of using more realistic regression functions in place of polynomials, and these commonly have parameters that do not enter linearly. My impression is that in the covariance problem these would most easily be handled by general nonlinear regression, regarding as nuisance parameters those that relate purely to concomitants. Here might seem to be a further reason for questioning continued use of the standard covariance computations; in view of the infrequency with which even quadratic functions have contributed much to the improvement of precision through covariance, I am disposed to attach little weight to this objection.

The interesting final section on partitioning, especially Case A, has evident connexions with bioassay and calibration that may be worth pursuing further. I shall resist the temptation to speak at length on this topic.

Though the point is somewhat apart from their main theme, I was glad to hear the authors' comment on nonlinear transformations. Too many textbooks carry unthinking rules about transforming data that show particular variance patterns, and fail completely to recognize that the original scale may be needed for useful interpretation. I was surprised by the authors' remarks on clinical trials. Is it not uncommon for a new clinical trial to be so similar to another recent trial that the question of using the same control subjects will arise? Usually a difference in purpose for the new trial, a change in diagnostic standards, a need for additional information no longer obtainable for the earlier controls, or some other similar factor will interfere. 'Historical controls' are too often a miscellaneous collection of patients from the investigator's past experience or from case records to which he has access. There are situations in which historical controls provide the only ethical or practicable way of proceeding. Although such controls should always be seen as a last resort, almost invariably open to severe criticism, statistical techniques that assist in interpreting them are needed. We should beware of conveying to our medical friends any notion that we have now cleverly devised ways of escape from difficulty; of course, the authors make no such suggestion, nor do they in any way advocate the use of historical controls when randomized contemporary controls are possible. Farewell and d'Angio, in the 1981 paper referred to by the authors, describe a most interesting example where a strong case for at least partial dependence on apparently very suitable historical controls seemed attractive; their finding of no advantage for the total economy of the research is discouraging, and has a similarity to work of my own (Finney, Holt and Sheffield, 1975) in a very different context.

I have the impression that, at least in the United Kingdom, concern for animal welfare is going to lead to much the same arguments about ethics of experimentation on animals as have been encountered in human medicine. This may prove both a challenge and an opportunity for statisticians. We need to be clear that an experiment skimped in numbers or weak in logic may be less ethical than one that exposes more animals to pain or risk yet in its logical soundness leads to a definitive conclusion. This is one reason among many for urging that in all fields of biological experimentation increased efforts be made to exploit covariance analysis. I am always surprised by the relative neglect of the technique in situations where a covariate potentially useful for increasing precision either is available or could have been for little extra cost.

While reading this paper, I looked again at the 1957 papers by Cochran and Fairfield Smith. How impressive these still are in their breadth of coverage! These earlier authors provided accounts of covariance that were the first clear explanations of purpose and limitations since Fisher's original work, and that remain valuable today. Perhaps nothing else so important has been published on this topic until the present paper.

References

Finney, D. J., Holt, L. B. and Sheffield, F. (1975) Repeated estimations of an immunological response curve. *Journal of Biological Standardization* **3**, 1–10.

Outhwaite, A. D. and Rutherford, A. A. (1955) Covariance analysis as an alternative to stratifica*tion in the control of gradients. Biometrics* **11**, *431–440.*

Professor R. R. Hocking (Institute of Statistics, Texas A&M University, College Station, Texas 77843, U.S.A.)

When I was asked to serve as a discussant on this paper, it was suggested that I concentrate on the linear-model aspects of the analysis of covariance As a long-time student and practitioner of linear-model analysis, I continue to be amazed by how slowly new developments find their way into either our textbooks on statistical methods or our computer packages devoted to linear-model analysis.

As Professor Cox and Dr McCullagh note, the analysis of the general covariance model

$$\mathrm{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} \qquad (1)$$

is particularly easy if the analysis of variance model

$$\mathrm{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \qquad (2)$$

has a simple structure. The computations for the analysis of the general model (1) are conveniently displayed in terms of quadratic and bilinear forms suggested by the ANOVA model (2). The student in statistical methods is then presented with an analysis of covariance table containing these quantities and is asked to accept the fact that certain manipulations with these quantities will test hypotheses on the coefficients of the concomitant variables or on the treatments being examined. While this approach is convenient for hand computations, it is, at best, confusing to the student. I support the authors' suggestion that this approach is less important now that flexible computer programs are becoming quite generally available. I would prefer that the computational problems be assigned to the computer while the textbooks devote space to the concepts of the analysis and interpretation of results.

To a consultant, it is interesting to note how those in various disciplines view the covariance model. The experimental scientist is generally concerned with the ANOVA portion of the model and views the covariates or quantitative variables as necessary evils to improve the precision of the analysis. On the other hand, the social scientist is primarily concerned with the regression portion of the model and views the design or qualitative variables as necessary evils to improve the analysis. It is of interest to note that both groups almost invariably assume parallelism. That is, they assume that there is no interaction between the qualitative and quantitative variables. In my experience, the possibility of nonparallel responses is rarely raised by either group.

The effect of this dichotomy of users is most apparent in our computer packages. The experimental scientist will typically use a general linear-model program, designed primarily for ANOVA models, but which will allow covariates. These programs have not incorporated the recent work on regression diagnostics even though it is certainly applicable. Residual plots are desirable even for ANOVA programs and indicators of multicollinearity and high-leverage observations should be an essential part of the output for a covariance analysis. The possibility of observations with unduly high influence is a reality. One of the convenient indicators of high leverage is the 'Hat' matrix described by Hoaghlin and Welsch (1978). The Hat matrix, **H**, is defined by the relation

$$\hat{\mathbf{Y}} = \mathbf{HY} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\boldsymbol{\gamma}}. \tag{3}$$

Thus, **H** indicates the effect of each response on the 'predicted' vector, $\hat{\mathbf{Y}}$. It is of interest to note that for the covariance model the Hat matrix is the sum of the two components. That is,

$$\mathbf{H} = \mathbf{H}_X + \mathbf{H}_{Z|X}. \tag{4}$$

Here, $\mathbf{H}_X$ is the Hat matrix associated with the design matrix and is given by

$$\mathbf{H}_X = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T. \tag{5}$$

The second term gives the leverage of the quantitative variables 'adjusted' for the qualitative variables. This is, in fact, the Hat matrix as derived from the so-called 'error' normal equations [equation (9) of Cox and McCullagh] and is given by

$$\mathbf{H}_{Z|X} = \mathbf{R}_X\mathbf{Z}(\mathbf{Z}^T\mathbf{R}_X\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{R}_X, \tag{6}$$

where

$$\mathbf{R}_X = \mathbf{I} - \mathbf{H}_X. \tag{7}$$

I suspect that $\mathbf{H}_{Z|X}$ might be a more appropriate indication of leverage for the quantitative variables.

The social scientist will typically use a regression program for the analysis and hence is likely to include regression diagnostics in his analysis. The analysis of the qualitative variables may be of secondary interest but recent papers on hypothesis testing in ANOVA models such as Hocking, Hackney and Speed (1978) suggest that this analysis may not be clearly understood. The primary reason for this misunderstanding is that the regression user typically generates the full-rank design matrix in a way which may lead to incorrect interpretation. For example, with two qualitative variables and their interaction the 'design' portion of the model is often written as

$$\mathrm{E}(y_{ijk}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}. \tag{8}$$

The usual full model arises by requiring the conditions

$$\alpha_. = \beta_. = (\alpha\beta)_{i.} = (\alpha\beta)_{.j} = 0, \tag{9}$$

while the regression programs frequently use the conditions

$$\alpha_1 = \beta_1 = (\alpha\beta)_{i1} = (\alpha\beta)_{1j} = 0. \tag{10}$$

The ANOVA hypotheses in the analysis of covariance are simply hypotheses about the intercepts of the several regression lines. With conditions (9), the 'main effect' hypotheses compare the intercepts for the different levels of, say, Factor A averaged over levels of Factor B. With condition (10), the 'main effect' hypothesis compares the intercept for the different levels of Factor A at only the first level of Factor B. The analysts are rarely aware of this subtle but important difference.

To summarize, I would like to thank Professor Cox and Dr McCullagh for their enlightening presentation, and charge the authors of text books and the writers of computer programs with the responsibility of clarifying and modernizing their treatment of analysis of covariance.

References

Hoaglin, D. C. and Welsch, R. E. (1978). The hat matrix in regression and ANOVA. *American Statistician* **32,** 17–22.

Hocking, R. R., Hackney, O. P. and Speed, F. M. (1978). The analysis of linear models with unbalanced data. In *Contributions to Survey Sampling and Applied Statistics*, H. A. David (ed.) New York: Academic Press.

The authors replied as follows:

We are grateful to the contributors, whose contributions add valuable clarification to our paper. There is little that is controversial and our reply can therefore be brief.

We agree with Professor Finney and Professor Hocking that analysis of covariance as an algorithm for least squares computations is no longer of critical importance: we have tried to emphasize other aspects of the topic. The difference in types of applications that Professor Hocking notes is largely that between experimental and observational studies: we consider that examination for parallelism, not necessarily by formal significance tests, should be more widely used in both kinds of application.

We are rather concerned that Professor Hocking in his penultimate paragraph appears to favour estimating and testing main effects in the presence of interaction, a thing we consider rarely physically meaningful.

We agree with Professor Finney's valuable remarks about nonlinear regression.

Professor Bhapkar has set out the framework of least squares analysis with empirically estimated weights for various hypotheses for categorical data. Usually conclusions from this method will be very close to those from maximum likelihood. It would certainly be interesting to know more about the circumstances under which appreciable differences between the two methods can arise and what is then the sensible interpretation.