Analysis of Covariance: A Delicate Instrument[1]

JANET D. ELASHOFF
Stanford University

Covariance analysis is a popular technique. It is widely used to "adjust" criterion scores such as achievement for the effects of a covariate such as ability in order to compare several treatments. However, analysis of covariance procedures must be applied and interpreted with care. As Kendall (1946) put it:

> . . . we would emphasize that the analysis of variance [and covariance], like other statistical techniques, is not a mill which will grind out results automatically without care or forethought on the part of the operator. It is a rather delicate instrument which can be called into play when precision is needed, but requires skill as well as enthusiasm to apply to the best advantage. The reader who roves among the literature of the subject will sometimes find elaborate analyses applied to data in order to prove something which was almost obvious from careful inspection right from the start; or he will find results stated without qualification as "significant" without any attempt at critical appreciation.

Here, I will attempt to describe more exactly just what covariance analysis does and doesn't do, to point out the advantages and limitations of the technique, and to discuss the conditions the data must satisfy for covariance analysis to be a valid technique.

Section II contains a brief review of covariance analysis, the underlying model and necessary assumptions.

Section III contains a discussion of the assumptions about the data which must be satisfied if covariance is to be a valid technique,

[1]This research was carried out at the Stanford Center for Research and Development in Teaching at Stanford University, pursuant to a contract with the United States Department of Health, Education and Welfare, Office of Education under the provisions of the Cooperative Research program.

a description of the effects on the covariance procedure if an assumption is not satisfied, and suggestions for checking the assumptions.

Section IV contains a discussion of covariance analysis in relation to other techniques: adjustments using between-groups regression, and matching or blocking.

MODEL AND ASSUMPTIONS

Covariance analysis is designed for the following type of experiment. Suppose that a total of nt individuals are selected at random from a population of interest. A measurement x is made on each individual. Then n individuals are assigned at random to each of t treatments. After the treatment has been applied a criterion measurement y is made for each individual. Thus we have two scores x_{ij} , y_{ij} for the i^{th} individual in the j^{th} treatment ($i = 1, \ldots, n, j = 1, \ldots, t$). The research questions are: are the average criterion scores significantly different for the t treatments? What are good estimates of the average criterion scores in each treatment?

The research questions could be answered by ignoring the x measurements and performing an ordinary analysis of variance on the criterion scores y_{ij}. However, since the x variable is thought to have a large influence on the criterion variable we might like to answer the same questions with y scores "adjusted" for x. We might then compare treatments using a covariance analysis to "adjust" criterion scores for the x scores if the necessary assumptions can be satisfied. The covariance procedure would reduce possible bias in treatment comparisons due to differences in the covariate x and increase precision in the treatment comparisons by reducing variability in criterion scores "due to" variability in the "covariate," x.

For example, suppose that the "treatments" are teachers of high school French. Students are given the Modern Language Aptitude Test (MLAT), the x variable, at the beginning of the course and a standard French achievement examination, the y variable, at the end of the first semester. The investigator wishes to estimate the effectiveness of each teacher using the class achievement scores but adjusting for the initial language ability of the class. (In this kind of investigation it often happens that assignment to treatments is not at random.)

The statistical model for analysis of covariance is composed of four independent terms

$$y_{ij} = \mu + \alpha_j + \beta(x_{ij} - \bar{x}) + e_{ij} \tag{1}$$

where μ is the mean of the criterion variable y across individuals

and treatments and α_j is the differential effect on the mean due to treatment j. In the covariance model, however, individual variability in achievement is not lumped into one term but divided into two parts—variability due to a linear regression on x, $\beta(x_{ij} - x)$, and unexplained variability e_{ij}. The e_{ij} are assumed to be an independent random sample from a normal distribution with mean zero and variance σ^2_e. The null hypothesis to be tested is that $\alpha_j = 0$ for all j—that is, there is no difference among treatments not due to differences in the covariate. For an introduction to covariance analysis and the details of the computations, see Cochran (1957), Dixon & Massey (1957), Myers (1966) or Winer (1962).

Thus, the analysis of covariance is a valid technique for testing for differences in average criterion scores among treatments if we can assume:

a) random assignment of individuals to treatments,
b) within each treatment, criterion scores have a linear regression on x scores,
c) the slope of the regression line is the same for each treatment (there is no slope-treatment interaction),
(2) d) for individuals with the same score x, in the same treatment, criterion scores, y, have a normal distribution,
e) the variance of the distribution of y scores for all students with the same x score in a particular treatment is the same for all treatments and x scores.
f) criterion scores are a linear combination of independent components: an overall mean, a treatment effect, a linear regression on x, and an error term.

Since the covariate model

$$y_{ij} = \mu + \alpha_j + \beta(x_{ij} - \bar{x} + e_{ij}$$

can be rewritten as

$$y_{ij}{}^a = y_{ij} - \beta(x_{ij} - \bar{x}) = \mu + \alpha_j + e_{ij},$$

covariance analysis could be viewed as an analysis of variance on adjusted scores $y_{ij}{}^a$ if β were known. However, since β is estimated from the data, the adjusted treatment means $\bar{y}_j' = \bar{y}_j - b(\bar{x}_j - \bar{x})$ are correlated, and the numerator of the covariance F test for differences in adjusted means is *not* identical to

$$\sum_{j=1}^{t} n \ (\bar{y}'_j - \bar{y})^2. \ / \ (t-1).$$

In a randomized experiment the gain in precision obtained by making treatment comparisons using covariance analysis based on linear regression rather than by using analysis of variance depends on the correlation between x and y within treatments. If the within-treatment correlation between x and y is ρ, then Cochran (1957) states "If σ^2_y is the experimental error variance when no covariance is employed, the adjustments reduce this variance to a value which is effectively about

$$(3) \qquad \sigma^2_y \ (1 - \rho^2) \ \left(1 + \frac{1}{f_e - 2}\right)$$

where f_e is the error number of d.f." Therefore, if ρ is smaller than 0.3 in absolute value the increase in precision of covariance analysis over analysis of variance will be negligible.

The validity of the analysis of covariance depends on how closely the data satisfy the necessary assumptions. Some of the assumptions—such as normality—are necessary for statistical convenience, others like no treatment-slope interaction are necessary for the covariance technique to be logically meaningful. In the next section we discuss each of the assumptions in detail, describe the effects on the analysis of covariance when an assumption is not satisfied, and suggest methods for checking the data for the validity of the assumptions. Each assumption will be considered separately—very little investigation has been made of situations where more than one assumption is unsatisfied.

EFFECTS OF ASSUMPTIONS

Randomization

The analysis of covariance is based on the assumption that individuals are randomly assigned to treatmnet groups and that all groups are treated exactly the same except for treatments. However, covariance adjustment procedures are often recommended for reducing bias due to the covariate in studies where the experimenter must work with intact groups. Covariance analysis can indeed be useful where assignment to groups is not random but the results must be interpreted with caution. Winer (1962) remarks, "At best covariance adjustments for initial biases on the covariate are poor substitutes for direct controls." (See the section on blocking at the end of this paper.)

Evans and Anastasio (1968) distinguish three separate cases: 1) individuals are assigned to groups at random and treatments are assigned to groups at random, 2) intact groups are used but treatments are assigned to groups at random, 3) intact groups and

treatments occur together naturally. Covariance analysis is appropriate for case 1, may be used with caution for case 2, and is likely to be misleading when used for case 3 (see page 388).

There are two major difficulties involved in the interpretation of covariance analysis when individuals are not randomly assigned to treatments. First, we can never be sure that the covariance adjustment has removed all bias—some bias may still be present from a disturbing variable which was overlooked. Secondly, when the x variable shows real differences among groups covariance adjustments may involve extrapolation.

To illustrate, suppose that teaching methods A and B are to be compared and that the class using method A was composed entirely of high ability students while the method B class consisted only of low ability students. A covariance analysis is performed and, on the basis of the within-class relationships between ability and achievement, we obtain mean achievement scores for methods A and B which are estimates of the achievement scores that would have been observed if methods A and B had been used on classes of equal and average ability.

Bias may still be present. Perhaps the high ability class volunteered to take the class, had fewer students, or the low ability students were repeating the subject. It may make no sense to think about comparing methods A and B for students of average ability, perhaps each has been designed specifically for the ability level it was used with, or neither method will, in future, be used for students of average ability. Statistically adjusting for ability is *not* the same as holding ability constant.

The regression slope used in the adjustment must be based entirely on high and low ability scores and there is no assurance that a linear regression holds throughout the entire ability range. The relationship between ability and achievement may be different for students of intermediate ability.

The farther apart the two groups are in mean ability the more imprecise is the estimate of the difference in the adjusted treatment means. The variance of the difference between the adjusted means of the high and low ability groups depends on the square of the difference between their mean ability scores. (Any imprecision in estimating the slope of a line makes more difference in the estimated position of the line the further one extends the line from the overall mean.) Accordingly, Cochran (1957) states ". . . the adjusted difference may become insignificant statistically merely because the adjusted comparisons are of low precision."

Lord (1963) expresses the issue strongly—the random assignment of individuals to treatments

> . . . is the logical (not merely the statistical) prerequisite to a controlled experiment. If the individuals are not assigned to the treatments at random, then it is not helpful to demonstrate statistically that the groups after treatment show more difference than would be expected by random assignment—unless, of course, the experimenter has special information showing that the non-random assignment was nevertheless random in effect. If, as often happens, randomized assignment is impossible, then there is often no way to determine what is the appropriate adjustment to be made for initial differences, and hence often no way to show convincingly by statistical manipulations that one treatment is better than another.

To ensure that the assumption of randomization is satisfied, assignment to groups and treatments must be made at random. There is no way to tell afterwards if assignment has indeed been random in effect. If other variables known or suspected to influence achievement have also been measured they may be included in the adjustment or separately checked for possible contributions to bias by comparing their distributions in each group (see Cochran (1965)).

Covariate Independent of Treatment

A basic postulate underlying the use of analysis of covariance to adjust treatment means for the effects of the covariate x is that the x variable is *statistically* independent of the treatment effect. In other words, this means that the distribution of covariate values is not affected by the treatments either through direct causation or through correlation with another affected character (and the x variable does not affect the treatment). To achieve this statistical independence, the x variable should be measured prior to the administration of treatments and treatments should be assigned to groups at random. Therefore, if the treatments themselves produce the differences in covariate means or the treatments are not manipulated as independent variables but are classifications of naturally occurring groups this assumption will not be valid. Evans and Anastasio (1968) comment that

> . . . when treatment and covariate are inherently related, there is likely to be a strong linear correlation between treatment effects and covariate means. This circumstance . . . makes the assumption of homogeneity of between-group and within-group regression quite generally untenable.

When the covariate x is affected by the treatment, the regression adjustment may remove part of the treatment effect or produce a spurious treatment effect. If, for example, achievement were adjusted for study time but amount of study time were an inherent characteristic of the teaching methods, then a comparison of such adjusted

achievement means is meaningless. Actually holding study time constant would have changed the treatments.

An analysis of variance of the covariate may be useful as an indication of whether or not the treatments are affecting the covariate. Analysis of covariance is inappropriate if the covariate is not independent of the treatment.

Covariate Measured Without Error

Analysis of covariance procedures are based on the assumption that the concomitant variable, x, is measured without error. Lord (1962) says,

> Making allowances for initial differences among groups on a poor measure of some variable is not the same thing as making allowances for initial differences on the variable itself. If the variable in question cannot be reliably measured, it should be controlled experimentally (by randomization) if possible. Otherwise, some special modification of the analysis of covariance is desirable (Lord, 1960).

Cochran (1968) discusses the effect of errors of measurement in the x variable on covariance analysis. Suppose that the covariance model (1) holds where x is an individual's "true" score. However, we can only measure X

$$X = x + d \tag{4}$$

where d is the error of measurement. If x and d are normally and independently distributed and the regression of y on x has slope β, then the regression of y on X will have slope $\beta \left(\frac{\sigma^2_x}{\sigma^2_x + \sigma^2_d}\right)$ — or β times the reliability of X.

The error of measurement in X will decrease the precision of the experiment by increasing the error variance to $\sigma^2_e + \beta^2\sigma^2_x (1 - R_X)$ where $R_X = \frac{\sigma^2_x}{\sigma^2_x + \sigma^2_d}$, the reliability of X. The difference in adjusted treatment means will have expected value

$$\alpha_2 - \alpha_1 + \beta (1 - R_X) (\mu_{2x} - \mu_{1x})$$

where μ_{jX} is the population mean of x for treatment j. Thus with random assignment to treatments the adjusted treatment means will still be unbiased but if real differences in the mean of x exist in different treatments the covariance adjustment will not remove all the bias due to x. (In large samples, a fraction, $1 - R_x$, of the bias due to differences in x will remain in y even after the covariance adjustment using X). Note that if y were assumed to have a linear regression

on X, or "measured ability" this problem would not arise, we would simply adjust for X and ignore x.

If the x variable cannot be measured reliably, additional assumptions must be made about the model or additional information about the magnitude of the measurement error must be obtained in order to derive an appropriate covariance procedure.

Lord (1960) presents a large-sample analysis of covariance procedure for two treatment groups where each individual has one measurement on the criterion variable and two duplicate measurements on a covariate which is subject to measurement error. Porter (1967) developed a covariance procedure for t groups using estimated true scores for x for situations where an estimate of the reliability of X is available. He also investigated the effects of different parameter values on Lord's procedure and his own "true score" procedure.

Linearity

There are many ways in which the criterion variable, y, and the covariate x could be related. This relationship must be known or estimated in order to "adjust" y scores for the effects of the x scores. The appropriate form of adjustment must be based on theoretical grounds, prior experimentation, or examination and analysis of x, y scatter plots. If an incorrect adjustment is made, the assumptions made about the residuals e_{ij} (i.e., normality, homogeneity of variances, etc.) will generally not hold, not to mention the questionable meaning of the "adjusted" scores.

While a covariance type of analysis may be developed for many models describing the relationship between x and y, the standard covariance analysis assumes that the covariate has a *linear* relationship with the criterion variable (see equation (1)). It is often argued that a linear model may provide an adequate fit for other relationships between x and y and is simple to use and interpret. But how seriously are the results of standard covariance analysis affected if the true relationship between x and y is not linear?

Atiqullah (1964) investigated the effect on treatment comparisons obtained using standard covariance analysis for the case where the true regression of y on x is quadratic:

$$(5) \qquad y_{ij} = \mu + \alpha_j + \beta(x_{ij} - \bar{x}) + \theta(x_{ij} - \bar{x})^2 + e_{ij}.$$

For comparisons of two treatments, the estimated mean difference, $\hat{\alpha}_1 - \hat{\alpha}_2$, derived under the standard covariance model (1) is biased unless the x observations are random samples from the same normal

population. The amount of bias depends on θ, $\bar{x}_1 - \bar{x}_2$, $\Sigma(x_{i1}-\bar{x}_1)^2 - \Sigma(x_{i2}-\bar{x}_2)^2$, $\Sigma\Sigma\ (x_{ij}-\bar{x}_j)^2$. (Note that the effect of nonlinearity is most severe when random assignment to groups is not possible or protection against non-normality in the y's is lowest).

Asymptotic results (for t large) indicate that comparisons of t treatment means (including the overall F test for adjusted treatment differences) are seriously biased even if the x observations have the same normal distribution in each group unless the coefficient of the quadratic term, θ, is quite small.

The simplest check for linearity is a carefully prepared set of x-y scatter plots for each treatment group. Gross departures from linearity will be easily discovered. A test for linearity of regression is given in Hays, Chapter 16. If that technique is not applicable a general regression program may be used to examine the increase in explained sum of squares due to the addition of a quadratic or cubic term in x.

If the linearity assumption is not satisfied, adjustments should be based on a more accurate model of the relationship between x and y. Transformation of the data may be helpful in producing a linear relationship (see Kruskal (1968)). In future experiments involving x and y, matching or blocking may be a useful alternative procedure, see the section on matching.

Homogeneity of Regression

The standard covariance analysis procedure rests on the assumption that the regression of y on x is linear, and that the slope is the same for all treatment groups (there is no treatment-slope interaction). The model is:

$$y_{ij} = \mu + \alpha_j + \beta(x_{ij} - \bar{x}) + e_{ij}. \tag{1}$$

If there is a treatment-slope interaction an alternative model is:

$$y_{ij} = \mu + \alpha_j + \beta_j(x_{ij} - \bar{x}) + e_{ij}. \tag{6}$$

If model (1) obtains and there is no treatment-slope interaction, then the treatment which is best on the average (has the largest α_j) is also best at each level of x by the same amount. However, if there is a treatment-slope interaction as in model (6) where $\beta_j \neq \beta_j'$ then the treatment which is best on the average may not be best at all x levels. If model (6) holds it is questionable whether a covariance-type of analysis is relevant. If the slopes differ markedly from group to group an investigation of the treatment-slope interaction and a comparison of treatments at each level of x may be appropriate.

The assumption that there is no treatment-slope interaction may be checked by comparing scatter plots of y versus x for each treatment group. A test for the equality of the slopes of the regression lines within the groups, i.e., a test of H_0: $\beta_1 = \beta_2 \ldots = \beta_t$, is given by Winer (1962), p. 587.

Suppose that covariance analysis still seems appropriate even though (6) is the correct model. What effect would this have on the treatment comparisons obtained using a standard covariance analysis?

Atiqullah (1964) derived $E(\hat{\alpha}_1 - \hat{\alpha}_2)$ and $Var\ (\hat{\alpha}_1 - \hat{\alpha}_2)$ for comparisons involving two treatment groups and found the approximate distribution of the overall F for t treatments. Atiqullah distinguishes three situations: (1) there are only two treatment groups, (2) we are comparing two treatment means but there are t treatment groups, (3) comparisons of t treatments ($t > 2$). He introduces notation

$$W_{jj} = \sum_i (x_{ij} - \bar{x}_j)^2 \quad and \quad W_2 = \Sigma W_{jj}.$$

When there are only two treatment groups and model (6) is correct, the difference in adjusted treatment means, $\bar{y}'_1 - \bar{y}'_2$, obtained by a standard covariance analysis is biased unless the mean of the concomitant variable x is the same in both groups or the variance of the x's is the same in both groups.

$$E\ (\hat{\alpha}_1 - \hat{\alpha}_2 \mid (6)) = \hat{\alpha}_1 - \hat{\alpha}_2 - (\beta_1 - \beta_2)\ (\bar{x}_1 - \bar{x}_2)\ \frac{(W_{11} - W_{22})}{2W_2} \tag{7}$$

$$Var\ (\hat{\alpha}_1 - \hat{\alpha}_2) = \sigma^2 \left(\frac{2}{n} + \frac{(\bar{x}_1 - \bar{x}_2)^2}{W_2} \right)$$

Although the formula for the variance of $\alpha_1 - \alpha_2$ obtained under model (1) is still correct for model (6) the estimate of σ^2 obtained using model (1) will be an overestimate if model (6) is true. Writing $r_{ij} = y_{ij} - \mu - \alpha_j - \beta(x_{ij} - \bar{x}) = (\beta_j - \beta)\ (x_{ij} - \bar{x}) + e_{ij}$ as the "true residuals" when model (6) holds but the model (1) approach is used, we find

$$\sigma^2_r = \Sigma\Sigma (\beta_j - \beta)^2\ (x_{ij} - \bar{x})^2 + \sigma^2_e \tag{8}$$

which is always larger than σ^2_e by an amount depending on the $(\beta_j - \beta)^2$. Under these circumstances use of model (1) when model (6) obtains can be expected to result in a conservative F test even when $\bar{x}_1 = \bar{x}_2$ or $W_{11} = W_{22}$. Atiqullah recommends "in the absence of a prior presumption that β_1 and β_2 are nearly equal, the model L_2 [(6)] should be used, separate regressions fitted and treatment effects estimated as a function of x."

When there are t treatment groups and model (6) holds, adjusted differences between pairs of treatments $(y'_k - y'_m)$ obtained from a standard covariance analysis are biased unless the x observations in each treatment group have the same mean as well as the same variance.

(9)

$$E\ (\hat{\alpha}_1 - \hat{\alpha}_2) = \alpha_1 - \alpha_2 + \frac{1}{W_2}\Big[\ ((\bar{x}_1 - \bar{x})\ W_{22} + (\bar{x}_2 - \bar{x})\ W_{11})\ (\beta_1 - \beta_2)$$

$$+ (\bar{x}_1 - \bar{x}) \sum_{l=3}^{t} W_{1l}\ (\beta_1 - \beta_l) - (\bar{x}_2 - \bar{x}) \sum_{l=3}^{t} W_{1l}\ (\beta_2 - \beta_l)\ \Big].$$

Atiqullah's asymptotic results indicate that the standard between-groups F-test for t treatments may yield misleading results when model (6) holds, even if y or x or both of them follow normal distributions. In the special case where x has the same normal distribution in all treatment groups, $\sigma^2_x/\sigma^2_e < 1$, and σ^2_β is of the order of $1/t^2$, (where

$$\sigma^2_\beta = \sum_{j=1}^{t} \frac{(\beta_j - \bar{\beta})^2}{t} \text{ and } \bar{\beta} = \sum_{j=1}^{t} \frac{\beta_j}{t}\ ),$$

the presence of slope-treatment interaction is not likely to affect the F-test seriously. (Note that in a pre-test, post-test situation it is unlikely that $\sigma^2_x/\sigma^2_e < 1$.)

Peckham (1968) made a Monte Carlo investigation of the distribution of the F-test on adjusted means for several sets of parameters β_j. In Phase I he generated x values for each treatment group from the same normal distribution with the restriction that $\bar{x}_j = \bar{x}_{j'}$ and $W_{jj} = W_{j'j'}$ all j, j'. In Phase II he allowed $\bar{x}_j \neq \bar{x}_{j'}$ but fixed $t = 2$. He found that the F-tests were increasingly conservative as σ^2_β increased.

Normality

Covariance analysis is based on the assumption that within each group at each ability level the criterion variable, y, has a normal distribution. That is, the residuals e_{ij} are normally distributed. How robust is the technique to non-normality?

The ordinary analysis of variance F-tests for differences in treatment means in the standard orthogonal cross-classifications are not strongly affected by non-normality (see Atiqullah 1962)). However, Atiqullah (1964) found that the analysis of covariance F-test is ". . . appreciably affected by non-normality even in balanced classifications. The degree of sensitivity to non-normality is determined by the distribution of the concomitant variables."

Atiqullah (1964) investigated one-way analysis of covariance classifications with one concomitant variable x. Assuming that the e_{ij} are uncorrelated, with mean zero, equal variances σ^2, and constant kurtosis γ_2 ($\gamma_2 = 0$ for a normal distribution) the effect of non-normality in the y scores can be judged approximately by the sizes of factors $(1 + \gamma_2 C_1)$ and $(1 + \gamma_2 C_2)$. The approximate formulas for the C_i involve the degrees of freedom for treatments and error, moments of the distribution of the x's, and whether x is fixed or random. Using these formulas, the C_i can be calculated and the effect of non-normality judged for any particular problem.

Using this approach, Atiqullah finds that non-normality in the distribution of y's, the criterion variable, has little effect on the F-test for differences in treatment means when the distribution of x's, the concomitant variable, is normal.

Note that the distribution of the e_{ij}, whatever it is, is assumed to be the *same* in each treatment group. (The next section is concerned with situations where the distribution of the e_{ij} may not have the same variance in each group). Note, also, that Atiqullah has not investigated the effects of skewness in the distributions of the residuals. (See Elashoff (1968) for suggestions about dealing with skewed distributions.)

The normality assumption may be examined by plotting x scores and estimated residual scores e_{ij} on normal probability paper or by computing skewness and kurtosis coefficients. (See sec. 5.5 in Dixon and Massey (1957)). In some cases transformations of the data may be useful in producing more nearly normal distributions. (See Kruskal (1968) for an introduction to the subject of transformations.)

Homogeneity of Variances

Covariance analysis relies on the assumption that the variance of y scores for a given x is the same for each treatment group and independent of x (that is, σ^2_e is constant). There are two major cases: the variance of the y scores depends on x but for a fixed x is constant across treatments; the variance of the y scores within a treatment

is the same for each x but the variances are unequal across treatments.

Potthoff (1965) derives analysis of covariance tests for differences in two treatment means when $\sigma^2_{e_j}$ does not depend on x but differs for the two treatments. His results suggest that the effect of inequality of variances across treatments on the standard analysis of covariance depends on the size of the factor $C^* = \frac{n_1\sigma_{x_1}}{n_2\sigma_{x_2}}$ where n_j is the sample size in treatment group j, and σ_{x_j} denotes the standard deviation of the x scores in treatment j. Thus the effect of inequality of variances in the y scores across treatments is minimized when $n_1 = n_2$ and $\sigma_{x_1} = \sigma_{x_2}$ (that is the treatments have equal sample sizes and the variance of the x scores is the same in each).

Inequality of variance independent of x may be detected by comparing the variances of the estimated residuals across treatments. See Scheffé (1959), Box and Andersen (1955) or Levene (1960) and Glass (1966) for several tests for several tests for homogeneity of variances which are robust against non-normality.

If the variance of the y scores depends on x in a uniform way across treatments, transformation of the data may be desirable (see Acton, Chapter 8, Johnston, section 8-2, or Kruskal (1968).) When y is a measurement such as trials to criterion, counts, or proportions the variance or standard deviation is frequently proportional to its mean (and therefore x) and a transformation would be indicated. Inequality of variances of this nature may be detected by inspecting scatter plots of the estimated residuals against their corresponding x values. Rutemiller and Bowers (1968) present a large sample iterative procedure for estimation and testing of regression coefficients when the variance of y scores depends on x. The procedure can also be used to test the assumption of homoscedasticity.

General Conclusions

To sum up, the assumptions that assignment to treatments has been at random, that the covariate is independent of the treatments, and that there is no treatment-slope interaction are crucial to the underlying rationale for the use of covariance analysis. The assumptions of linearity, normality, and homogeneity of variances are necessary for

statistical simplicity and the validity of standard statistical tests; transformations of the data may be useful for making the data satisfy these assumptions and alternative covariance procedures which do not depend on these assumptions have been developed for certain cases. Generally, violation of the assumption of linearity, homogeneity or regressions, normality, or homegeneity of variances will be less serious if individuals have been assigned to treatments at random and the x variable has a normal distribution.

RELATED TOPICS

Within-Group Versus Between-Group Regression

In covariance analysis treatment means are "adjusted" using a pooled within-group regression slope. Sometimes, however, adjustments of a similar nature are made using a between-groups regression coefficient.

There are two important cases: 1) there is only one observation in each treatment group, 2) there is more than one observation per group. Let us denote the within-groups regression slope by β_w, the between-groups regression slope by β_t and their corresponding estimates by b_w and b_t.

If there is only one observation per group then clearly there is no within-group regression and b_w is unobtainable. A between-treatment regression slope b_t may then be used to "adjust" treatment scores as an attempt to untangle treatment effects and the effects of the regression of the criterion variable on the covariate. At best, this approach provides only a rough approximation to the desired results.

There are two major reasons for this. One—there is no way to separate treatment effects from "error"—the ordinary fluctuation in scores occurring between individuals in the absence of treatment effects. Thus there is no way to assess the "significance" of differences between treatments. Second, the between-treatments regression slope b_t, calculated by least squares methods, may be unduly affected by a few extremely effective or ineffective treatments and therefore its use in adjustments may produce misleading results.

To illustrate, suppose that there are four treatments with only one observation each and we observe the following outcome:

Treatment	1	2	3	4
x	−3	−1	1	3
y	0	3	3	6
y_{adj}	2.7	3.9	2.1	3.3

The estimate of the slope of y on x is $b_t = .9$.

How much of the apparent difference between treatments (2.1 to 3.9) is due to actual differences between treatments and how much is due to differences among the individuals which are not measured by the x variable? How big a range is 1.8?

To illustrate the second point, we suppose that in the absence of treatment differences, the criterion score y is exactly equal to the covariate x, $y = x$, $\beta = 1$. Suppose also, that of the four treatments in the experiment, three have the same effect and one is considerably more effective than the rest, so that "true" treatment means are $T_1 = -1$, $T_2 = -1$, $T_3 = -1$, $T_4 = 3$. Suppose also that the four individuals available have x scores, —3, —1, +1, +3. No matter which individual is assigned to treament four, adjustments using b_t will give treatment four the highest score, but all of the 24 possible individual to treatment assignments would result in a misleading assessment of the amount by which treatment four excells, and would give the false impression that differences existed among treatments 1, 2, 3. See Table 1. Only "on the average" across all possible assignments do the results indicate that teacher four is best and the rest are equivalent; even "on the average" however, the size of the difference is not correct. If random assignment is not used, then even this dubious assurance is unavailable. In short, in any one experiment b_t will be a biased estimate of β and yield biased estimates of the treatment effects.

TABLE 1

Treatment		1	2	3	4	
true score		—1	—1	—1	3	
Assignments	x	—3	—1	+1	+3	
1-6	y	—4	—2	0	6	$b_t = 1.6$
	y_{adj}	.8	—.4	—1.6	1.2	
7-12	x	3	1	—1	—3	
	y	2	0	—2	0	$b_t = .4$
	y_{adj}	.8	—.4	—1.6	1.2	
13-18	x	—3	—1	3	1	
	y	—4	—2	2	4	$b_t = 1.2$
	y_{adj}	—.4	—.8	—1.6	2.8	
19-24	x	—3	+1	+3	—1	
	y	—4	0	2	2	$b_t = .80$
	y_{adj}	—1.6	—.8	—.4	2.8	
Average over all 24	$E[y_{adj}]$	—2/3	—2/3	—2/3	2	$\bar{b}_t = 1.0$

If there is more than one observation in each group then both β_w and β_t can be estimated. Under the model for which covariance analysis is appropriate, the pooled within-group regression slope b_w is the appropriate one to use in making adjustments. If individuals have been assigned to treatments at random then under the null hypothesis of no treatment effects, the means of the treatment groups will fall around the same regression line as do the individuals within treatments and $\beta_w = \beta_t$ (see Lord (1962)). Then b_w and b_t are estimates of the same quantity, but b_w will generally have higher precision than b_t; the existence of any treatment effects will tend to bias the estimate b_t in a particular experiment (see Table 1).

A test of the hypothesis $\beta_w = \beta_t$ is given by Winer (1962), p. 588 and Dixon and Massey (1957), p. 219 assuming that the regression slopes within groups are the same and that the relationship between group means y_j, x_j is linear.

H Fairfield Smith (1957) gives a lengthy discussion of the problem of comparing β_w and β_t. He points out that the basic assumption of analysis of covariance states that treatment effects are independent of x, and therefore a significant difference between b_t and b_w in a particular problem is either an indication of a chance and irrelevant association of treatment effect and the covariate, x, as in the example shown in Table 1, or that in fact the covariance model is inadequate for the problem at hand since there is a meaningful relationship between x and treatment effects for this class of problem. He then goes on to say that if a test of $\beta_w = \beta_t$ has meaning, the standard test cited by Dixon and Massey and Winer is not appropriate.

Kahneman (1965) points out that when there is unreliability in the measurement of x, the observed regression among group means will tend to be steeper than the observed within-group regression ($b_t > b_w$).

If the covariance model is appropriate and its assumptions are satisfied, the within-group regression coefficient b_w should be used in making adjustments. If there is only one observation per group b_t could be used as a poor substitute. If the covariance model is not appropriate because there is some interaction between the treatments and the covariate x, a new model should be constructed which describes the situation and an analytic procedure derived for that model. Constructing a new model is not an easy problem. It is not at all clear that a between-groups regression adjustment would ever be the appropriate solution. Smith (1957) comments:

No work seems to have been done on constructing mathematical models for situations where a between treatment regression may be of interest and dif-

ferent from an internal regression. The task seems to be less simple than anticipated. A suitable form may need to be derived for each case individually with attention to the physical or biological circumstances.

Covariance Versus Matching or Blocking

When treatment groups are to be compared on the basis of a criterion variable, y, how should a covariate, x, be treated in the design of the study? Should the experimenter use completely randomized assignment and covariance analysis or rank the nt individuals on x and form n blocks of t individuals each and assign at random to treatment groups from each block? There is no simple answer to this question. It depends on the size of the correlation between x and y and the likelihood that all the assumptions necessary for covariance analysis will be satisfied. Of course, either procedure should be used with caution if assignments to treatments were not at random.

Cox (1957) and Myers (1966) give good discussions of the relative advantages of matching and covariance. Under the null hypothesis of no treatment effects and the assumption that x and y follow a bivariate normal distribution with correlation coefficient ρ, Cox (1957) compared the precision of blocking versus covariance. He concluded that if $\rho < 0.4$, blocking is preferable to covariance analysis, if $\rho > 0.6$ covariance analysis is somewhat better, and if $\rho > 0.8$ covariance analysis is appreciably better.

There are other considerations involved in the decision, however. If the number of treatments is large it may be difficult to construct homogeneous blocks. If the distribution of x has long tails, the end-blocks will consist of individuals with widely different values of x and blocking will be relatively less effective.

On the other hand, blocking is a reasonably efficient procedure for any smooth regression curve between y and x, not just for linear regression. The treatments by blocks interaction may be of interest; if it is expected to be significant, covariance analysis would be inappropriate because the assumption of homogeneity of regression would be incorrect.

In short, unless a strong linear relationship between x and y can be expected independent of treatments, blocking is probably a preferable technique.

It is sometimes suggested that covariance analysis is preferable to blocking when x is measured with error. In fact, however, unreliability in the measurement of x has similar effects on the two procedures—neither removes all the bias due to x.

REFERENCES

ACTON, FORMAN S. *Analysis of Straight-line Data.* New York: Wiley, 1959. 267 pp.

ATIQULLAH, M. "The Estimation of Residual Variance in Quadratically Balanced Least-Squares Problems and the Robustness of the F-test." *Biometrika* 49: 83-92; June 1962.

ATIQULLAH, M. "The Robustness of the Covariance Analysis of a One-way Classification." *Biometrika* 51: 365-372; December 1964.

BOX, GEORGE E. P. and ANDERSEN, S. L. "Permutation Theory in the Derivation of Robust Criteria and the Study of Departures from Assumption." *Journal of the Royal Statistical Society* 17: 1-26; June 1955.

COCHRAN, WILLIAM G. "Analysis of Covariance: Its Nature and Uses." *Biometrics* 13: 261-281; September 1957.

COCHRAN, WILLIAM G. "The Planning of Observational Studies of Human Populations." *Journal of the Royal Statistical Society* 128: 234-266; June 1965.

COCHRAN, WILLIAM G. "Errors of Measurement in Statistics." *Technometrics* 10: 637-666; November 1968.

COX, DAVID R. "The Use of a Concomitant Variable in Selecting an Experimental Design." *Biometrika* 44: 150-158; June 1957.

COX, DAVID R. *Planning of Experiments.* New York: Wiley, 1958. 308 pages.

DIXON, WILFRID J., and MASSEY, FRANK J. *Introduction to Statistical Analysis.* New York: McGraw-Hill, 1957. 488 pp.

ELASHOFF, ROBERT M. "Effects of Errors in Statistical Assumptions." *International Encyclopedia of the Social Sciences.* New York: Macmillan and Free Press, 1968. pp. 132-142.

EVANS, SELBY H., and ANASTASIO, ERNEST J. "Misuse of Analysis of Covariance when Treatment Effect and Covariate Are Confounded." *Psychological Bulletin* 69: 225-234; April 1968.

FINNEY, DAVID J. "Stratification, Balance, and Covariance." *Biometrics* 13: 373-386; September 1957.

GLASS, GENE V. "Testing Homogeneity of Variances." *American Educational Research Journal* 3: 187-190; May 1966.

HARRIS, CHESTER W. *Problems in Measuring Change.* Madison: University of Wisconsin Press, 1963. 259 pp.

HAYS, WILLIAM. *Statistics for Psychologists.* New York: Holt, Rinehart and Winston, 1963. 719 pp.

JOHNSTON, JOHN. *Econometric Methods.* New York: McGraw-Hill, 1963.

KAHNEMAN, DANIEL. "Control of Spurious Association and the Reliability of the Controlled Variable." *Psychological Bulletin* 64: 326-329; November 1965.

KENDALL, MAURICE G. *The Advanced Theory of Statistics.* Volume II. London: Charles Griffin, 1946. 676 pp.

KRUSKAL, J. B. "Statistical Analysis: II Transformations of Data." *International Encyclopedia of the Social Sciences.* New York: Macmillan and Free Press, 1968. pp. 182-192.

LEVENE, HOWARD. "Robust Tests for Equality of Variances." *Contributions to Probability and Statistics.* Edited by I. Olkin, et al. Stanford: Stanford University Press, 1960. Chapter 25, pp. 278-292.

LORD, FREDERIC M. "Large-sample Covariance Analysis When the Control Variable Is Fallible." *Journal of the American Statistical Association* 55: 307-321; 1960.

LORD, FREDERIC M. "Elementary Models for Measuring Change." In Harris, *Problems in Measuring Change.* Madison: University of Wisconsin Press, 1962.

LORD, FREDERIC M. "A Paradox in the Interpretation of Group Comparisons." *Psychological Bulletin* 68: 304-305; November 1967.

MYERS, JEROME L. *Fundamentals of Experimental Design.* Boston: Allyn and Bacon, 1966. 407 pp.

PECKHAM, PERCY D. *An Investigation of the Effects of Non-Homogeneity of Regression Slopes upon the F-Test of Analysis of Covariance.* Laboratory of Educational Research, Report No. 16. University of Colorado, Boulder, Colorado, 1968.

PORTER, ANDREW C. *The Effects of Using Fallible Variables in the Analysis of Covariance.* Doctoral dissertation, University of Wisconsin. University Microfilms No. 67-12, 147. Ann Arbor, Michigan, 1967.

POTTHOFF, RICHARD F. "Some Scheffé-type Tests for Some Behrens-Fisher Type Regression Problems." *Journal of the American Statistical Association* 60: 1163-1190; December 1965.

RUTEMILLER, HERBERT C. and BOWERS, DAVID A. "Estimation in a Heteroscedastic Regression Model." *Journal of the American Statistical Association* 63: 552-557; June 1968.

SCHEFFÉ, HENRY. *The Analysis of Variance.* New York: Wiley, 1959. 477 pp.

SMITH, H. FAIRFIELD. "Interpretation of Adjusted Treatment Means and Regression in Analysis of Covariance." *Biometrics* 13: 282-308; September 1957.

WINER, B. J. *Statistical Principles in Experimental Design.* New York: McGraw-Hill, 1962. 672 pp.

(Received October, 1968)
(Revised January, 1969)

AUTHOR

ELASHOFF, Janet D. *Address*: School of Education, Stanford University, Stanford, California 94305. *Title*: Assistant Professor. *Age*: 27. *Degrees*: B.S., Stanford University; Ph.D., Harvard University. *Specialization*: Statistics, data analysis, design of experiments.

References

The Estimation of Residual Variance in Quadratically Balanced Least-Squares Problems and the Robustness of the F-Test

M. Atiqullah

Biometrika, Vol. 49, No. 1/2. (Jun., 1962), pp. 83-91.

Stable URL:

http://links.jstor.org/sici?sici=0006-3444%28196206%2949%3A1%2F2%3C83%3ATEORVI%3E2.0.CO%3B2-W

The Robustness of the Covariance Analysis of a One-Way Classification

M. Atiqullah

Biometrika, Vol. 51, No. 3/4. (Dec., 1964), pp. 365-372.

Stable URL:

http://links.jstor.org/sici?sici=0006-3444%28196412%2951%3A3%2F4%3C365%3ATROTCA%3E2.0.CO%3B2-R

Permutation Theory in the Derivation of Robust Criteria and the Study of Departures from Assumption

G. E. P. Box; S. L. Andersen

Journal of the Royal Statistical Society. Series B (Methodological), Vol. 17, No. 1. (1955), pp. 1-34.

Stable URL:

http://links.jstor.org/sici?sici=0035-9246%281955%2917%3A1%3C1%3APTITDO%3E2.0.CO%3B2-9

LINKED CITATIONS

Analysis of Covariance: Its Nature and Uses
William G. Cochran
Biometrics, Vol. 13, No. 3, Special Issue on the Analysis of Covariance. (Sep., 1957), pp. 261-281.
Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28195709%2913%3A3%3C261%3AAOCINA%3E2.0.CO%3B2-K

The Planning of Observational Studies of Human Populations
W. G. Cochran; S. Paul Chambers
Journal of the Royal Statistical Society. Series A (General), Vol. 128, No. 2. (1965), pp. 234-266.
Stable URL:
http://links.jstor.org/sici?sici=0035-9238%281965%29128%3A2%3C234%3ATPOOSO%3E2.0.CO%3B2-F

Errors of Measurement in Statistics
W. G. Cochran
Technometrics, Vol. 10, No. 4. (Nov., 1968), pp. 637-666.
Stable URL:
http://links.jstor.org/sici?sici=0040-1706%28196811%2910%3A4%3C637%3AEOMIS%3E2.0.CO%3B2-C

The Use of a Concomitant Variable in Selecting an Experimental Design
D. R. Cox
Biometrika, Vol. 44, No. 1/2. (Jun., 1957), pp. 150-158.
Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28195706%2944%3A1%2F2%3C150%3ATUOACV%3E2.0.CO%3B2-7

Stratification, Balance, and Covariance
D. J. Finney
Biometrics, Vol. 13, No. 3, Special Issue on the Analysis of Covariance. (Sep., 1957), pp. 373-386.
Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28195709%2913%3A3%3C373%3ASBAC%3E2.0.CO%3B2-Y

Testing Homogeneity of Variances
Gene V. Glass
American Educational Research Journal, Vol. 3, No. 3. (May, 1966), pp. 187-190.
Stable URL:
http://links.jstor.org/sici?sici=0002-8312%28196605%293%3A3%3C187%3ATHOV%3E2.0.CO%3B2-P

LINKED CITATIONS

Large-Sample Covariance Analysis When the Control Variable is Fallible
Frederic M. Lord
Journal of the American Statistical Association, Vol. 55, No. 290. (Jun., 1960), pp. 307-321.
Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28196006%2955%3A290%3C307%3ALCAWTC%3E2.0.CO%3B2-E

Some Scheffe-Type Tests for Some Behrens-Fisher-Type Regression Problems
Richard F. Potthoff
Journal of the American Statistical Association, Vol. 60, No. 312. (Dec., 1965), pp. 1163-1190.
Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28196512%2960%3A312%3C1163%3ASSTFSB%3E2.0.CO%3B2-Y

Estimation in a Heteroscedastic Regression Model
Herbert C. Rutemiller; David A. Bowers
Journal of the American Statistical Association, Vol. 63, No. 322. (Jun., 1968), pp. 552-557.
Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28196806%2963%3A322%3C552%3AEIAHRM%3E2.0.CO%3B2-1

Interpretation of Adjusted Treatment Means and Regressions in Analysis of Covariance
H. Fairfield Smith
Biometrics, Vol. 13, No. 3, Special Issue on the Analysis of Covariance. (Sep., 1957), pp. 282-308.
Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28195709%2913%3A3%3C282%3AIOATMA%3E2.0.CO%3B2-C