THE ANALYSIS OF COVARIANCE: STATISTICAL AND METHODOLOGICAL ISSUES

Jeroen G.W. Raaijmakers
Department of Psychology
University of Nijmegen
Nijmegen, The Netherlands

Jo P.M. Pieters
Rijks Psychologische Dienst
The Hague, The Netherlands

Abstract

This article discusses a number of problems and confusions concerning the proper use and interpretation of the analysis of covariance. It is argued that in discussions of a technique for data analysis statistical and methodological issues should be kept separate. It is shown that a number of socalled assumptions of ANCOVA that have been mentioned in the literature, are not in fact necessary from a statistical point of view. Part of the confusion concerning the use of ANCOVA seems to be due to the fact that ANCOVA is often misinterpreted as a kind of ANOVA on scores corrected for their relation to the covariate. The logic of ANCOVA gives no support for its use as a substitute for experimental control.

A detailed discussion is given of the effects of measurement error in the covariate on the results of ANCOVA. We criticize the position taken by Overall and Woodward (1977a, 1977b) who defended the use of ANCOVA in certain situations involving fallible covariates. It is shown that in such cases a conventional ANCOVA nearly always leads to a biased test for the hypothesis of no group differences. In addition, an extensive discussion is given of alternative approaches based on the formulation of ANCOVA as a functional or structural relationship model. It is shown how the standard null hypothesis might be tested in such an approach. On the basis of simulation results it is concluded that such a testing procedure is preferable to the conventional *F*-test of the ANCOVA null hypothesis.

Introduction

The analysis of covariance (ANCOVA) was developed as a means for increasing the precision of statistical evaluation of experimental effects and as a method for removing bias due to differences between experimental groups. Generally speaking, there are two procedures for reducing the error variance and thus increasing the power of the statistical test. First, there are a number of experimental designs that lead to a reduction in error variance (compared to a simple randomized groups design) by direct control of relevant factors. The most important of these, at least for the present discussion, is the randomized block design. In this design the error variance is reduced by grouping experimental units into blocks that are homogeneous with respect to the concomitant variable one wishes to control for. Second, one could try to control statistically for the effects of the concomitant variable. In such a procedure the effects of a concomitant variable are partialled out by assuming a particular functional relationship between the concomitant variable and the dependent variable. One of the most widely used procedures for achieving this kind of statistical control is the analysis of covariance. The usual ANCOVA model is based on the assumption of a linear relationship between the concomitant variable or covariate and the dependent variable. If the relation between these variables happens to be nonlinear, stratification of experimental units (i.e. a randomized block design) leads to a greater reduction in the experimental error than the linear adjustment used in ANCOVA. Stratification is essentially a function-free regression scheme.

ANCOVA has also been proposed as a method to adjust for initial differences between experimental groups. The need for such an adjustment is most prominent in situations where the experimental units cannot be assigned at random to the experimental conditions. This situation occurs frequently in quasi-experimental and observational type research (see Cook & Campbell, 1979). It should be clear that in such an application ANCOVA adjusts only for differences between experimental groups that can be attributed to the particular concomitant variable which is used in the experimental design. Firm conclusions may still be difficult due to remaining pre-experimental differences.

The rationale underlying the ANCOVA procedure may be illustrated by the ANCOVA-model for a single factor experiment. In this case one assumes that the following model is appropriate for the dependent variable y_{ij}

$$y_{ij} = \mu + \alpha_j + \beta(x_{ij}-\bar{x}) + \varepsilon_{i(j)}, \quad i=1,\dots,r;\ j=1,\dots,c \qquad (1)$$

where μ is the overall mean, α_j is the effect due to treatment condition j, β is the parameter of the regression of y on x (the concomitant variable or covariate), and $\varepsilon_{i(j)}$ refers to the error component. For the moment we will assume that the x_{ij}'s are fixed constants (i.e. no distribution is assumed for x) and that the expectation of $\varepsilon_{i(j)}$ is zero.

The usual F-test for the null hypothesis $\alpha_j = 0$ (all j) is based on a comparison of the variance explained by the linear model of Equation 1 with the variance explained by the reduced model

$$y_{ij} = \mu + \beta(x_{ij}-\bar{x}) + \varepsilon_{i(j)}, \quad i=1,\dots,r;\ j=1,\dots c.$$

In order for the F-statistic that is used to test the null hypothesis to have a variance ratio or F-distribution, the following assumptions have to be fullfilled.

1. the $\varepsilon_{i(j)}$ are independent random variables each having a normal distribution with mean zero and variance σ^2.
2. as is implied by the model of Equation 1, all groups must have the same regression coefficient, i.e. $\beta_1 = \beta_2 = \dots = \beta_c = \beta$.

In addition to these two assumptions, various other conditions have been mentioned in the literature as being necessary for a valid application of ANCOVA.

For example Winer (1971, p.764) mentions the requirement that not only the within-groups regression coefficients should be equal, but that also the common within-groups regression coefficient should be equal to the between-groups regression coefficient and to the total regression coefficient. According to Evans and Anastasio (1968) a necessary condition for a valid application of ANCOVA is that treatment effect and covariate are uncorrelated. This condition implies that

$$\Sigma \alpha_j \bar{X}_j = 0 ,$$

which in general implies that all covariate means are equal.

Yet another problem with the application of ANCOVA concerns the effect of measurement error in the covariate. It is frequently mentioned that measurement error in the covariate leads to a biased estimate for β and thus to an incorrect estimate of the treatment effect. On the other hand, Overall and Woodward (1977a) argue that ANCOVA leads to valid results in a number of cases even if the covariate is fallible. In addition, Overall and Woodward (1977b) suggest that measurement error in the covariate does not necessarily invalidate the results of ANCOVA even in the case when subjects are non-randomly assigned to treatment conditions, provided the assignment is based on the *observed* value of the covariate.

It is also frequently assumed that ANCOVA is equivalent to an ordinary analysis of variance (ANOVA) on the residuals from the overall regression line between the dependent variable and the covariate. For example, Kirk (1968, p.469) rewrites the ANCOVA-model as

$$y_{ij(adj)} = y_{ij} - \beta(x_{ij}-\bar{x})$$

$$= \mu + \alpha_j + \varepsilon_{i(j)} .$$

This could easily be (mis)interpreted as implying that ANCOVA is equivalent to ANOVA on the adjusted scores. Some textbooks (e.g. Kerlinger & Pedhazur, 1973, p. 267; Pedhazur, 1982, p. 497) in fact do give such an interpretation of ANCOVA.

In this paper we will review a number of these problems. First, we will show that several misconceptions regarding ANCOVA are due to a confounding of statistical and methodological issues related to the assumptions of ANCOVA. A major purpose will be to investigate the effects of measurement error in the covariate on the validity of the results of ANCOVA. We will not only consider the effects on the estimation of the parameters of the ANCOVA-model (α, β), but we will also consider the effects of a fallible covariate on the usual F-test. In addition, we will discuss some alternative procedures for the analysis of data when the covariate is measured with error. Finally, we will briefly discuss some methodological issues regarding the application of ANCOVA in experiments where subjects are non-randomly assigned to treatment conditions.

I. Common misconceptions regarding assumptions of ANCOVA

Before discussing the abovementioned 'assumptions' of ANCOVA that can be found in the psychological literature, we would like to stress the following points that are of general importance in any discussion on the validity of a specific model for the analysis of data. In order to fully understand the strengths and weaknesses of a particular method, one should clearly distinguish between the following points in the discussion.

First, one should clearly specify the mathematical model which is assumed to be satisfied by the data. Second, one should give a specification of the statistical assumptions concerning the model parameters such as distributional properties and covariances between parameters. Third, a distinction should be made between properties of estimators for the parameters (e.g. expectation) and the properties of statistics used to test various hypotheses with respect to the model parameters (e.g. significance tests). Last but not least, one should clearly distinguish between statistical issues concerning the validity of the application of a specific model and methodological issues concerning the application of a specific model. A major issue in this paper will be to show that several misconceptions concerning the validity of ANCOVA in certain situations are due to a confounding of arguments that relate to different aspects of the use of statistical methods. For example, it may well be the case that the application of a specific statistical method is completely sound from a statistical point of view but at the same time completely absurd from a methodological point of view.

Equality of between- and within-groups regression coefficients

It is frequently mentioned that a valid use of ANCOVA requires that the population values of the between- and pooled within-groups regression coefficients should be equal (e.g. Evans & Anastasio, 1968; Winer, 1971, p.764).

The ANCOVA model used by Evans and Anastasio (1968) is

$$y_{ij} = \mu + \alpha_j + \beta x_{ij} + \varepsilon_{i(j)}, \qquad i=1,\dots,r;\ j=1,\dots c$$

which is essentially the same as the model specified in Equation 1. Given this model there is only one parameter that measures the relation between the dependent variable and the covariate, i.e. parameter β. Hence, any reference to different regression parameters or population values of these parameters could on a priori grounds already be considered absurd. The only possible distinction that could be made is the distinction between various estimators of β and their properties. In the usual textbook approach to ANCOVA one can find three different estimators for the regression parameter β. These estimators are (1) the pooled within-groups regression coefficient b_W, (2) the between-groups regression coefficient b_B, and, (3) the total regression coefficient b_T. These coefficients are usually defined as

$$b_W = E_{xy}/E_{xx}, \qquad b_B = T_{xy}/T_{xx}, \qquad b_T = S_{xy}/S_{xx},$$

where $E_{..}$ denotes the within-groups sum of squares or crossproducts, $T_{..}$ denotes the between-groups sum of squares or crossproducts, and, $S_{..}$ denotes the total sum of squares or crossproducts (see e.g. Winer, 1971). Using the assumption that the covariance between the error parameters and the covariate is zero, it can be shown that the expectation of the pooled within-groups regression coefficient b_W is given by

$$E(b_W) = \beta ,$$

that the expectation of the between-groups regression coefficient b_B is given by

$$E(b_B) = \beta + r\Sigma\alpha_j(x_j-\bar{x})/T_{xx},$$

and that the expectation of the total regression coefficient b_T is given by

$$E(b_T) = \beta + r\Sigma\alpha_j(x_j-\bar{x})/S_{xx}.$$

These expressions imply that only the pooled within-groups regression coefficient is an unbiased estimator for the regression parameter β. The other estimators are only unbiased estimators for β under the null hypothesis $\alpha_j = 0$ (all j).

The test statistic that is commonly used to test this hypothesis is composed of the adjusted between- and within-groups sums of squares defined respectively as

$$T^*_{yy} = T_{yy} - (S^2_{xy}/S_{xx} - E^2_{xy}/E_{xx}) ,$$

and

$$E^*_{yy} = E_{yy} - E^2_{xy}/E_{xx} .$$

It has been shown by various authors (see e.g. Searle, 1971) that if the $\varepsilon_{i(j)}$'s are independent normally distributed random variables with mean zero and common variance σ^2, the random variable T^*_{yy}/σ^2 has a non-central chi-square distribution with *(c-1)* degrees of freedom and the random variable E^{*}_{yy}/σ^2 has a central chi-square distribution with *(N-c-1)* degrees of freedom where N stands for the total number of observations. Using the fact that the random variables T^{*}_{yy}/σ^2 and E^{*}_{yy}/σ^2 are independent, it can be shown that the statistic Q defined as the ratio of these random variables divided by their respective degrees of freedom has a non-central variance ratio or F-distribution with *(c-1)* and *(N-c-1)* degrees of freedom. Given that $\alpha_j = 0$ (all j) the statistic Q has a central F-distribution with *(c-1)* and *(N-c-1)* degrees of freedom. Hence, the statistic Q can be used as a test for the null hypothesis $\alpha_j = 0$.

The conclusion of this discussion should be clear. The requirement of equality of the various regression coefficients mentioned by a number of authors (e.g. Evans & Anastasio, 1968; Winer, 1971) does not prove to be necessary for a valid application of ANCOVA. The misconception is probably due to a confounding of arguments related to the properties of estimators for model parameters and the basic statistical assumptions related to the validity of estimation and testing procedures for the evaluation of treatment effects. For example, it has been shown that the regression coefficient b_T is not an unbiased estimator for the regression parameter β except under the null hypothesis of no treatment effect. The fact that b_T is a biased estimator does not imply that the model specified in Equation 1 is violated.

ANCOVA is ANOVA on residuals of the regression line

Many researchers, especially in the social sciences, think of ANCOVA as being similar to ANOVA performed on the residuals about the overall regression line which describes the relationship between the covariate and the dependent variable. This idea has its origin in the terminology used in almost all textbooks that discuss ANCOVA. Notably, the use of concepts like 'adjusted sums of squares' strengthens such ideas. Researchers which hold this opinion on ANCOVA are however in good company. The famous R.A. Fisher originally introduced

ANCOVA as a standard ANOVA on the residual scores $y_{ij}-bx_{ij}$ (see e.g. Cox & McCullagh, 1982). However, it was readily recognized that the ordinary *F*-test of ANOVA on residual scores was only an approximate test for the evaluation of treatment effects. Influenced by papers of Wishart (1934) and Wilsdon (1934), R.A. Fisher therefore designed a new procedure for ANCOVA that gave an exact *F*-test for treatment effects. This procedure was discussed in the preceding section.

Insert Figure 1 about here

The following example clearly shows the differences between these approaches. Using the data in Figure 1, we performed an ANOVA on the residual scores as well as the appropriate ANCOVA. These data can be thought of as being generated by Equation 1 with $\varepsilon_{i(j)}=0$. The results of these analyses, shown in Table 1, clearly demonstrate the difference between the two approaches. The ANCOVA analysis correctly estimates the error variance (MS_{within}) as nil, while the ANOVA on the residuals leads to quite a different and obviously incorrect estimate. Hence ANCOVA and ANOVA on the residuals are not the same and only ANCOVA leads to a valid test for the null hypothesis. In a similar way (for example using the data in Figure 2), it may be shown that use of the common within-groups regression coefficient as an estimator for the slope of the regression line does not alter this conclusion.

Insert Table 1 about here

Independence of treatment and covariate

Several authors, e.g. Evans and Anastasio (1968) and Winer (1971, p.754), state that a necessary condition for a valid application of ANCOVA is that treatment effect and covariate are uncorrelated. This implies that

$$\Sigma \alpha_j \bar{x}_j = 0$$

which in general implies that all $\bar{x}_j$'s are equal. If this requirement would indeed be necessary, quite a few applications of ANCOVA would prove invalid (see also Sprott, 1970), since in many cases there are substantial differences between groups on the covariate.

However, as mentioned in the introductory section, the ANCOVA-model makes no distributional assumptions with respect to the covariate, i.e. all x_{ij} are assumed fixed constants. In particular, the model does not assume that all x_{ij}'s are (independent) samples from the same distribution. As long as,

$$y_{ij} = \mu + \alpha_j + \beta x_{ij} + \varepsilon_{i(j)}, \quad i=1,\ldots,r;\ j=1,\ldots c$$

is the appropriate model, Evans and Anastasio's requirement is not a necessary assumption of the ANCOVA model nor is it mandatory for the validity of the *F*-test used to evaluate treatment effects (see also Sprott, 1970). From a purely statistical point of view, a valid application of ANCOVA does not require independence of covariate and treatment. In such a situation, ANCOVA does precisely what it is meant to do: adjust for the correlation between covariate and dependent variable. Thus if the ANCOVA model holds, it is the *best* procedure for evaluating treatment effects.

However, if the covariate is actually affected by the treatment or if covariate and treatment are correlated due to initial differences between the experimental groups, ANCOVA adjusts not only for the variance due to the term βx_{ij}. It may also remove part of the treatment effect. Statistically this is completely correct since ANCOVA is to remove *all* differences in the dependent variable that can be explained by differences in the covariate. Potential removal of part of the treatment effect is thus not a statistical issue but a methodological problem. The application of ANCOVA may not be valid not because of a violation of necessary assumptions but because it gives an answer to a question the researcher was not interested in! In such a case, ANCOVA leads to the conclusion that the observed differences can be (partially) explained by differences on the covariate. This does not imply however, that the observed differences *have* to be explained by the covariate. It may still be true that there is no *causal* relationship between the covariate and the dependent variable.

Insert Figure 2 about here

This can be clearly demonstrated by the following example. We generated artificial data such that the within-groups regression coefficient was equal to zero while the groups differed considerably on the covariate. In addition, the data were generated in such a way that the treatment had a large effect on the dependent variable (see Figure 2).

Insert Table 2 about here

The results of ANCOVA are shown in Table 2 together with the results of an ordinary ANOVA on the dependent variable ignoring the covariate. Indeed, ANCOVA in this situation removes not just part but allmost all of the treatment effect. Thus, ANCOVA would lead to the conclusion of no treatment effect. As we have argued before, this is not due to a violation of necessary assumptions of ANCOVA in the present situation, but it is due to the fact that we chose the wrong type of analysis. The question we wanted to be answered was

Are there treatment effects ? Instead we got an answer to the question: *Are there treatment effects that cannot be explained by differences on the covariate ?* If one is not interested in this question, one should not blame ANCOVA for giving the wrong answer, but instead use a more correct procedure that may give an answer to the appropriate question. Indeed, in such a situation ANCOVA should be used with great care. ANCOVA should only be used when it seems reasonable to assume a causal relationship between the covariate and the dependent variable. Once more we would like to stress that the decision is not based upon statistical but on methodological arguments.

The validity of ANCOVA in the case where treatment effect and covariate are correlated, was also discussed in some detail by Sprott (1970). He showed that Evans and Anastasio's requirement is not necessary for a valid application of ANCOVA. However, in discussing the difference between an unconditional ANOVA and ANCOVA, Sprott stated that the correct requirement for a valid use of ANCOVA is that the treatment has no effect on the covariate. In the remaining part of this section we will show that this statement is in its generality incorrect. Sprott uses the following ANCOVA model:

$$y_{ij} = \mu + \alpha_j + \beta x_{ij} + \varepsilon_{i(j)} \qquad (2)$$

where

$$x_{ij} = \mu_x + \theta_j + u_{i(j)} \quad .$$

In this model μ_x is the overall mean of the covariate, θ_j is the true effect of the treatment on the covariate ($\Sigma\theta_j = 0$), and the $u_{i(j)}$ are independent normally distributed random variables with mean zero and common variance σ_u^2.

It can be shown by simple algebra that under the null hypothesis of no treatment effect the total regression coefficient is an unbiased estimator for parameter β. As in the case of Equation 1, the pooled within-groups regression coefficient is always an unbiased estimator for parameter β. It is also easily shown that the ANCOVA estimator of the treatment effect, a_j, is an unbiased estimator of the true treatment effect since

$$E(a_j) = E((\bar{y}_j-\bar{y})-b_W(\bar{x}_j-\bar{x}))$$

$$= \alpha_j \quad .$$

In addition, the expectation of the adjusted within-groups sum of squares is given by

$$E(E^*_{yy}) = (N-c-1)\sigma^2$$

while the expectation of the adjusted between-groups sum of squares is given by

$$E(T^*_{yy}) = r\Sigma\ \alpha_j^2 - (r\Sigma\alpha_j\theta_j)^2/(r\Sigma\ \theta_j^2+(c-1)\sigma_u^2) + (c-1)\sigma^2 \quad .$$

As before the usual statistic Q has a non-central F-distribution with $(c-1)$ and $(N-c-1)$ degrees of freedom. Under the null hypothesis of no treatment effect, the statistic Q has again a central F-distribution with $(c-1)$ and $(N-c-1)$ degrees of freedom. This should come as no surprise since Equation 2 is equivalent to the conventional ANCOVA model (Equation 1).

It is clear, then, that the requirement $\theta_j = 0$ stated by Sprott (1970) is not a necessary condition for a valid application of ANCOVA. The problems caused by an effect of treatment on covariate are once again not of a statistical but of a methodological nature. ANCOVA provides the correct answer to the question it is supposed to answer. The technique should not be blamed if it is used to get answers to questions of an entirely different nature.

However real statistical problems arise in a slightly different situation. Suppose the data satisfy the model

$$y_{ij} = \mu + \alpha_j + \beta X_{ij} + \varepsilon_{i(j)}$$

where

$$x_{ij} = \mu_x + \theta_j + X_{ij} \quad .$$

In this model X_{ij} is the (fixed) 'true' score of the observed covariate x_{ij} (i.e. the covariate score unaffected by the treatment). Note the difference between this model and the previous one. In this case the linear relationship is not between the dependent variable and the *observed* value of the covariate, but between the dependent variable and the 'true' score of the covariate. It can be shown that in this situation the statistic Q is not a valid statistic for testing the null hypothesis $\alpha_j = 0$ (for all j) but instead it tests for the composite hypothesis $\alpha_j + \beta\theta_j = 0$ (for all j).

The general conclusion should be, then, that the independence of treatment and covariate is not a necessary requirement for a valid application of ANCOVA as long as the dependent variable is linearly related to the observed values of the covariate, that is, as long as Equation 1 applies. The problems mentioned in the literature concerning the application of ANCOVA in such a situation are based on misconceptions regarding the true nature of ANCOVA. These problems do not relate to statistical issues but to the possible inappropriateness of ANCOVA from a methodological point of view. We have shown that the confusion arises from a confounding of arguments that relate to different issues concerning the application of ANCOVA. Statistical analysis techniques that are valid from a statistical point of view (i.e. provide correct tests for the null hypotheses) may in a given situation be inappropriate from a methodological point of view.

II. Measurement error in the covariate

One of the most widespread confusions regarding ANCOVA concerns the issue whether a valid application of ANCOVA requires the covariate to be measured without error. Overall and Woodward (1977a, 1977b) are frequently cited as having shown that this requirement is not always necessary for a proper use of

ANCOVA. In this section we will show that their results are misleading because they apply only in a special case and, more importantly, we will show that the general conclusion reached by Overall and Woodward regarding the applicability of ANCOVA in case of a fallible covariate is wrong.

To investigate the implications of a fallible covariate, we will use the following model:

$$y_{ij} = \mu + \alpha_j + \beta T_{ij} + \varepsilon_{i(j)} \qquad (3)$$

and

$$x_{ij} = \mu_x + T_{ij} + \delta_{i(j)} \qquad (4)$$

where T_{ij} ($\Sigma T_{ij} = 0$) denotes the true score of the observed covariate x_{ij} and the $\delta_{i(j)}$'s are independent, normally distributed error variables with mean zero and common variance σ^2_δ. We will further assume that

$$cov(\varepsilon,\delta) = cov(T,\delta) = cov(T,\varepsilon) = 0 \quad .$$

Then it can be shown that the pooled within-groups regression coefficient can be written as

$$E(b_W) = \rho_{xx}\beta$$

where

$$\rho_{xx} = A \;/\; (A+(N-c)\sigma^2_\delta) \quad . \qquad (5)$$

In this formula, which gives the reliability of the covariate, A denotes the pooled within-groups sum of squares of the true score of the covariate, i.e.

$$A = \Sigma\Sigma\,(T_{ij} - \overline{T}_j)^2 \quad .$$

It is obvious that in this case b_W is a biased estimator of the regression parameter β. Overall and Woodward investigated the conditions in which the estimator of the treatment effect will be unbiased despite the presence of measurement error in the covariate. It is easily shown that the usual estimator for the treatment effect α_j, a_j,

$$a_j = (\bar{y}_j - \bar{y}) - b_W(\bar{x}_j - \bar{x})$$

has the following expectation:

$$E(a_j) = \alpha_j + \beta(1-\rho_{xx})\bar{T}_j \quad ,$$

where ρ_{xx} is given by Equation 5. It is clear then, that in general the estimator for the treatment effect will be biased. The estimator will only be unbiased if $\bar{T}_j$ is equal to zero which is one of the cases studied by Overall and Woodward (1977a). We will show however that even in this case application of the standard ANCOVA leads to misleading and incorrect results. Most users of ANCOVA are not so much interested in estimating treatment effects but in demonstrating treatment effects through the standard F-test. The question whether treatment effects are correctly estimated, is less relevant than suggested by Overall and Woodward. They circumvented this problem by a cursory reference to the robustness of the F-test. That such a naive confidence in the robustness of the F-test may be quite inappropriate in the case of a fallible covariate, will be shown in this section.

Using the assumption that ε, δ and T are independent, it can be shown that the expectation of the adjusted pooled within-groups sum of squares E^{*}_{yy} (given the assumption that $\bar{T}_j$=0) is equal to

$$E(E^{*}_{yy}) = \beta^2(1-\rho_{xx})\ \Sigma T^2_{ij} + (N-c-1)\sigma^2$$

while the expectation of the adjusted treatment sum of squares T^{*}_{yy} is equal to

$$E(T^{*}_{yy}) = r\ \Sigma\alpha^2_j + (c-1)\sigma^2 \quad .$$

Given the assumption that the $\varepsilon_{i(j)}$ and the $\delta_{i(j)}$ are independent normally distributed variables, it can be shown that the random variable E^{*}_{yy}/σ^2 has a non-central chi-square distribution with $(N-c-1)$ degrees of freedom, while the random variable T^{*}_{yy}/σ^2 has a non-central chi-square distribution with $(c-1)$ degrees of freedom. Hence, the statistic Q defined as

$$Q = (T^{*}_{yy}/(c-1))/(E^{*}_{yy}/(N-c-1))$$

has a doubly non-central F-distribution with $(c-1)$ and $(N-c-1)$ degrees of freedom (Johnson & Kotz, 1970). Under the null hypothesis, $\alpha_j = 0$ (for all j), the statistic Q has a non-central F-distribution with $(c-1)$ and $(N-c-1)$ degrees of freedom. A necessary requirement for the validity of the usual F-test is that the statistic Q has a *central* F-distribution under the null hypothesis. Otherwise the test will lead to biased results. It is clear, then, that in the case of a fallible covariate the usual F-test of ANCOVA is biased, even in the case $\bar{T}_j = 0$, considered by Overall and Woodward. The extent of this bias depends on the size of the non-centrality parameter λ,

$$\lambda = (\beta^2(1-\rho_{xx})\Sigma\ T^2_{ij})/\sigma^2 \quad .$$

Insert Table 3 about here

The extent of this bias was determined using Monte Carlo simulations. A typical example of the results is shown in the lefthand part of Table 3. These results are based on 5000 simulations of a two-group design with one covariate, using Equations 3 and 4 with $\alpha_1=\alpha_2=0$, $\bar{T}_1=\bar{T}_2=0$, $\mu=50$, $\mu_x=0$, $\beta=5$, $\rho_{xx}=.5$, $\sigma^2=\sigma^2_\delta=10$ and $r=10$. A frequency distribution of the obtained values of

Q was formed using equally probable class intervals, i.e. the class limits were determined from the percentile points of the appropriate F-distribution. Thus, if the usual F-test would be valid, we would expect 500 observations in each class. The results shown in Table 3 indicate that the departure from the expected values is quite substantial, showing that in the case $\bar{T}_j = 0$, the use of the conventional F-test is incorrect.

While in the case $\bar{T}_j=0$ the F-test is biased in the direction of acceptance of the null hypothesis, it can be shown that as the $\bar{T}_j$ depart from zero, the F-test will become increasingly biased in the opposite direction. As in the preceding case, the statistic Q has, under the null hypothesis, a non-central F-distribution with $(c-1)$ and $(N-c-1)$ degrees of freedom, with a slightly more complex non-centrality parameter. Monte Carlo simulation using the same parameters as before except $\bar{T}_1=-10$ and $\bar{T}_2=10$, resulted in the frequency distribution for the statistic Q shown in the righthand part of Table 3. It is clear then that in this case the F-test is biased toward rejection of the null hypothesis. Hence, the conclusion of Overall and Woodward (1977b, p.591)

> *Specifically, where subjects have been assigned at random, ANCOVA can appropriately be used whether or not the concomitant variable is measured with error. This case would seem to be generally accepted; however, unqualified statements about the inadequacy of ANCOVA corrections when covariate measurements are fallible have appeared in numerous articles in the psychological literature.*

must be rejected as just another unqualified statement. Instead the conclusion should be that in the case of a fallible covariate the ANCOVA procedure may lead to biased and incorrect results, both with respect to the estimation of treatment effects as well as with respect to the F-test for detecting significant treatment effects. Depending upon the size of the departure of the $\bar{T}_j$ from zero, the actual Type-I error rate will be negatively or positively biased compared to the nominal Type-I error rate of the F-test.

Nonrandom assignment and ANCOVA

Usually, ANCOVA is used in a situation where experimental units have been randomly assigned to treatment conditions. However, in quasi-experimental or observational-type research one often cannot use this type of randomization, for pratical and/or ethical reasons. The question arises then whether ANCOVA can be considered an appropriate method to adjust for initial group differences due to nonrandom assignment. Before addressing the more general models discussed by Overall and Woodward (1977a, 1977b), we will discuss two cases that are quite typical in quasi-experimental designs. First, we will discuss the socalled regression discontinuity design (see Cook & Campbell, 1979) in which subjects are assigned to treatment conditions using a fixed criterion based on the observed covariate scores. Second, we will discuss procedures that use a probabilistic mechanism for the assignment of subjects to treatment conditions based on the true score on the covariate.

The regression-discontinuity design is described by Cook and Campbell (1979) as one of the nonequivalent control group designs useful in quasi-experimental research. It is appropriate whenever the decision whether the subject should receive some sort of treatment depends on the score on a pretest. The logic behind this design is to classify subjects according to their scores on the pretest. Subjects which score above (c.q. below) a specified cutting point are assigned to the treatment group, while subjects which score below (c.q. above) the cutting point are assigned to the control group. If the treatment does have an effect, one should find a discontinuity at the cutting point in the regression line computed between pre- and posttest scores. It can be shown that ANCOVA is often appropriate for the testing of treatment effects in this design, even in those cases where the covariate is fallible.

As we have emphasized above, ANCOVA will lead to statistically correct results as long as the data satisfy the model in Equation 1. Thus, the important question is whether there is a linear relation between the dependent variable and the *observed* covariate scores. If such a relation exists (under the null hypothesis), there will be nothing *statistically* wrong with the application of ANCOVA. The question that remains to be answered is whether ANCOVA will lead to correct results when the linear relation does not hold between the dependent variable and the observed values of the covariate but between the dependent variable and the 'true' scores of the covariate. Suppose the covariate is measured with error and suppose furthermore that (under the null hypothesis) the dependent variable and the observed covariate scores have a bivariate normal distribution. Then it follows from the properties of the bivariate normal distribution (see e.g. Mood, Graybill & Boes, 1974) that the conditional scores on the dependent variable satisfy a linear regression equation, even though the theoretical relationship is between the dependent variable and the 'true' scores of the covariate. In this case, then, the ANCOVA model holds even though the covariate is measured with error.

The application of ANCOVA in case of nonrandom assignment has also been discussed by Overall and Woodward (1977a, 1977b). They primarily discuss the case where the dependent variable and the covariate have a bivariate normal distribution under the null hypothesis. However, they generalize the applicability of ANCOVA to all situations involving nonrandom assignment of subjects to treatment groups. According to Overall and Woodward, ANCOVA leads to valid results whenever there is a fixed assignment rule that specifies the probability of assignment to a specific treatment group, given the observed covariate score. It should be evident, however, that this statement cannot be generally correct. For example, suppose subjects are assigned with equal probability to all treatment groups. Although in this case a fixed assignment rule is used, it is in fact equivalent to random assignment of subjects to treatment groups. Hence, the results discussed previously still apply. In particular, the *F*-test will be severely biased.

As another example of a situation in which nonrandom assignment leads to incorrect results when ANCOVA is used, consider the case where the covariate scores are not normally distributed but are sampled from two separate distributions. Suppose the two distributions are both normal with equal variances.

If the two means of these distributions are more than, say, four standard deviations apart and if the cutting point used in the regression-discontinuity analysis lies halfway between the two means, then it is obvious that this analysis will lead to similar results as would have been obtained, had the assignment of subjects to treatment groups been based on random assignment from these two distributions.

Apart from these statistical considerations, regression-discontinuity designs may also criticized on methodological grounds. This may be clarified by the following example. Suppose that there is no relationship between the dependent variable and the covariate scores, i.e. the pooled within-groups regression coefficient is zero. If in such a case a regression-discontinuity design is used in combination with ANCOVA, part if not all of the treatment effect will be removed due to the fact that ANCOVA corrects for the artificially created correlation between treatments and covariate. This corresponds to the example shown in Figure 2. Hence, such designs are not statistically efficient with respect to the estimation of treatment effects. This has also been shown by Goldberger (see Cook & Campbell, 1979, p.205). Goldberger compared the power of the regression-discontinuity design with that of completely randomized designs. His results showed that the completely randomized design was 2.75 times as efficient in detecting treatment effects. This means that in order to obtain the same power, regression-discontinuity designs require approximately 2.75 as many subjects. Hence, the assertion of Overall and Woodward (1977b, P.594)

> *... random assignment is not an essential principle of good experimental design. The researcher is in a position to make strong statistical inferences, so long as he controls the assignment to treatment groups according to any of a variety of possible assignment rules, only one of which is random assignment.*

may be correct from a purely statistical point of view, but does not take the efficiency into account which is also part of good experimental design.

III. Alternatives to ANCOVA in case of fallible covariate measurements

In the previous section we have shown that ANCOVA does not lead to valid results when the covariate is measured with error. A natural question to ask, is whether there are feasible alternatives to ANCOVA for such a case that do lead to valid results. In this section we will present some results that are relevant to this issue. In discussing the alternatives to ANCOVA we will only present the major results since a complete derivation of the formulas would go beyond the scope of the present paper and will be presented elsewhere.

Suppose the data are in accordance with the following model

$$y_{ij} = \mu + \alpha_j + \beta(T_{ij}-\bar{T}) + \varepsilon_{i(j)} \qquad (6)$$

and

$$x_{ij} = T_{ij} + \delta_{i(j)} \quad , \qquad (7)$$

where T_{ij} denotes the true score on the covariate. It is assumed that the disturbance variables $\varepsilon_{i(j)}$ and $\delta_{i(j)}$ are normally distributed with expectation zero and variances equal to σ^2_ε and σ^2_δ, respectively. This is similar to a well-known model that has been studied extensively in econometrics and mathematical statistics (see e.g. Kendall & Stuart, 1967) under the heading of linear *functional* and *structural* relationships. The distinction between a functional and a structural relationship is that in the former approach no distributional assumptions are made regarding the T_{ij}, whereas in the structural case one generally assumes that the T_{ij} are sampled from normal distributions with means μ_j and variances σ^2_j. This implies that in the functional case *all* T_{ij} are parameters (fixed, unknown constants), whereas in the structural case the T_{ij} are random variables that are characterized by the parameters μ_j and σ^2_j. Note that this distinction is similar to the difference between fixed effects and random effects models in ANOVA.

In this section, we will discuss procedures for estimating the parameters for such models and methods for testing various hypotheses. We will give special attention to the question how the conventional ANCOVA null hypothesis, $\alpha_j = 0$ (for all j), should be tested. We will discuss and compare the usual F-test, likelihood ratio tests, and a newly designed test for this hypothesis. We will confine ourselves to a discussion of the distributions of these test statistics under the null hypothesis. Questions concerning the relative power of these tests are beyond the scope of the present paper.

Both the structural and the functional case will be discussed. We will present results concerning a number of special cases including $\alpha_j=0$, $\sigma^2_\varepsilon=\sigma^2_\delta$, and ρ_{xx} (the reliability of x) known. In each of these cases we will derive consistent and, if possible, maximum likelihood (ML) estimators for the parameters of the ANCOVA model.

ANCOVA as a functional relationship

In the functional relationship approach, the true scores T_{ij} are not assumed to be a random sample from a particular parent distribution, but are considered to be fixed, unknown constants (i.e. parameters). For simplicity, we will restrict ourselves to the two-group case with equal n's. In that case we may rewrite Eqs. 6 and 7 as follows

$$y_{ij} = \mu + \alpha Z_j + \beta(T_{ij}-\bar{T}) + \varepsilon_{i(j)}$$

$$x_{ij} = T_{ij} + \delta_{i(j)}$$

where $Z_j = 1$ for $j = 1$, and $Z_j = -1$ for $j = 2$.

This implies that the model has $(2n+5)$ parameters: T_{11}, $T_{21}, \ldots, T_{n1}$, $T_{12}, \ldots, T_{n2}$, μ, α, β, σ^2_ε, and σ^2_δ. The T_{ij}'s are usually referred to as *incidental* parameters, and the other five parameters are called *structural*. Incidental parameters are specific to individual observations, while the structural parameters are common to sets of observations. The number of structural parameters does not depend on the sample size. This kind of situation where the total number of parameters increases with the sample size, was investigated in a classic paper by Neyman and Scott (1948). The presence of incidental parameters poses a problem in statistical estimation since the standard definition of consistency in estimation becomes meaningless in this case. An estimator is called consistent, when the estimate converges with probability one to the true value as the sample size increases. The problem here is that the convergence of the estimates for the structural parameters depends on the assumptions one makes concerning the asymptotic behavior of the incidental parameters as sample size increases. It should perhaps be noted that the notion of consistency obviously does not apply to the incidental parameters themselves. However, even if we make the assumption that the variance of the T_{ij}'s (and hence the variance of the x_{ij}'s) converges to a fixed value, the ML estimators for the structural parameters are not necessarily consistent. An example of this will be shown when we discuss the ML estimators for the parameters of the functional model for ANCOVA. For the moment, it suffices to note that the standard theory of ML estimation does not necessarily apply to this kind of functional relationship model.

It should be noted that under the null hypothesis $\alpha=0$, the present model reduces to the famous linear functional relationship problem discussed by e.g. Kendall and Stuart (1967) and Anderson (1976, 1984). In this case the likelihood function is given by

$$L = (4\pi^2\sigma_\delta^2\sigma_\varepsilon^2)^{-n} \exp\{-(2\sigma_\delta^2)^{-1} \Sigma\Sigma\, (x_{ij}-T_{ij})^2 - (2\sigma_\varepsilon^2)^{-1} \Sigma\Sigma\, (y_{ij}-\mu-\beta(T_{ij}-\bar{T}))^2\}$$

It can be shown, however, that this likelihood function has no maximum. To illustrate this fact, suppose we let $T_{ij} = x_{ij}$. It is easy to see that with this substitution $L \to \infty$ as σ_δ^2 approaches zero. Since the likelihood function has no maximum, maximum likelihood estimators do not exist (see also Anderson, 1984; Anderson & Rubin, 1956). Similar results can be obtained for the present

ANCOVA model in case $\alpha \neq 0$. Estimation methods other than the ML-method do exist for this case, e.g. Geary's method of using product cumulants (see Kendall & Stuart, 1967). This method however becomes useless as the distribution of the true scores T_{ij} approaches a normal distribution (in that case all cumulants of order ≥ 3 are zero and the equation system used in estimating the parameters becomes unsolvable). Hence, it is to be expected that such a method will not be very useful in practice.

The assumption of equal variances

In order to obtain more meaningful results, we must either impose some restriction on the model or obtain additional information such as knowledge concerning the reliability of the covariate. The latter case will be discussed in the next section. The most common identifying restriction that is made in this situation is that the error variances of x and y are equal, i.e. $\sigma_\varepsilon^2 = \sigma_\delta^2$. It should be noted that the more general assumption that $\sigma_\varepsilon^2 = \lambda\sigma_\delta^2$ (with λ a known constant), is identical to the present restriction provided that we rescale the observations accordingly. It should be clear that the resulting estimates for μ, α and β, will also have to be rescaled.

With this assumption, the following ML-estimators are obtained:

$$\hat{T}_{ij} = \{x_{ij}+\hat{\beta}(y_{ij}-\bar{y}-\hat{\alpha}Z_j+\hat{\beta}\bar{x})\}/C$$

where $C = (1+\hat{\beta}^2)$, and $Z_j = 1$ for $j = 1$ and $Z_j = -1$ for $j = 2$,

$$\hat{\mu} = \bar{y},$$

$$\hat{\alpha} = (\bar{y}_1-\bar{y}) - \hat{\beta}(\bar{x}_1-\bar{x}),$$

$$\hat{\sigma}_\varepsilon^2 = (\hat{\beta}^2W_{xx}-2\hat{\beta}W_{xy}+W_{yy})/2NC \ ,$$

and

$$\hat{\beta} = \{W_{yy}-W_{xx}+\sqrt{(W_{yy}-W_{xx})^2+4W_{xy}^2}\}/2W_{xy} ,$$

provided that $W_{xy} \neq 0$. In these equations W_{xx}, W_{yy}, and W_{xy} denote the pooled within-groups sums of squares and cross-products.

What can be said about the properties of these estimators? First of all, let us consider the question of the consistency of these estimators. In order for this to be meaningful, we have to make some assumption concerning the asymptotic behavior of the incidental parameters. Let us assume that the pooled within-groups variance of T,

$$\Sigma\Sigma\ (T_{ij}-\bar{T}_j)^2/N ,$$

converges to a fixed value S_T^2. In that case the sample pooled within-groups variances and covariances converge to

$$W_{xx}/N \rightarrow S_T^2 + \sigma_\varepsilon^2$$

$$W_{xy}/N \rightarrow \beta S_T^2$$

$$W_{yy}/N \rightarrow \beta^2 S_T^2 + \sigma_\varepsilon^2 .$$

Upon insertion of these results into the equation for $\hat{\beta}$, we obtain the result that $\hat{\beta} \rightarrow \beta$, and hence, $\hat{\beta}$ is a consistent estimator for β. The estimator for the error variance σ_ε^2, however, is not consistent but converges to $\sigma_\varepsilon^2/2$. This illustrates the abovementioned fact that in the presence of incidental parameters ML-estimators are not always consistent. In this case, however, the inconsistency can be easily remedied by using $2\hat{\sigma}_\varepsilon^2$ as an estimator for σ_ε^2. In the remainder of this paper we will denote the estimator $2\hat{\sigma}_\varepsilon^2$ by $\hat{\sigma}^2$.

Insert Figure 3 about here

Although the present estimator for β was derived using the ML-method, it is of some interest to note that it is possible to give a least squares interpretation to this estimator. It can be shown that the within-groups regression lines based on $\hat{\beta}$ are such that they minimize the sum of squared distances of the observed points from the fitted lines. That is, the parameter estimate minimizes the expression

$$\phi = \Sigma\Sigma \{(y_{ij}-\bar{y}_j) - \beta(x_{ij}-\bar{x}_j)\}^2/(1+\beta^2).$$

Hence, $\hat{\beta}$ can be said to be a generalized least squares estimator for β. It is not very difficult to see that the minimum value of ϕ equals $\hat{\sigma}^2$, the consistent estimator for σ^2_ε. This is illustrated in Figure 3, which shows some observed points and the distances to the fitted regression lines. Such an estimation procedure is also known as "orthogonal regression". Note that this type of regression is a compromise between the conventional regression of y on x and that of x on y. The regression of y on x minimizes the distances parallel to the Y-axis, while the regression of x on y minimizes the distances parallel to the X-axis. In orthogonal regression, one minimizes the distances perpendicular to the regression line.

The present solution may be characterized in yet another way. It can be shown (see e.g. Anderson, 1984) that it corresponds to the first principal component of the pooled within-groups variance-covariance matrix; that is, the line is in the direction of the maximal scatter. This leads to the following quadratic equation in $\hat{\beta}$:

$$\hat{\beta}^2 W_{xy} + \hat{\beta}(W_{xx}-W_{yy}) - W_{xy} = 0 \quad .$$

The correct estimate is given by the positive root of this equation which corresponds to the formula given above (the negative root leads to inconsistencies with respect to the other parameter, i.e. negative variance estimates).

On comparing these estimators with those of the conventional ANCOVA model, we may note a number of similarities. First of all, both $\hat{\beta}$ and b_W (the regression coefficient in ANCOVA) are calculated from the pooled within-groups variance-covariance matrix. Hence, these estimates are not sensitive to the differences between groups. Next, the estimation equations for both $\hat{\mu}$ and $\hat{\alpha}$ are identical to those of the conventional ANCOVA model, provided $\hat{\beta}$ is substituted for b_W. Therefore, the present analysis procedure corresponds to a conventional ANCOVA analysis provided that in the estimation of the slope of the regression lines the measurement error in the covariate is taken into account. We will make use of this correspondence in the construction of a test statistic for the hypothesis $\alpha = 0$.

Since we are using the ML-method, the most natural test statistic would seem to be the traditional chi-square test based on the likelihood ratio statistic. In most statistical models a restriction on the parameters (such as $\alpha = 0$ instead of α free) may be tested by a comparison of the likelihood under the general model (the traditional alternative hypothesis) with that under the restricted model (the null hypothesis). Let us denote these by, respectively, L_1 and L_0. The likelihood ratio statistic λ refers to the ratio L_0/L_1. The maximum value of λ equals 1 which is attained when the two models fit the data equally well (this occurs when the estimates under the alternative hypothesis already obey the restriction). The minimum value is equal to 0 which value is approximated as the data deviate more and more from the null hypothesis. A famous theorem in mathematical statistics is that under quite general conditions the distribution of the statistic $-2ln\lambda$ approximates, as sample size increases, a chi-square distribution with the number of degrees of freedom equal to the difference in the number of estimated parameters. This approach, however, breaks down in this case due to the presence of incidental parameters. The reason for this is that the abovementioned theorem is based on the assumption that the number of parameters does not change with sample size. This condition is obviously violated in this case.

Insert Table 4 about here

This conclusion was verified by Monte-Carlo simulation of the present functional ANCOVA model (see Table 4). These results show that $-2ln\lambda$ does not approximate a chi-square distribution at all, not even when the sample size is quite large.

In order to obtain a meaningful test statistic we evidently have to take a different approach. A possibly fruitful angle to attack this problem is provided by a reconsideration of the test statistic in the ordinary ANCOVA model. It can be shown that in this model $\hat{\alpha}$ is normally distributed with mean α and variance $var(\hat{\alpha})$. By standard methods an estimate of $var(\hat{\alpha})$, $\widehat{var}(\hat{\alpha})$, may be obtained from the sample results. Hence,

$$t = (\hat{\alpha}-\alpha)/\sqrt{\widehat{var}(\hat{\alpha})}$$

follows a t-distribution with $(N-3)$ degrees of freedom ($N-3$ since 3 degrees of freedom are used up in the estimation of μ, α, and β). In view of the well known relation between the t-distribution and the F-distribution, t^2 follows an F-distribution with 1 and $(N-3)$ degrees of freedom. The important result for our purposes is that in the case of the null hypothesis $\alpha = 0$, t^2 is equivalent to the usual F-statistic in ANCOVA.

It happens to be the case that in the functional ANCOVA model, $\hat{\alpha}$ is asymptotically normally distributed as $N \rightarrow \infty$ (assuming the pooled within-groups variance of T converges to a fixed value S_T^2). Hence, we will consider the test statistic

$$t = \hat{\alpha}/\sqrt{\hat{var}(\hat{\alpha})}$$

for testing the null hypothesis $\alpha = 0$. The problem is thus reduced to finding a reasonable estimate for var($\hat{\alpha}$). A first-order approximation of this variance is

$$var(\hat{\alpha}) \approx \sigma_\varepsilon^2(1+\beta^2)/N + var(\hat{\beta})(\mu_1-\mu_2)^2/4 \quad .$$

Furthermore, it can be shown (see also Robertson, 1974) that

$$var(\hat{\beta}) \approx \sigma_\varepsilon^2\{(1+\beta^2)S_T^2+\sigma_\varepsilon^2\}/(NS_T^4) \quad . \qquad (8)$$

Unfortunately, these formula's are large-sample approximations that are not very good with small samples and/or large error variances (as was observed from Monte-Carlo simulations). This is probably related to the fact that the exact distribution of $\hat{\beta}$ has some peculiar characteristics (e.g. infinite moments, see Anderson & Sawa, 1982). In unfavorable circumstances, these formula's severely underestimate the variances obtained from Monte-Carlo simulations. It turns out, however, that a simple correction for bias reduces many of these problems considerably. It may be shown (see e.g. Robertson, 1974) that the expected value of $\hat{\beta}$ is approximately equal to

$$E(\hat{\beta}) \approx \beta\{1 + \sigma_\varepsilon^2\ ((1+\beta^2)S_T^2+\sigma_\varepsilon^2)/(N(1+\beta^2)S_T^4)\} \ .$$

Hence, this formula may be used to obtain an approximately unbiased estimate for β, $\hat{\beta}_C$:

$$\hat{\beta}_C = \hat{\beta}/C \ ,$$

where C is given by

$$C = 1 + \hat{\sigma}^2\{(1+\hat{\beta}^2)\hat{S}_T^2 + \hat{\sigma}^2\}/\{N(1+\hat{\beta}^2)\hat{S}_T^4\} \quad ,$$

with

$$\hat{S}^2_T = W_{xx}/N - \hat{\sigma}^2.$$

Insert Table 5 about here

Simulation results show that the variance of $\hat{\beta}_c$ is well approximated (even with relatively small sample sizes) by Equation 8. A corrected estimate for α is then given by:

$$\hat{\alpha}_c = (\bar{y}_1-\bar{y}) - \hat{\beta}_c(\bar{x}_1-\bar{x}) \ . \qquad (9)$$

Hence, we conclude that $\hat{\beta}_c$ and $\hat{\alpha}_c$ are approximately normally distributed with mean β and α, respectively, and variances var($\hat{\beta}$) and var($\hat{\alpha}$) as given above. As a final step, sample estimates have to be plugged into these formula's to obtain estimated variances for $\hat{\alpha}_c$ and $\hat{\beta}_c$. It turns out that the best approximation is provided by using $\hat{\beta}_c$, $\hat{S}^2_T$ and $N\hat{\sigma}^2/(N-3)$ in these formula's as estimators for, respectively, β, S^2_T and σ^2_ε. Table 5 gives some results showing how well the resulting test statistic for the hypothesis $\alpha = 0$ approaches a t-distribution with $(N-3)$ degrees of freedom. Note that the approximation becomes better as σ^2_ε decreases and as sample size increases. Although in some cases the approximation cannot be said to be very good, it should be noted that even in such cases the present test statistic is still always quite superior to the usual F-test. Hence, we may conclude that *the present test statistic is uniformly superior to the traditional F-test*. Table 6 gives a numerical example in order to illustrate the necessary calculations.

Insert Table 6 about here

Extensions of the basic model

Similar procedures can be developed for a number of extensions of the above model. The major problem with that model is the rather restrictive assumption concerning the error variances. The assumption of equality of σ_ε^2 and σ_δ^2 (or the equivalent assumption that the ratio of these variances is known) may not be realistic in many applications. This is especially so since σ_ε^2 will usually consist of two components, only one of which contributes to σ_δ^2. These two components are the measurement error and the error in the equation, i.e. the deviation of the error-free dependent variable score from the value predicted on the basis of the functional relationship. σ_ε^2 measures the combined effect of these two sources of variation, while σ_δ^2 consists of measurement error only. However, there is no way out of this predicament unless we have some additional information that allows us to identify the error variances separately.

In practice, if we do have additional information, it will usuallly be of a kind that allows us to determine or estimate the reliability of the covariate measurements. One such instance was analyzed (in a not widely known paper) by Lord (1960). This analysis (which is consistent with the general approach followed in this paper) assumes that two *parallel* measurements of the covariate are available. In effect, this assumption implies that replicated observations are available concerning the T_{ij}-scores. This of course allows us to estimate σ_δ^2 from the replicated observations, and hence enables us to obtain a separate consistent estimator for σ_ε^2 (see Lord, 1960). However, it should be noted that although we are in this case able to obtain consistent estimators for the parameters, these are *not* ML-estimators. The likelihood function for this situation still remains unbounded, for similar reasons as discussed above. The same difficulty arises in all other cases in which we have additional information that allows σ_δ^2 to be estimated. However, consistent estimators may be derived, based on the pooled within-groups variance-covariance matrix. As in the model discussed in the previous section, the resulting estimator for β may then be corrected for bias and an approximate *t*-test may be constructed for the hypothesis $\alpha = 0$.

It should be noted that the assumption of *parallel* measurements is not necessary. Consistent estimation is also possible when the "true" covariate is measured through two socalled *congeneric* tests. In this case a second covariate, z, is available that is known to be correlated with the true score of x, but is independent of the errors of x and y. Such a variable z is usually referred to as an *instrumental variable* and its use in the estimation of the parameters of functional relationships has been studied in the statistical and econometric literature (see e.g. Kendall & Stuart, 1967; Moran, 1971).

As an example, let us consider the case that the reliability of the covariate, ρ_{xx}, is known. In that case, σ_δ^2 may be estimated as

$$\hat{\sigma}_\delta^2 = (1-\rho_{xx})W_{xx}/N \quad .$$

Hence, $\hat{S}_T^2$ is given by

$$\hat{S}_T^2 = \rho W_{xx}/N \quad .$$

Consistent moment estimators for the remaining parameters may now be obtained as follows. The estimator for β is defined as:

$$\hat{\beta} = W_{xy}/N\hat{S}_T^2 \quad .$$

Estimators for α and σ_ε^2 are given by

$$\hat{\alpha} = (\bar{y}_1-\bar{y}) - \hat{\beta}(\bar{x}_1-\bar{x}) \quad ,$$

$$\hat{\sigma}_\varepsilon^2 = W_{yy}/N - \hat{\beta}^2\hat{S}_T^2 \quad .$$

Using standard methods, it may be shown that the expectation and variance of $\hat{\beta}$ are approximately equal to

$$E(\hat{\beta}) \approx \beta\{1 + 2S_T^2\sigma_\delta^2/N(S_T^2+\sigma_\delta^2)^2\} \quad ,$$

$$var(\hat{\beta}) \approx \{S_T^2(\sigma_\varepsilon^2+\beta^2\sigma_\delta^2)+\sigma_\varepsilon^2\sigma_\delta^2\}/NS_T^4 \; - 2\beta^2\sigma_\delta^4/N(S_T^2+\sigma_\delta^2)^2 \quad .$$

As in the previous case, an approximately unbiased estimator for β, $\hat{\beta}_c$, may be obtained as follows:

$$\hat{\beta}_c = \hat{\beta}/C$$

where C is given by

$$C = 1 + 2\hat{S}_T^2\hat{\sigma}_\delta^2/N(\hat{S}_T^2+\hat{\sigma}_\delta^2)^2 \quad .$$

The corrected estimator for α, $\hat{\alpha}_c$, is defined as before (see Equation 9). The variance of $\hat{\alpha}_c$ will be approximately equal to

$$var(\hat{\alpha}) \simeq (\sigma_\varepsilon^2+\beta^2\sigma_\delta^2)/N + var(\hat{\beta})(\mu_1-\mu_2)^2/4 \quad .$$

As before, these formulas are large sample approximations. Using similar arguments as in the previous case, a t-test may be constructed for the null hypothesis $\alpha = 0$. As an illustration Table 7 gives the necessary calculations when this procedure is applied to the numerical example given by Lord (1960). Since the number of observations is in this case different in the two groups, it is easiest to test the hypothesis $\alpha_1-\alpha_2=0$ instead of $\alpha=0$. The approximate variance of $\hat{\alpha}_1-\hat{\alpha}_2$ is then given by:

$$var(\hat{\alpha}_1-\hat{\alpha}_2) \simeq (\sigma_\varepsilon^2+\beta^2\sigma_\delta^2)N/(n_1n_2) + var(\hat{\beta})(\mu_1-\mu_2)^2 \quad .$$

Insert Table 7 about here

One of the assumptions of ANCOVA (see e.g. Equation 1) is the equality of the within-groups regression coefficients. In the ordinary ANCOVA model this assumption may be tested by comparing the residual sum of squares about the within-groups regression lines based on the pooled estimate for the within-groups regression coefficient with the residual sum of squares obtained by using a separate regression coefficient for each group (see e.g. Winer, 1971, p.772-773).

A similar approach may be followed in the present case. The ANCOVA model defined in Equations 6-7 is based on the assumption of a single, common, regression coefficient β. A test for the equality of the within-groups regression coefficients may be obtained, starting from a model with separate regression coefficients, i.e. Equation 6 with β_j substituted for β. Under the assumption that $\sigma_\varepsilon^2 = \sigma_\delta^2$, β_j may be estimated from the sums of squares and crossproducts in group j. It may be shown that a ML-estimator for β_j is given by

$$\hat{\beta}_j = \{W_{j,yy} - W_{j,xx} + \sqrt{(W_{j,yy} - W_{j,xx})^2 + 4W_{j,xy}^2}\}/2W_{j,xy} \ ,$$

where $W_{j,..}$ is the sum of squares or crossproducts in group j. As before, unbiased estimates for the β_j's may be obtained as follows:

$$\hat{\beta}_{j,c} = \hat{\beta}_j / C$$

where C is equal to

$$C = 1 + \hat{\sigma}^2\{(1+\hat{\beta}_j^2)\hat{S}_{j,T}^2 + \hat{\sigma}^2\}/\{n(1+\hat{\beta}_j^2)\hat{S}_{j,T}^4\} \quad .$$

In this equation $\hat{\sigma}^2$ and $\hat{S}_{j,T}^2$ are given by

$$\hat{\sigma}^2 = \Sigma\,(\hat{\beta}_j^2 W_{j,xx} - 2\hat{\beta}_j W_{j,xy} + W_{j,yy})/N(1+\hat{\beta}_j^2)$$

and

$$\hat{S}^2_{j,T} = W_{j,xx}/n - \hat{\sigma}^2 \quad .$$

In large samples, $\hat{\beta}_{j,c}$ will be approximately normally distributed with mean β_j and variance var($\hat{\beta}_j$),

$$var(\hat{\beta}_j) \approx \sigma_\varepsilon^2\{(1+\beta_j^2)S^2_{j,T} + \sigma_\varepsilon^2\}/nS^4_{j,T}$$

where sample estimates have to be substituted for population parameters, using $N\hat{\sigma}^2/(N-4)$ as an estimator for σ_ε^2. The null hypothesis $\beta_1 = \beta_2$ may be tested with the statistic

$$t = (\hat{\beta}_1 - \hat{\beta}_2)/\sqrt{2S^2/n}$$

where

$$S^2 = \{\widehat{var}(\hat{\beta}_1) + \widehat{var}(\hat{\beta}_2)\}/2 \quad .$$

This statistic is approximately t-distributed with $(N-4)$ degrees of freedom. It should be noted that if the assumption of equal β's has to be rejected, application of the ANCOVA model should be strongly discouraged, since the results will generally not be meaningful. In that case, a comparison between groups depends on the value of the covariate (see e.g. Tatsuoka, 1971).

ANCOVA as a structural relationship

In the structural relationship approach, the true scores T_{ij} are assumed to be randomly sampled from normal distributions with means μ_j and variances σ_j^2. Given this assumption, the data in each group follow a bivariate normal distribution. It is then possible to derive ML-estimates for the parameters of the model by maximizing the likelihood function. Since these ML-estimates may be obtained from the LISREL-program (Jöreskog & Sörbom, 1981), we will not present the likelihood function nor any of the equations that can be derived for the parameter estimates.

Although it is not entirely clear why, it turns out that the ML-estimates do not behave very well unless it is assumed that $\sigma_\varepsilon^2 = \sigma_\delta^2$ (except when the model fits the data perfectly). Perhaps this should come as no surprise since ML-estimators do not exist in the corresponding functional model. We will return to this problem later on in this paper.

As the LISREL model will be used extensively in the remainder of this section, we will briefly discuss its basic equations. For a more detailed introduction to the LISREL approach we refer to Lomax (1982, 1983). The LISREL model is based on the following structural equation system

$$\eta = B\eta + \Gamma\xi + \zeta$$

where B and Γ are coefficient matrices, η and ξ are vectors corresponding to the latent dependent and independent variables, and ζ is a vector of disturbance variables. The latent variables are related to the observed independent and dependent variables x and y, respectively, through the following measurement model:

$$x = \Lambda_x\xi + \delta$$

and

$$y = \Lambda_y\eta + \varepsilon \quad ,$$

where Λ_x and Λ_y contain the regression coefficients of x and y upon ξ and η, and δ and ε denote the measurement errors in the observed variables. The measurement error in x is assumed to be uncorrelated with the measurement error in y. In addition, it is assumed that the measurement errors are not correlated with the latent variables η and ξ, and the disturbance variables ζ. Furthermore it is assumed that the disturbance variables ζ are uncorrelated with ξ, and that the latent variables and the disturbance variables have zero expectation. The LISREL model is completely specified if the variance-covariance matrices of the measurement errors ε and δ, the disturbance variables ζ, and the latent variables ξ are suitably defined. Provided the model is identifiable, the parameters of the model may be estimated by fitting the variance-covariance matrix expected under the LISREL model to the observed variance-covariance matrix by using a suitable loss function (Jöreskog & Sörbom, 1981).

The reader familiar with the factor analysis model may have noted that the LISREL model is equivalent to a restricted factor analysis model in which the factors, η and ξ, satisfy a linear structural equation system as the one defined above. Extensions of the LISREL model given above include the multisample model and the socalled structured means model. The former extension allows one to specify a separate LISREL model in each group. In addition, one may specify equality constraints on the LISREL parameters over groups. The structured means model offers a relaxation of the assumption that the random variables in the model have zero expectation. A combination of these extensions leads to the most general LISREL model, the structured means multi-sample model. For a more complete description of the LISREL model and examples of its use, we refer to Lomax (1982, 1983) and Jöreskog & Sörbom (1981).

The case of equal variances

In this section we will discuss the analysis of the structural ANCOVA model using the LISREL approach. Using the LISREL terminology, the basic equations of the structural ANCOVA model, Equations 6 and 7, may be rewritten as follows:

$$\begin{bmatrix} \eta_{ij} \\ T_{ij} \end{bmatrix} = \begin{bmatrix} 0 & \beta \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \eta_{ij} \\ T_{ij} \end{bmatrix} + \begin{bmatrix} \mu+\alpha_j-\beta T \\ \mu_j \end{bmatrix} \begin{bmatrix} 1 \end{bmatrix} + \begin{bmatrix} 0 \\ T_{ij}-\mu_j \end{bmatrix}$$

$$\begin{bmatrix} y_{ij} \\ x_{ij} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \eta_{ij} \\ T_{ij} \end{bmatrix} + \begin{bmatrix} \varepsilon_{i(j)} \\ \delta_{i(j)} \end{bmatrix}$$

Since in each group the equations contain constant intercept terms, equal to, respectively, $\mu+\alpha_j-\beta T$ and μ_j, we have the structured means version of the LISREL model. This implies that the LISREL specification should contain an x-variable which is identical to 1, and thus coincides with ξ. The intercept terms are included in the matrix Γ. In this case, one should analyze the raw moment matrix instead of the covariance or correlation matrix. In addition, the fixed-x option of LISREL should be used (see Jöreskog & Sörbom, 1981).

Under the assumption that $\sigma_\varepsilon^2 = \sigma_\delta^2$, the model contains $3g+2$ free parameters: β, σ_ε^2, and for each of the g groups $\mu+\alpha_j-\beta T$, μ_j, and σ_j^2. The estimates of σ_ε^2 and σ_j^2 are given by the variance-covariance matrices of ε and ζ. The null hypothesis $\alpha_j = 0$ may be tested by constraining the intercept parameters $\mu+\alpha_j-\beta T$ to be equal. The appropriate likelihood ratio (LR) statistic for testing this hypothesis is formed by subtracting the chi-square value reported by LISREL under the unconstrained model from the chi-square value obtained in the restricted case. In large samples this statistic follows a chi-square distribution with $(g-1)$ degrees of freedom.

It should be noted, however, that the chi-square values reported by the LISREL program are incorrect. It turns out that the reported chi-square values should be multiplied by a factor $n/(n-1)$, assuming equal n's in each group. (No simple correction can be applied if the groups contain unequal numbers of observations, since the LISREL program does not report chi-square values for each group). The reason for this discrepancy is that the LISREL model is based on the assumption that the observed covariance matrix follows a Wishart distribution. The structural ANCOVA model, however, is based on the assumption that the observations in each group follow a multivariate normal distribution. Although these two assumptions are closely related, they are not equivalent. In particular, the likelihood functions are slightly different. This difference implies that the chi-square values in each group have to be corrected. This is generally the case if intercept terms are included in the model and the raw moment matrix is analyzed (see Jöreskog, 1973, p.93).

Insert Table 8 about here

An example of the LISREL approach to the analysis of a linear structural ANCOVA model is presented in Table 8. This table includes the appropriate parameter estimates and the corrected chi-square values. The example which will be discussed, is the same as was used to illustrate the functional solution (see Table 6). These data were generated in such a way that the observed statistics would correspond almost perfectly to a structural ANCOVA model. There are only two minor deviations from the model as presented above: (1) the two groups were slightly different with respect to the parameter β, and (2) σ^2_δ and σ^2_ε were also given different (but nearly equal) values. The parameter values were: $\mu = 0.0$, $\alpha = 1.6$, $\beta_1 = 2.1$, $\beta_2 = 1.9$, $\mu_1 = 5.0$, $\mu_2 = 8.0$, $\sigma^2_1 = 10.0$, $\sigma^2_2 = 15.0$, $\sigma^2_\varepsilon = 1.9$, $\sigma^2_\delta = 2.1$ and $n = 10$ in each group. The resulting moment matrices are given in Table 8.

Note that the parameter estimates closely resemble the corresponding functional estimates. Thus, the latter values may be used as initial estimates for the structural parameters. Initial values for the two remaining parameters, σ^2_1 and σ^2_2, may be obtained from the following formula:

$$\hat{\sigma}^2_j = (s^2_{x(j)} + 2\hat{\beta}s_{xy(j)} + \hat{\beta}^2 s^2_{y(j)} - \hat{\sigma}^2_\varepsilon(1+\hat{\beta}^2))/(1+\hat{\beta}^2)^2 \quad .$$

These initial parameter estimates are quite useful since the LISREL program may arrive at an incorrect solution when supplied with bad initial estimates. Incorrect solutions are not infrequent in LISREL and are characterized by negative variance estimates in the Ψ matrix. The LISREL program may generate such solutions because it does not restrict the variance estimates to non-negative values. It turns out however that in many cases the likelihood function has several (local) maxima, only one of which is in the correct parameter space.

This is a serious flaw in the LISREL program that one should be aware of. For example, when supplied with bad initial estimates, LISREL obtained an incorrect solution (under the alternative hypothesis) for the numerical data given above with a chi-square value of 0.37. It should be noted that the remark in the LISREL manual (Jöreskog & Sörbom, 1981, p. 31-32) that such a solution indicates "that the model is wrong or that the sample size is too small" is incorrect. Instead, it will always be necessary to check the appropriateness of the solution using different initial estimates.

Extensions and further tests of the structural model

Our solution to the fallible covariate problem in the structural case (as well as in the functional case) is based on the assumption $\sigma^2_\varepsilon=\sigma^2_\delta$. As mentioned in the discussion of the functional model, this a very restrictive assumption that is difficult to defend. It is not known to what extent violation of this assumption biases the results of the analysis. In particular, we do not know whether this leads to a substantial bias in the likelihood-ratio test for the ANCOVA null hypothesis. Although the present case seems quite similar to the functional case, there are some important differences. In the functional model, the parameters σ^2_ε and σ^2_δ are not both identifiable (unless additional information is present). In the structural model, on the other hand, these parameters cannot be said to be unidentifiable. The reason for this is that these parameters are estimated correctly when the observed covariance matrix (or moment matrix) fits the model *exactly*, i.e. if a perfect solution is possible. Problems arise however as soon as the observed matrix deviates slightly (and nonsignificantly) from the predicted structure.

Insert Table 9 about here

In order to demonstrate this point, we generated a number of moment matrices which violated the assumption $\sigma_\varepsilon^2=\sigma_\delta^2$ in varying degrees. In each case, σ_ε^2 was set equal to 2.0, while σ_δ^2 was varied between 2.0 and 17.0. The remaining parameters (except for β_1 and β_2) were equal to those used to generate the data in Table 8. When the regression coefficients in the two groups, β_1 and β_2, were set equal to each other (in which case a perfect solution is possible), the correct solution was always obtained and all the parameter estimates were equal to the true values. However, when β_1 and β_2 were given slightly different values, strange and unexpected results were obtained. Note that none of these datamatrices violate the structural ANCOVA model with $\sigma_\varepsilon^2\neq\sigma_\delta^2$ to any significant degree. Table 9 gives the most important results of this analysis. In this table, the estimates for σ_ε^2 and σ_δ^2 are given as well as the chi-square values from the LR-test for the hypothesis $\sigma_\varepsilon^2=\sigma_\delta^2$. These parameter estimates were not obtained with the LISREL program, but with a general purpose minimization program that allows upper and lower limits on the parameter values (James & Roos, 1975). This program was used because in these cases the LISREL estimates for σ_ε^2 and σ_δ^2 were often outside the admissible parameter space (LISREL does not restrict the parameter estimates to values within the admissible parameter space). For these parameter values, the results are very unstable and strongly dependent on small differences in β_1 and β_2.

The lefthand part of Table 9 gives the results for β_1=1.9 and β_2=2.1. In this case σ_ε^2 was always estimated as 0.0 (the lower bound), while σ_δ^2 was always overestimated in a systematic way: the estimated value for σ_δ^2 was equal to the true value plus a constant (0.499). More importantly, the LR-test does not seem to be very sensitive to changes in σ_δ^2 (although the chi-square values become slightly larger as σ_δ^2 deviates more and more from σ_ε^2). In all cases, the value of this statistic is quite small and never leads to rejection of the hypothesis $\sigma_\varepsilon^2=\sigma_\delta^2$. The righthand part of Table 9 gives the corresponding results for β_1=2.1 and β_2=1.9. Although the regression coefficients have not been changed very much, the pattern of the parameter estimates is completely different from the previous case. In this case, σ_ε^2 is grossly overestimated while σ_δ^2 is severely underestimated. Moreover, the chi-square values of the LR-test *decrease* with increasing differences between σ_ε^2 and σ_δ^2. Hence, we may conclude that the hypothesis $\sigma_\varepsilon^2=\sigma_\delta^2$ is not testable and that the separate estimation of σ_ε^2 and σ_δ^2 leads to unsatisfactory results.

More satisfactory results can only be obtained if we have additional information that allows the identification of both σ_ε^2 and σ_δ^2. Suppose for example, that we know (or have information that permits the estimation of) the reliability of the covariate, ρ_{xx}. In that case, σ_δ^2 might be fixed at $(1-\rho_{xx})W_{xx}/N$, as in the corresponding functional case. This allows σ_ε^2 to be estimated. Although this does not correspond to the conventional ANCOVA model, knowledge regarding ρ_{xx} allows one to estimate separate error variances within each group. It is usually assumed that the measurement error σ_δ^2 is equal in all groups. If there is reason to suspect that this assumption is not correct, separate reliabilities should be used for the estimation of these variances.

Insert Table 10 about here

As an example of such an analysis with different reliabilities, we reanalyzed Lord's numerical example (Lord, 1960). Three types of analysis were performed, the results of which are given in Table 10. Model I is the type of analysis we have just described, adapted to this situation, i.e. σ_δ^2 in each group is set equal to $(1-\rho_j)W_{j,xx}/n_j$. Model II is the correct analysis given the assumption of unequal measurement errors. In this analysis σ_δ^2 was set equal to $\sigma_j^2(1-\rho)/\rho$, where σ_j^2 is the true score variance in group j. Since the LISREL program does not allow such a restriction on the parameters, these estimates were obtained by direct minimization of the appropriate likelihood function using a general-purpose minimization routine. Another type of solution was presented by Sörbom (1978). Sörbom used the information concerning the reliabilities in a different way. Instead of fixing or restraining σ^2_δ, Sörbom created an artificial second covariate which was constructed in such a way that the two covariates were parallel measurements with a correlation equal to the observed reliability. In doing so, Sörbom followed the original approach taken by Lord (1960) who analyzed these data in a similar way. Model III gives the estimates for this analysis obtained with the LISREL program (for reasons unclear to us, the results deviate somewhat from those reported by Sörbom, 1978).

On comparing the results, it is evident that there are only minor differences in this case between the three approaches. However, if one does not have easy access to a general-purpose minimization routine and prefers to use the LISREL program, it is in our opinion advisable to use the Model I type of analysis instead of the Model III or Sörbom type of analysis. The behavior of the likelihood function may depend on the assumption of independent parallel measurements (which are in fact not available) and this might affect properties of the estimates such as their standard errors.

It should perhaps be noted that, as in the functional case, it is not necessary to know the reliability of the covariate or to have parallel measurements. Sufficient information to solve the problem is available when the covariate is measured by two congeneric tests. That is, when two independently observed covariate scores are available with true scores that are linearly related. Since we have already discussed the use of such instrumental variables in the functional case and since this case has been dealt with quite extensively by Sörbom (1978), we refer the interested reader to that paper for further details (see Carter & Fuller, 1980, for a discussion of the asymptotic distribution of such estimates for β).

Finally, a likelihood-ratio test for the assumption of equal within-groups regression coefficients can be obtained in a straightforward manner with the

LISREL program. However, contrary to the assertion of Sörbom (1978), it is not possible to test the null hypothesis α_j=0 with the LISREL program if the assumption of equal regression coefficients is not tenable. The reason for this is that in the model underlying the LISREL approach the test for α_j=0 is not independent of interval scale transformations of the covariate measurements. In the LISREL approach to ANCOVA, the model may be written as

$$y_{ij} = \mu_j + \beta_j T_{ij} + \varepsilon_{i(j)}$$

where

$$\mu_j = \mu + \alpha_j - \beta_j \overline{T} \quad .$$

The ANCOVA null hypothesis is tested in LISREL by comparing the μ_j's. The problem now is that these μ_j's may be different even though α_j=0 (for all j). Moreover, scale transformations of T affect the outcome of this likelihood-ratio test. However, as we have already mentioned earlier, it is best not to proceed with an ANCOVA type of analysis when the assumption of equal regression coefficients has been rejected.

Conclusions

In this paper we have critically examined several problems and confusions concerning the proper use and interpretation of the analysis of covariance. The most general conclusion is that if the ANCOVA model as stated in Equation 1 is correct, there is nothing *statistically* wrong with the application of ANCOVA. None of the other assumptions that are frequently mentioned as being necessary for a valid application of ANCOVA, such as equality of between- and within-groups regression coefficients or independence of treatment and covariate, are in fact required for a valid use of ANCOVA. The confusion concerning the situations in which ANCOVA should or should not be applied, seems to be due to an incorrect interpretation of the conclusions that can be drawn from such an analysis.

The application of ANCOVA in situations where intact or non-equivalent groups are used, does however present a *methodological* problem. In such situations, researchers often use ANCOVA as a means to correct for initial differences on the covariate. Such a (mis)use of ANCOVA is in general not correct and cannot be defended in view of the rationale underlying the ANCOVA F-test. As we have emphasized repeatedly, ANCOVA only tests whether there are differences between the various experimental conditions that cannot be explained on the basis of the covariate alone. In this respect, ANCOVA is equivalent to the test of a partial correlation coefficient. In such a test, the proportion of variance that could be attributed to both predictors, is not reflected in the partial correlation coefficient. Such a coefficient only indicates the contribution that is unique to a particular predictor. A nonsignificant partial correlation coefficient does not indicate that that particular variable has no causal influence on the dependent variable. It only indicates that knowledge of that variable does not improve the prediction of the dependent variable, given knowledge of the other predictor variable(s). Hence, *ANCOVA is no substitute for experimental control*, and the logic of ANCOVA gives no support for its use as a means to "equate" intact, non-equivalent groups.

Such an incorrect usage of ANCOVA has been criticized quite cogently in the past. Thus, Anderson (1963, p. 170) remarks: "One may well wonder what exactly it means to ask what the data would look like if they were not what they are". A similar point is made by Lord (1969) who argues that a statistical technique such as ANCOVA cannot, in principle, answer the question what the results would have been had the groups been comparable on the covariate. The answer to such a question obviously depends on the means used to achieve equality on the covariate, and, hence, no general answer is possible. It should be noted however that in this case the difficulty has nothing to do with the statistical rationale underlying ANCOVA, but that the problem is related to the correct interpretation of the results of such an analysis. The confusion seems to be due to the fact that ANCOVA is generally regarded as a kind of ANOVA on corrected scores, a view which we have shown to be incorrect.

A major part of this paper was devoted to a consideration of the effects of measurement error in the covariate. We have shown that, except in certain special cases, a conventional ANCOVA is never appropriate with fallible covariates. Contrary to statements voiced in the literature (Overall & Woodward, 1977a, 1977b), ANCOVA will not only lead to incorrect results when the assignment of subjects to groups is non-random but also when a random assignment procedure is used, that is, when the means of the various groups on the covariate are equal. The only exception that we are aware of occurs when the covariate and the dependent variable have a bivariate normal distribution (under the null hypothesis). In that case, there will be a linear relation between the dependent variable and the *observed* covariate scores and therefore the usual ANCOVA model applies. Note that contrary to statements in the literature, this has nothing to do with whether or not a fixed assignment rule has been used.

Generally speaking, when the covariate is measured with error, a functional or a structural relationship approach is called for. Examples of both types of analysis have been given for a simple two-group design. Several cases have been discussed and we have given special attention to issues of model identifiability. An approximate statistical test based on the functional relationship approach has been constructed. On the basis of our simulation results it may be concluded that this testing procedure is to be preferred to the conventional F-test of the ANCOVA null hypothesis. Further analysis of this approach to the problem of fallible covariates is obviously desirable, especially regarding its extension to more complicated ANCOVA designs. If one is willing to assume a normal distribution for the covariate scores in each group, the ANCOVA model may be formulated as a structural relationship problem. In this case, an analysis based on the LISREL methodology should be performed. It would seem that in both the functional and the structural case, knowledge of the reliability of the covariate is desirable. In most cases, one should try to obtain parallel measurements of the covariate. With such additional information, arbitrary assumptions concerning the error variances can be avoided.

A general conclusion that may be drawn from our analysis is that in discussions of the applicability of a certain statistical technique one should always try to keep statistical and methodological issues apart. Statistical issues can only be decided on the basis of a clear presentation of the statistical model that is assumed for the data. Methodological issues, on the other hand, require an awareness of the exact nature of the testing procedure on which conclusions are being based. Had such a strategy been taken in the past, the confusion concerning the proper use of ANCOVA probably would have been eliminated years ago.

Footnote

This article represents the equal contribution of the two authors. Requests for reprints should be sent to Jeroen G.W. Raaijmakers, Psychologisch Laboratorium, P.O. Box 9104, 6500 HE Nijmegen, The Netherlands.

References

Anderson, N.H. (1963) Comparison of different populations: Resistance to extinction and transfer. *Psychological Review*, *70*, 162-179.

Anderson, T.W. (1976) Estimation of linear functional relationships: Approximate distributions and connections with simultaneous equations in econometrics. *Journal of the Royal Statistical Society, Series B*, *38*, 1-36.

Anderson, T.W. (1984) Estimating linear statistical relationships. *The Annals of Statistics*, *12*, 1-45.

Anderson, T.W. & Rubin, H. (1956) Statistical inference in factor analysis. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, *5*, 111-150.

Anderson. T.W. & Sawa, T. (1982) Exact and approximate distributions of the maximum likelihood estimator of a slope coefficient. *Journal of the Royal Statistical Society, Series B*, *44*, 52-62.

Carter, R.L. & Fuller, W.A. (1980) Instrumental variable estimation of the simple errors-in-variables model. *Journal of the American Statistical Association*, *75*, 687-692.

Cook, T.D. & Campbell, D.T. (1979) *Quasi-experimentation: Design & analysis issues for field settings*. Boston: Houghton Mifflin.

Cox, D.R. & McCullagh, P. (1982) Some aspects of analysis of covariance. *Biometrics*, *38*, 541-561.

Evans, S.H. & Anastasio, E.J. (1968) Misuse of analysis of covariance when treatment effect and covariate are confounded. *Psychological Bulletin*, *69*, 225-234.

James, F. & Roos, M. (1975) MINUIT, a system for function minimization and analysis of the parameter errors and correlations. *Computer physics communications*, *10*, 343-367.

Johnson, N.L. & Kotz, S. (1970) *Continuous univariate distributions-2.* Boston: Houghton Mifflin.

Jöreskog, K.G. (1973) A general method for estimating a linear structural equation system. In A.S. Goldberger & O.D. Duncan (Eds.), *Structural equation models in the social sciences.* New York: Seminar Press.

Jöreskog, K.G. & Sörbom, D. (1981) *LISREL V: Analysis of linear structural relationships by maximum likelihood and least squares methods.* Chicago: International Educational Services.

Kendall, M.G. & Stuart, A. (1967) *The advanced theory of statistics. Vol. 2: Inference and relationship. (2nd ed).* London: Griffin.

Kerlinger, F.N. & Pedhazur, E.J. (1973) *Multiple regression in behavioral research.* New York: Holt, Rinehart and Winston.

Kirk, R.E. (1968) *Experimental design: Procedures for the behavioral sciences.* Belmont, Calif.: Brooks/Cole.

Lomax, R.G. (1982) A guide to LISREL-type structural equation modeling. *Behavior Research Methods & Instrumentation, 14*, 1-8.

Lomax, R.G. (1983) A guide to multiple-sample structural equation modeling. *Behavior Research Methods & Instrumentation, 15*, 580-584.

Lord, F.M. (1960) Large-sample covariance analysis when the control variable is fallible. *Journal of the American Statistical Association, 55*, 307-321.

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

Mood, A.M., Graybill, F.A. & Boes, D.C. (1974) *Introduction to the theory of statistics. (3rd ed).* New York: McGraw-Hill.

Moran, P.A.P. (1971) Estimating structural and functional relationships. *Journal of Multivariate Analysis, 1*, 232-255.

Neyman, J. & Scott, E. (1948) Consistent estimates based on partially consistent observations. *Econometrica, 16*, 1-32.

Overall, J.E. & Woodward, J.A. (1977a) Common misconceptions concerning the analysis of covariance. *The Journal of Multivariate Behavioral Research*, *12*, 171-185.

Overall, J.E. & Woodward, J.A. (1977b) Nonrandom assignment and the analysis of covariance. *Psychological Bulletin*, *84*, 588-594.

Pedhazur, E.J. (1982) *Multiple regression in behavioral research*. New York: Holt, Rinehart & Winston.

Robertson, C.A. (1974) Large-sample theory for the linear structural relation. *Biometrika*, *61*, 353-359.

Searle, S.R. (1971) *Linear models*. New York: Wiley.

Sörbom, D. (1978) An alternative to the methodology for analysis of covariance. *Psychometrika*, *43*, 381-396.

Sprott, D.A. (1970) Note on Evans and Anastasio on the analysis of covariance. *Psychological Bulletin*, *73*, 303-306.

Tatsuoka, M.M. (1971) *Multivariate analysis: Techniques for educational and psychological research*. New York: Wiley.

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

Winer, B.J. (1971) *Statistical principles in experimental design. (2nd ed)*. New York: McGraw-Hill.

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

Table 1

Comparison of ANCOVA to ANOVA on residuals of overall regression line (data of Figure 1)

Source	ANCOVA SS	ANOVA on residuals SS
Between groups	9.60	3.84
Within groups	0.00	5.76
Total	9.60	9.60

Table 2

Results of ANOVA and ANCOVA when $\beta=0$ and treatment and covariate are correlated

(data of Figure 2)

Source	ANOVA SS	ANCOVA SS
Between groups	22.5	0.957
Within groups	1.0	1.0
Total	23.5	1.957

Table 3

Frequency distribution of Q-statistic in case of measurement error in the covariate (expected frequency in each class is 500)

	Q-statistic	
class	$T_1=T_2$	$T_1 \neq T_2$
1	606	2
2	693	1
3	633	3
4	604	7
5	583	2
6	513	12
7	475	28
8	392	44
9	325	152
10	176	4749
chi-square (df=9)	463.88	40158.07

Table 4

Goodness-of-fit (χ^2, df=9) of likelihoodratio statistic to chi-square distribution (df=1)

		$T_1 = T_2 = 0$		$T_1 = -10,\ T_2 = 10$	
n	ρ	β=1	β=5	β=1	β=5
	.5	2501.89	2546.16	4259.84	2191.49
10	.7	2491.91	2534.63	3046.11	2134.68
	.9	2519.33	2537.54	2450.66	2086.76
	.5	1474.86	1399.33	3653.92	1676.22
50	.7	1477.79	1410.75	2305.60	1599.18
	.9	1465.79	1418.30	1654.21	1565.14
	.5	1362.81	1397.32	3484.50	1623.82
100	.7	1357.24	1373.61	2372.22	1559.05
	.9	1354.54	1377.94	1669.68	1478.92

Table 5

Goodness-of-fit values (chi-square, df=9) for proposed t-statistic in comparison to conventional F-test

		t-statistic				conventional F-test			
		$T_1=T_2$		$T_1 \neq T_2$		$T_1=T_2$		$T_1 \neq T_2$	
n	ρ	β=1	β=5	β=1	β=5	β=1	β=5	β=1	β=5
	.5	13.6	30.6	555.1	822.3	54.1	463.9	16975.4	40158.1
10	.7	12.9	20.0	210.6	339.7	33.7	123.6	6574.5	20921.2
	.9	17.6	23.5	47.4	91.2	6.7	13.5	771.7	2621.6
	.5	11.1	15.2	151.9	173.9	60.9	602.3	44940.0	45000.0
50	.7	10.0	16.4	45.2	72.9	40.4	195.7	43263.0	45000.0
	.9	11.5	15.9	12.1	17.5	8.6	21.4	18554.9	35816.3
	.5	7.7	15.9	94.5	136.9	79.4	623.5	45000.0	45000.0
100	.7	6.5	15.0	44.9	65.6	45.6	206.8	44980.0	45000.0
	.9	7.6	14.9	23.5	32.8	11.6	39.2	37251.6	44740.5

Note: Maximum value of chi-square is 45000

Table 6

Numerical example of the proposed test procedure

Pooled within-group covariance matrix:

	Y	X
Y	51.025	
X	24.750	14.600

$y_1=-1.55$ $y_2=1.25$

$x_1= 5.00$ $x_2=8.00$

$n_1=10$ $n_2=10$

Successive steps in calculating test statistic:

$\hat{\beta}=1.977$ $\hat{\beta}_c=1.961$

$\hat{\alpha}=1.566$ $\hat{\alpha}_c=1.542$

$\hat{\sigma}^2_\varepsilon=1.042$ $2N\hat{\sigma}^2_\varepsilon/(N-3)=2.451$

$\hat{\sigma}^2=2\hat{\sigma}^2_\varepsilon=2.084$ $\widehat{Var}(\hat{\beta})=0.049$

$\hat{S}^2_T=12.516$ $\widehat{Var}(\hat{\alpha})=0.705$

$t(17)=1.836$

Table 7

Application of test procedure to example discussed by Lord (1960)

Data:

	group 1	group 2
number of cases	119	93
dependent variable, Y		
mean	1.40	1.57
s.d.	0.75	0.61
covariate, X		
mean	4.07	5.34
s.d.	2.30	1.97
reliability of X	0.80	0.73

Computed values:

pooled reliability, ρ	0.7735
error variance of X, $\hat{\sigma}_\delta^2$	1.058
true score variance, $\hat{S}_T^2$	3.614
slope estimate, $\hat{\beta}$	0.241
error variance of Y, $\hat{\sigma}_\varepsilon^2$	0.269
corrected slope estimate, $\hat{\beta}_c$	0.241
estimate of group difference, $\hat{\alpha}_1-\hat{\alpha}_2$	0.136
variance of $\hat{\beta}_c$	0.000506
variance of $\hat{\alpha}_1-\hat{\alpha}_2$	0.007146
test statistic, t(209)	1.609

Table 8

Numerical example of the analysis of covariance

as a linear structural relationship

Observed moment matrices (input for LISREL):

	Group 1 (n=10)			Group 2 (n=10)		
	Y	X	1	Y	X	1
Y	48.403			57.613		
X	13.250	37.100		38.500	81.100	
1	-1.550	5.000	1.000	1.250	8.000	1.000

Maximum likelihood solution for parameters:

	H_0: $\alpha=0$	H_1: $\alpha\neq 0$
$\hat{\mu}$	-0.150	-0.150
$\hat{\alpha}$	0.0	1.566
$\hat{\beta}$	1.874	1.977
$\hat{\mu}_1$	5.586	5.000
$\hat{\mu}_2$	7.414	8.000
$\hat{\sigma}_1^2$	11.830	10.986
$\hat{\sigma}_2^2$	15.190	14.055
$\hat{\sigma}_\varepsilon^2$	2.555	2.084
LISREL χ^2	4.29	0.22
(corrected)	df=3	df=2

Test of H_0: χ^2=4.07, df=1

Table 9

ML-estimates for σ_ε^2 and σ_δ^2 and χ^2 value of likelihoodratio test for the hypothesis $\sigma_\varepsilon^2=\sigma_\delta^2$ as a function of true values of σ_δ^2, β_1 and β_2 and $\sigma_\varepsilon^2=2.0$

	$\beta_1=1.9,\ \beta_2=2.1$			$\beta_1=2.1,\ \beta_2=1.9$		
σ_δ^2	$\hat{\sigma}_\varepsilon^2$	$\hat{\sigma}_\delta^2$	χ^2	$\hat{\sigma}_\varepsilon^2$	$\hat{\sigma}_\delta^2$	χ^2
2	0.0	2.499	0.09	8.879	0.0	0.160
3	0.0	3.499	0.10	11.605	0.0	0.122
5	0.0	5.499	0.11	14.436	0.804	0.069
7	0.0	7.499	0.13	15.059	2.515	0.041
9	0.0	9.499	0.14	15.487	4.312	0.026
11	0.0	11.499	0.15	15.801	6.159	0.016
13	0.0	13.499	0.17	16.043	8.039	0.009
15	0.0	15.499	0.19	16.235	9.943	0.005
17	0.0	17.499	0.20	16.391	11.864	0.003

Table 10

Analysis of Lord's numerical example according to structural equation model with different reliabilities in each group (see text for explanation)

	Model I	Model II	Model III
$\hat{\mu}$	1.466	1.466	1.466
$\hat{\alpha}$	0.069	0.068	0.069
$\hat{\beta}$	0.243	0.241	0.242
$\hat{\mu}_1$	4.070	4.070	4.070
$\hat{\mu}_2$	5.340	5.340	5.340
$\hat{\sigma}^2_1$	4.326	4.196	4.282
$\hat{\sigma}^2_2$	2.580	2.792	2.662
$\hat{\sigma}^2_{\varepsilon 1}$	0.273	0.275	0.271
$\hat{\sigma}^2_{\varepsilon 2}$	0.249	0.250	0.256
$\hat{\sigma}^2_{\delta 1}$	1.058	1.049	1.033
$\hat{\sigma}^2_{\delta 2}$	1.058	1.046	1.084
χ^2 test for α=0	2.67(*)	2.64	2.92(*)

Note:

(*) uncorrected value from LISREL program

FIGURE CAPTIONS

Figure 1 Artificial data used to demonstrate that an analysis of covariance is not the same as an analysis of variance on the residual scores.

Figure 2 Artificial data generated with $\beta=0$ as an illustration of the effect of dependence between treatment and covariate.

Figure 3 Illustration of the principle of orthogonal regression.

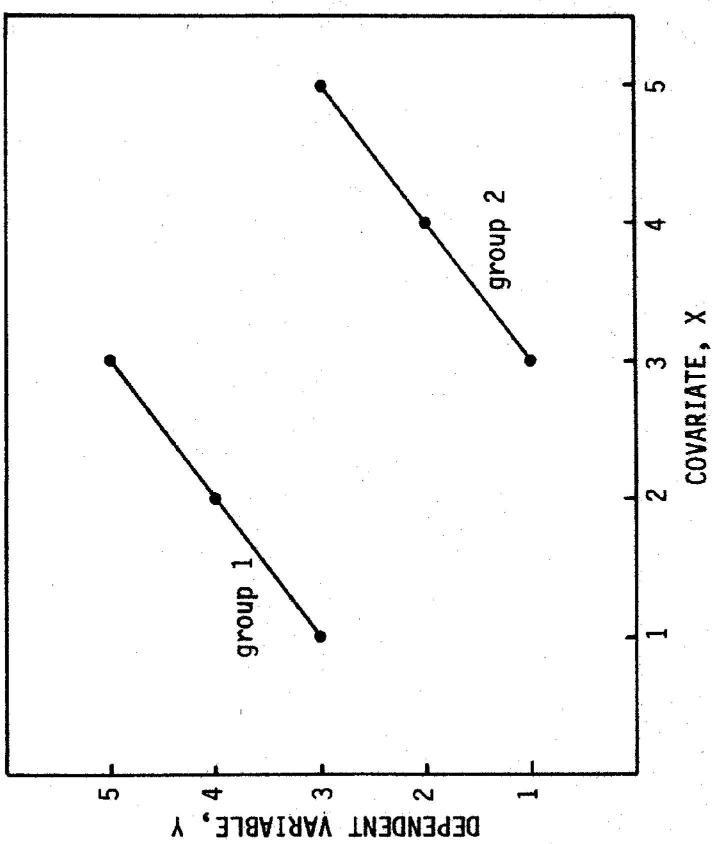

Figure 1

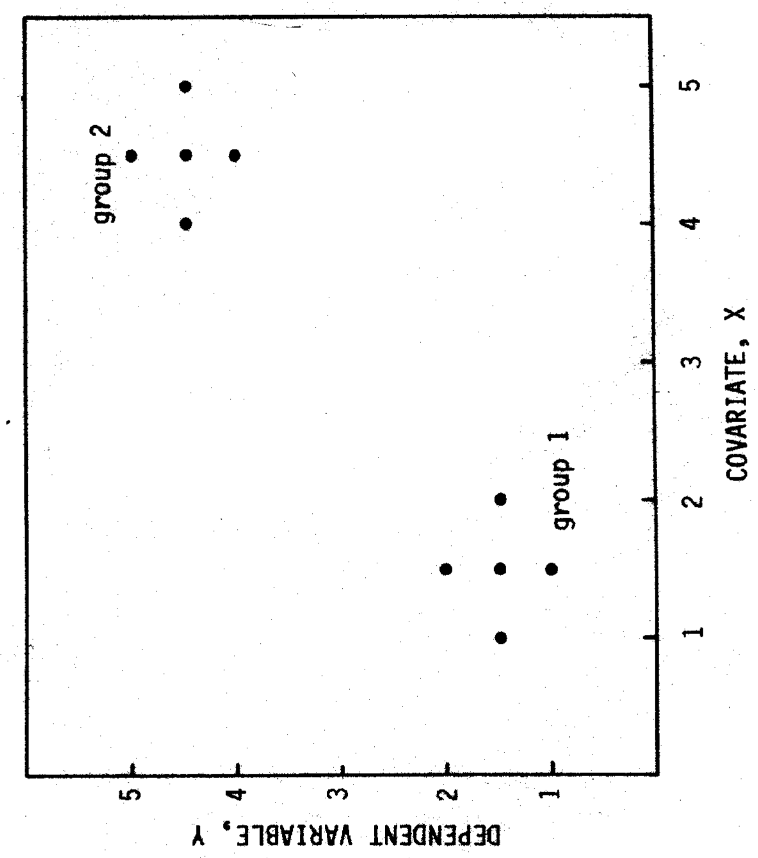
group 1
group 2
COVARIATE, X
DEPENDENT VARIABLE, Y
1 2 3 4 5
1 2 3 4 5

Figure 2

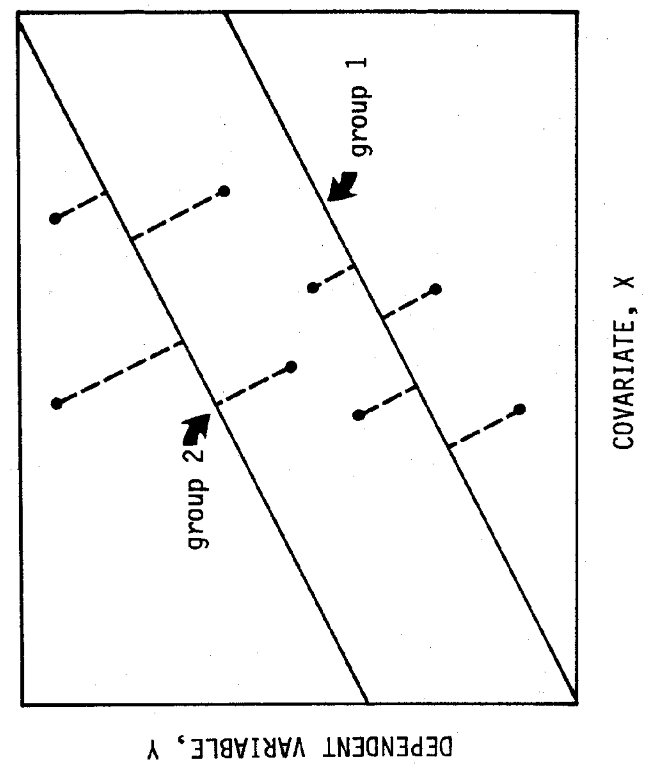

Figure 3