DOCUMENT RESUME

ED 336 412 TM 017 178

AUTHOR Blair, R. Clifford; Sawilowsky, Shlomo S.
TITLE Confounding Covariates in Nonrandomized Studies.
PUB DATE 91
NOTE 18p.; Paper presented at the Annual Meeting of the Midwestern Educational Research Association (12th, Chicago, IL, October 17-19, 1991).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Analysis of Covariance; Equations (Mathematics); Mathematical Models; *Research Design; Research Problems
IDENTIFIERS *Confounding Variables; *Nonrandomized Design

ABSTRACT

Analysis of covariance (ANCOVA) is a data analysis method that is often used to control extraneous sources of variation in non-equivalent group designs. It is commonly believed that as long as the covariate is highly correlated with the dependent variable there is nothing to lose in using ANCOVA, even in non-randomized studies. This paper examines some of the conditions that lead to successful and unsuccessful criterion source adjustments, and demonstrates that under certain circumstances, ANCOVA may perform in a manner that is antithetical to its intended purpose. Several hypothetical data sets were constructed, each with 70 observations, to illustrate two examples of appropriate ANCOVA use and two examples of inappropriate results. ANCOVA may serve to introduce confounding variables into the analysis when covariates represent differences between groups that are unrelated to outcome measures. Two tables present information from the analyses. A 26-item list of references is included. (Author/SLD)

CONFOUNDING COVARIATES IN NONRANDOMIZED STUDIES

R. Clifford Blair

Department of Pediatrics,
College of Medicine
and
Department of Epidemiology
and Biostatistics,
College of Public Health
University of South Florida

Shlomo S. Sawilowsky

Educational Evaluation
and Research
College of Education
Wayne State University

Paper presented at the annual meeting of the MidWestern Educational Research Association, Chicago, IL, 1991.

TM 017178

ABSTRACT

Analysis of Covariance (ANCOVA) is a data analysis method that is often employed to control extraneous sources of variation in non-equivalent group designs. It is commonly believed that so long as the covariate is highly correlated with the dependent variable there is nothing to lose in employing ANCOVA, even in non-randomized studies. This paper examines some of the conditions that lead to successful and unsuccessful criterion source adjustments, and demonstrates that under certain circumstances, ANCOVA may perform in a manner antithetical to its intended purpose.

CONFOUNDING COVARIATES IN NONRANDOMIZED STUDIES

INTRODUCTION

The analysis of covariance (ANCOVA), as employed in educational research practice, is routinely used for one or both of two purposes. The first of these purposes is to attain an increase in the power of a statistical test. As an example, a researcher might randomly assign students to various treatment groups with subsequent outcome measures being evaluated by means of an analysis of variance (ANOVA). In the event that ancillary information pertaining to the students is available in the form of measures that (a) correlate with the outcome measures and (b) do not reflect treatment effects, then the power of the ANOVA test may be augmented by the introduction of a covariate, similar to the use of "blocks" in different design contexts. Under random assignment covariate scores for students in various treatment groups are sampled from identical populations.

Covariates are also commonly employed to adjust criterion measures so as to ameliorate group differences that are unrelated to treatments. That is, the second use of ANCOVA is to control an extraneous variable. For example, an educational researcher may be unable to randomly assign students to treatments and subsequently become aware of differences between the groups in terms of intellectual ability. In the event that outcome measures are related to intellectual ability (e.g., scores on a reading test), then the researcher might employ IQ scores as a covariate in the model in order to control for group differences in intellectual ability. Unlike the first use of ANCOVA, in this instance groups do differ on the covariate measure and it is precisely because of this difference that the covariate is used. For a further discussion on this topic and related issues, see Cochran (1957, 1970), Elashoff (1969), Evans and Anastasio (1968), Fisher (1932), Harris (1963), Levin and Subkoviak (1977), Linn (1981), and Lord (1960).

The use of covariates is not a substitute for random assignment of experimental units, but its proliferation is apparently promoted by the belief that covariates in non-equivalent group designs can only improve the level of precision in the data analysis. The argument contends that, at best, in nonrandom assignment ANCOVA will control at least some of the sources of extraneous variation, and at worst, will not be biased from traditional ANOVA results. This in turn can only lead to greater confidence in the validity of results than would have been realized had the covariate not been employed in the model. While it has been clearly pointed out that, "ANCOVA provides the appropriate adjustment only under a very limited set of conditions" (Porter and Raudenbush, 1987, p. 390) and randomization is a primary condition, the propensity of usage in nonrandomized studies in education suggests that it is not commonly known that under certain circumstances *the use of ANCOVA will result in the introduction of extraneous influences into the analysis*. Not only does the ANCOVA fail to provide precision in this situation, but it will operate in a manner antithetical to its purpose.

PURPOSE OF THIS PAPER

The purpose of this paper is to explicate and focus attention on the problem of unsuccessful adjustments made in data analysis through misuse of covariates. It will be helpful to review some basic assumptions of psychometric test score theory. This background will form the basis for the subsequent discussion. Then, examples of successful and unsuccessful criterion score adjustments in situations where the null hypothesis of no treatment effect, as well as in situations where the null hypothesis is false, will be presented.

CLASSICAL TEST SCORE THEORY

Classical test score theory conceives of the raw score earned by a student on a given test as a function of two basic components, expressed as:

$$x_i = t_i + e_i \qquad (1)$$

Because of its error component (e_i), x_i is often referred to as the observed or fallible score earned by the ith student on the test. In contrast, t_i refers to the ith student's true score and represents the student's actual ability in the subject matter. If the student's ability does not change, then this quantity will remain stable from test to test so long as the tests measure the same subject matter on the same scale. The error component represents those random influences that inflate or deflate a test score, but are unrelated to the true score. The e_i are generally taken to have a mean of zero (Gulliksen, 1950).

There are other elements of test scores that fit neither of the above categories. These elements do not represent the student's ability in the subject matter, and yet are stable rather than random elements. Classical measurement theory expressed in (1) can therefore be expanded to:

$$x_i = t_i + c_{1i} + c_{2i} + \ldots + c_{ki} + e_i \qquad (2)$$

where c_{1i}, c_{2i}, ... c_{ki} represent the various components of this type that contribute to the fallible score of the ith student. (Some simplification of the theory has been made here.) A better understanding of a possible c_i can be gained by briefly examining a common example called "test-wiseness".

Test-wiseness has been defined as a student's ability to use the characteristics or format of a test to obtain a higher score (Millman, Bishop, and Ebel, 1965). It is independent of the student's knowledge of the subject matter. Test-wiseness can be seen in students who have had extensive experience with a particular type of test. For example, students taking multiple-choice tests soon learn to look for clues that will allow them to eliminate otherwise attractive foils. In

this situation a test score reflects not only a student's level of knowledge of the subject matter, but test taking skills as well. Other examples of score components that are unrelated to the intended object of a particular measurement process are discussed by Bajtelsmit (1975), Diamond and Evans (1972), Hall, Follman, and Fisher (1987), Kirkland and Hollandsworth (1980), Millman (1966), Rowley (1974), Sarnacki (1979), Stanley (1971), and Wigdor and Garner (1982).

EXAMPLES OF COVARIATE ADJUSTMENTS

The following examples of a single-factor ANCOVA are based on the linear model

$$y_{ij} = \mu + \alpha_j + \beta (x_{ij} - x) + e_{ij} \qquad (3)$$

$i = 1, \ldots, n$

$j = 1, \ldots, c$

where μ is the grand mean, α is the effect due to the treatment, β is the regression of y on x, and e refers to the error component.

Example 1: Successful Adjustment When H_0 Is True

Suppose that an educational researcher has designed a study to determine which of two methods of teaching arithmetic skills produces better results as measured by scores on a standardized arithmetic test. Practical considerations require the researcher to use two previously formed classes as a control (Class A) and experimental group (Class B). A pretest-posttest design is used with an intervening period of instruction being provided to the two classes. Suppose further that because of the manner in which the two classes were originally formed, Class A has a higher mean arithmetic ability than does Class B. Noting this, the researcher decides to test for treat-

ment effects through the use of an ANCOVA model in which the pretest arithmetic test score is to be used as the covariate. The attempt here is to control for the initial difference in arithmetic ability between the two classes.

Several hypothetical data sets were constructed with characteristics shown in Table 1 to illustrate this as well as the following analyses, . The information in this table indicates that the dependent and covariate measures for the two classes are made up of true and error score components only. The table also shows that true scores for Class A were sampled from a population with a mean of one, while those for Class B came from a population of a mean of zero. The column headed c (for other components) reflects the fact that the scores contained no other components. The last column shows that no treatment effect (in the form of a constant to produce a shift in location parameter) was added to the scores of either group. The e_{1i} and e_{2i} for all examples were sampled from distributions with a mean of zero and variance of one. All data sets were composed of 70 observations (35 per class) with random variates being sampled from normal populations with means and variances as shown in the table.

In order to further simply the discussion, Table 2 shows the various models based on (3) that were used in this example and subsequent analyses. In keeping with common practice, least squares regression models (with intercepts) were used for all analyses. In Table 2, dep and cov represent, respectively, the dependent and covariate variables. In the models, grp is a dummy vector (1 or 0) representing group membership, and int is the product of cov and grp.

Returning to the example, a test of β_1 in model (a) and β_3 in (c) produces (rounded) p values of .000 and .999, respectively. This indicates a relationship between the covariate and dependent variables and leaves as tenable the hypothesis of homogeneous regressions for the two classes. Of primary interest is the test of β_2 in (b) which yields p = .682, thereby leaving

tenable the hypothesis of no treatment effect. Had the covariate measures not been available, a test of β_1 in (d) (i.e., an independent samples t test) would have produced a significant result (p = .000), thereby leading to the erroneous conclusion of a difference between the effectiveness of instruction methods. In Example 1, therefore, the use of the covariate led to a correct assessment of no treatment effect, while an analysis performed without the covariate led to the opposite and erroneous conclusion.

Example 2: Successful Adjustment When H_0 Is False

The data used in this example are the same as those from Example 1 with the exception that a treatment effect (modeled, with a constant of 1.0, as a shift in location parameter) was added to the dependent variable scores of the members of Class B. Thus, the instructional method used in Class B was superior to that used in Class A. The ANCOVA analysis of this data reaches the same conclusions as in Example 1 regarding the preliminary tests, but produces p = .000 for the test of β_2 in (b). Thus, the ANCOVA analysis correctly detected the presence of a treatment effect. On the other hand, a test of β_1 in (d) (i.e., where the covariate is not taken into account) results in a nonsignificant p value of .857, thereby failing to detect the treatment effect.

In the above examples, the ANCOVA models performed appropriately. In the examples that follow, however, results generated by these methods were inappropriate.

Example 3: Unsuccessful Adjustment When H_0 Is True

Suppose in the situation outlined in Example 1 that no pretest scores were available for use as a covariate. Rather, scores from a *different* mathematics test (call this Test D) were available. Unlike the test used to measure the outcome variable, Test D does not assess arithmetic

skills by simple presentation of calculational problems, but instead uses word problems as the medium of presentation. Both tests measure arithmetic ability, but they do so through different forms of problem presentation.

Even though tests of the form represented by Test D measure what they purport to measure (i.e., arithmetic ability in this case), they also reflect a student's reading ability. (In addition, there might be other components not shared by these tests, such as reasoning ability and so forth.) This arises because problems must be read and understood before calculations can be carried out. In this example it is assumed that the two classes have approximately the same arithmetic skills, but Class A has higher mean reading ability than does Class B. Table 1 reflects this by indicating that reading components (c) of the covariate scores for Class A were sampled from a population with a mean of one and those for Class B had a population mean of zero. In this example reading ability is taken to be independent of arithmetic skill and is extraneous to the experiment.

The test of β_1 in (a) and β_3 in (c) results in p values of .000 and .949, respectively. The test of β_2 in (b) yields a p value of .027, leading to the incorrect conclusion that a treatment effect is present in the data. In this example, the correct conclusion of no treatment effect is obtained by the test on β_1 in (d) which gives p = .857.

Unlike the first two examples, use of a covariate in this case led to an incorrect conclusion, while analysis performed without the covariate resulted in a correct conclusion. The explanation is straightforward; the covariate adjustment was made on the basis of the difference in reading levels of the two classes. But, reading level is unrelated to the dependent variable in

this example and is therefore irrelevant to the study. The important point here is that rather than limiting or reducing the influence of some extraneous variable, the covariate acted to introduce a confounding variable into the data analysis.

Example 4: Unsuccessful Adjustment When H_0 Is False

In this example a constant (treatment effect) of .5 was added to the dependent variable scores of the students in Class B with scores being the same as those used in Example 3. Conclusions reached on the two preliminary tests are the same as those reached in the previous three examples. The test of β_2 in (b) is nonsignificant with $p = .814$, while that of β_1 in (d) yields $p = .021$. In this case, the model with a covariate failed to detect the presence of the treatment effect. It was detected, however, when the covariate was removed from the analysis.

COMMENTS

Rather than excluding the influence of confounding variables, ANCOVA may serve to introduce confounding variables into the analysis. This circumstance may occur when covariates reflect differences between groups that are unrelated to outcome measures. Educational researchers should be particularly aware of this problem when covariate and dependent variables are measured on different scales or when these measures are obtained under different sets of conditions. An example of the former situation occurs when one measure is obtained by observing students at some task, but the other is obtained through administration of a test that assesses knowledge of how the task is performed. The latter situation may occur when, for example, one measure is timed but the other is not. In this case one measure reflects not only a student's ability to perform, but also the rapidity with which the student performs.

The possibility of a covariate measure bringing about a confounding of results may be exacerbated by implementing techniques from the growing body of literature on the effect on ANCOVA when the covariate measure is fallible. This is the so-called "fallible covariable problem in which the reliability of the instrument from which the data were collected on the covariate is less than 1.00. (See, e.g., Carroll, Gallo, & Gleser, 1985; DeGracie & Fuller, 1972; Lord, 1960; Raaijmakers & Pieters, 1987; Rogers and Hopkins, in press, 1990; and Stroud, 1972.) Ironically, making adjustments to take into account reliability of an inappropriate covariate measure serves to increase the confounding effect of the covariate.

Finally, we note that educational researchers often exercise great care in collecting and scrutinizing dependent measures, but may fail to maintain the same level of care when dealing with covariate measures. This seems to stem from the mistaken belief that appropriate adjustments will be made whenever groups differ on a covariate that is highly correlated with the dependent measure. This paper has demonstrated otherwise.

REFERENCES

Bajtelsmit, J. W. (1975). Development and validation of an adult measure of secondary cue-using strategies on objective examinations: The test of obscure knowledge (TOOK). Annual Meeting of the Northeastern Educational Research Association, Ellenville, New York.

Carroll, R. J., Gallo, P., & Gleser, L. J. (1985). Comparison of least squares and errors-in-variables regression with special reference to randomized analysis of covariance. Journal of the American Statistical Association, 80, 929-932.

Cochran, W. G. (1957). Analysis of covariance: Its nature and use. Biometrics, 13 261-281.

Cochran, W. G. (1970). Some effects of errors of measurement on multiple correlation. Journal of the American Statistical Association, 65, 22-34.

DeGracie, J. S., & Fuller, W. A. (1972). Estimation of the slope and analysis of covariance when the concomitant variable is measured with error. Journal of the American Statistical Association, 67, 930-937.

Diamond, J. J., & Evans, W. J. (1972). An investigation of the cognitive correlates of test-wiseness. Journal of Educational Measurement, 9, 145-150.

Elashoff, J. D. (1969). Analysis of covariance: A delicate instrument. American Educational Research Journal, 6, 383-401.

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

Fisher, R. A. (1932). Statistical methods for research workers. (4th ed.). Edinburgh: Oliver and Boyd.

Gulliksen, H. (1950). Theory of mental tests. N.Y.: John Wiley & Sons.

Hall, B. W., Follman, J. C., & Fisher, S. K. (1987). An examination for affective correlates of test-wiseness. Annual Meeting of the National Council on Measurement in Education, Washington, D. C.

Harris, C. W. (1963). Problems in measuring change. Madison: University of Wisconsin.

Kirkland, K., & Hollandsworth, J. G. (1980). Effective test taking: skills-acquisition versus anxiety-reduction techniques. Journal of Consulting and Clinical Psychology, 48, 431-438.

Levin, J. R., & Subkoviak, M. J. (1977). Planning an experiment in the company of measurement errors. Applied Psychological Measurement, 1, 331-338.

Linn, R. L. (1981). Measuring pretest-posttest performance changes. Education evaluation methodology: The state of the art. Baltimore, MD: Johns Hopkins University Press.

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

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

Millman, J., Bishop, C. H., and Ebel, R. L. (1965). An analysis of test wiseness. Educational and Psychological Measurement, 25, 707-726.

Porter, A. C., & Raudenbush, S. W. (1987). Analysis of covariance: Its model and use in psychology research. Journal of Counseling Psychology, 34 383-392.

Rogers, W. T., & Hopkins, K. D. (in press). Power estimates in the presence of a covariate and measurement error. Educational and Psychological Measurement.

Rogers, W. T., & Hopkins, K. D. 1990). Quick power estimates incorporating the joint effects of measurement error and a covariate. Journal of Experimental Education, 57, 86-94.

Rowley, G. L. (1974). Which examinees are favored by the use of multiple choice tests? Journal of Educational Measurement, 11, 15-23.

Sarnacki, R. E. (1979). An examination of test-wiseness in the cognitive test domain. Review of Educational Research, 49, 252-279.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), Educational measurement. Washington, D. C.: American Council on Education.

Stroud, T. W. F. (1972). Comparing conditional means and variances in a regression model with measurement errors of known variances. Journal of the American Statistical Association, 67, 407-414.

Wigdor, A. K., & Garner, W. R. (1982). Ability testing: uses, consequences, and controversies. Washington, D. C.: National Academy Press.

TABLE 1

Characteristics of Data Sets Used In Examples 1-4.

		Score		μ; σ		
Example	Class	dep	cov	t	c	tr
1	A	t_i+e_{1i}	t_i+e_{2i}	1;1	-	0.0
	B	t_i+e_{1i}	t_i+e_{2i}	0;1	-	0.0
2	A	t_i+e_{1i}	t_i+e_{2i}	1;1	-	0.0
	B	t_i+tr+e_{1i}	t_i+e_{2i}	0;1	-	1.0
3	A	t_i+e_{1i}	$t_i+c_i+e_{2i}$	0;1	1;1	0.0
	B	t_i+e_{1i}	$t_i+c_i+e_{2i}$	0;1	0;1	0.0
4	A	t_i+e_{1i}	$t_i+c_i+e_{2i}$	0;1	1;1	0.0
	B	t_i+tr+e_{1i}	$t_i+c_i+e_{2i}$	0;1	0;1	0.5

NOTE: dep = dependent variable, cov = covariate, $\mu;\sigma(t)$ = mean and standard deviation of true score, $\mu;\sigma(c)$ = mean and standard deviation of component score, tr = treatment effect.

Table 2

Least Squares Models Used In Example Analyses.

Model

Designation	Model
(a)	$dep = \beta_0 + \beta_1 cov$
(b)	$dep = \beta_0 + \beta_1 cov + \beta_2 grp$
(c)	$dep = \beta_0 + \beta_1 cov + \beta_2 grp + \beta_3 int$
(d)	$dep = \beta_0 + \beta_1 grp$

NOTE: dep = dependent variable, cov = covariate, grp = group membership, int = product of covariate and group membership.