HOW ANALYSIS OF COVARIANCE CAN YIELD MISLEADING RESULTS IN EDUCATIONAL EXPERIMENTS - A MONTE CARLO STUDY

James E. McLean, University of Alabama
William B. Ware and James T. McClave, University of Florida

The problem of comparing two groups on the basis of change in mental test scores has confronted educational researchers for at least half a century. The widespread funding of compensatory education projects during the last decade has focused attention on a complex dimension of the problem--comparing two groups at different beginning levels. Many of the recent programs featuring innovative approaches to compensatory education are available only for the most needy students. For evaluation purposes, a comparison group is used which is sampled from the general population of students.

Until recently, the standard analysis recommended for such situations has been analysis of covariance (ANOCOV) using the pretest scores as covariants (e.g., Cochran and Cox, 1957; Winer, 1962; Campbell and Stanley, 1963; Kerlinger, 1973). Lord (1960, 1967, and 1969), Porter (1967), and Campbell and Erlebacher (1970) have warned of possible misleading results using ANOCOV when the covariate is fallible (i.e., not measured with perfect reliability). Campbell and Erlebacher (1970) state that they are reasonably certain that such a methodological error did occur in the Westinghouse/Ohio University study and possibly in other studies which showed no effects or even harmful effects from Head Start programs.

The purpose of this paper is to examine the robustness of the ANOCOV procedure with respect to the violation of certain assumptions which are likely to be violated in educational experiments. The three assumptions under study are the random assignment of subjects to groups, the covariates are measured with perfect reliability, and homogeneity of regression. In educational experiments, the random assignment of subjects to groups usually manifests itself in the inequality of pretest means because of sampling intact groups.

Method

The ANOCOV was examined under 48 sets of conditions on the basis of computer generated data. Six conditions of realibility, two levels of sample size, two levels of gain, and two different sets of pretest means were used. Both equal and unequal reliabilities were used in the study. On the basis of 2000 sets of computer generated data for each set of conditions, empirical alpha and power values (where appropriate) were calculated. These were compared to what the analytically calculated values would have been if all of the assumptions were met. The empirically derived alpha values and powers were then used as criterion variables in two factorial experiments to further examine the relationships among the controlled factors. Subsequent a posteriori analyses were performed where indicated.

A traditional measurement model (Gulliksen, 1950; Lord and Novick, 1968; O'Connor, 1972) was used in the generation of the scores. For each group, the generated scores were based on the models:

$$(1) \quad X = T + E_1$$

and

$$(2) \quad Y = T + G + E_2,$$

where X = observed pretest score,
Y = observed posttest score,
T = true pretest score,
G = true gain,
E_1= random measurement error in pretest score,
E_2= random measurement error in posttest score.

Based on the definition of reliability,

$$(3) \quad \rho_{XX} = \frac{\sigma_X^2 - \sigma_E^2}{\sigma_X^2}$$

and the standard assumptions for equations (2) and (3), (Lord, 1956), data can be generated with specified reliabilities by appropriately varying the variance components (Neel, 1970, p. 20-21; McLean, 1974, p. 17-18). For example, by setting σ_X^2 a priori to be 100, pretest scores with the desired reliability can be generated by choosing σ_T^2 according to the following formula:

$$(4) \quad \sigma_T^2 = 100\rho_{XX}.$$

The posttest reliabilities can be set in a similar manner.

Recall that one of the assumptions necessary for analysis of covariance is that the regression slopes of the dependent variable on the covariate must be equal for each treatment group. Let $\beta_{Y \cdot X}$ denote the regression slope for one treatment group.

Then,

$$(5) \quad \beta_{X \cdot Y} = \rho_{XY} \cdot \frac{\sigma_Y}{\sigma_X}, \quad \text{(Ferguson, 1971, p. 113).}$$

implies

$$(6) \quad \beta_{X \cdot Y} = \frac{\sigma_{XY}}{\sigma_T^2 + \sigma_{E_1}^2}$$

The covariance between X and Y is equal to the variance of T since it has been assumed that all the components of the pretest and posttest are independent except T with itself. Therefore

$$(7) \quad \beta_{Y \cdot X} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_{E_1}^2},$$

hence,

$$(8) \quad \beta_{Y \cdot X} = \rho_{XX}$$

by equation (3). Thus, the slope, Y on X, of any group is equal to the reliability of its pretest. If the reliabilities of the pretests were the same for both groups, the assumptions concerning slopes would be satisfied. Hence, choosing different pretest reliabilities for each group, as would likely occur in an educational experiment, would result in a violation of the homogeneity of regression assumption.

The true score mean was set a priori at 100 when both the experimental control groups have equal means. The situation in which the experimental group has a lower mean was also analyzed. In this case, the true score mean of the experimented group was set at 80. These values have also been empirically chosen based on Project Follow Through data. Based on the assumptions,

$$(9) \quad E(X) = E(T+E_1) = \mu_T,$$

hence the mean of the observed pretest scores is equal to the mean of the true scores. Also

$$(10) \quad E(Y) = E(T+G+E_2) = \mu_T+\mu_G.$$

Thus, the mean of the observed posttest scores is equal to the sum of the means of the true scores and the gain scores.

Clearly in the case of the control group and in both groups where no gain was used, the mean gain was zero. The selected value of μ_G for the gain situation was based on power considerations, that is, μ_G was chosen such that the power of an $\underline{F}$ test for ANOCOV was .50.

A linear model representation of the ANOCOV for two groups is

$$(11) \quad Y = \beta_0+\beta_1 X+\beta_2 W+\varepsilon$$

where X is the pretest score (covariate), Y is the posttest score, and W is a dummy variable designating group membership (W=1 if experimental group, 0 if control group). Testing the hypothesis that β_2 is equal to 0 in equation (16) is equivalent to the $\underline{F}$ test for treatments in the ANOCOV procedure. It can be shown that β_2 in equation (11) is equivalent to the mean gain, μ_G. The mean of the posttest for the experimental group is

$$(12) \quad E(Y_G) = \beta_0+\beta_1\mu_{X_G}+\beta_2$$

where Y_G is a posttest score for the experimental (gain) group and μ_{X_G} is the mean pretest score for the experimental group. Likewise, the mean of the posttest for the control (no-gain) group is

$$(13) \quad E(Y_{NG}) = \beta_0+\beta_1\mu_{X_{NG}}$$

where Y_{NG} is a pretest score for the control group, $\mu_{X_{NG}}$ is the mean pretest score for the control group. But it has been assumed that μ_{X_G} and $\mu_{X_{NG}}$ are equal. Furthermore, the experimental group has mean gain, μ_G and the control group has mean gain zero, thus

$$(14) \quad \mu_G = D(Y_G)-E(Y_{NG}) = \beta_2.$$

Hence, choosing the value of β_2 that yields a power of .50 is equivalent to choosing a value of μ_G to produce a power of .50 in the ANOCOV procedure.

This power can be obtained from the following probability statement:

$$(15) \quad Pr[\underline{t}^* > \underline{t}] = .50$$

where $\underline{t}^*$ is a noncentral $\underline{t}$ statistic. This expression can be approximated by the substitution of a $\underline{z}$ statistic for $\underline{t}^*$.

$$(16) \quad Pr \; \frac{\hat{\beta}_2-\beta_2}{\sigma_{\hat{\beta}_2}} = .50.$$

Thus a value of β_2 can be chosen such that the substituted $\underline{z}$ statistic is equal to $\underline{t}_{.025}$ with the appropriate degrees of freedom. This value of β_2 is the value of the average gain, μ_G, such that the power of analysis of covariance is .50 under the conditions of perfect reliability. In order to find this value of μ_G, a numerical expression for $\sigma^2_{\hat{\beta}_2}$ is needed.

The variance of $\hat{\beta}_2$, $\sigma^2_{\hat{\beta}_2}$, can be approximated in the following manner. Karmel and Polasek (1970, p. 245) state that

$$(17) \quad \sigma^2_{\hat{\beta}_2} = \sigma^2_Y \frac{\Sigma(X-\bar{X})^2}{[\Sigma(X-\bar{X})^2][\Sigma(W-\bar{W})^2]-[\Sigma(X-\bar{X})(W-\bar{W})]^2}$$

Dividing both the numerator and denominator by N^2 and substituting population variances for sample variances, $\sigma^2_{\hat{\beta}}$ is approximately equal to the following:

$$(18) \quad \sigma^2_{\hat{\beta}_2} \approx \frac{\sigma^2_Y}{N} \frac{\sigma^2_X}{\sigma^2_X \sigma^2_W - (\sigma_{XW})^2}$$

The quantity, σ_{XW}, the covariance between X and W has been assumed equal to zero thus, in equation (23), the σ^2_X's divide out. Hence

$$(19) \quad \sigma^2_{\hat{\beta}_2} \approx \frac{\sigma^2_Y}{N} \cdot \frac{1}{\sigma^2_W}$$

Under the conditions assumed for the model, the variance of Y, σ^2_Y, is equal to 4. The variance of W, σ^2_W can easily be computed to be 125. Thus, by substitution, $\sigma^2_{\hat{\beta}_2}$ equals .80 when N is equal to 20 and $\sigma^2_{\hat{\beta}_2}$ equals .80 when N is equal to 200.

Hence, for N = 20, μ_G can be found by the following expression:

(20) $\mu_{G_{20}} = \pm t_{.025,197}\sigma_{\hat{\beta}_2} = 2.11\sqrt{.80} = 1.88.$

Likewise, for N=200, μ_G can be found by the following:

(21) $\mu_{G_{200}} = \pm t_{.025,197}\sigma_{\hat{\beta}_2} = 1.97\sqrt{.08} = .56.$

The approximated values of μ_G were tested using Monte Carlo generated variables and found to indeed produce a power of .50. The approximated values of μ_G under each set of conditions are shown in Table 2.

The study required the use of computer generated normally distributed random variables with specified means and variances. Two thousand sets of variables were generated for each of the 48 sets of conditions. Muller (1959) identified and compared six methods of generating normal deviates on the computer. A method described by Box and Muller (1958) was judged most attractive from a mathematical standpoint. According to Muller (1959, p. 379), "Mathematically this approach has the attractive advantage that the transformation for going from uniform deviates to normal deviates is exact." This method was endorsed by Marsaglia and Bray (1964), who modified the algorithm to reduce central processing computer time without altering its accuracy.

The method first requires the generation of two independent uniform random variables, U_1 and U_2, over the interval (-1, 1). The variables

$$Z_1 = U_1[-2\ln(U_1^2+U_2^2) / (U_1^2 + U_2^2)]^{1/2}$$

$$Z_2 = U_2[-2\ln(U_1^2 + U_2^2) / (U_1^2 + U_2^2)]^{1/2}$$

will be two independent random variables from the same normal distribution with mean zero and unit variance. The variables were then transformed to have the desired means and variances.

After the 2000 sets of data were generated and analyzed using ANOCOV, the sample proportion of rejections were noted. For the cases in which the gain in both groups was the same, the proportions are, by definition, alpha values, i.e., sample probabilities of type I errors. For the cases where the experimental group had the larger gain, these values are, by definition empirically generated powers. The empirical alpha values and powers were then compared with analytical values. These empirically derived alpha values and powers were further used as dependent variables in factorial experiments to further examine the factors affecting the outcomes of ANOCOV.

Results and Discussion

The empirically derived alpha values and the conditions under which they were generated are found in Table 1. The true alpha value for each case should be .05 if all assumptions were met. Note that in every case where the pretest mean of the experimental group is less than that of the control group, there is a significantly higher proportion of rejection than would be expected by chance. This is a likely result of using ANOCOV for comparing groups at different beginning levels. This proportion of rejections becomes more pronounced as reliability is reduced. Also note that the inequality of pretest reliabilities (thus the nonhomogeneity of regression slopes) does not seem to have an effect.

The empirically derived powers and the conditions under which they were generated (including the gain in the experimental group) are presented in Table 2. The time power should be .50 in every case if all of the assumptions were met. The generated powers differed significantly from this in every case. Sometimes the empirical power was significantly below .50 and sometimes it was significantly above .50. Again, it deviates farther from the time power as the reliability decreases. The equality of pretest reliability again has little effect. In the case where the reliabilities were 1.00 (not shown) and the pretest means were different, the empirically generated alpha value was not significantly different from .50. The most dramatic result is that where the experimental group actually experienced a gain and the control group did not and the pretest mean of the experimental group was less than that of the control group, the adjusted posttest means indicated that the control group was better.

The factorial experiments using the empirically generated alpha values and powers did not provide any further information but did support the results previously stated. That is, the equality of pretest means and level of reliability are the most critical factors and they do interact. Also, the homogeneity of regression is not critical in these situations. The analysis of variance summary tables are not presented here due to a lack of space.

Summary

This study was designed to evaluate the effects of violating the assumptions of ANOCOV in educational experiments. The first assumption, subjects randomly assigned to treatment groups, seems to be the most crucial assumption. Violating this assumption further confuses the results when it is combined with a violation of the assumption, measuring the covariate with perfect reliability. Based on this study, a violation of the reliability assumption most radically affects the power of the ANOCOV procedure. Violating the homogeneity of regression assumption seems to have little effect on the outcome of the experiment when the violation is only moderate as used in this paper.

The results of this study point to the recommendation that ANOCOV be approached with caution when the reliabilities are below .90 and nonrandom samples are used.

TABLE 1

FRACTION OF SIGNIFICANT F's WHERE THE MEAN GAIN WAS ZERO IN BOTH GROUPS

Reliability of Treatment Group Scores	Reliability of Comparison Group Scores	Sample Size of Each Group	Pretest Mean of Treatment Group[a]	Fraction of Significant F's for Analysis of Covariance[b]
.90	.90	10	100	.048
.90	.90	10	80	.089*
.90	.90	100	100	.050
.90	.90	100	80	.540*
.70	.70	10	100	.048
.70	.70	10	80	.232*
.70	.70	100	100	.050
.70	.70	100	80	.980*
.50	.50	10	100	.056
.50	.50	10	80	.374*
.50	.50	100	100	.044
.50	.50	100	80	.999*
.76	.90	10	100	.056
.76	.90	10	80	.134*
.76	.90	100	100	.055
.76	.90	100	80	.794*
.60	.70	10	100	.050
.60	.70	10	80	.259*
.60	.70	100	100	.046
.60	.70	100	80	.994*
.42	.50	10	100	.054
.42	.50	10	80	.400*
.42	.50	100	100	.042
.42	.50	100	80	1.000*

[a]Pretest mean of comparison group is 100 in all cases.
[b]Monte Carlo derived alpha values.
*Significantly different from analytical alpha (.05) at .01 level.

TABLE 2

FRACTION OF SIGNIFICANT F's WHERE THE MEAN GAIN WAS POSITIVE ONLY IN THE TREATMENT GROUP

Reliability of Treatment Group Scores	Reliability of Comparison Group Scores	Sample Size of Each Group	Mean Gain of Treatment Group	Pretest Mean of Treatment Group[a]	Fraction of Significant F's for Analysis Covariance[b]
.90	.90	10	1.88	100	.123*
.90	.90	10	1.88	80	.057*
.90	.90	100	.56	100	.126*
.90	.90	100	.56	80	.318*
.70	.70	10	1.88	100	.074*
.70	.70	10	1.88	80	.130*
.70	.70	100	.56	100	.087*
.70	.70	100	.56	80	.958*
.50	.50	10	1.88	100	.074*
.50	.50	10	1.88	80	.277*
.50	.50	100	.56	100	.073*
.50	.50	100	.56	80	.998*
.76	.90	10	1.88	100	.101*
.76	.90	10	1.88	80	.070*
.76	.90	100	.56	100	.101*
.76	.90	100	.56	80	.668*
.60	.70	10	1.88	100	.074*
.60	.70	10	1.88	80	.154*
.60	.70	100	.56	100	.075*
.60	.70	100	.56	80	.983*
.42	.50	10	1.88	100	.081*
.42	.50	10	1.88	80	.280*
.42	.50	100	.56	100	.063*
.42	.50	100	.56	80	1.000*

[a]Pretest mean of comparison group is 100 in all cases.
[b]Monte Carlo derived powers.
*Significantly different from analytical power (.50) at .01 level.

REFERENCES

Box, G. E. P. and Muller, M. E. A Note on the Generation of Random Normal Deviates. Annals of Mathematical Statistics, 1958, 29, 610-611.

Campbell, D. T. and Erlebacher, A. How Regression Artifacts in Quasi-Experimental Evaluations can Mistakenly Make Compensatory Education Look Harmful, in Disadvantaged Child Vol. 3, Ed. Hellmuch, Jerome. New York: Brunner/Maryel, Inc., 1970.

Campbell, D. T. and Stanley, J. C. Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally and Company, 1963.

Cochran, W. G. and Cox, G. M. Experimental Design, Second Edition. New York: John Wiley and Sons, Inc., 1957.

Ferguson, G. A. Statistical Analysis in Psychology and Education, Third Edition. New York: McGraw-Hill Book Company, 1971.

Glass, G. V., Peckham, P. D., and Sanders, J. R. Consequences of Failure to Meet Assumptions Underlying Analysis of Variance and Covariance, Review of Educational Research, Summer, 1972, 42, 237-288.

Gulliksen, H. Theory of Mental Tests. New York: John Wiley and Sons, Inc., 1950.

Karmel, P. H. and Polasek, M. Applied Statistics for Economists, Third Edition. Bath, Great Britian: Pitman Publishing, 1970.

Kerlinger, F. N. Foundations of Behavioral Research, Second Edition. Holt, Rinehart and Winston, Inc., 1973.

Lord, F. M. A Paradox in the Interpretation of Group Comparisons. Psychological Bulletin, 1967, 68, 304-305.

Lord, F. M. Large-Sample Covariance Analysis When the Control Variable is Fallible, Journal of the American Statistical Association. 1960, 55, 307-321.

Lord, F. M. The Measurement of Growth. Educational and Psychological Measurement, 1956, XVI, 421-437.

Lord, F. M. Statistical Adjustments When Comparing Preexisting Groups. Psychological Bulletin, 1969, 72, 336-337.

Lord, F. M. and Novick, Melvin R. Statistical Theories of Mental Test Scores, Reading, Massachusetts: Addison-Wesley Publishing Company, 1968.

Marsaglia, G. and Bray, T. A. A Convenient Method for Generating Normal Variables, SIAM Review, July, 1964, 6, 260-264.

McLean, J. E. An Empirical Examination of Analysis of Covariance With and Without Porter's Adjustment for a Fallible Covariate. Dissertation, University of Florida, 1974.

Muller, M. E. A Comparison of Methods for Generating Normal Deviates on Digital Computers, Association for Computing Machinery Journal, 1959, 6, 376-383.

Neel, J. H. A Comparative Analysis of Some Measures of Change. Dissertation, University of Florida, 1970.

O'Connor, E. R., Jr., Extending Classical Test Theory to the Measurement of Change. Review of Educational Research, 1972, 42, 73-97.

Porter, A. C. The Effects of Using Fallible Variables in the Analysis of Covariance. Dissertation, University of Wisconsin, 1967.

Winer, B. J. Statistical Principles in Experimental Design. New York: McGraw-Gill Book Company, 1962.