Some Nonstandard Applications of the Analysis of Covariance Model

John T. Pohlmann
Southern Illinois University
Carbondale

This paper illustrates two applications of the ANCOVA model under problematic conditions: Johnson-Neyman significance region analysis and the analysis of a regression discontinuity design. The differences between intact and randomized group designs are discussed in the ANCOVA context. The analyses are demonstrated using the SAS REG program.

The analysis of covariance (ANCOVA) is used when a dependent variable and an antecedent independent variable are measured in multiple groups. The antecedent variable, or covariate, measures a source of variation that is to be statistically controlled. In order to accommodate the covariate, the ANCOVA model posits structural assumptions about the relationship between the dependent variable and the covariate. The simplest ANCOVA model assumes a homogeneous linear relationship between the dependent variable and covariate in each of the design groups. When the structural assumptions of ANCOVA are met, the ANCOVA can yield more powerful tests of significance than the analysis of variance (ANOVA). The power advantage of ANCOVA derives from a reduction of the error variance due to the effects of the covariate.

A number of authors (Elashoff, 1969; Lord, 1969; Mueller, 1990) have cautioned that the ANCOVA, performed with intact groups, does not control for preexisting group differences with the same level of rigor as does a randomized design. These authors argue that statistical controls cannot be considered as equivalent substitutes for randomization. Randomization will equate design populations for differences on the covariate along with any other differences that might exist prior to the experiment. Whereas, statistical adjustments can only be applied to variables that have been measured and only over the ranges of the variables observed in the sample. Statistical controls are also highly dependent on the model's structural assumptions. One can only statistically control for the relationships allowed by the model.

The ANCOVA, however, continues to be used with intact groups because of its convenience. Statistical adjustment is often the only control mechanism available to a researcher, and the ANCOVA may be the best statistical treatment of the data. The ANCOVA can be used under the same conditions that would justify the use of a partial correlation coefficient. The major difference between the ANCOVA and a partial correlation analysis is that the ANCOVA model is used when the independent variable is categorical.

This paper will treat the ANCOVA as a multiple group regression model solution. It is assumed that a researcher has measured a dependent variable and a covariate in each of J groups. The dependent variable is regressed separately on the covariate for each of the J groups. The structural assumptions of linearity and homogeneity of regression are tested. Then, tests of hypotheses about expected values of the dependent variable are demonstrated. The tests illustrated in this paper will not be used to test for group differences generally, but will instead examine group differences on the dependent variable at specific point values of the covariate. A regression discontinuity design and Johnson-Neyman significance region analysis will be used to illustrate this approach. These applications were chosen because they present alternatives that a researcher may use when structural or design problems are encountered.

Linear Models for the ANCOVA

The simplest form of the ANCOVA assumes a homogeneous linear relationship between the dependent variable (Y) and the covariate (X) for each of J groups. The linear model for the simple ANCOVA is given by Winer (1971, p. 757) as:

$$Y_{ij} = \mu + \alpha_j + \beta_w\left(X_{ij} - \mu_x\right) + \varepsilon_{ij} \qquad [1]$$

Y_{ij} and X_{ij} are the measures on the dependent and concommitant variables for case i in group j. The expected values of Y and X are denoted by μ and μ_x respectively. β_w is the within groups' regression coefficient, and is assumed to be the common slope of the regressions of Y on X for all groups. Group j's deviation on

Y is represented by α_j, and is called the treatment effect for group j. The only random term in the model is ε_{ij}, the error term, and is assumed to be NID$(0,\sigma^2)$. All other terms in the model are fixed. Essentially, model [1] fits J parallel regression lines, predicting Y from X for the J groups in the design.

For the purposes of this paper, model [1] will be reparameterized into the following form:

$$Y_{ij} = \alpha_j + \beta_w X_{ij} + \varepsilon_{ij} \qquad [2]$$

Model [2] expresses α_j as the Y-intercept and β_w as the common slope for all groups. Model [2] permits the regression lines to have different intercepts, but only one slope; the regression lines are assumed to be parallel. If heterogeneity of regression is detected, model [2] can be revised to reflect nonparallel regression lines by replacing β_w with β_j. Each group is then allowed to have its own slope parameter in addition to a unique intercept parameter. Rewriting [2] accordingly, yields:

$$Y_{ij} = \alpha_j + \beta_j X_{ij} + \varepsilon_{ij} \qquad [3]$$

In models [1] and [2] group differences on Y could be measured by differences in the α_j values. Since the regression lines were assumed to be parallel, the differences between the α_j could be generalized over the full range of the covariate. In model [3], however, the difference between any two groups on Y depends on the value of X. Specifically, when X = C, the difference between the expected Y values of groups k and l is developed as follows:

$$E(Y_k|X = C) = \alpha_k + \beta_k C,$$

$$E(Y_l|X = C) = \alpha_l + \beta_l C,$$

$$E(Y_k|X = C) - E(Y_l|X = C)$$

$$= (\alpha_k + \beta_k C) - (\alpha_l + \beta_l C) \qquad [4]$$

If one hypothesized that the E(Y) values were equal at C, this condition could be expressed as a statistical hypothesis:

$$H_o: (\alpha_k + \beta_k C) = (\alpha_l + \beta_l C) \qquad [5]$$

In regression parlance, expression [4] is the difference between the predicted Y values for populations k and l at the value C on X. Expression [5] will serve as the null hypothesis for the tests described in this paper.

Applications Of Model [3]

Two applications of model [3] will be presented in this paper: a regression discontinuity analysis (Campbell and Stanley, 1963, p. 61) and Johnson-Neyman significance regions (Pedhazur, 1982, p. 469-472). These applications are interesting since they both represent tests that are performed under what is traditionally thought to be an undesirable situation. The regression discontinuity design represents an extreme case of group differences on the covariate. Group differences on the covariate can confound treatment effects. The Johnson-Neyman technique applies when heterogeneous regressions are observed. Heterogeneous regressions preclude a straightforward analysis of the α_j values, since, as per model [3], the α_j values only assess differences on Y when X=0.

The regression discontinuity design is used to test for effects on a dependent variable when individuals are treated differently, depending on the value of the covariate. The Campbell and Stanley (1968) illustration presents a quasi-experimental design for determining if a scholarship award, given on the basis of performance on a selection test, positively influences academic achievement. The covariate is the selection test and the two design groups are students who received the award and students who did not receive an award. Achievement is regressed on the selection test separately for each group. Then, the difference between the predicted values at the award cut-off is tested to assess the effect of the award. A test of the hypothesis in formula [5] could be used to perform this analysis.

Figure 1 shows a situation where the regressions of achievement on the selection test are homogeneous. The diagonal lines in the figure represent the separate regression lines for the award and no award groups. The regression lines are represented in Figure 1 as being parallel. The use of model [3], however, does not require homogeneity of regression. Each group's regression line could have any equation, and hypothesis [5] is still testable.

The Johnson-Neyman significance region technique uses tests of hypotheses like expression [5] to define regions on the covariate where groups differ, or do not differ, significantly on Y. Figure 2 below illustrates a possible outcome of a Johnson-Neyman significant region analysis. Testing differences between predicted values on Y for groups 1 and 2 might show that for values of $X < X_1$, group 2's predicted values are significantly higher on Y than those of group 1. Between X_1 and X_2 there is no significant difference between the groups' regression lines. Finally, for $X > X_2$ there is a significant difference favoring group 1. These regions can be defined by testing hypotheses like expression [5] for a full range of values on X and then noting which regions permit a significant interpretation. It might be necessary to iterate on X for reasonably accurate values of X_1 and X_2.

Figure 1 Illustration of the Regression Discontinuity Design

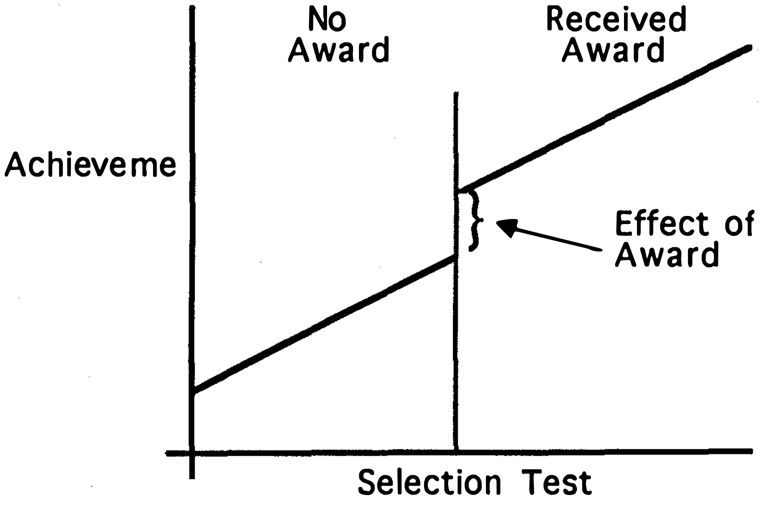
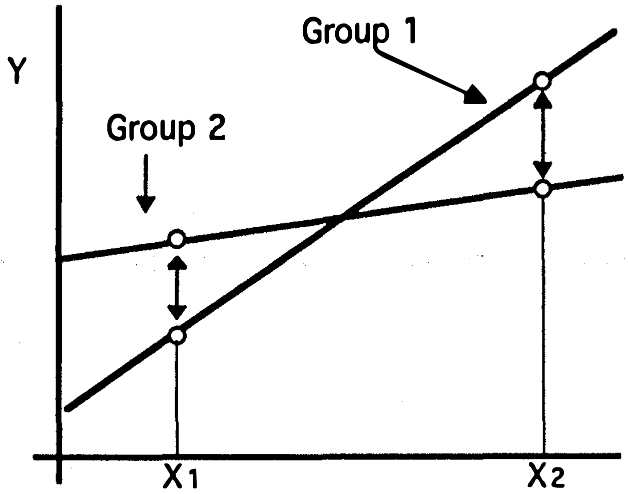


Figure 2 An Illustration of Johnson-Neyman Significance Regions



The two illustrations presented above have assumed a linear relationship between the covariate and the dependent variable. This assumption has been made to simplify the illustrations, not b●cause it is an essential assumption for the ANCOVA. For any linear model application, the essential assumption is that the model be correctly sp●cified.

Model [3] would be inappropriate if X and Y had a nonlinear relationship. Suppose the relationship between X and Y could be described by a quadratic model, then model [3] could be revised as follows:

$$Y_{ij} = \alpha_j + \beta_{j1} X_{ij} + \beta_{j2} X^2_{ij} + \varepsilon_{ij} \qquad [6]$$

Model [6] allows each group to have its own quadratic regression line. For group j, the quadratic model is describ●d by the regression constant α_j; β_{j1}, the coefficient for X; and β_{j2}, the coefficient for X^2.

Using The SAS REG Program For Nonstandard ANCOVA Tests

The SAS REG program will be used to illustrate the analyses. The SAS REG program is used for this purpose since it conveniently provides F-tests of any linear hypothesis for model parameters. PROC REG allows natural language expressions of linear hypotheses of the form H_O: Lß=c, where Lß=c represents a generalized system of linear combinations of the ß parameters (SAS Institute Inc., 1990, Chapter 6). Note that expression [5] is a linear combination of the α and ß parameters, and will serve as the basis for test requests in PROC REG.

Coding The Data For Analysis

Models [3] and [6] require that each of the J groups in the analysis have a separate model fitting the covariate to the dependent variable. Any computer model used for these analyses must posses this fundamental property. Now, there are an infinite number of ways to fit such models. For example, models [1] and [2] are isomorphic representations of the same structural model. In fact, any linear combination of the variables in these models will produce parameter estimates that will fit Y equally well. The approach shown here is not distinctive, in any important sense. Rather, its value lies in its simplicity.

Group Membership Coding

Binary variables will be used to code membership in the design groups. If a case is a member of group 1, then a binary variable, G1, will be coded 1 for that case. If a case is not a member of group 1, then that case's value on G1 will be 0. In a like manner each design group will be represented by a binary variable. For a J group design there will be J mutually exclusive and exhaustive binary variables; (G1, G2, ... , GJ)

Coding The Covariate

The covariate will be denoted as X in the following illustrations. In order to fit models like [3] and [6], X will be expanded to J variables. Cases in group 1 will have variable X1 coded with their value of X, while cases not in group 1 will have a 0 coded in X1. In a like manner X will be expressed as J variables (X1, X2, ... , XJ). In the case of model [6], X^2 will also be expanded the same way as X was expanded intd (X1, X2, ... , XJ) in the preceding coding scheme. A case in group 1 will be coded in the variable XSQ1 with its value of X^2. A case that is not a member of group 1, will receive a code of 0 in XSQ1. X^2 will thereby be expanded to (XSQ1, XSQ2, ...,XSQJ) variables.

Table 1 below illustrates this coding scheme for a three group problem fit by model [3].

If the data set contains only Y, X and the GROUP variables given in Table 1, the following data transformations can be used to generate G1 through X3:

```
IF GROUP = 1 THEN G1 = 1; ELSE G1 = 0;
IF GROUP = 2 THEN G2 = 1; ELSE G2 = 0;
IF GROUP = 3 THEN G3 = 1; ELSE G3 = 0;
X1 = G1*X;
X2 = G2*X;
X3 = G3*X;
```

PROC REG Commands

The following commands illustrate how model (3) would be estimated with a data set like that in Table 1. The MODEL statement will use the NOINT option, which means that the program is not to estimate a common intercept for the entire sample. The coding of G1, G2 and G3 will permit separate intercepts to be estimated for each group.

```
PROC REG;
MODEL Y = G1 G2 G3 X1 X2 X3 / NOINT;
```

The parameter estimates for G1 to X3 in this SAS model statement are interpreted as follows:

G1 = Y intercept for group 1,
G2 = Y intercept for group 2,
G3 = Y intercept for group 3,
X1 = the slope for the regression line for group 1,
X2 = the slope for the regression line for group 2,
X3 = the slope for the regression line for group 3.

If one wanted to test the hypothesis that the expected values of Y when X = 50 in populations 1 and 2 were equal, the null hypothesis would become:

$$H_o: (\alpha_1 + \beta_1 50) = (\alpha_2 + \beta_2 50) \qquad [7]$$

Table 1 Coding for a Three Group ANCOVA Data Set in PROC REG

Y	X	GROUP	G1	G2	G3	X1	X2	X3
6	3	1	1	0	0	3	0	0
8	2	1	1	0	0	2	0	0
5	4	2	0	1	0	0	4	0
9	5	2	0	1	0	0	5	0
8	4	3	0	0	1	0	0	4
7	6	3	0	0	1	0	0	6

The following TEST request could be inserted after the model statement to produce an F-test of this hypothesis.

TEST G1 + 50*X1 = G2 +50*X2;

If groups 1 and 2 were the award and no award groups in the Campbell and Stanley regression discontinuity design, and if the value X $\geq$ 50 qualifies one for an award, then the above statement would produce an appropriate test of the effect of the award on achievement.

The same type of test could be used for many values of X to locate regions on X for which there is a significant difference between the predicted Y values for groups 1 and 2. For example, if X was observed in the range (1,10), ten tests could be requested to locate the significance regions.

```
TEST G1 + 1*X1 = G2 + 1*X2;
TEST G1 + 2*X1 = G2 + 2*X2;
TEST G1 + 3*X1 = G2 + 3*X2;
TEST G1 + 4*X1 = G2 +4*X2;
TEST G1 + 5*X1 = G2 + 5*X2;
TEST G1 + 6*X1 = G2 + 6*X2;
TEST G1 + 7*X1 = G2 + 7*X2;
TEST G1 + 8*X1 = G2 + 8*X2;
TEST G1 + 9*X1 = G2 + 9*X2;
TEST G1 + 10*X1 = G2 +10*X2;
```

If the first three tests were significant and the last seven were not significant, the significance region would be $1 \leq X \leq 3$ and the nonsignificance region would be $4 \leq X \leq 10$. SASLOGs and listings with demonstrations of nonlinear ANCOVA extensions of these same tests can be obtained by writing the author.

Conclusion

The ANCOVA can provide a flexible approach to many analysis problems. Researchers are encouraged to use ANCOVA models that are structurally appropriate for their data and their research questions. This paper illustrated some simple tests of expected values that can be expressed as linear combinations of the model parameters. These simple tests were applied to the regression discontinuity design and the Johnson-Neyman significance region analysis. These applications were selected because they both are performed when a researcher encounters a problem with structural or design assumptions. The tests shown here illustrate how a researcher can articulate and test interesting hypotheses under these problematic conditions.

References

Campbell, D. T. and Stanley, J. C. (1968). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

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

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

Mueller, R. O. (1990). Teaching ANCOVA: The importance of random assignment. *Multiple Linear Regression Viewpoints, 17* (2), 1-14.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research* (2nd Ed), New York: Holt, Rinehart and Winston.

SAS Institute Inc. (1990). *SAS/STAT user's guide* (Vol. 2). Cary, N.C.: SAS Institute Inc.

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