Adjusted Means in Analysis of Covariance: Are They Meaningful?

Dennis L. Clason
New Mexico State University

Daniel J. Mundfrom
Eastern Kentucky University

In many data analytical applications involving the comparison of means from several populations, analyses are adjusted by including one or more continuous variables in the model. Such an analysis is often called an Analysis of Covariance. In this study, we examine the effect of such an adjustment on the precision of inferences regarding the adjusted means and the interpretation of those means. We consider the variance of the difference between a pair of adjusted means and show that the variance is minimized when the mean of the covariate is the same in both groups. However, when the covariate means differ across groups, the additional term in the model can dominate the variance of the comparison. We investigate the break-even point for this comparison and under what condition that occurs. Finally, we demonstrate, with several data sets, how these results may be manifested in practice.

There are a variety of analytical procedures that have been developed for use in comparing the performance of individuals in two or more groups. With only one variable and two groups, a simple *t*-test will suffice in most instances. With additional groups involved in the comparison, a straight-forward extension of the independent samples *t*-test to a one-way analysis of variance (ANOVA) is common (see Dowdy, Weardon, & Chilko, 2004; Glass & Hopkins, 1996). More advanced procedures allow for the inclusion of additional variables in the analysis, which are sometimes called independent or predictor variables. When a second categorical variable is included, the analysis may be referred to as a two-factor ANOVA, or a randomized complete block design, depending on the context of the analysis and the research design employed. When the additional predictor variable is numerical, the analysis is typically called an analysis of covariance (ANCOVA) and the numerical predictor is commonly called a covariate. These last two scenarios are easily extended conceptually to allow for the inclusion of additional predictors/covariates, if desired.

The focus of this research is on ANCOVA and, in particular, the interpretation of the adjusted means from such analyses. Most modern textbooks addressing experimental design include chapters or sections on the methods of ANCOVA (see Hicks & Turner, 1999; Kuehl, 2000; Montgomery, 2009). Fisher (1934) stated that ANCOVA "combines the advantages and reconciles the requirements of . . . regression and analysis of variance" (p. 281). Cochran (1957) summarized the theory, nature, and use of ANCOVA and identified some potential problems and misuses of the procedure. One of the most frequent uses Cochran identified was "to increase precision in randomized experiments" (p. 262). Typically, this increased precision was accomplished by making preliminary measurements on experimental units prior to applying treatments. He noted, however, that it was not uncommon for the covariate to be measured on a different scale than that of the response variable and used reading performance adjusted for initial IQ scores as an example. In situations such as these, it is important to verify that treatments have had no effect on the covariate. Cochran further explained that

> When the treatments do affect the [covariates] to some extent, the covariance adjustments take on a different meaning. They no longer merely remove a component of experimental error. In addition, they distort the nature of the treatment effect that is being measured. (p.264)

The same caution has been made more recently by Tracz, Nelson, Newman, and Beltran (2005) and Nimon and Henson (2010). A second common use of ANCOVA described by Cochran (1957, p. 264) was "to remove the effects of disturbing variables in observational studies," primarily in situations that are not easily investigated with randomized experiments. He identified two situations in observational studies that could be of concern: 1) bias and 2) extrapolation in the ANCOVA adjustment as a result of differences between the distributions of the covariate across the groups. In regard to situation 1, consider a typical one-way ANCOVA model:

$$Y_{ij} = \mu + \tau_i + \beta(X_{ij} - \bar{X}) + \varepsilon_{ij} \tag{1}$$

and the unadjusted difference between the two group means

$$\bar{Y}_{1.} - \bar{Y}_{2.} = \tau_1 - \tau_2 + \beta(\bar{X}_1 - \bar{X}_2) + \varepsilon_{1.} - \varepsilon_{2.} \tag{2}$$

If the two groups have not been matched for values of the covariate, the difference, $(\bar{X}_{1.} - \bar{X}_2)$, may be reflecting a real difference in the distributions of the covariate that is much larger than which the within-group variations can account. The term, $(\bar{X}_{1.} - \bar{X}_2)$, measures the nature of the bias and, if left unaccounted for, will invalidate any significance tests or confidence intervals.

To illustrate situation 2, Cochran proposed the following illustration.

> Suppose we were adjusting for differences in parents' income in a comparison of private and public school children, and that private-school incomes ranged from $10,000 - $12,000, while the public-school incomes ranged from $4,000 - $6,000. The [analysis of] covariance would adjust results so that they allegedly applied to a mean income of $8,000 in each group, although neither group has any observation in which incomes are at or even near this level. (pp. 265-266)

There are two consequences of this extrapolation that Cochran (1957) notes. First, for the ANCOVA to remove all the bias, the regression equation of the dependent variable on the covariate must hold in the region of scores where data are not observed and there is no empirical evidence to support that it does. Second, even if the regression does hold in this region, the standard errors of the adjusted means become large due to the fact that the standard error formula accounts for the extrapolation that takes place. According to Cochran, "Consequently, the adjusted differences may become insignificant statistically merely because the adjusted comparisons are of low precision" (p.266). He concludes by cautioning that when the groups differ widely in the values of the covariate, "the interpretation of an adjusted analysis is speculative rather than soundly based" (p. 266).

In a similar, related paper on the interpretation of adjusted treatment means in ANCOVA, Fairfield Smith (1957) illustrated, with several examples, that the interpretation of the adjusted means requires careful study. Cochran (1957) summarizes, "Sometimes these averages have no physical or biological meaning of interest to the investigator, and sometimes they do not have the meaning that is ascribed to them at first glance" (p. 267). Again, this point seems to be a primary focus shared recently by Tracz et al. (2005) and Nimon and Henson (2010).

A Theoretical Approach

To investigate the effect of an analysis of covariance adjustment on the precision of inferences regarding the adjusted means and the interpretation of those means, we examine the variance of the difference between a pair of adjusted means to determine, mathematically, the condition in which this variance is minimized.

Derivation

Consider the linear model:

$$\mathrm{y} = \mathrm{X}\beta + \mathrm{W}\mu + \varepsilon,$$

$$\mathrm{y} = [\mathrm{X} \quad \mathrm{W}] \begin{bmatrix} \beta \\ \mu \end{bmatrix} + \varepsilon$$

where **W** is an *n* by *t* incidence matrix defining the treatment structure (i.e., or population structure for observational studies) of a study and **X** is an *n* by *k* matrix of concomitant variables. In the following, we take **X** to have centered columns [i.e., $\mathbf{X} = \mathbf{X}^*(\mathbf{I}_n - \mathbf{1}_n\mathbf{1}'_n/n)$ if the columns of $\mathbf{X}^*$ are not centered] and both **W** and **X** to have full column rank. Performing the estimation in two stages, we have the following solutions to the normal equations:

$$\hat{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y},$$

$$\bar{y} = (\mathrm{W}'\mathrm{W})^{-1}\mathrm{W}'\mathrm{y},$$

$$\hat{\mu} = [\mathrm{I}_t - (\mathrm{W}'\mathrm{W})^{-1}\mathrm{W}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{W}]\bar{\mathrm{y}}$$

The estimate vector $\hat{\boldsymbol{\mu}}$ is often referred to as the estimate(s) of the mean(s) adjusted for the covariate(s), **X**, or simply the adjusted mean(s). The second term in the sum acts to estimate the response at the mean value of each covariate. Similarly, the regression partial slopes can be adjusted for the incidence structure as follows:

$$\hat{\beta}_{\mathrm{adj}} = \hat{\beta} - (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{W}(\mathrm{W}'\mathrm{W} - \mathrm{W}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{W})\big(\mathrm{y} - \mathrm{X}\hat{\beta}\big)$$

In the simplest case (i.e., t levels for the incidence matrix and a single covariate), this situation is expressed in scalar form as:

$$y_{ij} = \mu_i + \beta x_{ij} + \varepsilon_{ij},\ i = 1, 2, \dots t; j = 1, 2, \dots, n_i;\ \Sigma_{\mathrm{i}} n_i = n$$

The unadjusted means are:

$$\bar{y}_{i+} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

If we adjust for the (hypothesized linear) effect of the concomitant variable on the adjusted means (i.e., also called least-squares means and population marginal means, see Searle, Speed, & Milliken, 1980), we have:

$$\hat{\mu}_i = \bar{y}_{i+} - \hat{\beta} \bar{x}_{i+},$$

and the difference between any pair of adjusted means is:

$$\hat{\mu}_i - \hat{\mu}_{i\prime} = \bar{y}_{i+} - \bar{y}_{i\prime+} - \widehat{\beta}\,(\bar{x}_{i+} - \bar{x}_{i\prime+})$$

The variance of this difference is:

$$\begin{aligned}\mathrm{Var}(\hat{\mu}_i - \hat{\mu}_{i\prime}) &= \mathrm{Var}\left[\bar{y}_{i+} - \bar{y}_{i\prime+} - \widehat{\beta}\,(\bar{x}_{i+} - \bar{x}_{i\prime+})\right],\\ &= \mathrm{Var}(\bar{y}_{i+}) + \mathrm{Var}(\bar{y}_{i'+}) + \mathrm{Var}\left(\widehat{\beta}\,(\bar{x}_{i+} - \bar{x}_{i'+})\right),\\ &= \frac{\sigma^2}{n_i} + \frac{\sigma^2}{n_{i\prime}} + \sigma^2 \frac{(\bar{x}_{i+} - \bar{x}_{i\prime+})^2}{\sum\sum x_{ij}^2}\end{aligned}$$

Results

Clearly, this expression is minimized when the mean of the concomitant variable is the same in both incidence groups. On the other hand, when the means of the concomitant variable differ in the incidence groups, the added term can grow to dominate the variance of the comparison. The break-even point occurs when the analysis ignoring the concomitant variable and the analysis with the concomitant variable have the same variance for pairwise comparisons. This result occurs when:

$$\sigma_A^2 \left(\frac{1}{n_i} + \frac{1}{n_{i\prime}}\right) = \sigma_C^2 \left(\frac{1}{n_i} + \frac{1}{n_{i\prime}} + \frac{(\bar{x}_{i+} - \bar{x}_{i\prime+})^2}{\sum\sum x_{ij}^2}\right),$$

where σ_A^2 represents the variance ignoring the concomitant variable and σ_C^2 is the variance when the concomitant variable is included in the analysis[†]. Some rearrangement of terms shows the break-even condition occurs when:

$$\frac{\sigma_A^2}{\sigma_C^2} = 1 + \frac{(\bar{x}_{i+} - \bar{x}_{i\prime+})^2}{\sum\sum x_{ij}^2} \cdot \frac{n_i n_j}{n_i + n_j}.$$

If we further simplify the structure of the problem by assuming equal replication in the incidence groups, the break-even condition is:

$$\frac{\sigma_A^2}{\sigma_C^2} = 1 + \frac{(\bar{x}_{i+} - \bar{x}_{i\prime+})^2}{\sum\sum x_{ij}^2} \cdot \frac{n}{2}.$$

Conclusions

Thus, whenever the reduction in the experimental error is more than the right hand term in the sum, precision is increased by including the concomitant variable in the analysis. If the reduction in the experimental error is less than this quantity, we actually lose precision by including the concomitant variable in the analysis. This relationship is depicted in the three graphs in Figure 1 using Wishart's (1938) swine growth rate data to compare three diets. The response variable is growth rate (i.e., in pounds/week) and the covariate is the initial weight of the pig (i.e., in pounds).

We modeled the effect of separating the covariate means over the diets. Initially, the covariate means were not statistically significantly different across treatments. Two values of the classification variables were arbitrarily chosen and the covariate values were shifted apart by adding and subtracting a constant to the covariate values in those groups. Positive shifts increase the separation of the treatment means and negative shifts (i.e., down to about -5) decrease the separation. Shifts "below" -5 reverse the treatment effects and treatment separation increases again. The upper left panel shows, as expected, that the adjusted difference between treatment means is a linear function of the separation. The upper right panel shows that the standard error of the difference in means is an increasing function of the separation. The lower panel shows the adjusted, partial F-test for treatment effects. Note that the treatment differences are not statistically significant when the adjustment brings the means closer together. However, as the adjustment continues to pull them apart, the treatment effect eventually attains statistical significance. The implication of performing such adjustments is simply this: that naively adjusting the means on the basis of covariates is dangerous if there is a relationship between the treatment and the covariate.

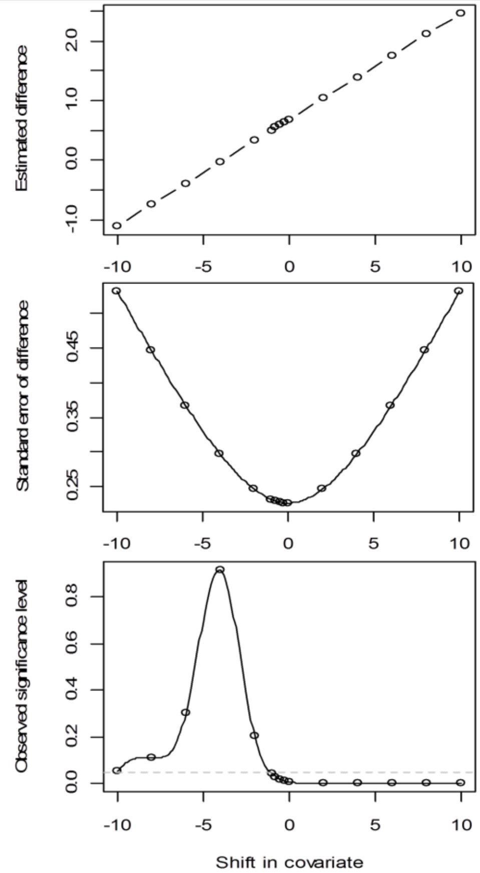

Figure 1. Effects of shifting covariate means on estimated differences, standard errors, and observed significance levels.

Simulated Demonstration

To illustrate these results, we performed several simulations with varying group sizes, sample sizes, regression slopes, and group separation to provide some examples of how these results may be encountered in practice. The simulations do not represent an exhaustive collection of possible scenarios, but rather just a few examples for illustration.

Methods

For these simulations, data were generated using SAS version 9.2 data steps and the resulting data were analyzed with SAS PROC GLM. The critical contrast was estimated using both the LSMEANS statement and a hand-coded ESTIMATE statement. The observed significance level (i.e., Ordinary Least Squares (OSL), *p*-value) and the estimated standard error of the comparison were saved in a separate file for analysis. Parameters of the simulation included two numbers of groups to be compared (2 and 8), three sample sizes per group (4, 8, and 16), four regression slopes (0, 0.2, 0.4, and 0.6) measured in σ_ε units, and five levels of separation of the critical groups (0, 0.25, 0.50, 0.75, and 1.50) measured in σ_ε units. The standard deviations of the error and of the concomitant variable X were both set to 1. The means of the critical groups were adjusted so that the critical groups had equal means before adjustment. Consequently, in each scenario, the analysis of variance is conducted under a true global null hypothesis (i.e., the ANCOVA has a true null for the unadjusted means and the alternate is true for the adjusted means). Other than for the pair of critical groups, all means were set to 0.

The OSLs were analyzed using a quantile-quantile plot of the empirical OSLs against the expected quantiles for a uniform distribution on [0,1]. If the test detected differences, the plot displayed as a straight line (Clason, 2012). If the test detected differences, the plot was strictly convex. Selected plots are shown in Figures 2 and 3. Kolmogorov single-sample tests of the OSLs showed that the uniform distribution on [0,1] fit the data extremely well. The standard errors for the critical comparisons were analyzed using box-and-whisker plots. These selected plots are shown in Figure 4.

Results

The plots in Figures 2 and 3 show that when there is no separation of the concomitant variables, or when the concomitant variable is unrelated to the response variable, the tests perform as expected. However, when the concomitant variable means are related to the group definitions, the test shows noticeable effects even for low slopes ($0.2\sigma_\varepsilon$) and small separations (i.e., as small as .25 of the standard deviation of the concomitant variable). The box plots in Figure 4 shows that the findings of the mathematical derivations are borne out (i.e., increasing separation of the concomitant variable means increases in the standard error of the difference of the adjusted means). What the theory did not directly predict is the greatly increased variability in the sampling distribution of the standard errors.

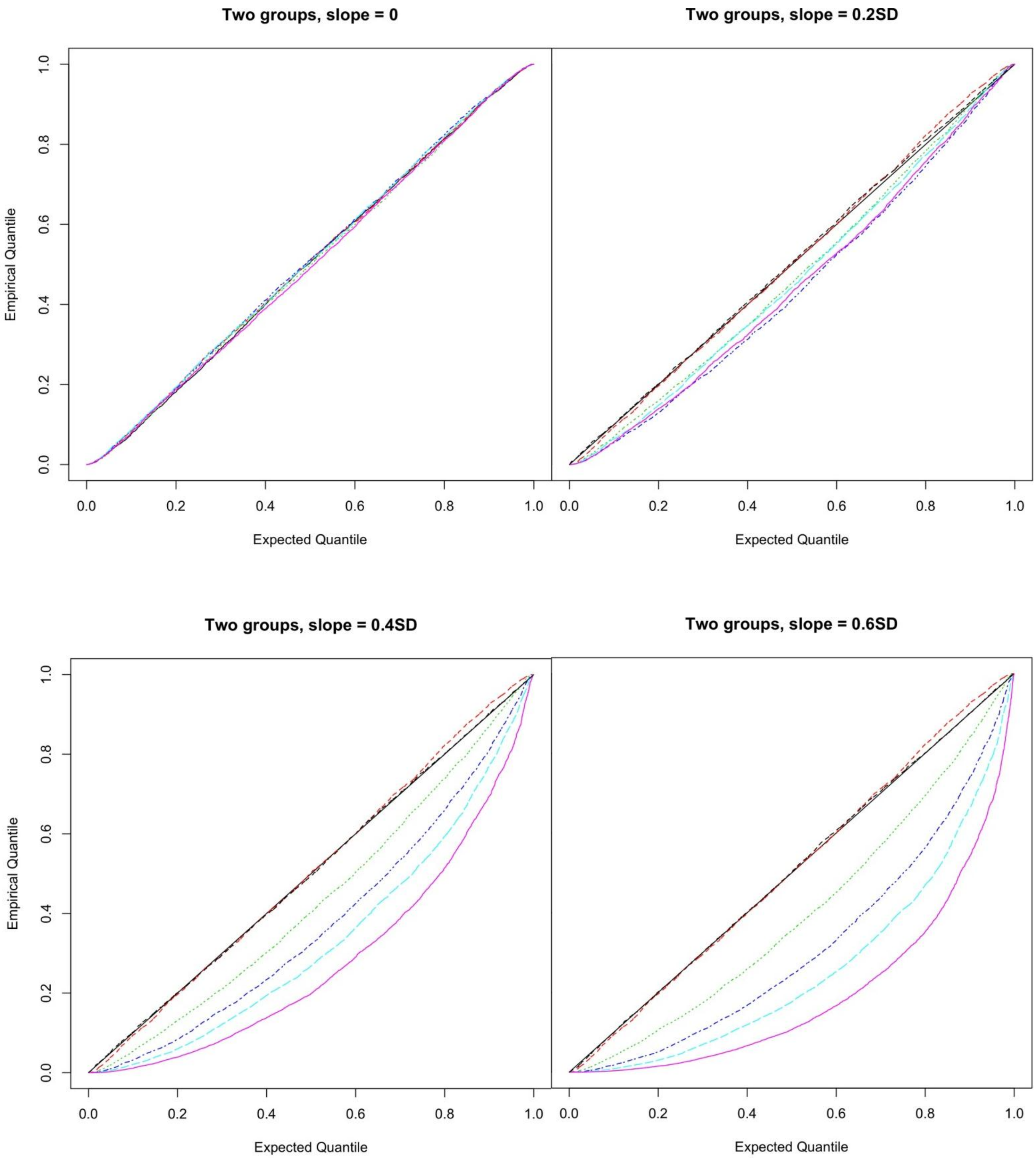

Figure 2. Q-Q Plot of empirical observed significance levels vs. expected quantiles of a uniform [0, 1] distribution with 4 observations/group.

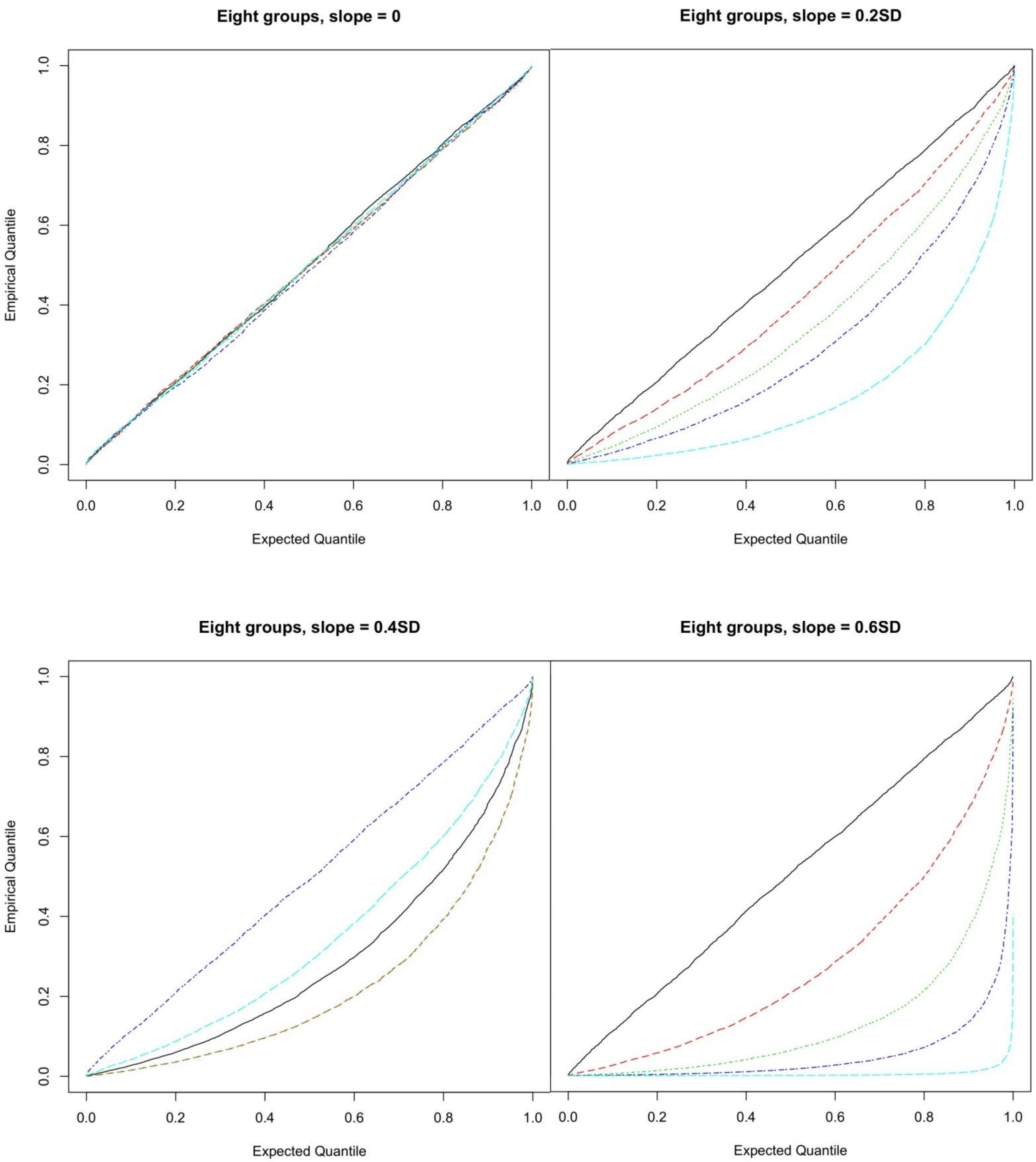

Figure 3. Q-Q Plot of empirical observed significance levels vs. expected quantiles of a uniform [0, 1] distribution with 16 observations/group.

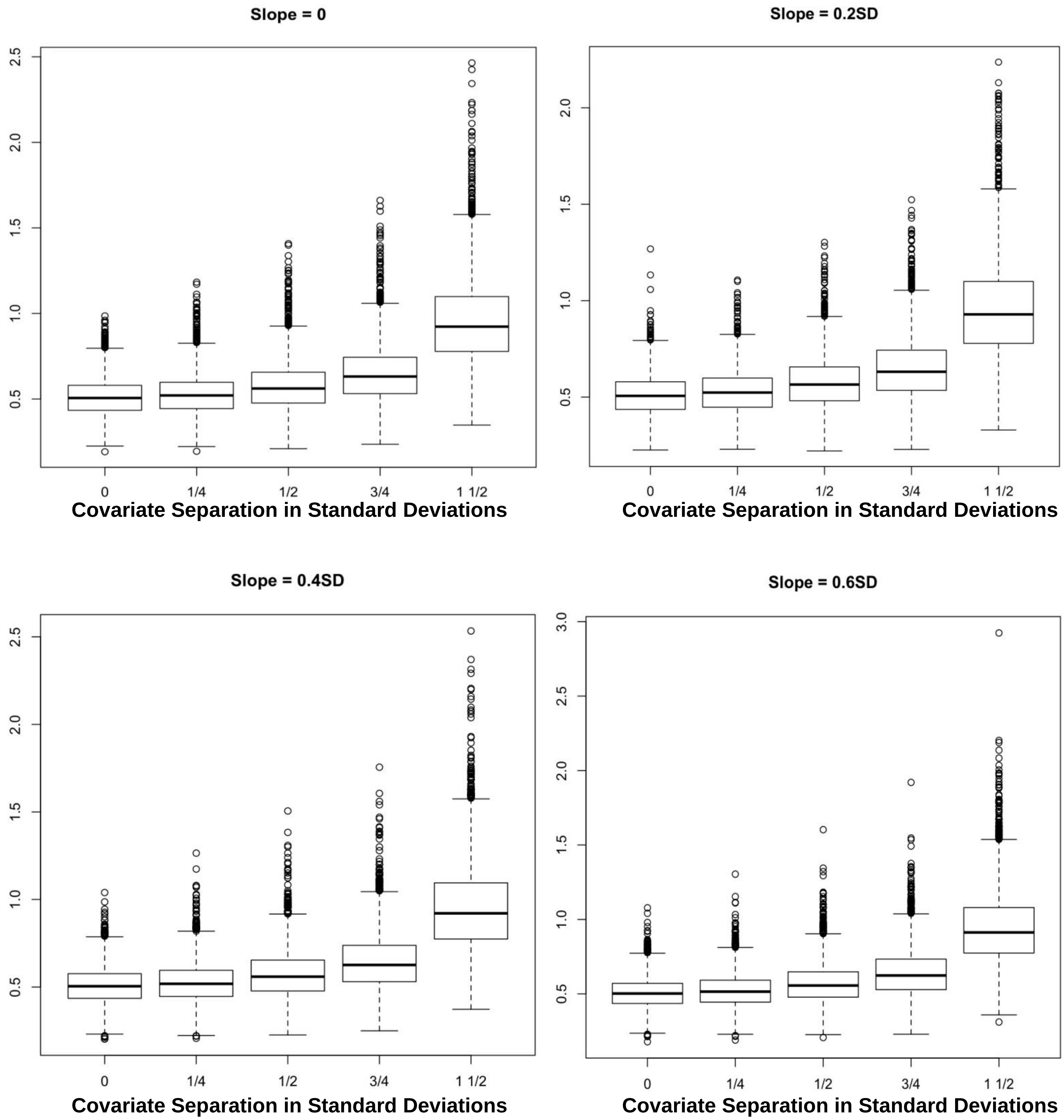

Figure 4. Standard errors for the difference between two adjusted means with 8 observations/group and slopes measured in standard deviations of Y.

Conclusions and Implications

The simulations point out the importance of carefully considering the contextual meanings of adjusting means in ANCOVA. If it is possible to adjust the distribution of the concomitant variable(s), then a comparison of adjusted means may make sense. If, however, the distribution of the concomitant variable cannot be adjusted, as in using a pre-test score as a covariate in analyzing post-test results, then the adjustment is comparing entities that not only do not exist, but (probably) cannot exist. Milliken (1980) described the term "analysis of covariance" as a misnomer; he preferred to consider the problem as "analysis of lines." There is much to recommend this approach. When analyzing lines, we rarely ask if we should adjust the mean responses to the mean of the explanatory variable. Rather, we ask at which values of the explanatory variable(s) we wish to make comparisons. Keeping this concept in the forefront of our thinking is critical, especially when analyzing observational studies rather than designed experiments. In the setting of a designed experiment, the randomization process assures us that there is, on average, no relationship between the concomitant variable and the group incidence. This level of control is not present in observational studies. This study limited attention to a single covariate. It is obvious from the matrix formulation of the linear model that including additional concomitant variables cannot improve the situation; it can only make it worse.

Footnote

[†]This could be problematic, as it is not clear that there exists a way for the errors to be normal under both analyses unless the covariate is uncorrelated with the response.

References

Clason, D. L. (2012). *The probability integral transformation and simulation studies*. Manuscript submitted for publication.

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

Dowdy, S., Weardon, S., & Chilko, D. (2004). *Statistics for research* (3rd ed.). Hoboken, NJ: Wiley & Sons.

Fairfield Smith, H. (1957). Interpretation of adjusted treatment means and regressions in analysis of covariance. *Biometrics*, *13*, 282-308.

Fisher, R. A. (1934). *Statistical methods* (5th ed.). Edinburgh, Scotland: Oliver and Boyd Ltd.

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Needham Heights, MA: Allyn and Bacon.

Hicks, C. R., & Turner, K. V. (1999). *Fundamental concepts in the design of experiments* (5th ed.). New York: Oxford Press.

Kuehl, R. O. (2000). *Design of experiments: Statistical principles of research design and analysis* (2nd ed.). Pacific Grove, CA: Brooks-Cole Publishing.

Montgomery, D. C. (2009). *Design and analysis of experiments* (7th ed.). Hoboken, NJ: Wiley & Son.

Nimon, K., & Henson, R. K. (2010, May). *Validity of a residualized dependent variable after pretest covariance correction: Still the same variable?* Paper presented at the Annual Meeting of the American Educational Research Association, Denver, CO.

Rao, C. R. (1973). *Linear statistical inference and its applications*. Hoboken, NJ: Wiley & Sons.

Searle, S. R., Speed, F. M., & Milliken, G. A. (1980). Population marginal means in the linear model: An alternative to least squares means. *American Statistician*, *34*, 216-221.

Tracz, S. M., Nelson, L. I., Newman, I., & Beltran, A. (2005). The misuse of ANCOVA: The academic and political implications of Type VI errors in studies of achievement and socioeconomic status. *Multiple Linear Regression Viewpoints*, *31*, 16-21.

Wishart, J. (1938). Growth rate determinations in nutrition studies with the bacon pig and their analysis. *Biometrika*, *30*, 16-28.

Send correspondence to: Dennis L. Clason
New Mexico State University
Email: dclason@nmsu.edu