Journal of Applied Psychology
February 1990 Vol. 75, No. 1, 95-96

Variance Homogeneity in Interactive Regression
A Clarifying Note About Data Transformations

Philip Bobko
Department of Management Rutgers University
Craig Russell
Department of Industrial Relations and Human Resources Rutgers University

ABSTRACT

A data transformation recently used by Stone and Hollenbeck (1989) is based on a faulty premise and should not be considered in future analyses. Specifically, these authors mistakenly equate the homoscedasticity assumption in regression analysis with the notion that subgroup variances need to be equal. We demonstrate that subgroup variances in regression can be legitimately different, owing to true main effects and interactions in the data. Therefore, any transformation addressing these differential subgroup variances may unwittingly remove true effects of the independent variables.

We wish to thank Gene Stone for his careful reading of this article. We appreciate his comments on, and agreement with, the contents of this article. We have no intention of making any ad hominem comments in this article. Rather, we (like Gene Stone) hope to see clarity in the issues surrounding the use of moderated regression techniques.
Correspondence may be addressed to Philip Bobko, Department of Management, Rutgers University, Avenue D and Rockafeller Road, New Brunswick, New Jersey, 08903.
Received: April 10, 1989
Revised: June 22, 1989
Accepted: June 23, 1989

Recently, Stone and Hollenbeck (1989) empirically addressed the sensitivity of moderated regression analysis to differences in both subgroup correlations and subgroup regression weights. They took issue with two data sets and conclusions initially presented by Arnold (1982) . These data sets were analyzed in both previous articles to help determine the potential differences between subgroup and moderator analyses. Stone and Hollenbeck's contention was that these data sets show "nontrivial violations of the assumptions connected with

regression analysis" (p. 5). They then suggested that, by making scale transformations that allegedly reduce the violation of these assumptions, many of the conclusions of Arnold can be dismissed. Although several of Stone and Hollenbeck's points are valid, these authors make one critical error in their data transformations that needs to be avoided by future researchers. Our intent is to demonstrate the error and clarify the use of this data transformation.

Regression Assumptions

In ordinary least squares multiple regression, it is assumed that the conditional variance of the criterion (Y) is the same regardless of the values of the predictors X_i. This assumption is often referred to as the assumption of *homoscedasticity.* Other usual assumptions for general linear model theory include independence and underlying normality of the error terms (see Scheffé, 1959 , or Stone & Hollenbeck, 1989 , p. 7).

Unfortunately, Stone and Hollenbeck added yet another assumption to their list of concerns: homogeneity of variance. For example, they stated that researchers must be concerned with "normality, homogeneity of variance, homoscedasticity, independence..." (p. 4). Note that the notion of homogeneity of variance is distinct from the notion of homoscedasticity. This distinction can be found throughout their reanalysis of Arnold's data.

As operationalized by Stone and Hollenbeck, the distinct notion of homogeneity of variance is not a statistically legitimate assumption. We now demonstrate this fact by considering Stone and Hollenbeck's use of the term and then a straightforward example.

Variance Homogeneity

Consider Arnold's (1982) original Data Set 1. The descriptive statistics for this data set appear in Stone and Hollenbeck's Table 1 . For our purposes, it is sufficient to note that the data are split into two subgroups (Subgroups A and B). The residual variance for Subgroup A (i.e., the variance *within* Subgroup A, *conditional* on the predictor X) is given as 601.35; the residual variance for Subgroup B is given as 25,359.64. Stone and Hollenbeck noted these differences in residual variances and indicated that the homoscedasticity assumption is violated. We have no quarrel with this logic.

However, Stone and Hollenbeck also pointed out that the variance on Y for Subgroup A S^2_a is 1,268.25, whereas the variance on Y for Subgroup B S^2_b is 26,792.84. Note that these variances are computed *within* each subgroup, but *across* all data points in that subgroup, regardless of the corresponding values on the predictor (X). Stone and Hollenbeck then stated, on the basis of the difference between these latter two variances, "As can be seen..., the assumption of homogeneity of variance was violated" (p. 5).

Stone and Hollenbeck then reanalyzed the Arnold data after first conducting several data transformations, one of which attempts to reduce the discrepancy between the two subgroup variances. Unfortunately, this is not an appropriate transformation. In fact, we demonstrate that these subgroup variances can be different, owing to true main effects and interactions in the data rather than because of any violations in assumptions.

An Example

The original Arnold data used a continuous dependent variable (Y), a continuous predictor (X), and a dichotomous subgrouping variable (either Group A or B). Essentially, the moderated regression analysis was a multiple regression of Y on two variables, X_1 and X_2, where X_1 was the continuous predictor and X_2 was the dichotomous grouping variable.

We assume a simplified version of this situation, where the predictor of interest X_1 is also assumed to be dichotomous. Then, the underlying model is equivalent to a two-way analysis of variance (ANOVA) with two levels of each factor. (See Cohen & Cohen, 1983 , or Pedhazur, 1982 , on how to code the general linear model to accommodate experimental designs.)

Hypothetical data for such an ANOVA (i.e., a regression with two dichotomous predictors) is presented in Table 1 . In this table, the original predictor of interest X_1 is represented by the two rows; the subgrouping variable X_2, by the two columns. Note that, within each of the four cells, the variance of the Y scores is the same (i.e., the residual variance, or mean-squared error in the ANOVA , is equal to 1.67). The only difference between the cells is that a constant value of 10 has been added to each score in the lower right-hand cell. This modification also creates a difference in the row and column marginal means (mean of 3 vs. mean of 8). Thus, there are two main effects and an interaction in the data. (See Bobko, 1986 , for a discussion of this effect.)

Now, the regression assumption of homoscedasticity has been met in this data because all residual, within-cell variances S^2_{res} are equal. However, notice that S^2_a and S^2_b are quite different in magnitude (1.54 vs. 28.46). According to Stone and Hollenbeck's operationalizations, this would mean that their notion of variance homogeneity has been violated.

However, the reason for the differences in these two variances has nothing to do with any assumption of ordinary least squares regression. Rather, these two variances are different precisely because there are main effects and an interaction present in the data. Looking at Table 1 , one can readily see that the differential effect of being in the lower right-hand cell causes the dispersion of all scores in the second column (i.e., across all scores in Group B) to be much larger than the dispersion across all scores in the first column (Group A). Again, this difference is due solely to legitimate mean differences (or, in a regression

sense, legitimate effects for the two predictors and their interaction). In fact, in the traditional ANOVA for this example, the expected values of the sum of squares for the two factors and their interaction are directly related to differences in these two variances.

Conclusions

Stone and Hollenbeck (1989) presented a variety of thoughtful considerations to the literature on interactive regression. However, the overarching implication of our analysis is clear: It makes no sense to transform regression data on the basis of what Stone and Hollenbeck call variance heterogeneity. In fact, controlling for such differences is, in effect, controlling for potential main effects and interactions, which is precisely what the regression analyses are attempting to search for!

Transformations may be useful when they address legitimate regression assumptions (e.g., within-cell variance stabilizing transformations) and when the choice of a transformation is based on its theoretical meaningfulness (see Scheffé, 1959 , chap. 10). However, transformations of the data can change the sample-based significance of interaction tests (e.g., see Busemeyer & Jones, 1983 ; Scheffé, 1959) and must therefore be used with care. Stone and Hollenbeck's within-subgroup "homogeneity of variance" transformation is simply inappropriate and should not be used by future researchers in this domain.

References

Arnold, H. J. (1982). Moderator variables: A clarification of conceptual, analytic, and psychometric issues. *Organizational Behavior and Human Performance, 29*, 143-174.

Bobko, P. (1986). A solution to some dilemmas when testing hypotheses about ordinal interactions. *Journal of Applied Psychology, 71*, 323-326.

Busemeyer, J. R. & Jones, L. E. (1983). Analysis of multiplicative combination rules when the causal variables are measured with error. *Psychological Bulletin, 93*, 549-562.

Cohen, J. & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.).(Hillsdale, NJ: Erlbaum)

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

Scheffé, H. (1959). *The analysis of variance.* (New York: Wiley)

Stone, E. F. & Hollenbeck, J. R. (1989). Clarifying some controversial issues surrounding statistical procedures for detecting moderator variables: Empirical evidence and related matters. *Journal of Applied Psychology, 74*, 3-10.

Table 1.

Table 1
Illustrative Data, Means, and Variances (Both Marginal and Within-Cell) for a 2 × 2 Analysis of Variance

		X_2 (Subgroup)	
		A	B	
X_1	A	1, 2, 3, 3, 3, 4, 5 $s^2_{within} = 1.67$ $M = 3.00$	1, 2, 3, 3, 3, 4, 5 $s^2_{within} = 1.67$ $M = 3.00$	$M = 3.00$
	B	1, 2, 3, 3, 3, 4, 5 $s^2_{within} = 1.67$ $M = 3.00$	11, 12, 13, 13, 13, 14, 15 $s^2_{within} = 1.67$ $M = 13.00$	$M = 8.00$
		$M = 3.00$ $s^2_w = 1.54$	$M = 8.00$ $s^2_w = 28.46$	