

## THE MODERATOR VARIABLE VIEWED AS HETEROGENEOUS REGRESSION<sup>1</sup>

WAYNE F. VELICER<sup>2</sup>

Testing and Research Services, Wisconsin State University at Oshkosh

While there has been a great variety of recent work dealing with the moderator variable, the meaning of this concept remains unclear. A useful framework for dealing with moderator variables is heterogeneous regression. The heterogeneous model and two useful tests of significance are described. A real data example is used as an illustration. It is also shown that this framework is not inconsistent with past work, and some guidelines for employing heterogeneous regression are presented.

A great deal of recent work in the area of psychological prediction has dealt with the concept of the moderator variable. Unfortunately, this term has been applied to a wide variety of approaches. Different methods for employing moderators have been proposed by Saunders (1956), Ghiselli (1956, 1960a, 1960b, 1963), Cleary (1966), and Rock, Barone, and Linn (1967). In general, all of the proposed systems are based on the concept that some individuals or subgroups of individuals differ in predictability. Various authors have conceived of this as meaning either that they differ in the degree of predictability or that different prediction patterns exist in the subgroups.

The position of this article is that the moderator variable is most profitably conceived of as a case of heterogeneous regression. Such a viewpoint possesses several advantages, not the least of which is simplicity. It also permits continued utilization of a well-known prediction model—multiple regression—and yields two very useful statistical tests. Using a known statistical model also permits us to outline some guidelines as to when moderators are likely to occur on the basis of past experience. It can also be shown that such a conception of moderator variables is not inconsistent with previous systems.

### HETEROGENEOUS REGRESSION MODEL

If we begin with the typical multiple-regression model,

$$Y = Xb,$$

<sup>1</sup>This article is published out of turn at the request of the Editor in order to appear with related articles.

<sup>2</sup>Requests for reprints should be sent to Wayne F. Velicer, Testing and Research Services, Wisconsin State University, Oshkosh, Wisconsin 54901.

where  $Y$  is an  $n \times 1$  vector of criteria scores, and  $X$  is an  $n \times p$  matrix of predictor scores for  $n$  people and  $p$  predictors, and  $b$  is a  $p \times 1$  vector of regression weights, we can obtain the usual least squares solution of the form

$$(X'X)^{-1}X'Y = b,$$

assuming  $(X'X)^{-1}$  exists. This can be written in deviation score form as

$$\Sigma_{xx}^{-1}\Sigma_{xy} = b,$$

where  $\Sigma_{xx}$  is the variance-covariance matrix for the predictors and  $\Sigma_{xy}$  is the covariance vector for the predictors and the criteria.

The total variance of the criteria  $SS_T - SS_Y$  can be divided into two orthogonal parts, the predictable variance,  $SS_{\text{regression}}$ , or

$$SS_R = SS_{\hat{y}} = b' \Sigma_{xy}$$

and the residual variance,  $SS_{\text{error}}$ , or

$$SS_E = SS_{y-\hat{y}} = SS_Y - SS_{\hat{y}}.$$

The theory of heterogeneous subgroups assumes that the total population, and the corresponding sample from that population, is, in reality, two or more subgroups or subpopulations that differ either in having a different predictor covariance matrix ( $\Sigma_{xx}$ ) or a different covariance relation between the predictors and criteria ( $\Sigma_{xy}$ ). This would result in different "best-fitting" regression lines in each subpopulation. Hopefully, by fitting the exact equations in each subgroup, prediction results are superior to just using one regression equation for all  $S$ s.

The regression equation can differ from one subgroup to another in one of two ways. First, the regression weights might differ: this would

be a question of the parallelism of the two lines. Second, it is possible that the two lines might be parallel but not coincidental: this would be a question of the location of the two lines. It follows that two tests are important: (a) a test of the difference of regressions (parallelism), and (b) a test of the difference of positions (coincidentalism). (This discussion follows the development in Williams [1959].)

Assume that we have  $k$  subsamples, consisting of  $n_1, \dots, n_k$  Ss, where

$$\sum_{i=1}^k n_i = N.$$

We can look at regression in the sample in three ways: overall regression, regression within the subsamples, and combined regression. (a) Overall regression is expressed when all elements are used as one sample, and there is just one regression equation. The total variance of the criterion is

$$SS_{T(o)} = SS_{R(o)} + SS_{E(o)}.$$

(b) Regression within the subsamples is expressed when a separate regression line is calculated for each of the  $k$  subsamples. The variance breakdown for the  $i$ th subsample is

$$SS_{T_i} = SS_{R_i} + SS_{E_i}$$

where  $i = 1, \dots, k$ . (c) Combined regression is expressed when the combined covariance matrix is defined as

$$\Sigma_{xx(c)} = \sum_{i=1}^k \Sigma_{xx_i}$$

and

$$\Sigma_{xy(c)} = \sum_{i=1}^k \Sigma_{xy_i}.$$

Then, by the same logic as the general regression model,

$$b(c) = \Sigma^{-1}_{xx(c)} \Sigma_{xy(c)}$$

and

$$SS_{R(c)} = b'(c) \Sigma_{xy(c)}$$

which leads to the regression breakdown for the combined,

$$SS_{T(c)} = SS_{R(c)} + SS_{E(c)}.$$

Using these definitions, it is now possible to calculate the variances for the difference of

regressions using

$$\sum_{i=1}^k SS_{R_i} - SS_{R(c)}$$

with  $(k-1)p$  degrees of freedom, where  $p$  is the number of predictors. The variance for the difference of positions would be

$$SS_{T(o)} - SS_{T(c)} - SS_{R(o)} + SS_{R(c)}$$

with  $k-1$  degrees of freedom. The combined residual would be

$$SS_{T(c)} - \sum_{i=1}^k SS_{R_i}$$

with  $N-kp-k$  degrees of freedom. The combined residual variance is used in the denominator for tests of both difference of regressions and difference of positions. (See Table 1 for a representation of this analysis of variance.)

The method for analyzing heterogeneous subgroups is only sketched here. More detailed mathematical derivations are available in Williams (1959) and Kullback and Rosenblatt (1957). At this time, it would be more profitable to consider an actual numerical application.

#### EXAMPLE EMPLOYING REAL DATA

The study involved 880 first-year students at Purdue University. Previous studies (Kline & Rock, 1968; Seashore, 1962) have already suggested that sex is a potential moderator variable. It also represents an ideal choice, since it involves no classification problems.

The criterion was first-semester grade point average, and the five predictors were (a) high school rank, (b) verbal score on the Scholastic Aptitude Test (SAT), (c) mathematics score on SAT, (d) English achievement score on the College Entrance Examination Board (CEEB), and (e) mathematics score on the CEEB. The criterion and predictors were chosen because of their widespread use.

The analysis of variance described in the preceding section was performed (see Table 2). Tests of significance for both difference of position and difference of regression use the combined residual mean square for the denominator. Both terms were significant, but the difference of positions ( $p < .001$ ) was ap-

TABLE 1  
ANALYSIS OF VARIANCE FOR HETEROGENEOUS REGRESSION

Source	df	SS
Overall regression	$p$	$SS_{R(O)}$
Difference of positions	$k - 1$	$SS_{T(O)} - SS_{T(C)} - SS_{R(O)} + SS_{R(C)}$
Difference of regression	$(k - 1)p$	$\sum_{i=1}^k SS_{R_i} - SS_{R(C)}$
Combined residual	$N - kp - k$	$SS_{T(C)} - \sum_{i=1}^k SS_{R_i}$
Total	$N - 1$	$SS_{T(O)}$

parently a more powerful effect than the difference of regression ( $p < .05$ ).

An examination of the regression values in each of the two subgroups showed a large difference between the multiple-regression coefficients in the two groups ( $R = .50$  for males,  $R = .69$  for females; see Table 3). An examination of the intercept values for the two groups ( $A = 183.11$  for the males,  $A = 76.76$  for females) showed the large difference of positions suggested by the analysis of variance. Some differences also existed between the regression weights; however, an examination of the residuals shows very little overall improvement in prediction by subgrouping.

The data were reanalyzed, adding sex as a predictor. This resulted in no improvement in prediction, with the multiple correlation remaining exactly the same. Sex is, therefore, clearly a moderator variable and not a predictor or suppressor variable.

TABLE 2  
ANALYSIS OF VARIANCE FOR  
MALE-FEMALE SUBGROUPS

Source	SS	df	MS
Overall regression	1461375	5	292275
Difference of positions	33137	1	33137**
Difference of regressions	47849	5	9569*
Combined residual	3186043	868	3670
Total	4728404	879	

\*  $p < .05$ .  
\*\*  $p < .001$ .

DISCUSSION

The concept of subgrouping is not unique to the present article. The method employed by Rock, Barone, and Linn (1967) is completely consistent with this concept. Ghiselli (1956, 1960a, 1960b, 1963) also employed subgroups but utilized them in a different manner. Recent criticisms (Velicer, 1972) suggest that the Ghiselli model may be extremely limited in application. Cleary's (1966) model may be conceived of as a subgrouping model where each individual forms a subgroup. The only model not consistent with the subgrouping concept is that of Saunders (1956). Ironically, since Saunders coined the term moderator, the Saunders model is actually a curvilinear regression model.

TABLE 3  
REGRESSION VALUES FOR SUBGROUPS  
AND OVERALL REGRESSION

	Male <sup>a</sup>	Female <sup>b</sup>	Overall <sup>c</sup>
$b_1$	1.49509	2.34581	1.80559
$b_2$	.62590	.68298	.59974
$b_3$	.63242	1.75510	.97073
$b_4$	.43707	1.19177	.97686
$b_5$	.98072	-.10663	.35131
$A$	183.11	76.76	151.04
$R$	.50	.69	.56
$SS_T$	3202298	1514580	4728404
$SS_R$	797163	733672	1461375
$SS_E^d$	2405134	780908	3267028

<sup>a</sup>  $N = 602$ .  
<sup>b</sup>  $N = 278$ .  
<sup>c</sup>  $N = 880$ .  
<sup>d</sup> Combined  $SS_E$  3186042.

Viewing moderator variables as classifiers in a heterogeneous regression system suggests that the best moderators would be discrete variables such as biodata items. A recent study bears this out. Gross (1970) discusses two types of moderator variables, perfect moderators and imperfect moderators. With perfect moderators there is no problem classifying people; the classes are discrete as in the preceding example. The imperfect moderator is one for which the classes overlap, and there is some chance of misclassification. Classification could be accomplished by means of a discriminant function analysis or a grouping procedure. Assuming multivariate normal distribution in the subgroups, Gross (1970) performed a Monte Carlo study which indicated that even with a relatively low probability of misclassification ( $p = .20$ ), an overall regression equation performed more satisfactorily than moderated regression for even widely diverse populations.

In addition, moderated regression should be considered a large sample method. Practitioners often recommend that a sample of several hundred be used if the investigator hopes to achieve stable regression weights. If we arbitrarily set the minimum sample size at 200, this means that we would need  $N \geq 200 \times k$  where  $k$  is the number of subgroups to achieve stable regression weights in all subsamples. We might need a much larger sample if one of the subsamples contains a disproportionately small number of Ss.

Viewing moderated regression simply as a type of heterogeneous regression provides several distinct advantages: (a) it brings the moderator variable within the realm of a known mathematical system; (b) it provides useful tests of significance; (c) it provides some indication of what type of variables are likely to be good moderators; and (d) it gives some idea of the necessary sample size. Heterogeneous

regression is the logical outgrowth of the hierarchical grouping procedures of Rock et al. (1967) and encompasses several other models as special cases.

#### REFERENCES

- CLEARY, T. A. An individual differences model for multiple regression. *Psychometrika*, 1966, 31, 215-224.
- GHISELLI, E. E. Differentiation of individuals in terms of their predictability. *Journal of Applied Psychology*, 1956, 40, 374-377.
- GHISELLI, E. E. Differentiation of tests in terms of the accuracy with which they predict for a given individual. *Educational and Psychological Measurement*, 1960, 20, 675-684. (a)
- GHISELLI, E. E. The prediction of predictability. *Educational and Psychological Measurement*, 1960, 20, 3-8. (b)
- GHISELLI, E. E. Moderating effects and differential reliability and validity. *Journal of Applied Psychology*, 1963, 47, 81-86.
- GROSS, A. L. A Monte Carlo study of moderated regression. Unpublished doctoral dissertation, Purdue University, 1970.
- KLEIN, S. P., & ROCK, D. A. Predicting multiple criteria of creative achievements with multiple moderators. Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago, February 1968.
- KULLBACK, S., & ROSENBLATT, H. M. On the analysis of multiple regression in  $k$  categories. *Biometrika*, 1957, 44, 67-83.
- ROCK, D. A., BARONE, J. L., & LINN, R. L. A FORTRAN computer program for a moderated stepwise prediction system. *Educational and Psychological Measurement*, 1967, 27, 709-713.
- SAUNDERS, D. R. Moderator variables in prediction. *Educational and Psychological Measurement*, 1956, 16, 209-222.
- SEASHORE, H. C. Women are more predictable than men. *Journal of Counseling Psychology*, 1962, 9, 261-270.
- VELICER, W. F. A comment on the general inapplicability of Ghiselli's moderator system for two predictors. *Journal of Applied Psychology*, 1972, 56, xxx-xx.
- WILLIAMS, E. J. *Regression analysis*. New York: Wiley, 1959.

(Received November 22, 1971)