MULTIPLE LINEAR REGRESSION VIEWPOINTS
Volume 2, Number 4
March, 1972

A publication of the Special Interest Group on Multiple Linear Regression

of the American Educational Research Association

Table of Contents

IMPORTANT NOTICE

The meeting times for the SIG in Multiple Linear Regression at the AERA Convention in Chicago have been changed to Friday, April 7. (They have been incorrectly listed as being on Wednesday in the AERA announcements). The two letters following this notice will give more details on the SIG activities.

Memo from Koplyay and Bottenberg, AFHRL/PH 8 March 1972
John D. Williams, University of North Dakota
Earl Jennings, University of Texas
Samuel R. Houston, University of Northern Colorado
Keith A. McNeil, Southern Illinois University

SUBJECT: SIG/Multiple Linear Regression

1. The SIG/MLR meetings have been rescheduled. Even though I had made careful arrangements to have the meetings on Friday, 7 April, they have been erroneously scheduled into the program for Wednesday, 5 April.

2. The correct meeting information is as follows:

Friday, April 7, 9:00 - 10:30 Session #27.5
Boardroom (La Salle)
Multiple Linear Regression Applied to
Automatic Interaction Detection and
to the Analysis of Covariance, Residual
Gains and Gain Scores

Friday, April 7, 10:45 - 12:15 #28.21
Boardroom (La Salle)
Business Meeting

3. I assume that all participants will be available on Friday morning, 7 April. I will contact everyone on Thursday for last minute details. Try to have hard copy handouts for your presentation rather than just talk. If you need other equipment, let me know.

Joe H. Ward Jr.

JOE H. WARD, JR.

Southern Illinois
University at Carbondale

CARBONDALE, ILLINOIS 62901

Department of Guidance and Educational Psychology

March 8, 1972

Dear SIG:MLR Members:

Block out 9:00-12:15 on Friday, April 7, for SIG:MLR activities. AERA Sessions 27.5 and 28.21 will both be held in the Boardroom of the La Salle Hotel. In our Business Meeting at 10:45 we will be electing a new President and a new Secretary/Treasurer. Discussion of the format of VIEWPOINTS and the paper sessions will also be in order. Frankly, I feel that we haven't made optimal use of the inexpensive communications of ideas afforded by VIEWPOINTS and John Williams' editorial contributions.

Your treasurer, Bill Connett, has become burdened with an excess balance. We have decided to solve his problem by liquidating some of the funds in a Social Hour immediately following the Business Meeting.

Several other sessions may be of interest to SIG:MLR members: 5.10; 5.11; 7.09; and 21.12. See you at AERA.

Keep regressing,

Keith A. McNeil

Keith A. McNeil

A Suggested Format for the Presentation of Multiple Linear Regression

Isadore Newman

University of Akron

An arguement that has been presented by researchers who prefer to use analysis of variance (ANOV) over multiple regression analysis is that ANOV is easy to present and has established a standard Sum of Squares table. This table is familiar to anyone who has read the educational and psychological literature and because of its common use there has been less ambiguity regarding the symbols used and interpretation of the table.

On the other hand, multiple regression, when presented in the literature, has always been formated in an idiosyncratic manner. In addition, the format rarely presents all the relevant information in a concise, easy to inspect manner. Instead, one tends to find himself thumbing through pages of the article to find the relevant information.

Table 1[1] is an example of a format which I would like to suggest as a somewhat standard form for the presentation of multiple regression models and the information required for their interpretation. The example chosen is from an unpublished dissertation by I. Newman (1971). This example may be somewhat longer than one someone would generally find, it therefore required an additional table, as stated in the table footnote, to explain the variables. In a smaller table, the variable explaination can be presented in a footnote at the bottom of the table.

---

[1] I first saw something very similar to this in a dittoed paper by Sommers, et al. (1969, Southern Illinois University).

I would appreciate any comments regarding the suggested format for the presentation of multiple regression and I would also appreciate any suggested modification that would enhance our ability to communicate multiple regression output in the most concise and easily interpretable form.

References

Newman, I. A multivariate approach to the construction of an attitude battery. Unpublished doctoral dissertation, Southern Illinois University, 1971.

TABLE # 1

Models, F-Ratings and R^2 For Predicting The Ratings Of The Author (Y_1)

Models	Models	R^2	df	F	P
Model 1 $Y_1 = a_0u + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5 + a_6x_6 + a_7x_7 + a_8x_8 + a_9x_9 + E_1$	Full	.11			
Restriction: $a_1=a_2=a_3=a_4=a_5=a_6=a_7=a_8=a_9$			8/300	4.46	.00004
Model 99 $Y_1 = a_0u + E_0$	Restricted	.00			
Model 1 $Y_1 = a_0u + a_1x_1 + \ldots a_9x_9 + E_1$	Full	.11			
Restriction: $a_9=0$ (interaction)			1/300	.22	.634
Model 2 $Y_1 = a_0u + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5 + a_6x_6 + a_7x_7 + a_8x_8 + E_2$	Restricted	.10			
Model 3 $Y_1 = a_0u + a_1x_1 + E_3$	Full	.002			
Restriction: $a_1=0$ (Black)			1/307	.64	.422
Model 99 $Y_1 = a_0u + E_0$	Restricted	.000			
Model 4 $Y_1 = a_0u + a_2x_2 + E_4$	Full	.001			
Restriction: $a_2=0$ (White)			1/307	.35	.555
Model 99 $Y_1 = a_0u + E_0$	Restricted	.000			
Model 5 $Y_1 = a_0u + a_3x_3 + E_5$	Full	.001			
Restriction: $a_3=0$ (N-infor.)			1/307	.25	.619
Model 99 $Y_1 = a_0u + E_0$	Restricted	.000			

(cont.)

TABLE #1 (cont.)

Model 6 $Y_1 = a_0u + a_4x_4 + a_5x_5 + a_6x_6 + a_7x_7 + a_8x_8 + a_9x_9 + E_6$	Full	.099			
Restriction: $a_4 = 0$ (affect$_1$)			1/303	3.119	.078
Model 7 $Y_1 = a_0u + a_5x_5 + a_6x_6 + a_7x_7 + a_8x_8 + a_9x_9 + E_7$	Restricted	.089			
Model 6 $Y_1 = a_0u + a_4x_4 + \ldots . a_9x_9 + E_6$	Full	.099			
Restriction: $a_5 = 0$ (affect$_2$)			1/303	.243	.622
Model 8 $Y_1 = a_0u + a_4x_4 + a_6x_6 + a_7x_7 + a_8x_8 + a_9x_9 + E_8$	Restricted	.098			
Model 6 $Y_1 = a_0u + a_4x_4 + \ldots . a_9x_9 + E_6$	Full	.099			
Restriction: $a_6 = 0$ (cognitive)			1/303	.045	.831
Model 9 $Y_1 = a_0u + a_4x_4 + a_5x_5 + a_7x_7 + a_8x_8 + a_9x_9 + E_9$	Restricted	.098			
Model 6 $Y_1 = a_0u + a_4x_4 + \ldots . a_9x_9 + E_6$	Full	.099			
Restriction: $a_7 = 0$ (response disposition)			1/303	25.713	.001
Model 10 $Y_1 = a_0u + a_4x_4 + a_5x_5 + a_6x_6 + a_8x_8 + a_9x_9 + E_{10}$	Restricted	.022			
Model 6 $Y_1 = a_0u + a_4x_4 + \ldots . a_9x_9 + E_6$	Full	.099			
Restriction: $a_8 = 0$ (response disposition$_2$)			1/303	.614	.433
Model 11 $Y_1 = a_0u + a_4x_4 + a_5x_5 + a_6x_6 + a_7x_7 + a_9x_9 + E_{11}$	Restricted	.097			

NOTE: The probability values (P) that are reported are for a two tail test of significance (see Table #4 for description of variables).

THE USE OF FACTOR REGRESSION IN DATA ANALYSIS

William E. Connett, Samuel R. Houston, and Dale G. Shaw

The educational researcher is often faced with preliminary sample data in which well-defined models or reasonably clear hypotheses dealing with interrelations between variables are lacking. His purpose at this point in time is to suggest interpretations that may be put to the test in later studies. Hence, the problem is exploratory, as opposed to descriptive or cause-effect, research; it is particularly associated with massive input-output research efforts in educational settings.

This paper describes a strategy for analyzing the relation between a dependent variable and a set of independent variables when the latter are not necessarily amenable to standard statistical treatment. This problem arises, for example, when our full regression model contains in the set of independent variables several variables which are highly intercorrelated. This is usually addressed as the problem of multicollinearity. Rather than give up on estimation procedures by classical regression methods, the educational researcher is encouraged to explore the possibilities of a principal components regression analysis or what we choose to call factor regression.

THE FACTOR REGRESSION PROCEDURE

The problem is to express the criterion variable as a function of a set of independent variables in which the intercorrelations between the various independent variables is near zero. The procedure involves restructuring the full regression model in such a way that the criterion variable is expressed as a function of several mutually orthogonal factor variables. This principal components-regression approach permits one then to investigate the unique contribution of each of the factor variables to explaining the dependent variable. For a detailed discussion of generating the restructured full regression model, see Massy (1965).

Obtaining factor scores

The procedure begins with the complete orthogonal factoring of the set of m predictors into an m X m factor matrix. For the sake of interpretation, this factor matrix may be rotated, but only with a rotation, such as Varimax, which preserves orthogonality. The next step is to standardize the predictor scores for each predictor variable. This matrix of standard scores is then postmultiplied by the factor matrix to obtain the factor scores.

The factor score matrix is orthogonal which means that the matrix of intercorrelations among the factor scores should be the identity matrix. Also, since the original predictor set was converted to z-scores, the means for the factor scores are all zero and the standard deviations are equal to the indices of factorization for the factors.

The regression model

If a regression model is cast, regressing some criterion variable onto the set of factor score predictors, several interesting properties are noted. The beta weight for a predictor is equal to the validity for that predictor. The R^2 value for any model is equal to the sum of the squares of the beta weights for the model. The exclusion of a factor score variable from the predictor set will result in a drop in R^2

equal to the square of the beta weight for the variable dropped. And, perhaps most important, the dropping of any predictor variable from the predictor set will not affect the beta weights (predictive contribution in this case) of any of the other variables. These properties are demonstrated in the following example. Slight discrepancies are attributed to rounding errors.

THE EXAMPLE

For each of the 120 college students 8 measures were obtained from their high school and college records. These 8 variables are described below.

Criterion variable (Y)

The last recorded college cumulative grade point average computed on a 4.0 scale was used as each student's criterion score.

The Predictors were:

1. A fluctuation score indicative of the variance of a student's high school grades.
2. The actual variance of high school grades computed as the average of the sum of the squares of deviations about the mean.
3. An adjusted variance score based on the actual variance score.
4. The student's high school GPA based on all high school courses.
5. The GPA based on only high school English, Social Science, Natural Science, and Mathematics grades.
6. The score made on the ACT test before admission to college.
7. The high school rank of each student within his graduating class.

Table 1 contains the varimax factor solution for the predictor set of 7 variables with the associated indices of factorization. Communalities were set equal to 1.00 and 7 factors were extracted to provide a complete factoring of the predictor set.

The varimax solution provided a simple structure which was easily interpreted. Factor 1 loaded highly on GPA measures, factor 2 on variance measures, and factor 3 on the ACT score. The remaining factors were interpreted as decomposition factors which broke down previous groupings of predictors.

TABLE 1

VARIMAX FACTOR SOLUTION

VAR. NO.	1	2	3	4	5	6	7
1	-.08	.43	-.05	.90	-.01	.00	.00
2	-.28	.86	-.13	.35	.01	.22	.00
3	.10	.96	-.08	.20	-.01	-.13	.00
4	.98	-.04	.09	-.07	.04	-.03	.11
5	.98	-.07	.13	-.06	.04	-.01	-.11
6	.15	-.12	.98	-.04	.02	-.01	.00
7	.78	-.01	.07	-.03	.62	.00	.00
INDEX	2.66	1.87	1.02	.98	.38	.07	.03

TABLE 2
SUMMARY STATISTICS FOR THE FACTOR SCORES

FACTOR NUMBER	MEAN	STD. DEV.	VALIDITY
1	0.00	2.66	0.52
2	0.00	1.87	0.11
3	0.00	1.02	-0.11
4	0.00	0.98	-0.10
5	0.00	0.38	0.03
6	0.00	0.07	0.10
7	0.00	0.03	0.07

The factor scores were then computed by post-multiplying the standardized raw predictor scores by the factor matrix. Table 2 contains the means, standard deviations, and validities for the 7 factor scores. Note that the standard deviations are identical to the indices of factorization reported in Table 1 and also that the largest validity is of course for the first factor since both the criterion and first factor score are GPA measures.

TABLE 3
INTERCORRELATIONS BETWEEN FACTOR SCORE VARIABLES

VAR. NO.	1	2	3	4	5	6	7
1	1.00						
2	-.01	1.00					
3	.00	-.01	1.00				
4	.00	.00	-.01	1.00			
5	.00	.00	.00	.02	1.00		
6	.00	.00	.00	..00	-.01	1.00	
7	.00	.00	.00	..00	.00	-.04	1.00

The matrix of intercorrelations between the factor scores is given in Table 3. Observe that the matrix is not quite a perfect identity matrix. This fact in connection with the type of iterative technique used to obtain the beta weights leads to slight discrepancies in the regression analysis.

The RSQ for the full model with all factor score variables included in the predictor set was .3236. This was identical to the RSQ obtained for the model in which Y was regressed onto all of the raw predictor variables as one would have done in a usual regression analysis. This indicates that the factor scores retain all of the information useful in a regression analysis possessed by the raw predictor scores when the factoring is done completely.

TABLE 4
SUMMARY DATA FROM REGRESSION ANALYSIS

VAR.	(1) BETA WT IN FULL MODEL	(2) $(\text{BETA WT})^2$	(3) RSQ	(4) DROP IN RSQ FROM FULL
1	.5245	.2751	.0489	.2747
2	.1102	.0121	.3113	.0123
3	-.1107	.0123	.3112	.0124
4	-.0987	.0097	.3138	.0098
5	.0344	.0012	.3224	.0012
6	.0988	.0098	.3139	.0097
7	.0717	.0051	.3185	.0052

Columns 1 and 2 in Table 4 are the beta weights and their squares obtained for the full model for the factor scores. Column 3 gives the RSQ value for the restricted model with the variable in that row removed from the predictor set. An entry in column 4, then is the drop in RSQ from the full model to the indicated restricted model which gives an indication of the unique contribution of that factor score to the prediction of the criterion. Note that columns 2 and 4 are almost identical supporting the premise that the drop in RSQ is equal to the square of the beta weight for the variable dropped.

Since the factor scores are orthogonal their intercorrelations are all zero. As soon as one computes the validities for the factor scores he is essentially done since the beta weights to be determined in the regression analysis are simply the validities. The RSQ value then for a model using any subset of the factor scores is simply the sum of the squares of the validities for those factor scores included in the predictor set.

SUMMARY COMMENTS

Regression upon principal components appears to be worthwhile during the exploratory phases of empirical research. It permits a systematic analysis of data in situations where the problem of multicollinearity would otherwise make data analysis quite difficult. As Massy (1965) has indicated, these procedures are not a substitute for the long established principals of statistical inference, those being hypothesis building and testing; rather, the procedures discussed here provide the data organization preliminary to hypothesis development and testing.

REFERENCES

Massy, W.F. Principal components regression in exploratory statistical research. American Statistical Association Journal, 1965,60,234-256.