DOCUMENT RESUME

ED 137 382 TM 006 188

AUTHOR Shrestha, Gambhir
TITLE Second Order Regression Model Applied to 1972-73 Florida Statewide Assessment Program.
PUB DATE [Apr 77]
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Academic Achievement; Correlation; *Educational Assessment; Elementary Education; Grade 6; Mathematical Models; Mathematics; *Multiple Regression Analysis; *Predictor Variables; *School Districts; State Programs; Statistical Analysis
IDENTIFIERS Florida Statewide Assessment Program

ABSTRACT

A stepwise regression technique was used to analyze assessment data while taking differences in nonschool variables across districts into account. The primary purpose of this investigation was to determine whether the inclusion of quadratic and/or interaction terms in a regression model would improve the prediction of school district average score. Results indicated that interaction and quadratic terms improve the prediction of district averages. The second purpose of the investigation was to illustrate certain concepts of regression techniques while attempting to determine the quadratic and/or interaction effects in the regresssion model. (Author)

SECOND ORDER REGRESSION MODEL APPLIED TO 1972-73 FLORIDA STATEWIDE ASSESSMENT PROGRAM

BY

GAMBHIR SHRESTHA

Introduction

For the past three years, the Florida Statewide Assessment Program has been gathering data on basic cognitive skills of mathematics and communication domain from a sample of students in each of the sixty-seven (67) districts in Florida. One of the primary uses of the assessment data is to determine how students in a given district are progressing towards the mastery of certain objectives. The basic question, 'How well is district X doing?' is answered by comparing the district percentage with the state average. The district performance indicator obtained in this way did not control for differing community and student background inputs across districts and might mistakenly or unjustly give the district blame or credit.

The community and student background inputs are measured by any number of socio-economic or socio-cultural variables such as family income, parents' educational levels and parents' occupations. Those represent 'hard-to-change' variables and in general are related to achievement to a greater degree than are manipuable variables such as class size, teacher experience, etc. Thus, any attempt to examine the effectiveness of a district's educational program must control for the non-school variables in order that meaningful interpretation can be made. On the basis of these findings, the Florida Statewide Assessment Program has begun analyzing the assessment data while taking into account of differences in non-school variables across districts.

Statement of the Problem

In the age of the electronic computer, many problems are being solved using multiple correlation and regression techniques which would never have been attempted had electronic computers not been available. However, with the computer doing the calculations, problems can be solved without the manipulation being fully understood by the person employing the technique. Therefore, there is a need for a discussion of certain concepts of multiple correlation and regression techniques to prevent the user of these techniques from reaching erroneous conclusions. This article attempts to fill this need.

In the course of analysis, typical of the kinds of problems that were encountered by the author was whether to create complex variables which would account for interactions between simple input variables or to use more easily explained variables. The purpose of this study was two-fold. Firstly, it was to determine whether the inclusion of quadratic and/or interaction terms in a regression model would improve the prediction of district score as represented by the multiple correlation. The second purpose of this investigation was to illustrate the step-wise regression procedure while attempting to determine the quadratic and/or interaction effects in the regression model.

Data for Analysis

The total mean score in grade 6 for mathematics was selected as the criterion (output) variable. There are sixty-seven (67) observed scores, one for each school district. The score for each district was calculated from the 1972-73 Statewide Assessment results and is shown together with its standard deviation in Table 1.

To obtain a pool of potential prediction (input) variables, the lists of variables contained in the Accreditation files, the Bureau of Finance files and U.S. Census data were scanned for those variables which might relate to achievement in the Statewide Assessment Program. The final selection of 13 predictor variables that were analyzed are listed in Table 2, and their means and standard deviations are given in Table 3.

Procedure for Analysis

In order to compare the usefulness of first-order and second-order regression equations, four models, I, II, III and IV, were developed as shown in Table 4. Each model was developed using a step-wise multiple regression program (BMDOZR) on an IBM 370 Computer. The computational details of the method is illustrated below using the results for the linear model I.

The first step is to select one of the thirteen (13) predictor variables. One way to choose the first variable would be to perform thirteen (13) separate simple regressions and compute the F-ratio using

$$F = \frac{n-2}{1} \quad \frac{R^2_{1.i}}{1-R^2_{1.i}}$$

$$F = \frac{n-2}{1} \quad \frac{\text{Sum of squares due to regression}}{\text{Sum of squares due to residual}}$$

$$= \frac{n-2}{1} \quad \frac{\text{Regr SS}_i}{\text{Resd SS}_i}$$

Where $R_{1.i}$ denotes the multiple correlation coefficient between the criterion variable X_1 and the predictor variable X_i, i = 2, 3, ..., 13, 14. The sum of squares (SS) in equation (1) has subscripts i to indicate X_i is the predictor variable. The F-value obtained from equation (1) can be used to test the null hypothesis H_0: $\beta_i = o$. The selection of one of thirteen (13) variables depends upon the magnitude of its F-value; a variable with the highest F-value would be used as the predictor variable. Table 5 indicates that X_2 is the variable with the highest F-value.

In order for a variable to be included in the analysis, the F-value for the variable must exceed some predetermined value. The preassigned F-value can be set quite low; sometimes it is set as low as F = 0.001 so that one is almost certain to get a variable included. In this analysis, significance of β_i was tested at an alpha (α) level of 0.05. In order to be significant with $\alpha = 0.05$, the F-value has to be higher than 4.00.

The next step in the analysis consists of choosing the second predictor variable to be included in the regression analysis. One way to do that is to compute the partial correlation coefficients $r_{1i.2}$, i=3,4, ... 13, 14 using the formula

$$r_{1i.j} = \frac{r_{1i} - (r_{1j})(r_{ij})}{\sqrt{1-r^2_{1j}}\ \sqrt{1-r^2_{ij}}}$$

The partial coefficients, $r_{1i.2}$, measure the relationship between the criterion variable X_1 and each of the remaining predictor variables, X_3, X_4 ... X_{14}, while controlling for the variable X_2. It was necessary to control for X_2 in order to take out its effect since the variable X_2 has already been included in the equation in the first step of the analysis. The second predictor variable to be included would be that variable which explains most of the remaining variation in the criterion variable X_1. This variable is the one with the highest partial correlation.

An equivalent way of choosing the second variable to be included in the analysis is to compute the multiple correlation coefficient $R^2_{1.2i}$ (i=3,4,...14) for each possible two variable regression models containing the variable X_2 and one additional variable X_i. The coefficient $R^2_{1.2i}$ is computed using the formula

$$R^2_{1.2i} = R^2_{1.2} + r^2_{1i.2} \quad (1-R^2_{1.2})$$

variation explained by X_2 and X_i = variation explained by X_2 + additional variation explained by X_i x variation unexplained by X_2

The variable with the highest multiple correlation is the one with the highest partial correlation. In addition, the variable with the highest multiple correlation is the one with the highest F-value. The F-ratio is computed using the formula

$$F = \frac{n-3}{1} \frac{R^2_{1.2i} - R^2_{1.2}}{1 - R^2_{1.2i}}$$

$$= \frac{n-3}{1} \frac{\text{Regr } SS_{2i} - \text{Regr } SS_2}{\text{Resd } SS_{2i}}$$

which is distributed as F with 1 and n-3 degrees of freedom. The correlations $r^2_{1i.2}$ and $R^2_{1.2i}$ are given, together with F-values, in Table 6. It can be seen from the table that both correlations (partial and multiple) and the F-value for X_6 are the highest. Thus, X_6 is the second variable included in the analysis since its F-value (21.63) exceeds the predetermined value, 4.00.

Having included the variable X_6 in the analysis, the next step in the procedure is to examine whether the variable X_2, included in the first step, is needed for the regression equation any longer. This is done by first regressing the criterion variable X_1 on X_6, resulting in $R^2_{1.6}$, and then examining whether adding the variable X_2 produces a significantly larger coefficient $R^2_{1.26}$. The increase in prediction is measured by the F-ratio

$$F = \frac{n-3}{1} \frac{R^2_{1.26} - R^2_{1.6}}{1 - R^2_{1.26}}$$

$$= \frac{n-3}{1} \frac{\text{Regr } SS_{26} - \text{Regr } SS_6}{\text{Resd } SS_{26}}$$

$$= 39.22$$

Since the F-value, 39.26, is greater than the predetermined value of F=4.00, the variable X_2 still contributes enough to be included in the analysis.

Having included X_2 and X_6, the step-wise procedure next computes $r^2_{1i.26}$ (i=3,4,5,7,8,...13,14). These coefficients measure the relationship between the criterion variable X_1 and each of the eleven remaining variables while controlling for the variables X_2 and X_6 which are already included in the analysis. The partial coefficients are listed in Table 7. It can be seen from the Table 7 that X_8 has the highest partial coefficient, -.2571. Since X_8 has the F-value, 4.459, greater than the pre-set value of 4.00, it is the third variable to be included in the analysis. Having included X_8, the procedure next examines whether X_2 and X_6 are needed any longer in the analysis. X_2 will be excluded from the analysis if the F-ratio

$$F = \frac{n-4}{1} \; \frac{R^2_{1.268} - R^2_{1.68}}{1 - R^2_{1.268}}$$

is smaller than the pre-set value, 4.0. Similarly, X_6 will be excluded if the F-ratio

$$F = \frac{n-4}{1} \; \frac{R^2_{1.268} - R^2_{1.28}}{1 - R^2_{1.268}}$$

is less than 4.0. In Table 7, the F-values for X_2 and X_6 are equal to 45.71 and 27.25 respectively. Since both of these values are greater than the pre-set value, 4.0, X_2 and X_6 are retained in the analysis after the inclusion of X_8.

This procedure of inclusion of the next variable and exclusion of possible variables already included continues until no new variable contributes enough to the multiple correlation to be included in the regression model. Of thirteen (13) predictor variables, only three variables, X_2, X_6 and X_8, contribute enough to the multiple correlation to be included in the model I. The three variables from Model I were forced to remain in the prediction equations in the Models

II, III and IV. This was necessary in order to make the statistical comparison of the linear model and other models designed to measure curvilinear relationships.

Model II was developed by including the squared terms of each of the predictor variables plus the forced linear terms X_2, X_6 and X_8 from Model I. The new variables included in Model II are the variable X_4 and the square of X_2 denoted by $X_2.X_2$. The Model III was investigated by including all possible interaction terms and the variables X_2, X_6 and X_8 from Model I. An interaction variable is the product of two predictor variables, denoted by $X_i.X_j$, where i, j = 2,3, ... 13, 14. The thirteen (13) predictor variables give rise to 76 possible interaction terms. Since the number of interaction variables X_iX_j exceeds the number of cases (n=67), interaction variables were systematically analyzed in groups of 25 variables along with the variables X_2, X_6 and X_8. This was necessary in order to avoid overfitting the regression equation. Model III included the interaction terms X_2X_5 and X_5X_{10} plus the three linear terms X_2, X_6, and X_8. The fourth model included the significant linear, quadratic, and interaction terms included in the previous models. Namely, the variables X_2, X_4, X_6, X_8, X_2X_2, X_2X_5 and X_5X_{10} were included in Model IV.

Comparison of four Models

Since the purpose of this study was to investigate whether the inclusion of square and/or interaction terms in a regression model would be an improved model in terms of predictability, the improvement was determined by comparing the result from the Model I against the results from the Models II, III and IV. There are several criteria which can be applied to make this comparison. One of the most common criteria is to examine the square of multiple correlation coefficient, R^2, defined by

$$R^2 = \frac{\text{Sum of squares due to regression}}{\text{Total sum of squares}}$$

It is often stated as a percentage, 100 R^2. The larger it is, the better the fitted equation explains the variation in the data. The value of R^2 resulting from each of the four models is compared in Table 8. Thus, we see a substantial increase in R^2 in the second-order model.

A second way of determining the predictability of the four models is to compare the standard error of estimate S, in relation to the mean of the 67 observed scores. The value of S as a percentage of $\overline{X}_1 = 58.4656$ for each of the four models is shown in Table 8. Examination of this statistic indicates that the inclusion of curvilinear effects in the linear model has reduced the standard error of estimate from 5.8 to about 5.3 percent of the mean observations.

Table 1

1972-73 Means and Standard Deviation for Grade 6 Mathematics

District No.	Mean Score	Standard Deviation
1	56.3	1.02
2	58.8	1.65
3	62.6	0.84
4	53.3	1.22
5	69.6	0.67
6	61.5	0.55
7	66.1	1.26
8	63.6	1.33
9	60.6	1.07
10	65.9	0.98
11	56.9	0.99
12	53.7	1.56
13	64.4	0.28
14	57.6	1.47
15	46.0	1.97
16	57.2	0.62
17	61.9	0.66
18	50.0	1.34
19	59.0	1.42
20	49.1	1.04
21	67.9	1.86
22	47.4	1.38
23	59.2	1.49

Table 1 Cont'd

District No.	Mean Score	Standard Deviation
24	52.1	1.81
25	50.0	1.23
26	56.9	1.54
27	62.0	1.06
28	51.0	1.29
29	60.1	0.59
30	61.6	1.13
31	58.3	0.84
32	59.1	0.98
33	47.0	1.69
34	55.6	1.93
35	60.3	1.02
36	57.9	0.90
37	58.1	1.06
38	56.0	1.18
39	56.9	1.87
40	50.0	1.07
41	61.6	0.89
42	55.7	0.88
43	54.9	1.21
44	62.6	1.05
45	59.7	1.28
46	66.0	0.88
47	55.1	1.39
48	63.6	0.69
49	62.0	1.27

Table 1 Cont'd

District No.	Mean Score	Standard Deviation
50	58.8	0.45
51	62.8	1.04
52	63.2	0.66
53	62.2	0.67
54	55.6	1.05
55	56.9	0.92
56	56.1	1.22
57	65.6	0.87
58	63.9	0.69
59	64.1	0.98
60	59.3	1.43
61	55.6	1.12
62	57.4	1.25
63	55.0	1.72
64	64.2	0.85
65	52.9	1.64
66	66.0	1.28
67	54.9	1.47

Trble 2: Description of Prediction Variables

Minority Enrollment. Percent of pupil enrollment that is non-white, Spanish speaking, Oriental or American Indian.

Source: Quantitative Report, ACC-1, Accreditation Section, DOE.

Variable Number: X_2 or simply 2. Variable Symbol: MNRE

Average Daily Attendance. The number of pupils in average daily membership, grades K - 12 for the year 1972-73.

Source: Quantitative Report, ACC-1, Accreditation Section, DOE.

Variable Number: X_3 or simply 3. Variable Symbol: ADM

Poverty Level. Approximate percent of the student body from families with an average annual income of less than $3,000.

Source: Quantitative Report, ACC-1, Accreditation Section, DOE.

Variable Number: X_4 or simply 4. Variable Symbol: FMI

White Collar Occupation. Approximate percent of the student body from families with 'white collar' occupations include professional, technical, clerical and kindred worker. A more detailed example can be found in the source.

Source: Quantitative Report, ACC-1, Accreditation Section, DOE.

Variable Number: X_5 or simply 5. Variable Symbol: OCP

Average Family Income. The combined income of all families divided by the number of families in the district.

Source: United States Census of Population, 1970: General Social and Economic Characteristics, Florida Summary. Series PC(1) - C11, Bureau of Census, United States Department of Commerce, April 1972.

Variable Number: X_6 or simply 6. Variable Symbol: AVGI

Per Capita Income. This is the mean income computed for every man, woman, and child in a particular group. It is derived by dividing the total income of a particular group by the total population in that group.

Source: U.S. Census of Population, 1970. Series PC(1) - C11, Bureau of Census, U.S. Department of Commerce.

Variable Number: X_7 or simply 7. Variable Symbol: INCP

Housing. Percent increase in housing units, 1960-70.

Source: Florida Statistical Abstract, 1971, Bureau of Economic and Business Research, University of Florida.

Variable Number: X_8 or simply 8. Variable Symbol: HSNG.

School Education. This is the median school years completed for the population 25 years of age and older of the district.

Source: U.S. Census of Population, 1970 Series PC(1) - C11, Bureau of Census, U.S. Department of Commerce.

Variable Number: X_9 or simply 9. Variable Symbol: SCHED

College Education. Percent of 1970 male population, with 1 to 3 years of college completed.

Source: U. S. Department of Commerce, Bureau of Census, PC(1) - C11.

Variable Number: X_{10} or simply 10. Variable Symbol: COLED

Post College Ed. Percent of 1970 male population, with 4 or more years of college completed.

Source: U. S. Department of Commerce, Bureau of the Census, PC(1) - C11.

Variable Number: X_{11} or simply 11. Variable Symbol: CRAD

Percent of Population Classified as Urban. The percent of the district's total resident population living in urban places and urban areas according to the 1970 census.

Source: U. S. Department of Commerce, Bureau of the Census, PC(1) - C11.

Variable Number: X_{12} or simply 12. Variable Symbol: URBN

Sixty-five Years and over. The percent of 1970 population with 65 years and over.

Source: U.S. Department of Commerce, Bureau of Census, PC(1) - B11.

Variable Number: X_{13} or simply 13. Variable Symbol: SXTY

Free Lunch. Approximate percent of the student body receiving free or reduced lunch.

Source: Food and Nutrition Service, Florida Department of Education.

Variable Number: X_{14} or simply 14. Variable Symbol: LNCH

Table 3

Means and Standard Deviations of the Predictor Variables in Table 2

Variable Name	Mean	Standard Deviation
Minority Enrollment	24.57	5.35
Average Daily Attendance	24.01	14.96
Poverty Index	23.40	42.90
White Collar Occupation	27.90	13.80
Average Family Income	6.23	13.13
Capita Income	2.47	1.63
Housing	39.50	0.60
School Education	10.80	35.80
College Education	8.60	1.30
Post College Education	9.40	3.50
Urban	42.80	5.50
Sixty-Five Years	13.60	30.70
Free Lunch	38.70	6.80

Table 4

Four Regression Models

Model I (first-order model)

Model 1 (First Order):

$$X_1 = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \text{-----------} + \beta_{14} X_{14} + \varepsilon_1$$

MODEL 2 (Quadratic):

$$X_1 = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \text{-----------} + \beta_{14} X_{14} + \sum_{i=2}^{14} \beta_{ii} X_i^2 + \varepsilon_2$$

MODEL 3 (Interaction):

$$X_1 = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \text{------------} + \beta_{14} X_{14} + \sum_{\substack{i=2 \\ i \neq j}}^{14} \sum_{j=2}^{14} \beta_{ij} X_i X_j + \varepsilon_3$$

MODEL 4 (Second-Order):

$$X_1 = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \text{------------} + \beta_{14} X_{14} + \sum_{i=2}^{14} \sum_{j=2}^{14} \beta_{ij} X_i X_j + \varepsilon_4$$

Where: X_1 is the criterion variable

β_1, β_2, ----- β_{14} and β_{ij} are unknown regression coefficients. These are estimated by the quantities b_1, b_2, -----b_{14} and bij by requiring the error sum of squares to be minimized.

X_2, X_3, -----X_{14} are the values of predictor variables.

X_i X_j $(i, j = 2, 3, \ldots. 14)$ is the product of the value corresponding to X_i and the value corresponding to X_j.

And ε_i is the residual for Model i.

Table 5

Data for selection of variables in Step #1

			Variable numbers as listed in Table 2									
	2	3	4	5	6	7	8	9	10	11	12	13	14
$R_{1.i}$	-.658	.325	-.617	.531	.563	.437	.273	.480	.476	.368	.416	.152	-.650
$R^2_{1.i}$	.433	.106	.381	.282	.317	.191	.075	.231	.226	.135	.173	.023	.423
$1-R^2_{1.i}$	.567	.894	.619	.718	.683	.809	.925	.769	.773	.865	827	.977	.577
^{2} $1.i/R^2_{1.i}$	.764	.118	.615	.393	.464	.236	.081	.299	.293	.156	209	.024	.732
F	49.6	7.67	39.9	25.5	30.2	15.3	5.23	19.5	19.1	10.1	13.6	1.53	47.5

Table 6

Data for selection of variables in Step #2

			Variable numbers as listed in Table 2									
	3	4	5	6	7	8	9	10	11	12	13	14
$r_{1i.2}$	.415	-.397	.382	.503	.346	.008	.414	.499	.404	.435	.038	-.309
$r^2_{1i.2}$	.173	.157	.146	.253	.119	.001	.172	.249	.163	.189	.002	.095
$R^2_{1.2i}$	.531	.522	.516	.576	.501	.433	.531	.574	.525	.541	.434	.487
F	13.3	11.9	10.9	21.6	8.74	0.01	13.2	21.3	12.5	15.0	0.09	6.78

Table 7

Data for selection of variables in step #3

```
STEP NUMBER   .   3
VARIABLE ENTERED      8

MULTIPLE R                 0.7771
STD. ERROR OF EST.         3.4030

ANALYSIS OF VARIANCE

                          DF     SUM OF SQUARES     MEAN SQUARE      F RATIO
          REGRESSION       3          1112.557         370.852       32.024
          RESIDUAL        63           729.561          11.580

                      VARIABLES IN EQUATION                .                     VARIABLES NOT IN EQUATION
                                                           .
   VARIABLE        COEFFICIENT   STD. ERROR   F TO REMOVE  .     VARIABLE     PARTIAL CORR.     TOLERANCE      F TO ENTER
                                                           .
  (CONSTANT          55.88690 )                            .
  MNRENR   2         -0.23298      0.03446     45.7068 (2) .     ADA      3      0.04329         0.4860        0.1164 (2)
  AVGINC   6          1.55221      0.29735     27.2504 (2) .     FMLINC   4     -0.12329         0.4259        0.9569 (2)
  HOUSNG   8         -0.02970      0.01407      4.4594 (2) .     OCUPTN   5     -0.00672         0.3913        0.0028 (2)
                                                           .     INCPTA   7     -0.13520         0.1679        1.1543 (2)
                                                           .     SCHLED   9     -0.00561         0.2359        0.0919 (2)
                                                           .     COLGED  10      0.19284         0.2121        2.3947 (2)
                                                           .     COLGRD  11      0.05048         0.3740        0.1584 (2)
                                                           .     URBAN   12      0.07637         0.3773        0.3638 (2)
                                                           .     SIXYRS  13      0.00905         0.8616        0.0051 (2)
                                                           .     FRLNCH  14     -0.09265         0.3382        0.5368 (2)
```

Table 8

Data for Comparison of Four Regression Models

MODEL #	VARIABLES IN THE MODEL	R^2	RESIDUAL Sum of Squares			STANDARD ERROR OF EST.		F-VALUE
			Sum of Squares	DF	Mean Square	S	$100S/\overline{X}_1$ $\overline{X}_1 = 58.46$	
I	X_2, X_6, X_8	60.4	729.56	63	11.58	3.41	5.80	32.02
II	$X_2, X_6, X_8,$ X_4, X_2X_2	67.1	606.09	61	9.94	3.15	5.39	24.88
III	$X_2, X_6, X_8,$ $X_2X_5,\ X_5X_{10}$	65.9	627.63	61	10.29	3.21	5.49	23.61
IV	$X_2,\ X_4,\ X_6,\ X_8,$ $X_2X_2,\ X_2X_5,$ X_5X_{10}	69.2	567.99	59	9.63	3.10	5.30	24.88