VOLUME 16, NUMBER 1

SPRING 1988

MULTIPLE LINEAR REGRESSION VIEWPOINTS

A publication of the Special Interest Group on Multiple Linear Regression of the American Educational Research Association

MLRV Abstracts appear in microform and are available from University Microfilms International
MLRV is listed in EBSCO Librarians Handbook.

ISSN 0195-7171

Library of Congress Catalog Card #80-648729

MULTIPLE LINEAR REGRESSION VIEWPOINTS

Chairman ..... Carolyn Benz
The University of Akron
Akron, OH 44325

Editor ..... Isadore Newman
The University of Akron
Akron, OH 44325

Assistant Editor ..... Rita Cowan, Diana Cook
Jeffrey Robson
The University of Akron
Akron, OH 44325

Executive Secretary ..... John Pohlman
Southern Illinois University
Carbondale, IL 62901

Cover Artist ..... David G. Barr

EDITORIAL BOARD

TABLE OF CONTENTS

JLTIPLE LINEAR REGRESSION VIEWPOINTS
VOLUME 16, NUMBER 1, SPRING 1988

A Perspective on Applications of Maximum Likelihood and Weighted Least Squares Procedures in the Context of Categorial Data Analysis

Andrew J. Bush
Baptist Memorial Hospital
Memphis, Tenn.

Pioneering technical contributions to the applied statistical literature by Grizzle, Starmer, and Koch (1969), Bishop (1969), Fienberg (1970), Goodman (1970), Koch and Reinfurt (1971), and, more recently, didactic contributions by Forthofer & Lehnen (1981) and by Kennedy (1983) have helped focus the attention of many research practitioners in the behavioral sciences on the potential for sophisticated analysis of categorical response data. In consequence, there is a growing awareness that a richer analysis can be performed on responses measured on the nominal or ordinal scale than is customarily permitted by simple crosstabulation and chi-square partitioning.

This awareness has led to the ever increasing popularity of strategies for the analysis of asymmetric, categorical data models--that is, models having at least one variable identified as a response variable. In particular, strategies that follow either the method of maximum likelihood (ML) in the Goodman tradition, such as log-linear (logit) and logistic regression analysis, or the method of weighted least squares in the Grizzle, Starmer, and Koch (GSK) tradition have been strongly gaining in acceptance.

Parenthetically, two points need now be made before proceeding to the main course of the narrative. First, the strategies mentioned above also allow for the analysis of symmetric models--that is, models for which a dependent or response variable has not been identified. However, for the purpose of discussion, the focus here will be on asymmetric models.

Secondly, the GSK strategy subsumes an approach that is known by some as Minimum Chi-Square Estimation (cf. Aldrich and Nelson, 1984) and is a specific, direct, weighted least-squares approach employing categorical independent variables only. This point is made to call attention to the fact that the label,

weighted least-squares, is a general descriptor for any weighted regression procedure using any weighting factor whatsoever. Since differential selection of weighting schemes will produce different regression results, all weighted regression procedures are not equivalently effective. But, because of an unfortunate tendency to group any and all weighted procedures under a single label, the GSK procedure has had some undeserved bad press, in the form of guilt by association, from those who disparage the regression analysis of categorical data in general. The upshot of this digression is to admit that the GSK approach is a weighted regression approach with the further admission that it is fundamentally sound.

As might be expected, since the ML and the GSK approaches use different mathematical bases in their foundation, and thus can lead to differing statistical judgments, some dispute regarding their relative merits has begun to appear. Advocates of ML based strategies typically highly value log-linear and logistic regression analysis but look askance at the use of linear regression for the analysis of categorical outcomes. This position is particularly likely to develop amongst analysts who pursue log-linear problems from the mental framework of the Deming-Stephan iterative proportional fitting (IPF) algorithm (see Kennedy, 1983, for a particularly lucid description of

the algorithm).

By employing the IPF technique, a sound strategy in and of itself, it is unfortunately quite possible to miss the point that log-linear analysis is essentially a linear modeling process. More specifically, it is altogether too easy to overlook the tautology that log-linear models really are, in fact, linear models, and as such they can be structurally coded and resolved as linear models. Those familiar with the alternative to IPF, the Newton-Raphson iteratively reweighted regression algorithm for achieving ML estimates (see Haberman, 1978, for a full description), recognize the truth of this perspective much more readily.

In reality, that which separates ML from GSK analysis is not that one employs linear models and the other does not, nor is it that one employs a regression strategy and the other does not. Both, in fact, are rooted in a regression basis. What really separates the two is that their methods of implementing the regression strategy differ.

On the one hand, GSK seeks to achieve parameter estimates through minimizing a model's residual chi-square. It does so noniteratively under the mechanism of weighted least squares regression by adopting a weighting matrix formed as the inverse of the variance of a researcher specified response function (see Forthofer and Lehnen, 1981, for a very thorough

description).

ML, on the other hand, seeks to achieve parameter estimates by maximizing the likelihood function and does so iteratively under the mechanism of reweighted least squares regression. Per force, the weighting matrix, the basis matrix, and the form of the response variable for ML differ from those used under GSK.

Both strategies avoid the well-known problems that plague ordinary least squares in this context by not making untenable distributional assumptions. Neither assumes normality nor homogeneity of variance of the residual. Both assume independence and both typically assume a product-multinomial parent data distribution for asymmetric problems.

A Technical Overview Of The GSK And ML Categorical Data Analysis Strategies

To help fix the idea that both the GSK and ML procedures for analyzing categorical data are, in fact, regression based techniques, a summary overview of both procedures is offered on the following four pages. The technical description of each is highly condensed and is meant to give a reference point to the reader rather than a full, didactic exposition. The text underscores

that both procedures rest solidly on the foundation of weighted least squares (WLS). Pages six and seven describe major aspects of the GSK strategy while pages eight and nine deal with the ML approach.

The GSK Approach

Weighted Least Squares (WLS) analysis, employs a mathematical model that adopts the following notation:

1. $\underline{p}$ a vector of proportions. Each p_{ij} is computed as the ratio of a response frequency f_{ij} to $f_i = \sum_{j=1}^{r} f_{ij}$ where the subscript i indexes a particular independent variable level or combination of levels, the subscript j addresses a particular level of the response measure, and r denotes the number of levels present in the response measure. The elements of $\underline{p}$ are arranged so that the r proportions corresponding to a value of i are contiguous and in ascending order of j.
2. $\underline{A}$ a vector of contrast coefficients with elements a_j.
3. $\underline{Y}$ a vector of contrasts such that $Y = Ap$ for <u>additive</u> models. Each Y_i is formed as $Y_i = \sum a_j p_{ij}$. Alternatively intrinsically <u>multiplicative</u> models can be formulated by first taking the <u>natural log of</u> the p_{ij}. In this case, the vector $\underline{Y}$ is formed as $Y = A \ln(p)$. For such models, $Y_i = \sum a_j \ln(p_{ij})$.
4. $\underline{X}$ an independent variable coding matrix. For, WLS results to approximate those of a log-linear analysis, the matrix $\underline{X}$ is coded using effect codes (i.e., 1,0,-1).
5. $\underline{\beta}$ a vector of regression weights.
6. $\underline{\varepsilon}$ a vector of residuals.
7. $\underline{W}$ a matrix of weights such that $\underline{W} = V(Y)^{-1}$.

In the case of an <u>additive</u> model, $V(Y_i) = \frac{1}{f_i}\left[\sum_{j=1}^{r} a_j^2\, p_{ij} - (\sum_{j=1}^{r} a_j\, p_{ij})^2\right]$

Should r=2 and A = [1 0] or A = [0 1], $V(Y_i) = \frac{p_{ij}(1-p_{ij})}{f_i}$ for j=1 or j=2.

In the case of a <u>multiplicative</u> model, $V(Y_i) = \sum_{j=1}^{r} \frac{a_j^2}{f_{ij}} - \frac{1}{f_i}(\sum_{j=1}^{r} a_j)^2$. Here, should r=2 and A = [1 -1] or A = [-1 1], (the logit function), then it follows that $V(Y_i) = \frac{1}{f_i p_{ij}(1-p_{ij})}$ for either j=1 or j=2.

Using these conventions, the regression model can be written as:

$$Y = X\beta + \epsilon$$
$$b = (X^T W X)^{-1}(X^T W Y)$$
$$V(b) = (X^T W X)^{-1}$$
$$\hat{Y} = Xb$$
$$V(\hat{Y}) = X(X^T W X)^{-1} X^T$$

The <u>residual</u> chi-square for such models is:

$$\chi^2 = (Y - Xb)^T W (Y - Xb)$$

with df = k - m

where k = the number of independent cells (i.e., rows in X)

and m = the number of parameters (i.e., columns in X)

Given a contrast matrix C that has dimensions c x m, component chi-squares (i.e., corresponding to the general linear hypothesis $C\beta = 0$) can be computed as:

$$\chi^2 = (Cb)^T[C(X^T W X)^{-1} C^T]^{-1} Cb$$

with df = c

Approximations to component chi-squares, can also be computed by taking the difference in residual chi-squares for competing models with <u>df</u> equal to the difference in the respective number of parameters. This approximation method is not as effective here as it is in log-linear analyses since the chi-square estimates are the classical Pearsonian rather than the maximum likelihood ratio chi-squares developed by Fisher and are, consequently, not precisely additive.

Iterative Weighted Least Squares (WLS) can be used to achieve Maximum Likelihood (ML) estimates. The strategy assumes the following notational conventions:

1. A diagonal matrix $\underline{F}$ of dimensionality (kr x kr) where k is the number of independent variable cells and r is the number of response variable levels. The elements of $\underline{F}$ are individual f_i where $1 <= i <= kr$. They are arranged on the major diagonal so that the order of rotaton is through the response levels for a particular independent variable cell before the next cell is represented.

2. A diagonal matrix $\underline{E}$ whose entries e_i are the expected frequencies for a given model in correspondence to the f_i.

3. A design matrix $\underline{X}$ of dimensionality (kr x m) where m is equal to the sum $(k-1)+(r-1)+(k-1)*(r-1)$. Note that m represents the total component degrees of freedom in a given model excluding the intercept (or grand mean) which is not coded. The design matrix $\underline{X}$ is composed of effect codes (1,0,-1) and is formed as:

   a. The first $k-1$ columns of $\underline{X}$ are effect codes on the independent variables-- each row of which is replicated contiguously r times.

   b. The next $r-1$ columns of $\underline{X}$ are formed by block replicating effect codes on the response measure k times. Each block is of dimensionality r x $r-1$.

   c. The remaining $(k-1)*(r-1)$ columns represent the independent-dependent variable interaction terms and are formed by multiplication of the corresponding prior columns.

4. The subscript c represents the current iteration and the subscript p represents the prior iteration.

5. Vector $\underline{Y} = \text{diag}\,[\ln(E_p) + (F - E_p)E_p^{-1}]$. On the first iteration, this procedure is replaced by computing each element $\underline{Y}$ to be $Y_i = \ln(e_i)$ where $e_i = f_i + .05$.

6. A matrix $\underline{D}$ of the same dimensionality as $\underline{X}$ formed by kr row replicates of the vector $\underline{d}$ with elements d_j $(1 <= j <= m)$

   where $$d_j = \frac{\sum_{i=1}^{kr} X_{ij}\, e_i}{\sum_{i=1}^{kr} e_i}$$

   given the e_i are from the prior iteration.

The iterative process, given $\underline{X}$ and $\underline{F}$, is as follows:

1. Compute $\underline{Y}_c$ as described.
2. Compute $\underline{D}_c$ as described.
3. Generate the matrix $\underline{A} = \underline{X} - \underline{D}_c$
4. Estimate the regression weights β as

$$b_c = [A_c^T E_p A_c]^{-1} A_c^T E_p Y_c$$

$$a_c = \ln \left| \frac{\sum_{i=1}^{kr} f_i}{\sum_{i=1}^{kr} e^{\sum_{j=1}^{m} b_j x_{ij}}} \right|$$

5. Estimate the i^{th} element of E_c as $e^{a_c + \sum_{j=1}^{m} b_j x_{ij}}$
6. If the estimates b_c converge on b_p then stop iteration otherwise return to step 1.

Given convergence, the following additional estimates can be made:

1. $V(b) = (A_c^T E_c A_c)^{-1}$
2. Standardized residuals are $(f_i - e_i)/\sqrt{e_i}$
3. Residual $L^2 = 2 \sum_{i=1}^{kr} f_i \ln\left(\frac{f_i}{e_i}\right)$
4. Residual $\chi^2 = \sum_{i=1}^{kr} \frac{(f_i - e_i)^2}{e_i}$

both with $\underline{df} = kr - m - 1$

As the reader can readily see, both approaches permit point and interval estimation of regression parameters. To help profile how the strategies compare with one another, their relative merit from the author's point of view will now be examined along several dimensions. Those dimensions are:

1) Ability to deal with symmetric models.
2) Facility for testing hypotheses.
3) Statistical properties of estimators.
4) Relative computational requirements.
5) Ease of interpretation of estimators.
6) Robustness with respect to extreme values.
7) Capacity for handling interval variables.

Symmetric Models. With regard to doing data analysis where no individual variable is perceived to be a response (dependent) variable, the ML method has a clear edge. In fact, log-linear analysis, having its roots in the field of sociological methodology, a field that does not often enjoy the luxury of experimental manipulation of independent variables, is exceptionally well-geared for coping with marginal and partial associations among variables.

In contrast, the GSK approach, an approach that emanates from the biostatistical world, is focused directly on exploring the effects of one or more independent on one or more dependent variables. Unlike the log-linear strategy, GSK forces selection of a response variable. This does not mean that the GSK approach can not handle symmetric problems--it can. However, an analyst must systematically rotate through a problem's variables choosing different variables, individually, as the response measure. Consequently, the GSK method is not as desirable in such a context.

Facility for Testing Hypotheses. Assuming the asymmetric environment for the remainder of this narrative, how do the strategies compare on the basis of testing hypotheses? In this writer's opinion, the GSK approach is probably stronger but not overpoweringly so. GSK, on the surface, appears to have far greater flexibility because the analyst is permitted to establish nearly any linear combination on nearly any transformation of the response measure. Such flexibility permits definition of a response function in terms of raw proportions, or logged proportions (the latter leading directly to odds ratios), or even exponentiated proportions.

In comparison, the log-linear approach forces a definition of the response function in terms of logged proportions. However, what is often overlooked during a log-linear analysis is that expected frequencies are generated and that the analyst is free to establish any desired transformation and linear combination on those frequencies he or she wishes. This implies that the ML method can be as rich analytically as the GSK method (cf. Haber, 1984). In fairness, though, the more extended mode of analysis under ML is not typical and is more mechanically difficult.

Statistical Properties. With respect to the statistical properties of the estimators produced by GSK and ML, a slight edge has to be awarded ML since the ML estimators are well-known to be asymptotically consistent and relatively efficient. What is not as well known is that the GSK estimators are similarly asymptotically consistent and, for that matter, aysmptotically equivalent to ML estimators. They are, in fact, best asymptotic normal estimators (BAN).

For fully saturated models of any sample size, the two methods deliver identical results. For unsaturated models on large samples, differences in the estimators tend to be trivial. However, as sample sizes decrease,

the GSK and the ML estimators can be disparate with the ML estimators tending to have smaller variance. The question of how large is large enough to feel fairly comfortable that similar results will be afforded by both strategies is not precisely known. However, it is generally recommended that samples be of sufficient size before employing either approach. For specific guidelines under GSK, the reader is referred to Forthofer and Lehnen (1981) and, for guidelines under ML, to Haberman (1978).

Computational Requirements. From a computational perspective, GSK has a clear edge. In the first place, it is non-iterative. In the second, its basis matrix is a factor of r*r smaller where r denotes the number of categories present in the response variable. For problems involving polytomous response measures, computational resource requirements heavily favor GSK. While such considerations may not be critical for mainframe applications, the resource implications for microcomputing are clear.

Ease of Interpretation. With regard to estimator interpretability, ML estimates are slightly easier for a novice to make sense of if a canned log-linear strategy is being employed. This is the case because

the parameters are conceptually well identified in the paradigm of analysis of variance effects on logged expected cell frequencies. If, however, the more flexible regression coding scheme afforded by the Newton-Raphson strategy is employed to deviate from traditional effect definitions, this edge evaporates and both ML and GSK estimates must be carefully identified by the analyst.

Robustness for Extreme Values. From the perspective of extreme values, the GSK and the ML strategies share common problems. Both must cope with empty cells by either making a numeric replacement or collapsing categories. Further, both rely on having large samples to effect robustness in the statistical properties of their estimators. From this author's viewpoint, neither procedure has an edge with regard to this problem. However, it should be noted that it is recommended that the GSK approach engage a log transformation on proportions when proportions are extreme rather than operating upon them in their native metric (see Forthofer & Lehnen, 1981). Intuitively, the same caveat should apply to followup contrasts on ML estimates.

Interval Independent Variables. With regard to interval independent variables, one variant of ML, namely logistic regression analysis, has a distinctive advantage. It has the capacity for coping with a mix of both categorical and continuous variables with the provision that the response measure be a dichotomous variable.

Neither the GSK nor traditional log-linear ML methods can duplicate this capacity. Even so, an analyst could approach the situation of interval variables with either log-linear or GSK analysis by meaningfully categorizing all interval variables present.

Synthesis. Given this profile, which procedure then is preferable? From the author's perspective neither completely dominates the other. Both are powerful and are well worth mastering.

Should the research purpose be to examine marginal and partial associations symmetrically, the ML approach embodied by log-linear analysis is preferable. Should the research purpose be to test hypotheses on response level proportions or on complex functions, the GSK approach is preferable. If interval level independent

variables are present and recoding is not desirable, the logistic regression ML approach is promising--providing no more than two levels are present in the response variable.

Should computing facilities be highly restricted, the GSK approach can be preferable. If the analyst is unsophisticated with respect to the analysis of linear models, a traditional log-linear analysis will be easier to pursue. If sample sizes are small or empty cells are present, neither strategy is particularly safe. If extreme proportions are present, both approaches should make appropriate adjustments.

In the final analysis, both approaches have specific strengths as well as detractions. Both offer strong analytic capabilities and both belong in our repertoire.

An Analysis of Hypothetical Data By ML And By GSK

For the purpose of illustrating the similarity of the two methods in an applied scenario and for the purpose of demonstrating their versatility, the following simple numeric example is offered. The data shown below were constructed by John J. Kennedy, of The

Ohio State University, as a didactic example to show how effect contrasts might be estimated through chi-square partitioning. With his kind permission, the data will be employed here to show (1) how both ML and GSK can be used to estimate linear and quadratic effects and (2) how both the ML and GSK procedures can pursue traditional log-linear effects.

The data are given in Table 1 and consist of frequency counts that have been crosstabulated on the basis of student sex (A_1= Males, A_2= Females), an unspecified treatment variable (B_1= Treatment, B_2= Control), and a trichotomous outcome measure (C_1= Poor, C_2= Satisfactory, and C_3= Good).

Table 1. A Hypothetical 2x2x3 Data Example.

		Outcome Description:			
		Poor	Satisfactory	Good	Sum
Sex	Treatment				
M	T	5	19	4	28
M	C	3	6	13	22
F	T	6	16	6	28
F	C	2	8	12	22
Sum		16	49	33	100

Page 19 demonstrates a linear and quadratic effect coding setup used as input to an author prepared Newton-Raphson ML program that has been designed to teach the flow of the ML procedure. The input consists of (1) the number of rows in the regression basis matrix, (2) the number of columns in that matrix--note the omission of a unit vector for the grand mean, (3) the basis matrix, itself, arranged in column order:

a) Sex vector.

b) Treatment vector.

c) Sex x Treatment.

d) Linear Response Contrast.

e) Quadratic Response Contrast.

f) Linear Effect of Sex.

g) Quadratic Effect of Sex.

h) Linear Effect of Treatment.

i) Quadratic Effect of Treatment.

j) Linear Effect of Sex x Treatment.

k) Quadratic Effect of Sex x Treatment.

and (4) the raw frequencies themselves with the response variable rotating most rapidly, followed by treatment, and sex in that order. Pages 20 and 21 show the ML analysis with page 21 being the more interesting since it delivers parameter estimates. Pages 22 and 23 show the corresponding GSK analysis with page 22 delivering the linear analysis and 23, the quadratic.

ML Analysis of 2x2x3 Data Set Using Linear & Quadratic Codings

```
12
11
 1   1   1 -.5 -.333333    -.5 -.333333   -.5 -.333333   -.5 -.333333
 1   1   1   0  .666667     0  .666667     0  .666667     0  .666667
 1   1   1  .5 -.333333     .5 -.333333     .5 -.333333     .5 -.333333
 1  -1  -1 -.5 -.333333    -.5 -.333333     .5  .333333     .5  .333333
 1  -1  -1   0  .666667     0  .666667     0 -.666667     0 -.666667
 1  -1  -1  .5 -.333333     .5 -.333333    -.5  .333333    -.5  .333333
-1   1  -1 -.5 -.333333     .5  .333333    -.5 -.333333     .5  .333333
-1   1  -1   0  .666667     0 -.666667     0  .666667     0 -.666667
-1   1  -1  .5 -.333333    -.5  .333333     .5 -.333333    -.5  .333333
-1  -1   1 -.5 -.333333     .5  .333333     .5  .333333    -.5 -.333333
-1  -1   1   0  .666667     0 -.666667     0 -.666667     0  .666667
-1  -1   1  .5 -.333333    -.5  .333333    -.5  .333333     .5 -.333333
5
19
4
3
6
13
6
16
6
2
8
12
```

ML Analysis of 2x2x3 Data Set Using Linear & Quadratic Codings

Cell Frequencies Iteration is 4

cell =>	1	obs freq =>	5.0000	exp freq =>	5.0000
cell =>	2	obs freq =>	19.0000	exp freq =>	19.0000
cell =>	3	obs freq =>	4.0000	exp freq =>	4.0000
cell =>	4	obs freq =>	3.0000	exp freq =>	3.0000
cell =>	5	obs freq =>	6.0000	exp freq =>	6.0000
cell =>	6	obs freq =>	13.0000	exp freq =>	13.0000
cell =>	7	obs freq =>	6.0000	exp freq =>	6.0000
cell =>	8	obs freq =>	16.0000	exp freq =>	16.0000
cell =>	9	obs freq =>	6.0000	exp freq =>	6.0000
cell =>	10	obs freq =>	2.0000	exp freq =>	2.0000
cell =>	11	obs freq =>	8.0000	exp freq =>	8.0000
cell =>	12	obs freq =>	12.0000	exp freq =>	12.0000

Dvector Iteration is 4

column =>	1	value =>	-0.0000
column =>	2	value =>	0.1200
column =>	3	value =>	0.0000
column =>	4	value =>	0.0950
column =>	5	value =>	0.1567
column =>	6	value =>	-0.0050
column =>	7	value =>	0.0100
column =>	8	value =>	-0.1050
column =>	9	value =>	0.1700
column =>	10	value =>	-0.0050
column =>	11	value =>	0.0500

Amatrix Iteration is 4

1.000	0.880	1.000	-0.595	-0.490	-0.495	-0.343	-0.395	-0.503	-0.495	-0
1.000	0.880	1.000	-0.095	0.510	0.005	0.657	0.105	0.497	0.005	0
1.000	0.880	1.000	0.405	-0.490	0.505	-0.343	0.605	-0.503	0.505	-0
1.000	-1.120	-1.000	-0.595	-0.490	-0.495	-0.343	0.605	0.163	0.505	0
1.000	-1.120	-1.000	-0.095	0.510	0.005	0.657	0.105	-0.837	0.005	-0
1.000	-1.120	-1.000	0.405	-0.490	0.505	-0.343	-0.395	0.163	-0.495	0
-1.000	0.880	-1.000	-0.595	-0.490	0.505	0.323	-0.395	-0.503	0.505	0
-1.000	0.880	-1.000	-0.095	0.510	0.005	-0.677	0.105	0.497	0.005	-0
-1.000	0.880	-1.000	0.405	-0.490	-0.495	0.323	0.605	-0.503	-0.495	0
-1.000	-1.120	1.000	-0.595	-0.490	0.505	0.323	0.605	0.163	-0.495	-0
-1.000	-1.120	1.000	-0.095	0.510	0.005	-0.677	0.105	-0.837	0.005	0
-1.000	-1.120	1.000	0.405	-0.490	-0.495	0.323	-0.395	0.163	0.505	-0

Analysis of 2x2x3 Data Set Using Linear & Quadratic Codings

s Iteration is 4

ercept is 1.917424 old value was 1.917424

umn =>	1	A	value =>	-0.018176	Change	0.000000
umn =>	2	B	value =>	0.131955	Change	0.000000
umn =>	3	AB	value =>	-0.051147	Change	-0.000000
umn =>	4	C1	value =>	0.758738	Change	0.000000
umn =>	5	C2	value =>	0.719449	Change	0.000000
umn =>	6	AC1	value =>	-0.137141	Change	-0.000000
umn =>	7	AC2	value =>	-0.016173	Change	-0.000000
umn =>	8	BC1	value =>	-0.870310	Change	-0.000000
umn =>	9	BC2	value =>	0.494252	Change	-0.000000
umn =>	10	ABC1	value =>	0.025570	Change	0.000000
umn =>	11	ABC2	value =>	0.249045	Change	0.000000

of changes 0.000000

iance Iteration is 4

0152 0.0017-0.0027 0.0044 0.0013-0.0130-0.0100-0.0023-0.0041 0.0151-0.0015
0017 0.0152-0.0002 0.0151-0.0015-0.0023-0.0041-0.0130-0.0100 0.0044 0.0013
0027-0.0002 0.0152-0.0023-0.0041 0.0151-0.0015 0.0044 0.0013-0.0130-0.0100
0044 0.0151-0.0023 0.1111 0.0195-0.0035-0.0066-0.0131-0.0226 0.0181 0.0034
0013-0.0015-0.0041 0.0195 0.0532-0.0066 0.0011-0.0226-0.0143 0.0034 0.0013
0130-0.0023 0.0151-0.0035-0.0066 0.1111 0.0195 0.0181 0.0034-0.0131-0.0226
0100-0.0041-0.0015-0.0066 0.0011 0.0195 0.0532 0.0034 0.0013-0.0226-0.0143
0023-0.0130 0.0044-0.0131-0.0226 0.0181 0.0034 0.1111 0.0195-0.0035-0.0066
0041-0.0100 0.0013-0.0226-0.0143 0.0034 0.0013 0.0195 0.0532-0.0066 0.0011
0151 0.0044-0.0130 0.0181 0.0034-0.0131-0.0226-0.0035-0.0066 0.1111 0.0195
0015 0.0013-0.0100 0.0034 0.0013-0.0226-0.0143-0.0066 0.0011 0.0195 0.0532

Iteration is 4

umn =>	1	value =>	3.2242
umn =>	2	value =>	12.4921
umn =>	3	value =>	4.1241
umn =>	4	value =>	6.0796
umn =>	5	value =>	14.6828
umn =>	6	value =>	0.7101
umn =>	7	value =>	2.3169
umn =>	8	value =>	-6.2753
umn =>	9	value =>	20.3232
umn =>	10	value =>	-0.9058
umn =>	11	value =>	4.5616

rsonian 0.0000
herian 0.0000

The Pattern Matrix X as Entered

1.00	1.00	1.00	1.00
1.00	1.00	-1.00	-1.00
1.00	-1.00	1.00	-1.00
1.00	-1.00	-1.00	1.00

The Parameter Coefficient Matrix:

0.25	0.25	0.25	0.25
0.25	0.25	-0.25	-0.25
0.25	-0.25	0.25	-0.25
0.25	-0.25	-0.25	0.25

The Frequencies as Entered

CATEGORY:		
1	2	3
5	19	4
3	6	13
6	16	6
2	8	12

CONTRAST: -1.00 0.00 1.00

PARAMETER	LOG EST	LOG SE	ODDS EST	ODDS SE	Z ESTIMATE
INTERCEPT	0.759	0.333	2.136	1.396	2.277
AC1	-0.137	0.333	0.872	1.396	-0.412
BC1	-0.870	0.333	0.419	1.396	-2.612
ABC1	0.026	0.333	1.026	1.396	0.077

PERFECT FIT --- SATURATED MODEL

RESIDUAL CHI-SQUARE = 0.000 DF = 0 ALPHA = 1.00

LOG-P FUNCTION	PREDICTED	RESIDUAL
-0.223	-0.223	0.000
1.466	1.466	0.000
0.000	0.000	0.000
1.792	1.792	0.000

GSK Quadratic Analysis p 23

The Pattern Matrix X as Entered

.00	1.00	1.00	1.00
.00	1.00	-1.00	-1.00
.00	-1.00	1.00	-1.00
.00	-1.00	-1.00	1.00

The Parameter Coefficient Matrix:

.25	0.25	0.25	0.25
.25	0.25	-0.25	-0.25
.25	-0.25	0.25	-0.25
25	-0.25	-0.25	0.25

The Frequencies as Entered

CATEGORY:

1	2	3
5	19	4
3	6	13
6	16	6
2	8	12

'RAST: -0.50 1.00 -0.50

METER RCEPT	LOG EST	LOG SE	ODDS EST	ODDS SE	Z ESTIMATE
	0.719	0.231	2.053	1.259	3.120
	-0.016	0.231	0.984	1.259	-0.070
	0.494	0.231	1.639	1.259	2.143
	0.249	0.231	1.283	1.259	1.080

PERFECT FIT --- SATURATED MODEL

RESIDUAL CHI-SQUARE = 0.000 DF = 0 ALPHA = 1.00

LOG-P FUNCTION	PREDICTED	RESIDUAL
1.447	1.447	0.000
-0.040	-0.040	0.000
0.981	0.981	0.000
0.490	0.490	0.000

Collecting the effect estimates from the runs just presented lets us produce Table 2. Note that two separate analyses had to be performed by GSK to produce first the linear and then the quadratic results.

Table 2. Summary of ML & GSK Analysis of Linear & Quadratic Effects in the 2x2x3 Example.

	ML			GSK		
Effect	b	SE	Page	b	SE	Page
AC_1	-.14	.33	21	-.14	.33	22
AC_2	-.02	.23	21	-.02	.23	23
BC_1**	-.87	.33	21	-.87	.33	22
BC_2*	.49	.23	21	.49	.23	23
ABC_1	.03	.33	21	.03	.33	22
ABC_2	.25	.23	21	.25	.23	23

** p .01

* p .05

Clearly the two sets of results are isomorphic with each revealing both a linear and quadratic effect for the treatment variable on the response frequencies. With respect to the linear trend, the odds favoring a response of "good" over a response of "poor" are better in the control group than in the treatment.

With respect to the quadratic trend, the treatment group average odds favoring a "satisfactory" response over the other two response categories are better than the corresponding odds for the control condition. Obviously, if this were a true research situation, an analyst would suddenly get gray hair but the data do serve the purpose of illustration.

Repeating the exercise with linear codings established to produce traditional log-linear parameters, the ML input file is shown on page 26 and follows exactly the same pattern as before. This time, however, the linear and quadratic codes give way to average effect codes.

Pages 27 and 28 reproduce the results from the ML analysis with page 28 being the more interesting. The GSK output is shown on pages 29, 30, and 31. This time three runs were made under GSK in order to directly estimate the parameters associated with the third level of the response variable. These could, admittedly, have been determined by subtraction. However, the variance estimates for the parameters on page 31 would have had to have been inferred rather than obtained from inspection.

ML Analysis of 2x2x3 Data Set Using Log-Linear Codings

```
12
11
 1   1   1   1   0    1   0    1   0    1   0
 1   1   1   0   1    0   1    0   1    0   1
 1   1   1  -1  -1   -1  -1   -1  -1   -1  -1
 1  -1  -1   1   0    1   0   -1   0   -1   0
 1  -1  -1   0   1    0   1    0  -1    0  -1
 1  -1  -1  -1  -1   -1  -1    1   1    1   1
-1   1  -1   1   0   -1   0    1   0   -1   0
-1   1  -1   0   1    0  -1    0   1    0  -1
-1   1  -1  -1  -1    1   1   -1  -1    1   1
-1  -1   1   1   0   -1   0   -1   0    1   0
-1  -1   1   0   1    0  -1    0  -1    0   1
-1  -1   1  -1  -1    1   1    1   1   -1  -1
 5
 19
 4
 3
 6
 13
 6
 16
 6
 2
 8
 12
 12
 11
```

nalysis of 2x2x3 Data Set Using Log-Linear Codings

Cell Frequencies Iteration is 4

=>	1	obs freq =>	5.0000	exp freq =>	5.0000
=>	2	obs freq =>	19.0000	exp freq =>	19.0000
=>	3	obs freq =>	4.0000	exp freq =>	4.0000
=>	4	obs freq =>	3.0000	exp freq =>	3.0000
=>	5	obs freq =>	6.0000	exp freq =>	6.0000
=>	6	obs freq =>	13.0000	exp freq =>	13.0000
=>	7	obs freq =>	6.0000	exp freq =>	6.0000
=>	8	obs freq =>	16.0000	exp freq =>	16.0000
=>	9	obs freq =>	6.0000	exp freq =>	6.0000
=>	10	obs freq =>	2.0000	exp freq =>	2.0000
=>	11	obs freq =>	8.0000	exp freq =>	8.0000
=>	12	obs freq =>	12.0000	exp freq =>	12.0000

tor Iteration is 4

mn =>	1	value =>	-0.0000
mn =>	2	value =>	0.1200
mn =>	3	value =>	0.0000
mn =>	4	value =>	-0.1900
mn =>	5	value =>	0.1400
mn =>	6	value =>	0.0100
mn =>	7	value =>	0.0200
mn =>	8	value =>	0.2100
mn =>	9	value =>	0.3600
mn =>	10	value =>	0.0100
mn =>	11	value =>	0.0800

rix Iteration is 4

.000	0.880	1.000	1.190	-0.140	0.990	-0.020	0.790	-0.360	0.990	-0.080
.000	0.880	1.000	0.190	0.860	-0.010	0.980	-0.210	0.640	-0.010	0.920
.000	0.880	1.000	-0.810	-1.140	-1.010	-1.020	-1.210	-1.360	-1.010	-1.080
.000	-1.120	-1.000	1.190	-0.140	0.990	-0.020	-1.210	-0.360	-1.010	-0.080
.000	-1.120	-1.000	0.190	0.860	-0.010	0.980	-0.210	-1.360	-0.010	-1.080
.000	-1.120	-1.000	-0.810	-1.140	-1.010	-1.020	0.790	0.640	0.990	0.920
.000	0.880	-1.000	1.190	-0.140	-1.010	-0.020	0.790	-0.360	-1.010	-0.080
.000	0.880	-1.000	0.190	0.860	-0.010	-1.020	-0.210	0.640	-0.010	-1.080
.000	0.880	-1.000	-0.810	-1.140	0.990	0.980	-1.210	-1.360	0.990	0.920
.000	-1.120	1.000	1.190	-0.140	-1.010	-0.020	-1.210	-0.360	0.990	-0.080
.000	-1.120	1.000	0.190	0.860	-0.010	-1.020	-0.210	-1.360	-0.010	0.920
.000	-1.120	1.000	-0.810	-1.140	0.990	0.980	0.790	0.640	-1.010	-1.080

ML Analysis of 2x2x3 Data Set Using Log-Linear Codings

Bwts Iteration is 4

intercept is 1.917425 old value was 1.917425

column =>	1	A	value =>	-0.018176	Change	0.000000
column =>	2	B	value =>	0.131955	Change	0.000000
column =>	3	AB	value =>	-0.051147	Change	-0.000000
column =>	4	C1	value =>	-0.619185	Change	-0.000000
column =>	5	C2	value =>	0.479633	Change	0.000000
column =>	6	AC1	value =>	0.073962	Change	0.000000
column =>	7	AC2	value =>	-0.010782	Change	-0.000000
column =>	8	BC1	value =>	0.270404	Change	0.000000
column =>	9	BC2	value =>	0.329501	Change	-0.000000
column =>	10	ABC1	value =>	-0.095800	Change	-0.000000
column =>	11	ABC2	value =>	0.166030	Change	0.000000

Sum of changes 0.000000

Variance Iteration is 4

```
 0.0152 0.0017-0.0027-0.0026 0.0008 0.0098-0.0067 0.0025-0.0027-0.0070-0.
 0.0017 0.0152-0.0002-0.0070-0.0010 0.0025-0.0027 0.0098-0.0067-0.0026 0.
-0.0027-0.0002 0.0152 0.0025-0.0027-0.0070-0.0010-0.0026 0.0008 0.0098-0.
-0.0026-0.0070 0.0025 0.0402-0.0183-0.0029 0.0019-0.0124 0.0107 0.0058-0.
 0.0008-0.0010-0.0027-0.0183 0.0236 0.0019 0.0005 0.0107-0.0064-0.0014 0.
 0.0098 0.0025-0.0070-0.0029 0.0019 0.0402-0.0183 0.0058-0.0014-0.0124 0.
-0.0067-0.0027-0.0010 0.0019 0.0005-0.0183 0.0236-0.0014 0.0006 0.0107-0.
 0.0025 0.0098-0.0026-0.0124 0.0107 0.0058-0.0014 0.0402-0.0183-0.0029 0.
-0.0027-0.0067 0.0008 0.0107-0.0064-0.0014 0.0006-0.0183 0.0236 0.0019 0.
-0.0070-0.0026 0.0098 0.0058-0.0014-0.0124 0.0107-0.0029 0.0019 0.0402-0.
-0.0010 0.0008-0.0067-0.0014 0.0006 0.0107-0.0064 0.0019 0.0005-0.0183 0.
```

XY Iteration is 4

column =>	1	value =>	3.2242
column =>	2	value =>	12.4921
column =>	3	value =>	4.1241
column =>	4	value =>	-12.1592
column =>	5	value =>	15.9446
column =>	6	value =>	-1.4202
column =>	7	value =>	2.7652
column =>	8	value =>	12.5507
column =>	9	value =>	36.7601
column =>	10	value =>	1.8116
column =>	11	value =>	7.7483

Pearsonian	0.0000
Fisherian	0.0000

GSK Log-linear: C1 odds p 29

The Pattern Matrix X as Entered

)0	1.00	1.00	1.00
)0	1.00	-1.00	-1.00
)0	-1.00	1.00	-1.00
)0	-1.00	-1.00	1.00

The Parameter Coefficient Matrix:

25	0.25	0.25	0.25
25	0.25	-0.25	-0.25
25	-0.25	0.25	-0.25
25	-0.25	-0.25	0.25

The Frequencies as Entered

CATEGORY:

1	2	3
5	19	4
3	6	13
6	16	6
2	8	12

RAST: 0.67 -0.33 -0.33

METER	LOG EST	LOG SE	ODDS EST	ODDS SE	Z ESTIMATE
RCEPT	-0.619	0.200	0.538	1.222	-3.090
	0.074	0.200	1.077	1.222	0.369
	0.270	0.200	1.310	1.222	1.349
	-0.096	0.200	0.909	1.222	-0.478

PERFECT FIT --- SATURATED MODEL

RESIDUAL CHI-SQUARE = 0.000 DF = 0 ALPHA = 1.00

LOG-P FUNCTION	PREDICTED	RESIDUAL
-0.371	-0.371	0.000
-0.720	-0.720	0.000
-0.327	-0.327	0.000
-1.059	-1.059	0.000

GSK Log-linear: C2 odds p 30

The Pattern Matrix X as Entered

1.00	1.00	1.00	1.00
1.00	1.00	-1.00	-1.00
1.00	-1.00	1.00	-1.00
1.00	-1.00	-1.00	1.00

The Parameter Coefficient Matrix:

0.25	0.25	0.25	0.25
0.25	0.25	-0.25	-0.25
0.25	-0.25	0.25	-0.25
0.25	-0.25	-0.25	0.25

The Frequencies as Entered

CATEGORY:		
1	2	3
5	19	4
3	6	13
6	16	6
2	8	12

CONTRAST: -0.33 0.67 -0.33

PARAMETER	LOG EST	LOG SE	ODDS EST	ODDS SE	Z ESTIMATE
INTERCEPT	0.480	0.154	1.615	1.166	3.120
AC2	-0.011	0.154	0.989	1.166	-0.070
BC2	0.330	0.154	1.390	1.166	2.143
ABC2	0.166	0.154	1.181	1.166	1.080

PERFECT FIT --- SATURATED MODEL

RESIDUAL CHI-SQUARE = 0.000 DF = 0 ALPHA = 1.00

LOG-P FUNCTION	PREDICTED	RESIDUAL
0.964	0.964	0.000
-0.027	-0.027	0.000
0.654	0.654	0.000
0.327	0.327	0.000

Again collecting the computed results produces Table 3. Once more the profile is consistent.

Table 3. Summary of ML & GSK Analysis of Log-linear Effects in the 2x2x3 Example.

	ML			GSK		
Effect	b	SE	Page	b	SE	Page
AC_1	.07	.20	28	.07	.20	29
AC_2	-.01	.15	28	-.01	.15	30
AC_3	-.06	.16		-.06	.16	31
BC_1	.27	.20	28	.27	.20	29
BC_2*	.33	.15	28	.33	.15	30
BC_3**	-.60	.16		-.60	.16	31
ABC_1	-.10	.20	28	-.10	.20	29
ABC_2	.17	.15	28	.17	.15	30
ABC_3	-.07	.16		-.07	.16	31

** $p < .01$

* $p < .05$

Once more we clearly have identical results but now in terms of log-linear estimates. By way of interpretation, the significant BC_2 term indicates that the geometric average odds favoring a "satisfactory" response over all possible response categories are

GSK Log-linear: C3 odds p 31

The Pattern Matrix X as Entered

1.00	1.00	1.00	1.00
1.00	1.00	-1.00	-1.00
1.00	-1.00	1.00	-1.00
1.00	-1.00	-1.00	1.00

The Parameter Coefficient Matrix:

0.25	0.25	0.25	0.25
0.25	0.25	-0.25	-0.25
0.25	-0.25	0.25	-0.25
0.25	-0.25	-0.25	0.25

The Frequencies as Entered

CATEGORY:		
1	2	3
5	19	4
3	6	13
6	16	6
2	8	12

CONTRAST: -0.33 -0.33 0.67

PARAMETER	LOG EST	LOG SE	ODDS EST	ODDS SE	Z ESTIMATE
INTERCEPT	0.140	0.165	1.150	1.179	0.846
AC3	-0.063	0.165	0.939	1.179	-0.383
BC3	-0.600	0.165	0.549	1.179	-3.639
ABC3	-0.070	0.165	0.932	1.179	-0.426

PERFECT FIT --- SATURATED MODEL

RESIDUAL CHI-SQUARE = 0.000 DF = 0 ALPHA = 1.00

LOG-P FUNCTION	PREDICTED	RESIDUAL
-0.594	-0.594	0.000
0.747	0.747	0.000
-0.327	-0.327	0.000
0.732	0.732	0.000

stronger for the treatment group than the controls. The significant BC_3 term indicates that the average odds favoring a "good" response are better in the control condition. The results are consistent with the findings from the linear-quadratic analysis but reveal a slightly different aspect of the data based on the differential coding. Again, thankfully, the results are fictitious.

Concluding Remarks

The author hopes that a relatively convincing case has been built for embracing both the ML and GSK technologies and for appreciating that both are fundamentally regression based strategies. Further, he hopes that the point has been adequately made that to argue which is better is, at best, a contextually bound issue which begs the question for a universal answer.

Certainly, much more could have been discussed regarding relative applications, for example, with respect to nested and blocking design or with respect to followups to omnibus tests. These matters are relevant and important but beyond the scope of the material presented here. Obviously the application arena is large and the application tools are superb.

References

1. Aldrich, J.H., & Nelson, F.D., *Linear probability, logit, and probit models*. Beverly Hills: Sage University Press, 1984.

2. Bishop, Y.M.M., Full contingency tables, logits, and split contingency tables. *Biometrics*, 1969, *25*, 383-400.

3. Fienberg, S.E., The analysis of multimensional contingency tables. *Ecology*, 1970, *51*, 419-433.

4. Forthofer, R.N., & Lehnen, R.G., *Public program analysis: A new categorical data approach*. Belmont: Lifetime Learning Publications, 1981.

5. Goodman, L.A., The multivariate analysis of qualitative data: Interactions among multiple classifications. *Journal of the American Statistical Association*, 1970, *65*, 226-256.

6. Grizzle, J.E., Starmer, C.F., and Koch, G.G., *Analysis of categorical data by linear models*. Biometrics, 1969, *25*, 489-504.

7. Haber, M., Maximum likelihood methods for linear and log linear models in categorical data. *Computational Statistics & Data Analysis*, 1985, *3*, 1-10.

8. Haberman, S.J., *Analysis of qualitative data: Introductory topics* (Vol 1). New York: Academic Press, 1978.

9. Koch, G.G., and Reinfurt, D.W., The analysis of categorical data from mixed models. *Biometrics, 1971, 27,* 157-173.

10. Kennedy, J.J., *Analyzing qualitative data: Introductory log-linear analysis for behavioral research.* New York: Praeger Publishers, 1983.

MULTIPLE LINEAR REGRESSION VIEWPOINTS
VOLUME 16, NUMBER 1, SPRING 1988

Predicting Statistics Achievement: A Prototypical Regression Analysis

Rodney J. Presley and Carl Huberty
University of Georgia

The purposes of the current study are: (a) to demonstrate a viable approach to the conduct of a multiple regression/correlation analysis; and (b) to illustrate the approach in the context of predicting achievement in an introductory statistical methods course. The analysis is proposed as being appropriate if the basic intent of a study is that of prediction as opposed to that of explanation. That is, the intent is to arrive at a model for predicting a criterion in as efficient a manner as the data on hand will allow. No model, causal or otherwise, is being posited or verified.

There are five dimensions of the suggested approach: 1) designing the study; 2) examining the data; 3) searching for an efficient prediction model; 4) using regression diagnostics; and 5) assessing the model(s). Each dimension of the study is presented in sections below, each of which includes an application in the context of predicting statistics achievement. [This list does not necessarily imply a sequential step-by-step analysis.]

An effective model for predicting statistics achievement may be useful in addressing three questions related to instruction and curriculum: 1) Can a fairly accurate rule be determined for predicting achievement in introductory statistics courses? 2) How effective are easily obtained graduate-level student test scores in predicting "high-achievers"? 3) In predicting "low-achievers"? Having some knowledge of predicted achievement

---

A special thanks is extended to Stephen Olejnik, David Payne, and John Stauffer (at The University of Georgia) for their cooperation in this study.

may be helpful in an obvious way to instructors. Furthermore, having rules for accurately predicting high and low achievers would possibly suggest either a special "advanced" section or some remedial pre-course experience.

Previous studies predicting achievement in introductory statistics courses have varied in predictor models used and in subject sample characteristics. Predictor variable domains employed in previous studies include computation skills, mathematics symbolism, previous mathematical experience, logical thinking, attitudes, anxiety, self appraisal, impulsiveness, arithmetic/mathematics achievement, and other biographical characteristics (e.g. gender, age, college major). Such predictor domains and others may be found in the studies by Bending and Hughes (1954), Bledsoe and Perkins (1976), Elmore and Vasu (1980), Feij (1976), Feinberg and Halperin (1978), Harvey, Plake, and Wise (1985), and Pruzek (1964). The size of the sample studied and the academic level of the students in the sample varied somewhat in these studies. For example, Bending and Hughes employed 71 undergraduate level students, while Elmore and Vasu (N=188) and Pruzek (N=112) employed graduate students; Feinberg and Halperin employed undergraduate (209) as well as graduate (94) level students, while Harvey et al. (1985) employed 47 and 41 undergraduate and graduate level students, respectively.

As might be expected most of the studies reviewed used a

multiple regression/correlation analysis. Typically, squared multiple correlation coefficients were reported (along with some type of "variable selection" results and some kind of regression weights). The percent of variance shared between statistics achievement and one or more variables (from predictor variable domains as listed above) has generally been in the range of 30 to 45 (based on unadjusted squared multiple correlation coefficients).

Designing the Study

In conducting a multiple regression/correlation study one must clearly define the population for which the prediction model is intended, select a meaningful criterion, and select a useful set of predictors.

Subjects

The target population of interest in this study is graduate students enrolled in the introductory statistical methods course. Students in eight sections of an introductory statistical methods course offered in The University of Georgia College of Education served as the experimental units. The first class enrolled in Summer Quarter 1984 and the last in Fall Quarter 1986. Most of the students were in College of Education graduate degree programs. [It is the opinion of the junior author, who has taught this course for several years, that these classes are representative of previous and subsequent classes in the same course.] Students in six of the classes (five of which were taught by the junior author) were administered equivalent tests

and examinations. Students from these classes constituted the design sample. Students from the two remaining classes constituted the "model assessment" sample.

Some descriptive information on all students who completed the course in the eight classes is given in Table 1. Only those students who had taken the Graduate Record Examinations prior to enrollment were considered in the final analysis. There were 122 students in the design sample (classes 1-6) and 51 students in the model assessment sample (classes 7 & 8).

Criterion

Since it is difficult to maintain contact with students after they complete the course, we decided to focus on an immediate criterion as opposed to an intermediate or ultimate criterion (Crocker & Algina, 1986, p. 225). The immediate criterion is end-of-course achievement in the introductory statistics class. Specifically the criterion variable, SCORE, is defined as a linear composite of Z transformations of the student scores on the in-class midterm and final examinations. The weights for midterm and final examination are 1.0 and 1.5, respectively: SCORE = 1.0 * ZMIDTERM + 1.5 * ZFINALEXAM. The raw-to-standard score transformation employed the mean and standard deviation based on classes 1-6.

Although four different textbooks (Glass & Hopkins, 1984; Hinkle, Wiersma, & Jurs, 1979; Iman & Conover, 1983; Wright, 1976) were used with the eight classes, the material covered in the course on introductory statistical methods was quite comparable across the classes. In classes 1-6 the midterm test (35 multiple-choice items) covered graphical and numerical

descriptors for data distributions. In the same six classes, the final examination (45 multiple-choice items) covered probability, probability distributions, estimation, and introduction to statistical testing. (Some test and examination items pertained to computation; however, the focus was on concepts and higher-level cognitive performance.) It may be argued that instructional performance was fairly constant, and that the six midterm and final examinations had comparable difficulty and internal consistency levels. For one administration of the midterm, the mean number of correct responses (total score of 35) was 21.8 and the Cronbach alpha value was .84; the respective values for one administration of the final examination (total score of 45) were 27.7 and .83. In essence it is assumed that a common scale of measurement was used for all six midterm examinations and for all six final examinations.

Predictors

In selecting predictor variables, Pedhazur (1982, p. 138) suggests attending to theoretical considerations and previous research evidence. There is some empirical evidence (e.g., Bledsoe & Perkins, 1976; Brown, 1933(I); Woelke & Leitner, 1980) that basic mathematical abilities can contribute to the prediction of introductory statistics achievement. Educators generally believe that previous relevant knowledge and skill will affect student achievement in new learning situations. Elmore and Vasu (1980) conducted a study examining the relationship between several affective variables and achievement in statistics. In their review of previous studies they noted that

the correlation between statistics achievement and affective variables was generally low. Elmore and Vasu did not consider measures of specific arithmetic and algebra skills in their study but did report significant correlations between two attitudinal variables and statistics achievement. Some type of specific arithmetic/algebra skill measures were included in most of the studies reviewed by these authors which reported low correlation between affective measures and statistics achievement. The present authors interpret this as indicating that affective variables contribute little to the prediction of statistics achievement when measures of specific arithmetic/algebra skills are also included as predictors. Based on previous research and instructional considerations, the current authors decided to consider predictor variables designed to measure mathematics/ algebra achievement or skill level in preference to affective predictors.

Various algebra and arithmetic achievement skills were sampled by a locally developed pre-statistics inventory. The seven scales of this inventory , the abbreviation as used throughout this paper, the content areas, and maximum number of points are listed below:

1) S1. Operations with integers, common fractions, and decimal fractions (25 points maximum),
2) S2. Proportions and percents (8 points),
3) S3. Squaring and extracting square roots (6 points),
4) S4. Operations with signed numbers (8 points),
5) S5. Operations with simple formulas and construction of simple formulas (8 points),

6) S6. Linear graphs (6 points), and

7) S7. Miscellaneous -- terms, inequalities, symbolism, etc. (13 points).

The sum of these seven scale scores, labeled TOTAL (74 points), was also considered as a predictor measure.

In addition to the seven scale scores and TOTAL score, three predictor measures were obtained from the Graduate Record Examinations; the Verbal score (GREV), Quantitative score (GREQ), and the product of the Verbal and Quantitative scores (GREVQ). Cohen (1978) has suggested the use of product scores in regression models to represent nonadditive or interaction effects between two variables. Because many statistics problems are presented in narrative form, the present authors believe that verbal and quantitative achievement may interact to effect achievement in statistics. It is interesting to note that in ten studies reviewed, the Graduate Record Examinations scores were used as predictor measures only by Elmore & Vasu (1986) and by Noble (1986). These scores are readily available for most students, being an admission requirement in many programs, and seem a natural choice for predictors with statistics achievement as the criterion. The GRE scores were selected because of their availability and their apparent relevance.

A matrix of correlations (see Table 2) among the predictors and between the predictors and the criterion may be useful in screening initially chosen measures. Predictors having near zero correlation with the criterion would be suspect as useful predictors. For the current study correlations of the predictors

/ith the criterion range from a minimum of r=.20 for GREV to a
iaximum of r=.50 for GREQ. Therefore no potential predictors
/ere eliminated at this point because of low correlation with the
:riterion. Predictors which correlate highly with one another
iay indicate redundancy of information. If two such variables
:re detected one may be eliminated from the analysis or when
ogically appropriate the items used to measure the two variables
iay be combined. For the current study the highest predictor
ntercorrelation was between GREV and GREVQ (r=.79). This is not
surprisingly strong correlation considering that GREVQ is a
unction of GREV. No other predictor intercorrelation approached
his magnitude. Therefore no variables were eliminated at this
tage because of redundancy.

Pedhazur (1982, pp. 32-36) discusses the assumptions
nderlying multiple regression analysis. He describes this
nalysis technique as robust. Stevens (1984, p. 335) has
uggested plotting the criterion values as a visual means of as-
essing approximate normalcy. Such a plot of the criterion
easures in this study suggest approximate normalcy (see Figure
). In addition, Stevens suggests plotting the predictor
ariables, not to check for normalcy, but as a visual aid in
etecting outliers in the predictor space.

Examining the Data

Errors in the data may seriously distort efforts at
rediction. Recording of data, transposing the data, and
ntering the data into the computer are all opportunities for
rrors. We used the computer to list the data as they were

entered and compared this listing with the original data. Also, we find the use of frequency histograms and stem-and-leaf plots of predictor and criterion measures useful in detecting extreme values which may be errors. In addition, these plots help to identify segments of the predictor range which are sparsely represented by the data sampled. If the data set is quite large and variables can only assume restricted values, then one may write computer statements to isolate all observations with variable values out of the allowed range of values. This approach may still allow errors into the data set. The best approach, though time consuming, is to list the data and make comparisons to the original observation records.

Searching for an Efficient Model

Two questions must be answered before the parameters of a linear regression model are estimated. First, what is the optimum number of the available predictors that should be retained in the model? Secondly, what is the best combination of predictors for a subset of chosen size? [This brings up a related question: How is one model deemed better than another? Cross-validation results may be the ultimate test of the appropriateness of a prediction model. The use of a validation or assessment sample in the current study is discussed later.] Three indices of model effectiveness will be examined at this time. A better model will account for more of the variability in the criterion variable and reduce the error in the predicted scores. Since the adjusted R-squared value reflects the proportion of variance in the criterion accounted for by the

model, one index of a good model is the adjusted R-squared value. The higher the adjusted R-squared value the better the model fits the sample data. The RSQUARE procedure in SAS (SAS Institute Inc., 1985) was used to calculate the adjusted R-squared values for all possible combinations of the predictor variables in all possible size subsets of the predictor variables. The adjustment formula used by SAS is

$$\text{adjusted R-squared} = 1-(1-\text{R-squared})(n-1)/(n-p)$$

where n is the number of units sampled and p is the number of parameters in the model including the intercept. The highest adjusted R-squared value for each predictor subset size may be plotted against the subset size (see Figure 2).

A second index is the Mean-Square Error which is equal to (Sum-of-Squares Error)/(n-p). The model with the lowest Mean-Square Error value has minimized the error and reflects a good fit of the model to the sample data. The lowest Mean-Square Error for each subset size may be plotted against the subset size (see Figure 3). A third index, Mallows' Cp statistic, is a measure of bias in estimating the parameters of the regression model (Chatterjee & Price, 1977, pp. 198-199). A model that is too simple (omits important predictors) may result in biased regression weights and biased prediction, while an overly complicated model (including predictors that add little or nothing in addition to the predictors already in the model) may result in large variance both in the regression weights and the predicted values (Myers, 1986, pp. 112-114). As Cp exceeds p the

bias in estimation of model parameters becomes more severe. Especially in the use of regression for prediction, one wishes to minimize the bias of estimating the model parameters. The values of Cp against p may also be plotted (see Figure 4). A good model will have a "low" value of Cp and one that is "close" to p.

These three indices, adjusted R-squared value, Mean Square Error, and Mallows' Cp, may be examined simultaneously to determine a good subset size. The three indices may not point to exactly the same subset size. After simultaneously considering the three indices one may decide to retain two or more predictor subset sizes. Examination of Figure 2 reveals that a model with three predictors will achieve the largest adjusted R-squared value. The smallest Mean-Square Error value is associated with a model of three predictors as can be seen in Figure 3. Examination of Figure 4 suggest that a model with more than three predictors may be desirable. As the predictor subset size is increased the value of Cp approaches p. But, at the same time the value of adjusted R-square begins to fall and the value of Mean-Square Error increases. It should be noted, as often happens, that neither of the three statistics indicates a predictor subset size that is greatly superior to others. Accordingly, we considered models of five and six predictors. [One additional model was considered; TOTAL score along with GREV and GREQ constituted the predictors of a third model. This model is simple and may reveal the advantages or disadvantages of summing the scale scores of the pre-statistics inventory into one score.]

Now that we have decided to look at models of five and six

predictors, we must decide which particular subset of variables to use in our model. In the SAS computer printout (see Table 3 for subset of six predictors) the combinations of variables in each subset size are ordered in accordance with the adjusted R-squared value. One might feel compelled to select the best combination of variables as indicated by the highest adjusted R-squared value (lowest Mean-Square Error, or Cp value closest to p). Examination of the actual values will reveal negligible difference in the adjusted R-squared value for the best and second best combination of variables in each subset size. Since the regression procedure capitalizes on sample specific relationships one need not feel bound to select the subset of variables with the highest adjusted R-squared value realizing that when the difference between the adjusted R-squared value for the best and second best subsets is negligible, the order of the best and second best set of variables of a given subset size may very well be reversed when a different sample is examined. With this in mind the present authors chose the models retaining the following variables for the five and six predictor variables models, respectively; S4, S5, S6, GREV, GREVQ and S1, S4, S5, S6, GREV, GREVQ. It was desirable from a substantive viewpoint to retain a variable subset with the GREV and GREVQ variables.

Using Regression Diagnostics

Regression diagnostic methodology is relatively new and the jury is still out on the relative usefulness of indices to detect influential data points and outliers. We restricted our

diagnostics to examination of the influence of single data points; the study of the influence of groups of data points is in its infancy, with very little practical guidance having been offered--see discussion by Atkinson and by Hoaglin and Kempthorne in Chatterjee and Hadi (1986). Also, little guidance has been suggested for the simultaneous consideration of predictor variable selection and outlier detection. [We selected predictors first and diagnosed second with an admission of potentially misleading results.]

In this section we will discuss the practical application of some of these techniques. After selecting the variables for models of five and six predictors the SAS PROC REG (regression procedure) was used to estimate a linear model relating the predictors to the criterion. Options were selected to print the actual criterion value and the predicted criterion value for each observation. The difference between the predicted value and the observed value is the simple residual value. These values were examined en masse and individually.

Assumptions Check

A plot of the residuals against the predicted score may reveal model underspecification (omission of important predictor variables), violation of the assumption of homogeneity of variance, departure from normalcy in the model errors, and extreme or suspect data points (Draper & Smith, 1981, pp. 141-147; Myers, 1986, p. 138). Consider the hypothetical plots in Figure 5. With an appropriately fitted linear regression model, the plot of the residual values against the predicted scores should look similar to plot 1 in Figure 5. A graph such as plot 2 in Figure

5 indicates that the variances are not constant suggesting a need for a weighted least squares analysis or a transformation of the criterion variable. A graph such as plot 3 in Figure 5 indicates an error in analysis; the departure from the fitted equation is systematic. This effect can also be caused by incorrectly omitting an intercept term in the model. A graph such as plot 4 in Figure 5 indicates an inadequate model--need for extra terms in the model (e.g. squares or crossproducts) or need for a transformation on the criterion values before analysis. After visually inspecting Figure 6, the graph of residuals against predicted scores for the five variable model, concerns of the type just discussed were set aside.

Outliers

An outlier is defined as an individual observation with a relatively large absolute value of residual score. We proceed to examine outliers individually. Since any model is an approximation of the data, outliers are not uncommon. Outlier observations may represent data error or they may be units that for some reason represent a population different than the majority of units in the sample. Outliers may have some characteristic in common that determines a different functional relationship between the predictor and criterion variables for them than for the majority of the sample. If this is so then one can search for the characteristic and determine if it is an important variable that should be included in future predictor models. Outliers may have an excessively strong influence on the estimation of regression weights compared to the influence of

other data points. If this is the case the outlier is also an influential observation point. Stevens (1984) (and others; e.g. Draper & Smith, 1981, p. 169, Weisberg, 1985, pp. 114-125, Chaterjee & Hadi, 1986, p. 380) point out that an outlier may or may not be an influential observation in determining estimates o regression parameters. Conversely, an observation may be influential and not be an outlier. We will identify outlier observations mindful of their impact on fit of the model to the sample data and their influence on estimation of the regression parameters. Also, observations which are not outliers but which are influential will be identified and examined. This will be discussed below. For a more technical discussion of regression diagnostics pertaining to outliers and influential data points see Cook and Weisberg (1982).

The simple residual, the standardized residual, and the studentized residual all are indicators of outliers in the crit rion space. We accept the argument of Stevens (1984, p. 336) that the studentized residual is a more sensitive detector of outliers. For more discussion on this and alternate names for these statistics, see Chatterjee and Hadi (1986). A studentize residual is referenced to the Student t distribution with N-p- degrees of freedom (Chatterjee & Hadi, 1986 p. 380). As the choice of alpha level in hypothesis testing is arbitrary, so i the choice of a critical value for studentized residuals. A stem-and-leaf plot of residuals may be constructed to identify data points which are outliers relative to other data points i the sample.

Observations may be outliers in the predictor space

(Stevens, 1984, p. 337) because of extreme values on one or more predictor measures or because they represent a rare combination of predictor values. Such observations will have a relatively large diagonal element in the so-called HAT matrix, h sub ii. These observations are also called high leverage points. High leverage points may or may not be influential. How large is a relatively large HAT diagonal element? A critical value of 2p/n has been suggested (Chatterjee & Hadi, 1986). For a discussion of critical values for influence indicators in general see Belsley, Kuh, and Welsh (1980). We prefer to consider the h sub ii values in context with the values for all observations by constructing a stem-and-leaf plot. An example will follow in the subsection, Illustration.

Influence Indicators

Several indicators of influence are reviewed by Chatterjee and Hadi (1986). Seven excellent comment reviews follow that article. There is some confusion about just what is being influenced in the influence measure. In addition there are only rule-of-thumb guidelines for the analyst to use in deciding when an influence measure is large enough to warrant concern. In regard to the latter, instead of adopting a rule-of-thumb critical value a stem-and-leaf plot may be constructed for each influence indicator. A visual inspection of those plots will reveal observations with influence indicator values that are large relative to others in the sample. This approach may be criticized as being arbitrary, as are the rule-of-thumb approaches. It is believed that these graphical approaches will

give the researcher a better feel for his/her data than employing rule-of-thumb values. The influence indicators considered here reflect influence on the b vector of regression weight estimates, the variance/covariance of the b vector, or a combination of both, and the influence on a single b value estimating a single model predictor parameter.

Cook's D or Cook's distance, sometimes abbreviated D sub i and C sub i (Chatterjee & Hadi, 1986, p. 383) measures the change in distance between the b vector as estimated with the ith observation in the model and the b vector as estimated with the ith observation removed from the model. It therefore indicates the influence of the ith observation on the parameter estimates of all the predictor weights (see comments by Hoaglin in Chatterjee and Hadi, 1986). The same information is also provided by Welsh's distance, and a modified Cook's distance. Different rule-of-thumb critical values are suggested for these influence indicators (Chatterjee & Hadi, 1986). Each of these indicators should identify influential observations in the same rank order.

The covariance ratio (CVR) and the Cook-Weisberg statistic provide information on the influence of the ith observation on the variability of the parameter estimates of the b vector elements. An index called DFFITS indicates influence on both the estimates of the b vector and the variance/covariance of the predictor parameter estimates.

Finally an observation may have strong influence on only one of the b values. This is indicated by an index called DFBETA. Plots of DFBETA against observation number are also referred to as partial regression leverage plots.

The numerous plots referred to above are not all reproduced herein. They are easily obtained from popular computer software packages such as SAS and SPSS. Regression diagnostics were conducted for the three models considered in this paper. For economy of space, only the diagnostics for the five variable model are discussed in detail. At the end of this discussion the reader is appraised of which observations we decided to eliminate from each model. Other researchers examining the exact same data and indicators of influence and outliers may reach slightly different decisions about eliminating observations. Finally it should be noted that observations which are outliers in the predictor space but, which are not excessively influential, may represent areas in which the sample data are sparse. Such observations may prompt the researcher to collect more data.

Illustration

We turn now to the predictor models studied in the context of predicting statistics achievement. Outliers and influential data points will be identified for one model (Model 2) and the decision to delete or not delete the associated observation will be addressed. The three models and their adjusted R-squared values are listed below;

Model 1	SCORE=GREV GREQ TOTAL	adj R**2=.2983
Model 2	SCORE=S4 S5 S6 GREV GREVQ	adj R**2=.3138
Model 3	SCORE=S1 S4 S5 S6 GREV GREVQ	adj R**2=.3093

The stem-and-leaf plot of the studentized residual (RSTUDENT) for Model 2 is given in Figure 7 (each stem-and-leaf plot is accompanied with a tabular listing of extreme

observations and their values). It is apparent that observation 215 and 176 have high studentized residual values relative to the sample . Observations 88 and 148 have relatively low studentized residual values. A small studentized residual value implies that the predicted criterion value for that observation is lower than the actual criterion value. Of these four observations only 215 is a relative outlier in the predictor space as indicated by the stem-and-leaf plot of h sub ii in Figure 8. At this point one may wonder if observation 215 is representative of the population from which it is believed the sample was drawn. In this study specifically, is there something about observation 215 that makes this person not representative of students enrolled in introductory statistics courses? This question is not addressed in this paper. Merely the point is made that regression diagnostics may lead the researcher to identify data points which have some characteristic different from the majority of the sample.

We now examine the influence indicators to identify observations which have an unusually strong influence on the paramaterization of the model. Examination of the stem-and-leaf plot of Cook's D (Figure 9) reveals that observation 215 and 176 are relatively influential in determining the estimates in the $\underline{b}$ vector. The stem-and-leaf plot for the DFFITS indicator is given in Figure 10. This suggests that observation 215 and 176 are influential in determining the $\underline{b}$ vector and/or the variance of the estimates in the $\underline{b}$ vector. Examination of the stem-and-leaf plot of COVRATIO (see Figure 11) reveals observation 215 but not

176 to be influential in increasing the variance of the b vector. In essence observation 215 receives a double indictment for its influential role in determining the b vector and its relatively strong contribution to lack of fit of the model to the sample data. Elimination of these two observation points and recalculation of the regression equation should improve the predictive accuracy of the model. In addition, the removal of observation 215 and to a lesser extent 176 should increase the fit of the model to the sample data.

In examining Figure 9 and Figure 10 the reader may have noticed that observation 144 is relatively influential in determining the b vector and/or the variance of the b vector. However, this observation is not a relative outlier in the criterion space or the predictor space. Examination of stem-and-leaf plots and frequency histograms of all the model variables does not indicate that observation 144 came from a sparse region of the data. No further consideration is given to deleting this observation at this time.

Plotting DFBETA for each predictor against observation number, the so-called partial regression leverage plot, did not indicate observations which were excessively influential in estimating the b value for one predictor.

Observation 215 and 176 were removed from the sample data and the regression equation for Model 2 was recalculated. The adjusted R-squared value rose from .3138 to .3759, an increase of over 6% explained variance.

After examining stem-and-leaf plots of the outlier measures and influence indicators for the other two models we decided to

drop observation 215 and 176 from Model 1 and observation 215, 176, and 144 from Model 3. The change in adjusted R-squared for Model 1 was from .2983 to .3761 and for Model 3 from .3093 to .4047.

Assessing the Model(s)

Information was gathered from classes 7 and 8 (N=29 and 22, respectively) in order to assess the usefulness of the models. Because the same criterion was not available for these two classes, this assessment differs from the traditional "cross validation" study. The instructors in these two classes were asked to rank-order their students based on performance. The regression models were applied to the predictor values for each student in these classes to obtain a predicted criterion score. These predicted criterion scores were rank-ordered and correlated with rankings assigned by each instructor. Using Model 2, the one discussed most extensively in this paper, the correlation for class 7 was r=.524 and for class 8 r=.607. Using Model 1 and Model 3 the respective correlations were all at least .60.

Finally we examined the use of Model 2 to predict high achievers who might benefit from accelerated instruction and low achievers who might benefit form remedial instruction. The junior author (five classes) plus the instructor of one other class identified those students who were judged to have been capable to benefit from an accelerated instructional experience in statistical methods. The judgments were based on such things

as completed work, perceived maturity in quantitative methods, work habits, persistence, etc., as well as on test performance. The judgments were made not knowing the predicted or actual SCORE value for each student.

Of the 122 design-sample students, 11 were judged to have been capable of succeeding in an accelerated course. [The junior author had taught two such course sequences prior to 1984.] Of these 11, nine obtained a predicted SCORE value (via Model 2) above +1.75. [The use of a cut-off value of +1.75 was judged reasonable, based on the junior author's use of SCORE with many other classes.] There was one false-positive, i.e., one student was empirically predicted to have been capable but was not judged capable by the instructor. And there were two false-negatives. [See Table 4.] With a false-positive error judged as being more serious, the resulting "hit-rate" was .82 (9/11). On the other hand, the hit-rate for predicting those students who might benefit from some remedial experience was extremely low (less than chance). It appears that Model 2, at least, has reasonable predictive validity in the sense that it is potentially useful for identifying those students who would be capable of benefiting from an accelerated course experience, whereas model validity is lacking for predicting remedial-instruction student candidates.

Discussion

In general one may question the representativeness of students enrolled in introductory statistical methods courses offered by the College of Education at The University of Georgia. The mean scores on the Graduate Record Examinations for these

students were near the national average. The variability in end-of-course achievement scores not accounted for by the models is typical of, if not lower than, that found in other studies with a similar purpose. One might hypothesize various factors that could account for this remaining variance--e.g., motivation, study habits, test taking skills, academic persistance, academic maturity, and research experience. It was assumed in this study that a serious effort was put forth in completing the pre-statistics inventory, and that the reported GRE scores were correct.

Predictive measures used in the models are readily obtainable and all contributed significantly to the obtained predictive acuracy. The effectiveness of each model was assesed in three ways: (1) an adjusted R-squared value; (2) correlation of instructor-judged rank orderings of two assesment classes against rank orderings of predicted SCORE; and (3) prediction of those students who might be advised to enroll in an accelerated course. The three assessment measures were considered "respectable": (1) adjusted R-squared values (after deletion of observations identified as outliers and/or influential) of .376, .376, and .405 for Models 1 through 3, respectively; (2) rank correlations of about .6; (3) and a ratio of 9 out of 11 students judged by instructors as capable of benefiting from an accelerated instructional experience correctly identified. Thus of the three questions posed at the outset of the paper concerning regression and statistics achievement, the first two may be answered in the affirmitive and the lattter negatively for

this study.

References

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). Regression diagnostics: identifying influential data and sources of collinearity. New York: Wiley.

Bendig, A. W., & Hughes, J. B. (1954). Student attitude and achievement in a course in introductory statistics. Journal of Educational Psychology, 45, 268-276.

Bledsoe, J. C., & Perkins, M. L. (1976). Prediction of success in elementary statistics: Three replications. Psychological Reports, 38, 723-726.

Brown, R. (1933). Mathematical difficulties of students of educational statistics. Contributions to Education, No. 569. New York: Teachers College, Columbia University.

Chatterjee, S., & Hadi, A. S. (1986). Influential observations, high leverage points, and outliers in linear regression (with comments). Statistical Science, 1, 379-416.

Chatterjee, S., & Price, B. (1977). Regression analysis by example. New York: Wiley.

Cohen, J. (1978). Partialed products are interactions; Partialed powers are curve components. Psychological Bulletin, 85, 858-866.

Cook, R. D., & Weisberg, S. (1982). Residuals and influence in regression. New York: Chapman and Hall.

Crocker, L, & Algina, J. (1986). Introduction to classical & modern test theory. New York: Holt, Rinehart & Winston.

Draper, N. R., & Smith, H. (1981). Applied regression analysis (2nd ed.). New York: Wiley.

Elmore, P. B., & Vasu, E. S. (1980). Relationship between selected variables and statistics achievement: Building a theoretical model. Journal of Educational Psychology, 72, 457-467.

Elmore, P. B., & Vasu, E. S. (1986). A model of statistics achievement using spatial ability, feminist attitudes and mathematics-related variables as predictors. Educational and Psychological Measurement, 46, 215-222.

Feij, J. A. (1976). Field independence, impulsiveness, high school training, and academic achievement. Journal of Educational Psychology, 68, 793-799.

Feinberg, L. B., & Halperin, S. (1978). Affective and cognitive correlates of course performance in introductory statistics. Journal of Experimental Education, 46, 11-18.

Glass, G. V., & Hopkins, K. D. (1984). Statistical methods in education and psychology. Englewood Cliffs, NJ: Prentice-Hall.

Harvey, A. L., Plake, B. S., & Wise, S. L. (1985, April). The validity of six beliefs about factors related to statistics achievement. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1979). Applied statistics for the behavioral sciences. Boston: Houghton Mifflin.

Iman, R. L., & Conover, W. J. (1983). A modern approach to statistics. New York: Wiley.

Myers, R. H. (1986). Classical and modern regression with applications. Boston: Duxbury.

Noble, R. F. (1986, April). Multiple regression analysis of six predictor variables of academic achievement in the course introduction to research. Paper presented at the annual meeting of the Americal Educational Research Association, San Francisco.

Pedhazur, E. J. (1982). Multiple regression in behavioral research (2nd ed.). New York: Holt, Rinehart and Winston.

Pruzek, R. M. (1964). Prediction of success in elementary statistics. Journal of Educational Measurement, 1, 165-167.

SAS Institute Inc. (1985). SAS user's guide: Statistics, version 5 edition. Cary, NC: SAS Institute Inc.

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. Psychological Bulletin, 95, 334-344.

Weisberg, S. (1985). Applied linear regression (2nd ed.). New York: Wiley.

Woelke, P. L., & Leitner, D. W. (1980). Gender differences in performance on variables related to achievement in graduate-level statistics. Psychological Reports, 47, 1119-1125.

Wright, R. L. D. (1976). Understanding statistics. New York: Harcourt Brace Jovanovich.

Table 1

Gender and Degree Program for Subjects

	Design Sample	Assessment	Sample
Class(es)	1-6	7	8
Gender			
F	87	13	20
M	35	9	9
Degree			
Master	87	15	18
Specialist	7	1	0
Doctorate	28	6	11

le 2

lictor/Criterion Correlations, Means, and Standard Deviations

S1	S2	S3	S4	S5	S6	S7	GREV	GREQ	GREVQ	Mean	SD
1.000										20.7	3.45
.387	1.000									5.8	2.65
.569	.335	1.000								3.6	1.95
.422	.287	.423	1.000							6.7	1.43
.289	.268	.222	.339	1.000						6.8	1.51
.364	.204	.343	.474	.293	1.000					3.3	1.90
.536	.279	.430	.594	.521	.576	1.000				9.8	2.55
.115	.048	.142	-.019	-.008	-.086	.027	1.000			516.0	98.80
.527	.307	.538	.448	.267	.520	.541	.003	1.000		535.2	84.10
.488	.233	.427	.259	.168	.263	.356	.791	.598	1.000	276200.8	72115.17
.355	.211	.330	.328	.228	.417	.378	.204	.497	.472	0.0	2.08

Fifteen "Best" Subsets of Size Five, Six, and Seven

IN	R-SQUARE	ADJUSTED	MSE	C(P)	VARIABLES IN MODEL
4	0.338703	0.315577	2.96871	-0.03907	S1 S6 GREQ GREVQ
4	0.338895	0.316282	2.96565	-.154462	S5 S6 GREQ GREVQ
4	0.338963	0.316363	2.96530	-.167647	S4 S6 GREQ GREVQ
4	0.339483	0.316902	2.96296	-.255735	S1 S6 GREV GREVQ
4	0.339583	0.317005	2.96252	-.272582	S5 S6 GREV GREVQ
4	0.340093	0.317532	2.96023	-.358947	S4 S6 GREV GREVQ
5	0.339834	0.311379	2.98692	1.68491	S5 S6 S7 GREV GREVQ
5	0.339940	0.311490	2.98644	1.66694	S1 S6 S7 GREV GREVQ
5	0.340046	0.311600	2.98596	1.649	S1 S2 S6 GREV GREVQ
5	0.340055	0.311609	2.98592	1.64758	S3 S5 S6 GREV GREVQ
5	0.340077	0.311632	2.98582	1.6438	S2 S5 S6 GREV GREVQ
5	0.340154	0.311712	2.98547	1.63086	S4 S6 GREV GREVQ GREVQ
5	0.340236	0.311798	2.98510	1.6169	S3 S4 S6 GREV GREVQ
5	0.340344	0.311910	2.98461	1.59868	S4 S6 S7 GREV GREVQ
5	0.340616	0.312194	2.98338	1.55267	S2 S4 S6 GREV GREVQ
5	0.340668	0.312249	2.98315	1.5438	S1 S4 S6 GREQ GREVQ
5	0.340869	0.312459	2.98224	1.50978	S1 S5 S6 GREQ GREVQ
5	0.341267	0.312873	2.98044	1.44252	S4 S5 S6 GREQ GREVQ
5	0.341824	0.313454	2.97792	1.34834	S1 S5 S6 GREV GREVQ
5	0.341938	0.313573	2.97740	1.32902	S1 S4 S6 GREV GREVQ
5	0.342108	0.313751	2.97663	1.30017	S4 S5 S6 GREV GREVQ
6	0.341591	0.307239	3.00488	3.38775	S2 S4 S5 S6 GREQ GREVQ
6	0.341828	0.307488	3.00379	3.34767	S1 S5 S6 S7 GREV GREVQ
6	0.341838	0.307499	3.00375	3.34591	S1 S3 S5 S6 GREV GREVQ
6	0.341889	0.307553	3.00351	3.33728	S1 S5 S6 GREV GREQ GREVQ
6	0.341947	0.307613	3.00325	3.32752	S1 S4 S6 GREV GREQ GREVQ
6	0.341951	0.307618	3.00323	3.32681	S1 S3 S4 S6 GREV GREVQ
6	0.341958	0.307625	3.00320	3.32568	S1 S4 S6 S7 GREV GREVQ
6	0.342072	0.307745	3.00268	3.30638	S1 S2 S5 S6 GREV GREVQ
6	0.342111	0.307786	3.00250	3.29973	S4 S5 S6 S7 GREV GREVQ
6	0.342215	0.307896	3.00203	3.28212	S4 S5 S6 GREV GREQ GREVQ
6	0.342221	0.307902	3.00200	3.28118	S3 S4 S5 S6 GREV GREVQ
6	0.342236	0.307918	3.00193	3.27855	S1 S2 S4 S6 GREV GREVQ
6	0.342357	0.308045	3.00138	3.25814	S2 S4 S5 S6 GREV GREVQ
6	0.342550	0.308248	3.00050	3.22552	S1 S4 S5 S6 GREQ GREVQ
6	0.343555	0.309306	2.99591	3.05546	S1 S4 S5 S6 GREV GREVQ
7	0.342227	0.301838	3.02830	5.28006	S3 S4 S5 S6 S7 GREV GREVQ
7	0.342251	0.301863	3.02820	5.27609	S1 S2 S4 S6 GREV GREQ GREVQ
7	0.342252	0.301864	3.02819	5.27593	S1 S2 S4 S6 S7 GREV GREVQ
7	0.342271	0.301884	3.02810	5.27263	S1 S2 S3 S4 S6 GREV GREVQ
7	0.342292	0.301907	3.02800	5.26901	S3 S4 S5 S6 GREV GREQ GREVQ
7	0.342359	0.301978	3.02770	5.25768	S2 S4 S5 S6 S7 GREV GREVQ
7	0.342422	0.302045	3.02741	5.24706	S2 S3 S4 S5 S6 GREV GREVQ
7	0.342471	0.302097	3.02718	5.23875	S2 S4 S5 S6 GREV GREQ GREVQ
7	0.342603	0.302237	3.02657	5.2164	S1 S3 S4 S5 S6 GREQ GREVQ
7	0.342681	0.302319	3.02622	5.20333	S1 S4 S5 S6 S7 GREQ GREVQ
7	0.342742	0.302384	3.02593	5.193	S1 S2 S4 S5 S6 GREQ GREVQ
7	0.343565	0.303257	3.02215	5.05378	S1 S3 S4 S5 S6 GREV GREVQ
7	0.343589	0.303283	3.02204	5.0497	S1 S4 S5 S6 GREV GREQ GREVQ
7	0.343682	0.303382	3.02161	5.03391	S1 S4 S5 S6 S7 GREV GREVQ
7	0.343684	0.303384	3.02160	5.03354	S1 S2 S4 S5 S6 GREV GREVQ

Table 4

Number of Students Predicted to Benefit from Accelerated Course

Model 2

		Prediction: Yes	Prediction: No	
Instructor Judgment	Yes	9	2	11
	No	1	110	111
		10	112	122

Note. Judgments/predictions are for the six design-sample classes.

Figure 1. Frequency histogram of SCORE.

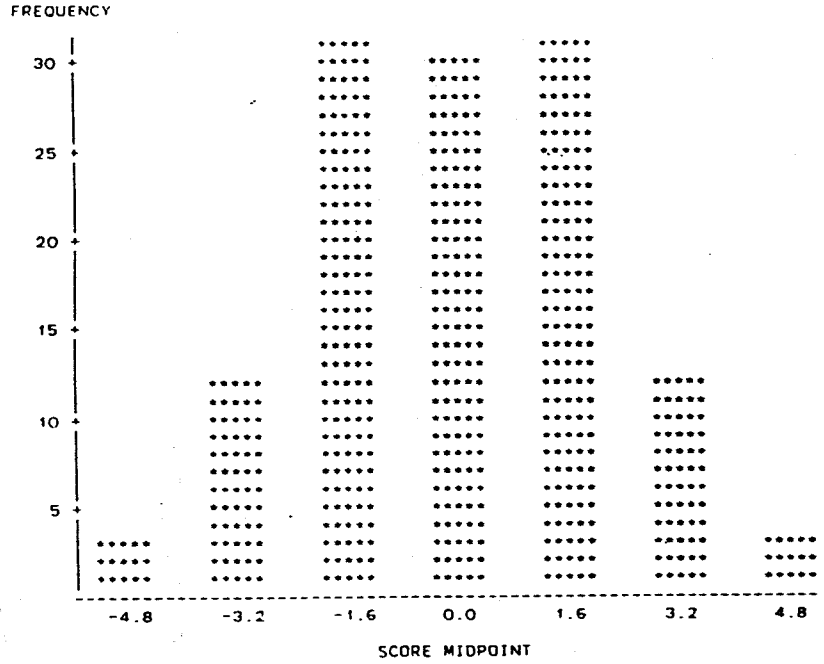


Figure 2. Plot of adjusted R^2 against sub set size.

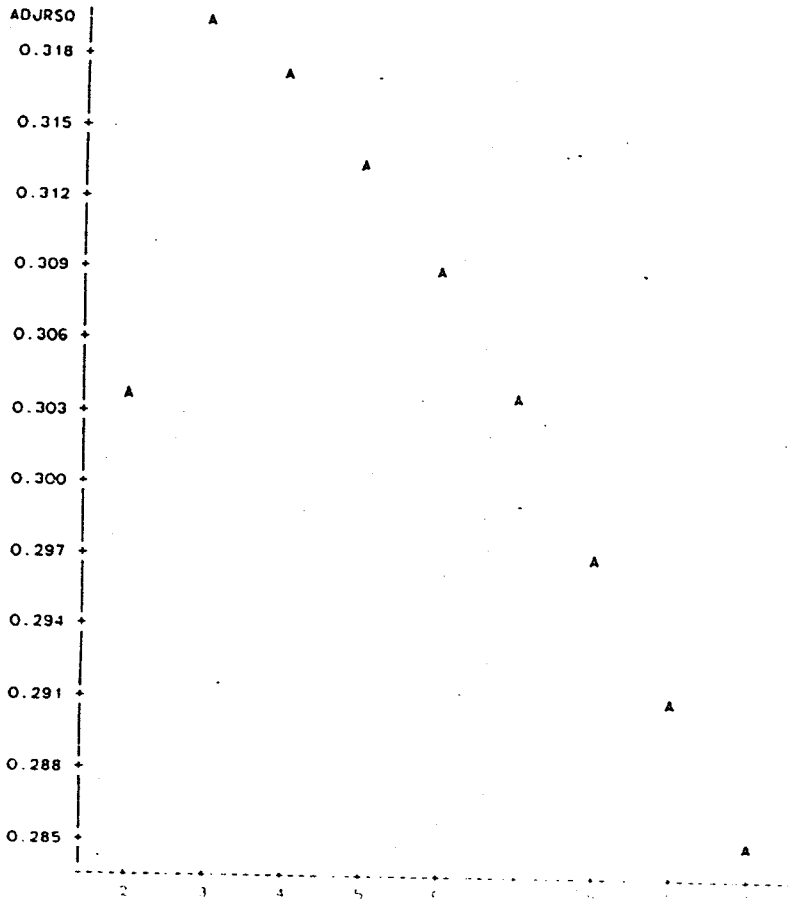


Figure 3. Plot of mean square error against sub set size.

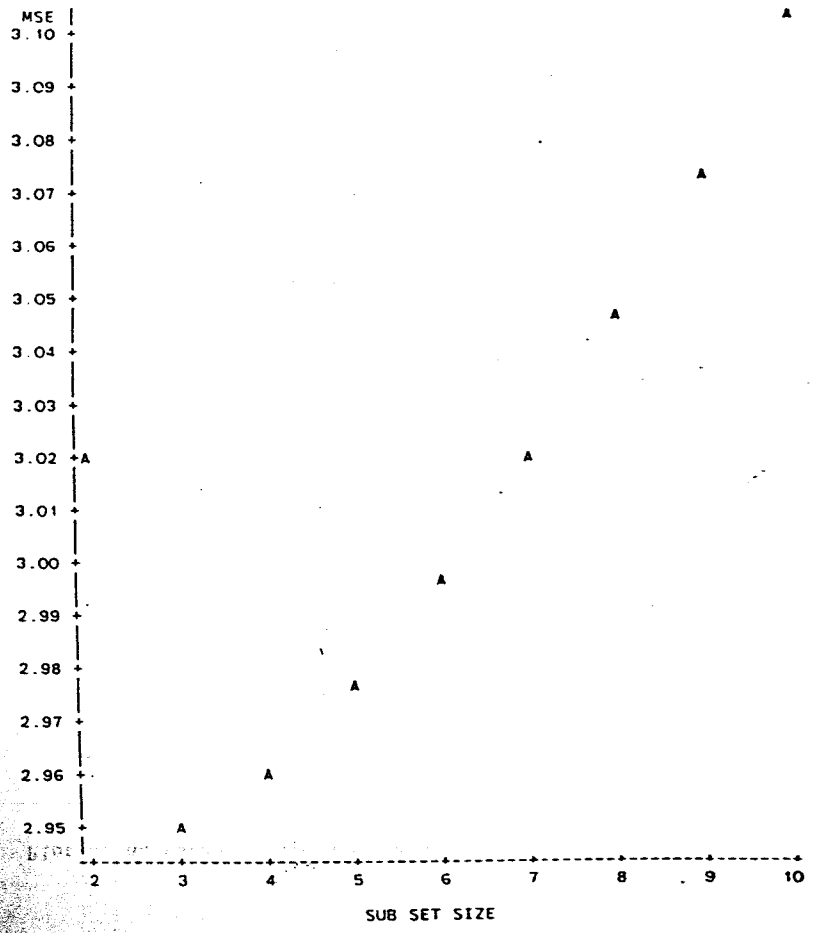


Figure 4. Plot of Cp against p.

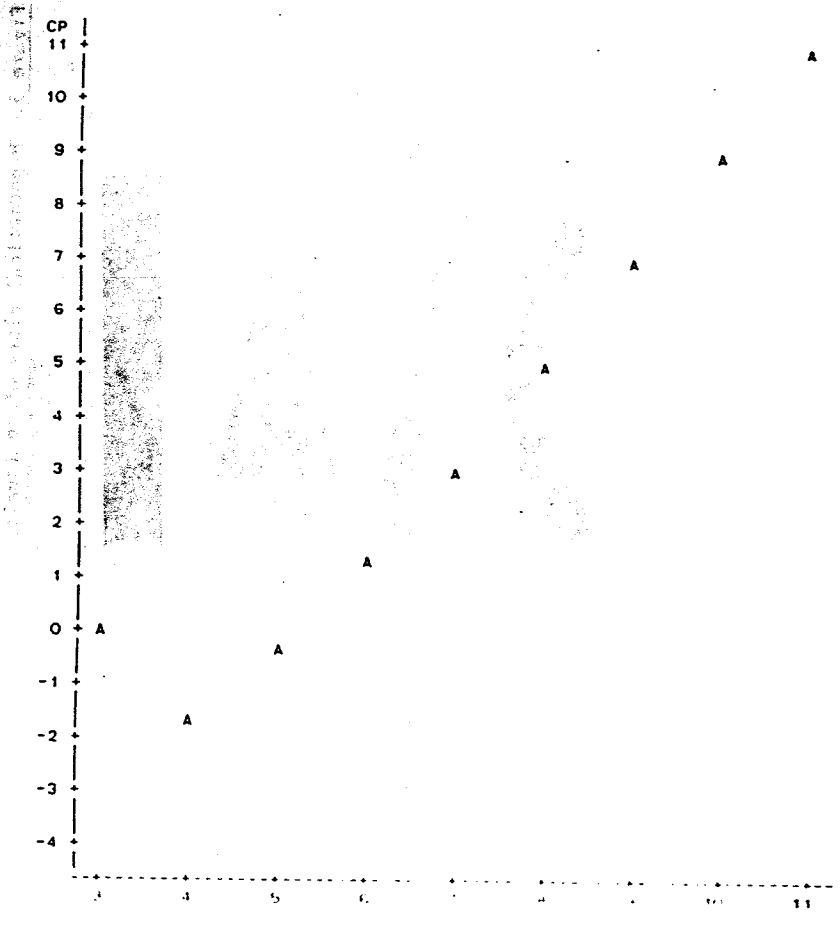


Figure 5. Hypothetical plots of residuals.

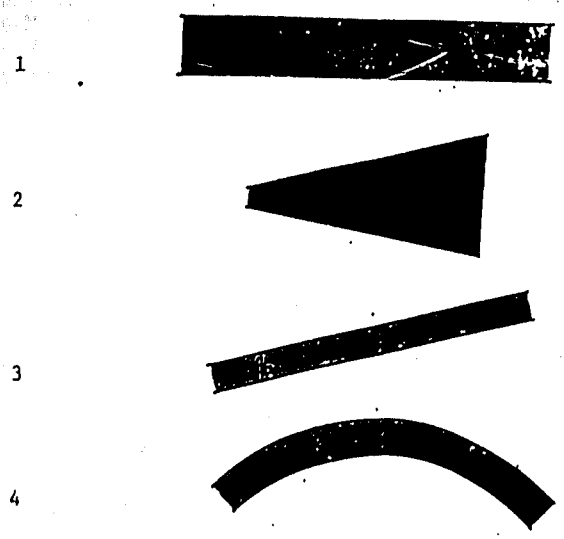
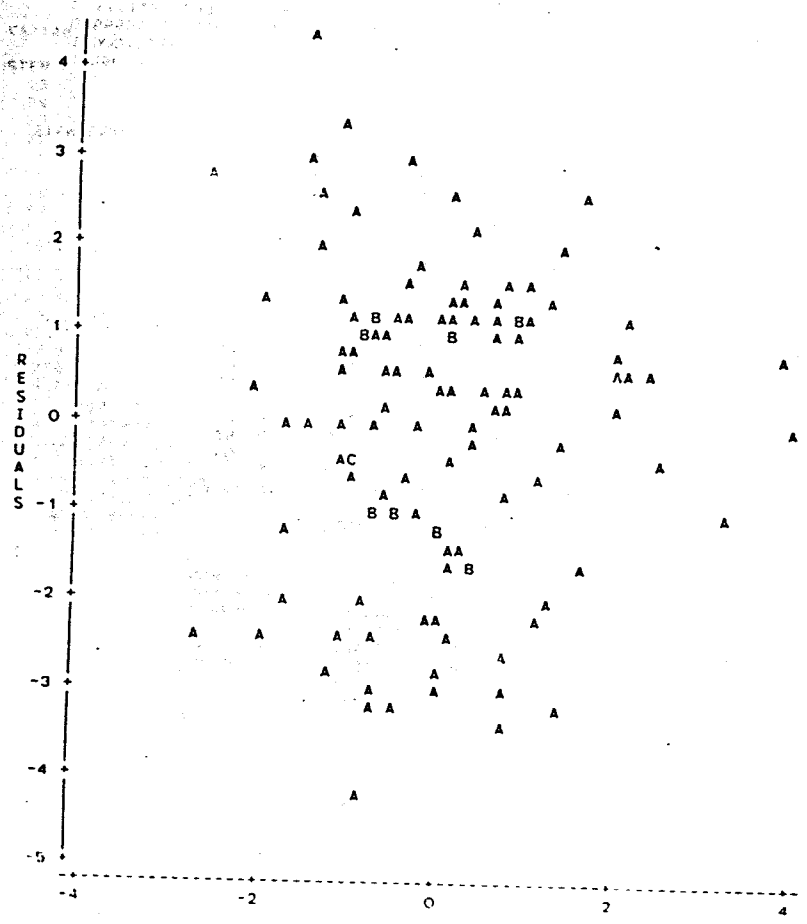


Figure 6. Plot of residuals against predicted value for Model 2.



VARIABLE=RSTUDENT STUDENTIZED RESIDUAL

MOMENTS

N	122	SUM WGTS	122
MEAN	0.00442784	SUM	0.540197
STD DEV	1.01898	VARIANCE	1.03831
SKEWNESS	-0.158568	KURTOSIS	-0.0695343
USS	125.638	CSS	125.636
CV	23012.9	STD MEAN	0.0922538
T:MEAN=0	0.0479963	PROB>|T|	0.961798
SGN RANK	162.5	PROB>|S|	0.678941
NUM ^= 0	122		

QUANTILES(DEF=4)

100% MAX	3.10708	99%	2.87259
75% Q3	0.704903	95%	1.60965
50% MED	0.179083	90%	1.16035
25% Q1	-0.661022	10%	-1.47887
0% MIN	-2.5126	5%	-1.82563
		1%	-2.40553
RANGE	5.61968		
Q3-Q1	1.36592		
MODE	-2.5126		

EXTREMES

LOWEST	ID	HIGHEST	ID
-2.5126(	88)	1.736(	216)
-2.04709(	148)	1.75613(	17)
-1.92431(	212)	1.79609(	144)
-1.92394(	151)	2.08756(	176)
-1.88357(	207)	3.10708(	215)

STEM	LEAF	#
3	1	1
2		
2	1	1
1	556788	6
1	00012234	8
0	5555566666666677777778888888899	32
0	111122222333333444	18
-0	43333222222000000	17
-0	999877666665555	15
-1	433222100	9
-1	9998877655555	13
-2	0	1
-2	5	1

BOXPLOT

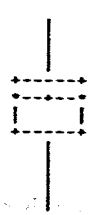


Figure 8. Diagonal elements of the HAT matrix.

VARIABLE=H H LEVERAGE

```
STEM LEAF                                #
  25 2                                   1
  24
  23
  22
  21
  20
  19 4
  18                                     1
  17
  16
  15
  14 7
  13 0                                   1
  12                                     1
  11 16
  10 3                                   2
   9 8                                   1
   8 67789                               1
   7 0133345677                          5
   6 011137889                          10
   5 000222456899                        9
   4 1122222223345567788                12
   3 00111222222334456677888899         19
   2 011111223344555567888899           26
   1 2356788999                         23
     ----+----+----+----+----+--        10
MULTIPLY STEM.LEAF BY 10**-02
```

VARIABLE=H H LEVERAGE

MOMENTS

N	122	SUM WGTS	122
MEAN	0.0491803	SUM	6
STD DEV	0.0339644	VARIANCE	0.00115358
SKEWNESS	2.85553	KURTOSIS	12.3066
USS	0.434665	CSS	0.139583
CV	69.0609	STD MEAN	0.00307499
T:MEAN=0	15.9937	PROB>|T|	0.0001
SGN RANK	3751.5	PROB>|S|	0.0001
NUM ^= 0	122		

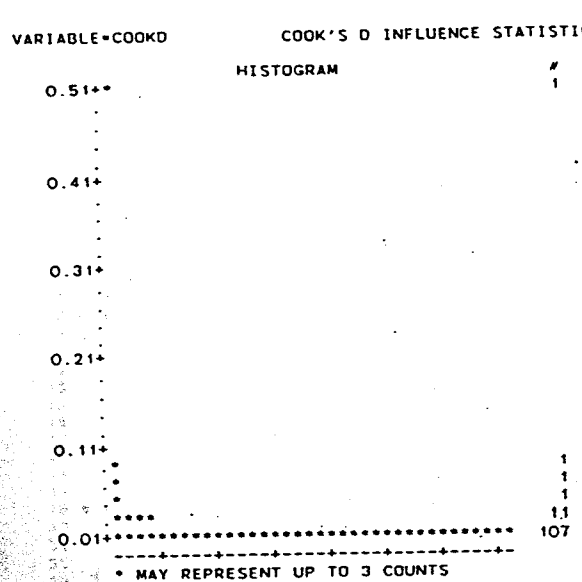
QUANTILES(DEF=4)

100% MAX	0.252143	99%	0.238775
75% Q3	0.0607935	95%	0.109586
50% MED	0.041395	90%	0.0866573
25% Q1	0.0283217	10%	0.0205924
0% MIN	0.0121976	5%	0.0178921
		1%	0.0123143
RANGE	0.239945		
Q3-Q1	0.0324718		
MODE	0.0121976		

EXTREMES

LOWEST	ID	HIGHEST	ID
0.0121976(	23)	0.115984(	172)
0.0127051(	15)	0.129566(	168)
0.0147533(	5)	0.14701(	144)
0.0162169(	178)	0.194024(	157)
0.0173934(	12)	0.252143(	215)

Figure 9. Cook's D.



VARIABLE=COOKD COOK'S D INFLUENCE STATISTIC

MOMENTS

N	122	SUM WGTS	122
MEAN	0.0119143	SUM	1.45355
STD DEV	0.0467903	VARIANCE	0.00218933
SKEWNESS	9.87782	KURTOSIS	103.964
USS	0.282228	CSS	0.26491
CV	392.723	STD MEAN	0.0042362
T:MEAN=0	2.81251	PROB>|T|	0.00573749
SGN RANK	3751.5	PROB>|S|	0.0001
NUM ^= 0	122		

QUANTILES(DEF=4)

100% MAX	0.504815	99%	0.409618
75% Q3	0.00929867	95%	0.0341537
50% MED	0.00351305	90%	0.0229737
25% Q1	.000854951	10%	.000172662
0% MIN	5.709E-08	5%	.000016249
		1%	6.158E-08
RANGE	0.504815		
Q3-Q1	0.00844372		
MODE	5.709E-08		

EXTREMES

LOWEST	ID	HIGHEST	ID
5.709E-08(	139)	0.0387193(	41)
7.660E-08(	12)	0.0448415(	81)
7.434E-07(	169)	0.0680216(	176)
.000001004(	150)	0.0909179(	144)
9.217E-06(	137)	0.504815(	215)

Figure 10. DFFITS.

VARIABLE=DFFITS DIFFERENCE IN FIT INFLUENCE

```
STEM LEAF                            #
  18 0                               1
  17
  16
  15
  14
  13
  12
  11
  10
   9
   8
   7 5                               1
   6 5                               1
   5
   4 16                              2
   3 268                             3
   2 001244444667                   12
   1 11122233333445555677889        23
   0 1122333344455577778888899      25
  -0 9999777666554330000            19
  -1 99998876554320                 14
  -2 9765443320                     10
  -3 65400                           5
  -4 96211                           5
  -5 2                               1
     ----+----+----+----+----+
MULTIPLY STEM.LEAF BY 10**-01
```

VARIABLE=DFFITS DIFFERENCE IN FIT INFLUENCE

MOMENTS

N	122	SUM WGTS	122
MEAN	0.0210004	SUM	2.56205
STD DEV	0.272165	VARIANCE	0.0740736
SKEWNESS	2.32027	KURTOSIS	14.5353
USS	9.01671	CSS	8.96291
CV	1296	STD MEAN	0.0246406
T:MEAN=0	0.852268	PROB>|T|	0.395749
SGN RANK	228.5	PROB>|S|	0.560204
NUM ^= 0	122		

QUANTILES(DEF=4)

100% MAX	1.80412	99%	1.56057
75% Q3	0.14309	95%	0.374837
50% MED	0.0324685	90%	0.240244
25% Q1	-0.147511	10%	-0.286331
0% MIN	-0.522705	5%	-0.400061
		1%	-0.514371
RANGE	2.32683		
Q3-Q1	0.290601		
MODE	-0.522705		

EXTREMES

LOWEST	ID	HIGHEST	ID
-0.522705(	81)	0.413054(	26)
-0.48647(	41)	0.457041(	10)
-0.456123(	88)	0.648031(	176)
-0.419952(	151)	0.745637(	144)
-0.409005(	162)	1.80412(	215)

Figure 11. Covariance ratio.

VARIABLE=COVRATIO COVARIANCE RATIO INFLUENCE

STEM	LEAF	#
130	2	1
128		
126		
124		
122		
120	0	1
118	0	1
116	0	1
114	086	3
112	567935	6
110	223770458999	12
108	034567788802345666667	21
106	012467778899922355558	21
104	22245666822256678899	20
102	5619	4
100	12351347	8
98	2522345	7
96	0279	4
94	0	1
92	2404	4
90	028	3
88	2	1
86	83	2
84		
82		
80		
78	0	1

----+----+----+----+--

MULTIPLY STEM.LEAF BY 10**-02

VARIABLE=COVRATIO COVARIANCE RATIO INFLUENCE

MOMENTS

N	122	SUM WGTS	122
MEAN	1.05385	SUM	128.569
STD DEV	0.0724998	VARIANCE	0.00525622
SKEWNESS	-0.602008	KURTOSIS	2.0435
USS	136.128	CSS	0.636003
CV	6.87955	STD MEAN	0.00656383
T:MEAN=0	160.554	PROB>|T|	0.0001
SGN RANK	3751.5	PROB>|S|	0.0001
NUM ^= 0	122		

QUANTILES(DEF=4)

100% MAX	1.30205	99%	1.28085
75% Q3	1.09568	95%	1.14665
50% MED	1.06704	90%	1.12601
25% Q1	1.01606	10%	0.946154
0% MIN	0.789554	5%	0.913275
		1%	0.807677
RANGE	0.5125		
Q3-Q1	0.0796142		
MODE	0.789554		

EXTREMES

LOWEST	ID	HIGHEST	ID
0.789554(	88)	1.15635(	219)
0.868347(	215)	1.17045(	155)
0.872611(	148)	1.19007(	172)
0.891833(	212)	1.20989(	168)
0.899888(	207)	1.30205(	157)

ULTIPLE LINEAR REGRESSION VIEWPOINTS
VOLUME 16, NUMBER 1, SPRING 1988

Some Parallels Between Predictive Discriminant Analysis and Multiple Regression

Dan Morris, Florida Atlantic University,
and Carl Huberty, University of Georgia

The purpose of this paper is to outline some important similarities in, and differences between, predictive discriminant analysis (DA) and multiple regression (MR). The areas covered are estimates of model accuracy, hypothesis testing, and non-least squares models. Some of the parallels are well known, some are less well known, and some appear to have not yet been considered at all.

It is well known that when (1) only two groups are involved, (2) the two population predictor covariance matrices are assumed equal, and (3) the two prior probabilities of group membership are taken to be equal, the popular "minimum chi-square rule" (Tatsuoka, 1971, p. 218) associated with discriminant analysis (DA) is equivalent to predicting a dichotomous criterion via multiple regression (MR) methods and classifying a subject into the group for which the predicted criterion is nearer the actual.

An especially enlightening examination of this and some other multivariate techniques from the general perspective of MR is provided by Flury and Riedwyl (1985).

However, a precaution about the equivalence of two-group classification and multiple regression with a dichotomous criterion is appropriate. In a two-group situation, there is one linear discriminant function (LDF) and there are two linear classification functions (LCFs); an LDF and an LCF are simply linear composites of the predictors. It is true in a two-group context that the regression weights are proportional to the single set of LDF weights. When a linear regression function (LRF) or an DLF is used for classification purposes a cut-off criterion needs to be determined--with an LRF it is midway between the two values by which the dichotomous criterion is coded, with an LDF it is midway between the LDF means for the two groups. With the use of LCFs, there is not cut-off per se; rather a unit is classified into the group with which is associated the larger LCF score. It turns out that the respective LCF weight differences are proportional to the corresponding LDF and (therefore) the LRF weights.

Input scores for an LRF, and LDF, and and LCF are typically predictor variable measures. [As stated above, any of the three linear composite types may be used for a two-group classification problem.] It turns out that another, still equivalent, approach

to two-group classification may be employed. Here, one uses LDF scores for each unit as input for an LCF; we thus have, in essence, a single predictor score for each unit.

When generalizing from a two-group problem to a k-group problem, it is advisable to forget the LRF and LDF approaches and focus on the LCF approach, with predictor measures as input scores.

Estimates of Model Accuracy

Estimation of the cross-validated accuracy of the prediction model offers similarities and differences between MR and DA methods. In both DA and MR the researcher must decide what type of cross-validated accuracy is of concern. For instance, is interest in simply estimating an accuracy index parameter from the associated statistic, that is, estimating the index of accuracy (R^2 or percent of "hits," respectively) that would obtain in the population from that same index in the sample, or is interest in the accuracy that would obtain on application of sample optimized weights to alternate samples from the same population? The concern in this paper will be with the latter type of accuracy.

As in an estimate of cross-validated R^2 in MR, a judgment of DA "hit-rate" based on the calibration sample is optimistically biased in reference to application to alternate samples. To estimate a cross-validated result in MR, another decision that must be made is whether interest is in relative accuracy, as manifested in the correlation of Y and Y, or in absolute accuracy,

as manifested in the MSE. In either case, several formula estimates are available (see Huberty & Mourad, 1980; Rozeboom, 1978). It is probable that most of the predictive uses of MR in the behavioral sciences, such as in personnel selection, are concerned with relative accuracy.

Unlike in MR, the concern in predictive DA is in classification accuracy; this is implicitly a concern of absolute accuracy. A formula estimate for cross-validated hit-rate in the general k-group case has largely eluded methodologists. However, a useful, although complicated, formula estimate for cross-validated hit-rate in the two-group case was derived by McLachlan (1957). According to that estimator, the hit rate, P_g for group g, where g = 1 or 2 is:

$$P_g = 1 - F(-D/2) - f(-D/2)\ (p - 1)/Dn_g)$$
$$+ D[4(4p - 1) - D^2]/32m) + (p - 1)(p - 2)/4Dn_g^2$$
$$+ (p - 1)[-D^3 + 8d(2p + 1) + 16/D]/(64mn_g)$$
$$+D[3d^6 - 4D^4(24p + 7) + 16d^2(48p^2 - 48p - 53)$$
$$+ 192\ (-8p + 15)]/(12288m^2)\ ,$$

where F is the standard normal distribution function i.e., F(-D/2) is the area to the "left" of -D/2, f is the standard normal density function, D is the Mahalanobis distance, p is the number of predictor variables, n_g is the number of subjects in group g, and $m = n_1 + n_2 - 2$. While the formula looks formidable, with patience it is calculable with hand-held calculator. Moreover, as the last term in the multiplier for f(-D/2) is usually very small,

one may choose to ignore it, making the formula even more tractable. If the researcher with an orientation toward MR notes that $D^2 = R^2N(N-2)/(1-R)^2n_1n_2$, then the McLachlan estimator of cross-validated hit-rate can be obtained from the R^2 resulting from regressing the dichotomous criterion on the predictors.

One slightly "unnerving" aspect of the McLachlan estimator is that it can yield estimated hit-rates that are <u>larger</u> than those that are estimated from the known positively biased process of reclassifying the calibration sample (Morris & Huberty, 1986; 1987). This is unlike the case in MR where the "shrunken" multiple correlation is necessarily less than the value of the multiple correlation derived from the calibration sample. The explanation for this apparent paradox between methods is that estimators of the cross-validated multiple correlation are functions of the corresponding calibration sample multiple correlations, and are therefore <u>guaranteed</u> to yield smaller values than the sample value. In this sense, the McLachlan hit-rate estimator is not parallel to the MR formula estimators. While it is an estimator of cross-validation hit-rate, it is not a function of the calibration sample generated hit-rate; rather, it is a function of the Mahalanobis distance between groups, as well as other variables. That is, it does not simply estimate a parameter from a function of the corresponding statistic as do the MR formula estimators.

An alternate nonparametric approach to estimating cross-validated hit-rate, which has a wide following in the DA literature, is the "leave-one-out" procedure (Huberty, 1984; Huberty & Mourad, 1980; Lachenbruch & Mickey, 1968; Mosteller & Tukey, 1968). In this method, a subject is classified by applying the rule derived from all Ss except the one being classified. This process is repeated "round-robin" for each subject with a count of the overall classification accuracy used to estimate the cross-validated accuracy.

Clearly the same "round-robin" procedure can be used to estimate either relative or absolute accuracy in the use of MR, and has appeared in that context, with perhaps the earliest reference due to Gollob (1967). In a system intended to select optimal MR predictor variable subsets, Allen (1971) coined the procedure "PRESS," and he appears to be the source most often cited in the MR literature.

The apparent computational difficulties due to the inversion of N matrices can be avoided in both MR and DA by using a matrix identity due to Bartlett (1951). This identity is cited an used explicitly in introducing the technique in the DA context by Lachenbruch and Mickey (1968), but was not mentioned by Allen in the first introduction of PRESS (1971) nor in its presentation in a later text (Allen & Cady, 1982, p. 254), although the same identity was implicitly used. Moreover, Allen doesn't cite the DA

literature and the parallel application of the PRESS procedure. It appears that this resampling process was "invented" independently in the MR and DA literatures.

Full vs. Restricted Model Hypothesis Testing

A technique that is well known and widely used by MR researchers is that of hypothesis testing through contrasting full and restricted prediction models. The power of this method, its generality, and its applicability to a very wide arena of theoretical questions in science is no doubt part of the reason for the establishment of the MLRSIG within AERA.

The same types of model contrast "explanatory increment" questions can be asked and seem to be of just as much potential interest when the criterion is classification accuracy. However, we know of no examples of this technique being used in the literature. There seems to be no reason not to test the difference in proportion of correct classifications (hit-rate) between full and restricted models to examine meaningful hypotheses, just as is done using the R^2 in MR. The appropriate test statistic is McNemar's (1947) contrast between correlated proportions. Moreover, as the index, "I," of increase in classification accuracy over chance (see Huberty, 1984, p. 168) is distributed similarly, it becomes apparent that such a test would also be applicable to that statistic.

An example of such a test from a study in which the subsequent high-school dropout of a sample of 76 children was

predicted from data available in fifth grade will now be presented. The six predictor variables were gender, race (two levels), number of elementary schools in which the child had been a student, the number of grades the child had repeated, the family structure (living with both parents, or not), and the child's total number of fifth grade absences. As we have evidence of the relationship between both gender and race and the criterion of high-school drop-out, the hypothesis to be tested concerned the significance of the increment to classification accuracy afforded by adding the four "non-organismic" variables (number of elementary schools, number of grades repeated, family structure, and the total number of fifth grade absences) to the prediction model containing only gender and race. Classifying the calibration sample, the proportion of correct classifications for the total model was 75% and for the model including only gender and race it was 65%. A 2x2 table illustrating the number of hits and misses for both models is:

All Predictors

		MISS	HIT
Gender and Race	HIT	9	39
	MISS	10	18

The test statistic, $z = 1.73$, would typically be considered non-significant ($P = .08$) and therefore offers no evidence that these other variables add to the classification accuracy afforded by just the demographics of race and gender.

While no significance tests were applied, the classification accuracies (again, derived from classifying the calibration sample) obtained with two other subsets of predictor variables are of some interest. The point of interest is that the classification accuracies for these two three predictor variable models (number of elementary schools, number of repeats, and family structure, 79%; number of elementary schools, number of repeats, and number of absences, 79%) were each greater than for the total six variable predictor model. Thus, unlike the multiple correlation coefficient in MR, even with non-cross-validated "internal" estimates of classification hit-rate, accuracy does not necessarily monotonically increase as one adds predictor variables. A different perspective concerning contrasting reduced and full model predictor variable subsets may therefore be necessary for DA applications.

One may argue, however, that the cross-validated estimate of accuracy should be used in any case. An illustration of the impact that using a cross-validated estimator might have is that the leave-one-out estimator for the hit rates involved in the hypothesis tested above were 64% for the full six-variable model, and 49% for the three variable model, with a resulting test statistic of $z = 2.45$, which is, of course, significant at the .02 level.

Non-Least Squares Models

Non-least-squares prediction strategies, particularly ridge regression, have received a great deal of attention in the MR literature (e.g., Darlington, 1978; Morris, 1982, 1982; Pagel & Lunneborg, 1985; Rozeboom, 1979), and some attention in DA (Campbell, 1980; DiPillo, 1976, 1977, 1979). As the benefit to predictive accuracy of such methods is a function of whether the context is relative or absolute accuracy, the results for DA tend to be a subset of those for MR. They appear to be largely parallel to the case of absolute accuracy in the MR case (Morris & Huberty, 1987); enhanced predictive accuracy is available under certain limited circumstances, however, reductions in accuracy are just as likely to occur without an informed decision about when to use the technique. Ridge methods are far from the panacea that they have been purported to be for either the MR or DA case. A suggested method for choosing between alternate predictor weighting algorithms, including ridge and least squares, has been presented for the DA case by Morris and Huberty (1987), and for the MR case by Morris (1986). Computer programs for both analysis types are available at no charge from:

John D. Morris
Institute for Research and Development in Teacher Education
College of Education
Florida Atlantic University
Boca Raton, FL 33431

References

Allen, D. A. (1971). The prediction sum of squares as a criterion for selecting predictor variables (Tech. Rep. No. 23). University of Kentucky, Department of Statistics.

Allen, D. A., & Cady, F. B. (1982). Analyzing experimental data by regression. Belmont, CA: Wadsworth.

Bartlett, M. S. (1951). An inverse matrix adjustment arising in discriminant analysis. Annals of Mathematical Statistics, 22, 107.

Campbell, N. A. (1980). Shrunken estimates in discriminant and canonical variate analysis. Journal of the Royal Statistical Society, 29, 5-14.

Darlington, R. B. (1978). Reduced variance regression. Psychological Bulletin, 85, 1238-1255.

DiPillo, P. J. (1976). The application of bias to discriminant analysis. Communications in Statistics,A5, 843-854.

DiPillo, P. J. (1977). Further applications of bias to discriminant analysis. Communications in Statistics, A6, 933-943.

DiPillo, P. J. (1979). Biased discriminant analysis: Evaluation of the optimum probability of misclassification. Communications in Statistics, A8, 1447-1457.

Flury, B. & Riedwyl, H. (1985). T^2 tests, the linear two-group discriminant function and their computation by linear regression. The American Statistician, 39, 20-25.

Gollob, H. F. (1967, September). Cross-validation using samples of size one. Paper presented a the meeting of the American Psychological Association, Washington, D.C.

Huberty, C. J. (1984). Issues in the use and interpretation of discriminant analysis. Psychological Bulletin, 95, 156-171.

Huberty, C. J., & Mourad, S. A. (1980). Estimation in multiple correlation/prediction. Educational and Psychological Measurement, 40, 101-112.

Lachenbruch, P. A., & Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. Technometrics, 10, 1-11.

McLachlan, G. J. (1975). Confidence intervals for the conditional probabilities of misallocation in discriminant analysis. Biometrics, 31, 161-167.

McNemar, Q. (1947). Note on the sampling error of the differences between correlated proportions or percentages. Psychometrika, 12, 153-157.

Morris, J. D. (1982). Ridge regression and some alternate weighting techniques: A comment on Darlington. Psychological Bulletin, 91, 203-210.

Morris, J. D. (1983). Stepwise regression: A computational clarification. Psychological Bulletin, 91, 363-366.

Morris, J. D. (1986). Microcomputer selection of a predictor weighting algorithm. Multiple Linear Regression Viewpoints, 15, 53-68.

Morris, J. D., & Huberty, C. J. (1986, April). A comparison of three methods of classification hit-rate estimation. Paper presented at the meeting of the American Educational Research Association, San Francisco.

Morris, J. D., & Huberty, C. J. (1987). Selecting a two-group classification weighting algorithm. Multivariate Behavioral Research (in press).

Mosteller, F., & Tukey, J. W. (1968). Data analysis, including statistics. In G. Lindzey, & E. Aronson (Eds.), Handbook of social psychology: Vol. 2. Reading, MA: Addison-Wesley.

Pagel, M. D. & Lunneborg, C. E. (1985). Empirical evaluation of ridge regression. Psychological Bulletin, 97, 342-355.

Rozeboom, W. W. (1978). Estimation of cross-validated multiple correlation: A clarification. Psychological Bulletin, 85, 1348-1351.

Rozeboom, W. W. (1979). Ridge regression: Bonanza or beguilement? Psychological Bulletin, 86, 242-249.

Tatsuoka, M. M. (1971). Multivariate analysis. New York: John Wiley.

JNEAR REGRESSION VIEWPOINTS
E 16, NUMBER 1, SPRING 1988

A Ten Year Study of Salary Differential by Sex Through a Regression Methodology

John D. Williams
University of North Dakota

Jole A. Williams
Northwestern Minnesota Mental Health Center

Stephen J. Roman
New Market Iowa Community School

Abstract

:n year study of salary differential by sex was completed,
ıg a multiple regression methodology, with rank, discipline,
-ee, years in department, years in current rank and sex as
lictors, focusing on the change in the value of the sex
ıable. The sex variable evidenced lower salaries for women
ı controlling for the other variables throughout the study
ıod for both proposed and actual salaries from $341 in 1978-79
'posed salary) to $1675 for 1981-82 (actual salary) to $504
1986-87 (proposed salary). This apparent drop in
:rimination by sex in salary at each rank was accompanied by
-nging differences in pay. The change is in the direction of
rket adjustments," i.e., paying lower salaries to those in
:iplines with higher proportions of women.

In a study of 1977-78 faculty salaries at the University of North Dakota (UND), using a regression approach, Martin and Williams (1978) found that women were underpaid $361 (in terms of the regression coefficient), on the average, taking into account a large number of variables. In that the ensuing years were supposed to be a time for eroding away sex discrimination, it was quite surprising that Anderson (1986) showed that the discrepancy in 1985-86 actual salaries may have become as large as $4619 at the same institution.

Subsequently, all UND faculty salary data for all years from 1977-78 to 1986-87 have been secured; these data are from public access files and thus contain no confidential information. The actual data are for nine complete years wherein the previous salary is given and the proposed salary for the following year is listed. Since it would be highly unusual for obvious, direct sex discrimination to take place without detection, the possibility of a secondary impact of discrimination is examined. If, for any given year, sex differences increase from proposed to actual salaries, it is important to document this process. The advantage of a long term data set (actual salaries from 1977-78 to 1985-86 and proposed salaries from 1978-79 to 1986-87) is that changes in the composition of the faculty can be monitored as well. One possibility is that arrivals and departures from the faculty may have devastating effects on sex discrimination measures. Other possibilities could be examined as well. The particulars of either the data set and/or the variables used

could have a major impact on outcomes. One cannot count out *a priori* another period of sex inequity in salary structure, though such inequity would of necessity be more subtle. First however, the particulars of the data should be addressed.

Obstacles to Salary Discrimination Research

Obtaining the data sets for analysis was a major obstacle in this study. Originally, Anderson's (1986) data was to be reanalyzed. She was agreeable to this, and the UND Vice President for Academic Affairs provided strong encouragement. However, because the Anderson data set was generated under the auspices of the university's Office of Institutional Research, the opinion of the university legal counsel was that her data should not be made available to outside researchers (despite the first author's being at that institution and having served on Anderson's doctoral committee!). Thus, the investigation was possible only through the use of public documents; all UND salary data (since at least 1926) are available at the university library. These data were secured for the academic years 1978-87 (the years following the studies by Martin and Williams, 1978, 1979). The quality of these salary data was shocking to these researchers. For some years several pages were missing, though these omissions were to some degree rectifiable. More important were obvious mistakes--mistakes that became apparent only as the data set was constructed. In several cases (perhaps 2-5%)

subsequent salary data suggested that earlier salary data were incorrect. For example, a person's salary history might read:

	Proposed Salary	Last Year Salary	Increase
1978-79	22000	21000	1000
1979-80	11500	22000	1000
1980-81	24000	23000	1000

This kind of "mistake" occurred when someone was on leave; the last year's salary for 1979-80 was actually a hypothetical salary, but was entered into salary history. The "mistake" shown here was a logical one; less logical or actual errors (perhaps due to the faculty member's negotiating a higher salary) also occurred, but became known only in the next year's budget. Thus, the proposed salary figures include persons who negotiated higher salaries than were budgeted, and also include those who resigned and didn't actually receive a salary. New faculty members usually don't show up at all in the proposed salary figures for their first year. In that sense, actual salary data is known (insofar as the public documents are concerned) only a year later.

Choice of Variables

The choice of variables in salary equity studies is particularly important; some variables such as academic rank have been viewed as biased themselves (Scott, 1977). She preferred a smaller set of variables that, from a practical point of view, tend to show more discrimination. The choice of variables is somewhat (if not wholly) political--and the choice of variables

surely influences the interpretation. For example, using a different selection of variables (including Scott's) Anderson (1986) found coefficients for sex favoring males from $1883 to $4619 for the 1985-86 actual salaries.

The original point of view for the present study was to incorporate variables similar to those used in Martin and Williams (1978), but deleting variables that had "suspect" outcomes. By "suspect" outcome is meant that the direction of the outcome for that variable is counter-intuitive; for example, that study found that serving on committees had a negative partial effect on salaries. Though different interpretations are possible, these sorts of variables may also incorporate sex inequity differences--in fact, women did have a higher tendency to serve on committees (Williams, 1978)--and including these variables helped cover over sex differences. Hence, committee membership was not included in the present analysis. Also, teaching in a graduate program had a <u>negative</u> impact on salary (Martin & Williams, 1978), an outcome that was counter-intuitive as well as counter-productive from a university's point of view. Publication information and teacher rating information are no longer available due to privacy considerations, and teacher rating information is no longer uniform as well. The variables finally selected are found in Table 1.

TABLE 1

Variables Included in the Regression Analysis Regarding Equity Adjustments to Salaries at the University of North Dakota

---

Degree Held
- Doctorate
- Bachelors/Professional
- (Masters, zero coded)

Years in Department

Sex
- Male = 1
- Female = 0

Rank
- Professor
- Associate Professor
- Assistant Professor
- (Instructor, zero coded)

Years in Current Rank
- Years in rank Professor
- Years in rank Associate Professor
- Years in rank Assistant Professor
- Years in rank Instructor

Discipline
(HEGIS Taxonomy)
- Biology
- Business
- Communication
- Computer Science
- Education
- Engineering
- Fine Arts
- Health Professions
- Languages and Humanities
- Library Science
- Mathematics
- Physical Sciences and Aviation
- Psychology
- Political Science
- Home Economics
- Law
- (Social Sciences, zero coded)

For the years 1978-79 through 1986-87 both proposed and actual previous salaries were used as criteria, using year appropriate data. In the case of promotion the rank would be one rank lower for proposed salary but is correct for actual salary. Table 2 gives results for the regression coefficient, F value, and biserial correlation for sex (with salary) along with R and the proportion of women for each year, in both the proposed and actual budget..

TABLE 2

Regression Coefficients, F Values, Biserial Correlations, R and Proportion of Women with Proposed and Actual Salaries

Proposed					Actual				
Reg. Coeff.	F	Point Bisl. Corr.	R	Prop. Women	Reg. Coeff.	F	Point Bisl. Corr.	R	Prop. Women
361.03	1.57	.268	.913	.145	537.55	2.71	.267	.870	.158
341.07	.80	.275	.849	.163	731.11	4.80	.286	.886	.156
689.32	2.62	.338	.854	.185	530.45	2.09	.313	.894	.189
572.27	1.56	.273	.840	.175	1250.23	6.27	.276	.842	.159
1351.95	6.28	.317	.838	.183	1674.58	10.35	.329	.850	.179
1542.32	7.96	.341	.848	.186	1007.74	3.91	.334	.861	.185
1293.57	5.56	.340	.836	.185	1362.68	5.30	.320	.834	.174
1110.44	4.19	.328	.841	.188	739.51	1.42	.286	.865	.190
849.79	2.23	.368	.861	.195	747.11	1.60	.375	.862	.200
504.12	.74	.392	.861	.211					

rom Martin and Williams (1978)

Table 2 yields some interesting outcomes. The actual amount of inequity by sex often exceeded the projected inequity by sex; also, the inequity by sex appeared to peak in the early 1980's (in terms of the regression coefficient for sex), and has appeared to drop to only about $140 higher than projected for

1977-78. However, the point biserial correlation has gone up considerably, indicating that real differences in mean salaries have sharply increased. It is useful to address salary differences by rank as shown in Table 3. The number of persons at each rank by sex are shown in Table 4.

TABLE 3

Mean Salaries by Sex and Rank for Projected and Actual Salaries, 1977-1987

	Proposed					Actual				
	Inst.	AsstP	AscP	Prof	Total	Inst.	AsstP	AscP	Prof	Tot
977-78*										
F		14606	17283	21389	16954	12883	15001	17143	21866	16
M		15524	18151	22164	19040	13085	15518	18263	22277	19
1978-79										
F	13395	15292	18002	23195	17008	13330	15180	18040	22786	17
M	14200	16370	19259	23335	20045	14158	16189	19275	23567	20
1979-80										
F	12813	15881	19422	24306	17286	13124	16109	18662	24393	18
M	15027	17207	20594	24951	21461	14400	16964	20403	25510	21
1980-81										
F	14648	16947	20148	25957	19420	16158	18560	22014	26219	21
M	15809	18512	21921	26868	23001	16683	20565	23316	28646	25
1981-82										
F	18112	20790	24318	29064	22757	16686	20271	24084	28141	22
M	21860	22438	26243	31896	27581	21864	22727	26058	31608	27
1982-83										
F	17997	20535	24901	27901	22996	17997	20398	24923	28922	23
M	21889	23243	27140	33153	28556	22172	23358	26710	32813	28
1983-84										
F	19272	20098	25229	29325	23335	19194	20598	24490	27727	23
M	21030	24190	27142	33000	28814	20294	23050	26650	32451	28
1984-85										
F	18393	21051	24952	27945	23275	17658	24255	24663	27540	23
M	21013	23245	26850	32568	28550	22943	23115	26341	32806	28
1985-86										
F	21556	22887	28083	31934	25997	22603	23127	26891	32116	26
M	23814	26848	29960	36743	32410	24380	26715	29677	36400	33
1986-87										
F	21922	24147	28084	34132	26819					
M	25202	27882	31134	38046	33788					

*Taken from Martin and Williams (1978)

TABLE 4

Number of Persons at Each Rank by Sex

Proposed					Actual				
Inst	AsstP	AscP	Prof	Total	Inst	AsstP	AscP	Prof	Total
*									
	14	20	6	40	9	15	24	8	56
	64	107	98	269	2	57	126	114	299
10	18	24	7	59	5	20	21	8	54
8	59	125	110	302	7	47	124	115	293
)									
13	27	22	7	69	9	22	25	11	67
6	59	125	114	304	3	45	125	115	288
8	21	25	11	65	1	22	22	9	54
5	61	125	115	306	5	43	117	121	286
8	29	24	9	70	5	21	29	8	63
11	57	124	121	313	8	50	111	121	290
3									
6	27	30	8	71	6	24	30	10	70
9	65	113	123	310	9	52	115	133	309
4									
8	26	30	10	74	7	17	28	11	63
9	62	122	134	327	5	46	111	138	300
5									
10	23	29	11	73	11	18	29	10	68
7	56	114	138	317	3	40	108	139	290
86									
11	24	30	10	75	6	19	35	9	69
4	54	111	140	309	1	39	101	138	277
87									
8	27	36	9	80					
3	48	108	140	299					

a from Martin and Williams (1978)

While there are some difficulties due to probable missing information (that is, information gone from the public documents), it seems clear that if women were "underranked" for the earlier years in the study, they are far more so for the most recent available year. Using projected data for 1977-78, 6 of 40

women or 15% are professors, as compared to 98 of 269 men or 36.43%. For 1986-87, 9 of 80 women or 11.24% are professors, as compared to 140 of 299 men or 46.82%. For those who might have hoped that these sorts of differences would dissipate during a period of supposed redressing of inequity, these outcomes confirm the dashing of those hopes. Further, salary differences by sex within ranks favored men by approximately $800 at each rank for projected 1977-78, compared to 1986-87 projected data where differences are in the range of $3000-$4000 at each rank, while salaries increased by only about $10000 for women and $13700 for men during the interim. This latter finding is particularly anomalous, considering the changes in the coefficient for sex (gender) shown in Table 2; it can be recalled that discrimination costs to women appeared to have reduced almost back to 1977-78 levels, after going much higher in the early 1980's.

Yet a different interpretation would be obtained from viewing the two-way ANOVA outcomes, suggesting it would be worthwhile to inspect changes in other variables in the regression analysis. Rather than attempt to give the entirety of the sets of regression analyses shown in Table 2, three analyses investigated are discussed. Table 5 records these analyses: the proposed salaries for 1978-79 and 1986-87 and the actual salaries from 1981-82. These years were chosen because they show the minimum effect for sex (proposed, 1978-79), maximum effect for sex (actual, 1981-82) and most recent outcome (proposed, 1986-87).

TABLE 5

Regression Analyses for Three Selected Years
(Proposed 1978-79, Actual 1981-82 and Proposed 1986-87)

e	Proposed 1978-79 Reg. Coeff.	F	Actual 1981-82 Reg. Coeff.	F	Proposed 1986-87 Reg. Coeff.	F
Held	802.08	6.18	1126.71	5.95	522.04	4.72
ate	1377.13	2.11	1680.21	1.51	3001.00	1.16
ors/Prof.						
n Dept.	-93.91	8.17	-106.51	5.93	-111.27	5.60
le=1, Female=0)	341.07	.80	1674.50	10.35	504.12	.74
sor	9999.24	134.02	8147.24	24.44	15884.11	64.54
iate Professor	5642.34	50.87	2883.28	3.27	9725.70	26.68
tant Professor	2188.97	7.62	241.56	.02	6045.03	10.28
in Current Rank						
ssor	197.58	17.17	374.05	32.63	433.98	39.67
iate Professor	159.98	7.93	332.66	19.53	313.60	15.54
tant Professor	266.46	12.73	277.91	5.70	192.64	2.54
uctor	157.60	.88	-949.04	2.32	874.97	1.51
line (HKGIS)	-869.94	1.42	38.13	.00	-392.59	.12
gy	1603.15	8.41	4059.71	21.31	6312.41	50.86
ness	533.33	.20	-633.56	.16		
nications	2410.42	3.77	3643.84	5.20	10927.30	38.99
ter Science	533.51	1.12	2469.74	9.06	1107.34	1.85
tion	392.07	.40	4773.05	21.36	6810.45	45.09
neering	-1220.63	3.82	1162.12	1.41	-437.15	.20
Arts	-1794.86	3.26	3401.56	5.37	1417.81	1.10
th Prof.	-761.19	2.11	571.01	.45	-48.01	.00
. and Hum.	1850.55	1.80	3441.30	3.01	5352.24	3.37
ary Science	392.85	.28	1360.86	1.36	104.04	.01
ematics	-47.98	.01	3011.09	11.84	4032.67	21.87
Sci. and Avtn.	760.22	1.04	735.67	.45	533.17	.18
hology	261.69	.09	2007.16	3.37	2486.40	2.74
tional Science	866.17	.56	2078.12	1.59	-176.89	.01
Economics	8205.57	97.43	16325.76	150.00	15109.78	153.88

Table 5 is clearly complex; simplistic interpretations would violate that complexity. Some interpretations, however, can be

made. The importance of discipline (HEGIS category) in salary becomes quite clear. Recent major gainers are computer science (up almost $7300, compared to social sciences, since 1981-82), business (with large comparative increases for the last two reported years), engineering (up more than $4000 from 1978-79 to 1981-82, and an additional $2000 for 1986-87), library science (up $1600 for 1981-82, and an additional $1900 for 1986-87) and political science (up $1750 for 1981-82 and an additional $500 for 1986-87). What is not apparent in the data is that these disciplines have higher proportions of males than do those whose climbs (vis-a-vis the social sciences which have a higher proportion of females) are not as marked. In the year 1985-86 in particular, an internal study allowed large individual deviations in salary based on "market" considerations. Those market considerations were achieved by comparing salaries in various categories to a regional average. Departments were compared to the mean of similar departments within that regional study with the intent of raising salaries to near the regional averages. This study, though of considerable importance in determining salaries, was not generally disseminated; within a college, results for affected departments might be known, but the overall texture for the university was not known. One case in point was the "statistics" department. Since the University of North Dakota has the only such grouping in the region, this department was exactly at the norm and thus needed no adjustment. The fallibility of the other data can only be conjectured---the data

were never made available for analysis. Nevertheless, on the basis of these data, one department in particular was the recipient of a windfall--political science (in the college of business). This department's salary changes from 1984-85 to 1985-86 included one individual going from $25975 to $37000 (a $11025 or 42.44% increase), while another went from $26450 to $37200 (a $10750 or 40.64% increase). The remaining five faculty received increases of $2120 to $6390 (8.37% to 20.52%); the mean increase within the university overall was 11.4%. These changes were a major source of internal departmental disagreement that eventually saw one faculty member moving to another department in the university, and newspaper articles on these increases in both the local and student newspapers. Last in all of this is that these so-called "market adjustments" helped validate even larger differences in pay between men and women, though additional losers were both men and women in the disciplines that had larger proportions of women than the university average. Ronda to the redressing of inequity had been circumvented in two ways--the market adjustments favored male dominated departments, and those faculty in departments receiving less favorable treatment could blame their treatment at least partially on their higher proportion of women.

Redressing inequity due to any cause (including gender based inequity) would seem not to be part of the immediate future at the University of North Dakota. Preliminary budgets for the 1987-89 biennium include pay increases totaling 2% for the entire

period, with that raise to come in 1988-89. Even this modest increase might still be eliminated; even worse, cutbacks in faculty and/or salaries are possible due to the financial woes of the state, which is largely dependent on two industries, agriculture and fossil fuels, both suffering in the present financial arena.

Comments on Choosing Variables Investigating Gender Bias in Salary

Scott (1977) suggested using a small number of variables, not including rank, in addressing possible sex bias. Her choice of not including rank was based upon rank's being a "contaminated" variable, that is, rank itself is accorded in a gender non-neutral way. The present study has used rank as a variable; perhaps to some degree, even to a large degree, Scott is correct in her assertion that rank is gender inequitable- surely the data on rank by sex in Table 4 would be more supportive than contradictive of her view. However, rank does have credence within a university setting, and its exclusion from consideration might render studies less acceptable in terms of redressing inequity.

The process of choosing variables is a political act; outcomes will be at least partially determined by the inclusion or exclusion of given variables. Generally speaking, the inclusion of more variables will tend to reduce the impact of a given variable (such as sex). Though not shown here, each

analysis shown in Table 2 was duplicated for each rank using a second degree term incorporating a quadratic regression for years in rank. Initially it was felt that a quadratic trend might possibly be occurring at the associate professor level and lower, the thinking being that those who failed to be promoted to the next rank might experience negative effects in regard to their salaries. While some second degree trends did exist for the data, almost without exception there were corresponding drops in the sizes of the coefficients for sex; one interpretation of this outcome is that for the lower ranks, women stay in a rank longer than men (this could be another result of possible discrimination), whereas at the professor rank men are in rank longer than women (obviously, if they get there sooner, they'll be there longer). Addressing inequity, whether due to gender related reasons or to some other cause, is a subtle process; different persons (whether researchers or not) will not often agree on the meaning of inequity or discrimination. The limits of regression as a technique for determining inequity should be apparent. If the researcher/activist is diligent in the choice of variables, he/she will be able to better show "what is." However, regression tells us nothing about "what should be." Too often, we misinterpret "what is" for "what should be." The former (what is) can be, to some degree, determined, depending on the ingenuity of the researcher in choosing variables. The latter (what should be) is fraught with personal meanings likely to differ for different individuals although consensus may

TIPLE LINEAR REGRESSION VIEWPOINTS
/OLUME 16, NUMBER 1, SPRING 1988

Multivariate Analysis Versus Multiple Univariate Analyses

Carl J. Huberty
University of Georgia

John D. Morris
Florida Atlantic University

Abstract

The argument for preceding multiple ANOVAs with a MANOVA to control for Type I error is challenged. Several situations are discussed in which multiple ANOVAs might be conducted. Three reasons for considering a multivariate analysis are discussed: to identify outcome variable system constructs, to select variable subsets, and to determine variable relative worth.

Paper presented at the annual meeting of the American Educational Research Association, Washington, April 1987.

Multivariate Analysis Versus Multiple Univariate Analyses

The analyses discussed in this paper are those used in research situations where analysis of variance techniques are called for. These analyses are used to study the effects of "treatment" variables on outcome variables (in ex post facto as well as experimental studies). With a single outcome variable we speak of univariate analysis of variance (ANOVA); with multiple outcome variables it is multivariate analysis of variance (MANOVA).

With multiple outcome variables, the typical analysis approach used in the group-comparison context, at least in the behavioral sciences, is to either: (1) conduct multiple ANOVAs, or (2) conduct a MANOVA followed by multiple ANOVAs. The thesis of the current author is that the latter approach is seldom appropriate, and the former approach is appropriate only in some special situations. The purpose of this paper is to provide a rationale for the stated thesis, and to present an argument for a truly multivariate analysis, when appropriate.

Type I Error Protection

An argument often given for conducting a MANOVA, as a preliminary to multiple ANOVAs, is to "control for Type I error probability" (see, e.g., Leary & Altmaier, 1980). The rationale typically given is that if the MANOVA yields

significance, then one has a "license" to carry out the multiple ANOVAs, with the data interpretation being based on the results of the ANOVAs. It may be intuitively appealing to conclude that one would incorrectly reject a null ANOVA hypothesis less frequently if the null MANOVA hypothesis is initially rejected than if the latter were not rejected. This is the notion of a "protected (ANOVA) F test" (Bock, 1975, p. 422), an extension of Fisher's protected t test idea as applied to the study of contrasts in an ANOVA context.

If a researcher has a legitimate reason for testing univariate hypotheses, then he/she might consider either of two testing procedures. One is a simultaneous test procedure (STP) advocated by Bird and Hadzi-Pavlovic (1983) and programmed by O'Grady (1986). For the STP, as applied to the current MANOVA-ANOVAs context, the referent distribution for the ANOVA F values would be based on the MANOVA test statistic used. Bird and Hadzi-Pavlovic (1983, p. 168), however, point out that for the current context, the overall MANOVA test is not really a necessary prerequisite to simultaneous ANOVAs. Ryan (1980) makes the same point for the ANOVA-contrasts context. These two contexts may be combined to a MANOVA-ANOVAs-contrasts context in which it would be reasonable to go directly to the study of univariate group contrasts, if univariate hypotheses are the main concern (see next section.)

A second procedure for testing univariate hypotheses is to employ the usual univariate test statistics with a Bonferroni adjustment to the overall Type I error probability. How

"overall" is defined is somewhat arbitrary. It could mean the probability of committing a Type I error across all tests conducted on the given data set. Or, it could mean the Type I error probability associated with an individual outcome variable when univariate questions are being studied. Whatever the choice (which can be a personal one, and one that is numerically nonconventional!), some error-splitting seems very reasonable. Assuming that Type I error probability for each in a set of m tests is constant, the alpha level for a given test may be determined by using either of two approaches. One approach is to use the additive Bonferroni inequality: for m tests, the alpha level for each test is given by the overall alpha level divided by m. A second approach is to use a multiplicative inequality: for m tests, the alpha level for each test is found by taking one minus the mth root of the complement of the overall alpha level. [See Games (1977).] The per-test alphas--constant across the m tests--found using the two approaches are, for most practical purposes, the same. Therefore, the simpler of the two approaches, namely the first one, is recommended when multiple tests are conducted.

In nearly all instances, outcome variables are interrelated. Thus, the ANOVA F tests are not independent; furthermore, contrast tests for individual outcome variables may not be independent. This lack of independence does not, however, present difficulties in determining the per-test alpha level to use. That this is the case may be seen by the following double inequality:

$$\text{overall alpha} < 1 - (1\text{-test alpha})^m < m \cdot \text{test alpha}.$$

It turns out that when conducting m tests, each at a constant alpha level, a considerably larger overall alpha level results. For example, 6 tests, each conducted using an alpha level of .05, yield an overall alpha level of .30 using the additive inequality, and about .26 using the multiplicative inequality (the middle of the double inequality above). The above double inequality ignores the extent of the outcome variable intercorrelations. If r is the constant correlation between all pairs of outcome variables, then the overall alpha level is approximately (Bird, 1975, p. 346)

$$1 - r^2(1 - \text{test alpha}) - (1 - r^2)(1 - \text{test alpha})^m.$$

Again, for 6 tests, each at an alpha level of .05, and a constant bi-variable correlation of .30, the overall alpha level is about .25.

While adjusting the individual test alphas in conducting multiple tests addresses the Type I error protection problem, a potential related problem emerges. For m tests and a test alpha equal to (1/m)th of the overall alpha, the statistical power of the multiple tests may be a concern if m is "large." One way of obtaining reasonable power values is to use an adequate sample size. Thus, in designing studies that incorporate multiple outcome variables, the sample size-to-variable ratio is an important consideration. The use of a liberal overall alpha is recommended; something like .20, or even higher in some situations. This whole issue becomes much more involved when group contrasts are studied for each outcome variable. Sound planning, good judgment, and reasonableness are clearly called for.

Merely conducting a MANOVA, obtaining significance at som level, and then conducting multiple ANOVAs, each at a conventional significance level, is hardly "controlling for Type I error probability." The notion that one completely controls for Type I error probability by first conducting an overall MANOVA or ANOVA is open to question (Bird & Hadzi-Pavlovic, 1983; Bray & Maxwell, 1982, p. 343; Ryan, 1980 since the alpha value for each follow-up test would be less than or equal to the alpha employed for the overall test only when the overall null hypothesis is true. (See, also, Wilkinson, 1975.) This notion does not have convincing empirical support in at least a MANOVA-ANOVAs context--the Hummel and Sligo (1971) and Hummel and Johnston (1986) studies notwithstanding.

When Multiple Univariate Analyses?

One situation in which multiple univariate analyses might be appropriate is as a means of screening outcome variables prior to a MANOVA. It behooves the researcher to screen out non-functional variables at the outset for various reasons; to enhance parsimony, to enhance estimated predictive accuracy, t abate collinearity, and so forth. Suppose a researcher has 15 sets of unimodally distributed outcome measures. A reasonable first analysis step would be to conduct 15 ANOVAs. A rule-of-thumb that seems appropriate is to delete any variable

m further analysis if the associated ANOVA F-value is less
n 1.00. In a two-factor design this rule would pertain to
"all-effects" test--the test of the equality of all design
l population means. A rationale for this rule is that such
F-value implies that the variable is contributing nothing
"noise" to the analysis. [An F-value of unity is
ivalent to an eta-squared value of $df_h / (df_h + df_e)$.]

A second situation that would call for the use of multiple
variate analyses is when the outcome variables are
nceptually independent" (Biskin, 1980). [This is the
ithesis of a situation involving a variable system, a notion
cussed in the next section.] In such a situation one would
interested in how a treatment variable affects each of the
come variables. Here, there would be no interest in seeking
linear combination of the outcome variables; an underlying
nstruct" is of no concern. In particular, an underlying
struct would perhaps be of little interest when each outcome
iable is from a different domain. Dossey (1976), for
mple, studied the effects of three treatment variables
aching Strategy, Exemplification, Student Ability) on four
come variables: Algebra Disjunctive Concept Attainment,
metric Disjunctive Concept Attainment, Exclusive Disjunctive
cept Attainment, and Inclusive Disjunctive Concept
ainment. Considering these outcome variables as
ceptually independent, four three-way ANOVAs were conducted.

The third situation in which multiple univariate analyses
ht be appropriate is when the research being conducted is

exploratory in nature. Such situations would exist when "new" treatment and outcome variables are being studied, and the effects of the former on the latter are being investigated so as to reach some tentative, nonconfirmatory conclusions. This approach might be of greater interest in status studies, as opposed to true experimental studies.

In the two latter situations it might be argued (via the "protected-test" argument) that the multiple tests on the individual outcome variables should be preceded by a MANOVA. As mentioned above, however, this is not necessary. If tests on individual outcome variables are the tests of basic interest, then going directly to the univariate analyses would seem reasonable. One can employ a simultaneous test procedure by referring to a MANOVA test statistic (with or without a Bonferroni adjustment), or multiple univariate analyses by referring to a univariate test statistic with a Bonferroni adjustment.

A fourth situation in which multiple univariate analyses may be appropriate is when some or all of the outcome variable under current study have been previously studied in univariate contexts. In this case separate univariate analysis results can be obtained for comparison purposes, in addition to a multivariate analysis if the latter is appropriate and desirable.

A fifth situation calling for multiple univariate analyse is where a researcher characteristic is considered. The researcher characteristic is a lack of understanding of, and/o

appreciation for, multivariate methods. A lack of training and experience in multivariate methods may very well account for the lack of understanding/appreciation. Attempting to use analysis procedures with inadequate understanding is futile indeed. One possible solution to the lack-of-understanding problem (for non-dissertation research) is to contact a knowledgeable methodologist, stimulate his/her interest in the topic being investigated, offer him/her co-authorship, and complete the collaboration.

Finally, there is an evaluation design situation in which multiple univariate analyses might be conducted. This is when some evidence is needed to show that two or more groups of units are "equivalent" with respect to a number of descriptors. These analyses might be considered in an in situ design for the purpose of a comparative evaluation of a project. In this situation evidence of comparability may be obtained via multiple informal ("eye-ball") tests, or formal statistical tests.

Some six situations are presented that would seem appropriate for multiple univariate analyses. Multiple univariate analyses might be conducted: (1) to screen outcome variables prior to a multivariate analysis; (2) to study the effects of some treatment variable(s) on conceptually independent outcome variables; (3) to explore new treatment-outcome variable bivariate relationships; (4) to re-examine bivariate relationships within a multivariate context; (5) when a researcher is multivariately naive; and (6) to select a "comparison" group in designing a study.

Why a Multivariate Analysis?

Of course, the analysis strategy employed by a researcher is dependent, among other things, upon the questions he/she ha of the data on hand. And these questions are, or at least should be, derived from beliefs or theories of the researcher. With questions in mind, it is assumed that the researcher has judiciously chosen a collection of outcome variables that are relevant to his/her investigation. The interrelationship of these variables is an important consideration in deciding upon an analysis strategy. More specifically, does the collection of variables constitute, in some substantive sense, a system? Or, perhaps, are there subcollections that may constitute multiple systems?

A "system" of outcome variables may be loosely defined as a collection of interrelated variables that, at least potentially, determines one or more meaningful underlying variates or constructs. In a system one has several outcome variables which represent a small number of constructs--typically one or two. For example, Watterson et al. (1980) studied a system of five outcome measures on attitudes (based on interview and questionnaire data) that lead to two meaningful variates, political attitude and freedom of expression; Hackman and Taber (1979) studied a system of 21 outcome measures on student performance (based on interview data) that determined two meaningful variates, academic performance and personal growth.

A goal of a multivariate analysis is to identify and interpret the underlying construct(s). For such potential constructs to be meaningful, the judicious choice of outcome variables to study is necessary; the conceptual relationships among the variables must be considered in light of some overriding "theory." A multivariate analysis should enable the researcher to "get a handle" on some characteristics of his/her theory: What are the "emerging variables"?

These emerging variables are identified by considering some linear composites of the outcome variables, called canonical variates or linear discriminant functions (LDFs). Correlations--sometimes called structure correlations--between each outcome variable and each LDF are found. Just as in factor analysis, the absolute values of these correlations, or "loadings," are used in the identification process: those variables with high loadings are "tied together" to arrive at a label for each construct.[2] [See, however, Harris (1985, p. 319), for an opposing point of view regarding such a use of loadings.]

Sometimes a researcher is interested in studying multiple systems, or subsystems, of variables. Those subsystems may be studied for comparative purposes (see, e.g., Lunneborg & Lunneborg, 1977), or simply because different (conceptually independent?) constructs--based on different variable domains--are present (see, e.g., Elkins & Sultmann, 1981). In this case, a separate multivariate analysis for each subsystem would be conducted.

A primary reason, then, for conducting a multivariate analysis is to identify the variates or constructs that underlie the collection of outcome variables chosen for analysis. By doing so, one analyzes the collection as a system, taking into consideration the intercorrelations of the variables. This approach enables a researcher to seek answers to more general (more interesting?), complex questions; questions that reflect the real world of behavioral (or any other) science.[3] [See Dempster, (1971) for more on data structure.]

There are two other potential reasons for conducting a multivariate analysis. Either of these reasons is considered when the intercorrelations of the outcome variables are to be kept in mind. One potential reason is to determine if fewer variables than the total number initially chosen can adequately define a meaningful system. This is the so-called variable selection problem, and is discussed in some detail by Huberty (1986). This problem might be considered so as to seek a parsimonous interpretation of a system. It should be noted that this is not an imposed parsimony--as one might get with multiple univariate analyses--but a parsimony taking into consideration the intercorrelations of the outcome variables.

Another potential reason for conducting a multivariate analysis is to make an assessment of the relative contribution of the outcome variables to the resultant group differences, or to the resultant effects of the "treatment" variable(s). This is the so-called variable ordering problem. Although the

assessment of variable importance is very difficult in all multi-variable analyses (including canonical correlation, factor analysis, cluster analysis), some reasonable indexes have been proposed for the MANOVA context (see Huberty, 1984). Of course, a meaningful ordering of variables that constitute a system can only be legitimately accomplished by taking the variable intercorrelations into consideration.

In a multiple-group situation, the study of system structure and of variable importance may lead to some interesting and informative conclusions. In the univariate case, group contrasts (pairwise or complex) are often of interest in addition to, or in lieu of, the omnibus inter-group comparison. Group contrasts may also be studied with multiple outcome variables--here we have multivariate group contrasts. The construct associated with one contrast may be characterized quite differently from that for another contrast. Also, the variable orderings for effects defined by two contrasts may be quite different. For a detailed discussion of this analysis strategy, see Huberty and Smith (1982).

None of the above three data analysis problems (system structure, variable selection, variable ordering) can be appropriately approached via multiple univariate analyses. As Gnanadesikan and Kettenring (1984, p. 323) put it, an objective of a multivariate analysis is to increase the "sensitivity of the analysis through the exploitation of the inter-correlations among the response variables so that indications that may not be noticeable in separate univariate analyses stand out more clearly in the multivariate analysis."

It should be pointed out that typically employed criteria for variate selection and variable ordering are sample- and system-specific. What is a good variable subset or a relatively good individual variable depends upon the collection of the variables in the system being studied. How well the proposed selection and ordering criteria "hold up" over repeated sampling needs further empirical study. Of course, replication is highly desirable. The rank-order position of a variable in a system of variables may change when new variables are added to the system. Similarly for the composition of a good subset of variables. Hence, a conclusion regarding the goodness of a variable subset and/or the relative goodness of individual variables must be made with some caution (see Huberty, 1986, for elaboration).

Additional Comments

Some apparently "funny" results can occur when comparing multivariate analysis with multiple univariate analyses. Significant univariate results do not necessarily imply significant multivariate results (see, e.g., Cramer, 1975), and vice versa (see, e.g., Tatsuoka, 1971, pp. 13-24). Of course, the meaning of "significant" in the two approaches may be different. Does rejecting a MANOVA null hypothesis lead to the same conclusion as rejecting one or more ANOVA null hypotheses? How does one compare a single P-value from MANOVA with the

multiple P-values from the ANOVAs? Furthermore, how does one compare the power of a multivariate test with the power of a set of univariate tests? These types of comparisons are problematic, particularly because of "inconsistent" MANOVA - ANOVA results that may occur.

Ignoring the interrelatedness of a collection of outcome variables can lead to obtaining redundant information. For example, suppose Variable 1 yields univariate significance, and that Variable 2 is highly correlated with Variable 1. Significance yielded by Variable 2, then, would not be a new result. Van de Geer (1971, p. 271) points out that, "with separate analyses of variance for each variable, we never know how much the results are duplicating each other."

In summary, if a collection of outcome variables constitutes a potentially meaningfull system, then a multivariate analysis called for. That is, a multivariate analysis should be conducted if interest is on potential underlying constructs. If not, then a multiple univariate analysis route would be taken (without a preliminary multivariate analysis). If control over Type I error is of concern when conducting multiple univariate analyses, it is suggested that Bonferroni-adjusted probability values be considered.

References

Bird, K. D. (1975). Simultaneous contrast testing procedures for multivariate experiments. Multivariate Behavioral Research, 10, 343-351.

Bird, K. D., & Hadzi-Pavlovic, D. (1983). Simultaneous test procedures and the choice of a test statistic in MANOVA. Psychological Bulletin, 93, 167-178.

Biskin, B. H. (1980). Multivariate analysis in experimental counseling research. The Counseling Psychologist, 8, 69-72.

Bock, R. D. (1975). Multivariate statistical methods in behavioral research. New York: McGraw-Hill.

Bray, J. H., & Maxwell, S. E. (1982). Analyzing and interpreting significant MANOVAs. Review of Educational Research, 52, 340-367.

Cramer, E. M. (1975). The relation between Rao's paradox in discriminant analysis and regression analysis. Multivariate Behavioral Research, 10, 99-107.

Dempster, A. P. (1971). An overview of multivariate data analysis. Journal of Multivariate Analysis, 12, 316-346.

Dossey, J. A. (1976). The relative effectiveness of four strategies for teaching algebric and geometric disjuncti concepts and for teaching inclusive and exclusive disjunctive concepts. Journal for Research in Mathemati Education, 7, 92-105.

Elkins, J., & Sultmann, W. F. (1981). ITPA and learning disability: A discriminant analysis. Journal of Learni Disabilities, 14, 88-92.

Gamos, P. A. (1977). An improved table for simultaneous control on g contrasts. Journal of the American Statistical Association, 72, 531-534.

Gnanadesikan, R., & Kettenring, J. R. (1984). A pragmatic review of multivariate methods in applications. In H. A. David & H. T. David (Eds.), Statistics: An appraisal (pp. 309-337). Ames, IA: Iowa State Universi Press.

Hackman, J. D., & Taber, T. D. (1979). Patterns of undergraduate performance related to success in college. American Educational Research Journal, 16, 117-138.

Harris, R. J. (1985). A primer of multivariate statistics. New York: Academic Press.

Huberty, C. J (1972). Regression analysis and 2-group discriminant analysis. Journal of Experimental Education. 41, 39-41.

Huberty, C. J (1984). Issues in the use and interpretation of discriminant analysis. Psychological Bulletin, 95, 156-171.

Huberty, C. J (1986, April). Problems with stepwise methods--Better alternatives. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Huberty, C. J, & Smith, J. D. (1982). The study of effects in MANOVA. Multivariate Behavioral Research, 17, 417-432.

Hummel, T. J., & Johnston, C. B. (1986, April). An empirical comparison of size and power of seven methods for analyzing multivariate data in the two-sample case. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Hummel, T. J., & Sligo, J. R. (1971). Empirical comparison of univariate and multivariate analysis of variance procedures. Psychological Bulletin, 76, 49-57.

Leary, M. R., & Altmaier, E. M. (1980). Type I error in counseling research: A plea for multivariate analyses. Journal of Counseling Psychology, 27, 611-615.

Lunneborg, C. E., & Lunneborg, P. W. (1977). Is there room for a third dimension in vocational interest differentiation? Journal of Vocational Behavior, 11, 120-127.

O'Grady, K. E. (1986). Simultaneous tests and confidence intervals. Behavior Research Methods, Instruments, & Computers, 18, 325-326.

Ryan, T. A. (1980). Comment on "Protecting the overall rate of Type I errors for pairwise comparisons with an omnibus test statistic." Psychological Bulletin, 88, 354-355.

Tatsuoka, M. M. (1971). Significance tests. Champaign, IL: Institute for Personality and Ability Testing.

Van de Geer, J. P. (1971). Introduction to Multivariate Analysis for the Social Sciences. San Francisco: Freeman.

Watterson, O. M., Joe, G. W., Cole, S. G., & Sells, S. B. (1980). Impression management and attitudes toward marihuana use. Multivariate Behavioral Research, 15, 139-156.

Wilkinson, L. (1975). Response variable hypotheses in the multivariate analysis of variance. Psychological Bulletin, 82, 408-412.

Footnotes

[1]In attempting to encourage graduate students to formally study multivariate methods, the current author has often been confronted with a response such as: "A researcher, keeping in mind some 'theory' underlying a research effort, poses his/her questions first, then seeks analyses to answer the questions. If multivariate analyses are imminent, then he/she can approach a 'statistician' for help." My argument, which only seldom is heeded, is that knowledge of multivariate techniques should enable the researcher to pose more interesting, relevant, and penetrating questions to begin with.

[2]It has been pointed out by Harris (1985, pp. 129, 257, 319) and proven by Huberty (1972) that in the two-group case, the squared LDF-variable correlations are proportional to the univariate F values. Thus, it might seem that if a system structure is to be identified via loadings, then multiple univariate analyses would suffice. In the multiple-group case where at least two LDFs result, however, the multiple constructs cannot be identified by multiple univariate analyses.

[3]The notion of a "construct" may be viewed as a varying one across different types of multivariate analyses. For the group-comparison or grouping-variable-effects situation on which we focus herein, the identified constructs are <u>extrinsic</u> to the set of outcome variables. That is, the optimization of the composites (i.e., LDFs) is based on something external to the outcome variables, namely, the maximization of effects.

Similarly for the optimization of composites (linear classification functions) in the context of predictive discriminant analysis (see Huberty, 1984) where classifica accuracy is maximized. On the other hand, in component analysis, for example, the identified constructs are intri to the set of outcome variables. That is, the optimizatio the composites (i.e., components) is based on something internal to the outcome variables, namely, the maximizatio accounted-for variance in the variable set. Furthermore, extrinsic-intrinsic, constructs-of-constructs situation co result when one conducts a MANOVA (or classification analy using component or factor scores as input.

MULTIPLE LINEAR REGRESSION VIEWPOINTS
VOLUME 16, NUMBER 1, SPRING 1988

Developmental Trends In Anadrogyny: Implications for Measurement

Bruce Thompson
University of New Orleans

Janet G. Melancon
Loyola University

ABSTRACT

The present study was conducted to investigate differences in item performance, reliability, and scale means of the Bem Sex-Role Inventory when comparisons are made across developmentally different groups. Analyses were conducted comparing results for adolescents with results for adults, and further analyses were conducted comparing results for the adolescents across various adolescent gender and age groups. The results tend to support the a conclusion that the BSRI has reasonable measurement integrity when used with adolescents, and thus indicate that the measure may be useful in exploring developmental changes in sex-role perceptions as they occur during adolescence.

In a seminal article in the literature on personality, Constantinople (1973) argued that persons could possess both characteristics that are stereotypically male as well as characteristics that are stereotypically female. Personality researchers have come to call such persons androgynous. Bem (1975, p. 634) has argued that "a non-androgynous sex role can seriously restrict the range of behaviors available to an individual as he or she moves from situation to situation." Kelly and Worrell (1977) summarize studies that have empirically tested the proposition that androgyny is an adaptive personality characteristic. Generally studies support Bem's position, though some studies (Heilburn, 1984) suggest that the trait may be more advantageous to females than to males.

Although several measures of androgyny have been developed, the Bem Sex-Role Inventory (BSRI) (Bem, 1974) "has been the most frequently used of the recent sex role instruments" (Koenigsberg, 1982, p. 2). However, the BSRI and the methods used to measure the androgyny construct have both been topics of heated academic discussion (e.g., Bem, 1979; Pedhazur & Tetenbaum, 1979).

Studies of the BSRI measure have been extraordinarily diverse in their methods and designs. Sample sizes have ranged from 44 (Bledsoe, 1983) to 894 (Sassenrath & Yonge, 1979). Powell (1979) employed 15 samples to cross-validate his results. Although many studies have used variations of common factor analysis to evaluate the measure, researchers have also employed multidimensional scaling (Koenigsberg, 1982), smallest space analysis (Ruch, 1984), confirmatory factor analysis (Marsh, 1985), analysis of the variance/covariance matrix (Belcher,

Crocker & Algina, 1984), and extraction of second-order factors (Edwards, Gaa & Liberman, 1978). Thompson (1986) presented a meta-analytic integration of the various factor analytic studies and concluded that the theoretically expected structure underlies BSRI data. Even seemingly contradictory results are generally supportive of the measure's validity once solutions are rotated into a common factor space.

Virtually all of these myriad studies have examined statistics that are a function of covariations (e.g., covariances, correlations) among item responses. However, these statistics are insensitive to the influence of central tendency. For example, two sets of scores can be perfectly correlated when: (1) both sets each have a mean of 5.0; or (2) both sets each have a mean of 1.0; or (3) one score set has a mean of 1.0 and the other score set has a mean of 5.0.

Since structure is a function of the relationships among items, a test may have a similar structure in diverse populations, but the populations may differ with respect to other aspects of item performance. For example, item means could be markedly different across populations even if the structures across the populations were identical. As Gorsuch (1983, p. 335) notes,

> To the extent that invariance can be found across systematic changes in either variables or the individuals, then the factors have a wider range of applicability as generalized constructs. The subpopulations over which the factor occurs could--

and probably would--differ in their mean scores or variances across the groups, but the pattern of relationships among the variables would be the same.

Knowledge regarding such a dynamic would be important from a measurement perspective because the process of summing item scores within a scale also assumes that all the items are reasonably homogeneous with respect to their mean values. This assumption is made with respect to both item characteristics within a given population and item performance across populations, if the test is to be employed in various populations.

A concrete example may clarify the essential character of this assumption. If the item means on a two item test in a population were both four on a seven-point scale, then a person who scored five on both scales is deviating from the expected item means by the same amount, and the scale score of 10 for the person represents a meaningful deviation from the known total score mean of eight. But say the population mean responses to items one and two were, respectively, six and two. The person who scores, respectively, six and two on the items is assigned a scale score of eight. The person who scores two and six is also assigned a scale score of eight, even though the two sets of item scores represent very different responses when compared with expected or average population responses.

It is unfortunate that central tendency has not been considered a noteworthy issue in most of the previous research on androgyny measures. The instruments that measure androgyny

typically produce Masculine and Feminine scale scores by summing relevant item scores. If means are not comparable across items within a given sample, then scores on items deviate about different means and adding item scores without considering these variations may distort total scores--the scores may lack measurement validity and studies using the measures may therefore be invalid. The process of adding item scores without considering variations in item means requires the critical assumption that the items are deviating about the same or at least comparable means so that one is not adding "apples and oranges", i.e., so that the addition process is itself valid.

Even if item means are comparable across items within given sample types, it is important to ascertain whether the item means are also comparable across sample types, e.g., developmentally active adolescent groups versus adult samples. If differences in scale means across developmental groups are due to a few items, the content of those items may have substantive implications or may raise questions about the validity of those items when used with certain types of samples.

However, most of the studies in this area have employed college students as subjects. The similar character of most of the samples limits ability to generalize about the validity of the BSRI. As Worell (1978, p. 783) notes, "restricting all of the sex-role research to college students, unfortunately, leaves us with many unanswered questions about the generality of results and the applicability to constrast populations." It is especially surprising that so few studies have employed adolescents as

subjects. Bem (1979, p. 1052) argues that even young children are aware of sex-roles. Marsh and Myers (1984) tested adolescent girls but school officials allowed the use of only a subset of BSRI items. Mills (1980) employed a sample of 418 adolescents, but primarily was concerned with the structure underlying BSRI responses rather than with central tendency of item responses.

The present study was conducted to investigate differences in BSRI results involving developmentally different subject groups. Three research questions were considered in the study. First, how comparable are item means across different developmental and sex groups? The influence of sex was considered since there are developmental differences across gender groups and since the BSRI measures sex-role perceptions that may also differ across groups as an interactive function of both developmental group and gender. Second, within a sample of adolescents, what are the influences of age and sex on BSRI reliability coefficients? If the test is reliable when used with younger subjects, the measure may be an important vehicle for investigating changes in adolescents' sex-role perceptions. Finally, what differences in the two BSRI scale means are there across adolescent age and sex groupings? The analysis of scale score means may provide some such insight regarding these changes.

Results

Several of the many BSRI validity studies in the literature report item means for biologically male subjects as against female subjects. Thus, five sets of item means from adult samples

were available from previous research. In order to provide a developmentally different comparison group, the present authors collected data from 256 adolescents (25% girls) ranging in age (mean = 12.9; SD = 1.86) from 9 to 17. These data were analyzed in several ways in order to address the study's first research question.

Figure 1 presents the item means reported in each of the previous studies. In order to facilitate comparisons, the means are graphically presented along the one to seven response scale employed on the instrument. Bem (1981) has proposed that a "short form" of her instrument can be constructed by only scoring 20 of the items on the BSRI. These items are underlined in Figure 1. Letters "A" through "E", respectively, represent: a) the means reported by Bledsoe (1983) in a study involving 44 female teachers; b) Hoferek's (1981) means from a nationwide survey of physical educators involving 189 women; c) Pedhazur and Tetenbaum's (1979) means for 489 female graduate education students; d) Hoferek's (1981) means for 102 men; and e) Pedhazur and Tetenbaum's (1979) means for 171 men. The means for the male adolescents in the study are repesented by pound signs ("#"); the means for female adolescents are represented by asterisks ("*"). The means for the two adolescent gender groups are presented within their 95% confidence intervals, represented by hyphens. The items are sorted first by scale; the 20 BSRI Feminine scale items follow the 20 Masculine scale items. For each item, the mean of the two means for adolescents and the mean of the five means for adults were computed, as was the deviation of these two statistics. This difference score is presented in parentheses for

each BSRI item. Within each scale, the items presented in Figure 1 have been arranged in order of descending differences across the two subject groups.

INSERT FIGURE 1 ABOUT HERE.

In order to compare the variability of item means across items and across the seven subject groups, on each of the 40 items a classical sex-by-age-group two-way analysis of variance was conducted using the item means as the dependent variable. Table 1 presents the 40 BSRI items in descending order of variability of the mean scores. Thus, for example, means on the item, "Feminine", tended to vary most across the seven subject samples. For each item, Table 1 also presents the sum of squares attributable to each effect and the percentage of each item's sum of squares that is attributable to each source of variance in the analysis.

INSERT TABLE 1 ABOUT HERE.

The reliability coefficients presented in Table 2 were computed in order to address the study's second research question. The table reports the alpha reliability coefficients for the two BSRI scales across various age and gender groups.

INSERT TABLE 2 ABOUT HERE.

Total scale scores within the various age and gender groups represented in the adolescents' data set were compared in order to address the study's third research questions. Table 3 presents

the cell means across the subject groups. Table 4 reports the results of a two-way analysis of variance for both the BSRI scales.

INSERT TABLES 3 AND 4 ABOUT HERE.

Discussion

The analyses reported in Figure 1 compared means of means in order to minimize the influence of disproportionate sample sizes in the various groups. The Figure 1 results indicate that adolescents tend to score lower across almost all of the BSRI items. In particular, with respect to Masculine items, the adolescent subjects perceived themselves to be less analytical, self-sufficient, self-reliant, forceful, independent, and forceful. The finding is not surprising, and primarily reflects perception of the reality that adolescents are dependent on others. The finding that adoloscents consider themselves less analytical may reflect a perceived obligation to be carefree.

With respect to the Feminine items, the adolescents perceived themselves to be less sensitive, compassionate, sympathetic, tender, warm and gentle. These results suggest a self-orientation that may be an adaptive effort to work through issues involving identity and role expectations.

These findings do not contradict a view that adolescence is a time of role exploration (Erikson, 1963, pp. 247-269), but suggest that this exploration may primarily be achieved by the "doing" of trying on roles rather than through the "thinking" of reflection. In fact, psychoanalytic theory (A. Frued, 1972, pp.

317-318) suggests that this doing may be an important component of adjustment:

> The character structure of a child at the end of the latency period... has to be abandoned to allow adult sexuality to be integrated into the individual personality. The so-called adolescent upheavels are no more than the external indications that such internal adjustments are in progress... We all know individual children who as late as the age of fourteen, fifteen, or sixteen show no such outer evidence of inner unrest... They are, perhaps more than any others, in need of therapeutic help.

The results presented in Table 1 provide further insight regarding the measurement characteristics of individual BSRI items--the magnitudes and the sources of variance in the mean scores from the various subject groups are presented. The variablitity (SOS=25.93; V=25.93/6=1.11; SD=1.05) of the seven means on the item, atheletic, was an artifact generated by including data from Hoferek's (1981) physical educators, who perceived themselves to be more atheletic than other subject groups. However, it is clear that there was disproportionate variability on two other items, feminine (SD=2.08) and masculine (SD=2.01). These standard deviations are especially noteworthy since the response format only ranges from one to seven.

It is disturbing that the vast preponderance of the variability on these items was associated with gender, as indicated by the effect sizes of sex for these items. Bem (1981, p. 14) has not included these items in the "short form" portion

of her measure:

> Note that the terms "feminine" and "masculine" have themselves been eliminated from the Short Form of the BSRI. These terms actually reflect "higher-order" traits and are constructs denoting clusters of traits themselves rather than behaviors.

However, a more parsimonious and thus more likely interpretation would argue that the these two items merely measure physical gender, as suggested by the present analyses. If so, the inclusion of these items seriously undermines the validity of the measure, since the measures purportedly evaluates psychological orientation regarding sex-roles and not physical gender. Thus the use of these items has been criticized previously on both theoretical and empirical grounds (Pedhazur & Tetenbaum, 1979).

Bem (1981, p. 5) notes that "the test is arranged so that the thirty short-form items appear first and, where time is limited, subjects may be instructed to stop after the item 'conventional.'" However, the savings in time from using the short form is very minimal. Many researchers will be tempted to employ the original "long form" so that their results will be more comparable with previous research and because they may presume that the long form will be more reliable since it is longer. However, the two forms are highly correlated (Bem, 1981, p. 15), and the "short form" Masculine scale is at least as reliable as the "long form" M scale and the "short form" Feminine scale is noticeably more reliable (Bem, 1981, p. 14) and may well be more valid. The use of the "short form" or of

the "long form" minus these two items is therefore strongly recommended for most research applications.

The remaining analyses presented in Table 1 support the previous interpretation of Figure 1. For example, large effect sizes for age were found for the items, sensitive, compassionate, analytical, and other variables noted previously. Nevertheless, the variability in item means across developmental groups was relatively small, was systematic rather than random, and involved theoretically interpretable differences. The analysis suggests that item means are reasonably comparable across subject groups, so that measurement concerns regarding this aspect of test performance are not appreciably warranted.

The analyses reported in Table 2 suggest that the BSRI has reasonable reliability even when used by younger subject groups. The Masculine scale reliability coefficient of .82 compares favorably with values of about .86 reported by Bem (1981, p. 14). The Feminine scale reliability coefficient of .78 compares favorably with values of about .78 reported by Bem (1981, p. 14) for several studies with adults. The Table 2 results also suggest that the measure can be reasonably employed even with younger age groups within the adolescent age range. The results must be interpreted with some caution, since some age groups included few subjects, but the pattern is consistent across the ages represented in the study.

The results presented in Tables 3 and 4 suggest that both gender groups tend to score somewhat higher on both scales as individuals grow older. However, the most noteworthy pattern is that males tend systematically to become more Masculine while

females tend to become systematically more Feminine as they age during adolescence. The tabled results also indicate that males and females are more comparable with respect to their Masculine scores than with respect to their Feminine scale scores. This suggests that females may be more likely to become androgynous than are their male peers. Males may find androgyny less advantageous during adolescence, just as some research suggests that androgyny may generally be more functionally advantageous for adult females (Heilburn, 1984).

In summary, the results of the present study generally support the conclusion that scores on the Bem Sex-Role Inventory (Bem, 1974) are reasonably reliable and valid even when subjects are young adolescents. Although the present results corroborate previous findings that the two items, masculine and feminine, do not have desirable measurement characteristics, variations in item performance across developmentally different groups generally were relatively small and were predictable. Thus, the BSRI measure may be helpful in exploring the development of sex-role perceptions during adolescence, or in tracking the effects of culture changes on the sex-role development process as societal expectations and norms change.

References

Belcher, M.J., Crocker, L.M., & Algina, J. (1984). Can the sam
instrument be used to measure sex-role perceptions of males an
females? Measurement and Evaluation in Guidance, 17, 15-23.

Bem, S.L. (1974). The measurement of psychological androgyny
Journal of Consulting and Clinical Psychology, 42, 155-162.

Bem, S.L. (1975). Sex role adaptability: One consequence o
psychological androgyny. Journal of Personality and Socia
Psychology, 31, 634-643.

Bem, S.L. (1979). Theory and measurement of androgyny: A reply t
the Pedhazur-Tetenbaum and Locksley-Colten critiques. Journal o
Personality and Social Psychology, 37, 1047-1054.

Bem, S. L. (1981). Bem sex-role inventory. Palo Alto, CA: Consultin
Psychologists Press.

Bledsoe, J.C. (1983). Factorial validity of the Bem Sex-Ro
Inventory. Perceptual and Motor Skills, 56, 55-58.

Constantinople, A. (1973). Masculinity-feminity: An exception to th
famous dictum. Psychological Bulletin, 80, 389-407.

Edwards, T.A., Gaa, J.P., & Liberman, D. (1978). A factor analys
of the BSRI and the PAQ. Paper presented at the annual meeti
of the American Psychological Association, Toronto. (ER
Document Reproduction Service No. ED 177 187)

Erikson, E. H. (1963). Childhood and society (2nd ed.). New York:
Norton.

Freud, A. (1972). Adolescence. In J. F. Rosenblith, W. Allnsmit
and J. P. Williams (Eds.), The causes of behavior. Boston: All
and Bacon, pp. 317-323.

Gorsuch, R.L. (1983). Factor analysis (2nd ed.). Hillsdale, N

Erlbaum.

Heilbrun, A.B. (1984). Sex-based models of androgyny: A further cognitive elaboration of competence differences. Journal of Personality and Social Psychology, 46, 216-229.

Hoferek, M.J. (1981). Factor analysis of the Bem Sex-Role Inventory using data from physical educators. Unpublished manuscript. (ERIC Document Reproduction Service No. ED 221 603)

Kelly, J.A., & Worell, J. (1977). New formulations of sex roles and androgyny: A critical review. Journal of Consulting and Clinical Psychology, 45, 1101-1115.

Koenigsberg, E. (1982). Perceptual-cognitive structure of sex-typed traits: Multidimensional scaling of the Bem Sex Role Inventory. Paper presented at the annual meeting of the American Educational Research Association, New York. (ERIC Document Reproduction Service No. ED 218 322)

Marsh, H. W. (1985). The structure of masculinity/femininity: An application of confirmatory factor analysis to higher-order factor structures and factorial invariance. Multivariate Behavioral Research, 20, 427-449.

Marsh, H.W., & Myers, M. (1984). Masculinity, femininity and androgyny: A methodogolical and theoretical critique. Collected papers of 1984 National Conference of the Australian Association for Research in Education. Perth, Western Australia: Australian Association for Research in Education, pp. 566-581. (ERIC Document Reproduction Service No. ED 242 758)

Mills, C.J. (1980). A factor-analytic comparison of two measures of masculinity-femininity in post and early adolescent populations.

Paper presented at the annual meeting of the Eastern Psychological Association, Hartford, CT. (ERIC Document Reproduction Service No. ED 189 495)

Pedhazur, E.J., & Tetenbaum, T.J. (1979). Bem Sex Role Inventory: A theoretical and methodological critique. Journal of Personality and Social Psychology, 37, 996-1016.

Powell, G.N. (1979, September). Factor analysis of the BSRI revisited: A comprehensive study. Paper presented at the annual meeting of the American Psychological Association, New York.

Ruch, L.O. (1984). Dimensionality of the Bem Sex Role Inventory: A multidimensional analysis. Sex Roles, 10, 99-117.

Sassenrath, J.M., & Yonge, G.D. (1979). Bem Sex Role Inventory Reexamined. Psychological Reports, 3, 935-941.

Thompson, B. (April, 1986). Performance on an androgyny measure of developmentally active versus other groups. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. (b)

Worell, J. (1978). Sex roles and psychological well-being: Perspectives on methodology. Journal of Consulting and Clinical Psychology, 46, 777-791.

Table 1
Classical SOS Decomposition

Variable	SOS Tot	SOS Sex	Effect Size	SOS Age	Effect Size	Inter	Effect Size
Feminine	25.93	25.183	97.12%	0.002	0.01%	0.055	0.21%
Masculine	24.36	23.989	98.48%	0.235	0.96%	0.175	0.72%
Atheletic	6.64	0.922	13.89%	0.138	2.08%	0.204	3.07%
Sensitive*	2.58	0.068	2.64%	2.224	86.20%	0.156	6.05%
Competitive	2.18	0.843	38.67%	0.012	0.55%	0.013	0.60%
Compassionate*	2.10	0.044	2.10%	1.773	84.43%	0.051	2.43%
Analytical	1.97	0.076	3.86%	1.910	96.95%	0.021	1.07%
Sympathetic*	1.84	0.137	7.45%	1.354	73.59%	0.161	8.75%
Childlike	1.79	0.253	14.13%	0.014	0.78%	0.058	3.24%
Self-sufficient	1.76	0.010	0.57%	1.537	87.33%	0.011	0.63%
Forceful#	1.68	0.163	9.70%	1.417	84.35%	0.150	8.93%
Tender*	1.61	0.136	8.45%	1.238	76.89%	0.030	1.86%
Self-reliant	1.53	0.000	0.00%	1.382	90.33%	0.005	0.33%
Eager soothe*	1.48	0.259	17.50%	0.803	54.26%	0.199	13.45%
Loves children*	1.46	0.287	19.66%	0.282	19.32%	0.679	46.51%
Affectionate*	1.42	0.228	16.06%	0.823	57.96%	0.152	10.70%
Acts as leader	1.36	0.311	22.87%	0.496	36.47%	0.001	0.07%
Independent#	1.33	0.003	0.23%	1.048	78.80%	0.000	0.00%
Gentle*	1.32	0.057	4.32%	0.934	70.76%	0.220	16.67%
Warm*	1.25	0.067	5.36%	0.979	78.32%	0.032	2.56%
Loyal	1.22	0.083	6.80%	0.913	74.84%	0.097	7.95%
Has Leadership#	0.90	0.233	25.89%	0.383	42.56%	0.007	0.78%
Take stand#	0.85	0.177	20.82%	0.633	74.47%	0.001	0.12%
Willing risk#	0.78	0.412	52.82%	0.000	0.00%	0.013	1.67%
Makes decisions	0.66	0.274	41.52%	0.356	53.94%	0.009	1.36%
Understanding*	0.65	0.060	9.23%	0.524	80.62%	0.019	2.92%
Soft-spoken	0.64	0.179	27.97%	0.274	42.81%	0.099	15.47%
Assertive#	0.63	0.002	0.32%	0.580	92.06%	0.018	2.86%
No harsh lang	0.62	0.062	10.00%	0.481	77.58%	0.001	0.16%
Aggresive#	0.62	0.271	43.71%	0.001	0.16%	0.001	0.16%
Individualist	0.53	0.001	0.19%	0.326	61.51%	0.177	33.40%
Defends belief#	0.50	0.004	0.80%	0.309	61.80%	0.038	7.60%
Dominant#	0.48	0.270	56.25%	0.110	22.92%	0.038	7.92%
Ambitious	0.43	0.063	14.65%	0.032	7.44%	0.088	20.47%
Gullible	0.36	0.216	60.00%	0.058	16.11%	0.033	9.17%
Cheerful	0.29	0.035	12.07%	0.123	42.41%	0.030	10.34%
Strong person#	0.23	0.007	3.04%	0.146	63.48%	0.000	0.00%
Flatterable	0.23	0.018	7.83%	0.021	9.13%	0.148	64.35%
Shy	0.19	0.013	6.84%	0.015	7.89%	0.023	12.11%
Yielding	0.04	0.001	2.50%	0.007	17.50%	0.012	30.00%

*
Scored as a Masculine item as part of the "short form."
#
Scored as a Feminine item as part of the "short form."

Table 2
Alpha Reliability Coefficients for Adolescents

	Both Sexes			Males			Females	
Age	n	M	F	n	M	F	n	M	F
9	8	.62	.85	8	.62	.85	--	--	--
10	16	.85	.07	13	.82	.04	3	.93	.58
11	30	.85	.78	22	.83	.78	8	.88	.37
12	54	.81	.59	46	.76	.49	8	.86	.82
13	64	.74	.81	50	.74	.79	14	.80	.88
14	37	.84	.84	23	.84	.84	14	.88	.80
15	20	.90	.81	12	.90	.80	8	.90	.54
16	16	.86	.87	10	.85	.84	6	.67	.87
17	11	.85	.86	8	.88	.84	3	.59	.29
All	256	.82	.78	192	.79	.74	64	.88	.82

Table 3
Cells Means for Two-Way Analysis

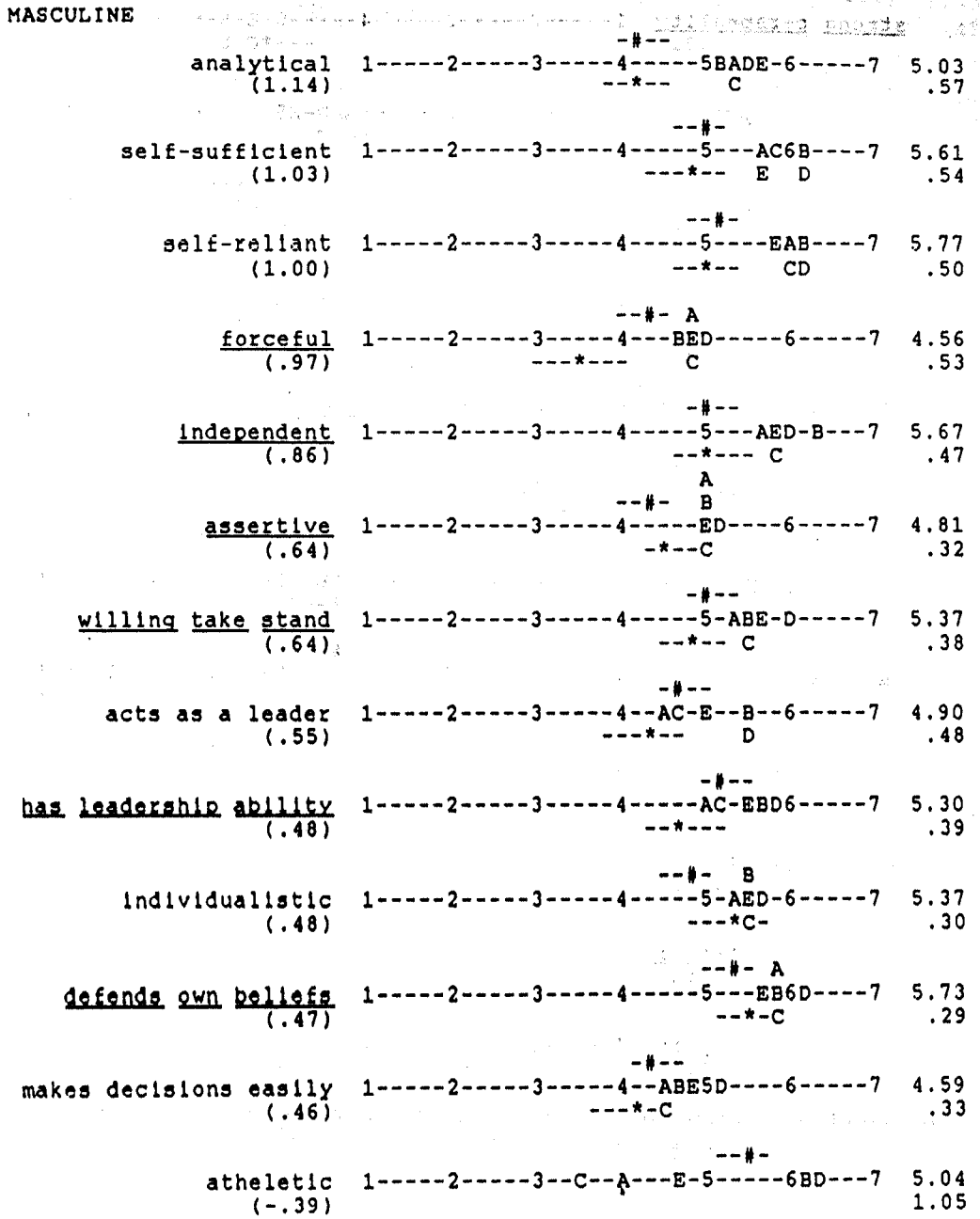
	Masculine			Feminine		
Age	Males	Females	Total	Males	Females	Total
9	94.5	--	94.5	83.6	--	83.6
10	93.0	81.0	90.8	79.5	81.7	79.9
11	98.7	87.1	95.6	82.2	98.2	86.5
12	99.0	78.5	95.9	81.9	88.4	82.8
13	96.1	97.6	96.4	83.4	92.2	85.4
14	100.6	98.6	99.8	82.0	97.2	87.7
15	102.9	91.4	98.3	76.0	93.0	82.8
16	103.3	85.5	96.6	89.9	106.8	96.2
17	103.5	94.7	101.1	91.0	111.7	96.6
Total	98.5	91.3	96.7	82.7	95.5	85.9

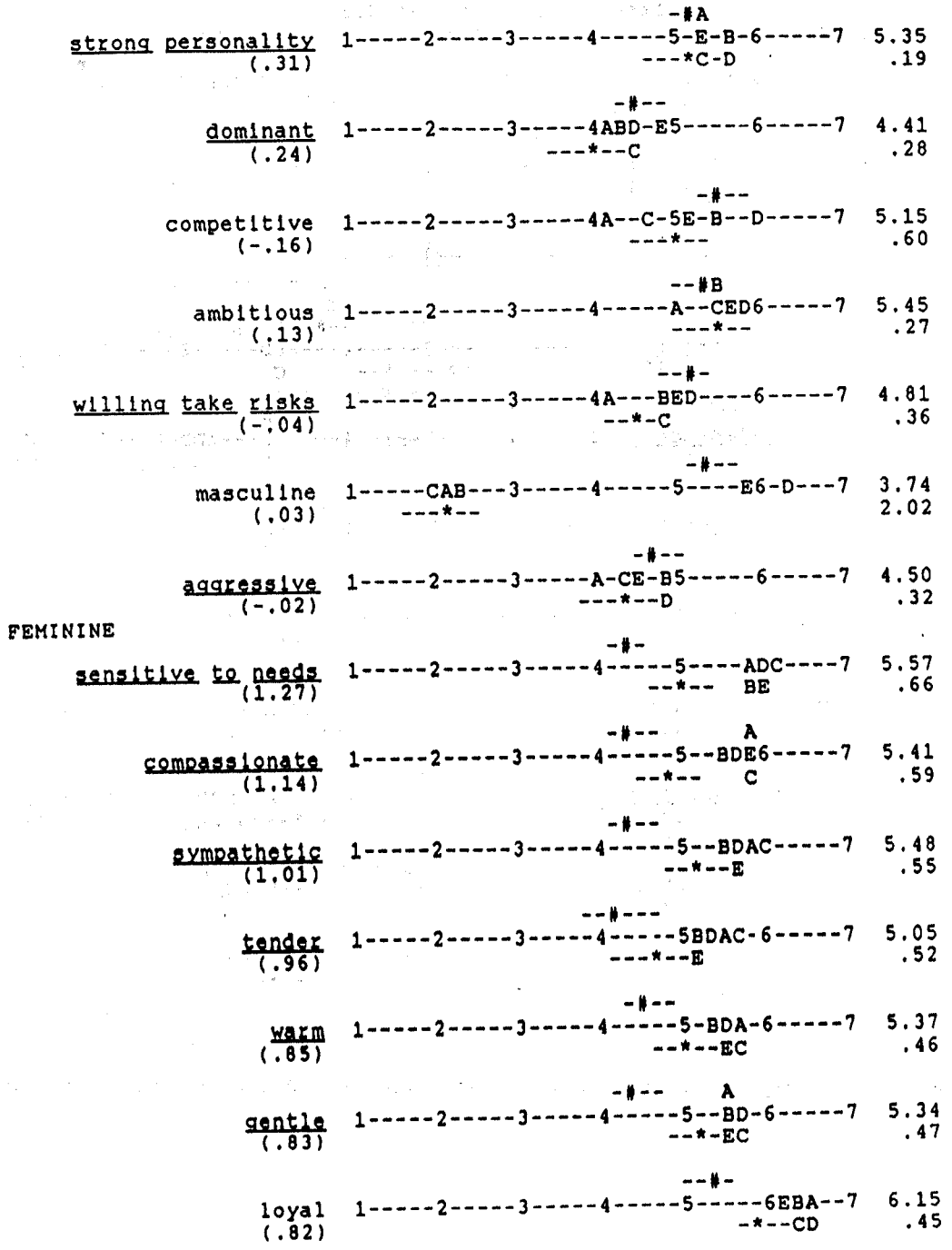
Table 4
Classic SOS Decomposition Across Scales

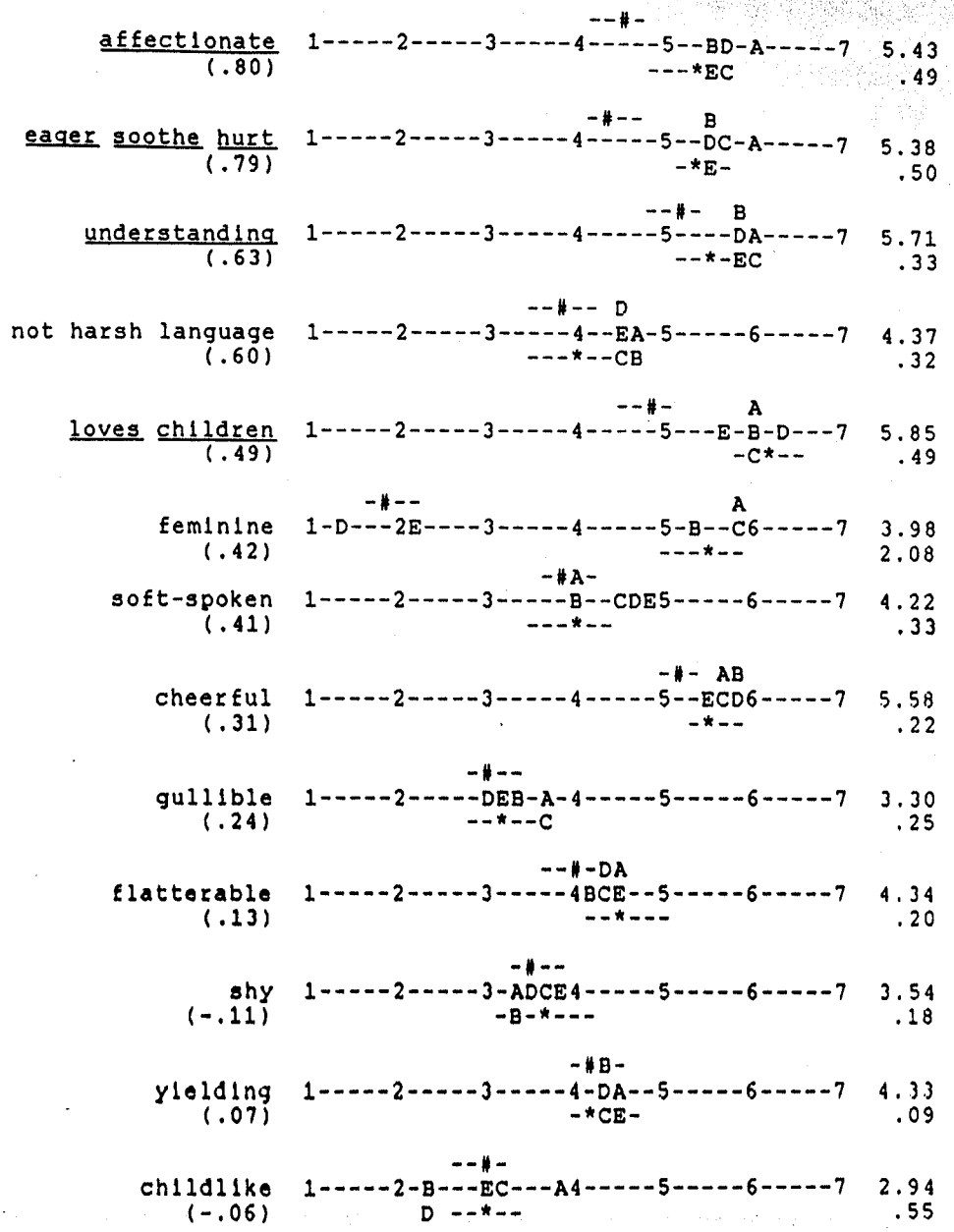
Masculine Source	SOS	df	MS	Fcalc	Effect Size
Age	1996.2	8	249.5	.99	3.0%
Sex	3177.2	1	3177.2	12.70	4.7%
Age*Sex	2875.4	7	410.8	1.64	4.3%
Residual	59772.4	239	250.1		
Total	67123.8	255	263.2		

Feminine Source	SOS	df	MS	Fcalc	Effect Size
Age	3542.0	8	442.7	2.47	6.4%
Sex	6969.6	1	6969.6	38.90	12.6%
Age*Sex	1114.3	7	159.2	.89	2.0%
Residual	42823.0	239	179.2		
Total	55347.5	255	217.0		

Figure 1
Item Means Across Studies




strong personality (.31)
1-----2-----3-----4-----5-E-B-6-----7 5.35
---*C-D .19
dominant (.24)
1-----2-----3-----4ABD-E5-----6-----7 4.41
---*--C .28
competitive (-.16)
1-----2-----3-----4A--C-5E-B--D-----7 5.15
---*-- .60
ambitious (.13)
1-----2-----3-----4-----A--CED6-----7 5.45
---*-- .27
willing take risks (-.04)
1-----2-----3-----4A---BED----6-----7 4.81
--*-C .36
masculine (.03)
1-----CAB---3-----4-----5----E6-D---7 3.74
---*-- 2.02
aggressive (-.02)
1-----2-----3-----A-CE-B5-----6-----7 4.50
---*--D .32
FEMININE
sensitive to needs (1.27)
1-----2-----3-----4-----5----ADC----7 5.57
--*-- BE .66
compassionate (1.14)
1-----2-----3-----4-----5--BDE6-----7 5.41
--*-- C .59
sympathetic (1.01)
1-----2-----3-----4-----5--BDAC-----7 5.48
--*--E .55
tender (.96)
1-----2-----3-----4-----5BDAC-6-----7 5.05
---*--E .52
warm (.85)
1-----2-----3-----4-----5-BDA-6-----7 5.37
--*--EC .46
gentle (.83)
1-----2-----3-----4-----5--BD-6-----7 5.34
--*-EC .47
loyal (.82)
1-----2-----3-----4-----5-----6EBA--7 6.15
-*--CD .45



Note. The confidence intervals for biologically male adolescents bound "#"; comparable values for females bound "*". The mean of the seven means is presented at the end of each scale; the SD is presented below the mean of the seven means.

If you are submitting a research article other than notes or comments, I would like to suggest that you use the following format if possible:

Title
Author and affiliation
Indented abstract (entire manuscript should be single spaced)
Introduction (purpose—short review of literature, etc.)
Method
Results
Discussion (conclusion)
References

All manuscripts should be sent to the editor at the above address. (All manuscripts should be camera-ready.)

It is the policy of the M.L.R. SIG-multiple linear regression and of *Viewpoints* to consider articles for publication which deal with the theory and the application of multiple linear regression. Manuscripts should be submitted to the editor as original, double-spaced, *camera-ready copy.* Citations, tables, figures and references should conform to the guidelines published in the most recent edition of the *APA Publication Manual* with the exception that figures and tables should be put into the body of the paper. A cost of $1 per page should be sent with the submitted paper. Reprints are available to the authors from the editor. Reprints should be ordered at the time the paper is submitted, and 20 reprints will cost $.50 per page of manuscript. Prices may be adjusted as necessary in the future.

A publication of the Multiple Linear Regression Special Interest Group of the American Educational Research Association, *Viewpoints* is published primarily to facilitate communication, authorship, creativity and exchange of ideas among the members of the group and others in the field. As such, it is not sponsored by the American Educational Research Association nor necessarily bound by the association's regulations.

"Membership in the Multiple Linear Regression Special Interest Group is renewed yearly at the time of the American Educational Research Association convention. Membership dues pay for a subscription to the *Viewpoints* and are either individual at a rate of $5, or institutional (libraries and other agencies) at a rate of $18. Membership dues and subscription requests should be sent to the executive secretary of the M.L.R. SIG."

THE UNIVERSITY OF AKRON
AKRON, OH 44325

ROGERS, BRUCE G.
DEPT OF ED PSYCH
UNIV OF NO IOWA
CEDAR FALLS, IOWA 50613

BOOKS - SPECIAL 4th CLASS RATE

TABLE OF CONTENTS

ISSN 0195-7171