Multiple Linear Regression Viewpoints

A Publication sponsored by the
American Educational Research Association's
Special Interest Group on Multiple Linear Regression

MLRV

Volume 21 • Number 1 • Fall 1994

Table of Contents

Editorial Board

Multiple Linear Regression Viewpoints (ISSN 0195-7171) is published by the AERA Special Interest Group on Multiple Linear Regression (SIG/MLR) through The George Washington University. Subscription and SIG membership information can be obtained from Steven Spaner, Executive Secretary, SIG/MLR, School of Education, University of Missouri, St. Louis, MO 63121. *MLRV* abstracts appear in CIJE, the ERIC system, and microform copies are available from University Microfilms International, 300 North Zeeb Road, Ann Arbor, MI 48106. *MLRV* is listed in the *EBSCO Librarians Handbook*. Second-class postage paid at Washington, DC. POSTMASTER: Send address changes to Ralph Mueller, Department of Educational Research, Graduate School of Education and Human Development, The George Washington University, Washington, DC 20052.

The Rise and Fall and Rise of Multiple Regression

Tianqi Han
Northern Illinois University

Dennis Leitner
Southern Illinois University

The rise and fall and rise of multiple regression is chronicled in the literature by examining its initial impetus and popularity, followed by the acknowledgement of potential problematic issues such as violation of assumptions and overzealous usage, and the subsequent resurgence of the technique as the problems are addressed and procedures clarified. Jacob Cohen brought to the attention of many researchers that multiple regression can be used as a general data-analytic system. With the increasing availability of mainframe computers and programs to perform statistical analysis, journal editors were inundated with an avalanche of regression analyses. The assumptions underlying the analyses were emphasized, considered, and often found to be unmet. Two major problems of using stepwise regression were identified: (1) incorrect degrees of freedom were specified when evaluating changes in explained variance and, (2) incorrect interpretation of stepwise resultswhen a few variables are selected from many. Subsequently, many different regression models have been developed for different situations, especially when assumptions are violated. These models include ridge regression, robust regression, and nonlinear regression.

Like the length of skirts or the cuffs on pant legs, the popularity of statistical tests rises and falls and rises. In his Presidential Address to the Mid-Western Educational Research Association, Leitner (1990) traced this rise and fall and rise of three statistical tests in the literature. First, the initial presentation and use was followed by a second period of the acknowledgement of potential problematic issues such as violation of assumptions and overzealous usage, which resulted in a third period characterized by a resurgence of the technique as the problems are addressed and procedures are clarified. The three statistical techniques he examined were the t-test/analysis of variance, factor analysis and meta-analysis. This same approach is uded in this paper to examine the rise and fall and rise of multiple regression.

While this review of literature is largely chronological, it is not strictly so. Some of the statistical aspects are reported in the mathematical and statistical literature long before they appear in the psychological and educational literature. It is the latter which forms the principal basis of the chronology.

The Initial Rise

In one of the first references to multiple regression in the social science literature, Goldberger (1964), having recognized the nature of multiple regression, pointed out that

> ...[T]he whole point of multiple regression as contrasted with simple regression is to try to isolate the effects of the individual regressors, by 'controlling' on the others. Still, when orthogonality is absent the concept of the contribution of an individual regressor remains inherently ambiguous. (p. 201)

A large impetus for the use of multiple regression came from the work in the late 1960's of the distinguished statistician, Jacob Cohen. Cohen (1968) pointed out that multiple regression and analysis of variance and covariance are special cases of the general linear model.

> If you should say to a mathematical statistician that you have discovered that linear multiple regression analysis and the analysis of variance (and covariance) are identical systems, he would mutter something like, 'Of course--general linear model,' and you might have trouble maintaining his attention. If you should say this to a typical psychologist, you would be met with incredulity, or worse. Yet it is true, and in its truth lie possibilities for more relevant and therefore more powerful exploitation of research data. (Cohen, 1968, p. 426)

He showed that through use of indicator variables (i.e., dummy variable coding), an equivalence between multiple regression and analysis of variance, in fact, exists. In addition, through use of contrast coding, powers and products of variables, and comparisons of appropriate regression equations, multiple regression can be used as a general data-analytic system.

At about the same time, Richard Darlington (1968)

emphasized that, besides providing the partial correlation between the dependent variable and each of the independent variables, regression weights rather than correlation coefficients have the interpretative advantage in prediction allowing statements like "Increasing X_j by 1 unit increases the dependent variable by β_j units" (p. 167). He discussed the logical fallacies involved in using variance-apportionment techniques for any purpose when the independent variables in a set are intercorrelated. He pointed out that the notion of "independent contribution to variance" is meaningless especially when multicollinearity is a problem (p.169).

In perhaps the first text devoted exclusively to the use of multiple regression, Kelly, Beggs, McNeil, Eichelberger, and Lyon (1969) took advantage of the growing presence of high-speed digital computers by freeing the researcher from simplistic designs that can be handled computationally with ease on a desk calculator. By forcing researchers to use "...a series of factorial designs, Type I, Type II models, etc., derived to ease computation with a desk calculator,"...the researchers were either "confused" or had to "impose such constraints on his design that he is forced to ask a limited research question." (Kelly et al., 1969, p. vii).

In addition, Kelly et al. (1969) emphasized that the availability of multiple regression procedures and programs allowed the researcher to ask meaningful research questions.

> The multiple regression analysis presented in this book is designed to prepare the research investigator to construct statistical models which will reflect his original research question rather than limiting that question. Regression analysis will be shown to be the generalized case of analysis of variance. These discussions shall be intimately related to a computer program so that the simple elegance of the generalized analysis of variance is not obscured and so that the investigator can circumvent the anachronistic desk calculator. (p. vii)

Four years later, another popular text of multiple regression was written by Kerlinger and Pedhazur (1973). The book, which listed a different computer program in the appendix than did the Kelly et al. (1969) text, promoted the advantages of multiple regression analysis.

> Multiple regression analysis [is] a most important branch of multivariate analysis... It is a powerful analytic tool widely applicable to many different kinds of research problems. It can be used effectively in sociological, psychological, economic, political, and educational research. It can be used equally well in experimental or nonexperimental research. It can handle continuous and categorical variables. It can handle two, three, four, or more dependent variables. In principle, the analysis is the same. Finally, multiple regression analysis can do anything the analysis of variance does ... (Kerlinger & Pedhazur, 1973, p. 2-3). [In addition], multiple regression analysis not only gives more information about the data, it also applicable to more kinds of data. (p. 6)

Multiple regression not only provides a way to analyze the relations of one variable with a set of variables, but it, with the stepwise method, also can be used for purposes of parsimony. Efroymson (1960) first advanced stepwise regression in an article in which he presented an algorithm which performed a true stepwise (as distinguished from FORWARD or BACKWARD methods) regression.

> An important property of the stepwise procedure is based on the facts that (a) a variable may be indicated to be significant in any early stage and thus enter the equation, and (b) after several other variables are added to the regression equation, the initial variable may be indicated to be insignificant. The insignificant variable will be removed from the regression equation before adding an additional variable. Therefore, only significant variables are included in the final regression. (p. 192)

Efroymson's (1960) article presented computer output from an example, as well as estimates of how much space and time would be needed to run problems based on the number of variables and sample size. Stepwise regression has received considerable attention in reducing the number of independent variables in the prediction equation or selecting the best subset of the variables from a set of independent variables.

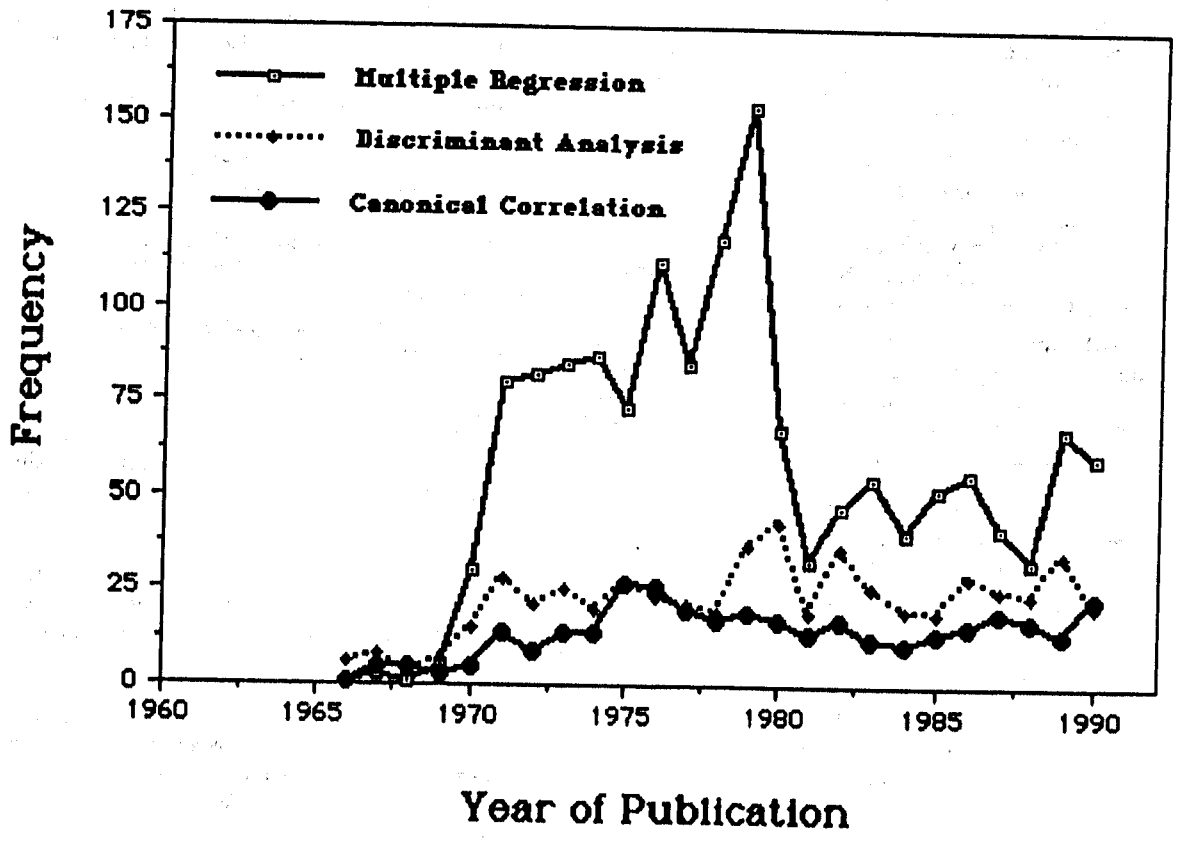
Following Cohen and Darlington's work, the 1970's saw a great increase in research on the theory as well as application of multiple regression. For example, see Heise (1969, 1970) who used multiple regression in causal relation research using social science and panel data.

As you will see in the next section, the middle of the 1970s saw the peak in the number of applications of multiple regression. Questions about assumptions being met and appropriate uses come to the forefront of researchers' use of the statistical methodology.

The Subsequent Fall

With the increasing availability of mainframe computers and programs to perform statistical analysis, journal editors were inundated with an avalanche of regression analyses. Figure 1 demonstrates the growth of multiple regression, discriminant analysis, and canonical correlation from article references by the Educational Resources Information Center (ERIC). The ERIC database consists of the Resources in Education (RIE) file of document citations and the Current Index to Journals in Education (CIJE) file of journal article citations from over 750 professional journals.
Questions were raised about whether assumptions were being met and the use of stepwise regression was strongly criticized. Attention was given to whether the regression models were correctly specified. Confusion between multiple correlation and prediction estimation began to be identified.

Figure 1 **Number of Citations of Multiple Regression, Discriminant Analysis, and Canonical Correlation in ERIC Journals from 1965 to 1991**



Violation of the Assumptions

During 1970s there were many criticisms related to the assumption of normal distribution of errors of measurement which is, in many cases, not likely to be true with variables in behavioral research.

The classical linear model $Y = \beta x + \varepsilon$ assumes that y, an N x 1 vector, is a random variable, X is an N x (k+1) matrix with fixed (not random) values', (i.e., X is matrix of known constants); β, a (k+1) x 1 vector, contains the k unknown parameters, or regression weights, plus an intercept parameter; and ε, an N x 1 vector, is a random variable. It further assumes that the errors have the properties of normality, linearity, independence, and homoscedasticity. This expression of the classical model is from Sockloff (1976), pp. 268-9.

It seemed that multiple regression does not have any requirement for the data except meeting those assumptions described above. Box (1966) alerted the mathematical community to a possible concern in treating data collected from "field research" (without controls on variables or manipulation of independent variables) in the same manner as data from "lab experiments" (with random assignment of subjects to groups).

The method of least squares is used in the analysis of data from planned experiments and also in the analysis of data from unplanned happenings. ... It is the tacit assumption that the requirements for the validity of least squares analysis are satisfied for unplanned data that produces a great deal of trouble. Whether the data are planned or unplanned the quantity ε, which is usually quickly dismissed as a random variable having the very specific properties mentioned above, really describes the effect of a large number of 'latent' variables $x_{k+1}, x_{k+2}, \ldots, x_m$, which we know nothing about. (Box, 1966, p. 625)

For the unplanned data, suppose k independent variables are input in the model, ε includes a combination of some latent variables, say, x_{k+1}, $x_{k+2}, \ldots x_m$. Therefore, the regression model contained two components:

$$Y = [\beta_0 + \beta_1 x_{1u} + \beta_2 x_{2u} + \ldots \beta_k x_{ku}] + [\beta_{k+1} x_{k+1} + \ldots \beta_m x_m]$$

$$Y = x_1 \beta_1 x_2 \beta_2$$

As an example of analysis of unplanned data, Box (1966) discussed a possible situation in industry.

> In the operation of an industrial process past experience often shows that certain variables are of major importance. In order to control fluctuations in the process, therefore, care is taken to hold precisely these variables very close to fixed values. As the "statistical significance" of any variable is greatly affected by the range it covers, there is a strong probability, therefore, that the most important variables will be dubbed "not significant" by a standard regression analysis. A further difficulty is that with unplanned data regression variables will frequently be highly correlated only because of operating policy. (p. 628)

Although presented here as a violation of the assumption of the errors being normally and identically distributed, the problem identified by Box (1966) may also be considered as misspecification of the regression model and multicollinearity resulting from unplanned data.

Some people questioned the robustness of least square estimation when the assumption(s) was(were) not met. Wainer and Thissen (1976) concluded:

> In this paper we have explored a variety of schemes for estimating coefficients of linear functions with respect to their ability to yield reasonable answers when the form of the data distribution ranges broadly. Our strongest finding is that the most commonly applied methodology, least squares estimators (LSE), are the worst performers in general. (pp. 30-31)

Earlier in this article, Wainer and Thissen discussed the assumptions in multiple regression and using equal weights (β s).

> The robustness of equal weights is beyond question, since their estimation does not involve the data at all; the shape of the sample distribution is irrelevant. Least squares estimates are another story. They are used without distributional assumptions and are identical to maximum likelihood estimates with Gaussian assumptions, provided that one assumes independence of error. If this assumption is violated the least squares estimates overestimate the betas. This is only one thing that can go wrong and is indicative of the "capitalization on chance" that has become the hallmark of least squares regression. (p. 12)

In another article advocating the use of robust regression methods, Wainer (1976) wrote:

> It is noted that the usual estimates that are optimal under a Gaussian assumption are very vulnerable to the effects of outliers. ... Normality assumptions are very useful theoretically, but have sometimes proved unrealistic in practice. (p. 285)

In a 1976 article, Sockloff noted that the assumptions under which analyses are conducted are not always specified.

> Recent works by Cohen (1968), Kelly, Beggs, McNeil, Eichelberger, and Lyon (1969), Kerlinger and Pedhazur (1973), and Bottemberg and Ward ... have attested to the flexibility of the General Linear Model. These publications have shown the capabilities of a single approach to the solution of correlation, regression, and the Fisherian analysis of variance problems. It is noteworthy that all six of these publications claim, more or less, to be using the General Linear Model, but in no case has the particular linear model and its assumptions been clearly specified and consistently applied.
>
> The General Linear Model is a name given to the family of models possessing a common characteristic, namely, linearity in the parameters of the equation specifying the model. The members of this family are distinguishable in terms of their various assumptions, and it is the contention of this author that the distinctions among these different linear models are of more than just passing interest.
>
> The above publications, plus those of Digman (1966) and of McNeil and Spaner (1971), have shown the capabilities of the General Linear Model in handling the analysis of nonlinear data....[T]he interest of this paper is to show that the analysis of nonlinearity via polynomial and product variables in a linear model has limitations far more stringent than have been realized by educational and psychological researchers. (pp. 267-268)

Sockloff (1976) distinguished between three linear models (fixed, random, and provisional) and emphasized the differences between a fixed model and a random model and the limitation of general linear model in handling nonlinear data. In the "fixed" model, the matrix X consists of "regressors that are observable and are fixed (determined a priori) values of random variables" (p. 269). In the random model, X is a matrix of regressors that are observable and random variables.

> The Random Normal Model requires the additional assumptions: (a) in the population, X and y are distributed multivariate normal, and X and ε are uncorrelated; and (b) in the sample, each multivariate observation corresponding to a row of X and y is randomly drawn. If X and y are distributed multivariate normal, the $\varepsilon = y - X\beta$ is independently distributed multivariate normal with common variance σ^2 as in the Fixed Normal Model, and X and ε are not only uncorrelated but also independent. The population to which inferences are made under the Random Normal

> Model covers the total multivariate population from which the validation sample is randomly drawn." (pp. 269-270).
>
> Kelly et al. (1969), Kerlinger and Pedhazur (1973), and Bottenberg and Ward devote most of their respective texts to multiple regression and capitalize on the similarity of computational procedures required for the solution of analysis of variance, multiple correlation, and polynomial regression problems. Whereas Bottenberg and Ward fail to specify models or assumptions, Kelly et al. and Kerlinger and Pedhazur work under an apparent Fixed Normal Model insofar as distributional assumptions are not made about the regressors. Although Kerlinger and Pedhazur never distinguish the two classical models, Kelly et al. make a distinction, but this distinction is made late in the book at which point the reader cannot easily determine the appropriate model for each of the problems presented earlier. (p. 272)

He pointed out that the computational similarity between the fixed and random models was the initial source of the confusion of the two models. He argued that "regarding the analysis of nonlinearity in observational data under the Random Model, the Random Normal Model cannot be used, and contrary to the various publications extolling the generality of the General Linear Model, the appropriate counterpart inferential model does not currently exist." (p. 288)

Multicollinearity

Statistical analysts using multiple regression have known for some time about the problems caused by intercorrelations among the independent variables. High intercorrelations among the predictors, but not complete linear dependency, has been called "collinearity" or "ill conditioning" of the correlation matrix, or for the purposes of this paper, "multicollinearity". Gordon (1969) alerted us to the potential problems:

> Although the warnings concerning multicollinearity are to be found in statistics texts, they are insufficiently informative to prevent the mistakes described here. This is because the problem is essentially one of substantive interpretation rather than one of mathematical statistics per se. (p. 592)

The effects of multicollinearity on the least squares estimates of the regression coefficients were pointed by Johnstone in 1972 as follows:

> 1. The precision of estimation falls so that it becomes very difficult, if not impossible, to disentangle the relative influence of various x variables. This loss of precision has three aspects; Specific estimates may have very large errors; these error may be highly correlated, one with another; and the sampling variances of the coefficients will be very large.
>
> 2. Investigators are sometimes led to drop variables incorrectly from an analysis because their coefficients are not significantly different from zero, but the true situation may be not that a variable has no effect but simply that the set of sample data has not enabled us to pick it up.
>
> 3. Estimates of coefficients become very sensitive to particular sets of sample data, and the addition of a few more observations can sometimes produce dramatic shifts in some of the coefficients. (p. 160)

Gordon (1969) concluded:

> ...[W]e have not been condemning the method of multiple regression in general. There remain many situations in sociology for which regression is an excellent tool of analysis. We do condemn, however, those applications of regression coefficients that seek to determine the relative importance of variables in the manner of the examples we have cited. (pp. 615-6)

Abuse of Stepwise Regression

One of the most common uses of regression has been model-building automatically, that is, determining the relative importance of variables by the order in which they are entered (or deleted) to find the "best" regression model. Pope and Webster (1972) pointed out that:

> The methods generally known as stepwise procedures are, however, the most widely used data analysis methods; in particular by non-professional statisticians. This has come about through the availability of computer programs.
>
> This paper was stimulated by this widespread use of the stepwise procedures and the lack of understanding (by the non-statistician) of their weaknesses. (p. 328)

Huberty (1989) listed three intended uses of stepwise regression.

> Stepwise analyses have basically been used for three purposes: (1) selection or deletion of variables, (2) assessing relative variable importance; or (3) both variable selection and variable ordering. (p. 45)

Stepwise regression has been commonly used for selecting the best subset for any specified number of retained independent variables. Among a total of $k(k+1)/2$ fits, "as observed by Gorman and Toman (1966), it is unlikely that there is a single best subset but rather several equally good ones" (Hocking, 1960, p. 9). Mantel (1970) criticized forward selection by illustrating a situation in which an excellent model would be overlooked because of the restriction of adding only one variable at a time and pointed out the disadvantage of forward selection needs $k(k+1)/2$ fits k where backward elimination only needs k fits for testing

among k variables. Hocking (1960) also expressed concern about the limited number of solutions for the "best" regression equation.

> Another criticism of FS [forward selection] and BE [backward elimination] often cited is that they imply an order of importance to the variables. This can be misleading since, for example, it is not uncommon to find that the first variable included in FS is quite unnecessary in the presence of other variables.... The lack of satisfaction of any reasonable optimality criterion by the subsets revealed by stepwise methods, although a valid criticism, may not be as serious a deficiency as the fact that typical computer routines usually reveal only one subset of a given size. (p. 9)

Pope and Webster (1972) pointed out the "pseudoness of the F-statistic" for testing the significance of independent variables in linear prediction equation" (p. 327). "Unfortunately, the most widely used computer programs print this statistic at each step without any warning that it does not have the F distribution under automated stepwise selection" (Wilkinson, 1979, p. 168). Using a Monte Carlo simulation, Wilkinson (1979) constructed the tables of the upper 95th and 99th percentage points of the sample R^2 distribution in forward selection. He examined 71 articles published in psychology from 1969 to 1977 which used stepwise regression.

> Out of these articles 66 forward selection analyses reported as significant by the usual F test were found. Of these 66 analyses, 19 were not significant [using Wilkinson table].. . (p. 172)

The severe consequences of abuse of stepwise regression were emphasized by Thompson in a 1989 editorial entitled, "Why Won't Stepwise Methods Die?"

> First, most researchers, thanks to "canned" computer programs, do not employ the correct degrees of freedom when evaluating changes in explained variance (i.e., usually changes in squared R or lambda). ... Second, some researchers incorrectly interpret stepwise results in which *q* predictor variables have been selected as indicating that the predictor variables are the best variables to use if the predictor variable set is limited to size *q*. ... Third, some researchers incorrectly consult order of entry information to evaluate the importance of various predictor variables." (pp. 146-147)

In one of the most serious and thorough critiques of stepwise regression, Huberty (1989) postulated that:

> (1) ..stepwise analysis should not generally be used for variable selection purposes. A basic detect of stepwise procedures is attributable to 'their consideration of variables one-at-time... direct tests for the additional information supplied jointly by several variables are not made' (McKay & Campbell, 1982, pp. 13, 45)
>
> (2)...order of variable entry in a stepwise analysis should not be used to assess relative variable contribution/importance." because "the inter-relationship of the response variables are completely ignored when the most 'important' [first variable entered] is determined... and the dependence [of following variable on preceding variable] or conditionality truly makes variable importance as determined by stepwise analysis very question". (pp. 46-47)

Kachigan (1986), warned researchers that sampling error can seriously distort stepwise results.

> There is a danger that we might selected variables for inclusion in the regression equation based on chance relationship. Therefore, as stressed in our discussion of multiple correlation, we should apply our chosen regression equation to a fresh sample of objects to see how well it does in fact predict values on the criterion variable. This validation procedure is absolutely essential if we are to have any faith at all in the future applications of the regression equation. (p. 265)

We will see in the Second Rise section that Huberty proposed alternative methods to address these problems.

Misspecification of Regression Models

Included in our definition of misspecification of regression models are specification errors by using the "wrong" independent variables as well as expressing the wrong relationship among the independent variables or the relationship between the independent variables with the dependent variable. This first type was identified in 1971 by Borhnstedt and Carter.

> When one has mistakenly either omitted or included variables in an equation assumed to capture the true causal structure to Y, or when the functional form chosen to represent the variables is incorrect, we say that one has made a specification error. (p. 128)

The second type would include following: (a) specifying a linear model though a nonlinear model is more appropriate, (b) postulating an additive model even though a nonadditive model is more appropriate, and (c) applying a linear additive model when a nonlinear or nonadditive one is called for (Pedhazur, 1982, pp. 225-229).

When any of the assumptions are violated or when the stepwise regression technique is not correctly used, misspecification of the regression model is an inevitable outcome. However, researchers often ignore such errors.

Gordon (1969) contended that the theoretical context of research should determine the nature of importance of the variables controlled. Since R^2 was the most often used criterion to judging prediction models and (partial) regression coefficients were often

used as indicators of the relative importance of variables, Gordon (1969) showed the interrelationship of the multicollinearity and misspecification problems. He provided the examples showing that:

> [S]mall variation among the correlations of a highly related set can be create large variations among their regression coefficients" (p. 612). In addition "the values of regression coefficients are not immutable and that they can be greatly affected by changes in the selection of independent variables to be included in an analysis" (p. 613). He warned us that "multiple regression is not an all-purpose methods for data reduction" (p. 163) and emphasized going "beyond simple examination of the regression coefficients". (p. 615)

Bohrnstedt and Carter (1971) discussed the effect of specification errors:

> specification errors can seriously affect our estimates of the true structural parameters operating in the system. ... if we hypothesize the wrong model, then our estimation of that model will yield meaningless estimates. (p. 141)

They concluded that "we can only come to the sobering conclusion, then, that many of the published results based on regression analysis... are possible distortions of whatever reality may exist" (p. 143).

Confusion Between Multiple Correlation and Prediction Estimation

The prediction model and the correlation model were seldom to be distinguished. Huberty and Mourad (1980) emphasized the difference of the parameters estimated in the multiple correlation and prediction estimation.

> All of the statistical techniques associated with the prediction model are applicable with the correlation model. However, from a correlation estimation viewpoint, different parameters are associated with the two models. With the correlation model, the population multiple correlation coefficient of interest is ρ, which reflects the correlation between Y and the optimal linear composite of $X_1, X_2, \ldots X_p$ in the population as a whole. The optimal linear composite is that composite determined so as to maximize this correlation in the population. With the prediction model, the population multiple correlation coefficient of interest is ρ_v which reflects the correlation between Y and the linear composite of the X's which is optimal for the calibration sample. With each calibration sample is associated a ρ_v, which is a type of validity coefficient. Values of ρ_v are coefficients of correlation between a criterion Y and a linear composite of the predictors, the weights of which will vary across repeated sampling. (p. 102)

They also criticized the deficiencies in reporting estimates of correlation coefficient in the literature and the inflated predictive validity of the studies, overestimation of the parameter ρ for prediction using R_w and R_e. They discuss two estimation procedures for the parameters ρ and ρ_v: cross-validation and usage of a "shrinkage" formula.

The Second Rise

In this period which these authors call the 'second rise', comparatively new techniques are recognized for handling the problems identified during the period of "the fall." Some of those techniques are robust regression, ridge regression and nonlinear regression. These methods were introduced to behavioral scientists in the late 70's and early 80's. Also, new methods using multiple and/or categorical dependent variables, such as canonical correlation and discriminant analysis, have been popularized.

Nonlinear Regression

When the assumption of linearity is violated, an appropriate nonlinear regression model should be considered. Since regression weights in nonlinear regression equations can be changed by changing the means of the independent variables, and the means are often chosen arbitrarily, the coefficients of nonlinear regression models can not be interpreted causally. A general solution to the importance of each independent variable in the linear and nonlinear models was attempted by Darlington and Rom (1972). For the sake of the difficulty of the interpretation of the nonlinear regression model, the effects on the transformation of polynomial regression equations into a format that is readily interpretable were made.

Robust Regression

In 1976, Howard Wainer wrote an article published in *Psychological Bulletin* entitled "Estimating Coefficients in Linear Models: It Don't Make No Nevermind." In his article, he stated:

> It is proved that under very general circumstances coefficients in multiple regression models can be replaced with equal weights with almost no loss in accuracy on the original data sample. It is then shown that these equal weighs will have greater robustness than least squares regression coefficients. (p 213)

The general conditions given are "all predictor variables should be oriented properly" and "the predictor variables should be intercorrelated positively" (Wainer, 1976, p 213).

Wainer's approach essentially ignores the sample data. A less radical solution to the problems with ordinary least squares solutions (OLS) to the estimate of parameters in multiple regression in light of non-normality or outlier problems has been addressed by Huyhn (1982), who referenced the sources of the alternatives for handling outliers and explained the concept and functions of Least Absolute Residual (LAR), first introduced by Gentle (1977):

> LAR estimates are the maximum-likelihood

> estimates when the errors follow a double exponential structure. Because large residuals are given smaller weights in LAR estimation than in OLS [ordinary least squares] estimation, LAR estimates are less influenced than OLS estimates by those residuals. (Huyhn, 1982, p. 506)

Huyhn reviewed each of the four robust regression techniques provided by Huber (M-estimate), Hample (psi function), Andrew (sine estimate) and Tukey (biweight estimate), respectively, provided an example of using these four robustness regression methods, and compared them with the results from employing the ordinary least square method. The reader should refer to Hogg (1979) for a discussion of the last four estimators. Huyhn (1982) summarized the conclusions about robust regression against OLS.

> First, if the data do not contain any outlying observations, then OLS and robust regressions provide estimates that do not differ markedly from each other. Second, for data with suspected or abnormal observations, OLS estimates may differ substantially from the robust estimates; third, observations considered as outliers by OLS regression may not be outliers at all under robust regressions. Fourth, robust regression procedures, as proposed by Hampel, Andrews, or Tukey, may be able to detect outliers automatically by giving each one a weight that is zero or very small as compared with other weights. (p. 511)

He re-emphasized the recommendations provided by Hogg (1979).

> Perform the usual OLS analysis along with a robust procedure such as that used by Andrews. If the resulting estimates are in essential agreement, report the OLS estimates and relevant statistics. If substantial differences occur, however, take a careful look at the observations with large robust residuals and check to determine whether they contain errors of any or if they represent significant situations under which the postulated regression model is not appropriate. (pp. 511-512)

Ridge Regression

Knowledge of the potential problems caused by multicollinearity has alerted researchers to avoid misinterpretations. Many alternatives have been proposed. A researcher might first try to eliminate the variables that contribute to the high degree of multicollinearity. However, we should not have considered a logically redundant variable initially. Removal of any one variable may lead to misspecification of the model. Pedhazur (1982) noted other remedies:

> One of the proposed remedies is the collection of additional data in the hope that this may ameliorate the condition of high multicollinearity. Another set of remedies relates to the grouping of variables either in blocks on the basis of a priori judgements or by the use of such methods as principal components analysis and factor analysis.... Another set of proposals...is to abandon Ordinary Least-Squares analysis and use instead other methods of estimation. One such method that has been gaining in popularity is Ridge Regression.... [N]one of the proposed methods of dealing with high multicollinearity constitutes a cure. High multicollinearity is symptomatic of insufficient, or deficient, information, which no amount of data manipulation can rectify. (p. 247)

Reduced variance regression, as a compromise between ordinary regression and some other techniques such as weighted least squares, was advocated for its potential solution of dealing with problems of multicollinearity, ratio of number of predictors to sample size, as well as validity issues. Ridge regression, introduced by Hoerl and Kennard in 1970, is an application of reduced variance regression. "Ridge regression is a controversial procedure that attempts to stabilize estimates of regression coefficients by inflating the variance that is analyzed" (Tabachnick & Fidell, 1989, p. 130).

In late 70's and early 80's, ridge regression was reemphasized in the psychology and social sciences. For example, Price (1977) and Darlington and Boyce (1982) highlighted the function of ridge regression in exploring and extracting information from multifactor data. Price (1977) gave an example of how to use ridge regression, introduced the criterion of choosing a value of k (see below) from inspection of the ridge trace, and emphasized the nature of ridge regression in reducing total mean square error by introducing some degree of bias.

Darlington and Boyce (1982) also provided the behavioral scientist with a very comprehensible explanation about ridge regression using the concept of regression to the mean.

> It is well known that estimates for many independent parameter values can be improved by regressing the unbiased estimates of those values toward the grand mean of all the values. ... If the investigator assumes that on the average, each observed correlation exceeds the true value by a proportion k, then the ratio between average observed and true values is (1+k) / 1. ... Ridge regression essentially consists of adjusting all the correlations in the matrix (both the X - X and the X - Y correlations) by this factor 1/(1+k), and then deriving regression weights in the ordinary way. ... Thus adjustment of the X - X correlations produces the largest increases in apparent independence (and hence increases in beta weights) for those regressors which correlate most highly with the other regressors. This is how ridge regression takes advantage of validity concentration -- regressors correlating

> highly with the total set of regressors are upgraded in importance relative to the others. (pp. 84 - 85)

They informed researchers that about a dozen formulae for estimating k have been proposed and the ridge trace was no longer recommended by the statisticians. The alternative for estimating k, an iteration procedure was introduced in this paper. They also provided recommendations about when ridge regression should be used.

Alternatives to Stepwise Regression

Concerning the possible distorted results from careless use of stepwise regression, many researchers tried to find better alternatives to stepwise regression. Huberty (1989) provided the alternative approaches and suggested that "a 'natural' criterion to use to determine the best subset size in the context of prediction and estimation is to minimize the residual sum-of-squares value" (p. 50). For selecting the variables from a set of initial variables, SAS PROC RSQUARE (SAS Institute, Inc., 1990) procedure was recommended to assess 2^{p-1} equations, where p is the number of predictors (Huberty, 1989, p. 50). For determining the final subset size of the independent variables, Huberty (1989) recommended adjusted R^2 or scree test --- "plot[ing] the adjusted R^2 values for the 'best' subset of each size (determined by the researcher using information from computer output plus sound judgment) against subset size" (p. 51).

Thompson (1989) proposed that a possible alternative to the misleading results of stepwise regression would be to "employ a cross-validation procedure such as one recommended by Huck, Cormier, and Bounds (1974, p. 159)". Huck, Cormier and Bounds (1974) proposed a four-step method.

> (1) The original group of people (for whom both predictor and criterion scores are available) is randomly divided into two subgroups. (2) Just one of the subgroups is used to develop the prediction equation.. (3) The equation is used to predict a criterion score for each person in the second subgroup, i.e., the subgroup that was not used to develop the prediction equation). (4) The predicted criterion scores for people in the second subgroup are correlated with their actual criterion scores. A high correlation (that is significantly different from zero) means that the prediction equation works for people other than those who were used to develop the equation. If the individuals in future studies are not too much different from those in the cross-validation procedure, the researcher is justified in using the prediction equation for groups other than the original. (pp. 159-160)

Henderson and Velleman (1981) illustrated the superiority of substantively guided data analysis over automatic model building. "Automated multiple regression model-building techniques often hide important aspects of data from the data analyst. Such feature as nonlinearity, collinearity, outliers, and points with high leverage can profoundly affect automated analyses, yet remain undetected." Henderson and Velleman (1981) proposed an alternate method integrating "interactive computing and exploratory methods to discover unexpected features of the data." (p 391). They illustrated their alternative method using two examples, one from Hocking (1973) involving variables on 32 automobiles and a second example on air pollution and mortality from McDonald and Schwing (1973).

Henderson and Velleman (1981) stated a fundamental axiom of their philosophy of data analysis "The data analyst knows more than the computer" (p. 391).

Checking for the Assumptions

Following the concern for possible violation of assumptions, methods to check for whether assumptions were tenable or not were developed using computer programs. Some of these methods were nicely summarized in a paper by Elmore, Woehlke, and Spearing (1990). They also compared the procedures available in SAS and SPSS^X. Leitner (1990) provided examples of how multicollinearity among independent variables can be detected using the SAS and SPSSX computer packages, and recommended procedures for reducing the extent of multicollinearity. In addition, Pohlmann (1990) presented some methods using SAS (version 6) check for outliers.

Multivariate Technique

Although it was originally developed in the 30's (Hotelling, 1935), canonical correlation was not realized as the most general case of the general linear model until the late 70's or early 80's.

> ...Baggaley (1981) has noted that canonical correlation analysis, and not regression analysis, is the most general case of the general linear model. Knapp (1978) demonstrated this in detail and concluded that "virtually all of the commonly encounter parametric tests of significance can be treated as special cases of canonical correlation analysis, which is the general procedure for investigating the relationships between two sets of variables." In a similar vein Fornell (1978) notes, "Multiple regression, MANOVA and ANOVA, and multiple discriminant analysis can all be shown to be special cases of canonical analysis...." (Thompson, 1984)

Extended from a single dependent variable in the model to multiple dependent variables, canonical correlation could be used at least to predict or explain a set of dependent variables by a set of independent variables. When the dependent variables are categorical, the procedure is called discriminant analysis. The roles of discriminant analysis include that separation, discrimination, and estimation of the populations of objects (Huberty, 1975). Since a great deal of research in the behavioral sciences involves these three aspects, discriminant analysis has been considered as, follow-up technique to MANOVA, one of the most significant

development in multivariate analysis.

Conclusion

While this journey through the literature was not exhaustive (although it may have been tiring to many readers) and strictly chronological, the authors feel that a similar trend of introduction, questioning, and resolution of the problems for the statistical technique of multiple regression existed as with t-test, factor analysis and meta-analysis. Perhaps other statistical procedures could similarly be documented.

References

Baggaley, A. R. (1981). Multivariate analysis: An introduction for consumers of behavioral research. *Evaluation Review, 5*, 123-131.

Bohrnstadt, R.A., & Carter, T.M. (1971). Robustness in regression analysis. In H.L. Coster, (Ed.) *Sociological methodology* . San Francisco: Jossey-Bass.

Box, G. E. P. (1966). Use and abuse of regression. *Technometrics, 8*, 625-629.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70*, 426-443.

Darlington, R. B., & Boyce, C. M. (1982). Ridge and other new varieties of regression. In G. Keren (Ed.), *Statistical And Methodological Issues In Psychology And Social Science Research*.. Hillsdale, NJ: Erlbaum.

Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin, 69*, 161-182.

Darlington, R. B., & Rom, J. F. (1972). Assessing the importance of independent variables. *American Educational Research Journal, 9*, 449-462.

Elmore, P., Woehlke, P., & Spearing, D. (1990, April) *Testing assumptions in multiple regression: a comparison of procedures available in SAS and SPSSX*. Paper presented at the Annual meeting of the American Educational Research Association, Boston, MA.

Efroymson, M. A. (1960). Multiple regression anlysis. In A. Ralston & H. S. Wilf (Eds.), *Mathematical Methods for Digital Computers*, New York: Wiley.

Fornell, C. (1978). Three approaches to canonical analysis. *Journal of the Market Research Society, 20*, 166-181.

Gentle, J. E. (1977). Least absolute values estimation: An introduction. *Communications in Statistics, B6*, 313-328.

Goldberger, A. S. (1964). *Econometric theory*. New York: Wiley.

Gordon, R. A. (1969). Issues in multiple regression. *American Journal of Sociology, 73*, 592-616.

Heise, D. R. (1970). Causal inference from panel data. In E. F. Borgatta, & G. W. Bohrnstedt (Eds.), *Sociological methodology: 1970*, pp. 3-27. San Francisco: Jossey-Bass.

Heise, D. R. (1969). Problems in path analysis and causal inference. In E. F. Borgatta & G. W. Bohrnstedt (Eds.), *Sociological Methodology: 1969*, (pp. 38-73). San Francisco: Jossey-Bass.

Henderson, H., & Velleman, P. F. (1981). Building multiple regression models interactively. *Biometrics, 37*, 391-411.

Hocking, R. R. (1960). Selection of the best subset of regression variables. In K. Enslein, A. Ralston, & H. S. Wilf (Eds.), *Statistical Methods For Digital Computers*. Vol. 3, (pp. 39-57). New York: Wiley.

Hogg, R. V. (1979). Statistical robustness: One view of its use in application today. *The American Statistician, 33*, 108-115.

Hotelling, H. (1935). The most predictable criterion. *Journal of Experimental Psychology, 26*, 139-142.

Huber, P. J. (1973). Robust regression: Asymptotics, conjecture and Monte Carlo. *Annals of Statistics, 1*, 799-821.

Huberty, C. J. (1975). Discriminant analysis. *Review of Educational Research, 45*, 543-598.

Huberty, C. J. (1989). Problems with stepwise methods–better alternatives. In B. Thompson (Ed.), *Advances In Social Science Methodology*. Vol. 1, (pp. 43-70). Greenwick, CT: JAI Press.

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

Huynh, H. (1982). A comparison of four approaches to robust regression. *Psychological Bulletin, 92*, 505-512.

Johnstone, J. (1972). *Econometric methods (2nd ed.)*. New York: McGraw-Hill.

Kachigan, S. K. (1986). Statistical analysis: An interdisciplinary introduction to univariate and multivariate methods (2nd ed.). New York: Holt, Rinehart and Winston.

Kelly, F. J., Beggs, D. L., McNeil, K. A., Eichelberger, T., & Lyon, J. (1969). *Multiple regression approach*. Carbondale, IL : Southern Illinois University Press.

Kerlinger, F., & Pedhazur, E. J. (1973). *Multiple regression in behavioral research*.. New York: Holt, Rinehart and Winston.

Knapp, T. R. (1978). Canonical correlation anlysis: A general parametric significance-testing system. *Psychological Bulletin, 85*, 410-416.

Leitner, D. W. (1990, April). Detecting multicollinearity in multiple regression: A comparison of procedures in SPSSX and SAS. Paper presented at the Annual meeting of the American Educational Research Association. Boston, MA.

Leitner, D. W. (1990). The rise and fall and rise of selected statistical procedures. *MWERA Researcher, 3*(1), 8-17.

Mantel, N. (1970). Why stepdown procedures in variable selection. *Technometrics, 12*, 621-625.

McDonald, G. C., & Schjwing, R. C. (1973). Instabilities of regession estimates relating air pollution to mortality. *Technometrics*, 463-481.

McNeil, K. A., & Spaner, S. D. (1971). Highly correlated predictor variables in multiple regression models. *Multivariate Behavioral Research, 6*, 117-125.

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

Pohlmann, J., (1990, April). Case influence statistics available in SAS (Version 5). Paper presented at the Annual meeting of the American Educational Research Association, Bosten, MA.

Pope, P. T., & Webster, J. T. (1972). The use of an F-statistic in stepwise regression procedure. *Technometrics, 14*, 327-340.

Price, B. (1977). Ridge regression: application to nonexperimental data. *Psychological Bulletin, 84*, 759-766.

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

Sockloff, A. L. (1976). The analysis of nonlinearity via linear regression with polyomial and product variables: An examination. *Review of Educational*

Research, 46, 267-291.

Tabachnick, B. G., & Fidell, L. S. (1989). *Using multivariate statistics (2nd ed.)*. New York: Harper & Row.

Thompson, B. (1984). Canconical correlation analysis: Use and interpretation. Beverly Hills, CA: Sage.

Thompson, B. (1989). Why won't stepwise methods die? *Measurement and Evaluation in Counseling and Development, 21*, 146-148.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83*, 213-217.

Wainer, H., & Thissen, D. (1976). Three steps towards robust regression. *Psychometrika, 41*, 9-34.

Wilkinson, L. (1979). Tests of significance in stepwise regression. *Psychological Bulletin, 86*, 168-174.

A Comparison of the Mallows Cp and Principal Component Regression Criteria for Best Model Selection in Multiple Regression

Randall E. Schumaker
University of North Texas

A cross validation comparison of the Mallows Cp subset model selection criteria using randomly generated data sets indicated that different subset models may be identified. The principal component regression method using Type II sum of squares with orthogonal principal component variables indicated a slightly different set of "best" variables. The two methods in the presence of multicollinearity can yield different subset models. It is recommended that researchers base regression models on substantive theory, model validation, and effect sizes for proper model testing and interpretation.

Multiple regression permits model testing wherein a set of independent variables are hypothesized to predict a dependent variable. Often when the set of variables selected does not significantly predict, the researcher searches for a "subset" of variables that provides the best prediction model. The statistical packages provide several stepwise methods for this purpose.

A review of the literature, however, indicates that most researchers misuse stepwise methods in determining the best predictor set or interpreting the importance of predictor variables (Huberty, 1989; Snyder, 1991; Thompson, 1989; Thompson, Smith, Miller, & Thomson, 1991; Welge, 1990). Tracz, Brown, and Kopriva (1991) summarized much of the literature to indicate that the results of stepwise procedures do not yield a "best" equation because different criteria can be used in the selection of different sets of variables; that when variables are intercorrelated, there is no satisfactory way to determine the relative contribution of the variables to R-squared because various subsets of variables could yield a similar R-squared value; that stepwise methods inflate Type I error rates by not using the correct degrees of freedom in calculating the change in R-squared; and that the order of variable entry is incorrectly interpreted as defining the importance of the variable or "best set" of predictors.

Current research literature indicates that the all possible subset approach is preferred over the stepwise methods for determining the best model (Berk, 1977; Cummings, 1982; Thayer, 1986; Davidson, 1988; Henderson & Denison, 1989; Welge, 1990; Thayer, 1990; Tracz, Brown, & Kopriva, 1991). Several criteria, however, are available for selecting the best subset model when using the all possible subset approach: R-squared, adjusted R-squared, mean squared error, Mallow's C_p, or a principal component regression. Constas and Francis (1992) presented a graphical method for selecting the best subset regression model using R-squared and adjusted R-squared. They plotted R-squared and adjusted R-squared against the number of predictors in the model. The maximum number of predictors for best subset model was determined at the point where the R-squared and/or the adjusted R-squared values descended.

The Mallows Cp criteria has also been recommended for selecting the best subset of predictor variables in contrast to the stepwise methods using a sample data set (Tracz, Brown, & Kopriva, 1991; Zuccaro, 1992). The Cp statistic measures the effect of underfitting (important predictors left out of the model) or overfitting (include predictors that make no contribution or are marginal). Mallows (1966; 1973) has suggested that the selection of the best subset model with the lowest bias is indicated by the smallest Mallows Cp criteria, especially in the presence of multicollinearity. The SAS package (Freund & Littell, 1991) currently prints the Mallows Cp value and a variance inflation factor (VIF) which can be used to determine which variables may be involved in the multicollinearity. Pohlmann (1983) had previously noted that multicollinearity among a set of predictor variables didn't affect the Type I error rate, but did affect the Type II error rate and width of the confidence interval. His findings suggest that sample size and model validity could compensate for multicollinearity effects, especially when certain research questions

require models with highly correlated predictors, for example, $Y = \beta_1 X_1 + \beta_2 X^2_1 + e$.

The principal component regression (PCR) has also been proposed as a criteria for selecting the best predictor model. This method appears to be useful when predicting values in one sample based upon estimates from another sample and when multicollinearity exists among a set of variables (Morrison, 1976). The indication for using a PCR approach is when the mean squared error of a biased estimate is smaller than the variance of an unbiased estimate. The PCR method, however, is not appropriate for multiple regression subset models containing interactions (Aiken & West, 1993). Since the PCR method creates a set of new variables called principal components, which are uncorrelated or orthogonal, it should not be used when models depict nonlinear, correlated predicter sets.

In summary, the all possible subset approach is recommended as an alternative over stepwise methods for selecting the best set of predictor variables. The Mallows Cp criteria or a principal components regression approach is advocated for determining the best subset model over the use of R-squared, especially when the predictors are correlated. The principal component regression method, which determines the best model for prediction by creating orthogonal variables, appears more useful when estimates from one sample are used to predict in another sample or when multicollinearity exists among the predictors.

How do these criteria compare when selecting the best subset model? When might a researcher choose one criteria over another for selecting the best model? A comparison of the Mallows Cp selection criteria upon cross validation and a comparison of the parameter estimates and standard errors between the multiple regression and the PCR approach should shed further light on their usefulness for subset model selection. An applied example will further elaborate the comparison of the two criteria.

Simulation

A SAS program was used to generate a heuristic population (n = 10,000 observations) with a dependent variable and ten correlated predictor variables. The program then randomly sampled the population data set for n = 200 observations. This data set was then randomly divided to create two separate data sets of equal size ($n_1 = n_2 = 100$ observations). The SAS programs used in this simulation are available from the author.

The population correlation matrix, variable means and standard deviations are in Table 1. The correlation matrix, variable means and standard deviations for the sample data set used to compute the parameter estimates are in Table 2. The correlation matrix, variable means and standard deviations for the cross validation data set are in Table 3. Parameter estimates, computed using the ordinary least squares criterion from the first data set, were used with the second data set to calculate R^2 and the Mallows Cp values, and to determine the best variable subset models.

Table 1 Population Correlation Matrix, Means, and Standard Deviations (n = 10,000)

	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X1	.44										
X2	.25	.10									
X3	.34	.13	.10								
X4	.43	.19	.10	.15							
X5	.42	.19	.11	.13	.19						
X6	.30	.13	.09	.11	.13	.12					
X7	.24	.11	.07	.06	.10	.08	.07				
X8	.50	.22	.13	.17	.21	.21	.16	.11			
X9	.28	.12	.08	.10	.12	.11	.09	.07	.15		
X10	.26	.11	.05	.07	.11	.12	.06	.08	.14	.08	
Mean	9.99	17.92	16.12	18.94	21.96	28.05	25.97	38.90	42.05	33.97	12.05
S.D.	2.00	4.44	8.21	6.00	4.66	4.95	6.61	8.61	4.12	6.95	8.12

Note. All values have been rounded to two decimal places.

Table 2 Sample Correlation Matrix, Means, and Standard Deviations for Estimation Sample (n_1 = 100)

	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X1	.41										
X2	.28	.02									
X3	.41	.05	.23								
X4	.38	.23	.01	.15							
X5	.24	-.01	.04	.02	.16						
X6	.33	.02	.16	.09	.08	.08					
X7	.25	.16	.08	.03	.01	.01	.10				
X8	.39	.22	.13	-.04	.19	.06	.21	.01			
X9	.33	.19	.07	.04	.24	-.15	.03	.22	.21		
X10	.46	.23	.08	.24	.21	.03	.10	.17	.11	.17	
Mean	10.18	18.40	15.37	20.49	22.76	28.41	25.88	39.55	41.89	34.27	11.04
S.D.	1.80	4.61	8.88	5.94	4.30	4.99	6.79	7.81	4.13	6.80	8.13

Note. All values have been rounded to two decimal places.

Table 3 Sample Correlation Matrix, Means, and Standard Deviations for Cross Validation Sample (n_2 = 100)

	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X1	.39										
X2	.28	.14									
X3	.34	-.05	-.08								
X4	.52	.03	.13	.20							
X5	.54	.17	.20	.28	.37						
X6	.26	.01	.01	.07	.18	.19					
X7	.14	.03	.05	.08	.07	.01	-.03				
X8	.55	.27	.11	.26	.26	.21	.06	.02			
X9	.32	.26	.18	-.09	.20	.07	.11	.09	.09		
X10	.31	.26	.07	.11	.12	.21	.11	.19	.09	.24	
Mean	9.94	17.91	16.55	19.26	21.37	28.40	25.34	39.23	41.92	33.93	10.38
S.D.	1.99	4.86	8.57	6.13	5.35	4.75	6.82	9.43	4.27	6.73	7.78

Note. All values have been rounded to two decimal places.

Table 4 R^2 and Cp Values for Sample$_1$ And Sample$_2$ Best Variable Subset Models ($n_1 = n_2 = 100$)

Subset Size	Variables in Subset Model	Sample$_1$ R^2	Sample$_1$ Cp
1	(10)	.21	102.92
2	(3),(8)	.33	74.44
3	(3),(8),(10)	.44	49.79
4	(1),(3),(8),(10)	.50	36.13
5	(1),(3),(6),(8),(10)	.54	27.42
6	(1),(3),(5),(8),(9),(10)	.58	19.74
7	(1),(3),(5),(6),(8),(9),(10)	.62	12.26
8	(1),(2),(3),(5),(6),(8),(9),(10)	.63	11.85
9	(1),(2),(3),(4),(5),(6),(8),(9),(10)	.64	11.27
10	(1),(2),(3),(4),(5),(6),(7),(8),(9),(10)	.65	11.00

		Sample$_2$ R^2	Sample$_2$ Cp
1	(8)	.30	101.79
2	(5),(8)	.49	50.44
3	(4),(5),(8)	.55	33.41
4	(1),(4),(5),(8)	.61	21.34
5	(1),(4),(5),(8),(9)	.63	17.05
6	(1),(3),(4),(5),(8),(9)	.65	13.37
7	(1),(3),(4),(5),(6),(8),(9)	.66	11.58
8	(1),(2),(3),(4),(5),(6),(8),(9)	.67	9.79
9	(1),(2),(3),(4),(5),(6),(7),(8),(9)	.68	9.96
10	(1),(2),(3),(4),(5),(6),(7),(8),(9),(10)	.68	11.00

Table 5 Cross Validation Comparison of R^2 and Cp Values: Sample$_1$ to Sample$_2$ for Best Variable Subset Models ($n_1 = n_2 = 100$)

Subset Size	Variables in Subset Model	Sample$_1$ Cp	Sample$_1$ R^2	Sample$_2$ Cp	Sample$_2$ R^2
1	(10)	.21	102.92	.15	159.53
2	(3),(8)	.33	74.44	.36	92.08
3	(3),(8),(10)	.44	49.79	.40	77.64
4	(1),(3),(8),(10)	.50	36.13	.45	65.44
5	(1),(3),(6),(8),(10)	.54	27.42	.47	55.78
6	(1),(3),(5),(8),(9),(10)	.58	19.74	.59	26.35
7	(1),(3),(5),(6),(8),(9),(10)	.62	12.26	.61	23.38
8	(1),(2),(3),(5),(6),(8),(9),(10)	.63	11.85	.62	20.82
9	(1),(2),(3),(4),(5),(6),(8),(9),(10)	.64	11.27	.63	10.34
10	(1),(2),(3),(4),(5),(6),(7),(8),(9),(10)	.65	11.00	.66	11.00

Table 4 indicates the model subset selection for each sample data set. Table 5 indicates a comparison between the R^2 and Mallows Cp values from the estimation sample data set to the cross validation sample data set using parameter estimates from the estimation sample. The Mallows Cp values were inflated because the parameter estimates applied to the second data set altered the residual sums of squares used in the formula to calculate them. Although the relative ordering of Cp values were the same, these values did not indicate the same single best variable subset model in the second data set.

Table 6 compares the parameter estimates using the Mallows Cp and the principal components regression method for each best variable subset model. The R^2 values will be the same regardless of which method is used. The real difference is seen when comparing the relative significance of the parameter estimates. The Mallows Cp method with correlated predictors indicated that all the parameter estimates were significant. This was not the case in the principal components regression approach. An applied example will further illustrate this distinction between the two methods.

Applied Example

Subjects

Participants in the study were a cohort of students accepted into the Texas Academy of Mathematics and Science (TAMS) at the University of North Texas in Fall, 1993. TAMS is an early college entrance program in which students earn approximately 60 hours of college credit by taking University of North Texas courses. Students enter TAMS at the beginning of their 11th year in high school. They live on campus in a special residence hall and take regular university courses in mathematics, science and the humanities. After two years, participants receive a special high school diploma and have amassed at least 60 hours of college credit. Each year approximately 200 high school sophomores, who have met the selection criteria and completed the 10th grade, are accepted into the Texas Academy of Mathematics and Science.

In the study year, TAMS accepted 204 students. Of these, 156 students attended an August orientation, which occurred a week prior to their first semester of college coursework, and completed the LASSI. There were 80 females and 76 males who participated in the study. The students who took the LASSI were similar in demographic background and academic ability as previous classes because of the academy's consistent admission requirements and pool of applicants. The participants' SAT-M and SAT-V means and standard deviations, respectively, were: M=651, SD=57; and M=530, SD=75.

Instrument

The LASSI is an English language assessment tool designed to measure college students' use of learning and study strategies. It was designed to provide assessment and pre-post achievement measures for students participating in a learning strategies and study skills project. A high-school version is available, but it was not recommended for use with accelerated students in these programs (Eldredge, 1990). The LASSI can be administered in a group setting in approximately 30 minutes. The carbonless test format allows participants to score their own assessment and take a copy of the results with them from the testing session.

The ten LASSI subscales focus on thoughts and behaviors related to successful learning. The ten subscales are (1) Attitude, (2) Motivation, (3) Time Management, (4) Anxiety, (5) Concentration, (6) Information Processing, (7) Selecting the Main Ideas, (8) Study Aids, (9) Self-testing, and (10) Test Strategies (for more details see Weinstein, 1987). Reliability studies reported Cronbach alpha internal consistency values ranging from .70 to .86 and test-retest reliabilities from .70 to .85. Validity studies have also reported normative data for high school and college students with different instruments for each group (Weinstein, Palmer, & Schulte, 1987). Students respond to individual items on each subscale using a five-point scale: (5) very typical of me; (4) fairly typical of me; (3) somewhat typical of me; (2) not very typical of me; and (1) not at all typical of me. Some item values are reverse keyed before being added to obtain a subscale score. The subscale scores are compared by graphing them onto a normal curve equivalent percentile chart.

According to the LASSI user's manual (Weinstein, 1987), students scoring above the 75th percentile do not need to improve that specific skill or strategy. Students scoring between the 75th percentile and the 50th percentile should consider improvement. Students scoring below the 50th percentile on a subscale need assistance to improve that skill or strategy. For example, students scoring below the 50th percentile on the anxiety subscale would be considered anxious about being in college. Likewise, students scoring below the 50th percentile on the motivation subscale lack appropriate motivation to do college level work effectively.

Research Question

The research question of interest was whether the ten LASSI subscales could predict a student's college grade point average after one semester of college coursework. A related question pertained to whether a "subset" of the ten LASSI subscales could better predict college grade point average for this sample of students. Students not maintaining at least a 2.50 grade point average after one semester of college coursework were dismissed from the Academy. Knowledge of which subscales are best predictors of college grade point average would aid staff in identifying potential at-risk students upon entering the Academy.

Data Analysis

The data were analyzed using a SAS statistical program. The student's college grade point average was predicted by the ten LASSI subscales using PROC REG with the SELECTION statement requesting the best subset model criteria. The PROC PRINCOMP procedure was used to create ten orthogonal principal component variables. The principal component variable parameter estimates were then computed using the

Table 6 Mallows Cp and Principal Components Regression Comparison (n_1 = 100)

Best Variable Subset Model	Mallows Cp				Principal Components				
	ß	$SE_ß$	t	p	ß	$SE_ß$	t	p	R^2
X10	.10	.02	5.00	.0001	.82	.16	5.13	.0001	.21
X3	.13	.03	4.33	.0001	.02	.15	.13	.90	.33
X8	.18	.04	4.50	.0001	1.05	.15	7.00	.0001	
X3	.10	.02	5.00	.0001	.98	.12	8.17	.0001	.44
X8	.16	.03	5.33	.0001	.42	.14	3.00	.0024	
X10	.07	.02	3.50	.0001	.21	.16	1.31	.1951	
X1	.10	.03	3.33	.0009	1.04	.11	9.45	.0001	.50
X3	.10	.02	5.00	.0001	.07	.12	.58	.59	
X8	.14	.03	4.67	.0001	.28	.15	1.87	.07	
X10	.06	.02	3.00	.0004	.14	.16	.88	.39	
X1	.11	.03	3.67	.0004	1.06	.10	.60	.0001	.54
X3	.10	.02	5.00	.0001	.11	.12	.92	.35	
X6	.06	.02	3.00	.0004	.07	.13	.54	.55	
X8	.12	.03	4.00	.0001	.19	.15	1.27	.20	
X10	.06	.02	3.00	.0004	-.02	.15	-.13	.90	
X1	.09	.03	3.00	.0004	.97	.10	9.70	.0001	.58
X3	.10	.02	5.00	.0001	.42	.11	3.92	.0004	
X5	.09	.02	4.50	.0001	.31	.12	2.58	.01	
X8	.12	.03	4.00	.0001	.22	.14	1.57	.11	
X9	.06	.02	3.00	.0004	-.11	.14	-.79	.43	
X10	.06	.02	3.00	.0004	.17	.15	1.13	.26	
X1	.10	.03	3.33	.0004	1.02	.09	11.33	.0001	.62
X3	.09	.02	4.50	.0001	.41	.11	3.73	.0002	
X5	.08	.02	4.00	.0001	-.10	.11	-.91	.37	
X6	.05	.02	2.50	.03	.09	.12	.75	.45	
X8	.10	.03	3.33	.0004	.16	.13	1.23	.24	
X9	.06	.02	3.00	.0004	.20	.14	1.43	.16	
X10	.05	.02	2.50	.03	.11	.14	.79	.44	

PROC REG procedure. The number of significant principal component parameter estimates were subsequently identified. These procedures are outlined in the *SAS System for Regression* manual (Freund & Littell, 1991).

Results

The correlation matrix, means and standard deviations of the ten LASSI subscales are in Table 7. The intercorrelations among the subscales indicated that Anxiety was not significantly correlated with Time Management, Information Processing, Support Techniques/Materials, and Self-Testing. The lowest subscale mean was on Selecting Main Ideas.

Mallows Cp

The Mallows Cp statistic is calculated as: Cp = (SSEp/MSE) - (n - 2p) + 1 (Freund & Littell, 1991) or Cp = [1/2 (RSSp) - n + 2p] (Mallows, 1973); where RSSp is the residual sum of squares from the best variable subset model, MSE and/or 2 is the mean square error from the full model with all predictor variables, n = sample size, and p = number of predictors.

The procedure for finding the optimum subset of all possible subset sizes requires computing 2^m equations. The ten subscale predictors in the model yielded 1024 regression equations (2^{10}) with associated selection criteria statistics {Note: the determination of the number of subset equations generated for p predictor variables from an m variable full model is: m!/[p!(m-p)!]. For example, the number of 2 variable subset equations generated from a 10 variable model would be 45}. Only the single best variable subset models of each size are reported.

Table 6 (cont.) Mallows Cp and Principal Components Regression Comparison (n_1 = 100)

Best Variable Subset Model	Mallows Cp				Principal Components				
	ß	$SE_ß$	t	p	ß	$SE_ß$	t	p	R^2
X1	.10	.03	3.33	.0004	1.03	.09	11.44	.0001	.63
X2	.02	.01	2.00	.05	.18	.10	1.80	.09	
X3	.09	.02	4.50	.0001	.03	.11	.27	.77	
X5	.08	.02	4.00	.0001	.30	.11	2.72	.01	
X6	.05	.02	2.50	.03	.01	.13	.08	.92	
X8	.09	.03	3.00	.0004	.12	.13	.92	.36	
X9	.05	.02	2.50	.03	.25	.14	1.78	.09	
X10	.05	.02	2.50	.03	-.05	.14	-.36	.75	
X1	.09	.03	3.00	.0004	.99	.08	12.38	.0001	.64
X2	.02	.01	2.00	.05	.24	.10	2.40	.02	
X3	.08	.02	4.00	.0001	.03	.11	.27	.77	
X4	.05	.03	1.67	.10	.10	.11	.91	.36	
X5	.07	.02	3.50	.0004	-.08	.13	-.62	.52	
X6	.05	.02	2.50	.03	.08	.13	.62	.52	
X8	.09	.03	3.00	.0004	.02	.14	.14	.91	
X9	.05	.02	2.50	.03	-.001	.14	.007	.99	
X10	.05	.02	2.50	.03	.33	.15	2.20	.04	
X1	.09	.03	3.00	.0004	.97	.08	12.13	.0001	.65
X2	.02	.01	2.00	.05	.27	.10	2.70	.008	
X3	.08	.02	4.00	.0001	.05	.10	.50	.60	
X4	.05	.03	1.67	.10	-.09	.11	-.82	.42	
X5	.07	.02	3.50	.0004	.06	.11	.55	.59	
X6	.05	.02	2.50	.03	.06	.12	.50	.60	
X7	.02	.02	1.00	.25	-.07	.12	.58	.57	
X8	.09	.03	3.00	.0004	.01	.14	.07	.94	
X9	.04	.02	2.00	.05	.23	.15	1.53	.12	
X10	.04	.02	2.00	.05	.19	.15	1.27	.21	

Note. Regression parameters have been rounded to two decimal places unless otherwise noted. The t value = ß / $SE_ß$.

Table 7 LASSI Subscale Inter-Correlations, Means, and Standard Deviations (n = 156)

LASSI Subscale	1	2	3	4	5	6	7	8	9	10
1 Attention										
2 Motivation	.59									
3 Time Mngmnt	.39	.60								
4 Anxiety/Worry	.32	.15	.09							
5 Concentration	.57	.62	.62	.33						
6 Information	.20	.15	.39	.03	.26					
7 Select Ideas	.25	.36	.31	.37	.47	.30				
8 Support	.24	.40	.47	.05	.38	.45	.40			
9 Class Prep.	.38	.50	.63	.06	.55	.56	.39	.64		
10 Test Strategy	.54	.47	.33	.50	.66	.20	.60	.21	.34	
Mean	34.33	33.12	24.91	28.38	28.56	28.94	18.32	26.03	27.36	31.46
SD	4.17	4.73	6.18	5.92	4.93	5.24	3.51	5.96	5.84	4.58

Note. The values have been rounded to two decimal places.

Table 8 Best Model Selection Criteria by Subset Size

Subset Size	Variables in Subset Model	R^2	Cp
1	(2)	.09	10.88
2	(2),(8)	.11	8.01
3	(2),(6),(8)	.14	5.16
4	(2),(4),(8),(9)	.17	2.72
5	(2),(4),(6),(8),(9)	.18	2.93
6	(2),(4),(6),(7),(8),(9)	.18	3.68
7	(1),(2),(4),(6),(7),(8),(9)	.19	5.10
8	(1),(2),(4),(6),(7),(8),(9),(10)	.19	7.05
9	(1),(2),(3),(4),(5),(6),(8),(9),(10)	.19	10.04
10	(1),(2),(3),(4),(5),(6),(7),(8),(9),(10)	.19	11.00

Note. The four variable subset model according to the Cp criteria would be selected as the best model.

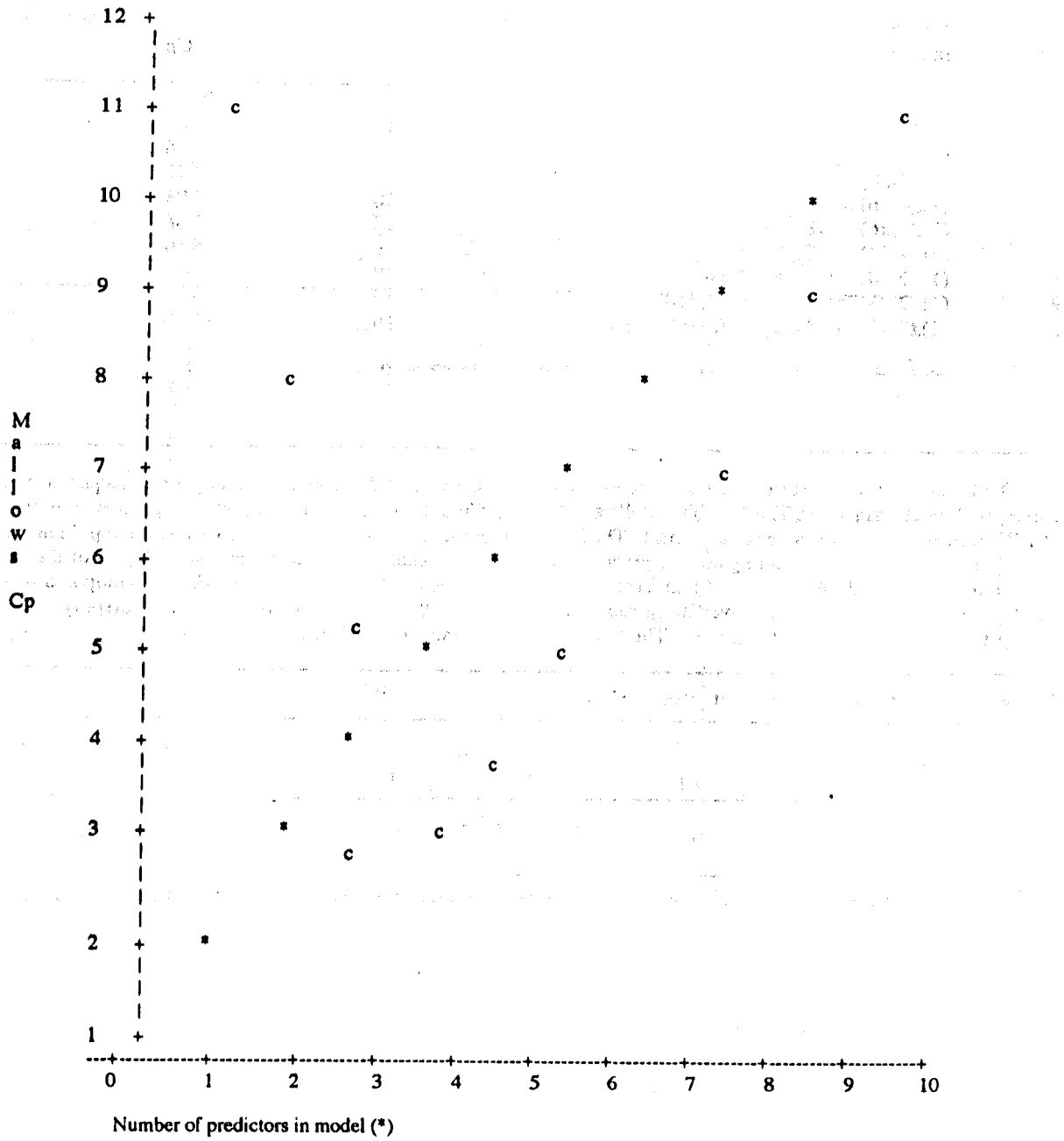
The best subset model for each subset size with the corresponding criteria are in Table 8. The Mallows Cp of 2.72 indicated a four variable subset model. The four variable subset model for predicting college grade point was Anxiety/ (4), Study Aids (8), and Self Testing (9). The Cp criteria also indicated the overfitting caused by having too many variables in the model. The large Cp values indicated equations with larger mean square error. If $Cp > (p + 1)$, for any subset size p, then bias was present. If $Cp < (p + 1)$, for any subset size p, then the model contained too many variables. A plot of the Cp values against the number of predictors, compared to a plot of the $(p + 1)$ values, visually displays this phenomenon in Figure 1.

Table 9 Principal Component Regression

Model	Type II SS	df	MS	F	p.	R^2
Regression	10.76	10	1.08	3.35	.001	.19
Model Components						
(1)	4.16	1				
(2)	.99	1				
(3)	1.13	1				
(4)	1.93	1				
(5)	.09	1				
(6)	.23	1				
(7)	.58	1				
(8)	1.33	1				
(9)	.29	1				
(10)	.03	1				
Error	46.58	145	.32			
Total	57.34	155				

Note. Adj. R^2 = .13, PCR $R^2_{1,4,8}$ = 69 % (7.42/10.76).

Figure 1 Overlay Plot of Cp and (p + 1) Values



The present pattern of Cp values for the various subsets of size p are typical when multicollinearity is present. The Cp values initially become smaller, but then start to increase. The plot of Cp values is similar to a "scree" plot in factor analysis and as such a multiple regression method might also be useful in determining the number of variables to retain (Zoski & Jurs, 1993). The best subset model is indicated when the Cp values begin to increase and cross the (p + 1) values (Figure 1).

Principal Components Regression

Principal components are obtained by computing eigenvalues from the correlation matrix. The correlation matrix is used so that variables are not affected by the scale of measurement as in the use of a variance-covariance matrix. Since eigenvalues are the variances of the principal component variables, the sum of the eigenvalues equal the number of variables in the full model, just as the sum of standardized variable

variances would equal the number of variables. This sum is the measure of the total variation in the data set. A wide variation in the eigenvalues would suggest the presence of multicollinearity among the variables. The number of eigenvalues greater than unity, as in factor analysis, would indicate the number of variables from the full model that would explain most of the variance in the data set. The eigenvectors, in contrast, contain the coefficients for each principal component variable. These coefficients are used to create the observed values of the original variables. These observed values are then used in multiple regression as orthogonal predictor values with no multicollinearity present.

Preliminary inspection of the model components (Type II SS) in Table 9 indicated three principal component variables (1, 4, and 8) that accounted for 69% of the variance in predicting college grade point average (7.42/10.76). The first model component alone explained 39 % of the variance (4.16/10.76).

A comparison of the full model parameter estimates in Table 10 between the original correlated predictors and the principal component regression variables sheds better insight into the best variable subset model selection criteria. The multiple regression analysis with correlated predictors identified Motivation (2) and Support (8) while the principal component method identified Attention (1), Anxiety/worry (4), and Support (8).

Summary

The Cp criteria identified a four variable predictor model as best: Motivation (2), Anxiety/worry (4), Support (8), and Class Preparation (9). This four variable subset model was further verified by examining where the plot of Cp values against the (p + 1) values crossed. The Cp criteria selected the smallest variable subset model in the presence of variable multicollinearity. The principal components approach identified Attention (1), Anxiety(4), and Study Aids (8). In examining the parameter estimates in the multiple regression analysis, only Motivation (2) and Study Aids (8) were significant relative to the other predictors in the model. The Mallows Cp and PCR criteria indicated slightly different sets of predictor variables depending upon whether the independent variables were correlated.

In using multiple regression it is important to have a theoretical basis for the regression model and to consider model validation. A common misconception in multiple regression is that the model with all the significant predictors included is the best model. This isn't always the case. The problem is that the beta values and R-squared values are data dependent due to the least squares criterion being applied to a specific sample of data. A different sample will usually result in different parameter estimates and variance explained. Although the standard errors of the beta values do provide the researcher with some indication of the amount of change expected from sample to sample, the fact remains that the estimates obtained from one sample may predict poorly when applied to a new set of sample data. The primary method to assess any change in estimates is to replicate the regression model using other sample data. The Mallows Cp criteria was similarly suspect because values were inflated upon cross validation and the best variable subset model in one sample was not identified in the other sample. Obviously, if the mean square error estimates and the residual sums of squares fluctuate, then model selection will be erroneous (see Mallows Cp formula).

The rationale behind a regression model is to estimate σ^2 (the true model's mean square error variance). Since σ^2 is not generally known, a researcher must estimate it from a knowledge of prior research ($\sigma^2 = \sigma^2_{y.x}$), obtain estimates from a model containing all theoretically relevant predictors, replicate the study, or use bootstrapping, jacknifing, and cross-validation methods. In this regard, effect size considerations, as recommended by Thompson et al. (1991), become important to consider in evaluating a regression model.

Table 10 Multiple Regression and Principal Component Parameter Estimate Comparisons

	Mallows Cp				Principal Components			
Variable	ß	SEß	t	p	ß	SEß	t	p
1	.01	.02	.68	.50	.08	.02	3.60	.001
2	.03	.02	2.29	.02	.06	.04	1.76	.081
3	.002	.01	.19	.84	-.08	.04	-1.88	.062
4	.02	.01	1.84	.07	.14	.06	2.45	.015
5	-.003	.02	-.17	.87	-.03	.06	-.53	.600
6	.01	.01	1.30	.20	.05	.06	.84	.404
7	-.02	.02	-1.02	.31	.10	.08	1.34	.182
8	-.03	.01	-2.82	.005	-.18	.09	-2.03	.044
9	.02	.01	1.28	.20	.09	.09	.95	.341
10	.005	.02	.27	.79	-.03	.10	-.31	.758

References

Aiken, L.S. & West, S.G. (1993). *Multiple regression: Testing and interpreting interactions.* Newbury Park, CA: SAGE

Berk, K.N. (1977). Tolerance and condition in regression computations. *Journal of the American Statistical Association, 72,* 863-866.

Constas, M.A. & Francis, J.D. (1992). A graphical method for selecting the best subset regression model. *Multiple Linear Regression Viewpoints, 19*(1), 16-25.

Cummings, Corenna, C. (1982, March). *Estimates of multiple correlation coefficient shrinkage.* Paper presented at the American Educational Research Association annual meeting. New York, NY.

Davidson, Betty, M. (1988, November). *The case against using stepwise research methods.* Paper presented at the Mid-South Educational Research Association annual meeting. Louisville, KY.

Eldredge, J.L. (1990). Learning and study strategies inventory: a high school version (test review). *Journal of Reading, 34,* 146-149.

Freund, R.J. & Littell, R.C. (1991). SAS System for Regression (2nd Ed.) SAS Institute: Cary, NC.

Henderson, Douglas, A. & Denison, Daniel R. (1989). Stepwise regression in social and psychological research. *Psychological Reports, 64*(1), 251-257.

Huberty, C.J. (1989). Problems with stepwise methods--better alternatives. In B. Thompson (Ed.), *Advances in Social Science Methodology, 1,* 43-70. Greenwich, CT: JAI Press.

Mallows, C.L. (1966). *Choosing a subset regression.* Paper presented at the Joint Statistical Meetings. Los Angeles, CA.

Mallows, C.L. (1973). *Some comments on* C_p. Technometrics, 15, 661-675.

Morrison, D.F. (1976). *Multivariate Statistical Methods* (2nd Ed.), New York: McGraw-Hill.

Pohlmann, J. (1983, April). *A perspective on multicollinearity.* Paper presented at the American Educational Research Association annual meeting. Montreal, Canada.

Snyder, P. (1991). Three reasons why stepwise regression methods should not be used by researchers. In B. Thompson (Ed.), *Advances in Educational Research: Substantive findings, methodological developments, 1,* 99-105. Greenwich, CT: JAI Press.

Thayer, Jerome, D. (1986, April). *Testing different model building procedures using multiple regression.* Paper presented at the American Educational Research Association annual meeting. San Francisco, CA.

Thayer, Jerome, D. (1990, April). *Implementing variable selection techniques in regression.* Paper presented at the American Educational Research Association annual meeting. Boston, MA.

Thompson, B. (1989). Why won't stepwise methods die? *Measurement and Evaluation in Counseling and Development, 21(4),* 146-148.

Thompson, B., Smith, Q.W., Miller, L.M., & Thomson, W.A.. (1991, January) *Stepwise methods lead to bad interpretations: better alternatives.* Paper presented at the Southwest Educational Research Association annual meeting. San Antonio, TX.

Tracz, S., Brown, R., & Kopriva, R. (1991). Considerations, issues, and comparisons in variable selection and interpretation in multiple regression. *Multiple Linear Regression Viewpoints, 18*(1), 55-66.

Weinstein, C.E. (1987). *LASSI user's manual.* Clearwater, FL: H&H Publishing Company, Inc.

Weinstein, , C.E., Palmer, D.R., Schulte, A.C. (1987). *Learning and Study Strategies Inventory.* Florida: H&H Publishing.

Welge, Patricia (1990, January). *Three reasons why stepwise regression methods should not be used by researchers.* Paper presented at the Southwest Educational Research Association annual meeting. Austin,TX.

Zoski, K.K. & Jurs, S.G. (1993). Using multiple regression to determine the number of factors to retain in factor analysis. *Multiple Linear Regression Viewpoints, 20*(1), 5-9.

Zuccaro, Cataldo (1992). Mallows' C_p statistic and model selection in multiple linear regression. *Journal of the Market Research Society,34*(2), 163-172.

The author wishes to acknowledge Dr. Panu Sittiwong in the Academic Computing Center at the University of North Texas for his assistance in coding the SAS programs for this study.

Testing Directional Research Hypotheses

Keith McNeil
New Mexico State University

Theory, literature review, and past research results will guide the development and testing of most research questions. This paper argues that most research questions will be directional, instead of nondirectional, particularly since most researchers want to make a directional conclusion. Although many researchers incorrectly make directional conclusions after finding significance with a nondirectional test, tests of directional hypotheses are the only ones that allow directional conclusions.

Most computer packages only report the nondirectional probability. Therefore, an adjustment is necessary when a directional research hypothesis has been tested. Exhibits are provided for testing both directional and nondirectional hypotheses regarding a) the difference between two means, b) single population correlation, c) traditional covariance, d) interaction between two dichotomous predictors, e) interaction between one continuous variable and one dichotomous variable, f) contribution of a variable, and g) selected non-linear hypotheses.

Researchers have a choice of various statistical tools; readers of this journal realize that most research hypotheses can be tested with the GLM. Each statistical tool can be used to test both nondirectional research hypotheses and directional research hypotheses. The researcher has to decide whether the research hypothesis is directional or nondirectional. The choice should not be difficult, as the decision is affected by theory, literature review, and past research. If these areas do not provide a clue, then the researcher should consider the desired conclusion. If the researcher is content with stating, "There is a difference between Treatment and Comparison," then the nondirectional research hypothesis is appropriate. But if all the forces point to desiring to make the directional conclusion, "Treatment is better than Comparison," then a directional research hypothesis is appropriate. The choice of a directional or nondirectional research hypothesis is not a statistical one. The choice is driven by the research base and tied to one's desired conclusion.

A sample of three recent statistics texts illustrates the confusion related to this issue. Grimm (1993) waffles on the use of directional research hypotheses.

> Research hypotheses (scientific hypotheses) are usually stated as predictions about the expected direction of an experimental effect. For Exhibit, persuasion technique A will induce greater attitude change than technique B; subjects' perceptions of control over a stressor will decrease stress reactions; or higher levels of physiological arousal will create stronger emotions. Researchers typically frame their statistical hypotheses in a nondirectional form. In other words, even though the research hypothesis makes a prediction about which of two means will be larger, the null and alternative hypotheses allow the investigator to discover if a treatment effect is opposite to the predicted effect. (p. 184)

His major concerns are that choice of the direction should be made before data are collected, a valid concern. But the other concern is that results in the opposite direction are ignored with a directional test. If one is theory building, then one may want to investigate those anomalous results to see if, in fact, they are replicable. Grimm (1993) does not treat directional hypotheses with statistical tests other than the difference between two means, although directional interpretations are often made with nondirectional tests of significance.

Sprinthall (1990) introduces directionality when discussing differences between two means, but treats the concept as a mechanical issue, "Remember, in terms of technique, the only difference between a one-tail and a two-tail *t* is how we look up the significance level" (p.185). He also doesn't discuss directional tests of significance for other tests of significance, but makes directional conclusions from several nondirectional research hypotheses. Several of his examples are stated as directional, but tested as nondirectional. Sprinthall (1990) points out that "the alternative hypothesis for *F* can never be directional. That is, if *t* is computed by taking the square root of *F*, then its significance must be evaluated against the critical values in the two-tailed *t* table "(p.275). That this is not so will be demonstrated later.

Shavelson (1988) is more in line with the essence of this paper. He introduces directional research hypotheses with the very first statistical test, even discussing the directional hypothesis before the nondirectional. In discussing most subsequent tests, he uses the same approach. He continuously emphasizes that "if both theory and empirical evidence suggest the outcome of a study, a directional research hypothesis should be used" (p. 251). He discusses directional hypothesis testing for a single mean, difference between two means, correlation, planned comparisons, and difference between two correlations. He does not discuss directional hypotheses in terms of ANOVA, ANCOVA, or multiple regression. Because he doesn't discuss the use of one degree of freedom F tests, he doesn't attend to the issue of computer-generated probabilities discussed in this paper.

Rationale for Directional Research Hypotheses

In the case of a new treatment, a researcher should show that it is more effective, costs less, is quicker to administrate, has longer lasting impact, etc. Who would care if the new treatment is worse than the existing comparison treatment? Any idiot can design a new treatment that is worse, costs more, is slower to administer, has a shorter lasting impact, etc. What would the research community learn from such findings? Over many years of experience with this issue, it has become apparent that nondirectional research hypotheses are only useful in dredging data in search of hypotheses for another researcher with some other data to verify. If a researcher has a good grasp of the content area, a directional research hypothesis will be desired.

Model Structure

An area of confusion is that both directional and nondirectional research hypotheses are tested by the same null hypothesis. For instance, if the research hypothesis is directional, "Treatment is more effective than Comparison," the statistical hypothesis is "Treatment is as effective as Comparison." If the research hypothesis is nondirectional, "Treatment and Comparison are not equally effective," the statistical hypothesis is "Treatment is as effective as Comparison." Most statistics texts illustrate this fact, but give primary coverage to the nondirectional hypothesis. Unfortunately, statistics texts do not emphasize the permissible conclusions of the two. Indeed, some statistics texts confuse the issue by making directional conclusions from nondirectional research hypotheses. Journal reviewers and editors reinforce the confusion by allowing only the statistical hypothesis to be reported. Why not force the author to state what is desired?

From a GLM perspective, the Full Model and Restricted Model are identical. The difference is the desired algebraic status of the weighting coefficient which will be identified as "want" in the following exhibits. Statistical packages (e.g., SAS, SPSS, BMDP) report only one probability value--that for the nondirectional research hypothesis. Consequently, many users mistakenly report that nondirectional probability when they have tested a directional research hypothesis.

Adjustment of Computed Probability

Statistics texts make the case that the required critical value depends upon whether one has a directional or nondirectional research hypothesis. We have all seen pictures of alpha in one tail of the t-distribution for a directional hypothesis, and alpha split between the two tails for a nondirectional research hypothesis. We also all remember that the relationship between t and F is $t^2 = F$. Thus the tails of the negative and positive sides of the t distribution both constitute the right-hand tail of the F distribution, as in Figure 1. What this means is that we would get a large F value half the time when sample mean_T > sample mean_C and half the time when sample mean_T < sample mean_C. If our research hypothesis was directional, then we would be interested only in one of the two halves of the F distribution in Figure 1. If the calculated F was 4.24, then the reported (nondirectional) probability would be .05. But if we had a directional research hypothesis, (say population mean_T > population mean_C) and the results were in line with our research hypothesis (sample mean_T = 15, sample mean_C = 10) instead of being exactly opposite, say (sample mean_C = 15, sample mean_T = 10), then we would obtain a t value of 2.06 and we would need to divide the reported probability by 2, as discussed in Figure 2.

On the other hand, if our results did turn out opposite to expectations, we would not want to say we had "significant results." Suppose our results produced a t value of -2.06 at @ in Figure 1. Although that t value translates to an F value of 4.24, one cannot rely on the F value (and the probability associated with it). One must check the data to see if the results are in the direction hypothesized. If the results are in the hypothesized direction (the shaded area in the bottom of the F distribution), then the computed probability must be divided by 2. If the results are not in the desired direction, then the computed probability must be divided by 2 and subtracted from 1.00. These procedures are outlined in Figure 2 and apply to each of the following exhibits.

Figure 1 Relationship Between t and F with Respect to Directional and Non-Directional Hypotheses

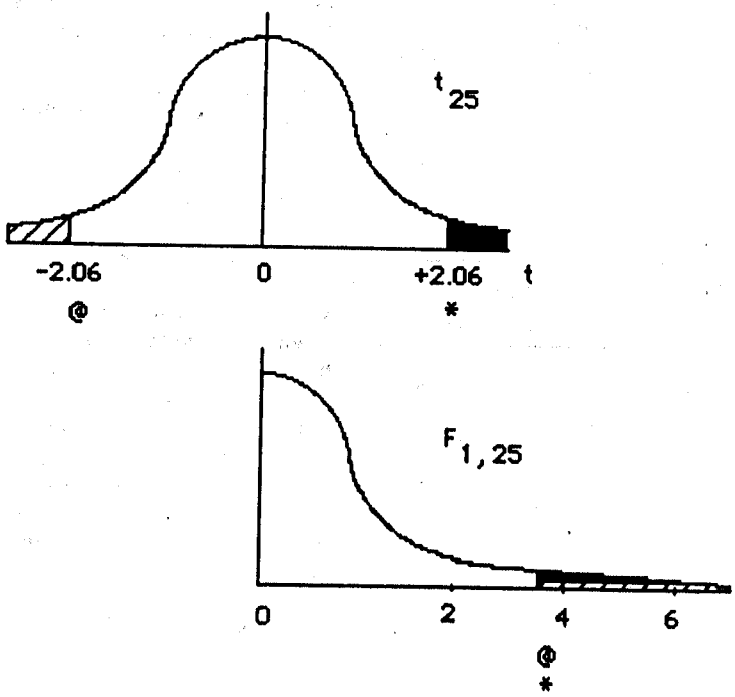


Figure 2 Procedures for Changing Computer-Generated Nondirectional Probability of F-tests to Directional Probabilities

Check to see whether **Condition I** or **Condition II** holds.

Condition I: If results (means, correlations, difference between means, etc.) are in the hypothesized direction: Divide nondirectional computer probability by 2.

Example: Nondirectional probability on printout is .08. Therefore the directional probability is (.08 / 2) .04, which is the probability that should be reported, and is indicated by the * in Figure 1.

Condition II: If results (means, correlations, differences between means, etc.) are opposite to the hypothesized direction, divide nondirectional computer probability by 2 and subtract the resulting value from 1.00.

Example: Nondirectional probability on printout is .08. Therefore the directional probability is 1 - (.08 / 2), or .96, which is the probability that should be reported, and is indicated by the @ in Figure 1.

Note. The directional research hypothesis could only have been tested when the numerator degrees of freedom are equal to 1.

Examples

Difference Between Two Means

Exhibit 1 contains both the directional research hypothesis and the nondirectional research hypothesis for testing the difference between two means. Notice that both research hypotheses use the same statistical hypothesis. The two Full Models are exactly the same, and the two Restricted Models are exactly the same. The difference is in the "want." The different wants require that different actions be taken on the computed probability, as discussed in the previous section. The different wants also impact the permissible conclusions.

Exhibit 1 Difference Between Two Population Means

Directional Research Hypothesis: For the population of interest, Group A has a higher mean than Group B on the criterion Y.

Nondirectional Research Hypothesis: For the population of interest, Group A and Group B are not equally effective on the criterion Y.

Statistical Hypothesis: For the population of interest, Group A and Group B are equally effective on the criterion Y.

Full Model: $Y = a0U + aGA + E3$

Want (for directional RH) $a > 0$; restriction: $a = 0$.
Want (for nondirectional RH) $a \neq 0$; restriction: $a = 0$.

Restricted Model: $Y = a0U + E4$

Where: Y = criterion; U = 1 for all subjects; GA = 1 if subject in Group A, 0 if subject in Group B; and a0 and a are least squares weighting coefficients calculated so as to minimize the sum of the squared values in the error vectors.

PROC REG; MODEL Y = GA;
TEST GA = 0;

Correlation

The above discussion is also appropriate to testing correlations. If a new testing instrument is developed, one would hope that it is reliable and valid. These conclusions require positive correlations, not correlations different from 0. If a theory posits that X and Y are related, the theory should specify if that relationship is positive or negative. If one is going to consider studying for a test, one needs to know if the relationship between studying and exam grade is positive or negative! Exhibit 2 provides the complete GLM solution of a research hypothesis regarding directional correlation.

Exhibit 2 Correlation

Directional Research Hypothesis: For some population, X is positively related with Y.

Nondirectional Research Hypothesis: For some population, X is related with Y.

Statistical Hypothesis: For some population, X is not related with Y.

Full Model: $Y = a0U + bX + E1$

Want (for directional RH) $b > 0$; restriction: $b = 0$.
Want (for nondirectional RH) $b \neq 0$; restriction: $b = 0$.

Restricted Model: $Y = a0U + E2$

Where: Y = criterion; U = 1 for all subjects; X = predictor score for subject; a0 and b are least squares weighting coefficients calculated so as to minimize the sum of the squared values in the error vectors.

PROC REG; MODEL Y = X;
TEST X = 0;

Analysis of Covariance

Assume that you have a Treatment and Comparison situation as previously described, and you want to adjust the posttest scores for initial differences in pretest scores. You would want the Treatment group to be higher than the Comparison group on the adjusted posttest scores. Again, who would be interested in a treatment that produced lower adjusted posttest scores? Exhibit 3 provides the GLM solution for both the nondirectional and directional analysis of covariance research hypothesis. The directional research hypothesis in ANCOVA is applicable only when there are two groups being compared, resulting in one degree of freedom in the numerator of the F. When there is more than one degree of freedom in the numerator, only a nondirectional research hypothesis can be tested.

Exhibit 3 Analysis of Covariance

Research Hypothesis: For a given population, Method A is better than Method B on the criterion Y, over and above the covariable C.

Nondirectional Research Hypothesis: For a given population, Method A and Method B are differentially effective on the criterion Y, over and above the covariable C.

Statistical Hypothesis: For a given population, Methods A and B are not differentially effective on the criterion Y, over and above the covariable C.

Full Model: Y = a0U + a2G2 + c1C + E1

Want (for directional RH) a2 < 0; restriction: a2 = 0
Want (for nondirectional RH) a2 not equal 0; restriction: a2 = 0

Restricted Model: Y = a0U + c1C + E2

Where: Y = criterion; U = 1 for each subject; G2 = 1 if subject received Method B, 0 if Method B; C = covariable score; and a0, a2, and c1 are least squares weighting coefficients calculated so as to minimize the sum of the squared values in the error vectors.

PROC REG; MODEL Y = G2 C;
TEST G2 = 0;

Interaction Between Two Dichotomous Variables

Suppose you have two treatments and two levels of motivation, and are interested in Posttest scores. Traditional analysis of variance tests for the interaction effect first, and then proceeds to the main effects if the interaction is not significant, and to simple effects if the interaction effect is significant. The interaction effect usually is treated as an assumption, or as an effect that is preferably not in existence. But the interaction effect may be the researcher's primary hypothesis, and it may be either directional or nondirectional. (In traditional analysis of variance it is always nondirectional, unless tested as an a priori contrast.)

Suppose that the treatment was designed to be particularly responsive to highly motivated students. Based on the assumption that there might be ways to increase student's motivation, you expect the directional interaction pictured in Figure 3. Your expectation is that "Students with high motivation will do better on the Posttest than students with low motivation, and the difference will be greater for the Treatment than for the Comparison." The focus of the directional interaction could just as well have been on treatments, with the expectation being "Treatment students will do better on the Posttest than Comparison students, and the difference will be greater for high motivated students than for low motivated students." The two statements are equivalent and both identify directional interaction. The complete GLM solution is provided in Exhibit 4. Notice again that the only difference between directional and nondirectional is in the "want," in the adjustment of the probability, and the permissible conclusion. Again, the directional interaction can be tested only if there is one degree of freedom in the numerator of the F.

Figure 3 Directional Interaction Between Two Dichotomous Predictors

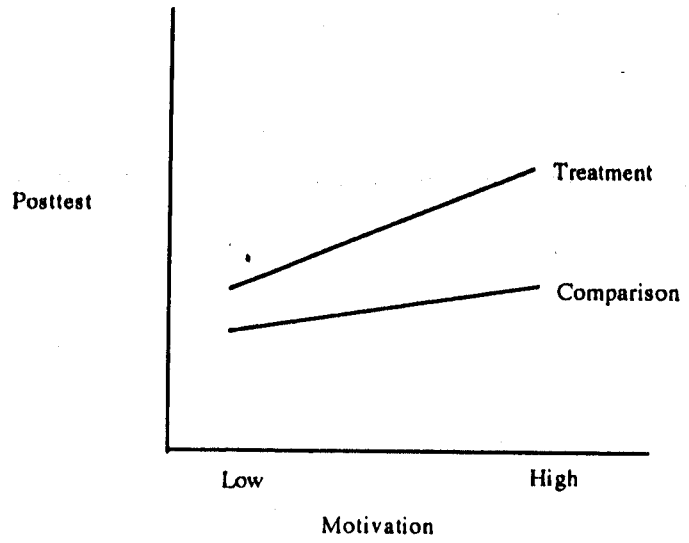


Exhibit 4 Directional Interaction Between Two Dichotomous Predictors

Directional Research Hypothesis: For a given population, the relative effectiveness of Method A (X10) as compared to Method B (X11) on the criterion of interest (X9) will be greater for Group A (X12) than for Group B (X13).

Nondirectional Research Hypothesis: For a given population, the relative effectiveness of Method A (X10) as compared to Method B (X11) on the criterion of interest (X9) will be different for Group A (X12) than for Group B (X13).

Statistical Hypothesis: For a given population, the relative effectiveness of Method A (X10) as compared to Method B (X11) on the criterion of interest (X9) will be the same for Group A (X12) as for Group B (X13).

Full Model: X9 = a0U + b(X10*X13) + c(X11*X12) + d(X11*X13) + E1

Want (for directional RH) (c) > (b - d);
restriction: (c) = (b - d)
Want (for nondirectional RH) (c) not equal (b - d);
restriction: (c) = (b - d)

Restricted Model: X9 = a0U + eX10 + fX12 + E2

PROC REG; MODEL X9 = X10*X13 X11*X12 X11*X13;

TEST (X11*X12) = (X10*X13 - X11*X13);

Interpretation: If the weighting coefficient c is numerically larger than (b - d), the directional probability is appropriate and the following conclusion can be made: For a given population, the relative effectiveness of Method A (X10) as compared to method B (X11) on the criterion of interest (X9) will be greater for Group A (X12) than for Group B (X13).

Interaction Between One Continuous Variable and One Dichotomous Variable

An extension of the previous section would be to consider motivation as a continuous variable instead of as a dichotomous variable. The same rationale applies, although now since motivation is being considered as a continuous variable two lines will be fit to the date, not four means. Figure 4 depicts the expected directional interaction. Note that Figure 4 appears very similar to Figure 3, the only difference is that motivation is considered as a continuous variable in Figure 4. The directional interaction research hypothesis would be, "As motivation increases, the relative superiority of Treatment over Comparison increases." Shavelson (1988) presents a directional example of this type, framed as the "test for difference between regression slopes from two independent samples." His presentation is in terms of a complicated t test. The GLM approach illustrates the similarity of all directional research hypotheses and relies on the same model comparisons as all the previous examples. Exhibit 5 contains the complete GLM solution for interaction between one continuous variable and one dichotomous variable.

Exhibit 5 Interaction Between One Continuous Variable And One Dichotomous Variable

Directional Research Hypothesis: For a given population, as X increases, the relative superiority of Method A over Method B on Y will linearly increase.

Nondirectional Research Hypothesis: For a given population, as X increases, the relative superiority of Method A over Method B on Y will linearly change.

Statistical Hypothesis: For a given population, as X increases, the difference between Method A and Method B on Y will remain the same.

Full Model: Y = aU + a1U1 + b1X1 + b2X2 + E1

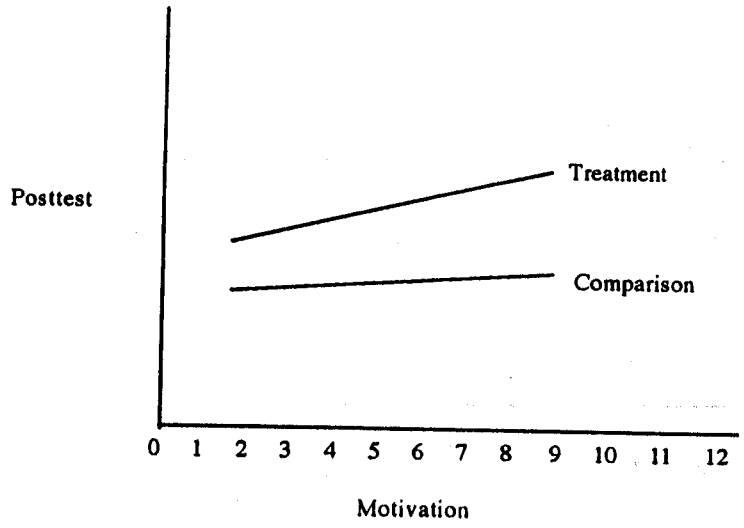
Want (for directional RH) b1 > b2;
restriction: b1= b2
Want (for nondirectional RH) b1 not equal b2;
restriction: b1 = b2

Restricted Model: Y = aU + b3X + E8

Where: Y = the criterion; U1 = 1 if the score on the criterion is from a subject in Method A, 0 otherwise; X = the continuous predictor variable; X1 = (U1*X) = the continuous predictor variable if the criterion is from a subject in Method A, 0 otherwise; U2 = 1 if the score on the criterion is from a subject in Method B, 0 otherwise; X2 = (U2*X) = the continuous predictor variable if the criterion is from a subject in Method B, 0 otherwise; and a, a1, b1, b2, and b3 are least squares weighting coefficients calculated so as to minimize the sum of the squared values in the error vectors.

PROC REG; MODEL Y = U1 X1 X2;
TEST X1 = X2;

Figure 4 Directional Interaction Between One Continuous Predictor (Motivation) and One Dichotomous Predictor (Type of Treatment)



Non-Linear Relationships

If all the predictor variables of interest are polynomial terms, the directional research hypothesis is still appropriate. Consider the case in which the linear and second-degree terms are under consideration. The second-degree curve can be either an inverted U or U-shaped. The U-shaped curve identifies a "trough" of minimum performance on the criterion, whereas the inverted U identifies a "peak" of maximum performance on the criterion. These are two very different conclusions and are a function of the sign of the second-degree term. The curves are identified in Figure 5 and the GLM solution is in Exhibit 6.

Exhibit 6 Non-linear Hypotheses

Directional Research Hypothesis: For a given population, there is a positive second degree effect of X on Y, over and above the linear effect of X.

Nondirectional Research Hypothesis: For a given population, there is a second degree effect of X on Y, over and above the linear effect of X.

Statistical Hypothesis: For a given population, there is not a positive second degree effect of X on Y, over and above the linear effect of X.

Full Model: Y = a0U + aX + bX16 + E1
Where: X16 = X*X

Want (for directional RH) b > 0; restriction: b = 0
Want (for nondirectional RH) b not equal 0; restriction: b = 0

Restricted Model: Y = a0U + aX + E2

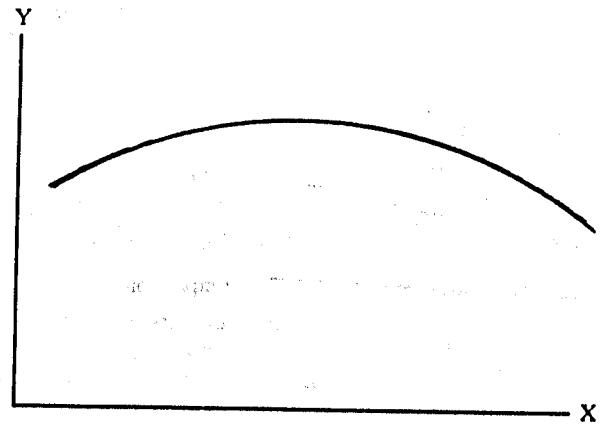
PROC REG; MODEL Y = X X16;
TEST X16 = 0;

Contribution of One Variable, Over and Above Other Variables

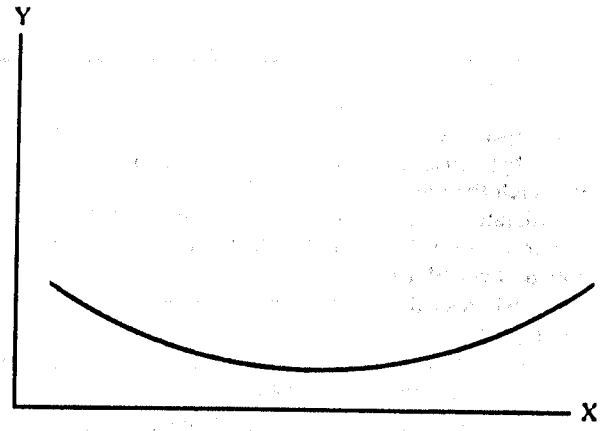
A researcher may be interested in how a variable is related to a criterion, after the effects of several other variables have been "statistically adjusted." If the variable is dichotomous (say study or not study), then this question is simply an extension of the analysis of covariance discussion into more than one covariable. The GLM solution would simply have the multiple covariables in the Full Model as well as in the Restricted Model as in Exhibit 3.

If the variable under concern is a continuous variable (say hours of studying), then whether the variable relates positively or negatively to the criterion after adjustment for the covariables would be of interest in the directional situation. Again, knowing that studying is predictive of the criterion (over and above the other variables) is not that informative; what is informative is knowing whether studying is positively related or negatively related to the criterion. If one wanted to use these results to recommend trying to increase the criterion, one would have to know the directional relationship between studying and the criterion. The GLM solution is provided in Exhibit 7.

Figure 5 U-Shaped Curves Resulting From Negative and Positive Weights of Second Degree Terms



$Y = aU + bX = cX^2 + E$, where c is negative.



$Y = aU + bX + cX^2 + E$, where c is positive.

Exhibit 7 General Over and Above

Directional Research Hypothesis: For a given population, X6 is positively predictive of the criterion Y, over and above X1, X2, X3, and X4.

Nondirectional Research Hypothesis: For a given population, X6 is predictive of the criterion Y, over and above X1, X2, X3, and X4.

Statistical Hypothesis: For a given population, X6 is not predictive of the criterion Y, over and above X1, X2, X3, and X4.

Full Model: Y = a0U + a1X1 + a2X2 + a3X3 + a4X4 + a6X6 + E1

Want (for directional RH) a6 > 0; restriction: a6 = 0
Want (for nondirectional RH) a6 not equal 0; restriction: a6 = 0

Restricted Model: Y = a0U + a1X1 + a2X2 + a3X3 + a4X4 + E2

Where: Y = the criterion; X1, X2, X3, X4, X6 = continuous or categorical information; and a0, a1, a2, a3, a4, and a6 are least squares weighting coefficients calculated so as to minimize the sum of the squared values in the error vectors.

PROC REG; MODEL Y = X1 X2 X3 X4 X6;
TEST X6 = 0;

Summary

Researchers often do not follow the knowledge base by stating a directional research hypothesis. Often, though, directional conclusions are made from testing non-directional research hypotheses. Since the statistical (or null) hypothesis is the same for directional and non-directional research hypotheses, researchers often overlook the distinction. In addition, all canned computer packages report only the non-directional probability. This paper has illustrated how the GLM can be used for directional hypothesis testing and for obtaining the correct directional probability.

All the previous exhibits are subsets of the same general situation described in Exhibit 7. The differences depend on the number of predictors, number of covariates (many, one, none), and whether the variable tested is continuous or dichotomous. In all the statistical tests discussed, a directional research hypothesis can be tested if there is a directional expectation. If there is a directional research hypothesis, there is only one want, one restriction, and one degree of freedom in the numerator of the F-test. In all cases the reported nondirectional probability must be adjusted based on how the sample results match the directional research hypothesis. These are all essential elements of a directional hypothesis.

References

Grimm, L. G. (1993). *Statistical applications for the behavioral sciences*. New York: Wiley.

Shavelson, R. J. (1988). *Statistical reasoning for the behavioral sciences*. (2nd ed.). Boston: Allyn and Bacon.

Sprinthall, R. C. (1990). *Basic statistical analyses*. Englewood Cliffs, NJ: Prentice Hall.

Orthogonal Comparisons A Teaching Example

Keith McNeil
New Mexico State University

When the omnibus one-way Analysis of Variance (ANOVA) is found to be significant, the research question that "at least two populations have different means" can be accepted, but is found to be lacking. (What most textbooks fail to mention is that this means that the one-way ANOVA question is a fruitless question.) Most textbooks turn to post-hoc analyzes as a way to determine "where the significance is." But that journey is often muddled by: a) discussion of a myriad of post-hoc procedures, b) insufficient parallel examples, c) downplay of the value of planned comparisons, and d) failure to tie orthogonal comparisons to the two-way ANOVA. This paper will attempt to alleviate the above issues, with various examples of four groups.

Suppose that a researcher is interested in comparing four different treatments, and is encouraged to "first conduct the one-way ANOVA." The research hypothesis being tested here is, "For the population, at least two of these treatments are differentially effective." Given that the omnibus F is significant, the researcher can conclude, "For the population, at least two of these treatments are differentially effective." Note that which treatments are different cannot be specified. Nor can the more effective treatment be specified. The omnibus one-way F can be called a non-specific, non-directional research hypothesis, yielding little (or no) information.

Post-Hoc Comparisons

The myriad of post-hoc comparisons have been developed to attempt to rectify the non-specificity problem. These procedures protect the Type I error, some with orthogonal comparisons. It is this family of orthogonal comparisons on which the remainder of the paper will focus.

Orthogonal Comparisons

A comparison is said to be orthogonal if the set of contrast coefficients sum to zero, and if the sum of cross products with all other orthogonal comparisons also sums to zero. The set of contrast coefficients for RH1 in Exhibit 1 meets both criteria, as the set of coefficients sums to 0 (1 + 0 + 0 + -1 = 0), and the sum of the cross products of set 1 with set 2 also sums to 0 [(1 x 0) + (0 x 1) + (0 x -1) + (-1 x 0) = 0]. Each orthogonal comparison is a t-test question, either comparing one group to another (as in RH1 and RH2), or some combination of groups to some other combination of groups (as in RH3). With four groups, there is three degrees of freedom associated with the Between groups sum of squares. The three orthogonal contrasts identify three ways this sum of squares can be partitioned. It should be noted here that there are many (infinite?) ways that the sum of squares can be partitioned--some more meaningful for how the four groups were determined.

An example of when research hypothesis 1 (RH1), RH2, and RH3 might be of interest is when a researcher is studying two classes of each of two teachers, one in the AM and one in the PM. Let's assume that M1 is Teacher A, AM; M4 is Teacher A, PM. RH1 could be: "There is a difference in the effectiveness of Teacher A in the PM from that in the AM." Further assume that M1 is Teacher B, AM and M3 is Teacher B, PM. RH2 could be: "There is a difference in the effectiveness of Teacher B in the PM from that in the AM." While RH1 and RH2 both compare teacher effectiveness of AM and PM, the comparisons are on different teachers, so what is found with RH1 (Teacher A) will not have a bearing on what is found with RH2 (Teacher B). In this case, the data to determine the answer to RH1 is different from that determining the answer to RH2. (The data doesn't have to be different in order for orthogonality to hold, as evidenced by RH3, but it certainly clarifies the issue). RH3 compares the effectiveness of Teacher A (averaged over AM and PM) with the effectiveness of Teacher B (averaged over AM and PM). Logically, the outcome of RH1 (the relative effectiveness of Teacher A at AM and PM), and the outcome of RH2 (the relative effectiveness of Teacher B at AM and PM) does not impinge on the overall effectiveness of Teacher A as compared to Teacher B.

Exhibit 1 One Possible Set Of Contrast Coefficients With Four Groups: Non-Directional Hypotheses

	M1	M2	M3	M4
RH1 Non-directional: M1 not equal M4 SH: M1 = M4 OR 1*M1-1*M4 = 0	1	0	0	-1
RH2 Non-directional: M2 not equal M3 SH: M2 = M3 OR 1*M2 -1*M3 = 0	0	1	-1	0
RH3 Non-directional: M1+M4 not equal M2+M3 SH: M1+M4 = M2+M3 OR 1*M1 + 1*M4 -1*M2 -1*M3 = 0	1	-1	-1	1

Directional, Planned Orthogonal Comparisons

The above research hypotheses were non-directional, which is to say that differences were expected, but not directionally specified. For RH1, if the orthogonal contrast if found to be significant, then the conclusion is simply a restatement of the research hypothesis, "There is a difference in the effectiveness of Teacher A in the PM from that in the AM." While we know now that "groups M1 and M4 are different," we do not know how they are different. A directional conclusion can be made if the direction was posited in the research hypothesis before the data were looked at (preferably before the data were collected). Orthogonal contrasts specified before data collection are referred to as planned comparisons, and may be directional. Directional conclusions cannot be made from any post-hoc comparisons, only from planned comparisons. Exhibit 2 contains the same set of orthogonal comparisons as in Exhibit 1, but here as planned comparisons with expectations: (RH1') Teacher A being more effective in the AM than the PM, (RH2'), Teacher B being more effective in the AM than the PM, and (RH3') Teacher A being more effective than Teacher B (averaging over AM and PM classes).

Notice that the statistical hypothesis (SH) is the same in Exhibit 1 and Exhibit 2, and the orthogonal coefficients are the same. Again, what is different is the expected direction, and the permissible conclusion.

RH4, RH5, and RH6 in Exhibit 3 are another set of three orthogonal contrasts. While RH5 and RH2 are exactly the same, RH4 and RH6 are different from RH1 and RH3. The coefficients within RH4, RH5, and RH6 all add up to zero, and the sum of the cross products add up to zero, thus RH4, RH5, and RH6 constitute a different set of three orthogonal contrasts. Which set a researcher should use depends on the design of the study and the questions one has of the groups. Indeed, there are many other sets of orthogonal contrasts. As in all research, the questions should guide the analysis. With post-hoc comparisons, the researcher is limited to one less question than there are groups.

An example of when RH4, RH5, and RH6 might be of interest is when a researcher is testing the effectiveness of three different New treatments (M1, M2, and M3) and one Comparison treatment (M4). Since there are four groups, three orthogonal questions can be asked, and if the questions are asked before inspection of the data, Directional Research Hypotheses can be tested. Indeed, if a New treatment is being researched,we should expect it to be better than the Existing treatment. RH4 determines if the avererage

Exhibit 2 One Possible Set Of Contrast Coefficients With Four Groups: Directional Hypotheses

	M1	M2	M3	M4
RH1' Directional: M1 > M4 SH: M1 = M4 OR 1*M1-1*M4 = 0	1	0	0	-1
RH2' Directional: M2 > M3 SH: M2 = M3 OR 1*M2 -1*M3 = 0	0	1	-1	0
RH3' Directional: M1+M4 > M2+M3 SH: M1+M4 = M2+M3 OR 1*M1 + 1*M4 -1*M2 -1*M3 = 0	1	-1	-1	1

Exhibit 3 Another Possible Set Of Contrast Coefficients With Four Groups

M1 = New Treatment #1 M2 = New Treatment #2 M3 = New Treatment #3 M4 = Existing Treatment

		M1	M2	M3	M4
RH4	Non-directional: (M1+M2+M3)/3 not equal M4 Directional: (M1+M2+M3)/3 > M4				
SH	(M1+M2+M3)/3 = M4 OR 1*M1 + 1*M2 + 1*M3 - 3*M4 = 0	1	1	1	-3
RH5	Non-directional: M2 not equal M3 Directional: M2 > M3				
SH:	M2 = M3 OR 1*M2 - 1*M3 = 0	0	1	-1	0
RH6	Non-directional: M1 not equal (M2+M3)/2 Directional: M1 > (M2+M3)/2				
SH:	M1 = (M2+M3)/2 OR 2*M1 -1*M2 -1*M3 = 0	2	-1	-1	0

of the three New treatments is better than the one Comparison treatment. RH5 tests if New treatment 2 is better than New treatment 3. Finally, RH6 tests if New treatment 1 is better than the average of the other New treatments. As should now be clear, the design of the research, and the desired conclusion(s) determine the choice of the hypotheses, and whether the hypotheses are directional or non-directional. No one choice is always correct; the choice will depend on the research questions!

Pictorial Representation of Orthogonal Comparisons

Notice that RH5 and RH2 are the same, in terms of contrast coefficients. Since the two Exhibits were discussed with different samples, the research hypotheses may have seemed different. But in both cases, M2 was contrasted with M3. The sum of squares due to the four groups, though, was partitioned in different ways, as depicted in the Venn diagram in Figure 1. Figure 1a illustrates the one-way partitioning of sum of squares, into Within groups and Between groups. Note that the Between groups is between the four groups. Figure 1b illustrates the contrasts in Exhibit 1. About one-half of the Between groups sum of squares is due to RH2, and about one-fourth is due to RH1 and one-fourth to RH3. If the contrasts in Exhibit 3 were applied to the same data as in Exhibit 1, then Figure 1c might result. Note that since RH5 and RH2 are the same contrast, the sum of squares attributable to those contrasts is the same. But since RH4 and RH6 are different from RH1 and RH3, the sum of squares partitioned to these hypotheses will likely be different. RH6 is shown to account for none of the sum of squares in Figure 1c, while RH4 accounts for one-half of the Between groups sum of squares.

Source Tables

Another way to comprehend the different comparisons depicted in the Exhibits and in Figure 1 is through the source tables in Tables 1 through 3. Table 1 contains the one-way results, with the Total sum of squares being partitioned into just Between and Within. The four groups account for 40% of the Total sum of squares.Table 2 contains the partitioning depicted in Exhibit 1. Notice that the Total and Within sum of squares is the same as in Table 1, but the sum of squares due to Between groups has been further partitioned into the three comparisons. The RH2 comparison accounts for half of the sum of squares due to groups (20/40--hence half the overlapped area in Figure 1b). Since all of the F values in Table 2 fall beyond the critical value, all of these comparisons would be significant. Table 3 reflects the contrasts in Exhibit 3 and Figure 1c. Again notice that the sum of squares for RH2 in Table 2 and RH5 in Table 3 is the same. RH4 and RH6 are different from RH1 and RH3, and therefore the sum of squares is different. RH6 accounts for none of the sum of squares and is therefore not significant.

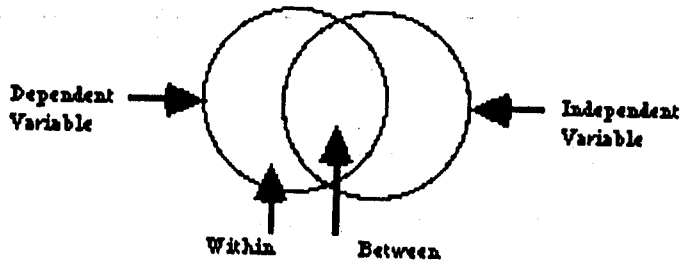
Example of Non-orthogonal Hypotheses

The reader may wonder why each of the New treatments in Exhibit 3 were not compared to the Existing treatment. These may be interesting research hypotheses, but they are not orthogonal. Exhibit 4 contains the hypotheses and orthogonal coefficients. While the coefficients do sum to zero within each of the hypotheses, the sum of the cross products is not zero. Think of it this way--if we start out by assuming all four treatments are equal, but find one inferior to another, isn't it likely that that one will be inferior to one of the others as well? In this case, the results from one hypothesis have a bearing on the results from another. Once we know the answer to one hypothesis, we have an inkling as to the answer to the other hypothesis. Additionally these hypotheses as a set are of little value, because they do not lead to a conclusive

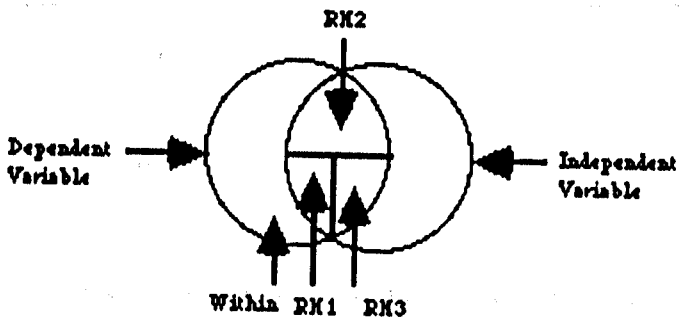
Figure 1 Hypothetical Sample Means and Venn Diagrams

	AM	PM
TEACHER A	7.5 (M1)	12.5 (M4)
TEACHER B	10.0 (M2)	20.0 (M3)

1A One Way Analysis



1B Exhibit 1 or 2 Analysis



1C Exhibit 3 Analysis

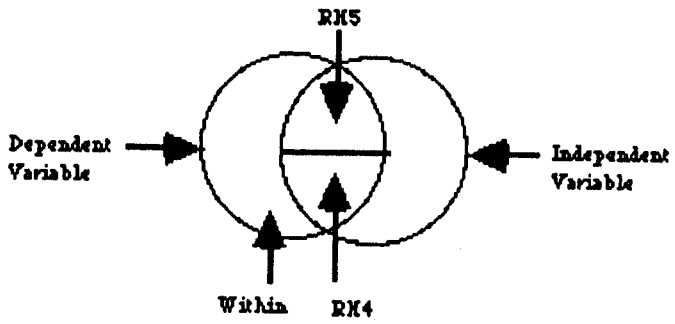


Table 1 One-way Source Table

SOURCE	SS	df	MS	F	E_{cv}
BETWEEN	40	3	13.33	13.33	2.76
WITHIN	60	60	1.00		
TOTAL	100	63			

Note. All E_{cv} are at alpha = .05.

Table 2 Exhibit 1 Source Table

SOURCE	SS	df	MS	F	E_{cv}
RH1	10	1	10	10.0	4.0
RH2	20	1	20	20.0	4.0
RH3	10	1	10	10.0	4.0
WITHIN	60	60	1		
TOTAL	100	63			

Note. All E_{cv} are at alpha = .05.

Table 3 Exhibit 3 Source Table

SOURCE	SS	df	MS	F	E_{cv}
RH4	20	1	20	20.0	4.0
RH5	20	1	20	20.0	4.0
RH6	0	1	0	0.0	4.0
WITHIN	60	60	1		
TOTAL	100	63			

Note. All E_{cv} are at alpha = .05.

answer. Suppose that all of the New treatments were better than the Existing treatment. Which New treatment would you recommend? The set of orthogonal hypotheses in Exhibit 3, on the other hand, lead to such a definite recommendation.

Trend Analysis

When the treatments are ordered on some underlying continuum, one may want to investigate the trends in the data as in Exhibit 5. That is, does the criterion increase linearly with an increase in the underlying continuum (as in RH10), or is there a minimum performance as in RH11? (By reversing all the weights in RH11, one could investigate maximum performance.) Finally, with four groups there may be a quadratic trend as in RH12. Note that the coefficients for RH10, RH11, and RH12 all add to zero, and that the cross products all add to zero. Therefore, RH10, RH11, and RH12 constitute another set of orthogonal contrasts for four groups.

Two Factors

Now suppose that the four groups differ not on just one underlying factor as in the above examples, but on two underlying factors. Exhibit 6 posits the following example of two groups getting the New treatment and two groups getting the Comparison treatment. Thus the first underlying factor is treatment: New vs. Comparison.

One of the New treatment groups is in the AM and one is in the PM. One of the Comparison treatment groups is in the AM and one is in the PM. Thus, the second factor is time of treatment: AM vs. PM.

What would be the research hypotheses of interest with this design? One probably would want to compare the New treatments to the Comparison treatments, and

Exhibit 4 Another Possible Set of Contrast Coefficients With Four Groups: Non-Orthogonal

M1 = New Treatment #1 M2 = New Treatment #2 M3 = New Treatment #3 M4 = Existing Treatment

		M1	M2	M3	M4
RH4	Non-directional: M1 not equal M4 Directional: M1 > M4				
SH	M1 = M4 OR 1*M1 + 0*M2 + 0*M3 - 1*M4 = 0	1	0	0	-1
RH5	Non-directional: M2 not equal M4 Directional: M2 > M4				
SH	M2 = M4 OR 1*M2 - 1*M4 = 0	0	1	0	-1
RH6	Non-directional: M3 not equal M4 Directional: M3 > M4				
SH	M3 = M4 OR 1*M3 -1*M4 = 0	0	0	1	-1

possibly the AM treatments to the PM treatments. These two hypotheses will be developed first, and then we will turn our attention to the third orthogonal comparison.

The Non-directional Research Hypothesis for treatment would be: "The two treatments, averaged across the two different time periods, are not equally effective," resulting in the orthogonal coefficients for RH13 in Exhibit 6. One could have stated this Research Hypothesis with a directional expectation, resulting in the same set of orthogonal coefficients. The Non-directional Research Hypothesis for time of treatment would be RH14: "The two time periods, averaged across the two different treatments, are not equally effective." Again, one could have stated this hypothesis with a directional expectation. Notice that the coefficients for RH14 are orthogonal to those for RH13. RH13 and RH14 are referred to as "main effects" hypotheses within the Analysis of Variance framework. Unless stated directionally a priori, they are always tested in a non-directional fashion.

Given the above two orthogonal contrasts, the third orthogonal contrast would have to be that specified in RH15. The non-directional research hypothesis associated with these coefficients is: "The difference between AM New treatment and PM New treatment is <u>different</u> from the difference between AM Comparison treatment and PM Comparison treatment." Again, one could have stated this hypothesis with a directional expectation. (For example, "The difference between AM New treatment and PM New treatment is <u>different</u> from the difference between AM Comparison treatment and PM comparison treatment." Again, one could have

Exhibit 5 One Possible Set of Contrast Coefficients With Four Groups: Trend Analysis

		M1	M2	M3	M4
RH10 linear trend	Non-directional: -3M1 -1M2 +1M3 +3M4 not equal 0 Directional: -3M1 -1M2 +1M3 +3M4 > 0				
SH	-3M1 -1M2 +1M3 +3M4 = 0 OR -3*M1 -1*M2 +1*M3 +3*M4 = 0	-3	-1	1	3
RH11 quadratic trend	Non-directional: M1 -M2 -M3 + M4 not equal 0 Directional: M1 -M2 -M3 + M4 > 0				
SH	M1 -M2 -M3 + M4 = 0 OR 1*M1 -1*M2 -1*M3 +1*M4 = 0	1	-1	-1	1
RH12 cubic trend	Non-directional: -M1 +3M2 -3M3 +M4 not equal 0 Directional: -M1 +3M2 -3M3 +M4 > 0				
SH	-M1 +3M2 -3M3 +M4 = 0 OR -1*M1 +3*M2 -3*M3 +1*M4 = 0	-1	3	-3	1

Exhibit 6 One Possible Set of Contrast Coefficients: Two-Way Analysis Of Variance

M1 = New treatment, AM
M2 = New treatment, PM
M3 = Comparison treatment, AM
M4 = Comparison treatment, PM

		M1	M2	M3	M4
RH13	Non-directional: The two treatments, averaged across the two different time periods, are not equally effective (M1+M2)/2 not equal (M3+M4)/2 Directional: The New treatment, averaged across the two different time periods, is more effective than the Comparison treatment (M1+M2)/2 > (M3+M4)/2				
SH	(M1+M2)/2 = (M3+M4)/2 OR (M1+M2) = (M3+M4) OR (M1+M2) - (M3+M4) = 0 OR 1*M1 +1*M2 -1*M3 -1*M4 = 0	1	1	-1	-1
RH14	Non-directional: The two time periods, averaged across the two treatments, are not equally effective (M1+M3)/2 not equal (M2+M4)/2 Directional: The AM period, averaged across the two different treatments, is more effective than the PM period (M1+M3)/2 > (M2+M4)/2				
SH	(M1+M3)/2 = (M2+M4)/2 OR (M1+M3) = (M2+M4) OR (M1+M3) - (M2+M4) = 0 OR 1*M1 -1*M2 +1*M3 -1*M4 = 0	1	-1	1	-1
RH15	Non-directional: The difference in effectiveness of the AM New treatment and the PM New treatment is different from the difference between the AM Comparison treatment and the PM Comparison treatment (M1 - M2) not equal (M3 - M4) Directional: The difference in effectiveness of the AM New treatment and the PM New treatment is greater than the difference between the AM Comparison treatment and the PM Comparison treatment (M1 - M2) > (M3 - M4)				
SH	The difference in effectiveness of the AM New treatment and the PM New treatment is the same as the difference between the AM Comparison treatment and the PM Comparison treatment (M1 - M2) = (M3 - M4) OR (M1 - M2) - (M3 - M4) = 0 OR 1*M1 -1*M2 -1*M3 +1*M4 = 0	1	-1	-1	1

stated this hypothesis with a directional expectation. (For example, "The difference between AM New treatment and PM New treatment is greater than the difference between AM Comparison treatment and PM Comparison treatment.") RH15 is referred to in the ANOVA literature as the "interaction" hypothesis.

An alternative way of stating this hypothesis is by looking at the differences within time, rather than within treatment: "The difference between AM New Treatment and AM Comparison Treatment is greater than the difference between PM New Treatment and PM Comparison Treatment. Both statements yield the same orthogonal coefficients, since they are the same question.

Summary

Discussing various sets of orthogonal comparisons for four groups should help illustrate the fact that there are many possible contrasts. The "appropriate contrast" depends on the design of the study and the research hypotheses of the researcher. While four groups were chosen for all the examples, the same conclusions can be developed for other numbers of groups. Four groups, though, does make the link to two-way ANOVA easy.

Few statistical texts make the link between orthogonal comparisons and the two-way ANOVA. Few also encourage directional hypothesis testing when there is one degree of freedom, as in the planned orthogonal comparisons. The reader is reminded that

although all these orthogonal comparisons (as well as many others) can be made on these four groups, only some of the comparisons make sense for any one design. For instance, trend analysis is appropriate to neither the teacher-time design in Exhibits 1-4, nor the two-way ANOVA design in Exhibit 6.

Information for Contributors

Multiple Linear Regression Viewpoints (*MLRV*), a publication sponsored by the American Educational Research Association's Special Interest Group on Multiple Linear Regression (SIG/MLR), is published once or twice per year to facilitate communication among professionals who focus their investigations on the theory, application, or teaching of multiple linear regression models or their extensions. Also, the journal accepts news items of interest to members of the SIG/MLR.

All manuscripts submitted to *MLRV* should conform to the language, style, and format of the *Publication Manual of the American Psychological Association* (3rd ed., 1983), available from Order Department, American Psychological Association, P. O. Box 2710, Hyattsville, MD 20784. Three copies of the manuscript, all double spaced (including equations, footnotes, quotes, and references) and accompanied by an abstract of 100 words or less, should be submitted to the editor at the address listed below. Mathematical symbols and Greek letters should be precise and clear and should leave no question as to interpretation. All figures must be camera ready. Manuscripts that do not conform to the above specifications may be returned to the author for style changes before the review process will begin. A submitted manuscript will receive a blind review from at least two members of the editorial board (except occasional invited contributions, letters to the editor, editorials, or news items). Any author identifying information should appear on the title page only. Efforts will be made to keep the review process to a maximum of eight weeks. The final version of an accepted manuscript should be submitted on a 3.5" disk, preferably in Apple Macintosh Microsoft Word, Version 5.1, although other formats might be acceptable. The editor reserves the right to make minor changes to an accepted manuscript in order to facilitate a clear and coherent publication.

Potential authors are encouraged to contact the editor to discuss ideas for contributions or to informally determine whether manuscripts might be appropriate for publication in *MLRV*. The editor also welcomes suggestions for debates, theme issues, other innovative presentation formats, and general inquiries about the journal. SIG/MLR news items should be sent to the editor as soon as they become available.

Manuscripts and other correspondents with the editor should be addressed to:

Ralph O. Mueller, Editor, *MLRV*
Department of Educational Leadership
Graduate School of Education and Human Development
The George Washington University
Washington, DC 20052

phone: (202) 994-4593
fax: (202) 994-5870
internet: RMUELLER @ GWIS2.CIRC.GWU.EDU

Graduate School of Education and Human Development
The George Washington University
Washington, DC 20052