STATISTICS FOR THE SOCIAL SCIENCES[1]

by
Henry F. Magalit[2]

FOREWORD

Some of the natural sciences have developed for centuries without the use of statistics. But this seems to be primarily a matter of good fortune or, to give these scientists credit for their own efforts, a relatively satisfactory control over disturbing elements of the environment. If carefully controlled laboratory conditions prevail there is less practical need for statistical techniques. To the social scientist, statistics is indispensable since he does not have control over all important relevant variables. He has to rely on statistical control since some variables are beyond his manipulative (or experimental) control. Unfortunately, in much social science research, control of the independent variables cannot be done experimentally (or by design) since his data is nonexperimental. The independent variables have already "occurred" and the investigator cannot control them directly by design.

This paper is an attempt to outline the statistical techniques that are available to the social scientists. However, since statistics may be considered a common kit of tools for describing and analyzing data of various disciplines, the statistical techniques outlined here can also be applied by the biologists and other natural scientists.

Introduction

The basic purpose of science is to explain natural phenomena

---

[1]SEARCA Professorial Chair Lecture paper, April 26, 1978, UPLB, College, Laguna.

[2]Associate Professor of Statistics, Department of Statistics and Statistical Laboratory, College of Arts and Sciences, University of the Philippines at Los Baños.

by discovering and studying the relations among the variables involved. Most social science research attempts to explain and/or predict a set of dependent variables, describing natural phenomena, using a set of "independent" variables. The set of independent variables is further divided into (a) variables that explain and/or predict the dependent variables and (b) control variables. The variables are related as given in figure 1.

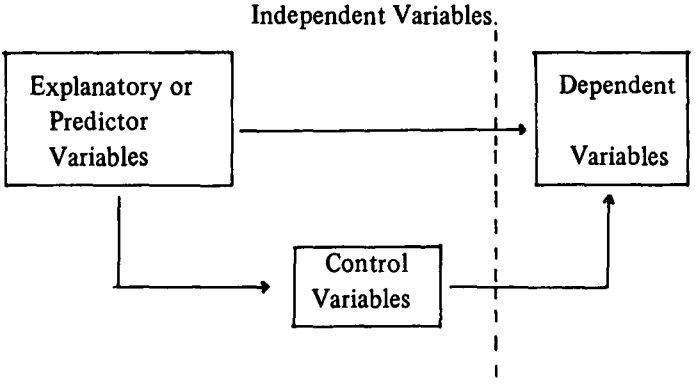

Figure 1

Since the statistical techniques that can be used depend on the levels of measurement, the researcher should endeavor to have his dependent variables measured numerically (interval or ratio scale) to be able to use powerful tests. Let us now have a preview of the topics presented in this paper.

We first considered the criteria in choosing an appropriate statistical test. Then the relation between measurements and statistics is discussed. Various statistical techniques are then presented and tabulated for convenience.

Various multivariate techniques are mentioned, starting with the most commonly used and often misused, the contingency tables. Multiple regression is next discussed using the generalized inverse procedure as well as the various techniques it uses such as all possible regression, stepwise regression, stagewise regression,

and autoregression. Problems of regression such as correlated predictor variables, nonlinearity, unequal variances and sampling design effect of complex survey data are discussed next.

The different techniques for reducing the number of variables are then presented; namely, principal component and factor analysis. Canonical correlation and multivariate analysis of variance are discussed very briefly. Strengths of multiple regression summarized. Finally, several examples are presented.

Choosing an Appropriate Statistical Test

In choosing a statistical test we must consider the manner in which the sample was drawn, the nature of the population from which the sample was drawn, the kind of measurement or scaling which was employed in the operational definitions of the variables involved, the hypothesis being tested and lastly, but not the least, the power of the test compared with alternative tests.

When we have asserted the nature of the population and the manner of sampling, we have established a statistical model. With every statistical test we associate a model and a measurement requirement. The model and the measurement requirement specify the conditions under which the test is valid. The conditions of the statistical model of a test are often called the "assumptions" of the test.

It is obvious that the weaker the assumptions that define a particular model, the less qualifying we need to do about our decision arrived at by the statistical test associated with the model.

However, the most powerful tests are those which have the strongest assumptions. The t or F tests both parametric tests, for example, are the most likely of all tests to reject H_o when H_o is false if the assumptions are tenable. That is, when research data may appropriately be analyzed by a parametric test, that test will be more powerful than any other in rejecting H_o when it is false. However, measurement must be at least in an interval scale for parametric tests to be valid. Different tests require measurement of different levels. The basic notions in the theory of measurement

must be known in order to understand the measurement test requirements of the various statistical tests.

Measurement and Statistics

The levels of measurement are: nominal, ordinal, interval and ratio. The arithmetic operations allowable on a given set of scores are dependent on the level of measurement.

It should be noted that these various levels of measurement themselves form a cumulative scale. An ordinal scale possesses all the properties of a nominal scale plus ordinality. An interval scale has all the properties of an ordinal scale plus a unit of measurement. The cumulative nature of these scales means that it is always legitimate to drop back one or more levels of measurement on analyzing the data. Sometimes this will be necessary when statistical technique are either unavailable or unsatisfactory in handling the variable of a high level of measurement. However, we lose information in doing so because we can no longer consider differences.

Can we go up the scale of measurement say, from ordinal to an interval scale? We are often tempted to do so since we would be able to make use of more powerful statistical techniques. The use of a particular statistical (mathematical) model presupposses that a certain level of measurement has been obtained. If statistical techniques which assume strong levels of measurement are used on weak levels of measurement, the result of which should be considered approximate rather than exact and some caution must be exercised in making conclusions.

Statistical Techniques

Let us now consider the two sets of tests; namely, parametric and nonparametric. A parametric statistics test is a test whose model specifies conditions about the parameters of the population from which the research sample was drawn. Since these conditions are not ordinarily tested, they are assumed to hold. The meaningfulness of the results of a parametric test depends on the

validity of these assumptions. A non-parametric statistical test uses a model which does not specify conditions about the parameters of the population from which the sample was drawn. Nonparametric tests do not require measurement so strong as that required for the parametric tests which is at least an interval scale. There are nonparametric tests for both ordinal and nominal scale. The t-test assumes that the observations in the sample come from a normally distributed population and measured at least in an interval scale. If these two assumptions are not tenable, for the one-sample case we have the Binomial and Chi-square tests when the measurement is nominal; and Kolmogorov-Smirnov test when the measurement is ordinal.

For the two-sample case and when the samples are related, we have the McNemar test, when the measurement is nominal; and the Wilcoxon signed-rank tests when the measurement is ordinal. Also when we are not willing to assume normality, we can use the Walsh test and the randomization test for matched pairs. We have the matched pairs t-test as the parametric counterpart of all these tests.

When the two samples are independent and the variances of the two populations are assumed equal we have the parametric t-test. When the variances of the populations are assumed unequal we have the weighted t-test of Cochran and Cox. We have a lot of nonparametric tests for this case. The most popular of these are the Mann-Whitney U and Kolmogorov-Smirnov tests.

For three or more populations and the samples are related we have the two-way classification analysis of variance (AOV) and when the samples are independent we have the completely randomized design or one-way classification AOV. Both are parametric and use analysis of variance technique. When the samples are related and the measurement is nominal we have Cochran Q test; and when the scale is ordinal we have Friedman two-way analysis of variance. When the samples are independent and the scale is nominal we can again use the Chi-square test; and when the scale is ordinal we have Kruskal-Wallis one-way analysis of variance.

For nonparametric measures of correlation we have the contingency coefficient when the scale is nominal; and Spearman and Kendall rank correlations when it is ordinal. If we want the correlation of two variables in the ordinal scale holding the rest of the variables fixed we have Kendall partial rank correlation. However, we do not have test of significance using Kendall's partial rank correlation, since its sampling distribution is not known.

The parametric measures of correlations are Pearson's product-moment (simple) correlation coefficient, partial correlation, multiple correlation coefficients, and canonical correlations. Factor analysis can also be used to examine patterns of correlation or dependence structure. Patterns of correlation are also examined using partial correlations; however, these are in terms of the observed variables rather than the conceptual variables used in factor analysis. Partial correlation will give us an estimate of the strength of association between any two variables removing the effect of the other variables. Multiple correlation measures the association between the dependent variable and a set of predictor variables. The correlation between two sets of variables are known as cannonical correlation.

For examining the dependence structure of the variables (dependent as well predictor) we have principal component analysis and factor analysis. The principal component is used to find the linear combination of the variables with "large" variance while factor analysis is a method of reducing a large number variables to a small number called latent variables or factor.

In problems of explanation and prediction where we are interested in one dependent variable as function of several predictor variables we have multiple (linear) regression which uses as tools (a) all possible regression (b) stepwise regression (c) stagewise regression (d) autoregression and (e) weighted regression. We also have nonlinear regression if the relation between y and the predictor variables is nonlinear. For discrimination classification problems we have discriminant analysis and cluster analysis. Finally when we have more than one dependent variable, multiple

NONPARAMETRIC STATISTICAL TESTS

LEVEL OF MEASUREMENT	*One-Sample Case*	*Two-sample case*		*K-sample case*		*NONPARAMETRIC MEASURES OF CORRELATION*
		Related Samples	*Independent samples*	*Related samples*	*Independent samples*	
NOMINAL	*Binomial* χ^2	*McNemar*	*Fisher* χ^2	Cochran Q	χ^2	Contingency
ORDINAL	Kolmogorov-Smirnov Runs	Sign Wilcoxon	Median Mann-Whitney U Kolmogorov Smirnov Wald-Wolfowitz Moses	Friedman two-way AOV	Median Kruskal-Wallis one-way AOV	Spearman rank correlation Kendall rank correlation Kendall partial rank correlation
INTERVAL OR		Walsh Randomization	Randomization			
RATIO		*PARAMETRIC STATISTICAL TESTS*				*MEASURES OF CORRELATION*
	Z-test t-test	paired t-test	t-test weighted t-test	RCB Two-way classification AOV	CRD or One-way classification AOV	Simple correlation Partial correlation Multiple correlation

MULTIVARIATE STATISTICAL TECHNIQUES

Multiple (Linear) Regression
- a) Multiclassification AOV
- b) All possible Regression
- c) Stepwise Regression

Multivariate Analysis of Variance
Multivariate Contingency Analysis
Principal Component Analysis
Factor Analysis

linear regression analysis may be generalized to multivariate analysis of variance.

Multivariate Contingency Tables

Data which are categorical are usually presented in two-way tables. Sometimes three-way, or even four-way tables, have been given but difficulties of tabulation, printing and especially interpretation have prevented, or at least restricted, tabulations by more than two variables at a time.

There are three fundamental problems in the analysis of multiple contingency. One is to set up a measure of relationship between two or more variables. This is usally done by the use of the χ^2 statistic. The other is to find one's way through a maze of possible hypotheses in a systematic manner. Lastly, there may be a shortage of cases and it is frequently necessary to control one relevant variable at a time. When we come to a three-way table, (say of variables A, B, C) there are 17 hypotheses to test, of which 4 are trivial. They may be exhibited as follows:

A	A,B	A,B,C	AB	AB,C	AB,AC	ABC
B	A,C		AC	AC,B	AC,BC	
C	B,C		BC	BC,A	BC,AB	

Here, for example, A,B refer to a hypothesis based on fixing the univariate margins of A and B. AB represents the hypothesis that the entire three-way table is determined by the joint distribution of A and B. AB,AC is a test fixing the two-way margins AB and AC. The test that a single variable A explains the entire table is trivial: it simply tests whether the frequencies in the same category of A are all equal within sampling limits. Similarly for B and for C. Likewise model ABC (which is added for completeness) requires no test because it fixes all the cells in the table; it is referred to as the "saturated" model. The other 13, however, may be of interest. For example, a test based on AB,C is similar to the test of partial correlation — are A and B dependent when the effect of C is abstracted?

The number of possibilities to be examined increases alarmingly with the number of dimensions. For four-way tables there are 167 and for five-way tables there are thousands.

The subject developed fast recently and there exists an extensive literature on the subject. Reference may be made to Goodman (1973), Plackett (1974), Kendall (1975), and Bishop et al (1975).

Multiple Regression

Consider the linear regression model in matrix notation.

$$Y = X\beta + \epsilon \qquad (1)$$

where Y is an n x 1 random observed vector of the dependent variable

X is an n x p matrix of known fixed quantities

β is a p x 1 vector unknown parameters

ϵ is a n x 1 random vector n is the sample size and $p \leqslant n$.

We assume the following for estimation purpose:

$$E(\epsilon) = 0, E(\epsilon \epsilon') = I\sigma^2$$

and X has rank $k \leqslant p$ and σ^2 is unknown.

By least squares the resulting normal equations are:

$$(X'X)\,\tilde{\beta} = X'Y. \qquad (2)$$

A solution for the above is

$$\hat{\beta} = (X'X)^{-}X'Y \qquad (3)$$

where $(X'X)^{-}$ is a generalized inverse of $X'X$ if $k < p$. If $k = p$ then $(X'X)^{-}$ is equal to $(X'X)^{-1}$, the ordinary inverse.

One advantage of using a generalized inverse in solving (2) is that no restrictions are needed even with nominal predictor variables to be able to get a solution. Another advantage is that it can handle multicollinearity problem of regression.

With nominal predictor variables we may solve (2) without using a generalized inverse by imposing p – k restrictions on it, (Magalit, 1977). It can be shown that these two methods are equivalent and that by imposing restrictions on (2) we can use any ordinary multiple regression computer program without the added complexity of having to use a generalized inverse procedure.

Let us take a look at one property of a generalized inverse. A generalized inverse has the following property.

$$(X' X)\ (X' X)^{-} = \begin{bmatrix} 1 & 0 & 0 \ldots 0 \\ 0 & 1 & 0 \ldots 0 \\ 0 & 0 & 1 \ldots 0 \\ 0 & \ldots & \ldots 0 \end{bmatrix} \qquad (4)$$

$p \times p \quad p \times p \qquad\qquad p \times p$

where X' X is a p x p matrix of rank $k < p$. The right-hand-side of equation (4) is a diagonal matrix with diagonal elements of k ones and p – k zeroes, while the off-diagonals are all zeroes.

With nominal predictor variables the estimate of each regression coefficient is biased. However, there are linear functions of $\hat{\beta}$ that are not biased and these are the functions that researchers are usually interested in anyway. For example, the contrast of the regression coefficients of a nominal variable is unbiased. (See Magalit, 1977).

Let us now discuss how we will select the numerical predictor variables in regression.

In a theoretical sense the all possible regression is best in that it enables us to "look at everything". It can be recommended if we have not more than five predictor variables. However, for 10 predictor variables, the amount of computer time and the sheer physical effort of examining all the computer printouts is enormous since there are 1023 equations to look at.

The stepwise is one of the best variable selection procedure and its use is recommended. However, it can easily be abused by the user. As with all the procedures, sensible judgment is still

required in the initial selection of variables and in the critical examination of the model through examination of residuals whenever feasible.

Complexity is encountered in realistic nonexperimental research where one needs to allow for reciprocal causation. We have assumed that the choice of dependent variables is not problematic and that there are no feedback effects from the dependent to predictor variables. Although we have allowed for the possibility of intercorrelated independent variables by using the stepwise procedure, we have not considered models that attempt to account for these intercorrelations by taking some of the "predictor" variables as functions of the others. These topics have been studied in considerable detail by econometrician in connection with simultaneous equations, (Johnton, 1972). This type of problem is handled by stagewise regression technique.

In time series data where the errors are correlated we have auto-regression technique, (Fuller, 1971).

For a comprehensive discussion on multiple regression in behavioral research, see Kerlinger and Pedhazur (1973).

If the functional relation between numerical predictor variables and the dependent variable is not known, the nonlinearity in regression may be tested by the following approximate F test.

Non-Linearity in Regression

If the regression equation happens to be linear in form, we can expect that the $\overline{Y}_{.j}$, mean for given X, will all fall approximately on the least-squares lines so that it will make little difference whether deviations are taken about the category means or the least-squares line. Kalton (1966) has shown that one loses very little precision by categorizing a numerical variable into a set of classifications. If, for instance, a predictor has a linear relationship with a dependent variable accounting for P percent of its variance, then a categorization into as few as five subclasses of the predictor variable will account for ninety-five percent of that potential P percent, and ten subclasses will account for ninety-nine percent of

the total possible. And if, in fact the relationship is not linear, a categorized variable may easily account for more of the total variance than a linear regression using the full numerical detail. If the relationship is non-linear, then for at least some of the categories, the sum of squares about the category mean will be quite a bit smaller than that the least-squares line. Thus the proportion of variation explained by the categories will be larger than the proportion explained by the least-squares line unless the true relationship is linear.

By fitting a categorized predictor variable we have

$$E^2 = \frac{\text{Among Categories SS}}{\text{Total SS}} \qquad (5)$$

and by fitting a least-squares line we have

$$r^2 = \frac{\text{Regression SS}}{\text{Total SS}} \qquad (6)$$

If the relationship between the predictor and the dependent variable is non-linear $E^2 > r^2$ and we can test for non-linearity by the following analysis of the variance table with one predictor variable.

Table 1. Analysis of Variance Test for Non-Linearity

SV	df	SS	MS	F
Total	$n-1$	$\Sigma(y-\bar{y})^2$		
Explained by linear model	1	$r^2\Sigma(y-\bar{y})^2$		
Additional explained by non-linear model	$k-2$	$(E^2-r^2)\Sigma(y-\bar{y})^2$	$\frac{(E^2-r^2)\Sigma(y-\bar{y})^2}{k-2}$	$\frac{(E^2-r^2)(n-k)}{(1-E^2)(k-2)}$
Error	$n-k$	$(1-E)^2\Sigma(y-\bar{y})^2$	$\frac{(1-E^2)\Sigma(y-\bar{y})^2}{n-k}$	

The above analysis can easily be extended to p predictor variables as in the example to be given later. However, it must be emphasized that the test is only an approximate one.

Correlated Predictor Variables

If predictors are positively correlated to one another and both positively (or negatively) correlated with y, they overlap since the two considered together would explain less than the two, each considered separately. Figure 2 presents this schematic form. The total area covered by the two circles is less than the sum of their individual areas because they overlap.

On the other hand, it works in the opposite fashion if the predictors are negatively correlated to each other and are positively (or both negatively) correlated to the dependent variable. Together they would explain more of the variations in the dependent variable than the two considered separately.

Both of the cases can be explained fully by the following equation.

$$R^2_{y.12} = \frac{r^2_{y1} + r^2_{y2} - 2r_{y1}\, r_{y2}\, r_{12}}{1 - r^2_{12}} \qquad (7)$$

where R_{y12} is the multiple correlation of Y with X_1 and X_2.

r_{y1} is the correlation between Y and X_1.

r_{y2} is the correlation between Y and X_2.

r_{12} is the correlation between X_1 and X_2.

From the equation we can see that as r_{12} approaches 1, $R^2_{y.12}$ will be greater than 1 and the error sum of squares will be negative. This is a collinearity problem and only one of the predictor variable can be used in the equation. This problem can be remedied by the use of a generalized inverse procedure in

solving the normal equations. The stepwise regression procedure may be used also for this type of problem.

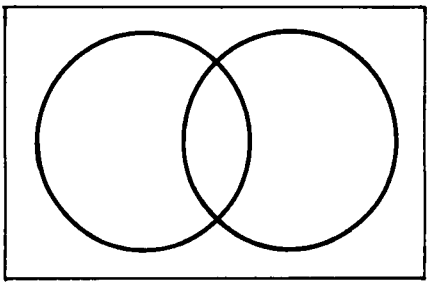

Figure 12
Correlated Predicators

In practice however, perfect multicollinearity, $r_{12} = 1$, is easier to detect and seldom happens. The more problematic case is where r_{12} is say .90 or .80, in which case $R^2_{y.12} < 1$, and are hard to detect.

Weighted Regression and Sampling Design Effect

Consider again the matrix linear model (1).

$$Y = X\beta + \epsilon$$

where Y is an n x p random observed vector of the dependent variable

X is an n x p matrix of known fixed quantities

β is a p x 1 vector of unknown coefficients (parameters)

ϵ is an n x 1 random vector

and n is the sample size.

$$\text{Let } W = \begin{pmatrix} w_1 & & & \\ & w_2 & & \phi \\ & & w_3 & \\ & \phi & & \ddots \\ & & & \ddots \\ & & & w_n \end{pmatrix} \qquad (8)$$

where W is a diagonal n x n matrix and

w_i is the weight associated with the i^{th} sample.

By least squares; that is, minimizing $\sum_i^n w_i (Y_i - \hat{Y}_i)^2$, we arrive at the following normal equations:

$$(X' WX)\beta = X' WY \tag{9}$$

A solution for the above equations is

$$\hat{\beta} = (X' WX)^- X'WY \tag{10}$$

where $(X'WX)^-$ is a generalized inverse of X'WX.

The above is encountered in survey data using unequal probabilities of selection of the population units. With unequal inclusion probabilities π_i, i = 1,2, . . , N and using a simple linear regression model one can easily arrive at the estimate of the slope of the line as

$$b = \frac{(\sum_i w_i)\ \sum^n w_i X_i Y_i - (\sum^n w_i X_i)\ (\sum^n w_i Y_i)}{(\sum^n w_i)\ \sum^n w_i X_i^2 - (\sum^n w_i X_i)^2} \tag{11}$$

and the intercept as

$$a = \frac{\Sigma w_i Y_i - b \Sigma w_i X_i}{\Sigma w_i} \tag{12}$$

where $w_i = \frac{1}{\pi_i}$

However, even for such a simple model the distributions of b and a are already extremely difficult to find. As can be seen from the above, inclusion of the survey design effect results in not only complicated estimates but whose distributions are unknown. However, there are now computer-based techniques for computing standard errors of the estimates as mentioned in David's (1977) paper. These are pseudo-replications (McCarthy, 1969), the balanced repeated replications (Kish and Frankel, 1974), and the Jacknife method (Cochran, 1977).

Another case of the weighted least squares is encountered when the observations have unequal variance and sometimes even correlated. Let V be the variance-covariance matrix of the ϵ Assuming that V^{-1} exist the least squares estimate of β is

$$\hat{\beta} = (X' V^{-1} X)^{-} X' V^{-1} Y \tag{13}$$

where $(X' V^{-1} X)^{-}$ is a generalized inverse of $X' V^{-1} X$. If the weighted least squares analysis were called for but an ordinary least squares were performed, the estimates obtained would still be unbiased but would not have minimum variance, since the minimum variance estimates are obtained from the correct weighted least squares analysis (Draper and Smith, 1966).

An interesting, but complicated problem is to combine both the survey design effects and the knowledge that the error variances are unequal or even correlated. A least squares estimate of β is

$$\hat{\beta} = (X' WV^{-1} X)^{-} X' WV^{-1} Y \tag{14}$$

In practical problems it is often difficult to obtain specific information on the form of V at first. For this reason it is sometimes necessary to make the (known to be erroneous) assumption $V = I\sigma^2$ and then attempt to discover something about the form of V by examining the residuals from the regression analysis. Anscombe and Tukey (1963) and Anscombe (1961) give some statistics for analyzing residuals. They deal with three types of discrepancies; namely, (a) variance not constant (b) linear effect of X not removed and (c) nonaddivity of the model. They use the following statistics (a) $T_{21} = \sum_i^n e_i^2 \hat{Y}^1$ (b) $T_{11} = \sum_i^n e_i \hat{Y}_i$ and (c) $T_{12} = \sum_i^n e_i \hat{Y}^2$ for examining the foregoing discrepancies, respectively. These statistics can also be used for complicated models. The normality assumption can also be tested by the use of residuals.

Principal Component Analysis

An examination of the individual correlations between pairs of predictor variables is not sufficient. What is required is an analysis of the whole set of correlations. One of the best method is to compute the correlations matrix of the predictors and to determine the constant known as latent roots or eigenvalues. Any zero eigenvalue will imply a linear relations among some of the predictors variables and therefore a redundancy among them. A small eigenvalue indicates near collinearity among them and warns that the coefficients are inflated. Should this situation arise it is preferable to delete some of the predictor variables.

One important use of the principal component technique is that of summarizing most of the variation in a multivariate system in fewer uncorrelated variables. By partitioning the total variance of all variables into successively small portions we determine the new set of fewer uncorrelated variables called components. The factorization of the covariance (or correlated) matrix is brought about by transformation rather than as the result of a fundamental model for covariance technique used in factor analysis. The components are not invariant under changes is scale of the variables.

Factor Analysis

This is a method for reducing a large number of variables to a small number of presumed underlying latent variables called factors. Factors are usually derived from the intercorrelations among the variables. If the correlations among the variables, are zero or near-zero no factors can emerge. If, on the other hand, the variables are substantially correlated, one or more factors can emerge. It is a powerful tool for discovering underlying relations among the variables and a multivariate method related to multiple regression analysis. Each factor is a linear combination of the variables.

Factor A, for instance, can be written as

$A = \sum_{i}^{k} a_i x_i$, a_i is factor loading of i^{th} variable.

Thus the factors can be called dependent variable and the variables as independent variables.

But the situation can be viewed differently. Any variable can be conceived as a linear combination of the factor, say

$$X_i = a_i A + a_2 B + \ldots\ldots\ldots$$

Here the factors are viewed as independent variables and the variables as the dependent variable. Again, the similarity to multiple regression is obvious.

Although factor analysis and multiple regression analysis resemble each other, in general the two methods have quite different purposes. Multiple regression's fundamental purposes are to predict dependent variables and test research hypotheses. Factor analysis, is used to discover unities or factors among many variables and thus to reduce many variables to fewer underlying variables or factors. Multiple regression explains a single known, observed and measured dependent variable through the predictor (independent) variables. Factor analysis explains many variables, usually without independent and dependent variable distinction, by showing their basic structure, how they are similar, and how they are different. In addition, factor analysis almost always seeks to name the components of the structure, the underlying factors. This is an important scientific purpose.

There is now a maximum-likelihood procedure for estimating the parameters from the model. Furthermore, the goodness of fit of a solution with just m factors could be tested vigorously by the generalized likelihood-ratio principle. Estimates of the parameters can also be derived from the model without assuming multi-normality in the model. Finally, for time series data a model for covariance structure that links the observed variables to a latent stochastic process is also possible, (Morrison, 1967).

Cannonical Correlation

Cannonical correlation is the generalization of multiple regression analysis to any number of dependent variables. This is not

a large conceptual step but rather a large computational step. Except for the simplest problems, it is so complex to make desk calculator calculations forbidding. Intelligent and critical reliance on the computer is necessary.

It is multiple regression analysis with p categorical variables and $m > 1$ dependent variables. Through least squares analysis, two linear composites are formed, one for the predictor variables, X_i, and one for the dependent variables, Y_j. The correlation between these two composites is the cannonical correlation, R_c. It's square, R_c^2 is an estimate of the variance shared by the two composites.

Like the coefficient of multiple correlation, the cannonical correlation coefficient is the maximum correlation possible between the two sets of variables. It also uses the least squares procedure that seeks the regression weights (coefficient) to be attached to each variables of both sets of variables. However, the weights become a problem when more than one cannonical correlation is calculated from the same set of data. The weights then must be interpreted with great caution.

Multivariate Analysis of Variance

Analysis of variance with any number of independent variables and any number of dependent variables is called multivariate analysis of variance. Like univariate analysis of variance, its purpose is basically to test statistical hypothesis about group means of more than one dependent variable. In univariate analysis of variance, the total sum of squares is partitioned into groups and within groups sum of squares. In multivariate analysis of variance, say for two independent variables the sum of products, $\Sigma\ Y_1\ Y_2$ is also partitioned according to the independent variables into between groups and within groups sums of product. The test of statistical significance is used to determine whether the mean of the two dependent variables, considered simultaneously are equal. A multivariate F test, test the significance of mean differences m dimensionally, in this case two-dimensionally.

Discriminant Analysis

In discrimination we are given the existence of two or more populations (groups) and a sample of individuals from each. The problem is to set up a rule, based on measurements from these individuals, which will enable us to allot some new individuals to the correct population (group) when we do not know from which it emanates. We set up a function which will enable us to allocate any freshly observed individual to the correct population (group).

The general problem is to set up a function which will give the smallest possible frequency of misclassification when used as a means of discrimination. Such a linear combination is termed a discriminant function of the form:

$$Z = \lambda_1 X_1 + \lambda_2 X_2 + \ldots\ldots + \lambda_p X_p$$

where Z is the standardized index

$X_1, X_2, \ldots\ldots, X_p$ are the variables measured and

$\lambda_1, \lambda_2, \ldots\ldots, \lambda_p$ are the corresponding weights.

A test can be made of the significance of the discriminant function by means of the F test or Chi-square test.

As in regression there is also a problem of determining the significance of the discriminating value of each X_i. This is done by partial F tests.

Strengths of Multiple Regression Analysis

It is able to explain and/or predict events in the world in which we live.

It is able to handle any number and kind (level of measurement) of predictor variables.

It is often the best method of analysis of non-experimental data. It can be used to control the effects of some variables while studying the effects of other variables on the dependent variable. Frequency and cross-tabulations analysis may be used but the most it can intelligibly do is to present relations among three variables at a time and such a three-way cross-tabulations are

difficult to grasp and interpret. It can answer certain questions in roundabout and often clumsy way.

It opens up research possibilities not available, at least not generally and readily available, in the past by the use of dummy variables in regression.

It is rich of various statistics to be used in the interpretation of the data: namely, (a) R^2 measures the overall relation between the dependent variable and the predictor variables, in terms of proportion of variations accounted for by all the predictor variables or any subject of them.

The proportion of variance of any subset of predictor variables can be tested for statistical significance using an F test. That is, any set of variables (on each variable) may be tested whether it significantly affects the dependent variable or not and (b) estimates of regression coefficients are available and corresponding t tests of their statistical significance.

Data in Examples

No data used in this paper consists of two independent random samples of 225 trainees and 225 non-trainees, a survey conducted by the National Manpower and Youth Council in 1977 under the direction of Ms. Racquel B. Goodrich.

The following socio-economic variables were gathered:

- Y – Wage after training (or Jan. 1977), the dependent variable.
- X_1 – Educational level
- X_2 – Work experience
- X_3 – Wage before training (before Jan. 1977)
- X_4 – Mobility
- X_5 – Need achievement score
- X_6 – Trainability
- X_7 – Type of industry
- X_8 – Migration Experience

X_9 – Access to government agency

X_{10} – Regional Background

X_{11} – Socio-economic status of trainees

X_{12} – Socio-economic status of guardian

In example 1, the regression model for the trainees is presented while the model for non-trainees is not included. Since the distribution of the dependent variable, Y, is highly skewed and has zero values, Y is transformed to log (Y + 1).

Example 1
Nonlinearity in Regression
Analysis of Variance

SV	*df*	*SS*	*MS*	*F*
Total	224	1933.6632		
Explained by linear model	12	1577.5055		
Additional explained by nonlinear model	33	344.6523	10.4440	162.42**
Error	179	11.5053	.0643	

E^2 = .9945 from nominal regression model

R^2 = .8158 from numerical regression model

Additional explained by nonlinear model SS:

$$(E^2 - R^2) \Sigma (Y - \bar{Y})^2 = 1922.1578 - 1577.5055 = 344.6523$$

Since nonlinearity is highly significant it is recommended to use nominal predictor variables rather than numerical predictors variables in the model. This result shows that if one is in doubt of the linearity of the relation between the predictor variables and

the dependent variable it is better to use nominal predictor rather numerical predictor variables in regression. This verifies Kalton's result.

Nominal Variables Regression Model for Trainees
Analysis of Variance

SV	*df*	*SS*	*MS*	*F*
Total	224	1933.6632		
Categorical Variables	45	1922.1578	42.7146	664.55**
Error	179	11.5054	.0643	

Since "due to regression" is highly significant, a closer look at the twelve nominal variables is necessary. This is given in the following analysis of variance, the procedure of which was discussed in Magalit (1977).

Analysis of Variance for
Each Nominal Variable

SV	*df*	*Adjusted SS*	*MS*	*F*
X1	3	.0863	.0288	<1
X2	4	.2018	.0504	<1
X3	4	4.5219	1.1305	17.59**
X4	4	.7495	.1974	2.92*
X5	4	.2507	.0627	<1
X6	2	.0692	.0346	<1
X7	5	835.5285	167.1057	2599.82**
X8	1	.0003	.0003	<1
X9	9	1.2753	.1417	2.20*
X10	1	.0324	.0324	<1
X11	4	.4470	.1118	1.74
X12	4	.7716	.1929	3.00*
Error	179	11.5053	.0643	

The following variables significantly affected the wage after training:

X_3 wage before training

X_4 mobility

X_5 type of industry

X_9 access to government agency

X_{12} socio-economic status of guardian

Scheffe's tests may be done for each of the above significant nominal variables, (Magalit, 1977). The t tests for the k categories of a nominal variable are not recommended because the tests are not independent.

Looking at X_4, mobility, no significant difference was found using Scheffe's tests while several significant differences were found using t tests.

Example 2
Discriminant Analysis

A discriminant function was derived for the two groups, trainees and non-trainees.

The discriminant function is as follows:

$$Z = -.6661\ X_5 + .3683\ X_7 - .2827\ X_1 + .2389\ X_3$$

$$- .2156\ X_6 - .1224\ X_4 - .1078\ X_8 + .0697\ X_9$$

To test for the significance of the function, the chi-square statistic was used.

$X^2_{comp} = 224.88$ $\qquad$ $X^2.05\ (8) = 15.507$

Since $X^2_{comp} > X^2$ tab, therefore the discriminant function could be used to discriminate between the two groups, trainee and non-trainee and that the function could be used as well as to

classify an unknown individual to the group to which it best belongs.

The dividing point between the two groups was determined by taking the mean of the groups (trainees is group 1 and non-trainees is group 2). The dividing point is 32.4625, mean of Z.

Therefore if an unknown individual has measurements $X_1^*, X_3^*, \ldots, X_9^*$ and we wish to know to which group (1 or 2) this unknown individual belongs we compute Z*:

$$Z^* = \sum_{i=1}^{9} \lambda_1 X_i^* \qquad , i \neq 2.$$

and compare it with the dividing point. If Z* is less than or equal to 32.4625, then the unknown individual belongs to group 1, trainees, and if Z* is greater than 32.4625, then it belongs to group 2, non-trainees.

Of the original 12 variables, only 8 were found to be significant and sufficient to discriminate between the two groups, trainees and non-trainees.

Selected Bibliography

Ascombe, F. I. (1961) – Examination of Residuals – *Proceeding of the Fourth Berkeley Symposium on Mathematical Statistics and Probability 1-36.*

Ascombe, F. I. and J.W. Tukey (1963) – *Technometrics* 5, 141-160.

Bishop, Y. M. et al (1975). *Discrete Multivariate Analysis.* MIT Press.

Cochran, W. G. (1977). *Sampling Techniques.* Third Edition. John Wiley.

David, I. P. (1977). *Analytic Use of Survey Data: Some Current Issues and Problems.* Paper for the Symposium on Household Economics, Manila, May 1977.

Draper, N. R. and H. Smith (1966). *Applied Regression Analysis.* Wiley and Sons. Inc.

Fuller, W. (1971). *Introduction to Statistical Time Series.* Ames: Iowa State University Bookstore.

Galant, A. R. (1975). *Nonlinear Regression.* American Statistician, Vol. 28.

Goodman, L. A. (1973). Guide and Unguided Methods for the Selection of Models for a Set of T Multidimensional Contingency Tables, 68: 165-175. *Journal of American Statistical Association.*

Johnston, J. (1972). *Econometric Methods.* Second Edition. New York: McGraw Hill Book Company, Inc.

Kalton, G. A. (1966). *A technique for Choosing the Number of Alternative Response Categories to Provide in Order to Locate an Individual's Position on a Continuum.* Memos dated, Feb. 10, 1967 and Mar. 10, 1967. Ann Arbor Michigan, Sampling Section, Survey Research Center.

Kendall, M. G. (1975). *Multivariate Analysis.* London: C. Griffin and Co.

Kerlinger, F. N. and Pedhazur (1973). *Multiple Regression in Behavioural Research.* New York: Holt, Rinehert and Winston, Inc.

Kish, L. and M. R. Frankel (1974). Interference from Complex Samples. Journal of the Royal *Statistical Society,* B, 36, 1-37.

Magalit, H. F. (1977), *Nominal Variables in Regression,* SEARCA Professorial Chair Lecture paper, April 22, 1977.

McCarthy, P. J. (1969). Pseudo-Replication: Half Samples. *Review of the International Statistical Institute* 37, 239-263.

Morrison, D. F. (1967). *Multivariate Statistical Methods.* New York: McGraw-Hill Book Co.

Plackett, R. L. (1974). The Analysis of Categorical Data. New York: Hafner Press.

Selected Bibliography

Ascombe, F. I. (1961) – Examination of Residuals – *Proceeding of the Fourth Berkeley Symposium on Mathematical Statistics and Probability 1-36.*

Ascombe, F. I. and J.W. Tukey (1963) – *Technometrics* 5, 141-160.

Bishop, Y. M. et al (1975). *Discrete Multivariate Analysis.* MIT Press.

Cochran, W. G. (1977). *Sampling Techniques.* Third Edition. John Wiley.

David, I. P. (1977). *Analytic Use of Survey Data: Some Current Issues and Problems.* Paper for the Symposium on Household Economics, Manila, May 1977.

Draper, N. R. and H. Smith (1966). *Applied Regression Analysis.* Wiley and Sons. Inc.

Fuller, W. (1971). *Introduction to Statistical Time Series.* Ames: Iowa State University Bookstore.

Galant, A. R. (1975). *Nonlinear Regression.* American Statistician, Vol. 28.

Goodman, L. A. (1973). Guide and Unguided Methods for the Selection of Models for a Set of T Multidimensional Contingency Tables, 68: 165-175. *Journal of American Statistical Association.*

Johnston, J. (1972). *Econometric Methods.* Second Edition. New York: McGraw Hill Book Company, Inc.

Kalton, G. A. (1966). *A technique for Choosing the Number of Alternative Response Categories to Provide in Order to Locate an Individual's Position on a Continuum.* Memos dated, Feb. 10, 1967 and Mar. 10, 1967. Ann Arbor Michigan, Sampling Section, Survey Research Center.

Kendall, M. G. (1975). *Multivariate Analysis.* London: C. Griffin and Co.

Kerlinger, F. N. and Pedhazur (1973). *Multiple Regression in Behavioural Research.* New York: Holt, Rinehert and Winston, Inc.

Kish, L. and M. R. Frankel (1974). Interference from Complex Samples. Journal of the Royal *Statistical Society,* B, 36, 1-37.

Magalit, H. F. (1977), *Nominal Variables in Regression,* SEARCA Professorial Chair Lecture paper, April 22, 1977.

McCarthy, P. J. (1969). Pseudo-Replication: Half Samples. *Review of the International Statistical Institute* 37, 239-263.

Morrison, D. F. (1967). *Multivariate Statistical Methods.* New York: McGraw-Hill Book Co.

Plackett, R. L. (1974). The Analysis of Categorical Data. New York: Hafner Press.