COURSE OBJECTIVES - REGRESSION MODELS AND COMPUTERS

This course is designed to develop capabilities in the use of a general regression model approach in the analysis of research problems with the assistance of a high-speed digital computer. The presence of the IBM 7040 computer at the Personnel Research Laboratory enables research workers to think of problems in a different way than is possible without the computer. Experience with electronic computers in the solution of research problems has revealed that it is not only necessary to have a precise language with which to communicate with computers, but also a rigorous yet flexible language for effectively formulating analysis of research problems. The purpose of the course is to provide formal instruction and intensive supervised practice in the use of vector concepts in multiple linear regression models and FORTRAN computer programming in behavioral science research problem analysis. The participants in this course will also work with a FORTRAN-oriented computer subroutine system named PERSUB. The most important objective of this course is to allow research workers to design a model that is appropriate to the analysis of their particular problem rather than requiring the research worker to merely fit problems into existing models. Each participant will carry out several complete analyses of research data. This will involve formulation of an appropriate regression model and programming the computer to obtain the final results. Emphasis will be placed on practical applications of general principles.

<u>Topics to be Covered</u>:

1. Vector concepts needed in research problem analysis such as linear combination of vectors and orthogonal decomposition of vectors.

2. Computational aspects of regression analysis.

3. Foundations of regression analysis.

4. Programming in the FORTRAN language with emphasis on subroutine development.

5. Relationships which exist between regression models and techniques commonly called by other names such as analysis of variance and co-variance, pattern analysis, correlation ratio, etc.

Use of Unit Vector and Other Comments on PERSUB Regression Program

By Robert A. Bottenberg

Questions occasionally arise regarding the necessity for specifying a unit vector as input data when using Persub regression routines. In order to clarify the role of the unit vector, it will be helpful to distinguish between (a) a conceptual regression model together with an implied computational procedure for its solution, and (b) the version of the model and the computational procedure used by the Persub routines. To review first, what appears in the usual regression model is a set of variables in their raw score form. These consist of observed values, transformations of them such as squares, cubes, products, square roots, logs, etc., and coded binary variables to represent group membership or the presence or absence of some characteristic. Also included in such a set of raw score variables is usually, though not necessarily, a "variable" which takes the value of 1 for all observations. This last variable is what is ordinarily referred to as a unit vector. Its standard deviation is, of course, 0. The "correlation" of such a variable with any other variable would in a mathematical sense be undefined because its s.d. is 0, but its correlation with other variables may, in fact, be treated as a defined quantity by a computer program when such a variable is specifically introduced as a member of an array of variables. Computer programs are ordinarily constructed so as to arbitrarily set the correlation of a variable having an s.d. of 0 with any other variable to 0.

Let the raw score conceptual model be denoted by M, and be of the form

$$X_c = A_1X_1 + A_2X_2 + \dots + A_pX_p + E,$$

where X_c is the criterion variable in its raw score form, $X_1, X_2, \dots, X_p$ are predicator variables in their raw score form, $A_1, A_2, \dots, A_p$ are raw score weights, and E is a raw score residual vector. Assume for the present that one of the predicators, X_p, is the unit vector. In order to minimize the sum of the squared elements in E, we would need to solve a system of p equations in the p unknowns $A_1, A_2, \dots, A_p$. The coefficients of the unknowns in these equations would be the sums of squares and cross products of the vectors $X_1, X_2, \dots, X_p$, and the right side of the i(th) equation would be the sum of cross products of the raw scores in X_i and X_c. One reason for including the unit vector X_p as a predictor vector is that the other predictors, $X_1, X_2, \dots, X_{p-1}$ are all continuous vectors. If the unit vector was not included, the prediction system would be required to predict a criterion score of 0 whenever the values on the continuous predictors are all 0. It may be desirable to avoid such a requirement as a general rule on the following grounds. If the unit vector is included, but expected criterion scores should in fact be 0 when the value on all continuous predictors is 0, the chances are good that the estimated weight associated with X_p will turn out to be 0, or nearly so. So no harm is done to the prediction system by the inclusion of the unit vector, even if it makes no contribution to the criterion. On the other hand, if it is arbitrarily excluded from the conceptual model, but it does in fact contribute to the criterion, the prediction system will turn out to be in error. There are special situations in which it is desired to have a model which does not include the unit vector. But as will be evident from what follows, the use of Persub to handle models which include only continuous vectors will require special methods. As mentioned later, the use of a unit vector in the conceptual model when the other vectors include an array of binary vectors which sum to the unit vector is no special problem with regard to the conceptual model. The Persub regression computing routines have been constructed so as to obtain one of the unlimited number of least squares solutions for a regression system when there are linear dependencies in the conceptual model.

The Persub regression routine does not actually tackle directly the problem of obtaining a solution for a regression system for a model, M, which includes raw score variables, $X_1, X_2, \ldots, X_{p-1}$, and the unit vector, X_p. Instead, it obtains the solution for another regression model, m, which is closely related to M. The model solved by Persub is of the form

$$x_c = a_1x_1 + a_2x_2 + \cdots + a_{p-1}x_{p-1} + e,$$

where x_c is the standardized form of X_c, that is an element of x_c is obtained by subtracting the mean of all elements in X_c for the corresponding value in X_c, and dividing the difference by the s.d. of elements in X_c. Similarly the variables $x_1, x_2, \ldots, x_{p-1}$ are the standardized forms of $X_1, X_2, \ldots, X_{p-1}$, the elements in e are residuals, and the p-1 a's are chosen so as to minimize the sum of squares of elements in e.

The solution for the a's in model m requires obtaining the solution to a set of p-1 equations in which the unknowns are $a_1, a_2, \ldots, a_{p-1}$. The coefficients of the unknowns in these equations are the sums of squares and cross products of the variables $x_1, x_2, \ldots, x_{p-1}$. But these sums of squares and cross products are simply N times the intercorrelations, and the right side of each equation is N times the validity coefficient. So if each equation is divided through by N, the equations to be solved have the inter-correlation matrix of the x_i as coefficients of the unknown a's, and the vector of p-1 validity coefficients appears on the right side. Note in passing that the intercorrelations and validities of the X_i are the same as the intercorrelations and validities of the x_i. The Persub routine finds a solution to a close approximation for this set of p-1 equations. The accuracy of the solution is shown by the size of the "equation errors" which are reported on the print-out. The equation error reported along with standard weight a_i on the print-out is obtained by putting the computed values of $a_1, a_2, \ldots, a_{p-1}$ into an equation in which the coefficients of the a's are the appropriate elements from the i(th) row of the correlation matrix of predictors. This weighted sum of the a's should, if the solution for the a's is exact, be identical to the validity coefficient of X_1 for X_c. Any difference between the weighted sum and the validity coefficient is due to the fact that the solution is approximate rather than exact, and the difference is referred to as an equation error. If the solution were exact, all equation errors would be 0. Ordinarily, equation errors which are reported are small, usually less than .005. The solution can be made more exact and the equation errors made generally smaller by specifying a more rigorous stop criterion, for example, .00001 rather than .0001.

There is a relationship between the a's in model m and the A's in model M. The relationship is that $A_i = a_i \cdot z_i$, for i = 1, 2, ..., p-1, where z_i is the ratio of the s.d. of X_c to the s.d. of X_i. Having obtained the values of $A_1, A_2, \ldots, A_{p-1}$, the value of A_p is given by the difference between the mean of X_c and the weighted sum of the means of $X_1, X_2, \ldots, X_{p-1}$ where the weights are $A_1, A_2, \ldots, A_{p-1}$. The above is correct only on the condition that there was a unit vector in model M. Note however, that if one version of M does not include a unit vector but does include a set of mutually exclusive and exhaustive binary vectors of which the unit vector is a linear combination, we will simply state that the conceptual raw score model did, in fact, include the unit vector as X_p since the introduction of this linear combination into the predictor set will not change the predictive efficiency of the system nor the obtained set of predicted socres and error scores.

The normal flow then is to input a set of raw data consisting of X_c and $X_1, X_2, \ldots, X_{p-1}$ (not the unit vector), compute an intercorrelation matrix for these variables, specify one of them as a criterion and identify a set of predictors and call the REGREB or REGRED regression subroutine.

To review what happens, the Persub regression routine uses an intercorrelation matrix as input, solves a system of equations to a close approximation to give a set of "standard partial weights", and then converts the standard weights into a corresponding set of raw score weights and a regression constant A_p. As you will have noted, both the standard weights and raw score weights are reported on the output.

The net effect of the above is that when you specify the set of variables which are to be intercorrelated and used as input to the regression routine, it is not necessary to list the unit vector as one of these variables. The fact that the regression routines uses correlations as input implies that your conceptual raw score model did contain a unit vector. If you do list the unit vector as a variable to be intercorrelated along with other predictors, its correlations, although mathematically undefined will all be set to 0. This being the case, its weight will never be corrected away from 0 during the iteration sequence, and its standard weight will therefore be reported as 0. So will its raw score weight. However, the regression constant will undoubtedly be non-zero, but this will be for the unit vector which was implicitly in the conceptual raw score model anyway. Thus, if you list a unit vector as part of your input variables to the correlation routine, you will not hurt anything since it will have absolutely no effect on the solution which will be obtained by the regression routine. On the other hand, the listing of a unit vector as an input variable to the correlation routine is simply not necessary. You will get a weight for a unit vector even though you have not listed it as input to the program, and this weight will be the reported regression constant.

CHAPTER FOUR

Assumptions Underlying the Fixed x Model

by

Dr Robert A. Bottenberg
6570th Personnel Research Laboratory
Lackland AFB Texas

The set of computational procedures which are sometimes referred to as multiple linear regression analysis can be derived from a group of assumptions about the data on which the analysis is to be performed. These assumptions involve the distribution form, variability, and independence of the variables from which a data sample is drawn.

The concept of expected value is basic to these assumptions. Expected value is a mathematical notion which simply refers to a weighted sum. The expected value of a variable y is the weighted sum of the values taken by y. The weighting factor associated with a particular value of y is the probability that y takes that value. This process of weighting by probabilities makes an expected value similar to an average or mean value. An expected value of a variable is frequently referred to as a measure of central tendency for that variable. Its analog in physics is the center of gravity. For any variable defined over a specified population, there is a probability distribution function, and the expected value of the variable will be determined by the probability distribution function.

The fixed x model assumes that the distribution form of each variable in question is normal. Suppose there is a sample of n observations, Y_1, Y_2,..., Y_n. The three major assumptions of the fixed *x* model can be stated as follows: (1) each of the Y_i is an observation drawn at random on an underlying variable y_i, and each of the underlying variables $y_1, y_2, \ldots, y_n$ is normally distributed; (2) each of the variables $y_1, y_2, \ldots, y_n$ has the same variance, σ^2; and (3) for every pair of variables y_i and y_j among these *n* the correlation is 0.0. These three assumptions are usually referred to as the assumptions of normality, homogeneity of variance, and independence.

Certain additional assumptions are made about the expected values of the variables under study. Let $E_1, E_2, \ldots, E_n$ represent the *n* expected values of $y_1, y_2, \ldots, y_n$ respectively. It is assumed that some of the E_i may differ from each other, although not necessarily that they are all different. Further, it is assumed that each E_i is an unknown value, but that it may be represented adequately as a sum of products. One factor in each product is assumed to be an unknown parameter and the other is an observable quantity. Let $\beta_1, \beta_2, \ldots, \beta_p$ be a set of *p* unknown parameter values, and let $X_{i,1}, X_{i,2}, \ldots, X_{i,p}$ be a set of *p* observable quantities characterizing the population over which the variable y_i is defined, then our assumption is that $E_i = X_{i,1}\beta_1 + X_{i,2}\beta_2 + \ldots + X_{i,p}\beta_p$. Note that the parameters $\beta_1, \beta_2, \ldots, \beta_p$ are components of each of the *n* expected values. On the other hand, there is no requirement that for any pair of expected values E_i and E_j, the corresponding observable components are equal. That is, $X_{i,1}$ and $X_{j,1}$ are not necessarily equal, though in special cases they may be so. Similarly for $X_{i,2}$ and $X_{j,2}$, $X_{i,3}$ and $X_{j,3}$, ..., $X_{i,p}$ and $X_{j,p}$. If corresponding observable quantities in E_i and E_j are identical, then of course $E_i = E_j$, and we may think of Y_i and Y_j as being single observations on the variables y_i and y_j which have identical distribution functions, or of Y_i and Y_j as being two observations made on a single underlying variable.

For any given sample of size _n_, there are _p_ sets of X values, with _n_ elements in each set. The first of these _p_ sets consists of $X_{1,1}$, $X_{2,1}$, ..., $X_{n,1}$; the second set of $X_{1,2}$, $X_{2,2}$, ..., $X_{n,2}$; and the last set of $X_{1,p}$, $X_{2,p}$, ..., $X_{n,p}$. The important point here is that all the elements of a set multiply the identical unknown β parameter in the expanded form of the _n_ expected values. Thus if we hold constant elements in all but one in the X sets, it is apparent that the expected values of the _n_ variables in question will differ if (a) the unknown parameter is not 0, and (b) there is variability among the observable values in the set.

The term _fixed x model_ is used because no assumptions are made regarding the distribution function for any set of the X values. The model states that we are given an _nxp_ array of observable quantities which constitute the observable components of the expected values of _n_ variables. These _n_ variables in turn are normally distributed, have a common variance, and are independent, and the sample of data to be analyzed consists of one observation drawn at random on each variable.

The x's among the n in a given set may be either measured values such as father's height, muzzle velocity, temperature at which a chemical reaction is carried out, etc. Or, they may be coded values which represent the presence or absence of some qualitative characteristic such as sex, membership in a particular group, etc. Ordinarily, such coded "variables" are assigned the number 1 and 0. Thus, $X_{1,1}$, $X_{2,1}$ and $X_{3,1}$ might all equal 1 and $X_{4,1}$, $X_{5,1}$ and $X_{6,1}$ be 0 if Y_1, Y_2 and Y_3 are observations made on male subjects, while Y_4, Y_5 and Y_6 are observations made on female subjects, and the "variable x_1" is defined as 1 if the observation is from a male, 0 otherwise.

Hypotheses which can be tested by standard multiple regression procedures, including analysis of variance, are linear in the unknown parameters. We may make a test of a simple hypothesis involving a single parameter, such as

$$\beta_i = k,$$

where k is a specified constant, or we may test a compound hypothesis by imposing r linear restrictions simultaneously on a set of r of the parameters. A test for the independent contribution of x_i can be made by imposing the restriction $\beta_i = 0$. If the hypothesis is true, then other things equal, the expected values do not differ from each other as a function of differences in the x_i variable. In this event the x_i variable may be said to make no independent contribution.

Let e_1 be the error sum of squares reflecting the squared differences obtained between the observed values of the Y_i and the estimated values of the E_{i} when maximum likelihood estimators are substituted for the unknown parameters. After an hypothesis, either simple or compound, is specified and the appropriate substitutions have been introduced into the model, thus revising the definitions of the E_i, let e_2 be the error sum of squares obtained when maximum likelihood estimators are obtained for the restricted model. It can be shown that

$$\frac{(e_2 - e_1)/s}{e_1/(n - t)}$$

is the likelihood ratio test statistic, and has the F distribution with s and (n - t) df if the hypothesis implied by the set of restrictions is true, where t is the rank of the nxp array of observable quantities for the n expected values, and s is the difference between the ranks of the nxp matrix and the matrix of constants which results when the restrictions of the hypothesis are imposed on the model.

Considerable research, both empirical and theoretical, has been devoted to the robustness of this statistical testing procedure. That is, to the question of how rapidly the procedure becomes invalid as departures are made from the normality, homogeneity and independence assumptions. In general, the procedure is robust against non-normality. It is somewhat less robust against violations of the homogeneity and independence assumptions, but it is a surprisingly accurate test despite obvious non-normality and non-homogeneity which is present in most actual data.

Since distribution assumptions regarding the variables which are components of the expected values do not enter the assumptions of the fixed x model, it is entirely appropriate to define the elements of x_j as being some functional form of the corresponding elements of x_i, such as the squares, cubes, square roots, sines, or whatever seems conceptually reasonable in the context of the problem under investigation. When the assumptions of the fixed x model form the basis for the analysis procedure, the question of whether an x_i variable is normally distributed is not relevant. The fixed x model is not concerned with the distribution of elements on a given x_i variable. From the standpoint of estimation of a parameter, in general the estimation procedure provides better estimates if a given x_i has a wide rather than a narrow range of values when these are measured values. But the model makes no specific demand regarding the distribution of such a variable.

THE COMPUTATION OF THE F STATISTIC

By

Joe H. Ward, Jr.

The following material discusses the computation of the F statistic in terms of predictive accuracy and the dimension of the vector spaces associated with the prediction systems.

1. THE LARGEST MODEL

Consider the prediction of y from s(1), s(2), ..., s(n) where each vector s(i) has its i-th element equal to 1 and all other elements equal to 0. These n vectors are linearly independent and the dimension of the vector space generated by these vectors is n.

Then $y = u_1 s(1) + u_2 s(2) + \ldots + u_n s(n) + h$

Let $L = u_1 s(1) + u_2 s(2) + \ldots + u_n s(n)$

Then $y = L + h$

In this problem the least squares regression coefficients are

$$w_i = y_i \quad (i = 1,n)$$
and $$y = L$$

The residual vector h is the NULL vector

Now let $\sum_{i=1}^{n} y_i^2 = t_y$ associated with this value consider the dimension of the vector y and call the dimension $d(t_y)$ with value n, i.e. $d(t_y) = n$

$\sum_{i=1}^{n} L_i^2 = p_L$ associated with this value consider the number of Independent vectors in the set s(1), s(2), ..., s(n). This is the dimension of the space, $d(p_L) = n$

COMPUTATION OF THE F STATISTIC

$\sum_{i=1}^{n} h_i^2 = q_h = 0$ since h is the NULL vector the sum of squares $q_h = 0$. Associated with this vector h we defind $d(q_h) = 0$

Then write

$t_y = p_L + q_h$ $d(t_y) = d(p_L) + d(q_h)$

$t_y = p_L + 0$ $n = n + 0$

$t_y = p_L$ $n = n$

2. THE UNRESTRICTED MODEL

Now consider the prediction of y from a set of Independent vectors x(1), x(2), ..., x(k) - (where $k<n$). Each of these vectors can be expressed as a linear combination of s(1), s(2), ..., s(n).

Then $y = a_1x(1) + a_2x(2) + \ldots + a_kx(k) + e$

Let $u = a_1x(1) + a_2x(2) + \ldots + a_kx(k)$

Then $y = u + e$

Assume that a_i $(i = 1,k)$ are least squares values

Then let $\sum_{i=1}^{n} u_i^2 = p_u$ associated with this value consider the number of Independent vectors in the set x(1), x(2), ..., x(k). This is the dimension of the space, $d(p_u) = k$

$\sum_{i=1}^{n} e_i^2 = q_e$ associated with this residual vector is the number $d(q_e)$ with value equal to $n - k$

Then we can write

$t_y = p_u + q_e$ $d(t_y) = d(p_u) + d(q_e)$

$n = k + (n-k)$

$n = n$

COMPUTATION OF THE F STATISTIC

3. THE RESTRICTED MODEL

Finally, consider the prediction of y from a set of Independent vectors z(1), z(2), ..., z(c) (where $c<k<n$). Assume that each of the vectors can be expressed as a linear combination of x(1), x(2), ..., x(k).

Then $y = b_1z(1) + b_2z(2) + \ldots + b_cz(c) + f$

Let $r = b_1z(1) + b_2z(2) + \ldots + b_cz(c)$

Then $y = r + f$

Assume that $b_i (i = 1, c)$ are least squares values.

Then let $\sum_{i=1}^{n} r_i^2 = p_r$ associated with this value consider the number of Independent vectors in the set z(1), z(2), ..., z(c). This is the dimension of the space, $d(p_r) = c$

$\sum_{i=1}^{n} f_i^2 = q_f$ associated with this residual vector is the number $d(q_f)$ with value equal to n-c

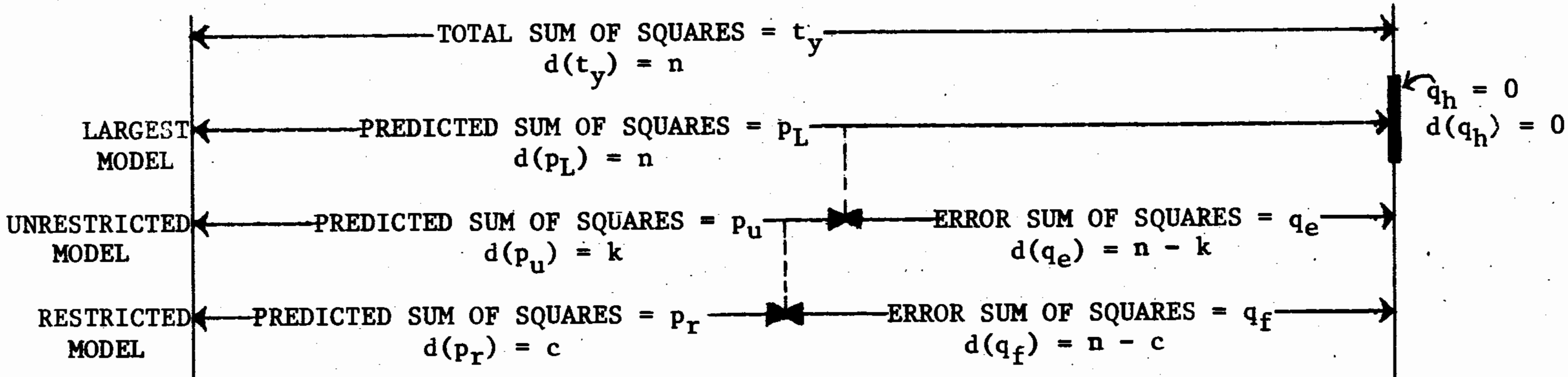
Then we can write

$t_y = p_r + q_f$ $\qquad d(t_y) = d(p_r) + d(q_f)$

$n = c + (n-c)$

$n = n$

PICTORIAL REPRESENTATION OF THE F STATISTIC



WE CONSIDER:

1. The lost predictive accuracy per vector given up between LARGEST and UNRESTRICTED models.

2. The lost predictive accuracy per vector given up between UNRESTRICTED and RESTRICTED models.

$$\frac{p_L - p_u}{n - k} = \frac{q_e}{n - k} \qquad \frac{p_u - p_r}{k - c} = \frac{q_f - q_e}{(n-c)-(n-k)} = \frac{q_f - q_e}{k - c}$$

Then we compare these two values as a ratio and have:

$$F = \frac{(q_f - q_e)/(k - c)}{q_e/(n - k)} = \frac{(R^2_{UNRES.} - R^2_{RES})/(k - c)}{(1 - R^2_{UNRES.})/(n - k)}$$

/lpe
2/2/67

PROBLEM SET 1

GENERATION OF VECTORS AND LINEAR COMBINATIONS

The following data is available for 3 Lt, 4 Capt, 5 Maj. Each officer has three part scores on a test.

One Lt, One Capt and two Maj	have	part	1	scores	of	50;
Two Lt, One Capt and one Maj	"	"	"	"	"	60;
One Capt and one Maj	"	"	"	"	"	70;
One Maj	has	"	"	score	"	80;
One Capt	"	"	"	"	"	90.

The Lt with part 1 score of 50 and the Maj with part 1 score of 80 have part 2 scores of 30; the Capt with part 1 score of 50 and the Capt with part 1 score of 90 have part 2 scores of 35; one Maj with part 1 score of 50 and the Capt with part 1 score of 70 and a Lt with part 1 score of 60 have part 2 scores of 40; the remaining officers have part 2 scores of 45. The Capt with part 1 score of 90, the Maj with part 1 score of 80 and the Lt with part 1 score of 60 and part 2 score of 45 have part 3 scores of 20. All others have part 3 scores of 25.

Each of the 12 officers is identified by one of the numbers (1, 2, ..., 12) according to the following rules:

Lt numbers are less than Capt numbers and the Capt numbers are less than the Maj numbers. Within a particular rank the identification number of an officer is small than another officer if his part 1 score is smaller. If the officers within a rank have equal part 1 scores, then a similar test is made for part 2 scores for determining identification number.

<u>Prob #1</u> - Generate an identification vector $x^{(1)}$ whose $\underline{i}$th element is equal to $\underline{i}$ $(i = 1, 2, \ldots, 12)$. For further purposes this vector will be considered to consist of the identification numbers of the officers in this problem.

Generate 6 other vectors $x^{(2)}, x^{(3)}, \ldots, x^{(7)}$ which reflect the rank status and part score information described above. Be sure that the elements of each vector correspond to the identification number in $x^{(1)}$.

<u>Prob #2</u> - Generate a vector $\underline{t}$ (the total test score) whose elements are the sum of the part scores and express $\underline{t}$ as a linear combination of the part score vectors from Prob #1.

<u>Prob #3</u> - Generate a weighted composite vector $\underline{c}$ whose elements are 2 times the part 1 score plus 3 times the part 2 score minus 1.2 times the part 3 score, and express $\underline{c}$ as a linear combination of the part score vectors from Prob #1.

<u>Prob #4</u> - Generate a composite vector $\underline{p}$ whose elements are 100 times the vector representing rank of Lt from Prob #1 plus 110 times the vector representing rank of Capt from Prob #1 plus 120 times the vector representing rank of Maj from Prob #1.

PROBLEM SET 1

Prob #5 - Find a vector d so that it is possible to express the vector t of Prob #2 as the linear combination

$$t = 1(p) + 1(d).$$

Prob #6 - Find the vector u which is the sum of the vectors of Prob #1 which represent rank information.

Prob #7 - Generate a vector s whose elements are the squares of the first part scores.

Prob #8 - Generate a vector r in which each element is the ratio of the 3rd part score to the 2nd part score.

Prob #9 - Generate a vector $z^{(1)}$ in which an element is 1 if the officer is a captain with part 1 score of 50; and 0 otherwise.

Prob #10 - Express the vector representing the rank of Captain as a linear combination of the other two rank vectors and the vector u of Prob #6.

MULTIPLE REGRESSION PROBLEM SET 2

Assume that the following information is available on 22 high school students:

Grade point average for Freshman Boys	70, 75
" " " " Freshman Girls	65, 75, 90
" " " " Sophomore Boys	60, 85
" " " " Sophomore Girls	80, 85
" " " " Junior Boys	65, 65, 80, 90
" " " " Junior Girls	75, 85, 95
" " " " Senior Boys	50, 70, 75, 80
" " " " Senior Girls	70, 80

Prob #1 - Define and write explicitly 9 vectors that reflect these data.

Prob #2 - Express the grade point average vector $\underline{y}$ as a linear combination of the 8 other data vectors and a residual vector. This is the full model.

Prob #3 - State the following four hypotheses in mathematical form; for each class the expected grade point average for the two sexes is equal.

Prob #4 - Introduce the restrictions of problem three into the full model and rewrite the restricted model in the simplest form.

Prob #5 - Indicate computation required for testing the hypothesis above.

Prob #6 - State the following four hypotheses in mathematical form:

a. The expected grade point average for freshman girls is 1.1 times that of freshman boys.

b. The expected grade point average for sophomore girls is 1.2 times that of sophomore boys.

c. The expected grade point average for junior girls is 1.3 times that of junior boys.

d. The expected grade point average for senior girls is 1.4 times that of senior boys.

Prob #7 - Introduce the restrictions of problem 6 into the full model of problem 2 and rewrite the restricted model in the simplest form.

Prob #8 - Indicate computations required for testing the hypotheses in problem six.

Card Columns	
1-2	Code Sequence
3	blank
4-5	Y Independent Groups t-test-EDWARDS- Statistical Methods for the Behavioral Sciences. P. 276. Problem No. 134. (N=20)
6	blank
7	1 if control -
8	blank
9	1 if Experimental -
10	blank
11	Group No. (2 x 3 design in Edwards - Experimental Design in Psychological Research: (1960) P. 221)
	1= A1
	2= A2
	3= A3
	4= B1
	5= B2
	6= B3
12	blank
13-14	Y for 2x3 (N=60)
15	blank
16	Group No. -(Equated Groups T- test Edwards, P. 286)
17	blank
18-19	Y (N=25)
20	blank
21	X (Covariable)
22	blank
23-24	Pair No. -(Paired Observations t - test

IN A 2 groups study, there will be 4 predictor vectors

RESTRICTED MODEL WILL HAVE 7 PREDICTOR VECTORS: 6 MEMBERSHIP VECTORS AND 1 COVER VECTOR

12 vectors since there are 6 groups: 6 MEMBERSHIP & 6 COVER.
WED. DEC 13
EXPERIMENT

Covariance analysis assuming no interaction

NOTE
It is not worth the trouble to match pairs and you must assume no interaction when you do it. Just see that the full range of the controlled variable is included in both samples.

	Edwards, P. 280)
25	blank
26	Y (N=20)
27	blank
28	Group of Time Number
29	blank
30-31	Y-(Simple linear Regressin - Edwards, P. 138, Problem No. 77 (N=16)
32	blank
33-34	X - Predictor
35	blank
36	Group No. - 3
37	blank
38-40	Y (N=30) Criterion Achievement
41	blank
42-44	Aptitude (Covariable)
45	blank
46	Y-(Correlation Ratio Problem, Edwards, P. 208, Problem 10.6)
47	blank
48	X - Predictor
49	blank
50	Polynominal Study - Finger dexterity (N=38)
51	blank
52-53	Hours of Practice
54	blank
55-56	Y - Situation 1
57	blank
58-59	Y - Situation 2
60	blank
61-62	Y - Situation 3

63	blank
64-65	Y - Situation 4
66	blank
67-68	Y - Situation 5