Principal Components Regression in Exploratory Statistical Research

William F. Massy

Journal of the American Statistical Association, Vol. 60, No. 309. (Mar., 1965), pp. 234-256.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28196503%2960%3A309%3C234%3APCRIES%3E2.0.CO%3B2-A

Journal of the American Statistical Association is currently published by American Statistical Association.

PRINCIPAL COMPONENTS REGRESSION IN EXPLORATORY STATISTICAL RESEARCH*

WILLIAM F. MASSY
Stanford University

Regression upon principal components of the percentage points of the income and education distributions for 1950 census tracts in the city of Chicago led to the estimation of "beta coefficient profiles" for television receiver and refrigerator ownership, for central heating system usage, and for a measure of dwelling unit overcrowding. The betas are standardized coefficients of regression of a dependent variable upon the proportions of families in the classes of the marginal income and education distributions. They measure the relative contribution of families in these classes to the over-all per cent saturation of the dependent variable in the tract. The coefficients were estimated by techniques developed in the first portion of the paper; estimation by classical regression methods would have been impossible because of multicollinearity. The empirical results are in substantial agreement with findings from regressions of the dependent variables upon the mean values of income and education, and their squares. The statistical devices appear to be useful in exploratory empirical research.

EXPLORATORY empirical work poses significant problems for the statistician. When well-defined models or reasonably clear hypotheses dealing with the interrelations between variables are lacking, it is often necessary to use a preliminary sample of data to suggest interpretations that may be put to the test in later studies. This is the problem of exploratory, as opposed to descriptive or cause-effect, research [3, pp. 21–55]; it is particularly pertinent to areas of economics, marketing, and business research at the present time.

This paper explores some of Kendall's [7] ideas for analyzing the relation between a dependent variable and a set of independent variables when the latter are not necessarily amenable to standard statistical treatment. He tentatively expanded on some of Stone's [9] results by obtaining the principal components of a set of standardized explanatory variables, calculating their regression upon a dependent variable, and projecting the resulting parameters back into the terms of the original variates. In discussing this approach, he states:

> "It is possible to 'orthogonalize' a regression situation for any arbitrary set of variates X. This possibility does not seem to have been much discussed in the literature; but it throws some new light on certain old but unsolved problems; particularly (a) how many do we take?, (b) how do we discard the unimportant ones?, and (c) how do we get rid of multicollinearities in them?" [7, p. 70].

The aim of this paper is to examine his suggestion by pulling together the major algebraic results needed for combining regression and principal components analysis, attempting to make some additional contributions to the statistical machinery, and trying out the resulting techniques upon a set of

* The calculations for this study were conducted on the IBM 7090 computer installed at the Stanford University Computation Center; machine time was in part financed by the center under National Science Foundation Grant No. NSF-GP948.

real data. Empirical analysis is needed to begin to deal with the rather complicated interpretation problems arising when the methods are applied.

The study examines the relations of the incidence of television and refrigerator ownership, of central heating usage, and of a condition of overcrowding in dwelling units with a set of twenty-three income and education variables. The data pertain to census tracts in the city of Chicago, as of 1950. Percentage points in the marginal distributions for income and education were used as independent variables; they were reduced to principal components, the components related to each of the dependent variables given above, and the results projected back into terms of the original income and education variates through methods to be discussed in the paper. Classical least-squares regressions of the dependent variables upon the means of income and education and their squares were calculated to provide comparisons for the results obtained from the principal components-regression analysis. The agreement was good in almost all cases, and it was concluded that the principal components regression method is a useful approach in exploratory studies of complex relations between variables.

1. DEVELOPMENT OF STATISTICAL METHODS

Review of Principal Components Analysis. The method of principal components has been known for many years and has been discussed in different ways by a variety of authors, so the present review need not be extensive. The reader may consult Anderson [1, Ch. 11] or Kendall [7, Ch. 1], and the references cited therein, for a more comprehensive discussion of the subject. The material to be presented in this section is limited to the mathematical development that is needed for understanding the sequel.

The objective of principal components analysis is to find a linear transformation of a set of n variates of X into a new set denoted by P, where the new set has certain desirable properties. These properties, which provide the rationale for using the p's rather than the original x's, are: (i) the elements of P are uncorrelated with each other in the sample (orthogonality); and (ii) each element of P, progressing from p_1 to p_2, etc., accounts for as much of the combined variance of the x's as possible, consistent with being orthogonal to the preceding p's. The new variables correspond to the principal axes of the ellipsoid formed by the scatter of sample points in the n dimensional space having the elements of X as a basis. The principal components transformation is thus a rotation from the original x coordinate system to the system defined by the principal axes of this ellipsoid.

As is well-known, the principal axes of the space spanned by the elements of X are not invariant to changes in the scales in which the variables are measured. In practice, this means that a change in the unit of measurement for even one variable (e.g., a change from inches to feet) can change the pattern of principal components. A discussion of the relation between units of measurement and principal components analysis is beyond the scope of this paper, but the problem can be side-stepped if the analyses are confined to the principal axes of the x elements, as standardized through division by the square roots of their respective sums of squares.

Defining $\boldsymbol{Z}$ as the matrix of standardized x variates, the transformation to principal components is given by

$$P = M'Z. \tag{1}$$

Providing that the rank of $\boldsymbol{Z}$ is equal to n (note that $R(\boldsymbol{Z}) = R(\boldsymbol{X})$), both $\boldsymbol{P}$ and $\boldsymbol{Z}$ contain n rows and T columns, where T is the number of observations in the sample. The objective of the principal components procedure is to find $\boldsymbol{M}'$, the $n \times n$ coefficient matrix for the transformation, working only from a knowledge of $\boldsymbol{Z}$.

To see how $\boldsymbol{M}$ is determined, postmultiply equation (1) by $\boldsymbol{P}'$ and substitute the transpose of its right-hand side:

$$PP' = M'ZZ'M. \tag{2}$$

Note that $\boldsymbol{ZZ}' = \boldsymbol{R}$ is the matrix of simple correlations among the x's and that $\boldsymbol{PP}' = \boldsymbol{D}$, which is the variance-covariance matrix for the principal components. Substituting these results into (2) yields:

$$M'RM = D. \tag{3}$$

$\boldsymbol{D}$ should be diagonal by virtue of requirement (i) above. Therefore, equation (3) is an orthogonal similarity transformation diagonalizing the symmetric matrix $\boldsymbol{R}$. A well-known theorem of matrix algebra states that the transformation matrix $\boldsymbol{M}$ has an orthonormal set of Eigenvectors of $\boldsymbol{R}$ as its columns, and that $\boldsymbol{PP}'$ has the eigenvalues of $\boldsymbol{R}$ as its diagonal elements [4, p. 248]. Since the columns of $\boldsymbol{M}$ are orthonormal, moreover, it is known that $\boldsymbol{M}' = \boldsymbol{M}^{-1}$; therefore, the equation for the z's in terms of the p's can be written simply as $\boldsymbol{Z} = \boldsymbol{MP}$. If the columns of $\boldsymbol{M}$ are ordered so that the first diagonal element of $\boldsymbol{PP}'$ contains the largest eigenvalue of $\boldsymbol{R}$, the second the next largest, etc., the principal components will be ordered as specified in requirement (ii) above. Furthermore, it can be shown that the first component accounts for a maximum amount of the combined variance of the x's, the second for the maximum amount consistent with being uncorrelated with the first, etc. [7, pp. 15–6].

As normally calculated and utilized, principal components are left in the form given in equation (1) above, in which case their variances are equal to their respective eigenvalues. However, experience with the related technique of factor analysis has demonstrated that it is useful to scale the components so that all of them have equal variance. This is accomplished by premultiplying both sides of equation (1) by the inverse of the diagonal matrix of standard deviations for the p's, namely $\boldsymbol{D}^{-\frac{1}{2}}$. These scaled principal components will be denoted by $\boldsymbol{F}$ rather than $\boldsymbol{P}$; they are defined by

$$F = (D^{-\frac{1}{2}}M')Z. \tag{4}$$

The coefficient matrix $(\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{M}')$ is the inverse of the one ordinarily obtained in the factor analysis equation $\boldsymbol{Z} = \boldsymbol{AF}$. $\boldsymbol{A}$ is called the "principal components loading matrix." It is computed by multiplying each orthonormal eigenvector of $\boldsymbol{R}$ by the square root of its Eigenvalue; thus,

$$A = (D^{-\frac{1}{2}}M')^{-1} = MD^{+\frac{1}{2}}. \tag{5}$$

One advantage of scaling the principal components to unit variance is that the elements of $\boldsymbol{A}$ have a natural and useful interpretation; each a_{ij} is the simple correlation of the ith variate with the jth component. Empirical work with principal components analysis begins with the calculation and examination of $\boldsymbol{A}$.

Regression on Principal Components. One reason for transforming a set of variates into principal components is so that their relations with yet another variable may be explored more easily. If the original independent variables (the x's) are highly collinear with one another, or if there are a great number of potential explanatory variables, it may be appropriate to simplify their sample space by a transformation to principal components. The dependent variable can then be regressed upon the resulting principal components rather than upon the original variates.

The principal components transformation can be regarded in the same manner as any other transformation that is used to prepare variates for regression. One needs to know the properties of the transformation and its inverse, and the conditions in which their use is appropriate. The major contention of this paper is that the principal components transformation can be very useful in exploratory statistical research; the paper concentrates upon a partly mathematical and partly empirical examination of the properties of principal components transformation.

Consider a $1 \times T$ vector of observed values for a single variate, which is to be predicted and hopefully explained by the set of n independent variables considered above. Writing this vector as y', with each element standardized through division by the dependent variables' standard deviation, the regression of y on the scaled principal components of x is denoted by

$$y' = \gamma' F + e', \tag{6}$$

where γ' is a vector of regression coefficients, and the elements of e' are the error terms in the regression. Working from the principal components loadings matrix, it would be possible to calculate the $n \times T$ matrix F in order to provide the input information for (6):

$$F = A^{-1}Z = D^{-1}A'Z. \tag{7}$$

Once the values of F are known for every observation in the sample, the parameters of (6) can be estimated by ordinary methods.

The work required to calculate F can be eliminated by substituting (7) into (6) directly, as shown in (8) and (9). Despite the simplicity of this short-cut procedure, it does not seem to be currently in use: the standard library computer programs for principal components-regression analysis with which the author is familiar use the long method discussed in the last paragraph. The effort saved by using the short cut becomes significant when the sample size (T) is large. This is true even when a computer is being used for the calculations, since for samples that are too large to be held in core memory, the long method requires four reading and writing operations for each of the n variates or components over all T of the observations, as opposed to just one such operation for the short method. These input-output operations involve the expenditure of relatively large amounts of time on most machines.

The short method can be presented in terms of the least-squares solution for (6), as given by equation (8). Substitution of (7) into (8) and use of the relation $FF'=I$ yields the final result, equation (9).[1]

$$\gamma = (FF')^{-1}(Fy'); \tag{8}$$

$$\gamma = D^{-1}A'Zy' = D^{-1}A'r_{zy}. \tag{9}$$

The last term is the vector of simple correlations between y and the z's. In scalar terms the regression coefficient of y on the kth principal component is given by

$$\gamma_k = \frac{1}{d_{kk}} \sum_{j=1}^{n} a_{jk} r_{zy_j}. \tag{10}$$

Since the f's have been standardized, the γ_k are the "beta" or standardized coefficients of regression of y on the principal components. Moreover, the orthogonality of the f's implies that the γ's can be interpreted as correlation coefficients between y and the components. In fact, the γ's could be regarded as an extra row added to the loadings matrix A, corresponding to the dependent variable y and relating it to the principal components in exactly the same way as the a_{ij} relate the z's to the components. (Note, however, that the orthogonality of A would be destroyed by the addition of this row.)

The coefficient of determination and error variance for the regression are easily calculated from the γ's:

$$R^2 = \sum_{k=1}^{n} \gamma_k^2; \qquad \sigma^2 = \left(\sum_{t-1}^{T} y_t^2\right)\left(\frac{1-R^2}{T-n-1}\right).$$

Again, following well-known regression theory, the covariance matrix for the γ's is:

$$\Sigma_\gamma = (FF')^{-1}\sigma_e^2 = \sigma_e^2 I.$$

Subject to all of the usual restrictions on the regression error terms, the γ's are independently distributed normal variates with identical variances σ_e^2.

The inverse of the principal components transformation is also easily obtained. Given estimates of the γ's, it is possible to obtain values for the regression coefficients of the original variables by applying the same transformation that takes F back to Z:

$$y = \gamma' F = (\gamma D^{-1}A')Z = \beta' Z. \tag{11}$$

Since the least-squares solutions for the last two equalities and y must be unique, the vector of beta coefficients for the regression of y on the z's is equal to the following:

$$\beta = AD^{-1}\gamma = AD^{-1}A^{-1}r_{zy}. \tag{12}$$

[1] Equation (9) and the sequel are not valid in factor analysis, where "communalities" are used instead of ones on the principal diagonals of the correlation matrix from which roots are extracted. This occurs because the use of communalities destroys the equality between Zy' and r_{xy}.

It may be noted in passing that the transformation from variables to principal components, the regression of y on the principal components, and the inverse transformation of the regression coefficients back into the z domain yields the usual least-squares solution for the beta coefficients of regression:

$$\gamma = A^{-1}r_{zy},$$

by (9) and (5); and

$$\beta = A^{-1\prime}\gamma = (A^{-1\prime}A^{-1})r_{zy} = (AA')^{-1}r_{zy} = R^{-1}r_{zy}.$$

The last quantity is the classical least-squares equation for beta.

The variance-covariance matrix for the β's can be calculated, provided that the sample principal components are viewed as predetermined summary variates, rather than as estimates of "true" components in the underlying population. Then the elements of AD^{-1} are known coefficients in the linear relation linking the betas to the gammas, which in turn are independently distributed random variables with zero mean. Thus,

$$\Sigma_\beta = E(\beta\beta') = E(AD^{-1}\gamma\gamma'D^{-1}A') = \sigma_e^2(AD^{-2}A'). \tag{13}$$

The covariance for the ith and jth beta coefficients is therefore given by

$$\sigma_{\beta_i\beta_j} = \sigma_e^2 \sum_{k=1}^{n} \frac{a_{ik}a_{jk}}{d_{kk}^2}\cdot$$

All of these quantities are easily calculated on a computer once the loadings matrix and correlation coefficients among the dependent and independent variables are known.

Properties of the Principal Components-Regression Parameters. Use of the procedures discussed above would hardly be necessary when the beta vector could be estimated directly by classical methods. At least two situations arise, however, in which ordinary multivariate regression is not appropriate: (i) when the independent variables are collinear with one another, making inversion of the correlation matrix impossible and the elements of beta indeterminate; and (ii) when, because of high (but not complete) collinearity or for some other reason, it is desirable to collapse the independent variable space by deleting one or more principal components from the regression relationship. The two cases will be considered separately below.

(i) Collinear independent variables. Heretofore, it has been assumed that the rank of Z is equal to n, the number of independent variables. Consider now the case when the rows of Z are subject to one or more linear restrictions reducing its rank to $m<n$. The classical regression solution is indeterminate under these conditions, but by using principal components regression technique, it is possible to estimate the parameters of regression upon the projections of the original variables into the m-flat of the space E^n spanned by the rows of Z.[2]

Given the rank of Z, m is the maximum number of nonzero eigenvalues that can be extracted from the correlation matrix. Thus, A contains n rows but

[2] This situation may also be handled by solving the classical regression problem with parameters subject to $m\times m$ linear constraints, as discussed in Johnston [6]

only m columns. While $\boldsymbol{A}^{-1}$ is not defined for $\boldsymbol{A}$ not square, it is possible to prove that $\boldsymbol{D}^{-1}\boldsymbol{A}'$ is the *left inverse* of $\boldsymbol{A}$, where $\boldsymbol{D}$ is now of order $m\times m$. (See Perlis [8, pp. 58–9] for a discussion of left and right inverses.) This means that $\boldsymbol{D}^{-1}\boldsymbol{A}'\boldsymbol{A}=\boldsymbol{I}_m$; the result is seen intuitively because, for the z's collinear, the principal components transformation amounts to a rotation of axes in the subspace of E^n that is spanned by the columns of $\boldsymbol{Z}$, and such a rotation is freely reversible. Thus, the logic given by equations (1)–(5) holds unaltered for $\boldsymbol{Z}$ collinear.

The definition of linear dependence and rank provides that, for $\boldsymbol{Z}$ of rank less than n, there must be a matrix $\boldsymbol{\lambda}$ containing $n-m$ non-null rows of order n such that $\boldsymbol{\lambda Z}=\boldsymbol{0}$. Premultiplying the equation defining the eigenvectors of the correlation matrix $\boldsymbol{RM}=\boldsymbol{MD}$ (or equivalently $(\boldsymbol{ZZ}')\boldsymbol{M}=\boldsymbol{MD}$) by $\boldsymbol{\lambda}$ and applying equation (5) yields the relations:

$$(\boldsymbol{\lambda Z})\boldsymbol{Z}'\boldsymbol{M} = (\boldsymbol{0}) = \boldsymbol{\lambda MD},$$

and

$$(\boldsymbol{0}) = \boldsymbol{\lambda A D}^{\frac{1}{2}}.$$

Since all the row elements in a given column of $\boldsymbol{A}$ are multiplied by the same diagonal element of $\boldsymbol{D}^{\frac{1}{2}}$, it is apparent that the rows of $\boldsymbol{A}$, taken by themselves, are linearly dependent. Furthermore, the coefficients defining the dependence are the same as those defining it for the rows of $\boldsymbol{Z}$. Carrying through this same procedure, equation (12) is premultiplied by $\boldsymbol{\lambda}$ to obtain:

$$\boldsymbol{\lambda\beta} = (\boldsymbol{\lambda A})\boldsymbol{D}^{-1}\boldsymbol{\gamma} = (\boldsymbol{0})\boldsymbol{D}^{-1}\boldsymbol{\gamma} = (\boldsymbol{0}). \tag{14}$$

Thus, the n betas estimated by the principal components-regression procedure lie in the same subspace of E^n as did the original z's. Exactly the same results hold for the betas' variance-covariance matrix, which is obtained according to equation (13).

The preceding paragraphs dealt with the reduction of the number of components used in a principal components regression analysis for cases where extreme multicollinearity has reduced the rank of the data matrix from n to m. This amounts to a process of: (a) transforming the sample variates to principal components; and then, (b) dropping the $n-m$ components that fail to account for any of the variance of the original variates. (The components that are dropped have zero eigenvectors.) This simplification of the sample space of the x's is required for the subsequent regression.

(ii) Deletion of Components. The possibility of simplifying the x-sample space through deletion of components exists even when the rank of the data matrix is equal to n, since one or more *nonzero* principal components may be dropped from the regression. It is desirable to examine the effects of dropping components upon the results given earlier in this paper, recognizing that unlike the previous case, the sacrifice of nonzero components represents a reduction in the amount of information that is provided as input to the subsequent regression.

The first question that must be considered is, "What components should be deleted in order to simplify the statistical analysis without destroying whatever basis may exist in the explanatory data for predicting the dependent variable?" There are at least two alternative criteria for deleting components:

a. Delete the components that are relatively unimportant as predictors of the *original independent variables* (X) in the problem; i.e., the components having the smallest eigenvectors should be dropped.

b. Delete the components that are relatively unimportant as predictors of the *dependent variable* (y) in the problem. In this case the components having the smallest values of gamma (the correlation between the components and y) should be dropped.

Hotelling [5] has pointed out that in general there is no reason why components that are important as far as the independent variables of a problem are concerned will be highly correlated with the dependent variable in a regression, so criteria *a* and *b* above are likely to lead to different results. Furthermore, it is easily shown that y need not be highly correlated with components having large eigenvalues in order for the explanatory power of the complete principal components regression to be high.

The choice of criteria must rest with the purpose of the analysis, as well as the degree to which the principal components results can be interpreted in terms of the structure of the process underlying the data for the independent variables. If the first few principal components can be related to something "real," as is hopefully the case in factor analysis, for example, then it may make sense to retain them as explanatory variables in a principal components-regression analysis, regardless of their correlation with the dependent variable. (In the author's experience, components with large eigenvalues are the ones most likely to yield natural interpretations.) Conversely, if the emphasis is on finding the correlates of y rather than testing its relation to any particular structural concepts, it would seem to make more sense to adopt criterion *b* and retain those components with the highest values of gamma. This is often the case in purely exploratory studies. The latter approach has been adopted in the empirical work to be reported later, but the results given in the following paragraphs hold, regardless of the criterion used for deleting components.

Removal of components causes the loadings matrix to become non-square, and reduces the order of many of the other matrices discussed above. It has been seen, however, that equations (1)–(5) hold, whether $\boldsymbol{A}$ is square or not. It can be shown that they hold, regardless of which components are retained, or of the order in which the columns of $\boldsymbol{A}$ are arranged. The fact that other non-null components exist has no bearing upon the mathematics given in (1)–(5) and hence cannot affect any of the later results.

Deletion of one or more nonzero principal components amounts to partitioning the set of independent variables into two groups of n variates each, assuming $\boldsymbol{Z}$ is of rank n. (If a selection criterion based on the gammas is used, one of the two groups can be regarded as related to the dependent variable and the other as relatively unrelated.) This might be expressed as follows:

$$y = \sum_{i=1}^{n} \beta_i' z_i' + \left[\sum_{i=1}^{n} \beta_i'' z_i'' + e \right].$$

The effects of the n singly primed variables (z') are included in the regression, while those of the doubly primed set are thrown into the error term. Estimates of the beta coefficients for the z_i' are obtained by working through equation (12), using only those components f_j' that have been included in the regression. The values of the z_i' can be estimated by the equation $\boldsymbol{Z}^{(')}=\boldsymbol{A}^{(')}\boldsymbol{F}^{(')}$, if desired. The proportion of the variance of each of the original variables z_i that is accounted for by z_i' is given by the sum of squares of the elements a_{ij} of $\boldsymbol{A}^{(')}$, taken over j. (This sum is called a communality in factor analysis.)

Thus, the regression of a dependent variable upon a reduced set of principal components amounts to partitioning the n-space spanned by the original independent variables into two orthogonal subspaces. If the gamma criterion is used, one space of dimension m (the number of retained components) is defined by principal axes that lie comparatively close to the dependent variable vector in the full sample space E^T. The artificial z' variables defined in the preceding paragraph represent the projections of the original variables into the E^m subspace, and the z'' are projections into its compliment.

Only m of the z' and $n-m$ of the z'' can be linearly independent; this follows as a direct consequence of the dimensions of the subspaces in which they are defined. Therefore, the estimated values of the β_is will be subject to $n-m$ linear restrictions, or more, if the original variables are not independent. No prior information about the nature of these two subspaces is needed before conducting the analysis; not even their respective dimensions need be known. On the other hand, it may be difficult to interpret the z' and z'' in terms of the substance of the empirical problem.[3]

2. EMPRICAL RESULTS

The statistical procedures developed in the previous section were applied to a set of real data in order to gain some insight into the practical problems to be expected in their application. The data were selected so as to compare the results from using the new technique against those from classical multiple regression in a situation where the use of the two approaches seemed to be reasonably compatible. In addition, it was hoped that the empirical findings would prove to be interesting in their own right.

Description of Data. Data on the distributions of income for families and unrelated individuals (in 1949) and educational achievement for male family heads were obtained from the census tract statistics of the city of Chicago for 1950.[4] These data make up the set of independent variables in the study. They consist of the proportions of a tract's population that fell into each of

[3] While not related to the empirical analysis reported in this paper, principal components regression may prove useful for exploring certain "errors in variables" regression problems were: (i) the errors in the independent variables are almost uncorrelated with the dependent variable *and* its errors; and (ii) the errors in the independent variables are either much more or much less correlated with one another than are the independent variables themselves. Condition (ii) is necessary to insure that the errors will be separated from the true variables by the principal axis method, while (i) implies that the appropriate components will be selected for inclusion in the regression upon y. Conditions other than (ii) may yield a separation of the variables' and errors, subspaces via principal components, but their nature is not well understood at the present time. The approach is related to Ragnar Fritsch's confluence analysis, which is well-explained in the paper by Sonte [9].

[4] U. S. Bureau of the Census, *17th Census of the United States*; *1950 Population, Volume 3,* Census Tract Statistics. Cross classifications of the dependent variables by income and education classes are not available from published census material.

the fourteen income classes (ranging from "$0–500" to "over $10,000") and nine education classes (ranging from "no formal education" to "college graduate or better") reported in the census. Only the marginal distributions for income and education were available; thus the number of independent variables was 23. The total number of census tracts included in the sample was 1090. The independent variables were subject to two linear restrictions, since the fourteen income and nine education variables had to sum to one, separately, for each tract.

Four dependent variables were chosen from the same source. Their names, definitions, and mean saturation[5] levels for the sample are as follows.

> *Television Ownership.* The proportion of the households in the tract that owned one or more television sets, whether or not the set was in working order at the time of the enumeration. Mean saturation = .31.
>
> *Refrigerator Ownership.* The proportion of households in the tract that had the use of a mechanical refrigerator for the refrigeration of food in the home. (Home ice chests, etc., were excluded.) Mean saturation = .90.
>
> *Central Heating.* The proportion of households in the tract that were domiciled in a dwelling unit that was heated by piped steam, hot water, or a central warm air furnace. Mean saturation = .72.
>
> *Overcrowding.* The proportion of households in the tract domiciled in a dwelling unit having an average occupancy rate of more than 1.01 persons per room. (Excluded in the definition of a room were bathrooms, halls, closets, porches, and other rooms "not suitable for living accommodations.") Mean saturation = .15.

In addition to the variables utilized in the principal components-regression analysis, the means of the income and education distributions were calculated for use as exogenous variables in some classical regression runs. They were computed using the midpoints of the class intervals as multipliers for the proportions; the two upper classes were open-ended, so the values $12,500 and 16 grades were chosen more or less arbitrarily for them.

The Demand Models. The purpose of the investigation reported here is to compare the results obtained from principal components-regression analyses utilizing the data described above with those from classical least-squares analysis based on the usual type of summary statistics that may be derived from the data. The following expressions for the demand for stocks of the four products being studied were used in the two analyses.

(i) Principal components-regression analysis. The flexibility of the principal components-regression approach permits utilization of the census information on the distribution of income and education within each tract in its most disaggregative form, regardless of the problem of collinearity. These distributions may be related to the tract's saturation (for each of the dependent variables) in the following manner:

$$Y_i = \sum_{j=1}^{14} b_{I_j}(P_{I_{ij}} - \bar{P}_{I_j}) + \bar{Y}_i + e_i,$$

[5] "Saturation" refers to the proportion of families in a given tract who own or exhibit the indicated attribute.

and (15)

$$Y_i = \sum_{j=1}^{9} b_{E_j}(P_{E_{ij}} - \bar{P}_{E_j}) + \bar{Y}_i + u_i.$$

In the equations, $P_{E_{ij}}$ and $P_{I_{ij}}$ refer to the percentage of households in the jth education and income class, respectively, in the ith tract. The e_i and u_i are the random error terms in the equations. The definitions of all the other variables and parameters should be obvious from the context. The sum of the above equations is also an equation. It has the desirable property of combining the income and education information into a single expression that is amenable to analysis by the principal components-regression technique. This gives the following demand function:

$$Y_i = \sum_{j=1}^{14} (\tfrac{1}{2}b_{I_j})(P_{I_{ij}} - \bar{P}_{I_j}) + \sum_{j=1}^{9} (\tfrac{1}{2}b_{E_j})(P_{E_{ij}} - \bar{P}_{E_j}) + \bar{Y}_i + v_i, \tag{16}$$

where the error v_i is equal to one-half the sum of e_i and u_i. The regression coefficients represent the effect of deviations in the income and education distributions of a given tract, from those for the sample as a whole, upon the saturation level for the tract. The empirical results reported are all based on beta coefficients of regression, which are related to the parameters of (16) by the expression $\beta_i = (\frac{1}{2}b_i)(\sigma_{x_i}/\sigma_y)$, where x_i is any member of the independent variables' vector.

The manner of reporting the beta coefficients in this study is noteworthy. The "beta profiles" presented in Figure 1 show these results in the form of a connected graph, which implies some continuity of effects between adjacent income and education classes. The procedure is valid because adjacent classes represent simple cuts along the same income or education distribution; thus the observed values of beta for two or more adjacent classes might be expected to be more nearly the same than would be the case for widely separated classes. It must be emphasized that this continuity of the several income and education dimensions is *not* a necessary condition for the use of principal components-regression analysis. The methods discussed in this paper will also work on sets of variables that do not share a logical commonality, although they may still be highly correlated among themselves. Pairs of variables like "average income" and "percentage of working wives" would also be acceptable, for instance.

The particular demand model and its associated data base used in this study were chosen because: (i) some information about the nature of the saturation-income and saturation-education effects—that is, the shapes of the beta profiles—was available on *a priori* grounds; (ii) the principal components-regression results could be compared with those of ordinary regression using summary information from the same data base in a fairly direct fashion; and (iii) there was an opportunity to investigate the stability of the principal components-regression procedure in terms of departures from the expected smooth shapes of the beta profiles. The properties of the estimation method depend only on the correlation properties of the data and not on the special nature of the model chosen for analysis, so the findings should also be valid for less well-defined sets of data.

(ii) Classical Regression. Some ordinary regression runs were made to provide a check against the results obtained by using the principal components-regression technique. It is clear that equation (16) is not appropriate for use in ordinary regression, since both the income and education variable sets are collinear. (The variables in each of them must sum to one.) The regression plane would be highly unstable even if one of the variables from each set were deleted to destroy the singularity of the system. Therefore, it is necessary to derive some summary variables from the data in order to perform the classical regression. It may be noted in passing that in the present case the principal components-regression method "works" precisely because it, too, calculates summary measures from the original data set prior to regression. In the classical case used here, summary measures are defined on *a priori* grounds rather than as summary principal components.

Many different sets of summary statistics could be calculated from the income and education distribution data. The simplest possible set was chosen; namely, the means of the distributions and their squares, as calculated for each tract in the sample.[6] The use of means has obvious intuitive appeal. In addition, they are highly correlated with the medians of the distributions, which have been used successfully in linear and quadratic regressions on durables goods ownership (cf. Dernberg [2]). The squared terms were required to provide degrees of freedom adequate to allow a reasonable comparison with the beta profiles obtained from the principal component-regression analyses. Thus, the classical regression used is:

$$Y_i = c_0 + c_1\bar{I}_i + c_2\bar{I}_i^2 + c_3\bar{E}_i + c_4\bar{E}_i^2 + e_i. \tag{17}$$

Once the parameters of (17) were estimated, the predicted value of y was calculated for each of a range of values of $\bar{I}$ and $\bar{E}$. These predictions were compared with the profile of the beta coefficients obtained from the principal components-regression analysis.

A preliminary word about the comparison of the two statistical procedures is in order at this point. The complete principal components-regression analysis contains a total of 21 free parameters which are to be estimated from the data (i.e., the total number of income and education classes minus the two linear restrictions mentioned earlier). The classical regression based on summary statistics of the detailed class information contains only four parameters. It has already been noted that the method of summarization is predetermined in the classical regression and is left to be determined from the data in the principal components-regression procedure. If the coefficient of determination (R^2) were adopted as the criterion for the relative success of the two approaches, the result would seem to be biased in favor of the principal components analysis, since it appears to have the larger number of degrees of freedom. But this way of thinking about the problem is not appropriate because:

a. components will be deleted so that some of the regressions in this category

[6] In retrospect, it would seem to have been more appropriate to calculate the means and the standard deviations of income and education from the class data, and use the latter in the regressions in places of the squares of the means. For most tracts, however, the two statistics are highly correlated with one another, so the difference is probably not important.

will contain four or even fewer parameters, and

b. the theoretical advantage of the principal components-regression technique over classical methods lies in its greater facility for handling large and highly interrelated sets of data directly—without the need for specifying summarization methods on *a priori* grounds.

The comparison is therefore based on the use of both techniques in a relatively "natural" manner: variations in the number of free parameters are merely an embodiment of the difference between the fields of applicability of the two approaches. In fact, the comparisons are biased in favor of the classical technique, since in the present instance the nature of the data, as well as a body of economic theory and previous empirical work, strongly indicated what summary measures should be chosen. This is an advantage that is not always present in exploratory statistical research.

The Principal Components and Their Correlations with the y's. Table 1 gives the sample means, standard deviations, and the loadings matrix for 20 principal components obtained from the set of 23 independent variables. The rank of Z was 21, but owing to its extremely small eigenvalue, the last component could not be evaluated because of rounding errors accumulated in the computer program. In addition to the loadings, the last row of the table gives the eigenvalue (d_{kk}) associated with each component. The gammas, or correlations between the principal components and dependent variables, are presented in Table 2, along with the standard errors of the gammas and the coefficient of determination for each regression.

The first two components account for most of the relationship between the y's and z's in all four cases. They also explain some 56 per cent of the variation of the z's themselves. As can be seen from Table 1, the first component is negatively loaded on the middle income and high education variables, and positively correlated with the low and high income and low education variables. Component number two is positively loaded on the middle education and low to middle income classes and negatively loaded on the high and low education and high income variables. Television ownership is negatively related to the first component and positively related to the second, and thus would seem to increase with the proportion in the high education classes. For the other classes, the effects of the two components run in opposite directions, making only the most tortuous interpretations possible. Similar conditions hold for the other dependent variables. It is therefore desirable to transform the results back into terms of the z's to make more definitive conclusions possible.

Beta Estimates for the Original Variates. The beta coefficients and their variance-covariance matrices were calculated for each of the dependent variables according to equations (12) and (13). Four sets of betas were obtained for each study in order to assess the effect of deleting different collections of principal components. Selection of the components to be included in each set was accomplished by examination of the ratio γ/σ_{γ}, which is an expression for the significance of the correlation between each principal component and the respective dependent variable. The four sets are based upon components with ratios of at least 2, 1, $\frac{1}{2}$, and 0 (the latter condition admits all the components regardless of their correlation with the dependent variable). The resulting

TABLE 1. LOADINGS MATRIX FOR STANDARDIZED PRINCIPAL COMPONENTS

(23 Income and Education Distribution Variables)

Income ($000)	Sample*		Principal Components*																			
	$\bar{X}$	σ_x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0– .5	028	029	35	24	57	52	16	21	−06	−16	−08	08	−14	13	00	14	−06	−06	−13	04	01	00
.5–1.0	064	045	68	22	47	13	−05	−04	−06	00	08	−18	09	−10	−03	−21	07	−01	26	−18	00	01
1.0–1.5	064	034	70	34	26	−19	−05	−16	03	09	11	−12	16	−07	−00	−11	18	03	−22	27	02	00
1.5–2.0	047	022	57	43	06	−30	−14	−13	04	10	−10	−23	−06	23	32	28	−03	−02	02	−08	−00	02
2.0–2.5	213	070	−14	77	−07	−16	−01	−03	05	29	−07	26	−31	−20	−12	−01	−10	−00	−02	−04	00	−05
2.5–3.0	156	042	−15	73	−23	−11	03	04	−02	−26	−14	−09	22	29	−23	08	−22	11	02	04	−00	−02
3.0–3.5	181	071	−84	−09	−24	−06	01	07	−01	−25	04	06	05	−03	04	04	22	−15	−04	−06	09	21
3.5–4.0	061	044	−68	−62	−03	−00	02	06	05	−08	10	−06	03	01	10	01	03	01	−01	−01	−07	−20
4.0–4.5	052	061	−57	−73	06	03	−00	07	09	08	08	−07	00	05	10	−02	−05	02	02	05	−09	−11
4.5–5.0	041	030	56	−46	−06	16	03	−05	−01	29	−01	29	42	−03	00	20	−12	03	06	01	05	05
5.0–6.0	022	019	60	−47	−27	21	−03	−15	−09	16	−23	−04	−09	15	−10	−11	12	−00	−08	−04	−27	10
6.0–7.0	020	017	59	−51	−30	17	−05	−13	−12	05	−13	−05	−14	16	−01	−08	10	04	−00	−00	32	−11
7.0–10.0	022	016	64	−47	−19	−01	03	03	−05	−12	−00	−21	−15	−18	−09	10	−18	−21	17	23	−01	04
10.0–up	035	020	75	−27	−13	00	−00	00	−06	−26	16	01	−15	−18	10	05	−07	39	−04	−03	−02	08
Education (grades)																						
0	136	094	84	−23	−03	−14	−04	−04	−04	−16	05	−03	09	−11	−16	09	−04	−13	−18	−18	−01	−10
1–4	169	071	78	08	−02	−14	−06	13	04	−17	−03	41	−05	14	18	−12	11	−03	16	11	−05	−03
5–6	127	044	32	41	−29	08	66	20	22	12	−09	−19	04	−09	09	−00	10	04	02	−00	00	−01
7	113	039	−06	52	−29	20	−40	50	−27	21	15	−15	02	−02	01	04	04	01	00	04	−01	−00
8	080	033	−23	44	−18	45	−33	−29	53	−10	03	−03	01	−05	01	03	00	00	01	02	−00	−00
9–11	138	051	−51	36	−07	13	26	−41	−29	07	43	03	−07	17	00	03	−02	−01	04	03	−02	01
12	081	038	−67	19	00	12	−02	−20	−33	−06	−35	−02	10	−23	31	−14	−13	−01	−05	03	−00	−00
13–15	098	052	−76	−09	26	−06	−03	−09	−10	−01	−26	−03	−02	−05	−22	25	24	17	13	07	−01	−02
16–up	059	089	−52	−64	28	−12	−04	10	19	19	−00	−09	−08	11	−02	−14	−19	−05	−03	00	09	16
Eigenvalues			8.09	4.82	1.35	.92	.86	.80	.72	.66	.61	.59	.52	.49	.45	.38	.38	.31	.28	.25	.22	.18

* Decimal points are to the left of the lead digit. Loadings can be interpreted as correlations between variables and principal components.

TABLE 2. CORRELATIONS BETWEEN PRINCIPAL COMPONENTS AND THE DEPENDENT VARIABLES

(The gamma coefficients)

Dependent Variables	Sample*		Principal Components*																			
	σ_a	R^2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Television	065	64	−59	43	17	01	06	07	00	12	−06	04	−03	09	−09	−06	−12	−00	−01	00	11	05
Refrigerators	072	71	−80	18	−08	01	02	13	04	−03	02	09	−05	02	01	−01	−05	01	−02	01	07	−02
Central Heating	186	55	−55	−32	−28	−05	−01	03	−03	−17	01	−01	08	−03	01	04	07	−04	−04	−00	08	11
Overcrowding	051	76	80	−17	−11	−07	−02	−03	−03	−20	01	−10	15	02	−00	04	05	01	03	−01	03	02

* Decimal points are to the left of the lead digit in all cases.

TABLE 3. PRINCIPAL COMPONENTS-REGRESSION SUMMARIES

Dependent Variable	Components Used in the Regressions			P.C.—Regression Results		R^2 for Classical Regressions	
	Criterion	Number	% Variance	R^2	$\bar{\sigma}_\beta$	Means and Means Squared	Means Only
Television	2σ#	3	62.0	.56	.014		
	1σ	9	75.0	.63	.050	.53	.51
	$1/2\sigma$	14	86.5	.64	.069		
	all	20	99.5	.64	.088		
Refrigerators	2σ#	2	56.1	.67	.009		
	1σ	5	68.0	.70	.029	.56	.46
	$1/2\sigma$	8	72.9	.71	.051		
	all	20	99.5	.71	.098		
Central heating	2σ	1	35.1	.30	.016		
	1σ#	3	62.0	.49	.040	.49	.43
	$1/2\sigma$	5	65.6	.53	.099		
	all	20	99.5	.55	.251		
Overcrowding	2σ	4	64.2	.70	.019		
	1σ#	7	73.7	.75	.027	.66	.61
	$1/2\sigma$	13	85.8	.76	.049		
	all	20	99.5	.76	.069		

group of 16 sets of 23 coefficients each was too large for presentation, but some of the results are given in Table 3 and Figure 1.

Summary statistics are presented in Table 3. The left-hand portion of the table shows the number of components whose correlations with y met the gamma criterion, the proportion of the variances of the z's that is accounted for by these components (per cent variance), the coefficient of determination for the regression (R^2), and the average of the standard deviations of the beta coefficients resulting from the principal components-regression analysis ($\bar{\sigma}_\beta$). The last quantity provides a simple summary of the precision of the beta estimates; it was calculated as the square root of the average of the diagonal elements of Σ_β. Two results emerge from the left-hand portion of Table 3: (i) most of the explanatory power of the regressions is concentrated in a relatively small number of principal components; and (ii) the average standard errors of the beta coefficients go up sharply as the number of principal components in the regression is increased. The first point has already been noted in connection with Table 2. The second will be considered further below.

The solid lines in Figure 1 are profiles of the beta coefficients calculated for the regressions denoted by (#) in Table 3. According to the television profile, tracts with a relatively high incidence of families with incomes above $4000 per year tended to have lower television saturations than those with a larger proportion of low income families. Since these are essentially regression results, the income coefficients must be interpreted as if the education variables were held constant at their mean values. For income constant, it appears that increases in educational attainment are associated with increased television saturation up until the highest education class is reached, when the relationship is reversed. It must be recalled that, according to the arguments given in the first part of this paper, the television profiles presented in Part A of Figure 1 are based on only three degrees of freedom, since only three principal components were utilized in the underlying regression. But it will be seen that these conclusions are compatible with those obtained from the classical regression and, moreover, that they are not greatly affected when the number of degrees of freedom is increased by the addition of more principal components.

From the beta profile for refrigerators it appears that for education held constant, increments in mid-range of the income distribution produce disproportionate increases in percentage saturation. (Once again, this finding is not substantially altered by the addition of more principal components.) The findings for education are more nearly in line with expectations: ownership of refrigerators increases rapidly at the lower end of the education distribution and then levels off at the higher levels where saturation is virtually 100 per cent. Roughly the same pattern is observed for the education variables in the central heat case: the beta profile first rises sharply and then levels off. The central heat-income relationship is more reasonable than its counterpart for refrigerators, since tracts with a high incidence of low income families exhibit very low saturations, but the tendency for the profile to reach a peak at the three to four thousand dollar class is still very noticeable.

Education is the most important determinant of dwelling unit overcrowding. There also seems to be some tendency for overcrowding to increase along with the proportion of households in the higher income brackets, which is contrary to expectations. It will be seen, however, that the same result is obtained with the classical least-squares method, so the anomaly cannot be attributed to the use of principal components. The sharp oscillation of the profile in the two to three thousand dollar income range is due to the relatively large number of components used in this regression, as will be discussed below. Further discussion of the principal components-regression results will be postponed until after the classical results are considered.

Comparison with the Classical Regression Results. Coefficients of regression for each of the dependent variables upon tracts' mean income and education and their squares were estimated by classical least squares. The results are presented in Table 4 and summarized in the right-hand portion of Table 3. The squared terms are highly significant in all cases, as can be seen from the F tests for the addition of these two terms to the regression. Since they were highly correlated with mean income and education, it is not surprising that the coefficients for the latter were changed markedly and reduced in significance by the addition of the squared terms to the regression. The first set of results is generally consistent with that reported by Dernberg [2] in his study of television ownership.

While the coefficients of determination are fairly large, they are smaller than the ones obtained by using the principal components-regression technique. This is true even where the number of components is reduced to two or three. Hence it may be concluded that the first few principal components contain

TABLE 4. SIMPLE REGRESSION RESULTS

(*t* ratios for coefficients in parentheses)

Television	
$TV = -0.336 + .170I - .0289I^2 + .070E - .00163E^2$ (2.79) (4.44) (10.12) (4.03)	$R^2 = 0.53$
$TV = + .440 - .109I + .042E$ (−21.81) (32.16)	$R^2 = 0.31$
F ratio for inclusion of squared terms: $F(2, 1085)^* = 21.1$	
Refrigeration	
$RF = -0.591 + .2361 - .024I^2 + .171E - .0071E^2$ (3.21) (3.06) (20.29) (14.55)	$R^2 = .56$
$RF = +0.568 - 0.021I + 0.049E$ (3.25) (28.87)	$R^2 = .456$
Heat	
$H = 4.90 + 1.894I - 0.172I^2 + .101E - .0033E^2$ (11.58) (9.84) (5.40) (3.05)	$R^2 = .49$
$H = -0.855 + 0.269I + 0.040E$ (19.66) (11.17)	$R^2 = .431$
$F(2, 1085)^* = 71.5$	
Overcrowding	
$OC = 0.656 + 0.041I - .00146I^2 - 0.113E + 0.0040E^2$ (0.84) (0.28) (20.14) (12.2)	$R^2 = .66$
$OC = 0.331 + 0.046I - 0.046E$ (10.89) (41.11)	$R^2 = .614$
$F(2, 1085)^* = 76.1$	

* $F(2, \infty)_{.995} = 5.30$.

more information about each of the dependent variables than do the means of the distributions and their squares. The conclusion about the information content of the means can be refined if their coefficients of determination are compared with those obtained by using only the two best principal components in each case. The R^2 for the two component regressions can be calculated by taking the sum of squares of the first two gammas in each row of Table 2; they are .53, .67, .40, and .67, respectively. Central heating is the only dependent variable for which the two-component regression fails to do better than the classical regression on the means alone, as can be seen by comparing the numbers given above with those in the last column of Table 3.

The partial regression curves are given in Figure 1. Those for income were calculated by substituting the ordinate values of income into equation (17) while holding education constant at its over-all mean value, and similarly for the education curves. The comparison with the beta profiles that are also presented in these charts is rough, since the two curves mean somewhat different things, the number of degrees of freedom differs, and the ordinates refer to average income in one case and to particular income class limits in the other. Nevertheless, the two sets of estimates should exhibit roughly the same shape, and indeed they do so for most of the cases. There is good agreement between the television-education, refrigerators-education, and overcrowding-education curves. Allowing for the differences reflected in the ordinates of the curves, there is good agreement between them for central heat and income. Those for television-income and central heat-education are reasonably consistent, while the divergence between the beta profile for overcrowding and income and the classical regression curve is not especially significant, for reasons to be discussed below. Only the refrigerator-income curves do not fit together; even though the classical regression parabola would slope downward to match the beta profile at the upper income levels, this extension is not valid because it would run well beyond the range of the observed means for the tract.

It was thought that the divergence of the refrigerator-income curves might be due to the education effect swamping that of income in the principal components-regression analysis. Alternatively, it was possible that with only two components in the regression, the two degrees of freedom available for estimating the betas were insufficient to allow an adequate fit to the data. Both hypotheses were explored and both were found to be false. The regressions incorporating 5, 8, and 20 components showed roughly the same shape as did the one for 2 components (except for decreased stability of the type to be discussed below). A principal components-regression run for refrigerators on 12 principal components of the income variables by themselves produced betas that were nearly the same as those for the 8-component income-education case. The refrigerator-income beta profile thus remains a puzzle, especially since the classical regression results appear to be reasonable.

Effect of Adding Components to the Regression. The principal component-regression results reported so far are all based upon a relatively small number of components. The curves become much less stable as more components are added, and there seems to be a tendency for the betas to fluctuate about a fairly smooth trend as the number increases. The profiles given in Figure 1 are reasonably good representations of this trend in all of the cases, which is one reason why they were presented in preference to the other alternatives.

A comparison of two beta profiles for television, based on 3 and 14 principal components, respectively, is presented in Figure 2. Two standard deviation confidence limits have been added to the estimated profiles. The 3-principal components curve falls within the confidence limits for the 14-component one for all but 2 of the 23 variables (Figure 2, Part B). Discounting the irregular fluctuations in the 14 component curve, their shapes are fairly equivalent, except for the highest income and education levels.

A noteworthy facet of the two charts is that the confidence limits on all the beta coefficients increase rapidly as components are added to the regression, a fact already implied by the average standard deviations given in Table 3.

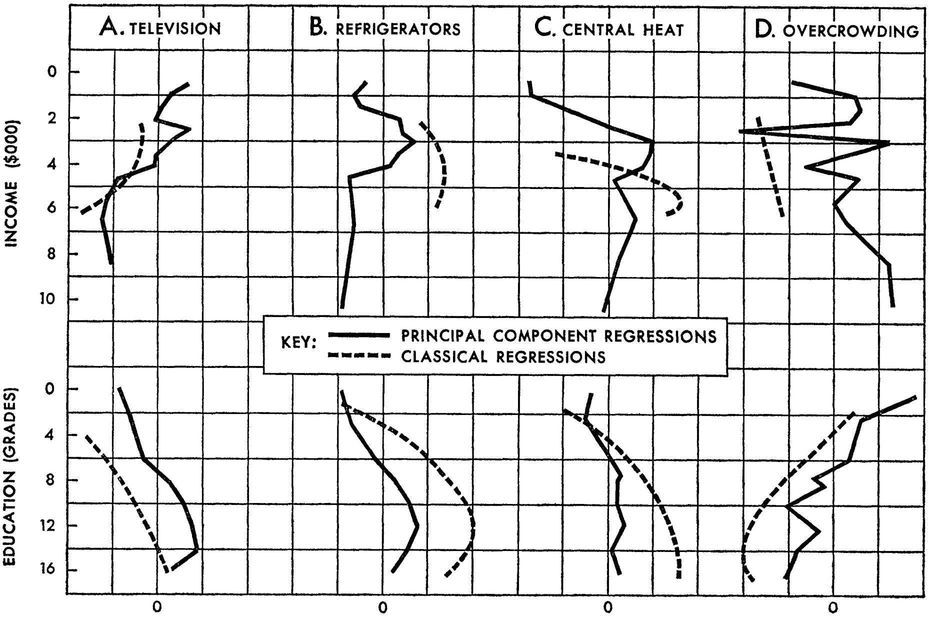

Fig. 1. Beta Coefficient Profiles and Classical Regression Saturation Estimates*
Values of Beta and Estimates of Saturation

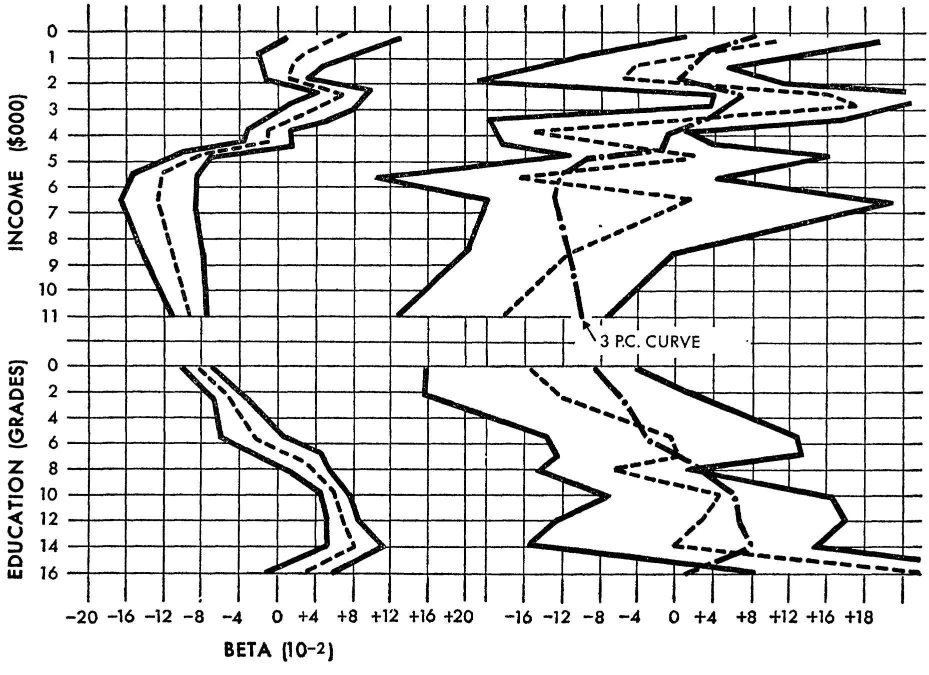

Fig. 2. Two Sigma Confidence Limits on Beta Profiles—Television; A. 3 Principal Components, B. 14 Principal Components.

The *X* scale is not relevant for comparing the two types of regression methods. (All four scales are the same.)

There is a good reason for this result: as components are added to the regression, linear restrictions on the betas are removed without a commensurate decrease in σ_e^2. Since as far as the principal components-regression technique is concerned these restrictions represent *a priori* information about the beta profiles, their uncompensated loss must result in a reduction of estimating efficiency. This shows that, other things being equal, it is possible to get more precise estimates of the betas in restricted subspaces, although they are also more highly correlated among themselves. (And any group of more than m betas is perfectly collinear.) Of course the various beta profiles refer to projections of the original variates into different subspaces, so that they are not entirely equivalent quantities. Nevertheless, the degree of identity between the projections is strong.

3. CONCLUSIONS

The empirical research reported in this paper was aimed at finding whether, from the same set of data, the principal components-regression analysis of complex data can yield results that are comparable with those from traditional regression analysis that uses well-accepted summary statistics. In addition, the effect of increasing the number of components used in the regressions to include some that are not strongly correlated with the dependent variable was examined empirically. The following general conclusions were obtained:

a. The principal components-regression results are in substantial agreement with those obtained from regressions on the tracts' mean income and education levels and their squares in almost all of the cases studied, as judged by comparing the shapes of the beta profiles from the former with the partial regression slopes obtained from the latter. That is, most conclusions about the effects of income and education upon the saturation levels of the four dependent variables would be about the same as those obtained by traditional means—although the degree of subjectivity required is higher.

b. Coefficients of determination for the principal components-regression analyses are almost always higher than those for the regressions upon the tracts' means and their squares. While the differences are not large, the principal components-regression technique yielded the larger R^2 while employing fewer regressors than their classical counterparts in three out of the four cases.

c. Only those components that are strongly related to the dependent variables can be included in the principal components-regression analysis if the beta profiles are to be used as the basis for conclusions about structure. The addition of components not so related caused the beta profile to become unstable and difficult to interpret, although the general shape of the profile is not substantially altered. The exact number of principal components to be included was found by trial and error.

These results are important because they show that principal components-regression analysis can lead to reasonable results, even in situations where the input data are highly collinear. If theoretically appropriate summary variables for the original data set had not been available, the results reported above might have been used to define such a set of statistics. The new variables would presumably be highly correlated with those original variates for which the beta profiles assumed extreme values, while being more directly interpretable in terms of the structure of the problem than are principal components. These variables, and hypotheses based on them, would in turn have formed the basis for subsequent statistical analysis.

Regression upon principal components appears to be worthwhile during the exploratory phases of empirical research. The methods explored in this paper are useful because: (i) they permit rapid calculation of the correlations between the dependent variable and each of the components, and (ii) they refer the regression results back to the projections of the original independent variables into the space spanned by the components included in the given regression. Furthermore, the characteristics of the betas can be traced through different subspaces quickly and easily, without the need to recalculate any principal components. This flexibility is very important in exploratory research.

While most methods for summarizing a group of variables in a space of smaller dimension involve a loss of information, this strategy is often desirable. Where the basic set of variables is unworkable in its original form—whether because the number of variables is too great for available analysis methods, because of multicollinearity, or for any other reason—some kind of summarization is necessary, and the only question is how it should be accomplished. Summarization of the income and education distribution variables used in this study was accomplished by calculating the distribution means and their squares for each of the tracts; in this case, the information content of the original set was collapsed into a particular subspace of dimension four. But the basis for summarization is not always obvious *a priori*. Transformation to principal components boasts the two advantages discussed earlier in this paper: (i) the components are orthogonal, allowing easy exploration of alternative subspaces in relation to the dependent variable's vector in the sample space; and (ii) each component contains a maximum amount of information, consistent with being uncorrelated with the previous ones. If the correlation with the dependent variable is used as the criterion for retaining components in the regression, the subspace into which the independent variables are to be projected is chosen in part on the basis of its proximity to the vector that is to be predicted or explained.

A major problem in ordinary principal components analysis is to give substantive meaning to the components after they have been discovered. When combined with regression, it is necessary to identify the components in order to give meaning to the regression coefficients. In the present case this problem is partially overcome by referring the regression results back to the terms of the original set of independent variables, as projected into the summary subspace.

The methods explored in this paper are not intended to be substitutes for the established principals of statistical inference: hypothesis building and testing. The author bows in advance to all charges of espousing measurement without theory, but submits that this is necessary in the early states of empirical research. Hypotheses may be developed after certain kinds of summary variables have been suggested by looking at data and relating them to fragments of theory. These hypotheses should then be tested upon a fresh sample. While supplanted by more precise techniques at a later stage of research, the ability to rapidly and conveniently explore relationships between variables in a complex set of data can be of great importance. More theoretical and empirical work on methods for untangling complex relationships by means of flexible statistical procedures, including the ones discussed here, is urgently needed.

REFERENCES

[1] Anderson, T. W., *Introduction to Multivariate Statistical Analysis*. New York: John Wiley and Sons, Inc., 1958.

[2] Dernberg, T. F., "Consumer response to innovation: television," in Dernberg, Rossett, and Watts, *Studies in Household Economic Behavior*. Yale Studies in Economics, IX. New Haven: Yale University Press, 1958.

[3] Frank, R. E., Kuehn, A. A., and Massy, W. F., *Quantitative Techniques in Marketing Analysis*. Homewood, Illinois: Richard D. Irwin, Inc., 1962.

[4] Hadley, G., *Linear Algebra*. Reading, Massachusetts: Addison-Wesley Publishing Company, 1961.

[5] Hotelling, H., "The relations of the newer multivariate statistical methods to factor analysis," *British Journal of Statistical Psychology*, 10 (1957), 69–79.

[6] Johnston, J., *Econometric Methods*. New York: McGraw-Hill Book Company, 1963. P. 131.

[7] Kendall, M. G., *A Course in Multivariate Analysis*. London, England: Charles Griffin and Company, Ltd., 1957.

[8] Perlis, S., *Theory of Matrices*. Cambridge, Massachusetts: Addison-Wesley Publishing Company, 1952.

[9] Stone, J. R. N., "The analysis of market demand," *Journal of the Royal Statistical Society, New Series*: 108 (1945), 286–382.