CANONICAL CORRELATION ANALYSIS OF SELECTED DEMOGRAPHIC AND HEALTH PERSONNEL VARIABLES*

James W. Dunn and Gerald A. Doeksen

The number of employed health care personnel in an area is the combined result of factors affecting both demand and supply for health care personnel. When service areas are compared, differences between each area's number of health care personnel are related to differences in health facilities' ability and desire to attract these persons.

At any point in time, the potential supply of health care personnel is fixed. The choice facing this segment of the labor force is between alternative locations or non-participation. The location decision is a function of salary, working conditions and such non-working conditions as cost of living in the community, schools, cultural opportunities, employment opportunities for other family members and general amenities associated with a community.

Demand for a certain type health care personnel is a function of the cost of obtaining these individuals, demand for health care in general and cost of obtaining substitute and complementary health personnel and facilities. Demand for health care in general is affected by such items as income, education, present health, age, ethnic background and other determinants of tastes and preferences.

Demographic factors influence both supply and demand. It is hypothesized that individual variables are primarily related to demand, while a combination of several variables influence supply. Those influencing supply are characteristics that make a community a more attractive place to work and live. Such a community, in the eyes of most health professionals, would be an economically thriving one with good schools, educated people, little poverty and few minorities.

Separation of these many factors into a concise, meaningful model using traditional methods would require data that are not available. One method of reducing the problem to a manageable size is with canonical correlation. In particular, what is proposed here is to use canonical correlation to compare per capita numbers of health care personnel with selected demographic variables to determine what relationships exist between them and how demographic variables affect health care variables. While health care variables affect demographic variables in the long run, major short run effects are on health care personnel rates by the demographic factors.

CANONICAL CORRELATION[1]

Canonical correlation analyzes interrelationships between two sets of measurements on population. Linear combinations of the two data sets having the largest correlation between them are found. Subject to the condition of orthogonality to all previously derived canonical variates, subsequent pairs of linear compounds with the next highest correlation between them are found, with total number of correlate pairs equal to number of variables in the smallest set of variables.

VARIABLES

For the purposes of this analysis one variable set, composed of demographic variables, is viewed as the predictor. The other variable set, per capita numbers for various health personnel, is viewed as the criterion set. More specifically, the demographic and health

James W. Dunn is Research Assistant, Department of Agricultural Economics, Oklahoma State University and Gerald A. Doeksen is Economist, EDD, ERS, Stillwater, Oklahoma.

*Oklahoma Agricultural Experiment Station Article 3283.

[1] A further explanation of canonical correlation may be found in Morrison, [1, Chapter 6].

personnel variables are:

Demographic Variables

(1) Population (POP) = 1970 population of the county
(2) Young population (PLT15) = percent of the 1970 population of the county less than 15 years of age
(3) Middle aged population (POPMID) = percent of the 1970 population of the county greater than 45 years of age and less than 65 years of age
(4) Elderly population (PGT65) = percent of the 1970 population of the county greater than 65 years of age
(5) Non-white population (NWPOP) = percent of the 1970 population of the county which is non-white
(6) Mean family income (AVINC) = mean family income of the 1970 population of the county
(7) Education (EDUC) = median school years completed by persons 25 years and older in the county as of 1970
(8) Family size (FAMSIZE) = mean family size in the county, 1970
(9) Low income families (POV) = percent of the families in the county with less than poverty level of income in 1970

Health Personnel Variables

(1) Physicians (PHYSPC) = number of physicians per capita in county, 1970
(2) Dentists (DENPC) = number of dentists per capita in county, 1970
(3) Registered nurses (RNPC) = number of registered nurses per capita in county, 1970
(4) Licensed practical nurses (LPNPC) = number of licensed practical nurses per capita in county, 1970
(5) Pharmacists (PHARMPC) = number of pharmacists per capita in county, 1970
(6) Radiologists (RADPC) = number of radiologic technicians per capita in county, 1970
(7) Dieticians (DIETPC) = number of dieticians per capita in county, 1970
(8) Physical therapists (PTPC) = number of physical therapists per capita in county, 1970.[2]

The 77 counties of Oklahoma were the spatial unit used to divide the population. These counties are of a rather uniform geographic size, yet demographically are quite heterogeneous. This heterogenity may offer sufficient breadth of observations to support generalizations of findings to other areas of the United States, particularly the Great Plains and the South.

The Bartlett chi-square test, with a five percent significance level, was used to determine how many of the variate pairs to analyze carefully.

EMPIRICAL RESULTS

An initial step in canonical correlation analysis is an inspection of the correlation matrix (Table 1). Proper analysis begins with a simple examination of the correlations' significance. For the degrees of freedom available with our data, any correlation coefficient with an absolute value greater than or equal to 0.27 is significant at the five percent level. In general, there is significant correlation within groups, i.e., between one health personnel variable and others, and between one demographic variable and other demographic variables. Among predictor set variables, 23 of the 36 correlations are significant, along with 19 of the 28 correlations among criterion set variables. Insignificant correlation exists between health personnel and demographic variables, as only 25 of the 72 correlations between predictor variables and criterion set variables are significant.

The canonical model was estimated for the variable sets using the Statistical Analysis System (SAS) canonical correlation routines. Three of the eight canonical variate pairs proved significant at the five percent confidence level. The coefficient, or weight, for each variable in these significant variates is given in Table 2. For greater ease of interpretation, these weights are a transformation of the SAS weights. In particular, they represent weights appropriate for normalized data, with a value of ±1.0 given to that weight having the largest absolute value: All others are scaled accordingly. This does not affect the results, since correlation is unaffected by a linear transformation of one or both variates, and since the relative magnitude of the coefficients is of interest and absolute magnitude is not.

The first canonical variate pair had a correlation

[2]Data were obtained from Profile of Regional Health Variables—County Detail [2], published by the Oklahoma State Health Planning Agency in 1972, with the exception of some of the demographic variables, which were updated to 1970, using the 1970 Census of the Population, General Social and Economic Characteristics—Oklahoma [3]. Variables selected for inclusion included nearly all listed in the Profile of Regional Health Variables, which fell into the general categories suggested by the data set titled. All variables except population were adjusted to rates rather than absolute levels to ensure the observed relationships were more than simple agglomeration effects.

TABLE 1. CORRELATION COEFFICIENTS BETWEEN THE VARIABLES

	POP	PLT15	POPMID	PGT65	MWPOP	AVINC	EDUC	FAMSIZE	POV	PHYSPC	DENPC	RNPC	LPNPC	PHARMPC	RADPC	DIETPC	PTPC
POP	1.000	0.246	-0.320	-0.441	0.072	0.433	0.289	0.181	-0.301	0.596	0.305	0.431	0.022	0.008	0.392	0.326	0.179
PLT15		1.000	-0.481	-0.610	0.320	0.084	-0.078	0.303	0.060	-0.073	-0.142	-0.078	-0.336	-0.316	0.199	-0.166	0.081
POPMID			1.000	0.805	-0.167	-0.211	-0.296	-0.805	0.182	0.014	-0.150	-0.193	0.238	0.058	-0.260	-0.360	-0.148
PGT65				1.000	0.008	-0.520	-0.451	-0.605	0.430	-0.147	-0.140	-0.290	0.157	0.209	-0.283	-0.284	-0.257
MWPOP					1.000	-0.419	-0.482	0.388	0.585	-0.179	-0.197	-0.360	-0.221	-0.177	-0.093	-0.147	-0.188
AVINC						1.000	0.740	-0.015	-0.838	0.412	0.397	0.550	0.037	0.122	0.224	0.292	0.222
EDUC							1.000	0.068	-0.910	0.367	0.584	0.723	0.124	0.256	0.193	0.404	0.262
FAMSIZE								1.000	0.068	-0.113	0.050	0.071	-0.095	-0.086	0.088	0.364	0.022
POV									1.000	-0.387	-0.493	-0.640	-0.124	-0.187	-0.211	-0.294	-0.262
PHYSPC										1.000	0.555	0.622	0.386	0.316	0.546	0.270	0.451
DENPC											1.000	0.621	0.212	0.495	0.288	0.351	0.070
RNPC												1.000	0.300	0.377	0.547	0.425	0.416
LPNPC													1.000	0.180	0.267	-0.062	0.353
PHARMPC														1.000	0.106	0.074	0.030
RADPC															1.000	0.220	0.490
DIETPC																1.000	0.127
PTPC																	1.000

of 0.863, with an observed significance level of 0.0001, using Bartlett's chi-square test mentioned previously. The canonical weights for the first variate pair indicate that a high median education (EDUC), large average family size (FAMSIZE), a large percentage of the population living in poverty (POV), and a large percentage of the elderly population (PGT65), is associated with a large number of registered nurses per capita (RNPC), a small number of radiologists per capita (RADPC), and a large number of dieticians per capita (DIETPC). Interpretation of a variable's importance based on the weights can be misleading, since variables within a data set are not independent. A more accurate interpretation must consider the correlation between variables in a data set and that set's canonical variate. These correlations, or loadings, provide information about the relative contributions of variables to each independent canonical relationship. Loadings for the first three canonical variates are given in Table 3.

TABLE 2. CANONICAL VARIATE COEFFICIENTS FOR THE FIRST THREE CANONICAL VARIATES

Variables	Variate One	Variate Two	Variate Three
Demographic			
POP	0.272	1.000	-0.365
PLT15	-0.039	0.108	0.589
POPMID	0.264	0.431	-1.000
PGT65	0.317	-0.815	0.359
MWPOP	-0.190	0.252	0.201
AVINC	0.109	-0.151	0.036
EDUC	1.000	-0.613	0.331
FAMSIZE	0.468	-0.708	-0.617
POV	0.385	-0.482	-0.208
Health Supply			
PHYSPC	-0.002	1.000	-0.934
DENPC	0.232	-0.186	0.848
RNPC	1.000	0.074	0.381
LPNPC	0.183	-0.329	-1.000
PHARMPC	0.003	-0.398	-0.204
RADPC	-0.493	0.105	0.501
DIETPC	0.369	-0.282	-0.514
PTPC	-0.030	-0.170	0.638
Correlation	0.863	0.710	0.638
Chi-Square	217.150	124.410	76.800
Prob > Chi-Square	0.0001	0.0001	0.0009

The sum of squared loadings of the canonical variables divided by the number of variables in the data set indicates the proportion of total variance of that data set explained by that variate. For the first canonical variate, 24.8 percent of the predictor variables' variance is explained by that variate and 28.5 percent of the criterion variables' variance.

Of the individual variables, EDUC loaded the heaviest of the predictor set (0.878), followed by POV (−0.733) and AVINC (0.605), with RNPC (0.878), DENPC (0.717), DIETPC (0.615) and PHYSPC (0.522) leading the ordering for the criterion variables.

Here arises an instance illustrating the value of considering both weights and loading. The poverty variable has a positive weight and a negative loading, so an analysis using only one of these measures would be suspect in terms of accuracy. In this instance, consideration of the loadings seems to render most satisfactory results, but this is certainly not always the case.

The first canonical variate pair may be viewed as a comparison of a general index of health personnel and services and the index of demographic variables most highly correlated with it. Demographic characteristics associated with a high level of health care

TABLE 3. PROPORTION OF TOTAL VARIANCE OF THE VARIABLE SETS EXTRACTED BY THE FIRST THREE CANONICAL VARIATES

	First Variate			Second Variate			Third Variate		
	r	r^2	% of Total Variance Explained[a]	r	r^2	% of Total Variance Explained	r	r^2	% of Total Variance Explained
Demographic									
POP	.405	.164	07.4	.796	.634	56.3	-.015	.000	00.0
PLT15	-.317	.100	04.5	.395	.156	13.9	.674	.454	30.7
POPMID	-.176	.031	01.4	.043	.002	00.2	-.612	.374	25.3
PGT65	-.279	.057	02.6	-.303	.092	08.2	-.552	.305	20.6
MWPOP	-.425	.180	08.1	.063	.004	00.4	.031	.001	00.1
AVINC	.605	.367	16.5	.367	.134	11.9	.252	.063	04.3
EDUC	.878	.771	36.6	.049	.002	00.2	.379	.144	09.7
FAMSIZE	.148	.022	01.0	-.248	.061	05.4	.192	.037	02.5
POV	-.733	.537	24.1	-.202	.041	03.6	-.321	.103	07.0
$\Sigma r^2/P$		0.248			.125			.165	
Personnel									
PHYSPC	.522	.272	11.9	.699	.489	57.4	-.189	.036	06.5
DENPC	.717	.514	22.6	.078	.006	00.7	.204	.042	07.6
RNPC	.878	.770	33.8	.233	.054	06.3	.215	.046	08.3
LPNPC	.300	.090	04.0	-.055	.003	00.4	-.544	.296	53.7
PHARMPC	.409	.167	07.3	-.260	.067	07.9	-.062	.004	00.7
RADPC	.194	.038	01.7	.426	.181	21.2	.280	.079	14.3
DIETPC	.615	.378	16.6	-.065	.004	00.5	-.051	.003	00.5
PTPC	.222	.049	02.2	.219	.048	05.6	.211	.045	08.2
$\Sigma r^2/q$		0.285			.106			.069	

[a]May sum to other than 100% due to rounding.

personnel are a high education level, a low number of poor families and a high average income. These results are consistent with variables hypothesized to affect the supply of health personnel. Additionally, these variables were discussed as those positively influencing the demand for health care, since communities having these characteristics should exhibit both a desire for good health care and an ability to pay for it. This first variate is one, therefore, in which supply and demand factors are intermingled, but in which the hypothesized effects are in the same direction for a locally financed health care system.

A correlation of 0.710, with an observed significance level of 0.0001, was exhibited by the second canonical variate pair. Variables in the predictor set having large weights were POP, PGT65, FAMSIZE, EDUC, POV and POPMID. Of these, FAMSIZE, PGT65, EDUC and POV had negative weights. The criterion variable with heaviest weight was PHYSPC bearing a positive coefficient with PHARMPC and LPNPC each having considerably smaller weights of the opposite sign.

The second variate pair explained, respectively, 12.5 percent and 10.6 percent of the predictor and criterion variables' variance. Loadings reveal that only one variable from each set is highly correlated to its respective variate, POP (0.796) from the demographic variables, and PHYSPC (0.699) from the health care personnel variables. In this variate pair, the criterion variate may be viewed as a physician availability index, adjusted by the relative numbers of support personnel. The correlation indicates that the primary criterion in physician availability is population. This is expected since physicians, particularly specialists, require large populations to support them, and those located in non-metropolitan areas have considerably higher costs per patient than those in more populous areas. This would reduce demand and, *ceteris paribus*, per capita numbers.

An observed significance level of 0.0009 and a correlation of 0.638 characterized the third and last significant canonical variate pair. Of the predictor variables, POPMID and FAMSIZE drew large negative weights with smaller positive weights going to PLT15 and PGT65 and EDUC and a smaller negative weight on POP. In the criterion variable set, large negative weights were placed on LPNPC and PHYSPC. Smaller weights are placed on DEMPC, PTPC, DIETPC, RADPC and RNPC, with only DIETPC receiving a negative weight.

This canonical pair explained 16.5 percent of the variation of the demographic variables, and 6.9 percent of the health care personnel variables. Variables from the predictor variables set receiving large loadings were the three age distribution variables, PLT15, POPMID and PGT65, these receiving loadings of 0.674, −0.612 and −0.552, respectively. LPNPC received the only large loading in the second variable set with a −0.544.

The variate pair represents a demand difference

where a population with a large percentage of older people requires more per capita health care than does a younger population. The variate indicates that the health manpower group sensitive to such changes is composed of licensed practical nurses, because they are the ones with fewest size economies or diseconomies due to low salaries and greater willingness to work part time.

Altogether, 53.8 percent of the variance of the demographic variable set was explained by the first three variates and 46.0 percent of the health personnel set. A further evaluation of the relationship between the two variable sets can be made by examining the redundancy, or informational overlap, of the criterion set given the predictor set. This is the proportion of the variance explained in the criterion set times that of shared variance between the two sets, i.e., the squared canonical correlation coefficient. Hence, for the first canonical pair the informational overlap is $(0.863)^2$ (0.285) or 21.2 percent. Similarly, the redundancy is 5.3 percent for the second variate pair and 2.8 percent for the third. Summing these three redundancy rates, a total redundancy of 29.3 percent is obtained, i.e., the total amount of informational overlap on the criterion variable set, given the predictor variable set, is 29.3 percent. Thus, the explained variance measure of 46.0 percent overstated the actual explanatory power of the model.

SUMMARY AND CONCLUSIONS

In summary the analysis indicates three things. To support a high level of health care personnel and services in general, a prosperous area is shown to be a primary criterion. "Prosperous area" is defined as an area with a relatively large, well-educated high income population, such an area generally having a relatively small poor and minority population. The communities able to support relatively high physician rates, relative to other personnel rates, are characterized by large populations. Conversely, those with few physicians relative to other health personnel are characterized by low populations. Lastly, those areas with larger proportions of older residents have unusually large numbers of licensed practical nurses, reflecting their role as the most divisible health personnel type, as seen by their heavy usage in homes for the elderly and convalescent.

The theoretical relationships between demographic variables and levels of health care personnel are supported by the canonical variates. The method yields additional and different results from the overused regression analysis in this situation. While regression procedures have their advantages, they would not indicate how general economic welfare affects, through both supply and demand, general health care personnel levels, since in regression only a single variable may be regressed per equation. Additionally, regression would not reveal that the variance in physician rates, not correlated with other health personnel rates, varies almost exclusively with population.

Beyond the support of certain health economics supply and demand theory, the analysis shows that canonical correlation is a potentially valuable tool in economics.

REFERENCES

[1] Morrison, Donald F. *Multivariate Statistical Methods*, New York, McGraw-Hill Book Company, 1967.

[2] Oklahoma State Health Planning Agency. *Profile of Regional Health Variables*, County Detail, Oklahoma City, 1972.

[3] U.S. Department of Commerce. *1970 Census of Population—General Social and Economic Characteristics—Oklahoma*, Washington, 1970.