AN AD HOC PROCEDURE FOR REDUCING THE NUMBER OF VARIABLES TO BE INCLUDED IN A LINEAR MODEL

Jackie Smith and Luther Tweeten*

Department of Agricultural Economics, Oklahoma State University, Stillwater, Oklahoma

INTRODUCTION

In explaining phenomena in the social sciences, theory almost always supplies more admissible hypotheses than statistical procedures such as multiple regression analysis can accommodate without degeneration of estimates due to multicollinearity. In the case of a linear model that is being estimated by the least-squares technique, the inclusion of strongly related independent variables results in multicollinearity, causing problems in determining the relative influence of the explanatory variables. This paper presents an ad hoc procedure for reducing the number of variables that are included in a model while preserving structural integrity of the theoretical model. The procedure is applied to an economic development research problem. A limited number of variables were chosen from a larger set of variables using a grouped-variable technique.The paper identifies some elements accounting for the outlays by the Economic Development Administration to generate jobs through industrial development projects.

MULTICOLLINEARITY

Multicollinearity involves the existence of a linear relationship among the explanatory variables. When an exactly linear or nearly linear relationship exists, it is difficult to estimate the parameters associated with the explanatory variables in regression analysis. Coefficient estimates are unstable and standard errors of the coefficients are large (1, p. 153). When multicollinearity is present, removing from the regression one of two or more highly correlated variables does not markedly reduce the proportion of variance in the dependent variable accounted for. Eliminating some of the independent variables, obtaining new data, or utilizing *a priori* information concerning the coefficients are possible solutions to the problem. The last two methods were judged to be unsatisfactory for this study. This paper relates how variables were grouped into closely related sets and how one of these variables was chosen as a "representative" from the homogeneous group so that the effect of this most significant variable (i.e. the variable that most effectively conveys the influence of that group of variables on the dependent variable) is not unduly distorted because of multicollinearity problems.

AN AD HOC PROCEDURE

The study evaluates the efforts of the Economic Development Administration (EDA) to generate jobs in private industry in economically lagging areas (2). It examines the influence on cost-effectiveness (EDA cost per job generated) of several variables including unemployment, underemployment, skill levels, region, and proximity to population centers. In 1970 the Programs Analysis Division of EDA and the Boise Cascade Center for Community Development estimated direct jobs generated and EDA cost per project for 250 EDA job-development projects (3). Data from these two independent evaluations (125 projects by each evaluator) were used in this study. First applying regression analysis to just the successful projects (projects where at least one job was created), this study estimates the impact of explanatory variables on the cost-effectiveness measured by direct jobs generated in industry per unit of public outlays. Then using all observations, successful and unsuccessful, a measure of the probability of success in generating any jobs is determined. The measure of the probability of success and the cost-effectiveness coefficients for projects successful in generating at least some jobs are combined

*Research Assistant and Regents Professor, respectively. Professional paper P-313.

Proc. Okla. Acad. Sci. 57: 185-188 (1977)

to show the full impact of the variables on the cost-effectiveness of the EDA projects.

A number of economic development theories suggest variables to include in a model predicting the success of a job-creating development project in a region. Location theory suggests that industry moves to areas with the greatest profit potential. Although market incentives may be distorted in depressed areas needing assistance to expand the employment base, firms are expected to respond to profit potential as evident in availability of low-cost inputs, agglomeration economies, adequate transportation, and nearness to large markets. Central place theory, growth center theory, and growth pole theory are related to neoclassical theory but heavily stress agglomeration economies growing out of external and internal economies of scale, availability of business credit, and other supportive public and private services, especially as found in larger cities. Some variables measuring basic ingredients of economic growth — natural resources, institutions, and attitudes of people — are included in "reduced form" rather than in a form directly tied to one of the particular economic development theories mentioned. But all of the variables are considered to be relevant on theoretical grounds to explain cost-effectiveness.

All the theoretically admissible variables were selected and arranged in seven reasonably homogeneous groups for convenience in applying the ad hoc procedure to reduce the number of closely related variables included in the final model. The variables within each group, shown in Table 1, are closely related in theory and tend to be highly correlated with each other. The relative location group is made up of six variables which are the distances to cities of various sizes and to an interstate highway. The specific location group includes four different variables describing the population pattern and intensity of the city and county of each project. The employment group measures the availability and skills of labor in the region. The income group contains various measures of the income in the area involved in the EDA project. The institutional group is comprised of two variables designed to measure government involvement in the county. Two demographic variables, percent of population over 65 years and mean years education,

TABLE 1. *List of Variable by Groups*

Relative Location
L1 = Distance[a] in miles from project location to city of over 25,000
L2 = Distance in miles from project location to city of over 50,000
L3 = Distance in miles from project location to city of over 100,000
L4 = Distance in miles from project location to city of over 250,000
L5 = Distance in miles from project location to city of over 750,000
LIS = Distance in miles from project location to nearest interstate highway
Specific Location
P1 = Population of location of project (city)
P21 = Population per square mile of county (1967 Sales Management Survey of Buying Power)
P4 = Population per square mile of county (1960 census)
P5 = Percent urban (county—1960)
Employment
E11 = Percent nonagricultural employment in the county (1960)
E2 = Percent underemployment in the county (1960)
E3 = Percent unemployment in the county (1960)
E5 = Percent of the county labor force in manufacturing (1960)
E6 = Percent of the county labor force in white collar jobs (1960)
Income
Y1 = Household income (1967 Sales Management Survey of Buying Power)
Y21 = County per capita income (1960)
Y3 = County median income (1960)
Y4 = Percent in county with income less than $3,000 (1960)
Institutional
PAR1 = Percent public assistance recipients in county (1960)
G1 = Total local government expenditures per person in county (1960)
Demographic
AGE = Percent of county residents over 65 years of age (1960)
ED = Mean years of education in county (1960)
Dummy[b]
Development Regions, Variables D1, D2, D3, D4, D5, D6, D7*
EDA or Boise Cascade Evaluation, Variables D8, D9*
Industrial or Public Works Project, Variable D10, D11*
Census Regions, Variables D12, D13, D14, D15*

[a] Minimum highway mileage.
[b] Dummy variable with asterisks excluded in estimation procedure.

were considered jointly. The final group included four sets of dummy variables.

In applying the grouped-variable technique to reduce multicollinearity while preserving some structural validity of the theoretical model, a separate regression first was estimated for each of the six groups with cost-effectiveness the dependent variable and only the variables within each respective group as the independent variables. Two different methods were used to select the variable from each group. Model I was formed by regressing variables in a given group on cost-effectivenesss, then choosing the one independent variable having the largest t statistic on its coefficient. Model II was formed by choosing the variable from each group that accounted for the highest proportion of explained variance in the dependent variable in a stepwise procedure applying a maximum R^2 improvement technique. The variable from each group was included in the final model. Because the demographic group included only two variables and these seemed to pose no multicollinearity problems, both variables were included in the final OLS model containing all the groups of variables with JBDL (jobs created per thousand dollars) as the dependent variable and the representative variables from the respective groups as the independent variables.

In the first regression, only the observations from development projects where at least some jobs were generated were used to determine the impact of explanatory variables on the cost-effectivenesss — direct jobs generated in industry per unit of public outlays. Then using all observations, from successful and unsuccessful projects, a measure of the probability of success in generating any jobs was determined. The measure of the probability of success and the cost-effectiveness coefficients for proj-

TABLE 2. *A Comparison of the Two Models*

	Model I		Model II	
	Variable[a]	Coefficient (Significance Level)	Variable[b]	Coefficient (Significance Level)
A. *Successful Observations, Cost-Effectiveness*	ED	−.357 (.02)	ED	−.179 (.30)
	P5	+.014 (.03)	P4	−.001 (.15)
	L2	+.004 (.13)	L1	−.005 (.10)
	E11	−3.286 (.15)	E6	−.004 (.83)
	AGE	−.046 (.19)	AGE	−.074 (.05)
	Y21	+.00005 (.48)	Y21	+.0002 (.74)
	G1	+.001 (.52)	G1	+.002 (.27)
	D1	+.098 (.82)	D12	+.513 (.21)
B. *All Observations, Probability of Success*	D10	−.334 (.001)	D10	−.317 (.0003)
	E3	−.039 (.002)	E3	−.033 (.01)
	P1	.000002 (.01)	P21	−.0002 (.32)
	G1	−.002 (.02)	G1	−.0012 (.32)
	L3	+.009 (.15)	L4	−.0004 (.42)
	AGE	+.009 (.50)	AGE	−.001 (.93)
	Y1	+.00002 (.72)	Y3	−.00001 (.99)
	ED	−.012 (.79)	ED	+.006 (.91)

[a] Variables are listed in order of significance in Model I; for example, in part A, the most significant is ED (.02) and the least significant is D1 (.82).

[b] Variables in Model II are listed in such an order that the variables chosen from the same groups in the two models are listed together. See numbers in parentheses for significance levels.

ects successful in generating at least some jobs were combined to show the full impact of the variables on the cost-effectiveness of the EDA projects.

The results of the two methods used to select the variables were quite similar although in the regressions run with all the observations (Table 2, part A) the variables chosen were the same for only two of the six groups (not including the demographic group). Even though the actual variable chosen from a given group was not necessarily the same for Models I and II, it contributed much the same influence to the dependent variable. The four most significant variables came from the same groups in both models. The R^2 values were very similar. According to these results the variables having the most influence on cost-effectiveness were from the location group, the specific location group, and the demographic group. The regressions run on all the observations to get a measure of the probability of success in generating at least one job yielded almost identical results by both methods (Table 2, part B). The most significant variables were chosen from the same groups by each method.

When the above two regressions were combined to get the full impact on cost-effectiveness of the variables the results again were very similar. The variables that most influence the overall cost-effectiveness of EDA funds were L1, L2, ED and D10. These four variables suggest the location and type of a cost-effective job-creating project: projects close to cities of 25,000 and away from cities of 50,000 (L1 and L2), use of a more direct type assistance to industry (D10) in an area with low average schooling attainment (ED). The latter variable probably reflects advantages for industry that utilizes a blue-collar labor force, with relatively low wage rates. Such industry is not necessarily the most desirable for any given community, however.

The two statistical measures used to select the variable from each group yielded similar equations. The procedure of grouping the variables into homogeneous sets of variables and subsequently choosing one of the variables to represent the group reduced the multicollinearity by including only one of a group of closely related variables. The regression equations containing only one variable from each group had only slightly lower (5-10%) R^2 values than the equations that included all the variables. Little was lost except the multicollinearity.

SUMMARY

This paper offers a suggestion for the researcher who finds that the theory is too imprecise to limit the set of variables to a workable number. In the example presented, the variables were placed into groups in which the variables were strongly correlated, with only one variable in the group being selected. The inclusion of only one of a group of related variables should significantly reduce the multicollinearity in the final equation without undue loss of explanatory power.

REFERENCES

1. H. THEIL, *Principles of Econometrics,* J. Wiley & Sons, Inc., New York, New York, 1971.
2. U. S. DEPARTMENT OF COMMERCE, "Creating New and Permanent Employment: *Annual and Final Report of the ARA,"* Washington, D. C., December 1965.
3. U. S. DEPARTMENT OF COMMERCE, Economic Development Administration, *Public Works Program, An Evaluation,* Vol. II, Washington, D. C., 1970.