Title

Improving Accuracy in Multiple Regression Estimates of Population Using Principles from Causal Modelling

Journal

Demography, 17(4)

Author

Swanson, David A

Publication Date

1980

Peer reviewed

DEMOGRAPHY© Volume 17, Number 4 November 1980

IMPROVING ACCURACY IN MULTIPLE REGRESSION ESTIMATES OF POPULATION USING PRINCIPLES FROM CAUSAL MODELLING

David A. Swanson
Washington State Board for Community College Education, Olympia, Washington 98504 and Demographic Research Laboratory, Department of Sociology, Western Washington University, Bellingham, Washington 98225

Abstract—This paper reports a mildly restricted procedure for using a theoretical causal ordering and principles from path analysis to provide a basis for modifying regression coefficients in order to improve the estimation accuracy of the ratio-correlation method of population estimation. The modification is intended to take into account temporal changes in the structure of variable relationships, a major element in determining the accuracy of post-censal estimates. The modification of coefficients is conservative in that it uses rank-ordering as a basis of change. Empirical results are reported for counties in Washington state that demonstrate the increased accuracy obtained using the proposed procedure.

INTRODUCTION

Most procedures intended to improve regression estimation accuracy are based on modifications introduced during model construction. These procedures include variable selection, variable transformation, higher-order forms, data augmentation and ridge regression. Most of these types of procedures have been applied to the "ratio-correlation method," a regression technique introduced by Schmitt and Crosetti (1954) and now commonly used for making post-censal estimates of subnational populations. For example, variable selection is discussed by Goldberg, Rao and Namboodiri (1964) and Zitter and Shryock (1964); variable transformation by Schmitt and Grier (1966), O'Hare (1976) and Swanson (1978b); data augmentation by Rosenberg (1968), Pursell (1970), Namboodiri and Lalu (1971) and Martin and Serow (1978); and ridge regression by Swanson (1978a) and Spar and Martin (1979). One limiting feature common to these types of procedures is that they are only applied to the data used in constructing a given regression model (i.e., to the "model data set") and do not incorporate subsequent post-censal information, say, for example, from the predictor variables substituted into a given regression model in order to produce an actual post-censal estimate (i.e., information from the "estimation data set"). The issue of utilizing more fully information available subsequent to the construction of a given model is important. Studies by Namboodiri and Lalu (1971), Ericksen (1974), Namboodiri (1972), O'Hare (1976), Martin and Serow (1978), Swanson (1978a), Spar and Martin (1979), Tayman (1979) and Mandell (1980) have all implicated the temporal stability of model coefficients as a primary element in the accuracy of post-censal estimates.

One example of a procedure that does incorporate information subsequent to that available from the "model data set" is given by Ericksen (1974). This procedure, the "regression sample method," incorporates current sample data and symptomatic information in order to produce more accurate estimates of post-censal populations. However, this procedure has

not caught on, probably because of the increased complexity and data collection burden it requires.

In this paper, an alternative procedure for improving regression estimation accuracy is introduced which, like Ericksen's, utilizes the information external to the model data set used in constructing a regression model but, unlike his, is gained solely from the "estimation data set." The procedure is subject to two limitations: it can only be applied to situations in which reasonable assurance of the correct specification of a causal ordering of variables is obtained and in which each correlation between the predictor variables and the dependent variable can be assumed to be positive. While these two restrictions may be incapacitating in some estimation problems, they do not preclude using the procedure in the majority of problems relating to county, state or other local area population estimates using the symptomatic indicators typically found in conjunction with the ratio-correlation method (see, e.g., U.S. Bureau of the Census, 1973; 1976). For example, it is reasonable to assume that a model which specifies county population as causally prior to county voters, employment and school enrollment is a correctly postulated theoretical structure. It is physically impossible to have voters, employment and school enrollment without a population; on the other hand, while it may be unlikely, it is possible to have a county population without voters, or employed persons, or students.[1]

The second limitation, that of a positive correlation between the dependent variable and each of the predictor variables, while more constraining than the idea of a causally prior population variable, does not hinder the procedure for most states.[2] The U.S. Bureau of the Census (1976, pp. 70–74) reports only ten states in which the relationship between the dependent variable is negative for state-specific ratio-correlation models used for county population estimates. In these states (Colorado, Delaware, Maryland, Minnesota, Mississippi, Nevada, New Mexico, Rhode Island, Vermont, and West Virginia) the procedure could still be used by eliminating the variable or variables in question and re-constructing the model. It would, of course, be advisable to weigh the costs of variable reduction against the additional information gained by using the procedure.

THE RATIO-CORRELATION METHOD

The ratio-correlation method uses proportional numbers, which means that the county populations must sum to a state total population, which in an estimation year is determined independently of the regression-estimated county populations. The model is designed to estimate the temporal change in county population proportions using the observed temporal changes in the county proportions of symptomatic indicators such as school enrollment, voters, and the like. The temporal change is measured simply by taking a ratio of the proportions at two points in time for each variable—hence the name ratio-correlation. Since enumerated county populations for an entire state are found only for federal decennial census years, the model is always constructed for two points in time that are ten years apart. The data underlying this construction are termed the "model data-set." For example, in Washington state, a ratio-correlation model used to estimate annual county populations from 1971 to 1979 was constructed using the ratio of 1970 to 1960 proportions.

Once a model is constructed, the actual estimation is accomplished by algebraically manipulating the estimation of change in proportions into actual county population numbers. The estimated county population numbers are then adjusted to the independently determined total state population and, often, smoothed to provide orderly transitions from earlier estimates.

An example of this procedure for the year 1972 follows: first, 1972 over 1970 ratios of symptomatic indicator proportions

are substituted into the model that was constructed using ratios of 1970 over 1960 data. These ratios of proportions are the "independent" variables in the model. Next, the model is run and estimated ratios of 1972 over 1970 county population proportions are generated. Since the 1970 county populations are known, these estimates can be algebraically manipulated into estimated 1972 proportions. These, in turn, are then multiplied by the independently derived state total and adjusted and smoothed where necessary. A more formal description of the ratio-correlation model is given below.

$$Y = a_0 + a_1x_1 + a_2x_2 + , \cdots , + a_\kappa x_\kappa + E \tag{1}$$

Where

a_i = coefficient to be estimated
E = error term

In the ratio-correlation form:

$$Y = \frac{\text{(Population in County, time} = T_{o+x})}{\text{State Total Population}} \div \frac{\text{(Population in County, time} = T_o)}{\text{State Total Population}} = \frac{Yx}{Yo}$$

$$X_i = \frac{\text{Symptomatic (Indicator in County, time} = T_{o+x})}{\text{State Total Indicator}} \div \frac{\text{Symptomatic (Indicator in County, time} = T_o)}{\text{State Total Indicator}} = \frac{X_{ix}}{X_{io}}$$

Although equation (1) is readily understandable, it is more convenient to use matrix notation to represent it. (For a good introduction to the matrix approach to regression, see Draper and Smith, 1966.) Equation (1) in matrix form is denoted by

$$\mathbf{Y} = \mathbf{XB} + \boldsymbol{\varepsilon} \tag{2}$$

where the $n \times p$ matrix $\mathbf{X}$ contains the values of p predictor variables at each of the n data points. $\mathbf{Y}$ is the vector of values for the dependent variable, $\mathbf{B}$ is the $p \times 1$ vector of regression coefficients, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of stochastic errors, where $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, and $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\mathbf{I}_n$.

Further, it is also convenient to consider equation (2) in correlation form. That is, in a form where the variables are standardized by centering each observation on its mean and scaling it by dividing by the standard deviation of the variable in question. (See, again, Draper and Smith, 1966.)

In its correlation form, the estimate of B is given by

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{3}$$

with $(\mathbf{X}'\mathbf{X})$ being the zero-order correlation matrix for the independent variables, and $(\mathbf{X}'\mathbf{Y})$ the vector of zero-order correlations between each of the independent variables and the dependent variable.

Throughout this paper, the intent of the procedures described is to produce a modified $\hat{\mathbf{B}}$ vector, primarily by exploiting information for $(\mathbf{X}'\mathbf{X})_e$, where the subscript "e" identifies an "estimation data set." An established procedure for producing modified $\hat{\mathbf{B}}$ vectors is ridge regression (Hoerl and Kennard, 1970) which is used in the presence of multicollinearity in order to stabilize the estimate of the regression coefficient vector. However, ridge regression is designed to deal with instability caused by multicollinearity not instability caused by structural changes over time, although in some applications it may be difficult to distinguish between these two causes of instability. In any event, ridge regression and other procedures mentioned earlier do not exploit the information always available from the zero-order correlation matrix $(\mathbf{X}'\mathbf{X})_e$ for the predictor variables in an estimation data set. In these procedures, predictor values are simply plugged into the coefficients previously calculated from the

model data set and the estimated values for the dependent variable are generated. Ignoring the relationships found in the zero-order correlations among the predictor variables in the estimation data set disregards information that can be used to modify $\hat{\mathbf{B}}$ and improve the accuracy of estimated values.

CAUSAL MODELLING AND PATH ANALYSIS CONCEPTS

The key to exploiting the information contained in the zero-order correlations found in an estimation data set is taken from Land (1969, Chapter IV), work that is based on the fundamental theorem underlying path analysis as developed by Wright (1921). As stated in the Introduction and footnote 1, it involves a theoretical "reversal" of the dependent variable in the regression model, the population variable, as an unmeasured, causally prior variable and a "just-identified" structure—a minimum of three predictor variables (in the regression model), the covariance of which is assumed to be due to the fact that they are all causally related to the population variable. The path diagram specifies this theoretical "reversal."Let

Z_i = a variable in standard form (subtracted from its mean and divided by its standard deviation)

Z_a = the population variable

Z_1 = predictor variable 1

Z_2 = predictor variable 2

Z_3 = predictor variable 3

Z_b, Z_c, and Z_d = residual effects

r_{ij} = the zero-order correlation between variable i and variable j.

If the causal structure is adequately specified, then the following relations hold:

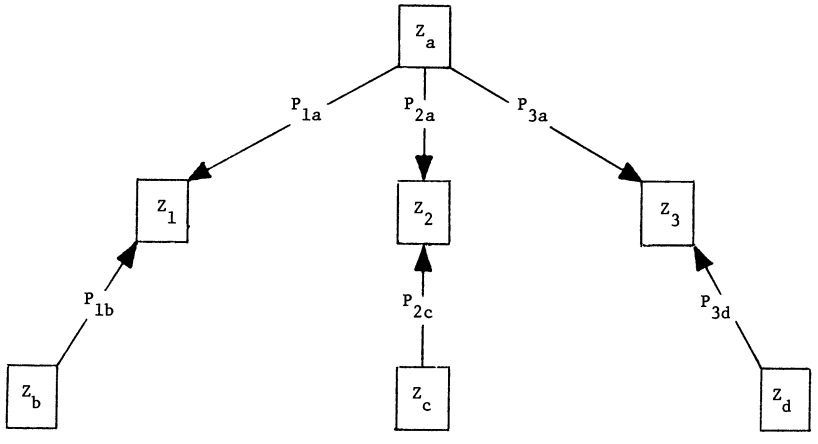
$$\begin{aligned} r_{12} &= (P_{1a})(P_{2a}) \\ r_{13} &= (P_{1a})(P_{3a}) \\ r_{23} &= (P_{2a})(P_{3a}) \end{aligned} \tag{4}$$

Further, from the above system, if P_{1a}, P_{2a}, and P_{3a} can be assumed to be positive, a system of three linear equations in three unknowns can be made using logarithmic transformation in order to solve for $\ln(P_{1a})$, $\ln(P_{2a})$ and $\ln(P_{3a})$.

$$\ln(r_{12}) = (1)\,|\ln(P_{1a})| + (1)\,|\ln(P_{2a})| + (0)\,|\ln(P_{3a})|$$

$$\ln(r_{13}) = (1)\,|\ln(P_{1a})| + (0)\,|\ln(P_{2a})| + (1)\,|\ln(P_{3a})|$$

$$\ln(r_{23}) = (0)\,|\ln(P_{1a})| + (1)\,|\ln(P_{2a})| + (1)\,|\ln(P_{3a})|$$



Once the unknowns are found they are transformed into estimates of P_{1a}, P_{2a}, and P_{3a}, which are also estimates of r_{1a}, r_{2a}, and r_{3a}, respectively since $P_{1a} = r_{1a}$, $P_{2a} = r_{2a}$,

and $P_{3a} = r_{3a}$. These estimates are the basis for modifying an original regression model's coefficients. The actual process of modification will be discussed in the subsequent section.

If more predictor variables are available, i.e., more than three, using the theoretical causal structure specified above leads to an over-identified model which in turn can be used to test the adequacy of the specified causal structure. This point is pursued by Land (1969) and Heise (1969).

At this point, one might think that if reasonable estimates of the correlations are available between the dependent variable and the predictor variables for an estimation data set, why not simply use these to re-estimate the original coefficient vector, $\hat{\mathbf{B}}$. This approach is usually blocked by two factors. The first is that the coefficient vector, $\hat{\mathbf{B}}$, is extremely sensitive to estimation errors in a given $(\mathbf{X}'\mathbf{Y})$ matrix. Even small error could lead to a significantly different coefficient vector which could produce less accurate estimates than an unmodified, original coefficient vector. The second factor is that in order to transform the modified estimated (standardized) coefficient vector into the unstandardized regression coefficients (and, consequently, to have, by definition an intercept in the regression equation) needed for an actual estimation of the dependent variable, an estimate of the variance of the dependent variable must be available. This variance is not available for an estimation data set, since by definition, an estimation data set excludes any observations of the dependent variable. Further, it does not appear to be stable enough to estimate its value from that contained in the model data set although this must be determined empirically for specific estimation problems.

THE RANK-ORDER PROCEDURE

At this point it is useful to provide some notation.

Let

$(\mathbf{X}'\mathbf{X})_m$ = the zero-order correlation matrix for the p predictor variables in the model data set, the subscript "m" refers to the model data set,

$\hat{\mathbf{B}}_m = (\mathbf{X}'\mathbf{X})_m^{-1}(\mathbf{X}'\mathbf{Y})_m$, the coefficient vector estimated from the model data set,

$(\mathbf{X}'\mathbf{X})_e$ = the zero-order correlation matrix for the p predictor variables in the estimation data set, which is always available in an estimation situation,

$\hat{\mathbf{B}}_e = (\mathbf{X}'\mathbf{X})_e^{-1}(\mathbf{X}'\mathbf{Y})_e$, the coefficient vector that would result if in an estimation data set the dependent variable could be observed. (In actual practice, $\hat{\mathbf{B}}_e$ is unknown because $(\mathbf{X}'\mathbf{Y})_e$ is unknown.) And

$(\mathbf{X}'\mathbf{Y})_e^*$ = the zero-order correlation vector for each of the p predictor variables with the (unknown) dependent variable that is *estimated* for the estimation data set using the rank-order procedure.

Further, let the p unique correlation coefficients appearing in $(\mathbf{X}'\mathbf{Y})_m$ and (the unknown) $(\mathbf{X}'\mathbf{Y})_e$ be ranked in descending order in $\mathbf{R}_m$ and $\mathbf{R}_e$, respectively. The correlation coefficients that are estimated using the rank-order procedure and contained in $(\mathbf{X}'\mathbf{Y})_e^*$ are placed in descending rank-order in $\mathbf{R}_e^*$. By either calculating to a high number of significant digits or using a procedure to decide among any ties within $\mathbf{R}_m$ and $\mathbf{R}_e^*$, respectively, there are p^2 pairwise comparisons of ranks that can be made for p predictor variables between $\mathbf{R}_m$ and $\mathbf{R}_e^*$ and a minimum of zero and a maximum of p re-orderings of the ranked coefficients in $\mathbf{R}_m$ such that their rank order becomes equivalent to their corresponding rank-order in $\mathbf{R}_e^*$. This re-ordering of ranks in going from $\mathbf{R}_m$ to $\mathbf{R}_e^*$ is the basis of the procedure. The key is in estimating $(\mathbf{X}'\mathbf{Y})_e^*$ and, consequently, placing these correlations in descending rank-order in $\mathbf{R}_e^*$. This key is found in

the causal structure outlined earlier. While, as it was stated earlier, it is usually not feasible to use $(\mathbf{X'X})_e$ in conjunction with the algebraic manipulations possible from the specified causal structure to directly estimate $(\mathbf{X'Y})_e$, and, consequently, $\hat{\mathbf{B}}_e$, it is reasonable to assume that the estimated correlations, $(\mathbf{X'Y})_e$* *are in the same rank-order* as those in $(\mathbf{X'Y})_e$. The reasonableness of this assumption is directly related to the reasonableness of the causal structure specified. In turn, the rank-orders in $\mathbf{R}_m$ and $\mathbf{R}_e$* are the same as the rank-orders in $\hat{\mathbf{B}}_m$ and $\hat{\mathbf{B}}_e$, respectively. Consequently, the change in rank-order, if any, observed in going from $\mathbf{R}_m$ to $\mathbf{R}_e$* can be used to modify $\hat{\mathbf{B}}_m$ such that it conforms to the (unobserved) rank-order in $\hat{\mathbf{B}}_e$. Further, since the variance of each of the variables in the model data set is known, $\hat{\mathbf{B}}_m$ (which is, remember, in standardized form), modified to conform to the rank-order found in $\mathbf{R}_e$*, can be manipulated into a modified set of unstandardized regression coefficients, which can be used to provide an alternative estimate of the dependent variable.

Although others are feasible, the modification of coefficients proposed here is intended to provide a conservative approach. The actual modification proposed is as follows. First, $\mathbf{R}_m$ and $\mathbf{R}_e$* are determined. Since $\mathbf{R}_m$ preserves the same rank-ordering found in $\hat{\mathbf{B}}_m$, the coefficients in $\hat{\mathbf{B}}_m$ can be incremented by selected values such that their rank-order is modified to conform to the rank-order in $\mathbf{R}_e$*. By using "minimum" increments, this procedure assumes even more of a conservative approach to modifying the coefficients in $\hat{\mathbf{B}}_m$.

The use of "minimum" increments has two major issues associated with it that require clarification. The most obvious one is the selection of a value for the increment. Should it be in tenths, hundredths, thousandths, or even more detailed? In an actual application the selection of a useful increment should be determined both by the magnitude of change required and by the judgment of the user. The other issue is how to apply the selected increment to a set of coefficients. A set of two hypothetical examples will help with the required clarification. In the first hypothetical example, suppose that there are only two predictor variables, X_1 and X_2, whose standardized regression coefficients in the original model are $B_{1a} = .52$ and $B_{2a} = .48$, respectively. Suppose further that using the subsequent information derived from the zero-order correlations in the estimation data set, $(\mathbf{X'X})_e$, B_{1a} is indicated to be less than B_{2a}. How does one now select—and apply—an increment of change to the original coefficients such that the modified B_{1a} is made to be less than B_{2a}? Obviously, by arbitrarily changing B_{1a} to .40 and leaving B_{2a} at .48 the desired new rank-ordering will be achieved. Such an arbitrary approach, however, could decrease the accuracy of an estimation because it ignores certain theoretical and empirical regularities of the ratio-correlation form of multiple regression.

These regularities can help provide a useful set of guidelines for conservatively changing the original coefficients. First, recall that in a multiple regression equation the sum of the standardized model coefficients is finite. Further, in the ratio-correlation form, this sum consistently approximates 1.00 regardless of the number of predictor variables. Also, under the limitations given for this entire procedure, each independent variable is positively correlated with the dependent variable. These regularities imply that a change in one coefficient should be balanced by a corresponding change in the opposite direction. In the hypothetical example given, the arbitrary reduction of B_{1a} from .52 to .40 results in a coefficient vector that is "too short" in terms of its expected sum. Where the original sum was .52 + .48 = 1.00, the sum of the modified set is only .40 + .48 = .88. In order to conform to its expected sum, the modified coefficient vector should, in this example, be equal to 1.00. This can be accomplished by matching each decrease in B_{1a} with a corresponding increase in B_{2a}.

At this point one question still remains

unanswered. What value should be selected as an increment for modifying the set of coefficients? If .1 is used, at the first step B_{1a} will be reduced to .52 − .10 = .42 and B_{2a} will be correspondingly increased from .48 by .10 to .58. If .00001 is used, the point where the desired rank-ordering is achieved is where B_{1a} = .49999 and B_{2a} = .50001. Here, a useful guideline to follow in order to achieve a meaningful yet conservative change is to define the desired point of reversal in terms of empirical referents. Typically, this will probably be in terms of hundredths for many ratio-correlation models but some degree of experimentation is advised for any given application. When the number of coefficients exceeds two (which will usually be the case in actual practice), additional consideration must be given to the manner in which the sum of the original coefficient vector is preserved. For example, in a three coefficient system, there are exactly 3! = 6 possible outcomes for the new rank-ordering. In one of these outcomes, the same rank-ordering found in the original set is preserved; consequently, no modifications are required. In three of these possible outcomes, only two of the coefficients change ranks; consequently, the same procedure for making corresponding changes that was outlined in the hypothetical example for two coefficients can be used. That is, since the only possible changes are in (1) B_{1a} and B_{2a} or in (2) B_{1a} and B_{3a} or in (3) B_{2a} and B_{3a}, the balance necessary to preserve the original coefficient sum can be maintained by making changes in only the two coefficients in question; the third coefficient does not require any change. However, in the two remaining of the six possible outcomes, where all three coefficients change ranks, the balance must be maintained by making changes in all three coefficients.

This, however, is still not a very complicated matter. In both of these two outcomes, one coefficient changes rank by going from either the highest to the lowest rank or from the lowest to the highest. The remaining two coefficients each change one rank in the direction opposite to the change in the other one. This implies that the increment (or decrement) used for the single coefficient undergoing the maximum change of rank can be balanced by splitting its value equally between the other two and moving them in the opposite direction. For example, assume that the original coefficients are B_{1a}

Table 1.A.—Correlations between Variables in the "Model" Data Set

	$(X'X)_m$			$(X'Y)_m$
	X_1	X_2	X_3	Civ. Pop. < 65 Yrs. X_a
X_1 (employment)	1.000	.69379	.66318	.73138
X_2 (voters)	.69379	1.000	.81169	.93228
X_3 (enrollment)	.66318	.81169	1.000	.91942

= .5, B_{2a} = .3 and B_{3a} = .2, and that these need to be changed so that B_{3a} becomes the largest, B_{1a} the second largest, and B_{2a} the smallest. In the first step, B_{3a} is increased by .1 to .30; the corresponding changes in B_{1a} and B_{2a} are .50 − .05 = .45 and .30 − .05 = .25, respectively. In the second step, B_{3a} is again incremented by .1 from .3 to .4; B_{1a} and B_{2a} are decremented by .05 to .40 and .20, respectively. In the final step, B_{3a} is incremented by .1 to .5 while B_{1a} and B_{2a} are each decremented by .05 to .35 and .15, respectively. This last step gives the coefficient values required to conform to the desired rank-ordering using the "minimum" increment, conservative procedure. As the number of coefficients increases, the problem becomes more complicated but it is still manageable since it reduces to a series of counting rules. In actual practice, the number of coefficients is usually under five and consequently it is possible to solve the majority of outcomes by hand—although this may become tedious. For those situations involving a number of coefficients a computer algorithm can be easily developed from basic counting rules and the allocations that were demonstrated in the two preceding examples.

It is important to bear in mind that the procedure outlined here is not the only one that could be used as a basis for coefficient modification. It is suggested because, in the absence of additional testing, it is likely to produce more accurate estimates than an unmodified model, especially if a great deal of change has occurred in the structure of the relationships for the variables in question; no claim is made for it producing an optimally accurate estimate.

EMPIRICAL RESULTS

Table 1.A gives the zero-order correlations relating to a 1960–1950 based ratio-correlation model for estimating county civilian population under sixty-five years from employment, voters, and grades 1-8 enrollment for the state of Washington.[3] Characteristics of the model constructed from these data are given in Table 1.B. In Tables 2.A and 2.B similar results are given for the 1970–1960 period. This set

Table 1.B.—Model Results

	$\hat{B}_m$ Standardized Regression Coefficients	Unstandardized Regression Coefficients
X_1	.07533	.066786
X_2	.51085	.550365
X_3	.45481	.356083
Constant	------	.046618

Multiple Correlation Coefficient, R = .97443
R^2 = .94953
Adjusted R^2 = .94519
S.E.E. = .05022

Table 2.A.—Correlations between Variables in the "Estimation" Data Set

	$(X'X)_e$			$(X'Y)_e$
	X_1	X_2	X_3	X_a
X_1 (employment)	1.000	.48155	.61864	.65478
X_2 (voters)	.48155	1.000	.73827	.83726
X_3 (enrollment)	.61864	.73827	1.000	.88966

forms the estimation data over which the procedure will be used and tested.

Although full knowledge of the estimation data set is available, the procedure is used as if this were not the case and the known results in Table 2 are shown for purposes of comparison. Of course, what is known in Table 2—and for any estimation problem—is $(\mathbf{X'X})_e$ which is used in conjunction with the theorem to estimate the rank-order of the (unknown) coefficients in R_e and, consequently, provide the re-ordering to which $\hat{\mathbf{B}}_m$ will be modified for conformation. Observe that in Table 1.A, the correlations in $(\mathbf{X'Y})_m$ are ranked:

Table 2.B.—Model Results

	$\hat{B}_e$ Standardized Regression Coefficients	Unstandardized Regression Coefficients
X_1	.15351	.081743
X_2	.38825	.473637
X_3	.50805	.492240
Constant	------	-0.559175

Multiple Correlation Coefficient, R = .93680
R^2 = .87759
Adjusted R^2 = .86709
S.E.E. = .05077

R_m	Rank
r_{2a}	(1)
r_{3a}	(2)
r_{1a}	(3)

Let the causal structure (as in the path diagram) be specified for the model data set (Table 1.A) and estimated data set. Then, for the model data set, the system of equations to be solved would be:

$$\ln(.69379) = (1)|\ln(P_{1a})| + (1)|\ln(P_{2a})| + 0$$

$$\ln(.66318) = (1)|\ln(P_{1a})| + 0 + (1)|\ln(P_{3a})|$$

$$\ln(.81169) = 0 + (1)|\ln(P_{2a})| + (1)|\ln(P_{3a})|$$

Solving the preceding system leads to:

$$\hat{P}_{1a} = \hat{r}_{1a} = .75290$$

$$\hat{P}_{2a} = \hat{r}_{2a} = .92146$$

$$\hat{P}_{3a} = \hat{r}_{3a} = .88082$$

These values correspond very closely to the correlations actually computed for the model data set which are $r_{1a} = .7314$, $r_{2a} = .9323$, and $r_{3a} = .9194$. The causal structure specified appears to be reasonable for the model data set. In an actual application, the preceding comparison is recommended. If the estimated correlations are at least not in the same rank-order as the calculated correlations then the procedure may not be adequate since the causal structure specified is not adequate.

Using the same causal structure and system of equations given above for the estimation data set leads to the following estimate, i.e., $(\mathbf{X'Y})_e^* =$

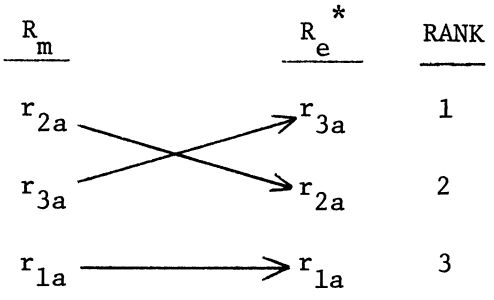
$$\hat{P}_{1a} = \hat{r}_{1a} = .635210$$

$$\hat{P}_{2a} = \hat{r}_{2a} = .758055$$

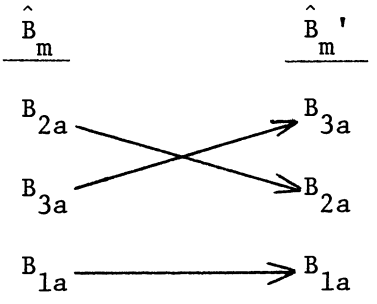
$$\hat{P}_{3a} = \hat{r}_{3a} = .973848$$

Notice that these estimated values of the true (and, remember, in actual practice, unobservable) correlations are less accurate than those estimated for the model data set. This decline in accuracy is related to the changing structure of relationships for these variables. In spite of the decline in accuracy, $(\mathbf{X'Y})_e^*$ still preserves the correct rank-ordering in $(\mathbf{X'Y})_e$.

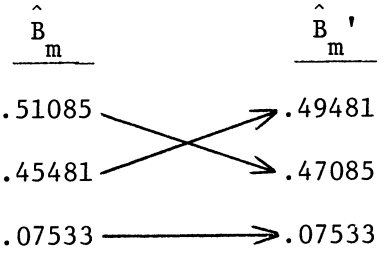
Ordering the correlations in $(\mathbf{X'Y})_m$ and $(\mathbf{X'Y})_e^*$ by descending rank in $\mathbf{R}_m$ and $\mathbf{R}_e^*$, respectively, leads to the following "map" of indicated changes:



Using the changes in rank-order observed in going from $\mathbf{R}_m$ to $\mathbf{R}_e^*$, $\hat{\mathbf{B}}_m$ can be modified in the following manner in order to provide the type of change described earlier,



The actual values are:



Note that here the increment selected was .01, as was suggested earlier. Some experimentation may be useful in other applications.

With these modified estimates of the standardized regression coefficients and the standard deviations of the variables in

the model data set, which are .2145, .2420, .1991, and .2740 for X_a, X_1, X_2, and X_3, respectively, the modified unstandardized coefficients can be estimated using the relationship $\hat{\mathbf{b}}'_{mi} = \hat{\mathbf{B}}'_{mi}(s_{ma}/s_{mi})$. The modified model in raw form becomes, then

$$\hat{Y} = .046618 + .066787X_1 + .50727X_2 + .38736X_3.$$

Estimates for 1970 of the county civilian population under sixty-five years of age (adjusted to the independently estimated state total) resulting from the preceding modified model and its unmodified version are presented in Table 3 along with the actual enumerated populations. Examination of the individual estimates produced by the original and modified versions of the model reveal that the modified model provides a higher degree of accuracy. Summary statistics bearing out this impression are given in Table 4.

The mean of the Absolute Percent Error is noticeably lower for the estimates provided by the modified version. The Index of Misallocation (IM) (see Palit et al., 1974 or Swanson, 1980) is a measure of total allocation accuracy relative to the entire state. Unlike the mean of the absolute percentage error—which is referenced to the individual counties—IM is referenced to the entire state. It is interpreted as the percentage of the entire state total that would have to be re-allocated in order to have the estimated county populations conform to the actual county populations. Although the difference may not seem great, 1.560 percent must be re-allocated to achieve zero mis-allocation using the original model while 1.493 percent must be re-allocated using the modified version, this level of improvement can mean a greal deal when, as is often the case, these estimates are used to allocate funds.

DISCUSSION

Perhaps the first point that deserves attention is the similarity of the unmodified estimates to the modified estimates. The ratio-correlation model for the empirical period under consideration already achieves a relatively high degree of accuracy. However, because data are only available at ten-year intervals for evaluating the estimates produced by different models, the actual estimation accuracy of a given model over virtually 90 percent of its use is unknown. While the relative gain in accuracy achieved using the modified version may seem modest, it is nevertheless an observable gain.

If a procedure consistently produces more accurate estimates over the entire inter-censal period, relatively slight gains in accuracy tend to accumulate if for no other reason than that later estimates tend to be smoothed or adjusted to earlier estimates. Furthermore, even slight gains in accuracy are worth pursuing because population estimates produced by many state demographic centers and the U.S. Bureau of the Census are used as a basis for allocating funds (Engels, 1978; Rosenberg and Myers, 1977; and U.S. Bureau of the Census, 1973).

Recall that a number of studies were cited above in which the temporal stability of model coefficients was found to be a major factor in the accuracy of post-censal estimates using the ratio-correlation method. Examination of the regression coefficients in Table 1.B and the corresponding ones in Table 2.B provides an example of change in the structure of relationships for Washington state. The modified model provides a tractable means for attempting to cope with this type of instability. Further, recall also that it was earlier suggested that perhaps one reason why Ericksen's (1974) "regression sample method" has not gained general acceptance is because of the difficulties involving coordination—and collection—of the data sets it requires. Compared with Ericksen's procedure, one advantage of the one presented here is that it does not impose any additional data collection and data coordination burden on a user. The procedure relies only on information that is contained

wholly within the data necessary for a ratio-correlation procedure.

An alternative method that could be used to deal with the instability problem is ridge regression (see, e.g., Hoerl and Kennard, 1970), although this procedure is generally thought of as a means to cope with the instability of coefficients due to the large mean square errors associated with multicollinearity. Nonetheless, ridge regression provides a means of coefficient modification and it deserves some attention. A ridge procedure was conducted using the same data underlying the model

Table 3.—Estimation Accuracy of the Two Models' Civilian Population under Sixty-Five Years by County, State of Washington, 1970

County	Enumerated Population	Estimated Population		Percentage Difference	
		Unmodified	Modified	Unmodified	Modified
Adams	11,102	11,387	11,458	2.570	3.207
Asotin	11,862	11,819	11,814	-0.360	-0.408
Benton	63,144	67,823	67,511	7.411	6.916
Chelan	35,862	36,373	36,177	1.426	0.879
Clallam	30,023	31,371	31,294	4.489	4.232
Clark	116,663	111,312	111,437	-4.587	-4.480
Columbia	3,771	4,222	4,161	11.953	10.340
Cowlitz	62,586	61,636	61,581	-1.518	-1.606
Douglas	15,287	16,375	16,252	7.116	6.313
Ferry	3,336	3,408	3,397	2.152	1.825
Franklin	23,983	24,770	24,631	3.281	2.700
Garfield	2,546	2,770	2,761	8.814	8.435
Grant	38,921	42,750	42,606	9.839	9.469
Grays Harbor	52,583	52,173	52,114	-0.779	-0.891
Island	20,589	22,215	22,148	7.897	7.572
Jefferson	9,235	9,551	9,473	3.423	2.579
King	1,054,271	1,035,704	1,037,937	-1.761	-1.549
Kitsap	86,529	85,989	85,821	-0.625	-0.818
Kittitas	22,764	19,972	19,863	-12.266	-12.744
Klickitat	10,729	11,968	11,923	11.552	11.132
Lewis	39,265	40,124	40,122	2.187	2.183
Lincoln	8,168	9,185	9,107	12.452	11.494
Mason	18,411	17,867	17,827	-2.956	-3.172
Okanogan	22,952	23,656	23,591	3.068	2.786
Pacific	13,310	12,834	12,795	-3.580	-3.872
Pend Oreille	5,185	5,919	5,893	14.162	13.648
Pierce	339,048	346,430	346,728	2.177	2.265
San Juan	3,089	2,947	2,918	-4.603	-5.544
Skagit	45,703	48,868	48,758	6.924	6.683
Skamania	5,330	5,360	5,358	0.554	0.527
Snohomish	245,193	231,238	231,996	-5.691	-5.382
Spokane	251,057	256,882	256,723	2.320	2.257
Stevens	15,178	15,814	15,780	4.189	3.969
Thurston	68,719	69,613	69,540	1.301	1.194
Wahkiakum	3,137	3,409	3,397	8.677	8.293
Walla Walla	36,608	38,323	38,271	4.686	4.543
Whatcom	72,111	70,801	70,670	-1.817	-1.998
Whitman	34,843	32,510	32,409	-6.696	-6.984
Yakima	128,960	136,455	136,283	5.812	5.679

Table 4.—Summary Indices of Estimation Accuracy

Index	Model	
	Unmodified	Modified
Misallocation (IM)	1.560	1.493
Absolute Percent Error		
Mean	5.068	4.886
s.d.	3.813	3.639
N $\geq$ 10.00	5	5
N $\leq$ 3.00	15	16

data set used earlier. The optimally accurate ridge model was found where the bias level was equal to $k = .10$ (using increments of .01 from 0.00 to .10 and increments of .10 from .10 to .90). The optimally accurate model produced estimates for 1970 with absolute percentage error that exceeded 10 percent seven times (out of thirty-nine) and had a mean and standard deviation of 5.001 and 3.754, respectively. The Index of Misallocation of this model was 1.519 percent. The model clearly produced less accurate results than the model using the rank-order procedure, although it did provide more accurate estimates than did the original. Additional research comparing the rank-order procedure and ridge regression may be useful within the ratio-correlation context, especially where the effects of instability due to temporal change can be distinguished from those due to multicollinearity.

Even in a situation where the entire rank-order procedure is not used, it can still provide some valuable information. For example, it may show that there has been no change in the rank-ordering of a model's coefficients over time. This gives some indication that the temporal instability issue may not be a major source of estimation error and that a given regression model can be used with some degree of confidence. The rank-order procedure may be of more importance in an area that has experienced rapid population change and for which it is obvious that an original model can not be expected to produce the most accurate estimates possible. A trade-off that must be evaluated for each situation is the one mentioned in the first section: the cost of reducing the number of independent variables, in order to conform to the limitations described earlier, against utilizing information about temporal change. While this may not be a problem for many users it is a potential source of difficulty that must be mentioned.

Less conservative approaches than that used in the "minimum" increments rank-order procedure introduced here may be possible but, as discussed earlier, one major hurdle to be surmounted is controlling for the sensitivity of $\hat{\mathbf{B}}_e^*$ to errors in $(\mathbf{X}'\mathbf{Y})_e^*$. It may be of interest to explore this issue in relation to producing optimally accurate estimates within the con-

fines of the approach to coefficient modification described here. The rank-order procedure described in this paper, remember, is only intended to provide estimates that are more accurate than those available from an original model for which it is likely that structural changes have taken place over time. It does not necessarily provide an optimally accurate set of estimates available within the context of the information that can be gained from an estimation data set using the theoretical causal specification and path-analysis algebra exploited here.

NOTES

[1] Throughout the paper it is important to realize that while the proposed procedure relies upon the *conception* that a population must be causally prior to its symptomatic indicators, this is a theoretical construct that is used to exploit algebraic relationships in order to modify model coefficients. That is, this conception is a theoretical reversal that relies upon being able to assume that the dependent variable in the regression model is causally prior to its symptomatic indicators in a theoretical causal structure. In the actual estimation, the intent is still to produce a population estimate using the symptomatic indicators as a set of predictor variables.

[2] This second restriction is like the first in that it is necessary in order to exploit algebraic relationships, in this case, logarithms. If values are negative, then the logarithms required cannot be calculated.

[3] These data are in the official Washington State Population Data Base (File Base 526), available on request.

ACKNOWLEDGMENTS

The author is grateful for comments made by Jeff Tayman and anonymous reviewers.

REFERENCES

Draper, N. R. and H. Smith. 1966. Applied Regression Analysis. New York: Wiley.

Engels, R. A. 1978. Local Area Population Research and Federal Programs. Review of Public Data Use 6:12–21.

Ericksen, E. P. 1974. A Regression Method for Estimating Population Changes of Local Areas. Journal of the American Statistical Association 69:867–875.

Goldberg, D., V. R. Rao, and N. K. Namboodiri. 1964. A Test of the Accuracy of Ration-Correlation Population Estimates. Land Economics 40:100–102.

Heise, D. R. 1969. Problems in Path Analysis and Causal Inference. In E. F. Borgatta (ed.), Sociological Methodology 1969. San Francisco: Jossey-Bass.

Hoerl, A. E. and R. W. Kennard. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 12:55–67.

Land, K. C. 1969. Explorations in Mathematical Sociology. Unpublished Ph.D. dissertation. University of Texas, Austin.

Mandell, M. P. 1980. Further Comparisons of the Ratio-Correlation and Difference-Correlation Methods of Population Estimation. Unpublished M.S. thesis. Florida State University.

Martin, J. H. and W. J. Serow. 1978. Estimating Demographic Characteristics Using the Ratio-Correlation Method. Demography 15:223–234.

Namboodiri, N. K. 1972. On the Ratio-Correlation and Related Methods of Subnational Population Estimation. Demography 9:443–453.

——— and N. M. Lalu. 1971. The Average of Several Regression Estimates as an Alternative to the Multiple Regression Estimate in Postcensal and Intercensal Estimates: A Case Study. Rural Sociology 36:187–194.

O'Hare, W. 1976. Report on a Multiple Regression Method for Making Population Estimates. Demography 13:369-380.

Palit, C. D. et al. 1974. Procedures for Estimating M.C.D. Populations for State Revenue Sharing. Proceedings of the American Statistical Association, Social Statistics Section:396–399.

Pursell, D. 1970. Improving Population Estimates with the Use of Dummy Variables. Demography 7:87–92.

Rosenberg, H. 1968. Improving Current Population Estimates Through Stratification. Land Economics 44:331–338.

——— and G. C. Myers. 1977. State Demographic Centers: Their Current Status. The American Statistician 31:141–146.

Schmitt, R. C. and A. H. Crosetti. 1954. Accuracy of the Ratio-Correlation Method for Estimating Postcensal Population. Land Economics 30:279–281.

——— and J. M. Grier. 1966. A Method of Estimating the Population of Minor Civil Divisions. Rural Sociology 31:355–361.

Spar, M. A. and J. H. Martin. 1979. Refinements to Regression-Based Estimates of Postcensal Population Characteristics. Review of Public Data Use 7:16–22.

Swanson, D. A. 1978a. Preliminary Results of an Evaluation of the Utility of Ridge Regression for Making County Population Estimates. Paper presented at the Annual Meeting of the Pacific Sociological Association, Spokane, Washington.

———. 1978b. An Evaluation of "Ratio" and "Difference" Regression Methods for Estimating Small, Highly Concentrated Populations: The Case of Ethnic Groups. Review of Public Data Use 6:18–27.

———. 1980. Allocation Accuracy in Population Estimates: An Overlooked Criterion with Fiscal Implications. Paper prepared for presentation at the Conference of the American Statistical Association, Committee on Small Area Statistics, Houston, Texas.

Tayman, J. 1979. Confidence Intervals for Postcensal Estimates of State Populations: A Regression Approach Using Time-Series Data on Age-Specific Death Rates. Unpublished Ph.D. dissertation. Florida State University.

U.S. Bureau of the Census. 1973. Federal-State Cooperative Program for Local Population Estimates: Test Results–April 1, 1970. Current Population Reports P-26, No. 21 (April).

———. 1976. Estimates of the Population of Counties: July 1, 1973 and 1974. Current Population Reports P-25, No. 620 (February).

Wright, S. 1921. Correlation and Causation. Journal of Agricultural Research 20:557–585.

Zitter, M. and H. S. Shryock. 1964. Accuracy of Methods of Preparing Postcensal Population Estimates for Local Areas. Demography 1:227–241.