Title
Improving the Measurement of Temporal Change in Regression Models Used For County Population Estimates

Journal
Demography, 21(3)

Authors
Swanson, David A
Tedrow, Lucky

Publication Date
1984

Peer reviewed

DEMOGRAPHY© Volume 21, Number 3 August 1984

IMPROVING THE MEASUREMENT OF TEMPORAL CHANGE IN REGRESSION MODELS USED FOR COUNTY POPULATION ESTIMATES

David A. Swanson
Economics and Statistics Department, Pacific Gas and Electric Company, San Francisco, California 94106
Lucky M. Tedrow
Demographic Research Laboratory and Department of Sociology, Western Washington University, Bellingham, Washington 98225

Abstract—The ratio-correlation method of population estimation is shown to contain an inconsistent temporal relationship between the model's empirical structure and its actual application. A simple transformation of the model's variables is provided that eliminates the inconsistency. Two tests of the relative accuracy of the original and transformed models show that the transformed model achieves accuracy levels equal to or higher than the original. In one test, all nine years show a higher degree of accuracy, of which four are statistically significant. Several possible reasons are given for the increased accuracy shown by the transformed model. The transformation, termed the "rate-correlation" model, is recommended as a logical starting point in the examination of coefficient stability and spatial autocorrelation as well as a method for estimating small populations.

The accurate estimation of county populations within a given state is important for many reasons, not the least of which is the role an annual set of estimates plays in allocating general revenue sharing and other funds to county and subcounty governmental units. These estimates are commonly made using the ratio–correlation form of multiple regression (Namboodiri, 1972; O'Hare, 1976; Schmitt and Crosetti; 1954; Serow and Martin, 1977; U.S. Bureau of the Census, 1973); although the difference-correlation form of multiple regression (O'Hare, 1976; Schmitt and Grier, 1966; and Swanson, 1978, 1980, 1981) and other procedures and combinations of procedures may be used (Bogue, 1950; Namboodiri and Lalu, 1971; U.S. Bureau of the Census, 1973; Zitter and Shryock, 1964). Although the U.S. Bureau of the Census uses the ratio-correlation method in its program of population estimating for federal revenue sharing, more exclusive reliance upon ratio-correlation is found among individual states, many of which use it to generate population estimates for state-funded revenue sharing programs.[1]

This report addresses a logical inconsistency found in both the ratio-correlation and difference-correlation forms and, further, suggests a solution to this inconsistency. For the sake of brevity, the report focuses exclusively on the ratio-correlation form since it is the most commonly used regression method for estimating county populations. However, it is understood that remarks concerning the ratio-correlation form can be generalized, taking into account certain adaptations, to include the difference-correlation and, possibly, other forms of regression used for estimating annual postcensal (or intercensal) county populations.

In the report that follows a description of the ratio-correlation and its logical inconsistency is first given. Following this description are a proposed transformation of the ratio-correlation form that is designed to eliminate the logical inconsistency, an empirical evaluation of a series of estimates produced by the original and transformed regression forms, and a discussion and summary of the results.

RATIO-CORRELATION AND ITS LOGICAL INCONSISTENCY

The ratio-correlation method is designed to estimate the temporal change in county population proportions using (observed) temporal changes in county proportions of symptomatic indicators such as automobile registration, employment, registered 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 population numbers for all counties in a given state are found only in federal decennial census years, models are always constructed using two points in time that are ten years apart. The ratio-correlation model as constructed using two successive decennial census dates, is formally described as:

$$Y_i = a_0 + \Sigma_{ai}X_i + \varepsilon$$

where

a_i = coefficient to be estimated
ε = error term

and

$$Y_i = \left[\frac{P_{i,}\ t + 10}{\Sigma P_{i,}\ t + 10}\right] \Big/ \left[\frac{P_{i,}\ t}{\Sigma P_{i,}\ t}\right]$$

$$X_i = \left[\frac{S_{i,}\ \mathrm{t} + 10}{\Sigma S_{i,}\ t + 10}\right] \Big/ \left[\frac{S_{i,}\ t}{\Sigma S_{i,}\ t}\right]$$

where

p = population
s = symptomatic indicator
i = county
t = time

Once a model is constructed, the actual estimation is accomplished by algebraically manipulating the model's estimation of change in proportion into actual county population numbers. The estimated county population numbers are then "adjusted" to an independently determined total state population.

The ratio-correlation method offers several advantages over other potential approaches. The most important is probably the use of a measure of proportional change. This characteristic is useful because it controls for differential county population sizes (typical in many states) and yet allows the algebraic manipulation of the model's estimates into an actual population number. This is, indeed, a powerful feature of this technique. A second advantage is that ratio-correlation, being a regression model, possesses characteristics such as R^2 and S.E.E. that allow one to make an objective selection of a given model. Many other techniques of population estimation are not based in statistical theory and, consequently, demand a higher level of subjectivity in terms of evaluation and selection. A third advantage is its widespread use. Many state demographic centers and the U.S. Bureau of the Census have used it for many years, often in conjunction with the other methods, and consistently have had satisfactory results.

One feature of this method that is inconsistent with its implicit logic is, however, found in its measure of temporal change. The inconsistency is not that a *ratio* is used as a measure of temporal change; it is that temporal change is *inconsistently* measured in moving from model construction to actual use of the model for population estimation. In the construction of a model, a ratio is taken that refers to a ten-year period of change—the ten years between census dates. However, in an actual estimation, the ratios of change in the independent variables always refer to a number of years that is less than the ten-year period referenced in model construction. Consider a 1972 set of estimates. The model coefficients are referenced to a ten-year period of change *but* the ratios of 1972 to 1970 proportions for the symptomatic indicators are referenced to a two-year period of change. In this case, the model, referenced to a ten-year period of change, is now expected to produce an estimate using data that are referenced to a two-year period of change. This is a logical inconsistency that exists independently of any considerations regarding the applicability of observed change from 1960 to 1970 to post-1970 events. Why should the model be expected to perform optimally when it is applied to periods that are referenced to less than ten years? This is important because the temporal stability of coefficients has been identified as a critical issue in the accuracy of estimates produced using the ratio-correlation model (Mandell and Tayman, 1982; O'Hare, 1980; Swanson, 1980). Although the temporal stability of coefficients is a much broader issue than that addressed in this paper, one aspect pertaining to it is certainly the logic of applying a model referenced to a ten-year period of change to a period less than ten years. This situation may not be crippling in actual practice but, nonetheless, the issue exists and deserves both a discussion of its potential effects and a remedy.

RESOLVING RATIO-CORRELATION'S INCONSISTENCY

A solution that comes immediately to mind is to construct a model that is referenced to an annual *rate of change* rather than a decennial *amount* of change. Such a model would be logically consistent with any estimation because an annual rate of change can be applied to find the change expected over any number of years. This solution can be accomplished using one of several approaches to measuring rates of change: linear, geometric, or exponential (Shryock and Siegel, 1973; pp. 377–380). Constructing such a model would require additional transformations of the variables in order to estimate an annual rate of change, but in return, a model so constructed could be appropriately referenced to each of the postcensal (or intercensal) years for which an estimate is desired. A measure of an annual rate of change that is a natural extension of the ratio-correlation logic is based on an exponential approach to change, which can be calculated by taking the natural logarithm of the "ratio" and dividing the result by the total number of years between two successive census dates. In addition to being a natural extension of the ratio-correlation logic, this approach may also provide a distinct benefit over others: Malec and Katzoff (1983) provide evidence that a logarithmic transformation reduces spatial auto-correlation.

Formally described, the modified ratio-correlation model as constructed using the notation described earlier is:

$$Y_i^* = a_0 + \Sigma_{ai} X_i^*$$

where

$$Y_i^* = \frac{\ln(Y_i)}{10}$$

$$X_i^* = \frac{\ln(X_i)}{10}$$

When this model is applied the actual estimated product will be the rate of annual *change in the county population proportions, which can be easily manipulated into the appropriate expected amount of change over the period in question.* This amount of change simply brings one back to the same result found in a "ratio-correlation" procedure, which can be algebraically manipulated into a current proportion and, given an independently determined total state population, the estimated population in each county.

Although this simple transformation resolves the logical inconsistency in using the ratio-correlation method, its utility must also be judged on other criteria. The most important criterion is the relative gain in accuracy obtained when the ratio-correlation method is extended to become an "average annual rate of change" method, or as referred to in shorthand form for the remainder of this paper, the "rate-correlation" method. The issue of gains in accuracy is, however, difficult if not impossible to answer because the noncensus years for which this transformation is designed are the very years for which complete census data are not available. Consequently, in this paper a surrogate for population—school enrollment—is used in one evaluation to examine the relative accuracy of both the ratio-correlation and rate-correlation models. In the final analysis, such an evaluation cannot be regarded as complete. However, neither can it be said such an evaluation is less satisfactory than others that have been conducted using population data measured at decennial census dates. On the one hand, the use of enrollment data as the dependent variable allows a look at the estimation accuracy of a model for a set of noncensus years that could not be assessed by using population as the dependent variable; on the other hand, population figures are the desired, ultimate use of these models, but the only time for which these data are available is once every ten years.

EMPIRICAL RESULTS

In this section two distinct evaluations are conducted. The first involves using county enrollments in grades one to eight as the object of estimation, while the second involves actual county populations.

Enrollment Models

Using Washington state's county enrollment in grades 1–8, estimation models are constructed over 1970 and 1960 using three symptomatic indicators: employment (covered by unemployment insurance); registered voters; and registered private automobiles.[2] The ratio-correlation model and the rate-correlation model constructed using the same 1970–1960 data are presented in Table 1.

These two models, as would be expected, exhibit similar characteristics but the ratio-correlation model appears to give a better fit than does the rate-correlation model. This is also consistent with expectations since the rate-correlation model attempts to use the *amount* of observed decennial change in order to estimate a *rate* of change.

Using these two models, estimates of enrollment were generated for each year from 1971 to 1979 and their respective levels of accuracy were judged relative to the actual enrollments reported for these years. The index of accuracy used is the mean of the absolute percentage error. Following the work of Mandell and Tayman (1982), paired observation *T*-tests were used to determine whether significant differences exist between the mean errors of the two models. The results, shown in Table 2, reveal that the rate-correlation model exhibits a lower level of error for each year, and in addition, in four of these years, 1976 through 1979, the difference is statistically significant ($\alpha = .05$).

Both models exhibit a similar pattern in the trend of errors: the degree of accuracy declines as the number of years between 1970 and the estimate year increases. This observation is not unexpected given the changes in fertility and migration that have occurred in Washington over the past two decades. However, the results of the statistical tests indicate that the relative gain in accuracy of the rate-correlation model becomes significant as time goes by.

The rate-correlation method shows a consistently higher degree of accuracy but the relative gain in accuracy is not dramatic. This is due in part to the relatively high levels of accuracy already existing for the ratio-correlation method. For example, in 1971 the absolute mean

Table 1.—1970–1960 Based Multiple Regression Models for Enrollment.

Symptomatic Indicator	Ratio-Correlation Form: Unstandardized Regression Coefficient	Ratio-Correlation Form: Standardized Regression Coefficient	Rate-Correlation Form: Unstandardized Regression Coefficient	Rate-Correlation Form: Standardized Regression Coefficient
(Intercept)	-.01469	--	-.00578	--
Employment	.07633	.13889	.07976	.13290
Votes	.22181	.17617	.24852	.18156
Autos	.66428	.62861	.64918	.60234
	R^2 = .7532		R^2 = .7123	
	Adj. R^2 = .7321		Adj. R^2 = .6877	
	S.E.E. = .0744		S.E.E. = .0085	

percentage error for the ratio-correlation method is only 2.08. This is an exceedingly high degree of accuracy that does not leave much room for improvement. Consequently, while the rate-correlation does improve it, the relative gain for five of the nine years could be considered marginal.

While it appears that the rate-correlation method is consistently more accurate than the ratio-correlation method in estimating enrollment, this evaluation is not directly transferable to the problem of estimating population. It can, however, be taken as an indication that transforming a ratio-correlation model into a rate-correlation model has the potential to improve estimation accuracy. A conservative approach would be that the rate-correlation method may or may not improve the population estimation accuracy expected from a ratio-correlation model, but it is not expected to provide less accurate population estimates. As empirical support for this argument, the following section contains an evaluation

Table 2.—Accuracy of 1971 through 1979 Enrollment Estimates Using the 1970–1960 Based Regression Models.

Year	Mean Absolute Percentage Difference: Ratio-Correlation Model	Mean Absolute Percentage Difference: Rate-Correlation Model	One-Tailed Probability (df = 38)
1971	2.08	1.99	$.10 < p < .25$
1972	3.56	3.49	$.10 < p < .25$
1973	4.41	4.39	$.40 < p < .45$
1974	5.31	5.20	$.10 < p < .25$
1975	5.56	5.54	$.40 < p < .45$
1976	6.04	5.78	$.001 < p < .005$*
1977	6.87	6.62	$.001 < p < .005$*
1978	6.93	6.74	$.01 < p < .05$*
1979	7.16	6.92	$.001 < p < .005$*

*Statistically significant.

of the relative accuracy of the two methods in estimating county population levels at a census date, 1970.

Population Models

Again using Washington state data, each county's civilian population under 65 years of age is the object of estimation. Here the census data are for 1970, and the years for which data are taken to construct the models are 1960 and 1950. The three symptomatic indicators used are employment, registered voters and enrollment. The ratio-correlation model and its corresponding rate-correlation model are presented in Table 3.

As in the case of the two models for enrollment presented in Table 1, these two models have structural similarities that are consistent with the transformation of a decennial amount of change to a rate of change. The models exhibit very high levels of explained variation, although the ratio-correlation model displays a slight advantage. Using these two models, estimates were generated for 1970 in order to compare the estimates produced by each model with the actual enumerated populations. The results of this evaluation are presented in Table 4.

In this comparison, it is obvious that both methods produce similar results. The mean errors are virtually identical and no statistical difference is shown by the paired observation *T*-test. Given that the pattern of change found during the 1950 to 1960 period is roughly similar in the 1960 to 1970 period, both models *should* produce virtually identical estimates. It is not surprising that they do this. A key point here is that the rate-correlation model did not provide *less* accurate estimates than the ratio-correlation model.

DISCUSSION AND SUMMARY

Perhaps one point that emerges from the empirical results is the similarity of the ratio-correlation and rate-correlation population estimates for 1970 using the 1960–1950 based models. When data are only available at ten-year intervals for evaluating estimates, the actual estimation accuracy of a given model over virtually 90 percent of its use is unknown. For example, while the relative gains in accuracy achieved using the rate-correlation model to estimate annual enrollments are modest, they are nevertheless observable gains and, in four of the nine years, statistically significant. Whether the rate-correlation model for *population* would exhibit similar gains in accuracy during intercensal years is a question that may never be answered empirically. The intercensal data are not

Table 3.—1960–1950 Based Multiple Regression Models for Civilian Population under 65 Years of Age.

	Ratio-Correlation Form		Rate-Correlation Form	
Symptomatic Indicator	Unstandardized Regression Coefficient	Standardized Regression Coefficient	Unstandardized Regression Coefficient	Standardized Regression Coefficient
(Intercept)	.04662	--	.00216	--
Employment	.06679	.07633	.04987	.07171
Voters	.55037	.51085	.54197	.50093
Autos	.35608	.45481	.34361	.47023
	R^2 = .9495		R^2 = .9434	
	Adj. R^2 = .9452		Adj. R^2 = .9385	
	S.E.E. = .0502		S.E.E. = .0051	

Table 4.—Accuracy of 1970 Estimates of the Civilian Population under 65 Years of Age Using the 1960–1950 Based Regression Models.

Year	Mean Absolute Percentage Difference		One-Tailed Probability (df = 38)
	Ratio-Correlation	Rate-Correlation	
1970	5.068	5.062	$.20 < p < .25$

sufficiently complete. However, any procedure that shows indications of a gain in accuracy is worth considering, especially when, as in the rate-correlation procedure, it represents virtually no increase in the cost of generating population estimates. Given the importance of these estimates of revenue-sharing and planning activities in both the private and public sectors, it is important for a demographic center to thoroughly examine the relative accuracy of its potential methods. Failure to do so could lead to undesirable consequences, including a loss of credibility and, in extreme cases, litigation (Martin and Serow, 1979).

In addition to the evidence provided by the enrollment data, there are several reasons why it may be justified to assume that a rate-correlation model will achieve a higher degree of accuracy for population estimates. One reason is that it uses the same *trends* as the ratio-correlation model. If these trends deteriorate during a number of estimation years, then the ratio-correlation model will suffer more because it extrapolates a structural relationship based upon a decennial amount of change.

A second reason why the rate-correlation model could be expected to be more accurate than a corresponding ratio-correlation model is related to the reduction in variance caused by transforming ratios into rates. Small areas such as counties are more likely to exhibit higher variance as time goes by and their patterns of change in symptomatic variables and population begin to diverge. Given this tendency, one could expect that coefficients constructed using the rate-correlation model will have less potential for divergence because it restricts the potential for variance more than a corresponding ratio-correlation model. By restricting the variance, the rate-correlation model being used will have coefficient values closer to a "true" best-fitting (but unknown) rate-correlation model than one would find for a corresponding ratio-correlation model being used if it could be compared to a "true" best-fitting (but unknown) ratio-correlation model. The comparison of results between the enrollment data and population data support this argument. The enrollment data represent a small "population" that underwent substantial change. By the late 1970s the rate-correlation model exhibits significant improvements in accuracy over the ratio-correlation model. This suggests that the rate-correlation model may be preferable for estimating small populations, whether as a whole across the counties in a state with a small total population, or as compositional groups such as age, race and sex.

It may also be the case that the rate-correlation model contributes to greater structural stability than does the ratio-correlation model. Notice the apparently anomalous results where the rate-correlation model provides higher accuracy in enrollment estimates late in the 1970s—far from the 1970 census point. The logic underlying the rate-correlation model implies just the opposite—that it would perform better earlier in the 1970s, closer to the 1970 census point. Viewed from this standpoint, the results for the enrollment data suggest that the reduction in variance described earlier is manifested as an increase in coefficient stability that is more apparent in dealing with small populations.

A third reason for improved accuracy may be due to the evidence provided by Malec and Katzoff (1983), which, as mentioned earlier, indicates that the effects of spatial autocorrelation (Ord, 1975) may be reduced by a logarithmic transformation. While time and other considerations did not permit an examination of this issue in this study, it would be worthwhile to evaluate this in future work involving the rate-correlation method.

In a general sense this paper has approached the problem of regression estimation accuracy from a single standpoint: the modification of a well-established technique. It should be noted that substantial gains in estimation accuracy are not likely to come from the modification of techniques; rather, they are more likely to come from reductions in measurement error and the acquisition of new data as symptomatic indicators. However, the costs associated with error reduction and new data are much higher than those associated with developing or refining techniques that use existing data. Consequently, it is likely that substantial data gains in regression estimation accuracy are not forthcoming. Instead, there may be only a series of small gains resulting from refining existing techniques and developing new ones. Given these considerations, the rate-correlation method may provide a good starting point for further research on the aspects of accuracy that are known to be related to specific techniques. Both the temporal stability of coefficients and spatial autocorrelation are examples of such aspects of accuracy. If researchers continue to evaluate these issues solely with respect to the ratio-correlation model, its full potential for accuracy may not be realized. The rate-correlation model is, after all, only a simple transformation of the ratio-correlation model, but it appears to offer the prospect of a relatively inexpensive way to achieve increased accuracy, and may be especially suitable for use with small populations.

NOTES

[1] As an example, in 1979 official county population estimates for Washington produced with the aid of the Ratio-Correlation Regression Method formed a basis for the allocation of $8 million directly to county governments, and $79 million to subcounty governmental units in 1980 (State of Washington, 1979: i).

[2] The data used in this and subsequent tables were taken from the official Washington State Data Base maintained by the Population, Enrollment, and Economic Studies Division, Office of Financial Management, State of Washington. The authors are grateful to Theresa J. Lowe and the late John R. Walker for their assistance in providing the data.

ACKNOWLEDGMENTS

This is a revision of a paper presented at the 1983 Annual Meeting of the American Statistical Association in the session, "Spatial Modelling of Demographic Data (Canada and the United States)." The authors are grateful to participants in the session as well as anonymous reviewers for providing comments that improved the paper.

REFERENCES

Bogue, D. J. 1950. A Technique for Making Extensive Population Estimates. Journal of the American Statistical Association 45:149–163.

Malec, D. and M. J. Katzoff. 1983. Application of Some Common Spatial Models to Population Data. Paper presented at the annual meeting of the American Statistical Association, Toronto, Ontario.

Mandell, M. and J. Tayman. 1982. Measuring Temporal Stability in Regression Models of Population Estimation. Demography 19:135–146.

Martin, J. H. and W. J. Serow. 1979. Virginia's State-Local Cooperative Program: Conflict and Cooperation in Producing Population Estimates. State Government Autumn:182–186.

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 Simple Regression Estimates as an Alternative to the Multiple Regression Estimate in Post Censal and Intercensal Population Estimation: A Case Study. Rural Sociology 36:187–194.

O'Hare, W. 1976. Report on a Multiple Regression

Method for Making Population Estimates. Demography 13:369–379.

———. 1980. A Note on the Use of Regression Methods in Population Estimates. Demography. 17:341–343.

Ord, K. 1975. Estimation Methods For Models of Spatial Interaction. Journal of the American Statistical Association 70:120–126.

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.

Serow, W. J. and J. H. Martin. 1977. Estimating Demographic Characteristics Using the Ratio-Correlation Method. Demography 15:223–233.

Shryock, H. S. and J. S. Siegel. 1973. The Methods and Materials of Demography. Second Printing (Revised). Washington, D.C.: U.S. Government Printing Office.

State of Washington. 1979. Population Trends, 1979. Olympia, Washington: Office of Financial Management.

Swanson, D. A. 1978. 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. Improving Accuracy in Multiple Regression Estimates of Population Using Principles from Causal Modelling. Demography 17:413–427.

———. 1981. Allocation Error in Population Estimates: An Overlooked Criterion With Fiscal Implications. Pp. 13–21 in Small Area Population Estimates: Methods and Their Accuracy and New Metropolitan Area Definitions and Their Impact on the Private and Public Sector, Series GE-41, No. 7. Washington, D.C.: U.S. Bureau of the Census.

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

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