Survey Methodology, December 1989
Vol. 15, No. 2, pp. 271–280
Statistics Canada

Confidence Intervals for Postcensal Population Estimates: A Case Study for Local Areas

DAVID A. SWANSON[1]

ABSTRACT

This paper presents a technique for developing appropriate confidence intervals around postcensal population estimates using a modification of the ratio-correlation method termed the rank-order procedure. It is shown that the Wilcoxon test can be used to decide if a given ratio-correlation model is stable over time. If stability is indicated, then the confidence intervals associated with the data used in model construction are appropriate for postcensal estimates. If stability is not indicated, the confidence intervals associated with the data used in model construction are not appropriate, and, moreover, likely to overstate the precision of postcensal estimates. Given instability, it is shown that confidence intervals appropriate for postcensal estimates can be derived using the rank-order procedure. An empirical example is provided using county population estimates for Washington state.

KEY WORDS: Population estimation; Confidence intervals; Ratio-correlation regression.

1. INTRODUCTION

A method of generating confidence intervals for postcensal estimates was not available until Espenshade and Tayman (1982) introduced a time-series regression estimation technique utilizing age-specific postcensal death rates. The Espenshade-Tayman technique represents an important breakthrough in estimation technology; however, like most breakthroughs it has limitations, of which two are notable:

1. The technique is likely to be unsatisfactory at the subprovincial or substate level (Espenshade and Tayman 1982); and
2. It is a major departure from the standard regression technique used in Canada and the United States for estimating county-equivalent populations, namely, ratio-correlation. This departure is a particularly salient issue in terms of data requirements and the experience of people responsible for making county-equivalent and other subprovincial level population estimates. (Statistics Canada 1987). The term "county equivalent" is defined as a Census Division in Canada (Statistics Canada 1987) and as a county in nearly all U.S. states; notable exceptions in the U.S. include Alaska, in which county-equivalents are Census Areas, Louisiana, where Parishes functions as counties, and Virginia, in which "independent cities" are included as county-equivalents.

This paper presents a means of developing confidence intervals for postcensal county-equivalent populations using the rank-order procedure, a modification of the ratio-correlation method introduced by Swanson (1980) that exploits causal modeling concepts to take into account postcensal structural changes in a given ratio-correlation model.

There are three issues relevant to the development of confidence intervals made using the ratio-correlation method. The first has to do with model stability over time. If the structure of associations among model variables is invariant over time, then the confidence intervals

[1] David A. Swanson, Department of Sociology, Pacific Lutheran University, Tacoma, Washington 98447, U.S.A.

constructed in regard to the model data set will apply to the population estimates generated by the model from the estimation data set. Although it has been consistently documented that it is not prudent to assume model invariance (D'Allesandro and Tayman 1980; Ericksen 1973, 1974; Mandell and Tayman 1982; Namboodiri 1972; O'Hare 1976, 1980; Smith and Mandell 1984; Spar and Martin 1979; Swanson 1980; Swanson and Prevost 1986; Swanson and Tedrow 1984; Tayman and Schafer 1982; Verma *et al.* 1983), it would be useful to have a testing procedure for stability. This leads to the second issue, namely, the use of a statistical test. If the test indicates that stability can not be assumed, and yet confidence intervals associated with, say, a model constructed using 1960-70 data, are applied to estimates generated for, say, 1979, they are likely to overstate the level of precision in the 1979 estimates. Thus, the third issue is the need for a procedure that will generate appropriate confidence intervals.

In the report that follows, a description of ratio-correlation is provided along with the modification that forms the basis for developing appropriate confidence intervals. Next, the logic for developing these confidence intervals is formally described, followed by an empirical example showing both the test for instability and the generation of both "inappropriate" and "appropriate" confidence intervals.

2. METHODOLOGY FOR POPULATION ESTIMATION

Ratio-correlation is a regression method designed to measure the temporal change in county-equivalent population proportions using observed temporal change in proportions of symptomatic indicators such as registered voters, covered employment and public school enrollment. The temporal change is measured by simply taking a ratio of proportions at two points in time.

Since enumerated population numbers for all county-equivalents are available only from the federal census, a ratio-correlation regression model is always constructed using two points in time separated by a regular number of years. It is formally described as

$$R_{it} = a_o + \sum_{j=1}^{k} (b_j)\ (X_i)_{jt} + \epsilon$$

where

a_o = the intercept term to be estimated

b_j = the regression coefficient to be estimated

ϵ = the error term

j = symptomatic indicator, $(1 \leq j \leq k)$

i = county-equivalent $(1 \leq i \leq n)$

t = the year of the most recent census

and

$$R_{it} = \left[\frac{P_{i,t}}{\sum P_{i,t}}\right] \div \left[\frac{P_{i,t-z}}{\sum P_{i,t-z}}\right] \qquad (1.A)$$

$$(X_i)_{t,j} = \left[\frac{S_{i,t}}{\sum S_{i,t}}\right] \div \left[\frac{S_{i,t-z}}{\sum S_{i,t-z}}\right]_j \quad (1.B)$$

where

Z = the number of years between each census

P = Population

S = Symptomatic Indicator

Once a model is constructed, it is used to develop a postcensal estimate for time $t + x$ by substituting $(S_{i,t+x}/\sum S_{i,t+x})_j$ into the numerator of the right-hand side of equation [1.B] while $(S_{i,t}/\sum S_{i,t})_j$ is substituted into the denominator of the right-hand side of equation [1.B]. This means that once $\hat{R}_{i,t+x}$ is obtained, an actual population for area i at time = $t + x$ is developed by introducing an independently estimated total population, P_{t+x}, into equation [1.A] and algebraically solving equation [1.A] for $P_{i,t+x}$. Since $\sum \hat{P}_{i,t+x}$ does not usually equal the independently derived total, P_{t+x}, an adjustment is made to force the summed population figures to the independently estimated total.

One limitation of ratio-correlation is that its structure is invariant over time, which is why the rank order procedure was introduced by Swanson (1980). The rank-order procedure is based on the fact that information contained in the zero-order correlations found in an estimation data set can be exploited due to work by Land (1969, Chapter IV); work that is based on the fundamental theorem underlying path analysis as developed by Wright (1921). 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.

3. METHODOLOGY FOR CONFIDENCE INTERVALS ESTIMATION

If the relationships found among the variables in the model data set remain stable over time (as shown through the rank-order procedure) then the same relationships should be found among the variables in the estimation data set. This stability would indicate that the S.E.E. associated with the model data set is appropriate for generating confidence intervals for the estimation data set. However, if stability does not exist, then the S.E.E. associated with the model data set is not appropriate, and may, in fact, generate confidence intervals that overstate the precision of postcensal estimates. These considerations lead to the question of determining stability through statistical inference.

In answering the question just posed, consider that we are examining related pairs of variables. This implies that the Wilcoxon matched-pairs signed rank test could be used (Mosteller and Rourke 1973). In using this test, the null hypothesis is that there are no differences between the population estimates (scores) produced by the unmodified and modified regresion models.

The key to developing confidence intervals for postcensal county equivalent population estimates is found in the fact that the rank-order procedure generates a set of regression coefficients for the estimation data set. From these coefficients, estimates of R^2 and the S.E.E. for the estimation data set can be developed, and the estimated S.E.E. leads directly to the development of confidence intervals. First, recall that the coefficient of multiple determination, R^2, is simply the sum of the products of each zero-order correlation between an independent variable and the dependent variable, and the standardized regression coefficient for each independent variable (Hayes 1973), so that S.E.E. is (Hayes 1973)

$$\text{S.E.E.} = \left[\frac{(n)\ (S_y^2)\ (1 - R^2)}{n - 2}\right]^{1/2}$$

where

n = number of cases (county-equivalents)

S_y^2 = variance of the dependent variable

R^2 = coefficient of multiple determination

The formula for generating a confidence interval around a given estimated value for a point on a (population) regression line is provided by Kmenta (1971)

$$Y_i \pm (t_{n-2,\alpha/2})\ (\text{S.E.E.})$$

An important point to realize is that the confidence interval is not directly generated for a population estimate, rather it is for the estimated ratio of proportions, or R_{it+x}. However, as shown by Espenshade and Tayman (1982), a confidence interval around one variable can be translated for another variable algebraically substituted for the first. Thus, by finding the lower and upper confidence boundaries of R_{it+x}, these lower and upper confidence boundaries can be translated into the population values:

$$(R_{it+x}) \pm (t_{n-2,\alpha/2})\ (\text{S.E.E.})$$

$$= \left[\frac{P_{it+x}}{\sum P_{it+x}}\right] \div \left[\frac{P_{it}}{\sum P_{it}}\right] \pm (t_{n-2,\alpha/2})\ (\text{S.E.E.})$$

which leads to

$$\text{L.L.}\ (\hat{P}_{it+x}) =$$

$$\left[\frac{P_{it}}{\sum P_{it}}\right] (\sum P_{it+x}) \left[(\hat{R}_{it+x}) - (t_{n-2,\alpha/2})\ (\text{S.}\hat{\text{E}}\text{.E.})\right]$$

and

$$\text{U.L.}\ (\hat{P}_{it+x}) =$$

$$\left[\frac{P_{it}}{\sum P_{it}}\right] (\sum P_{it+x}) \left[(\hat{R}_{it+x}) + (t_{n-2,\alpha/2})\ (\text{S.}\hat{\text{E}}\text{.E.})\right]$$

4. EMPIRICAL STUDY

Table 1.A in Swanson (1980) gives the zero-order correlations relating to a ratio-correlation model for estimating county civilian populations under sixty-five years from employment, voters, and grades 1-8 enrollment for the state of Washington, for the period 1950-1960. Characteristics of the model constructed from these data are given in Table 1.B. while Tables 2.A and 2.B provide similar results for the 1960-1970 period as found in Swanson (1980). This latter set forms the estimation data over which the procedure will be described.

Although full knowledge of the estimation data set is available, the procedure is used as if this were not the case. Of course, what is known in any estimation problem is the zero-order correlation matrix for the independent variables, which is used in conjunction with the fundamental theorem of path analysis to estimate the coefficients for the modified model. Using the complete rank-order procedure, the modified model (Swanson 1980) is:

$$Y = 0.046618 + 0.066786X_1 + 0.50727X_2 + 0.38736X_3.$$

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 are presented in Table 1 along with the actual enumerated populations.

The Wilcoxon test was conducted for the Washington data using the procedure in the SPSSx NPAR Tests command (SPSS 1986). To save space, the unmodified and modified population estimates are not presented. They can be found in Table 3 of Swanson (1980). Under the null hypothesis, the probability of obtaining $Z = -3.2096$ is 0.0013. Thus, the null hypothesis is rejected and it is assumed that instability exists for Washington counties in going from the model constructed using 1960/1950 data to the true unknown model associated with 1970/1960 data.

As a note of interest, the Chow test (Chow 1960) validated the results of the Wilcoxon test by showing that the difference between the "true" 1970-1960 ratio-correlation model and the 1960/1950 ratio-correlation model was statistically significant.

Had the results of the Wilcoxon test led us not to reject the null hypothesis, we would have used the unmodified coefficients from the 1960/1950 model data set to generate 1970 population estimates for Washington counties. Further, the S.E.E. for this same model (0.05022) would have been used to generate confidence intervals for the 1970 estimates. However, the results of the Wilcoxon test led us to reject the null hypothesis in this case. This indicates the modified coefficients developed using the rank-order procedure should be used in lieu of the unmodified model. Further, it indicates the need for a revised S.E.E., one that is not likely to overstate the precision of the 1970 estimates.

Using the estimated values found in the 1970 example data for Washington state (Swanson 1980) we find

$$\hat{R}^2 = (0.07533)\,(0.75290) + (0.47085)\,(0.92146) + (0.49481)\,(0.88082) = 0.926$$

and

$$(\text{S.}\hat{\text{E}}\text{.E.}) = \left[\frac{(39)\,(0.2145)^2\,(1 - 0.926)}{39\text{-}2}\right]^{1/2} = 0.0599$$

Table 1

90% Confidence Interval for the Estimated Civilian Population Under Sixty-Five Years by County, State of Washington 1970

County	Enumerated Population	Lower Limit	Estimated Population	Upper Limit	90% Confidence Interval (in percent)
Adams	11102	10335	11458	12581	± 9.80
Asotin	11862	10469	11814	13154	± 11.38
Benton	63144	60405	67511	74616	± 10.53
Chelan	35862	31733	36177	40620	± 12.28
Clallam	30023	28063	31294	34525	± 10.32
Clark	116663	101183	111437	121690	± 9.20
Columbia	3771	3683	4161	4639	± 11.49
Cowlitz	62586	55170	61581	67992	± 10.41
Douglas	15287	14569	16252	17935	± 10.36
Ferry	3336	2963	3397	3831	± 12.78
Franklin	23983	21960	24631	27302	± 10.84
Garfield	2546	2447	2761	3075	± 11.37
Grant	38921	37561	42606	47651	± 11.84
Grays Harbor	52583	46294	52114	57935	± 11.17
Island	20589	20512	22148	24040	± 7.39
Jefferson	9235	8440	9473	10506	± 10.90
King	1054271	935664	1037937	1140203	± 9.85
Kitsap	86529	77022	85821	94619	± 10.25
Kittitas	22764	17649	19863	22077	± 11.15
Klickitat	10729	10440	11923	13406	± 12.44
Lewis	39265	35747	40122	44497	± 10.90
Lincoln	8168	7939	9107	10275	± 12.83
Mason	18411	16057	17827	19596	± 9.93
Okanogan	22952	21002	23795	25688	± 10.97
Pacific	13310	11270	12795	14320	± 11.92
Pend Oreille	5185	5147	5893	6639	± 12.86
Pierce	339048	314272	346728	379184	± 9.36
San Juan	3089	2636	2918	3201	± 9.66
Skagit	45703	43255	48758	54261	± 11.29
Skamania	5330	4787	5358	5929	± 10.66
Snohomish	245193	213164	231996	250827	± 8.12
Spokane	251057	227372	256723	286072	± 11.43
Stevens	15178	13869	15780	17692	± 12.11
Thurston	68719	63644	69540	75436	± 8.48
Wahkiakum	3137	3033	3397	3761	± 10.72
Walla Walla	36608	33727	38271	42812	± 11.87
Whatcom	72111	63218	70670	78122	± 10.54
Whitman	34843	28960	32409	35858	± 10.64
Yakima	128960	120347	136203	152219	± 11.69

Note, that from Table 2 in Swanson (1980), the actual R^2 and S.E.E. values are 0.878 and 0.05077, respectively. In comparison with the actual S.E.E. of 0.05077, the estimated S.E.E. is higher. This is appropriate given that we are more uncertain about the precision of estimates generated by the rank-order procedure than we would be about the precision associated with the "true" model, if in fact, the true model was obtainable. With the rank-order procedure, we can now generate a confidence band from the following formula:

$$Y_i \pm (t_{37,\alpha/2})\ (0.0599)$$

In Table 1 an empirical example using a 90% confidence interval is given for the 1970 estimated county population figures presented also in Table 1. Here, the 90% confidence interval is given by:

$$\left[\frac{P_{i1960}}{2522141}\right] (3032053) \left[(\hat{R}_{i1970}) \pm (1.69)\ (0.0599)\right]$$

In examining the confidence intervals given in Table 1 in combination with the enumerated populations provided, it is found that in only one county (Kittitas) is the enumerated population outside of the 90% confidence interval. In this instance, the enumerated population exceeds the upper limit by 687 people. At a 90% level of confidence, the intervals are fairly wide, with a mean of 10.81, a minimum of ±7.39 percent for Island county and a maximum of ±12.83 percent in Lincoln County. Compare these with the mean of the absolute percent errors associated with the 1970 estimates, which is 4.89 (Swanson 1980). This comparison suggests that the 90% level generates intervals that are too broad for practical use. Given this, it is of interest to consider which level of confidence would be more appropriate. It is also of interest to consider the effect of using the unmodified S.E.E. (0.05022) from the 1960/1950 model. We would expect that the confidence intervals generated by the unmodified model would be too optimistic. That is, at a given level of confidence, there would be fewer than expected counties for which the interval encompassed the actual population. To explore these issues, Table 2 was constructed.

In Table 2, two distinct sets of information are provided. For both sets, however, a comparison is made between the unmodified and modified estimates and their associated confidence intervals. In regard to the issue of expecting optimistic confidence intervals for the 1970 estimates generated by the unmodified model, Table 2 indicates that at varying levels of confidence ranging from 90% down to 50%, the intervals are, indeed, optimistic in that for only two of the six levels examined are the expected number of county estimates within the specified level of precision. At the 80% level, for example, only 28 (72 percent) of the counties have enumerated 1970 populations within the confidence interval specified around the estimates; at the 60% level, only 22 (56%) of the counties have enumerated 1970 populations within the confidence interval specified around the estimates.

The second aspect of Table 2 is the mean interval associated with a given level of confidence. At the 90% level, the mean of the intervals associated with the unmodified model is 9.10 percent; for the modified model it is 10.81 percent. At the 50% level, the means are 3.66% and 4.35%, respectively. Thus, it is clear that the 60% and 50% levels of confidence generate a mean interval that is more in line with the mean absolute percent error, which is 4.88 for the modified model.

Table 2

Number (%) of Counties in Which Actual 1970 Population was Inside the Confidence Interval

Level of Confidence	Unmodified S.E.E. (0.05022)	Modified S.E.E. (0.0599)
90%	35 (89.7%)	38 (97.4%)
80%	28 (71.8%)	33 (84.6%)
70%	24 (61.5%)	29 (80.6%)
66.66%	24 (61.5%)	26 (66.66%)
60%	22 (56.4%)	23 (59.0%)
50%	20 (51.3%)	22 (56.4%)
	Mean Interval (in percent)	
	Unmodified S.E.E. (0.05022)	Modified S.E.E. (0.0599)
90%	9.10	10.81
80%	7.02	8.38
70%	5.66	6.75
66.66%	5.59	6.40
60%	4.59	5.47
50%	3.66	4.35

In examining the issue of confidence intervals, it appears that a procedure is needed for generating confidence intervals that are not misleading in terms of the precision of postcensal county-equivalent population estimates. However, guidance is also needed on selecting a given level of confidence that is appropriate for the estimates. Of interest in this regard is the work of Stoto (1983) on empirical confidence intervals for population projections. One of Stoto's (1983:18) findings is the high and low population projections produced for the United States by the Bureau of the Census (1977) correspond to a 66.66% confidence interval. It may be the case that for county-equivalent postcensal populations, that the 66.66% confidence level is also appropriate, although in this test this level of confidence generates a mean interval of 6.4 percent for the modified estimates, which is somewhat above their mean percent error (4.9). Another consideration is the length of time between the year for which a postcensal estimate is desired and the preceding census. In the example, the maximum period of postcensal time in the United States was used, 10 years. For each county, we have, in essence, a situation in which maximum uncertainty exists in regard to estimates. From this perspective, the relatively wide interval generated for each county at a 90 percent level of confidence is appropriate. We would expect that structural model changes occur relative to time. Hence, a narrower band would likely be generated in the first year following the end-census year of model construction than in the second year; and so on through the intercensal period.

5. CONCLUSION

At this point it should be clear that the rank-order procedure is not being presented as a fully-validated technique for constructing confidence intervals around postcensal county-equivalent population estimates. However, it appears to offer a reasonable starting point. Even with its limitations, the use of the Wilcoxon test and the confidence intervals developed using the rank-order procedure appears capable of providing benefits to those responsible for making such postcensal population estimates. In the first place, as noted by Espenshade and Tayman (1983), it is important to provide the users of postcensal population estimates some notion of their accuracy as do both the Wilcoxon test and the confidence intervals. Second, with the selection of appropriate confidence intervals, a formal means is available for resolving disputes over the population of a given county-equivalent by using hypothesis testing procedures. Third, S.E.E. can be used as a basis for selecting one model over another. This means that a set of different ratio-correlation models could be considered for any given postcensal estimation year and, further, that a formal criterion is available for selecting one model over another. This feature could be useful in the event that the ratio-correlation estimates generated by a federal, provincial or state demographic center, are challenged in a given postcensal year, an event that has become more frequent, especially in the U.S. (D'Allesandro 1987).

ACKNOWLEDGMENTS

The Author is grateful to Jeff Tayman, anonymous referees, and the editorial staff for comments and suggestions. Peggy Jobe typed a draft of this paper and Carey Taylor typed the final version.

REFERENCES

CHOW, G. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrics*, 28, 591-605.

D'ALLESANDRO, F. (1987). Should applied demographers take out liability insurance? Paper presented at the Annual Meeting of The Population Association of America.

D'ALLESANDRO, F., and TAYMAN, J. (1980). Ridge regression for population estimation: Some insights and clarification. *Staff Document No.* 56. Office of Financial Management, State of Washington: Olympia, Washington.

DRAPER, N.R., and SMITH, H. (1981). *Applied Regression Analysis, 2nd Edition*. New York: Wiley.

ERICKSEN, E.P. (1973). A method for combining sample survey data and symptomatic indicators to obtain population estimates for local areas. *Demography*, 10, 137-160.

ERICKSEN, E.P. (1974). A regression method for estimating population changes of local areas. *Journal of the American Statistical Association*, 69, 867-875.

ESPENSHADE, T.J., and TAYMAN, J. (1982). Confidence intervals for postcensal state population estimates. *Demography*, 19, 191-210.

HAYS, W.L. (1973). *Statistics for the Social Sciences*. New York: Holt, Rinehart and Winston.

KMENTA, J. (1971). *Elements of Econometrics*. New York: Macmillan.

LAND, K.C. (1969). Explorations in mathematical sociology. Unpublished Ph.D. dissertation. University of Texas, Austin.

MANDELL, M., and TAYMAN, J. (1982). Measuring temporal stability in regression models of population estimation. *Demography*, 19, 1351-46.

MOSTELLER, F., and ROURKE, R. (1973). *Sturdy Statistics*. Reading, Massachusetts: Addison-Wesley.

NAMBOODIRI, N.K. (1972). On the ratio-correlation and related methods of subnational population estimation. *Demography*, 9, 443-453.

O'HARE, W. (1976). Report on a multiple regression method for making population estimates. *Demography*, 13, 369-379.

O'HARE, W. (1980). A note on the use of regression methods in population estimates. *Demography*, 17, 341-343.

SMITH, S., and MANDELL, M. (1984). A comparison of local population estimates: The housing unit method versus component II, regression, and administrative records. *Journal of the American Statistical Association*, 99, 292-289.

SPAR, M., and MARTIN, J. (1979). Refinements to regression-based estimates of postcensal population characteristics. *Review of Public Data Use*, 7, 16-22.

SPSS, Inc. (1986). *SPSSx User's Guide*. Chicago: SPSS, Inc.

STATISTICS CANADA (1987). *Population Estimation Methods, Canada*. Catalogue No. 91-528E, Statistics Canada.

STOTO, M.A. (1983). The accuracy of population projections. *Journal of the American Statistical Association*, 78, 13-20.

SWANSON, D. (1980). Improving accuracy in multiple regression estimates of population using principles from causal modeling. *Demography*, 17, 413-427.

SWANSON, D., and PREVOST, R. (1986). Identifying extreme errors in ratio-correlation estimates of population. Presented at the Annual Meeting of the Population Association of America.

SWANSON, D., and TEDROW, L. (1984). Improving the measurement of temporal change in regression models used for county population estimates. *Demography*, 21, 373-381.

TAYMAN, J., and SCHAFER, E. (1985). The impact of coefficient drift and measurement error on the accuracy of ratio-correlation population estimates. *The Review of Regional Studies*, 15, 3-10.

U.S. BUREAU OF THE CENSUS (1977). Projections of the Population of the United States, 1977 to 2050. *Current Population Reports. Series P-25 No. 704*. Washington, D.C.: U.S. Government Printing Office.

VERMA, R.V.P., BASAVARAJAPPA, K.G., and BENDER, R.K. (1983). The regression estimates of population for subprovincial areas in Canada. *Survey Methodology*, 9, 219-240.

WRIGHT, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20, 557-585.