PREPRINT

OCT 1971

9

Morris L. Eaton

Error and Variance in the Population Prediction Process

UNIVERSITY OF COPENHAGEN
INSTITUTE OF
MATHEMATICAL STATISTICS

Morris L. Eaton

ERROR AND VARIANCE IN THE POPULATION PREDICTION PROCESS

Preprint 1971 No. 9

INSTITUTE OF MATHEMATICAL STATISTICS
UNIVERSITY OF COPENHAGEN

October 1971

Introduction.

This paper is a revision and expansion of my remarks on Professor Keyfitz's paper "On Future Population" presented at the Honolulu Symposium. The discussion below has grown from thoughts on Professor Keyfitz's paper and the stimulating discussion at the Symposium regarding the whole prediction process. In particular, I will draw on comments made by Paul Meier concerning sources of error in population predictions and a suggestion due to Lincoln Moses that the demographers might employ subjective (Bayesion Type) methods together with the classical δ-method for population predictions.

The paper basically consists of three parts. First, comments on Professor Keyfitz's work and a general discussion of the prediction processes are presented. Second, sources of error in population prediction and the incorporation of these errors into the prediction model are discussed. Lastly, based on remarks of Lincoln Moses, we present a somewhat technical discussion of the δ-method and the use of Bayesian methods for calculating an error variance for predictions.

The population model used below is, to say the least, unrealistic. However, it is not my purpose to explicitly develop useful formulae or models, but only to use this simplistic model to illustrate the basic notions of the prediction process, prediction error, and variance. The use of these ideas in the development of realistic models and computationally useful formulae is undoubtedly straightforward, although notationally cumberssome. To do so here would simply obscure the basic points I wish to make.

The Projection Process.

Professor Keyfitz's paper is basically a plea for demographers (and statisticians) to not only be more explicit concerning assumptions used to produce a projection or a prediction, but also to take a somewhat broader view of methods available for producing projections and predictions. The projection process described in the Keyfitz paper is, in essense, as follows: Let P_0 be the population size at time $t = 0$, B_0 be the birth rate at time $t = 0$ (we assume only one birth rate, for simplicity) and let D_0 be the death rate at time $t = 0$ (one death rate, for simplicity). The quantities P_0, B_0, and D_0 are calculated (with various adjustments) from observations on the current population. To compute the projected population size at time $t = 1$, say P_1, we assume that the population evolves through time according to some model. Based on this model, we then compute P_1 as some function of B_0, D_0, and P_0, say

$$P_1 = f(B_0, D_0, P_0). \tag{1.1}$$

The function f is a result of the model assumptions we have made and simply represents the calculations performed to get P_1 from B_0, D_0, and P_0. To compute the projected population size at time $t = 2$, say P_2, we use the same reasoning as above to obtain

$$P_2 = f(B_1, D_1, P_1) = f(B_1, D_1, f(B_0, D_0, P_0)). \tag{1.2}$$

However, at time $t = 0$, B_1 and D_1 are not available. In some cases, it is reasonable to assume that $B_1 = B_0$ and $D_1 = D_0$, while in other cases B_1 and D_1 are adjustments of B_0 and D_0 resulting from expected declining or increasing birth or death rates, or other considerations. We assume that B_1 and D_1 are arrived at by the demographer and are available in computing P_2. Similarly

$$(1.3)\quad P_3 = f[B_2, D_2, P_2] = f[B_2, D_2, f\{B_1, D_1, P_1\}] = f[B_2, D_2, f\{B_1, D_1, f(B_0, D_0, P_0)\}],$$

and so on for P_t, $t = 1, 2, \ldots$.

In terms of the symbols introduced above, Professor Keyfitz asks the demographer to specify the following: (a) the data and computation of B_0, D_0, and P_0; (b) the function f used to compute projections; and (c) the assumptions made and computations involved in arriving at B_i, D_i, $i = 1, \ldots$. Certainly, this is a reasonable request and (a), (b), and (c) are necessary information if one is to understand and evaluate the demographic projection. When one changes from projection to prediction*), further discussion of (1.1), (1.2), and (1.3) is required. The projection model given above is a deterministic model rather than a model involving random variables - that is, the quantities, $B_0, D_0, P_0, B_1, D_1, P_1$, etc., are treated as expectations (parameters) and not random variables. However, when one uses P_1 as a prediction of future population on which to make decisions, it is essential that we have some idea of the amount of error in P_1 as a predictor of the true population size at time $t = 1$. Thus, prediction necessitates that we think of P_1, and hence B_0, D_0, P_0, as variables which are subject to error. From a decision making point of view, a prediction together with some estimate of its standard deviation is perhaps the least one can demand. Professor Keyfitz goes even further in asking that the prediction

---

*) Throughout this paper, the distinction between projection and prediction is that made in Professor Keyfitz's paper "On Future Population".

of future population be accompanied with a probability distribution for its error - certainly a reasonable request, but a bit beyond our present knowledge and expertise.

With a prediction P_1 of future population at time t = 1, we would like the demographer to supply a number σ_1 - his estimate of the standard deviation of P_1. Certainly the number σ_1 will reflect his expertise and subjective evalutions of errors in B_0, D_0, P_0. My only basic criticism of the Keyfitz paper is that, although Professor Keyfitz asks for a complete specification of assumptions used to calculate P_1 as a projection, he does not go on to ask the demographer for assumptions and subjective evaluations that go into the calculation of σ_1. I would say that these latter assumptions are more important than the former in the evaluation of P_1 as a prediction of future population. For example, I would like to know Professor Keyfitz's assumptions and subjective evaluations for his confidence interval (240 million, 320 million) for the population of the United States in the year 2000.

Prediction Errors.

In this section we discuss various sources of error that give rise to prediction errors. This question has been discussed previously by Pollard (1966), was commented on by Paul Meier of the Symposium, and has undoubtedly been discussed by others. My breakdown of the errors is slightly different than those of Pollard or Meier. This error decomposition is the following:

I: Initialization Errors.

(a) Model errors - errors in f in (1.1).

(b) Parameter errors - errors in $B_0, D_0,$ and P_0.

II: Future Errors.

(a) Ordinary fluctuations in parameters in the future - ordinary fluctuations in B_i, D_i, P_i, $i = 1, \ldots$.

(b) Gross fluctuations in parameters in the future caused by major events (war, famine, large changes in economic conditions, etc.).

The breakdown of future errors into II(a) and II(b) is somewhat artificial since given II(b) has occured, one cannot separate II(a) from II(b), except perhaps on the basis of past ordinary fluctuations. However, it is important to note that there are fluctuations in B_i, D_i, P_i, $i = 1, \ldots$ in the absence of II(b). These are caused by advances in medical care, family planning, changes in attitude toward family size, etc.

We now want to incorporate errors into our model for population projection to obtain a model for prediction. Assume that the errors (I.a) can be ignored, so we proceed as if the model, f, is correct. To account for errors (II.b) write $P_0 + p_0$ for the true population size at time $t = 0$ where p_0 is a random variable with mean 0 and variance $\rho_0 \geqq 0$. Also write $B_0 + b_0$ for the true birth rate of time $t = 0$ where b_0 is a random variable with mean 0 and variance β_0, and $D_0 + d_0$ for the true birth rate at time $t = 0$ where d_0 is a random variable with mean 0 and variance δ_0. Then, set

$$\tilde{P}_1 = f(B_0 + b_0, D_0 + d_0, P_0 + p_0)$$

so $\tilde{P}_1$ is a random variable which represents the population size at time $t = 1$. The prediction process consists of <u>predicting</u> the random variable $\tilde{P}_1$

by $P_1 = f(B_0, D_0, P_0)$. More generally, let

$$\tilde{P}_{i+1} = f(B_i + b_i, D_i + d_i, P_i + p_i)$$

$i = 0, 1, 2, \ldots$ where $B_i + b_i, D_i + d_i, P_i + p_i$ are the true birth rate, death rate, and population size respectively at time $t = i$, and b_i, d_i, and p_i are random variables with mean 0 and variances β_i, δ_i, and ρ_i respectively. Then, prediction consists of predicting the random variable $\tilde{P}_{i+1}$ by $P_{i+1} = f(B_i, D_i, P_i)$. To obtain some idea of the variation in $\tilde{P}_{i+1}$, we would like to estimate (at least roughly) ρ_{i+1}. This is the subject of the next section.

The δ-method and Prediction Variance.

Again consider the prediction model of the previous section

$$\tilde{P}_{i+1} = f(B_i + b_i, D_i + d_i, P_i + p_i) \tag{3.1}$$

where $Var(b_i) = \beta_i$, $Var(d_i) = \delta_i$, and $Var(p_i) = \rho_i$. It is assumed that p_i is uncorrelated with b_i and d_i. Let $\gamma_i = Cov(d_i, b_i)$.

Using the so called δ-method, (Cramer, 1946), expand f in a Taylor series about the point (B_i, D_i, P_i), and discard terms higher than the linear term. Then

$$\tilde{P}_{i+1} \doteq f(B_i, D_i, P_i) + f_1(B_i, D_i, P_i)b_i + f_2(B_i, D_i, P_i)d_i + f_3(B_i, D_i, P_i)p_i \tag{3.2}$$

where $f_j(B_i, D_i, P_i)$ is the partial derivative of f with respect to its j^{th}

argument evaluated at (B_i, D_i, P_i). Computing the variance of both sides of (3.2),

$$\rho_{i+1} \doteq [f_1(B_i, D_i, P_i)]^2\beta_i + [f_2(B_i, D_i, P_i)]^2\delta_i + 2\, f_1(B_i, D_i P_i)\; f_2(B_i, D_i, P_i)\gamma_i + [f_3(B_i, D_i, P_i)]^2\rho_i. \tag{3.3}$$

Now, let $u_i = [f_1(B_i, D_i, P_i)]^2\beta_i + [f_2(B_i, D_i, P_i)]^2\delta_i + 2\, f_1(B_i, D_i, P_i) f_2(B_i, D_i, P_i)\gamma_i$ and $v_i = [f_3(B_i, D_i, P_i)]^2$. Then (3.3) becomes

$$\rho_{i+1} \doteq u_i + v_i\rho_i, \quad i = 0, 1, 2, \ldots . \tag{3.4}$$

Thus $\rho_1 \doteq u_0 + v_0\rho_0$, $\rho_2 \doteq u_1 + v_1\rho_1 = u_1 + v_1(u_0 + v_0\rho_0) = u_1 + v_1u_0 + v_1v_0\rho_0$. Continuing this process we find that

$$\rho_{i+1} \doteq u_i + \sum_{j=0}^{i-1} \left(\prod_{k=j+1}^{i} v_k\right)u_j + \left(\prod_{j=0}^{i} v_j\right)\rho_0. \tag{3.5}$$

From (3.5) ρ_{i+1} is determined by the numbers $\rho_0, u_0, \ldots, u_i$ and $v_0, \ldots, v_i$. Thus if one specifies the numbers $\rho_0, u_0, \ldots, u_i$ and $v_0, \ldots, v_i$, then ρ_{i+1} can be calculated. Since f (and thus f_j, $j = 1, 2, 3$) is known and B_i, D_i, and P_i are computed, to specify $\rho_0, u_0, \ldots, u_i, v_0, \ldots, v_i$, one must specify values for $\rho_0, \beta_0, \ldots, \beta_i$, $\delta_0, \ldots, \delta_i$, and $\gamma_0, \ldots, \gamma_i$.

Recall that ρ_0 can be interpreted as the variance of the error in P_0 as an estimate of the true population size at time $t = 0$. Further $\Sigma_i = \begin{pmatrix} \beta_i & \gamma_i \\ \gamma_i & \delta_i \end{pmatrix}$, $i = 0, 1, \ldots$, can be interpreted as the covariance matrix of the vector $\begin{pmatrix} B_i \\ D_i \end{pmatrix}$ as an estimate of the true birth and death rates of the population at time i. If the demographer is engaged in producing a population projection together with some estimate of its error, then, he has some idea as to the magnitude of error in P_0 and (B_i, D_i), $i = 0, 1, \ldots$ as estimates. Hence, if the demographer will transform his subjective ideas (of error in P_0 and (B_i, D_i), $i = 0, 1, \ldots$) into numerical values for ρ_0 and Σ_i, $i = 0, 1, \ldots$, then he can mechanically produce an estimate of the variance in his population projection. Further, the user of a population projection can evaluate this projection in terms of the numerical quantities ρ_0 and Σ_i, $i = 0, 1, 2, \ldots$.

The numerical values for ρ_0 and Σ_0 will reflect the demographers' assessments of errors (I.b). However, Σ_i, $i = 1, \ldots$ will reflect not only errors (I.b) but will reflect errors (II.a) and (II.b). Thus, with the method presented here, the demographer is able to incorporate his subjective feelings and uncertainty about the future via his values for Σ_i, $i = 1, \ldots$, and use these values to calculate an estimated value for the variance of his projection. Of course, it is the inclusion of these subjective feelings that makes the method smack of Bayesianism. Once the estimate of ρ_{i+1} has been computed, then approximate confidence intervals for population size at time i+1 can be easily computed by assuming normality or some other distribution for the population size.

Summary and Conclusions.

The subject of this paper has been the prediction process in demography with an emphasis on errors in predictions. Two broad categories of errors, intialization errors and future errors, were discussed. Then, the model (3.1) was introduced in an effort to compute an error variance for population projections. Let (B_i, D_i), $i = 0, 1, \ldots$ denote the birth and death rates used by the deomgrapher to calculate the population projection for time t = i+1. If the demographer specifies a subjective numerical value for the covariance matrix of (B_i, D_i), i = 0,1,... and a numerical value for ρ_0 (the variance of the estimated population size at time t = 0), then it has been demonstrated that ρ_{i+1} (the variance of the projected population at t = i+1) could be computed. Although this demonstration was based on a simplistic projection model, it appears that a direct application of the method will yield similar computational formula for projection models in current use.

The usefulness of the method discussed here depends on two issues: (a) the development of computational formulae for realistic projection models, and (b) the evaluation of the method in the light of past and current demographic data. Both (a) and (b) are areas for possible future research in demography.

References.

Keyfitz, N. (1972). On future population. To appear in the Proceedings of the Honolulu Symposium.

Pollard, J.H. (1966). The direct matrix product and stochastic population models. Biometrika 53, 397-415.

Cramer, H. (1946). Mathematical Methods of Statistics. Princeton University Press, Princeton, New Jersey.