Regression Estimates of Stochastic Compartmental Parameters Using Net Balance and Cumulative Flow Data

Richard L. Patterson

School of Natural Resources
The University of Michigan
Ann Arbor, Michigan 48109

ABSTRACT

A compartmental network is formulated in which each of N compartments represents a possible behavioral state of a discrete population immigrating from outside sources. The network contains absorbing states (sinks) sufficient to maintain an accounting record of every individual that enters the network. Time dependent distributions of counts of individuals in every compartment are given for the case of random (Poisson) immigration, from which formulae for expected net balances in compartments and expected cumulative flows into compartments are derived. These formulae are used in a regression model from which parameter estimates are obtained for the compartmental model. Model parameters that can be estimated, given data, are: (1) immigration intensities, (2) immigration to individual compartments, (3) emigration intensities from individual compartments, (4) residence time distributions in individual compartments, (5) compartmental transfer probabilities.

INTRODUCTION

An $M(t)/G/\infty$ queue was applied by Patterson [1] to model the abundance of the ichthyoplankton species *Dorosoma cepedianum* (gizzard shad) in western Lake Erie. A linear regression model of the conditional expectation of larval density was developed using calendar time as a single independent variable. That model was limited in the following respects:

(1) a single compartment (life stage) only was considered;
(2) the Poisson distributed larval birth process was assumed to be a piecewise constant function of time;
(3) no regressor variables other than mathematical functions of time were used.

This study removes restrictions (1) and (2).

APPLIED MATHEMATICS AND COMPUTATION 29:245–254 (1989)

655 Avenue of the Americas, New York, NY 10010 0096-3003/89/$03.50

A discrete population is assumed to enter a network of behavioral states (compartments) in which movement from state to state by individuals is described by a Markov renewal process $(X_n, T_n;\ T>0;\ n=1,2,\ldots)$ with stochastic transition matrix $P=(p_{ij})$ $(i,j=1,2,\ldots,N)$ and residence time distribution function matrix $W=(w_{ij}(z))$ $(i,j=1,2,\ldots,N)$. There are N states of the process, and X_n is the state of an individual at its nth transition. Associated with the Markov renewal process is a semi-Markov process $(X(t)=X_n;\ t_n\leqslant t<t_{n+1})$ in which $X(t)$ is the state of an individual at its most recent change of state. Movement of an individual from state to state constitutes a sequence of random delays as it undergoes behavioral changes. Absorbing states (sinks) are permitted for purposes of accounting for the total number of individuals that enter the system and, additionally, to account for the numbers of individuals that enter certain states from which they do not depart (death, for example). Associated with the matrices P and W is a third matrix called an interval transition probability function matrix $F=(f_{ij}(t))$ $(i,j=1,\ldots,N)$. The element $f_{ij}(t)$ is the conditional probability that $X(t)=j$ at time t, given that $X(t=0+)=i$. F is a stochastic matrix, since P is assumed to be stochastic. Elements of F are determined in terms of elements of P and W by conditioning on the number of changes of state of an individual prior to time t:

$$
\begin{aligned}
f_{ij}(t) &= \sum_{l=0}^{\infty} \Pr(X(t)=j \mid X(0+)=i,\ l \text{ changes of state in } (0,t)) \\
&\quad \times \Pr(l \text{ changes of state in } (0,t) \mid X(0+)=i) \\
&= \delta_{ij} h_i(t) + p_{ij} \int_0^t w'_{ij}(z) h_j(t-z)\,dz \\
&\quad + \sum_{l=2}^{\infty} \left[\sum_{q_1=1}^{N} p_{iq_1} \sum_{q_2=1}^{N} p_{q_1q_2} \cdots \sum_{q_{k-2}=1}^{N} p_{q_{k-3}q_{k-2}} \sum_{q_{k-1}=1}^{N} p_{q_{k-2}q_{k-1}} p_{q_{k-1}q_l} \right. \\
&\qquad \left. \times \left(\int_0^t (w_{iq_1} * w_{q_1q_2} * \cdots * w_{q_{k-2}q_{k-1}} * w_{q_{k-1}q_l})'(z) h_j(t-z)\,dz \right) \right] \\
&\qquad\qquad (i,j=1,\ldots,N), \quad (1)
\end{aligned}
$$

where

$q_l = j$;

$\int_0^t (w_i q_1 * \cdots * w_{q_{k-1}q_l})'(z) h_j(t-z)\,dz$ is an l-fold convolution density convolved with $h_j(t)$ and multiplied by the probability $p_{iq_1} \cdots p_{q_{k-1}q_l}$ that the l-step sequence of changes of state $(i, q_1; \ldots; q_{k-1}, q_l)$ occurs;

$h_i(t) = 1 - \sum_{k=1}^{N} p_{ik} w_{ik}(t)$;

δ_{ij} is the Kronecker delta;

state j is assumed to be accessible from state i, for otherwise $f_{ij}(t)=0$.

Equation (1) provides a computing formula for the probabilities $f_{ij}(t)$.

Let C denote a discrete population of a given class in which the behavioral states of individuals are in one-to-one correspondence with states of (X_n, T_n). Let $S = (P, W, C)$ symbolize the network governing transfers of individuals among states, independently, after they immigrate to S from external sources. The conditional probability that an individual is in state j at time t, given that it initially entered state i of S at time z $(0 < z < t)$, is $f_{ij}(t - z)$.

Let $(y_{ij}(t_m);\ i, j = 1,\ldots, N;\ m = 1,\ldots, M)$ denote an N by M array of counts or averages of counts of individuals in states $1,\ldots, N$ at times $0 < t_1 < t_2 < \cdots < t_M$. The index i denotes the initial state of entry of individuals into S. The regression problem is to model, and estimate, immigration into S and flows among states of S, given the system specification and data array $(y_{ij}(t_m))$. No lagged or cross-correlated variates are to be employed. Mathematical functions of time which are expectations of numbers of individuals in states at given times t are to be used as regressor variables.

NUMBERS OF INDIVIDUALS IN S UNDER POISSON IMMIGRATION

Let $Y_{ij}(t)$ $(i, j = 1,\ldots, N)$ be random variables denoting numbers of individuals in states $1,\ldots, N$ at any time $t > 0$ whose initial state of entry into S is i. The joint probability function of $Y_{i1}(t),\ldots, Y_{iN}(t)$ is

$$\Pr\left(Y_{i1}(t) = y_{i1},\ldots,\ Y_{iN}(t) = y_{iN}\right) = \prod_{j=1}^{N} \frac{\left(\int_0^t a_i(z) f_{ij}(t-z)\,dz\right)^{y_{ij}}}{y_{ij}!} e^{-\int_0^t a_i(z) f_{ij}(t-z)\,dz} \quad (2)$$

where

$a_i(z)$ is the nonnegative Poisson input intensity for immigrants into initial state i;

$y_{ij} = 0, 1, 2, \ldots.$

The conditional expected number of individuals in state j at time t, given that all immigrants initially enter state i, is

$$E\left(Y_{ij}(t)\right) = m_{ij}(t) = \int_0^t a_i(z) f_{ij}(t-z)\,dz \qquad (j = 1,\ldots, N). \quad (2.1)$$

The right side of Equation (2.1) is symmetric in the integrand with respect to the two nonnegative functions, and can be used as the basis for estimating

either $a_i(z)$ or $f_{ij}(z)$. It is only necessary to expand the chosen function as a linear combination of known functions in which unknown multiplicative coefficients must be estimated. This is the simplest possible case. The expectation is then set equal to the data elements $y_{ij}(t_m)$.

For the special case in Equation (2) where $N=1$ and $a_i(t)$ is a piecewise constant function over the interval $(0,t)$, Equation (2.1) reduces to Equation (6) in [1]. When immigrants initially enter states $1,\ldots,N$ with probabilities $q_1,\cdots,q_N$,

$$\Pr(Y_1(t)=y_1,\ldots,Y_N(t)=y_N)=\prod_{j=1}^{N}\frac{m_j(t)^{y_j}}{y_j!}e^{-m_j(t)}, \qquad (2.2)$$

where

$$m_j(t)=\sum_{i=1}^{N} q_i\int_0^t a_i(z)f_{ij}(t-z)\,dz, \qquad j=1,\ldots,N;$$

and

$$E(Y_j(t))=m_j(t). \qquad (2.3)$$

INTERNAL ARRIVALS TO AND DEPARTURES FROM STATES

Let $c_j(t)$ denote the Poisson arrival intensity of individuals to state j including both arrivals from other states and external immigrants, where state j is the initial state of entry to S. Let $b_{ji}(t)$ denote the Poisson intensity of departures from state j which arrive next at state i, where $\sum_{i=1}^{N} b_{ji}(t)=b_j(t)$ is the intensity of the stream of departures from state j. Then

$$\int_0^t b_j(z)\,dz=\int_0^t\sum_{i=1}^{N} b_{ji}(z)\,dz$$

$$=\int_0^t\left(\sum_{i=1}^{N} p_{ji}\int_0^z c_j(x)w_{ji}(z-x)\,dx\right)dz, \qquad j=1,\ldots,N. \quad (3)$$

Equation (3) gives the expected number of departures from state j in the time interval $(0,t)$.

Arrivals into state j are Poisson distributed with expectation

$$\int_0^t c_j(z)\,dz = \int_0^t \left(\sum_{i=1}^{N} p_{ij} \int_0^z c_i(x) w_{ij}(z-x)\,dx \right) dz + \int_0^t a_j(z)\,dz,$$

$$j = 1,\ldots,N. \quad (4)$$

The expectation of the Poisson distributed number of individuals in state j at time t can now be written in terms of $c_j(z)$:

$$E\big(Y_j(t)\big) = \int_0^t c_j(z) \left(1 - \sum_{i=1}^{N} p_{ji} w_{ji}(t-z) \right) dz, \qquad j = 1,\ldots,N. \quad (5)$$

Any of the equations giving expected numbers of individuals in states at time t provides a basis for a regression model from which Poisson intensity functions, interval transition probability functions, or residence time distribution functions can, in principle, be estimated.

ESTIMATION OF PARAMETERS

Represent the net balances $Y(t) = (Y_1(t),\ldots,Y_N(t))^T$ of individuals in states $1,\ldots,N$ at time t as

$$Y(t) = \alpha(t;\beta_1,\ldots,\beta_v) + \epsilon(t), \quad (6)$$

where

$\alpha(t;\beta_1,\ldots,\beta_v) = (\alpha_1(t),\ldots,\alpha_N(t);\beta_1,\ldots,\beta_v)^T$ is an N-dimensional column vector of expectations of numbers of individuals in states $1,\ldots,N$ at time t;

$\beta_1,\ldots,\beta_v$ are unknown parameters to be estimated;

$\epsilon(t) = (\epsilon_1(t),\ldots,\epsilon_N(t))^T$ is an N-dimensional column vector of random variables which are assumed to have means equal to zero for all $t > 0$.

In particular, $\alpha_j(t)$ may be set equal to the right hand sides of Equation (2.1), (2.3) or (5).

By setting

$$E\big(Y(t_i)\big) = \alpha(t;\beta_1,\ldots,\beta_v) = y(t_i) + r(t_i) \qquad (i = 1,\ldots,M), \quad (7)$$

where

$y(t_i)$ is an N-dimensional column vector of observations of numbers (or averages of numbers) of individuals in states $1,\ldots,N$ at time t_i, and

$r(t_i)$ is an N-dimensional column vector of residuals assumed to be sampled from $\epsilon(t_i)$,

a set of $M \times N$ equations in the unknowns $\beta_1,\ldots,\beta_v$ are obtained. The parameter estimation problem is stated as an optimization problem:

$$\min_{\beta \in B} \sum_{i=1}^{M} \sum_{j=1}^{N} \left|\gamma_{ij} \cdot r_j(t_i)\right|^d = J(\beta_0) \tag{8}$$

subject to

$$E(Y(t_i)) - r(t_i) = y(t_i) \qquad (i = 1,\ldots,M),$$

where

$\gamma_L < \gamma_{ij} < \gamma_U$ are importance weighting vectors;
B is a set of permissible vectors containing one or more optima β_0;
d is an exponent ($d > 0$)

Equation (8) is in generic form. When $d = 1$ the problem is one of minimizing the sum of absolute deviations of the computed vectors $E(Y(t_i))$ from observed or measured vectors $y(t_i)$. If a linear goal program is employed to find the optimum vector β_0, the residuals $r(t_i)$ are written as the difference of two nonnegative vectors so as to implement the simplex algorithm. The problem need not be linear in β, however. If it is, and if linear goal programming is used to find an optimum β_0, then sensitivity analysis is a convenient numerical technique for finding limits on observed vectors $y(t_i)$ within which the optimum β_0 does not vary. Additionally, uniqueness of the solution vector β_0 can be established through the solution to the dual program. Thus, when the optimization problem is linear in β and $d = 1$, linear goal programming offers advantages over other methods for finding the optimum β_0.

If $d = 2$, the problem is one of least squares. If the problem is linear in β, ordinary least squares methods can be used to obtain trial solutions for the optimum β_0. Difficulties can arise in that one or more elements of β may be negative. If this is so, the physical feasibility of negative elements of β must be checked. For instance, immigration intensities cannot be negative.

A second difficulty with least squares is the nonexistence of a solution due to collinearity of constraints. When Equation (2.1) or (2.3) is used to

represent $\alpha(t)$, collinearity can occur if the matrix P contains 1's off the diagonal and zeros on it.

A third problem which occurs, and is not unique to the case of $d = 2$, is bias due to missing data. When observations on numbers of individuals in given states are missing, flow rates into those states can be estimated nonetheless, at a risk of bias in the estimates. Manipulation of the importance weights may be able to reduce the bias somewhat.

The constraint set in Equation (8) is minimal. If the problem is one of estimating interval transition probabilities $f_{ij}(t)$, additional restraints must be specified to insure that $f_{ij}(t)$ lies in the positive unit interval and that $f_{i1}(t) + \cdots + f_{iN}(t) = 1$, for all positive t, and for all i. Such restraints are particularly easy to specify for $d = 1$ and when the optimization problem is linear in β.

CASE STUDY

A four-compartment model of flows of chlorides through the lower Great Lakes was developed [2]. Equation (2.1) was used to specify $\alpha(t; \beta_1, \ldots, \beta_6)$, in which the six unknown parameters were optimized and found to be

$$\beta_1 = c_{11} = 0.158, \qquad \beta_2 = c_{12} = 0.004, \qquad \beta_3 = c_{21} = 0.504,$$

$$\beta_4 = c_{22} = 0.048, \qquad \beta_5 = c_{31} = 0.000, \qquad \beta_6 = c_{32} = 0.000.$$

The parameter optimization problem was set up according to Equation (8) as an ordinary least squares exercise and gave unsatisfactory results. Not all parameters were nonnegative, which violated physical requirements of nonnegative immigration intensities. There occurred instances in which multicollinearity of restraints prevented a solution. A linear goal programme ($d = 1$) was subsequently formulated, from which the estimates given above were obtained. From the information given above it is clear which parameter estimates were negative when the least squares formulation was employed.

As explained in [2], residence time distributions were exponential with means $u_1^{-1} = 22.6$ yr, $u_2^{-1} = 2.6$ yr, $u_3^{-1} = 7.9$ yr. These estimates of hydraulic residence times for Lakes Huron, Erie, and Ontario, respectively, were obtained from a hydrological balance of the great lakes and watershed. Using these values optimized estimates of $\beta_1, \ldots, \beta_6$ were obtained, which led directly to estimates of immigration intensity functions $a_1(t)$, $a_2(t)$, and $a_3(t)$. The estimated intensities were then substituted into Equation (2.1), and Equation (8) was reformulated to reestimate the mean residence times

u_1^{-1}, u_2^{-1}, and u_3^{-1}. The distributions $w_{ij}(t)$ were written as

$$w_{23}(z) = \beta_1(1 - e^{-2.6z}) + \beta_2(1 - e^{-(2.6+K)z}) \qquad (K > 2),$$

$$\beta_1 + \beta_2 = 1, \qquad \beta_i > 0;$$

$$w_{34}(z) = \beta_3(1 - e^{-7.9z}) + \beta_4(1 - e^{-(7.9-K)z}) \qquad (K > 2),$$

$$\beta_3 + \beta_4 = 1, \qquad \beta_i > 0.$$

Values of K for each distribution were selected, and optimized values of $\beta_1, \ldots, \beta_4$ were obtained from a linear goal program, yielding optimized estimates of the residence time distribution functions with means 3.6 and 4.7 yr for lakes Erie and Ontario, respectively. The revised estimates of the residence time distribution functions were resubstituted into the original formulation of Equation (8), from which revised estimates of the immigration intensities $a_1(t)$, $a_2(t)$, and $a_3(t)$ were computed. The revised estimates differed only very slightly from the original ones, indicating that additional rounds of optimizing residence time distributions and immigration intensities would not change those functions.

DISCUSSION

Equations (2.1), (2.3), and (5) are all formulae for expected net balances in an arbitrary state j. They differ in the ease with which each facilitates estimation of given subsets of model parameters. The same unconditional net balance data are indicated for Equations (2.3) and (5), while Equation (2.1) limits net balance data to reflect immigration into compartment i only. Use of Equation (5), in contrast to (2.1) or (2.3), does not require prior derivation of interval transition probabilities $f_{ij}(z)$. The choice of Equation (2.1), (2.3), or (5) for use in the regression model (8) must necessarily depend upon which parameters are to be estimated and available data. As illustrated by Patterson and Ma [2], use of either Equation (2.1) or (2.3) in the model (8) leads to simultaneous equations involving parameters of inputs to more than a single compartment, yielding an interdependent system of constraints. Equation (5), as it stands, involves state j only, but by substituting the right hand side of Equation (4) into Equation (5) it can be made to yield an interdependent system of regression constraints, so that simultaneous counts of individuals in multiple compartments can be used to estimate model parameters.

ESTIMATES BASED ON CUMULATIVE INFLOWS OR OUTFLOWS

The left hand sides of Equations (2.1), (2.3), (3), (4), and (5) indicate which of three types of data are to be used with the respective equations when estimating parameters: (i) counts of individuals in states at given instants in time, (ii) cumulative counts of arrivals into states in subintervals of time, or (iii) counts of departures from states in subintervals of time. Equations (3) and (4) are to be used with counts of departures and arrivals, respectively. Unlike Equation (5) in structure, Equations (3) and (4) are simultaneous in the unknowns $c_1(t),\ldots,c_N(t)$, so that counts of arrivals and/or departures from a given state will influence estimates of parameters of inputs to other states as well, as determined by the transfer probabilities p_{ij}. This is true even though the parameters of input intensities $c_1(t),\ldots,c_N(t)$ remain lumped. Since the transfer probabilities p_{ij} and p_{ji} are not necessarily equal, and in fact p_{ij} may equal zero while p_{ji} may equal one, Equations (3) and (4) will, in general, yield estimates of different sets of parameters. For instance, assume that the cumulative counts of departures from j in time intervals $t_1, t_2,\ldots,t_M$ are $B_1, B_2 - B_1,\ldots, B_M - B_{M-1}$. Define $c_j(x)$ empirically as

$$c_j(x) = \beta_{1j}g_{1j}(x) + \cdots + \beta_{k_jj}g_{k_jj}(x),$$

where the β coefficients are unknowns to be estimated and the functions $g(x)$ are known. The regression constraints using Equation (3) are

$$B_1 = \beta_{1j}S_{1j}(t_1) + \cdots + \beta_{k_jj}S_{k_jj}(t_1),$$

$$\vdots$$

$$B_M = \beta_{1j}S_{1j}(t_M) + \cdots + \beta_{k_jj}S_{k_jj}(t_M),$$

where the functions $S(t_i)$ are computed following the right hand side of Equation (3) after $c_j(x)$ as defined above is substituted.

An important assumption underlying the above discussion is that the input intensities are represented empirically even though Equation (4) shows that, in principle, they can be solved for explicitly in terms of elements of the matrices P and W. If explicit solutions for $c_j(t)$ are required in terms of parameters of the Markov renewal process defined by P and W, Equations (2.1) and (2.3) provide those solutions.

INFERENCES ABOUT MODEL PARAMETERS

A model of the measurement vector $Y(t)$ incorporating a random component to account for unexplained variability [Equation (6)] permits subsequent testing of hypotheses about parameters, based upon assumptions about statistical behavior of the random term. Without formal specification of the statistical properties of the random component, no inferences about estimates are possible, even though Equations (7) and (8) may still be used to compute numerical values of estimators. Statistical behavior of the random component of the model of the measurement vector should be based upon experience with actual residuals obtained after applying Equation (8) to compute parameter estimates. Different assumptions about behavior of the random term of a measurement model are possible, and they are likely to be in error unless prior experience has suggested what they should be. The random term is typically assumed to be normally distributed, but actual plots of residuals may show that some other distribution more adequately describes the situation. Monte Carlo methods are often required to determine appropriate assumptions about residuals. Statistical behavior of residuals for the case of $d = 1$ in Equation (8) is discussed in [1] on the basis of previous studies by others (references cited).

CONCLUSION

Equation (8) remains valid for all three types of observational data: (i) net balances, (ii) inflow, and (iii) outflow. It permits a variety of optimizing criteria, and the constraints may be linear or nonlinear in the unknown parameters. The case study demonstrates that the optimization problem is easily solved as a linear program when $d = 1$. Least squares can provide useful information about optimum solutions even in the presence of possible difficulties of multicollinearity and negative estimates of parameters. By iterative application of the optimization model successive estimates of parameters of immigration and residence times in compartments can be obtained, which converge as demonstrated in [1].

REFERENCES

1 Richard L. Patterson, Regression estimates of inputs to an $M(t)/G/\infty$ service system, *Appl. Math. Comput.* 24:47–63 (1987).

2 R. L. Patterson and Zhenkui Ma, Statistical distributions of compartmentalized populations governed by continuous time, discrete state semi-Markov processes, *Appl. Math. Comput.*, 30(1):1–23.