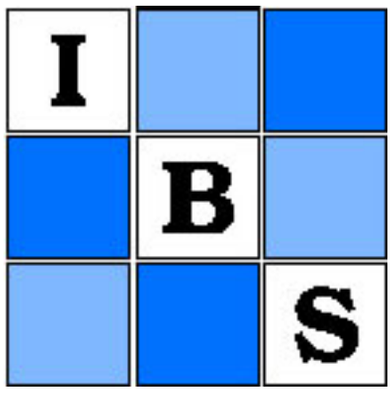

BIOMETRICS 44, 163–173
March 1988

Models for the Statistical Analysis of Infectious Disease Data

Michael Haber, Ira M. Longini, Jr., and George A. Cotsonis
Department of Statistics and Biometry, Emory University,
Atlanta, Georgia 30322, U.S.A.

SUMMARY

The Longini–Koopman model (1982, *Biometrics* **38,** 115–126) describes the process underlying the transmission of an infectious disease in terms of household and community level transmission probabilities. This model is generalized by allowing for different transmission probabilities that may correspond to various levels of risk factors on both the household and community levels. Two types of models are considered: (i) models for household data, where the numbers of susceptible and infected members in each household are known along with the values of household level risk factors; and (ii) models for individual data, where the infection status and risk factor level are known for each individual in the household. Although the type (i) models can be expressed as special cases of the type (ii) models, they deserve special attention as they can be represented and analyzed as log-linear models. Both types of models can be analyzed using maximum likelihood methods, while the type (i) models, when expressed as log-linear models, can also be analyzed by the weighted least squares method. Data from influenza epidemics in Tecumseh, Michigan and Seattle, Washington are used to illustrate these methods.

1. Introduction

An important objective of many infectious disease epidemiologic studies is to assess the impact of host and environmental risk factors on the transmission of infectious agents in the household and the community. Longini and Koopman (1982) formulated a probabilistic model that partitions the sources of infection transmission into those within the household and those from the general community. This model has been used to measure and compare the transmissibility of various infectious agents in a variety of community and household settings (see Longini et al., 1982; Longini et al., 1984a, 1984b).

In this paper, we introduce models that can provide a general framework for assessing the impact of risk factors on the two sources of transmission of infectious agents. This approach leads to methods of parameter estimation and hypothesis testing for reduced models that can be applied to the analysis of infectious disease data. More specifically, these models can be used to identify those risk factors that affect transmission within the household as opposed to risk factors that are associated with the transmission throughout the community. Previous studies (Monto and Ross, 1977; Fox et al., 1982b) suggest that the household and community sources of virus transmission may be related to different hosts, and to various demographic, socioeconomic, and environmental factors. It is therefore important to develop and investigate models that reflect these two sources of transmission and their association with various risk factors.

Two kinds of data are considered in this work: (i) household data, where for each household the number of susceptible and infected persons is known as well as the values of risk factors on the household level, such as geographic location or number of rooms;

Key words: Infection from the community; Infection within the household; Log-linear model; Maximum likelihood; Reed–Frost model; Weighted least squares.

(ii) individual data, where for each susceptible person we know whether he or she has been infected and the values of individual risk factors, such as age or the number of people who share the same bedroom. For both types of data, models are developed to describe the association of household and community transmission probabilities with risk factors. Data on epidemics of influenza in Tecumseh, Michigan and Seattle, Washington are used to illustrate the methods that are derived from these models.

2. A Probability Model for the Transmission of Infectious Diseases

Longini and Koopman (LK, 1982) derived a probability model for the final-size distribution for the number of household cases during a given epidemic period. The assumptions underlying this model are as follows:

(i) A person may become infected at most once during the course of the epidemic.
(ii) All persons are members of a closed "community." In addition, each person belongs to a single "household." A household may consist of one or several individuals.
(iii) The sources of infection from the community are distributed homogeneously throughout the community. Household members mix at random within the household.
(iv) Each person can be infected either from within the household or from the community. The probability that a person is infected from the community is independent of the number of infected members in his or her household.

Obviously, individuals who are immune at the beginning of the epidemic do not affect the spread of the disease and can be excluded from the analysis. We denote by B the probability that a susceptible individual is *not* infected from the community during the course of the epidemic. Further, let Q denote the probability that a susceptible person escapes infection from a *single* infected household member. For a household with s initial susceptibles and j initial infectives, LK have shown that the probability that exactly k additional individuals will become infected is given by

$$P(k \mid s,j) = \binom{s}{k} P(k \mid k,j) B^{(s-k)} Q^{(j+k)(s-k)}, \quad k = 0, 1, \ldots, s-1;$$
$$P(s \mid s,j) = 1 - \sum_{k=0}^{s-1} P(k \mid s,j). \tag{2.1}$$

If there is spread only within the household ($B = 1$), this reduces to the final-size distribution of the Reed–Frost model (see Bailey, 1975, p. 248, eq. 14.7). On the other hand, if there is no spread of infection among family members ($Q = 1$), the LK model (2.1) reduces to a binomial distribution. The probability $1 - B$ provides a measure of the community involvement in the disease transmission and has been referred to as the "community probability of infection" (CPI) by Longini et al. (1982). The probability $1 - Q$ measures the secondary transmission of infection within households. Thus, the value of $100 \times (1 - Q)$ provides an accurate measure of the household secondary attack rate (SAR) as described in Longini et al. (1982).

For the rest of this work, we assume that $j = 0$, i.e., at the beginning of the epidemic season one may become infected only from the community. LK derived an iterative numerical method to calculate maximum likelihood (ML) estimators of the transmission probabilities B and Q. The LK model has been found to provide an adequate fit to influenza and rhinovirus epidemic data (Longini et al., 1982; Longini et al., 1984a) and to rotavirus data (Koopman and Monto, unpublished manuscript). In addition, the LK model has been successfully applied to the analysis of dengue fever epidemic data (Dantes et al., 1988). The estimated SARs and CPIs have been used to classify and compare the transmissibility of the above-listed viruses in the household and the community across different household

types and community settings. Simulations have been conducted to investigate the robustness of the ML estimator for the SAR under a variety of epidemic situations. The estimator has been found to be quite robust (Longini et al., 1984b) even when the assumption (iii) of community homogeneity is violated.

3. Log-Linear Models for Household Data

Consider a household with s initial susceptibles. We now denote the probability that exactly k ($0 \leqslant k \leqslant s$) of these individuals become infected by π_{sk} (thus, $\pi_{sk} = P(k \mid s, 0)$ in the notation of Section 2).The quantities π_{sk} can be viewed as cell probabilities underlying a triangular contingency table of counts n_{sk} ($s = 1, \ldots, S$; $k = 0, 1, \ldots, s$) when S is the maximal household size (in terms of the number of susceptibles). The LK model can be written in the following log-linear form:

$$\log \pi_{sk} - \log \pi_{kk} - C_{sk} = (s-k)\beta + k(s-k)\gamma, \qquad k = 0, \ldots, s-1; \quad s = 1, \ldots, S; \tag{3.1}$$

where

$$C_{sk} = \log\binom{s}{k}, \quad \beta = \log B, \quad \gamma = \log Q, \quad \pi_{00} = 1$$

and

$$\sum_{k=0}^{s} \pi_{sk} = 1 \quad \text{for all } s.$$

Thus, each of the functions of the cell probabilities on the left side of (3.1) is a linear combination of the two parameters β and γ. The weighted least squares (WLS) method (Grizzle, Starmer, and Koch, 1969) can be used to assess the goodness of fit of the model (3.1) and to estimate the parameters along with their variance–covariance matrix.

The ML and WLS methods have been applied to data from the influenza A(H3N2) 1977–1978 epidemic in Tecumseh, Michigan (see Monto, Koopman, and Longini, 1985). Only households with 4 or fewer initial susceptibles have been included. The data and the estimates produced by the two methods are presented in Table 1. While the ML and WLS estimates are very close to each other, the standard error of WLS CPI is somewhat smaller than that of the ML CPI, although this may not be true in general. One should note that the model (3.1) is very sensitive to small frequencies, especially when they appear in a cell with $k = s < S$. When an observed frequency is zero, a constant must be added (to all the frequencies) in order to use the WLS method.

Suppose now that each household can be classified into one of R mutually exclusive strata. Each stratum may correspond to a level of a single risk factor or the combination of the levels of several risk factors. Let B_r and Q_r be the values of B and Q at the rth stratum ($r = 1, \ldots, R$). If π_{rsk} denotes the probability of k infected individuals in a household with s initial susceptibles at the rth stratum, then the model (3.1) is generalized as follows:

$$\log \pi_{rsk} - \log \pi_{rkk} - C_{sk} = (s-k)\beta_r + k(s-k)\gamma_r, \qquad k = 0, 1, \ldots, s-1; \quad s = 1, \ldots, S; \quad r = 1, \ldots, R; \tag{3.2}$$

with $\pi_{r00} = 1$ for all r and $\sum_{k=0}^{s} \pi_{rsk} = 1$ for all r, s. This model can be fitted using the WLS method. An alternative representation of the model (3.2) is:

$$\log \pi_{rsk} - C_{sk} = \alpha_{rk} + (s-k)\beta_r + k(s-k)\gamma_r, \qquad k = 0, 1, \ldots, s; \quad s = 1, \ldots, S; \quad r = 1, \ldots, R; \tag{3.3}$$

Table 1
Observed and expected frequencies on a household level from the influenza A(H3N2) *epidemic season* (1977–1978) *in Tecumseh, Michigan*[a]

Number infected	Number of susceptibles per household							
	1		2		3		4	
	Obs	Exp[b]	Obs	Exp[b]	Obs	Exp[b]	Obs	Exp[b]
0	65	67.1	88	78.4	27	31.8	22	24.1
1	13	10.9	14	21.6	15	11.1	9	9.4
2	—		4	6.0	4	5.2	9	5.7
3	—		—		4	1.9	3	3.4
4	—		—		—		1	1.4
Total	78	78.0	106	106.0	50	50.0	44	44.0
	Probability model:		$\widehat{\text{CPI}} = .140 \pm .015$, $\widehat{\text{SAR}} = 15.5 \pm 3.5$					
	(ML)		Goodness of fit $\chi^2(8\text{df}) = 12.14$, $P = .145$					
	Log-linear model:		$\widehat{\text{CPI}} = .123 \pm .006$, $\widehat{\text{SAR}} = 17.3 \pm 3.4$					
	(WLS)		Goodness of fit $\chi^2(8\text{df}) = 12.28$, $P = .139$					

[a] Only households with no more than 4 susceptibles are included.
[b] Expected frequencies are calculated from the probability model.

with $\alpha_{r0} = 0$ for all r. Let n_{rsk} be the observed frequencies. It is assumed that the totals $N_{rs} = \sum_{k=0}^{s} n_{rsk}$ are fixed by design. The model (3.3) can be written in terms of the expected frequencies $m_{rsk} = N_{rs}\pi_{rsk}$ as follows:

$$\log\left\{ m_{rsk} \Big/ \left[N_{rs}\binom{s}{k} \right] \right\} = \alpha_{rk} + (s-k)\beta_r + k(s-k)\gamma_r, \qquad (3.4)$$
$$k = 0, 1, \ldots, s; \quad s = 1, \ldots, S; \quad r = 1, \ldots, R.$$

The model (3.4) cannot be treated as an ordinary log-linear model as it does not include a term that corresponds to the fixed marginal totals N_{rs}. Thus, it is necessary to ensure that the expected frequencies satisfy $\sum_{k=0}^{s} m_{rsk} = N_{rs}$ for all r and s. The ML estimators of the expected frequencies m_{rsk} in (3.4) subject to these linear constraints can be calculated using an iterative method decribed in Haber and Brown (1986). Once the expected frequencies m_{rsk} have been estimated, ML estimates of the parameters α_{rk}, β_r, γ_r, and hence of the stratum-specific transmission probabilities can be obtained using standard methods.

Table 2 presents the frequencies of households in Tecumseh and Seattle (see Fox et al., 1982a) by the initial number of susceptibles and the number of household members infected during the 1977–1978 influenza A(H3N2) epidemic. In this case, the location of the household, i.e., Tecumseh vs Seattle, is the risk factor. We can see that while the secondary attack rates in the two cities are very similar, the probability of becoming infected from the community in Seattle is more than twice that in Tecumseh. This could be expected, since Seattle is a city (pop. cir. 500,000) with diverse sources of community transmission, while Tecumseh is a small town (pop. cir. 10,000) with far fewer sources of community transmission than Seattle. The WLS method provides standard errors for the parameters and facilitates significance testing; it can also be readily used to test composite hypotheses (e.g., H_0: $\beta_1 - \beta_2 = 0$ and $\gamma_1 - \gamma_2 = 0$) using a single test statistic (see §5). The ML method is useful in obtaining the estimated expected frequencies and examinations of discrepancies between observed and fitted counts.

Table 2
Observed and expected frequencies on a household level from the influenza A(H3N2) *epidemic seasons* (1977–1978) *in Tecumseh, Michigan and Seattle, Washington*[a]

Number infected		Number of susceptibles per household							
		Tecumseh (1)				Seattle (2)			
		1	2	3	4	1	2	3	4
0	Obs	65	88	27	22	29	19	23	9
	Exp[b]	67.1	78.4	31.8	24.1	30.4	23.1	20.0	8.2
1	Obs	13	14	15	9	12	14	11	9
	Exp[b]	10.9	21.6	11.1	9.4	10.6	12.8	13.2	5.7
2	Obs	—	4	4	9	—	9	11	2
	Exp[b]		6.0	5.2	5.7		6.1	10.0	5.2
3	Obs	—	—	4	3	—	—	4	3
	Exp[b]			1.9	3.4			5.8	4.8
4	Obs	—	—	—	1	—	—	—	4
	Exp[b]				1.4				3.2
Total		78	106	50	44	41	42	49	27

$\widehat{\text{CPI}}_1 = .123 \pm .006$, $\widehat{\text{SAR}}_1 = 17.3 \pm 3.4$ (WLS estimators)
$\widehat{\text{CPI}}_2 = .255 \pm .021$, $\widehat{\text{SAR}}_2 = 17.8 \pm 4.4$
Goodness of fit $\chi^2(16\text{df}) = 19.58$, $P = .24$

Tests:		
H_0:	$\text{CPI}_1 = \text{CPI}_2$	$Z = 5.56$, $P < .001$
H_a:	$\text{CPI}_1 \neq \text{CPI}_2$	
H_0':	$\text{SAR}_1 = \text{SAR}_2$	$Z = .08$ (n.s.)
H_a':	$\text{SAR}_1 \neq \text{SAR}_2$	

[a] Only households with no more than 4 susceptibles are included.
[b] Expected frequencies are calculated via the ML method.

4. Models for Individual Data

When individuals of the same household belong to different strata (for example, when strata are defined by age), the models discussed in Section 3 are inappropriate. We now describe an alternative approach that is useful in such situations. It is assumed that the risk level and the infectious status of each individual in a household are known. Consider a household with s initial susceptibles whose risk levels are $r_1, \ldots, r_s$. Let $x_i = 1$ for an infected individual and $x_i = 0$ for a noninfected ($i = 1, \ldots, s$). Denoting by $k = \sum x_i$ the number of infected in the household and using the arguments leading to (2.1), the probability of the outcome $(x_1, \ldots, x_s)$ is given by

$$P(x_1, \ldots, x_s \mid r_1, \ldots, r_s)$$

$$= \begin{cases} \prod_{i=1}^{s} B_{r_i} & \text{for } k = 0, \\ P(\mathbf{1}_k \mid r_{j_1}, \ldots, r_{j_k}) \prod_{i:x_i=0} B_{r_i} Q_{r_i}^k & \text{for } k = 1, \ldots, s-1, \\ 1 - \sum_{x_1,\ldots,x_s:\,\Sigma x_i<k} P(x_1, \ldots, x_s \mid r_1, \ldots, r_s) & \text{for } k = s, \end{cases} \tag{4.1}$$

where $j_1, \ldots, j_k$ are the indices of the k infected individuals and $\mathbf{1}_k$ denotes the array $(1, \ldots, 1)$ of order k. For example, the probability of the outcome $x_1 = x_2 = 1, x_3 = 0$ in a

Table 3

Observed and expected frequencies for households from the Influenza A(H3N2) *epidemic season* (1977–1978) *in Tecumseh, Michigan by the numbers of susceptible and infected adults* (*age* 18+) *and children* (*age* 0–17)[a]

No. susceptible/household		No. infected/household		No. of households	
Adults	Children	Adults	Children	Observed	Expected
1	0	0	0	64	66.3
		1	0	10	7.7
0[b]	1[b]	0	0	1	2.9
		0	1	3	1.1
2	0	0	0	84	80.4
		1	0	13	17.0
		2	0	3	2.7
1	1	0	0	4	3.9
		0	1	1	1.3 (a)
		1	0	0	.4 (a)
		1	1	1	.4 (a)
3	0	0	0	3	4.3
		1	0	3	1.3 (b)
		2	0	0	.4 (b)
		3	0	0	.1 (b)
2	1	0	0	19	19.3
		0	1	6	6.0
		1	0	5	3.2
		1	1	2	3.2 (c)
		2	0	1	.4 (c)
		2	1	0	.8 (c)
1	2	0	0	5	5.2
		0	1	0	2.8 (d)
		0	2	1	1.3 (e)
		1	0	1	.4 (d)
		1	1	0	.6 (e)
		1	2	4	.6 (e)
4[b]	0[b]	0	0	2	1.3
3[b]	1[b]	0	0	1	.5
2	2	0	0	19	16.6
		0	1	6	8.2
		0	2	5	3.5
		1	0	1	2.2
		1	1	4	3.4
		1	2	2	3.1
		2	0	0	.2 (f)
		2	1	1	.7 (f)
		2	2	1	1.1 (f)
1[b]	3[b]	0	1	1	.1
		1	0	1	0.0
4[b]	1[b]	0	1	1	.1
3[b]	2[b]	0	0	1	.8
		0	1	1	.2
2	3	0	0	2	2.2
		0	1	1	.4 (g)
		0	2	1	.3 (g)
		1	1	1	.1 (g)
		1	2	1	.1 (g)
		1	3	1	.3 (g)
		all other		0	3.6 (g)
1[b]	4[b]	0	1	1	0.0
Total				289	289.0

Overall goodness-of-fit $\chi^2(13) = 11.36$, $P = .58$

Maximum likelihood estimates:

Adults: $\widehat{CPI}_1 = .104 \pm .016$, $\widehat{SAR}_1 = 8.7 \pm 3.8$; Children: $\widehat{CPI}_2 = .272 \pm .041$, $\widehat{SAR}_2 = 21.3 \pm 7.1$

Hypothesis tests:

H_0: $CPI_1 = CPI_2$, $Z = 3.82$, $P < .001$; H_0: $SAR_1 = SAR_2$, $Z = 1.56$, $P \approx .12$

[a] Only households with no more than 5 susceptibles are included.

[b] These combinations were not included in the goodness-of-fit test. Only outcomes that have been observed at least once are shown.

(a)–(g): Outcomes with the same letter were pooled for the goodness-of-fit test.

household of size 3 is the product of the probability $P(1, 1 \mid r_1, r_2)$ that the first two individuals become infected and the probability $B_{r_3}Q_{r_3}^2$ that the third individual escapes infection from the community and from the two infected individuals in the household. Thus, for each household, the probability of the observed outcome $(x_1, \ldots, x_s)$ can be expressed in terms of the $2R$ parameters $B_1, \ldots, B_R, Q_1, \ldots, Q_R$. The LK model (2.1) is a special case of model (4.1) when $R = 1$, while the probability mass function of model (3.2) is a special case of model (4.1) when all the members in a household are at the same risk level. The likelihood function for (4.1) can be derived as the product of the probabilities of the observed outcomes for all the households in the sample. Numerical methods can be used to calculate the maximum likelihood estimators of the parameters. Hypotheses concerning these parameters (e.g., $B_1 = \cdots = B_R$) can be tested using the likelihood-ratio statistic for comparing two models. More complex hypotheses, such as lack of interaction between two risk factors, require the estimation of the covariance matrix of the parameters. This matrix can also be evaluated using numerical methods.

The goodness of fit of the model (4.1) can be assessed as follows: For a given household size s and a given combination of risk levels $r_1 \leqslant r_2 \leqslant \cdots \leqslant r_s$, there are at most 2^s different outcomes. The observed and expected frequencies of each outcome can be determined and a Pearson χ^2 statistic can be calculated. The sum of these χ^2 statistics over all combinations of risk levels can serve as an overall goodness-of-fit criterion. For many of these combinations, the number of households in the sample may be rather small. Thus, it may become necessary to pool outcomes in order to avoid sparse data. It is always advantageous to pool equivalent outcomes, such as the outcomes $(0, 1, x_3)$ and $(1, 0, x_3)$ when $s = 3$ and $r_1 = r_2$ (Haber, unpublished paper presented at Biometric Society ENAR meeting, Richmond, Virginia, 1981). The number of degrees of freedom for the goodness-of-fit test is calculated as the total number of outcomes for all risk factor combinations (after pooling) minus the number of risk factor combinations minus the number of estimated parameters ($2R$ in the most general model).

The analysis of individual infection data is illustrated in Table 3, using the influenza 1977–1978 A(H3N2) epidemic in Tecumseh. (The total number of households here exceeds that of Table 1, as households of size 5 are not included in the earlier analysis.) In this example, the risk factor is age. Comparison of age-specific transmission probabilities reveals that children are about 2.5 times more likely than adults to be infected from community sources. This age difference in the CPIs is consistent with past findings (see Longini et al., 1982; Longini et al., 1984a), as it is well known that schools, day-care centers, and other groups where children congregate play an important role in the community transmission of influenza. Children also seem more likely to be infected from another infected household member, though the difference in the SARs is not statistically significant.

5. Computations

Maximum likelihood and weighted least squares methods are used throughout this work. The ML methods for the models (2.1) and (3.4) are described in detail elsewhere [Longini and Koopman (1982) and Haber and Brown (1986), respectively]. The household model (3.2) can be written as $\mathbf{f}(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{f}(\boldsymbol{\pi})$ are given functions of the cell probabilities of a contingency table, $\mathbf{X}$ is a known matrix, and $\boldsymbol{\beta}$ are unknown parameters. Thus, the WLS method can be applied in order to obtain BAN estimators of the parameters:

$$\mathbf{b} = (\mathbf{X}'\mathbf{V}_\mathbf{f}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_\mathbf{f}^{-1}\mathbf{f}(\mathbf{p}),$$

where the elements of $\mathbf{f}(\mathbf{p})$ are the values of $\mathbf{f}$ evaluated at $\boldsymbol{\pi} = \mathbf{p}$ (the relative frequencies) and $\mathbf{V}_\mathbf{f}$ is the estimated covariance matrix of $\mathbf{f}(\mathbf{p})$, which can be obtained using the "delta

method" (Grizzle et al., 1969). The goodness-of-fit test statistic is given by

$$\chi^2 = [\mathbf{f(p)} - \mathbf{Xb}]' \mathbf{V}_\mathbf{f}^{-1} [\mathbf{f(p)} - \mathbf{Xb}].$$

Now, to test hypotheses regarding the elements of $\boldsymbol{\beta}$, one simply uses the fact that the WLS estimators $\mathbf{b}$ are asymptotically normally distributed with mean $\boldsymbol{\beta}$ and covariance $(\mathbf{X}' \mathbf{V}_\mathbf{f}^{-1} \mathbf{X})^{-1}$. Given that the model provides an acceptable fit to the data, one tests the hypothesis H_0: $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$, where $\mathbf{C}$ is the appropriately constructed matrix of linear contrasts. One should note that some standard computer versions of the WLS method (such as CATMOD of SAS) cannot be applied to the models and data discussed here. The household model (3.2) involves the constants C_{sk} and the relevant data form a set of triangular contingency tables. Although the statistical software GLIM (Generalized Linear Models; see McCullagh and Nelder, 1983) can be applied to the household model (3.2), we have developed special software to facilitate the analysis of the model presented in this work. (We will be happy to send copies of our software to interested researchers.)

The MLEs for the individual model (4.1) were found using nonlinear regression software AR in the BMDP statistical package (Ralston, 1985). This package employs a derivative-free, pseudo–Gauss-Newton algorithm. To obtain starting values for the parameters, it can first be assumed that the transmission probabilities are constant across strata and the simpler model (2.1) can be fitted to the data.

Sparse contingency tables cause difficulties in the analysis and interpretation of log-linear models. For household data, it may become necessary to limit the analysis to households that do not exceed a given size. (This has been done in Tables 1 and 2, where only households with 4 or fewer susceptibles have been included.) The maximum likelihood method used for the individual model (4.1) is less sensitive to sparse data.

6. Discussion

The log-linear models presented here provide a statistical framework for the analysis of infectious disease data. Such data present several characteristics that necessitate the use of special statistical models in their analysis. These characteristics are as follows:

(i) Dichotomous response variables (i.e., infected/not infected or ill/not ill)
(ii) Cluster sampling (e.g., sampling is done by households, rather than by individuals)
(iii) Correlated response variables within clusters

Characteristic (iii) results from the fact that a contagious process is being studied. The statistical analysis proposed here accounts for characteristics (ii) and (iii) by incorporating the infection process directly into the models (3.2) and (4.1). A standard statistical procedure, such as logistic regression, may not be appropriate for the analysis of infectious disease data since the usual assumptions of random sampling and independent responses are violated by characteristics (ii) and (iii), respectively.

Another advantage that the models presented here have over more standard statistical procedures, is that the parameters, i.e., the SAR and CPI, have direct epidemiologic interpretations in terms of transmission probabilities at different levels of risk factors. Thus, the parameter estimates can be used to evaluate the contribution made by various associated risk factors to disease transmission on the household and community levels. Such information can be useful in planning public health measures to break or decrease transmission of the infectious agents under study.

As described in Section 2, the LK model has been used to compare differences in the SAR and CPI at different levels of a risk factor measured on the household level. This was accomplished using large-sample tests based on the assumption of asymptotic normality of

the ML estimates for the SAR and CPI. However, the log-linear model (§3) provides a much more general framework for hypothesis testing including multiple tests and tests for trend in the SAR and CPI, at various levels of the risk factors, via the appropriately constructed contrasts. The methods presented here can be easily modified to include two or more risk factors so that one can explore their main effects and interactions with regard to the transmission mechanism. For the continuous risk factor, the model for individual data (4.1) can be modified by assuming a linear or log-linear trend in the transmission parameters, e.g., $B_r = A_0 + B_0 r$ or $B_r = A_0 B_0^r$ where r is the value of the risk factor.

The LK model cannot be used to directly evaluate the effects of risk factors on an individual level. Thus, the model for individual data, described in Section 4, constitutes a major advance in this area. For many viral agents, the most effective, and often the only means of disease control is through the application of prophylactic measures to individuals (e.g., vaccines, virucidal nasal tissues, interferon). The model for individuals can be used to evaluate the efficacy of such measures in carefully constructed household-level field trials. Even when the risk factors are equal for all the individuals in the same household, it may be advantageous to use the individual model. The analysis of individual data is less sensitive to small frequencies, compared to the household model. Also, the individual model is not affected by the absence of data on households with few susceptibles, while the household model cannot be used in its present form when, for instance, there are no data on households of size 1.

A number of modifications and extensions of the models presented here are being planned in order to accommodate a wide variety of infectious disease data sets. Complications in data analysis arise in the cases of incomplete and truncated data. The latter case occurs frequently when households enter the survey only after an initial infective appears, especially in the case of symptom data. Then the sample is truncated for households with no infected individuals. The LK model can be written in the zero-truncated form by dividing the right-hand side of equation (2.1) by the factor $1 - B^s$, and then finding the ML estimator of the CPI and SAR (see Longini and Koopman, 1982). However, the problem becomes computationally difficult. Accordingly, the model for individuals (4.1) can be easily fit to such data.

In some cases, the cluster sampling may be for units larger than households (e.g., schools or communities) and only a sample of individuals may be taken from each unit. The "chain binomial" form of the LK model will no longer be appropriate and will have to be replaced by a large-sample approximation. A Poisson approximation has been suggested (see Saunders, 1980; Longini, 1986), but the final-value approximation to (2.1), for large S, has yet to be derived.

Another extension involves the case where the data are sequential over time. Such data usually take the form of onset times of symptomatic illness (and duration of illness) for small sampling units, such as families, or larger units, such as schools or whole communities. Although several ML procedures have been proposed for the estimation of a single infection contact parameter (e.g., see Bailey, 1975; Becker and Hopper, 1983; Becker, 1983; Saunders, 1980), there has been little progress in developing methods for relating the degree of infectious contact to measured risk factors. Becker (1986) has applied GLIM to estimate the infectious contact parameter for sequential data. This approach could be generalized to include risk factors. Perhaps an approach similar to that used in survival analysis (see Cox, 1972) may be most appropriate. However, the inclusion of characteristics (ii) and (iii) above leads to technical difficulties when applying the survival analysis approach.

The methods presented in this paper provide a means for the systematic approach to the analysis of infectious disease epidemic data that arise in a number of epidemiologic situations. Such methods hitherto have been lacking in infectious disease epidemiology.

ACKNOWLEDGEMENTS

The authors wish to thank James S. Koopman for his comments on an earlier version of this manuscript. This research was supported by NIH Grant 1-R01-AI22877-01.

RÉSUMÉ

Le modèle de Longini–Koopman (1982, *Biometrics* **38,** 115–126) décrit le processus sous-jacent à la transmission de maladies infectieuse en termes de probabilités de transmission qui dépendent des foyers et des niveaux des facteurs de risque. On généralise ce modèle en autorisant différentes probabilités de transmission suivant le niveaux des facteurs de risque associés aux foyers et aux groupes sociaux. Deux types de modèle sont considérés: (1) modèle pour les données de foyers, ou les nombre de membres menacés et infectés dans chaque foyer sont connus et associés à la valeur des facteurs de risque par foyer; et (2) modèle pour les données individuelles, ou on connait pour chaque individu de chaque foyer son état infectieux et le niveau du facteur de risque qui lui est associé. Bien que les modèles de type (1) puissent être exprimés comme cas particuliers du modèle (2), ils demandent une attention spéciale car ils peuvent être représentés et analysés à l'aide de modèle log-linéaire. Des méthodes du maximum de vraisemblance permettent d'analyser les deux types de modèle, alors que le modèle de type (1), sous forme log-linéaire, peut aussi être analysé par des méthodes de moindres carrés pondérés. Pour illustrer ces méthodes, on utilise des données sur des épidémies de grippes à Tecumseh, Michigan et Seattle, Washington.

REFERENCES

Bailey, N. T. J. (1975). *The Mathematical Theory of Infectious Disease and Its Applications*, 2nd edition. New York: Hafner.

Becker, N. G. (1983). Analysis of data from a single epidemic. *Australian Journal of Statistics* **25,** 191–197.

Becker, N. G. (1986). A generalized linear modeling approach to the analysis of a data from a single epidemic. In *Pacific Statistical Congress*, I. S. Francis, B. F. J. Manly, and F. C. Lam (eds), 464–467. New York: Elsevier.

Becker, N. G. and Hopper, J. L. (1983). Assessing the heterogeneity of disease spread through a community. *American Journal of Epidemiology* **117,** 362–374.

Cox, D. R. (1972). Regression models and life-tables (with Discussion). *Journal of the Royal Statistical Society, Series B* **34,** 187–220.

Dantes, H. G., Koopman, J. S., Addy, C. L., et al. (1988). Dengue transmission on the Pacific coast of Mexico. *International Journal of Epidemiology* **17,** in press.

Fox, J. P., Coney, M. K., Hall, C. E., and Foy, H. M. (1982a). Influenzavirus infections in Seattle families, 1975–1979. I. Study design, methods and the occurrence of infections by time and age. *American Journal of Epidemiology* **116,** 212–227.

Fox, J. P., Coney, M. K., Hall, C. E., and Foy, H. M. (1982b). Influenza infections in Seattle families, 1975–1979. II. Pattern of infection in invaded households and relation of age and prior antibody to occurrence of infection and related illness. *American Journal of Epidemiology* **116,** 228–242.

Grizzle, J. E., Starmer, C. F., and Koch, G. G. (1969). Analysis of categorical data by linear models. *Biometrics* **25,** 489–504.

Haber, M. and Brown, M. B. (1986). Maximum likelihood methods for log-linear models when expected frequencies are subject to linear constraints. *Journal of the American Statistical Association* **81,** 477–482.

Longini, I. M. (1986). The generalized discrete-time epidemic model with immunity: A synthesis. *Mathematical Biosciences* **82,** 19–41.

Longini, I. M. and Koopman, J. S. (1982). Household and community transmission parameters from final distributions in households. *Biometrics* **38,** 115–126.

Longini, I. M., Koopman, J., Monto, A. S., and Fox, J. P. (1982). Estimating household and community transmission parameters for influenza. *American Journal of Epidemiology* **115,** 736–751.

Longini, I. M., Monto, A. S., and Koopman, J. S. (1984a). Statistical procedures for estimating the community probability of illness in family studies: Rhinovirus and influenza. *International Journal of Epidemiology* **13,** 99–106.

Longini, I. M., Seaholm, S. K., Ackerman, E., Koopman, J. S., and Monto, A. S. (1984b). Simulation studies of influenza epidemics: Assessment of parameter estimation and sensitivity. *International Journal of Epidemiology* **13,** 496–501.

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models*. London: Chapman and Hall.

Monto, A. S., Koopman, J. S., and Longini, I. M. (1985). Tecumseh study of illness. XIII. Influenza infection and disease, 1976–1981. *American Journal of Epidemiology* **121,** 811–822.

Monto, A. S. and Ross, H. (1977). Acute respiratory illness in the community: Effect of family composition, smoking and chronic symptoms. *British Journal of Preventive and Social Medicine* **31,** 101–108.

Ralston, M. (1985). Derivative-free nonlinear regression. In *BMDP Statistical Software Manual*, W. J. Dixon (ed.), 305–309. Berkeley, California: University of California Press.

Saunders, I. W. (1980). An approximate maximum likelihood estimator for chain binomial models. *Australian Journal of Statistics* **22,** 307–316.

Received September 1986; *revised April* 1987.