Estimating the Death Rate of an Emerging Infectious Disease by Time Series Analysis

Kung-Sik Chan[1,*] and Howell Tong[2]

[1] *Department of Statistics & Actuarial Science, University of Iowa*
Iowa City, IA 52242, U.S.A.
[2] *Department of Statistics, London School of Economics*
London WC2A 2AE, U.K.

SUMMARY

We introduce a novel time-series method for estimating the death rate of an emerging, infectious disease with censored aggregate data. Our approach is based on a Generalized Linear Mixed Model for the (daily) number of deaths in terms of the current and past (daily) number of newly confirmed cases. We illustrate the new method with data from the SARS outbreak in Hong Kong during 2003; the new method outperforms the simple WHO estimator of dividing the number of deaths by the number of confirmed cases, being less biased and converging more quickly to the death rate computed from the complete data.

KEY WORDS: distributed lag model; GLMM; SARS; transfer function model.

*Correspondence to: Department of Statistics & Actuarial Science, University of Iowa, Iowa City, IA 52242, U.S.A.

Contract/grant sponsor: National Science Foundation; contract/grant number: DMS-0405267

Contract/grant sponsor: Research Grants Council of Hong Kong; contract/grant number: HKU 7111/02P

1. INTRODUCTION

The world is constantly under threat of emerging or re-emerging infectious diseases, e.g. the SARS outbreak in 2003 and the pending influenza pandemic. An important epidemiological parameter of a new infectious disease is its death rate (also known as the case fatality rate), which may be difficult to estimate in a timely fashion and accurately, based on incomplete data from the initial course of the disease. Yet, clearly a timely and accurate estimate of the death rate as an emerging infectious disease is *evolving* is of urgent importance as it is of great concern to the general public and pivotal for formulating appropriate public health measures.

Here, we develop a novel time-series approach to estimating the death rate of an emerging, infectious disease. Our new approach circumvents the problem that the outcomes of many confirmed cases are censored, i.e. not known at the end of the study period. We do this by first expressing the conditional mean daily (or any basic sampling time unit) number of deaths in terms of a linear function of current and past daily number of (newly) confirmed cases. The coefficients of the daily number of confirmed cases and its lags constitute the (defective) probability mass function of time to death due to the disease. These coefficients are assumed to be smooth in that their second differences are small. The smoothness constraints can be incorporated by formulating a Generalized Linear Mixed Model (GLMM) for the daily number of deaths, which can be estimated by various methods, e.g. maximum likelihood estimation. An advantage of the time-series method is that it only requires aggregate data and not individual data as are necessary in epidemiological analysis via survival analysis; aggregate data are often more readily available.

In section 2, we formulate the GLMM for estimating the death rate of an infectious disease. Then, we illustrate our approach using the Hong Kong SARS data in section 3. We briefly conclude in section 4.

2. MODEL FORMULATION

We assume the following simple probabilistic model. Let p_n be the probability that a confirmed case ends up in a death on the nth day after the confirmation of the disease. Similarly, let q_n be the corresponding probability of being recovered and discharged from the hospital on the

nth day. Then $p = \sum_{n=0}^{\infty} p_n$ is the death rate of the disease, and $q = 1 - p = \sum_{n=0}^{\infty} q_n$ is the recovery rate. All cases are assumed to be independent of each other and have identical death rate and probability distribution of time to death (discharge).

In order to better appreciate the new time series approach, we first consider two fairly standard estimation methods. In the first approach, let K be the number of cases whose outcomes are known at the end of the study period, out of which there are D deaths. One may then proceed by assuming that conditional on K, D has a binomial distribution with probability p and number of trials K so that the death rate can be estimated by D/K, i.e. number of deaths over the sum of the number of deaths and that of discharges. Unfortunately, this approach is generally flawed because the conditional probability of death among the cases of known outcomes need not equal the unconditional probability of death, owing to possible "selection" bias. For example, this is the case if the time to discharge differs from the time to death on average, resulting in differential probabilities of censoring for the two types of outcome over a finite study period. The selection bias, however, vanishes if the study period is sufficiently long. A second approach circumvents the selection bias by adopting a cohort approach. In this approach, cases from the same day constitute a cohort. A cohort is *complete* if all of its cases have known outcome at the end of the study period. The binomial analysis is then restricted to the sample consisting of all cases from complete cohorts. The cohort approach does suffer from two problems, namely it requires detailed data that are often inaccessible and more importantly it is inefficient as it discards substantial amount of data.

We now explain the time series approach. Let C_t be the number of confirmed cases on the tth day, the corresponding number of deaths by D_t and that of recovered and discharged cases by R_t. Each death on day t must come from an earlier confirmed case. The probability that it is from day $t - n$ equals p_n and there are C_{t-n} cases on that day, so the conditional mean of D_t given $\mathcal{F}_t = \{C_{t-j}, j = 0, 1, 2, 3, ...\}$ equals $\sum_{j=0}^{\infty} p_j C_{t-j}$. For finite data, the infinite sum becomes a finite sum given by $\sum_{j=\ell_1}^{\ell_2} p_j C_{t-j}$, where $\ell_1, \ell_2 \geq 0$ are known lower and upper bounds of the time to death. Since D_t are counts, their conditional distributions may be specified as, e.g. Poisson or negative binomial. Thus, it is a Generalized Linear Model [8]

with an identity link function and the particular conditional mean function:

$$E(D_t|\mathcal{F}_t) = \sum_{j=\ell_1}^{\ell_2} p_j C_{t-j}, \tag{1}$$

where $C_t = 0$ for $t < 0$. Theoretically, D_t are serially dependent because they come from the earlier confirmed cases. Serial correlation among the D_t may be accounted for by including a latent process on the right side of (1) that may be specified as some Auto-Regressive Moving-Average (ARMA) process. The more general model then specifies that, conditional upon the counts of confirmed cases and the latent process ϵ, D_t are independent and have the conditional mean given by

$$\mu_t = \sum_{j=\ell_1}^{\ell_2} p_j C_{t-j} + \epsilon_t. \tag{2}$$

For disease data with low death rate, serial dependence may be negligible, which justifies the omission of the latent process from the model. A prudent approach consists of fitting a model without the latent process and then checking whether the residuals are roughly white noise, i.e. serially uncorrelated. In the case of significant residual serial correlation, an ARMA process can be specified for the latent process based on the residual serial dependence structure. Needless to say, other covariates may be incorporated in the model if needed, as will be demonstrated in the application below.

If the conditional response distribution is normal, the preceding model is simply the transfer function model, also known as the distributed lag model [4, 1]. The lags ℓ_1 and ℓ_2 are chosen to sufficiently span the range of non-zero p_j's. Often, a natural choice for ℓ_1 is 0. The coefficients p_j are expected to vary smoothly with the lag j. Indeed, the incorporation of this smoothness assumption in the estimation is essential because the model contains a large number of lags of C_t that can easily introduce multicollinearity. Hence, an unconstrained fit will generally yield very erratic estimates of the p's. The smoothness assumption may be effected by postulating some parsimonious parametric class of models, e.g. rational transfer function model [4] or polynomial model [1]. In practice, the functional form of the p's is seldom known, and a nonparametric approach may be more appealing as it lets the data tell us the functional form of the p's. Here, we adopt the nonparametric approach in that the second differences of the p's be independent and identically normally distributed with zero mean and standard deviation ζ; see [9]. The model may be estimated via a number of approaches, e.g. penalized likelihood

[7], Bayesian approach [9], or mixed-effect model [11, 12]. With suitable tuning parameters, estimators from these approaches are identical. The mixed-effect formulation is perhaps most amenable to statistical analysis, given the availability of standard software such as the nlme library of R [3, 12].

Some details of the mixed-effect approach follow; see also [12]. Define the m-dimensional parameter vector as $\theta = (p_{\ell_1}, \cdots, p_{\ell_2})^T$. Let $\mu = (\mu_1, \mu_2, \cdots, \mu_n)^T$ be the vector of the conditional mean responses. Let R be a constant $(m-2) \times (\ell_2 - \ell_1 + 1)$ matrix such that $R\theta$ equals $(p_{\ell_1} - 2p_{\ell_1+1} + p_{\ell_1+2}, \cdots, p_{\ell_2-2} - 2p_{\ell_2-1} + p_{\ell_2})^T$, the second differences of the p's. The constraint that the second differences of p's be small is then equivalent to requiring that the square-norm of $R\theta$ be small. Sometimes, the p_j's may be expected to taper off when j approaches one or both end-points of the interval $[\ell_1, \ell_2]$. As $p_j = 0$ for j outside the range $[\ell_1, \ell_2]$, the tapering-off of the p's as $j \to \ell_1$ is equivalent to requiring $p_j - 2p_{j+1} + p_{j+2}$ to be small for $j = \ell_1 - 2, \ell_1 - 1$, which can be effected by augmenting R by the two rows $(1, 0, \cdots, 0)$ and $(-2, 1, 0, \cdots, 0)$ and requiring $\theta^T R^T R\theta$ to be small. (Tapering-off at the other end-point or at both end-points can be similarly incorporated in the estimation.) The smoothness constraints can be implemented by requiring that θ has a multivariate Normal distribution of zero mean and precision matrix equal to $\lambda R^T R$, the inverse of which, if it exists, is the covariance matrix. The smoothness parameter $\lambda > 0$ controls the smoothness of θ, and is estimated from the data.

Consider the simple case with ϵ being absent, and let X be the design matrix. Then

$$\mu = X\theta, \tag{3}$$

where θ is multivariate normal with zero mean and precision matrix $S = \lambda R^T R$. Often, S is singular, e.g. for the unaugmented R, $S\theta = 0$ for any linear θ, i.e. $p_j = \beta_0 + \beta_1 j$, for $\ell_1 \le j \le \ell_2$. On the other hand, it can be checked that imposing the smoothness constraint on the p's across either end of the interval $[\ell_1, \ell_2]$ implies that S is invertible. In the general case, consider a singular value decomposition of $R = UDV^T$ where U and V are orthogonal matrices and D is a diagonal matrix whose diagonal elements are in descending magnitude. Let D be decomposed into two block diagonal matrices D_R and D_F, where the diagonal elements of D_R are non-zero and D_F is a zero matrix. Similarly partition $V = [V_R, V_F]$. Define $\theta_R = D_R V_R^T \theta$ and $\theta_F = V_F^T \theta$. Similarly define $X_R = XV_R D_R^{-1}$ and $X_F = XV_F$. Clearly $S = \lambda V D^2 V^T$. Then

(3) becomes

$$\mu = X_F\theta_F + X_R\theta_R, \tag{4}$$

where θ_F is the fixed-effect parameters and θ_R consists of iid random effects that are centered normal random variables of variance $1/\lambda$. Thus, the response vector $D = (D_1, D_2, \cdots, D_n)^T$ follows a Generalized Linear Mixed Model (GLMM), namely, D is conditionally Poisson distributed with mean vector given by (4). As remarked earlier, serial auto-correlation may be accounted for by further adding an ARMA process to the right side of (4). Maximum likelihood estimation of the GLMM can be done via the glmmPQL [10, 5] function in R. In the data analysis reported in the next section, maximum likelihood estimation is done via the gamm function (with a new smooth class of functions called "tfb"; available from the authors) of the library mgcv [12] of R, which uses the glmmPQL function to do approximate maximum likelihood estimation.

3. EMPIRICAL ANALYSIS OF THE DEATH RATE OF SARS IN HONG KONG

According to the World Health Organization (WHO), during the recent outbreak of severe acute respiratory syndrome (SARS) in 2003, a total of 8,098 people were sick with SARS. Of these, 774 died, resulting in a global death rate of approximately 9.56%. The disease originated in Southern China, reached Hong Kong in late February and from there spread to many countries. During the first few months when SARS struck Hong Kong, there was a controversy [2] on how best to estimate the death rate of SARS, which is of paramount public health concern. The death rate ranged from 2% in March, 2003 to 7.2% in May, 2003, based on the WHO formula of estimating the death rate by the ratio of the number of known deaths to the total number of confirmed cases. However, this formula is likely to underestimate the true death rate because the outcomes of many cases were still unknown at the time these figures were computed. The death rate of SARS may depend on a number of factors [6], e.g. age, sex, region, and time. The covariate region is a proxy for the medical treatment policy and the dominant strain of SARS while time is a proxy for the mutation effects of the primary coronavirus causing SARS. Here, we shall not address the dependence of the death rate on these specific factors. Instead, we focus on the problem of estimating the overall death rate of

SARS in Hong Kong. Reliable daily SARS data are available over the study period that begins on March, 11, 2003, designated as day 1, and ends at day 116 when the last SARS death case occurred in Hong Kong. During the study period, Hong Kong has accounted for about 22% of all the globally confirmed SARS cases.

There are altogether 1755 confirmed SARS cases in Hong Kong, out of which 299 cases resulted in death with the last death on day 116. This represents a death rate of 17.0% versus the much lower 9.6% overall death rate of SARS worldwide. Initially, we fitted the model with the conditional mean death on the t-th day as $\mu_t = \beta + \sum_{j=\ell_1}^{\ell_2} p_j C_{t-j}$ where $\ell_1 = 0$ and $\ell_2 = 50$ (the upper bound may be determined by some information criteria such as AIC and BIC; results reported below are robust to the use of other larger upper bounds), with the constraint that p_j tapers off when j approaches 50. Thus, we effectively assume that the time to death is at most 50 days long. The intercept term β should be zero under the model. However, the fitted model using all data from day 1 to day 116 shows that there is an outlier on the 9th day, see Fig. 1a. Hence, an indicator variable for the outlier is included in the model:

$$\mu_t = \beta_0 + \beta_1 O_t + \sum_{j=\ell_1}^{\ell_2} p_j C_{t-j}, \tag{5}$$

where $O_t = 1$ for $t = 9$ and 0 otherwise. The intercept term is estimated to be 0.038 with standard error 0.078, hence insignificant as expected. On the other hand, the coefficient estimate of the outlier equals 3.76 with standard error 1.96, with p-value equal to 0.058. The overall death rate of SARS (in Hong Kong) can be estimated by $\sum_{j=0}^{50} \hat{p}_j$ which equals 16.5% with standard error 0.0108, so an approximate 95% confidence interval equals $(14.4\%, 18.6\%)$. The fitted values appear to track the daily number of deaths relatively well; see Fig. 1a. The estimates of p_j provide an estimate of the probability mass function estimate of the time to death, which appears to be unimodal, see Fig. 1b. The Pearson residuals versus fitted values plot (Fig. 1c) shows no strong systematic pattern but the fitted values appear to underfit the observations, but this could also be related to the skewness of the Poisson distributions. There seems to be no other outliers, based on the Bonferroni rule. Furthermore, the Pearson residuals appear to have no serial correlation, as its auto-correlation function (acf, not shown) is significant only at lag 11, out of the first 20 lags. We conclude that the model defined by (5) provides a good fit to the data.

Figure 1 is about here

As the motivation of the transfer function model is to provide a less biased method for estimating the death rate of a new disease with incomplete data, it is of interest to check the empirical performance of the transfer function model using rolling data. Specifically, we fitted the model defined by (5) for days 1 to k, with $k = 30, 31, \cdots, 116$, and computed rolling death rate estimates. Fig. 1d plots the rolling death rate estimates based on the transfer function model, and their 95% confidence band, which shows that the death rate estimates rapidly approach the overall death rate, and indeed are quite close to the target by day 40 or later, although there are some small oscillations over the period of day 40 to day 60. In comparison, the simple rolling ratios of cumulative number of deaths over that of confirmed cases converge monotonically to the overall death rate but they are rather biased downward and only get close to the target by day 70 or later, see Fig. 1d.

4. CONCLUSION

We have demonstrated the usefulness of the new approach for providing relatively timely and accurate estimates of the death rate of an emerging disease with censored aggregate data. Moreover, the method does not assume the functional form of the probability mass function of time to death due to the disease. It is relatively straightforward to include covariates in the model. As the probability mass function is non-negative, the coefficients p_j must be non-negative, which was not enforced in the SARS data analysis. An interesting future research problem concerns how to carry out the estimation of the GLMM defined by (1) subject to the constraint that $p_j \geq 0$.

REFERENCES

1. Almon S. The distributed lag between capital appropriations and expenditures. *Econometrica* 1965; **33**(1): 178-196.
2. Altman KL. The SARS epidemic: the front-line research; study suggests a higher rate of SARS death. *New York Times* 2003, May 7.
3. Pinheiro JC and Bates DM. *Mixed-Effects Models in S and S-PLUS*. Springer-Verlag: New York; 2000.

4. Box GEP, Jenkins GM and Reinsel GC. *Time Series Analysis: Forecasting and Control 3rd ed.*. Prentice Hall; 1994.
5. Breslow NE and Clayton DG. Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association* 1993; **88**, 9-25.
6. Donnelly CA, Ghani AC, Leung GM, Hedley AJ, Fraser C, Riley S, Abu-Raddad LJ, Ho LM, Thach TQ, Chau P, Chan KP, Lam TH, Tse LY, Tsang T, Liu SH, Kong JH, Lau EM, Ferguson NM, Anderson RM. Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong. *The Lancet* 2003; **361**(9371): 1761-6.
7. Green PJ and Silverman BW. *Nonparametric Regression and Generalized Linear Models*. CRC Press: Boca Raton; 1993.
8. McCullagh P and Nelder JA. *Generalized Linear Models 2nd ed.* CRC Press: Boca Raton; 1999.
9. Shiller RJ. A distributed lag estimator derived from smoothness priors. *Econometrica* 1973; **41**: 775-788.
10. Venables WN and Ripley BD. *Modern Applied Statistics with S. 4th edition*. Springer-Verlag: New York; 2002; pp 297-8.
11. Wang Y. Mixed effects smoothing spline analysis of variance. *J.R. Statist. Soc. B* 1998; **60**: 159-174
12. Wood SN. *Tensor product smooth interaction terms in Generalized Additive Mixed Models.* 2004; Technical Report of the Department of Statistics, University of Glasgow, UK; `http://www.maths.bath.ac.uk/~sw283/simon/papers/tgamm.pdf`.

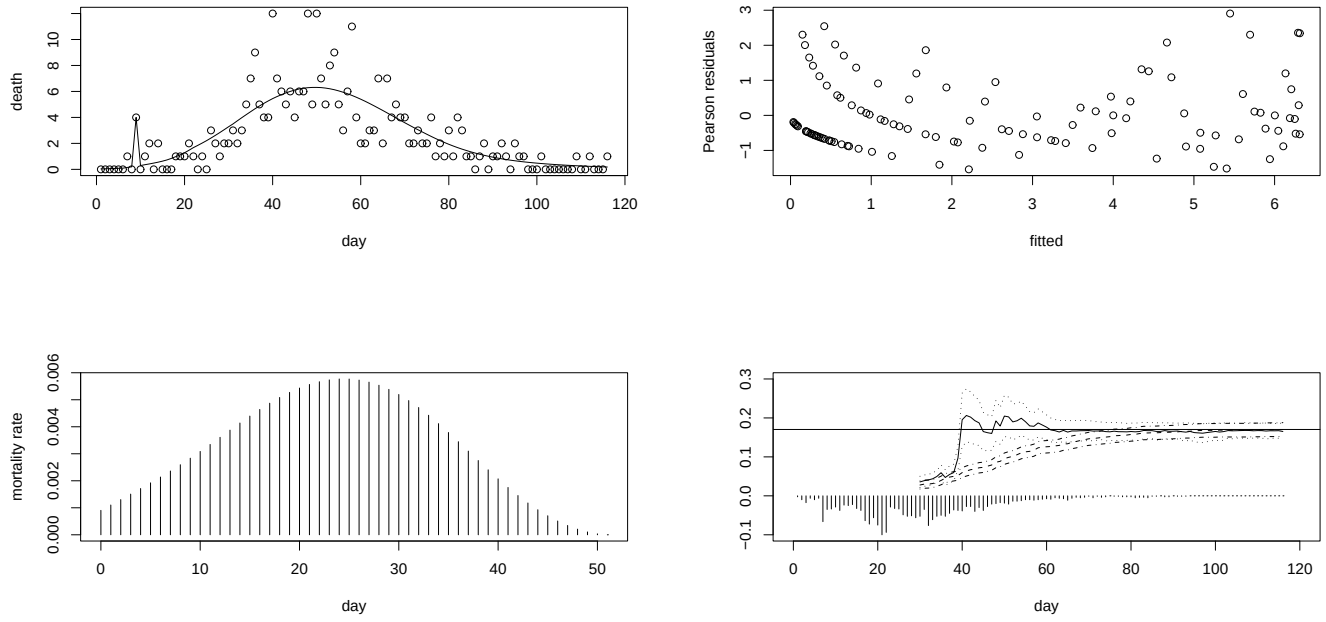

Figure 1. Hong Kong SARS data. (a) Time Plot of the daily number of deaths; raw data – open circle. The fitted curve is from the transfer function model using all data from day 1 (March 11, 2003) to day 116. (b) Maximum Likelihood Estimates of $p_j, j = 0, 1, \cdots, 50$. (c) The Pearson residuals versus the fitted values plot of the fitted model. (d) The estimates of the death rates estimated from the model using rolling data from day 1 to day $k, k = 30, 31, \cdots, 116$. Solid curve – estimates of the death rate of SARS based on the transfer function model, dotted curves – corresponding 95% confidence limits, dashed curve – sample proportion of death cases, dotdash curves – corresponding 95% confidence limits, horizontal line – SARS death rate based on all data. The vertical bars are proportional to negative daily number of deaths.