Estimation Of SARS Daily New Cases

H Liang

Citation

H Liang. *Estimation Of SARS Daily New Cases*. The Internet Journal of Infectious Diseases. 2005 Volume 5 Number 2.

Abstract

We used an auto-regression model to fit the daily new case number from the 2003 severe acute respira¬tory syndrome outbreak in Beijing, and demonstrated that the conventional model selection criteria are inappropriate for a selection of the model order. An improved AIC procedure was suggested for over¬coming the deficiency of these criteria. The resulting model indicated that we may use the cases of the previous 15 days to estimate the new case number in the current day. The conclusion of our modeling may give insights into ongoing outbreaks that may facilitate public health responses.

INTRODUCTION

In the past two years, mathematicians, statisticians, and biologists proposed a variety of models to analyze severe acute respiratory syndrome (SARS) cases since it occurred in the Southeast Asian countries in 2003. Riley et al. ([1]) and Lipsitch et al. ([2]) used general dynamic models to study the respective transmission dynamics of SARS in Hong Kong and Singapore. Their models may be too complicated to be used in practice as pointed out by Hsieh, Chen, and Hsu ([3]). The latter authors ([3]) used a linear system of equations and applied three-stage least squares to estimate the parameter, which can delineate the rapid epidemic growth. Zhou and Yan ([4]) used Richards model ([5]), a logistic-type model, to fit the cumulative number of SARS cases reported daily in Singapore, Hong Kong, and Beijing, and properly confirmed that the epidemic might be brought under control if the current intervention measures were continued. Hsieh and Cheng ([6]) further used a variation of the single-equation Richards model to fit the daily cumulative case data from the 2003 SARS outbreak in Toronto, and the authors estimated the turning points and case numbers during the 2 phases of this outbreak. Cauchemez et al. ([7]) proposed a Bayesian statistical framework for estimating the reproduction number early in an epidemic, and applied their approach to the SARS epidemic that started in February 2003 in Hong Kong. Intuitively SARS cases in the current day are strongly related to the situation of the previous days. In this article, we use an auto-regression (AR) model ([8]) to fit the SARS cases from April 21 to June 7 in Beijing.

On April 21, 2003, the World Health Organization (WHO) reported 3,861 probable severe acute respiratory syndrome (SARS) cases with 217 deaths globally ([9]); At that time, Beijing already had 588 probable cases containing 143 new cases, the capsheaf of transmission. The daily new cases gradually decreased and reached zero on June 2, 2003. Although there were two cases appeared on June 7 and 11, 2003, SARS transmission in Beijing was basically controlled.

METHOD

The AR model is of form:

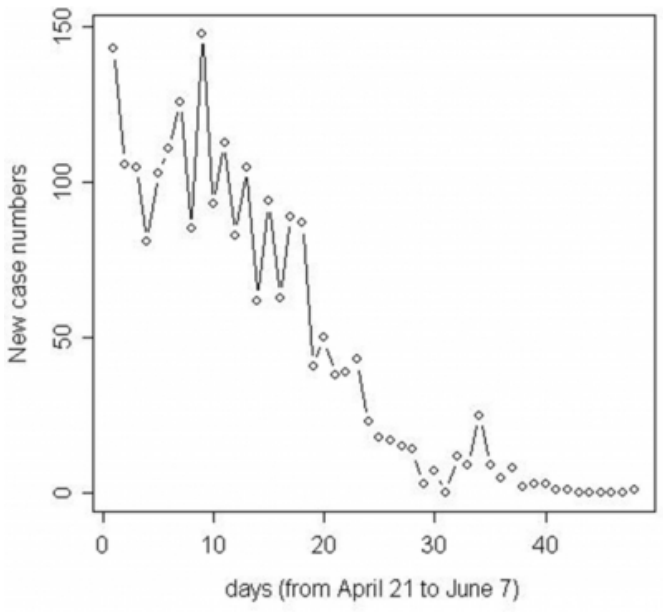
Figure 1

$$X(t)= \sum_{j=1}^{p} b_j X(t-j)+ \varepsilon_t$$

where X(t) ..., X(t-p), denotes the observations of the t,...,(t-p)-th days, p is the order of the model, b_1,... ,b_p are the unknown auto-regression coefficients, and ε_t is the measurement error. This model means that the current term of the series can be estimated by a linear weighted sum of previous terms in the series. The weights are the auto-regression coefficients. The ultimate goal is to derive an appropriate model, which may be used to forecast the cases of the proceeding day using the cases of the prior days. We present the SARS case numbers in Beijing from April 21 to June 7, 2003 in Figure 1.

Figure 2

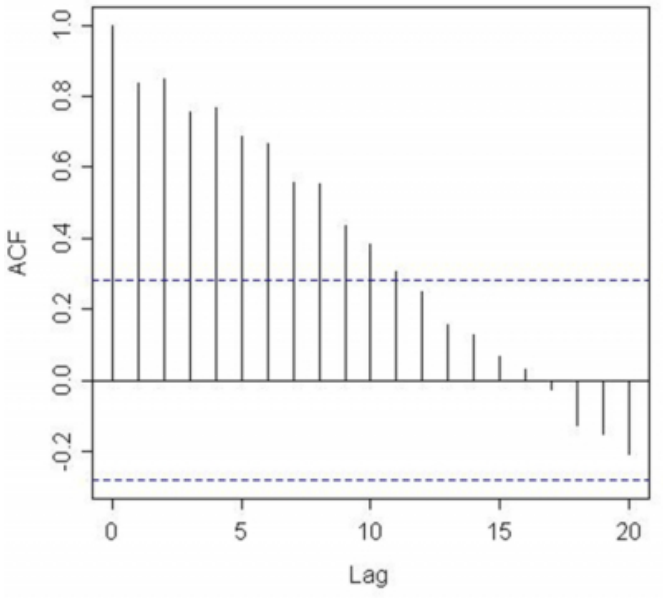
Figure 1: The Scatter plot of the SARS daily new case numbers versus the time from April 21 to June 7, 2003.



Around April 21, the case numbers vibrated but explicitly went down. A total of 48 case numbers will used in our analysis. To assess the degree of dependence in the data, we calculate the sample autocorrelation function (ACF) of the data and show it in Figure 2, in which the vertical bars show the corresponding sample ACF at lags 0,1,..,20 and dotted horizontal lines are the bounds ±1.96/√48. 1.96 is the .975 quantile of the standard normal distribution.

Figure 3

Figure 2: The Sample ACF of the SARS data



If the data are independent, we would expect roughly (20 * 0.05=)1 value to fall outside the bounds, while this plot shows us that the first ten of forty-eight values outside the bounds ±1.96/√48. This feature reflects that the observations are consecutively dependent and a time series analysis of these SARS data is worthwhile.

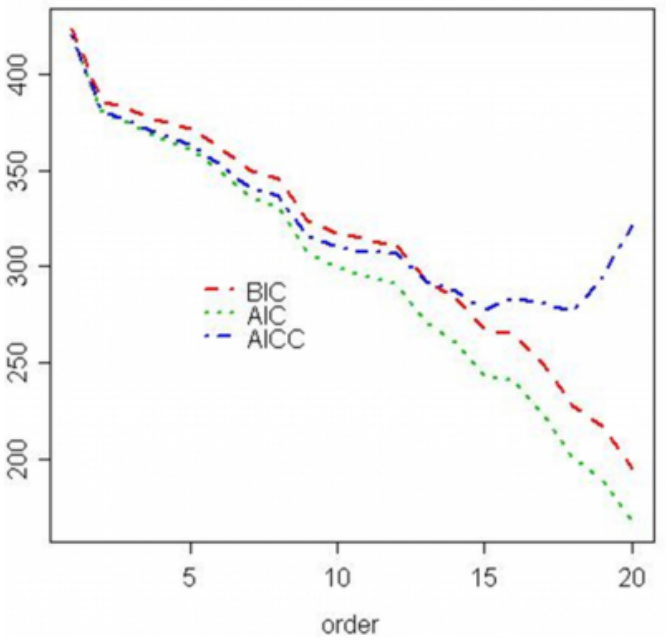
How to select an appropriate p is therefore important for fitting model. In traditional literature of time series, conventional approaches, such as AIC ($_{10}$) and BIC ($_{11}$), are widely used for variable selection, and can be easily implemented in common commercial software such as Splus, SAS, and Matlab. Their deficiency in small sample was pointed out by Hurvich and Tsai ($_{12}$). The authors showed that AIC may be drastically biased for time series, and developed a modified version, denoted AICC, which is nearly unbiased and provides better model choices than AIC and BIC in small samples. In this article, we use the AICC for selection of model order and estimated the auto-regression coefficients after identifying an appropriate order. We use the AR model to fit the data by first letting p = 20; The candidate models are those whose orders are 1, ...,p_o . For given an order, we fit the candidate model and calculate the AIC, BIC, and AICC.

RESULTS

Based on the rule that the smaller the criterion value, the better the model, it was found that both AIC and BIC attach minimum at p = 20, and indicated a tendency of overfitting model. While AICC select the best model of p = 15. We show the criterion values of AIC, BIC, and AICC in Figure 3. As p increases, AIC and BIC gradually decrease to p = 20. However, AICC reaches a global minimum at p = 15 and then increases.

Figure 4

Figure 3: The values of AIC, BIC, and AICC for different order



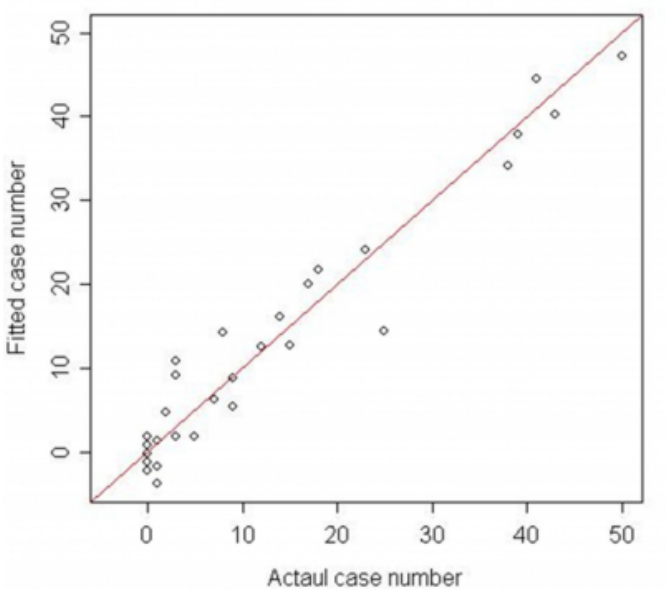
On a basis of the AICC criterion, we obtain the "best" model to the SARS data, of form

X(t)=0.484X(t-1)+0.377X(t-2)-0.294X(t-3)-0.02X(t-4)+0.599X(t-5)
-0.207X(t-6)-0.227X(t-7)+0.308X(t-8)+0.043X(t-9)-0.272X(t-10)
+0.142X(t-11)+0.101X(t-12)-0.344X(t-13)-0.045X(t-14)+0.2X(t-15).

The coefficient of determination R^2 , which is often as a convenient measure of how well the model describes the data with values of R^2 close to one indicating a good fit, is 0.96. The estimated SARS case numbers based on this model against the actual SARS case numbers are presented in Figure 4. Both the value of R^2 and fitted residuals indicate that the final model works well. A further observation found that the coefficients of several lower order terms are not significant. We remain these lower order terms in the model because the highest order term, /3is, is statistically significant.

Figure 5

Figure 4: The fitted versus actual SARS case numbers.



DISCUSSION

The derived model suggests that the number of average days of one SARS patient transmitting to other people is within 15 days. This result is partially because (i) most deaths occurred within a fortnight; (ii) Beijing is hot after June 7 and the SARS virus hardly survived; and (iii) the Chinese government made all possible efforts to control transmission. The result may be not true for other cities or other periods. However, this modeling procedure is still available to fit the SARS data from other cities/countries. Given the limited data available, the model may be not perfect. With more and better data, more realistic model may be feasible. In this article, we didn't consider patients' demographic factors such as age and gender, profession and family history. This needs a further investigation. Observing the scatter presented in Figure 1, one may see that the variance of is heteroscedastic. We fitted the data using weighted least squares approach and found that the results are similar to those given above.

ACKNOWLEDGMENT

This research was partially supported by two grants from the NIAID/NIH.

CORRESPONDENCE TO

Hua Liang, Ph.D. Department of Biostatistics and Computational Biology University of Rochester Medical Center 601 Elmwood Avenue, Box 630 Rochester, NY 14642 Fax: 585-273-1031 Email: hliang@bst.rochester.edu

References

1. Riley S, Fraser C, Donnelly C, Ghani AC, Abu-Raddad LJ, Hedley AA, et al. Transmission dynamics of the etiological agent of SARS in Hong Kong: Impact of Public Health Interventions. Science 2003;300:961-6.
2. Lipsitch M, Cohen T, Cooper B, Robins JM, Ma S, James L, et al. Transmission dynamics and control of severe acute respiratory syndrome. Science 2003; 300:1966-70.
3. Hsieh YH, Chen WS, Hsu SB. SARS outbreak, Taiwan, 2003. Emerg Infect Dis 2004; 10:201-6.
4. Zhou Q, Yan G. Severe acute respiratory syndrome epidemic in Asia. Emerg Infect Dis 2003;9:1608-10.
5. Richards FJ. A flexible growth function for empirical use. J Exper Botany. 1959; 10:290-300.
6. Hsieh YH, Cheng YS. Real-time forecast of multiphase outbreak. Emerg Infect Dis 2006;12:122-7.
7. Cauchemez S, Boelle PY, Donnelly CH, Ferguson NM, Thomas G, Leung GM, Hedley AA. Anderson RM, Valleron AA. Real-time estimates in early detection of SARS. Emerg Infect Dis 2006; 12:110-23.
8. Brockwell PJ, Davis RA. Introduction to time series and forecasting. New York: Springer 1996.
9. Cumulative number of reported probable cases of severe acute respiratory syndrome (SARS). Geneva: WHO, 2003. Available at http://www.who.int/csr/sars/country/2003_04_21/en/
10. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control 1974; AC-19:716-23.
11. Schwarz G. Estimating the dimension of a model. Ann Statl978;6: 461-4.
12. Hurvich CM, Tsai CL. Regression and time series model selection in small samples. Biometrika 1989; 76: 297-307.

Author Information

Hua Liang, Ph.D.
Department of Biostatistics and Computational Biology, University of Rochester Medical Center