RESEARCH REPORT

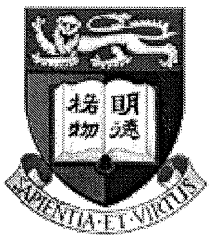
Serial No. 508

November 2013

AN ESSAY ON STATISTICS AND DYNAMICAL PHENOMENA TO HIGH SCHOOL CHILDREN

by

Howell Tong



THE UNIVERSITY OF HONG KONG
DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE

Pokfulam Road, Hong Kong. E-Mail: saas@hku.hk
Tel.: (852) 2859 2466/7 Fax: (852) 2858 9041

An Essay on Statistics and Dynamical Phenomena To High School Children

Howell Tong

London School of Economics & Political Science

1 Introduction

My friend who is an expert in statistical genomics told me the following story. Once at a dinner party, an attractive lady asked him, 'And what do you do for a living?' He replied, 'I model.' As my friend is a handsome man, the lady believed him and continued, 'What do you model?' 'Genes.' She then looked at him up and down and said, 'Mmmm, you must be very much in demand.' 'Yes very much so, especially after I helped discover a new culprit gene for a common childhood disease.' The lady looked puzzled.

I hope to convince you that Statistics (with a capital S) can be very exciting as the above story has hinted at. There are many exciting aspects of Statistics and I shall focus on just one small subset of them, specifically that part dealing with phenomena that change over time, in other words, dynamical events.

2 Chaos and Autoregressive Models

Consider the following doubling map (also known under many other names), where n denotes an integer:

$$y_0 = y > 0; \text{ for all } n \geq 0, \ y_{n+1} = 2y_n \mod 1. \tag{1}$$

In words, we double the previous y-value, throw away the integer part and keep only the decimal part. The map is also called a saw-tooth map because it is equivalent to the iteration defined by the saw-tooth function

$$f(y) = \begin{cases} 2y, & 0 \leq y < 0.5 \\ 2y - 1, & 0.5 \leq y < 1. \end{cases} \tag{2}$$

If we plot the graph of the function, we will see that it comprises two straight lines with a break (threshold) at 0.5. Such a function is nonlinear although piecewise

linear. Now, select any positive number as a starting point and iterate forward to get $y_1, y_2, \ldots$. We will quickly realize that the output shows no regularity no matter how long we run it, that is the output looks quite chaotic (hence we need chaos theory), almost indistinguishable from a random series. This seemingly puzzling feature of generating randomness from a wholly deterministic mechanism can be understood as soon as we realize that the function f is highly sensitive to initial values, in the sense that two initial values differing only in say the 8th decimal place will quickly diverge upon repeated applications of f. The famous French mathematician, Henry Poincaré (1854-1912), listed such sensitivity as one of the sources of randomness. Another curious thing about the doubling map is that if we think of n as designating time and run the map backward in time, we will discover that we need to introduce external randomness denoted by ε_n, namely

$$X_n = 0.5X_{n-1} + \varepsilon_n, n = 1, 2, \ldots, \tag{3}$$

where $X_n = y_{-n}$ and ε_n equals 0 and 0.5 with equal probability. The ε_n term appears because the inverse of f maps one point to two possible points equally likely. Here is yet another curious feature: we started with a deterministic nonlinear map (2) and finish with a random (a more fancy word is stochastic) linear equation (3).

Now, an equation like (3) defines what is called a time series model in Statistics, and it describes the dynamics over discrete time (n). In general, the ε_n can have a more general probability distribution, e.g. the Gaussian distribution (after the famous German mathematical genius Carl Friedrich Gauss (1977-1855)) over the set of real numbers. More generally, an equation of the form

$$X_n = a_1X_{n-1} + \ldots + a_pX_{n-p} + \varepsilon_n \tag{4}$$

is called a (linear) autoregressive (AR for short) model, which was invented by the British statistician Udny Yule in 1927 when he studied the annual sunspot numbers. Here, the a_js are the defining parameters to be estimated from observations.

The amazing thing is that the AR model, either in its original form or its multi-dimensional or piecewise linear generalizations, has been used in diverse fields as we shall see in the following examples from real applications.

3 Some Real Examples

- US hog data

  Professors George Box and George Tiao of USA analyzed the above data (See fig. 1.) by using a 5-dimesional AR model and concluded that

$$\frac{H_p H_s}{(R_p R_s)^{0.75} W^{0.50}}$$

  is approximately independently distributed about a fixed mean. Here, H_p, H_s, R_p, R_s, W denote hog price, hog supply, corn price, corn supply and farmer's wage, respectively. In other words, using real observations, they have vindicated that {return to the farmer} / {farmer's expenditure} follows a stable economic law! It is a remarkable historical fact that such a simple empirical relationship had never been discovered for the above classic data set, until the two statisticians invented a new time series method based on another statistical technique called canonical correlations in 1977. In 1987, two American econometricians re-discovered special cases of the Box-Tiao method and were hugely rewarded-a Nobel memorial prize in Economics in 2003. (I think of Econometrics as the product of an Economics father and a Statistics mother.)

- Economics and Finance

  The piecewise linear deterministic model (2) has its stochastic cousin by adding ε_n. The result is commonly called the threshold AR (or TAR) model first introduced by the author in 1978. In a slightly more general form,

$$X_n = \begin{cases} a_0 + a_1 X_{n-1} + \ldots + a_p X_{n-p} + \varepsilon_n, & X_{n-d} < r \\ b_0 + b_1 X_{n-1} + \ldots + b_p X_{n-q} + c\varepsilon_n, & X_{n-d} \geq r. \end{cases} \tag{5}$$

  Here, r is called the threshold parameter and has to be estimated along with a_js, b_js, and p, q, d, c and variance of ε_n, from the observed data. The model has made an enormous impact in economics and finance. In 2011, the US econometrician, Professor Bruce Hansen, published a comprehensive review of 75 papers related to TAR models and published in the econometrics and economics literature, many of which are themselves highly cited. In empirical economics, he has listed output growth, forecasting, interest

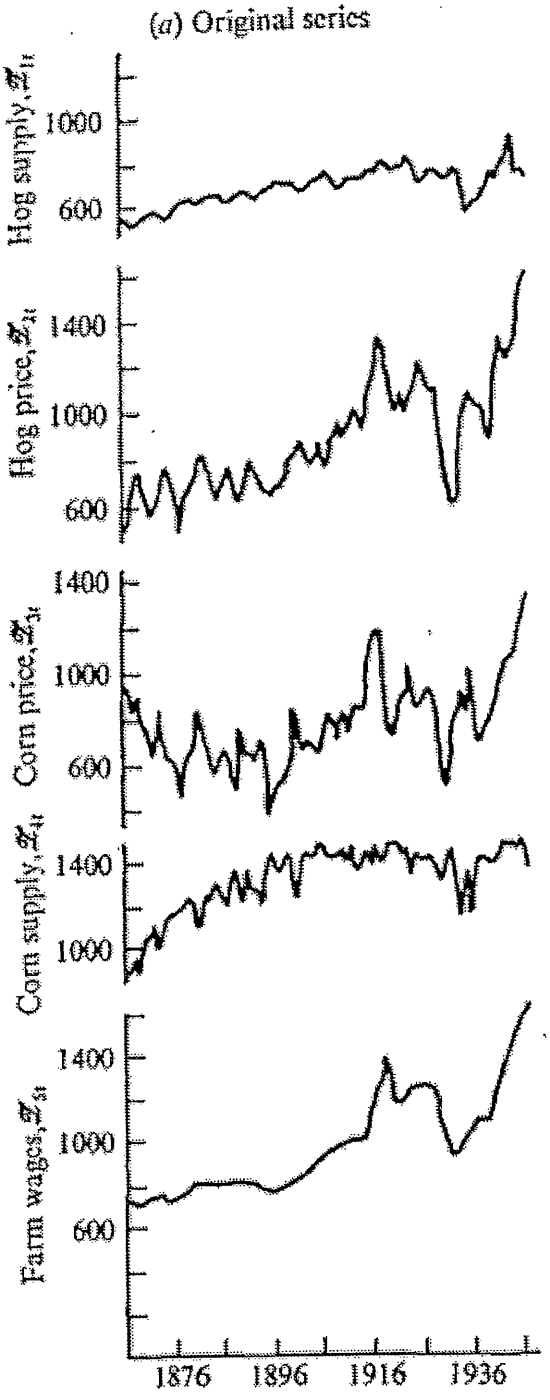

(a) Original series
Hog supply, x_{1t}
1000
600
Hog price, x_{2t}
1400
1000
600
Corn price, x_{3t}
1400
1000
600
Corn supply, x_{4t}
1400
1000
Farm wages, x_{5t}
1400
1000
600
1876
1896
1916
1936

Figure 1

rates, prices, stock returns and exchange rates. As typical examples, the TAR model has been used to model aggregate output as measured by GNP growth rates. For instance, the US GNP has been shown to be subject to floor and ceiling effects. The use of the TAR model to study the relationship between long and short interest rates has helped to reveal the strong asymmetric response of interest rate changes to the spread between the long and short rates. Economic arbitrage requires that the prices of related goods should move in tendon, but transaction costs are shown to be such that only deviations above a certain threshold will have an effect on price movements.

- Plagues in Kazakhstan

  Early this century, the statistician Kung-Sik Chan and his then doctoral student, Noelle Samia, worked with a team of biologists/epidemiologists led by Professor Nils Chr. Stenseth, formerly President of the Norwegian Academy of Science and Letters, on an extensive scale study of plague epidemics in Central Asia. The project was funded by numerous organisations including the European Union, the Welcome Trust, the National Science Foundation of the USA, and others. Concerning the bacterium Yersinia pestis that causes bubonic plague, they concluded in their report in the Proceedings of the National Academy of Sciences (USA) in 2006 that *"Y. pestis prevalence in gerbils increases with warmer springs and wetter summers... Climatic conditions favouring plague apparently existed in Central Asia at the onset of the Black Death as well as when the most recent plague pandemic arose in the same region, and they are expected to continue or become more favorable as a result of climate change."* This conclusion is based on a modified TAR model that the participating statisticians developed and fitted from observed data. In the model to be given below, $N_{t,\ell}$ is the number of great gerbils examined at time t in "large square" ℓ (The large squares are the result of dividing Kazakhstan into non-overlapping $40 \times 40km^2$ squares.) The number of great gerbils testing positive under a bacteriological test does not follow a Gaussian distribution. Instead it is more likely to follow a binomial distribution and hence we modify the TAR model slightly while retaining the piecewise linear structure. The binomial distribution model has parameters $(N_{t,\ell}, P_{t,\ell})$, where if t is a spring the true

prevalence rate $P_{t,\ell} = 0$ when the lag-d^s occupancy (i.e. how many burrows are occupied by the great gerbils d^s units of time ago), namely $X_{t-d^s,\ell}$, is below the spring threshold r_ℓ^s (the superscript s signifies spring) but otherwise follows a logistic regression model shown below. A similar specification applies to the fall data. We should not worry about the detail as long as we get the message that the dynamics underlying bacterium Yersinia pestis operates in 4 regimes depending on the season (spring or fall) and the lag-d^s occupancy rate (above or below the seasonal threshold).

$$P_{t,\ell} = \begin{cases} \begin{cases} 0, & \text{if } X_{t-d^s,\ell} < r_\ell^s \text{ and } t \text{ is a spring} \\ \text{logit}^{-1}\{(\beta_0^s + b_{0,\ell}^s) + (\beta_1^s + b_{1,\ell}^s)T_{sp,t} + b_{2,\ell}^s R_{sp,t} + \epsilon_{t,\ell}\}, & \text{if } X_{t-d^s,\ell} \geq r_\ell^s \text{ and } t \text{ is a spring;} \end{cases} \\ \begin{cases} 0, & \text{if } X_{t-d^f,\ell} < r_\ell^f \text{ and } t \text{ is a fall} \\ \text{logit}^{-1}\{(\beta_0^f + b_{0,\ell}^f) + \beta_1^f R_{su,t} + \beta_2^f X_{t-1/2,\ell} + \epsilon_{t,\ell}\}, & \text{if } X_{t-d^f,\ell} \geq r_\ell^f \text{ and } t \text{ is a fall;} \end{cases} \end{cases} \tag{6}$$

Here, the superscript f signifies fall, X denotes the great gerbil occupancy, $T_{sp,t}$ is the spring temperature, $R_{sp,t}$ is the log spring rainfall, and $R_{su,t}$ is the log summer rainfall. The βs are tuning parameters to be estimated from the observed data. The logistic function $\text{logit}^{-1}(x) = \frac{1}{1+\exp(-x)}$ is just a mathematical trick to transform the binomial model setting into a TAR model format.

4 Looking to the Future

The present information age poses many exciting challenges to Statistics. Data collection is so fast and plentiful that suddenly we humans find ourselves flooded with data. Let us start with a simple illustration. We saw the example of the hog data at one site. Suppose we have similar data at say 100 sites. Does a similar empirical law hold for all of them or what? Similar panel time series can and do occur in many situations, e.g. a panel of stock price time series across different

stock markets, a panel of death rates due to a particular infectious disease at different locations in the world, and so on. Many interesting questions then need to be answered, e.g. are the different stock markets equally volatile or do they cluster in some way? Is the infectious disease spreading? It is clear that new statistical methodologies will be necessary to cope with the new challenges and that is where fresh thinking is needed. To do so, we may have to re-examine existing statistical methodologies and even our philosophy. Statistics as a scientific discipline has a rather recent history, not much longer than 100 years. Over these 100 years or so, the discipline has been dominated by the concept called likelihood. It is founded on the assumption that a true model is known and the only thing unknown are its tuning parameters. The observed data are then used to enable us to estimate them. Two schools of thought, the frequentist school and the Bayesian school, have been the pillars of Statistics. Despite their sometimes heated and colourful polemics, they share the common ground of likelihood. What if the true model does not exist? What if we know that the model given to us by our scientist friend is wrong but it is the best available? I think these are legitimate questions that we need to address. Some opening shots have been fired by people like Professor Laurie Davies in Germany, Professor Yingcun Xia in Singapore, the author and others. Professor Rudolf Beran has put it very well in his discussion of Laurie Davies's paper, 'The *Kepler Challenge* for Statistics is to develop a general compression or pattern recognition algorithm that has cogent theoretical properties, that works well in case studies, and that, when applied to data like Brahe's, yields Kepler's three laws. The challenge demands a rethinking of Statistics. Mathematics offers powerful languages besides probability theory.' The information age has brought upon us the Kepler Challenge; to bring about a revolution of Statistics young brains like yours are needed. Come along and join the fun!

5 Further Readings

1. Box, G.E.P. and Tiao, G.C. (1977). A canonical analysis of multiple time series. *Biometrika*, **64**, 355-365.

2. Chan, K-S. and Tong, H. (2001). *Chao: A statistical perspective.* Springer-

Verlag.

3. Davies, P.L. (2008). Approximating data. *J. Korean Stat. Soc.*, **37**, 191-211.

4. Hansen, B. (2011). Threshold autoregression in economics. *Statistics and Its Interface*, **4**, 123-128.

5. Stenseth, N.C., Samia, N.I., Vilugrein, H., Kausrud, K.L., Begon, M. Davis, S., Leirs, H., Dubyanskiy, V.M., Esper, J., Ageyev, V.S., Klassovskiy, N.L., Pole, S.B. and Chan, K.S. (2006). Plague dynamics are driven by climate variation. *Proc. Natl. Acad. Sci. USA*, **103—, 13110-13115.**

6. Tong, H. (1978). On a threshold model. *Pattern recognition and signal processing.* NATO ASI Series E: Applied Sc. **29**, ed. C.H.Chen. The Netherlands: Sijthoff & Noordhoff, 577-586.

7. Xia, Y. and Tong, H. (2011). Feature matching in time series modeling. *Statisticl Science*, **26**, 21-46.

Current Departmental Research Reports

Serial No.	Date	Research Report Title	Author(s)
496	Jul-12	A geometric process maintenance model and optimal policy	Yeh Lam
497	Oct-12	Test for homogeneity in gamma mixture models using likelihood ratio	T.S.T. Wong and W.K. Li
498	Oct-12	Buffered threshold autoregressive time series models	Guodong Li, Bo Guan, Wai Keung Li and Philip L.H. Yu
499	Nov-12	On a rescaled fractionally integrated GARCH model	Muyi Li, Wai Keung Li and Guodong Li
500	Nov-12	On the sphericity test with large-dimensional observations	Qinwen Wang and Jianfeng Yao
501	Nov-12	On generalized expectation based estimation of a population spectral distribution from high-dimensional data	Weiming Li and Jianfeng Yao
502	Nov-12	A note on multivariate CUSUM charts	Hualong Yang and Jianfeng Yao
503	Apr-13	A geometric process credibility model	Yeh Lam
504	Jun-13	On mixture double autoregressive time series models	Zhao Liu and Guodong Li
505	Jun-13	A simple formula for mixing estimators with different convergence rates	Stephen M.S. Lee and Mehdi Soleymani
506	Nov-13	On singular value distribution of large-dimensional autocovariance matrices	Zeng Li, Guangming Pan and Jianfeng Yao
507	Nov-13	Threshold models in time series analysis-some reflections	Howell Tong
508	Nov-13	An essay on statistics and dynamical phenomena to high school children	Howell Tong

❖❖❖ *The complete listing can be found at* ❖❖❖

http://www.saasweb.hku.hk/research/staff_research_report.php

Requests for off prints may be sent to saas@hku.hk by e-mail