Batanero, C (Ed.), Training Researchers in the Use of Statistics, 129-137.

Printed in Granada, Spain.

TOSHIRO SHIMADA

PRECAUTION AGAINST ERRORS IN USING STOCHASTIC SOFTWARE

There are many statistics packages available that make it easy to perform stochastic procedures. Therefore, today's students may think they can handle their data processing needs, and obtain stochastic results simply by clicking a PC button. However, without being aware of it, they can make many mistakes, and treat their data incorrectly.

In this paper we compare generalised logistic curves with simple logistic curves and explain their characteristics to help students and researchers avoid mistakes.

1. INTRODUCTION

Many researchers are using information software which they have created themselves, and therefore they do not make mistakes in analysing their data. Moreover, there are many statistics packages available that help them to perform stochastic procedures although the standard options in these packages are not always the best for each specific research problem.

However, some researchers might believe they can solve their data analysis needs, and obtain statistics results simply by clicking a PC button. Without being aware of it, they can make mistakes, and analyse their data incorrectly, which may have serious consequences for their research.

2. EXAMPLES

Figure 1. TSE: Moving Average of Daily Volume on the Tokyo Stock Exchange (Sep.-Nov., 1956, Millions of Shares)

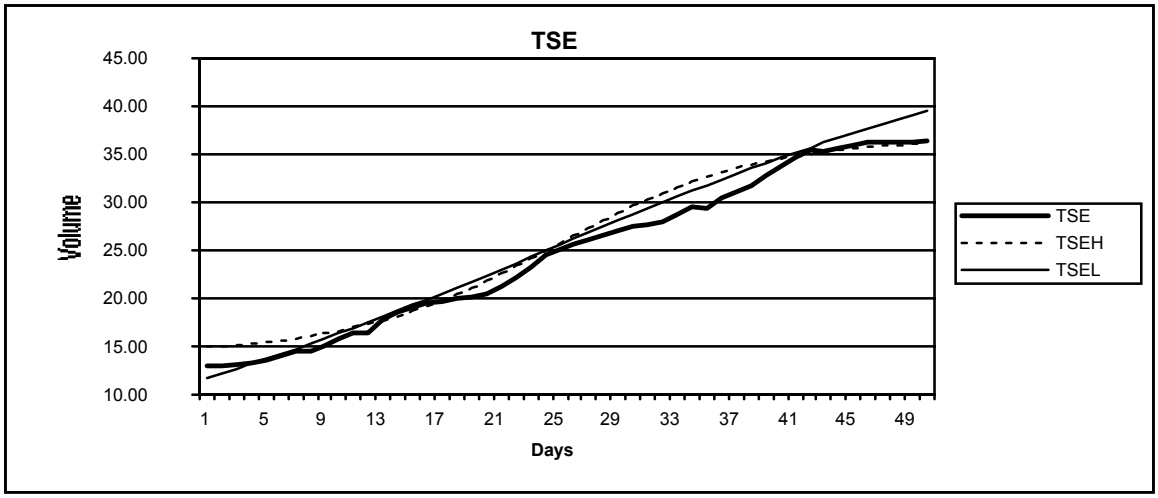
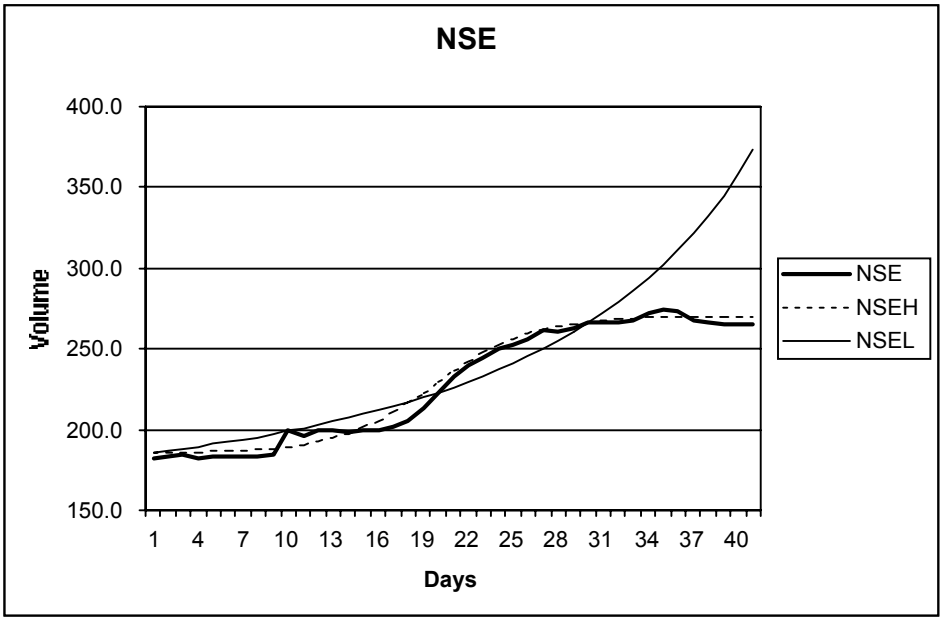


Figure 2: NSE: Moving Average of Daily Volume for the Dow Jones 30 Industrial Stocks of the New York Stock Exchange, (Dec. 26, 1956, Millions of Shares)



In this paper we will be discussing some examples of incorrect use of software in the particular case of logistic curves. Figures 1 and 2 present the two examples we are going to discuss.

3. GENERALIZED LOGISTIC FUNCTION

Graphs of the TSE and NSE data have apparent asymptotes which are different from the x axis. These are called generalised logistic curves. These curves are expressed by the following equation,

$$y = c + \frac{k}{1 + me^{-ax}} \tag{3.1}$$

The common logistic equation is:

$$y1 = \frac{k}{1 + me^{-ax}} \tag{3.2}$$

These equations are very simple, but both are non linear. Many teachers can easily handle *y1* type curves, but the (3.1) type curve is a little different and is rarely treated in stochastic packages. We will determine the four parameters c, m, a and k of (3.1) for TSE and NSE. Since (3.1) is non linear we will first try to linearise this equation. The expression

$$F(x,y;c,m,a,k) = y - c - \frac{k}{1 + me^{-ax}} = 0 \tag{3.3}$$

will represent the estimated curve.

We assume that $c_0, m_0, a_0,$ and k_0 are the approximations of the parameters. (The method of determining them will be discussed later.) The differences between the estimated values and the assumed approximations of the parameters, so-called residuals, are represented as (3.4).

$$C = c_0 - c,\ M = m_0 - m,\ A = a_0 - a,\ K = k_0 - k \tag{3.4}$$

Let Y be the daily volume for n days, and y the estimated daily volume. Then (3.3) may be rewritten as follows:

$$\begin{aligned} &F(x_i, y_i; c, m, a, k) = y_i - c - \frac{k}{1 + me^{-axi}} = 0 \\ &(i = 1, 2, \text{K}\ , n) \end{aligned} \tag{3.5}$$

The residual of y is given by (3.6).

$$Vy = Y_i - y_i \tag{3.6}$$

From (3.4) and (3.6) we obtain:

$$y_i = Y_i - Vy,\ c = c_0 - C,\ m = m_0 - M,\ a = a_0 - A,\ k = k_0 - K \tag{3.7}$$

When (3.7) is substituted into (3.5), the equation (3.5) becomes:

$$\begin{aligned} &F(x_i, Y_i - Vy, c_0 - C, m_0 - M, a_0 - A, k_0 - K) = 0 \\ &(i = 1, 2, \text{K}\ , n) \end{aligned} \tag{3.8}$$

By expanding (3.8) using Taylor's series and by omitting the terms containing powers of the residuals higher than the second degree, we obtain:

$$\begin{aligned} &F_0 - VyFy - FcC - FmM - FaA - FkK = 0 \\ &(i = 1, 2, \text{K}\ , n) \end{aligned} \tag{3.9}$$

where

$$F_0 = F(x_i, Y_i; c_0, m_0, a_0, k_0) \tag{3.10}$$

and Fy, Fc, etc, are the values of partial derivatives of the function F for the values (x_i,Y_i; c_0, m_0, a_0, k_0) (i=1,2,...,n), namely,

$$\begin{aligned} &Fy = 1,\ Fc = -1 \qquad Fm = \frac{k_0}{(1 + m_0 e^{-a_0 x})^2} \\ &Fa = \frac{k_0 m_0 x e^{-a_0 x}}{1 + me^{-a_0 x}} \qquad Fk = \frac{-1}{(1 + m_0 e^{-a_0 x})^2} \end{aligned} \tag{3.11}$$

Expressions (3.9) are the linearised equations, so we may solve the problem that the sum of the squares of the residuals is $\sum Vy_i^2$ to be a minimum under auxiliary conditions (3.9). Then we can find the normal equations for fitting a generalised logistic curve (3.1)[(2)]. They are

$$\left.\begin{aligned} &[cc]C+[cm]M+[ca]A+[ck]K=[co]\\ &[mc]C+[mm]M+[ma]A+[mk]K=[mo]\\ &[ac]C+[am]M+[aa]A+[ak]K=[ao]\\ &[kc]C+[km]m+[ka]A+[kk]K=[ko] \end{aligned}\right\} \tag{3.12}$$

where [cc]=∑FcFc, [cm]=[mc]=∑FcFm, [co]=∑FcFo, etc.

By solving the normal equations (3.12) and determining the values of C, M, A and K, we can estimate the parameters c, m, a, and k using (3.7). Then the problem of the trend line can be solved.

4. THE APPROXIMATIONS OF THE PARAMETERS

After examining Table 1 (given in the Appendix 2) and Fig. 1 we select $c_0 = -12.89$, and write

$$\eta = Y - c_0 = \frac{k_0}{1 + m_0 e^{-a_0 x}} \tag{4.1}$$

Then we obtain $a_0, m_0,$ and k_0, which will satisfy (3.1) for the three values of x (10, 25, and 40) as follows:

$$\left.\begin{aligned} 1+m_0e^{-10a_0} &= \frac{k_0}{\eta_{10}}\\ 1+m_0e^{-25a_0} &= \frac{k_0}{\eta_{25}}\\ 1+m_0e^{-40a_0} &= \frac{k_0}{\eta_{40}} \end{aligned}\right\} \tag{4.2}$$

These equations seem complicated, but if we choose the above mentioned values of x, (4.2) can be solved without difficulty if we write

$$z = e^{-5a_0}$$

then

$$z = \left(\frac{\eta_{10}(\eta_{40}-\eta_{25})}{\eta_{40}(\eta_{25}-\eta_{10})}\right)^{1/3} = 0.5411 \tag{4.3}$$

And hence

$$a_0 = 0.1228$$

Using (4.2), the value of m_0 and k_0 can be easily obtained. Thus,

$$c_0 = 12.89,\ m_0 = 21.10,\ a_0 = 0.1228,\ k_0 = 25.13 \tag{4.4}$$

Using these values, we can find the normal equations (3.12). By solving these we get residuals,

$$\text{C=-1.369, M=-15.24, A=-0.02319, K=2.704} \tag{4.5}$$

From (3.7), (4.4) and (4.5),

$$\text{c=14.26, m=36.34, a=0.1460, k=22.43} \tag{4.6}$$

and hence,

$$TSEH = 14.26 + \frac{22.43}{1 + 36.34e^{-0.1460x}} \tag{4.7}$$

(Sep. 21, 1956; x units: 1 day; y is daily volume in millions of shares.)

This curve (4.7) is shown in Fig. 1 as TSEH.

Many years ago I used FORTRAN and a simultaneous equations package, and easily obtained the above results. My students also understood this procedure without problems.

The generalised logistic curve of NSE shown in Fig. 2 was obtained by the same procedure as that used for TSE. In this case, the approximations of the parameters were

$$c_0 = 183.5,\ m_0 = 362.6,\ a_0 = 0.3085,\ k_0 = 87.96 \tag{4.8}$$

Residuals are

$$\text{C=-1.93, M=-58.7, A=0.00278, K=2.866} \tag{4.9}$$

Hence the estimation of NSE becomes

$$NSEH = 185.4 + \frac{85.09}{1 + 362.6e^{-0.3035x}} \tag{4.10}$$

5. COMMON LOGISTIC CURVES

TSEL In Fig. 1 and NSEL in Fig. 2 are the common logistic curves of TSE and NSE. Their equations are as follows:

$$TSEL = \frac{50.07}{1 + 3.420e^{-0.05095x}} \tag{4.11}$$

$$NSEL = \frac{151.46}{1 - 0.1784\, e^{-0.02937\, x}} \tag{4.12}$$

These parameters are rough values obtained from the procedures as the approximations of the generalised logistic curves. We can see good fitting of TSEH and NSEH to the original data, but TSEL and NSEL are unsatisfactory. In particular, NSE has high values, so after x=30, NSEH goes up very high.

6. CONCLUSION

In this paper we compared generalised logistic curves with simple ones and showed their characteristics to help students to avoid mistakes. I once heard a research report at a yearly meeting of an academic society, given by a researcher from a large Japanese computer maker. He expected the Christmas sales trend to follow a common logistic curve, but the data were clearly of a generalised logistic curve. I asked him about this, and he answered that he did not know the details of the software he was using, but he simply accepted the results.

We must be very careful when using statistical software and teach basic principles to beginners to avoid these kinds of mistakes.

APPENDIX 1.

Table 1. Sample Data Used for Figures 1 & 2

x	TSE	TSEH	TSEL	NSE	NSEH	NSEL
1	13.02	14.95	11.78	182.2	185.6	185.5
2	12.98	15.05	12.24	183.9	185.7	186.8
3	13.07	15.17	12.72	184.3	185.9	188.1
4	13.24	15.31	13.21	182.0	186.0	189.4
5	13.53	15.47	13.71	183.9	186.3	190.9
6	14.02	15.65	14.22	183.0	186.6	192.3
7	14.60	15.85	14.75	183.9	187.0	193.9
8	14.59	16.08	15.28	183.4	187.6	195.5
9	14.94	16.34	15.83	184.3	188.4	197.3
10	15.76	16.63	16.39	199.1	189.4	199.1
11	16.39	16.96	16.95	196.1	190.8	200.9
12	16.37	17.33	17.53	199.3	192.6	202.9
13	17.76	17.73	18.11	199.5	194.9	205.0
14	18.52	18.19	18.71	198.1	197.8	207.2
15	19.12	18.68	19.31	199.5	201.4	209.5
16	19.52	19.22	19.92	199.9	205.8	211.9
17	19.66	19.81	20.53	201.3	210.9	214.5
18	19.94	20.44	21.13	205.8	216.7	217.2
19	20.17	21.12	21.77	213.4	223.0	220.0
20	20.54	21.83	22.40	223.0	229.4	223.0

21	21.20	22.58	23.04	232.8	235.9	226.2
22	22.19	23.36	23.67	239.8	241.9	229.6
23	23.28	24.16	24.31	245.0	247.4	233.2
24	24.46	24.97	24.94	249.8	252.2	237.0
25	25.17	25.79	25.58	252.4	256.2	241.0
26	25.58	26.60	26.22	255.5	259.4	245.4
27	26.08	27.41	26.85	261.3	262.1	250.0
28	26.52	28.19	27.49	261.0	264.1	254.9
29	27.00	28.95	28.12	262.8	265.7	260.2
30	27.47	29.67	28.74	265.8	266.9	265.9
31	27.66	30.35	29.36	266.2	267.8	272.1
32	28.00	30.99	29.98	266.6	268.5	278.7
33	28.69	31.59	30.59	268.0	269.0	285.9
34	29.46	32.14	31.19	272.4	269.4	293.6
35	29.44	32.65	31.79	275.0	269.6	302.1
36	30.49	33.11	32.38	273.5	269.9	311.3
37	31.03	33.53	32.95	267.2	270.0	321.4
38	31.74	33.90	33.52	266.1	270.1	332.6
39	32.80	34.24	34.08	265.2	270.2	344.8
40	33.78	34.54	34.63	264.9	270.3	358.5
41	34.62	34.81	35.17	265.2	270.3	373.7
42	35.40	35.04	35.70			
43	35.35	35.25	36.21			
44	35.58	35.44	36.72			
45	35.91	35.60	37.21			
46	36.21	35.74	37.69			
47	36.30	35.86	38.16			
48	36.31	35.97	38.62			
49	36.25	36.07	39.04			
50	36.33	36.15	39.49			

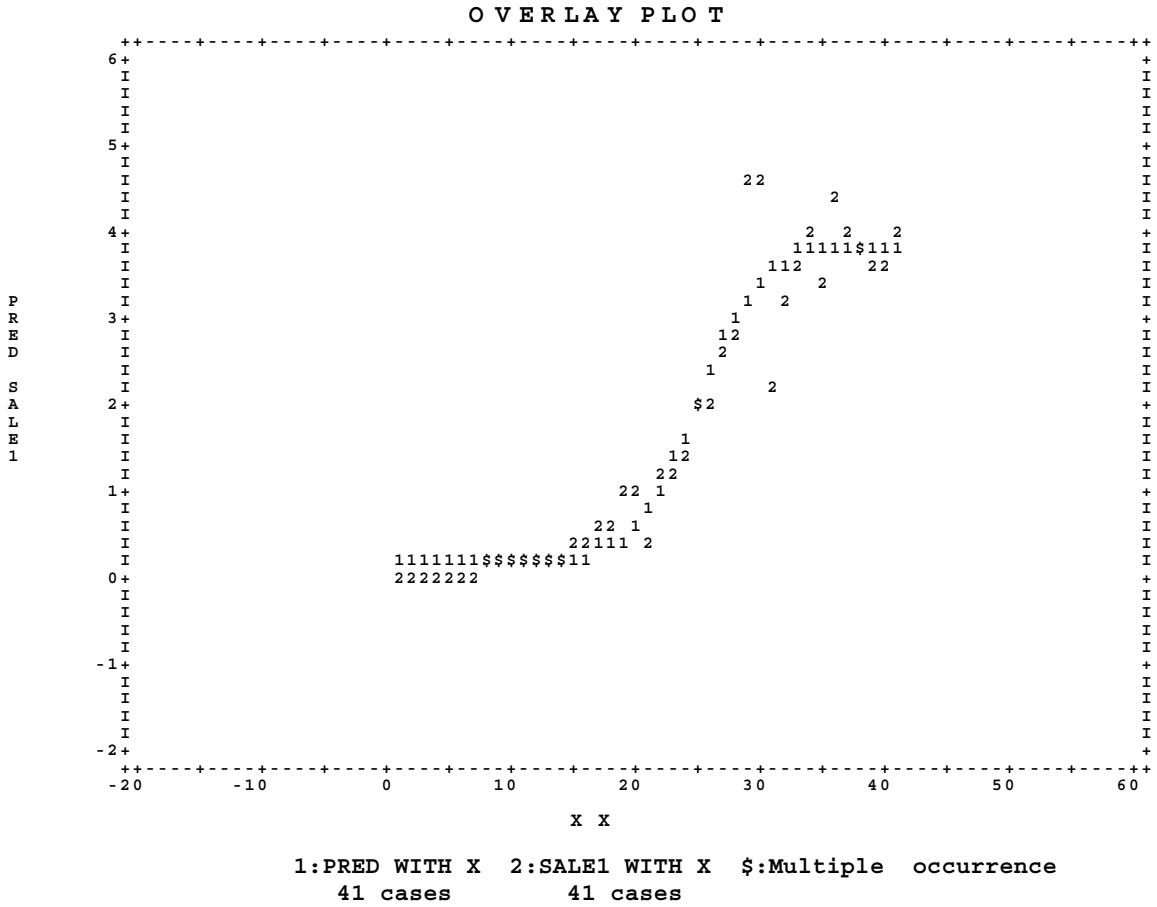
APPENDIX 2.

Another example of the generalised logistic equation uses summer sales of a Japanese department store treated with SPSS version 4. This simple procedure is similar to those of TSE and NSE.

```
//BB00001S JOB (SIMADA,200),'SIMADA',CLASS=S,MSGCLASS=1
// EXEC SPSSX
//SYSIN DD *
RUN NAME       SUMMER SALES (NONLINEAR REGRESSION)
DATA LIST FREE/ X SALE1
BEGIN DATA
 1 .0610  2 .0650  3 .0740  4 .0840  5 .0920
 6 .0690  7 .0270  8 .1740  9 .1690 10 .1950
11 .2190 12 .2510 13 .2640 14 .1450 15 .4860
16 .4960 17 .5480 18 .5500 19 .9570 20 .9620
```

```
21 .3280 22 1.189 23 1.203 24 1.421 25 2.096
26 2.037 27 2.538 28 2.826 29 4.560 30 4.560
31 2.265 32 3.109 33 3.609 34 4.055 35 3.362
36 4.471 37 3.914 38 3.778 39 3.583 40 3.572
41 3.947
END DATA
MODEL PROGRAM  C=0.06 K=113.05 M=681.1 A=0.1960
COMPUTE        PRED=C+K/(1+M*EXP(-A*X))
DERIVATIVES
COMPUTE        D.C=1
COMPUTE        D.M=-K*EXP(-A*X)/(1+M*EXP(-A*X))**2
COMPUTE        D.A=K*M*X*EXP(-A*X)/(1+M*EXP(-A*X))**2
COMPUTE        D.K=1/(1+M*EXP(-A*X))
NLR           SALE1 WITH X/ SAVE PRED
PLOT FORMAT=OVERLAY / PLOT=PRED WITH X; SALE1 WITH X
FINISH
```

Figure 3. Summer Sales (Nonlinear Regression)



The first three steps are Job instructions which will be different for each computer system. The next steps comprise the SPSS data and instructions program for the generalised logistic equation. The variables are x and sales, which are sales indices of daily summer sales for 41 days.

The MODEL PROGRAM line gives the initial parameter values, which I assumed.

The first COMPUTE line shows the generalised logistic equation.

NLR means the NONLINEAR REGRESSION program of SPSS.

The Results of the Summer Sales Regression.

Iteration was stopped after 51 model evaluations. The parameter estimates were c=.2004, k=3.634, m=29800, a=.4127. Fig. 3 shows the original data and the regression results for the summer sales.

REFERENCES

Biehler, R (1997). Software for learning and for doing statistics, *International Statistical Review,* 65, 2, 167-189.

Shimada, T. (1961). Short term trend lines of daily volume of trading in stocks. *Journal of Commercial College of Meiji University,* 44(7), 1-21.

SOFTWARE

(1) SAS, SAS Institute Inc., Cary, NC, USA
(2) S and Splus, MathSoft, Inc.
(3) SPSS, SPSSX User's Guide, Second Edition, 1985, SPSS Inc.,

Toshiro Shimada
Professor Emeritus Meiji University,
8-9-3-314, Kokuryo-machi Chofu-shi,
Tokyo, 182-0022, Japan.
E-mail: simfuji0@mtj.biglobe.ne.jp