When are customers in the market? Applying survival analysis to marketing challenges

Received (in revised form): 21st September, 2001

Tim Drye

is Managing Director of DataTalk (Statistical Solutions) Ltd, a statistical analysis company specialising in marketing consultancy. He is known for his innovative application of statistical techniques from other industries to provide a new insight into marketing data. He was previously Head of Statistics and Computing at Anglia Polytechnic.

Graham Wetherill

is a Statistical Consultant with over 15 years' experience of the pharmaceutical industry. He was previously a lecturer at both the Liverpool John Moores and Aberdeen Universities delivering specialised courses in generalised linear models and survival analysis. He is a regular contributor to professional journals.

Alison Pinnock

is a Statistical Analyst with DataTalk (Statistical Solutions) Ltd. Having achieved an honours degree in Mathematics at St Andrews University, Alison spent five years in the water industry developing hydraulic system models and applying statistical techniques to improve a wide range of operational issues.

Abstract Identifying when customers want an organisation's products and services has been high on the list of priorities for a significant length of time. Many data are bought and sold to identify key times in a customer's lifecycle. The mantra of marketing is that the right message needs to be communicated to the right person, at the right time.

Yet in this situation, the techniques of survival analysis, widely used in non-marketing contexts to identify the time to occurrence of critical events, is hardly used. The authors outline how survival analysis has been used in a number of case studies and illustrate the dangers of not using the insights contained within survival analysis when trying to understand time-based issues.

Customer lifetime analysis uses customer information like the typical time between purchases to predict the potential value of a customer. Simple means and summary statistics can be seriously biased and misleading for this type of data. Survival analysis is a good, sound and flexible tool to analyse length of time-to-next-purchase and cancellation data. As a result the output from survival analysis can be fed into various forms of customer lifetime analysis to provide more reliable and accurate conclusions. This can lead to a better understanding of customer lifecycles and value and, as a result, more efficient use of the available budgets.

Dr Tim Drye
Managing Director, DataTalk Ltd, 27a New Street, St Neots, Cambridgeshire, PE19 1AE, UK.

Tel: +44 (0)1480 381 352; Fax: +44 (0)1480 356 186; e-mail: timd@DataTalkOnline.co.uk

INTRODUCTION

Regular customers make up a significant proportion of business for most organisations. They not only provide financial support for the organisation, but also generate exposure of the brand and act as advocates of the products. Acquiring, retaining and developing these customers is recognised as an important focus for business development.

Early identification of potential lapsed customers can not only save money, time and effort on unwanted repeated communications, but can also enable a more suitable mailing/marketing approach to encourage decisions.

Sending the correct type of information to customers is also important. Customers may get put off by repeated requests for help that are not practical for them. In order to improve the targeting of communications with customers, a better understanding of the customer base is vital. This should be based on actual information rather than 'perceived general knowledge' which can sometimes be quite wrong. Indeed, analyses can sometimes be most useful in showing the shortcomings of received 'wisdom' and current strategies.

Analyses may also shed light on strategy. If some customers are dissatisfied, or simply do not feel well served or well informed, then this might not show until customers start to lapse. Analyses may point to areas where strategy/activity is weak before the situation becomes serious, which often makes the problem much easier to solve. Changes in strategy may go deeper than simply changing the style, method or frequency of communications to individual customers.

RENEWAL SCHEMES

A simpler situation is when an organisation has a renewal scheme, for example within the general insurance sector. A key indicator is the length of time that a customer is expected to remain in contact beyond their current expiry date. (This is called the expected residual lifetime). Identifying those customers most likely to stay the longest may help to focus recruitment campaigns towards those customers who are most likely to maintain their interest, support and membership for the longest, and as a result have a significantly higher lifetime value.

Another aspect is the propensity for an existing member to lapse. Probabilities can be estimated in order to identify those customers who are most likely to lapse. Different strategies can then be tested to increase the numbers of customers maintaining their membership and/or to reduce the expenditure on customers who have not, and will not, maintain their membership. Critical events can be identified to track loyalty scheme membership. These may be based upon the regular use of the scheme or the actual cancellation of an automatic payment.

CASE STUDIES

This paper details two case studies. Each case study considers the length of membership, that is the time until the customer terminates their membership/no longer maintains their membership, and the time between successive purchases. The first is taken from a consumer marketing context and shows behaviour relating to a small range of available services. There is a period when the customer does not require these services as they are in the process of consumption, for example a period following the purchase of a car when the new product is being enjoyed. The second is taken from a business-to-business context, but also shows the type of behaviour typical with organisations such as mail order, with a broad range of products on offer to a customer.

SIMPLE MEANS?

Simple means are biased (giving an unfair picture) for time to event data since, time to event information will not be available for all cases. (For some

customers, the event in question will not have happened within the time period of the data, which is why the analysis is being done.) In the second case study, the time to next payment, by definition, will never be available for lapsed customers.

The bias in simple means can be quite serious, leading to large differences in estimates. In a recent example, the simple mean estimated the average membership period to be just over four years, whereas the unbiased figure was nearly seven years. This arises because members who will eventually become long-term members but were only recruited a short time ago, would conventionally count as current, this being a short membership period. As a result the data are biased towards shorter timescales than eventually occur.

Such biases will vary between subgroups, and could obviously affect conclusions. There are examples where the bias has resulted in simple means being virtually equal between subgroups, whereas the actual results differed by over 100 per cent.

SIMPLE TABLES?

There will be various factors that influence whether or not customers will continue their purchases. Some of these will be known — for example, there may be a difference in support from men and women. In order to investigate another factor (such as region) and the proportions of males and females vary by region, then the comparison of regions will have to be adjusted for the different proportions of men and women.

Simple tabulations will be of limited value, as they can only easily look at one or two factors at a time, and have difficulty taking into account imbalance and interactions which often occur. An example of imbalance would be if different regions have different proportions of men and women. An example of an interaction would be if the response rates from men and women were more similar in some regions than in other regions.

Statistical analyses enable a variety of factors to be included and also allow testing of whether different strategies should be used for different types of customers. In the current context, where time to event data are of interest, this should be analysed using survival analysis as this takes account of the incomplete data.

WHAT IS SURVIVAL ANALYSIS?

Survival analysis looks at the time between events. The methods were originally devised to study the time between medical intervention and death (which only occurs once), and were used to demonstrate the benefits of certain interventions. The methods have now been expanded to analyse other events, and to analyse events that occur many times for the same individual.

Historically most applications of survival analysis have concentrated on negative events, and a lot of the terminology reflects this. For example the terms ‘hazard’ and ‘survival’ both assume that the event is undesired. In the first case study, the event is the termination of a membership which is negative and which marketing activity would seek to discourage. In the second example, however, the event is a donation which is positive and marketing activity would seek to encourage. In both examples, an important discriminator is the time between events.

CENSORED OBSERVATIONS

It is important to realise that the incomplete nature of the data is not a

question of data quality nor of poor recording. The incomplete data are an integral part of the data to be measured and are closely related to the desired outcome. For example, in the second case study, at the time the data are captured customers will have a period of time following their last donation in which no donation was made. For such intervals there is not an exact 'time to next donation'; instead it is known that the next donation occurred after the final date. The distinctive feature about survival analysis methods is that these intervals can be included in the analysis. The conventional statistical terminology for these intervals is 'censored observations', indicating that the particular interval has been 'censored' before the event of interest had occurred. In the second case study repeat customers will tend to generate relatively short time intervals between purchases, whereas lapsed customers are likely to generate relatively long intervals that are censored because no further purchase has taken place during the time span of the dataset. The survival analysis will take these issues into account appropriately.

Methods that do not take censored data into account will be of limited use for analysing customer data, as they will not include data about when a customer lapses — which is what is to be predicted.

WHY SURVIVAL ANALYSIS?

Survival analysis allows investigation of the probability of lapsing and purchasing. Other factors and past history can be used to predict which customers are most likely to lapse. This information can then be compared with current policies and used to develop alternative policies. Alternative policies can be further tested by carrying out a small trial and analysing the results. Since survival analysis makes such efficient use of the data available, these trials can be limited to very small sizes, a few 100 customers for example can provide statistically significant results. This is a result of their development within a pharmaceutical context where initial trials of drug effectiveness need to be conducted on 20 to 30 patients.

Survival analyses shift the emphasis away from measures such as means towards non-parametric measures such as the median and quartiles.

Survival analysis can then better identify the most regular customers, and those most likely to repeat their support. Survival analysis can also identify those who are unlikely to repeat their support.

WHAT INFORMATION IS REQUIRED?

Whereas little is known about prospects, there is already information about current and lapsed customers, which should enable types of customer to be differentiated. Often more is known about customers than is realised. Simple financial transactions can be made in so many ways:

— cheque, bank giro, standing order, direct debit
— different frequencies of purchase
— the purchase may be regular or irregular (in time and in amounts)
— the purchase may be linked to special events/communications/times of year.

Alternatively, other inbound customer communications and activity could act as triggers for future purchasing. These events could be used to identify the start of intervals to investigate the lead time to purchase and enable appropriate timing of communications.

The minimum information required is a sequence of dates (of activity) with the corresponding customer IDs. Classification variables for different characteristics are, however, useful. The analysis techniques can use static information (eg gender), and dynamic (changing) information (eg current age, number of previous purchases, time since first contact/registration).

Available information can be supplemented by area property price, unemployment, income, financial risk and other information available via postcode matching databases.

The analysis can be much more informative about the effectiveness of current policies and mailings if the date and type of communication is available for each customer/group of customers.

WHAT CAN THE TECHNIQUE DELIVER?

The technique can deliver survival charts, log survival charts, hazard charts and life tables — examples appear in the case studies below.

— survival charts show the probability that the event in question has occurred by a given time. In the second case study, lapsed customers appear as the proportion of customers who have not made a further donation, even after a considerable period of time
— log survival charts are used to show the shape of the curve. If the log survival chart is a straight line, then this indicates that time since the last event is unimportant (does not affect the chance of a further event). If the log survival chart is more complex, then this indicates that time is important, and may indicate different phases of activity. For example, activity shortly after the last donation might be different from activity in the long term
— hazard charts show the likelihood of the event occurring by time since the last event. In the simplest situation this is constant through time. A more complex situation will show patterns that may point toward windows of opportunity for marketing activities
— a life table contains the information from the above graphs in a table.

These charts can help to identify the most suitable times to make requests. Sometimes showing the inaccuracies and inadequacies of current thinking can also be quite important.

Survival analysis techniques are very well developed, for example using Cox regression to identify the impact of different factors. These are beyond the scope of this paper.

CASE STUDY 1: RENEWAL

This organisation, an insurance company, has an annual renewal with records of current and lapsed members going back 17 years.

The number of lapsed members was tabulated by number of years' membership — thus 50,805 members lapsed after their first year's membership.

The number of current members was tabulated by total number of years' membership (paid for). These are listed in column 3 in Table 1. These are censored observations, it is not known whether or not they will maintain their membership next time around.

Next the number at risk is calculated. A total of 357,818 had one or more year's membership and thus could potentially subscribe for a second year. Because 82,379 members are still in their first year, and 50,805 failed to maintain their membership after their first year,

Table 1: Life-table for case 1

Number of years membership	Number not maintaining	Number of current members	Total failed and current to end	Probability of failing (lapsing)	Probability of maintaining at the end of each year	Probability that membership lasts $>t$ years
T	Failed	Censored	At risk	Hazard %	%	Survival %
1	50,805	82,379	357,818	14.2	85.8	85.8
2	35,629	16,481	224,634	15.9	84.1	72.2
3	24,451	8,265	172,524	14.2	85.8	62.0
4	16,139	5,490	139,808	11.5	88.5	54.8
5	14,823	3,875	118,179	12.5	87.5	47.9
6	11,134	2,652	99,481	11.2	88.8	42.6
7	11,047	2,110	85,695	12.9	87.1	37.1
8	10,204	1,926	72,538	14.1	85.9	31.9
9	8,729	2,016	60,408	14.5	85.5	27.3
10	6,448	1,964	49,663	13.0	87.0	23.7
11	5,976	1,768	41,251	14.5	85.5	20.3
12	5,185	1,216	33,507	15.5	84.5	17.1
13	8,623	573	27,106	31.8	68.2	11.7
14	7,282	246	17,910	40.7	59.3	6.9
15	4,837	0	10,382	46.6	53.4	3.7
16	3,785	0	5,545	68.3	31.7	1.2
17	1,760	0	1,760	100.0	0.0	0.0

only 224,634 members started their second year.

The probability of lapsing after the first year is 50,805/357,818 = 14.2 per cent. Thus 85.8 per cent of members maintained at the end of their first year. If a customer maintains their membership for two years, then the probability of lapsing at the end of the second year is 35,629/224,634 = 15.9 per cent. The remaining rows are calculated in a similar way.

The next column, which is the probability of maintaining each year, is simply one minus the hazard.

The probability of surviving (ie membership lasting) beyond the first year is the probability of maintaining at the end of the first year. The probability of surviving beyond the second year is given by the probability of surviving (membership lasting) beyond the first year, multiplied by the probability of maintaining at the end of the second year. That is 85.8 per cent × 84.1 per cent = 85.8 per cent × 84.1/100 = 72.2 per cent. The remaining rows are calculated in a similar way.

Note that the current membership only goes back 14 years, whereas past membership was up to 17 years. This may be due to the situation changing over time, or due to data problems/inconsistencies. Either of these two columns going to zero, or changing dramatically, is an indication of data peculiarities that should be investigated. Advice should be sought about the appropriate ways of dealing with any such 'features' in the data. In the current example, the graphs have only been drawn up to 12 years, and the estimates up to that point should not be unduly affected by the peculiarities noted.

A graph of the probability of remaining in membership against years since joining showed a smooth curve (see Figure 1). This is the shape of an exponential curve, as might be expected. To check this, plot the log of the probability against time (see Figure 2);

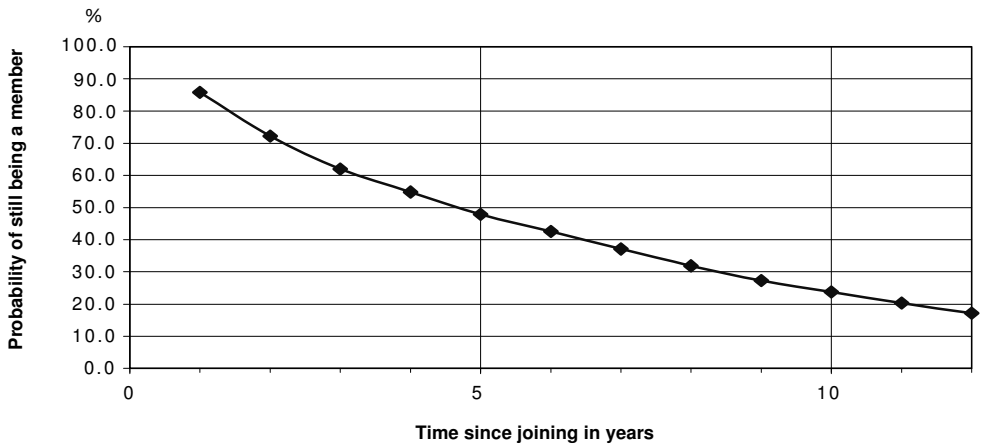

Figure 1 Probability of still being a member against time since joining

the straight line indicates that the data do follow an exponential curve.

Also for an exponential curve, the hazard (the probability of lapsing at the end of any year) is constant. This can be tested by a graph of the hazard against time (Figure 3). The graph is approximately constant over time, but there are some minor variations. The situation was not quite as simple as was first presented. A few people take out two or five-year memberships when first joining, and this may be related to the peaks at two and five years. Other differences may be due to random fluctuations, or could be due to different subpopulations.

The dominant feature of these data is the flat nature of the hazard function. This suggests that the length of time in contact so far is not a guide as to how long the customer will remain in contact. As a result, associated marketing strategies should be designed to develop an ongoing relationship without any specific time component. Around this general strategy, there may be some adaptations at the relative peaks and troughs, for example after two years in contact.

The analysis can be developed further by identifying those characteristics known at the time of recruitment to predict their expected length of membership and therefore derive estimates of their overall lifetime value.

Although little is currently known about the characteristics of the members, further analyses using postcode information have shown differences between subgroups, and have been used to inform and target recruitment campaigns.

CASE STUDY 2: REPURCHASE

This organisation conducts business-to-business marketing of a range of business development and training products. They conduct a significant amount of marketing communications to a dynamic universe of potential customers. This has meant they can benefit from accurately tracking customer and prospect behaviour and keeping an eye on new entrants into the market. Survival analysis was implemented straightforwardly and provided some useful insights. The analysis looked particularly at the different levels of follow-on behaviour from previous customers and recently acquired customers and the way that customers responded to direct mail communications.

Within the portfolio of survival

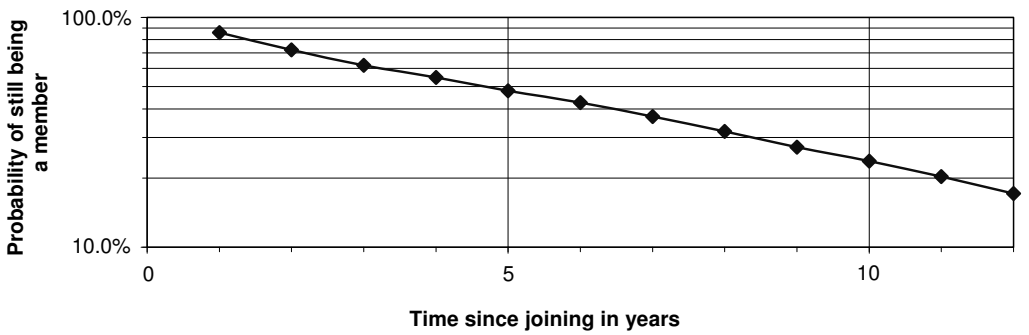

Figure 2 Log survival chart

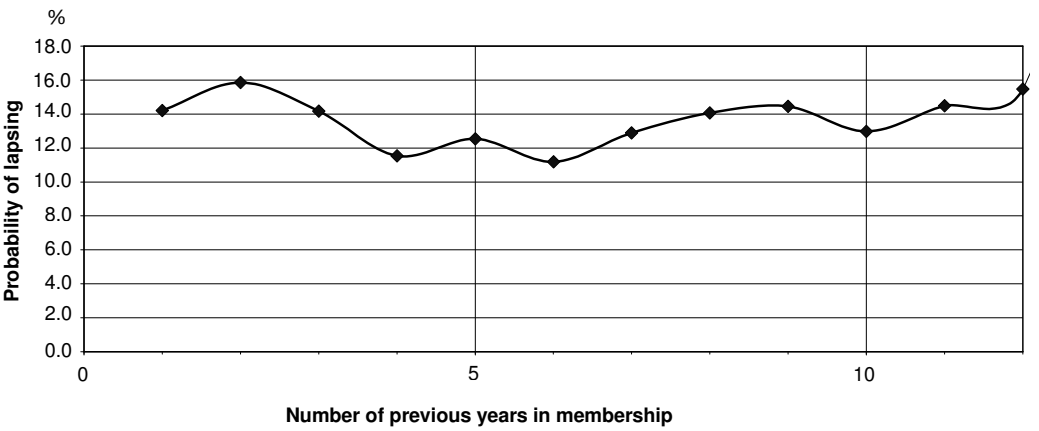

Figure 3 Probability of lapsing by number of previous years in membership

analysis techniques a method called Gehan's generalised Wilcoxon test has been developed to identify different behaviour between groups. This is equivalent to analysis of variance techniques that are commonly used for normally distributed behaviour. The method adapts the non-parametric versions of these tests to accommodate the types of distributions obtained from time-based data.

Table 2 shows the significant differences in the long-term behaviour of previous customers compared to new customers, as shown in Figure 4.

Figure 4 demonstrates that just under 90 per cent of new customers will not make a follow on purchase in the long term compared to 65 per cent of previous customers. As can be seen censored rates are high in this type of data making the difficulties experienced with more conventional techniques even more apparent.

Survival analysis was also used to understand the response of customers to direct mail communications. The hazard function that shows the conditional probability of response after a given period of time following the mailing is shown in Figure 5. The chart demonstrates the expected peak in response immediately following the communication but also indicates an

Table 2: Comparison of survival experience using the Wilcoxon (Gehan) statistic

Survival Variable NEXTDATE NextDateDiff_c grouped by NRESPONS NTITLES of RESPONSE
Overall comparison statistic 88.481 D.F. 1 Prob. 0.0000

Group label	Total	Uncensored	Censored	Censored %	Mean score
2 (new)	15,816	1,193	14,623	92.46	69.4750
4 (previous)	1,223	217	1,006	82.26	−898.4595

*Output from SPSS v 10.0 Advanced models module

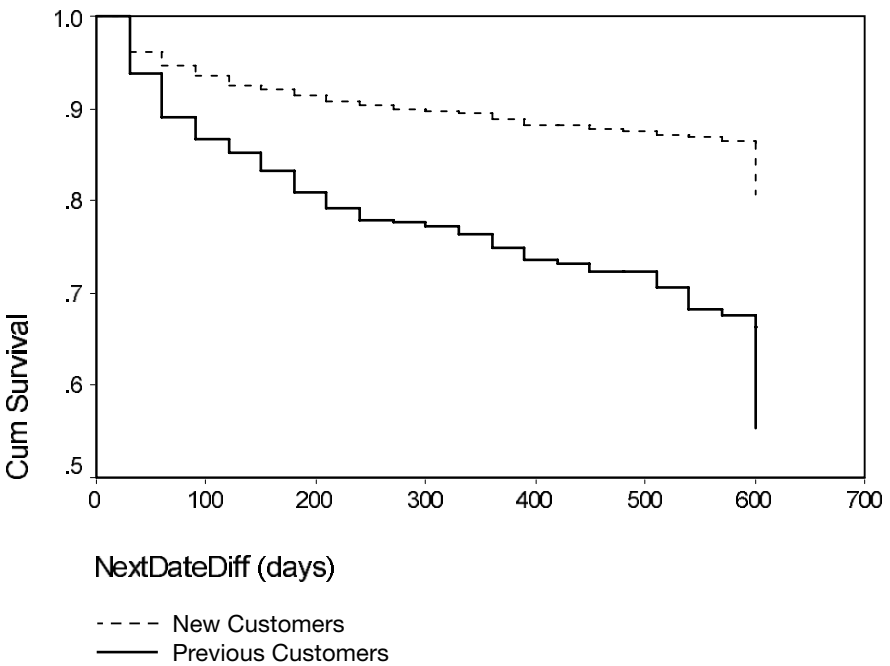

Figure 4 New and previous customers compared: Survival function

increase in response around three and six months following. This provides the impetus for investigating follow-up campaigns and timed and targeted outbound calls to coincide with these times of interest.

CONCLUSION

Survival analysis is the appropriate tool for looking at data on the length of contact and time to the next purchase. Survival analysis can help to provide insights into the data, and can also test the success of new procedures in encouraging and maintaining the customer base.

As has been shown in a number of other areas, survival analysis is the method to use with time-to-event data. As experience of applying the technique within the marketing arena grows, understanding how to handle the nuances of marketing information will develop. A key contrast with other application areas is the existence of long-term censoring.

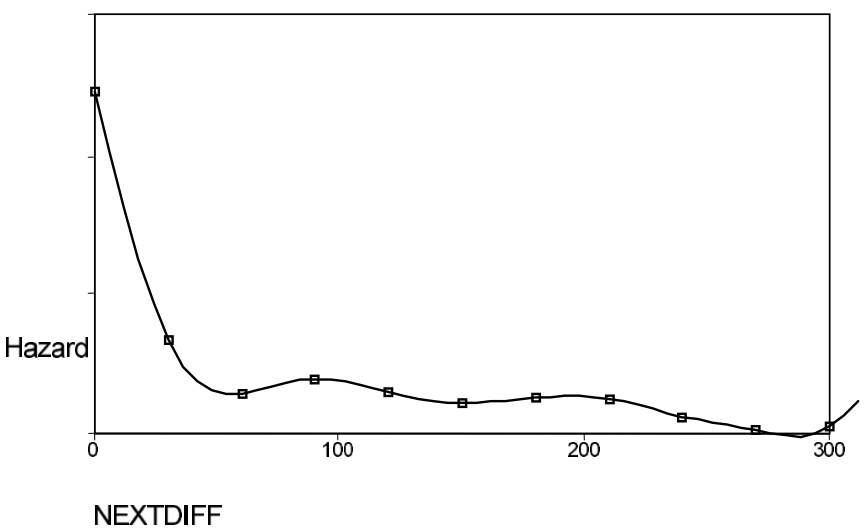

Figure 5 Response to communications: When to time follow-up activity?

That is, with most applications, the event (eg death) will eventually occur in all individuals, whereas in the marketing context the event (eg next donation) does not occur in all individuals. With the current descriptive application of the technique those differences are not important. As we apply more advanced areas, these considerations will become more relevant.

References

SPSS Advanced Models v10.0 Manual. Copyright © 1999 by SPSS Inc, USA.

Modern Applied Statistics with S-plus, W.N. Venables & B.D. Ripley. Copyright © 1994 by Springer-Verlag New York Inc.