MEDWELL PUBLISHING
Journal of Animal and Veterinary Advances 8 (1): 177-182, 2009
ISSN: 1680-5593

An Algorithm to Sort Complex Pedigrees Chronologically without Birthdates

[1]Zhiwu Zhang, [2,3]Changxi Li, [4]Rory J. Todhunter, [5]George Lust,
[3,6]Laksiri Goonewardene and [3]Zhiquan Wang
[1]Institute for Genomic Diversity, Cornell University, Ithaca, New York 14853, USA
[2]Agriculture and Agri-Food Canada, Lacombe Research Centre,
6000 C and E Trail, Lacombe, AB T4L 1W1, Canada
[3]Department of Agricultural, Food and Nutritional Science,
University of Alberta, Edmonton, AB T6G 2P5, Canada
[4]Department of Clinical Sciences, College of Veterinary Medicine,
Cornell University, Ithaca, New York 14853, USA
[5]Baker Institute for Animal Health, College of Veterinary Medicine Cornell University,
Ithaca, New York 14853, USA
[6]Alberta Agriculture and Rural Development, #204-7000-113 Street,
Edmonton, AB, T6H 5T6, Canada

Abstract: Information on the additive genetic relationship among individuals in a pedigree is essential in genetic analyses for the estimation of variance components, prediction of breeding values and association mapping for Quantitative Trait Loci (QTL), so that background polygenic effects can be estimated. Calculation of the additive genetic relationship from a pedigree requires individuals to be chronologically ordered such that parents appear before their progeny. This can be accomplished by sorting individuals by their birth date. However, in practice, birth dates may not be available for some individuals due to missing records or inaccurate/errors in recording. In this study, we derived a Pyramid algorithm to obtain chronologically ordered data with the feature that parents appear before their progeny based on their parental information embodied in the pedigree. A software package (SeqPed) was developed based on the algorithm. The chronological order of individuals and the additive relationship matrix calculated from a dog pedigree based on the full birth date information was the same as those calculated from the pedigree and the chronological order obtained from the software without requiring birth date information.

Key words: Pyramid algorithm, pedigree, chronological order, additive genetic relationship

INTRODUCTION

Additive genetic relationships, i.e. kinship and inbreeding coefficients are essential parameters for genetic analyses as these parameters define the variance and covariance structure for the random effect in Dr. Henderson's Mixed Model Equation (MME). The solutions to the individual's random additive genetic effect from MME, called the Best Linear Unbiased Prediction (BLUP), is used as a breeding value for the genetic improvement in animals and plants (Henderson, 1963, 1973, 1976). The mixed model or animal model approach which accounts for all known genetic relationships among individuals has also been recommended for gene association analyses as the model can provide unbiased estimates of single gene effects on quantitative traits (Kennedy *et al*., 1992). The calculation of the additive genetic relationships requires the pedigree to be chronologically ordered such that parents must precede their progeny (Quaas, 1989). In genetic analyses, the raw data must be formatted in a way that parents appear before their progeny. Such ordering is a requirement for input data for software such as the Multiple Trait Derivative-Free Restricted Maximum Likelihood (MTDFREML) program (Boldman *et al*., 1993), Average Information Restricted Maximum Likelihood (ASReml) (Gilmour *et al*., 2000) and Statistical Analysis Software (SAS) Family procedure (SAS, 1989) in order to implement the mixed model. Ordering a pedigree chronologically is usually accomplished by sorting

Corresponding Author: Zhiquan Wang, Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, AB T6G 2P5, Canada

individuals by their birth date. However, birth date information may not always be available (Mc Parland *et al*., 2007) and in some dairy and beef cattle populations 2.4-12.9% of the birth dates may be missing (Mc Parland *et al*. (2007) personal communications). In practice, individuals with missing birth dates are simply deleted from the analyses which compromises the utilization of the full data set. In addition, erroneous birth dates will result in an incorrect pedigree chronological order which may cause bias in estimates of kinship and inbreeding coefficients if individuals are sorted by their birth date. The objective of the study, was to derived an efficient algorithm to sort individuals in a pedigree in a chronological order without requiring birth dates. Furthermore, the effect of birth date errors on calculating kinship and inbreeding coefficients was also assessed when the chronological order of individuals in a pedigree was obtained by sorting their birth dates.

MATERIALS AND METHODS

Raw pedigree format: For each individual in a pedigree, we use three columns to present the parentage relationship. The first column is each Individual Identification (ID). The second and the third columns are the IDs of the two parents (father and mother). Each founder with unknown parents is required to appear in the individual column with an unknown indicator in the parent columns. An example of the raw pedigree format is given in the first three columns in Table 1.

Algorithm: Based on the raw pedigree data, the algorithm was to examine the parental status of each individual and to identify the generation information in an iterative process. The process initially considers all individuals in a pedigree as parents and sets all individuals' parental indicator to 1 and the generation indicator to zero. Then the algorithm identifies the individuals that do not appear as parents of others and sets their parental indicators to 0 and keeps their generation indicator unchanged. For parents, the algorithm increases their generation indicator by 1 and keeps their parental indicator unchanged, i.e. 1. The algorithm iterates the process among potential parents until there are no parents left. The chronological order of individuals in the pedigree can be easily achieved by simply sorting individuals on their generation indicator in ascending order which will place parents before their progeny in the data set because a progeny's numerical generation indicators are always smaller than their parent's numerical generation indicators. This process is illustrated in Fig. 1. As the iteration progresses, the number of potential parents at each stage gets smaller and smaller, therefore, this algorithm is named the Pyramid algorithm. Mathematically, the algorithm is implemented through the following four steps:

Step 1: Let p be a pedigree with three columns indicating the individual, father and mother. Assume that each column has n rows. Let t be an empty pedigree. Initialize each individual's generation indicator g to 0: g(i) = 0 for all i in individual column of p.

Table 1: Illustration of application of the Pyramid algorithm to obtain a data set with a correct chronological order[a]

			Iteration												Output			
			0		1		2		3		4		5					
Original																Original		
ID	Father	Mother	S*	G*	S	G	S	G	S	G	S	G	S	G	G	ID	Father	Mother
1	4	12	1	0	1	1	1	2	0	2	0	2	0	2	4	3	0	0
2	11	13	1	0	1	1	1	2	0	2	0	2	0	2	4	9	0	0
3	0	0	1	0	1	1	1	2	1	3	1	4	0	4	3	4	3	9
4	3	9	1	0	1	1	1	2	1	3	0	3	0	3	3	11	3	9
5	14	15	1	0	1	1	1	2	0	2	0	2	0	2	3	12	0	0
6	5	10	1	0	1	1	0	1	0	1	0	1	0	1	3	13	0	0
7	6	8	1	0	0	0	0	0	0	0	0	0	0	0	3	14	0	0
8	2	1	1	0	1	1	0	1	0	1	0	1	0	1	3	15	3	9
9	0	0	1	0	1	1	1	2	1	3	1	4	0	4	2	1	4	12
10	11	13	1	0	1	1	1	2	0	2	0	2	0	2	2	2	11	13
11	3	9	1	0	1	1	1	2	1	3	0	3	0	3	2	5	14	15
12	0	0	1	0	1	1	1	2	1	3	0	3	0	3	2	10	11	13
13	0	0	1	0	1	1	1	2	1	3	0	3	0	3	1	6	5	10
14	0	0	1	0	1	1	1	2	1	3	0	3	0	3	1	8	2	1
15	3	9	1	0	1	1	1	2	1	3	0	3	0	3	0	7	6	8
			a													b		

[a]The example data is from the JV Pedigree (Goddard *et al*., 1996). The raw pedigree data is shown as the first three columns containing the individual's original IDs and IDs of their father and mother. At the start of the iteration (a), all individuals are considered parents and their parent indicator S* are coded as "1" and generation indicator G* as "0". Iteration 1 identifies individual 7 as not a parent of any other individuals in the pedigree and changes its parent indicator S to 0 and keeps the generation indicator unchanged. For other individuals, their generation indicator is increased by 1 and the parent indicator "S" is left unchanged. Each of following iterations repeats the process among individuals with parent indicator as "1" until no parents can be identified. It took five iterations to finish the process. At the end of the iteration, individuals are sorted by generation indicator in a descending order. In the sorted data set (b), all parents precede their progeny

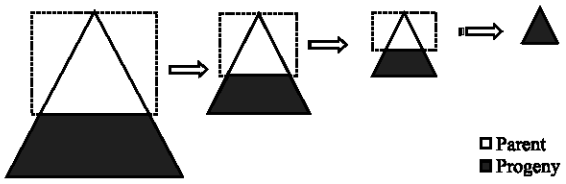

Fig. 1: Pyramid algorithm to chronologically sort pedigrees. At each stage, parents (who have progeny) are separated from progeny (who are not parents of any other individuals in the data set) and the generation indicator of the parent in increased by 1. The parents are kept as new data and the process is repeated until no individual can be claimed as parent. A correct chronological order of individual in the pedigree can be achieved by sorting all individuals on their generation indicator in descending order

Step 2: For each element i in the individual column of p, let f be the father of i and m be the mother of i. If f is not in the individual column in t, copy the row in p containing f in the individual column to t, increase the generation for f: g(f) = g(f)+1. If m is not in the individual column in t, copy the row in p containing m in the individual column to t, Increase the generation for m: g(m) = g(m)+1.

Step 3: If t is empty, go to step 4, otherwise let p = t and t = empty, go back to step 2.

Step 4: Sort all individuals in p by their generation indicator g in a descending order.

A numerical example: The JV Pedigree was used as a numerical example to demonstrate the iteration process of the Pyramid algorithm. The pedigree was graphically illustrated by Goddard *et al.* (1996) for JV family from France based on a Caucasian data set (Nakura *et al.*, 1994). For illustration purpose, the 15 individuals in the pedigree were given arbitrary IDs from 1-15 (Fig. 2). Birth date information was not available for all the individuals. The pedigree was displayed numerically by the first three columns in Table 1 section a with an unsorted chronological order for the application of the Pyramid algorithm iteration. At the start of the iteration (0) (Table 1), all individuals are considered parents and their parent indicator S* are coded as 1 and generation indicator G* as 0. Iteration 1 identifies individual 7 as not a parent of any other individuals in the pedigree and changes its parent indicator S to 0 and keeps the generation indicator unchanged. For other individuals, their generation indicator is increased by 1 and the parent indicator remains unchanged. Each of the following

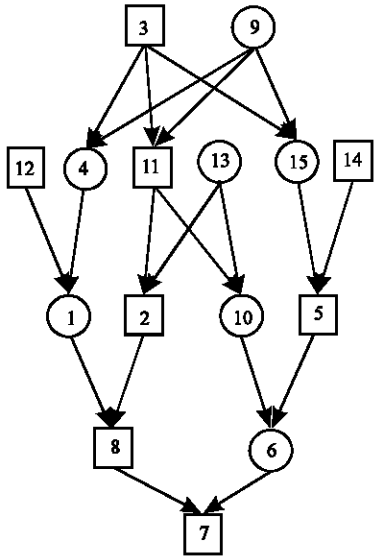

Fig. 2: The JV Pedigree (Goddard *et al.*, 1996). The individuals were arbitrarily named with sequential ID 1-15

iterations repeats the process among individuals with parent indicator as 1 until no parents can be identified. At the end of the iteration, individuals are sorted by their generation indicator in descending order and the chronologically ordered IDs are presented in Table1 section b featuring the parents before their progeny in the sorted data set.

An empirical example: We used a dog pedigree to demonstrate the efficiency and effectiveness of the proposed Pyramid algorithm in formatting members of a pedigree in chronological order without birth date information and to assess the impact of different birth date error rates on the calculation of kinship and inbreeding coefficients when the chronological order is obtained through sorting the birth dates. The pedigree contained 266 dogs that were sampled from a research colony maintained at the Baker Institute for Animal Health at Cornell University for over 30 years. This pedigree included 104 Labrador Retrievers and their 12 founders with unknown parents, 143 crossbred progeny between 8 Labrador Retriever founders (4 males and 4 females) and 7 Greyhound founders (2 males and 5 females) over four generations (F1 × both Greyhound and Labrador Retriever founders, F2 and ¾ × ¾ Labrador Retriever). Birth dates were recorded for all the dogs except founders. The birth date of a founder was assigned to January 1 of the year which was 2 years before having the first progeny. No birth date error was identified as parents have earlier birth dates than the progeny.

The pedigree data was formatted in a chronological order using the proposed algorithm in this study (Analysis 1) and subsequently the kinship and inbreeding coefficients were calculated. For comparison, the kinship and inbreeding coefficients were also calculated based on the data set ordered chronologically by sorting the birth dates (Analysis 2).

To evaluate the impact of birth date error on the calculation of kinship and inbreeding coefficients, birth date error was simulated at 3 levels of error rate (r = 1, 5 and 10%). Part of the 266 dogs were sampled at for each error rate their birth dates were shuffled to create errors in their birth dates. Kinship and inbreeding coefficients were calculated from the data sets in a chronological order which was obtained by sorting the birth dates with simulated errors. The simulation was repeated 1000 times for each of the three levels of error (Analysis 3).

The three analyses were compared for the properties of the matrix containing kinship as off diagonals and one plus inbreeding coefficients as diagonals. This matrix is called additive relationship matrix or numerical relationship matrix. Six matrix properties were evaluated: The number of non-zero elements, the number of inbred dogs, the average inbreeding coefficients of the 266 dogs, the average additive relationship between pairs of individuals, the logarithm of the determinant, the trace.

RESULTS AND DISCUSSION

The Pyramid algorithm proposed in this study succeeded to sort the raw pedigree data in a correct chronological order. For the numerical example from the JV Pedigree, all parents preceded their progeny in the output data set (Table 1). The kinship and inbreeding coefficients calculated for the empirical dog example were identical between the two data sets with the chronological order obtained by Pyramid algorithm (Analysis 1) and by sorting birth dates without error (Analysis 2). The above mentioned 6 properties of the numerical relationship matrix calculated from the 2 methods were presented in Table 2.

Although, the kinship and inbreeding coefficients were identical, the position of individuals in the two chronological ordered data sets may not be the same. Actually the order of individuals within the same generation does not have an effect on calculations of kinship and inbreeding coefficients as long as parents appear before their progeny in the data set.

The individuals may be chronologically ordered incorrectly when the chronological order is obtained by sorting birth dates with errors (Analysis 3). The effect of birth date errors on the six properties of the numerical relationship matrix is shown in Table 2. It can be seen that as the error rates increase, the number of non-zero elements, the number of inbred dogs, the average inbreeding coefficient, the average additive relationship between pairs of individuals, the logarithm of the determinant and the trace of the numerical relationship matrix each deviates more from the true value. Average kinship (off diagonals) among the 266 dogs decreased 1.5, 8.4 and 11.7% for the birth date error rate at 1, 5 and 10%, respectively. Average inbreeding coefficient (diagonals) of the 266 dogs decreased 1.0, 4.6 and 5.9% for birth date error rate at 1, 5 and 10%, respectively.

The details of the deviation from the true properties are showed in Fig. 3. These deviations could result in a biased estimation of genetic parameters in genetic analyses.

As the proposed pyramid algorithm is independent of birth dates and it only uses the information that is embedded in the pedigree data itself, it is robust in obtaining the required chronological order (parents before offspring) even when birth dates are not available and/or there are errors in the birth dates.

Recording birth dates for animals in pedigrees is a common practice in some animal species in their breeding programs. However, errors in birth dates may occur during the data collection. If errors occurred in the data collection, the genetic analyses based on the chronologically ordered data set which is obtained by sorting birth dates will result in an incorrect additive

Table 2: Properties[a] of the numerator relationship matrices for the 266 dogs based on the data set with a chronological order obtained by different methods[b]

		Error rate* (III)		
Analysis	I and II	1%	5%	10%
Non-zero elements	16496	16382 (377)	15834 (866)	15134 (1184)
Number inbred	36	35.76 (1.23)	34.07 (4.06)	31.99 (5.67)
Diagonal average	1.0133	1.0131 (0.0008)	1.0120 (0.0020)	1.0106 (0.0026)
Off diagonal average	0.0590	0.0584 (0.0016)	0.0557 (0.0036)	0.0524 (0.0045)
-Log determinant	136.61	135.86 (1.83)	132.36 (4.22)	128.05 (5.35)
Trace	269.53	269.48 (0.21)	269.19 (0.53)	268.83 (0.68)

[a]Properties include average statistics and standard deviation in brackets. [b]Methods include the Pyramid algorithm (Analysis I), the data set with a chronological order obtained by sorting on birth dates (Analysis II) and the data set with chronological order obtained by sorting on birth dates with errors (Analysis III). The results from Analysis I and II were identical. *Birth date errors of the 266 dogs were simulated at three rates: 1, 5 and 10%. Kinship matrix or additive numerator matrix was calculated based on a data set with a chronological order obtained by sorting on birth dates with the simulated errors. The simulations were repeated 1000 times

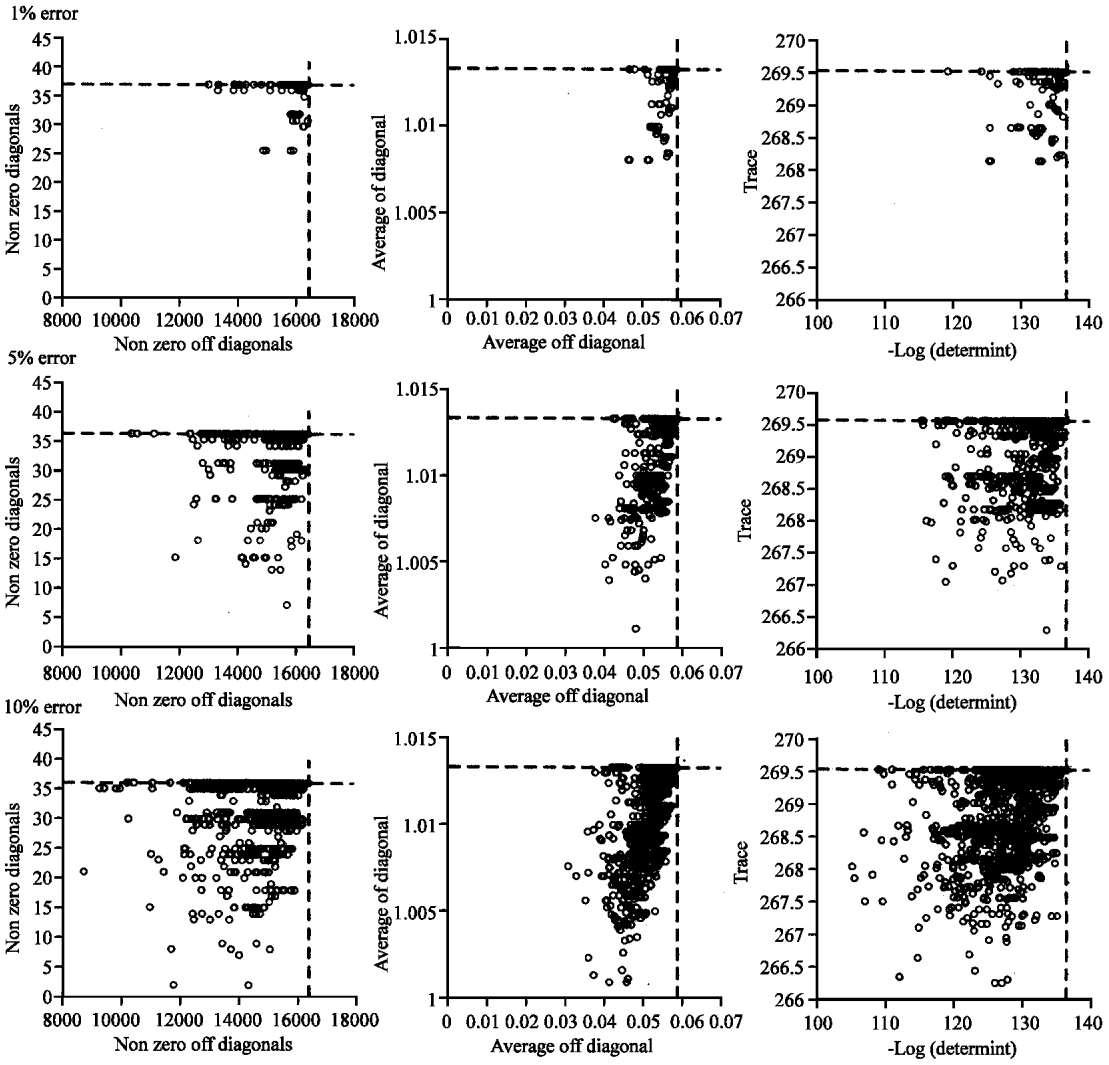

Fig. 3: Plots of the properties of kinship matrices generated on pedigrees of 266 dogs sorted on birth dates which were simulated at three levels of error rates (1, 5 and 10%) and the simulation was repeated 1000 times. The dashed lines indicate the true values from the kinship matrix generated from the pedigree sorted on birth dates without error

genetic relationship matrix, hence will lead to incorrect inferences. In addition, as computational technologies and genomics advance, animal models have been widely adapted to analyze genomics data in other species such as crops and forest trees in order to obtain more accurate estimates of genetic parameters or for genetic association studies. However, birth dates may not be available in those species due to the nature of breeding system practices in crops and trees, but the pedigree or parental data may be known. The Pyramid algorithm will provide an effective and robust method to generate a data set in correct chronological order when the pedigree or parental information is available. It is a tool to add precision to a genetic analysis where the pedigree information is missing or incomplete.

The computing time of the Pyramid algorithm is as efficient to the approach of sorting pedigree on birth date. An algorithm is efficient if the computing time is a polynomial function of the size of problem. Computing time is a linear function of nlog (n) to sort a pedigree by birth date, where n is the number of individuals in the pedigree. With the Pyramid algorithm, Step 2 takes n iterations. Step 3 repeat Step 2 t times, where t is number of generations. As t is always smaller than n, the overall computing time is a function of n^2 at the maximum. Consequently, the proposed algorithm is efficient.

CONCLUSION

The Pyramid algorithm is efficient and robust to obtain chronologically ordered data with parents appearing before their progeny based on parental information embodied in the pedigree without requiring birth dates. The algorithm is superior to the traditional method of obtaining a chronological order by sorting on birth dates as it is able to recover all parental relationships among individuals in a pedigreed population when missing and/or wrong birth dates are present. A software package of SeqPed was developed for the algorithm. Its trial version is free to download at http://www.aiivis.com/software/seqped.

REFERENCES

Boldman, K.G., L.A. Kriese, L.D. Van Vleck, L.A. Van Tassel and S.D. Kachman, 1993. A manual for use of MTDFREML, a set of programs to obtain estimates of variance and covariances. USDA-ARS, Clay Center, Nebraska, USA.

Gilmour, A.R., B.R. Cullis, S.J. Welham and R. Thompson, 2000. ASREML reference manual. IACR-Rothamsted Experimental Station, Harpedden, UK.

Goddard, K.A.B., C.E. Yu, J. Oshima, T. Miki, J. Nakura, C. Piussan, G.M. Martin, G.D. Schellenberg and E.M. Wijsman, 1996. Toward localization of the Werner syndrome gene by linkage disequilibrium and ancestral haplotyping: Lessons learned from analysis of 35 chromosome 8p11.1-21.1 markers. Am. J. Hum. Genet., 58: 1286-1302.

Henderson, C.R., 1963. Selection Index and Expected Genetic Advance. In: Hanson, W.D. and H.F. Robinson (Eds.), Statistical Genetics and Plant Breeding. NAS-NRC Publication No. 982, Washington/DC, pp: 141-163.

Henderson, C.R., 1973. Sire evaluation and genetic trends. Proceeding of the Anim Breed Genet Symp in Honor of JL Lush. Am. Soc. Anim. Sci, Am. Dairy Sci. Assn., Champaign/IL, pp: 10-41.

Henderson, C.R., 1976. A simple method for computing the inverse of a numerator relationship matrix used for prediction of breeding values. Biometrics, 32: 69-79.

Kennedy, B.W., M. Quinton and J.A.M. Van Arendonk, 1992. Estimation of effects of single genes on quantitative traits. J. Anim. Sci., 70: 2000-2012.

Mc Parland, S., J.F. Kearney, M. Rath and D.P. Berry, 2007. Inbreeding trends and pedigree analysis of Irish dairy and beef cattle populations. J. Anim. Sci., 85: 322-331.

Nakura, J., E.M. Wijsman, T. Miki, K. Kamino, C.E. Yu, J. Oshima, K.I. Fukuchi *et al.*, 1994. Homozygosity mapping of the Werner syndrome locus (WRN). Genomics, 23: 600-608.

Quaas, R.L., 1989. Transformed mixed model equations: A recursive algorithm to eliminate A^{-1}. J. Dairy Sci., 72: 1937-1941.

SAS, 1989. SAS/STAT User's Guide (Version 6) 4th Edn. SAS Inst. Inc., Cary, NC.