J. Ind. Soc. Agril. Statist.
62(2), 2008 : 138-148

Statistico-Genetic Considerations in Longitudinal Data Analysis

Prem Narain
B-3/27A, Lawrence Road, Keshav Puram, Delhi-110035

SUMMARY

The methodology of longitudinal data analysis (LDA) has been discussed with particular reference to applications in studies on nutrition and animal breeding. It is based on the concept of *intra-individual variation* first advocated by Sukhatme in nutrition studies. First the process view of nutrition is discussed with an auto-regressive Markov process for analysing data on protein or energy intake. The general theory of linear models with correlated errors is then used, in the context of half-sib mating design used in animal breeding, to develop the structure of covariance matrix. Its elements are in terms of four components of variation and one serial correlation coefficient. The observational components of variance are related to the causal components of variation based on genetic considerations. Intra-individual heritability (h_w^2) in a narrow sense, in contrast to the usual heritability (h^2) used in quantitative genetics literature, is introduced that depends on the process variance and the average serial correlation coefficient. As a consequence, a useful test for the existence or otherwise of additive × local environmental interaction effects has become available. A significant process variance with a significant autocorrelation function or its associated variogram indicates a significant h_w^2. The heritability of k repeated measurements is derived and used to develop a new formula for the heritability of the progeny test used in animal breeding. This formula indicates that the LDA leads to increased accuracy in predicting the breeding value of the male on the basis of offspring's performance. The estimation of the parameters of the linear model with correlated errors, particularly the covariances, by restricted maximum likelihood method is also described.

1. INTRODUCTION

Soon after the Indian Statistical Institute was established in 1931 at Calcutta (now Kolkata) by Professor PC Mahalanobis, it attracted a number of research workers who made great names in the field of statistics and related subjects. Dr. K Kishen was one of them who joined the Institute in 1936 and worked with Prof. RC Bose on the problem of confounding in general symmetrical factorial designs using galois fields and projective geometry. This work, published in *Sankhya* in 1940 and establishing interesting relationship between the fundamental simplex at infinity in a space of n dimensions and the components of main effects and interactions in a factorial experiment, is a landmark paper in the field of design of experiments that attracted favourable comments and praise from its architect, the well known Professor RA Fisher. But Dr. Kishen, although he continued his interests in the design of experiments, was more known as an agricultural statistician due to his enormous contributions in the Department of Agriculture, UP where he had joined as Statistician in 1940 and worked for several decades before his retirement from the same Department. He had a close association with the Statistical Wing of the ICAR, now known as IASRI, and the Indian Society of Agricultural Statistics and worked with Drs. PV Sukhatme and VG Panse on several problems of agricultural surveys and crop yield estimation.

My association with Dr. K Kishen started soon after I obtained my M.Sc. degree in Mathematical Statistics in 1954 from Lucknow University and took up an appointment in an ICAR Scheme on Cost of Cultivation of Sugarcane through Sample Survey at the Cane Commissioner's Office, Lucknow that was under his technical control. The scheme, under the guidance of Dr. VG Panse, was launched in UP and Bihar as a sequel to the success of the sample survey methodology for determining the cost of cultivation of a crop that was

attempted by Dr. Panse for cotton in Akola district of Maharashtra a few years ago. My work involved inspecting the field work of kamdars posted in different randomly selected villages of the State to collect data and supervising the subsequent data analysis at the Cane Commissioner's Office. Both Drs. Panse and Kishen used to conduct training programmes for the field staff regularly in one of the villages of the scheme to which we were also invited. It was in those meetings that I had a first hand experience in the techniques of sugarcane crop production and the associated field work. Dr. Kishen's zeal was enormous and later on he took fancy of me on my excellent output of data analysis. My association with Dr. Kishen became more intimate after my joining the IARS (now IASRI) and later on becoming the In-Charge of training programmes. I used to have discussions with him often on research problems in statistics. I recall one such discussion on partially balanced incomplete block designs - the truncated triangular designs with five associate classes that he had invented and presented, along with AN Shukla, in one of the conferences of the ISAS in 1974. I used this device in the construction of partial diallel crosses and published it, with AS Arya (who worked with me for his Ph.D.), in the journal of ISI, the *Sankhya* in 1981.

I pay my humble tribute to the memory of Dr. K Kishen by discussing an important topic, the statistico-genetic considerations in longitudinal data analysis (LDA).

2. NATURE OF LONGITUDINAL DATA

Longitudinal data are characterized by the fact that individuals have observations recorded *over time* for the *same* randomly selected unit – usually a short time series – in contrast to a cross-sectional data in which there is a single observation for each unit. Such data require special statistical methods since each observed unit now provides with a vector of observations that have a certain stochastic dependence structure. Longitudinal Data Analysis (LDA) therefore necessarily involves looking at the data as one realization of a stochastic process. There are numerous instances in which a characteristic can provide repeated measurements over time. I will deal with two situations, one from the field of nutrition and the other from the field of animal breeding.

In a nutrition study, the data could be *daily* N-balance on fixed intake measured on each of several individuals engaged in similar activities. Dr. PV Sukhatme viewed such data as an auto-regressive (AR) stochastic process and introduced the concept of intra-individual variation in protein requirement. Later on he and myself showed that the intra-individual variability in calorie or protein intake had a genetic interpretation in terms of the genotype × environment interaction. I will describe this *process view of nutrition* in Section 3 to illustrate one aspect of the special nature of the LDA.

In animal breeding, particularly dairy cattle and buffalo breeding, milk yield is one characteristic repeated in time when recorded in successive lactations and has been intensively used for genetic improvement studies. The usual practice is to correct the lactation records to the first lactation basis, using Sanders factors, before estimating the repeatability of the milk yield. However, when appropriate theory of LDA for applicability to data on successive lactation records is made available, Sander's procedure may not be necessary. Another instance relating to milk is its protein content. In order to determine it for a given cow, milk is collected on a number of times on different days, analyzed for its protein content and an average taken of the values on the protein content. This, however, ignores the interrelation between the repeated determinations on different days, thus losing valuable information. Instead one can adopt the method of LDA to the original data set. Litter size in animals is another example of a character repeated in time when it is recorded in successive pregnancies where LDA method could be used. Yet another example from animal breeding is body weight of a number of pigs measured in successive weeks in a growth study. The repeated measurements in each of the cases mentioned above constitute the longitudinal data. When such data are based on controlled matings, LDA requires statistico-genetic considerations as we discuss in this paper.

We consider a specific situation of sire evaluation and progeny testing in animal breeding as given in Narain (1990). Each of a set of males is randomly mated to a set of females and one offspring is considered from each such mating. This gives rise to half-sib families for a given male. We now assume that a set of records repeated over time for the character under study for each offspring is available for a given period, the objective being to evaluate the breeding value of the male on the basis of its offspring's measurements for the character repeated over time. We have thus a problem of LDA involving

genetic information, each observation on a number of individuals that are half-sibs, being now a vector of measurements that may be correlated. In Section 5, I discuss this feature of LDA using the general theory of linear model with correlated errors with particular reference to quantitative genetics problems presented in Section 4.

3. SUKHATME'S PROCESS VIEW OF NUTRITION

Sukhatme along with Sheldon Margen of the Department of Nutritional Sciences, University of California, Berkeley developed in 1978, the concept of protein requirements of individuals and indicated the method by which it can be extended to those of populations. According to the joint FAQ/WHO Ad-hoc Expert Committee on Energy and Protein Requirements— the safe level of protein intake is defined as the average requirement plus twice the standard deviation. According to them an individual eating below this level, though not malnourished, runs the risk of developing protein deficiency and this risk increases as the intake falls below the safe level. All along it was assumed that requirements remain constant in an individual. Sukhatme's approach was to take into account the intra-individual variability in requirement, not as a random noise due to measurement error but in a manner represented by an auto-regressive (AR) stochastic process.

When we have time series data on daily N-balance in man maintaining body weight on fixed intake and on the assumption that energy intake is not a limiting factor in the diet, we can represent the series as

$$w_t = \rho\, w_{t-1} + e_t \qquad (1)$$

where w_t is the balance on the t-th day, ρ is the serial correlation of order one between w_t and w_{t-1} and e_t is a random variable with mean zero and variance σ_e^2. This model represents an auto regressive Markov process, comprising of two components – one a short-term component arising from the current value of the process at the previous time point and the other a long-term component in the form of errors of measurement. In such a process, the errors get incorporated into the motion of the process to determine the balance on any given day and are not cancelled out as they would do in a purely random process with $\rho = 0$. The expected value of w_t is found to be zero and the variance of w_t is $\sigma_e^2/(1-\rho^2)$ which is independent of t and therefore remains constant. Such a process is known as stationary stochastic process. The observed value of balance on any given day will then be distributed around mean zero within limits $\pm 2\,\sigma_e/\sqrt{(1-\rho^2)}$ which are known as homeostatic limits.

Using daily data on N-balance on fixed intake, Sukhatme found that for intakes in the range of 3.5 to 12 gms. N/day, the day to day N-balances were not random but were serially correlated in an auto-regressive process as described above. The daily N-balance is regulated according to a probabilistic generating mechanism constant over time. At very high or negligible N-intake, this regulation is shown to break down i.e. homeostasis can no longer be maintained. It was shown that the magnitude of stationary variance is comparable to the variation between individuals. This was found to hold true even when the daily requirement was averaged over several days. Sukhatme and Margen (1978) concluded:

> *Protein deficiency may he defined as a failure of the process to be in statistical control, and not defined in the manner that assumes requirements to be fixed whereby if an individual consumes protein below this level, he suffers from protein deficiency.*

3.1 Genetic Interpretation of Intra~individual Variation

Sukhatme and Narain (1982, 1983, 1984, 1996-97) as well as Narain (1982, 1984, 1990, 1993, 1998, 2000) showed that the intra~individual variability in calorie or protein intake is enhanced due to interaction between the genotype of the individual and the environment as he advances in time.

Assuming that we have data on energy balance or protein intake for *k* subjects, recorded at successive *n* days, the model describing the response of subject *i* on *t*-th day is given by

$$Y_{it} = \mu + b_i + w_{it} \qquad (2)$$

where Y_{it} is the corresponding response with μ as overall mean, b_i's are independently and identically distributed with variance σ_b^2, independently of w_{it} and w_{it}'s for the

same individual are *n* consecutive random variables following an auto-regressive (AR) Markov process of order one given by

$$w_{it} = \rho\, w_{i(t-1)} + e_{it} \qquad (3)$$

where ρ is the serial correlation coefficient of order one and e_{it}'s are independently distributed with mean value zero and variance σ_e^2. It was shown that the expected value of the 'between subjects' mean square (S_b^2) would be

$$E(S_b^2) = [n - (n-1)\theta]\ \sigma_e^2 + n\sigma_b^2 \qquad (4)$$

where

$$\theta = [1 - 2(n+1)\rho / n + (n+1)\rho^2/(n-1) - 2\rho^{(n+1)}/ n(n-1)]/ (1-\rho^2) \qquad (5)$$

The variance of the mean of the individual when averaged over *n* different days can be expressed alternately as

$$V_{P(n)} = \sigma_b^2 + (\sigma_e^2/n)[\{(1+\rho)/(1-\rho)\} - 2\rho(1-\rho^n)/n(1-\rho)^2] = \sigma_b^2 + \bar{r}\,\sigma_e^2 + (1-\bar{r})\,\sigma_e^2/n \qquad (6)$$

where $\bar{r}$ is the average correlation between observations of a given individual and is related to ρ approximately as

$$\bar{r} \simeq 2\rho/n(1-\rho) \qquad (7)$$

The effect b_i in the above model reflects genetic effects of the i-th individual as well as certain environmental effects permanently associated with the individual's development such as intra-uterine and external environment experienced by him. Its variance would therefore contain the genetic component of variance (V_G) as well as common environmental component of variance (V_{Eg}) so that

$$\sigma_b^2 = V_G + V_{Eg} \qquad (8)$$

In so far as σ_e^2 is concerned, it reflects only the variability due to local environmental effects (V_{Es}) provided the genotype does not interact with the environment. If it is not so, another component of variance due to the interaction (V_{GEs}) would enter in the within individual component so that when the observations are averaged for several days, it does not bring about the reduction in the variance of the mean of the individual to the extent it would do if the genetico-physiological process of calorie or protein metabolism had been the same on each day. We, therefore, get

$$\bar{r}\,\sigma_e^2 = V_{GEs} \qquad (9)$$

$$(1-\bar{r})\,\sigma_e^2 = V_{Es} \qquad (10)$$

giving

$$V_{P(n)} = V_G + V_{Eg} + V_{GEs} + V_{Es}/n \qquad (11)$$

and

$$\bar{r} = V_{GEs}/(V_{GEs} + V_{Es}) \qquad (12)$$

The average correlation $\bar{r}$ can then be given a genetic interpretation as '*heritability of the individual'* in a *broad sense* in a manner similar to the concept of '*heritability'* in a *broad sense* quite frequently used in quantitative genetics literature. It is the fraction of the total intra-individual variability which is due to interaction between the genotype and environment and could take any value between 0 and 1. The existence of the genotype × environment interaction thus enhances the intra-individual variability with stabilization of variance as we increase the period of time over which the data are collected. The strength of this interaction can be measured in terms of the serial correlation coefficient signifying the degree of auto-regulatory mechanism.

4. GENERAL LINEAR MODEL WITH CORRELATED ERRORS

Let y_{ijt}, t = 1, 2, ... *k* be the set of observations on the j-th offspring of the i-th male with j = 1, 2, ... *n* and i = 1, 2, ... *m*, there being *n* half-sibs for each of the *m* males. Let the corresponding values of the *p* explanatory variables be x_{ijtl}, l = 1,2.....*p*. We assume that y_{ijt} are realizations of random variables Y_{ijt} which follow the model given by

$$Y_{ijt} = \Sigma\beta_l\, x_{ijtl} + \varepsilon_{ijt} \qquad (13)$$

where ε_{ijt} are random sequences of length *k* associated with each of the *n* offspring and correlated within offspring.

In terms of matrices, let $\mathbf{y}_{ij}$ denote the vector of *k* dimensions, representing the observations pertaining to the j-th offspring of the i-th male. Let $\mathbf{y} = ((\mathbf{y}_{ij}))$ be the matrix of order $m \times n$ representing the whole set of data of N = *mnk* observations. Let X be an $N \times p$ matrix of explanatory variables and for covariance structure let Σ be a block-diagonal matrix of order $nk \times nk$ with a non-

zero 'block' sub-matrix for observations from each male that would be explored in the next section. Then the general linear model for longitudinal data regards y as a realization of a multivariate normal random vector Y with

$$\mathbf{Y} \sim \text{MVN}\ (\mathbf{X}\beta,\ I_m \otimes \Sigma) \qquad (14)$$

where I_m is an $m \times m$ identity matrix and β is a vector of p dimensions.

In order to parametrize the mean and covariance structures separately, we express (14) as

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \qquad (15)$$

with

$$\varepsilon \sim \text{MVN}\ (\mathbf{0},\ I_m \otimes \Sigma) \qquad (16)$$

4.1 Parametric Model for Covariance Structure

We assume that the sequence Y_{ijt}, t = 1, 2, ... k, are sampled from independent copies of an underlying continuous-time stochastic process $\{Y(t), t \in R\}$. We explicitly model the stochastic dependence structure of this process by defining the covariance matrix. For this purpose we take the stochastic process as stationary and assume that it results in a correlation between pairs of observations on the same offspring that depends on the time separation between the pairs of observations. The correlation becomes weaker as the time separation increases. Such a type of correlation is known as *serial correlation* or *autocorrelation*.

The random effects of offspring for each male will give rise to a variance component and likewise the random effects of males themselves will lead to another variance component. The measurement process on individual offspring over time will add a further component of variation of its own. All these three variance components will have to be incorporated into the stochastic dependence structure of serial correlation that will have its own variance component. We model them as below.

(a) Serial correlation

Let the time-varying stochastic process operating within each offspring gives rise to a component of ε_{ijk} with zero mean, variance σ^2 and correlation function $\rho(u)$ for pairs of measurements on the same offspring with u time units apart. The variogram of the stationary process is then given by

$$\gamma(u) = \sigma_e^2 + \sigma^2 \{1 - \rho(u)\} \qquad (17)$$

where σ_e^2 is the variance of the component of ε_{ijk} due to measurement errors within each offspring. This component is estimable only when *at least* duplicate measurements are taken at each time instant.

The average correlation over the k measurements for each individual is given by

$$\bar{\rho} = [\Sigma\ (k - r)\ \rho\ (r)] / \Sigma\ (k - r) \qquad (18)$$

the summation being over r from 1 to k–1.

The average variogram over the k measurements for each individual is given by

$$\bar{\gamma} = [\Sigma\ (k - r)\ \gamma\ (r)] / [\Sigma\ (k - r)] \qquad (19)$$

the summation being over r with the usual limits from 1 to $(k - 1)$, the relation between $\bar{\rho}$ and $\bar{\gamma}$ being given by

$$\bar{\gamma} = \sigma_e^2 + \sigma^2\ (1 - \bar{\rho}) \qquad (20)$$

(b) Random effects

When we consider a particular time instant, the longitudinal data becomes cross-sectional with a design following half-sib structure. This gives rise to two variance components σ_m^2 and σ_o^2 corresponding to male and offspring effects respectively. In addition, if we visualize that a number of measurements is made instantaneously at the *same* time point on each individual offspring's each repeated measurement, the total *within* offspring variability would be, say $\sigma_{wo}^2 = \sigma^2 + \sigma_e^2$. This would have two components (a) $\sigma^2\ \bar{\rho}$, the *uniform* covariance between pairs of duplicate measurements within each measurement at the given time instant or the variance of the means of measurements, and (b) $\sigma^2(1 - \bar{\rho}) + \sigma_e^2$, the within measurement component of variance due to random errors of the measuring instrument.

The variance of each ε_{ijk} is then finally given by

$$\text{Var}\ (\varepsilon_{ijk}) = \sigma_m^2 + \sigma_o^2 + \sigma^2 + \sigma_e^2 = \sigma_T^2, \text{ say} \qquad (21)$$

The covariance matrix will thus involve five parameters – the four variances plus one correlation function.

4.2 Covariance Matrix

The $nk \times nk$ matrix $\sum$ will have a block structure of sub-matrices as depicted below

$$\Sigma = \begin{pmatrix} \Sigma^* & \sigma_m^2 I_k & \sigma_m^2 I_k & \cdots & \sigma_m^2 I_k \\ \sigma_m^2 I_k & \Sigma^* & \sigma_m^2 I_k & \cdots & \sigma_m^2 I_k \\ \sigma_m^2 I_k & \sigma_m^2 I_k & \Sigma^* & \cdots & \sigma_m^2 I_k \\ \vdots & & & & \\ \sigma_m^2 I_k & \sigma_m^2 I_k & \sigma_m^2 I_k & \cdots & \Sigma^* \end{pmatrix} \qquad (22)$$

where I_k is an $k \times k$ identity matrix and the $k \times k$ matrix Σ^* is given by

$$\Sigma^* = \begin{pmatrix} \sigma_T^2 & \sigma_s^2 + \sigma^2\rho(1) & \sigma_s^2 + \sigma^2\rho(2) & \cdots \sigma_s^2 + \sigma^2\rho(k-1) \\ \sigma_s^2 + \sigma^2\rho(1) & \sigma_T^2 & \sigma_s^2 + \sigma^2\rho(1) & \cdots \sigma_s^2 + \sigma^2\rho(k-2) \\ \sigma_s^2 + \sigma^2\rho(2) & \sigma_s^2 + \sigma^2\rho(1) & \sigma_T^2\ \sigma_s^2 + \sigma^2\rho(1) & \cdots \sigma_s^2 + \sigma^2\rho(k-3) \\ \vdots & \vdots & \vdots & \vdots\ \vdots \\ \sigma_s^2 + \sigma^2\rho(k-1) & \sigma_s^2 + \sigma^2\rho(k-2) & \cdots & \sigma_T^2 \end{pmatrix} \qquad (23)$$

where $\sigma_s^2 = \sigma_m^2 + \sigma_o^2$. It may be noted that when there is no autocorrelation i.e. $\sigma^2 = 0$ and $\rho(u) = 0$, Σ^* reduces to

$$\sigma_e^2 I_k + J_k^T \sigma_s^2 J_k \qquad (24)$$

where J_k is a row vector of k dimensions. This situation is common in animal breeding examples and gives rise to estimation of repeatability of a character. When, however, the character is not repeated over time so that $k=1$ and σ_o^2 is also zero, this matrix reduces to the scalar $\sigma_m^2 + \sigma_e^2 = \sigma_T^2$ and the reduced $n \times n$ matrix $\sum$ then has σ_T^2 as diagonal elements and σ_m^2 as off diagonal elements, a typical case of half-sib analysis of variance for the estimation of heritability of a character.

4.3 Relation between Observational and Causal Components of Variance

The variance components discussed above are *observational* components that are estimated from the data generated during the study and need to be related to *causal* components derived from genetic considerations based on the genetic design adopted for the study (Narain 1990). Let the additive genetic variance, additive × specific or local environment interaction variance, specific or local environmental variance, general or common environmental variance including non-additive genetic components of variance, environmental variance (including non-additive genetic components) and phenotypic variance be denoted respectively by V_A, V_{AEs}, V_{Es}, V^*_{Eg}, V_E and V_P. Then we have the relations as given below

$$\begin{aligned} &1.\ \sigma_m^2 = (¼)\, V_A \\ &2.\ \sigma_o^2 = (¾)\, V_A + V^*_{Eg} \\ &3.\ \sigma_s^2 = \sigma_m^2 + \sigma_o^2 = V_A + V^*_{Eg} = V_G + V_{Eg} \\ &4.\ \sigma_{wo}^2 = \sigma^2 + \sigma_e^2 = V_{AEs} + V_{Es} \qquad (25) \\ &5.\ \sigma_{wo}^2\ \bar{r}_o = \sigma^2\ \bar{\rho} = (¼)\, V_{AEs} \\ &6.\ \sigma_{wo}^2 (1 - \bar{r}_o) = \sigma^2(1 - \bar{\rho}) + \sigma_e^2 = (¾)\, V_{AEs} + V_{Es} \\ &7.\ \sigma_T^2 = V_P = V_A + V_{AEs} + (V_{Es} + V^*_{Eg}) \\ &\qquad\quad = V_A + V_{AEs} + V_E \end{aligned}$$

where $\bar{r}_o = (\sigma^2\ \bar{\rho} / \sigma_{wo}^2)$ denotes the intra-offspring correlation based on the *average* autocorrelation $\bar{\rho}$ in the same spirit as discussed for $\bar{r}$ in Section 3. However, Sukhatme's model is not the same as is considered in this section. His model has a Markovian structure for errors as given by equation (3) and therefore, cannot be extended to accommodate measurement errors and random effects of the half-sib analysis. Here, the serial correlation process is not Markovian so as to allow the variance component approach to be adopted. The algebraic results for serial correlation, however, happen to be the same in both the cases.

In the above relations, the male effect variance (σ_m^2) is equal to one-quarter of the additive genetic variance (V_A) since each male passes half of his genes to each offspring. When squared to compute the variance, the one-half additive genetic effect (breeding value) becomes one-quarter of the additive genetic variance. The offspring effect variance (σ_o^2) will therefore include the remaining three-fourth of the additive genetic variance, plus a variance (V^*_{Eg}) due to general or common environmental effect that includes non-additive genetic effects as well as environmental effects common to all observations of the offspring. The within offspring variability (σ_{wo}^2) includes the variance component (σ^2) due to serial

correlation, plus the error component of variance (σ_e^2), only estimable, as already noted above, if *at least* duplicate observations are recorded at each time point instantaneously. Due to longitudinal nature of data, the repeated observations taken at each time point for the same offspring will include local environmental effects due to time parameter with a variance (V_{Es}), plus an interaction effect due to additive genetic effects of the offspring interacting with the local environmental effect with a variance (V_{AEs}). As already discussed in the section on random effects, the within offspring variance has two components – one due to the part explained by the serial correlation ($\sigma^2\ \bar{\rho}$) and the other containing the remainder, plus that due to errors of measurement ($\sigma^2(1-\bar{\rho}) + \sigma_e^2$). The former is equal to one-quarter of the variance due to additive × local environmental effects ((¼) V_{AEs}) since each offspring carries only one-half of the male's genes and on computing the variance it becomes one-fourth. The other component will therefore include the remaining three-fourth of the interaction variance ((¾) V_{AEs}), plus the variance due to local environmental effects (V_{Es}). In this way, it may be noted, the balance sheet of total variability, at the observational as well as causal levels, is made intact with $\sigma_T^2 = V_P$.

4.4 Heritability

With longitudinal data, the heritability of the trait is of two kinds – one, the usual one used in quantitative genetics literature, is at the population level and the other, introduced in the Section 3, is at the individual level.

(a) Heritability at the population level

It is the fraction of the phenotypic variance (V_P) which is due to the additive effects of genes and can take any value between 0 and 1. Symbolically

$$h^2 = V_A / V_P \tag{26}$$

This heritability is stated to be in a *narrow sense*. We can easily see that it can be estimated by

$$h^2 = 4\,\sigma_m^2 \ /\ \sigma_T^2 = 4\,\sigma_m^2 \ /\ (\sigma_m^2 + \sigma_o^2 + \sigma^2 + \sigma_e^2) \tag{27}$$

We have therefore to estimate the four variance components σ_m^2, σ_o^2, σ^2 and σ_e^2 from the longitudinal data for estimating this heritability.

(b) Heritability at the individual level

It is the fraction of the total variance ($V_{AEs} + V_{Es}$) at the *individual level* given phenotypically by ($\sigma^2 + \sigma_e^2$) which is due to the additive × local environmental interaction effects and can take any value between 0 and 1. Symbolically

$$h_w^2 = V_{AEs} / (V_{AEs} + V_{Es}) \tag{28}$$

It is given the name *intra-individual heritability* or *heritability of the individual* in the same spirit as in Section 3. However, here it is defined in a *narrow sense* whereas in Sukhatme's model it is defined in a *broad sense* and given by (12). The difference is obvious since in that model we do not have a mating design like half-sib design used here and the data generated there gives only between individual variability that includes genetic variance (V_G) and common environmental variance (V_{Eg}). Here the half-sib design allows the additive genetic component of variance (V_A), a subdivision of V_G, to be included into the analysis. This heritability can be estimated by

$$\begin{aligned} h_w^2 = 4\ \bar{r}_o &= 4\,(\sigma^2\ \bar{\rho} / \sigma_{wo}^2) \\ &= 4\,[\sigma^2\ \bar{\rho} / (\sigma^2 + \sigma_e^2)] \end{aligned} \tag{29}$$

or else, in terms of variogram, by

$$h_w^2 = 4\,[1 - \bar{\gamma} / (\sigma^2 + \sigma_e^2)] \tag{30}$$

However, we need a model for $\rho(u)$ or $\gamma(u)$ to be able to determine $\bar{\rho}$ or $\bar{\gamma}$. If we take

$$\rho(u) = \rho^u \text{ with } |\rho| < 1 \tag{31}$$

the type studied under Sukhatme's model in Section 3, $\bar{\rho}$ is given by

$$\begin{aligned} \bar{\rho} &= [2\rho/(k-1)(1-\rho)][1-(1-\rho^k)/k(1-\rho)] \\ &\simeq 2\rho/[k(1-\rho)] \end{aligned} \tag{32}$$

This is substituted in (29) before estimating h_w^2. The behaviour of $\bar{\rho}$ that depends on k and ρ in a characteristic manner is as shown in Fig. 9.1 of Narain (1990).

We thus see that to be able to estimate h_w^2, we will need to estimate ρ, σ^2, and σ_e^2, k being given. It is also important to see that if the process variance σ^2 is found to be very small compared to σ_e^2 or σ_o^2, the increasing curve of the variogram is squeezed between the two horizontal lines corresponding to σ_o^2 and σ_e^2 and therefore disappears. That means h_w^2 becomes negligible,

indicating thereby that additive × local environmental effects are not present. The LDA then reduces to the usual type of the analysis of variance with variation over time within offspring providing an estimate of the experimental error variance σ_e^2 and hence as a measure of V_{Es} only. It is therefore apparent that the non-existence of additive × local environmental interaction effects implies non-existence of serial correlation and vice versa. LDA, in the context of a half-sib mating design, therefore provides with a test for the presence or absence of such interaction effects.

4.5 Heritability of *k* Repeated Measurements

When we have *k* repeated measurements on a an individual of the population following the above discussed model of serial correlation, we can determine the heritability (h_w^2) of the mean of the *k* phenotypic measurements by working out the regression coefficient of the breeding value (*A*) of the individual on the mean $\bar{P} = \Sigma P_i / k$, the summation being over i = 1 to i = *k*. Then

$$\text{Cov}(A, \bar{P}) = (1/k)\,\Sigma\,\text{Cov}(A, P_i) = V_A \qquad (33)$$

$$\begin{aligned}\text{Var}(\bar{P}) &= (1/k^2)\,[\text{Var}(\Sigma P_i)] \\ &= (1/k^2)[\,k^2\,\sigma_S^2 + k(\sigma^2 + \sigma_e^2) \\ &\quad + k(k-1)\,\sigma^2\,\bar{\rho}] \\ &= [\,\sigma_S^2 + \sigma_{wo}^2\,\{1 + (k-1)\,h_w^2\}/k] \\ &= [RV_P + (1-R)V_P\{1 + (k-1)\,h_w^2\}/k] \qquad (34)\end{aligned}$$

where R is the repeatability of the character given by

$$R = (\sigma_S^2/\sigma_T^2) = (V_A + V^*_{Eg})/V_P = (V_G + V_{Eg})/V_P \qquad (35)$$

Then

$$\begin{aligned}h_w^2 &= \text{Cov}(A, \bar{P})/\text{Var}(\bar{P}) \\ &= k\,h^2/[\{1 + (k-1)R\} + (k-1)(1-R)\,h_w^2] \qquad (36)\end{aligned}$$

When $h_w^2 = 0$, the expression for h_w^2 reduces, as it should, to that given by (9.29), with 'n' replaced by 'k', of Narain (1990) for the heritability of the repetitions of the same character without involving serial correlation. When $k = 1$, it becomes h^2 as it should. The accuracy in determining the breeding value of the individual is thus increased with longitudinal data with serial correlation.

5. CORRELATION BETWEEN THE BREEDING VALUE OF A MALE AND THE AVERAGE VALUE OF HIS OFFSPRING'S MEASUREMENTS (PROGENY TEST)

The *breeding* value, denoted by *A*, of an individual's phenotypic trait is an important concept in quantitative genetics literature. When an individual is mated to a number of individuals taken at random from a given population and an offspring is scored for the character from each mating, the mean deviation of the offspring's value from the population mean measures half the breeding value (½ *A*) of the parent (Narain 1990). In dairy cattle and buffalo breeding this forms the basis of *progeny testing* where a finite number of daughter's milk records are used to assess the breeding worth of the sire who does not express the milk characteristic. The square of the correlation between *A* and the progeny test- the average of the number of daughter's records – termed the *heritability* of the progeny test and denoted by h_{pr}^2 tends to unity as the number of progenies is increased indefinitely.

When we have several, say *k*, measurements over time on each offspring of the male, we can work out the h_{pr}^2 for each measurement separately, the usual formula being given by $n/(n + a)$ based on the equation (12.14) of Narain (1990) where $a = (4 - h^2)/h^2$. This can be summed up over *k* measurements and divided by *k* to give the progeny test on per measurement basis. Since each measurement is contributed by the same set of genes, the formula remains same as that for an individual measurement. This, however, does not take into account the covariance structure of the longitudinal data. In what follows, we therefore, examine the correlation between *A* and the progeny test when the covariance structure of the longitudinal data is taken into account.

Let the breeding value of the individual for a given trait, expressed as deviation from the population mean, be denoted by *A*. Let the phenotypic values of the trait of *n* offsprings from this male with *k* measurements over time, expressed again as deviation from the same population mean and denoted by D_{ij} , i = 1, 2 ... *n*, j = 1, 2, ... *k* be standardized to have unit variance so that the heritability of the trait, denoted by h^2 is the same as additive genetic variance V_A. We assume that for each offspring of a given male, same genes affect the character at the *k* different time points with the possibility of

additive × genotype interactions. The covariance structure discussed above and given by (22) and (23) indicates the necessary random effects due to male, progeny and measurement errors as well as the serial correlation effects. We consider the mean of the k measurements of each offspring, say $\bar{D}_i$ for the i-th offspring, with heritability given by (36). The breeding value of an individual with respect to this mean would have variance equal to this heritability times the Var ($\bar{D}_i$) that is same for all $i = 1, 2, \ldots n$. Then for $i = 1, 2, \ldots n$, we have

$$\text{Cov (A, } \bar{D}_i) = (1/2)\, h_k^2 \text{ Var } (\bar{D}_i) \tag{37}$$

$$\text{Cov } (\bar{D}_i, \bar{D}_j) = (1/4)\ h_k^2 \text{ Var } (\bar{D}_i) \tag{38}$$

$$\text{Var } (\bar{D}_i) = [R + (1 - R)\{1 + (k - 1)\, h_w^2\ \}/k] \tag{39}$$

Let the average of all the phenotypic values for the means of the offspring of a given male, n in number, be denoted by $\bar{D}$ so that

$$\bar{D} = \Sigma \bar{D}_i / n \tag{40}$$

We then have

$$\begin{aligned} \text{Cov } (\bar{D}, A) &= (1/n)\, \Sigma \text{ Cov } (\bar{D}_i, A) \\ &= (1/2)\, h_k^2 \text{ Var } (\bar{D}_i) \end{aligned} \tag{41}$$

$$\begin{aligned} \text{Var } (\bar{D}) &= (1/n^2)\, \Sigma \text{ Cov } (\bar{D}_i, \bar{D}_j) \\ &= (1/n)[1 + \{(n - 1)/4\}\ h_k^2\ ] \text{Var } (\bar{D}_i) \end{aligned} \tag{42}$$

$$\text{Var } (A) = h_k^2 \text{ Var } (\bar{D}_i) \tag{43}$$

Then the expression for the heritability of progeny test for LDA is obtained as

$$\begin{aligned} h_{pr}^2 \text{ (LDA)} &= [\text{Cov}(\bar{D}, A)]^2/[\text{Var}(\bar{D})\ \text{Var}(A)] \\ &= (n\, h_k^2/4)/[1 + (n - 1)\, h_k^2\ /4] \\ &= [n/(n + a_w)] \end{aligned} \tag{44}$$

where

$$a_w = a - 4(k - 1)(1 - h_w^2\ )(1 - R)/kh^2 \tag{45}$$

$$a = (4 - h^2)/\, h^2 \tag{46}$$

If we take $k = 1$, a_w equals a and h_{pr}^2 (LDA) becomes h_{pr}^2, the well known formula for the heritability of the progeny test (Narain 1990). When n tend to infinity h_{pr}^2 (LDA) tends to unity showing that the correlation attains the maximum value of one, thus preserving the well known result of h_{pr}^2. Further, since the extra term in a_w over and above the term a viz. $4(k-1)(1 - h_w^2\ )(1-R)/kh^2$ is necessarily positive, a_w is *less* than a, indicating that the heritability of the progeny test with longitudinal data is increased compared to its value with cross-sectional data. However, when V_{Es} is very small, making R very large and close to one or else h_w^2 is close to one, the a_w is close to a and the heritability with longitudinal data is nearly the same as that with cross-sectional data. It is only when R is low or h_w^2 has a small value that h_{pr}^2 (LDA) has advantage over h_{pr}^2 . It may be noted that h_w^2 depends on the process variance σ^2 and average serial correlation $\bar{\rho}$ by the relation (29) for which the model for $\rho(u)$ given by (31) is invoked.

An alternative form for the expression for the heritability of progeny test with longitudinal data, by substituting for h_k^2 from (36) and for R from (35), is given by

$$\begin{aligned} h_{pr}^2 \text{ (LDA)} &= (nkh^2/4)/[1+\{(n-1)k/4 \\ &\quad + (k-1)(1 - h_w^2\ )\}h^2 + (k-1)\ h_w^2 \\ &\quad + (k-1)(1 - h_w^2\ )V^*_{Eg}] \end{aligned} \tag{47}$$

6. ESTIMATION

The parameters for the estimation in the general linear model with correlated errors given by (15) and (16) are β, a vector of p dimensions, and the matrix $I_m \otimes \Sigma$ that is a function of five parameters, σ_m^2, σ_o^2, σ^2, σ_e^2 and ρ. While β, pertains to fixed effects like herd, season of calving, order of lactation etc., in the animal breeding example, the covariance matrix involves components of variance and serial correlation coefficient. We denote the latter by a vector α of 5 dimensions. The covariance matrix is then denoted by $\sigma_T^2 V(\alpha)$ where each element of the matrix has been divided by the total variance. The most versatile method of estimation is that of *restricted maximum likelihood* (REML) introduced by Patterson and Thompson (1971) in connection with variance components estimation. Here it is to be adopted for general linear model with correlated errors. For given α, the estimating equations for the fixed effects are

$$\hat{\beta}(\alpha) = (X^T V(\alpha)^{-1} X)^{-1} X^T V(\alpha)^{-1} y \quad (48)$$

Then the REML estimator for σ_T^2 is given by

$$\hat{\sigma}_T^2(\alpha) = RSS(\alpha) / (mnk - p) \quad (49)$$

where

$$RSS(\alpha) = \{y - X\hat{\beta}(\alpha)\}^T V(\alpha)^{-1} \{y - X\hat{\beta}(\alpha)\} \quad (50)$$

The REML estimate $\hat{\alpha}$ maximizes

$$L(\alpha) = -1/2 [mnk \log \{RSS(\alpha)\} + \log | V(\alpha) | + \log | X^T V(\alpha)^{-1} X|] \quad (51)$$

and the resulting REML estimate of β is $\hat{\beta} = \hat{\beta}(\hat{\alpha})$.

7. DISCUSSION

LDA requires more sophisticated statistical methods than that for cross-sectional data analysis. Here we have demonstrated this by discussing two specific areas of application viz. nutritional studies and animal breeding data analysis.

We have built up over the method given by Sukhatme and Narain on *intra-individual variability* in nutrition studies. By utilizing the characteristics of the longitudinal data in terms of serial correlation and associated components but without invoking the Markov process, it has been possible to give the concept of *intra-individual heritability* in a *narrow sense*. It turns out that this heritability is a function of the serial correlation coefficient and enables one to test for the existence or otherwise of the additive × local environmental interaction effects.

We have used a model for serial correlation function, given by (31), as $\rho(u) = \rho^u$ with $|\rho|<1$, so that it decreases as the distance in time, u, increases. This is a special case of the exponential correlation function model given by

$$\rho(u) = \exp(-\kappa u^\nu) \quad (52)$$

where $\rho = \exp(-\kappa)$ with $\nu = 1$. This more general model is particularly useful when time parameter is continuous.

When confronted with LDA, the first step is to explore the association among repeated observations for an individual to determine the type of serial correlation model to be used. This is exploratory data analysis (EDA) to visualize the patterns in data. The effects of explanatory variables, if any, are first removed by regressing the response **y** on **X** and residuals r_{ij} are obtained. Scatter plot matrix in which r_{ij} is plotted against r_{il} for all $j < l = 1, 2, \ldots k$ is then obtained. Such a graphical display can indicate the nature of correlation between repeated observations and the manner in which it decreases over time. This EDA is then followed by confirmatory analysis as required.

Of particular interest, in this paper, is the accuracy of the progeny test with longitudinal data, given by the square-root of the expression (44) or (47) that seems to be new in the literature of quantitative genetics. While for $k = 1$ they reduce to the familiar result in animal breeding, when k tends to be large at a given value of n, they tend to the limit given by

$$h^2_{pr}\ (\text{LDA}, k \to \infty) = n/[n + a - 4\ (1 - h^2_w\ )(1 - R)/h^2\ ]$$

$$= (nh^2/4)/[\{(n-1)\ h^2/4\} + h^2_w\ + R(1 - h^2_w\ )] \quad (53)$$

When h^2_w is close to one, this becomes h^2_{pr} but when h^2_w is zero this becomes $(nh^2/4)/[(nh^2/4) + \sigma^2_o\ ]$ if R is expressed as $(h^2/4 + \sigma^2_o\ )$. In the later case it indicates the effect of repeatability when the measurements repeated indefinitely provide with a perfect estimate of σ^2_e. It should be noted that in order that the effects of serial correlation and random effects due to male and offspring are distinguishable, k should be greater than two.

ACKNOWLEDGEMENTS

This work was supported by the Indian National Science Academy, New Delhi under their program 'INSA Honorary Scientist'.

REFERENCES

Narain, P. (1982). Genetic interpretation of the auto-regulatory mechanism in nitrogen balance. *Proc. Golden Jubilee Conference on Human Genetics and Adaptation*, I.S.I., Calcutta, 1-10.

Narain, P. (1984). On contributions of P.V. Sukhatme in the field of nutrition. In: *Impact of P.V. Sukhatme on Agricultural Statistics & Nutrition*, Ed. P. Narain, I.S.A.S., Delhi, 24-45.

Narain, P. (1990). *Statistical Genetics*. John Wiley, New York (Wiley Eastern Ltd., New Delhi, reprinted in 1993. Published by the New Age International Pvt. Ltd., New Delhi in 1999).

Narain, P. (1993). Interface among statistics, cybernetics and genetics. Dr. Rajendra Prasad Memorial Lecture delivered during 46th Annual Conference of Indian Society of Agricultural Statistics at Bhubaneswar. *J. Ind. Soc. Agril. Statist.,* **45,** 48-75.

Narain, P. (1998). Genetic relationship underlying the intra-individual variation in nutrition studies. P.V. Sukhatme Memorial Volume. *J. Ind. Soc. Agril. Statist.,* **51(2 & 3),** 129-134.

Narain, P. (2000). Biographical memoir of P.V. Sukhatme. In: *Biographical Memoirs of Fellows* of The Indian National Science Academy, **22,** 155-181.

Patterson, H. D. and Thompson, R.R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58,** 545-554.

Sukhatme, P.V. and Margan, S. (1978). Models for protein deficiency. *Amer. J. Clinical Nutrition*, **31,** 1237-1256.

Sukhatme, P.V. and Narain, P. (1982). A possible genetic interpretation of auto-regulatory mechanism in models for protein deficiency. *Proc. Indian National Science Academy,* **B48,** 748-754.

Sukhatme, P.V. and Narain, P. (1983). Intra-individual variation in energy requirement and its implications. *Ind. J. Med. Res.,* **78,** 857-865.

Sukhatme, P.V. and P. Narain (1984). The genetic significance of intra-individual variation in energy requirement. In : *W.G. Cochran's Impact on Statistics*, Ed. P.S.R.S. Rao and J. Sedransk, John Wiley & Sons, New York.

Sukhatme, P.V. and Narain, P. (1996-97). Intra-individual variation in energy requirement and its genetic significance. *J. Ind. Soc. Agril. Statist.*, **49** (Golden Jubilee Number), 1-10.