Research Journal of Mathematical and Statistical Sciences _____ISSN 2320-6047
Vol. **6(6),** 1-5, June **(2018)** *Res. J. Mathematical and Statistical Sci.*

Review Paper

A study of effective statistical tools for longitudinal data analysis

K.N. Krishnamurthy* **and K.B. Murthy**
Department of Ag. Statistics, Applied Mathematics and Computer Science, UAS, GKVK, Bengaluru, India
kkmurthy13@gmail.com

Available online at: www.isca.in, www.isca.me
Received 9th March 2018, revised 19th May 2018, accepted 10th June 2018

Abstract

Longitudinal studies play a very important role in human life, plant science and social sciences. In such studies, data are collected from the respondents over a period of time or periodical intervals. Consequently, observations are correlated and effective statistical methods/techniques are required for the analysis of such data. Other names given to such studies are the analysis of repeated measurements and growth curves. The main focus of such data analysis is to study the changes caused by development, aging and other factors such as application of different treatments over a period of time. Such studies typically have unbalanced designs, missing data and time varying covariates and thus not tenable to standard statistical methods. This paper gives an overview of literature and important references which lead for further effective studies.

Keywords: Longitudinal study, repeated measures, growth curve model.

Introduction

Longitudinal studies are defined as the studies in which response of each individual is observed over a period of long time. These studies offer investigator an opportunity of controlled and uniform measurement of exposure history and other factors related to effective outcome. Such studies are particularly useful when one is interested in studying the changes over time due to development, aging and other factors such as application of different treatments which affect changes. Studies of this type have broad application, especially in the life and social sciences. As examples, most clinical trials of new pharmaceutical drugs, different agronomic investigations and business surveys are characterized by repeated measurements over time on the basic experimental units.

The longitudinal studies give more efficient estimates, comparisons and predictions than the corresponding cross sectional studies with the same number of observations. Recently, methods have been developed to accommodate special complexities of these studies arising due to auto correlations, missing observations and time varying covariates. If the main purpose is to study the variability over units the longitudinal studies may not be very efficient. These studies are also time consuming while several 'classical' statistical methods exist and are very useful, application oriented of these methods regardless of underlying assumptions in common.

In addition, confusion between multivariate and univariate repeated measures approaches, the distinction between growth curves and repeated measures models which are seldom clear even in the minds of professional statisticians. An attempt is made to distinguish them in the following way:

Repeated measure studies: The basic goal of these studies is to detect differential treatment effects or factors or combination of factor levels at differential times. Here the investigator is interested in knowing whether the treatments differ as a whole over the entire period of time or not. Attention is focused on tests of significance between treatments, rather than on the relationship between the effects at times, when treatments were applied.

Growth curve studies: Generally there is a relationship between treatment effects and time, but the fundamental relationship is often neither stated nor derived explicitly. This function may be approximated by a polynomial structure. In such studies, the coefficients in the polynomial representation usually have physical interpretations. These parameters, along with variance and covariance, must be estimated from the data on such studies. Test of significance, hypothesis testing, as well as predictions are of interest to investigator.

A distinction between repeated measurement and growth curve models is the analogues to the perceived difference between analysis of variance (ANOVA) and regression models. However, ANOVA may be viewed as a case of regression, and the repeated measurements model is a particular case of growth curve model.

Repeated measurement data analysis

Under a very simplified version of the model let y_{ii} be the j^{th} measurement of the i^{th} individual then we may write the model as

$$y_{ij} = \mu_{ij} + \alpha_{ij} + e_{ij}$$

where, μ_{ii} is the j^{th} measurement on the i^{th} individual, α_{ii} is the effect of the i^{th} individual at the j^{th} occasion and varies independently over the population and e_{ii} is an independent error. If all individuals are on equal footing, then one can drop subscript 'i' and take μ_{ii} as μ_i. But, otherwise, when they are on different treatment groups, then the subscript 'i' indicate the special treatment factor affecting the observation of that individual. The vector $\mu_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{ip})'$ corresponds to the 'p' measurements on the i^{th} individual, and will be referred to as the μ - profile of that individual. One may take the following assumptions:

$$E(\alpha_{ij}) = 0, \;\; V(\alpha_{ij}) = \sigma^2_{\alpha_{ij}}, \;\; Cov(\alpha_{ij}, \alpha_{ij'}) = 0$$

$$E(e_{ij}) = 0, \; V(e_{ij}) = \sigma^2_{ij}, \;\; Cov(e_{ij}, e_{ij'}) = 0, \; Cov(\alpha_{ij}, e_{ij}) = 0$$

and α_{ij} , e_{ij} are normally distributed. This gives

$$Cov(y_{ij}, y_{ij'}) = \sigma_{\alpha_{ij}} \;\; , \;\; V(y_{ij}) = \sigma_{\alpha_{ij}} + \sigma^2_{ij}$$

This indicate that while observations from different individuals are independent, but there is correlation among the observations within each individual, and can be stated as

$$\rho = \frac{\sigma_{\alpha_{jj'}}}{\sqrt{(\sigma_{\alpha_{jj'}} + \sigma^2_{e_j})(\sigma_{\alpha_{jj'}} + \sigma^2_{e_{j'}})}}$$

and under the assumption that

$$\sigma^2_{e_j} = \sigma^2_e \; and \;\; \sigma^2_{\alpha_{jj'}} = \sigma^2_\alpha \, , \; correlation \;\; is \;\; given \;\; as$$

$$\rho = \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \sigma^2_e} \qquad (1)$$

which ranges from zero to one, and is known as intra-class correlation. The above model is known as mixed model.

For example, suppose that there are 'g' groups of animals each on different treatments, and there are n_j (i = 1, 2, …) animals in the i^{th} group such that $\sum_{i=1}^{g} n_i = n.$ Suppose weekly record on growth over certain weeks for each animal are available and it is of interest to compare the 't' treatments for the growth over 'p' weeks.

Split plot analysis: Under the compound symmetry model in (1), groups may be considered as main plot treatments, weeks within treatment may represent sub plot treatments and analysis of data proceeds on the line of split plot design.

F-statistic is valid if compound symmetry in (1) hold. According to Mauchly[1] there exists test for testing the assumption of compound symmetry. For a general description see mixed models by Scheffe[2]. Some recent results on the test of compound symmetry are established by Grieve[3].

Table-1: Split plot analysis of variance.

Source	Degrees of freedom
Groups	g – 1
Within groups	n – g
Weeks	p - 1
Weeks X groups	(p-1)(g – 1)
Weeks X Within groups	(p – 1)((n – g)
Total	np - 1

Multivariate Models without special co-variance structure: When the assumption of compound symmetry in (1) does not hold, then one may have to consider a general covariance matrix Σ of the order 'p'.

$$V(Y) = \sum = (\sigma_{jj'})$$

and the testing hypothesis about the components of mean 'μ_p' becomes the usual multivariate analysis of variance (MANOVA) problem. Here the main problem is to test the equality of mean vectors 'μ'.

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_g$$

for different treatment groups. Another problem is to test for certain linear combination of μ over treatment groups. For example, one may be interested whether growth rate over 'p' weeks is same over all the groups or not. Such hypothesis can be tested by usual MANOVA techniques. For more details see Rao[4].

Growth curve model

Suppose that there are 'g' different groups (treatments) and a single growth variable 'y' is measured at 'p' time points t_1 , t_2 , … t_p on n_j specimens chosen from the j^{th} group (j = 1, 2, … g). We specify a polynomial of degree (q – 1) for y on time interval 't'. Thus, for the j^{th} group.

$$E(y_t) = \xi_{j0} t^0 + \xi_{j1} t^1 + \ldots + \xi_{jq-1} t^{q-1} \qquad for \;\; (p > q-1)$$

Let $\xi'_j = (\xi_{j0}, \xi_{j1}, \ldots \xi_{jq-1})$ denote the vector of the regression coefficients (or the growth curve coefficients) for the j^{th} group. The observations $y_{t_1}, y_{t_2}, \ldots y_{t_p}$ being on the same specimen are correlated, and we shall denote their variance-

covariance matrix by Σ. We assume that Σ is same for all 'g' groups. Let Y_j denote the p x n_j matrix of the observations for the j^{th} group and let.

$Y_{p \times N} = (Y_1 , Y_2 , \ldots Y_g)$ with $N = n_1 + n_2 + \ldots + n_g$ be the total number of units then

$$E(Y_j) = \lfloor B\xi_{j1}, B\xi_{j2} , \ldots B\xi_{jg} \rfloor$$

$$where \quad B = \begin{bmatrix} t_1^0 & t_1^1 & \ldots & t_1^{q-1} \\ t_2^0 & t_2^1 & \ldots & t_2^{q-1} \\ \ldots & \ldots & \ldots & \ldots \\ t_p^0 & t_p^1 & \ldots & t_p^{q-1} \end{bmatrix}$$

$$and \quad E(Y) = E(Y_1 , Y_2 , \ldots Y_g)$$
$$= \left[\left| B\xi_1, \ldots, B\xi_1 \right|, \ldots, \left| B\xi_g, \ldots, B\xi_g \right| \right]$$
$$= B\xi A$$
$$where \quad \xi = \left| \xi_1, \ldots, \xi_g \right| \quad and \quad A_{g \times N} = Diag\left[E_{|n|}, \ldots, E_{|ng|} \right]$$
$$V(Y) = I_N \otimes \sum$$

Some times, it is convenient to take matrix B to be

$$B_0 = \begin{bmatrix} p_0^{t_1} & p_1^{t_1} & \ldots & p_{q-1}^{t_1} \\ p_0^{t_2} & p_1^{t_2} & \ldots & p_{q-1}^{t_2} \\ \ldots & \ldots & \ldots & \ldots \\ p_0^{t_p} & p_1^{t_p} & \ldots & p_{q-1}^{t_p} \end{bmatrix}$$

where $p_l(t)$ are orthogonal polynomials of degree l. At this point some historical comments may be appropriate. Wishart[5] is credited for using such growth curve models for the first time. The analysis begins by replacing large number of observations on each individual by a few coefficients of orthogonal polynomial. The mean of growth rate and its change are compared by univariate analysis of variance. Box[6] suggested the use of analysis of variance to the first differences of successive observations when the assumption of uniform co-variance matrix is valid. On the growth curve analysis, the object is to make inference about ξ and to test the adequacy of the degree of polynomial (q-1); estimate ξ and obtain variance co-variance matrix of ξ ; to test linear hypothesis about ξ,

i.e., H_0: L ξ M = 0
H_0: $\xi_1 = \xi_2 = \ldots = \xi_k$

Or to test sub hypothesis about ξ_1 and to make confidence intervals. Potthof and Roy[7] obtained the maximum likelihood estimator of ξ as

$$\bar{\xi} = (B' \sum^{-1} B)^{-1} (B' \sum^{-1} \bar{Y})$$
$$where \quad \bar{Y} = (\bar{Y}_1 , \bar{Y}_2 , \ldots , \bar{Y}_g) \quad and \quad \sum \quad is \quad known.$$

Khatri[8] obtained the maximum likelihood estimator when Σ is not known and is given by,

$$\hat{\xi} = (B' S^{-1} B)^{-1} (B' S^{-1} \bar{Y})$$
$$where \quad S = \sum (Y_j Y_j^{'} - n_j \bar{Y}_j \bar{Y}_j^{'})$$

and estimate of the variance of $\hat{\xi}$ is given by

$$V(\hat{\xi}) = k \; Diag\left(\frac{1}{n_1} , \frac{1}{n_2} , \ldots , \frac{1}{n_g} \right) \otimes (B' S^{-1} B)^{-1}$$
$$with \quad k = \left(\frac{N-g-1}{N-g-p+q-1} \right) (N-g-p+q)$$

For all testing problems MANOVA can be applied. When there is only one group (g = 1) then,

$$S = YY' - n\bar{Y}\bar{Y}'$$
$$\xi = (B' S^{-1} B)^{-1} (B' S^{-1} \bar{Y})$$

Hypothesis of the type Σ ξ = 0 can be tested by using Hotelling's T^2 statistic, which yields F-test. For details see Potthoff and Roy[7], Khatri[8] and Geisser[9].

Multidimensional growth curve model: In this, several response variables are simultaneously measured at different time points. For example, in practice both systolic and diastolic blood pressures or height and weight are included in medical study. Potthoff and Roy's model for one growth variable was generalized by Reinsel[10].

Time moving covariate model: Patel[11] introduced a growth curve model useful in repeated measurement designs where the covariate changes with respect to time. This arises in long term clinical trials where environment factors *viz*., air pollution, diet, exercise and smoking influence the diseases. Quantitative information on such influencing factors over the periods, between successive clinical visits, is utilized in the model. In clinical trials we often find two types of covariates, the first being between patient covariate, which remains unchanged during the trial such as initial weight, age, etc., while the second type of covariate called within patient covariate, varies to the successive next visit for the same patient. For such covariates even the baseline obtained before the start of each treatment period changes with time. Patel's time moving covariate model is

$$E(Y) = \xi A + \Gamma_1 X_1 + \ldots + \Gamma_r X_r$$

Here A is a design matrix including between covariates, Γ_i is a p x p diagonal matrix with diagonal elements as V_{i1}, V_{i2} ,… , V_{ip} (I = 1, 2, …, r); X_r is a p x N matrix of the values of i^{th} within patient covariates.

Here as in analysis of variance, we have the hypothesis,

H_0: L ξ M = 0 against H_1: L ξ M ≠ 0

Structure covariance matrices: It is already viewed with some simple structures of the variance co-variance matrix Σ, as intra-class correlation structure. Some structures are: Σ σ^2 G, where G is a given matrix.

Intra class correlation structure can be expressed as

$\Sigma = \sigma^2 (1-p)I + \rho II'$

Serial correlation structure is given by

$\Sigma = \sigma^2$ C, where $C = \left[\rho^{|i-j|}\right],\ (i, j = 1, 2, \ldots, p)$

Khatri[12] has derived tests for testing such structures of Σ.

Random Coefficient Regression (Two stage model): Here, the model can be considered in two stages which will imply a particular type of structure of Σ.

In the first stage for the i^{th} individual,

$Y_i = X \beta_i + e_i$

where, β_i are themselves random and over the population in the second stage.

$E(\beta_i) = \beta$, $V(\beta_i) = F$; and $Cov(\beta_i, \beta_j) = 0$ for $i \neq j$

And thus we have a 'random coefficient regression' (RCR) model. This model implies a structure of Σ as $\Sigma = XFX' + \sigma^2 I$.

Rao[4] obtained estimates of β_i, β and also test of hypothesis regarding the above structure of Σ. Rao[13] has considered empirical Bayes (EB) estimator for RCR model for estimation of β_i's where F, β and σ^2 are to be estimated from the estimates available.

Rao[14] also suggested a effective general variance co-variance structure.

$\Sigma = XFX' + Z\Delta Z' + \sigma^2 I$

Where Z is any p x (p-q) matrix of rank (p-q) such that X'Z = 0 and F and Δ are any arbitrary positive definite matrices. Under this structure the unweighted least square estimator is the best linear unbiased estimator. Swamy[15] considered a more general version of RCR model by allowing 'X' to be different for each unit 'i' but the dimension of 'X_i' are same for each 'i'.

Choice of model: All the above specified models defined a structure among means and dispersion matrix. The structure among means can be taken to be same, but the models differ considerably in the structure of dispersion matrix Σ. Pottoff and Roy[7] do not put any structure on Σ, but estimates can be very inefficient when number of time points is large as the number of parameters increase considerably. RCR models impose a particular structure on the error component. Rao[4] has given a method for testing this structure against unstructured dispersion matrix Σ.

Once a family of models is chosen, the maximum likelihood method can be used to estimate the parameters. In some case interactive techniques have to be used for this purpose. It has been shown that the estimates of mean parameters are not very sensitive to the refinement of estimating techniques. Therefore, it may not be very worthwhile to use very efficient and time consuming techniques for the estimation of mean parameters.

Prediction: Estimation and testing are of course, important aspect of the analysis, but sometimes main interest of the investigation is in prediction. It would be proper to make the distinction between estimation and prediction. Estimation is a procedure of determining the value of fixed parameters whereas prediction is concerned with the value of a random event itself. Some work on prediction in linear models is also reported in the literature. Consider the model

$Y = X \beta + e$

Where 'Y' is a p component random vector with X β and variance $\sigma^2 V_{11}$. The purpose is to predict Y, as a linear function of 'Y', where Y_* is $Y_* = x_*' \beta + e_*$

$$V(Y_*) = \sigma^2 V_{22},$$
$$Cov(Y, Y_*') = \sigma^2 V_{12}$$

Let $\hat{y}_* = l'y$ be a linear predictor of y_*. It can be shown that the best linear unbiased predictor of y_* is

$$\hat{y}_* = x_*' \beta + V_{12}' V_{11}^{-1} (Y - X\beta)$$
$$= E(y_* / Y)$$

If β and V are not known then it is natural to put the estimated values of these parameters. Rao[4,14], Geisser[9], Copas[16], Reinsel[17-20], Rao and Bourdreau[21] have done work on predictions. Suppose that N units are observed at 'p' time points and observations are represented by 'p' component vectors Y_i (I = 1, 2, …, N). Two types of prediction problems has been considered in literature. For $(N+1)^{th}$ unit, vector Y_{N+1} is observed at p_1 (< p) time points and one wishes to predict the remaining p_2 (= p – p_1) values. This is called a case of

conditional prediction. Another type of problem is of predicting the values of measurement at $(p+h)^{th}$ time point for these units. This is called the problem of predicting the future observations. Geisser[9], Rao[13] and Fern[22] have established the conditional prediction.

Growth curves with incomplete data: Growth curve experiments, repeated measurement designs and longitudinal studies are constructed so that data are taken repeatedly on the same experimental units. This process often results in an incomplete or unbalanced data. Most statistical methods that are appropriate for the analysis of data from such experiments will have the existence of full records (usually multivariate normal).

Conclusion

The present paper gives an overview of statistical tools necessary for longitudinal data analysis. The models based on repeated measures, split plot analysis, multivariate models without special co-variance structure, multidimensional growth curve models are reviewed. The choice of the model and the prediction under such models are suggested.

References

1. Mauchly J.W. (1940). Significance Test for Sphericity of a Normal n-Variate Distribution. *The Ann. Math. Stat*., 11(2), 204-209.

2. Scheffe Henry (1959). The analysis of variance. New York, Wiley.

3. Grieve A.P. (1984). Tests of Sphericity of normal distributions and the analysis of repeated measures designs. *Psychometrika,* 49(2), 257-267.

4. Rao C.R. (1965). The Theory of Least-Squares When the Parameters are Stochastic and Its Applications to the Analysis of Growth Curves. *Biometrika,* 52(3-4), 447-458.

5. Wishart John (1938). Growth rate determinations in nutrition studies with the bacon pig, and their analysis. *Biometrika*, 30, 16-28.

6. Box G.E.P. (1950). Problems in the analysis of growth and wear curves. *Biometrics,* 6, 362-389.

7. Pottoff R.F. and Roy S.N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51, 313-326.

8. Khatri C.G. (1966). A note on a MANOVA model applied to problems in Growth Curve. *Ann. Inst. Statist. Maths.*, 18, 75-86.

9. Geisser S. (1981). Sample reuse procedures for prediction of the unobserved portion of a partially observed vector. *Biometrika*, 68, 243-250.

10. Reinsel G. (1982). Multivariate repeated measurements for growth curve models with multivariate random effects covariance structure. *J. Amer. Statist. Assoc*., 77, 190-210.

11. Patel H.I. (1986). Analysis of repeated measures designs with changing covariates in clinical trials. *Biometrika*, 73(3), 707-715.

12. Khatri C.G. (1973). Testing some covariance structures under a growth curve model. *J. Multi. Anal.*, 3(1), 102-116.

13. Rao C.R. (1975). Simultaneous estimation of parameters in different linear models and applications to biometric problems. *Biometrics*, 31, 545-554.

14. Rao C.R. (1967). Least squares theory using estimated dispersion matrix and its application to measurement of signals. *In Proceedings of the 5th Berkeley Symposium on Math. Statist. and Prob., L. Le Cam and J. Neyman. eds*., 1, 355-372.

15. Swamy P.A.V.B. (1971). Statistical inference in random coefficient regression models. Springer-Verlag, Berlin.

16. Copas J.B. (1983). Regression, prediction and shrinkage. *J. Roy. Statist. Soc., Series* B, *(Methodological),* 45(3), 311-354.

17. Reinsel G. (1984). Estimation and prediction in a multivariate random effects generalized linear model. *J. Amer. Statist. Assoc*., 79, 406-414.

18. Reinsel G. (1984). A note on conditional prediction in the multivariate linear model. *J. Roy. Statist. Soc*., *Series B*, 46, 107-117.

19. Reinsel G. (1984). Effects of the estimation of covariance matrix parameters in the generalized multivariate linear model. *Comm, Stat., (Th. and Meth.),* 13(5), 639-650.

20. Reinsel G. (1985). Mean squared error properties of empirical Bayes' estimators in a multivariate random effects general linear model. *J. Amer. Statist. Assoc*., 80, 642-650.

21. Rao C.R. and Boudreau R. (1985). Prediction of future observations in factor analytic type growth model. *Multi. Anal.*, 4, 449-466.

22. Fearn T. (1975). A Bayesian approach to growth curves. Biometrika, 62(1), 89-100.