

APPROACH OF LINEAR MIXED MODEL IN LONGITUDINAL DATA ANALYSIS USING SAS

Pankaj Tiwari* and Gaurav Shukla

*LES & IT Divison, Ivri, Izzatnagar, Bareilly

E-Mail:panks.stat@gmail.com, saigaurav83@gmail.com

Abstract

Linear mixed model is one of the best methodologies for analysis of the longitudinal (repeated measures) data. One major advantage of this methodology is that it accommodates the complexities of typical longitudinal data sets. The analysis of Linear mixed model methodology for the analysis of repeated measurements is becoming increasingly common due to development of widely available software. This paper reviews and summarizes the methodology of Linear Mixed Model approach for the analysis of repeated measurements data using SAS Software. PROC MIXED in SAS provides a very flexible environment in which model can be many type of repeated measures data. It can be repeated in time, space or both. Correlation among measurements made on same subject or experiment unit can be modeled using random effect and through the specification of a covariance structure. PROC MIXED provides a useful covariance structures or modeling both time and space, including discrete & continuous increments of time and space.

Key Words: ML, REML, Random Effect, Variance Components, Covariance Structure, AIC, BIC, Likelihood Function.

1. Introduction

Experiment units are often measured more than once if precision of single measurements is not adequate or if changes expected over time. Variability among measurements on the same experimental unit can be homogeneous, but may alternatively be expected to change through time. Typical examples are milk yield during lactation, hormone concentration in blood, or growth measurements over same period. In repeated measurement design the effect of a treatment is tested on experimental unit that have been measured repeatedly over time. The term “repeated measurements” refers broadly to data in which the response of each experimental unit or subject is observed on multiple occasions or under multiple conditions. The “Longitudinal Data” is also often used to describe repeated measurements data. In longitudinal data, the response for each experimental unit in the study is observed on two or more occasions. The defining feature of a longitudinal data set is repeated observations on experimental units.

The goal of longitudinal data research is

1. To characterize patterns subject (e.g. growth, decline in lung cancer, increase in blood pressure) over “time”
2. To investigate the effects of important covariates on these patterns. There are two types of covariates in the longitudinal studies.
 - i. Non-time varying co-variates (e.g. gender, race) - between subject.
 - ii. Time-varying covariates (e.g. week, age weight, income, smoking status, exposure) within subject.

The Mixed model allows specification of matters determined by subject matters consideration methodology. It also explicits modeling and analysis of variation between subjects and within subjects. Henderson (1953) developed widely used analogous techniques for unbalanced data as maximum likelihood estimating of the fixed effects. Paterson and Thompson (1971) proposed the alternative Restricted Maximum Likelihood (REML) approach. The REML estimation approach applies ML estimation technique to the likelihood function associated with a set of “error constants” rather than that of associated with observations. This accounts for the loss of degree of freedom resulting for the loss of fixed effects and gives less biased estimates of variance components.

PROC MIXED is the procedure to fit the Linear MIXED model in the SAS System. It estimates parameters by likelihood or moment –based techniques. Diagnostic and Influence analysis for observations or group of observations can be carried out through mixed model. It gives solutions for different kinds of repeated measure problems. Random effects are used to make hierarchical models correlating measurements made on the same level of random factor, including subject specification regression models. Different kinds of covariance and correlation structures can be specified for the residuals in the mixed model. Random and Repeated Statements are used in PROC MIXED procedure for random effects and covariance structures respectively.

2. Methodology

Let $y_i = (y_{i1}, y_{i2}, y_{i3}, \dots, y_{it_i})'$ be the $t_i \times 1$ vector of responses from subject i for $i = 1, 2, 3, \dots, n$. The General Linear mixed model for longitudinal data is

$$y_i = X_i \beta + Z_i \gamma_i + \varepsilon_i, \quad i = 1, 2, \dots, n. \quad (1)$$

where X_i is a $t_i \times b$ model (design) matrix for the subject i , β is a $b \times 1$ vector of regression coefficients, γ_i is a $g \times 1$ vector of random effects for subject i , Z_i is a $t_i \times g$ design matrix for the random effects and ε_i is a $t_i \times 1$ vector of within- subject errors.

The γ_i and ε_i vectors are assumed to be independent $N_g(0_g, B)$ and $N_{t_i}(0_{t_i}, W_i)$ variates. In addition, y_i are independent $\sim N_{t_i}(X_i \beta, V_i)$

$$\text{When } V_i = Z_i B Z_i' + W_i \quad (2)$$

These matrices X_i, Z_i and W_i are subjected specific. This model is very general because subjects can have varying numbers of observation times and can differ among subjects. The within subject covariance matrix W_i is assumed to depend on i only through its dimension t_i ; unknown parameters in W_i do not depend on i . A wide variety of covariance structures for γ_i and ε_i can be considered.

Laired and Ware (1982) considered the linear mixed model as two random-effects model. In the first stage, they assumed that the model for the i^{th} subject is

$$y_i = X_i \beta + Z_i \gamma_i + \varepsilon_i \quad (3)$$

The vectors $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_i, \dots, \varepsilon_n$ are assumed to be independently distributed as $N\left(\begin{matrix} 0 \\ t_i \\ W_i \end{matrix}\right)$. The vector regression coefficient β and the subject specific vectors γ_i are considered to be fixed. In the second stage $\gamma_1, \gamma_2, \dots, \gamma_i, \dots, \gamma_n$ are assumed to be independent $N\left(\begin{matrix} 0 \\ g \\ B \end{matrix}\right)$ variates and γ_i and ε_i are assumed independent. Thus $y_1, y_2, \dots, y_i, \dots, y_n \sim N(X_i\beta, Z_iBZ_i + W_i)$.

2.1 Covariance Structures

(i) **Unstructured:** This is a completely general covariance matrix.

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{pmatrix}$$

(ii) **Compound Symmetry:** This structure has constant variance and constant covariance.

$$\begin{pmatrix} \sigma^2 + \sigma_1^2 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma^2 + \sigma_1^2 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma^2 + \sigma_1^2 \end{pmatrix}$$

(iii) **Compound Symmetry (Heterogeneous):** This covariance structure has heterogeneous variances and heterogeneous correlations.

$$\begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{21}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{31}\sigma_1\sigma_3 & \rho_{32}\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix}$$

(iv) **Autoregressive (AR(1)):** This is a first-order autoregressive structure with homogenous variances. The correlation between any two elements is equal to rho (ρ) for adjacent elements, ρ^2 for elements that are separated by a third, and so on. ρ is constrained so that $-1 < \rho < 1$.

$$\tilde{C} \begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix}$$

(v) **Autoregressive (Heterogeneous):** This is a first-order autoregressive structure with heterogeneous variances. The correlation between any two elements is equal to rho

(ρ) for adjacent elements, ρ^2 for two elements separated by a third, and so on. ρ is constrained to lie between -1 and 1 .

$$\begin{pmatrix} \sigma_1^2 & \rho\sigma_2\sigma_1 & \rho^2\sigma_3\sigma_1 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 & \rho\sigma_2\sigma_3 \\ \rho^2\sigma_3\sigma_1 & \rho\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix}$$

2.2 Likelihood Functions

As we know that -2 times log likelihood of the MLE is

$$-2\text{lMLE}(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \log |\mathbf{V}| + \mathbf{r}^T \mathbf{V}^{-1} \mathbf{r} + n \log 2$$

and -2 times log likelihood of the REML is

$$-2 \text{lREML}(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \log |\mathbf{V}| + \mathbf{r}^T \mathbf{V}^{-1} \mathbf{r} + \log |\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}| + (n-p) \log 2$$

where 'n' is the number of observations and 'p' is the rank of fixed effects design matrix. The key components of the likelihood functions are

$$l_1(\boldsymbol{\Sigma}) = \log |\mathbf{V}|$$

$$l_2(\boldsymbol{\Sigma}) = \mathbf{r}^T \mathbf{V}^{-1} \mathbf{r}$$

$$l_3(\boldsymbol{\Sigma}) = \log |\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}|$$

Therefore, in each estimation iteration, we need to compute $l_1(\boldsymbol{\Sigma})$, $l_2(\boldsymbol{\Sigma})$ and $l_3(\boldsymbol{\Sigma})$ as well as their 1st and 2nd derivatives with respect to $\boldsymbol{\Sigma}$. Where \mathbf{V} is $n \times n$ Covariance matrix of \mathbf{y} , \mathbf{y} is $n \times 1$ vector of dependent variable, \mathbf{r} is $n \times 1$ vector of residual and \mathbf{X} is $n \times p$ design matrix of fixed effect.

3. Information Criteria

Information criteria are used for model comparison. The following criteria are given in shorter and better form. Let 'l' be the log-likelihood of (REML or ML), n the total number of cases (or total of case weights if used) and 'd' the number of model parameters, then the formulae for various criteria are:

$$\text{Akaike information criteria (AIC)} = -2l + 2d$$

$$\text{Finite sample corrected (AICC)} = -2l + 2d \times n / (n - d - 1)$$

$$\text{Bayesian information criteria (BIC)} = -2l + d \times \log(n)$$

For REML, the value of 'n' is chosen to be total number of cases minus number fixed effect parameters and 'd' is number of covariance parameters. For ML,

the value of 'n' is total number of cases and 'd' is number of fixed effect parameters plus number of covariance parameters.

Some other main references who used this techniques are Jennrich and Schluchter (1986), Laird et.al (1987), Diggle (1988), Lindstrom and Bates (1988), Jones and Boadi – Boateng (1991), Jones (1993), Zimmerman and Nunez Anton (1997), Pourahmadi (1999) and Nuzen Anton and Zimmerman (2000) etc

4. Illustrations

Illustrations for proper understanding of repeated measures analysis through linear mixed model are taken from Davis (2004). In a randomized, multicentre, double –blind, placebo- controlled trial of botulinum toxin type B (Bot B) in the female patients with cervical dystonia, eligible subjects were randomized to one of three groups: placebo, 5000 units of Bot B, or 10,000 units of Bot B. The primary outcome variable was the Total Score on the Toronto Western Spasmodic Torticollis Rating Scale (TWSTRS-Total). The TWSTRS-Total, which measure severity, pain, and disability of cervical dystonia, is numerical score ranging from 0 to 87; high scores indicate impairment. The TWSTRS-Total was administered at baseline (week 0) and week 2, 4, 8, 12, and 16 for the above treatment. For Analyzing the data, we have coded placebo =1, 5000 units =2 and 10,000 units =3

Step 1:- First, we are considering the **unstructured** structure of the variance and covariance matrix for the selection of appropriate model.

```
proc mixed data = female_data covtest;
class treat week;
model score = treat|week;
repeated week /type=un subject=id r; run;
```

Section 4.1.1

The SAS System

The Mixed Procedure

Covariance Parameter Estimates

Cov Parm	Subject	Standard Estimate	Z Error	Value	Pr Z
UN(1,1)	ID	120.19	38.0079	3.16	0.0008
UN(2,1)	ID	118.32	45.5834	2.60	0.0094
UN(2,2)	ID	229.27	72.5023	3.16	0.0008
UN(3,1)	ID	123.47	47.4714	2.60	0.0093
UN(3,2)	ID	221.04	72.7157	3.04	0.0024
UN(3,3)	ID	248.15	78.4711	3.16	0.0008
UN(4,1)	ID	126.89	46.5952	2.72	0.0065
UN(4,2)	ID	207.63	69.0020	3.01	0.0026
UN(4,3)	ID	223.64	72.9452	3.07	0.0022

UN(4,4)	ID	227.30	71.8800	3.16	0.0008
UN(5,1)	ID	121.42	43.0617	2.82	0.0048
UN(5,2)	ID	167.85	59.4952	2.82	0.0048
UN(5,3)	ID	169.14	61.1289	2.77	0.0057
UN(5,4)	ID	186.85	62.1153	3.01	0.0026
UN(5,5)	ID	185.89	58.7834	3.16	0.0008
UN(6,1)	ID	122.12	42.3167	2.89	0.0039
UN(6,2)	ID	158.07	56.9470	2.78	0.0055
UN(6,3)	ID	166.08	59.4717	2.79	0.0052
UN(6,4)	ID	177.87	59.6517	2.98	0.0029
UN(6,5)	ID	173.81	55.9188	3.11	0.0019
UN(6,6)	ID	173.90	54.9934	3.16	0.0008

Section 4.1.2

Fit Statistics

-2 Res Log Likelihood	798.6
AIC (smaller is better)	840.6
AICC (smaller is better)	850.0
BIC (smaller is better)	864.4

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
20	212.48	<.0001

Section 4. 1.3

Type 3 Tests of Fixed Effects

Effect	Num Den		F Value	Pr > F
	DF	DF		
Treat	2	20	0.91	0.4200
Week	5	20	4.37	0.0075
Treat*Week	10	20	1.50	0.2113

Results and Discussion

Section 4.1.1 explains the results of estimates of covariance parameters, standard error of the each estimate, z values for the estimates and eventually provides P-values for significant testing of estimates of covariance parameters. Section 4.1.2 provides us Fit Statistics which will be discussed later. Section 4.1.2 indicates that unstructured covariance matrix is preferred to the diagonal matrix of the ordinary least squares null model. Section 4.1.3 provides the Analysis of Variance table for the fixed effects in the model which informs us that week variable has significant effect on subjects at 0.05 level of significance.

Step 2:- Second, we are going to take **Compound Symmetry** variance covariance matrix for the selection of appropriate model.

```
proc mixed data = pankaj.female_data covtest;
class treat week;
model score = treat| week;
repeated week /type=cs subject=id r;run;
```

Section 4.2.1

The SAS System
The Mixed Procedure

Covariance Parameter Estimates

Cov Parm	Subject	Standard Estimate	Z Error	Value	Pr Z
CS	ID	164.28	53.7042	3.06	0.0022
Residual		33.1704	4.6910	7.07	<.0001

Section 4.2.2

Fit Statistics

-2 Res Log Likelihood	865.5
AIC (smaller is better)	869.5
AICC (smaller is better)	869.6
BIC (smaller is better)	871.8

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
1	145.56	<.0001

Section 4.2.3

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Treat	2	20	0.91	0.4200
Week	5	100	7.06	<.0001
Treat*Week	10	100	1.96	0.0460

Results and Discussion

Section 4.2.1 explains the results of estimates of covariance parameters, standard error of the each estimate, z values for the estimates and eventually provides P-values for significant testing of estimates of covariance parameters and also shows the significant result of Compound Symmetry Structure for the variance covariance matrix. Section 4.2.2 provides us Fit Statistics to be discussed later. Section 2.2

indicates that Compound Symmetry covariance matrix is preferred to the diagonal matrix of the ordinary least squares null model. Section 4.2.3 provides the Analysis of Variance table for the fixed effects in the model which informs us that week variable and interaction effect in between treatment and week have significant effect on subjects at 0.05 level of significance.

Step 3:- Third, we are considering the **Compound Symmetry Heterogeneous** Structure of variance covariance matrix for the selection of appropriate model.

```
Proc mixed data = pankaj.female_data covtest;
class treat week;
model score = treat| week;
repeated week /type=csh subject=id ;run;
```

Section 4.3.1

The SAS System
The Mixed Procedure
Covariance Parameter Estimates

Cov	Parm	Standard Subject	Z Estimate	Error	Value	Pr Z
	Var(1)	ID	133.19	42.5845	3.13	0.0009
	Var(2)	ID	232.17	73.1892	3.17	0.0008
	Var(3)	ID	250.06	78.7707	3.17	0.0008
	Var(4)	ID	213.52	66.5508	3.21	0.0007
	Var(5)	ID	181.23	56.8031	3.19	0.0007
	Var(6)	ID	169.24	53.0308	3.19	0.0007
	CSH	ID	0.8407	0.04787	17.56	<.0001

Section 4.3.2

Fit Statistics

-2 Res Log Likelihood	857.2
AIC (smaller is better)	871.2
AICC (smaller is better)	872.2
BIC (smaller is better)	879.2

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
6	153.83	<.0001

Section 4.3.3

Type 3 Tests of Fixed Effects

Num Effect	Den DF	DF	F Value	Pr > F
Treat	2	20	0.91	0.4178
Week	5	100	5.91	<.0001
Treat*Week	10	100	1.73	0.0849

Results and Discussion

Section 4.3.1 explains the results of estimates of six covariance parameters, standard error of the each estimate, z values for the estimates and eventually provides P-values for significant testing of estimates of covariance parameters Section 4.3.2 provides us Fit Statistics which will discuss later and explains that Compound Symmetry Heterogeneous covariance matrix is preferred to the diagonal matrix of the ordinary least squares null model. Section 3.3 provides the Analysis of Variance table for the fixed in the model which informs us that effect of week variable has significant effect on subjects at 0.05 level of significance.

Step 4:- Fourth, we are considering the **Autoregressive** structure of variance covariance matrix for the selection of appropriate model.

```
proc mixed data = pankaj.female_data covtest;
class treat week;
model score = treat| week;
repeated week /type=ar(1) subject=id r ;run;
```

Section 4.4.1

Covariance Parameter Estimates

Cov Parm	Subject	Standard Estimate	Z Error	Value	Pr Z
AR (1)	ID	0.8694	0.03281	26.49	<.0001
Residual		171.86	40.1462	4.28	<.0001

Section 4.4.2

Fit Statistics

-2 Res Log Likelihood	853.4
AIC (smaller is better)	857.4
AICC (smaller is better)	857.5
BIC (smaller is better)	859.7

Null Model Likelihood Ratio Test

	DF	Chi-Square	Pr > ChiSq
1	157.63	<.0001	

Section 4.4.3

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Treat	2	20	1.15	0.3359
Week	5	100	5.76	0.0001
Treat*Week	10	100	1.93	0.0490

Results and Discussion

. Section 4.4.1 explains the results of estimates of covariance parameters, standard error of the each estimate, z values for the estimates and eventually provides P-values for significant testing of estimates of covariance parameter and also shows the significant result of Auto regressive Structure for the variance covariance matrix. Section 4.4.2 provides us Fit Statistics which will discuss later. Section 4.4.2 indicates that Autoregressive covariance matrix is preferred to the diagonal matrix of the ordinary least squares null model. Section 4.4.3 provides the Analysis of Variance table for the fixed in the model which informs us that week variable and inteaction effect in between treatment &week (very near to 0.05) have significant effect on subjects at 0.05 level of significance.

Step 5:- Fifth, we are considering **Autoregressive Heterogeneous** variance covariance matrix for the selection of appropriate model.

```
proc mixed data = pankaj.female_data covtest;
class treat week;
model score = treat| week;
repeated week /type=arh(1) subject=id r ;run;
```

Section 4.5.1

Covariance Parameter Estimates

Cov	Parm	Standard Subject	Z Estimate	Error	Value	Pr Z
	Var(1)	ID	190.57	63.0430	3.02	0.0013
	Var(2)	ID	305.15	97.7766	3.12	0.0009
	Var(3)	ID	243.63	75.6260	3.22	0.0006
	Var(4)	ID	194.25	58.1497	3.34	0.0004
	Var(5)	ID	144.43	41.4129	3.49	0.0002
	Var(6)	ID	131.11	36.7111	3.57	0.0002
	ARH(1)	ID	0.9007	0.02879	31.29	<.0001

Section 4.5.2

Fit Statistics

-2 Res Log Likelihood	841.7
AIC (smaller is better)	855.7
AICC (smaller is better)	856.7
BIC (smaller is better)	863.7

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
6	169.30	<.0001

Section 4.5.3

Type 3 Tests of Fixed Effects

Num Den Effect	DF	DF	F Value	Pr > F
Treat	2	20	0.94	0.4078
Week	5	100	4.52	0.0009
Treat*Week	10	100	1.22	0.2846

Results and Discussion

We can easily see that diagonal elements are variances and non diagonal elements are covariances of the matrix. Section 4.5.1 explains the results of estimates of covariance parameters, standard error of the each estimate, z values for the estimates and eventually provides P-values for significant testing of estimates of covariance parameters. Section 4.5.2 provides us Fit Statistics which will discuss later. Section 4.5.3 indicates that Autoregressive Heterogeneous covariance matrix is preferred to the diagonal matrix of the ordinary least squares null model. Section 4.5.4 provides the Analysis of Variance table for the fixed in the model which informs us that week variable and interaction effect in between treatment and week (very near to 0.05) have significant effect on subjects at 0.05 level of significance.

Model Comparison (Comparison with Compound Symmetry):

Model	AIC	-2Residual Log Likelihood	Parameter (df+1)	Difference -2Residual Log Likelihood (vs CS)	Difference in df (vs CS)
UN	840.6	798.6	21	66.9	19
CS	869.5	865.5	2	-	-

CSH	871.2	857.2	7	8.3	5
AR(1)	857.4	853.4	2	12.1	0
ARH(1)	855.7	841.7	7	23.8	5

The two most useful structures are **Autoregressive Heterogeneous Variances** and **Unstructured** since these two models have the smallest AIC values and the -2 Log Likelihood scores are significantly smaller than the -2 Log Likelihood scores of other models.

5. Conclusion

Linear mixed approach for the analysis of repeated measurement data with continuous response variable is more appropriate as compared to traditional methods like separate one-way ANOVA at each time point, two-way ANOVA (Variance Components-Covariance Structure), Repeated measures ANOVA and Multivariate approach (Unstructured-Covariance Structured). It gives analysis for Random and Fixed effects separately. The drawback of this approach is that it is used when response variable is in continuous.

References

1. Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19, p. 716-723.
2. Albert, P.S. (1999). Longitudinal data analysis (repeated measures) in clinical trials. *Statistics in Medicine* 18, p. 707-732
3. Diggle, P.J (1988). An approach to the analysis repeated measurements. *Biometric*, 44, p. 951-971
4. Henderson, C.R (1953). Estimation of variances and covariances components, *Biometrics*, 9, p.226-252.
5. Laird, N.M. and Ware, J.H. (1982). Random-effects models for longitudinal data, *Biometrics*, 38, p. 963-974
6. Littell, R. C., Milliken, George A., Stroup, Walter W., and Wolfinger, R. D. (1996). *SAS System for Mixed Models*, Cary, NCSAS Institute, Inc.
7. Schwarz, G. (1978). Estimating the Dimension of a Model, *Annals of Statistics*, 6, p. 461-464.
8. Wolfinger, R., Tobias, R. and Sall, J. (1994). Computing Gaussian likelihoods and their derivatives for general linear mixed models, *SIAM Journal on Scientific Computing*, 15:6, p. 1294-1310
9. Zimmerman, D.L. (2000). Viewing the Correlation Structure of Longitudinal data through a PRISM, *The American Statistician*, 54, p. 310 - 318