STRATHMORE UNIVERSITY

124384 - Luycer Bosire & 135214 - Ruth Katam

10/08/2021

# Question 1 (20 Marks)

1. Repeated measures data and longitudinal data are used extensively in various fields of science. Describe each design, providing any features that distinguish the two designs.

-Repeated Measures: It is an experimental design where each person in the experiment is tested for multiple conditions overtime or under different conditions overtime. -Longitudinal design: A group of individuals is taken and a study on something specific is done over a period of time. The feature that distinguishes the two designs is that longitudinal measures a specific item over the time period while Repeated Measure take into account multiple conditions on the same observation over the period.

1. Explain when and why researchers would prefer longitudinal studies over cross-sectional studies.

Both longitudinal and cross-sectional studies measure the changes over time in a particular outcome of interest. A researcher would prefer longitudinal study over cross-sectional when the researcher wishes to use the same sample over the years to conduct the study. Longitudinal study is good because the researcher will be able to measure the outcome taken at each point of interest on a single sample.

Other advantages of longitudinal studies are: >+ *Economizes on subjects.* >+ *Subjects serve as own control.* >+ *Between-subject variation excluded from error.* >+ *Can provide more efficient estimators than cross-sectional designs with same number and pattern of observations.* >+ *Can separate aging effects (changes over time within individuals) from cohort effects (differences between subjects at baseline)) cross-sectional design can’t do this.* >+ *Can provide information about individual change.*

1. Consider the linear mixed effects model (in matrix notation),
2. , expressed two level hierarchical model:

$$y|u{\sim}N(X\beta+Zu,R) \\ u \sim N(0,R)$$

## Theory of linear mixed models.

Where y is a

column vector, the outcome variable; X is a

matrix of the p predictor variables;

is a

column vector of the fixed-effects regression coefficients (the

); Z is the

design matrix for the q random effects and J groups; u is a

vector of q random effects (the random complement to the fixed

for J groups; and

is a

column vector of the residuals, that part of y that is not explained by the model,

.

1. Re-write the marginal model version of this model.

The general linear mixed model implies the marginal model:

$$Y\_i \sim N(X\_i\beta, Z\_iDZ\_i^` + \Sigma\_i)$$

The marginal model for the linear mixed effects model in matrix notation therefore is:

ii.Derive an expression for the Maximum Likelihood Estimator or weighted Least Squares Estimator of

.

## Solution

Let

$${E[\epsilon] = E[Zu +e] = ZE[u]+E[e] = X\beta \\ Var[\epsilon] = Var [Zu+e]=ZVar[u]Z^{T} + Var[e] = ZGZ^T + \Sigma}$$

such that

.

Under these circumstances, the MLE for

is:

1. Derive an expression for the best linear unbiased estimator of u.

Given the model specifications, the joint distribution of y and u is:

$$\mathbf \displaystyle \begin{bmatrix} y \\ u \end{bmatrix} \sim MVN \displaystyle \begin{bmatrix}X\beta \\ 0 \end{bmatrix}, \begin{bmatrix} V & ZG \\ GZ^T & G \end{bmatrix}$$

From the properties of multivariate normal distribution, we have that:

$$E[u |y ]= E[u ] +Cov[u, y^T ]Var ^{-1} [ y]( y-E[ y]) \\ =GZ^T V^{-1}(y-X\beta) \\ = GZ^T (ZGZ^T + \Sigma)^{-1}(y-X\beta)$$

The fixed effects

are replaced by their estimates,so that predictions are made based on the expression:

The solution

and

discussed before require

. Henderson (1959) presents the mixed model equations (MME) to estimate our parameters without the need for computing

.

The MME are derived by maximizing the parameters the joint density of y and u expressed as:

The logarithm of this function is:

$$\displaystyle ℓ =log[p(y,u| \beta,G,\Sigma)] ∝ |\Sigma| + |G| +(y-x\beta-Zu) + u^TG^{-1}u \\ = |\Sigma| + |G| + y^T\Sigma^{-1}y - 2y^T\Sigma^{-1}Zu + \beta X^T \Sigma^{-1}Zu+ u^T Z^T \Sigma^{-1} + u^TG^{-1}u$$

The derivatives of ℓ regarding

and u are:

$$\mathbf \displaystyle \begin{bmatrix} dℓ/d\beta \\ dℓ/du \end{bmatrix} = \displaystyle \begin{bmatrix}X^T\Sigma^{-1}y-X^T\Sigma^{-1}X\beta - X^T\Sigma^{-1}Zu \\Z^T\Sigma^{-1}y-Z^T\Sigma^{-1}X\beta - Z^T\Sigma^{-1}Zu -G^{-1}u\end{bmatrix}$$

Equating them to zero and solving for them can be expressed as:

$$\mathbf \displaystyle \begin{bmatrix} X^T\Sigma^{-1}X & X^T\Sigma^{-1}Z \\ Z^T\Sigma^{-1}X & Z^T\Sigma^{-1}Z+G^{-1}\end{bmatrix} \begin{bmatrix} \hat\beta \\ \hat u\end{bmatrix} = \begin{bmatrix} X^T\Sigma^{-1}y \\ X^T\Sigma^{-1}y\end{bmatrix}$$

known as the mixed model equations.

We solve the equation to obtain:

which is the best linear unbiased predictor (BLUP) of u.

# Question 2: Consider the sleep deprivation data, *sleepstudy*, in the R library *lme4*. In a sleep deprivation study, the average reaction time per day for subjects was measured. On day 0 the subjects had their normal amount of sleep. Starting that night they were restricted to 3 hours of sleep per night. The observations represent the average reaction time on a series of tests given each day (for 10 days) to each of the 18 subjects.

1. Give the two-stage representations of the random-intercept and the random coefficients models that can be used to model this data.
2. For each model in part (a), provide a single expression of the linear mixed effects model, clearly showing the fixed and random effects, with all relevant distributional assumptions provided.
3. Fit random-intercept and random-coefficients models to this data then:
4. Compute the intraclass correlation coefficient for the random intercept model and comment on it.
5. For the random slope model, determine the proportion of variance attributable to each variance component and comment on your results.
6. Carryout a Likelihood ratio test and determine the best fitting model. Comment on the relationship between the outcome and predictor in the model.
7. What is the significance of the correlation term given in the random effects part of results for the random-coefficients model?

Outcome: Reaction; The Average reaction time (ms) Predictors: Days; Number of days of sleep deprivation and Subject; Subject number on which the observation was made.

library(lme4)

## Loading required package: Matrix

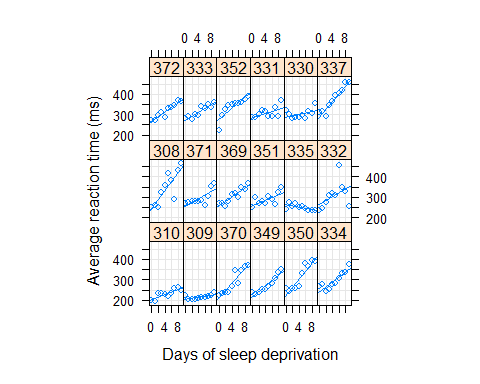
data("sleepstudy")  
str(sleepstudy)

## 'data.frame': 180 obs. of 3 variables:  
## $ Reaction: num 250 259 251 321 357 ...  
## $ Days : num 0 1 2 3 4 5 6 7 8 9 ...  
## $ Subject : Factor w/ 18 levels "308","309","310",..: 1 1 1 1 1 1 1 1 1 1 ...

require(lattice)

## Loading required package: lattice

xyplot(Reaction ~ Days | Subject, sleepstudy, type = c("g","p","r"),  
 index = function(x,y) coef(lm(y ~ x))[1],  
 xlab = "Days of sleep deprivation",  
 ylab = "Average reaction time (ms)", aspect = "xy")



(fm1 <- lmer(Reaction ~ Days + (Days|Subject), sleepstudy, subset=Days>=2))

## Linear mixed model fit by REML ['lmerMod']  
## Formula: Reaction ~ Days + (Days | Subject)  
## Data: sleepstudy  
## Subset: Days >= 2  
## REML criterion at convergence: 1404.094  
## Random effects:  
## Groups Name Std.Dev. Corr   
## Subject (Intercept) 31.507   
## Days 6.766 -0.25  
## Residual 25.526   
## Number of obs: 144, groups: Subject, 18  
## Fixed Effects:  
## (Intercept) Days   
## 245.10 11.44

## independent model  
(fm2 <- lmer(Reaction ~ Days + (1|Subject) + (0+Days|Subject), sleepstudy, subset=Days>=2))

## Linear mixed model fit by REML ['lmerMod']  
## Formula: Reaction ~ Days + (1 | Subject) + (0 + Days | Subject)  
## Data: sleepstudy  
## Subset: Days >= 2  
## REML criterion at convergence: 1404.626  
## Random effects:  
## Groups Name Std.Dev.  
## Subject (Intercept) 28.843   
## Subject.1 Days 6.285   
## Residual 25.747   
## Number of obs: 144, groups: Subject, 18  
## Fixed Effects:  
## (Intercept) Days   
## 245.10 11.44

Each subject’s reaction time increases approximately linearly with the number of sleepdeprived days. However, subjects also appear to vary in the slopes and intercepts of these relationships, which suggests a model with random slopes and intercepts.

fm1 <- lmer(Reaction ~ Days + (Days | Subject), sleepstudy)  
summary(fm1)

## Linear mixed model fit by REML ['lmerMod']  
## Formula: Reaction ~ Days + (Days | Subject)  
## Data: sleepstudy  
##   
## REML criterion at convergence: 1743.6  
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -3.9536 -0.4634 0.0231 0.4634 5.1793   
##   
## Random effects:  
## Groups Name Variance Std.Dev. Corr  
## Subject (Intercept) 612.10 24.741   
## Days 35.07 5.922 0.07  
## Residual 654.94 25.592   
## Number of obs: 180, groups: Subject, 18  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) 251.405 6.825 36.838  
## Days 10.467 1.546 6.771  
##   
## Correlation of Fixed Effects:  
## (Intr)  
## Days -0.138

We may use double-bar notation to fit a model that excludes a correlation parameter:

fm2 <- lmer(Reaction ~ Days + (Days || Subject), sleepstudy)  
summary(fm2)

## Linear mixed model fit by REML ['lmerMod']  
## Formula: Reaction ~ Days + ((1 | Subject) + (0 + Days | Subject))  
## Data: sleepstudy  
##   
## REML criterion at convergence: 1743.7  
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -3.9626 -0.4625 0.0204 0.4653 5.1860   
##   
## Random effects:  
## Groups Name Variance Std.Dev.  
## Subject (Intercept) 627.57 25.051   
## Subject.1 Days 35.86 5.988   
## Residual 653.58 25.565   
## Number of obs: 180, groups: Subject, 18  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) 251.405 6.885 36.513  
## Days 10.467 1.560 6.712  
##   
## Correlation of Fixed Effects:  
## (Intr)  
## Days -0.184

# Question 3 (20 Marks)

1. Describe each of the following covariance pattern models, explaining how they arise and how they can be employed in longitudinal data analysis:

Definition: Covariance pattern model  
The CPM for the

response vector y for individual i can be written as

.

is the

dependent variable vector for individual i,

is the with

individuals and

observations for individual i.

1. Compound Symmetry Structure  
   This is a simple form of the variance covariance matrix which specifies equal variances and covariances. In matrix form it is expressed as:

In longitudinal data analysis, the compound symmetry covariance structure can be specified for the R-sided matrix.  
The compound symmetry variance-covariance structure is the simplest and the most parsimonious pattern model designed for longitudinal analysis. A linear regression model including a residual variance-covariance matrix of CS can yield statistically efficient and consistent parameter estimates.

1. First-Order Auto-regressive Structure  
   Also known as a first order Markov process, this method is extensively used in time series analysis. It has homogenous variances, pairs of repeated measures separated by the one measurement occasion e.g laged by 1 are correlated equally. Repeated measures with larger lags are correlated to a lesser degree based on an exponential function of the one-lag correlation. Example repeated measures that are one lag apart (e.g., Time 1 and Time 2) are correlated at some estimated value , and repeated measures that are two lags apart (e.g., Time 1 and Time 3) would be correlated at . It is used for diminishing lags between timepoints. In LDA its used when there are many repeated measures and the spacing between measurement occasions is equal or nearly equal. The covariances for timepoints j and j’
2. where
3. is the AR(1) parameter and
4. is the error variance.

The structure in matrix formulation is given as:

1. Toeplitz or Banded Structure: It is used for diminishing lags between timeepoints, but not as rigid as AR(1). Each lag has its own correlation. Similar to the first-order autoregressive structure, all measures separated by one measurement occasion are equally correlated. However, for measures lagged by two, the Toeplitz structure estimates a separate correlation rather than simply squaring the lag 1 correlation. Toeplitz structure does not require that the change in correlation follow a specific function of the lag 1 correlation. In LDA it is used when their is a moderate number of repeated measures but correlations are not expected to decrease exponentially over time
2. Unstructured Form It assume a general unstructured form that allows all of the parameters of the variance-covariance matrix to be different every element is uniquely estimated. IN LDA it is typically reserved for very few measurement occasions or when the measurement occasions have an unorthodox structure.
3. In a clinical trial of patients suffering from epilepsy, patients were randomized to receive either a placebo or the drug progabide, in addition to standard therapy. A baseline count of the number of epilepsy seizures in an 8-week period prior to randomization was obtained. In addition, counts of the number of epilepsy seizures in each of four successive 2-week (post-baseline) treatment periods were obtained. The goal of the analysis was to make a comparison between the two treatment groups in terms of changes in the rates of epileptic seizures throughout the duration of the study. The raw data (epilepsy) for this question are found in the R library HSAUR3.
4. Fit a random intercept model within the generalized linear mixed effects paradigm to this data set. Present the results of the analysis using an appropriate summary table. [Hint: Use lme4 and specify the poisson family

We start by calculating the means and variances of the number of seizures for all treatment and the period interactions

## Loading required package: tools

## placebo.1 Progabide.1 placebo.2 Progabide.2 placebo.3 Progabide.3   
## 9.357143 8.580645 8.285714 8.419355 8.785714 8.129032   
## placebo.4 Progabide.4   
## 7.964286 6.709677

## placebo.1 Progabide.1 placebo.2 Progabide.2 placebo.3 Progabide.3   
## 102.75661 332.71828 66.65608 140.65161 215.28571 193.04946   
## placebo.4 Progabide.4   
## 58.18386 126.87957

Some of the variances are considerably larger than the corresponding means,which for a Poisson variable may suggest that over-dispersion may be a problem.

Next we fit a generalized linear mixed model:

1. Using a likelihood ratio test compare the random intercept model and the random coefficients model fitted within the generalized linear mixed effects paradigm.

Random Intercept model

Random coefficients model

We perform the likelihood ratio test:

rand(epy\_lmer)

## ANOVA-like table for random-effects: Single term deletions  
##   
## Model:  
## seizure.rate ~ base + treatment + age + period + (1 | subject)  
## npar logLik AIC LRT Df Pr(>Chisq)   
## <none> 7 -791.69 1597.4   
## (1 | subject) 6 -825.03 1662.0 66.667 1 3.214e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

\*\*\*Interpret the likelihood ratio obtained.

1. Fit generalized estimating equations to this data with unstructured, exchangeable, independent and AR correlation structures. For your best fitting model discuss the meaning of your findings. Comment on how you arrived at the best fitting model.

To compare the models we use QIC to select a good model in the generalized linear model. QIC is used to select a correlation structure. It compares models that have the same working correlation matrix and QIC form but varying mean specifications.Models with smaller values of QIC, CIC, QICu, or QICC are preferred.CIC has been suggested as a more robust alternative to QIC when the model for the mean may not fit the data very well and when models with different correlation structures are compared.

library(geepack)  
model\_unstructured <- QIC(gee\_unstructured)

model\_exchangeable<- QIC(gee\_exchangeable)

model\_independence <- QIC(gee\_independence)

model\_ar1 <- QIC(gee\_ar1)

model\_sum <- data.frame(model\_unstructured,model\_exchangeable,model\_independence,model\_ar1)

model\_sum%>% knitr::kable(digits = 2, "simple", caption = "GEE MODEL SUMMARY")

GEE MODEL SUMMARY

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | model\_unstructured | model\_exchangeable | model\_independence | model\_ar1 |
| QIC | -5893.1 | -5892.07 | -5883.67 | -5891.09 |
| QICu | -5890.7 | -5891.13 | -5891.13 | -5889.61 |
| Quasi Lik | 2949.4 | 2949.56 | 2949.56 | 2948.81 |
| CIC | 2.8 | 3.53 | 7.73 | 3.26 |
| params | 4.0 | 4.00 | 4.00 | 4.00 |
| QICC | -5892.1 | -5891.81 | -5883.49 | -5890.83 |
| Based on the | CIC values from the | result above, we pref | er the unstructured c | orrelation structure model as it has the smalled CIC value at 2.8. |

# References

<https://si.biostat.washington.edu/sites/default/files/modules/Seattle-SISG-18-MM-Lecture03.pdf>  
<https://www.mathworks.com/help/stats/estimating-parameters-in-linear-mixed-effects-models.html>  
Pan, W. (2001). Akaike’s information criterion in generalized estimating equations. Biometrics, 57, 120-125.  
Hardin, J.W. and Hilbe, J.M. (2012). Generalized Estimating Equations, 2nd Edition, Chapman and Hall/CRC: New York.  
Hin, L.-Y. and Wang, Y-G. (2009). Working-correlation-structure identification in generalized estimating equations, Statistics in Medicine 28: 642-658.  
Thall, P.F. and Vail, S.C. (1990). Some Covariance Models for Longitudinal Count Data with Overdispersion. Biometrics, 46, 657-671.