**Applied Mathematical Sciences, Vol. 3, 2009, no. 40, 1979 - 1989**

# Statistical Models for Longitudinal Data Analysis

**Michikazu Nakai and Weiming Ke**

Department of Mathematics and Statistics
South Dakota State University
Brookings, SD 57007, USA
Weiming.Ke@sdstate.edu

**Abstract**

Longitudinal data analysis has become popular as one of statistical methods. In this paper we introduce four common statistical models for handling longitudinal data. First, we introduce what longitudinal data are and the purpose of doing such an analysis. Then, using SAS examples, we focus on acquiring more applicable skills and ideas of applying these statistical models to longitudinal data analysis. In the end, we summarize and discuss characteristics of each model.

**Mathematics Subject Classification**: 62J10, 62J12

**Keywords**: SAS, longitudinal analysis, ANOVA, mixed-effect regression, generalized estimating equation.

## 1. Introduction:

Longitudinal studies are increasingly common in many scientific research areas. The longitudinal data are defined as the data resulting from the observations of subjects (human beings, animals, or laboratory samples, etc.) which are measured repeatedly over time [2]. The purpose of conducting longitudinal study

is to look at the change of treatments across time period. When change itself is the object of study, the only way to investigate the change is by collecting repeated measurement. For example, in medical area, patients may be assigned to take different treatments at the start of the study and to see what kind of effects they have with each treatment by week or by year. The advantage of doing such a longitudinal study is that it can provide information about individual change. That is, by collecting data over time, it can separate changes over time within individual sample from differences between subjects at baseline. Thus, these longitudinal studies give tremendous information on the subjects.

We have three interests of conducting the longitudinal study; (a) How treatment means differ (b) How treatment means change over time (c) How differences between means of treatments change over time [13].

Collecting longitudinal data can be challenging. First, we require measuring a sample of $N$ subjects at least twice or more, which costs more than cross-sectional data. Also, we have to ensure cooperation at each time of those who participated at baseline, as we cannot replace subjects who refuse (dropout) or are dead with others who did not participate at the previous measurement (attrition). Even in well-defined, controlled investigation, these situations invariably occur in longitudinal studies. These problems are referred as unbalanced design or missing value. Also, by collecting longitudinal data, since we measure same subject repeatedly, the observations are not independent. That is one of special remarks about longitudinal analysis. Some important references in the field of longitudinal data analysis can be found in [3, 5, 7, 8]. In this paper, we will introduce and discuss some statistical models for longitudinal data analysis.

## 2. Univariate and Multivariate Analysis of Variance

First, we use univariate and multivariate analysis of variance (ANOVA and MANOVA for short) for longitudinal studies. These methods are well-understood and most developed. Both models assume interval measurement and normally distributed errors that are homogeneous across groups. The weak aspect of these methods is that they only estimate and compare the group means and not informative about individual growth. Furthermore, as an assumption, these methods must have fixed time points. That is, each subject should have evenly or unevenly spaced time points. The ANOVA model is given by: $y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$ at $i = 1 \ldots N$, $j = 1 \ldots n$ . where $\mu$ = grand mean, $\alpha_i$ = individual difference

component for subject $i$ (constant over time), $\beta_j$ = effect of time (same for all subjects) and $\varepsilon_{ij}$ = error for subject $i$ and time $j$. It is assumed that the random components are distributed as $\alpha_i \sim N(0, \sigma^2_\alpha)$ where $\sigma^2_\alpha$ is the between-subjects variance, and $\varepsilon_{ij} \sim N(0,\sigma^2_\varepsilon)$ , with $\sigma^2_\varepsilon$ is the within-subject variance. The variance-covariance structure for $y_{ij}$ is compound symmetry for ANOVA, which assumes the covariate to be of the form $\sigma^2 + \gamma\delta_{ij}$ for unknown parameters $\sigma^2$ and $\gamma$ and where $\delta_{ij}$ equals one for $i = j$ and zero otherwise.

**Example**: Data are drawn from test results on file in the records office of the Laboratory School of the University of Chicago. They consist of scores, obtained from a cohort of pupils at the eighth through eleventh grade level on alternative forms of the vocabulary section of the Cooperative Reading Tests. There are 64 students in all, 36 male, 28 female (ordered). Data can be found at [7].

The data consist of repeated measurements, and all 64 students have completed scores. At first, we need to set up the data for the univariate repeated measures ANOVA, using data step in SAS software [10] with DO loop, DO TIME = 1 TO 4; VOCAB = V(TIME); OUTPUT; END. Then, we use PROC GLM with class variables of "TIME" and "SUBJECT". The SAS codes are given below.

```
PROC GLM;
CLASS TIME SUBJECT;
MODEL VOCAB = SUBJECT TIME;
RUN;
```

The CLASS statement is used to define all variables that are regarded as categorical factors. MODEL statement specifies the response variable and the fixed effects. The fixed effects can include both discrete covariates defined in the CLASS statement and quantitative covariates that are excluded from the CLASS statement [4, 13].

Now, R-square is equal to 0.87 which indicates that the dependent variables fit and predict 87 percent of independent variable in the GLM model. The p-values of "subject" and "grade" (i.e., "time") variables are both significant (less than 0.01). Therefore, the result shows that the scores of the vocabulary section of the Cooperative Reading Tests at the eighth through eleventh grade do differ. If we explore this analysis in deep, we may find out the average of each score to see whether their vocabulary skills have increased or decreased. Using PROC MEAN, we can create the chart below. Obviously, the mean is increasing over time. That is, the pupils have improved better the vocabulary section of the cooperative reading tests with grade.

| | Vocab 1 | Vocab 2 | Vocab 3 | Vocab 4 |
|---|---|---|---|---|
| Mean | 1.137 | 2.542 | 2.989 | 3.472 |
| SD | 1.889 | 2.085 | 2.169 | 1.926 |

The main difference between ANOVA and MANOVA is that MANOVA approach must discard all missing data. The MANOVA needs to be with complete data because it treats the repeated measures as one vector and the entire data vector must be complete for the subject to be included in the analysis. Also, the MANOVA assumes a general form for the correlation of repeated measurements over time, whereas the ANOVA assumes the much more restrictive compound-symmetric form. The MANOVA model in one sample case is given by:

$\boldsymbol{y}_i = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_i$

where $\boldsymbol{\mu}$ defines $n\times 1$ mean vector for time points and $\boldsymbol{\varepsilon}_i$ defines $n\times 1$ vector of errors with $\boldsymbol{\varepsilon}_i \sim N(0, \Sigma)$ where $\Sigma$ is the variance-covariance matrix. Compared to ANOVA, the variance-covariance structure for $\boldsymbol{y}_i$ is unstructured for MANOVA, which assumes the covariate to be a general $n\times n$ positive definite covariance matrix.

**Example**: We use the same dataset as in ANOVA. For MANOVA, DO loop step is not needed in the analysis. The SAS codes are given below.

```
PROC GLM;
MODEL VOCAB1--VOCAB4 = / NOUNI SOLUTION;
REPEATED TIME POLYNOMIAL / SUMMARY PRINTE;
RUN;
```

The NOUNI option prevents separate tests. The SOLUTION option presents estimates of the fixed effect parameters. The SUMMARY option provides univariate repeated-measures results. At last, the PRINTE option gives sphericity information. The test for sphericity provides a test of the hypothesis that all measured variables are independently distributed and they have the same variance $\sigma^2$, which is assumed to be unknown. If the sphericity test holds, then F-tests are valid. If it does not hold, then F-tests are generally too liberal [13]. Here, since p-value = 0.2767, we conclude that F-tests are valid. The parameter "time" (group) indicates significant of p-value (<0.001). Therefore, we may end up same conclusion as ANOVA example.

### 3. Mixed-effect Regression Model (MRM)

Next, we discuss Mixed-effect Regression Model (MRM). MRM explicitly models individual change across time. In additions, MRM is more flexible in term of repeated measures and does not require restrictive assumptions concerning missing data across time and the variance-covariance structure of the repeated measures. Also, each subject does not need to have same number of observations per subject. That is, it can handle subjects measured incompletely or at different time points. MRM can also be used for incomplete longitudinal data. A two-level MRM model is given by: $\boldsymbol{y_i} = \boldsymbol{x_i\beta} + \boldsymbol{z_i v_i} + \boldsymbol{\varepsilon_i}$ with $i = 1 \ldots N$ individuals and $j = 1 \ldots n_i$ observations for individual $i$. Here $\boldsymbol{y_i}$ is a $n_i \times 1$ response vector for individual $i$, $\boldsymbol{x_i}$ is a $n_i \times p$ design matrix for the fixed effect, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown fixed parameter, $\boldsymbol{z_i}$ is a $n_i \times r$ design matrix for the random effects, $\boldsymbol{v_i}$ is a $r \times 1$ vector of unknown random effect, and $\boldsymbol{z_i} \sim N(0, \Sigma_v)$. At last, $\boldsymbol{\varepsilon_i}$ is a $n_i \times 1$ residual vector with $\boldsymbol{\varepsilon_i} \sim N(0, \sigma^2 I_{n1})$.

**Example**: Drug Plasma Levels and Clinical response. Riesby and associate [11] examined the relationship between Imipramine (IMI) and Desipramine (DMI) plasma levels and clinical response in 66 depressed inpatients (37 endogenous and 29 non-endogenous). As an outcome variable, Hamilton Depression Riesby Scores (HDRS) were measured. As an independent variable, Endog has endogenous or non-endogenous. IMI is antidepressant and DMI is a metabolite of imipramine. Both are drug-plasma levels. The data can be found in [7]. The SAS codes are given below.

```
PROC MIXED METHOD=ML COVTEST;
CLASS ID;
MODEL HAMD = WEEK /SOLUTION;
RANDOM INTERCEPT  /SUBJECT=ID TYPE=UN G
RUN;
```

In PROC MIXED, the RANDOM statement is used to define all effects that are considered to be random. Specially, the RANDOM statement is used to define the covariates in the design matrix for the random effects. The SOLUTION option requests a listing of the solution of the mixed model equation for $\beta$ . The METHOD option specifies estimation method. Here, we have ML (Maximum likelihood) and COVTEST which provides estimates of the standard errors of the estimated variance components. The TYPE option specifies covariance structure. UN calls for an unstructured (2×2) covariance matrix. Also, *G* requests that the

estimates of the variances and covariance of the random effects be displayed [4, 13].

This model is a random intercepts model (model I). By adding "week" variable in RANDOM statement, it becomes a random tread model (model II). Also, to look at group effects with model II, "ENDOG" and "ENDWEEK" are added MODEL statement (model III). At last, to look at the suspecting quadratic "week" trend, with model II, "WEEK*WEEK" is added in both MODEL and RANDOM statement (Model IV).

Also, to look at which models fit best, it is a good idea to compare AIC values for each model. AIC stands for Akaike's Information Criterion. Given a data set, several competing models may be ranked according to their AIC values, with the one having the lowest AIC being the best.

| Model | I | II | III | IV |
|---|---|---|---|---|
| AIC | 2293.2 | 2232.0 | 2230.9 | 2227.6 |

According to AIC number, Model IV fits the best. Notice that "week" variable and intercept are significant in all models. The intercept being significant just indicates the HDRS scores are different than zero at baseline. So, it is not particularly meaningful. In Model III, the variables "ENDOG" and "ENDWEEK" are not significant. That is, to see the relationship between Imipramine (IMI) and Desipramine (DMI) plasma levels and clinical response, the HDRS score does not affect whether it is endogenous patient or not. Also, the HDRS score does not affect when patients stop the plasma levels. At last, Model IV tests whether the HDRS score has quadratic trend over time. P-value of "WEEK*WEEK" variable has 0.5621, which is not significant. The estimation of "Week" parameter is -2.633 (Model IV). Therefore, HDRS score has linearly negative relationship with plasma levels.

| | Week0 | Week1 | Week2 | Week3 | Week4 | Week5 |
|---|---|---|---|---|---|---|
| Endogenous | 24.0 | 23.0 | 19.3 | 17.3 | 14.5 | 12.6 |
| N | 33 | 34 | 37 | 36 | 34 | 31 |
| Non-Endogenous | 22.8 | 20.5 | 17.0 | 15.3 | 12.6 | 11.2 |
| N | 28 | 29 | 28 | 29 | 29 | 27 |
| Pooled SD | 4.5 | 4.7 | 5.5 | 6.4 | 7.0 | 7.2 |

The table gives a descriptive statistics (mean, sample size, and pooled sd). It shows that the means decreases over time. Lower scores on the HDRS reflect less depression. Thus, we can conclude that patients are improving over time.

## 4. Generalized Estimating Equation (GEE)

Last model is called generalized estimating equations (GEEs) introduced by Liang and Zeger [9]. They are extension of generalized linear model (GLM) to longitudinal analysis using quasi-likelihood estimation. A basic premise of GEE approach is that one is primarily interested in the regression parameter and is not interested in the variance-covariance matrix of the repeated measures. As such, generalized estimating equations treat covariance structure as a nuisance and they are not concerned about variance of each data. They have consistent and asymptotically normal solutions by relying on the independence across subjects to estimate constantly the variance of the regression coefficient even when the assumed correlation structure is incorrect.

GEE has a "working" correlation $R$ of the repeated measurements. This working correlation matrix is of size $n \times n$ because one assumes that there is a fixed number of time-points $n$ that subjects are measured at. A given subject does not have to be measured at all $n$ time-points. Each individual's correlation matrix $R_i$ is of size $n_i \times n_i$ with appropriate rows and columns removed if $n_i < n$. It is generally

recommended that choice of $R$ should be consistent with the observed correlations. If the choice of $R$ is incorrect, efficiency such as statistical power is reduced. However, the loss of efficiency is lessened as the number of subjects gets large. Some important references in the field of generalized estimating equation can be found in [1, 3, 6, 7, 9].

**Example:** Here are repeated measures data from the "Six Cities" study of the health effects of air pollution [14]. There are 16 selected cases for children respiratory disease and mother's smoking in [10]. Mother's smoking status was determined at the first interview. The binary response is the wheezing status of 16 children at ages 9, 10, 11, and 12 years. Whether the child had respiratory infection in the year prior to each exam was reported by the mother. The mean response is modeled as a logistic regression model using the explanatory variables: city of residence, age, and maternal smoking status at the particular age. The binary responses for individual children are assumed to be equally correlated, implying an exchangeable correlation structure. The data can be found at [12]. In SAS, PROC GENMOD procedure is used for GEE. The codes are given below.

```
PROC GENMOD DATA=SIX ;
CLASS CASE CITY ;
MODEL WHEEZE = CITY AGE SMOKE / DIST=BIN;
REPEATED SUBJECT=CASE / TYPE=EXCH COVB CORRW;
RUN;
```

Although PROC GENMOD is primarily a procedure for fitting generalized linear models to a single response, the REPEATED statement invokes the GEE method, specifies the correlation structure, and controls the displayed output from the GEE model. The DIST option specifies the default canonical link function and variance function that happen to be associated with particular exponential family distribution. The REPEATED statement distinguishes the fitting of a generalized linear model for a single univariate responses via maximum likelihood from the fitting of a marginal model to a vector of correlated responses using the GEE method. The statement is used to specify the assumed structure of the within-subject association among the repeated measurements. The option SUBJECT= CASE specifies that individual subjects are identified in the input data set by the variable case. The TYPE=EXCH option specifies an exchangeable working correlation structure, and for a complete discussion, the "working" correlation structure for exchangeable is $R'_{jj} = \rho$, that is, all of the correlations are equal. The COVB option specifies that the parameter estimate covariance matrix be displayed, and the CORRW option specifies that the final working correlation

be displayed [4, 12].

The result shows all parameters are not significant. That is, with provided data, we conclude that there is no relationship between children's respiratory disease and mother's smoking from "Six Cities"

**Example**: Dataset, schzx1.dat, has severity of illness scores on 437 schizophrenics measured across time. Subjects were randomized to one of four treatments: placebo, chlorpromazine, fluphenazine, or thioridazine. Here the drug groups have been combined into one group. The data file contains, in order, Patient ID, IMPS79 (7-point severity scale), IMPS79b (binary version of IMPS79), IMPS79o (ordinal version of IMPS79), intercept (a column of ones), treatment group (0=placebo, 1=drug), week (week 0 to week 6, though most of the measurement occurred on weeks 0, 1, 3, & 6), SWEEK (square root of week, helping to linearize the relationship of IMPS79 over time), and TXSWK (treatment group by square root of week). The data can be found at [7]. In all, 102 out of 437 did not complete the trial and treat as missing data. This example illustrates how to use GEE for different "working" covariate structure. The SAS codes are given below:

```
PROC GENMOD DESCENDING;
 CLASS ID WEEK;
 MODEL  IMPS79b = TX SWEEK TXSWK / LINK=LOGIT   DIST=BIN;
 REPEATED SUBJECT=ID / WITHIN=WEEK CORRW TYPE=EXCH;
 RUN;
```

This is a logistic regression model for binary outcomes as LINK option and DIST option indicate. The LINK statement specifies the choice of built-in link function relating the mean response to the linear predictor. The TYPE option determines which "working" correlation we choose. For example, EXCH stands for Exchangeable Structure, AR(1) for first-order Autoregressive Structure, MDEP(6) for Toepliz (banded) structure, and UN for Unstructured [4,12]. All Models showed "SWEEK" to be significant with negative estimation. Therefore, severity of illness scores decrease slowly across time.

## 5. Discussion

Analysis of longitudinal data is a crucial statistical approach that is widely used in the health, social, and biological sciences. In this article we discussed four statistical models for longitudinal data analysis-ANOVA, MANOVA, MRM, and

GEE. ANOVA and MANOVA are well-known and easy to manipulate in SAS, and both models assume interval measurement and normally distributed errors that are homogeneous across groups. ANOVA assumes compound symmetry which has little validity for longitudinal. And, MAOVA does not permit missing data. They only estimate and compare the group mean and not informative about individual growth. MRM models are quite widely used for analysis of longitudinal data. These models can be applied to ordinal outcomes or nominal or count outcomes that have a Poisson distribution, which we have not discussed in this paper. The advantage of MRM is that missing data are ignorable if the missing responses can be explained either by covariates in the model or by the available responses from a given subject. The disadvantage is that full-likelihood methods are more computationally complex than quasi-likelihood methods such as GEE. When the scientific interest is in estimation and inference of the regression parameters and not of the variance-covariance structure, GEE provides standard errors that are robust to mis-specifiction of the variance-covariance structure. Also, as stated, GEE is often used as a general and computationally convenient method. In fact, software for performing GEE analysis is available in most of the major statistical software packages. The disadvantage is that missing data are only ignorable if the missing data are explained by covariates in the model. This is a more stringent assumption than MRM, and therefore GEE models have somewhat limited applicability to incomplete longitudinal data.

## References

[1] G. A. Ballinger, Using Generalized Estimating Equations for Longitudinal Data Analysis, Organizational Research Methods, **7** (2004), 127-150.

[2] C. Bijleveld, L. van der Kamp, A. Mooijaart, W. van der Kloot, R. van der Leeden, E. van der Burg, Longitudinal Data Analysis: Designs, Models & Methods, Sage publications, 1999.

[3] P.J. Diggle, P. Heagerty, K.-Y. Liang, and S.L. Zeger, Analysis of Longitudinal Data, 2nd edition. Oxford University Press, New York, 2002.

[4] G.M. Fizmaurice, N.M, Laird, & J.H. Ware, Applied Longitudinal Analysis, Wiley, New Jersey, 2004.

[5] G.M. Fizmaurice, M. Davidian, G. Verbeke, G. Molenberghs, Longitudinal Data Analysis: A Handbook of Modern Statistical Methods, Chapman & Hall/CRC, Florida, 2008.

[6] J.W. Hardin & J.M. Hilbe, Generalized Estimating Equations, Chapman & Hall, Florida, 2003.

[7] D. Hedeker & R.D. Gibbons, Longitudinal Data Analysis, Wiley-InterScience, New Jersey, 2006.

[8] N.M. Laird and J.H. Ware, Random-effects models for longitudinal data, Biometrics, **38** (1982) ,963-974.

[9] K.-Y. Liang and S.L. Zeger, Longitudinal Data Analysis Using Generalized Linear Models, Biometrika, **73** (1986), 13-22.

[10] S.R. Lipsitz , G.M.Fitzmaurice, E.J. Orav, N.M. Laird, Performance of generalized estimating equations in practical situations. Biometrics, **50** (1994), 270–278.

[11] N. Reisby, L. Gram, P. Bech, A. Nagy, G. Petersen, J. Ortmann, I. Ibsen, S. Dencker, O. Jacobsen, O. Krautwald, I. Sndergaard, and J. Christiansen, Imipramine: Clinical effect and pharmacokinetic variability, Psychopharmacology, **54** (1997), 263-272.

[12] SAS Institute Inc. *SAS/STAT® User's Guide, Version 8*, Cary, NC, 1999.

[13] SAS Institute Inc. SAS® for Mixed Models: Second Edition, Cary, NC, 2006.

[14] J.H. Ware, D.W. Dockery, A. III Spiro, F.E. Speizer, B.G. Jr. Ferris, Passive Smoking Gas Cooking, and Respiratory Health of Children Living in Six Cities, American Review of Respiratory Diseases, **129** (1984), 366-374.

Received: November, 2008