Analyzing Longitudinal Data

Douglas G. Bonett
University of California, Santa Cruz

3/4/2014

Overview

- Comparison of traditional and modern methods
- Data file structures
- Time-varying and time-invariant covariates
- Modeling nonlinearity and interactions
- Modeling treatment effects
- Error covariance structures
- Models with random coefficients

A Comparison of Traditional and Modern Methods

Traditional Methods

paired-samples *t*-test and repeated measures ANOVA

MANOVA/MANCOVA

trend analysis

Advantages

tests and confidence intervals are “exact” in small samples

easy to use in SPSS, SAS and R

Disadvantages

requires equally spaced time intervals (trend analysis)

list-wise deletion is required with missing data

cannot accommodate time-varying covariates

A Comparison of Traditional and Modern Methods

Modern Methods

covariance pattern (CP) models
random coefficient (RC) models

Advantages

will use all available data
can handle unequally spaced time intervals
not all participants must be measured at all time periods
can accommodate time-varying covariates
can describe variability in model parameters (RC)

Disadvantages

computer programs are more difficult to use

tests and confidence intervals are only approximate in small samples

Data File Structure for CP and RC Models

CP and RC analyses require the data to be in a “long” format.

Suppose we have data for 30 participants who are each measured at 3 time points. The traditional “wide” format would have the following form and would have 30 rows and 3 columns (the Participant column is not analyzed). If data are in a wide format, the *Restructure* option in SPSS (under *Data* tab) can be used to convert the data file into a long format.

Participant	Time 1	Time 2	Time 3
1	24	29	30
2	16	15	18
…	…	…	…
30	20	21	25

Data File Structure for CP and RC Models

For a CP or RC analysis, these data would be entered in a "long" format as shown below. The data file has 90 rows and 3 columns. The Participant variable is used in a CP and RC model specification.

Participant	Time	Score
1	1	24
1	2	29
1	3	30
2	1	16
2	2	15
2	3	18
…	…	…
30	1	20
30	2	21
30	3	25

Time Invariant and Time Varying Covariates

Participant	DV	Grade	Sex	TEE
1	54	1	1	17
1	56	2	1	8
1	67	3	1	4
1	60	4	1	10
2	47	1	0	9
2	42	2	0	14
2	48	3	0	12
2	49	4	0	5
…	…	…	…	…

- Sex is a *time-invariant covariate*

- TEE (Teacher Emotional Exhaustion) is a *time-varying covariate*

Modeling Nonlinearity

Participant	DV	Day	Day^2	Day^3
1	54	-2	4	-8
1	56	-1	1	-1
1	67	0	0	0
1	60	1	1	1
1	64	2	4	8
2	47	-2	4	-8
2	42	-1	1	-1
2	48	0	0	0
2	49	1	1	1
2	52	2	4	8
…	…	…	…	…

Time for linear trend

$Time^2$ for quadratic trend

$Time^3$ for cubic trend

Note: Mean centering the Time variable reduces multicollinearity

Modeling Interaction Effects

Participant	DV	Grade	Sex	Grade x Sex
1	54	-2	1	-2
1	56	-1	1	-1
1	67	0	1	0
1	60	1	1	1
1	51	2	1	2
2	47	-2	0	0
2	42	-1	0	0
2	48	0	0	0
2	49	1	0	0
2	54	2	0	0
…	…	…	…	…

Note: Mean centering the Grade variable reduces multicollinearity

Modeling Treatment Effects

Example: 2-group repeated measured design

Participant	DV	Month	Treatment
1	54	1	1
1	56	2	1
1	67	3	1
1	60	4	1
2	47	1	0
2	42	2	0
2	48	3	0
2	49	4	0
…	…	…	…

Modeling Treatment Effects (*continued*)

Example: one-sample multiple pretest-posttest design

Participant	DV	Month	Treatment
1	54	1	0
1	56	2	0
1	67	3	1
1	60	4	1
2	47	1	0
2	42	2	0
2	48	3	1
2	49	4	1
…	…	…	…

Modeling Treatment Effects (*continued*)

Example: two-sample multiple pretest-posttest design

Participant	DV	Month	Treatment
1	54	1	0
1	56	2	0
1	67	3	1
1	60	4	1
2	47	1	0
2	42	2	0
2	48	3	0
2	49	4	0
…	…	…	…

Participant 1 is in treatment group

Participant 2 is in control group

The General Covariance Pattern (CP) Model

A CP model for participant i measured at t time periods:

$$y_{ij} = b_0 + b_1 t_{ij} + b_2 x_{2ij} + \cdots + b_q x_{qij} + e_{ij}$$

for j = 1 to t.

Unlike the traditional multiple regression model, it is not reasonable to assume that the prediction errors ($e_{i1}, e_{i2}, \ldots, e_{it}$) are uncorrelated and have equal variances. With longitudinal data, the errors will be usually be correlated and could also have unequal variances.

CP Model Error Covariance Structures

Unstructured $(t = 4)$

$$\begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 & \rho_{14}\sigma_1\sigma_4 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 & \rho_{24}\sigma_2\sigma_4 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 & \rho_{34}\sigma_3\sigma_4 \\ \rho_{14}\sigma_1\sigma_4 & \rho_{24}\sigma_2\sigma_4 & \rho_{34}\sigma_3\sigma_4 & \sigma_4^2 \end{bmatrix}$$

This error structure allows the $t = 4$ prediction error variances (diagonal elements) to be unequal and the covariances (off-diagonal elements) to be unequal.

CP Model Error Covariance Structures (*continued*)

Compound Symmetric (t = 4)

$$\begin{bmatrix} \sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \sigma^2 \end{bmatrix}$$

This error structure assumes that the t = 4 prediction error variances (σ^2) are all equal and that all covariances ($\rho\sigma^2$) are equal.

CP Model Error Covariance Structures (*continued*)

***AR*(1)** (t = 4)

$$\begin{bmatrix} \sigma^2 & \rho\sigma^2 & \rho^2\sigma^2 & \rho^3\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \rho^2\sigma^2 \\ \rho^2\sigma^2 & \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ \rho^3\sigma^2 & \rho^2\sigma^2 & \rho\sigma^2 & \sigma^2 \end{bmatrix}$$

This error structure assumes that the t = 4 prediction error variances (σ^2) are all equal and that the correlation between two prediction errors becomes smaller as the time separation increases.

CP Model Error Covariance Structures (*continued*)

***ARMA*(1,1)** (t = 4)

$$\begin{bmatrix} \sigma^2 & \gamma\sigma^2 & \gamma\rho\sigma^2 & \gamma\rho^2\sigma^2 \\ \gamma\sigma^2 & \sigma^2 & \gamma\sigma^2 & \gamma\rho\sigma^2 \\ \gamma\rho\sigma^2 & \gamma\sigma^2 & \sigma^2 & \gamma\sigma^2 \\ \gamma\rho^2\sigma^2 & \gamma\rho\sigma^2 & \gamma\sigma^2 & \sigma^2 \end{bmatrix}$$

This error structure assumes that the t = 4 prediction error variances (σ^2) are all equal and that the correlation between two prediction errors becomes smaller as the time separation increases, but the decreases are more gradual than in an *AR*(1) structure.

CP Model Error Covariance Structures (*continued*)

Toeplitz (t = 4)

$$\begin{bmatrix} \sigma^2 & \rho_1\sigma^2 & \rho_2\sigma^2 & \rho_3\sigma^2 \\ \rho_1\sigma^2 & \sigma^2 & \rho_1\sigma^2 & \rho_2\sigma^2 \\ \rho_2\sigma^2 & \rho_1\sigma^2 & \sigma^2 & \rho_1\sigma^2 \\ \rho_3\sigma^2 & \rho_2\sigma^2 & \rho_1\sigma^2 & \sigma^2 \end{bmatrix}$$

This error structure assumes that the t = 4 prediction error variances (σ^2) are all equal and that the correlations are equal only within each band that is parallel to the main diagonal.

CP Model Error Covariance Structures (*continued*)

***Heteroscedastic AR*(1)** has an AR(1) correlation structure but does not assume equal variances.

Heteroscedastic Toeplitz has a Toeplitz correlation structure but does not assume equal variances.

The *AR*(1), *ARMA*(1,1) and *Toeplitz* structures (with or without the equal variance assumption) both assume the time periods are approximately equally spaced. The *Unstructured* form does not require equally spaced time periods.

Choosing a CP Model Error Covariance Structure

If there are no missing data and the number of participants is greater than the number of time periods plus number of model parameters, then the *Unstructured* or *Heteroscedastic Toeplitz* can be used. These two structures provide the most realistic description of the true error covariance structure. The *Heteroscedastic AR*(1) error structure might also be a realistic option.

If the number of time periods is greater than the number of participants and there is substantial missing data, then it might be necessary to use a very simple error covariance structure such as *AR*(1), *ARMA*(1,1), or *Compound Symmetric*.

SPSS Example

A mood questionnaire was given to a sample of 30 assembly line workers on Monday, Wednesday and Friday ($t = 3$).

The *Mixed Model – Linear* option in SPSS will be used to analyze the data using a CP model with an *Unstructured* error covariance matrix and day of week as a predictor of mood.

SPSS will be used to estimate the intercept and slope of the following CP model:

$$y_{ij} = b_0 + b_1 t_j + e_{ij}$$

RC Models for Longitudinal Data

Consider the following "level-1" model for participant i

$$y_{ij} = b_{0i} + b_{1i}t_{ij} + e_{ij}$$

where the prediction errors (e_{ij}) are assumed to have equal variances and be uncorrelated.

The n participants are assumed to be randomly sampled from some population. This implies that the coefficients b_{0i} and b_{1i} can be treated as random variables.

RC Models for Longitudinal Data (*continued*)

We can write a "level-2" statistical model for each of these random coefficients.

$$b_{0i} = g_{00} + r_{0i}$$
$$b_{1i} = g_{01} + r_{1i}$$

Substituting these equations into $y_{ij} = b_{0i} + b_{1i}t_{ij} + e_{ij}$ gives the following "composite model"

$$y_{ij} = g_{00} + g_{01}t_{ij} + r_{0i} + t_{ij}r_{1i} + e_{ij}$$

r_{0i} and r_{1i} are the random effects and g_{00} and g_{01} are the fixed effects, hence the name "mixed-effects models".

RC Models for Longitudinal Data (*continued*)

The composite model can be written as

$$y_{ij} = g_{00} + g_{01}t_{ij} + e^*_{ij}$$

where $e^*_{ij} = r_{0i} + t_{ij}r_{1i} + e_{ij}$ is the composite prediction error.

After some covariance algebra, it can be shown that

$$var(e^*_{ij}) = var(r_0) + var(r_1)t_j^2 + 2cov(r_0, r_1)t_j + var(e_0)$$

$$cov(e^*_{ij}, e^*_{ij'}) = var(r_0) + var(r_1)t_j t_{j'} + cov(r_0, r_1)(t_j + t_{j'})$$

RC Models for Longitudinal Data (*continued*)

We started with the model $y_{ij} = b_{0i} + b_{1i}t_{ij} + e_{ij}$ where e_{ij} were assumed to be uncorrelated and have equal variances. But if b_{0i} and b_{1i} are random coefficients, we obtain the following model

$$y_{ij} = g_{00} + g_{01}t_{ij} + e_{ij}^{*}$$

where the prediction errors are *not* assumed to be uncorrelated or have equal variances.

The variances and covariances of the prediction errors (e_{ij}^{*}) are not completely unrestricted but are specific functions of time.

RC Models with Interaction Effects

Consider again the model $y_{ij} = b_{0i} + b_{1\mathrm{i}}t_{ij} + e_{ij}$ where b_{0i} and b_{1i} are random coefficients, but suppose that these random coefficients are believed to be related to a predictor variable (x).

For example, if y represents a "social skills" measurement of preschool children taken four times during the school year, we might suspect that the intercept value and slope value for a child is related to that child's vocabulary size (x) at the beginning of the year.

RC Models with Interaction Effects (*continued*)

The level-2 models for the two random coefficients are now

$$b_{0i} = g_{00} + g_{10}x_i + r_{0i}$$
$$b_{1i} = g_{01} + g_{11}x_i + r_{1i}$$

Substituting these equations into $y_{ij} = b_{0i} + b_{1i}t_{ij} + e_{ij}$ gives

$$y_{ij} = g_0 + g_{10}x_j + g_{01}t_{ij} + g_{11}(x_j t_{ij}) + e^*_{ij}$$

where the prediction errors (e^*_{ij}) will be correlated and have unequal variances. Note that g_{11} describes the interaction effect of time and vocabulary.

RC Models with Interaction Effects (*continued*)

When the random slope for the time variable (t) is assumed to be predictable from some other variable (x), this implies an interaction between t and x.

When analyzing a RC model using a mixed-effect statistical program, product variables are entered as predictor variables in exactly the same way they would have been specified in a CP model.

General RC Models

A more general level-1 RC model for participant i measured at t time periods is:

$$y_{ij} = b_{0i} + b_{1i}t_{ij} + b_{2i}x_{2ij} + \cdots + b_{qi}x_{qij} + e_{ij}$$

where b_{0i} and b_{1i} are assumed to be random coefficients, but now we also allow the prediction errors of the level-1 model (e_{ij}) to be correlated and have unequal variances. As in a CP model, the predictor variables can be polynomial functions of time, product variables, time-varying covariates, or time-invariant covariates.

General RC Models (*continued*)

In theory, the prediction errors of the level-1 RC model can have any of the covariance structures defined above for the CP model.

In practice, the parameters of these covariance structures can be so highly correlated with the variances and covariances of the random coefficients that they cannot be estimated. In these situations, SPSS will give a “Hessian matrix not positive definite” error message.

SPSS Random Coefficient Example

Let's reanalyze the longitudinal mood study in SPSS using the following RC model

$$y_{ij} = b_{0i} + b_{1i}t_j + e_{ij}$$

where the intercept (b_0) and slope (b_{1i}) coefficients are assumed to be random (i = 1 to n; j = 1 to t).

y = mood

t = day of week

Issues in Using CP and RC Models

CP and RC statistical packages provide options for maximum likelihood estimation (MLE) and restricted maximum likelihood estimation (RMLE). RMLE gives more accurate standard errors and is the recommended method.

MLE is needed to perform certain nested model comparison tests where the models differ in their error structures *and* the number of predictor variables.

CP models with *AR*(1), *ARMA*(1,1), and *Toeplitz* error structures require approximate equal spacing of time periods. Time periods can be unequally spaced with RC models or CP models with an *Unstructured* error covariance matrix.

Issues in Using CP and RC Models (*continued*)

RC models can handle studies where the number of time periods exceeds the number of participants and not all participants are measured at every time point. CP models can also handle this kind of data if a very simple error structure such as *AR*(1), *ARMA*(1,1), or *Compound Symmetry* is used.

RC models provide useful information about the variance of person-level intercept and slope coefficients in the population – however, if the number of participants is small, the confidence intervals can be uselessly wide.

Issues in Using CP and RC Models (*continued*)

Recall that a random intercept and a random slope implies the following error covariance structure:

$$var(e_{ij}^*) = var(r_0) + var(r_1)t_j^2 + 2cov(r_0, r_1)t_j + var(e_0)$$

$$cov(e_{ij}^*, e_{ij'}^*) = var(r_0) + var(r_1)t_j t_{j'} + cov(r_0, r_1)(t_j + t_{j'})$$

This covariance error structure is parsimonious (a function of only 4 parameters) but it might not be realistic. A CP model with a similarly parsimonious but more realistic error structure could be a better choice.

Suggested Readings

Davis, C.D. (2002). *Statistical methods for the analysis of repeated measures.* New York: Springer. (Chapters 1 - 6)

Singer, J.D. & Willett, J.B. (2003). *Applied Longitudinal Data Analysis: Modeling change and event occurrence*. New York: Oxford. (Chapters 1 – 8)

Heck, R.H., Thomas, S.L., Tabata, L.N. (2014). *Multilevel and longitudinal modeling with IBM SPSS, 2nd ed*. New York: Routedge. (Chapters 5 & 6)

Related Topics for Future Seminars

- Latent growth curve models
- Analyzing longitudinal binary responses
- Survival models
- Analyzing longitudinal data using R

Self-study Materials on CSASS Website

- PowerPoint slides
- Step-by step SPSS instructions for analyzing three different types of longitudinal studies
- SPSS .sav file for each example
- SPSS syntax for each example

Thank you.

Questions or comments?