BSTT537: Longitudinal Data Analysis

Don Hedeker

Advantages of Longitudinal Studies (chapter 1)

- 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)
  $\Rightarrow$ cross-sectional design can't do this
- Can provide information about individual change

Challenges of Longitudinal Data Analysis

- Observations are not, by definition, independent $\Rightarrow$ must account for dependency in data
- Analysis methods not as well developed, especially for more sophisticated models
- Lack and difficulty of using software
- Computationally intensive
- Unbalanced designs, missing data, attrition
- Time-varying covariates
- Carry-over effects (when repeated factor is condition or treatment, not time)

Notation

- $i = 1, \ldots, N$ subjects
- $j = 1, \ldots, n_i$ observations (n for balanced designs)
- total number of observations $= \Sigma_i^N n_i$
- $\boldsymbol{y}_i \ = \ n_i \times 1$ vector of responses
- $\boldsymbol{x}_{ij} \ = \ p \times 1$ covariate vector for subject i at time j
  - time-invariant or time-independent covariates (between-subjects)
  - time-varying or time-dependent covariates (within-subjects)
- $\boldsymbol{X}_i = n_i \times p$ matrix of covariates for subject i usually includes an intercept term

Data Layout

subject	observation	response	covariates		
1	1	y_{11}	x_{111}	$\ldots$	x_{11p}
1	2	y_{12}	x_{121}	$\ldots$	x_{12p}
.	.	.	.	. .	
1	n_1	y_{1n_1}	x_{1n_11}	$\ldots$	x_{1n_1p}
.	.	.	.	. .	
.	.	.	.	. .	
.	.	.	.	. .	
.	.	.	.	. .	
N	1	y_{N1}	x_{N11}	$\ldots$	x_{N1p}
N	2	y_{N2}	x_{N21}	$\ldots$	x_{N2p}
.	.	.	.	. .	
N	n_N	y_{Nn_N}	x_{Nn_N1}	$\ldots$	x_{Nn_Np}

- n_i varies by subjects (some analyses won't allow this)
- above is "univariate layout"
- different layout for repeated measures MANOVA ("multivariate layout")
- if x_r is time-invariant (between-subjects) $x_{i1r} = x_{i2r} = x_{i3r} = \ldots = x_{in_ir}$

Analysis Considerations

- Response variable
  - continuous (normal or non-normal)
  - categorical (dichotomous, ordinal, nominal, counts)
- Number of subjects N
- Number of observations per subject n_i
  - $n_i = 2$ for all: change score analysis or ANCOVA
  - $n_i = n$ for all: balanced design - ANOVA or MANOVA for repeated measures
  - n_i varies: more general methods

- Number & type of covariates - $E(\boldsymbol{y}_i)$
  - one sample
  - multiple samples
  - regression (continuous or categorical covariates)
  - time-varying covariates
- Type of variance-covariance structure - $V(\boldsymbol{y}_i)$
  - homogeneous or heterogeneous variances
  - homogeneous or heterogeneous covariances

General Approaches

- Derived variable: not really longitudinal, per se, reduce the repeated observations into a summary variable
  - average across time
  - change score
  - linear trend across time
  - last observation
- Longitudinal Analysis
  - ANOVA for repeated measures
  - MANOVA for repeated measures
  - Mixed-effects regression models
  - Covariance pattern models
  - Generalized Estimating Equations (GEE) models

Simplest Longitudinal Analysis

Paired t-test can be used to address whether there is significant average change between two timepoints

- $i = 1, \ldots, N$ subjects
- y_{i1} = pre-test
- y_{i2} = post-test
- $d_i = y_{i2} - y_{i1}$ = post to pre change score

$H_0 : \mu_{y_1} = \mu_{y_2}$ same as $H_0 : (\mu_{y_2} - \mu_{y_1}) = 0$

test statistic

$$
\begin{aligned}
t &= \bar{d} \;/\; \left(s_d/\sqrt{N}\right) \\
&= \bar{d} \;/\; \left(\sqrt{\left[\sum_i d_i^2 - (\sum_i d_i)^2/N\right]/(N-1)} \;/\; \sqrt{N}\right) \\
&\overset{H_0}{\sim} t_{N-1}
\end{aligned}
$$

Notice, can do the same test using regression model

$$d_i = \beta_0 + e_i$$

and testing $H_0 : \beta_0 = 0$

Change Score analysis

Suppose there is a grouping variable

- $x_i = 0$ for controls
- $x_i = 1$ for treatment group

$$d_i \; = \; \beta_0 \; + \; \beta_1 x_i \; + \; e_i$$

- testing $H_0 : \beta_0 = 0$ tests whether the average change is equal to zero for the control group
- testing $H_0 : \beta_1 = 0$ tests whether the average change is equal for the two groups

notice

$$d_i = \beta_0 + \beta_1 x_i + e_i$$

$$y_{i2} - y_{i1} = \beta_0 + \beta_1 x_i + e_i$$

$$y_{i2} = y_{i1} + \beta_0 + \beta_1 x_i + e_i$$

$\Rightarrow$ change score analysis assumes that the slope for $y_{i1} = 1$

Analysis of covariance of post-test scores

$$y_{i2} = \beta_0 \ + \ \beta_1 x_i \ + \ \beta_2 y_{i1} \ + \ e_i$$

- testing $H_0 : \beta_0 = 0$ tests whether the average post-test is equal to zero for the control group subjects with zero pre-test
- testing $H_0 : \beta_1 = 0$ tests whether the post-test is equal for the two groups, given the same value on the pre-test (*i.e.*, conditional on pre-test)
- testing $H_0 : \beta_2 = 0$ tests whether the post-test is related to the pre-test, conditional on group

Change score analysis and ANCOVA answer different questions

- change score: is average change the same between the groups
- ancova: is post-test average the same between groups for sub-populations with the same pre-test values (*i.e.*, is the conditional average the same between the groups)

Which to use?

- depends on the question of interest
- often yield similar conclusions for group effect
- if subjects randomized to group, then ANCOVA is more efficient (*i.e.*, more powerful)
- must be careful in non-randomized settings, where groups are not necessarily similar in terms of pre-test scores
  - Lord's paradox (Bock, 1975; Allison, 1990)

group	September	June
Males	·	·
Females	·	·

ANCOVA of change scores

$$d_i = \beta_0 + \beta_1 x_i + \beta_2 y_{i1} + e_i$$

$$y_{i2} - y_{i1} = \beta_0 + \beta_1 x_i + \beta_2 y_{i1} + e_i$$

$$y_{i2} = \beta_0 + \beta_1 x_i + (1 + \beta_2) y_{i1} + e_i$$

$\Rightarrow$ yields equivalent results for testing $H_0 : \beta_1 = 0$ as ordinary ANCOVA model

Comparison of Pre Post models

X_i = pre, Y_i = post, G_i = group (0=control, 1=test)

Post t-test

$$Y_i = \beta_0 + \beta_1 G_i + \epsilon_i$$

Change score t-test

$$(Y_i - X_i) = \beta_0 + \beta_1 G_i + \epsilon_i$$

ANCOVA

$$Y_i = \beta_0 + \beta_1 G_i + \beta_2 X_i + \epsilon_i$$

$H_0 : \beta_1 = 0$ is test of interest in all cases

Simulation results: tests of $H_0 : \beta_1 = 0$

- 10000 datasets with 100 subjects in each of 2 groups
- mean difference of 0 at pre, .4 at post
- variance = 1 at both timepoints for both groups
- correlation = .4, .45, .5, .55, .6 between pre and post measurements

correlation	model	rejection rate
0.400	ttest	0.81
0.400	change	0.73
0.400	ancova	0.87
0.450	ttest	0.81
0.450	change	0.77
0.450	ancova	0.89
0.500	ttest	0.81
0.500	change	0.81
0.500	ancova	0.91
0.550	ttest	0.81
0.550	change	0.85
0.550	ancova	0.92
0.600	ttest	0.81
0.600	change	0.88
0.600	ancova	0.94

Example - The Television School and Family Smoking Prevention and Cessation Project (Flay, *et al.*, 1988); a subsample of this project was chosen with the characteristics:

- *sample* - 1600 7th-graders - 135 classrooms - 28 LA schools
  - between 1 to 13 classrooms per school
  - between 2 to 28 students per classroom
- *outcome* - knowledge of the effects of tobacco use
- *timing* - students tested at pre and post-intervention
- *design* - schools randomized to
  - a social-resistance classroom curriculum (CC)
  - a media (television) intervention (TV)
  - CC combined with TV
  - a no-treatment control group

Change across time?

From SAS PROC MEANS:

Variable	N	Mean	Std Dev	Minimum	Maximum
PRETHKS	1600	2.06938	1.26018	0	6.00000
POSTHKS	1600	2.66188	1.38293	0	7.00000
THKSdelt	1600	0.59250	1.57932	-5.00000	6.00000

From PROC UNIVARIATE on THKSdelt (change score):

```
   Location             Variability

 Mean 0.592500     Std Deviation 1.57932

          Tests for Location:  Mu0=0

 Test             -Statistic- -----p Value------

 Student's t    t 15.00646  Pr > |t| < .0001
```

From PROC REG of THKSdelt (with no regressors):

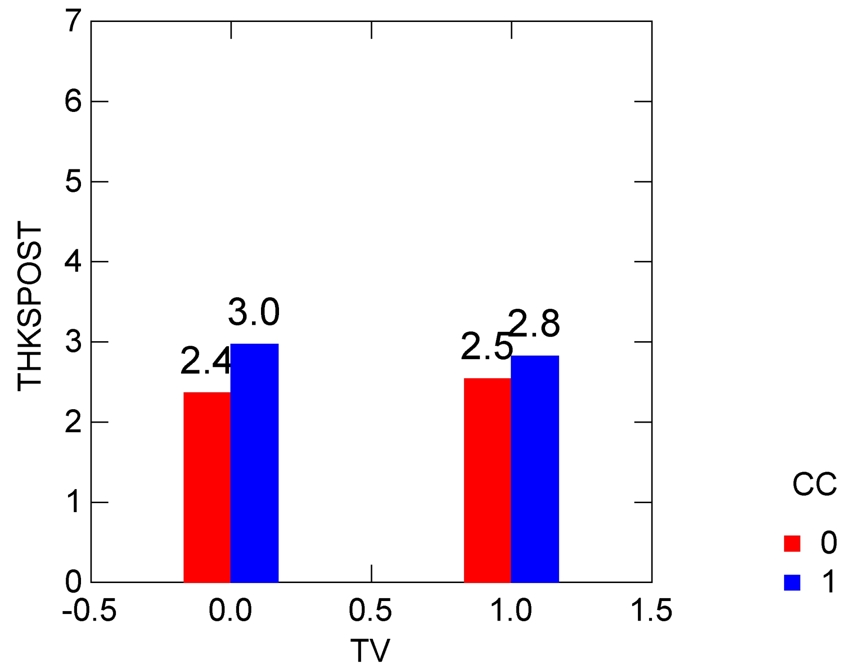
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.59250	0.03948	15.01	$< .0001$

Tobacco and Health Knowledge Scale - Subgroup Descriptives
Pretest, Post-Intervention, and Difference

	CC = no		CC = yes	
	TV = no	TV = yes	TV = no	TV = yes
N	421	416	380	383
Pretest mean	2.152	2.087	2.050	1.979
sd	1.182	1.288	1.285	1.286
Post-Int mean	2.361	2.539	2.968	2.823
sd	1.296	1.437	1.405	1.312
Difference	0.209	0.452	0.918	0.844

Does change across time vary by CC, TV, or both?

Regression of PostTHKS scores



Model with CC, TV, CC $\times$ *TV* $(R^2 = .029, \hat{\sigma}^2 = 1.86)$

Variable	Estimate	Std Error	t Value	Pr > |t|
Intercept	2.36105	0.06646	35.52	<.0001
CC	0.60738	0.09649	6.29	<.0001
TV	0.17742	0.09427	1.88	0.0600
CCTV	-0.32338	0.13652	-2.37	0.0180

Model adding PreTHKS $(R^2 = .117, \hat{\sigma}^2 = 1.69)$

Variable	Estimate	Std Error	t Value	Pr > |t|
Intercept	1.66126	0.08436	19.69	<.0001
PRETHKS	0.32518	0.02585	12.58	<.0001
CC	0.64055	0.09210	6.95	<.0001
TV	0.19871	0.08996	2.21	0.0273
CCTV	-0.32162	0.13025	-2.47	0.0136

Regression of Difference scores

Model with CC, TV, CC $\times$ TV $(R^2 = .034, \hat{\sigma}^2 = 2.41)$

Variable	Estimate	Std Error	t Value	Pr > |t|
Intercept	0.20903	0.07573	2.76	0.0058
CC	0.70939	0.10995	6.45	<.0001
TV	0.24290	0.10742	2.26	0.0239
CCTV	-0.31798	0.15556	-2.04	0.0411

Model adding PreTHKS ($R^2 = .323, \hat{\sigma}^2 = 1.69$)

Variable	Estimate	Std Error	t Value	Pr > |t|
Intercept	1.66126	0.08436	19.69	<.0001
PRETHKS	-0.67482	0.02585	-26.10	<.0001
CC	0.64055	0.09210	6.95	<.0001
TV	0.19871	0.08996	2.21	0.0273
CCTV	-0.32162	0.13025	-2.47	0.0136

Notice, $1 - .67482 = .32518$