

# An Introduction to Modeling Longitudinal Data

## Session I: Basic Concepts and Looking at Data

Robert Weiss

Department of Biostatistics  
UCLA School of Public Health  
robweiss@ucla.edu

August 2010

## Five Sessions

- 1 Session I: Basic Concepts; Looking at Data. Pages 1-44
- 2 Session II: Continuous Data; Time trends. Pages 1-52
  - ▶ Session II: Example. Pages 1-18
- 3 Session III: Covariance Models. Pages 1-32
  - ▶ Session III: Example. Pages 1-10
- 4 Session IV: Discrete Outcomes. Pages 1-14.
  - ▶ Session IV Example. Pages 1-8.
- 5 Session IV: Discrete Outcomes. Pages 15-24.
  - ▶ Session IV Example. Pages 9-15.
- 6 Session V: Two Longitudinal Outcomes. Pages 1-28

# An Introduction to Longitudinal Data Analysis: Part I

- Basic Concepts, Time
  - ▶ Introduction to Longitudinal data
  - ▶ Basic Concepts
  - ▶ Time
- Inspecting data
  - ▶ Looking at predictors
  - ▶ Plotting subject profiles
  - ▶ Empirical profile plots

# What is Longitudinal Data

The same outcome measured on subjects repeatedly across time.

- A special type of repeated measures.
- Observations on a subject separated by time.
- Time may be measured in minutes, days, months, years.
- Examples are legion: Height, weight, blood pressure, depression, anxiety, days of drug use, number of sex partners.
- Multiple subjects (one subject defines *time series*).
- Univariate longitudinal data: 1 outcome measured across time.
- Multivariate longitudinal data: 2 or more outcomes measured across time.

Outcomes can be of many types: Continuous, Count, binary.

- Continuous Example. Psychometric scales (Negative Coping Style; Brief Symptom Inventory), usually treated as continuous.
- Count Example. Counts of drug and sex behaviors in past 3 months.
- Binary Examples. Drug use yes/no, behavior yes/no. HIV-status disclosure to sex partner yes/no.

# Randomized Intervention Studies

- First analysis: Intervention effect on main outcome(s)
- Treatment groups *should be* similar at baseline
- Need to model time trend in each intervention groups
- Secondary analyses:
  - ▶ Secondary outcomes
  - ▶ Other predictors (and outcomes)
  - ▶ Subgroup analyses: predictor by treatment interactions

# Observational Studies

- Comparison of groups defined by a predictor – groups can be different at baseline.
- Time trends are of interest
- And time by predictor interactions.
- Interest in predictor effects: associations between outcome and predictors.
- Predictor by time trend interactions.

In an intervention study, analyses of predictor effects are essentially observational studies.

## Choosing Times of Observations

- Total amount of time: How long between baseline and last observation?
- Often dictated by nature of funding.
- When do we collect observations?
- More frequently earlier when change is expected to happen?
- Less frequently later when outcome not expected to change as much?
- *Equally spaced*, ie 0, 3, 6, 9, 12 months. Or yearly. Ease of interpretation.
- *Logarithmic spacing*, ie baseline, then at 1, 2, 4, 8, and 16 months. Or baseline, 3, 6, 12, and 24 months.

Longitudinal data is more complex than linear regression.

- Multiple observations over time from same person are correlated
- Must model the correlations among observations
- Modeling population mean trend across time can range from annoying to difficult
- Predictor effects
- Predictor by time interaction
- Linear regression on steroids

## CLEAR: Choosing Life: Empowerment, Action, Results!

- An intervention for young HIV+ individuals.
- Counseling intervention to maintain health, HIV/STI transmission reduction, improve quality of life.
- Three groups: Phone versus In-person intervention versus control.
- Nominal visits at baseline, 3, 6, 9, 15 and 21 months.
- Control group is a delayed intervention taking place after 15 months.
- Month 21 data omitted from most analyses.

# Definitions of Words Relating to Time

- **Nominal Times** The times when data are supposed to be collected by design.
- Nominally at baseline, after 3 months, 6 months, 9 months, 15 months and 21 months.
- **Actual Time** Specific amount of time since baseline when subject actually gives data.
- **Baseline** is day 0 by definition. The first data collected on a subject.
- **Baseline predictor** A predictor measured at baseline.

# CLEAR Study Nominal Times

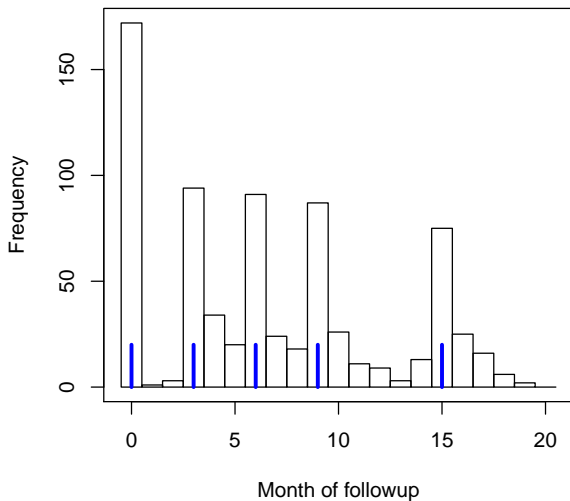
How many subjects at each of the nominal visit times in our data?

	Nominal Visit Month					
	0	3	6	9	15	21
Control ( <i>n</i> )	54	44	42	44	42	42
Interv ( <i>n</i> )	118	100	90	100	96	91
Control (%)	98	80	76	80	76	76
Interv (%)	98	83	75	83	80	76

Participants may skip a visit and participate in a later visit.

# CLEAR Study Actual Times

## Follow Up Times



When did visits actually take place?

# CLEAR Study Actual Times

When did visits actually take place? A table.

Nominal	0	3	6	9	15	21
Mean Actual	0.0	3.4	6.3	9.5	15.5	21.7
SD Actual	0.0	0.7	0.7	1.1	1.1	2.0

Mean (and SD) in months since baseline of actual visit date.

All follow up visits are significantly ( $p < .0001$ ) longer than the nominal time. Can compare inter-visit times. Only 1st follow up is significantly ( $p < .0001$ ) longer than 3 or 6 months after previous visit.

*Nobody* looks enough at their data. (Myself included).  
Why Not?

- Plotting data takes time.
- Lack of training.
- Complex data structures.
- Haste to publish.
- Crappy graphics programs. Use R instead.
- R is a fantastic graphing program.
- R graphics defaults are designed to provide high quality graphics.

R is at <http://www.r-project.org/>.

# The Importance of Looking at Your Data

- Value of plotting data: What you publish much more likely to be sensible.
- Will discover the oddities in your data.
- Costs: Apparent loss of sanctity of  $p$ -values. (Already not happening anyway.)
- Benefits: What you publish will actually be justified by your data.
- Benefits: Avoidance of Type III errors. (Accurately solving the wrong problem.)

# About Data Graphics

- Creating a well designed graph takes time: An hour or more for a good graphic, particularly the first time.
- Graphics should be created for the data analyst.
- Most graphics should show the data: *exploratory data graphics*. Not to be shown to readers of your journal article (at this time).
- Exploratory data graphics are for constructing your statistical model.
- A modest number of graphics can illustrate inferences. These might be included in a journal article.
- Practice & experience viewing graphics is extremely helpful.

# Looking at Data: Predictors

## Predictor, $x$ , Covariate, Independent variable, Explanatory Variable, Control Variable

- Important to inspect your data to understand the data and the population.
- Plots of continuous variables
- Tables of categorical variables
- Looking for relationships between (pairs of) important covariates, particularly intervention and other covariates.
- Is there balance between intervention groups?
- Many plots and tables are inspected.
- Most plots tend to show little of interest, but important for data analysts to inspect.

## Sexual Preference by Intervention

	MSM	MSMW	MSW	FEM	Tot
Control ( <i>n</i> )	16	23	3	12	54
Interv ( <i>n</i> )	40	49	3	26	118
Control (%)	29	32	50	32	31%
Interv (%)	71	68	50	68	69%
Overall (%)	33	42	3	22	100%

Column percents in rows 3, 4. Row percents in row 5. MSM: Men who have sex with Men.

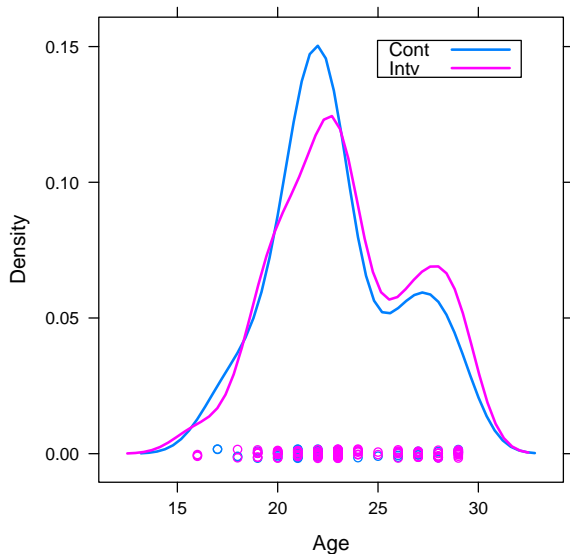
# Density Plots for Continuous Variables

We investigate the distribution of important continuous predictors in the data set.

- A kernel density estimate is an estimate of the density of a continuous variable.
- Think histogram but smooth.
- Where curve is highest, we have the greatest number of subjects, where curve is low, we have the fewest subjects.
- With an important grouping variable, look at the density of the variable within each group.

Next page: Kernel density plot of baseline age in the control and intervention groups.

# Kernel Density Plot: Age by Treatment



Baseline ages are slightly but not significantly ( $p = .1$ ) older in the intervention group.

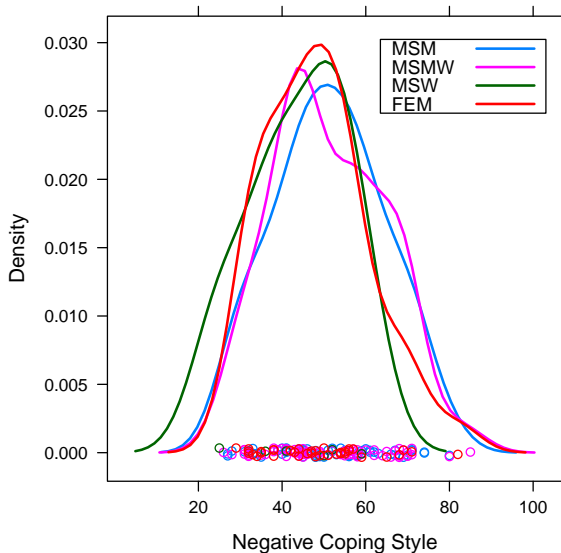
# Kernel Density Plot: Baseline Age and Treatment

- We see two modes, one around age 23, another around age 28.
- A couple of participants under age 18.
- Baseline ages are slightly older in the intervention group.
- But not significant ( $p = .1$ ).
- If we did not already know, we see from the dot plot along the bottom of the graphic that ages are integer valued.

# Plot: Negative Coping Style and Sexual Preference

- Next plot shows baseline negative coping style
- Negative coping style is the sum of several unfortunate coping style scales.
- Depression/withdrawal, self-destructive, passive problem solving, and so on.
- Treated as a continuous variable.
- No apparent serious differences across the different sexual preference groups.
- Negative coping is both an outcome and a predictor.

# Plot: Negative Coping Style and Sexual Preference



Sexual preference groups MSM, MSMW, MSW, FEM. Men who have Sex with Men (MSM) [Men and Women (MSMW), Women (MSW)], Females (FEM).

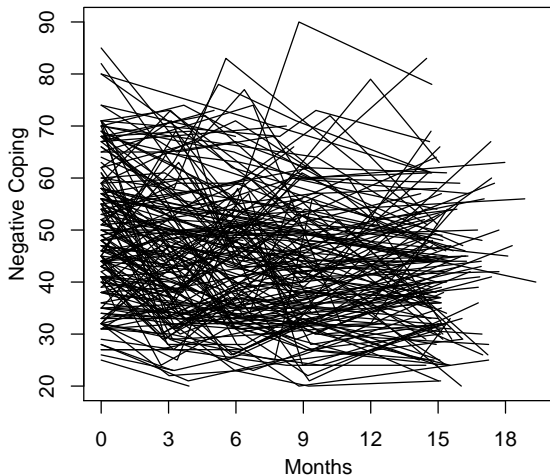
# Profile plots

Profile plots show the response over time within and across subjects.

- Connect the dots. Sequential data points from same subject are connected by line segments.
- For continuous outcomes profile plots are valuable.
- Shows how the response changes over time for each subject.
- For discrete data, less informative.

# Profile Plot: Negative Coping

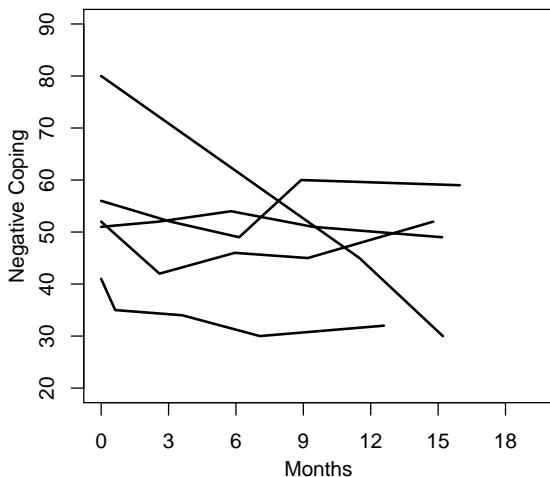
Profile Plot



Plotting all subjects' profiles can be messy. Still useful. Possible downward trend?

# Profile Plot: Negative Coping

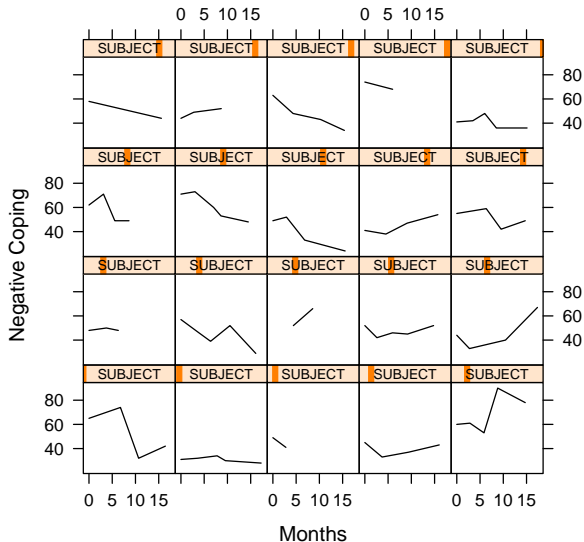
Profile Plot



Plotting a few subjects' profiles makes for a cleaner looking plot. 4 out of 5 subjects appear to be decreasing?

# Profile Plot: Negative Coping: 20 Subjects

## First 20 Subjects



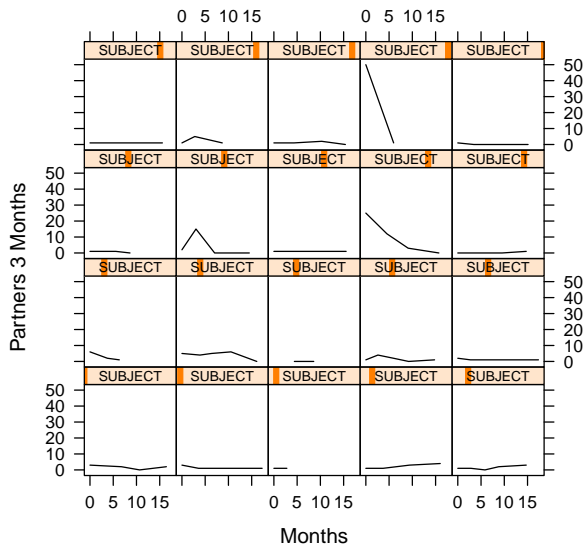
We can look at all individual profiles – inspect all profiles as part of data analysis.

# Profile Plots

- We look at many of these profile plots. Takes 9 pages to look at 180 subjects.
- I have seen more than 1000 subjects usefully plotted on 2 pages.
- Decrease in negative coping appears to be from baseline to first follow up, with possibly no further drops.
- Can compare two groups with color/line types or with separate plots.
- Not as useful for count or binary data.

# Profile Plot: Sex Partners Last 3 Months: 20 Subjects

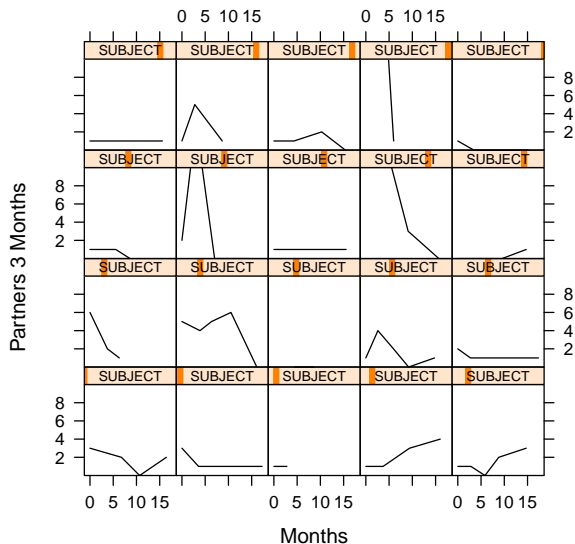
## First 20 Subjects



Sex partners last 3 month (Part3m): Range is enormous, part3m highly skewed. Tick marks range is 0 to 50.

# Profile Plot: Sex Partners Last 3 Months: 20 Subjects

## First 20 Subjects



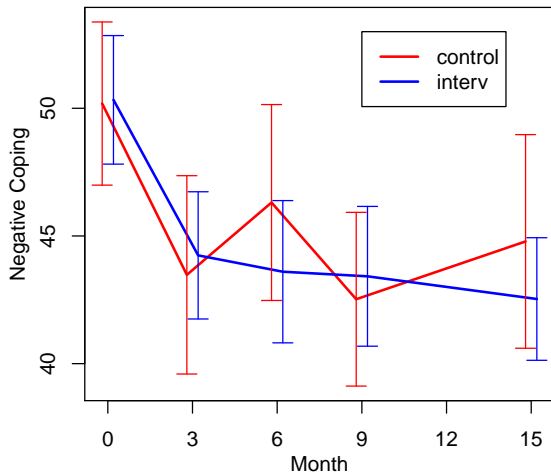
Part3m: These data are very noisy, hard to see patterns. Restricting the range of the y axis allows focus on the many low values. Tick mark range is 0 to 8.

## Plotting means over time. Can add error bars

- Plot means of outcomes as functions of nominal times
- Add error bars to each mean
- On one plot: separate means for each level of predictor
- Median (or other) split for continuous predictors.
- **Look at for all important predictors and all outcomes.**

# Empirical Summary Plot

## Empirical Summary Plot



Negative Coping:  
decrease from baseline  
to month 3. No apparent  
effect of intervention.

## Pluses

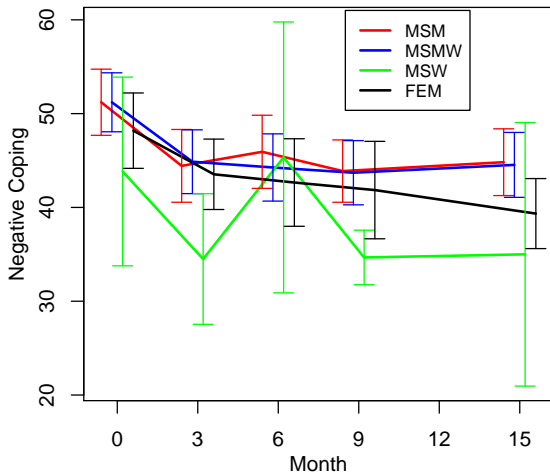
- Gives a *quick* visual estimate of
  - ▶ The population mean over time
  - ▶ The predictor effect and predictor by time interaction
  - ▶ How certain we are about the estimates
- Allows us to focus on modeling data as it is rather than consider all possible models for the population mean as a function of time.
- If two confidence intervals at same time do not overlap, difference is significant.
- Some overlap can still indicate significant differences

## Issues

- For data with nominal times, we take means and sds for all observations at each nominal time.
- For data with randomly observed times, can cluster observations into windows for averaging.
- Not as robust as results from fitting a model.
  - ▶ Plot requires data to be missing completely at random
  - ▶ Model inferences only require data to be missing at random.

# Negative Coping by Sexual Preference

## Empirical Summary Plot



Few MSW (Men who have sex with women).  
Females may trend lower by month 15.

# Empirical Summary Plots

- Can be used with both time-fixed grouping variables and time-varying groups.
- **Meth3m** is methamphetamine use last 3 months: 1 if meth used, 0 if not used in the last three months.
- Meth3m is a *time-varying predictor*.
- **Methever** Ever used methamphetamine is 1 if reported using methamphetamine last 3 months ever at visits 0 through month 15.
- Methever is a *time fixed predictor*.
- Meth3m can be a predictor in some models and an outcome in other models.

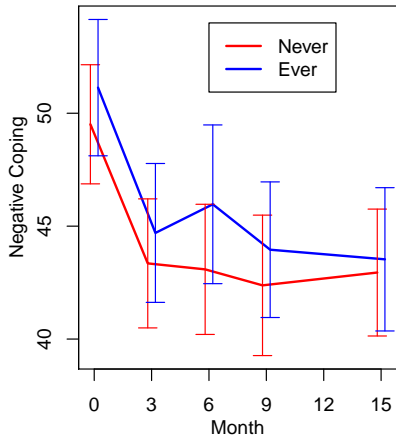
# Empirical Summary Plots

Meth use last 3 months and ever used.

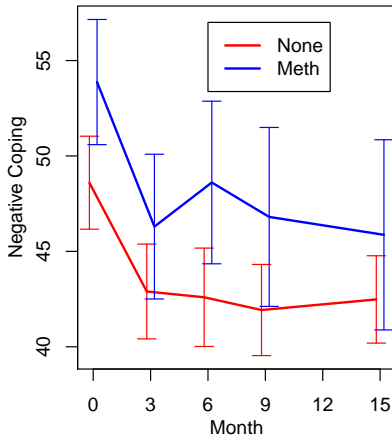
	<i>n</i>		<i>%</i>	
	not	used	not	used
0	117	55	68	32
3	97	47	67	33
6	91	41	69	31
9	108	36	75	25
15	108	30	78	22
21	99	34	74	26
methever	90	85	51	49

# Negative Coping and Meth Usage

## Ever Use Meth

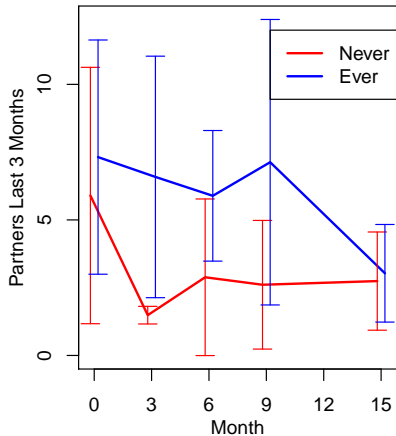


## Meth Last 3 Months

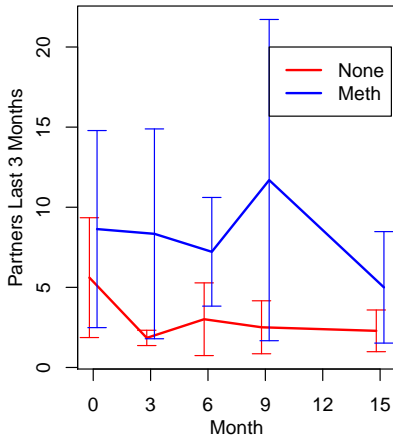


# Partners 3 Month and Meth Usage

## Ever Use Meth



## Meth Last 3 Months



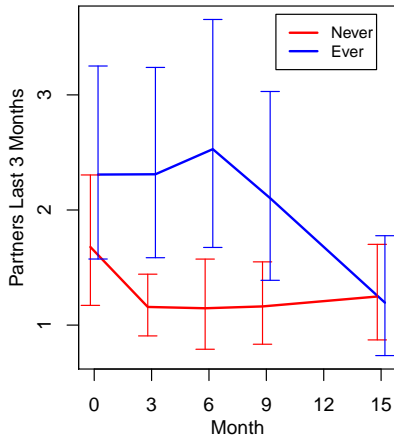
# Skewed variables and a log transformation

Highly skewed variables may be log transformed  $\log Y$  or if zeros are present,  $\log(Y + c)$  – typically choose  $c$  to be the smallest non-zero value in the data. For count data, take  $c = 1$ .

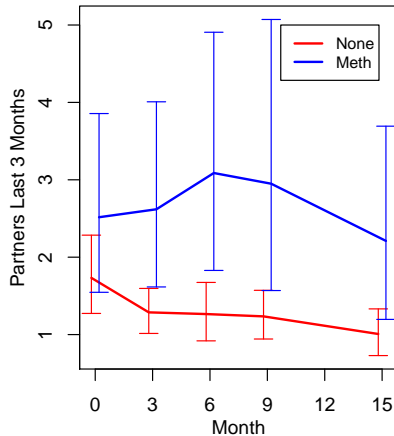
- Transform Part3m to  $\log(\text{Part3m} + 1)$ , calculate empirical summary plot means, and confidence limits.
- Can plot those, or can *backtransform* means and confidence limits by the inverse transformation.
- $\exp(\text{endpoint}) - c$  or  $\exp(\text{mean}) - c$ .
- Means and confidence limits are means and limits on the  $\log(Y + c)$  scale.
- Backtransformation aids in interpretation.

# Part3m Logged and Backtransformed

## Ever Use Meth

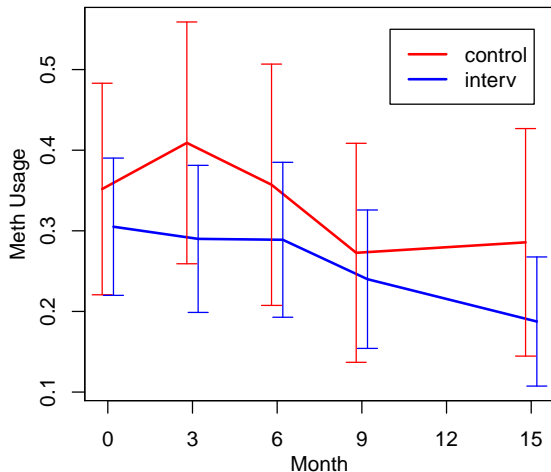


## Meth Last 3 Months



# Empirical Summary Plot: Meth3m by Intervention

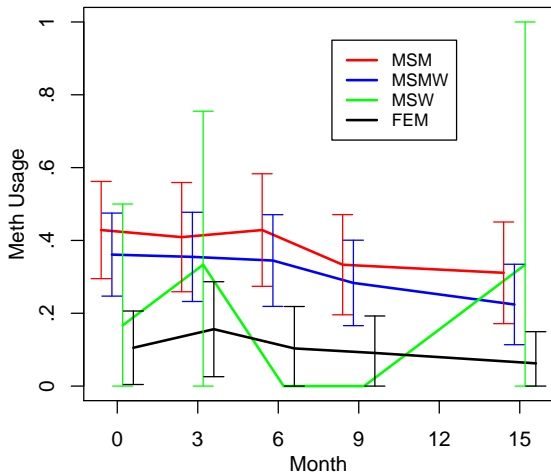
## Empirical Summary Plot



Binary variable:  
Probability of meth use  
last 3 months, by  
intervention group.  
Differences do not  
appear significant.

# Empirical Summary: Meth3m by Sexual Preference

## Empirical Summary Plot



Probability of meth use last 3 months, by sexual preference group. Women have lower use. Possibly general decrease over time.

# An Introduction to Modeling Longitudinal Data

## Session II: Continuous Data and Time Trends

Robert Weiss

Department of Biostatistics  
UCLA School of Public Health  
robweiss@ucla.edu

August 2010

## Part II: Continuous Outcomes

- Notation: Indexes, Outcomes
- Time
- Predictors
- Normal Model, Means and Residuals
- Modeling Population Time Trends
  - ▶ Constant mean over time
  - ▶ Linear time trend
  - ▶ Quadratic time trends
  - ▶ SAS code
  - ▶ Unstructured mean
- More Complex Time Trends
  - ▶ Bent Line Model
  - ▶ Spline Model

## Indexes $i$ and $j$

- Subjects are indexed by  $i$ ,  $i = 1, \dots, n$  subjects
- Little  $j$  indexes observations within a subject from 1 to  $J_i$ .
- $Y_{ij}$  is the  $j$ th observation on subject  $i$ .
- Because of missing data, not all subjects will have the same number of observations.
  - ▶ Subject  $i$  provides us with  $J_i$  observations.
  - ▶ If all subjects have the same number of observations then all  $J_i = J$ .
- The multiple observations per subject are correlated.

# What Time?

Outcome  $Y_{ij}$  is observed on subject  $i$  at time  $t_{ij}$ .

- Observations are ordered by time.
- Observation  $Y_{i(j-1)}$  occurs before  $Y_{ij}$  occurs before  $Y_{i(j+1)}$ .
- That is

$$t_{i(j-1)} < t_{ij} < t_{i(j+1)}$$

# What Exactly is Time?

Recall: Two kinds of time  $t_{ij}$ : Nominal and Actual.

- Baseline is typically the  $j = 1$  observation, and  $t_{i1} = 0$  for both nominal and actual times.
- Nominal time are the intended times of data collection.
- Actual times are the actual time since baseline.
- May use *both* definitions of time in the same statistical model!
- Other definitions of time are possible: time since an important event:
  - ▶ Time since natural disaster
  - ▶ Time since parental death

Time: May be nominal or actual, continuous or categorical.

- Needs more information to be fully specified.
- Nominal time may be treated as a continuous variable or as a categorical variable.
- Nominal time treated as categorical indicates that we will have a separate mean parameter in our model for every possible nominal time value.
- Actual time usually treated as continuous.
- Must identify which time(s) is(are) being talked about in a given circumstance.

Balanced Data means that all subjects have the same number of observations and same nominal times.

- Too stringent – some people will miss visits.
- **Balanced with missing data** – The intended design of the nominal times is balanced, but some people miss visits, causing unbalanced data.
- All nominal times are a subset of a small number of times.

**Random Times:** No consistent day for visits, no consistent number of visits.

- As distinguished from **random times** when people may be observed at any time and virtually nobody is observed at exactly the same times  $t_{ij}$ . For example, doctor visits at an HMO.
- Studies with random times don't have nominal times.
- We will assume balanced with missing data for our studies.

# How Much Data?

Outcome  $Y_{ij}$  from subject  $i$  at time  $t_{ij}$ .

- Index  $i = 1, \dots, n$ .
- Index  $j = 1, \dots, J_i$ .
- Total number of observations is  $N = \sum_{i=1}^n J_i$ .
- Two sample sizes are measures of data set size.
  - ▶ Little  $n$  is the number of people.
  - ▶ Big  $N$  is the number of observations.
- Having two sample sizes distinguishes longitudinal data from linear regression data.

Predictors  $x_{ij}$  for subject  $i$  and observation  $j$ .  
Also called **Fixed Effects**.

- A full set (vector) of capital  $K$  predictors  $x_{ij}$  for subject  $i$ , observation  $j$ .
- Example: intercept, actual\_time, intervention (either 0 or 1) and actual\_time\*intervention interaction.
- $K$  is 4 in this example.

# Specifying Predictors

We may specify predictors in a model as a list.

Outcome = time intervention time\*intervention

The outcome is on the left of the equals sign. Predictors are listed on the right.

- The intercept is assumed part of the model.
- time\*intervention is an interaction.
- Typical of notation for SAS and other statistical programs.
- There are many notations for specifying the predictors in a statistical model.

For continuous data we may write

$$\text{Data} = \text{Mean} + \text{Residual}$$

alternatively

$$Y_{ij} = \mu_{ij} + \epsilon_{ij}$$

- The mean  $\mu_{ij}$  will be a linear function of predictors  $x_{ij}$  and unknown regression parameters  $\alpha$ .
- In the normal longitudinal model, the function of predictors and regression coefficients is called the *linear predictor*.

# Normal Longitudinal Model: The Residual

The *residual*,  $\epsilon_{ij}$ , has three(!) jobs in longitudinal models:

- 1 The residual determines how far the data  $Y_{ij}$  is from its mean.

$$\epsilon_{ij} = Y_{ij} - \mu_{ij}$$

- 2 It determines the distribution, typically assumed normal.

$$\epsilon_{ij} \sim \text{Normal}(0, \text{Variance}_{ij})$$

- 3 And the residual determines the correlations with other observations!

- $\epsilon_{ij}$  and  $\epsilon_{ik}$  are correlated.

Too many jobs; not feasible for count and binary outcomes.

# The Multivariate Normal Model

$$Y_{ij} = \mu_{ij} + \epsilon_{ij}$$
$$\epsilon_i \sim N(\mathbf{0}, \Sigma_i)$$

- The vector of residuals for subject  $i$  is  $\epsilon'_i = (\epsilon_{i1}, \dots, \epsilon_{iJ_i})$ .
- $\epsilon_i$  has a *multivariate normal distribution* with mean zero, and covariance matrix  $\Sigma_i$ .
- Subscript  $i$  on  $\Sigma_i$  is due to missing data. If all subjects had equal number and times of observations, it would be  $\Sigma$ , and  $\Sigma$  would be  $J \times J$ .
- We will discuss models for the covariance matrix later.

# Steps in Modeling a Continuous Outcome

$$Y_{ij} = \mu_{ij} + \epsilon_{ij}$$

The steps are

- Modeling the mean  $\mu_{ij}$ .
  - ▶ Time trend
  - ▶ Predictor effects
  - ▶ {Predictor by time trend interactions}
- Choosing the covariance model among the observations
- Choosing the distribution (ie Normal)

# Constant Mean Over Time

- The simplest model is a constant mean over time.
- Time  $t_{ij}$  is not included in the model.
- If all covariates are time-fixed, the mean for each subject is constant over time.
- Each subject may have a different mean, if the predictors allow for that.
- Suppose sexual preference as only predictor, only 4 distinct groups, one for each sexual preference group.
- Sexual preference is not a significant predictor of negative coping style.

# Linear time trend

We enter continuous time  $t_{ij}$  as a covariate in our analysis. This would model a linear decrease or increase over time.

- In SAS notation,

Outcome = time.

- SAS automatically assumes an intercept unless you tell it otherwise.
- May use nominal or actual time.
- I prefer actual time, particularly when actual times differ substantially from nominal times.

# Quadratic Time Trend

The quadratic time trend model has predictors (i) time and (ii) time-squared  $t_{ij}^2$ .

Outcome = time time\*time.

- Particularly useful if modest deviation from linear trend.
- Higher order polynomials: cubic, possible but problematic to interpret.

# Unstructured Mean

Time  $t_{ij}$  is *nominal time*.

- Months 0, 3, 6, 9, 15 for the CLEAR data.
- Unstructured mean has a separate mean parameter for each *nominal time*.
- Interpret: differences from baseline.
- For example, look at the mean at month 3 minus the mean at baseline.
- There are 4 differences from baseline: months 3, 6, 9 and 15.
- Plotting means is very helpful.

# Interpreting a time trend

Two ways to interpret time trends:

- Plot the time trend.
- Estimate gain over time in study.
- The gain is the difference (mean outcome at study end) minus (mean outcome at baseline).
- For unstructured mean model, there is a gain at each nominal time point.
- Can compare gains for different groups.

## Quadratic Time Trend

```
Title1 'quadratic time';  
proc mixed data=clearcaldar covtest;  
class SUBJECT follow;  
model cpneg = time time*time / s;  
repeated follow / sub=SUBJECT type=cs;  
estimate 'last mean' intercept 1 time 15  
         time*time 225;  
estimate 'first mean' intercept 1 time 0  
         time*time 0;  
estimate 'last - first' time 15 time*time 225;  
run;
```

## A Brief Introduction to SAS Proc Mixed

```
Title1 'quadratic time';  
proc mixed data=clearcaldar covtest;
```

### Title1 Documentation

**proc** The call to Proc Mixed.

- ▶ dataset is called “clearcaldar”.
- ▶ covtest – asks SAS to produce standard errors,  $t$  statistics and  $p$ -values for covariance parameters.

## Quadratic Time Trend

```
class SUBJECT follow;
```

**class** statement – tells SAS that certain variables are categorical

- SUBJECT – subject identification number.
- follow – nominal times: 0, 3, 6, 9 or 15.

# Proc Mixed SAS Code

```
model cpneg = time time*time / s;
```

**model** tells SAS the outcome variable and the predictors.

- **cpneg** – negative coping style is the outcome.
- **time time\*time** – two predictors time and time-squared.
- The intercept is assumed present.
- Forward slash indicates the start of keywords
- keyword **s** asks SAS to print estimates, SEs,  $t$  and  $p$ -values for the coefficient estimates.

```
repeated follow / sub=SUBJECT type=cs;
```

- **repeated** is one of two ways to define the covariance model.
- This particular model is the *compound symmetry* model, abbreviated **cs**.
- `follow` is nominal time, treated as a categorical variable.
- I often start with **cs** or **ar(1)**.
- The variable named **SUBJECT** identifies the subject `sub=SUBJECT` that each observation belongs to.
- More later.

# Proc Mixed SAS Code

```
estimate 'last mean' intercept 1 time 15  
    time*time 225;  
estimate 'first mean' intercept 1 time 0  
    time*time 0;  
estimate 'last - first' time 15 time*time 225;
```

- **estimate** – asks SAS to estimate a *linear combination* of the regression coefficients.
- Each estimate statement has a list of predictors `intercept`, `time`, `time*time` and a value for each predictor 1, 15, 225.
- Three linear combinations estimate population quantities:
  - ▶ Mean at time  $t_{ij} = 15$ .
  - ▶ Mean at time  $t_{ij} = 0$
  - ▶ The difference (outcome at time 15) minus (outcome at time 0).

# Proc Mixed SAS Code

```
estimate 'last mean' intercept 1 time 15  
        time*time 225;  
estimate 'first mean' intercept 1 time 0  
        time*time 0;  
estimate 'last - first' time 15 time*time 225;
```

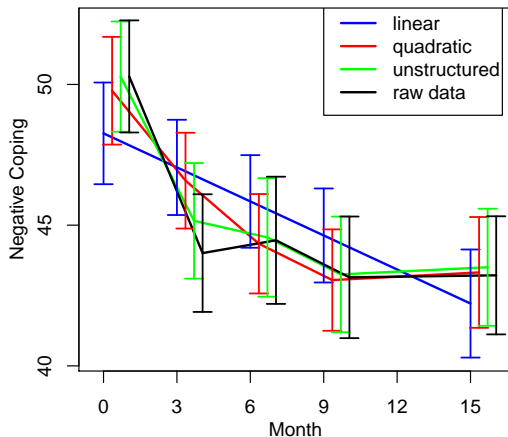
- Third line: Difference (outcome at time 15) minus (outcome at time 0).
- The intercept contributes equally to both estimates and disappears in the difference.
- 15 is the value of time at month 15 minus the value of time at month 0.
- $225 = 15^2 - 0$  is the difference in time squared at  $t = 15$  minus that at  $t = 0$ .
- The third estimate is the **gain from beginning to the end** of the study.

## Graphical Interpretation of Results

- Calculate estimated means and SEs at important time points.
- Draw a fitted value version of the empirical summary plot.
- *Inferential Summary Plot*
- Here we compare fitted values and 95% confidence intervals from the linear, quadratic, and unstructured models to the information from the empirical summary plot.
- Make predictions for several sets of predictor values for comparison and inferential purposes.

# Comparing Fits to Each Other and Raw Data

## Summary Plot



Linear, quadratic, unstructured fits, and empirical summary plot.

## Three fits, and empirical summary plot

- Linear model fits the data summary worst
- Quadratic may fit passably ok, but unstructured is slightly better fit.
- Unstructured is closest but not identical to empirical data.
- Unstructured fills in missing data according to the model.
- Empirical summary plot ignores the missing data entirely.
- Model estimates trade off covariance assumptions for assumptions about the missing data.

## Linear Time Trend

- Treatment groups should be similar at baseline.
- Interest lies in continuous time by treatment interaction.
- One intercept and slope for each treatment group.

Outcome = Time Treatment **Time\*Treatment**

- Intercept is assumed present.
- Include main effect for Treatment – to check if groups are similar at baseline.

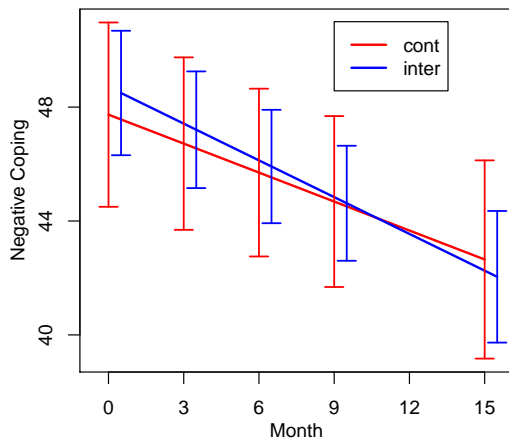
# Negative Coping: Time By Treatment SAS Output

Effect	Estimate	SE	t Value	Pr >  t
time	-.43	.07	-6.2	<.0001
time*intv	.091	.13	0.72	0.47

- $-.43$  is the slope in the treatment group.
- $.091$  is the difference in slopes (control minus treatment).
- Not significant,  $p = .47$ .
- SAS output, rounded
- Irrelevant information, words, symbols are omitted.
- Can also plot the fitted lines for a visual summary.

# CLEAR treatment: Fitted Lines

CLEAR: Linear Fit Summary Plot



Lines are not significantly different from each other, neither slopes, nor intercepts nor means at given times.

## Unstructured Mean

- Same model for time in each group.
- Interest is in treatment\*time interaction
- $F$ -statistic for testing the intervention by time interaction has  $J - 1$  numerator degrees freedom for  $J$  time points.
- Additional interest:
  - ▶ Difference in group response at each follow-up time minus difference at baseline.

# Type 3 Tests of Fixed Effects

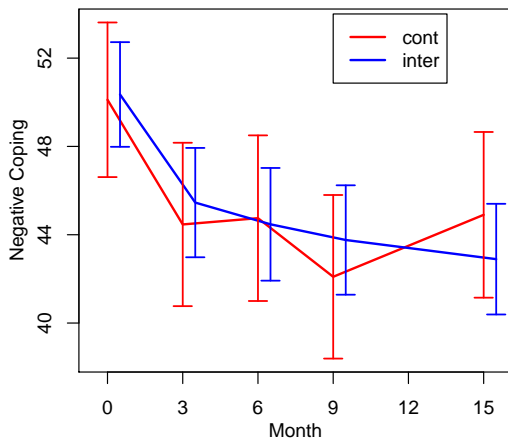
## Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
intv	1	173	0.00	0.9451
follow	4	547	16.63	<.0001
intv*follow	4	547	0.85	0.4968

- Only interested in the interaction – not significant.
- This is raw SAS output.
- This output needs further formatting.

# CLEAR treatment: Fitted Lines

## CLEAR: Unstructured Fit Summary Plot



CIs at each nominal time point overlap.

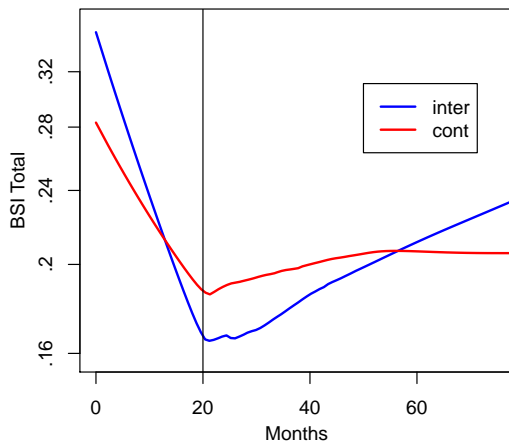
## TALC: Teens and Adults Learning to Communicate

- Illustrates a linear spline.
- Enrolled low-income HIV+ parents and their adolescent children.
- Intervention on coping with disease and communicating between parents and adolescents.
- Over 5 years of data collection.
- Every 3 months for years 1-2
- Every 6 months for years 3-5 and beyond
- Nonlinear time trend.

- Nominal times of data collection quite disparate from actual times.
- Define new nominal times with 3-month rounded times.
- Our outcome: Brief Symptom Inventory (BSI).
- 53 item scale.
- Skewed outcome: use  $\log_2(y + 1/53)$ .
- Time trend in intervention and control groups is non-linear.

# TALC: Smoothed Lowess Curves

TALC: Smoothed Lowess Curves



Smoothed averages of log BSI over time by intervention. A *lowess* curve. Uses actual times  $t_{ij}$  of observation.

# TALC Lowess Curves

- Separate smooth fitted curves for control and treatment.
- Baseline is time  $t = 0$ .
- Uses actual times of observations post baseline, not nominal times.
- We see a linear decrease over the first 20 months.
- Then the trend changes direction and is more-or-less linearly increasing afterwards.

# Lowess Curve

A *lowess curve* fits a smooth curve to the outcome as a function of time (or other  $X$  variable).

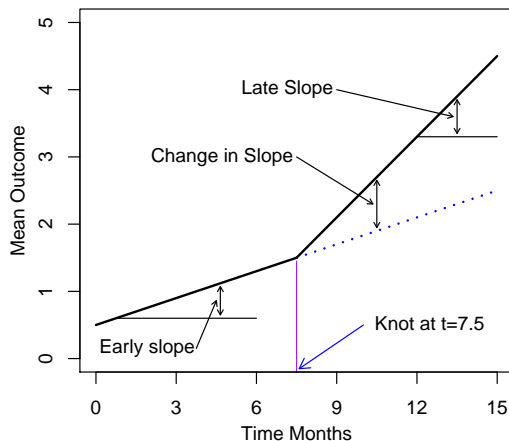
- For getting a general idea of the time trend.
- Not for inferential purposes – does not account for correlations among observations.
- *lowess* – LOcally Weighted Smoothing of Scatterplots.
- At each value  $X^*$  along the  $X$  axis: Estimate  $\hat{Y}^*$ -hat.
- Fit a curve to  $Y$  as a function of  $X$  using only data points with  $X$  values near  $X^*$ .
- Weight points higher if they are closer to  $X^*$ .
- There is a different fit  $\hat{Y}^*$  for every value of  $X^*$ .
- Connect the dots between the points  $(X^*, \hat{Y}^*)$  to get the lowess curve.

# Spline: Bent Line

- Bent line: Piece-wise linear time trend.
- Time trend: linear on intervals.
- Slope changes at specific known time point(s) called *knot*(s).
- May have several knots.
- Slope is different before and after each *knot*.
- Knot  $t_*$  : a time point  $t_*$  where the slope changes.

# Bent line Example

## Bent Line Example



Illustrative plot. One knot at  $t = 7.5$ , a mildly increasing slope before  $t = 7.5$ , and a steeper slope after  $t = 7.5$ .

# Spline: Bent Line With One Knot

Typical parameterization has three parameters:

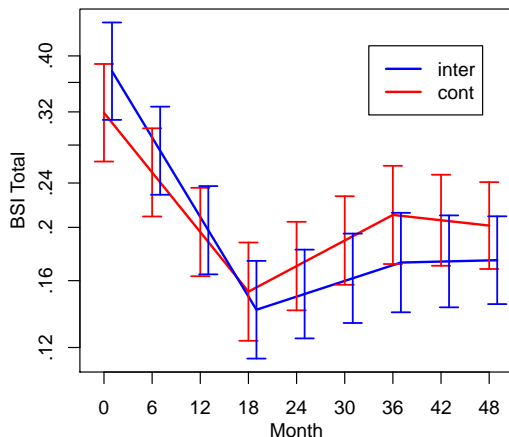
- Intercept
- Early slope
- Change in slope
- Later slope is equal to Early slope plus Change in slope.
- There is one change in slope for each knot.

After much exploration, we decided on a bent line model with 2 knots.

- Final Model: Knots at 18 months and 36 months.
- Based on exploring 1 or 2 knots at times around 20 months and 40 months
- An independent group identified the same times for knots!

# Bent line Example

TALC: Spline Fit Summary Plot



Fitted values and confidence intervals for intervention and control.

*Splines* are handy for modeling non-linear time trends.

- Particularly handy for trends over long time periods.
- Not so handy for 4 nominal time points.
- Bent line model is a *linear spline*.
- A linear spline is linear in between knots.
- Number of parameters: 2 plus 1 per knot.
- Can have quadratic splines or CUBIC splines.
  - ▶ A *Cubic Spline* is a cubic function of time in between knots.
  - ▶ The cubic function changes at each knot.
  - ▶ Number of parameters: 4 plus 1 per knot.

# General Splines

- The most popular splines: linear and cubic.
- Choice of knots:
  - ▶ Pick knots where the mean trend changes “a lot”.
  - ▶ Or, pick knots equally spaced.
  - ▶ May test significance of knots with  $t$ -test: coefficient estimate divided by its standard error.
  - ▶ Delete if knot not significant.

# Adjusting for Covariates

- In observational studies, subjects in different groups are not exchangeable.
- Add covariates important as predictors of the primary outcome.
- Include covariates significantly different between the groups of interest.

# Predictor Interactions

- In observational studies, primary interest is often in
  - ▶ Main effects of predictors,
  - ▶ Time trend.
- In randomized studies, may be interested in covariate by treatment interactions: Is treatment effect different in different groups?
- Difficult:
  - ▶ Many studies are not powered for interactions, only for main effects.
  - ▶ But still can check for interactions.
  - ▶ With many interactions of interest, multiple comparisons can be a problem.

# Baseline Value of Outcome Measure is an Outcome!

## The Baseline Value of the Outcome Measure is an Outcome.

- DO NOT include baseline outcome value as a predictor.
- DO NOT subtract baseline value from all other responses.
- Rather, include the baseline as part of the outcome, model it.
- Adjust at the inference stage.
- Adjust for baseline by subtracting fitted value at (say)  $t_{\max}$  minus fitted value at baseline.
- Correct longitudinal modeling of baseline and follow-up values correctly models the gain at follow-up from baseline.
- Getting the longitudinal model correct: Correct mean model and covariance model.

- A *kitchen sink* model, with a number of interesting predictors.

## SAS Proc Mixed Code

```
Title1 'Kitchen Sink';  
proc mixed data=clearcaldar covtest;  
class SUBJECT sexpref2 follow;  
model cpneg = follow intv intv*follow part3m trade3m IDU meth3m  
          sexpref2 haart3m/ s;  
repeated follow / sub=SUBJECT type=arma(1,1);  
run;
```

## Predictors in this model

- follow** – Nominal time. SAS includes 5 indicators for each nominal time.
- intv** – Intervention 0-1 treated as continuous because it is not included in the class statement.
- intv\*follow** – Time by Intervention interaction. Includes 5 indicators of time when intv=1.
- part3m** – Number of partners last 3 months.
- trade3m** – Traded sex (for drugs, money, room, board) in the last 3 months.
- IDU** – Ever used intravenous drug (baseline) .
- meth3m** – Meth use last 3 months.
- sexpref2** – Sexual preference in 3 groups: MSM, MSMW and FEM/MSW combined.
- haart3m** – On HAART last 3 months.

**repeated follow / sub=SUBJECT type=arma(1,1);**

Type defines the covariance model. Arma(1,1) was better than CS model and is used here.

# SAS Proc Mixed Output

## Complete Listing

Kitchen Sink

21:17 Sunday, July 18, 2010 22

The Mixed Procedure

Model Information

Data Set	WORK.CLEARCALDAR
Dependent Variable	cpneg
Covariance Structure	Autoregressive Moving Average
Subject Effect	SUBJECT
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

### Basic information about the model fit.

REML – one of two major estimation methods for continuous data. Other methods are available and widely used for count and binary data.

## Class Level Information

Class	Levels	Values
SUBJECT	175	11 12 13 14 15 16 17 18 19 20 22 23 25 28 29 30 31 32 33 34 36 37 38 39 40 41 43 46 47 48
<b>17 Lines deleted</b>		...
		592 596 603 604 606 607 610 613 617
sexpref2	3	1 2 3
follow	5	0 3 6 9 15

**Lists all levels of categorical variables. Helpful for basic sanity checking when results come out badly or model does not run or for other simple problems. May be omitted when SAS code writing and model fitting for this data set has become routine.**

## Dimensions

Covariance Parameters	3
Columns in X	20
Columns in Z	0
Subjects	175
Max Obs Per Subject	5

**Summary numerical information about the model and data set. Number of subjects n=175, maximum number of observations per subject is 5.**

Kitchen Sink

21:17 Sunday, July 18, 2010 23

## The Mixed Procedure

## Number of Observations

Number of Observations Read	730
Number of Observations Used	730
Number of Observations Not Used	0

Summary information about the data set. Observations used and not used. If there is missing data in predictors, the first two rows will not be equal. The second and third rows add up to the first row.

#### Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	5704.39279230	
1	4	5416.88922317	.
2	1	5403.33556801	0.00179178
3	1	5398.97874087	0.00026113
4	1	5398.39273615	0.00000718
5	1	5398.37778036	0.00000001

**Iteration history shows how hard SAS is working to produce answers for this model and data set. A large number of iterations may imply a problem with the data or model or algorithm. Possible solutions include using a different or simpler covariance model, and removing predictors that contribute to colinearity. Longitudinal models,**

**even appropriate models, are difficult to fit to longitudinal data. Linear, count or logistic regression models fit to independent data are much easier to fit.**

Convergence criteria met.

**An important piece of information. If this statement does not occur, you cannot trust the results.**

**Also: Check the *log file* for the statement:**

NOTE: Convergence criteria met.

**This is also important, and if this statement does not occur, you cannot trust the results. Other errors may be printed in the log file.**

## Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr >  Z
Rho	SUBJECT	0.9154	0.02857	32.04	<.0001
Gamma	SUBJECT	0.6436	0.03359	19.16	<.0001
Residual		159.56	12.4552	12.81	<.0001

**Covariance models are more complicated in longitudinal data analysis as compared to linear regression. SAS prints the parameter names estimates. Because we have the keyword 'covtest' in the model statement, SAS also prints the standard error and a z-value (estimate divided by standard error) and p-value.**

**We will discuss covariance models in the next section of the course.**

## Fit Statistics

-2 Res Log Likelihood	5398.4
AIC (smaller is better)	5404.4
AICC (smaller is better)	5404.4
BIC (smaller is better)	5413.9

**Fit statistics are particularly useful for comparing non-nested models. We will use them for comparing different covariance models. SAS prints these in “smaller is better” form.**

## Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
2	306.02	<.0001

**This test is testing the null hypothesis of independence of observations within subject. Would it have been equally sensible to use the independence covariance model? The answer is usually an emphatic “no”!**

## Solution for Fixed Effects

Effect	sexpref2	follow	Estimate	Standard Error	DF	t Value	Pr >  t
Intercept			42.0836	2.3394	170	17.99	<.0001
follow	0		4.5259	1.8831	543	2.40	0.0166
follow	3		-0.6446	1.8858	543	-0.34	0.7326
follow	6		-0.7195	1.8114	543	-0.40	0.6914
follow	9		-2.9412	1.6765	543	-1.75	0.0799
follow	15		0	.	.	.	.
intv			-1.6911	2.2279	170	-0.76	0.4489
intv*follow	0		2.2177	2.2474	543	0.99	0.3242
intv*follow	3		2.6512	2.2524	543	1.18	0.2397
intv*follow	6		1.7689	2.1746	543	0.81	0.4163
intv*follow	9		3.0821	2.0130	543	1.53	0.1263
intv*follow	15		0	.	.	.	.
part3m			0.05819	0.02557	543	2.28	0.0233
trade3m			2.0898	1.2008	543	1.74	0.0824
IDU			-1.0463	1.8680	170	-0.56	0.5762
meth3m			2.6167	1.0485	543	2.50	0.0129
sexpref2	1		3.0934	2.1405	170	1.45	0.1502
sexpref2	2		2.6752	2.0341	170	1.32	0.1902
sexpref2	3		0	.	.	.	.
haart3m			-0.1302	0.8969	543	-0.15	0.8846

**Table of  
regression coefficient estimates,  
standard errors,  
degrees of freedom,  
t-statistic,  
p-value**

**The number of predictors can be enormous in longitudinal models.**

***Class or categorical variables* follow and sexpref2 are identified in the class statement. Class variables get one indicator variable for every level of the variable. If we have K levels for the predictor, we only need K-1 indicator variables in the model. Thus SAS constructs one extra indicator variable for each class variable.**

**The extra indicator is then forced to be equal to zero. This is called *aliasing*. See the row:**

follow                    15                    0

**It can cause awkwardness in interpretation, but is a rigorous approach to specifying the**

**predictors. It is helpful in constructing estimate statements.**

**Intv (intervention) was not put in the class statement, it is still a 0-1 binary predictor. SAS won't construct two indicator variables for Intv.**

**Inspection shows that part3m, meth3m and with further exploration that follow – nominal time are all significantly associated with negative coping style.**

## Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
follow	4	543	5.24	0.0004
intv	1	170	0.02	0.8838
intv*follow	4	543	0.65	<b>0.6259</b>
part3m	1	543	5.18	0.0233
trade3m	1	543	3.03	0.0824
IDU	1	170	0.31	0.5762
meth3m	1	543	6.23	0.0129
sexpref2	2	170	1.20	<b>0.3045</b>
haart3m	1	543	0.02	0.8846

**The “type 3 tests of fixed effects” are particularly used to test whether variables in the class statement are significant predictors of the outcome.**

**Sexpref2 is not a significant predictor, p=.3.**

**With an interaction in the model, we do not test the main effects. We see that intervention\*follow is not significant, p=.6.**

**Depending on interests, one might stay with this model, or one might drop the non-significant predictors.**

**General comments: when you are the only consumer of a statistical analysis, you will inspect the SAS output directly. When producing results for a research group, it is important to summarize the results in compact form to allow the research group to understand the results quickly and not waste time plowing through pages of output.**

Effect	Est	SE	t	p
Intercept	42.1	2.3		
follow 0	4.5	1.9	2.4	.02
follow 3	-0.6	1.9	-0.3	.73
follow 6	-0.7	1.8	-0.4	.69
follow 9	-2.9	1.7	-1.8	.08
intv	-1.7	2.2	-0.8	.45
intv*follow 0	2.2	2.2	1.0	.32
intv*follow 3	2.7	2.3	1.2	.24
intv*follow 6	1.8	2.2	0.8	.42
intv*follow 9	3.1	2.0	1.5	.13
<b>part3m</b>	<b>0.1</b>	<b>0.0</b>	<b>2.3</b>	<b>.02</b>
trade3m	2.1	1.2	1.7	.08
IDU	-1.0	1.9	-0.6	.58
<b>meth</b>	<b>2.6</b>	<b>1.0</b>	<b>2.5</b>	<b>.01</b>
sexpref2 1	3.1	2.1	1.5	.15
sexpref2 2	2.7	2.0	1.3	.19
haart3m	-0.1	0.9	-0.2	.88

	Num	Den		
Effect	DF	DF	F	p
intv*follow	4	543	0.7	.63
sexpref2	2	170	1.2	.30

**Rounding, bolding of important results, only copying important results, elimination of unneeded text.**

**For example:**

Effect	Est	SE	t	p
<b>part3m</b>	<b>.06</b>	<b>.03</b>	<b>2.3</b>	<b>.02</b>
trade3m	2.1	1.2	1.7	.08
IDU	-1.0	1.9	-0.6	.58
<b>meth</b>	<b>2.6</b>	<b>1.0</b>	<b>2.5</b>	<b>.01</b>
haart3m	-.1	.9	-0.2	.88

# An Introduction to Modeling Longitudinal Data

## Session III: Covariance Models for Continuous Outcomes

Robert Weiss

Department of Biostatistics  
UCLA School of Public Health  
robweiss@ucla.edu

August 2010

### Outline

- General remarks on covariance models
- Basic covariance models
- More advanced covariance models
- Choosing among covariance models
- CLEAR data covariance model selection.

The covariance model determines possible variance and correlations between observations.

- Needs to match data!
- Correct choice of covariance model produces accurate
  - ▶ Fixed Effects standard errors,
  - ▶ Fixed Effects confidence intervals
  - ▶ Hypothesis test p-values
  - ▶ Predictions/imputations of data
- Inappropriate choice of covariance model will cause mistaken inferences.

## Game Plan

- Take a guess at the covariance model.
- We have seen `type=CS` and `type=ARMA(1,1)` in our examples.
- Select a working set of fixed effects
  - ▶ Important: get the time trend right.
  - ▶ Include VERY important predictors
  - ▶ For example: gender for psychometric outcomes like depression or anxiety.
  - ▶ Including or excluding insignificant, or barely significant predictors usually won't affect covariance model choice.
- Choose a best covariance model from among available options.

Goal is to choose a *parsimonious* covariance model that fits the data.

- Complex models  $\rightarrow$  more parameters  $\rightarrow$  better fit to data.
- Simpler covariance model  $\rightarrow$  fewer parameters  $\rightarrow$  more assumptions.
- Too many parameters  $\rightarrow$  increased variance in parameter estimates.
- Too few parameters  $\rightarrow$  bias in covariance model.
- Sufficiently complex for data set at hand, but not more complex than needed.
- Parsimonious in other words.

# Equally Spaced Times

Assume equally spaced times of observations for this discussion.

- Times (for now) assumed to be at nominal times.
- Times  $t_{ij} = j - 1$ , for  $j = 1, \dots, J$ . Equally spaced.
- Define: **Lag**: Time span between two times.
- **Lag 1**: Span between observations 1 time unit apart.
- Lag 2: Span between observations 2 time units apart.
- Observations at 0, 3, 6, 9 and 12 months, observations correspond to  $j = 1, 2, 3, 4, 5$  respectively. Observations 3 months apart are lag 1 apart; observations 6 months apart are lag 2 apart.

## Number of covariance parameters, $q$

- $J$  is number of nominal time points.
- Some covariance models have a large number  $q$  of unknown parameters.
- *Unstructured*: UN has  $J * (J + 1)/2$  parameters.
- UN is the most general covariance model.
- Some models have a small number  $q$  of unknown parameters.
- *Compound Symmetry* (CS), *Autoregressive* (AR(1)) have  $q = 2$ .
- Few parameters – easier for SAS to fit model, might induce bias.
- Many parameters –harder to fit model.

# Covariance Model Characteristics

Covariance model defines variances of observations and correlations between observations.

- Always: correlation  $\rho_{jk} = \text{Corr}(Y_{ij}, Y_{ik})$  a function of times  $t_{ij}$  and  $t_{ik}$ .
- Often: correlation  $\rho_{jk}$  only a function of the absolute difference in time  $|t_{ij} - t_{jk}|$ .
- With equal spacing  $t_{ij} = j - 1$ , then  $|t_{ij} - t_{jk}| = |j - k|$ .
- Two families:
  - ▶ **Autoregressive Family** Correlations higher between observations closer in time and lower between observations farther apart in time. (AR(1), ARMA(1,1))
  - ▶ **Random Effects Family:** Also **CS Family** Covariance does not necessarily fall off with increases in time. (CS, RI, RIAS)

# Autoregressive AR(1) Model

AR(1): correlation decreases as time between observations increases.

- Two parameters  $q = 2$ , variance  $\sigma^2$ , correlation parameter  $\rho$ .
- $\rho_{jk} = \rho^{|t_{ij} - t_{ik}|}$ , a function of absolute differences in times  $|t_{ij} - t_{ik}|$ .
- Good for equally spaced data.
- Extension for unequally spaced, unbalanced data available in SAS.
- Constant variance  $\sigma^2$  at all times.

```
class time;  
repeated time / type=ar(1) subject=id r;
```

Key word `r` prints the estimated correlation matrix. For balanced data, this can be compared to the raw data correlation matrix.

# Compound Symmetry

CS: correlation  $\rho$  constant as time between observations increases.

- Two parameters  $q = 2$ , variance  $\sigma^2$ , correlation parameter  $\rho$ .
- $\rho_{jk} = \rho$  a constant function of time difference  $|t_{ij} - t_{ik}|, j \neq k$ .
- Unequally spaced data ok.
- Unbalanced data ok (therefore balanced data also ok).
- Constant variance  $\sigma^2$  for all times.

```
class time;  
repeated time / type=cs subject=id r;
```

# Unstructured Covariance Matrix

Unstructured is the most general covariance model.

- Balanced data (with missing data ok).
- Unequal spacing ok.
- Number of parameters  $q = J(J + 1)/2$  parameters  $\sigma_j^2$  and  $\rho_{jk}$ .
- Nonconstant variance  $\sigma_j^2$ .
- More general than AR(1) and CS.

```
class time;  
repeated time / type=un subject=id r;
```

# Autoregressive Moving Average ARMA(1,1)

## ARMA(1,1) more general than AR(1), CS.

- Balanced with equal time spacings.
- $q = 3$  parameters: Variance  $\sigma^2$  two correlation parameters  $\rho, \gamma$ .
- That is, AR(1) and CS are special cases of ARMA(1,1).
- Less general than UN, ie nested inside UN.
- $\rho_{jk} = \gamma$  for lag 1.
- $\rho_{jk} = \gamma * \rho$  for lag 2.
- $\rho_{jk} = \gamma * \rho^2$  for lag 3.
- $\rho_{jk} = \gamma * \rho^{|j-k|-1}$

class time;

repeated time / **type=arma(1,1)** subject=id r;

## A laundry list of additional models in two families.

- Other options in the repeated statement include: ANTE(1), ARH, CSH, UN, SP(EXP), TOEP, TOEPH, ARMA(1,1), FA1(1), FA(1), FA1(2), FA(2).
- Some models generalize AR(1): ANTE(1), ARH, TOEP, TOEPH, SP(EXP).
- Some models generalize CS: CSH, FA1(1), FA1(2), FA(1), FA(2).
- ARMA(1,1), UN generalize both.

Some models add **non-constant variance** to a **constant variance** model. Observations at time  $j$  now have variance  $\sigma_j^2$ .

- Adds  $J - 1$  variance parameters to original model.
- ARH(1) to AR(1).  $q = J + 1$  instead of  $q = 2$ .
- CSH to CS.  $q = J + 1$  instead of  $q = 2$ .
- Toeplitz: TOEPH to TOEP.  $q = 2J - 1$  instead of  $q = J$ .

## AR(1) Family Models

- *Antedependence* ANTE(1) generalizes AR(1). Correlation decreases with increasing lag.  $q = 2J - 1$ .
- ANTE(1) has nonconstant variance.
- ANTE(1): lag 1 correlation at time  $j$  to  $j + 1$  is  $\rho_j$ .
- $J$  variance parameters,  $J - 1$  correlation parameters.
- *Toeplitz* TOEP generalizes AR(1). Also known as *banded*. Correlation changes (usually decreases) with increasing lag, but not in a fixed pattern.  $q = J$  parameters
- Lag  $s = |k - j|$  has correlation  $\rho_s$ .
- TOEP: equally spaced data, ANTE: balanced; missing ok.

## Random Effects Models Family (or CS Family).

- Random effects models can be used for numerous data sets other than continuous longitudinal data.
- For example: discrete longitudinal data, hierarchical data sets, aka nested data, etc.
- For longitudinal data, random effects models (REMs) compete with patterned covariance models.
- Patterned covariance models: repeated statement.
- Random effects models: random statement.

Each subject has own time trend, a **subject-specific** time trend.

- Model subject's underlying time trend by a simple polynomial function added to the population mean.
- Examples: flat line, line (ie not flat), quadratic, linear spline
- Independent *Residual error* keeps data from exactly following the underlying time trend.
- Independent residual variance is  $\sigma^2$ .
- The subject-specific time trend can not be determined exactly; it can be estimated (poorly) from the data.
- Random intercept (RI)
- Random intercept and slope (RIAS)

# Random Intercept Model

$$Y_{ij} = \mu_{ij} + \beta_i + \delta_{ij}$$

- $\mu_i$  is the population mean
- $\beta_i$  is a *subject specific effect*, the *random intercept*.
- If  $\beta_i$  is positive, all of subject  $i$ 's observations tend to be above the population mean.
- If  $\beta_i$  is negative all the  $Y_{ij}$  tend to be below the population mean.
- $\delta_{ij}$  is independent error.
- We put a distribution on  $\beta_i$

$$\beta_i \sim N(0, D)$$

## Why can we put a distribution on the random effect $\beta_i$ ?

- Subjects form a population.
- Characteristics about the subject then form a population of characteristics.
- The random effect  $\beta_i$  in this model is a characteristic of subject  $i$ .
- And therefore, the  $\beta_i$  also form a population.
- Why normal? Convenience, often not thought to matter much.
- A thought which I would tend to echo.
- With exceptions.

Random effects models split the residual job into two parts.

- We had:  $\text{Data} = \text{Mean} + \text{Residual}$
- In the random effects model, we split the residual  $\epsilon_{ij}$  into
- $\text{Residual} = \text{random effects terms} + \text{independent\_residual}$
- The random effects  $\beta_i$  carry the correlation between observations.
- The independent\_residual  $\delta_{ij}$  is error uncorrelated with other observations.
- Random effects models can be used for discrete data models.

## SAS Code for RIAS: random intercept and slope.

```
Title1 'RIAS: Small Sink';  
proc mixed data=clearcaldar covtest;  
class SUBJECT follow;  
model cpneg = follow part3m meth3m/ s;  
random intercept time/ sub=SUBJECT type=UN;  
run;
```

Always use **type=UN** keyword for longitudinal data RIAS random effects model.

## How Many Random Effects?

- Random intercept (RI) model is identical to CS covariance model.
- Random intercept and slope (RIAS) generalizes RI.
- RI has 1 random effect
- RIAS has 2 random effects
- RIASAQ (RIAS + random quadratic) has 3 random effects.

## Compound Symmetry Family: Factor Models

- Factor analysis models **FA1(1)**, **FA1(2)**, **FA(1)**, **FA(2)** generalize random effects models.
- FA and FA1 models are semi-parametric random effects models.
- Factors = Random Effects.
- **FA1(1)** and **FA(1)** have 1 factor, and generalize random effects models with 1 random effect.
- **FA1(2)** and **FA(2)** have 2 factors, and generalize random effects models with 2 random effects.
- FA1(s) has constant  $\sigma^2$  residual variance.
- FA(s) has non-constant residual variance and s factors.

## Several models are feasible when time is continuous.

- `type=SP (EXP) (time)` is the continuous time version of AR(1).
- Replace `time` with the name of the continuous time variable in your data set.
- Random effects (RE) models
- Independence (IND) model
- Can create more models by combining two covariance models:
  - ▶ SP(EXP)(time) and random effects.
  - ▶ SP(EXP)(time) and independent error.
- Both are popular models.

# Goodness of Fit Statistics

We use log likelihood and penalized log likelihood to select a covariance model.

SAS fit statistics output:

```
                Fit Statistics
-2 Res Log Likelihood           543.0
AIC (smaller is better)        547.0
AICC (smaller is better)       547.0
BIC (smaller is better)        551.2
```

We will use the `-2 Res Log Likelihood` and `AIC` for covariance model selection.

`-2 Res Log Likelihood` will often be called by “-2 log likelihood”, `-2ll`, or similar name.

# Selecting a Covariance Model

- Nested models  $M_1$  and  $M_2$  (models m-one and m-two) can be formally tested using a likelihood ratio test.
- Test statistic is the difference in minus 2 log likelihoods.
- An approximate chi-square test with  $df = q_1 - q_2$ , where model 2 has fewer parameters.
- Most covariance models are not nested in each other: can not be compared using a hypothesis test.
- Information criteria AIC (Akaike's Information Criterion) and BIC (Bayesian Information Criterion) may be used to select the best model.
- SAS default prints AIC and BIC in *smaller is better* form.
- Select the covariance model with the smallest AIC.

# Selecting a Covariance Model

- Parameterized covariance models can be tested against the independence model. (No repeated statement at all!)
- Test is often asymptotically chi-square. With many exceptions.
- Degrees of freedom is  $q - 1$ .
- Parameterized covariance models can also be compared to UN, unstructured model. If you can fit it.
- Nominal degrees of freedom is  $J(J + 1)/2 - q$ .
- Test statistic is the difference in  $-2\ln$  ( $-2$  res log likelihoods).

## Information Criteria

- AIC and BIC are information criteria.
- Both are log likelihood with a penalty for the number of parameters in the model.
- AIC is  $-2 \log \text{likelihood} + 2 * q$ , twice the number of covariance parameters.
- BIC uses  $(\log \# \text{ observations})/2$  instead of  $q$ .
- Pick model with smallest AIC.
- Use REML log likelihood (method=REML, which is the default).
- To compare two covariance models, must have same predictors for the mean.

## Overview

- SAS has a large number of choices for the covariance model.
- Best choice determined by what fits the data best while still being **parsimonious**.
- Differing numbers  $q$  of covariance parameters for different models.
- Some models suitable for:
  - ▶ Equally spaced data;
  - ▶ Both equally and unequally spaced data;
  - ▶ Balanced not necessarily equally spaced data;
  - ▶ Random times and unbalanced data.

# Documentation

- Documentation for Proc Mixed is available on line at <http://www.ats.ucla.edu/stat/sas/sasdoc.htm>.
- Click on `STAT` under `Version 9.2`,
- Click on `Procedures` in left hand column
- Scroll down for `The Mixed Procedure` for documentation.
- For extensive lists of covariance models see tables 56.13, 56.14 and 56.15 after clicking on the `repeated` statement.
- I frequently consult these tables.

- Computer labs for learning Proc Mixed available on line from <http://rem.ph.ucla.edu/rob/mld/index.html> .
- Click on 9 Computer labs for ...
- UCLA's Academic Technology Services has posted code for most of the examples in my book:  
[http://www.ats.ucla.edu/stat/sas/examples/mld\\_weiss/default.htm](http://www.ats.ucla.edu/stat/sas/examples/mld_weiss/default.htm)
- My book has the most extensive discussion of all these models available.

## What is the Best Covariance Model?

- Will try 16 different models.
- Treat data as equally spaced to simplify the illustration.
- ARMA(1,1) comes out best.
- TOEP is a close second.
- TOEP not significantly better than ARMA(1,1). (chisq=3.7, df=2).
- Similarly, ARMA(1,1) is also best for the BSI data from Project TALC.
- Random effects models are not the best for either outcome.

## Choosing the Covariance Model

**SAS code is given for fitting a variety of covariance models to the negative coping style for the CLEAR data. All have the same set of predictors: follow (categorical nominal times), partners in last 3 months, meth use in last 3 months.**

- (1) Independence (IND) model which never fits,**
- (2) Compound symmetry CS**
- (3) Random intercept (RI)**
- (4) Random intercept and slope (RIAS)**
- (5) Autoregressive (AR(1))**
- (6) Many others, using repeated statement, only Title1 and repeated statement are given.**

```
Title1 'IND: Independence never fits';  
proc mixed data=clearcaldar covtest;  
class SUBJECT follow ;  
model cpneg = follow part3m meth/ s;  
run;
```

```
Title1 'CS: Compound Symmetry';  
proc mixed data=clearcaldar covtest;  
class SUBJECT follow ;  
model cpneg = follow part3m meth/ s;  
repeated follow / sub=SUBJECT type=cs;  
run;
```

```
Title1 'RI: Random Intercept';  
proc mixed data=clearcaldar covtest;  
class SUBJECT follow ;  
model cpneg = follow part3m meth/ s;  
random intercept / sub=SUBJECT type=UN;  
run;
```

```
Title1 'RIAS: Random Intercept and Slope';  
proc mixed data=clearcaldar covtest;  
class SUBJECT follow ;  
model cpneg = follow part3m meth/ s;  
random intercept time/ sub=SUBJECT type=UN;  
run;
```

```
Title1 'AR(1): Autoregressive';  
proc mixed data=clearcaldar covtest;  
class SUBJECT follow ;  
model cpneg = follow part3m meth/ s;  
repeated follow / sub=SUBJECT type=ar(1);  
run;
```

```
Title1 'ARMA(1,1): Autoregressive moving average';  
repeated follow / sub=SUBJECT type=arma(1,1);
```

```
Title1 'UN: Unstructured';  
repeated follow / sub=SUBJECT type=UN;
```

```
Title1 'ANTE(1): Antedependence';  
repeated follow / sub=SUBJECT type=ANTE(1);
```

```
Title1 'TOEP: Toeplitz (banded)';  
repeated follow / sub=SUBJECT type=TOEP;
```

```
Title1 'FA(1): Factor Anal non-constant var, 1 factor';  
repeated follow / sub=SUBJECT type=FA(1);
```

```
Title1 'FA(2): Factor Anal non-constant var, 2 factors';  
repeated follow / sub=SUBJECT type=FA(2);
```

```
Title1 'FA1(1): Factor Analysis constant var, 1 factor';  
repeated follow / sub=SUBJECT type=FA1(1);
```

```
Title1 'FA1(2): Factor Analysis constant var, 2 factor';  
repeated follow / sub=SUBJECT type=FA1(2);
```

```
Title1 'CSH: Non-constant var compound symmetry';  
repeated follow / sub=SUBJECT type=csh;
```

```
Title1 'ARH(1): Non-constant var autoregressive';  
repeated follow / sub=SUBJECT type=arh(1);
```

```
Title1 'TOEPH: Non-constant var Toeplitz';  
repeated follow / sub=SUBJECT type=toeph;
```

**Fit statistics****Independence (IND)**

-2 Res Log Likelihood	5756.9
AIC (smaller is better)	5758.9
BIC (smaller is better)	5763.4

**Compound Symmetry (CS)**

-2 Res Log Likelihood	5446.7
AIC (smaller is better)	5450.7
BIC (smaller is better)	5457.0

**Autoregressive AR(1)**

-2 Res Log Likelihood	5481.5
AIC (smaller is better)	5485.5
BIC (smaller is better)	5491.8

**And so on. For these three models, IND is much worse than CS and CS is better than AR(1). Smaller is better and CS has the best so far.**

**Rather than interpreting raw SAS output, we put the information into a convenient table**

**and sort from best to worst among our covariance models.**

**On next page:**

- (1) Model – name of covariance model**
- (2) # parms – number of parameters in covariance model**
- (3) -2LL -2 Residual log likelihood**
- (4) AIC**
- (5) BIC**

Model	# parms	-2LL	AIC	BIC
ARMA ( 1 , 1 )	3	5435.0	5441.0	5450.5
TOEP	5	5431.3	5441.3	5457.1
RIAS	4	5438.7	5446.7	5459.3
TOEPH	9	5430.9	5448.9	5477.4
CS	2	5446.7	5450.7	5457.0
RI	2	5446.7	5450.7	5457.0
FA1 ( 2 )	10	5433.3	5453.3	5484.9
FA1 ( 1 )	6	5444.4	5456.4	5475.3
FA ( 2 )	13	5430.9	5456.9	5498.0
UN	15	5428.1	5458.1	5505.6
CSH	6	5446.5	5458.5	5477.5
FA ( 1 )	10	5439.1	5459.1	5490.8
AR ( 1 )	2	5481.5	5485.5	5491.8
ARH ( 1 )	6	5480.5	5492.5	5511.5
ANTE ( 1 )	9	5478.4	5496.4	5524.9
IND	1	5756.9	5758.9	5763.4

**Important: Sort table by AIC from smallest to largest.**

**Three models: (1) continuous time AR(1) models (uses the SP(EXP) key word), (2) continuous time AR(1) model plus random intercept, and (3) continuous time AR(1) model plus independent error. The variable 'time' needs to be the actual variable in the data set.**

```
Title1 'SP(EXP): Spatial exponential';  
proc mixed data=clearcaldar covtest;  
class SUBJECT follow ;  
model cpneg = follow part3m meth/ s;  
repeated follow / sub=SUBJECT type=sp(exp)(time);  
run;
```

```
Title1 'SP(EXP) + RI: Spatial exponential + random  
intercept';  
proc mixed data=clearcaldar covtest;  
class SUBJECT follow ;  
model cpneg = follow part3m meth/ s;  
repeated follow / sub=SUBJECT type=sp(exp)(time);  
random intercept / sub=SUBJECT type=un;  
run;  
*Need to define an observation index variable  
counterid;  
DATA clearcaldarnew;
```

```
    SET clearcaldar;
counterid = _N_;
RUN;
Title1 'SP(EXP) + IND: Spatial exponential +
Independent'; * Note: takes 28 minutes to run;
proc mixed data=clearcaldarnew covtest;
class SUBJECT follow ;
model cpneg = follow part3m meth/ s;
repeated follow / sub=SUBJECT type=sp(exp)(time);
random intercept / sub=counterid(SUBJECT) type=un;
run;
```

## Fit Statistics

**SP(EXP)**

-2 Res Log Likelihood	5500.6
AIC (smaller is better)	5504.6
BIC (smaller is better)	5510.9

**SP(EXP) + RI**

-2 Res Log Likelihood	5445.2
AIC (smaller is better)	5451.2
BIC (smaller is better)	5460.7

**sp(exp)(time) + IND (takes 28 minutes to run!)**

-2 Res Log Likelihood	5436.3
AIC (smaller is better)	5442.3
BIC (smaller is better)	5436.3

**The spatial AR(1) + independent error fits about as well as the ARMA(1,1) model. The basic difference is that spatial AR(1) uses exact times of observations, while ARMA(1,1) uses nominal times. The slight improvement in fit suggests that perhaps the nominal time is better to use than the actual time.**

# An Introduction to Modeling Longitudinal Data

## Session IV: Discrete Outcomes

Robert Weiss

Department of Biostatistics  
UCLA School of Public Health  
robweiss@ucla.edu

August 2010

### Outline

- Issues for Discrete Data
- Binary and Count Outcomes
- Random Effects
- Glimmix
- Examples

# Binary Outcomes

- Not all outcomes are continuous.
- Many outcomes are 0-1, yes/no, present/absent.
- Adolescent problem behaviors:
  - ▶ yes/no used drugs in past 3 months
  - ▶ yes/no smoked cigarettes 30 days
- Meth3m – used meth in the past 3 months
- Sex behaviors: unprotected sex at last act.

**Counts are discrete outcomes.** Responses take non-negative integer values, 0, 1, 2, etc.

Some questions from various surveys:

- Number of cigarettes smoked yesterday.
- Number of times used crack cocaine in the past 3 months.
- Number of sex partners last 3 months.
- Number of sex acts last 3 months.

# Bernoulli Distribution for Binary Data

The Bernoulli distribution describes 0/1 outcomes: failure/success, didnot/did. Let  $\pi_{ij}$  be the probability that  $Y_{ij} = 1$

$$Y_{ij} = \begin{cases} 0 & \text{with probability } 1 - \pi_{ij}, \\ 1 & \text{with probability } \pi_{ij}. \end{cases} \quad (1)$$

The mean and variance of  $Y_{ij}$  are

$$E[Y_{ij}] = \pi_{ij} \quad (2)$$

$$\text{var}(Y_{ij}) = \pi_{ij}(1 - \pi_{ij}). \quad (3)$$

In normal linear regression and in normal longitudinal models, the mean and variance are not connected. Here the variance is a function of the mean, and modeling binary data requires a special family of models.

# Modeling Discrete Outcomes

- The subject of **generalized linear models**.
- Outcomes  $Y_{ij}$  within subject  $i$  are correlated.
- There are several approaches to dealing with correlated binary data.
- We will look at **generalized linear mixed models. GLMM**.
- *Mixed models* include both fixed effects and random effects together in the same model.
- A historically derived name still in use.

# Predictors and the Mean

The usual linear model would set the mean of  $Y_{ij}$  to be a function of predictors  $X_{ijk}$  and coefficients  $\alpha_k$ :

$$E[Y_{ij}] = \pi_{ij} = \alpha_1 X_{ij1} + \alpha_2 X_{ij2} + \alpha_3 X_{ij3} + \dots \quad (4)$$

but  $0 \leq \pi_{ij} \leq 1$ , while the right hand side could be negative or greater than 1. This doesn't work.

A nonlinear *logit* transformation of  $\pi_{ij}$  allows us to set the logit function of  $\pi_{ij}$  equal to the *linear predictor*

$$\log \frac{\pi_{ij}}{1 - \pi_{ij}} = \alpha_1 X_{ij1} + \alpha_2 X_{ij2} + \alpha_3 X_{ij3} + \dots \quad (5)$$

# Logit function

The logit function is the *link* function between the mean and the linear predictor. The odds are the ratio

$$\frac{\pi_{ij}}{1 - \pi_{ij}}. \quad (6)$$

As  $\pi_{ij}$  takes on values from 0 to 1, the odds takes on values from 0 to plus infinity.

The *logit* function is the short name for the log odds function

$$\text{logit}(\pi_{ij}) = \log \frac{\pi_{ij}}{1 - \pi_{ij}} \quad (7)$$

which takes on values from minus infinity to plus infinity as  $\pi_{ij}$  varies from 0 to 1.

The inverse of the *logit* function is the *expit* function.

- The expit function takes a value of the linear predictor and converts it to a probability.
- Solving for  $\pi_{ij}$  as a function of the linear predictor

$$\pi_{ij} = \frac{\exp(\alpha_1 X_{ij1} + \alpha_2 X_{ij2} + \dots)}{1 + \exp(\alpha_1 X_{ij1} + \alpha_2 X_{ij2} + \dots)} = \text{expit}(\alpha_1 X_{ij1} + \alpha_2 X_{ij2} + \dots). \quad (8)$$

- This is *logistic regression*.
- Need to introduce correlations among longitudinal observations.

# Correlation in the Logistic Regression Model

$$Y_{ij} \sim \text{Bernoulli}(\pi_{ij}) \quad (9)$$

$$\text{logit}(\pi_{ij}) = \alpha_1 X_{ij1} + \alpha_2 X_{ij2} + \dots \quad (10)$$

- We have the
  - ▶ Distribution (Bernoulli) of the outcome  $Y_{ij}$ ,
  - ▶ Mean  $\pi_{ij}$  of the outcome, and
  - ▶ Linear predictor  $\alpha_1 X_{ij1} + \alpha_2 X_{ij2} + \dots$ .
- We are missing the correlation among observations.
- We can induce correlations across observations within a subject by adding *subject specific effect*  $\beta_i$  to the linear predictor.

# Logistic Random Effects Models

- Random effects models generalize directly to generalized linear models.
- The random effects are added as part of the linear predictor.
- Suppose we add a random intercept  $\beta_i$  to the linear predictor.

$$\text{logit}\pi_{ij} = \alpha_1 X_{ij1} + \alpha_2 X_{ij2} + \cdots + \beta_i \quad (11)$$

- $\beta_i$  increases or decreases the level of the linear predictor uniformly across all observations  $j$  for subject  $i$ .
- That increases or decreases the probability of each  $Y_{ij}$  equalling 1.

# Random effects have distributions

As with the normal random effects model, we put a distribution on the  $\beta_i$ .

To reiterate:

- Subjects are randomly selected from the population of subjects.
- There is one  $\beta_i$  per subject.
- The  $\beta_i$  are randomly selected from the population of subject specific effects.
- Hence, the  $\beta_i$  may be treated as random, hence *random effects*.
- The Normal distribution for modeling  $\beta_i$  is chosen out of convenience.

$$\beta_i \sim N(0, G) \tag{12}$$

# More Complex Random Effects

One can have more complex random effects.

- We saw the random intercept on the previous slide.
- Can also have a random intercept and slope (RIAS).

$$\text{logit}\pi_{ij} = \alpha_1 X_{ij1} + \alpha_2 X_{ij2} + \cdots + \beta_{i1} + \beta_{i2} t_{ij} \quad (13)$$

- Binary data can not support complex correlation models.
- RIAS is rarely useful for binary data.
- Stick to random intercept model for binary data.
- Count and continuous data can have more complex random effects models.
- If you have lots and lots of binary data, can consider more complex correlation structures.

# Meth3m as a Binary Outcome

For Glimmix Code and analysis Meth3m, we have glimmix code and results.

## Glimmix Examples

First example predicts binary methamphetamine (meth) use in the past 3 months (meth3m) for our CLEAR subjects.

Predictors are

**follow** – categorical nominal times of observation 0, 3, 6, 9, 15

**gender** – 1 = male, 0=female

**ethnic** – white/other = 1, black = 2, Hispanic = 3

**trade3m** – traded sex for money/drugs/room etc, past three months

This analysis uses a slightly different version of the data set (cleardata2).

```
title1 'Meth use last three months RI';
proc glimmix data=cleardata2;
  class subject follow ethnic;
  model meth3m = follow gender ethnic trade3m
              / dist=binomial link=logit s;
  random int / subject=subject;
run;
/* Ethnic: 1=white or other, 2=black, 3=Hispanic
*/
/* Gender: 1 = male, 0 = female */
```

**Comment statements to remind me of coding of the categorical predictors.**

```
title1 'Meth use last three months RI';
```

**Always document your runs.**

```
proc glimmix data=cleardata2;
```

**Call to glimmix and tell glimmix what data set is to be used.**

```
class subject follow ethnic;
```

**Tell SAS which predictors are categorical.**

```
model meth3m = follow gender ethnic trade3m /  
      dist=binomial link=logit s;
```

**The meat of the model specification. The outcome is meth3m meth use, last 3 months. The four predictors are then listed. After the forward slash /, key words are given.**

**dist=binomial defines the distribution as binomial, really Bernoulli, as the n is 1 for our binomial here. Say meth3m/n on the left of the equal sign if variable n in the data set is the binomial sample size for each observation and meth3m is the number of ‘successes’ out of n.**

**link=logit defines the link function.**

**s tells SAS we want the estimates and standard errors printed too.**

```
random int / subject=subject;  
run;
```

**We have a random intercept model and the variable subject is constant within subject, and changes from subject to subject.**

## Results. First SAS output, then a reasonably well formatted table of the results. Which is easier to read????

### Solutions for Fixed Effects

Effect	follow-up month	ethnic	Estimate	Standard Error	DF	t Value	Pr >  t
Intercept			-3.5601	0.5711	171	-6.23	<.0001
follow	0		0.5721	0.3487	550	1.64	0.1015
follow	3		0.7090	0.3597	550	1.97	0.0492
follow	6		0.5240	0.3755	550	1.40	0.1635
follow	9		0.1937	0.3691	550	0.52	0.5999
follow	15		0	.	.	.	.
gender			1.4257	0.4916	550	2.90	0.0039
ethnic		1	1.3342	0.3974	550	3.36	0.0008
ethnic		2	0.000073	0.4638	550	0.00	0.9999
ethnic		3	0	.	.	.	.
trade3m			0.8070	0.3552	550	2.27	0.0235

### Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
follow	4	550	1.33	0.2587
gender	1	550	8.41	0.0039
ethnic	2	550	6.78	0.0012
trade3m	1	550	5.16	0.0235

**Outcome: meth3m**

Effect	Est	SE	t	p
Intercept	-3.56	.57		
Month 0	.57	.35	1.6	.10
Month 3	.71	.36	2.0	.049
Month 6	.52	.38	1.4	.16
Month 9	.19	.37	.5	.60
Month 15 (ref)			F=1.3, p=.26	
Male	1.43	.49	2.9	.004
White	1.33	.40	3.4	.001
Black	.00	.46	.0	1.00
Hispanic (ref)			F=6.78, p=.001	
Sex trading 3 months	.81	.36	2.3	.024

Properly formatted table of regression results.

## Summary of results

**Time trend is not significant!**

**Males use much more meth than females. Odds ratio for females to males, if you prefer, is  $OR = 4.2$  with a 95% CI of (1.6, 11.1).**

**Whites are much more frequent users of meth than blacks or Hispanics, who use meth with about the same frequency.  $OR=3.8$  (1.7, 8.4),  $p=.001$ .**

**Sex trading in the past 3 months is associated with greater frequency of meth use in the past 3 months,  $est(se) = .81(.36)$ ,  $t=2.3$ ,  $p=.024$ .**

**Can also convert to odds ratios and 95% confidence intervals as well.**

Effect	OR	95% CI		p	
Month 0	1.77	.88	3.6	.10	
Month 3	2.03	.99	4.2	.049	
Month 6	1.69	.80	3.6	.16	
Month 9	1.21	.58	2.5	.60	
Month 15 (ref)			F=1.3, p=.26		
Male	4.16	1.56	11.1	.004	
White	3.80	1.71	8.4	.001	
Black	1.00	.40	2.5	1.00	
Hispanic (ref)			F=6.78, p=.001		
Sex trading 3 months	2.24	1.10	4.6	.024	

**Same results, but converted to odds ratios and 95% CI for the odds ratio. Depends on your preference, but has virtually the same information as the previous table.**

We follow the same game plan for count data.

- Choose a distribution (Poisson, Negative Binomial)
- Choose a link function
- Linear predictor
- Random effects added to the linear predictor.

Poisson regression is used to model count data.

- Outcome  $Y_{ij}$  is a non-negative integer,  $0, 1, 2, \dots$
- $E[Y_{ij}] = \lambda_{ij}$  at time  $t_{ij}$ .
- $\text{var}[Y_{ij}] = \lambda_{ij}$ .
- Mean equals variance for Poisson random variables.
- Mean  $\lambda_{ij}$  is positive.
- Log link is the usual link function.
- $\log \lambda_{ij} = \alpha_1 X_{ij1} + \alpha_2 X_{ij2} + \dots + \beta_j$ .

# Interpretation of Coefficients in Poisson Regression

Suppose  $X_{ij2}$  can take values 1 and 0 only and  $X_{ij1} = 1$  is the intercept. Then

$$\log \frac{\lambda_{ij}(X_{ij2} = 1)}{\lambda_{ij}(X_{ij2} = 0)} = \log \frac{\exp(\alpha_1 + \alpha_2 + \beta_i)}{\exp(\alpha_1 + \beta_i)} = \log(\exp(\alpha_2)) = \alpha_2. \quad (14)$$

Then  $\exp(\alpha_2)$  is a ratio of means for the two values of predictor  $X_{ij2}$ .

## The good news about the Poisson.

- Can consider more complicated random effects such as the random intercept and slope (RIAS) model.
- Not in our particular data set though.
- Software will fit this model to many data sets.
- Well known, and commonly used.

# The Problem with the Poisson

The Poisson distribution assumes that the variance is equal to the mean.

- Even with a random intercept in the model, given the random effect, the Poisson assumes mean equal to variance.
- Not true in data sets such as counts of sexual behaviors, amount of drug use.
- Need a distribution where the variance can be larger than the mean.
- The *Negative Binomial* distribution has this property.
- Using the Poisson ignores variability in the data, causing
  - ▶ Overly small standard errors
  - ▶ Overly small confidence intervals
  - ▶ Inappropriately small p-values

# The Negative Binomial Distribution

The negative binomial distribution is a distribution on the non-negative integers  $0, 1, 2, \dots$ .

- The negative binomial should be utilized preferentially over the Poisson for many sexual behavior analyses and counts of days of drug usage.
- Has an extra parameter that determines the variance – the variance is larger than the mean.
- Accommodates the extra-Poisson variability typical in data sets like these.
- Extra Poisson variability also known as *over-dispersion*.
- Downside: Fewer significant results.

Poisson models come with a goodness of fit statistic.

- The Generalized chi-square statistic can be inspected for lack of fit.
- It should be near 1 if the model fits correctly.
- The negative binomial has an extra *over-dispersion* parameter that is estimated by the data to make the goodness of fit statistic be exactly 1.
- The Poisson model chi-square goodness of fit statistic should be near 1 if Poisson model is correct.

## Count of sex acts, last three months

- We expect counts of sex acts in the past three months in high risk populations to be overdispersed compared to the Poisson.
- We fit a model with predictors  
`follow intv intv*follow gender ethnic trade3m.`
- Poisson goodness-of-fit statistic is 16, much much larger than 1.
- 2 would be large, much less 16
- In the CLEAR data set for this outcome, the Poisson model does not fit.
- We are not surprised, as this is a common problem with self-reported sex and drug behaviors in high-risk populations.
- We rely on the negative binomial results for this data set.

## Some Results Differ Between the Poisson and Negative Binomial Model.

- The intervention by time interaction is highly significant in the Poisson model  $p < .0001$ .
- It is not significant  $p = .54$  in the negative binomial model.
- I believe the negative binomial model.
- Sex trading last 3 months has essentially equal statistical significance in both models,  $\text{est} = .45(.21)$ ,  $p = .038$ .
- $\exp(.45) = 1.54$ , indicating that according to the negative binomial model:
- Sex trading in the past 3 months is associated with 54% (1%, 134%) more sex partners in the past 3 months compared to when not.

# Poisson and Negative Binomial Example

In the glimmix example we predict the number of sex acts in the past 3 months for our CLEAR study participants. The examples give details of the Poisson and negative binomial models at work including glimmix code and output.

## Count outcome examples: Negative Binomial model.

```
title1 'mmp1 Sex acts last 3m RI, Negative Binomial';  
proc glimmix data=cleardata2 method=mmp1;  
  class subject follow ethnic;  
  model act3m = follow intv intv*follow gender  
ethnic trade3m/ dist=negbin link=log s;  
  random int / subject=subject;  
run;
```

This negative binomial model needs `method=mmp1` to run. The `method=` key word selects an approach to estimation. The Poisson model and the binomial model run with the default `method=` and thus we don't include the keyword for those models. Negative binomial doesn't run using the default method, so we use the mmp1 method.

`random int` abbreviates intercept in the random statement.

Other parts of the model specification are similar to what we have seen before. The negative binomial distribution is specified by the keyword `dist=negbin`.

## The Poisson Model.

```
title1 'Sex acts last 3m RI, Poisson';  
proc glimmix data=cleardata2 ;  
  class subject follow ethnic;  
  model act3m = follow intv intv*follow gender  
ethnic trade3m/ dist=poisson link=log s;  
  random int / subject=subject;  
run;
```

`dist=poisson link=log` specify the Poisson distribution with log link. Other parts of the analysis follow along our previous model specifications.

**Fit statistics: How might we guess that the Poisson model doesn't fit well? Look at the fit statistics.**

### Fit Statistics Negative Binomial

-2 Log Pseudo-Likelihood	2893.23
Generalized Chi-Square	729.96
Gener. Chi-Square / DF	1.00

### Fit Statistics Poisson

-2 Res Log Pseudo-Likelihood	12378.34
Generalized Chi-Square	11821.81
Gener. Chi-Square / DF	16.49

The Pseudo-Likelihoods are not comparable unless we use the same method= , which we could not in this data set. However, the generalized chi-square divided by its degrees of freedom is a rough measure of goodness of fit, and should be in the ballpark of 1 if things are going well. If it is larger than 1, it is sign

that the model does not fit well. In the Poisson case, the generalized chisquare divided by its df is 16, which is huge, suggesting a distinct lack of fit. The negative binomial distribution has an *over-dispersion* parameter and its goodness of fit statistic is automatically 1.

## A comparison of Negative Binomial and Poisson hypothesis tests.

Effect	Num	Den	Poisson		Negative Binomial	
	DF	DF	F	p	F	p
Follow	4	547	66.4	<.0001	1.77	.13
Intv	1	547	.05	.83	.26	.61
Intv*Follow	4	547	18.0	<b>&lt;.0001</b>	.77	.54
Gender	1	547	.13	.72	2.41	.12
Ethnicity	2	547	.81	.45	1.23	.29
Trade3m	1	547	4.38	<b>.037</b>	4.27	<b>.039</b>

The intervention by time interaction is significant under the Poisson model, but not significant under the negative binomial model. Sex trading (Trade3m) is significant in both models.

The coefficient of Trade3m is estimated at .43 (.21) in the negative binomial model.

Exponentiating,  $\exp(.43) = 1.54$ , indicating that when participants are engaging in sex trading in the past three months, their number of sex partners tends to be 1.54 times higher, that is, 54% higher, than when they are not sex trading.

In the negative binomial model, if we omit Intv\*Follow and Intv, then Follow is still not significant.

# An Introduction to Modeling Longitudinal Data

## Session V: Two Longitudinal Outcomes

Robert Weiss

Department of Biostatistics  
UCLA School of Public Health  
robweiss@ucla.edu

August 2010

# An Introduction to Longitudinal Data Analysis: Part V

## Two Longitudinal Outcomes

### Outline

- Two longitudinal outcomes
- The bivariate random effects model
- More complex models
- Examples

My colleagues frequently are interested in the relationship between two outcomes, both measured over time.

- TALC: BSI Somatic symptoms in adolescents, and
  - ▶ Life events, parental bonding
  - ▶ Parental somatic, BSI, depression
- HIV+ risk behaviors (sex, drug) and mental health states (STAI, BDI) (Comulada et al 2010 Psych Addictive Behaviors)
- Child growth and its relationship to nutrition
- Pain rating and pain tolerance (Weiss 2005, book chap 13)

# Two Longitudinal Variables

Why not include one variable  $W_j$  as a time-varying predictor for the other  $Y_j$ ?

- Dropping the  $i$  for subject to reduce clutter.
- $W_j, Y_j$  are two outcomes at time  $j$ .
- Which variable is the predictor, which is the outcome?
- There is only 1 regression coefficient.
- Only one type of relationship.
- Hides details of the interrelationship.
- Assumes  $W_k$  does not have independent correlation with  $Y_j$  except through  $W_j$ .

Need bivariate longitudinal model.

# Modeling Two Longitudinal Outcomes

- A bivariate longitudinal model treats both variables as outcomes.
- The correlations between the two outcomes are of interest, not the means.
- The covariance model models the variances and correlations of both variables over time in a single large model.
- The covariance model can be chosen to appropriately model the interrelationships (read: correlations) among the two variables.
- A statistical model allows for subtle and complex relationships to be discovered.
- Some models allow early  $Y$  to predict later  $W$ s but not vice-versa, suggesting possible causality.

# How Might Two Longitudinal Variables Be Interrelated?

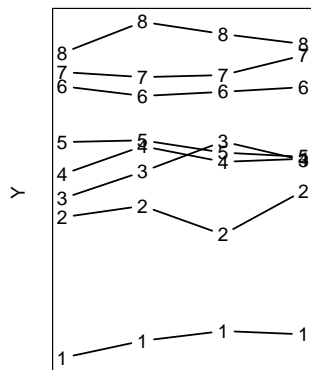
There are many ways that two longitudinal variables  $Y_j$  and  $W_j$  can be interrelated.

- The average values over time may be correlated.
- Visit to visit variation may be correlated.
- Both visit to visit and average values may be correlated but with different strengths (or signs!).
- The time trend (slope) in one variable may be correlated with the level of the other variable.
- And so on.

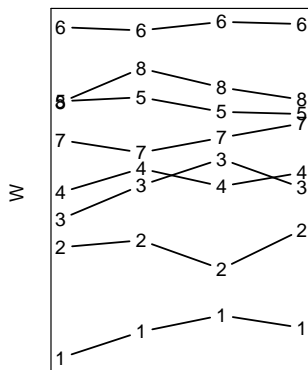
Lets see examples of the first two possibilities.

# Bivariate Longitudinal Data: Correlated Average Responses

## Strongly Correlated Random Effects



Time  
(a)



Time  
(b)

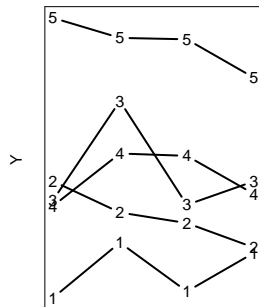
# Two Longitudinal Outcomes: Average Values Correlated

Previous profile plots of longitudinal  $Y$ s on the left,  $W$ s on the right.

- The average values over time are correlated.
- Subjects 1 through 8 numbered from lowest to highest on  $Y$  on left.
- Subjects 1 and 2 are near bottom for both outcomes  $Y$  and  $W$ .
- Subjects 5 through 8 at top of  $W$ .
- If look carefully, also can see visit to visit variation is positively correlated as well.

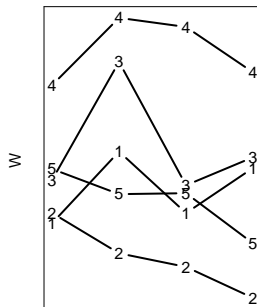
# Bivariate Longitudinal Data: Correlated Residuals

## Strongly Correlated Residuals



Time

(a)



Time

(b)

# Two Longitudinal Outcomes: Correlated Visit to Visit Changes

Previous profile plot: Longitudinal  $Y$ s on the left,  $W$ s on the right.

- Average values over time are uncorrelated.
- Visit to visit changes are highly correlated on the two outcomes.
- Subject 2 decreasing across time on both outcomes.
- Subject 4 up then flat then down on both outcomes.
- Subjects 1 and 3 both up, then down, then up.

Example: Two psychometric scales  $Y$ ,  $W$  whose levels have little meaning, but changes within subject are meaningful.

For each covariance there is a correlation.

- Recall that correlation is covariance scaled to a  $(-1, 1)$  scale.
- Will try to stick with correlation, however:
- I tend to use the words correlation and covariance interchangeably.
- Testing a covariance equal to zero is equivalent to testing the correlation equal to zero.
- Can switch back and forth from covariances to correlations.

# Let's Talk Correlations

Interest is in the *cross correlations* between the two longitudinal variables.

- $Y_j$  – measured at time  $t_j = j$
- $W_k$  – measured at time  $t_k = k$

The *cross correlations* of the  $Y_j$ s and  $W_k$ s are

$\text{Corr}(Y_j, W_j)$  same times  $j$  and  $j$

$\text{Corr}(Y_j, W_k)$  different times  $j \neq k$

Positive? Negative? Zero?

# Cross Correlations Between $Y$ and $W$

With  $J$  time points, there are  $J^2$  cross-correlations  $\text{Corr}(Y_j, W_k)$ ,  $j$  and  $k$  may be equal.

We are interested in questions such as

- Are all  $J^2$   $\text{Corr}(Y_j, W_j)$ ,  $\text{Corr}(Y_j, W_k)$  zero? All positive? All negative?
- If not zero, what are they?
- Are equal time  $\text{Corr}(Y_j, W_j)$  higher than  $j \neq k$   $\text{Corr}(Y_j, W_k)$ ?
- Are simpler parameterizations possible with fewer than  $J^2$  parameters? (Yes)

## Several models are available in SAS

- Unstructured covariance model
- Product correlation model
- **Bivariate random effects models**

## Unstructured Covariance Model

- If there are only a few time points  $J$ , the unstructured covariance model can be fit to the bivariate longitudinal data.
- Has  $J^2$  unstructured correlation parameters to test jointly equal to zero.
- Has the unstructured covariance model for each individual outcome.
- (Very) low power.
- Need to test all  $J^2$  correlation parameters equal to zero.
- Advantage: can inspect individual correlation parameters.

## Product Correlation Models

- Product correlation model has only 1 cross-correlation parameter!
- High power, 1 degree of freedom, but stringent assumptions.
- Alternative hypothesis is that  $Y$  and  $W$  are independent.
- Equivalent to letting  $W$  be a scalar predictor of  $Y$ .

We will concentrate on random effects models.

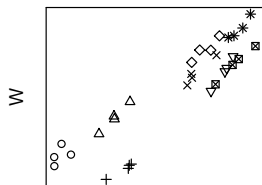
# The Bivariate Random Intercept Model

Two random intercepts  $\beta_W, \beta_Y$  one for each variable.

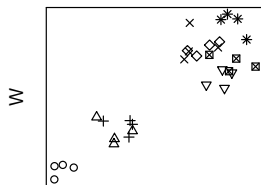
$$Y_j = \mu_{Yj} + \beta_Y + \delta_{Yj}$$
$$W_j = \mu_{Wj} + \beta_W + \delta_{Wj}$$

- Each outcome has a population mean  $\mu_{Yj}, \mu_{Wj}$  a function of predictors and regression coefficients.
- The mean is NOT of primary interest.
- Interest lies in the correlations between  $Y_j$  and  $W_j$ .
- The correlations  $\text{Corr}(Y_j, W_k)$  are driven by the
  - ▶ Correlation  $\text{Corr}(\beta_Y, \beta_W)$  between the random intercepts, and the
  - ▶ Correlation  $\text{Corr}(\delta_{Yj}, \delta_{Wj})$  between the residuals.
- Residuals at different times are uncorrelated.

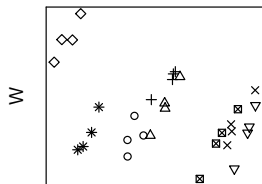
# Bivariate Longitudinal Data: Constructed Example, 8 Subjects, 4 Longitudinal Measures



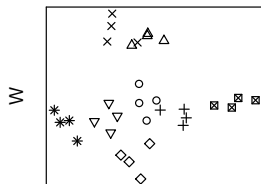
(a)



(b)



(c)



(d)

Four examples of bivariate longitudinal data. Correlated within subject (left column), correlated between subjects (top row). Time is *not* plotted

## Random Intercept Correlation

- Each subject's observations  $Y_j$  fall consistently above or below the population mean  $\mu_Y$  by the amount  $\beta_Y$ .
- Similarly for  $W_j$ .
- Is a subject typically higher on average on both  $W$  and  $Y$ ?
- Or lower on average on both  $W$  and  $Y$ ?
- Then the  $\beta_Y, \beta_W$  random effects are positively correlated.
- Weight and height over time – random intercepts are positively correlated. Residuals likely uncorrelated.

## Residual Correlation

- $\mu_{Y_j} + \beta_{Y_j}$  is the subject specific mean for the  $Y$ s.
- Residuals  $\delta_{Y_j}$  and  $\delta_{W_j}$  determine if observations  $Y_j$ ,  $W_j$  are above or below their own subject mean.
- Positive residual correlation: Observations  $Y_j$ ,  $W_j$  both tend to be above or below the subject specific means at the same times.
- Negative residual correlation: Observation  $Y_j$  is over,  $W_j$  is under the subject specific means at the same times.
- Recent extra drug use positively associated with depression in HIV+ young people.

# Bivariate Random Intercept and Slope

Each subject has a random intercept and random slope (RIAS) for each outcome.

- Random intercepts – where subjects start at baseline.
- Random slopes – how subjects trend (increase/decrease) over time.
- Four random effects – all can be correlated.
- Residuals – visit to visit variation.

## Bivariate Longitudinal RIAS Between BSI Somatic Symptoms and

- Adolescent variables
  - ▶ Life Events (log  $x + 1$  scale)
  - ▶ Parent Care (Parker Parental Bonding)
  - ▶ Parent Overprotection (Parker Parental Bonding)
- Parent variables
  - ▶ BSI Total, Depression, Somatization

# Adolescent BSI Somatization

- BSI somatization log transformed due to skewness
- As are parent BSI variables.
- Fit bivariate random intercept and slope model to these data.
- More data about adolescent variables than parent variables.
- More information about residual correlation than random effects correlation.
- Reporting intercept correlations, slope correlations and residual correlations.

# Adolescent Somatization and Adolescent Variables

	Log Life Events		Parent Care		Parent Overprotection	
	Corr	<i>p</i>	Corr	<i>p</i>	Corr	<i>p</i>
Intercepts	.52	<.0001	-.37	<.0001	.17	.01
Slopes	.60	<.0001	-.16	.08	.04	.72
Residuals	.16	<.0001	.01	.55	.04	.04

Life events intercepts and slopes and residuals highly correlated with somatization intercepts and slopes and residuals.

Parent Care is good, parent overprotection is bad. Intercepts correlated, but trends and residuals not related to soma trend.

# Adolescent Somatization and Parent Variables

	Parent BSI		Parent Depression		Parent Somatization	
	Corr	$p$	Corr	$p$	Corr	$p$
Intercepts	.26	.0001	.25	.0003	.28	<.0001
Slopes	.15	.15	.13	.38	.04	.81
Residuals	.04	.06	.00	.90	.04	.04

Parent variables all have similar relationships with adolescent somatization. Intercepts are correlated. Residual correlations are small, but borderline (not) significant.

# Discrete Data

- Can analyze bivariate longitudinal discrete data using SAS Proc Glimmix.
- Can analyze mixed data types, binary/continuous, binary/count, count/continuous.
- Won't have residual correlations, but can check for intercept and slope correlations.

That's All Folks!

Thank you for listening

# References

Weiss RE (2005). *Modeling Longitudinal Data*. Springer.

Fitzmaurice, GM, Laird, NM, Ware, JH (2004). *Applied Longitudinal Analysis*. Wiley.

Comulada WS, Rotheram-Borus MJ, Pequegnat W, Weiss RE, et al (2010). Relationships over Time between Mental Health Symptoms and Transmission Risk Among Persons Living with HIV. *Psychology of Addictive Behaviors*, 24, 109–118.

Rotheram-Borus MJ, Weiss RE, Alber S, and Lester P (2005). Adolescent Adjustment Before and After HIV-related Parental Death. *Journal of Consulting and Clinical Psychology*, 73, 221–228.

Singer, JD (1998). Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models. *Journal of Educational and Behavioral Statistics*, Vol. 23, No. 4, 323–355.