

**Modern Methods for Longitudinal Data
Analysis: Are Developments in Statistical
Software Finally Catching Up?**

Garrett Fitzmaurice

Laboratory for Psychiatric Biostatistics
McLean Hospital

Department of Biostatistics
Harvard School of Public Health

**Developments in methods for
longitudinal analysis**

Remarkable developments in methods for
longitudinal analysis during past 25 years.

The 1980's and 1990's might be considered
"golden years" for methods development.

Are these methods being widely applied?

Longitudinal Studies: Basic Concepts

- Defining feature of longitudinal studies is that measurements of the same individuals are taken repeatedly through time.
- Longitudinal studies allow direct study of change over time.
- Objective: primary goal is to characterize the change in response over time and the factors that influence change.

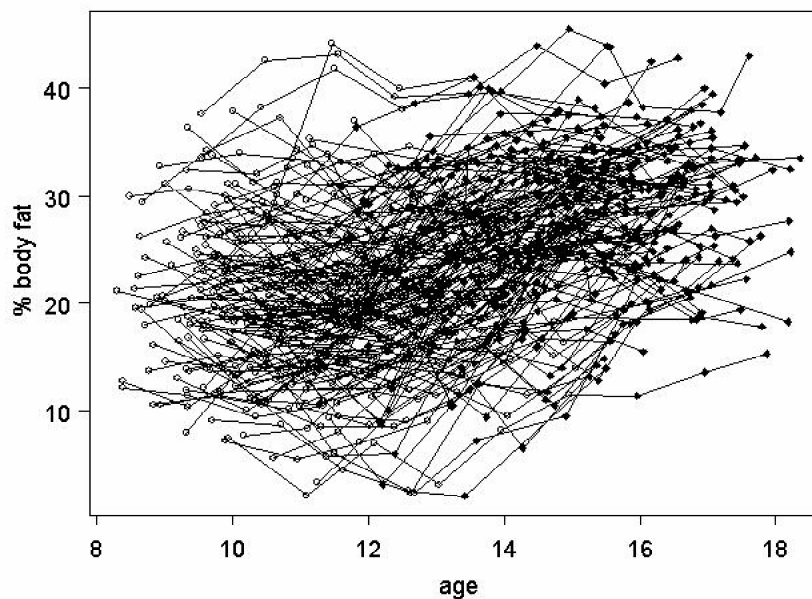
Longitudinal Data: Complications

- Repeated measures on individuals are correlated.
- Variability is often heterogeneous across measurement occasions.

Correlation and heterogeneous variability must be properly accounted for to obtain valid inferences.

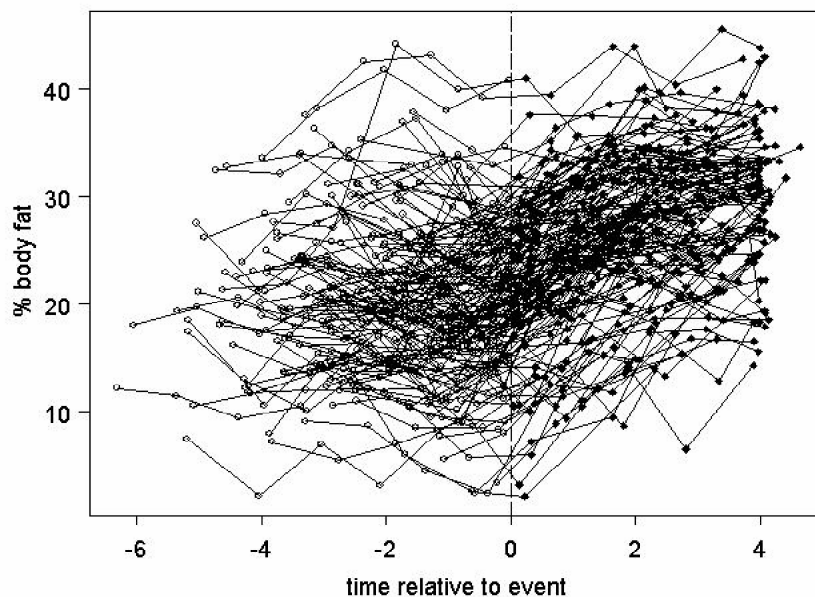
Example 1: Influence of Menarche on Changes in Body Fat

- Prospective study on body fat accretion in a cohort of 162 girls from the MIT Growth and Development Study.
- At start of study, all girls were pre-menarcheal and non-obese.
- Followed over time according to schedule of annual measurements until 4 years after reported date of menarche.
- At each examination, a measure of body fatness was obtained based on bioelectric impedance.



Example 1: Influence of Menarche on Changes in Body Fat

- Goal of analysis: compare changes in percent body fat before and after menarche.
- For the purposes of these analyses time must be coded as time since menarche.
- Note: Timing of measurements is regular or “balanced” when defined as time since baseline measurement.
- It is inherently irregular or “unbalanced” when timing of measurements is defined as time since menarche.



Example 2: Oral Treatment of Toenail Infection

- Randomized, double-blind, study of 294 patients comparing 2 oral treatments (denoted A and B) for toenail infection.
- Outcome variable: Binary variable indicating presence of onycholysis (separation of the nail plate from the nail bed).
- Measurement schedule: Baseline (week 0), weeks 4, 8, 12, 24, 36, and 48.
- Interested in the rate of decline of the proportion of patients with onycholysis over time and the effect of treatment on that rate.

Example 3: Clinical Trial of Anti-Epileptic Drug Progabide

- Randomized, placebo-controlled study of treatment of epileptic seizures with progabide.
- Patients randomized to treatment with progabide, or to placebo in addition to standard therapy.
- Outcome variable: Count of number of seizures.
- Measurement schedule: Baseline measurement during 8 weeks prior to randomization. Four measurements during consecutive 2-week intervals.

Features of Longitudinal Data

In longitudinal studies the outcome variable can be:

- continuous (e.g., percent body fat)
- binary (e.g., presence/absence of onycholysis)
- count (e.g., number of epileptic seizures)

The data set can be incomplete (missing data).

Subjects may be measured at different occasions (e.g., due to mistimed measurements).

Modern methods can handle all of these cases.

Developments in methods: Some highlights from the 1980's

Linear Mixed Effects Models:

Widely used method for analyzing longitudinal data.

Early development of mixed effects models for correlated data can be traced to ANOVA paradigm and seminal paper by Harville (1977).

Their usefulness for analyzing longitudinal data was highlighted by Laird and Ware (1982).

Some more highlights from the 1980's

Generalized Estimating Equations:

Liang and Zeger (1986) introduced generalized estimating equations (GEE).

They proposed a family of generalized linear models (e.g., logistic and loglinear regression) for fitting repeated observations of binary and counted data.

These models are referred to as "marginal models".

Some more highlights from the 1980's

Generalized Linear Mixed Models:

Extended conceptual approach of linear mixed effects models to repeated binary and count data.

Their conceptual foundation was laid in the 1970's (e.g., Ashford and Sowden, 1970; Pierce and Sands, 1975; Korn and Whittemore, 1979).

Work that followed in the 1980's and 1990's focused on the thorny problem of estimation.

Some highlights from the 80's and 90's

Generalized Linear Mixed Models:

Because no simple analytic solutions were available, Stiratelli, Laird, and Ware (1984) proposed an approximate method of estimation.

This led to development of general approach for fitting GLMMs known as penalized quasi-likelihood (PQL).

For example, Schall (1991), Breslow and Clayton (1993), and Wolfinger (1993).

Developments in methods for longitudinal analysis

Many developments in methods for longitudinal analysis occurred between the early 1980-1990's.

Despite important advances, these methods have been somewhat slow to move into the mainstream.

To bridge gap between methods development & application, appropriate statistical software is essential.

Statistical software for longitudinal data analysis

In general, methods available via commercially available software tend to lag behind recent advances in statistical methods.

Longitudinal analysis is not exceptional in this regard.

However, relatively recent introduction of new procedures for analyzing longitudinal/correlated data has made these methods accessible to practitioners.

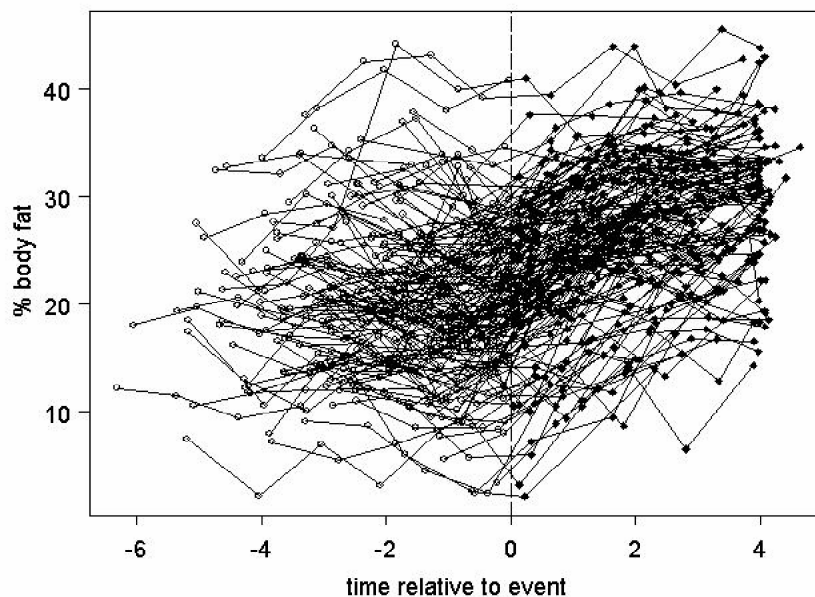
SAS procedures for longitudinal data analysis

In particular, SAS has produced a versatile set of procedures that can be used for the analysis of longitudinal data:

- ❑ PROC MIXED (Version 6.09 *circa* 1994-5)
- ❑ PROC GENMOD (Version 6.12 *circa* 1997)
- ❑ PROC NLMIXED (Version 8 *circa* 1999)
- ❑ PROC GLIMMIX (Version 9 *circa* 2006)

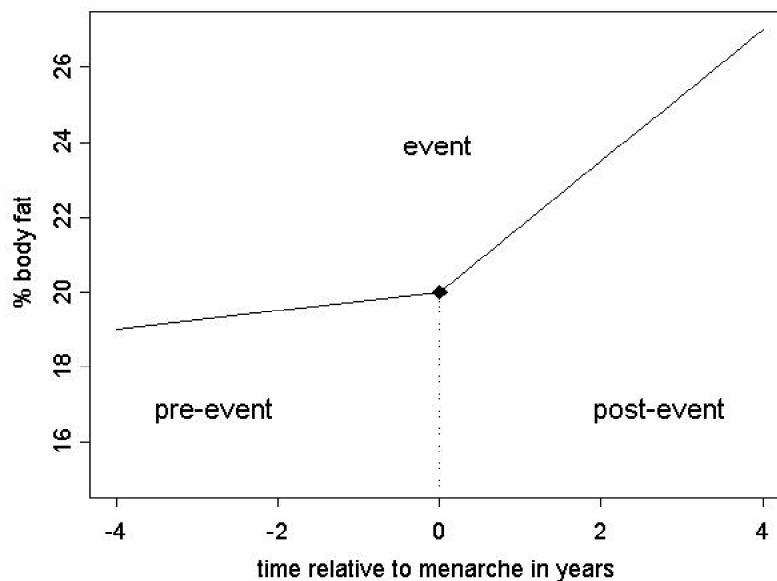
Fitting Linear Mixed Effects Models: Influence of Menarche on Changes in Body Fat

- Prospective study on body fat accretion in a cohort of 162 girls from the MIT Growth and Development Study.
- At start of study, all girls were pre-menarcheal and non-obese.
- Followed over time according to schedule of annual measurements until 4 years after reported date of menarche.
- At each examination, a measure of body fatness was obtained based on bioelectric impedance.



Fitting Linear Mixed Effects Models

- Consider hypothesis that %body fat increases linearly with age, but with different slopes before/after menarche.
- Assume each girl has a piecewise linear spline with a knot at the time of menarche (see Figure)
- Each girl's growth curve described with an intercept and two slopes, one slope for changes in response before menarche, another slope for changes in response after menarche.
- Note: knot is not at a fixed age for all subjects.



Fitting Linear Mixed Effects Models

Let t_{ij} denote time (years) of the j^{th} measurement on i^{th} subject before or after menarche (i.e., $t_{ij} = 0$ at menarche).

Consider the following linear mixed effects model:

$$E(Y_{ij} | b_i) = \beta_0 + \beta_1 t_{ij} + \beta_2 (t_{ij})_+ + b_{0i} + b_{1i} t_{ij} + b_{2i} (t_{ij})_+ + e_{ij}$$

where $(t_{ij})_+ = t_{ij}$ if $t_{ij} > 0$ and $(t_{ij})_+ = 0$ otherwise.

Interpretation of Parameters

Intercept, β_0 , is average %body fat at menarche (when $t_{ij} = 0$).

Slope, β_1 , is the average rate of change in %body fat (per year) during the pre-menarcheal period.

The average rate of change in %body fat (per year) during the post-menarcheal period is $\beta_1 + \beta_2$.

Goal: Assess whether population slopes differ before and after menarche, i.e., $H_0: \beta_2 = 0$.

Fitting Linear Mixed Effects Models

SAS commands:

```
data fat;
  infile 'fat.dat';
  input id age agemen time pbf;
  time_0=max(time,0);

title1 'Mixed Effects Model for %Body Fat with Random
  Intercepts and Slopes before and after Menarche';
proc mixed;
  class id;
  model pbf = time time_0 / s;
  random intercept time time_0 / subject=id type=un g;
run;
```

Selected SAS Output

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	21.3614	0.5646	161	37.84	<.0001
time	0.4171	0.1572	161	2.65	0.0087
time_0	2.0471	0.2280	159	8.98	<.0001

Estimated G Matrix					
Row	Effect	id	Col1	Col2	Col3
1	Intercept	1	45.9413	2.5263	-6.1096
2	time	1	2.5263	1.6311	-1.7505
3	time_0	1	-6.1096	-1.7505	2.7497

Interpretation of Results

Estimate of intercept, β_0 , is 21.36 and has interpretation as average %body fat at menarche.

Estimate of pre-menarcheal slope, β_1 , is 0.42 and indicates annual rate of growth is less than 0.5%.

Estimate of post-menarcheal slope, $\beta_1 + \beta_2$, is 2.46 and indicates annual rate of growth is approx. 2.5%.

Comparing estimate of β_2 to its standard error, slopes before and after menarche differ (at the 0.05 level).

Fitting Marginal Models via GEE: Oral Treatment of Toenail Infection

- Randomized, study of 294 patients comparing 2 oral treatments (denoted A and B) for toenail infection.
- Outcome: Binary variable indicating presence of onycholysis (separation of nail plate from nail bed).
- Measurement schedule: Baseline (week 0), weeks 4, 8, 12, 24, 36, and 48.
- Interested in rate of decline of the proportion of patients with onycholysis over time and the effect of treatment on that rate.

Fitting Marginal Models via GEE

Let t_{ij} denote time of the j^{th} measurement on i^{th} subject (in months).

Assume *marginal* probability of onycholysis follows the logistic model:

$$\text{Logit}\{\Pr(Y_{ij} = 1)\} = \beta_0 + \beta_1 t_{ij} + \beta_2 \text{trt} * t_{ij}$$

where $\text{trt}=1$ if treatment group B and 0 otherwise.

We assume $\text{Var}(Y_{ij}) = \Pr(Y_{ij} = 1) * \Pr(Y_{ij} = 0)$ and estimate all possible pairwise correlations.

Fitting Marginal Models via GEE

SAS commands:

```
data toenail;
  infile 'toenail.dat';
  input id y trt month visit;

title 'Marginal Model for Onycholysis Data';
proc genmod descending;
  class id visit;
  model y = month trt*month / dist=bin link=logit;
  repeated subject=id / withinsubject=visit type=un corrw;
run;
```

Analysis Of GEE Parameter Estimates						
Empirical Standard Error Estimates						
Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > Z
Intercept	-0.6981	0.1217	-0.9365	-0.4596	-5.74	<.0001
month	-0.1401	0.0261	-0.1913	-0.0889	-5.36	<.0001
trt*month	-0.0806	0.0415	-0.1620	0.0008	-1.94	0.0523

Working Correlation Matrix							
	Col1	Col2	Col3	Col4	Col5	Col6	Col7
Row1	1.0000	0.9061	0.7239	0.5031	0.2433	0.1758	0.1521
Row2	0.9061	1.0000	0.8389	0.6012	0.2877	0.2041	0.2233
Row3	0.7239	0.8389	1.0000	0.7680	0.3034	0.2093	0.2198
Row4	0.5031	0.6012	0.7680	1.0000	0.3798	0.2809	0.2657
Row5	0.2433	0.2877	0.3034	0.3798	1.0000	0.5182	0.4447
Row6	0.1758	0.2041	0.2093	0.2809	0.5182	1.0000	0.7310
Row7	0.1521	0.2233	0.2198	0.2657	0.4447	0.7310	1.0000

Interpretation of Results

- Suggestion of a difference in the rate of decline in the two treatment groups (p-value = 0.052).
- Over 12 months, odds of onycholysis decreased by factor of 0.19 [$\exp(-0.14 \times 12)$] in group A.
- Over 12 months, odds of onycholysis decreased by factor of 0.07 [$\exp(-0.221 \times 12)$] in group B.
- Odds ratio comparing 12 month decreases in risk between groups A & B is approx 2.6 (or $e^{12 \times 0.081}$).
- Overall, significant decline over time in prevalence of onycholysis for all randomized patients.

Fitting Generalized Linear Mixed Models

Let t_{ij} denote time of the j^{th} measurement on i^{th} subject (in months).

Assume *conditional* probability of onycholysis follows the logistic model:

$$\text{Logit}\{\Pr(Y_{ij} = 1|b_i)\} = \beta_0 + \beta_1 t_{ij} + \beta_2 \text{trt} * t_{ij} + b_i$$

where $\text{trt}=1$ if treatment group B and 0 otherwise.

We assume $\text{Var}(Y_{ij}|b_i) = \Pr(Y_{ij} = 1|b_i) * \Pr(Y_{ij} = 0|b_i)$ (Bernoulli) and $b_i \sim N(0, \sigma^2)$.

Fitting Generalized Linear Mixed Models

SAS commands:

```
data toenail;
  infile 'toenail.dat';
  input id y trt month visit;
proc sort; by id visit;
title 'Generalized Linear Mixed Model for Onycholysis Data';
proc nlmixed qpoints=50;
  parms beta0=-0.70 beta1=-0.14 beta2=-0.08
    sigma2=0 to 20 by 2;
  lp = beta0 + beta1*month + beta2*trt*month + u;
  p=(exp(lp))/(1+exp(lp));
  model y~binary(p);
  random u~normal(0, sigma2) subject=id;
run;
```

Parameters				
beta0	beta1	beta2	sigma2	NegLogLike
-0.7	-0.14	-0.08	0	909.881784
-0.7	-0.14	-0.08	2	714.551991
-0.7	-0.14	-0.08	4	692.500709
-0.7	-0.14	-0.08	6	687.431386
-0.7	-0.14	-0.08	8	687.111323
-0.7	-0.14	-0.08	10	688.56761
-0.7	-0.14	-0.08	12	690.76761
-0.7	-0.14	-0.08	14	693.280334
-0.7	-0.14	-0.08	16	695.905441
-0.7	-0.14	-0.08	18	698.543725
-0.7	-0.14	-0.08	20	701.144582

Parameter Estimates									
Parameter	Estimate	Standard Error	DF	t Value	Pr > t	Alpha	Lower	Upper	Gradient
beta0	-1.6972	0.3298	293	-5.15	<.0001	0.05	-2.3463	-1.0481	-2.01E-6
beta1	-0.3885	0.04330	293	-8.97	<.0001	0.05	-0.4737	-0.3033	0.000295
beta2	-0.1424	0.06493	293	-2.19	0.0291	0.05	-0.2702	-0.01464	0.000012
sigma2	16.0349	3.0395	293	5.28	<.0001	0.05	10.0529	22.0169	2.772E-6

Interpretation of Results

- Significant difference in rate of decline of risk for individuals in two treatment groups ($p < 0.05$).
- Over 12 months, odds of onycholysis decreased by factor of 0.01 [or $\exp(-0.389 \times 12)$] for an individual receiving treatment A.
- Over 12 months, odds of onycholysis decreased by factor of 0.002 [$\exp(-0.531 \times 12)$] for an individual receiving treatment B.
- Odds ratio comparing 12 month decreases in risk between groups A & B is approx. 5.5 (or $e^{12 \times 0.142}$).

Summary

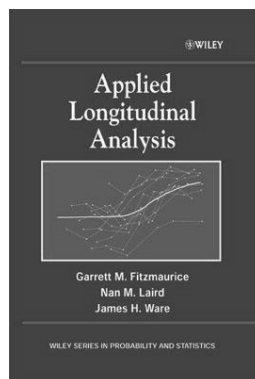
Remarkable developments in methods for longitudinal analysis during 1980's and 1990's.

Although these methods are not yet being routinely applied, the availability of software to fit these models is certainly helping to bridge the gap.

Has software for longitudinal data analysis finally caught up? YES!

Further Reading

Fitzmaurice, G.M., Laird, N.M. and Ware, J.H. (2004).
Applied Longitudinal Analysis. Wiley.



Also, see web site for book: www.biostat.harvard.edu/~fitzmaur/ala

References

- Ashford, J. R. and Sowden, R. R. (1970). Multivariate probit analysis. *Biometrics*, 26, 535-546.
- Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9-25.
- Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72, 320-338.
- Korn, E. L. and Whittemore, A. S. (1979). Methods for analyzing panel studies of acute health effects of air pollution. *Biometrics*, 35, 795-802.
- Laird, N.M. and Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics*, 38, 963-974.
- Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.
- Pierce, D.A. and Sands, B.R. (1975). Extra-Bernoulli variation in binary data. Technical Report 46, Department of Statistics, Oregon State University.
- Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 40, 719-727.
- Stiratelli, R., Laird, N.M. and Ware, J.H. (1984). Random effects models for serial observations with binary response. *Biometrics*, 40, 961-971.
- Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika*, 80, 791-795.