Biostatistics Workshop 2008

$\sim$ Longitudinal Data Analysis $\sim$

Session 2

GARRETT FITZMAURICE

Harvard University

MODELLING LONGITUDINAL DATA

Longitudinal data present two aspects of the data that require modelling:

(i) mean response over time, i.e., model $E(Y_{ij}|X_{ij}) = \mu_{ij}$ over time

(ii) covariance (variances at each occasion and pairwise correlations), i.e., model $\text{Var}(Y_{ij}) = \sigma_j^2$ and $\text{Cov}(Y_{ij}, Y_{ik}) = \sigma_{jk}$.
Note: $\text{Corr}(Y_{ij}, Y_{ik}) = \dfrac{\text{Cov}(Y_{ij}, Y_{ik})}{\sqrt{\text{Var}(Y_{ij})\text{Var}(Y_{ik})}} = \dfrac{\sigma_{jk}}{\sigma_j \sigma_k}$.

Longitudinal models must jointly specify models for mean and covariance.

We review three basic approaches:
(1) Parametric and Semi-Parametric Curves (Session 2)
(2) Covariance Models (Session 3)
(3) Linear Mixed Effects Models (Session 4)

For the next three lectures the focus is on methods for longitudinal data where the response variable is continuous.

The main emphasis is on linear models for longitudinal continuous responses.

Although we discuss likelihood-based analyses for normal responses, the normality assumption can be relaxed without substantial penalty.

Many of the key ideas extend and generalize to discrete outcomes, e.g., binary responses and count data:

$\Longrightarrow$ generalized linear models for longitudinal data (Sessions 5 and 6)

Finally, although the emphasis is on *longitudinal* data; generalizations to *cluster-correlated* data should be apparent (e.g., data from family studies, multilevel/hierarchical data).

MODELLING THE MEAN: PARAMETRIC AND SEMI-PARAMETRIC CURVES

Fitting parametric or semi-parametric curves to longitudinal data can be justified on substantive and statistical grounds.

Substantively, in many studies true underlying mean response process changes over time in a relatively smooth, monotonically increasing/decreasing pattern.

Fitting parsimonious models for mean response results in statistical tests of covariate effects (e.g., treatment $\times$ time interactions) with greater power.

Polynomial Trends in Time

Describe the patterns of change in the mean response over time in terms of simple polynomial trends.

The means are modelled as an explicit function of time.

This approach can handle highly unbalanced designs in a relatively seamless way.

For example, mistimed measurements are easily incorporated in the model for the mean response.

Linear Trends over Time

Simplest possible curve for describing changes in the mean response over time is a straight line.

Slope has direct interpretation in terms of a constant rate of change in mean response for a single unit change in time.

Consider two-group study comparing *treatment* and *control*, where changes in mean response are approximately linear:

$$E\left(Y_{ij}\right) = \beta_0 + \beta_1 \text{Time}_{ij} + \beta_2 \text{Group}_i + \beta_3 \text{Time}_{ij} \times \text{Group}_i,$$

where $\text{Group}_i = 1$ if i^{th} individual assigned to treatment, and $\text{Group}_i = 0$ otherwise; and Time_{ij} denotes measurement time for the j^{th} measurement on i^{th} individual.

Model for the mean for subjects in control group:

$$E\left(Y_{ij}\right) = \beta_0 + \beta_1 \text{Time}_{ij},$$

while for subjects in treatment group,

$$E\left(Y_{ij}\right) = \left(\beta_0 + \beta_2\right) + \left(\beta_1 + \beta_3\right) \text{Time}_{ij}.$$

Thus, each group's mean response is assumed to change linearly over time (see Figure 1).

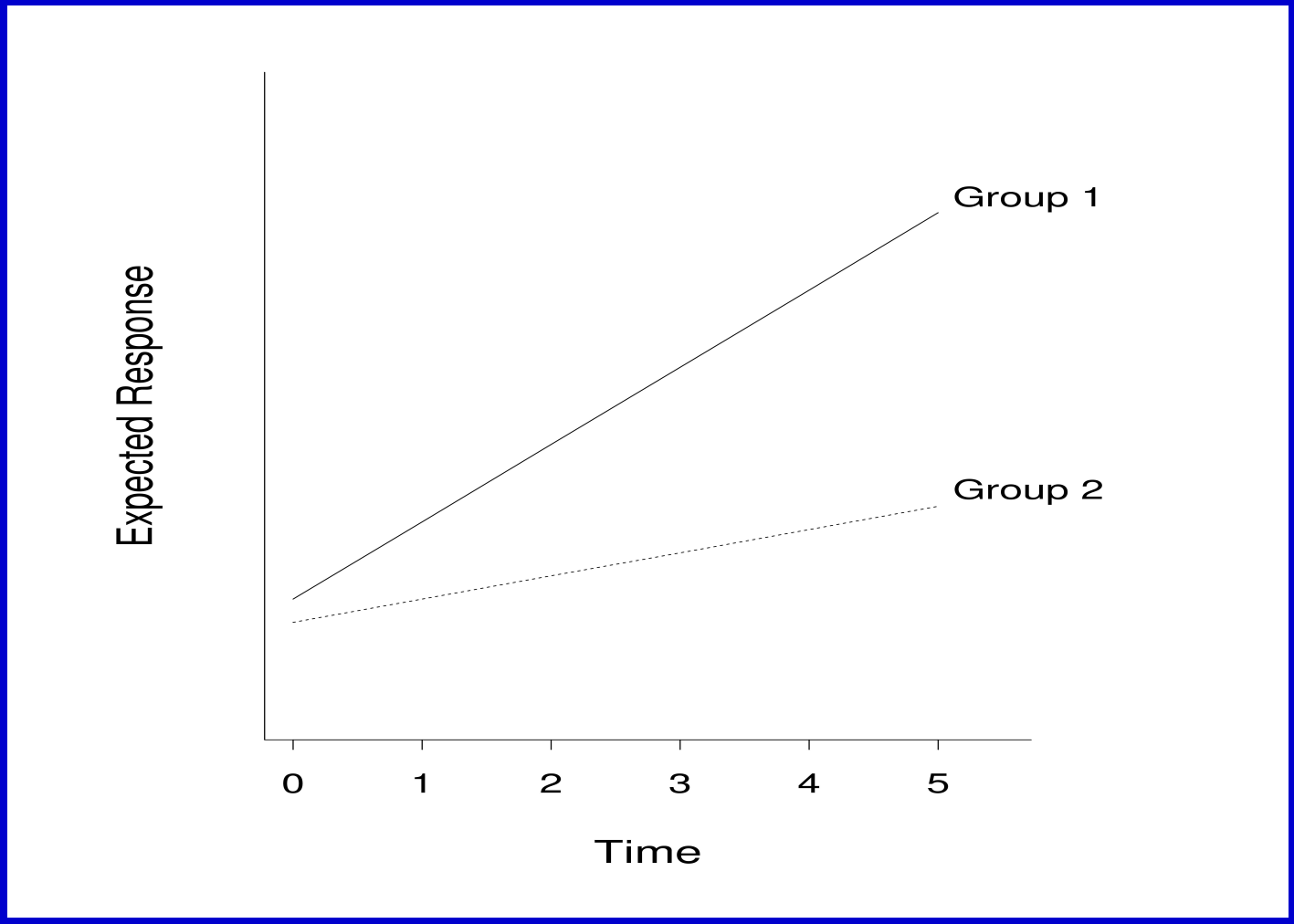

Figure 1: Graphical representation of model with linear trends for two groups.

Quadratic Trends over Time

When changes in the mean response over time are not linear, higher-order polynomial trends can be considered.

For example, if the means are monotonically increasing or decreasing over the course of the study, but in a curvilinear way, a model with quadratic trends can be considered.

In a quadratic trend model the rate of change in the mean response is not constant but depends on time.

Rate of change must be expressed in terms of two parameters.

Consider two-group study example:

$$\begin{aligned} E\left(Y_{ij}\right) &= \beta_0 + \beta_1 \text{Time}_{ij} + \beta_2 \text{Time}_{ij}^2 + \beta_3 \text{Group}_i \\ &\quad + \beta_4 \text{Time}_{ij} \times \text{Group}_i + \beta_5 \text{Time}_{ij}^2 \times \text{Group}_i. \end{aligned}$$

Model for subjects in control group:

$$E\left(Y_{ij}\right) = \beta_0 + \beta_1 \text{Time}_{ij} + \beta_2 \text{Time}_{ij}^2;$$

while model for subjects in treatment group:

$$E\left(Y_{ij}\right) = \left(\beta_0 + \beta_3\right) + \left(\beta_1 + \beta_4\right) \text{Time}_{ij} + \left(\beta_2 + \beta_5\right) \text{Time}_{ij}^2.$$

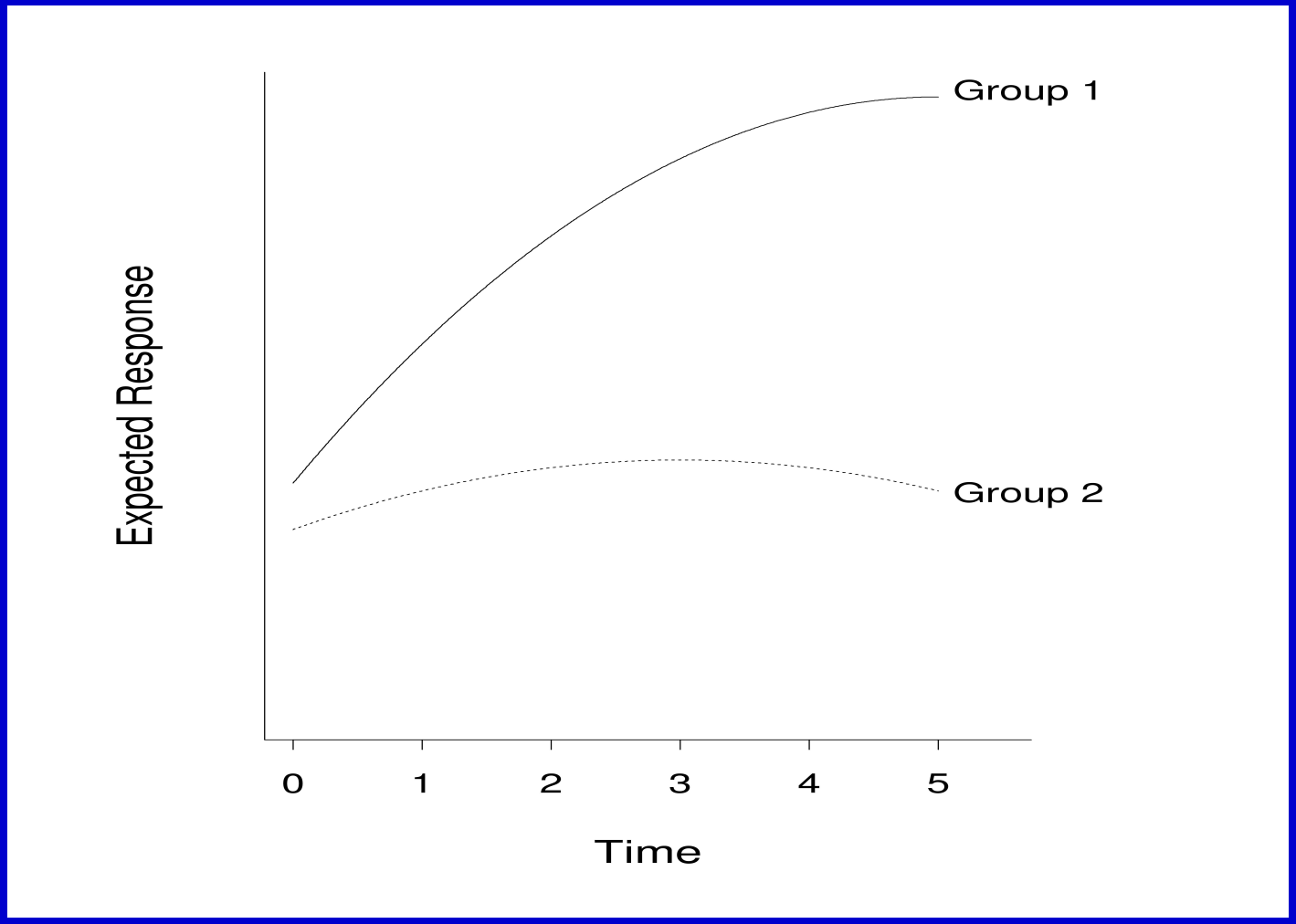

Figure 2: Graphical representation of model with quadratic trends for two groups.

Note: mean response changes at different rate, depending upon Time_{ij}.

Using calculus, instantaneous rate of change in control group is $\beta_1 + 2\beta_2 \text{Time}_{ij}$

Thus, early in the study when $\text{Time}_{ij} = 1$, rate of change is $\beta_1 + 2\beta_2$; while later in the study, say $\text{Time}_{ij} = 4$, rate of change is $\beta_1 + 8\beta_2$.

Regression coefficients, $(\beta_1 + \beta_4)$ and $(\beta_2 + \beta_5)$, have similar interpretations for treatment group.

Linear Splines

If simplest curve is a straight line, then one way to extend the curve is to have sequence of joined line segments that produces a piecewise linear pattern.

Linear spline models provide flexible way to accommodate many non-linear trends that cannot be approximated by simple polynomials in time.

Basic idea: Divide time axis into series of segments and consider piecewise-linear trends, having different slopes but joined at fixed times.

Locations where lines are tied together are known as "knots".

Resulting piecewise-linear curve is called a spline.

Piecewise-linear model often called "broken-stick" model.

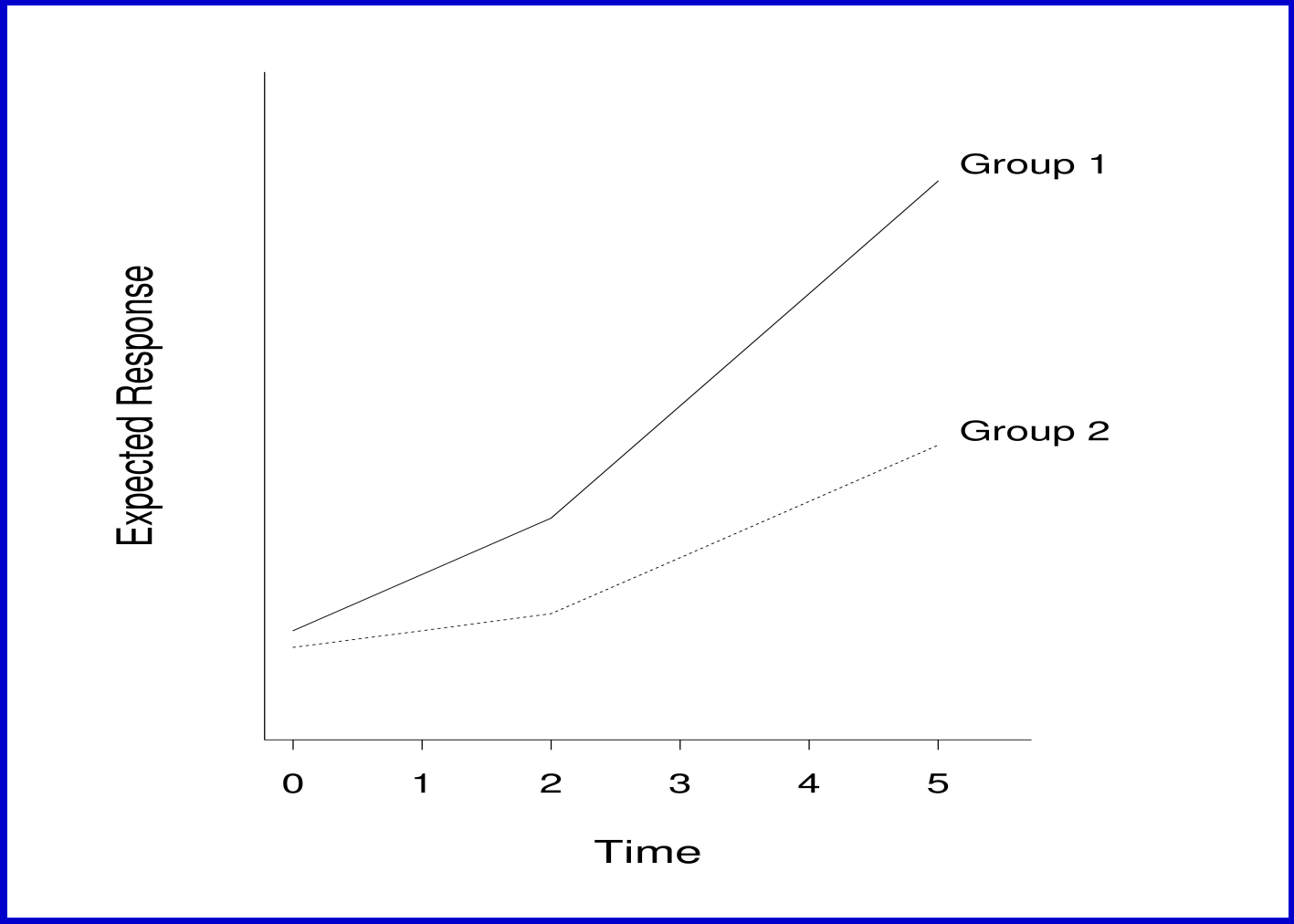

Figure 3: Graphical representation of model with linear splines for two groups, with common knot.

The simplest possible spline model has only one knot.

For two-group example, linear spline model with knot at t^*:

$$\begin{aligned} E(Y_{ij}) &= \beta_0 + \beta_1 \text{Time}_{ij} + \beta_2 (\text{Time}_{ij} - t^*)_+ + \beta_3 \text{Group}_i \\ &\quad + \beta_4 \text{Time}_{ij} \times \text{Group}_i + \beta_5 (\text{Time}_{ij} - t^*)_+ \times \text{Group}_i, \end{aligned}$$

where $(x)_+$ is defined as a function that equals x when x is positive and is equal to zero otherwise.

Thus, $(\text{Time}_{ij} - t^*)_+$ is equal to $(\text{Time}_{ij} - t^*)$ when $\text{Time}_{ij} > t^*$ and is equal to zero when $\text{Time}_{ij} \leq t^*$.

Model for subjects in control group:

$$E\left(Y_{ij}\right) = \beta_0 + \beta_1 \text{Time}_{ij} + \beta_2 (\text{Time}_{ij} - t^*)_+.$$

When expressed in terms of mean response prior/after t^*:

$$\begin{aligned} E\left(Y_{ij}\right) &= \beta_0 + \beta_1 \text{Time}_{ij}, && \text{Time}_{ij} \leq t^*; \\ E\left(Y_{ij}\right) &= (\beta_0 - \beta_2 t^*) + (\beta_1 + \beta_2)\text{Time}_{ij}, && \text{Time}_{ij} > t^*. \end{aligned}$$

Slope prior to t^* is β_1 and following t^* is $(\beta_1 + \beta_2)$.

Model for subjects in treatment group:

$$E\left(Y_{ij}\right) = (\beta_0 + \beta_3) + (\beta_1 + \beta_4)\text{Time}_{ij} + (\beta_2 + \beta_5)(\text{Time}_{ij} - t^*)_+.$$

When expressed in terms of mean response prior/after t^*:

$$\begin{aligned} E\left(Y_{ij}\right) &= (\beta_0 + \beta_3) + (\beta_1 + \beta_4)\text{Time}_{ij}, && \text{Time}_{ij} \leq t^*; \\ E\left(Y_{ij}\right) &= [(\beta_0 + \beta_3) - (\beta_2 + \beta_5)t^*)] \\ &\quad + (\beta_1 + \beta_2 + \beta_4 + \beta_5)\text{Time}_{ij}, && \text{Time}_{ij} > t^*. \end{aligned}$$

Case Study: Vlagtwedde-Vlaardingen Study

Epidemiologic study on prevalence of and risk factors for chronic obstructive lung disease.

Sample participated in follow-up surveys approximately every 3 years for up to 19 years.

Pulmonary function was determined by spirometry: FEV_1.

We focus on a subset of 133 residents aged 36 or older at their entry into the study and whose smoking status did not change over the 19 years of follow-up.

Each study participant was either a current or former smoker.

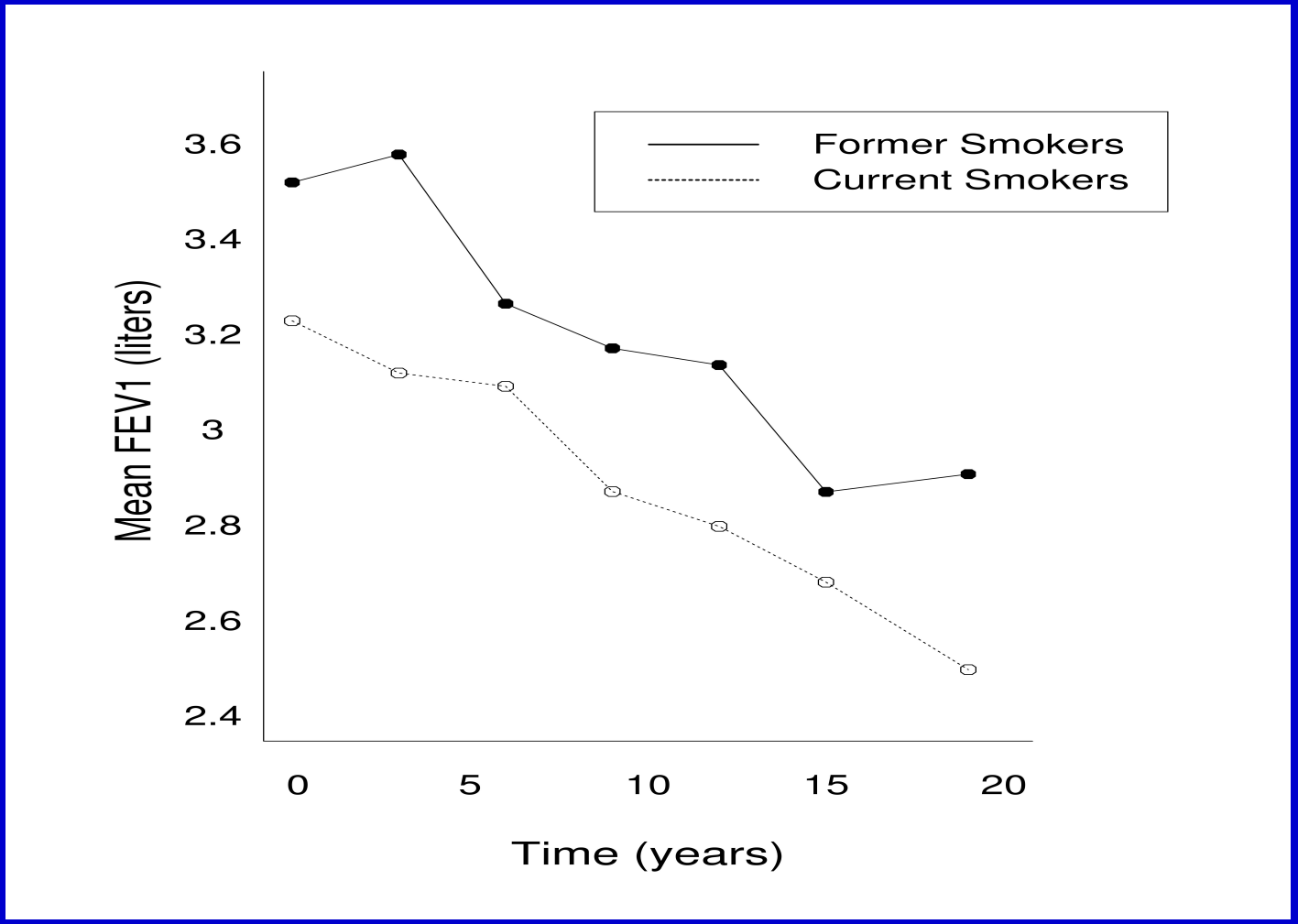

Figure 4: Mean FEV_1 at baseline (year 0), year 3, year 6, year 9, year 12, year 15, and year 19 in the current and former smoking exposure groups.

First we consider a linear trend in the mean response over time, with intercepts and slopes that differ for the two smoking exposure groups.

We assume an unstructured covariance matrix.

Based on the REML estimates of the regression coefficients in Table 1, the mean response for former smokers is

$$E\left(Y_{ij}\right) = 3.507 - 0.033\ \texttt{Time}_{\texttt{ij}},$$

while for current smokers,

$$\begin{aligned} E\left(Y_{ij}\right) &= (3.507 - 0.262) - (0.033 + 0.005)\ \texttt{Time}_{\texttt{ij}} \\ &= 3.245 - 0.038\ \texttt{Time}_{\texttt{ij}}. \end{aligned}$$

Table 1: Estimated regression coefficients for linear trend model for FEV_1 data from the Vlagtwedde-Vlaardingen study.

Variable	Smoking Group	Estimate	SE	Z
Intercept		3.5073	0.1004	34.94
$\texttt{Smoke}_\texttt{i}$	Current	−0.2617	0.1151	−2.27
$\texttt{Time}_\texttt{ij}$		−0.0332	0.0031	−10.84
$\texttt{Smoke}_\texttt{i} \times \texttt{Time}_\texttt{ij}$	Current	−0.0050	0.0035	−1.42

Thus, both groups have a significant decline in mean FEV_1 over time.

But there is no discernible difference between the two smoking exposure groups in the constant rate of change.

That is, the $\texttt{Smoke}_\texttt{i} \times \texttt{Time}_\texttt{ij}$ interaction (i.e., the comparison of the two slopes) is not significant, with $Z = -1.42$, $p > 0.15$.

But is the rate of change constant over time?

Adequacy of linear trend model can be assessed by including higher-order polynomial trends.

For example, we can consider a model that allows quadratic trends for changes in FEV_1 over time.

The maximized log-likelihoods for the models with linear and quadratic trends are presented in Table 2.

LRT comparing quadratic and linear trend models, produces $G^2 = 1.3$, with 2 degrees of freedom ($p > 0.50$).

Thus, when compared to quadratic trend model, linear trend model appears to be adequate.

Note: likelihood ratio test is based on the ML, not REML, log-likelihood.

Table 2: Maximized (ML) log-likelihoods for models with linear and quadratic trends for FEV_1 data from the Vlagtwedde-Vlaardingen study.

Model	-2 (ML) Log-Likelihood
Quadratic Trend Model	237.2
Linear Trend Model	238.5
$-2 \times$ Log-Likelihood Ratio: $G^2 = 1.3,\ 2$ df $\quad (p > 0.50)$	

STATISTICAL SOFTWARE: FITTING LINEAR MODELS TO LONGITUDINAL DATA

SAS and *Stata*, which are widely available, can perform all analyses presented in these lectures.

Alternative software packages, e.g. *SPSS* and *S-PLUS*, can also be used.

Caveat: Statistical software is constantly evolving.

Longitudinal Data Structure

PROC MIXED in SAS and xtmixed in Stata are very general and versatile procedure for fitting linear models to longitudinal and clustered data.

Before discussing specifics of command syntax for PROC MIXED in SAS or xtmixed in Stata, we must discuss the appropriate way to structure the data set.

Software for longitudinal analyses require each repeated measurement in a longitudinal data set to be a separate "record".

For example, in the TLC trial, the data are recorded as follows:

(ID	Group	Baseline	Week 1	Week 4	Week 6)
79	P	30.8	26.9	25.8	23.8
8	A	26.5	14.8	19.5	21.0
44	A	25.8	23.0	19.1	23.2
11	P	24.7	24.5	22.0	22.5
69	A	20.4	2.8	3.2	9.4
29	A	20.4	5.4	4.5	11.9
⋮	⋮	⋮	⋮	⋮	⋮
55	P	31.1	31.2	29.2	30.1

with a single "record" of the 4 repeated measurements for each child in the study.

The data set is in a *multivariate* mode (or "wide format").

Prior to analysis, these data must be converted to a data set with 4 records for each child, one for each measurement occasion.

In the latter form, data set is in a *univariate* mode (or "long format").

This can be accomplished using the illustrative SAS and Stata commands in Tables 3 and 4 which produced the following:

(ID	Group	Time	Y)
79	P	0	30.8
79	P	1	26.9
79	P	4	25.8
79	P	6	23.8
8	A	0	26.5
8	A	1	14.8
8	A	4	19.5
8	A	6	21.0
⋮	⋮	⋮	⋮
55	P	0	31.1
55	P	1	31.2
55	P	4	29.2
55	P	6	30.1

Table 3: Illustrative commands in SAS for transforming data set with single record for each individual to data set with multiple records for each measurement occasion.

```
DATA lead;
   INFILE 'tlc.dat';
   INPUT id group $ y0 y1 y4 y6;
   y=y0; time=0; OUTPUT;
   y=y1; time=1; OUTPUT;
   y=y4; time=4; OUTPUT;
   y=y6; time=6; OUTPUT;
   DROP y1-y4;
```

Table 4: Illustrative commands in Stata for transforming (reshaping) data set with single record for each individual to data set with multiple records for each measurement occasion.

```
. infile id str1 group y0 y1 y4 y6 using tlc.dat
. reshape long y, i(id) j(time)
```

Parametric Curves using PROC MIXED in SAS

Table 5: Illustrative commands for a linear trend model using PROC MIXED in SAS.

```
PROC MIXED;
   CLASS id group t;
   MODEL y=group time group*time / SOLUTION CHISQ;
   REPEATED t / TYPE=UN SUBJECT=id R RCORR;
```

Note that the CLASS statement includes a variable t. This variable is an additional copy of the variable *time*.

The difference is that while t is declared as a categorical variable on the CLASS statement, *time* is not and is treated as a quantitative covariate in the MODEL statement.

It is good practice to include, wherever possible, a REPEATED effect.

This ensures covariance is estimated correctly when the design is balanced but incomplete due to missingness or when repeated measures are not in same order for each subject in data set.

Table 6: Illustrative commands for a quadratic trend model using PROC MIXED in SAS.

```
PROC MIXED;
   CLASS id group t;
   MODEL y=group time timesqr group*time group*timesqr /S CHISQ;
   REPEATED t / TYPE=UN SUBJECT=id R RCORR;
```

Table 7: Illustrative commands for a spline model, with knot at $time = 4$, using PROC MIXED in SAS.

```
PROC MIXED;
   CLASS id group t;
   MODEL y=group time time_4 group*time group*time_4 /S CHISQ;
   REPEATED t / TYPE=UN SUBJECT=id R RCORR;
```

The MODEL statement includes $time$ and $time\_4$, where $time\_4$ is a derived variable for $(time - 4)_+$.

The latter variable can easily be computed in SAS as

$$time\_4 = max(time - 4, 0);$$

FURTHER READING

Diggle, P.J., Heagerty, P., Liang, K-Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data* (2nd ed.). Oxford University Press. (See Chapter 4).

Fitzmaurice, G.M., Laird, N.M. and Ware, J.H. (2004). *Applied Longitudinal Analysis*. Wiley. (See Chapter 6).