Lesson 9
Longitudinal Analysis

BIS 505b

Yale University
Department of Biostatistics

Date Modified: 04/28/2021

Goals for this Lesson

Addressing a Research Question

1. How to recognize longitudinal data
2. How to analyze the change in the mean response over time and determine how the changes relate to covariates

Contents

Progress this Unit

One Observation per Subject

- The methods discussed so far in this course focus on the analysis of one response measurement from each subject only measures once at the end of study Y
- Methods for analyzing these data assume the outcomes represent a set of independent observations

Outcome	Technique	Mathematical Model	Yields
Continuous	Linear Regression	$\mu = \alpha + \beta\, x$	Mean Difference
Binary	Logistic Regression	$\log\left(\frac{p}{1-p}\right) = \alpha + \beta\, x$	Odds Ratio
Count	Poisson Regression	$\log(\mu) = \alpha + \beta\, x$	Mean Ratio
Rate		$\log\left(\frac{\mu}{t}\right) = \alpha + \beta\, x$	Rate Ratio
Time-to-event	Cox Regression	$\log(h(t;x)) = \log(h_0(t)) + \beta\, x$	Hazard Ratio

Multiple Observations per Subject over Time

- In many settings, we are interested in studying patients over a period of time
- Longitudinal data occur when the subjects are studied over a period of time and multiple outcome measurements (over time) are recorded for each subject
  - Measuring subjects at multiple time points allows for research questions that involve comparisons between groups and study of changes over time
- When subjects have an outcome measured multiple times, the observations for each subject are not independent (i.e., the measurements are correlated)

coz multiple outcomes come form a same subject

Longitudinal vs. Cross-Sectional Data

- In a cross-sectional study design, the response is measured at a single occasion
  - Can only obtain estimates of between-individual differences in the response
  - Does not provide any information about how individuals change during the period
- In a longitudinal study design,
  - Can capture within-individual changes to explore how individuals change over a period of time
  - Can determine if these within-individual changes in the response are related to selected covariates (e.g., exposures)

Menarche and Body Fat (MBF): Example

- Body fat percentage in girls is thought to increase just before or around menarche leveling off approximately 4 years after menarche
- Suppose we are interested in modeling the increase in body fat percentage in girls after menarche

MBF in a *Cross-Sectional* Study Design

- Collect percent body fat on **two** separate groups of girls:
  1. 10-year old girls (pre-menarcheal 月经初潮 cohort) and
  2. 15-year old girls (post-menarcheal cohort)
- Compare average percent body fat in two groups using two-sample (unpaired) t-test
  - Does not provide an estimate of the change in body fat percentage as girls age from 10 to 15 years
  - The effect of growth, or aging, is an inherently within-individual effect
  - Cannot be estimated from a cross-sectional study that does not measure how individuals change with time
  - Effect of aging is potentially confounded with possible cohort effects

MBF in a *Longitudinal* Study Design

- Measure a **single** cohort of girls
  1a. Age 10
  1b. Age 15
- The analysis is based on a paired t-test, using the difference or change in percent body fat of each girl as the outcome variable
  - This within-individual comparison provides a valid estimate of the change in body fat as girls age from 10 to 15 years

t-Test Review

- Recall that a paired t-test was used to compare means of two *dependent* groups
  - Pre-intervention vs. post-intervention
  - $d_i = y_{1i} - y_{2i}$
  - $t = \dfrac{\bar{d}}{s_{\bar{d}}}$

Accounts for correlation between repeated measures

$s_{\bar{d}} = \sqrt{\text{Var}(Y_1 - Y_2)}$

- $\text{Var}(Y_1 - Y_2) = \text{Var}(Y_1) + \text{Var}(Y_2) - 2\,\text{Cov}(Y_1, Y_2)$

covariance

  - If $\text{Cov}(Y_1, Y_2) > 0$ (i.e., Y_1 and Y_2 are positively correlated), pairing is beneficial: lower variance than for the unpaired case
  - Reduces variance, more powerful test — harder to reject H_0
  - Two-sample unpaired t-test assumes independent groups, $\text{Cov}(Y_1, Y_2) = 0$

Modeling Considerations

- With longitudinal data, the usual assumptions for standard regression analysis do not hold
- Two aspects of longitudinal data that complicate their statistical analysis:
  1. Repeated measures on the same individual are usually positively correlated: violates independence
  2. The variance of the repeated measurements usually is not constant over the follow-up period (heterogeneous variance across measurement occasions) (e.g., the variance of baseline measurements is often smaller than post-baseline measurements): violates homoscedasticity

Correlation

- There may be patterns to these correlations
  - For example, a pair of repeated measures that have been obtained close together in time are expected to be more highly correlated than a pair of repeated measures further separated in time $t_1 \& t_2$ vs $t_1 \& t_{100}$
- In general, the correlation among repeated measures is expected to decline with increasing time separation

Notation

- Let Y_{ij} denote the continuous response variable for the i^{th} individual $(i = 1, \ldots, n)$ at the j^{th} occasion $(j = 1, \ldots, m)$
- For now, assume all subjects have the same number of repeated measurements (m) at the same set of occasions

$$\mathbf{y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{im} \end{pmatrix} \quad \in R^m$$

- Balanced longitudinal data: Subjects measured at the same set of occasions (roughly the same points in time)

Notation

Table: Longitudinal Data Structure

	Occasion (j)				
Individual (i)	1	2	3	...	m
1	y_{11}	y_{12}	y_{13}	...	y_{1m}
2	y_{21}	y_{22}	y_{23}	...	y_{2m}
$\vdots$					
n	y_{n1}	y_{n2}	y_{n3}	...	y_{nm}

- Interested in modeling the mean response over time and how the mean depends on covariates

Correlation between Repeated Measures

Covariance between measurements at time j and time k:

$$Cov(Y_j, Y_k) = \sigma_{jk} = \frac{1}{N}\sum_{i=1}^{N}(y_{ij} - \mu_j)(y_{ik} - \mu_k)$$

- When covariance equals 0, there is no correlation between the responses at the two occasions

Correlation between measurements at time j and time k:

$$Corr(Y_j, Y_k) = \rho_{jk} = \frac{\sigma_{jk}}{\sigma_j \sigma_k}$$

Correlation between Repeated Measures

- Can define the covariance and correlation matrix between the pairs of repeated measures:

covariance matrix

$$Cov\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{im} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1m} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \dots & \sigma_m^2 \end{pmatrix}$$

correlation matrix

$$Corr\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{im} \end{pmatrix} = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1m} \\ \rho_{21} & 1 & \dots & \rho_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{m1} & \rho_{m2} & \dots & 1 \end{pmatrix}$$

Correlation between Repeated Measures

$$Cov\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{im} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1m} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \dots & \sigma_m^2 \end{pmatrix}$$

$\neq 0$

- With longitudinal data, heterogeneity of the variance over time can be accounted for by allowing the elements on the main diagonal of the covariance matrix to differ
- The lack of independence among repeated measures is accounted for by allowing the off-diagonal elements of the covariance matrix to be non-zero
- Longitudinal analyses account for within-subject correlation between repeated measurements over time

Why Standard Methods are Not Appropriate

- Standard regression models assume that all observations are independent and, if applied to longitudinal data, may produce invalid standard errors
- Correlation among repeated measures is a feature of longitudinal data that must be accounted for in the analysis in order to make appropriate inferences
  - Correlation among longitudinal data enables us to estimate changes in the mean response, and their relation to covariates, with more precision than would be possible if the data were uncorrelated
  - Important to take this correlation into account in the analysis, although correlation not usually of interest (not main focus of the analysis)
  - Main interest in any longitudinal study is in describing changes in the mean response over time, and how these changes are related to covariates of interest

Treatment of Lead-Exposed Children (TLC) Trial: Example

- Example: Exposure to lead during infancy is associated with substantial deficits in tests of cognitive ability. CDC has concluded that children with blood lead levels above 10 micrograms per deciliter are at risk of adverse health effects.
- Chelation treatment usually requires injections and hospitalization
- Succimer is chelating agent that enhances urinary excretion of lead; given orally
- Placebo-controlled randomized trial examined changes in blood lead level in children during course of follow-up; baseline blood lead levels of 20-44 μg/dL
  - Mean age = 2 years; Mean blood lead levels = 26 μg/dL
  - 100 children randomized to placebo or Succimer
  - Measures of blood lead level at baseline, 1, 4, and 6 weeks

TLC Trial: Example

R Code

baseline ↑ Week 1 4 6

```
> head(tlc)
  id       trt   y0   y1   y4   y6
1  1   Placebo 30.8 26.9 25.8 23.8
2  2 Treatment 26.5 14.8 19.5 21.0
3  3 Treatment 25.8 23.0 19.1 23.2
4  4   Placebo 24.7 24.5 22.0 22.5
5  5 Treatment 20.4  2.8  3.2  9.4
6  6 Treatment 20.4  5.4  4.5 11.9

# Scatterplot matrix
> pairs(tlc[tlc$trt=="Placebo", 3:6])

# Correlation matrix
> cor(tlc[tlc$trt=="Placebo", 3:6])
```

- Blood lead levels (μg/dL) at baseline, week 1, week 4, and week 6
- Data are in wide format make plot easy
- TLC Trial is a balanced data set
  - 4 measurements per subject
  - Measurements taken at exactly the same time points

TLC Trial: Example

Figure: Scatterplot Matrix - Placebo Group

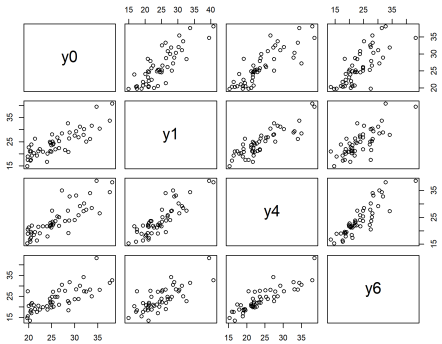


Table: Correlation Matrix - Placebo Group

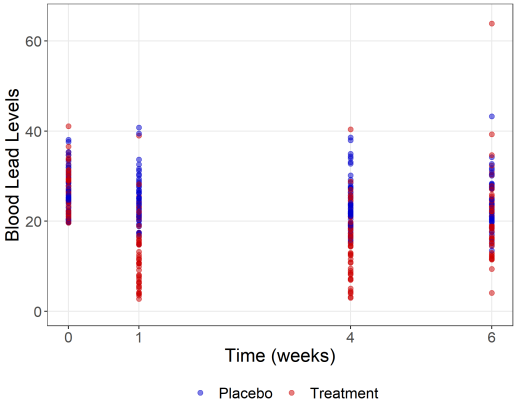
	Time 0	Time 1	Time 4	Time 6
Time 0	1	0.829	0.839	0.756
Time 1	0.829	1	0.861	0.759
Time 4	0.839	0.861	1	0.870
Time 6	0.756	0.759	0.870	1

- Examination of the correlations confirms that they are all positive and tend to decrease with increasing time separation

Progress this Unit

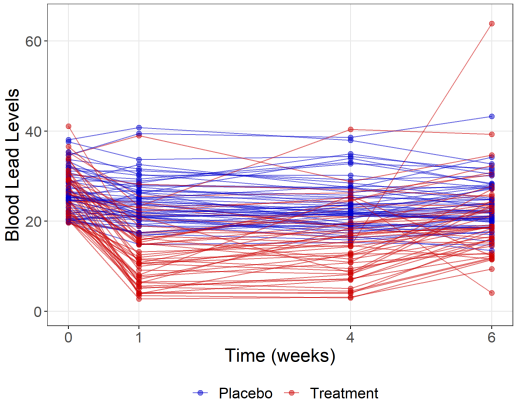
Time Plot

- Time Plot: A scatterplot with the responses on the y-axis and measurement times on the x-axis
- Not always helpful or interpretable –
  1. Many overlapping data points. Does
  2. not indicate which measurements are from the same individual.



Spaghetti Plot

- Spaghetti Plot: Time plot where successive repeated measurements on the same individual are joined with straight lines
- Difficult to discern the "signal" (trend in the mean response over time) from the noise in the data



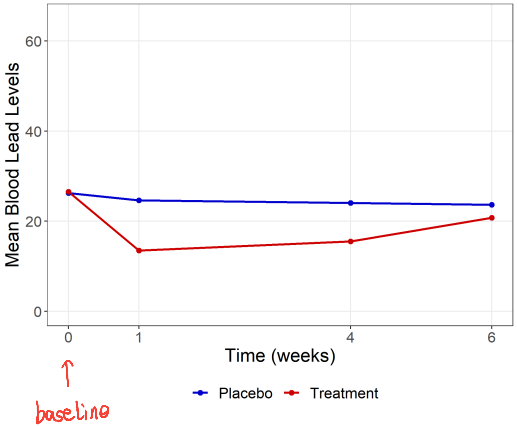
Mean Response Profile

good plot

- Mean Response Profile: Time plot of the average or mean response with successive points on the graph joined by straight lines

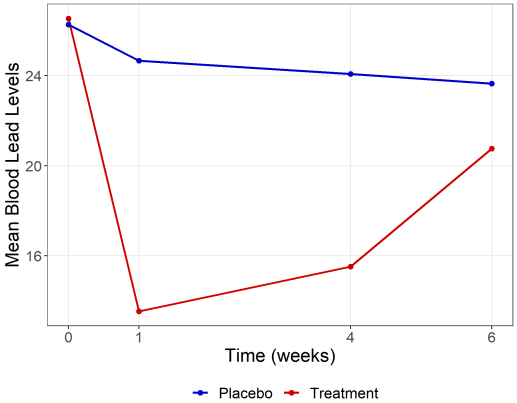
Table: Mean Blood Lead Levels (SD)

Group	Baseline	Week 1	Week 4	Week 6
Succimer	26.5	13.5	15.5	20.8
	(5.0)	(7.7)	(7.8)	(9.2)
Placebo	26.3	24.7	24.1	23.6
	(5.0)	(5.5)	(5.7)	(5.6)



TLC Trial: Example

- Graphical display of the mean response can provide basis for choosing an appropriate model for the analysis of change over time
- Levels in placebo arm relatively flat over time
- Large initial drop in drug arm; Rebound in blood levels thought due to lead stored in bones being mobilized



Modeling Longitudinal Data

- Regression methods permit inference about the average response trajectory over time and how changes are related to patient characteristics (covariates)
- In the TLC Trial, the investigators are interested in how blood lead levels changed over time and whether these changes were related to the treatment assigned
- Treatment group interaction effects with time have direct interpretation in terms of how the underlying rate of change in mean blood lead levels differs between the two treatment groups

Modeling Longitudinal Data

- We consider two approaches for modeling the mean:
  1. Analysis of response profiles: Compare two groups in terms of all post-baseline changes in mean blood levels from baseline
  2. Modeling a linear trend over time: Compare two treatments in terms of the rate of decline in blood levels over time (rate = slope)

problematic sometimes

Analysis of Response Profiles

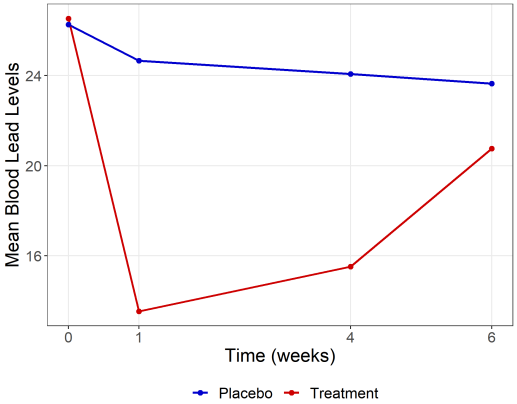
- Basic idea: Compare groups of subjects in terms of mean response profiles over time
- Data can be summarized by the mean response profile when all individuals are measured at the same set of occasions and the number of occasions is small
- When no specific a priori pattern for the differences in the response profiles between groups can be specified
  - Allows arbitrary patterns in the mean response over time (no specific time trend is assumed)
  - Measurement times are regarded as levels of a factor variable

Analysis of Response Profiles

- Analysis of response profiles can be extended to handle more than a single group factor
- Analysis of response profiles can also handle missing data
- Goal: Characterize the patterns of change in the mean response over time in the groups to determine whether the shapes of the mean response profiles differ among the groups

TLC Trial: Example

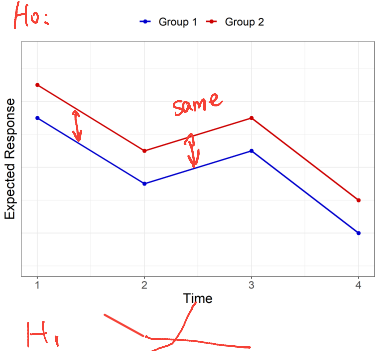
- In the TLC trial, the major question of scientific interest is whether changes in the mean blood lead levels are the same for the treatment and placebo groups
- Translates into a hypothesis about the interaction between the group factor and time



Hypotheses Concerning Response Profiles

- Given a sequence of m repeated measures, and a group/exposure of interest, there are three main questions: 3Q

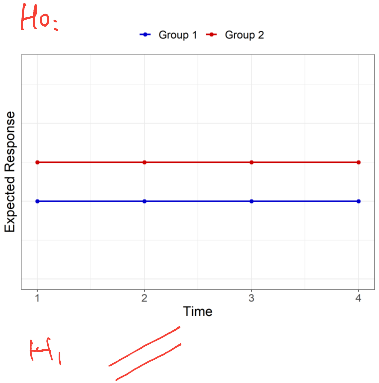
1. Does the treatment effect differ by group over time?



- This question looks at the group $\times$ time interaction
- Are the mean response profiles similar in the groups, in the sense that the mean response profiles are parallel?
- Graphical representation of the null hypothesis of parallel mean response profiles
- Effect of group constant over time

Hypotheses Concerning Response Profiles

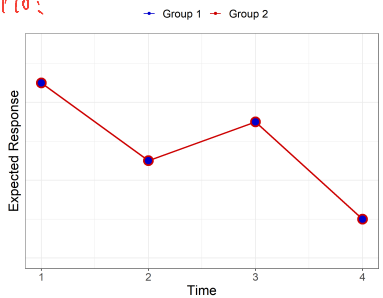
2. Assuming that the population mean response profiles are parallel, is there a difference in the mean response over time?



- This question looks at the time effect
- Assuming mean response profiles are parallel, are the means constant over time, in the sense that the mean response profiles are flat?
- Graphical representation of the null hypothesis of no effect of time

Hypotheses Concerning Response Profiles

3. Assuming that the population mean response profiles are parallel, is there a difference in the mean response in the two groups?



- This question looks at the group effect
- Assuming that the population mean response profiles are parallel, are they also at the same level in the sense that the mean response profiles for the groups coincide?
- Graphical representation of the null hypothesis that the mean response profiles are at the same level

General Linear Model Formulation

- The main focus of analysis looking at the effect of treatment or exposure is a test of the null hypothesis that the mean response profiles are not different over time
- In testing this hypothesis, both group and time are regarded as categorical variables
- The analysis of response profiles can be specified as a regression model with "indicator variables" for group and time
- Unlike standard regression, the correlation and variability among repeated measures on the same individuals must be properly accounted for

Longitudinal Data Structure: Example

R Code

```
> head(tlc)
  id       trt   y0   y1   y4   y6
1  1   Placebo 30.8 26.9 25.8 23.8
2  2 Treatment 26.5 14.8 19.5 21.0
3  3 Treatment 25.8 23.0 19.1 23.2
4  4   Placebo 24.7 24.5 22.0 22.5
5  5 Treatment 20.4  2.8  3.2  9.4
6  6 Treatment 20.4  5.4  4.5 11.9
```

- Blood lead levels (μg/dL) at baseline, week 1, week 4, and week 6
- Data are in wide format
- TLC Trial is a balanced data set

Longitudinal Data Structure: Example

R Code: Wide to Long Transformation

```
# Convert wide-formatted data into long
> tlclong <- reshape(tlc,
            varying = c("y0", "y1", "y4", "y6"),
            v.names = "lead",       # new y-variable name
            timevar = "time.num",  # new time variable name
            times = 1:4,     # consecutive integers for model
            idvar = "id",           # subject id in tlc
            direction = "long")
> tlclong$week[tlclong$time.num == 1] <- 0
> tlclong$week[tlclong$time.num == 2] <- 1
> tlclong$week[tlclong$time.num == 3] <- 4
> tlclong$week[tlclong$time.num == 4] <- 6
> tlclong$week_factor <- factor(tlclong$week) # factor week
# Sort data by id then by time
> tlclong <- tlclong[order(tlclong$id, tlclong$time.num),]
# Clear row names
> row.names(tlclong) <- NULL
```

R Code: Long Data

```
> print(head(tlclong, n = 10), row.names = F)
 id       trt time.num lead week week_factor
  1   Placebo        1 30.8    0           0
  1   Placebo        2 26.9    1           1
  1   Placebo        3 25.8    4           4
  1   Placebo        4 23.8    6           6
  2 Treatment        1 26.5    0           0
  2 Treatment        2 14.8    1           1
  2 Treatment        3 19.5    4           4
  2 Treatment        4 21.0    6           6
  3 Treatment        1 25.8    0           0
  3 Treatment        2 23.0    1           1
```

- For analysis, data must be in "long" format; will contain 4 records for each child, one for each measurement occasion

Response Profile Model Formulation

- Response profiles are modeled using the general linear model,

$$E(Y|x) = \mu_{y|x} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots$$

- A design with 2 groups and m repeated measures will require $2 \times m$ parameters for the 2 response profiles
  - For example, in the TLC trial, there are 2 groups measured on $m = 4$ occasions, so there are 8 mean parameters $(\mu_0(P), \mu_1(P), \ldots, \mu_6(S))$
- Tests of the group $\times$ time interaction and the main effects of time and group are possible once the covariance of Y_i has been specified
- Covariance of Y_i usually assumed to be unstructured (no constraints, each variance and covariance estimated from the data)

$$Cov \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{pmatrix}$$

Response Profile Model Formulation: Example

- Use indicator variables for treatment group and time:
  - Group: $x_1 = \begin{cases} 1 & \text{if child randomized to Succimer treatment} \\ 0 & \text{if child randomized to placebo} \end{cases}$
  - Time:

$$x_2 = \begin{cases} 1 & \text{if measurement at week 1} \\ 0 & \text{otherwise} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{if measurement at week 4} \\ 0 & \text{otherwise} \end{cases}$$

$$x_4 = \begin{cases} 1 & \text{if measurement at week 6} \\ 0 & \text{otherwise} \end{cases}$$

- Placebo and baseline (week 0) are the reference groups

Response Profile Model Formulation: Example

- Analysis of response profiles model can be expressed as:

$$\begin{aligned}\text{Intercept} &+ \text{Main effect of Group} \\ &+ \text{Main effect of Time} \\ &+ \text{Interaction of Group} \times \text{Time}\end{aligned}$$

$$\begin{aligned}y = \alpha &+ \beta_1 x_1 \\ &+ \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \\ &+ \beta_5 x_1\, x_2 + \beta_6 x_1\, x_3 + \beta_7 x_1\, x_4 + \epsilon\end{aligned}$$

$$\begin{aligned}y = \alpha &+ \beta_1\ \mathrm{S} \\ &+ \beta_2\ \mathrm{W1} + \beta_3\ \mathrm{W4} + \beta_4\ \mathrm{W6} \\ &+ \beta_5\ \mathrm{S\ W1} + \beta_6\ \mathrm{S\ W4} + \beta_7\ \mathrm{S\ W6} + \epsilon\end{aligned}$$

Response Profile Model Formulation: Example

$$E(Y) = \alpha + \beta_1 \text{ S} + \beta_2 \text{ W1} + \beta_3 \text{ W4} + \beta_4 \text{ W6} + \beta_5 \text{ S W1} + \beta_6 \text{ S W4} + \beta_7 \text{ S W6}$$

S=0

- Mean response for placebo group is:

$$\mu(P) = \begin{pmatrix} \mu_0(P) \\ \mu_1(P) \\ \mu_4(P) \\ \mu_6(P) \end{pmatrix} = \begin{pmatrix} \alpha \\ \alpha + \beta_2 \\ \alpha + \beta_3 \\ \alpha + \beta_4 \end{pmatrix}$$

S=1

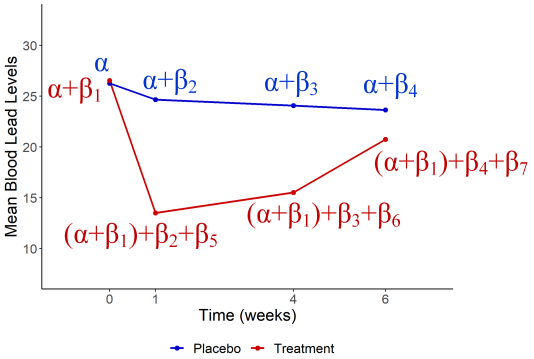
- Mean response for Succimer group is:

$$\mu(S) = \begin{pmatrix} \mu_0(S) \\ \mu_1(S) \\ \mu_4(S) \\ \mu_6(S) \end{pmatrix} = \begin{pmatrix} (\alpha + \beta_1) \\ (\alpha + \beta_1) + \beta_2 + \beta_5 \\ (\alpha + \beta_1) + \beta_3 + \beta_6 \\ (\alpha + \beta_1) + \beta_4 + \beta_7 \end{pmatrix}$$

- Test of group $\times$ time interaction: $H_0 : \beta_5 = \beta_6 = \beta_7 = 0$
- Test of no time effect: $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ in a model with no interaction
- Test of no group effect: $H_0 : \beta_1 = 0$ in a model with no interaction

Response Profile Model Parameters

$$E(Y) = \alpha + \beta_1 \text{ S} + \beta_2 \text{ W1} + \beta_3 \text{ W4} + \beta_4 \text{ W6} + \beta_5 \text{ S W1} + \beta_6 \text{ S W4} + \beta_7 \text{ S W6}$$



- If $\beta_5 = \beta_6 = \beta_7 = 0$, then trajectories would be parallel
- β_1: Main effect of group
- $\beta_2, \beta_3, \beta_4$: Main effects of time

Response Profile Analysis: Example

R Code: Mean Response Profile

```
# a*b includes a + b + a:b (interaction of a and b)
> mod.mrp <- gls(lead ~ trt*week_factor, data = tlclong,
                 # Covariance structure: unstructured
                 correlation = corSymm(form = ~ time.num | id),
                 # Variance structure: different for each time point
                 weights = varIdent(form = ~ 1 | week_factor))
```

correlation differ over time
within subject id

- `gls()` fits analogous models to `lm()`, but can account for correlation structure in the data
- The analysis of response profiles assumes an unstructured variance-covariance matrix and estimates separate variances for each occasion (4 variances) and six pairwise covariances
  - `correlation = corSymm()` specifies form of covariance matrix (unstructured); correlation between observations differs over time (`time.num`: numerical sequence of consecutive integers, whereas factor version of time in the model); observations correlated within same subject id
  - `weights = varIdent()` allows for heterogeneity in variance at each time point (allows residual variance to differ)

Response Profile Analysis: Example

R Code: Mean Response Profile

```
# Summary of model fit
> summary(mod.mrp)
# F-test of interaction
> anova(mod.mrp, type = "marginal")
Denom. DF: 392
                numDF   F-value p-value
(Intercept)         1 1368.0793  <.0001
trt                 1    0.0712  0.7898
week_factor         3    3.8731  0.0095
trt:week_factor     3   35.9293  <.0001
```

R Code: Estimated Covariance Matrix

```
# Estimated variance-covariance matrix
> getVarCov(mod.mrp)
Marginal variance covariance matrix
       [,1]   [,2]   [,3]   [,4]
[1,] 25.226 19.107 19.700 22.202
[2,] 19.107 44.346 35.535 29.675
[3,] 19.700 35.535 47.377 30.620
[4,] 22.202 29.675 30.620 58.651
```

variance

- Test of group $\times$ time interaction: $H_0 : \beta_6 = \beta_7 = \beta_8 = 0$
  - F-test of the group $\times$ time interaction: $p < 0.0001$
  - Reject H_0 and conclude the patterns of change from baseline are not the same in the two groups
  - Because this is a global test, it indicates that groups differ but does not tell us how they differ

Response Profile Interpretation of Slopes: Example

Table: Estimated regression coefficients based on analysis of response profiles

Variable	Group	Week		Estimate	SE	t	p-value
Intercept			a	26.272	0.710	36.99	$<.0001$
Group	S		b_1	0.268	1.005	0.27	0.79
Week		1	b_2	-1.612	0.792	-2.04	0.04
Week		4	b_3	-2.202	0.815	-2.70	0.008
Week		6	b_4	-2.626	0.889	-2.96	0.004
Group $\times$ Week	S	1	b_5	-11.406	1.120	-10.18	$<.0001$
Group $\times$ Week	S	4	b_6	-8.824	1.153	-7.66	$<.0001$
Group $\times$ Week	S	6	b_7	-3.152	1.257	-2.51	0.01

- Children treated with Succimer have greater decline in mean blood lead levels from baseline at all occasions compared to children treated with placebo

Response Profile Interpretation of Slopes: Example

Variable	Group	Week		Estimate
Intercept			a	26.272
Group	S		b_1	0.268
Week		1	b_2	-1.612
Week		4	b_3	-2.202
Week		6	b_4	-2.626
Group × Week	S	1	b_5	-11.406
Group × Week	S	4	b_6	-8.824
Group × Week	S	6	b_7	-3.152

$$\begin{pmatrix} \mu_0(P) \\ \mu_1(P) \\ \mu_4(P) \\ \mu_6(P) \end{pmatrix} = \begin{pmatrix} \alpha \\ \alpha + \beta_2 \\ \alpha + \beta_3 \\ \alpha + \beta_4 \end{pmatrix}$$

$$\begin{pmatrix} \mu_0(S) \\ \mu_1(S) \\ \mu_4(S) \\ \mu_6(S) \end{pmatrix} = \begin{pmatrix} (\alpha + \beta_1) \\ (\alpha + \beta_1) + \beta_2 + \beta_5 \\ (\alpha + \beta_1) + \beta_3 + \beta_6 \\ (\alpha + \beta_1) + \beta_4 + \beta_7 \end{pmatrix}$$

- The 6-week change from baseline in the Succimer group is estimated to be 3.152 units less than the 6-week change from baseline in placebo
  - Succimer, $\mu_6(S) - \mu_0(S) = [\alpha + \beta_1 + \beta_4 + \beta_7] - [\alpha + \beta_1] = \beta_4 + \beta_7$
  - Placebo, $\mu_6(P) - \mu_0(P) = [\alpha + \beta_4] - [\alpha] = \beta_4$
  - Succimer − Placebo, $\beta_4 + \beta_7 - \beta_4 = \boxed{\beta_7}$

Response Profile Analysis: Raw Means vs. Fitted Means

Table: From the raw means

Time	Succimer	Placebo	Succimer: Change from baseline	Placebo: Change from baseline	Succimer vs. Placebo
0	26.540	26.272	.	.	.
1	13.522	24.660	-13.018	-1.612	-11.406 $= b_5$
4	15.514	24.070	-11.026	-2.202	-8.824 $= b_6$
6	20.762	23.646	-5.778	-2.626	-3.152 $= b_7$

Table: From the model

$$\begin{pmatrix} \hat{\mu}_0(P) \\ \hat{\mu}_1(P) \\ \hat{\mu}_4(P) \\ \hat{\mu}_6(P) \end{pmatrix} = \begin{pmatrix} a \\ a+b_2 \\ a+b_3 \\ a+b_4 \end{pmatrix} = \begin{pmatrix} 26.272 \\ 26.272 - 1.612 = 24.660 \\ 26.272 - 2.202 = 24.070 \\ 26.272 - 2.626 = 23.646 \end{pmatrix}$$

- Fitted means identical to sample means (saturated model: 8 parameters to estimate 8 means)

Response Profile Analysis: Raw Means vs. Fitted Means

R Code: Raw Means vs. Fitted Means

```
# Raw means
> rawmeans <- aggregate(cbind(y0, y1, y4, y6) ~ trt, data = tlc, FUN = mean, na.rm = TRUE)
> print(rawmeans, row.names = FALSE)
      trt     y0     y1     y4     y6
   Placebo 26.272 24.660 24.070 23.646
 Treatment 26.540 13.522 15.514 20.762
> pred.x <- expand.grid(week_factor = levels(tlclong$week_factor), trt = unique(tlclong$trt))
# Predicted means
> pred.x$predicted <- predict(mod.mrp, newdata = pred.x, type = "response")
> print(pred.x, digits = 5, row.names = FALSE)
 week_factor       trt predicted
           0   Placebo    26.272
           1   Placebo    24.660
           4   Placebo    24.070
           6   Placebo    23.646
           0 Treatment    26.540
           1 Treatment    13.522
           4 Treatment    15.514
           6 Treatment    20.762
```

Analysis of Response Profiles

- Strengths
  - Allows arbitrary patterns in the mean response over time (no time trend assumed) and arbitrary patterns in the covariance
  - Analysis has a certain robustness since potential risks of bias due to misspecification of models for mean and covariance are minimal
  - Can accommodate an arbitrary pattern of missingness
- Drawbacks
  - Requires balanced longitudinal design (cannot incorporate mistimed measurements)
  - Analysis ignores the time-ordering (time trends) of the repeated measures in a longitudinal study

Parametric Curve

- Basic idea: Assume a parametric curve (e.g., linear or quadratic trend) for the mean response over time
- Can greatly reduce the number of model parameters
- Provides a parsimonious description of trends in the mean response over time and of the covariate effects on the mean response over time
  - A linear trend in the mean response can be characterized by 1 parameter (slope)
  - Interpreted as a constant rate of change in the mean response over time

Parametric Curve

- 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 than in analysis of response profiles

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

$$E(Y) = \alpha + \beta_1 \text{ Time} + \beta_2 \text{ Group} + \beta_3 \text{ Time} \times \text{Group}$$

continuous variable

- $\text{Group} = \begin{cases} 1 & \text{if assigned to treatment} \\ 0 & \text{if assigned to control} \end{cases}$

Linear Trends over Time

$$E(Y) = \alpha + \beta_1 \text{ Time} + \beta_2 \text{ Group} + \beta_3 \text{ Time} \times \text{Group}$$

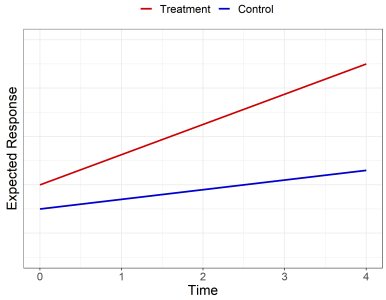
- Mean response for the control group:

$$E(Y) = \alpha + \beta_1 \text{Time}$$

- Mean response for the treatment group:

$$E(Y) = (\alpha + \beta_2) + (\beta_1 + \beta_3)\text{Time}$$

- Each group's mean response is assumed to change linearly over time



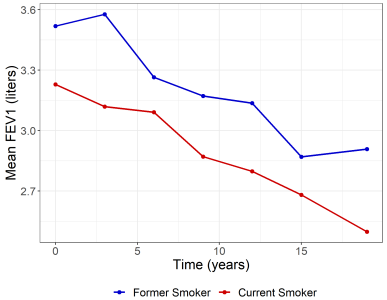
Vlagtwedde-Vlaardingen (COPD) Study: Example

- 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 21 years
- Pulmonary function was determined by spirometry: FEV1
- 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

COPD Study: Example

Table: Mean FEV1 (N)

Group	Baseline	Year 3	Year 6	Year 9	Year 12	Year 15	Year 19
Former Smoker	3.519	3.578	3.265	3.172	3.136	2.870	2.908
	23	27	28	30	29	24	28
Current Smoker	3.229	3.119	3.091	2.871	2.798	2.681	2.498
	85	95	89	85	81	73	74



- Not all subjects are measured at all time points
- Analysis goal: Describe changes in lung function over the 19-year follow-up and determine whether the time trends differ for current and former smokers
- Consider a linear trend in the mean response over time

Linear Trend Analysis: Example

R Code: Linear Trend

```
# Quantitative "time" variable
> mod.lin <- gls(fev1 ~ smoker_factor*time, data = smoke,
                 # Covariance structure: unstructured
                 correlation = corSymm(form = ~ time.num | id),
                 # Variance structure: different for each time point
                 weights = varIdent(form = ~ 1 | time))
```

- "Long" data structure required
- Again requires numeric time sequence of consecutive integers for time variable in covariance structure (`time.num`)
- Quantitative `time` used in model (not factor version as in mean response profile analysis)

Linear Trend Interpretation of Slopes: Example

Table: Estimated regression coefficients based on linear trend model

Variable	Group		Estimate	SE	t	p-value
Intercept		a	3.507	0.100	34.91	$<.0001$
Smoker	Current	b_1	-0.262	0.115	-2.27	0.0233
Time		b_2	-0.033	0.003	-10.84	$<.0001$
Smoker $\times$ Time	Current	b_3	-0.005	0.004	-1.42	0.16

- Estimated mean response in former smokers:

$$\hat{y} = 3.507 - 0.033 \text{ Time}$$

- Estimated mean response in current smokers:

$$\begin{aligned}\hat{y} &= (3.507 - 0.262) - (0.033 + 0.005) \text{ Time} \\ &= 3.245 - 0.038 \text{ Time}\end{aligned}$$

- No significant difference in the rate of change over time in current vs. former smokers ($p = 0.16$)

Linear Trend Interpretation of Slopes: Example

Table: Estimated regression coefficients from linear trend model without interaction

Variable	Group		Estimate	SE	t	p-value
Intercept		a	3.552	0.095	37.35	$<.0001$
Smoker	Current	b_1	-0.321	0.107	-2.99	0.003
Time		b_2	-0.037	0.001	-24.28	$<.0001$

- FEV1 is decreasing in both groups by 0.037 liters per year on average
- Average FEV1 in current smokers: 0.321 liters lower than in former smokers over study period
- Estimated mean response in former smokers: $\hat{y} = 3.552 - 0.037$ Time
- Estimated mean response in current smokers: $\hat{y} = 3.231 - 0.037$ Time
- Able to summarize trends over time and relation to covariates using a small number of parameters

Unstructured Covariance

- Have assumed an unstructured covariance matrix
  - Requires no assumptions about the pattern of variances and covariances
- Works well when number of measurement occasions small and individuals measured on same set of occasions; does not require complete data
- As number of measurement times increases, number of covariance parameters increases and estimation becomes unstable
- Mistimed repeated measurements cannot be accommodated in an unstructured covariance matrix

Alternatives

- When the sample size is not sufficiently large to estimate the unstructured covariance, but the design is balanced, can impose some structure on the covariance
  - Covariance patterns (e.g., autoregressive, compound symmetry)
- When have imbalanced longitudinal data, linear mixed effects models can be used
  - Well-suited for analyzing inherently unbalanced longitudinal data
  - Correlation among repeated measurements is incorporated through inclusion of random effects in the model

Lesson Summary

- When dealing with repeated measurements from subjects over time, it is important to take the correlation in the multiple outcome measurements provided by each subject
- Longitudinal analysis methods can investigate the changes in the response over time
  - Analysis of response profiles
  - Parametric (e.g., linear) trend over time