Multivariate Statistics & Methodology Using R

Block on Linear Mixed Models

Lecture 1

Antje Nuthmann

Structure of the LMM Module

- Lecture 1:
  - Classic procedures for repeated measures, ANOVA
  - Extended linear models: multiple groups and unbalanced data
- Lecture 2:
  - Random coefficient models
  - Linear mixed-effects models (theory and how to run in R)
- Lecture 3:
  - Model estimation
  - Model evaluation and selection
- Lecture 4:
  - Random-effects structure
  - Why use linear mixed-effects models?
- Lecture 5:
  - Example: factorial design analysed with repeated-measures ANOVA vs. LMM
  - Example: analysis of response accuracies with GLMM

Outline for Lecture 1

1. Characteristics of repeated measures data
2. Classical procedures for repeated measures
   - 2.1 Analysis of summary statistics
   - 2.2 Univariate repeated measures ANOVA
3. Flexible linear models for longitudinal data
   - 3.1 The ANOVA linear model
   - 3.2 A general regression structure
4. Extended linear models: multiple groups and unbalanced data
   - 4.1 Comparing two or more groups
   - 4.2 Imbalance: missing data
   - 4.3 Imbalance: individual measurement designs

1. Characteristics of Repeated Measures Data

- Narrow definition: multiple measures over time to investigate change over time
  - Data collected at several occasions over time for a given subject
- Repeated Measures Study vs. Longitudinal Study
- How does the mean performance of a group of subjects, and of individual responses of a particular subject, change over the course of the investigation?

Characteristics of Repeated Measures Data

- Alternative setup: Individuals (more generally, experimental units) are studied under a series of related conditions
- Example from (modern) psycholinguistics: repeated measurement data with subjects and items as crossed random effects (analyzed with mixed-effects models)

Repeated Measure as opposed to Single Measure Studies

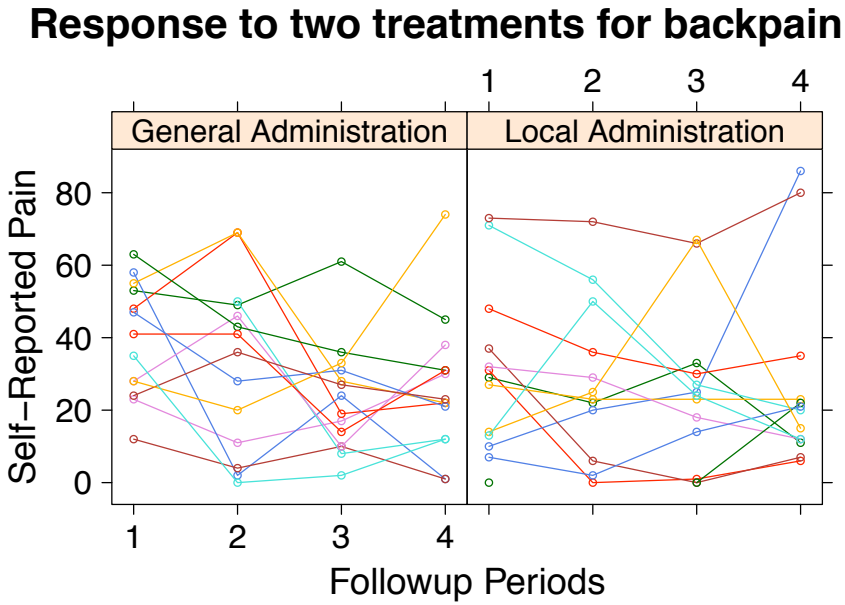
- Basic idea: serial measurements of an individual
- Key issue: measurements from a subject are correlated because they come from one individual

Objectives of Repeated Measure Studies

- Description
  - What occurs to subjects over time? Etc.
- Inference
  - Is a treatment effective? Does substantial change occur over time? Etc.
- Prediction

2. Classical Procedures for Repeated Measures
2.1 Analysis of Summary Statistics

- Good way to begin the study of repeated measures is to graph the data
  - When a graphical summary suggest "no effect" there generally is no effect

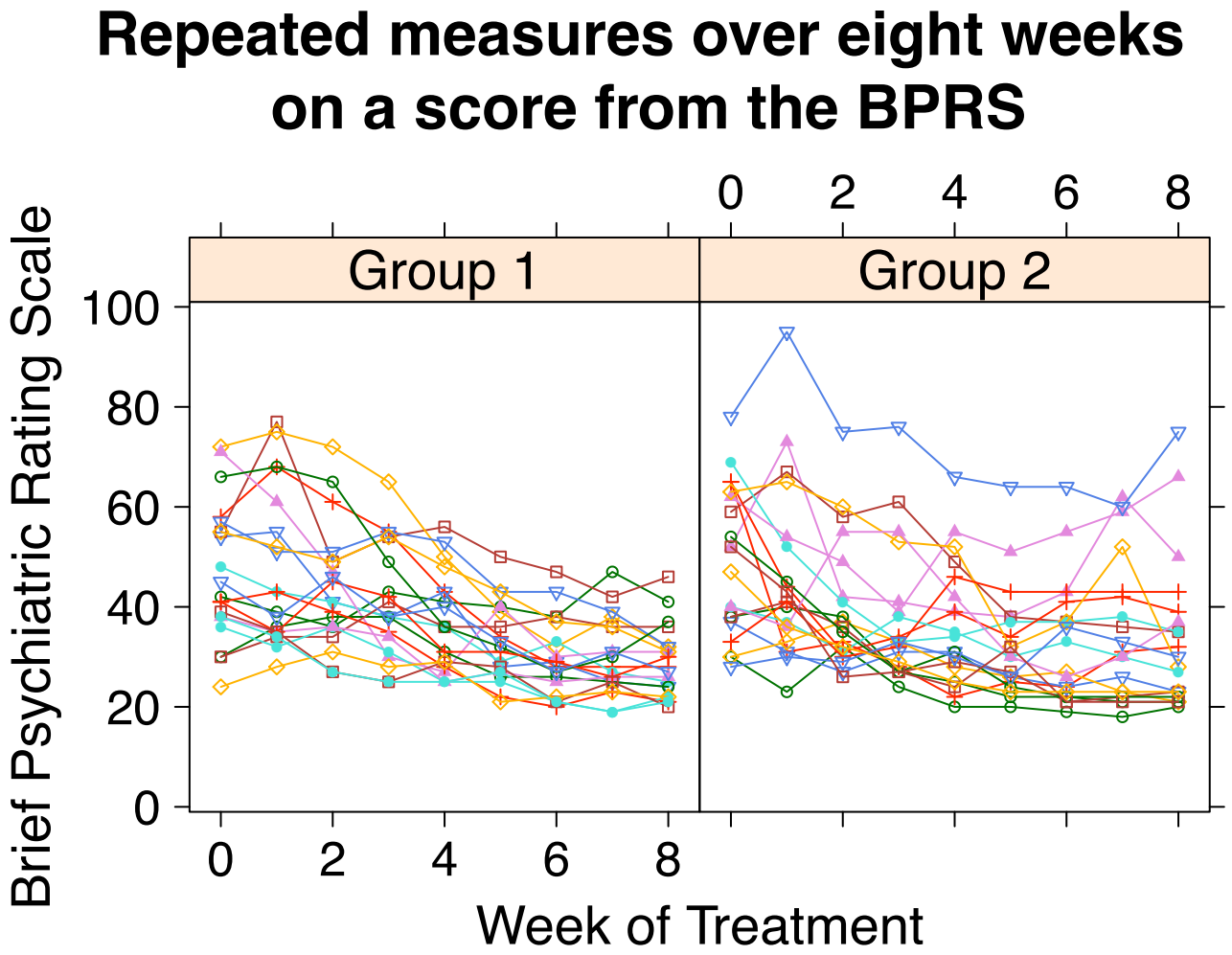


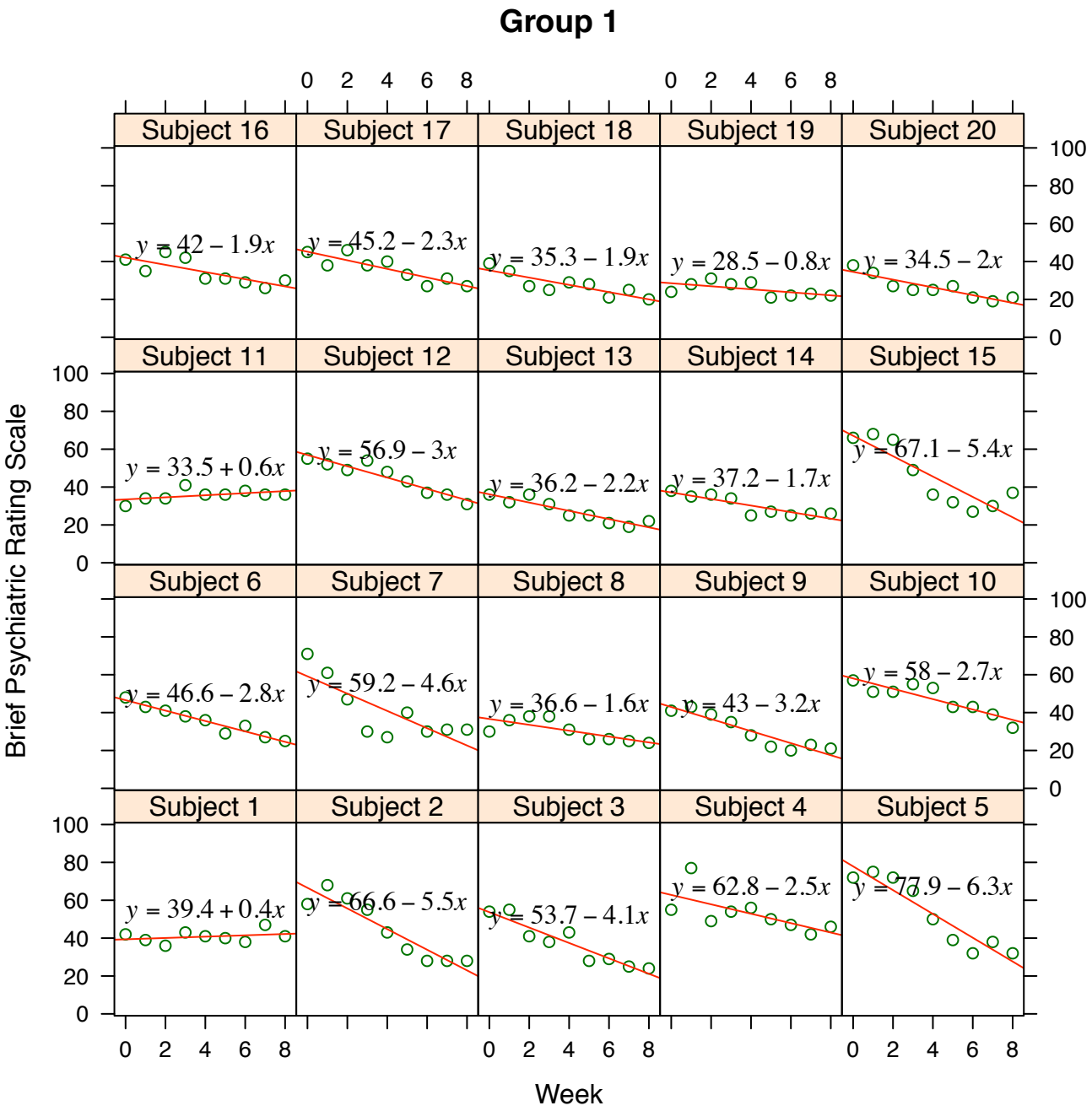
Analysis of Summary Statistics

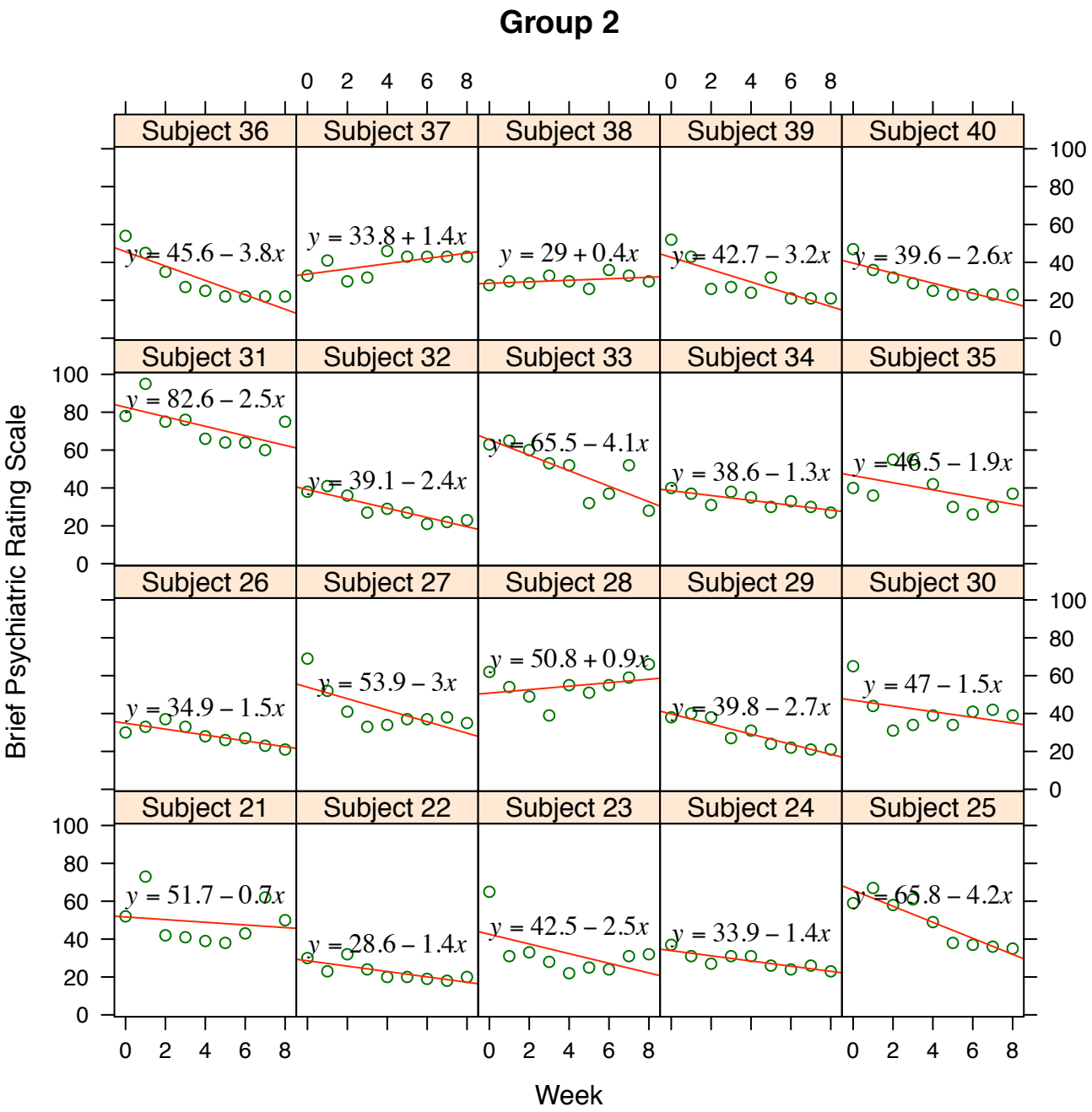
- Good way to begin the study of repeated measures is to graph the data
- Statistical summary
  - Quantification of the figure
  - Numerical description for effects or null effects

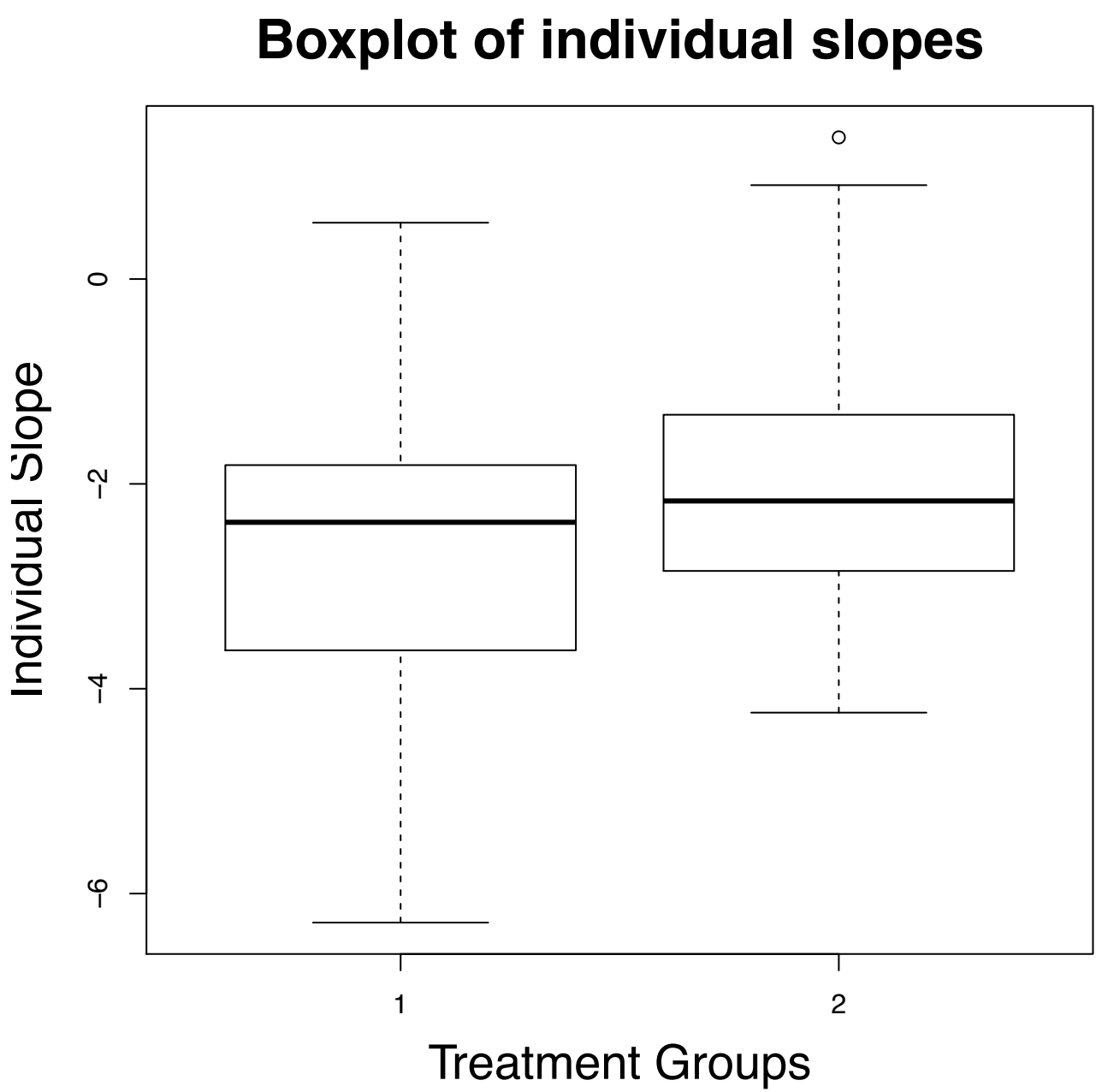
Regression Slope as Summary Measure

- Regression slope describes rate of change during course of the study; Q: Does the rate of change differ between groups?
- Example:
  - 40 men underwent treatment of a psychiatric disorder
  - *Between-subject* factor: treatment type (2)
  - *Within-subject* factor: time – measurements over eight weeks


Repeated measures over eight weeks
on a score from the BPRS
Group 1
Group 2
Brief Psychiatric Rating Scale
Week of Treatment
0
20
40
60
80
100
0
2
4
6
8


Group 1
Subject 16
y = 42 – 1.9x
Subject 17
y = 45.2 – 2.3x
Subject 18
y = 35.3 – 1.9x
Subject 19
y = 28.5 – 0.8x
Subject 20
y = 34.5 – 2x
Subject 11
y = 33.5 + 0.6x
Subject 12
y = 56.9 – 3x
Subject 13
y = 36.2 – 2.2x
Subject 14
y = 37.2 – 1.7x
Subject 15
y = 67.1 – 5.4x
Subject 6
y = 46.6 – 2.8x
Subject 7
y = 59.2 – 4.6x
Subject 8
y = 36.6 – 1.6x
Subject 9
y = 43 – 3.2x
Subject 10
y = 58 – 2.7x
Subject 1
y = 39.4 + 0.4x
Subject 2
y = 66.6 – 5.5x
Subject 3
y = 53.7 – 4.1x
Subject 4
y = 62.8 – 2.5x
Subject 5
y = 77.9 – 6.3x
Brief Psychiatric Rating Scale
Week
0 2 4 6 8
0 20 40 60 80 100


Group 2
Subject 36
y = 45.6 – 3.8x
Subject 37
y = 33.8 + 1.4x
Subject 38
y = 29 + 0.4x
Subject 39
y = 42.7 – 3.2x
Subject 40
y = 39.6 – 2.6x
Subject 31
y = 82.6 – 2.5x
Subject 32
y = 39.1 – 2.4x
Subject 33
y = 65.5 – 4.1x
Subject 34
y = 38.6 – 1.3x
Subject 35
y = 46.5 – 1.9x
Subject 26
y = 34.9 – 1.5x
Subject 27
y = 53.9 – 3x
Subject 28
y = 50.8 + 0.9x
Subject 29
y = 39.8 – 2.7x
Subject 30
y = 47 – 1.5x
Subject 21
y = 51.7 – 0.7x
Subject 22
y = 28.6 – 1.4x
Subject 23
y = 42.5 – 2.5x
Subject 24
y = 33.9 – 1.4x
Subject 25
y = 65.8 – 4.2x
Brief Psychiatric Rating Scale
Week
0 2 4 6 8
0 20 40 60 80 100


Boxplot of individual slopes
Individual Slope
0
−2
−4
−6
1
2
Treatment Groups

2.2 Univariate Repeated Measures ANOVA

- One *between-subjects* factor with $q \geq 1$ levels (treatment groups)
- *within-subjects* factor:
  Each subject is measured
  - with the same variable on n occasions, or
  - on each of n different variables

Model Decomposition

- For *i*-th individual, *g*-th group, *j*-th occasion

$$y_{igj} = \mu + \tau_g + \gamma_j + (\tau\gamma)_{jg} + b_{ig} + e_{igj} \qquad 1 \le j \le n; 1 \le g \le q; 1 \le i \le N_g$$

μ	the overall mean
τ_g	the effect associated with group g $(1 \le g \le q)$
γ_j	the effect associated with the j-th repeated measure $(1 \le j \le n)$
$(\tau\gamma)_{jg}$	interaction effect for group g at occasion j
b_{ig}	the random effect for subject i in the g-th group
e_{ijg}	random error for the i-th individual in group g on occasion j

- The two random terms are independent with distributions

$$b_{ig} \sim N(0, \varphi) \qquad e_{ijg} \sim N(0, \sigma^2_{e\,igj})$$

Example

- Example: $q = 3$ groups, $n = 4$ repeated measures, and $N_g = 10$ subjects in each condition

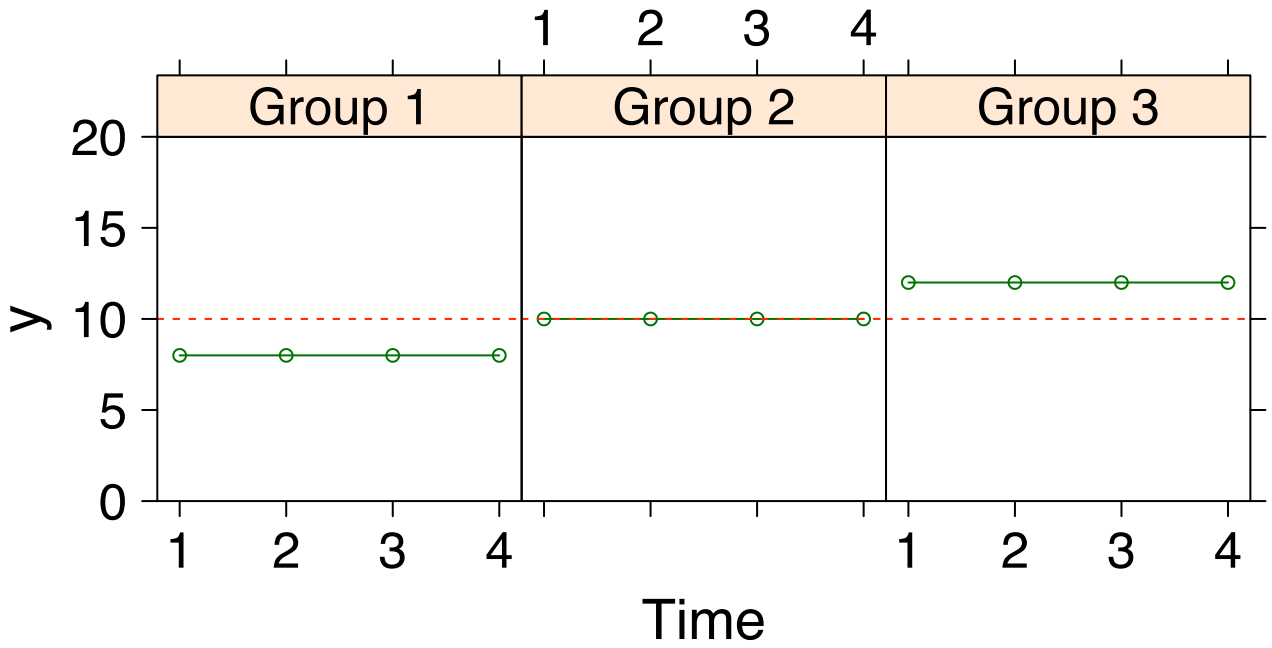
$$\mu = 10 \quad \tau = \begin{pmatrix} -2 \\ 0 \\ 2 \end{pmatrix} \quad \gamma = \begin{pmatrix} 0 \\ 2 \\ 4 \\ 6 \end{pmatrix} \quad (\tau\gamma) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -2 & -4 \\ 0 & -4 & -8 \\ 0 & -6 & -12 \end{pmatrix} \quad \sigma_e^2 = 1 \quad \varphi = 2$$

Reduced Model: Group Effect

$$y_{igj} = \mu + \tau_g, \quad j = 1, \ldots, 4$$

μ the overall mean
τ_g the effect associated with group g $(1 \leq g \leq q)$

- There is no change over time for any group, no differences between subjects, and no variability from random errors.

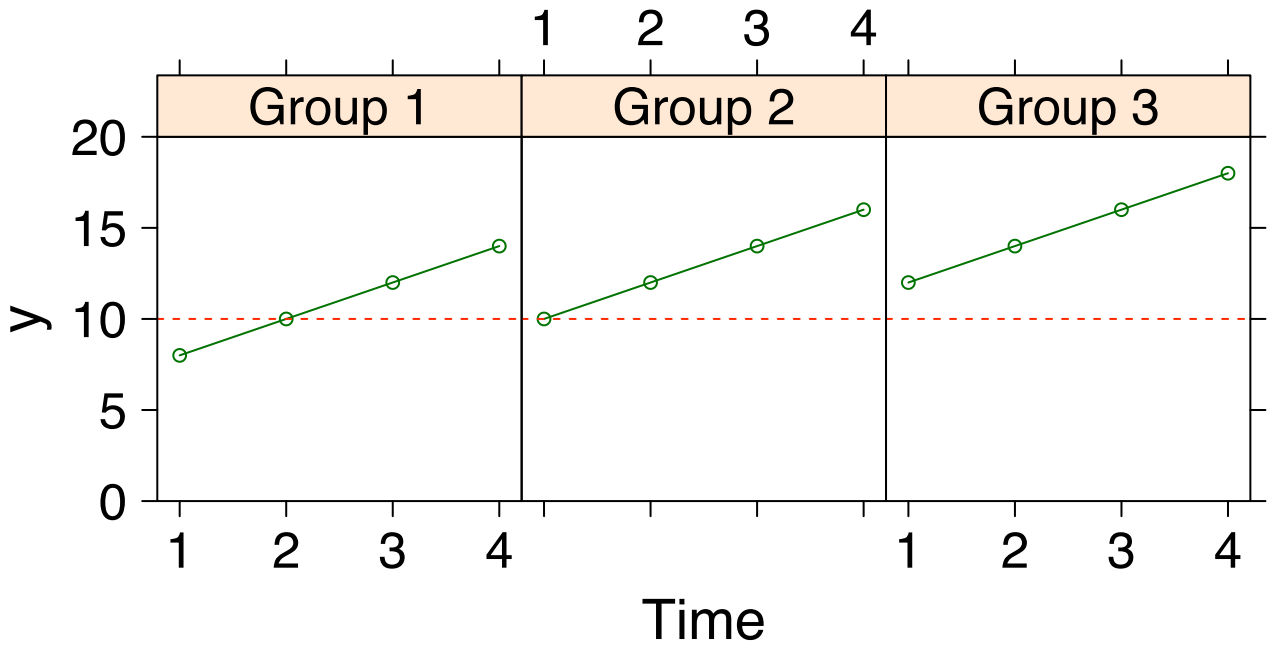


Add Repeated Measures Effect

$$y_{igj} = \mu + \tau_g + \gamma_j, \quad j = 1, \ldots, 4$$

μ	the overall mean
τ_g	the effect associated with group g $(1 \leq g \leq q)$
γ_j	the effect associated with the j-th repeated measure $(1 \leq j \leq n)$

- Produces an increasing trend that applies to all subjects in each group

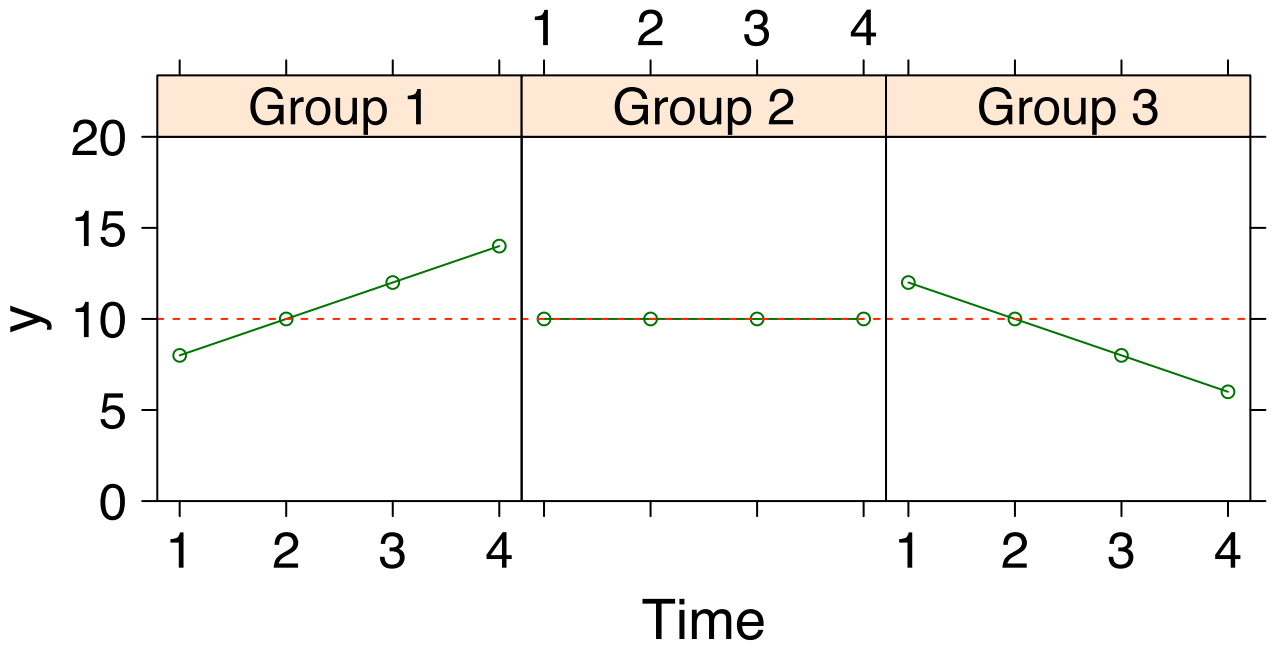


Add Interaction Effect

$$y_{igj} = \mu + \tau_g + \gamma_j + (\tau\gamma)_{jg}, \quad j = 1, \ldots, 4$$

μ	the overall mean
τ_g	the effect associated with group g $(1 \leq g \leq q)$
γ_j	the effect associated with the j-th repeated measure $(1 \leq j \leq n)$
$(\tau\gamma)_{jg}$	interaction effect for group g at occasion j

- The trend differs according to group.

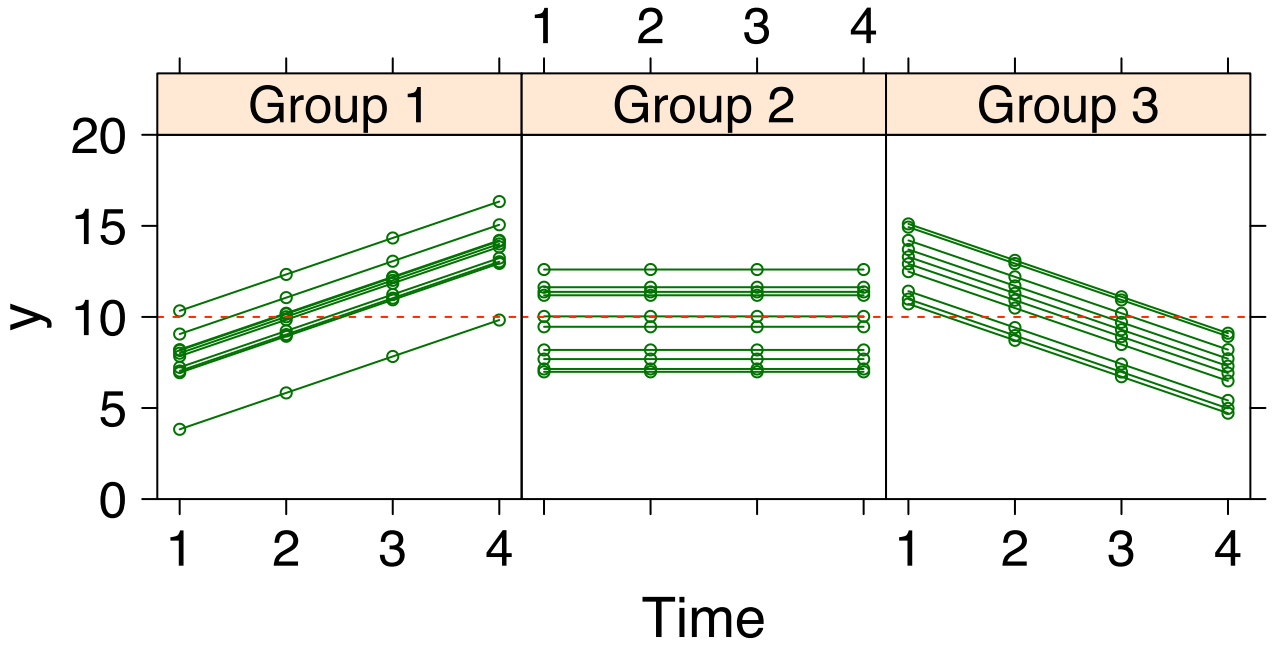


Add Subject Effects

$$y_{igj} = \mu + \tau_g + \gamma_j + (\tau\gamma)_{jg} + b_{ig}, \quad j = 1, \ldots, 4$$

μ	the overall mean
τ_g	the effect associated with group g $(1 \leq g \leq q)$
γ_j	the effect associated with the j-th repeated measure $(1 \leq j \leq n)$
$(\tau\gamma)_{jg}$	interaction effect for group g at occasion j
b_{ig}	the random effect for subject i in the g-th group

- Produces a series of parallel lines within each group.

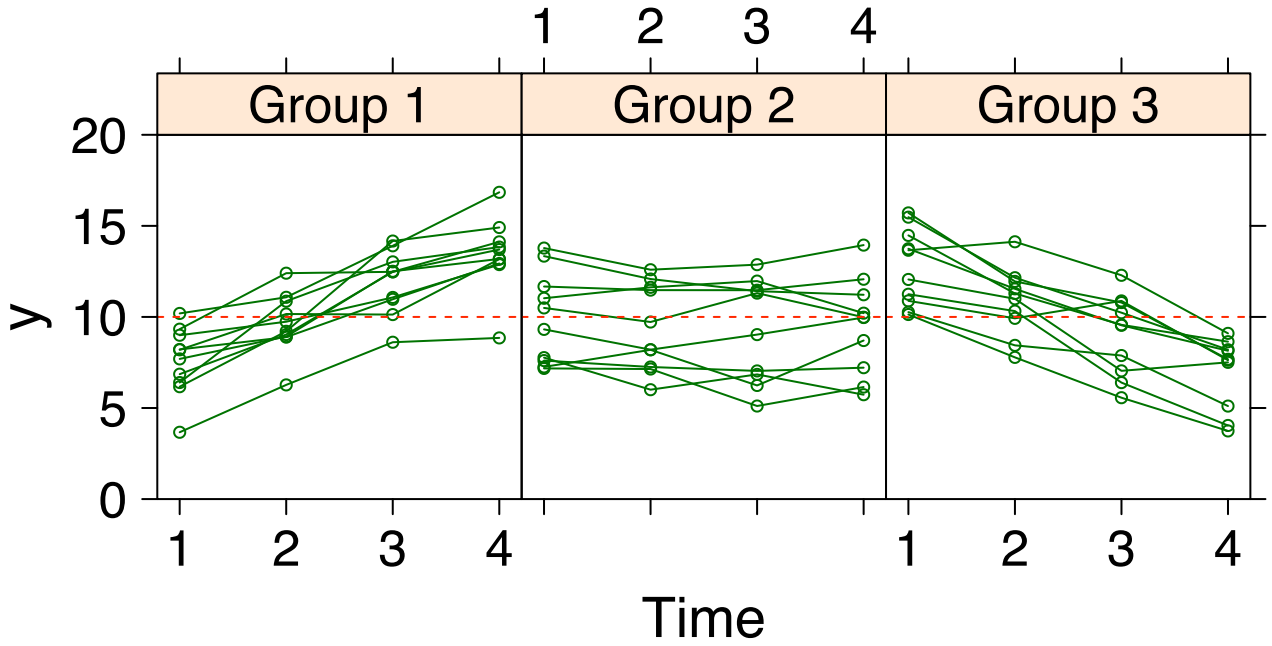


Add Random Error for Subjects

- Full model

$$y_{igj} = \mu + \tau_g + \gamma_j + (\tau\gamma)_{jg} + b_{ig} + e_{igj}, \quad j = 1, \ldots, 4$$

μ	the overall mean
τ_g	the effect associated with group g $(1 \le g \le q)$
γ_j	the effect associated with the j-th repeated measure $(1 \le j \le n)$
$(\tau\gamma)_{jg}$	interaction effect for group g at occasion j
b_{ig}	the random effect for subject i in the g-th group
e_{ijg}	random error for the i-th individual in group g on occasion j



Forms of Statistical Models

- Univariate repeated measures ANOVA describes the change process for scores of an individual
- A related model can be derived that pertains to the mean vector for the collection of *n* scores from individual *i*
- Another derivation gives the covariance matrix between all pairs of scores for an individual

Forms of Statistical Models

- Model for an individual

$$y_{igj} = \mu + \tau_g + \gamma_j + (\tau\gamma)_{jg} + b_{ig} + e_{igj}, \quad j = 1, \ldots, 4$$

- Model for the mean vector
  - assume $n = 4$ repeated measures

$$\begin{pmatrix} y_{ig1} \\ y_{ig2} \\ y_{ig3} \\ y_{ig4} \end{pmatrix} = \begin{pmatrix} \mu + \tau_g + \gamma_1 + (\tau\gamma)_{1g} \\ \mu + \tau_g + \gamma_2 + (\tau\gamma)_{2g} \\ \mu + \tau_g + \gamma_3 + (\tau\gamma)_{3g} \\ \mu + \tau_g + \gamma_4 + (\tau\gamma)_{4g} \end{pmatrix} + \begin{pmatrix} b_{ig} \\ b_{ig} \\ b_{ig} \\ b_{ig} \end{pmatrix} + \begin{pmatrix} e_{ig1} \\ e_{ig2} \\ e_{ig3} \\ e_{ig4} \end{pmatrix}$$

$$\mathbf{y}_{ig} = \boldsymbol{\mu}_g + \mathbf{1}b_{ig} + \mathbf{e}_{ig}$$

Model for the Mean Vector

- The expected pattern of change for the *g*-th group, ignoring individual differences, is

$$\boldsymbol{\mu}_g = E(\mathbf{y}_{ig}) = \begin{pmatrix} \mu + \tau_g + \gamma_1 + (\tau\gamma)_{1g} \\ \mu + \tau_g + \gamma_2 + (\tau\gamma)_{2g} \\ \mu + \tau_g + \gamma_3 + (\tau\gamma)_{3g} \\ \mu + \tau_g + \gamma_4 + (\tau\gamma)_{4g} \end{pmatrix}$$

Model for the Covariance Matrix

- Covariance matrix of repeated measurements describes
  - (i) the variability of scores for individuals within each group
  - (ii) the covariance between pairs of scores
- In ANOVA, covariance pattern is assumed to be the same for all subjects in each group across the collection of scores.

$$\mathbf{\Sigma} = cov(\mathbf{y}_{ig}) = \begin{pmatrix} \varphi + \sigma_e^2 & & & \\ \varphi & \varphi + \sigma_e^2 & & \\ \varphi & \varphi & \varphi + \sigma_e^2 & \\ \varphi & \varphi & \varphi & \varphi + \sigma_e^2 \end{pmatrix}$$

where Σ is of order *n* x *n*. This pattern is called *compound symmetry*.

Summary: Repeated Measures ANOVA

- The repeated measures ANOVA model for the scores is

$$y_{igj} = \mu + \tau_g + \gamma_j + (\tau\gamma)_{jg} + b_{ig} + e_{igj} \qquad 1 \le j \le n; 1 \le g \le q; 1 \le i \le N_g$$

- The individual terms and within-subject variability have distributions

$$b_{ig} \sim N(0, \varphi) \qquad e_{ijg} \sim N(0, \sigma^2_{eigj})$$

- The expected pattern of change for the *g*-th group is

$$\boldsymbol{\mu}_g = E(\mathbf{y}_{ig}) = \begin{pmatrix} \mu + \tau_g + \gamma_1 + (\tau\gamma)_{1g} \\ \mu + \tau_g + \gamma_2 + (\tau\gamma)_{2g} \\ \mu + \tau_g + \gamma_3 + (\tau\gamma)_{3g} \\ \mu + \tau_g + \gamma_4 + (\tau\gamma)_{4g} \end{pmatrix}$$

- and the covariance matrix is

$$\boldsymbol{\Sigma} = cov(\mathbf{y}_{ig}) = \begin{pmatrix} \varphi + \sigma_e^2 & & & \\ \varphi & \varphi + \sigma_e^2 & & \\ \varphi & \varphi & \varphi + \sigma_e^2 & \\ \varphi & \varphi & \varphi & \varphi + \sigma_e^2 \end{pmatrix}$$

- Consequently, the distribution of scores for an individual is also normal

$$\mathbf{y}_{ig} \sim N(\boldsymbol{\mu}_g, \boldsymbol{\Sigma})$$

ANOVA Hypothesis Tests

- No group difference
  - H_0: $\tau_g = 0$
- No change over time
  - H_0: $\gamma_j = 0$
- No group-by-occasion interaction
  - H_0: $(\tau\gamma)_{gj} = 0$

Shortcomings of Repeated Measures ANOVA

- Time variable not included directly
- Requires balanced data (measurements on each individual occur at the same occasions) without missing values
- Model imposes a uniform structure on the mean vector
- Restrictive assumptions about the variances and covariances of the variables

3. Flexible Linear Models for Longitudinal Data

- The beauty of ANOVA is that one size fits all.
- The flipside is the approach cannot be customized.
- How can we extend this basic linear model to make it more flexible?

3.1 The ANOVA Linear Model

$$y_{igj} = \mu + \tau_g + \gamma_j + (\tau\gamma)_{jg} + b_{ig} + e_{igj}$$

- μ: overall mean
- τ_g: effect associated with group g $(1 \le g \le q)$
- γ_j: effect associated with occasion j $(1 \le j \le n)$
- $(\tau\gamma)_{jg}$: interaction effect for group g, occasion j
- b_{ig}: random effect for subject i $(1 \le i \le N_g)$
- e_{igj}: random error

- The individual terms and within-subject variability have distributions

$$b_{ig} \sim N(0, \varphi) \qquad e_{ijg} \sim N(0, \sigma_e^2)$$

- Means of b_{ig} and e_{ijg} are zero; model is additive → mean over individuals of the scores on occasion *j* for group *g* is the sum of the first four terms

$$\begin{aligned} \mu_{jg} &= E_i(y_{igj}) \\ &= \mu + \tau_g + \gamma_j + (\tau\gamma)_{gj} \end{aligned}$$

- This is the *model of the mean*, or equivalently, the *mean structure*.
- How can we set up the mean structure as we need it?

Linear Models for the Mean Vector

- Let the n x 1 vector $\boldsymbol{\mu}_g = (\mu_{1g},\ldots,\mu_{ng})'$ denote the collection of cell means for the g-th group, $1 \leq g \leq q$, of a particular design.
- A *linear model for the means* is an expression in which the means for each of the q groups are a linear function of a fixed design matrix, X_g (n x p), and more fundamental parameters, $\theta = (\theta_1,\ldots,\theta_p)'$.
- In general, a linear model for the means of g groups has the form

$$\begin{pmatrix} \mu_{1g} \\ \mu_{2g} \\ \vdots \\ \mu_{ng} \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & & x_{2p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \theta_1 \\ \vdots \\ \theta p \end{pmatrix}$$

$$\boldsymbol{\mu}_g = \mathbf{X}_g\boldsymbol{\theta}$$

Example

$q = 2$ groups and $n = 3$ repeated measures layout of the cell means:

	Measures		
Groups	1	2	3
1	μ_{11}	μ_{12}	μ_{13}
2	μ_{21}	μ_{22}	μ_{23}

simple model with group effect only:

$\mu_{jg} = \mu + \tau_g \qquad g = 1, 2; j = 1, 2, 3$

There are only three fundamental parameters that make up the parameter vector $\boldsymbol{\theta} = (\mu, \tau_1, \tau_2)'$

The linear model is written as

Group 1 Model

$$\begin{pmatrix} \mu_{11} \\ \mu_{21} \\ \mu_{31} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \end{pmatrix}$$

$$\boldsymbol{\mu}_1 = \mathbf{X}_1\boldsymbol{\theta}$$

Group 2 Model

$$\begin{pmatrix} \mu_{12} \\ \mu_{22} \\ \mu_{32} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \end{pmatrix}$$

$$\boldsymbol{\mu}_2 = \mathbf{X}_2\boldsymbol{\theta}$$

Mean Structure for ANOVA

In the mean structure for ANOVA, there appears to be a total of $(q+1)(n+1)$ parameters used to describe only qn cell means (e.g., 12 p. for 6 means).

μ	τ_g	γ_j	$(\tau\gamma)_{gj}$
1	q	n	qn
1	2	3	2*3=6

Not all of these parameters can be estimated; the restrictions on the system are:

$$\sum_{g=1}^{q} \tau_g = 0 \qquad \sum_{j=1}^{n} \gamma_j = 0 \qquad \sum_{g=1}^{q} (\tau\gamma)_{gj} = 0 \qquad \sum_{j=1}^{n} (\tau\gamma)_{gj} = 0$$

Consequently, there are exactly qn parameters in the model related to the qn cell means (e.g., 6 parameters for 6 means).

μ	τ_g	γ_j	$(\tau\gamma)_{gj}$
1	$q-1$	$n-1$	$(q-1)(n-1)$
1	2-1=1	3-1=2	1*2=2

Example

example: $q = 2$ groups and $n = 4$ repeated measures

$qn = 2*4 = 8$ parameters

design matrix X_g (n x p)

$$\begin{pmatrix} \mu_{11} \\ \mu_{21} \\ \mu_{31} \\ \mu_{41} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 \end{pmatrix}$$

$$\begin{pmatrix} \mu_{12} \\ \mu_{22} \\ \mu_{32} \\ \mu_{42} \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 & 0 & 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 & 1 & 0 & 0 & -1 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 \end{pmatrix}$$

$$\begin{pmatrix} \mu \\ \tau_1 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ (\tau\gamma)_{11} \\ (\tau\gamma)_{12} \\ (\tau\gamma)_{13} \end{pmatrix}$$

Unattractive Features of ANOVA Model

- The mean structure for ANOVA is saturated
  - number of cell means and number of parameters are the same; model has same complexity as data
    → no parsimony
- Design matrixes for μ_1 and μ_2 always have the described form
  - no possibility for informed model-building

3.2 A General Regression Structure

The collection of repeated measures for individual *i* is treated as a unit.
Let y_i denote the *n* observations for the *i*-th subject.
Suppose that all subjects are measured at the same time points, with x_j being the number of days that have elapsed since the beginning of the experiment to the *j*-th occasion.
Define the *n* x 2 matrix X as

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

A simple linear model in which y_i is tied directly to time measurement is

$$\mathbf{y}_i = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}_i$$

where $e_i = (e_{i1},\ldots, e_{in})'$ are regression residuals.

Classical Regression

Matrix form of classical, simple regression

$$\begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ \vdots \\ e_N \end{pmatrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

Sample consists of randomly selected individuals.

A score from one subject is statistically independent of scores from all other subjects.

Independence assumption implies that the distribution of the collection of residuals is $e \sim N(0,\boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 & & \\ & \ddots & \\ & & \sigma^2 \end{pmatrix}$$

$$= \sigma^2 \mathbf{I}_N$$

Regression for Repeated Measures

- In a repeated measures study, the information in y_i comes from one subject.

$$\mathbf{y}_i = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}_i$$

- Elements of e_i are not independent; as they all come from the same individual, residuals are usually correlated to some degree
- Covariance structure for residuals

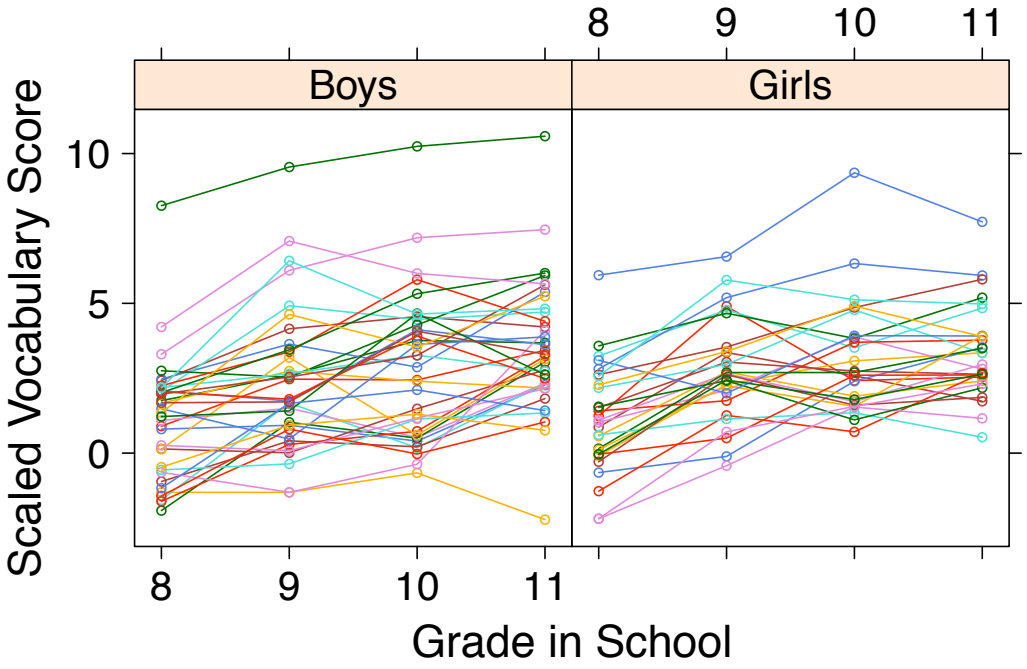
$$\boldsymbol{\Sigma} = cov(\mathbf{e}_i) = \begin{pmatrix} \sigma_1^2 & & & \\ \sigma_{21} & \sigma_2^2 & & \\ \vdots & \vdots & \ddots & \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix}$$

Four Discussion Points

- The repeated measures have particular means in the population; these are directly related to the time of the *j*-th measurement
- With the new approach, the means can be summarized by only a few parameters (ANOVA: *n* parameters for *n* means)
- Prediction is easy and follows immediately
- In contrast, predictions for unmeasured time points not possible in ANOVA; time does not appear in ANOVA model

Example

- Does growth of vocabulary slow in adolescence?
  - 64 high school students were assessed in grades 8, 9, 10, and 11
  - results: decelerating pattern of growth across the years



Bock (1985)

Example

- Does growth of vocabulary slow in adolescence?
  - 64 high school students were assessed in grades 8, 9, 10, and 11
  - results: decelerating pattern of growth across the years
- How to quantify average performance over time?
  - discrete variable grade level used as continuous indicator of age
  - quadratic model: $y_{ij} = \beta_0 + \beta_1 g_j + \beta_2 g_j^2 + e_{ij}$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & 8 & 8^2 \\ 1 & 9 & 9^2 \\ 1 & 10 & 10^2 \\ 1 & 11 & 11^2 \end{pmatrix}$$

  - equivalent: $[\mathbf{X}]_{j\bullet} = (1, g_j, g_j^2)$

Model Parameters

- Focus shift from evaluating means or variances (as in ANOVA) to evaluating parameters.
- Parameter estimates for the example

$$\hat{\beta}_0 : -24.4(4.4) \qquad \hat{\beta}_1 : 4.98(.97) \qquad \hat{\beta}_2 : -.222(.05)$$

- Test of $H_0 : \beta_2 = 0$
  - Does vocabulary have a nonlinear mean pattern across grade levels?

Predictions from the Model

- Parameters of the model allow us to estimate the mean vocabulary score at <u>any</u> grade level <u>within</u> the range covered in the study (grade 8 to 11).
  - point estimate of μ_8 is $y_8 = 1.14$
    (sample mean of vocabulary scores in grade 8)
  - model estimate: $\hat{\mu}_{g=8} = -24.4 + (4.98 \cdot 8) - (.222 \cdot 64) = 1.23$

    $\hat{\beta}_0 : -24.4(4.4) \qquad \hat{\beta}_1 : 4.98(.97) \qquad \hat{\beta}_2 : -.222(.05)$
  - Which of the two values would be preferable?

4. Extended Linear Models: Multiple Groups and Unbalanced Data

- so far: <u>one</u> design matrix, X, is appropriate for <u>all</u> subjects

$$\mathbf{y}_i = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}_i \text{ with } \mathbf{e}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$$

$$\mathbf{y}_i \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$$

$$E(\mathbf{y}_i) = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$$

- next step: generalization of the model by allowing different design matrices for individuals according to group (index *g*)

$$\mathbf{y}_i = \mathbf{X}_g\boldsymbol{\beta} + \mathbf{e}_i \text{ with } \mathbf{e}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$$

$$E(\mathbf{y}_i) = \boldsymbol{\mu}_g = \mathbf{X}_g\boldsymbol{\beta}$$

$$\mathbf{y}_i \sim N(\mathbf{X}_g\boldsymbol{\beta}, \boldsymbol{\Sigma})$$

4.1 Comparing Two or More Groups
- Parallel Mean Profiles -

- *Parallel means model* as preliminary model to explore similarities between two groups
- Group 1 = control group: each occasion *j* has its own mean
- Group 2 = experimental group: same, but constant, κ, added

$$y_{ij} = \begin{cases} \mu_j + e_{ij} & \text{Subject } i \text{ in group 1} \\ \mu j + \kappa + e_{ij} & \text{Subject } i \text{ in group 2} \end{cases}$$

- Regression coefficients in case of $n = 4$ measurements:

$$\boldsymbol{\beta} = (\mu_1, \mu_2, \mu_3, \mu_4, \kappa)'$$

- Two design matrices, one for each group

$$\mathbf{X}_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

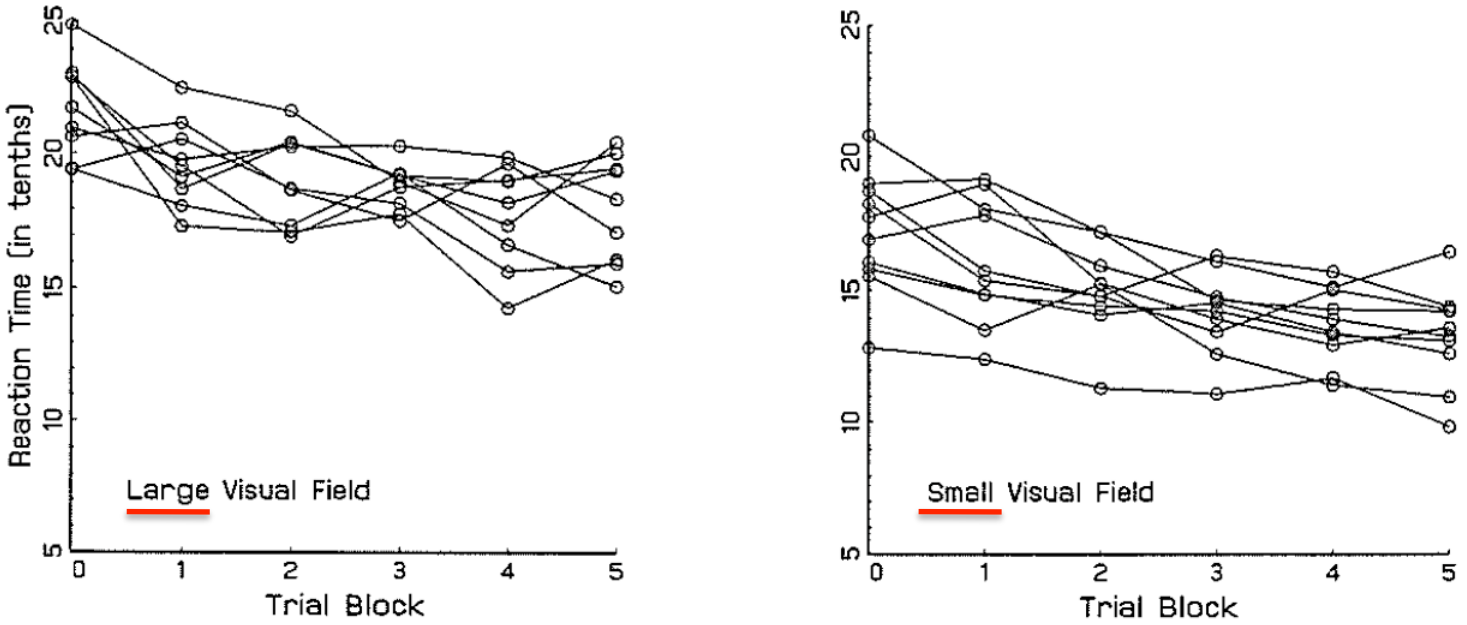
Group 1

$$\mathbf{X}_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

Group 2

Example: Learning in a Visual Search Task

- Phase 1: geometrical object presented on *large* or *small* computer display
- Phase 2: presentation of many other objects
- Did original object appear in second group of objects?



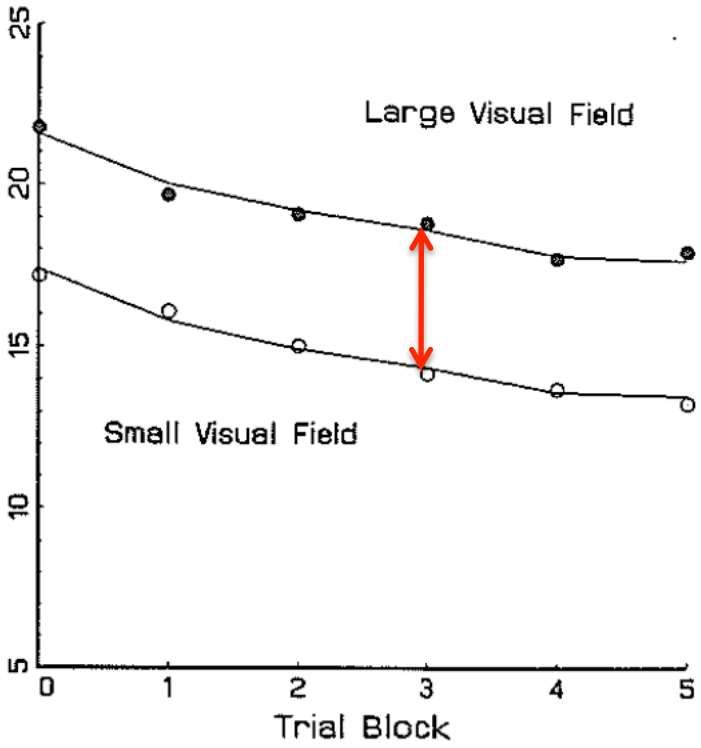
Chaiken (1994)

Example: Learning in Visual Search Task

- Parameter estimates for parallel means model:

$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$	$\hat{\mu}_5$	$\hat{\mu}_6$	$\hat{\kappa}$
21.6(.73)	20.0(.72)	19.2(.64)	18.6(.54)	17.8(.64)	17.7(.70)	−4.21(7.1)

- κ is offset coefficient for the mean profile difference of the *Small* Visual Field compared to the *Large*



4.1 Comparing Two or More Groups

- Parameterizing Group Differences Directly -

- *mean initial status* of the groups is the *same* at $t = 0$ but
- *rates of improvement differ* between the groups (conditions)
- model specification:

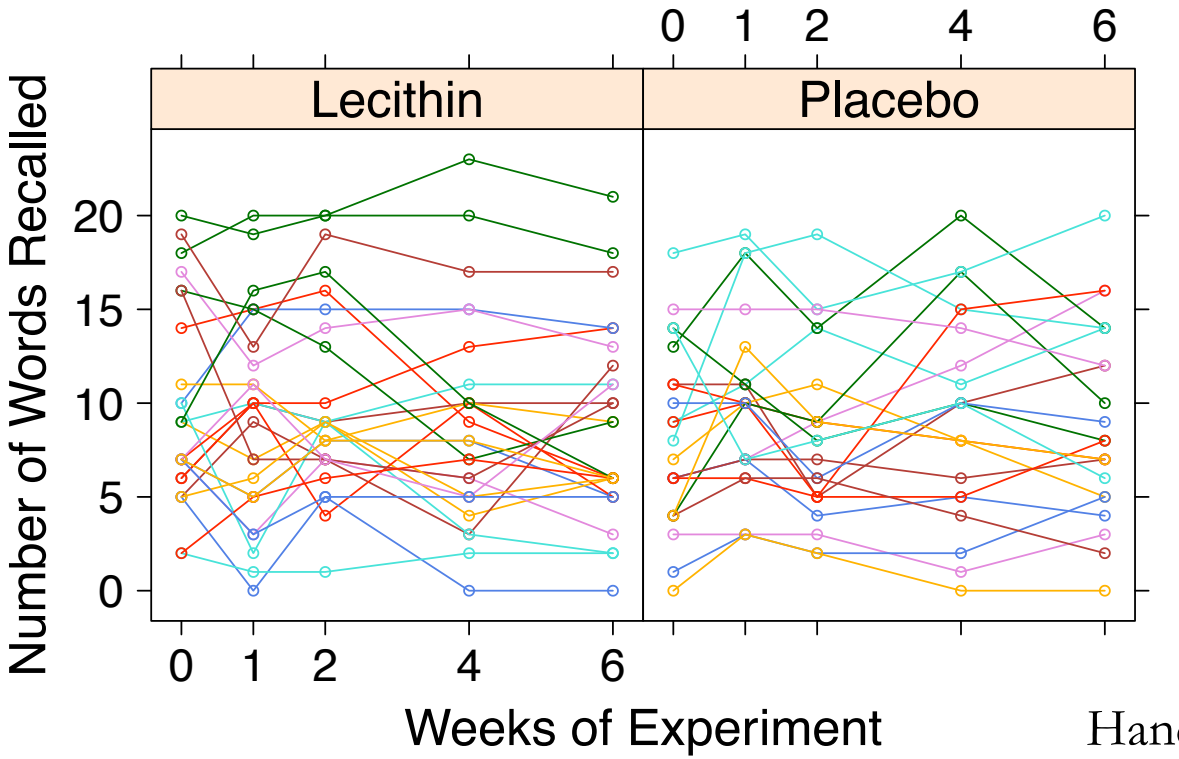
$$\mu_{ij} = \begin{cases} \beta_0 + \alpha t_j & \text{Subject } i \text{ in group 1} \\ \beta_0 + \gamma t_j & \text{Subject } i \text{ in group 2} \end{cases}$$

  - t_j is elapsed time between beginning of experiment and j-th occasion
- Group 1 = control group: $y_{ij} = \beta_0 + \alpha t_j + e_{ij}$
- Group 2 = experimental group, slope defined as $\gamma = \alpha + \delta$
  - δ represents increment for the experimental group over and above the slope of the control group (α)

    $y_{ij} = \beta_0 + (\alpha + \delta)t_j + e_{ij}$
- Linear model with 3 parameters: $\boldsymbol{\beta} = (\beta_0, \alpha, \delta)'$
  - parameter δ has direct relationship to the question of differential rate of change between groups

Example: Alzheimer's Treatment

- Can lecithin, a food supplement, improve short-term memory in patients?

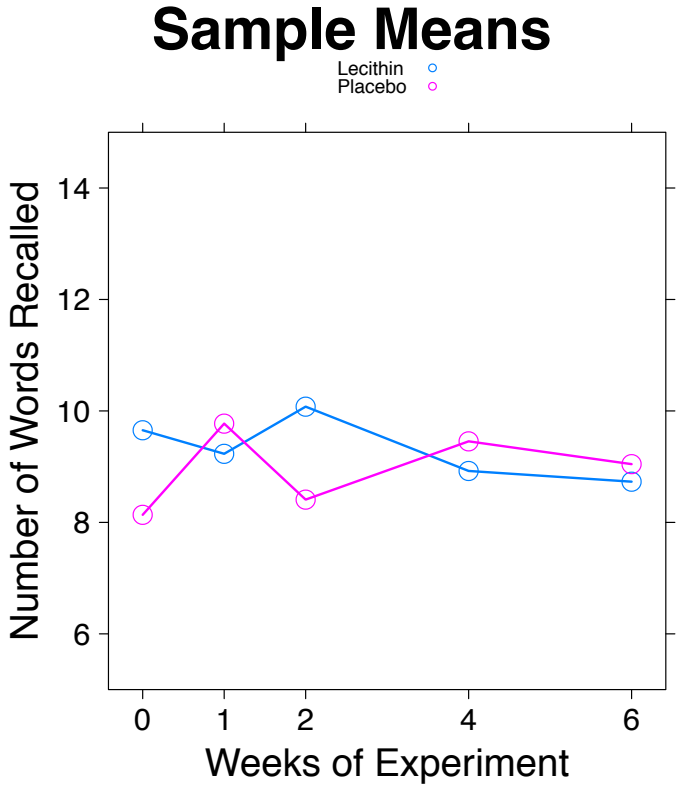


- 48 patients, randomly assigned to either lecithin or placebo groups
- memory test at 5 occasions

Hand & Taylor (1987)

Example: Alzheimer's Treatment

- Can lecithin, a food supplement, improve short-term memory in patients?



- 48 patients, randomly assigned to either lecithin or placebo groups
- memory test at 5 occasions

Hand & Taylor (1987)

Extension: Differences in Intercept and Slope

- Model with distinct *slopes* and *intercepts* for the two groups:

$$y_{ij} = f_{ij} + e_{ij}$$

$$\text{where} \quad f_{ij} = \begin{cases} (\beta_0 + \delta_0) + (\beta_1 + \delta_1)t_j & \text{lethicin} \\ \beta_0 + \beta_1 t_j & \text{placebo} \end{cases}$$

- Vector of regression coefficients: $\boldsymbol{\beta} = (\beta_0, \beta_1, \delta_0, \delta_1)'$
- Parameter estimates:

$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\delta}_0$	$\hat{\delta}_1$
$8.47(.96)$	$-.044(.13)$	$1.34(1.3)$	$-.0361(.17)$

- Hypotheses/ interpretations:
  - $\beta_1 \neq 0$ indicates change over time in *control* group
  - $\delta_0 \neq 0$ indicates *group* difference in *intercept*
  - $\delta_1 \neq 0$ indicates *group* difference in *slope* (due to treatment)
  - example: all parameter estimates small with respect to their standard errors → no effects

4.2 Imbalance: Missing Data

Now we let the genie out of the bottle…

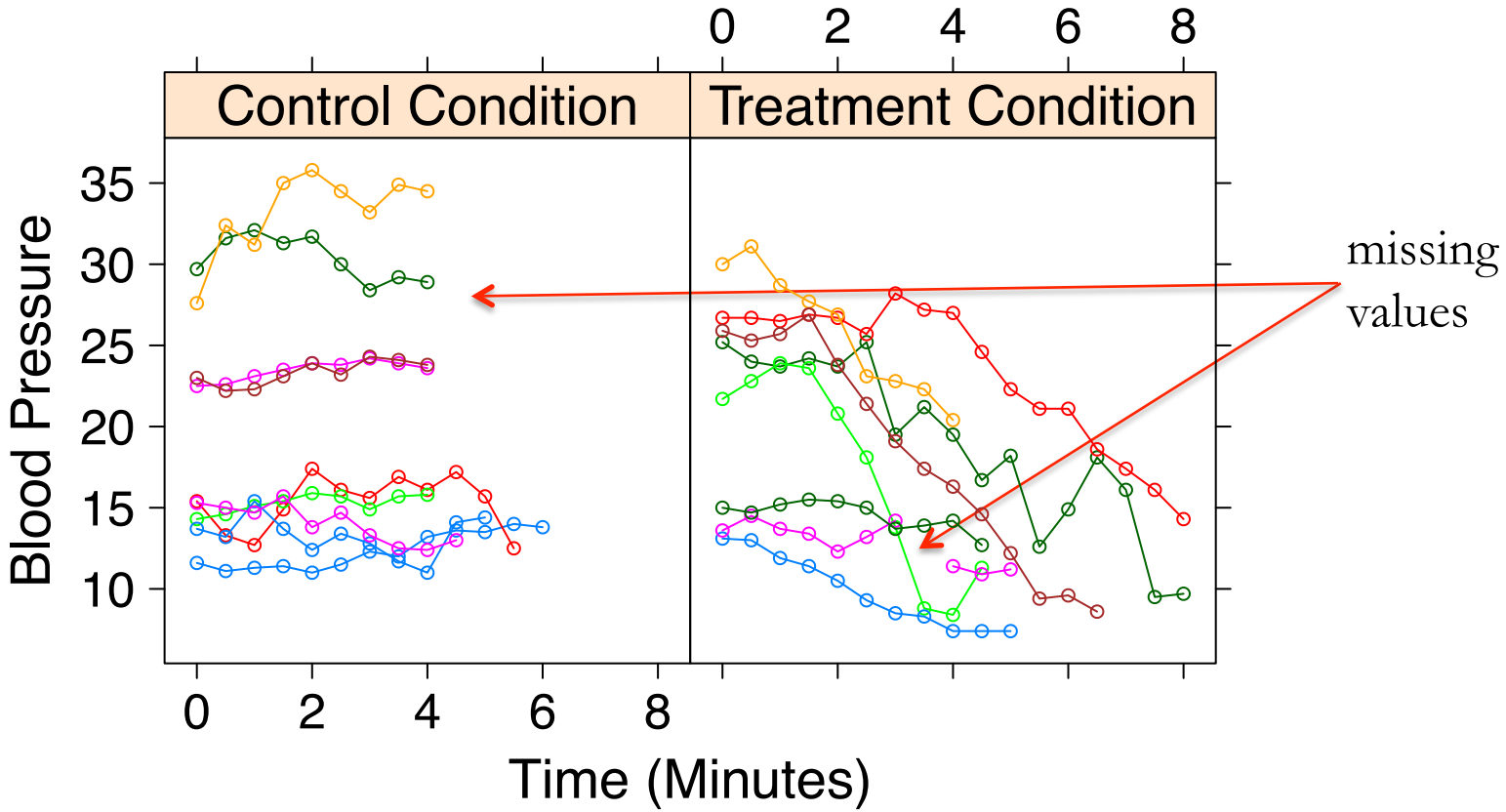
- Previously: two design matrices to handle two experimental conditions
- Next step: design matrices specified on completely individualized basis
  - X_i as design matrix for *i*-th subject
- We can now drop any requirement of balance in the data!
  - Example 1 (blood pressure): unequal number of observations for each subject
  - Example 2 (heart rate): different values of the independent variable used as predictors

Example: Treatment for High Blood Pressure

- Medication to reduce high blood pressure tested against placebo
- Blood pressure was recorded at baseline and at every 30 seconds for eight minutes → 17 planned measurements
  (Nevens et al., 1996)

Example: Treatment for High Blood Pressure

- Only two subjects completed the full protocol
- Most subjects have five or more missing values



Incomplete Data

- Complete data:
  - n measurements taken from subject i — $\mathrm{y}_i = (y_{i1},\ldots,y_{in})'$ for all i
  - x_i is the <u>same</u> for all N subjects (drop i) — $\mathrm{x} = (x_1,\ldots, x_n)'$ for all i
- Incomplete data
  - n_i, measurements for subject i: — $\mathrm{y}_i = (y_{i1},\ldots,y_{in_i})'$
  - $1 \le n_i \le n$
  - separate x_i for each of the N subjects — $\mathrm{x}_i = (x_{i1},\ldots, x_{in_i})'$
  - x_i is a subset of x
- Example: measurement protocol is $\mathbf{x} = (0, 5, 9, 14, 21)'$
  - 3 subjects with incomplete data were assed on days

    $\mathbf{x}_1 = (0, 9)'$ or $\mathbf{x}_2 = (5, 14, 21)'$ or $\mathbf{x}_3 = (9)'$, but not $\mathbf{x}_i = (2, 8)'$

Incomplete Data cont'd

- Design matrices allowed to differ for each subject

$$\mathbf{X}_i = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n_i 1} & \cdots & x_{n_i p} \end{pmatrix}$$

example: subject 2

$$\mathbf{X}_2 = \begin{pmatrix} 1 & 5 & 5^2 \\ 1 & 14 & 14^2 \\ 1 & 21 & 21^2 \end{pmatrix}$$

$x_2 = (5,14,21)'$

- *3 rows for 3 measurements*
- *3 columns for 3 parameters (intercept, linear, quadratic)*

- Covariance matrix has subscript *i* to distinguish the various possibilities individuals may have with observed and missing data; $\mathbf{\Sigma}_i$ is n_i x n_i
  - subject with complete data:

$$\mathbf{\Sigma}_i = \begin{pmatrix} \sigma_1^2 & & & & \\ \sigma_{21} & \sigma_2^2 & & & \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & & \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 & \\ \sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_5^2 \end{pmatrix}$$

subject 2 (incomplete data):

$$\mathbf{\Sigma}_2 = \begin{pmatrix} \sigma_2^2 & & \\ \sigma_{42} & \sigma_{42}^2 & \\ \sigma_{52} & \sigma_{54} & \sigma_5^2 \end{pmatrix}$$

Interim Summary

- for the i-th case with $n_i \leq n$ repeated measures, $y_i = (y_{i1},\ldots,y_{in_i})'$, the extended model uses the n_i x p design X_i made up from the independent variable values x_i in the model

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{e}_i \text{ with } \mathbf{e}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_i)$$

- Implications:
  - Enables completely open-ended designs where subjects do not need to have same number of data points
  - Everyone contributes, no one is excluded (even if they only contribute 1 measurement)
  - *Planned missingness* possible to alleviate costs and keep subjects motivated

Segmented Mean Structure (Blood Pressure Example)

- Control condition: no change over time

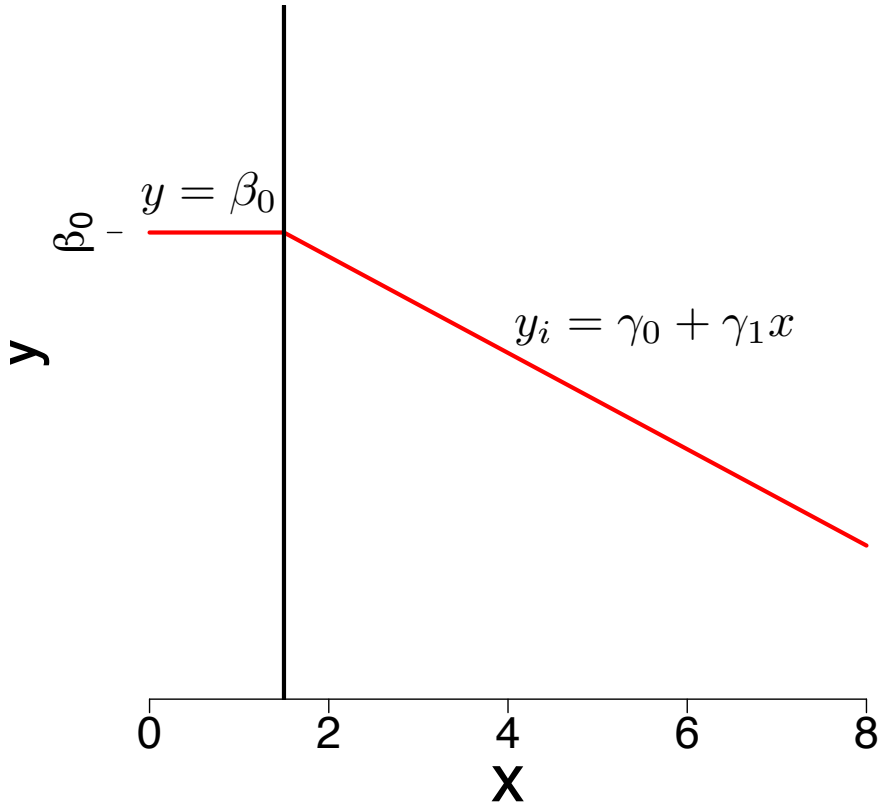
$$y_{ij} = \beta_0 + e_{ij} \quad j = 1, \ldots, n_i$$

- Treatment condition: two-phase response to drug
  - Phase 1: no change ("flat response") for first 1.5 min
  - Phase 2: linear decrease in blood pressure
  - Two-part model: $y_{ij} = f_{ij} + e_{ij}$

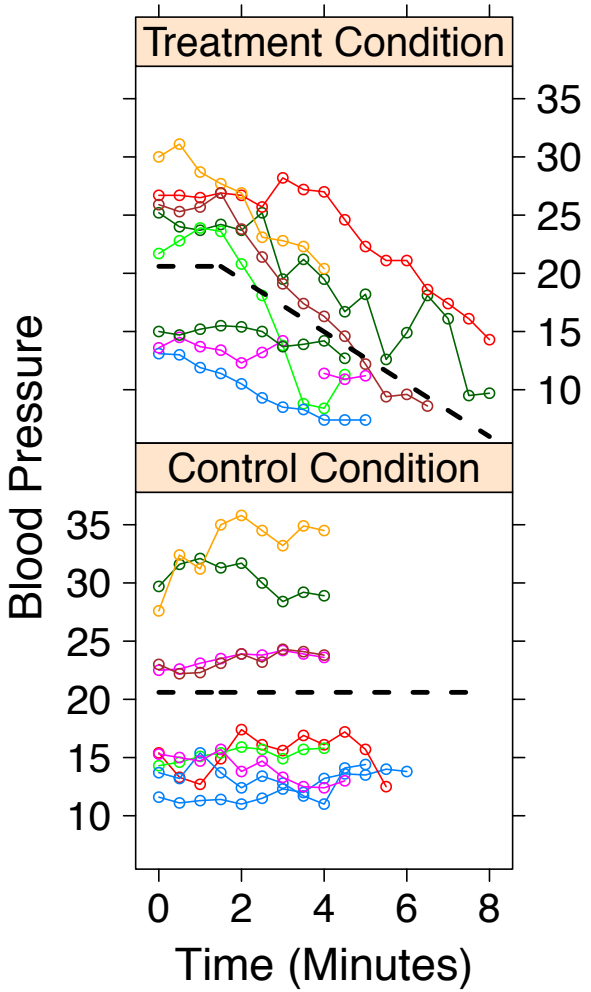
    where the systematic part is composed of

$$f_{ij} = \begin{cases} \beta_0 & x_j \leq 1.5 \\ \gamma_0 + \gamma_1 x_j & x_j > 1.5 \end{cases}$$

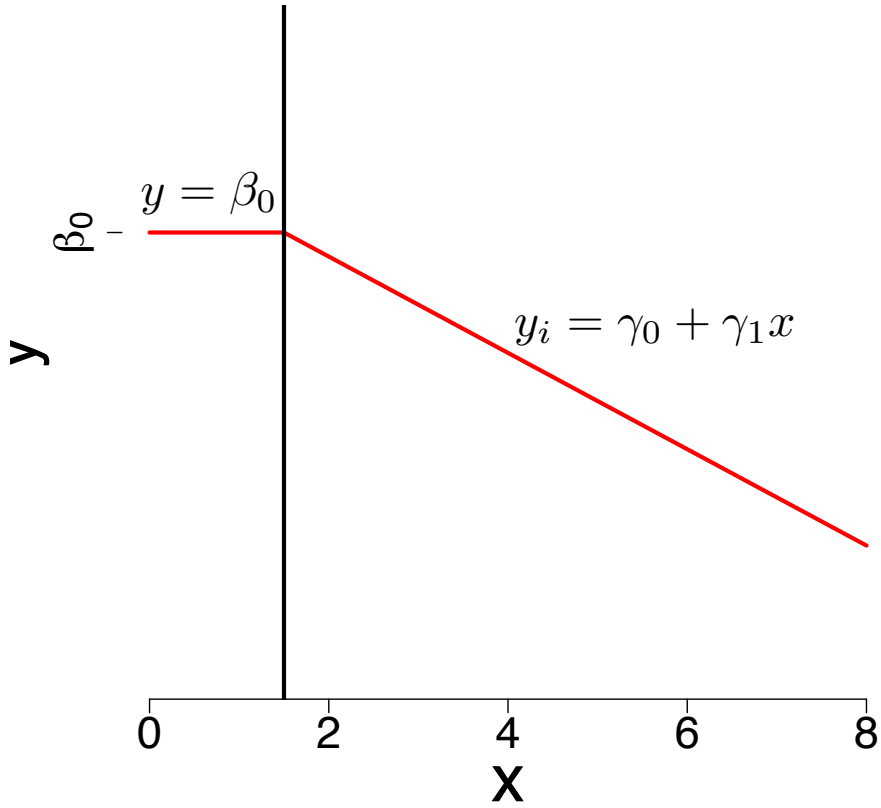
Segmented Mean Structure (Blood Pressure Example)



- the two segments joint at $x = 1.5$ min



Segmented Mean Structure (Blood Pressure Example)



- the two segments joint at $x = 1.5$ min

$$\beta_0 = \gamma_0 + 1.5\gamma_1$$

$$\gamma_0 = \beta_0 - 1.5\gamma_1$$

rewrite equation for phase 2 as:

$$\begin{aligned} \gamma_0 + \gamma_1 x_j &= \gamma_0 + \gamma_1 x_j \\ &= (\beta_0 - 1.5\gamma_1) + \gamma_1 x_j \\ &= \beta_0 + \gamma_1 (x_j - 1.5) \end{aligned}$$

$$f_{ij} = \begin{cases} \beta_0 & x_j \leq 1.5 \\ \beta_0 + \gamma_1 (x_j - 1.5) & x_j > 1.5 \end{cases}$$

- (only) two model parameters with meaningful interpretation
  - β_0 mean response during first 1.5 min
  - γ_0 slope in the second phase

Covariance Matrix of Residuals (Blood Pressure Example)

- Recall: 17 planned measurements
  - Subjects with complete data require Σ_i to be of order 17 x 17 → large number of parameters
- For now: lets assume *compound symmetry* (equal variances, equal covariances)

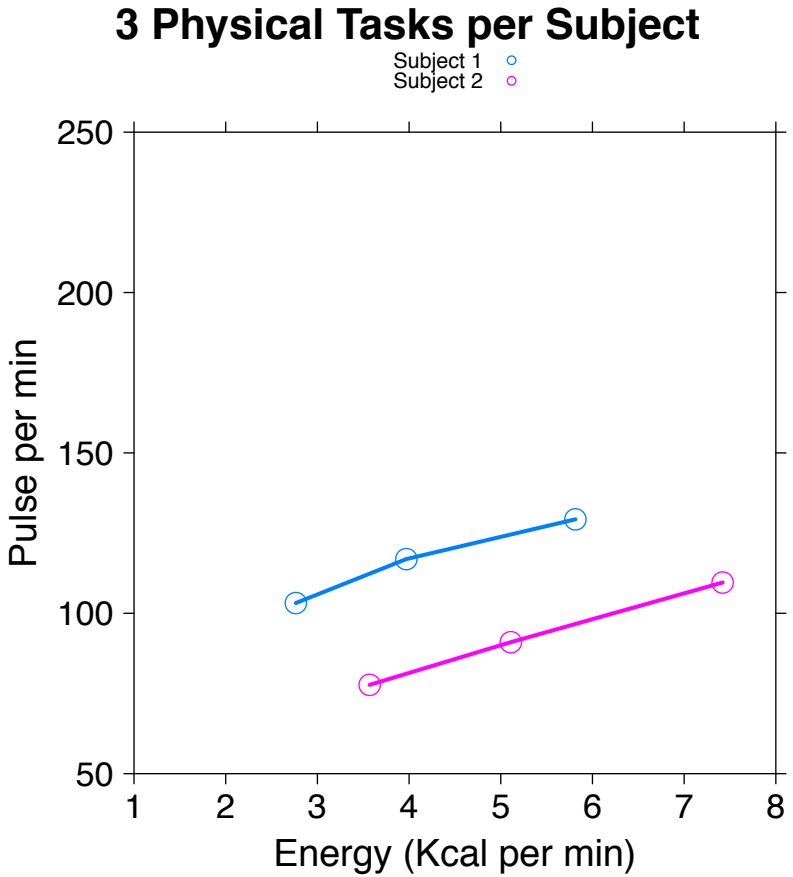
$$\boldsymbol{\Sigma}_i = \mathbf{J}_{n_i} c + \mathbf{I}_{n_i} (\sigma^2 - c)$$

4.3 Imbalance: Individual Measurement Designs

- Often difficult to assess each individual on exactly the same schedule as all others
- Heart rate example: different values of the independent variable used as predictors
  - y (DV, response): heart rate, measured during three exercises
  - x (IV, predictor): energy needed to complete the task
  - participants: 10 students
  - heart rate (H) linearly increases with energy use (E)

$$H_{ij} = \beta_0 + \beta_1 E_{ij} + e_{ij}$$

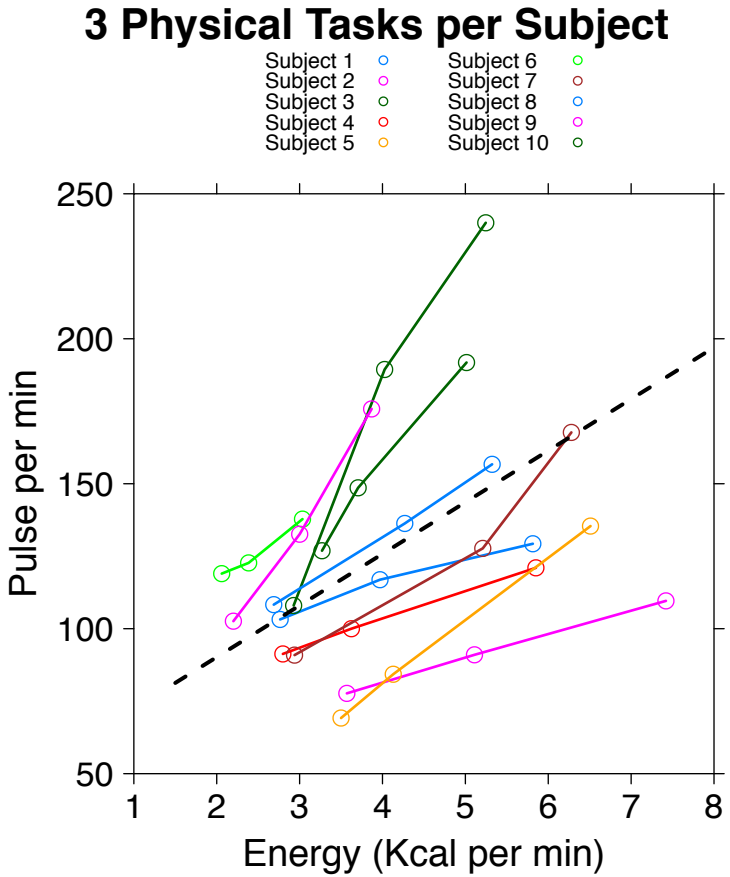
Imbalance: Individual Measurement Designs



- design matrix for the first two subjects (x is energy use, rather than time)

$$\mathbf{X}_1 = \begin{pmatrix} 1 & 2.765 \\ 1 & 3.970 \\ 1 & 5.814 \end{pmatrix} \qquad \mathbf{X}_2 = \begin{pmatrix} 1 & 3.570 \\ 1 & 5.110 \\ 1 & 7.420 \end{pmatrix}$$

Imbalance: Individual Measurement Designs



$$H_{ij} = \beta_0 + \beta_1 E_{ij} + e_{ij}$$

$$\hat{\beta} = \begin{pmatrix} 54.6 \\ 17.8 \end{pmatrix}$$

- mean function based on these parameters shown as dashed line in figure

Imbalance: Individual Measurement Designs

- Fixed measurement schedule: values of x common to all subjects
  - conditional mean: mean y-value for the scores over selected set of x-values
- Individual measurement designs: every individual has unique values on x
  - yet we still evaluate a *linear function* of x computed at any x-value
  - covariance matrix: each subject has n residuals (as many as nr. of measurements)

Summary

- ANOVA: one size fits all, assumptions oftentimes not met
- Regression models: contemporary approaches to the analysis of repeated measures data
  - Classical, simple regression: single design matrix applies to all subjects; common distribution of the residuals
  - Extensions for multiple groups and unbalanced data
  - Unique design matrix for each subject possible!
  - To be continued…

Literature

- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis*. New York: Oxford University Press.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.