Analysis and Sensitivity Analysis of Incomplete Data

Geert Molenberghs

geert.molenberghs@med.kuleuven.be

geert.molenberghs@uhasselt.be

Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat)

KU Leuven, Belgium & Universiteit Hasselt

www.ibiostat.be

RBras, Salvador, May 2016

Contents

1 Related References 1

I Longitudinal Data 8

2 The Rat Data 9

3 The Toenail Data 13

4 A Model for Longitudinal Data 17

5 The General Linear Mixed Model 22

6 Estimation and Inference 25

7 Generalized Estimating Equations 28

8 Generalized Linear Mixed Models (GLMM) 36

9 Fitting GLMM's in SAS 43

10 Marginal Versus Random-effects Models 46

II Incomplete Data 52

11 A Gentle Tour 53

12 Direct Likelihood / Ignorable Likelihood 75

13 Multiple Imputation 82

14 The Analgesic Trial 86

15 Creating Monotone Missingness 96

16 Case Study: The Dataset 109

17 Case Study: Weighted Generalized Estimating Equations 113

18 Case Study: Multiple Imputation 119

Chapter 1
Related References

- Aerts, M., Geys, H., Molenberghs, G., and Ryan, L.M. (2002). Topics in Modelling of Clustered Data. London: Chapman & Hall.

- Brown, H. and Prescott, R. (1999). *Applied Mixed Models in Medicine*. New York: John Wiley & Sons.

- Carpenter, J.R. and Kenward, M.G. (2013). *Multiple Imputation and its Application*. New York: John Wiley & Sons.

- Crowder, M.J. and Hand, D.J. (1990). *Analysis of Repeated Measures.* London: Chapman & Hall.

- Davidian, M. and Giltinan, D.M. (1995). *Nonlinear Models For Repeated Measurement Data.* London: Chapman & Hall.

- Davis, C.S. (2002). *Statistical Methods for the Analysis of Repeated Measurements.* New York: Springer.

- Demidenko, E. (2004). *Mixed Models: Theory and Applications.* New York: John Wiley & Sons.

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

- Fahrmeir, L. and Tutz, G. (2002). *Multivariate Statistical Modelling Based on Generalized Linear Models (2nd ed)*. New York: Springer.

- Fitzmaurice, G.M., Davidian, M., Verbeke, G., and Molenberghs, G.(2009). *Longitudinal Data Analysis. Handbook.* Hoboken, NJ: John Wiley & Sons.

- Fitzmaurice, G.M., Laird, N.M., and Ware, J.H. (2004). *Applied Longitudinal Analysis.* New York: John Wiley & Sons.

- Gałecki, A. and Burzykowski, T. (2013). *Linear Mixed-Effects Models Using R.* New York: Springer.

- Goldstein, H. (1979). *The Design and Analysis of Longitudinal Studies.* London: Academic Press.

- Goldstein, H. (1995). *Multilevel Statistical Models.* London: Edward Arnold.

- Hand, D.J. and Crowder, M.J. (1995). *Practical Longitudinal Data Analysis.* London: Chapman & Hall.

- Hedeker, D. and Gibbons, R.D. (2006). *Longitudinal Data Analysis*. New York: John Wiley & Sons.

- Jones, B. and Kenward, M.G. (1989). *Design and Analysis of Crossover Trials.* London: Chapman & Hall.

- Kshirsagar, A.M. and Smith, W.B. (1995). *Growth Curves.* New York: Marcel Dekker.

- Leyland, A.H. and Goldstein, H. (2001) *Multilevel Modelling of Health Statistics.* Chichester: John Wiley & Sons.

- Lindsey, J.K. (1993). *Models for Repeated Measurements.* Oxford: Oxford University Press.

- Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D., and Schabenberger, O. (2005). *SAS for Mixed Models (2nd ed.).* Cary: SAS Press.

- Longford, N.T. (1993). *Random Coefficient Models.* Oxford: Oxford University Press.

- McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models* (second edition). London: Chapman & Hall.

- Molenberghs, G., Fitzmaurice, G., Kenward, M.G., Tsiatis, A.A., and Verbeke, G. (2015). *Handbook of Missing Data*. Boca Raton: Chapman & Hall/CRC.

- Molenberghs, G. and Kenward, M.G. (2007). *Missing Data in Clinical Studies.* Chichester: John Wiley & Sons.

- Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data.* New York: Springer.

- Pinheiro, J.C. and Bates D.M. (2000). *Mixed effects models in S and S-Plus*. New York: Springer.

- Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data. With Applications in R.* Boca Raton: Chapman & Hall/CRC.

- Searle, S.R., Casella, G., and McCulloch, C.E. (1992). *Variance Components*. New-York: Wiley.
- Senn, S.J. (1993). *Cross-over Trials in Clinical Research.* Chichester: Wiley.
- Tan, M.T., Tian, G.-L., and Ng, K.W. (2010). *Bayesian Missing Data Problems.* Boca Raton: Chapman & Hall/CRC.
- van Buuren, S. (2012). *Flexible Imputation of Missing Data.* Boca Raton: Chapman & Hall/CRC.
- Verbeke, G. and Molenberghs, G. (1997). *Linear Mixed Models In Practice: A SAS Oriented Approach*, Lecture Notes in Statistics 126. New-York: Springer.
- Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics. New-York: Springer.

- Vonesh, E.F. and Chinchilli, V.M. (1997). *Linear and Non-linear Models for the Analysis of Repeated Measurements*. Basel: Marcel Dekker.

- Weiss, R.E. (2005). *Modeling Longitudinal Data.* New York: Springer.

- West, B.T., Welch, K.B., and Gałecki, A.T. (2007). *Linear Mixed Models: A Practical Guide Using Statistical Software*. Boca Raton: Chapman & Hall/CRC.

- Wu, H. and Zhang, J.-T. (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis.* New York: John Wiley & Sons.

- Wu, L. (2010). *Mixed Effects Models for Complex Data.* Boca Raton: Chapman & Hall/CRC.

Part I

Longitudinal Data

Chapter 2
The Rat Data

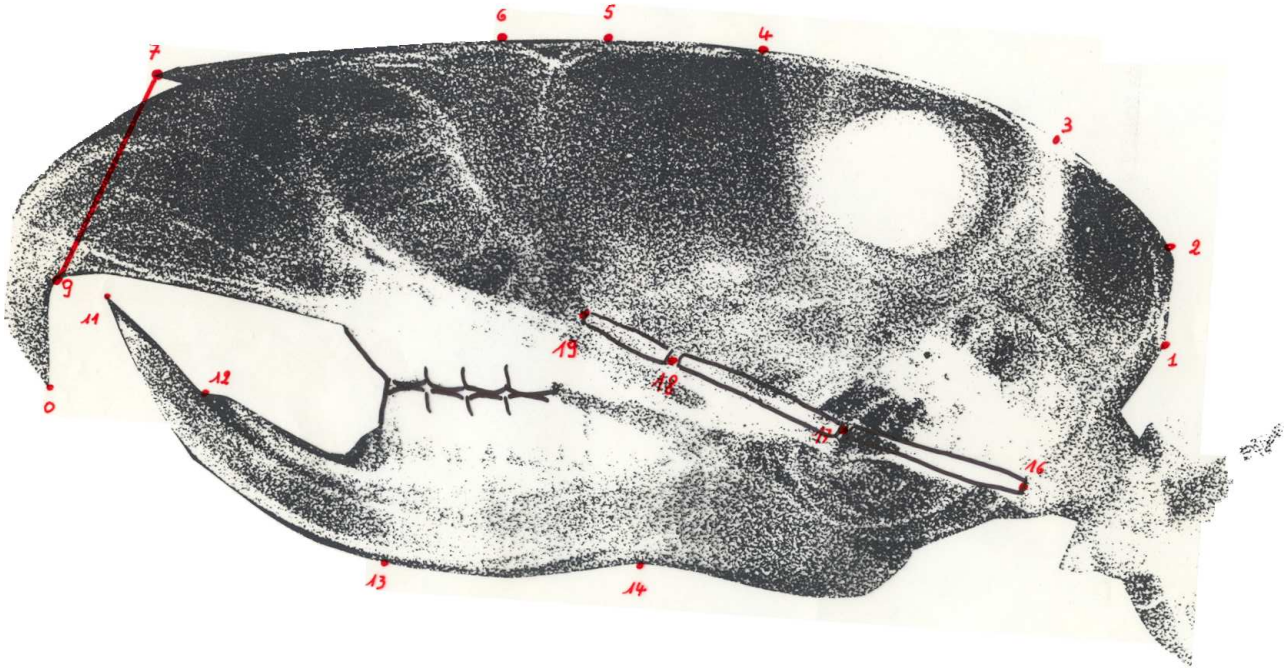
- Research question (Dentistry, K.U.Leuven):

> How does craniofacial growth depend on testosteron production?

- Randomized experiment in which 50 male Wistar rats are randomized to:
  - ▷ Control (15 rats)
  - ▷ Low dose of Decapeptyl (18 rats)
  - ▷ High dose of Decapeptyl (17 rats)

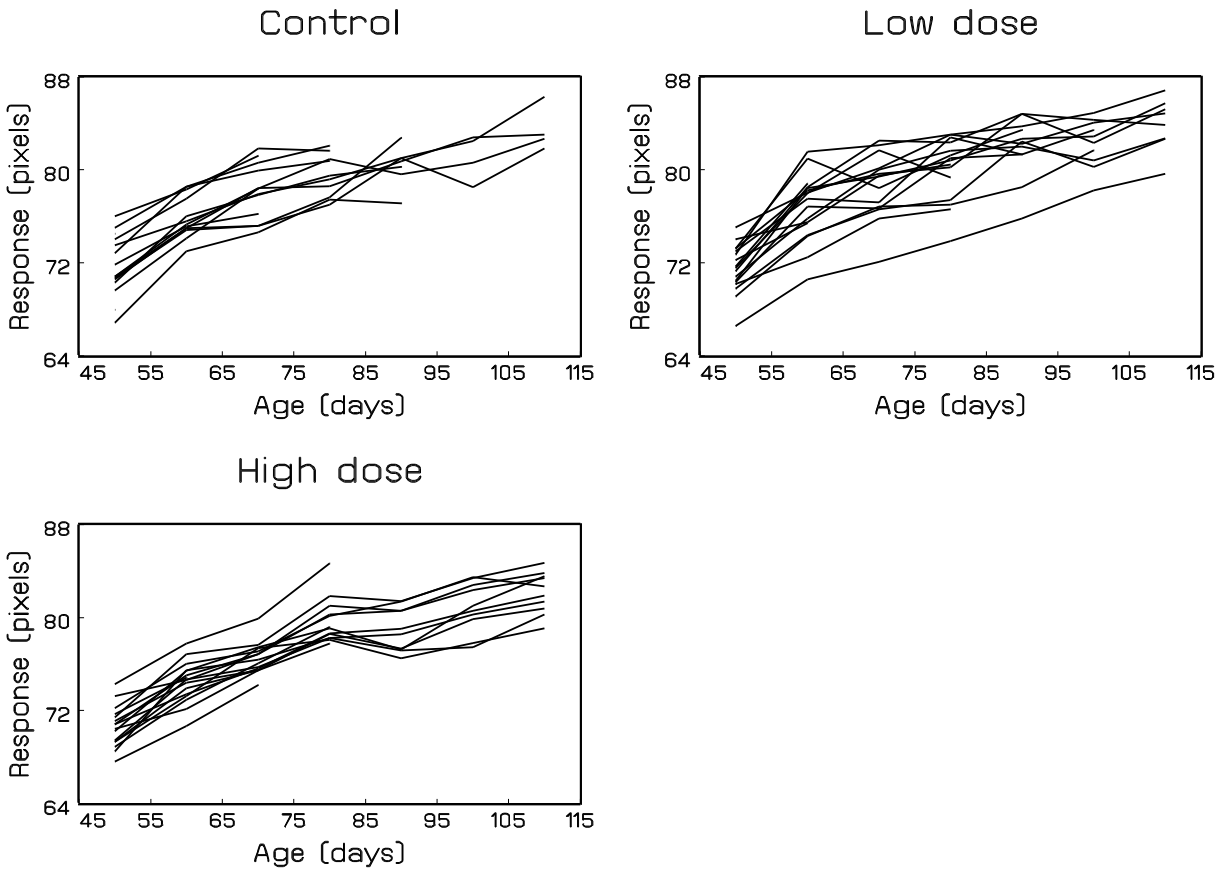
- Treatment starts at the age of 45 days; measurements taken every 10 days, from day 50 on.

- The responses are distances (pixels) between well defined points on x-ray pictures of the skull of each rat:



- Measurements with respect to the roof, base and height of the skull. Here, we consider only one response, reflecting the height of the skull.

- Individual profiles:



- Complication: Dropout due to anaesthesia (56%):

	# Observations			
Age (days)	Control	Low	High	Total
50	15	18	17	50
60	13	17	16	46
70	13	15	15	43
80	10	15	13	38
90	7	12	10	29
100	4	10	10	24
110	4	8	10	22

- Remarks:
  - ▷ A lot of variability between rats, much less variability within rats
  - ▷ Fixed number of measurements scheduled per subject, but not all measurements available due to dropout, for known reason.
  - ▷ Measurements taken at fixed time points

Chapter 3
The Toenail Data

- **T**oenail **D**ermatophyte **O**nychomycosis: Common toenail infection, difficult to treat, affecting more than 2% of population.

- Classical treatments with antifungal compounds need to be administered until the whole nail has grown out healthy.

- New compounds have been developed which reduce treatment to 3 months

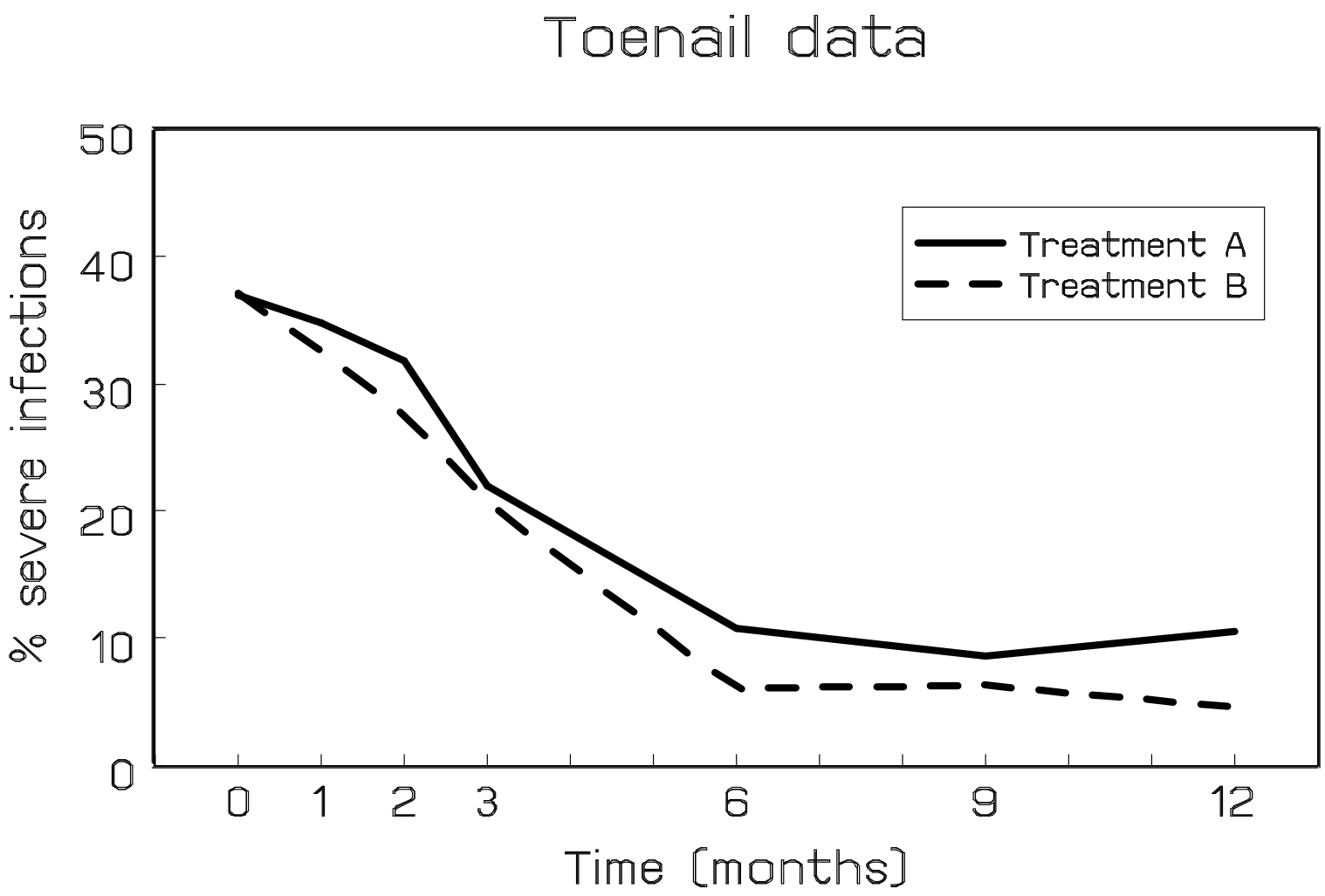
- Randomized, double-blind, parallel group, multicenter study for the comparison of two such new compounds (A and B) for oral treatment.

- Research question:

> Severity relative to treatment of TDO ?

- 2×189 patients randomized, 36 centers
- 48 weeks of total follow up (12 months)
- 12 weeks of treatment (3 months)
- measurements at months 0, 1, 2, 3, 6, 9, 12.

- Frequencies at each visit (both treatments):



3.1 Repeated Measures / Longitudinal Data

Repeated measures are obtained when a response is measured repeatedly on a set of units

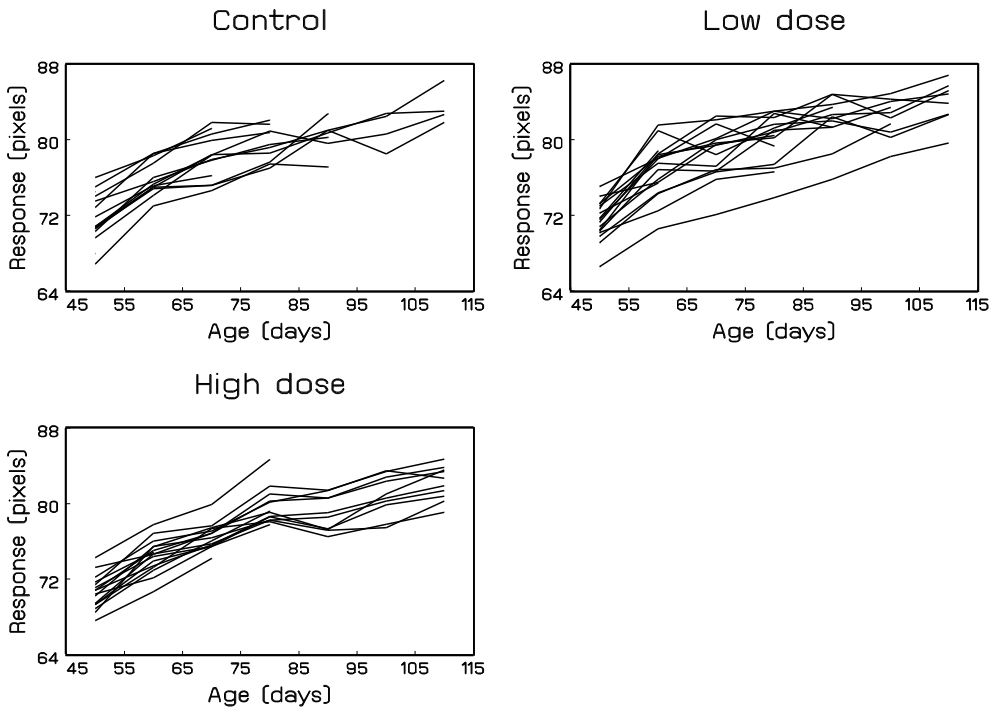
- Units:
  - ▷ Subjects, patients, participants, ...
  - ▷ Animals, plants, ...
  - ▷ Clusters: families, towns, branches of a company,...
  - ▷ ...
- Special case: **Longitudinal data**

Chapter 4
A Model for Longitudinal Data

- In practice: often unbalanced data:
  - ▷ unequal number of measurements per subject
  - ▷ measurements not taken at fixed time points
- Therefore, multivariate regression techniques are often not applicable
- Often, subject-specific longitudinal profiles can be well approximated by linear regression functions
- This leads to a 2-stage model formulation:
  - ▷ **Stage 1:** Linear regression model for each subject separately
  - ▷ **Stage 2:** Explain variability in the subject-specific regression coefficients using known covariates

4.1 Example: The Rat Data

- Individual profiles:



- Transformation of the time scale to linearize the profiles:

$$\text{Age}_{ij} \longrightarrow t_{ij} = \ln[1 + (\text{Age}_{ij} - 45)/10)]$$

- Note that $t = 0$ corresponds to the start of the treatment (moment of randomization)

- **Stage 1 model:** $Y_{ij} = \beta_{1i} + \beta_{2i} t_{ij} + \varepsilon_{ij}, \quad j = 1, \ldots, n_i$

- In the second stage, the subject-specific intercepts and time effects are related to the treatment of the rats

- **Stage 2 model:**

$$\begin{cases} \beta_{1i} = \beta_0 + b_{1i}, \\ \\ \beta_{2i} = \beta_1 L_i + \beta_2 H_i + \beta_3 C_i + b_{2i}, \end{cases}$$

- L_i, H_i, and C_i are indicator variables:

$$L_i = \begin{cases} 1 & \text{if low dose} \\ 0 & \text{otherwise} \end{cases} \quad H_i = \begin{cases} 1 & \text{if high dose} \\ 0 & \text{otherwise} \end{cases} \quad C_i = \begin{cases} 1 & \text{if control} \\ 0 & \text{otherwise} \end{cases}$$

- Parameter interpretation:
  - ▷ β_0: average response at the start of the treatment (independent of treatment)
  - ▷ β_1, β_2, and β_3: average time effect for each treatment group

4.2 The Linear Mixed-effects Model

- **Stage 1 model:** $$Y_{ij} = \beta_{1i} + \beta_{2i} t_{ij} + \varepsilon_{ij}, \quad j = 1, \ldots, n_i$$

- **Stage 2 model:** $$\begin{cases} \beta_{1i} = \beta_0 + b_{1i}, \\ \beta_{2i} = \beta_1 L_i + \beta_2 H_i + \beta_3 C_i + b_{2i}, \end{cases}$$

- **Combined:** $$Y_{ij} = (\beta_0 + b_{1i}) + (\beta_1 L_i + \beta_2 H_i + \beta_3 C_i + b_{2i}) t_{ij} + \varepsilon_{ij}$$

$$= \begin{cases} \beta_0 + b_{1i} + (\beta_1 + b_{2i}) t_{ij} + \varepsilon_{ij}, & \text{if low dose} \\ \beta_0 + b_{1i} + (\beta_2 + b_{2i}) t_{ij} + \varepsilon_{ij}, & \text{if high dose} \\ \beta_0 + b_{1i} + (\beta_3 + b_{2i}) t_{ij} + \varepsilon_{ij}, & \text{if control.} \end{cases}$$

Chapter 5
The General Linear Mixed Model

$$\boldsymbol{Y_i} = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i} + \boldsymbol{\varepsilon_i}$$

$$\boldsymbol{b_i} \sim N(\mathbf{0}, D),$$

$$\boldsymbol{\varepsilon_i} \sim N(\mathbf{0}, \Sigma_i),$$

$$\boldsymbol{b_1}, \ldots, \boldsymbol{b_N}, \boldsymbol{\varepsilon_1}, \ldots, \boldsymbol{\varepsilon_N}$$
independent

Terminology:

- ▷ Fixed effects: $\boldsymbol{\beta}$
- ▷ Random effects: $\boldsymbol{b_i}$
- ▷ Variance components:
  elements in D and Σ_i

5.1 Hierarchical versus Marginal Model

- The general linear mixed model is given by:

$$\boldsymbol{Y_i} = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i} + \boldsymbol{\varepsilon_i}$$

$$\begin{cases} \boldsymbol{b_i} \sim N(\mathbf{0}, D), \\ \boldsymbol{\varepsilon_i} \sim N(\mathbf{0}, \Sigma_i), \\ \boldsymbol{b_1}, \ldots, \boldsymbol{b_N}, \boldsymbol{\varepsilon_1}, \ldots, \boldsymbol{\varepsilon_N} \text{ independent}, \end{cases}$$

- It can be rewritten as:

$$\boldsymbol{Y_i}|\boldsymbol{b_i} \sim N(X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i}, \Sigma_i)$$

$$\boldsymbol{b_i} \sim N(\mathbf{0}, D)$$

- It is therefore also called a hierarchical model:
  - ▷ A model for $\boldsymbol{Y_i}$ given $\boldsymbol{b_i}$
  - ▷ A model for $\boldsymbol{b_i}$

- Marginally, we have that $\boldsymbol{Y_i}$ is distributed as: $\boldsymbol{Y_i} \sim N(X_i\boldsymbol{\beta}, Z_iDZ_i' + \Sigma_i)$

- Hence, very specific assumptions are made about the dependence of mean and covariance on the covariates X_i and Z_i:
  - ▷ **Implied mean :** $X_i\boldsymbol{\beta}$
  - ▷ **Implied covariance :** $V_i = Z_iDZ_i' + \Sigma_i$

- Note that the hierarchical model implies the marginal one, **not** vice versa

Chapter 6
Estimation and Inference

- Notation:
  - ▷ $\boldsymbol{\beta}$: vector of fixed effects (as before)
  - ▷ $\boldsymbol{\alpha}$: vector of all variance components in D and Σ_i
  - ▷ $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$: vector of all parameters in marginal model

- In most cases, $\boldsymbol{\alpha}$ is not known, and needs to be replaced by an estimate $\widehat{\boldsymbol{\alpha}}$

- Two frequently used estimation methods for $\boldsymbol{\alpha}$:
  - ▷ Maximum likelihood
  - ▷ Restricted maximum likelihood

6.1 Inference

- Inference for $\boldsymbol{\beta}$:
  - ▷ Wald tests, t- and F-tests
  - ▷ LR tests (not with REML)
- Inference for $\boldsymbol{\alpha}$:
  - ▷ Wald tests
  - ▷ LR tests (even with REML)
  - ▷ Caution: Boundary problems !
- Inference for the random effects:
  - ▷ Empirical Bayes inference based on posterior density $f(\boldsymbol{b_i}|\boldsymbol{Y_i} = \boldsymbol{y_i})$
  - ▷ 'Empirical Bayes (EB) estimate': Posterior mean

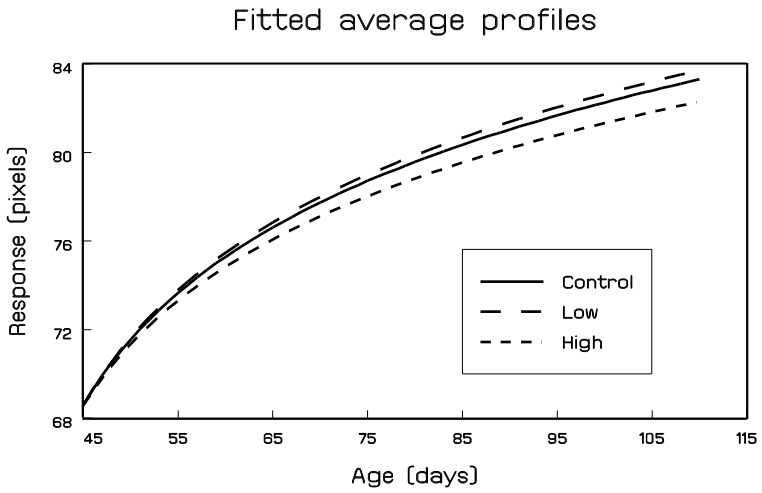
6.2 Fitting Linear Mixed Models in SAS

- A model for the rat data: $Y_{ij} = (\beta_0 + b_{1i}) + (\beta_1 L_i + \beta_2 H_i + \beta_3 C_i + b_{2i}) t_{ij} + \varepsilon_{ij}$

- SAS program:

```
proc mixed data=rat method=reml;
class id group;
model y = t group*t / solution;
random intercept t / type=un subject=id ;
run;
```

- Fitted averages:



Chapter 7
Generalized Estimating Equations

- Univariate GLM, score function of the form (scalar Y_i):

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} v_i^{-1}(y_i - \mu_i) = \mathbf{0}, \quad \text{with } v_i = \text{Var}(Y_i).$$

- In longitudinal setting: $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N)$:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} D_i' \left[\boldsymbol{V}_i(\boldsymbol{\alpha})\right]^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$$

  where

  - ▷ D_i is an $n_i \times p$ matrix with (i,j)th elements $\dfrac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}}$
  - ▷ Is V_i $n_i \times n_i$ diagonal?
  - ▷ $\boldsymbol{y}_i$ and $\boldsymbol{\mu}_i$ are n_i-vectors with elements y_{ij} and μ_{ij}

7.1 Large Sample Properties: Naive Approach

As $N \to \infty$

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim N(\mathbf{0}, I_0^{-1})$$

where

$$I_0 = \sum_{i=1}^{N} D_i'[V_i(\boldsymbol{\alpha})]^{-1} D_i$$

- **(Unrealistic) Conditions:**
  - ▷ $\boldsymbol{\alpha}$ is known
  - ▷ the parametric form for $\boldsymbol{V}_i(\boldsymbol{\alpha})$ is known
- **Solution: working correlation matrix**

7.2 Unknown Covariance Structure

Keep the score equations

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} [D_i]' [V_i(\boldsymbol{\alpha})]^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$$

BUT

- suppose $\boldsymbol{V}_i(.)$ is not the true variance of $\boldsymbol{Y}_i$ but only a plausible guess, a so-called working correlation matrix
- specify correlations and not covariances, because the variances follow from the mean structure
- the score equations are solved as before

7.3 Asymptotic Properties: Robust Approach

The asymptotic normality results change to (**sandwich estimator**)

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim N(\mathbf{0}, I_0^{-1} I_1 I_0^{-1})$$

$$I_0 = \sum_{i=1}^{N} D_i'[V_i(\boldsymbol{\alpha})]^{-1} D_i$$

$$I_1 = \sum_{i=1}^{N} D_i'[V_i(\boldsymbol{\alpha})]^{-1} \mathsf{Var}(\boldsymbol{Y}_i)[V_i(\boldsymbol{\alpha})]^{-1} D_i.$$

- The estimators $\hat{\boldsymbol{\beta}}$ are consistent even if the working correlation matrix is incorrect
- An estimate is found by replacing the unknown variance matrix $\mathsf{Var}(\boldsymbol{Y}_i)$ by

$$(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)'.$$

7.4 Application to the Toenail Data

7.4.1 The model

- Consider the model:

$$Y_{ij} \sim \text{Bernoulli}(\mu_{ij}), \qquad \log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_0 + \beta_1 T_i + \beta_2 t_{ij} + \beta_3 T_i t_{ij}$$

- Y_{ij}: severe infection (yes/no) at occasion j for patient i
- t_{ij}: measurement time for occasion j
- T_i: treatment group

7.4.2 Standard GEE

- **SAS Code:**

```
proc genmod data=test descending;
class idnum timeclss;
model onyresp = treatn time treatn*time
                / dist=binomial;
repeated subject=idnum / withinsubject=timeclss
                         type=exch covb corrw modelse;
run;
```

- SAS statements:
  - ▷ The REPEATED statements defines the GEE character of the model.
  - ▷ 'type=': working correlation specification (UN, AR(1), EXCH, IND,...)
  - ▷ 'modelse': model-based s.e.'s on top of default empirically corrected s.e.'s
  - ▷ 'corrw': printout of working correlation matrix
  - ▷ 'withinsubject=': specification of the ordering within subjects

- **Selected output:**
  - ▷ Regression paramters:

Analysis Of Initial Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square
Intercept	1	-0.5571	0.1090	-0.7708	-0.3433	26.10
treatn	1	0.0240	0.1565	-0.2827	0.3307	0.02
time	1	-0.1769	0.0246	-0.2251	-0.1288	51.91
treatn*time	1	-0.0783	0.0394	-0.1556	-0.0010	3.95
Scale	0	1.0000	0.0000	1.0000	1.0000	

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	-0.5840	0.1734	-0.9238	-0.2441	-3.37	0.0008
treatn	0.0120	0.2613	-0.5001	0.5241	0.05	0.9633
time	-0.1770	0.0311	-0.2380	-0.1161	-5.69	<.0001
treatn*time	-0.0886	0.0571	-0.2006	0.0233	-1.55	0.1208

Analysis Of GEE Parameter Estimates
Model-Based Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	-0.5840	0.1344	-0.8475	-0.3204	-4.34	<.0001
treatn	0.0120	0.1866	-0.3537	0.3777	0.06	0.9486
time	-0.1770	0.0209	-0.2180	-0.1361	-8.47	<.0001
treatn*time	-0.0886	0.0362	-0.1596	-0.0177	-2.45	0.0143

- The working correlation:

Exchangeable Working Correlation

Correlation 0.420259237

Chapter 8
Generalized Linear Mixed Models (GLMM)

- Given a vector $\boldsymbol{b_i}$ of random effects for cluster i, it is assumed that all responses Y_{ij} are independent, with density

$$f(y_{ij}|\theta_{ij},\phi) \;=\; \exp\left\{\phi^{-1}[y_{ij}\theta_{ij} - \psi(\theta_{ij})] + c(y_{ij},\phi)\right\}$$

- θ_{ij} is now modelled as $\theta_{ij} \;=\; \boldsymbol{x_{ij}}'\boldsymbol{\beta} \;+\; \boldsymbol{z_{ij}}'\boldsymbol{b_i}$

- As before, it is assumed that $\boldsymbol{b_i} \;\sim\; N(\mathbf{0}, D)$

- Let $f_{ij}(y_{ij}|\boldsymbol{b_i}, \boldsymbol{\beta}, \phi)$ denote the conditional density of Y_{ij} given $\boldsymbol{b_i}$, the conditional density of $\boldsymbol{Y_i}$ equals

$$f_i(\boldsymbol{y_i}|\boldsymbol{b_i}, \boldsymbol{\beta}, \phi) = \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{b_i}, \boldsymbol{\beta}, \phi)$$

- The marginal distribution of $\boldsymbol{Y_i}$ is given by

$$f_i(\boldsymbol{y_i}|\boldsymbol{\beta}, D, \phi) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{b_i}, \boldsymbol{\beta}, \phi)\ f(\boldsymbol{b_i}|D)\ d\boldsymbol{b_i}$$

  where $f(\boldsymbol{b_i}|D)$ is the density of the $N(\boldsymbol{0}, D)$ distribution.

- The likelihood function for $\boldsymbol{\beta}$, D, and ϕ now equals

$$\begin{aligned} L(\boldsymbol{\beta}, D, \phi) &= \prod_{i=1}^{N} f_i(\boldsymbol{y_i}|\boldsymbol{\beta}, D, \phi) \\ &= \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{b_i}, \boldsymbol{\beta}, \phi)\ f(\boldsymbol{b_i}|D)\ d\boldsymbol{b_i} \end{aligned}$$

- Under the normal linear model, the integral can be worked out analytically.

- In general, approximations are required:
  - ▷ Approximation of integrand: Laplace approximation
  - ▷ Approximation of data: Taylor series
  - ▷ Approximation of integral: (Adaptive) Gaussian quadrature

- Predictions of random effects can be based on the posterior distribution

$$f(\boldsymbol{b_i}|\boldsymbol{Y_i} = \boldsymbol{y_i})$$

- 'Empirical Bayes (EB) estimate':
  Posterior mode, with unknown parameters replaced by their MLE

8.1 Example: Toenail Data

- Y_{ij} is binary severity indicator for subject i at visit j.
- Model:

$$Y_{ij}|b_i \sim \text{Bernoulli}(\pi_{ij}), \qquad \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + b_i + \beta_1 T_i + \beta_2 t_{ij} + \beta_3 T_i t_{ij}$$

- Notation:
  - $\triangleright$ T_i: treatment indicator for subject i
  - $\triangleright$ t_{ij}: time point at which jth measurement is taken for ith subject
- Adaptive as well as non-adaptive Gaussian quadrature, for various Q.

- Results:

	Gaussian quadrature				
	$Q = 3$	$Q = 5$	$Q = 10$	$Q = 20$	$Q = 50$
β_0	-1.52 (0.31)	-2.49 (0.39)	-0.99 (0.32)	-1.54 (0.69)	-1.65 (0.43)
β_1	-0.39 (0.38)	0.19 (0.36)	0.47 (0.36)	-0.43 (0.80)	-0.09 (0.57)
β_2	-0.32 (0.03)	-0.38 (0.04)	-0.38 (0.05)	-0.40 (0.05)	-0.40 (0.05)
β_3	-0.09 (0.05)	-0.12 (0.07)	-0.15 (0.07)	-0.14 (0.07)	-0.16 (0.07)
σ	2.26 (0.12)	3.09 (0.21)	4.53 (0.39)	3.86 (0.33)	4.04 (0.39)
-2ℓ	1344.1	1259.6	1254.4	1249.6	1247.7
	Adaptive Gaussian quadrature				
	$Q = 3$	$Q = 5$	$Q = 10$	$Q = 20$	$Q = 50$
β_0	-2.05 (0.59)	-1.47 (0.40)	-1.65 (0.45)	-1.63 (0.43)	-1.63 (0.44)
β_1	-0.16 (0.64)	-0.09 (0.54)	-0.12 (0.59)	-0.11 (0.59)	-0.11 (0.59)
β_2	-0.42 (0.05)	-0.40 (0.04)	-0.41 (0.05)	-0.40 (0.05)	-0.40 (0.05)
β_3	-0.17 (0.07)	-0.16 (0.07)	-0.16 (0.07)	-0.16 (0.07)	-0.16 (0.07)
σ	4.51 (0.62)	3.70 (0.34)	4.07 (0.43)	4.01 (0.38)	4.02 (0.38)
-2ℓ	1259.1	1257.1	1248.2	1247.8	1247.8

- Conclusions:
  - ▷ (Log-)likelihoods are not comparable
  - ▷ Different Q can lead to considerable differences in estimates and standard errors
  - ▷ For example, using non-adaptive quadrature, with $Q = 3$, we found no difference in time effect between both treatment groups ($t = -0.09/0.05, p = 0.0833$).
  - ▷ Using adaptive quadrature, with $Q = 50$, we find a significant interaction between the time effect and the treatment ($t = -0.16/0.07, p = 0.0255$).
  - ▷ Assuming that $Q = 50$ is sufficient, the 'final' results are well approximated with smaller Q under adaptive quadrature, but not under non-adaptive quadrature.

- Comparison of fitting algorithms:
  - ▷ Adaptive Gaussian Quadrature, $Q = 50$
  - ▷ MQL and PQL

- Summary of results:

Parameter	QUAD	PQL	MQL
Intercept group A	−1.63 (0.44)	−0.72 (0.24)	−0.56 (0.17)
Intercept group B	−1.75 (0.45)	−0.72 (0.24)	−0.53 (0.17)
Slope group A	−0.40 (0.05)	−0.29 (0.03)	−0.17 (0.02)
Slope group B	−0.57 (0.06)	−0.40 (0.04)	−0.26 (0.03)
Var. random intercepts (τ^2)	15.99 (3.02)	4.71 (0.60)	2.49 (0.29)

- Severe differences between QUAD (gold standard ?) and MQL/PQL.

- MQL/PQL may yield (very) biased results, especially for binary data.

Chapter 9
Fitting GLMM's in SAS

- **GLIMMIX:** Laplace, MQL, PQL, adaptive quadrature

- **NLMIXED:** Adaptive and non-adaptive quadrature
  $\longrightarrow$ not discussed here

9.1 Example: Toenail data

- Re-consider logistic model with random intercepts for toenail data
- SAS code (PQL):

```
proc glimmix data=test method=RSPL ;
class idnum;
model onyresp (event='1') = treatn time treatn*time
                          / dist=binary solution;
random intercept / subject=idnum;
run;
```

- MQL obtained with option 'method=RMPL'
- Laplace obtained with option 'method=LAPLACE'

- Adaptive quadrature with option 'method=QUAD(qpoints=5)'

- Inclusion of random slopes:

```
random intercept time / subject=idnum type=un;
```

Chapter 10
Marginal Versus Random-effects Models

- We compare our GLMM results for the toenail data with those from fitting GEE's (unstructured working correlation):

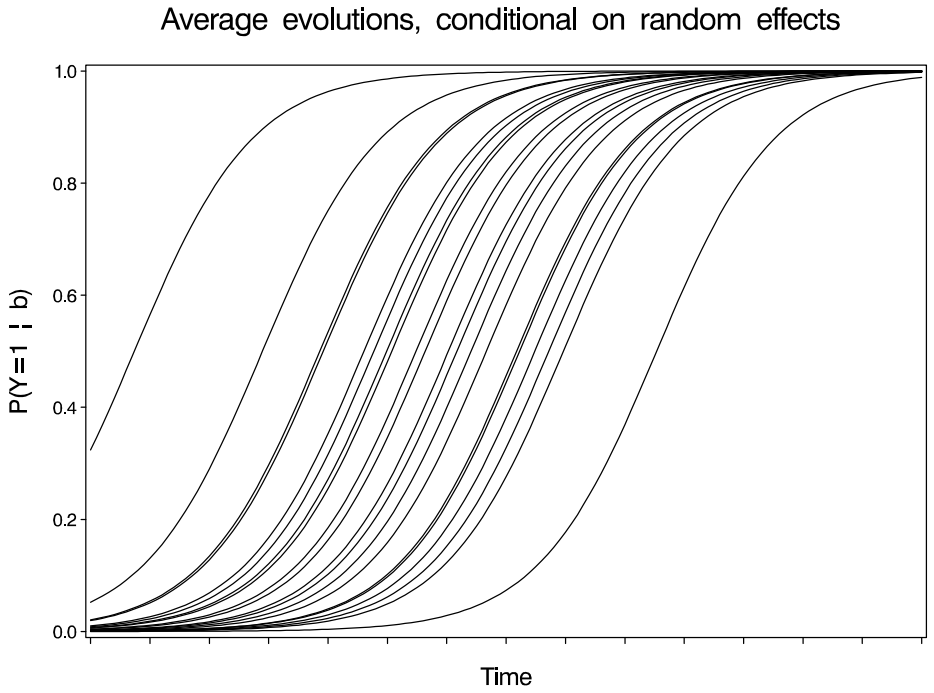
	GLMM	GEE
Parameter	Estimate (s.e.)	Estimate (s.e.)
Intercept group A	−1.6308 (0.4356)	−0.7219 (0.1656)
Intercept group B	−1.7454 (0.4478)	−0.6493 (0.1671)
Slope group A	−0.4043 (0.0460)	−0.1409 (0.0277)
Slope group B	−0.5657 (0.0601)	−0.2548 (0.0380)

- The strong differences can be explained as follows:
  - ▷ Consider the following GLMM:

$$Y_{ij}|b_i \sim \text{Bernoulli}(\pi_{ij}), \qquad \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + b_i + \beta_1 t_{ij}$$

  - ▷ The conditional means $E(Y_{ij}|b_i)$, as functions of t_{ij}, are given by

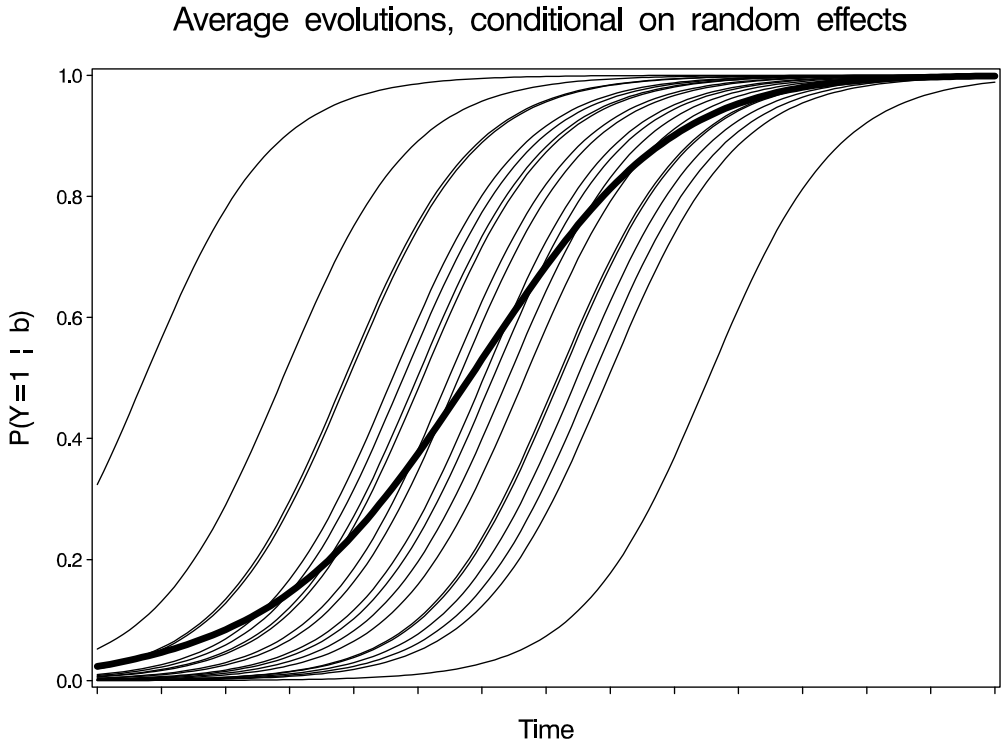
$$E(Y_{ij}|b_i) = \frac{\exp(\beta_0 + b_i + \beta_1 t_{ij})}{1 + \exp(\beta_0 + b_i + \beta_1 t_{ij})}$$



▷ The marginal average evolution is now obtained from averaging over the random effects:

$$E(Y_{ij}) = E[E(Y_{ij}|b_i)] = E\left[\frac{\exp(\beta_0 + b_i + \beta_1 t_{ij})}{1 + \exp(\beta_0 + b_i + \beta_1 t_{ij})}\right]$$

$$\neq \frac{\exp(\beta_0 + \beta_1 t_{ij})}{1 + \exp(\beta_0 + \beta_1 t_{ij})}$$



- Hence, the parameter vector $\boldsymbol{\beta}$ in the GEE model needs to be interpreted completely different from the parameter vector $\boldsymbol{\beta}$ in the GLMM:
  - ▷ GEE: marginal interpretation
  - ▷ GLMM: conditional interpretation, conditionally upon level of random effects

- For logistic mixed models, with normally distributed random random intercepts, it can be shown that the marginal model can be well approximated by again a logistic model, but with parameters approximately satisfying

$$\frac{\widehat{\boldsymbol{\beta}}^{\mathsf{RE}}}{\widehat{\boldsymbol{\beta}}^{\mathsf{M}}} = \sqrt{c^2\sigma^2 + 1} > 1, \qquad \sigma^2 = \text{variance random intercepts}$$

$$c = 16\sqrt{3}/(15\pi)$$

- For the toenail application, σ was estimated as 4.0164, such that the ratio equals $\sqrt{c^2\sigma^2 + 1} = 2.5649$.

	GLMM	GEE	
Parameter	Estimate (s.e.)	Estimate (s.e.)	Ratio
Intercept group A	−1.6308 (0.4356)	−0.7219 (0.1656)	2.2590
Intercept group B	−1.7454 (0.4478)	−0.6493 (0.1671)	2.6881
Slope group A	−0.4043 (0.0460)	−0.1409 (0.0277)	2.8694
Slope group B	−0.5657 (0.0601)	−0.2548 (0.0380)	2.2202

- Note that this problem does not occur in linear mixed models:
  - ▷ Conditional mean: $E(\boldsymbol{Y_i}|\boldsymbol{b}_i) = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i$
  - ▷ Specifically: $E(\boldsymbol{Y_i}|\boldsymbol{b}_i = \mathbf{0}) = X_i\boldsymbol{\beta}$
  - ▷ Marginal mean: $E(\boldsymbol{Y_i}) = X_i\boldsymbol{\beta}$

- For GLMM: $\boxed{E[g(Y)] \neq g[E(Y)]}$

10.1 Toenail Data: Overview

- Overview of all analyses on toenail data:

Parameter	QUAD	PQL	MQL	GEE
Intercept group A	−1.63 (0.44)	−0.72 (0.24)	−0.56 (0.17)	−0.72 (0.17)
Intercept group B	−1.75 (0.45)	−0.72 (0.24)	−0.53 (0.17)	−0.65 (0.17)
Slope group A	−0.40 (0.05)	−0.29 (0.03)	−0.17 (0.02)	−0.14 (0.03)
Slope group B	−0.57 (0.06)	−0.40 (0.04)	−0.26 (0.03)	−0.25 (0.04)
Var. random intercepts (τ^2)	15.99 (3.02)	4.71 (0.60)	2.49 (0.29)	

- Conclusion:

$$|\text{GEE}| < |\text{MQL}| < |\text{PQL}| < |\text{QUAD}|$$

Part II

Incomplete Data

Chapter 11
A Gentle Tour

- Orthodontic growth data
- Commonly used methods
- Survey of the terrain

11.1 Growth Data: An (Un)balanced Discussion

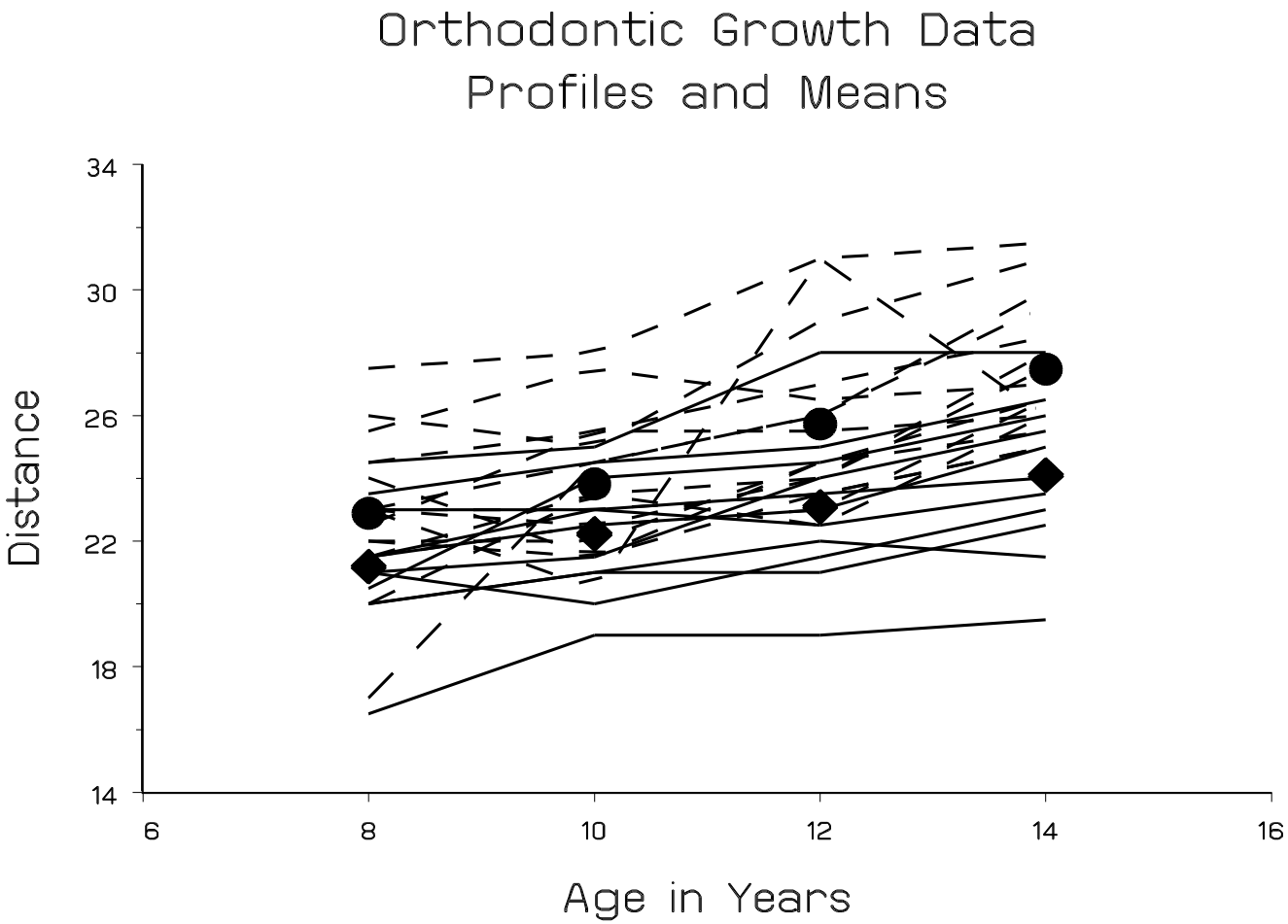
- Taken from Potthoff and Roy, Biometrika (1964)

- Research question:

> Is dental growth related to gender ?

- The distance from the center of the pituitary to the maxillary fissure was recorded at ages 8, 10, 12, and 14, for 11 girls and 16 boys

- Individual profiles:
  - ▷ **Un**balanced data
  - ▷ Balanced data



11.2 LOCF, CC, or Direct Likelihood?

Data:

20	30
10	40

75
25

LOCF:

20	30
10	40

75	0
0	25

$\Longrightarrow$

95	30
10	65

$\Longrightarrow \quad \widehat{\theta} = \frac{95}{200} = 0.475 \ [0.406; 0.544]$ (biased & too narrow)

CC:

20	30
10	40

0	0
0	0

$\Longrightarrow$

20	30
10	40

$\Longrightarrow \quad \widehat{\theta} = \frac{20}{100} = 0.200 \ [0.122; 0.278]$ (biased & too wide)

d.I.(MAR):

20	30
10	40

30	45
5	20

$\Longrightarrow$

50	75
15	60

$\Longrightarrow \quad \widehat{\theta} = \frac{50}{200} = 0.250 \ [0.163; 0.337]$

11.3 Direct Likelihood/Bayesian Inference: Ignorability

$$\boxed{\text{MAR}}: \; f(\boldsymbol{Y}_i^o|\boldsymbol{X}_i, \boldsymbol{\theta}) \; \cancel{f(\boldsymbol{r}_i|\boldsymbol{X}_i, \boldsymbol{Y}_i^o, \boldsymbol{\psi})}$$

Mechanism is MAR

$\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ distinct

Interest in $\boldsymbol{\theta}$

(Use observed information matrix)

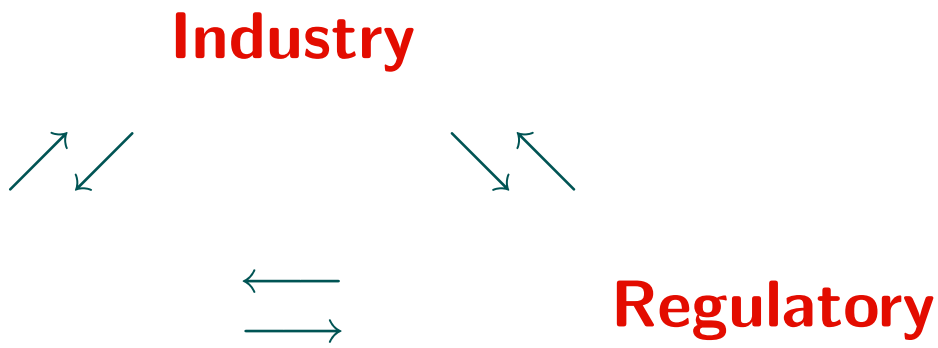
$\Longrightarrow$ Lik./Bayes inference valid

Outcome type	**Modeling strategy**	**Software**
Gaussian	Linear mixed model	SAS MIXED
Non-Gaussian	Gen./Non-linear mixed model	SAS GLIMMIX, NLMIXED

11.4 Rubin, 1976

- Ignorability: Rubin (Biometrika, 1976): 35 years ago!
- Little and Rubin (1976, 2002)
- Why did it take so long?

11.5 A Vicious Triangle



- **Academe:** The R^2 principle
- **Regulatory:** Rigid procedures $\longleftrightarrow$ scientific developments
- **Industry:** We **cannot / do not want to** apply new methods

11.6 Terminology & Confusion

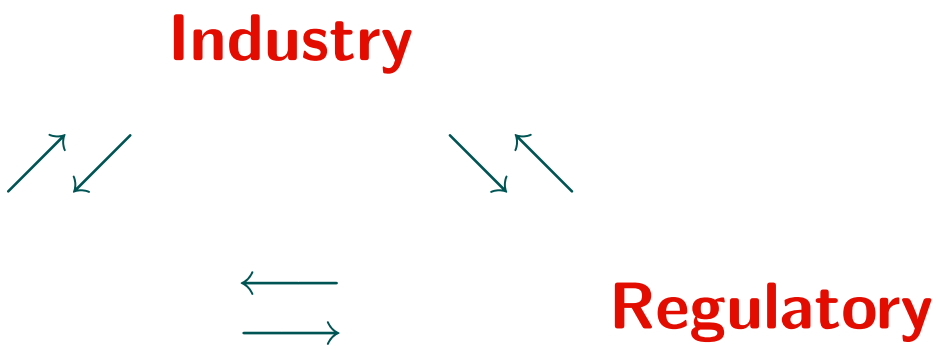
- The Ministry of Disinformation:

$\longleftarrow$ **All directions**

Other directions $\longrightarrow$

- **MCAR, MAR, MNAR:** "What do the terms mean?"
- **MAR, random dropout, informative missingness, ignorable, censoring,. . .**
- **Dropout from the study, dropout from treatment, lost to follow up,. . .**
- **"Under MAR patients dropping out and patients not dropping out are similar."**

11.7 A Virtuous Triangle



- FDA/Industry Workshops
- DIA/EMA Meetings
- **The NAS Experience**

11.8 The NAS Experience: A Wholesome Product

- **FDA $\longrightarrow$ NAS $\longrightarrow$ the working group**
- **Composition**
- **Encompassing:**
  - ▷ terminology/taxonomy/concepts
  - ▷ prevention
  - ▷ treatment

11.9 Taxonomy

- **Missingness pattern:** complete — monotone — non-monotone
- Dropout pattern: complete — dropout — intermittent
- **Model framework:** SEM — PMM — SPM
- **Missingness mechanism:** MCAR — MAR — MNAR
- **Ignorability:** ignorable — non-ignorable
- **Inference paradigm:** frequentist — likelihood — Bayes

11.10 The NAS Panel

Name	Specialty	Affiliation
Rod Little	biostat	U Michigan
Ralph D'Agostino	biostat	Boston U
Kay Dickerson	epi	Johns Hopkins
Scott Emerson	biostat	U Washington
John Farrar	epi	U Penn
Constantine Frangakis	biostat	Johns Hopkins
Joseph Hogan	biostat	Brown U
Geert Molenberghs	biostat	U Hasselt & K.U.Leuven
Susan Murphy	stat	U Michigan
James Neaton	biostat	U Minnesota
Andrea Rotnitzky	stat	Buenos Aires & Harvard
Dan Scharfstein	biostat	Johns Hopkins
Joseph Shih	biostat	New Jersey SPH
Jay Siegel	biostast	J&J
Hal Stern	stat	UC at Irvine

11.11 Modeling Frameworks & Missing Data Mechanisms

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, \boldsymbol{\theta}, \boldsymbol{\psi})$$

Selection Models: $f(\boldsymbol{y}_i | X_i, \boldsymbol{\theta}) \boxed{f(\boldsymbol{r}_i | X_i, \boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{\psi})}$

MCAR $\longrightarrow$ **MAR** $\longrightarrow$ **MNAR**

$f(\boldsymbol{r}_i | X_i, \boldsymbol{\psi})$ $\qquad$ $f(\boldsymbol{r}_i | X_i, \boldsymbol{y}_i^o, \boldsymbol{\psi})$ $\qquad$ $f(\boldsymbol{r}_i | X_i, \boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{\psi})$

Pattern-mixture Models: $f(\boldsymbol{y}_i | X_i, \boldsymbol{r}_i, \boldsymbol{\theta})\, f(\boldsymbol{r}_i | X_i, \boldsymbol{\psi})$

Shared-parameter Models: $f(\boldsymbol{y}_i | X_i, \boldsymbol{b}_i, \boldsymbol{\theta})\, f(\boldsymbol{r}_i | X_i, \boldsymbol{b}_i, \boldsymbol{\psi})$

11.12 Frameworks and Their Methods

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, \boldsymbol{\theta}, \boldsymbol{\psi})$$

Selection Models: $f(\boldsymbol{y}_i | X_i, \boldsymbol{\theta}) \boxed{f(\boldsymbol{r}_i | X_i, \boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{\psi})}$

MCAR/simple	$\longrightarrow$	**MAR**	$\longrightarrow$	**MNAR**
CC?		**direct likelihood!**		joint model!?
LOCF?		**direct Bayesian!**		sensitivity analysis?!
single imputation?		**multiple imputation (MI)!**		
⋮		**IPW $\supset$ W-GEE!**		
		d.i. + IPW = double robustness! (consensus)		

11.13 Frameworks and Their Methods: Start

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, \boldsymbol{\theta}, \boldsymbol{\psi})$$

Selection Models: $f(\boldsymbol{y}_i | X_i, \boldsymbol{\theta}) \boxed{f(\boldsymbol{r}_i | X_i, \boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{\psi})}$

MCAR/simple $\longrightarrow$ **MAR** $\longrightarrow$ **MNAR**

direct likelihood!

direct Bayesian!

multiple imputation (MI)!

IPW $\supset$ **W-GEE!**

d.I. + IPW = double robustness!

11.14 Frameworks and Their Methods: Next

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, \boldsymbol{\theta}, \boldsymbol{\psi})$$

Selection Models: $f(\boldsymbol{y}_i | X_i, \boldsymbol{\theta}) \boxed{f(\boldsymbol{r}_i | X_i, \boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{\psi})}$

MCAR/simple $\longrightarrow$ **MAR** $\longrightarrow$ **MNAR**

~~joint model!?~~

sensitivity analysis!

PMM

MI (MGK, J&J)

local influence

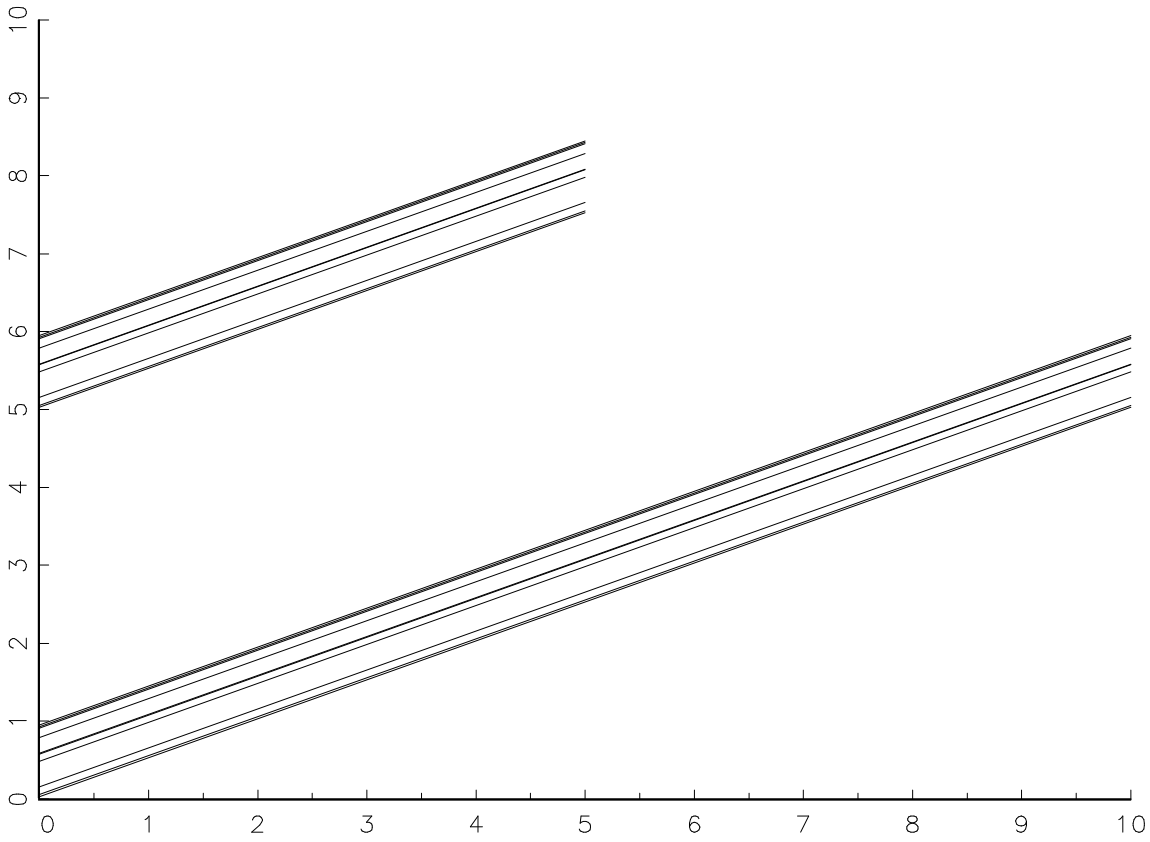
interval ignorance

IPW based

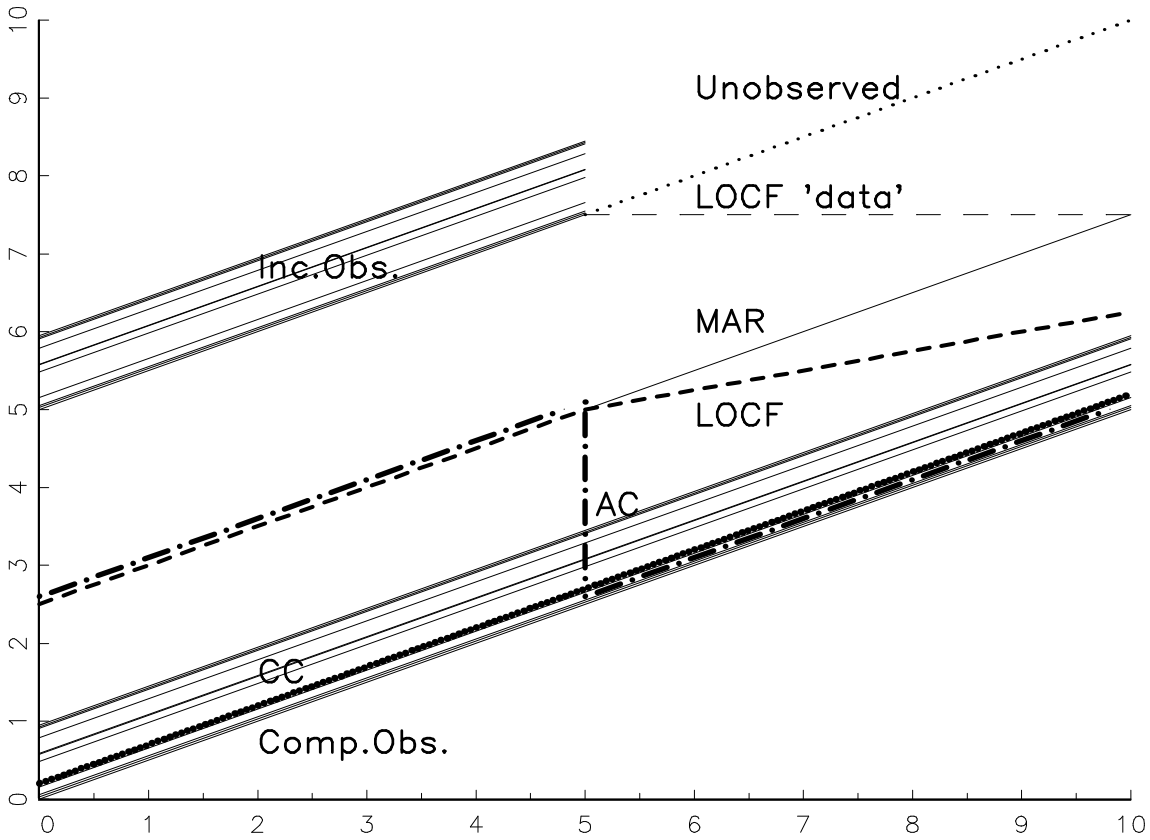
11.15 Overview and (Premature) Conclusion

MCAR/simple	CC	biased
	LOCF	inefficient
	single imputation	not simpler than MAR methods
MAR	direct lik./Bayes	easy to conduct
	IPW/d.r.	Gaussian & non-Gaussian
	multiple imputation	
MNAR	variety of methods	strong, untestable assumptions
		most useful in sensitivity analysis

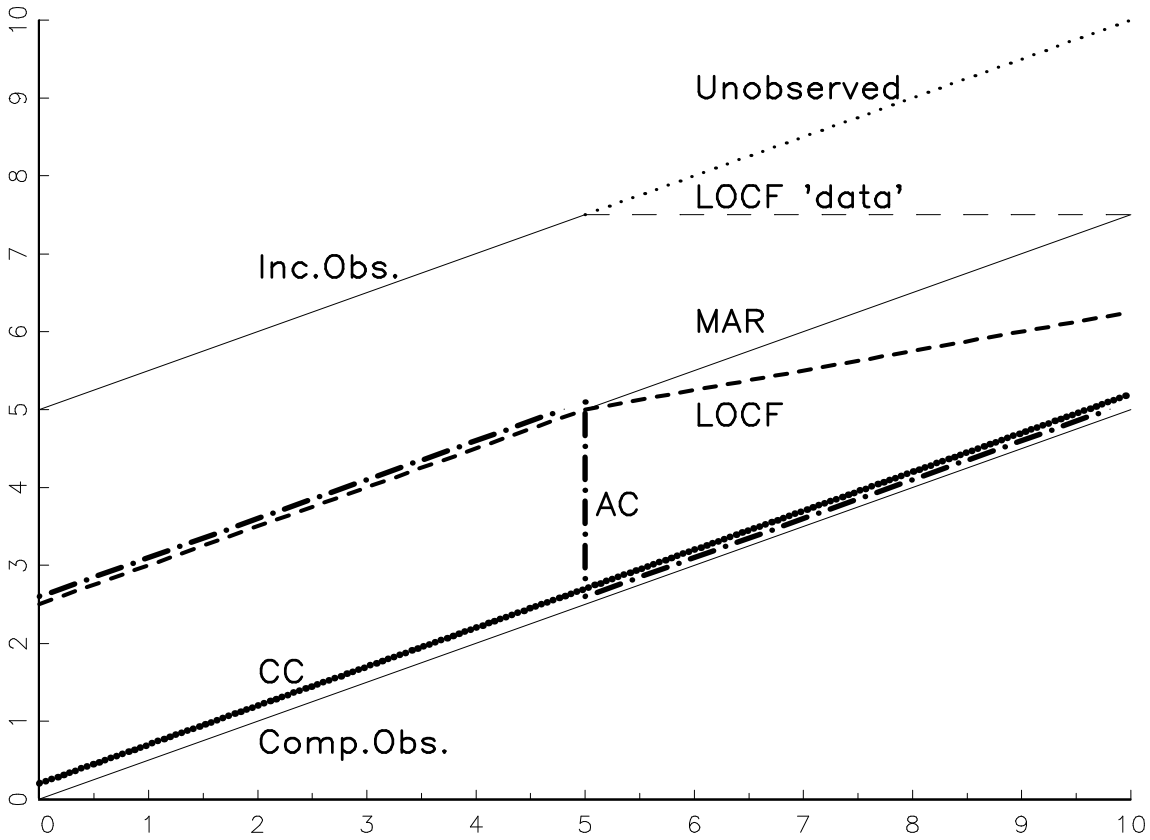
11.16 Incomplete Longitudinal Data



Data and Modeling Strategies

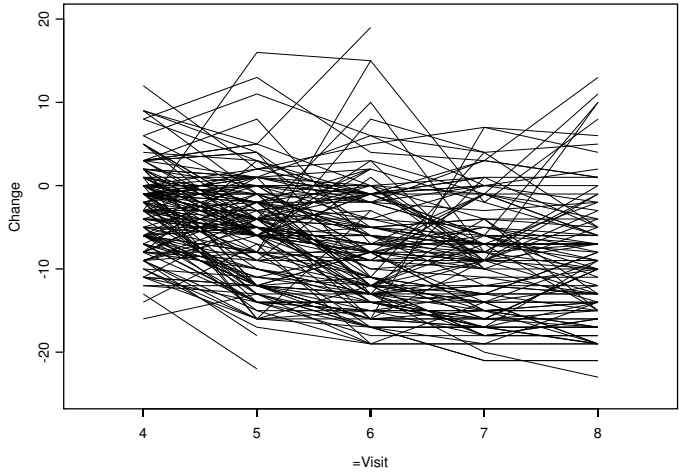


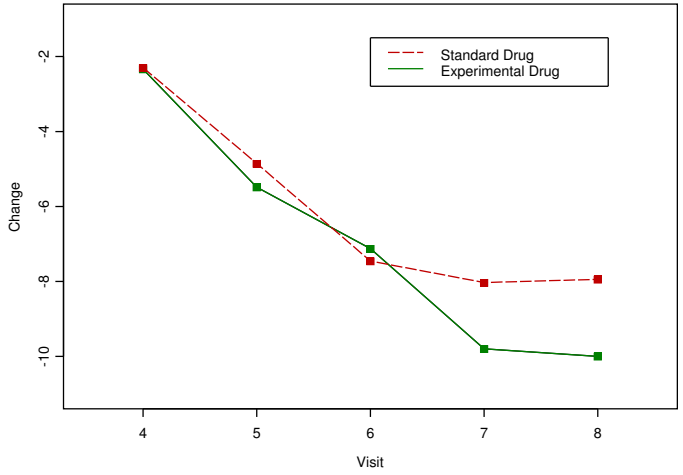
Modeling Strategies



11.17 The Depression Trial

- Clinical trial: experimental drug *versus* standard drug
- 170 patients
- Response: change versus baseline in $HAMD_{17}$ score
- 5 post-baseline visits: 4–8





11.18 Analysis of the Depression Trial

- Treatment effect at visit 8 (last follow-up measurement):

Method	Estimate	(s.e.)	p-value
CC	-1.94	(1.17)	0.0995
LOCF	-1.63	(1.08)	0.1322
MAR	-2.38	(1.16)	0.0419

Observe the slightly significant p-value under the MAR model

Chapter 12
Direct Likelihood / Ignorable Likelihood

- Simple methods
- Direct likelihood / ignorability

12.1 Simple Methods

MCAR

Complete case analysis:

$\Rightarrow$ **delete** incomplete subjects

- Standard statistical software
- Loss of information
- Impact on precision and power
- Missingness $\neq$ MCAR $\Rightarrow$ bias

Last observation carried forward:

$\Rightarrow$ **impute** missing values

- Standard statistical software
- Increase of information
- Constant profile after dropout: unrealistic
- Usually bias

12.2 Ignorability

Likelihood/Bayesian + MAR

&

Frequentist + MCAR

12.3 Direct Likelihood Maximization

$$\boxed{\text{MAR}}: \; f(\boldsymbol{Y}_i^o|\boldsymbol{\theta}) \cancel{f(\mathcal{D}_i|\boldsymbol{Y}_i^o, \boldsymbol{\psi})}$$

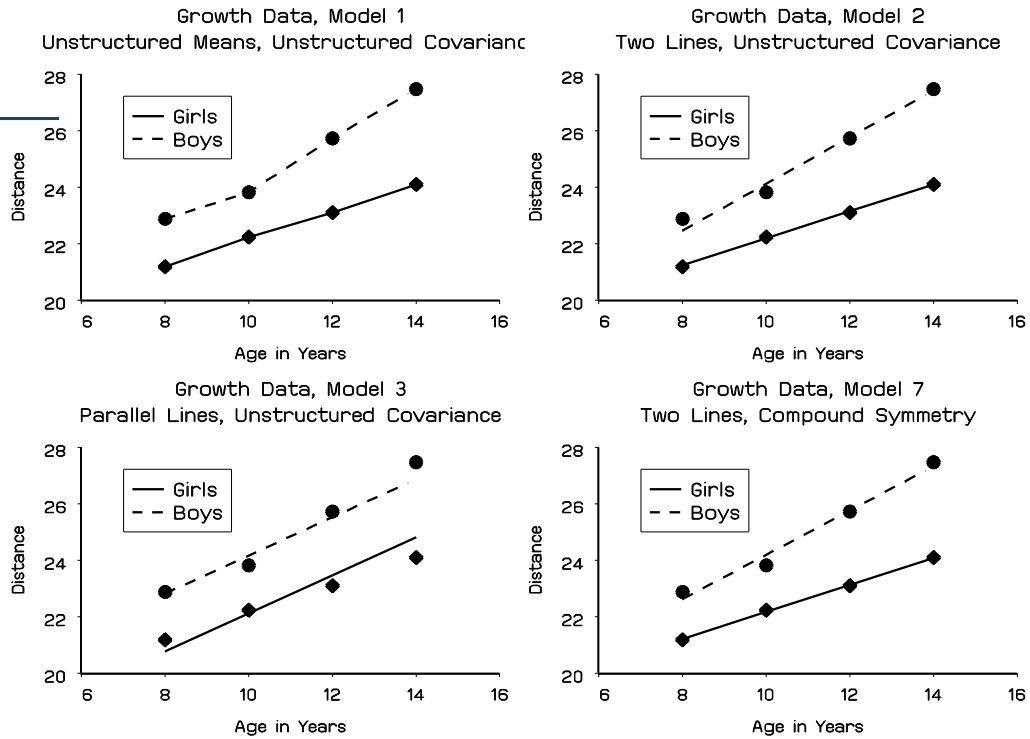
Mechanism is MAR
$\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ distinct
Interest in $\boldsymbol{\theta}$
(Use observed information matrix)

$\Longrightarrow$ Likelihood inference is valid

Outcome type	Modeling strategy	Software
Gaussian	Linear mixed model	SAS proc MIXED
Non-Gaussian	Generalized linear mixed model	SAS proc GLIMMIX, NLMIXED

12.4 Original, Complete Orthodontic Growth Data

	Mean	Covar	# par
1	unstructured	unstructured	18
2	$\neq$ slopes	unstructured	14
3	$=$ slopes	unstructured	13
7	**$\neq$ slopes**	**CS**	**6**



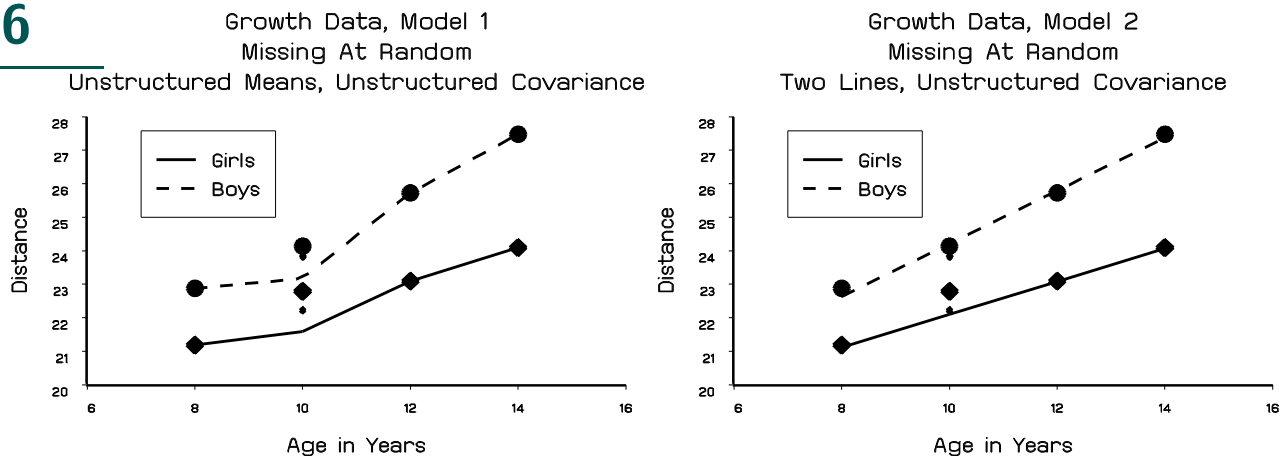
12.5 Trimmed Growth Data: Simple Methods

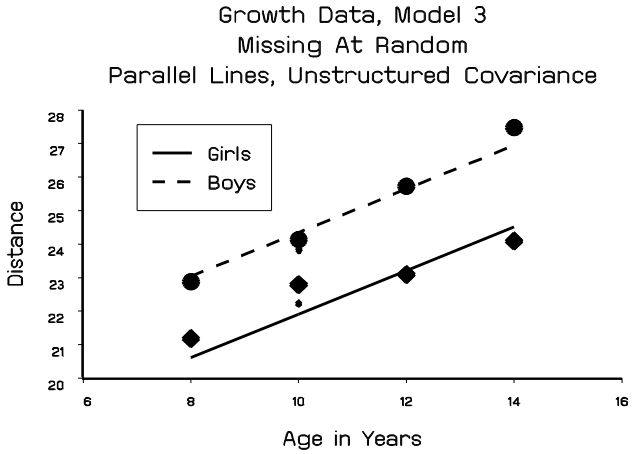
Method	Model	Mean	Covar	# par
Complete case	7a	= slopes	CS	5
LOCF	2a	quadratic	unstructured	16
Unconditional mean	7a	= slopes	CS	5
Conditional mean	1	unstructured	unstructured	18

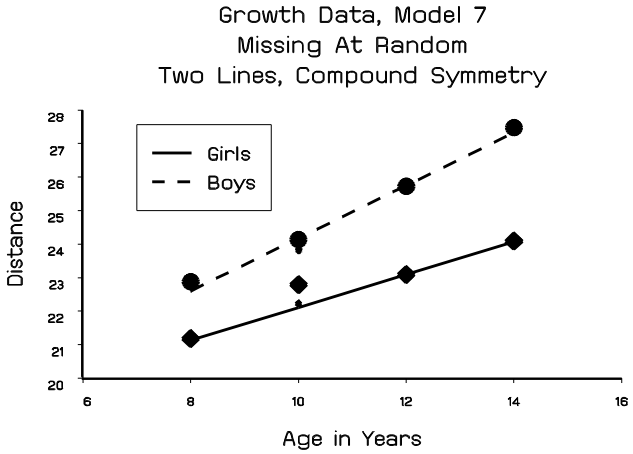
distorting

12.6 Trimmed Growth Data: Direct Likelihood

	Mean	Covar	# par
7	**$\neq$ slopes**	**CS**	**6**





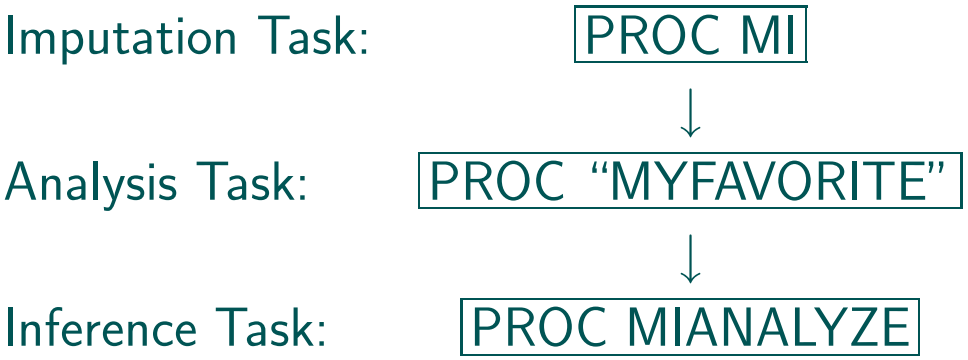


Chapter 13
Multiple Imputation

- Valid under MAR
- An alternative to direct likelihood and WGEE
- Three steps:
  1. The missing values are filled in M times $\Longrightarrow$ M complete data sets
  2. The M complete data sets are analyzed by using standard procedures
  3. The results from the M analyses are combined into a single inference
- Rubin (1987), Rubin and Schenker (1986), Little and Rubin (1987)

13.1 Use of MI in Practice

- Many analyses of the same incomplete set of data
- A combination of missing outcomes and missing covariates
- As an alternative to WGEE: MI can be combined with classical GEE
- MI in SAS:



13.2 Generalized Estimating Equations

Liang and Zeger (1986)

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} [D_i]^T [V_i(\boldsymbol{\alpha})]^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$$

- $\boldsymbol{V}_i(.)$ is not the true variance of $\boldsymbol{Y}_i$ but only a plausible guess
- the score equations are solved in a standard way
- Asymptotic distribution:

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \ \sim \ N(\mathbf{0}, I_0^{-1} I_1 I_0^{-1})$$

$$I_0 = \sum_{i=1}^{N} D_i^T [V_i(\boldsymbol{\alpha})]^{-1} D_i \qquad I_1 = \sum_{i=1}^{N} D_i^T [V_i(\boldsymbol{\alpha})]^{-1} \mathsf{Var}(\boldsymbol{Y}_i) [V_i(\boldsymbol{\alpha})]^{-1} D_i$$

13.3 Weighted GEE

$$\pi_i = \prod_{\ell=2}^{n_i}(1-p_{i\ell})$$

$$\pi_i' = \left[\prod_{\ell=2}^{d_i-1}(1-p_{i\ell})\right]\cdot p_{id_i}$$

$$p_{i\ell} = P\left(D_i=\ell \middle| D_i \geq \ell, Y_{i\overline{\ell}}, X_{i\overline{\ell}}\right)$$

$R_i = 1$ if subject i is complete

$R_i = 0$ if subject i is incomplete

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N}\frac{R_i}{\pi_i}\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'}V_i^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_i) = \mathbf{0}$$

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N}\frac{1}{\pi_i'}\frac{\partial \boldsymbol{\mu}_i^o}{\partial \boldsymbol{\beta}'}(V_i^o)^{-1}(\boldsymbol{y}_i^o-\boldsymbol{\mu}_i^o) = \mathbf{0}$$

Chapter 14
The Analgesic Trial

- single-arm trial with 530 patients recruited (491 selected for analysis)
- analgesic treatment for pain caused by chronic nonmalignant disease
- treatment was to be administered for 12 months
- we will focus on Global Satisfaction Assessment (GSA)
- GSA scale goes from 1=very good to 5=very bad
- GSA was rated by each subject 4 times during the trial, at months 3, 6, 9, and 12.

- Research questions:
  - ▷ Evolution over time
  - ▷ Relation with baseline covariates: age, sex, duration of the pain, type of pain, disease progression, Pain Control Assessment (PCA), ...
  - ▷ Investigation of dropout

- Frequencies:

GSA	Month 3		Month 6		Month 9		Month 12	
1	55	14.3%	38	12.6%	40	17.6%	30	13.5%
2	112	29.1%	84	27.8%	67	29.5%	66	29.6%
3	151	39.2%	115	38.1%	76	33.5%	97	43.5%
4	52	13.5%	51	16.9%	33	14.5%	27	12.1%
5	15	3.9%	14	4.6%	11	4.9%	3	1.4%
Tot	385		302		227		223	

- Missingness:

Measurement occasion					
Month 3	Month 6	Month 9	Month 12	Number	%
Completers					
O	O	O	O	163	41.2
Dropouts					
O	O	O	M	51	12.91
O	O	M	M	51	12.91
O	M	M	M	63	15.95
Non-monotone missingness					
O	O	M	O	30	7.59
O	M	O	O	7	1.77
O	M	O	M	2	0.51
O	M	M	O	18	4.56
M	O	O	O	2	0.51
M	O	O	M	1	0.25
M	O	M	O	1	0.25
M	O	M	M	3	0.76

14.1 Analysis of the Analgesic Trial

- A logistic regression for the dropout indicator:

$$\begin{aligned}\text{logit}[P(D_i = j | D_i \geq j, \cdot)] = \; & \psi_0 + \psi_{11} I(\text{GSA}_{i,j-1} = 1) + \psi_{12} I(\text{GSA}_{i,j-1} = 2) \\ & + \psi_{13} I(\text{GSA}_{i,j-1} = 3) + \psi_{14} I(\text{GSA}_{i,j-1} = 4) \\ & + \psi_2 \text{PCA0}_i + \psi_3 \text{PF}_i + \psi_4 \text{GD}_i\end{aligned}$$

  with

  - ▷ $\text{GSA}_{i,j-1}$ the 5-point outcome at the previous time
  - ▷ $I(\cdot)$ is an indicator function
  - ▷ PCA0_i is pain control assessment at baseline
  - ▷ PF_i is physical functioning at baseline
  - ▷ GD_i is genetic disorder at baseline are used

Effect	Par.	Estimate (s.e.)
Intercept	ψ_0	-1.80 (0.49)
Previous GSA$= 1$	ψ_{11}	-1.02 (0.41)
Previous GSA$= 2$	ψ_{12}	-1.04 (0.38)
Previous GSA$= 3$	ψ_{13}	-1.34 (0.37)
Previous GSA$= 4$	ψ_{14}	-0.26 (0.38)
Basel. PCA	ψ_2	0.25 (0.10)
Phys. func.	ψ_3	0.009 (0.004)
Genetic disfunc.	ψ_4	0.59 (0.24)

- There is some evidence for MAR: $P(D_i = j | D_i \geq j)$ depends on previous GSA.
- Furthermore: baseline PCA, physical functioning and genetic/congenital disorder.

- GEE and WGEE:

$$\text{logit}[P(Y_{ij} = 1|t_j, \text{PCA0}_i)] = \beta_1 + \beta_2 t_j + \beta_3 t_j^2 + \beta_4 \text{PCA0}_i$$

Effect	Par.	GEE	WGEE
Intercept	β_1	2.95 (0.47)	2.17 (0.69)
Time	β_2	-0.84 (0.33)	-0.44 (0.44)
Time^2	β_3	0.18 (0.07)	0.12 (0.09)
Basel. PCA	β_4	-0.24 (0.10)	-0.16 (0.13)

- A hint of potentially important differences between both

14.2 Analgesic Trial: Steps for WGEE in SAS

1. Preparatory data manipulation:

```
%dropout(...)
```

2. Logistic regression for weight model:

```
proc genmod data=gsac;
   class prevgsa;
   model dropout = prevgsa pca0 physfct gendis / pred dist=b;
   ods output obstats=pred;
   run;
```

3. Conversion of predicted values to weights:

```
...
%dropwgt(...)
```

4. Weighted GEE analysis:

```
proc genmod data=repbin.gsaw;
   scwgt wi;
   class patid timecls;
   model gsabin = time|time pca0 / dist=b;
   repeated subject=patid / type=un corrw within=timecls;
   run;
```

14.3 Analgesic Trial: Steps for WGEE in SAS, Using PROC GEE

- Experimental in SAS 9.4 (SAS/STAT 13.2)
- Preparation:

```
data gsaw;
   set gsaw;
   by patid;
   prevgsa = lag(gsa);
   if first.id then prevgsa = 1;
   time = time-1;
   timeclss = time;
run;
```

- Weighted GEE analysis:

```
ods graphics on;
proc gee data=gsaw plots=histogram;
   class patid timecls prevgsa;
   model gsabin = time|time pca0 / dist=bin;
   repeated subject=patid / within=timecls corr=un;
   missmodel prevgsa pca0 physfunt gendist / type=obslevel;
run;
```

Chapter 15
Creating Monotone Missingness

- When missingness is non-monotone, one might think of several mechanisms operating simultaneously:
  - ▷ A simple (MCAR or MAR) mechanism for the intermittent missing values
  - ▷ A more complex (MNAR) mechanism for the missing data past the moment of dropout
- Analyzing such data are complicated, especially with methods that apply to dropout only

- Solution:
  - ▷ Generate multiple imputations that render the datasets monotone missing, by including into the MI procedure:

```
mcmc impute = monotone;
```

  - ▷ Apply method of choice to the so-completed multiple sets of data

- Note: this is different from the monotone method in PROC MI, intended to fully complete already monotone sets of data

15.1 Example: Creating Monotone Missingness to Then Apply Weighted GEE

- Consider again the analgesic trial

- Multiple imputation to create monotone missingness:

```
proc mi data=m.gsa4 seed=459864 simple nimpute=10
        round=0.1 out=m.gsaimput;
title 'Monotone multiple imputation';
mcmc impute = monotone;
var pca0 physfct gsa1 gsa2 gsa3 gsa4;
run;
```

- Preparation of the data in vertical format, so that the data can be used in ordinary GEE:

```
data m.gsaw;
set m.gsa4;
array y (4) gsa1 gsa2 gsa3 gsa4;
do j=1 to 4;
   gsa=y(j);
   time=j;
   timecls=time;
   gsabin=.;
   if gsa=1 then gsabin=1;
   if gsa=2 then gsabin=1;
   if gsa=3 then gsabin=1;
   if gsa=4 then gsabin=0;
   if gsa=5 then gsabin=0;
   output;
end;
run;
```

- Standard GEE:

```
proc gee data=m.gsaw plots=histogram;
   title 'Standard GEE for GSA data';
   class patid timecls;
   model gsabin = time|time pca0 / dist=bin;
   repeated subject=patid / within=timecls corr=un;
run;
```

- Steps to prepare the data for weighted GEE, including definition of the 'previous' outcome:

```
data m.gsaimput02;
set m.gsaimput;
array y (4) gsa1 gsa2 gsa3 gsa4;
do j=1 to 4;
   gsa=y(j);
   time=j;
   timecls=time;
```

```
   patid2=1000*_imputation_+patid;
   output;
end;
run;

proc sort data=m.gsaimput02;
by _imputation_ patid2;
run;

data m.gsaimput03;
   set m.gsaimput02;
   by patid2;
   prevgsa = lag(gsa);
   if time=1 then prevgsa = 1;
   timeclss = time;
run;
```

```
data m.gsaimput03;
   set m.gsaimput03;
   if gsa<=3.5 then gsabin=1;
   if gsa>3.5 then gsabin=0;
   gsabin=gsabin+gsa-gsa;
   run;
```

- Weighted GEE, where weights are created at observation level:

```
ods graphics on;
proc gee data=m.gsaimput03 plots=histogram;
   title 'Weighted GEE for GSA Data Based on Multiple
          Imputation to Monotonize - OBSLEVEL';
   by _imputation_;
   class patid timecls;
   model gsabin = time|time pca0 / dist=bin covb;
   repeated subject=patid / within=timecls corr=un ecovb;
   missmodel prevgsa pca0 physfct  / type=obslevel;
   ods output GEEEmpPEst=gmparms parminfo=gmpinfo
              modelinfo=modelinfo GEERCov=gmcovb;
run;

proc mianalyze parms=gmparms  parminfo=gmpinfo covb=gmcovb;
title 'Multiple Imputation Analysis After Weighted GEE for GSA Data';
modeleffects intercept time time*time pca0;
run;
```

- To use weights at subject rather than observation level:

```
missmodel prevgsa pca0 physfct  / type=sublevel;
```

- Evidently, using these monotonized data, also standard GEE can be used:

```
ods graphics on;
proc gee data=m.gsaimput03 plots=histogram;
   title 'Standard GEE for GSA Data Based on
          Multiple Imputation to Monotonize';
   by _imputation_;
   class patid timecls;
   model gsabin = time|time pca0 / dist=bin covb;
   repeated subject=patid / within=timecls corr=un ecovb;V
   ods output GEEEmpPEst=gmparms parminfo=gmpinfo
              modelinfo=modelinfo GEERCov=gmcovb;
run;
```

- Files:
  - ▷ `analg11(met-proc-gee).sas`
  - ▷ `analg11(met-proc-gee).lst`
- Overview of results:

Effect	Par.	Est.(s.e.)	p-value	Est.(s.e.)	p-value
		Standard GEE			
		Without MI		After MI	
Intercept	β_0	2.90(0.46)		2.87(0.45)	
Time	β_1	-0.81(0.32)	0.0124	-0.83(0.32)	0.0087
Time^2	β_2	0.17(0.07)	0.0083	0.18(0.06)	0.0058
PCA$_0$	β_3	-0.23(0.10)	0.0178	-0.21(0.10)	0.0253
		Weighted GEE (after MI)			
		Observation level		Subject level	
Intercept	β_0	2.74(0.46)		2.62(0.60)	
Time	β_1	-0.76(0.33)	0.0231	-0.71(0.40)	0.0747
Time^2	β_2	0.17(0.07)	0.0155	0.16(0.08)	0.0444
PCA$_0$	β_3	-0.19(0.10)	0.0384	-0.21(0.12)	0.0853

- The dropout model is similar but slightly different than the one used with PROC GENMOD.

- Weighted GEE leads to increased standard errors, as observed before.

- This effect is less pronounced when weigths are constructed at observation level, rather than at subject level.

- A typical output for one of the imputed datasets takes the form (first imputation out of ten; with weights at observation level):

Parameter Estimates for Response Model
with Empirical Standard Error

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	2.4299	0.5890	1.2755	3.5843	4.13	<.0001
TIME	-0.5881	0.3912	-1.3548	0.1787	-1.50	0.1328
TIME*TIME	0.1392	0.0794	-0.0165	0.2949	1.75	0.0796
PCA0	-0.1797	0.1173	-0.4096	0.0501	-1.53	0.1254

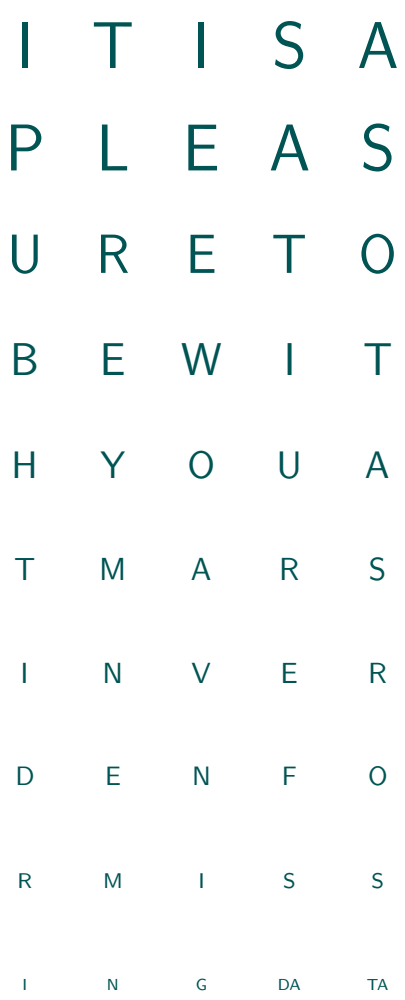
Parameter Estimates for Missingness Model

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	3.1335	0.4060	2.3377	3.9293	7.72	<.0001
prevgsa	-0.1974	0.0822	-0.3585	-0.0363	-2.40	0.0163
PCA0	-0.2495	0.0956	-0.4370	-0.0621	-2.61	0.0091
PHYSFCT	-0.0079	0.0037	-0.0151	-0.0007	-2.16	0.0311

Chapter 16
Case Study: The Dataset

- Visual Acuity
- Age-related Macular Degeneration
- Missingness

16.1 Visual Acuity



16.2 Age-related Macular Degeneration Trial

- Pharmacological Therapy for Macular Degeneration Study Group (1997)
- An occular pressure disease which makes patients progressively lose vision
- 240 patients enrolled in a multi-center trial (190 completers)
- Treatment: Interferon-α (6 million units) versus placebo
- Visits: baseline and follow-up at 4, 12, 24, and 52 weeks
- Continuous outcome: visual acuity: # letters correctly read on a vision chart
- Binary outcome: visual acuity versus baseline ≥ 0 or ≤ 0

- Missingness:

Measurement occasion					
4 wks	12 wks	24 wks	52 wks	Number	%
Completers					
O	O	O	O	188	78.33
Dropouts					
O	O	O	M	24	10.00
O	O	M	M	8	3.33
O	M	M	M	6	2.50
M	M	M	M	6	2.50
Non-monotone missingness					
O	O	M	O	4	1.67
O	M	M	O	1	0.42
M	O	O	O	2	0.83
M	O	M	M	1	0.42

Chapter 17
Case Study: Weighted Generalized Estimating Equations

- ▷ Model for the weights
- ▷ Incorporating the weights within GEE

17.1 Analysis of the ARMD Trial

- Model for the weights:

$$\text{logit}[P(D_i = j | D_i \geq j)] = \psi_0 + \psi_1 y_{i,j-1} + \psi_2 T_i + \psi_{31} L_{1i} + \psi_{32} L_{2i} + \psi_{34} L_{3i} + \psi_{41} I(t_j = 2) + \psi_{42} I(t_j = 3)$$

  with

  - $y_{i,j-1}$ the binary outcome at the previous time $t_{i,j-1} = t_{j-1}$ (since time is common to all subjects)
  - $T_i = 1$ for interferon-α and $T_i = 0$ for placebo
  - $L_{ki} = 1$ if the patient's eye lesion is of level $k = 1, \ldots, 4$ (since one dummy variable is redundant, only three are used)
  - $I(\cdot)$ is an indicator function

- Results for the weights model:

Effect	Parameter	Estimate (s.e.)
Intercept	ψ_0	0.14 (0.49)
Previous outcome	ψ_1	0.04 (0.38)
Treatment	ψ_2	-0.86 (0.37)
Lesion level 1	ψ_{31}	-1.85 (0.49)
Lesion level 2	ψ_{32}	-1.91 (0.52)
Lesion level 3	ψ_{33}	-2.80 (0.72)
Time 2	ψ_{41}	-1.75 (0.49)
Time 3	ψ_{42}	-1.38 (0.44)

- GEE:

$$\text{logit}[P(Y_{ij} = 1|T_i, t_j)] = \beta_{j1} + \beta_{j2} T_i$$

  with

  - ▷ $T_i = 0$ for placebo and $T_i = 1$ for interferon-α
  - ▷ t_j $(j = 1, \ldots, 4)$ refers to the four follow-up measurements
  - ▷ Classical GEE and linearization-based GEE
  - ▷ Comparison between CC, LOCF, and GEE analyses

- SAS code:

```
...

proc genmod data=armdhlp descending;
class trt prev lesion time;
model dropout = prev trt lesion time  / pred dist=b;
ods output obstats=pred;
ods listing exclude obstats;
run;
```

```
...

proc genmod data=armdwgee;
title 'data as is - WGEE';
weight wi;
class time treat subject;
model bindif = time treat*time / noint dist=binomial;
repeated subject=subject / withinsubject=time type=exch modelse;
run;

proc glimmix data=armdwgee empirical;
title 'data as is - WGEE - linearized version - empirical';
weight wi;
nloptions maxiter=50 technique=newrap;
class time treat subject;
model bindif = time treat*time /  noint solution dist=binary ;
random _residual_ / subject=subject type=cs;
run;
```

- Results:

Effect	Par.	CC	LOCF	Observed data	
				Unweighted	WGEE
			Standard GEE		
Int.4	β_{11}	-1.01(0.24;0.24)	-0.87(0.20;0.21)	-0.87(0.21;0.21)	-0.98(0.10;0.44)
Int.12	β_{21}	-0.89(0.24;0.24)	-0.97(0.21;0.21)	-1.01(0.21;0.21)	-1.78(0.15;0.38)
Int.24	β_{31}	-1.13(0.25;0.25)	-1.05(0.21;0.21)	-1.07(0.22;0.22)	-1.11(0.15;0.33)
Int.52	β_{41}	-1.64(0.29;0.29)	-1.51(0.24;0.24)	-1.71(0.29;0.29)	-1.72(0.25;0.39)
Tr.4	β_{12}	0.40(0.32;0.32)	0.22(0.28;0.28)	0.22(0.28;0.28)	0.80(0.15;0.67)
Tr.12	β_{22}	0.49(0.31;0.31)	0.55(0.28;0.28)	0.61(0.29;0.29)	1.87(0.19;0.61)
Tr.24	β_{32}	0.48(0.33;0.33)	0.42(0.29;0.29)	0.44(0.30;0.30)	0.73(0.20;0.52)
Tr.52	β_{42}	0.40(0.38;0.38)	0.34(0.32;0.32)	0.44(0.37;0.37)	0.74(0.31;0.52)
Corr.	ρ	0.39	0.44	0.39	0.33

Chapter 18
Case Study: Multiple Imputation

- Settings and Models
- Results for GEE
- Results for GLMM

18.1 MI Analysis of the ARMD Trial

- $M = 10$ imputations
- GEE:

$$\text{logit}[P(Y_{ij} = 1|T_i, t_j)] = \beta_{j1} + \beta_{j2}T_i$$

- GLMM:

$$\text{logit}[P(Y_{ij} = 1|T_i, t_j, b_i)] = \beta_{j1} + b_i + \beta_{j2}T_i, \qquad b_i \sim N(0, \tau^2)$$

- $T_i = 0$ for placebo and $T_i = 1$ for interferon-α
- t_j $(j = 1, \ldots, 4)$ refers to the four follow-up measurements
- Imputation based on the continuous outcome

- Results:

Effect	Par.	GEE	GLMM
Int.4	β_{11}	-0.84(0.20)	-1.46(0.36)
Int.12	β_{21}	-1.02(0.22)	-1.75(0.38)
Int.24	β_{31}	-1.07(0.23)	-1.83(0.38)
Int.52	β_{41}	-1.61(0.27)	-2.69(0.45)
Trt.4	β_{12}	0.21(0.28)	0.32(0.48)
Trt.12	β_{22}	0.60(0.29)	0.99(0.49)
Trt.24	β_{32}	0.43(0.30)	0.67(0.51)
Trt.52	β_{42}	0.37(0.35)	0.52(0.56)
R.I. s.d.	τ		2.20(0.26)
R.I. var.	τ^2		4.85(1.13)

18.2 SAS Code for MI

1. Preparatory data analysis so that there is one line per subject

2. The imputation task:

```
proc mi data=armd13 seed=486048 out=armd13a simple nimpute=10 round=0.1;
   var lesion diff4 diff12 diff24 diff52;
   by treat;
   run;
```

   Note that the imputation task is conducted on the continuous outcome 'diff·', indicating the difference in number of letters versus baseline

3. Then, data manipulation takes place to define the binary indicators and to create a longitudinal version of the dataset

4. The analysis task (GEE):

```
proc genmod data=armd13c;
   class time subject;
   by _imputation_;
   model bindif = time1 time2 time3 time4
                  trttime1 trttime2 trttime3 trttime4
                  / noint dist=binomial covb;
   repeated subject=subject / withinsubject=time type=exch modelse;
   ods output ParameterEstimates=gmparms parminfo=gmpinfo CovB=gmcovb;
run;
```

5. The analysis task (GLMM):

```
proc nlmixed data=armd13c qpoints=20 maxiter=100 technique=newrap cov ecov;
   by _imputation_;
   eta = beta11*time1+beta12*time2+beta13*time3+beta14*time4+b
         +beta21*trttime1+beta22*trttime2+beta23*trttime3+beta24*trttime4;
   p = exp(eta)/(1+exp(eta));
   model bindif ~ binary(p);
   random b ~ normal(0,tau*tau) subject=subject;
   estimate 'tau2' tau*tau;
   ods output ParameterEstimates=nlparms CovMatParmEst=nlcovb
              AdditionalEstimates=nlparmsa CovMatAddEst=nlcovba;
   run;
```

6. The inference task (GEE):

```
proc mianalyze parms=gmparms covb=gmcovb parminfo=gmpinfo wcov bcov tcov;
modeleffects time1 time2 time3 time4 trttime1 trttime2 trttime3 trttime4;
run;
```

7. The inference task (GLMM):

```
proc mianalyze parms=nlparms covb=nlcovb wcov bcov tcov;
modeleffects beta11 beta12 beta13 beta14 beta21 beta22 beta23 beta24;
run;
```

18.3 Example of Sensitivity Analysis

- We apply a **shift** to the treatment group:

```
proc mi data=m.armd13 seed=486048 simple out=m.armd13as1
        nimpute=10 round=0.1;
   title 'Shift multiple imputation';
   class treat;
   var lesion diff4 diff12 diff24 diff52;
   fcs reg;
   mnar adjust (diff12 / shift=10 adjustobs=(treat='2'));
   mnar adjust (diff24 / shift=15 adjustobs=(treat='2'));
   mnar adjust (diff52 / shift=20 adjustobs=(treat='2'));
   by treat;
run;
```

- Expanded results (for GLMM only):

			GLMM	
Effect	Par.	GEE	MAR	shift
Int.4	β_{11}	-0.84(0.20)	-1.46(0.36)	-1.42(0.36)
Int.12	β_{21}	-1.02(0.22)	-1.75(0.38)	-1.67(0.38)
Int.24	β_{31}	-1.07(0.23)	-1.83(0.38)	-1.83(0.39)
Int.52	β_{41}	-1.61(0.27)	-2.69(0.45)	-2.77(0.44)
Trt.4	β_{12}	0.21(0.28)	0.32(0.48)	0.24(0.48)
Trt.12	β_{22}	0.60(0.29)	0.99(0.49)	0.91(0.50)
Trt.24	β_{32}	0.43(0.30)	0.67(0.51)	0.65(0.51)
Trt.52	β_{42}	0.37(0.35)	0.52(0.56)	0.60(0.55)
R.I. s.d.	τ		2.20(0.26)	2.20(0.25)
R.I. var.	τ^2		4.85(1.13)	4.83(1.08)

- A shift applied to one sex group in the orthodontic growth data:

```
proc mi data=m.growthmi seed=459864 simple nimpute=10
        round=0.1 out=outmishift;
title 'Shift multiple imputation';
class sex;
by sex;
monotone method=reg;
mnar adjust (meas10 / shift=10 adjustobs=(sex='2'));
var meas8 meas12 meas14 meas10;
run;
```

- A sensitivity analysis following PMM ideas: NCMV-based imputation:

```
proc mi data=m.growthmi seed=459864 simple nimpute=10
        round=0.1 out=outmincmv;
title 'NCMV multiple imputation';
by sex;
monotone method=reg;
mnar model (meas10 / modelobs=ncmv);
var meas8 meas12 meas14 meas10;
run;
```

- The latter is available only for monotone missingness.

18.4 Overview and (No Longer Premature) Conclusion

MCAR/simple	CC	biased
	LOCF	inefficient
	single imputation	not simpler than MAR methods
MAR	direct lik./Bayes	easy to conduct
	IPW/d.r.	Gaussian & non-Gaussian
	multiple imputation	
MNAR	variety of methods	strong, untestable assumptions
		most useful in sensitivity analysis