Bayesian Methods for Missing Covariates in Longitudinal Studies

Nicole Erler[1] **& Emmanuel Lesaffre**[2]

[1]Department of Biostatistics

Erasmus MC, Rotterdam, the Netherlands

n.erler@erasmusmc.nl

[2]Interuniversity Institute for Biostatistics and statistical Bioinformatics

KU Leuven & University Hasselt, Belgium

emmanuel.lesaffre@kuleuven.be

IBC2022 in Riga — July 10, 2022

Session 1
What, why and how?

1.1 What is this course about?

- Aims of the course:
  - Examine the impact of missing values on the statistical analysis and conclusions of a statistical analysis
  - Focus is on missing values in covariates of (especially) observational studies for longitudinal data

Treatment of missing data in medical literature

- Missing data are common in all (longitudinal) studies
- They are often inadequately handled in experimental and observational research
- Much attention has been paid to the treatment of missing responses in RCTs, see review of Wood et al. (2004):
  - 71 published medical papers in BMJ, JAMA, Lancet and NEJM
  - 89% had partly missing outcome data
  - 37% trials with repeated outcome measures: 46% complete case analysis
  - Only 21% reported sensitivity analyses
- But also missing covariates in observational studies is a common and serious problem, see Karahalios et al. (2012):
  - 82 published medical papers (2000-2009)
  - 54 papers used complete case analysis
  - 5 used multiple imputation
  - 1 used maximum likelihood

Causes for missing data

- Data can be missing because of a variety of reasons:
  - One forgets to fill in an item in a questionnaire, or one is not willing to fill in that item (missing covariate values)
  - Subject was ill and could not come for a medical exam, or was on holiday (missing response and/or covariate(s))
  - By design, subjects are excluded at a time point when (a) vital parameter(s) exceeds a threshold (missing response)
  - Subject dies (missing response)
  - ...

1.2 Why this course could be useful

- ▷ There have been many courses on how to deal with missing data
- ▷ Many focused on the treatment of missing responses
- ▷ Other courses explored the many ways of multiple imputation (MI) for missing data in covariates with MICE as the preferred choice
- ▷ Most, if not all, courses on missing data make (essentially) use of the frequentist approach
- ▷ The maximum likelihood & Bayesian approach for missing covariates provide a valuable alternative to MI
- ▷ We consider here the specific case of missing covariates in longitudinal studies

A more specific treatment of missing values is needed

- Multiple imputation is a powerful procedure and is widely applied
- Nowadays MICE is the preferred approach, but:
  - It is an omnibus procedure that imputes missing values irrespective of the main response model
  - Thereby, imputed values may distort the fit of the data to the main model
- In this course we wish to illustrate the use of the R package JointAI:
  - Focus on mixed effect models
  - For dealing with (time-varying) covariates
  - Making use of the Bayesian approach (hence making use of the likelihood of covariates)

Some general literature on how to deal with missing data

- Well-known sources for the treatment of missing data:
  - P. Allison, *Missing Data*, 2001
  - M.J Daniels & J.W. Hogan, Missing Data in Longitudinal Studies. Strategies for Bayesian Modeling and Sensitivity Analysis, 2008
  - J.W. Graham, *Missing Data Analysis and Design*, 2012
  - S. van Buuren, *Flexible imputation of missing data (2nd ed.)*, 2018
  - J.R. Carpenter, *Missing Data*, 2020
  - R.J.A. Little & D. B. Rubin, *Statistical analysis with missing data (3rd ed.)*, 2020

Some references on how to deal with missing data in covariates

- Here focus on methods to handle missing data in epidemiological studies:
  - A. Karahalios, L. Baglietto, J.B. Carlin et al., *A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures*, 2012
  - O.U. Carroll, *How are missing data in covariates handled in observational time-to-event studies in oncology? A systematic review*, 2020
  - A.C. Sterne, I.R. White, J.B. Carlin et al., *Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls*, 2009
  - K.J.M. Jansen, A.R.T. Donders, F.E. Harrell et al. *Missing covariate data in medical research: To impute is better than to ignore*, 2010
- And of course:
  - E.N. Erler, R. Rizopoulos, J. van Rosmalen et al., *Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian approach*, 2016
  - E.N. Erler, R. Rizopoulos, V. Jaddoe et al., *Bayesian imputation of time-varying covariates in linear mixed models*, 2017

1.3 How we organized the course

- **Part 1:**
  - ▷ We start with introducing the motivating data sets:
    - ○ Two simple longitudinal data sets that missing values in the responses
    - ○ Two more complex longitudinal data sets with missing covariates
  - ▷ Then we review some general theory and results on missing data
  - ▷ Finally, we review the Bayesian approach, but briefly
- **Part 2:**
  - ▷ Using the two last data sets, illustrate the use of JointAI with hands-on computing

Learning objectives of this course

- ▷ Understand the difficulties of standard (Bayesian) imputation methods in complex data structures, with a focus on longitudinal data and multi-level data
- ▷ Focus on dealing with missing values in covariates
- ▷ Fit Bayesian models on data sets with incomplete data using the R package **JointAI**

Course program

Session 1: What, why and how? (just finished)

Session 2: Examples of longitudinal data

Session 3: Frequentist methods for longitudinal studies

Session 4: Missing data processes

Session 5: Bayesian Analysis of Incomplete Data with JointAI

Session 6: The use of the R package JointAI

- Practicals

Session 7: Summary & Extensions

Sources

- **General Bayesian textbook**: Lesaffre, E & Lawson, A: *Bayesian Biostatistics*, John Wiley & Sons, New York, 2012
- **Bayesian longitudinal studies**: Daniels, M J & Hogan, J W: *Missing Data in Longitudinal Studies. Strategies for Bayesian Modelling and Sensitivity Analysis*, Chapman & Hall, Boca Raton, 2008
- **NEJM paper to recommend appropriate approaches to deal with missing data**: `http://www.nejm.org/doi/full/10.1056/NEJMsr1203730#t=article`
- **Series of Ibrahim, Chen & Lipsitz papers**: Ibrahim (1990); Ibrahim, Chen, Lipsitz (1999a,b); Lipsitz et al. (1999); Herring, Ibrahim, Lipsitz (2002); Stubbendick & Ibrahim (2003); Huang, Chen, Ibrahim (2005); Ibrahim, Chen, Lipsitz, Herring (2005); Chen & Ibrahim (2006); Ibrahim, Chen, Lipsitz (2008)
- **Software upon which course is based**: N.S. Erler, D. Rizopoulos, E.Lesaffre: *JointAI: Joint Analysis and Imputation of Incomplete Data in R*, Journal of Statistical Software, 100(20), 1–56, 2021

End of Session 1

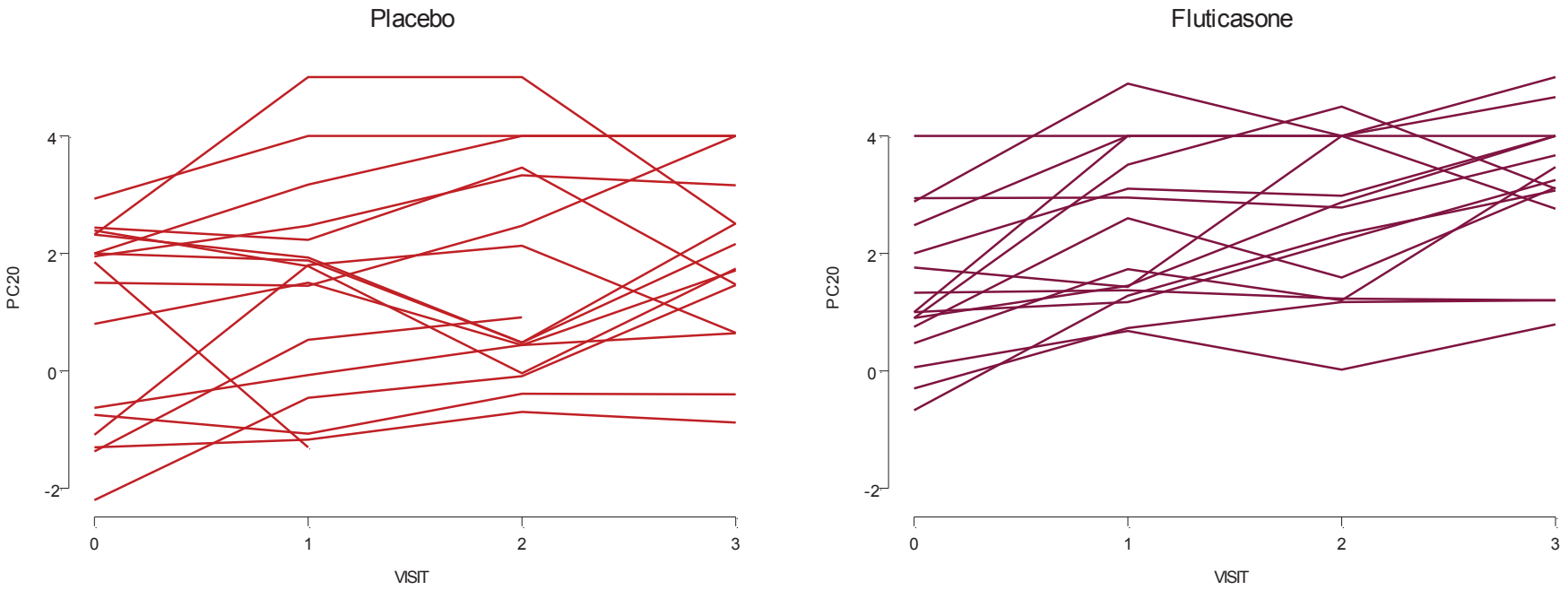
Session 2
Examples of longitudinal data

- Example 1: **Asthma RCT**: continuous outcome measured at regular time points but dropouts
- Example 2: **Jimma study**: growth curve study with a continuous outcome measured at irregular time points and missing data
- Example 3: **ELTR study**: registry transplant study exploring survival of patient and graft
- Example 4: **Hepatitis C**: liver fibrosis in patients with chronic hepatitis C infection

2.1 Asthma RCT

- Subjects: Children with mild to moderate asthma (N = 35)
- Outcome: (log2 transformed) PC20 measured after 1, 2 and 3 months
  - FEV1: volume of air exhaled under forced conditions in 1 second
  - PC20 = 2mg/mL: FEV1 decreased with 20% after inhalation of a histamine of 2mg/mL during 2 mins
- Covariates: visit number, treatment TRT=0 (Placebo) or TRT=1 (fluticasone)
- Design: Parallel randomized double blind
- **Question: Effect of fluticasone on PC20?**
- Reference: Hoekstra, Grol, Bouman et al., Respir. Crit. Care Med., 1996

Asthma RCT: Individual profiles



Three patients dropped out (2 at 2nd visit, 1 at 3rd visit)

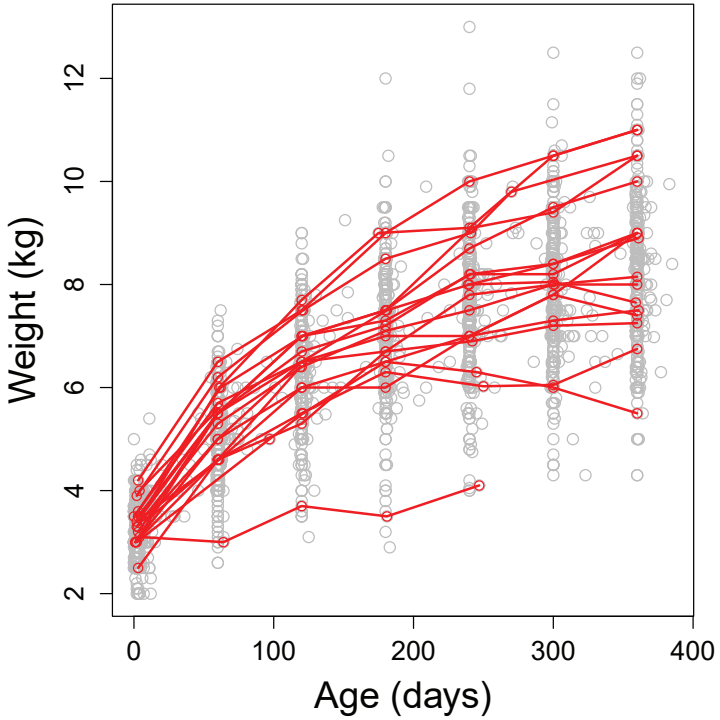
No missing values in covariates (only treatment and visit)

- ▷ This is a classical RCT with a few dropouts causing missing responses
- ▷ Typically we would fit a multivariate normal distribution (or a linear mixed model) to the data
- ▷ Missing responses are then implicitly imputed

2.2 Jimma Infant Survival Longitudinal Study

- Subjects: Newborns from the Jimma town (Ethiopia) born in period: 11-9-1992 till 10-9-1993 (subgroup of N = 495)
- Covariates: Gender, age of mother in first year, cultural practices, etc.
- Outcome: Various growth parameters, here height (but could also be weight, arm circumference) examined at birth and approximately each 2 months
- Design: Epidemiological cohort study
- Missing data: Children die, were lost-to-follow-up, missed visits + visits are only approximately taken at equal time lags
- **Question: Impact of demographic and other variables on child's first year's weight/height gain?**
- Reference: Lesaffre, Todem, Verbeke and Kenward, Biometrical Journal (2000)

Jimma study: Individual profiles & missing data



Visit	participating	drop-outs	deaths
live births	1501		
1	1480	0	21
2	1420	37	23
3	1342	50	28
4	1294	34	14
5	1246	29	19
6	1199	28	20
7	1152	30	16

Missing covariate values when response is missing

- ▷ This observational longitudinal study suffers from many dropouts and intermittent missing values
- ▷ Typically we would fit here a linear mixed model to the data
- ▷ Again missing responses are then implicitly imputed

2.3 ELTR study

- Subjects: (all) 1 549 first-transplant patients from the European Liver Transplant Registry (ELTR) for primary sclerosing cholangitis (PSC) between 1980 and 2015
- Outcome: patient and graft survival after recurrence of PSC (rPSC)
- Covariates: rPSC, recipient data (age, gender and blood type), donor data of the first graft (age, gender, blood type, graft type, date of transplant and total ischaemic time (warm and cold combined)), retransplant (time-varying)
- Design: Registry study, with livers nested in patients
- **Question: Impact of rPSC and other covariates on patient and graft survival?**
- Reference: Visseren, Erler, Polak et al., Transplant International, 2021

ELTR study: Descriptive statistics and missing data

Table 1. Patient and donor characteristics at time of first transplantation.

	Total *n* = 1,549	Free of rPSC *n* = 1,290	rPSC *n* = 259
Male recipient	1,045 (67.5%)	855 (66.3%)	190 (73.4%)
Recipient age (years)	42.5 [23.0, 62.6]	43.4 [23.1, 63.0]	38.7 [22.9, 58.6]
Donor age (years)	43.0 [16.0, 68.0]	43.0 [17.0, 69.0]	46 [16.0, 66.0]
Missing	89 (5.7%)	75 (5.8%)	14 (5.4%)
Donor gender			
Male	831 (56.9%)	696 (57.3%)	135 (4.9%)
Female	630 (43.1%)	519 (42.7%)	111 (5.1%)
Missing	88 (5.7%)	75 (5.8%)	13 (5.0%)
Graft type			
DBD full graft	1,318 (88.6%)	1088 (87.9%)	230 (2.0%)
Living donor	78 (5.2%)	74 (6.0%)	4 (1.6%)
DBD split graft	80 (5.4%)	65 (5.3%)	15 (6.0%)
DCD full graft	12 (0.8%)	11 (0.9%)	1 (0.4%)
Missing	61 (3.9%)	52 (4.0%)	9 (3.5%)
Total ischaemic time (hours)	8.7 [3, 14.6]	8.6 [2.7, 14.9]	9.0 [4.2, 14.1]
Missing	170 (11.0%)	145 (11.2%)	25 (9.7%)
Calendar year of LT	2004 [1991, 2013]	2005 [1991, 2013]	2002 [1992, 2011]

DBD, donation after brain death; DCD, donation after circulatory death; rPSC, recurrence of primary sclerosing cholangitis; LT, liver transplantation.

Characteristics of 1,549 patients who underwent a liver transplantation for PSC, as a total and divided in two groups: free of rPSC and ever diagnosed with rPSC. Shown are numbers (%) or median (2.5% and 97.5% quantile).

Missing values in several covariates

ELTR study: Bayesian survival analyses

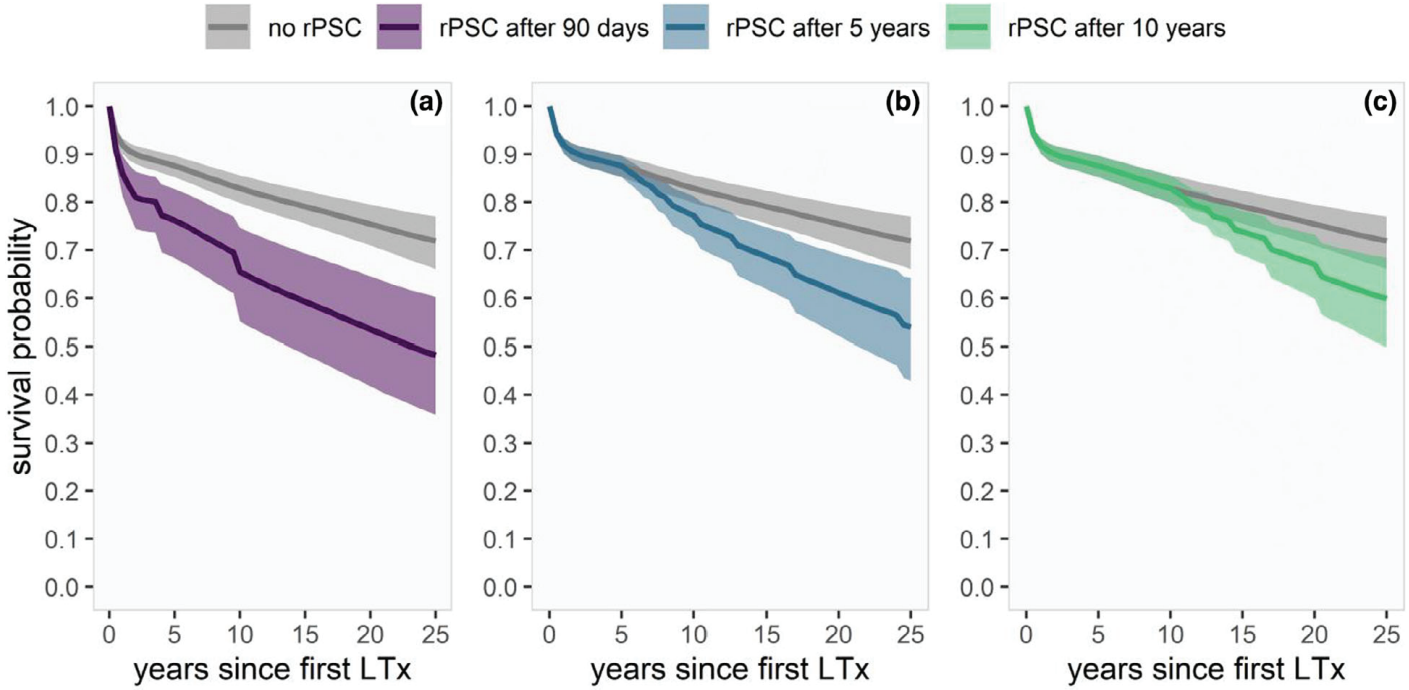


Figure 4 Expected patient survival with and without rPSC. Expected patient survival and corresponding 95% CIs in scenarios with and without rPSC. The curves show the scenarios when rPSC is diagnosed after 90 days (a), 5 years (b) and 10 years (c). The effect of rPSC is more detrimental when diagnosed early after LT, compared with a later onset.

- ▷ This survival study suffers from an appreciable number of missing covariates
- ▷ The percentage of missingness for each covariate is small, but combined it can affect the conclusions considerably
- ▷ This study will be used in part 2 of the course

2.4 Liver Fibrosis in Patients with Chronic Hepatitis C

- Chronic Hepatitis C:
  - Retrospective international cohort study
  - Patients with chronic hepatitis C diagnosed 1990 – 2017 with early stage liver disease
  - Follow-up until death or time of data extraction (0 - 20 years)
- Subjects: approx. 4000 patients from 4 hospitals
- Outcome(s): FIB4 (non-invasive fibrosis index), time to severe fibrosis (FIB4 $>$ 3.25), time to death / clinical events
- Covariates: baseline characteristics (age, sex, BMI, history of alcohol abuse, ...), repeatedly measured lab values (AST, ALT, ...)
- **Question: Pattern of fibrosis progression; predictors for severe fibrosis; FIB4 as predictor for clinical outcomes**
- Reference: (under review)

Chronic Hepatitis C: Descriptive statistics and missing data

Baseline characteristics		
	median [range] or [proportion]	missing
Age	47 [21.5 - 70.0]	(complete)
Sex (male)	60.0%	(complete)
BMI	25 [18.5 - 37.5]	23.2% NA
HBV anti-core positive	21.7%	9.2% NA
History of alcohol abuse	15.4%	5.9% NA
Diabetes mellitus	8.7%	(complete)
FIB4	1.3 [0.5 - 2.8]	(complete$^{(*)}$)
Bilirubin	9.0 [14.0 - 26.0]	11.5% NA
Albumin	42 [33 - 51]	21.6% NA
Creatinine	70 [47 - 142]	22.5% NA

$\Rightarrow$ approx. 19 % complete cases

(*) "baseline" = first available FIB4 measurement

Repeated lab-measurements

- ▷ median 5 [1 - 45] measurements per patient
- ▷ unbalanced measurements
- ▷ missing values (not all lab values determined each time)
- ▷ bilirubin (16% NA), albumin (28.7% NA), creatinine (18.6% NA)

Chronic Hepatitis C: Bayesian analyses

- Bayesian linear mixed model for FIB4 over time
- Bayesian proportional hazards model for time to severe fibrosis (FIB4 $>$ 3.25) ($\Rightarrow$ imputation of missing values in baseline covariates)
- Bayesian proportional hazards model for time to death/clinical outcomes with time-dependent FIB4 ($\Rightarrow$ imputation of missing values in baseline covariates and baseline lab using mixed models)

▷ **This data set involves multiple analyses**

▷ **The data used in the practicals is simulated from this data**

End of Session 2

Session 3
Frequentist methods for longitudinal data

- Classes of models
  - ▷ Non-likelihood methods: ANOVA, Least-Squares methods, GEE approaches, ... (omitted here)
  - ▷ Likelihood methods: model-based methods

3.1 Likelihood methods

- Model is specified for the responses given the covariates, e.g.
  - multivariate normal (MVN) model
  - mixed-effects models
- Parameters are estimated via maximum likelihood
- In this course we focus on mixed-effects models

3.1.1 Multivariate normal model

- Suppose n subjects with $d_i \leq d$ regular planned repeated measurements, with Gaussian responses & covariate design matrices $\boldsymbol{y}_i, \boldsymbol{X}_i \quad (i = 1, \ldots, n)$
- Gaussian likelihood $L(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y})$:

$$L(\boldsymbol{\beta}, \boldsymbol{\alpha} \mid \boldsymbol{y}_i) = \frac{1}{(2\pi)^{n_i/2}|\boldsymbol{V}_i(\boldsymbol{\alpha})|^{1/2}} \exp\left[(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T \boldsymbol{V}_i(\boldsymbol{\alpha})^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right] \text{ (i}^{th} \text{ subject)}$$

$$L(\boldsymbol{\beta}, \boldsymbol{\alpha} \mid \boldsymbol{y}) = \prod_{i=1}^{n} L(\boldsymbol{\beta}, \boldsymbol{\alpha} \mid \boldsymbol{y}_i) \quad \text{total likelihood}$$

- Note:
  - Dimension of $\boldsymbol{V}_i(\boldsymbol{\alpha})$ varies with subject
  - $\boldsymbol{V}_i(\boldsymbol{\alpha})$ is part of maximal covariance matrix $\boldsymbol{V}(\boldsymbol{\alpha})$ (dimension d)
  - Missing responses are (implicitly) estimated from $\boldsymbol{X}_i\widehat{\boldsymbol{\beta}}$
- **Example:** Asthma RCT

3.1.2 Linear mixed (effects) model (LMM)

- LMM popular for analyzing longitudinal studies with irregular time points + Gaussian response
- Two-stage definition:
  - Each subject has his/her longitudinal profile
  - Longitudinal profiles (intercept, slope, ...) are related to covariates
- Marginal interpretation of regression coefficients
- **Example:** Jimma Infant survival study

Definition of LMM

$$Y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \boldsymbol{z}_{ij}^T \boldsymbol{b}_i + \boldsymbol{\varepsilon}_{ij}$$

$$\boldsymbol{Y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{b}_i + \boldsymbol{\varepsilon}_i$$

- Y_{ij} response for jth observation on ith subject $(i = 1, \ldots, n)$
- $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^T$: $m_i \times 1$ vector of responses
- $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}^T, \ldots, \boldsymbol{x}_{im_i}^T)^T$: $m_i \times (d+1)$ design matrix
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_d)^T$ a $(d+1) \times 1$: fixed effects
- $\boldsymbol{Z}_i = (\boldsymbol{z}_{i1}^T, \ldots, \boldsymbol{z}_{im_i}^T)^T$: $m_i \times q$ design matrix of random effects
- $\boldsymbol{b}_i = q \times 1$ vector of random effects $(i = 1, \ldots, n)$
- $\boldsymbol{\varepsilon}_i = (\boldsymbol{\varepsilon}_{i1}, \ldots, \boldsymbol{\varepsilon}_{im_i})^T$: $m_i \times 1$ vector of measurement errors

Distributional assumptions:

- $\boldsymbol{b}_i \sim \mathsf{N}_q(\mathbf{0}, \mathsf{G})$, G: $q \times q$ covariance matrix
- **G** with (j,k)th element: σ_{b_j,b_k} $(j \neq k)$, $\sigma^2_{b_j}$ $(j = k)$
- $\boldsymbol{\varepsilon}_i \sim \mathsf{N}_{m_i}(\mathbf{0}, \mathsf{R}_i)$, R_i: $m_i \times m_i$ covariance matrix often $\mathsf{R}_i = \sigma^2 \mathsf{I}_{m_i}$
- $\boldsymbol{b}_i$ statistically independent of $\boldsymbol{\varepsilon}_i$ $(i = 1, \ldots, n)$

Two formulations of LMM:

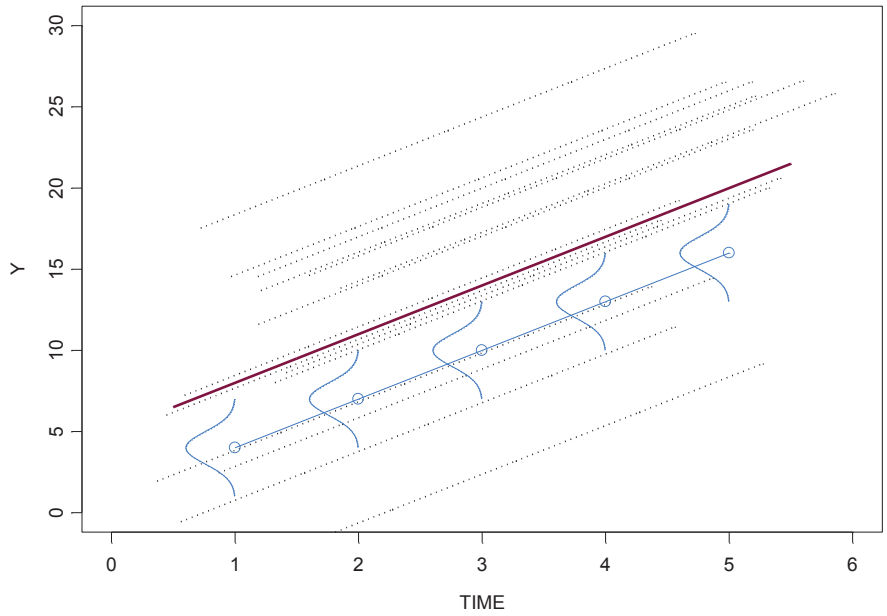
- Hierarchical/conditional: $\boldsymbol{Y}_i \mid \boldsymbol{b}_i \sim \mathsf{N}_{m_i}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i, \mathsf{R}_i) \quad \boldsymbol{b}_i \sim \mathsf{N}_q(\mathbf{0}, \mathsf{G})$
- Marginal: $\boldsymbol{Y}_i \sim \mathsf{N}_{m_i}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{Z}_i\mathsf{G}\boldsymbol{Z}_i^T + \mathsf{R}_i)$

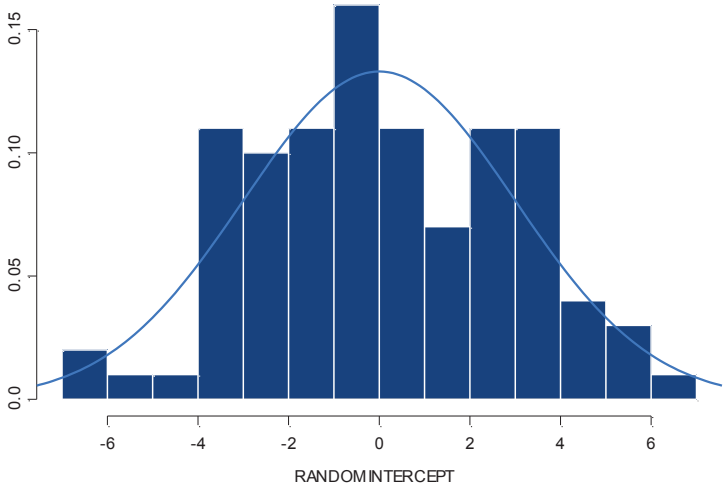
- ML and REML estimation of fixed effects
- Marginal likelihood is maximized $\Rightarrow$ random effects are integrated out before estimation

Distributional assumptions

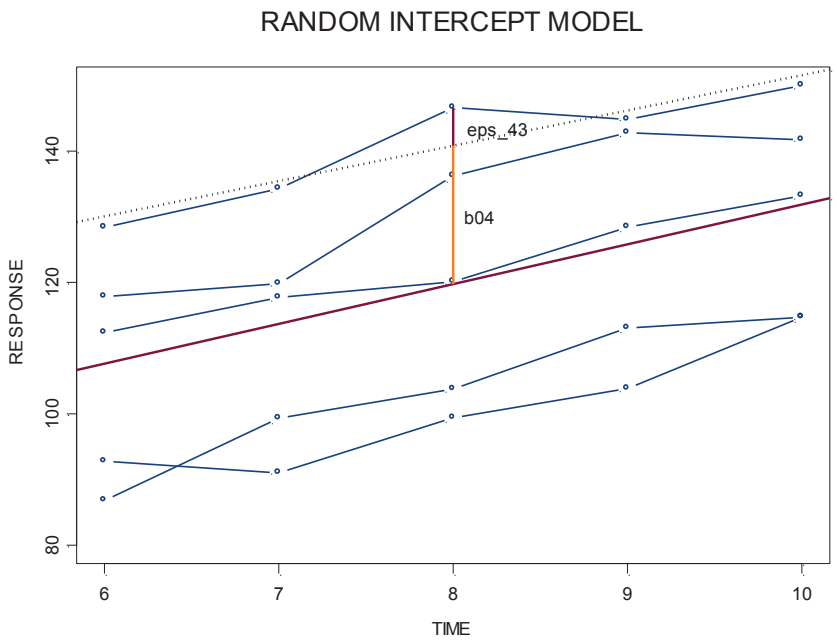
$$Y_i \mid \mathbf{b}_i \sim N_{m_i}\left(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \mathbf{R}_i\right)$$

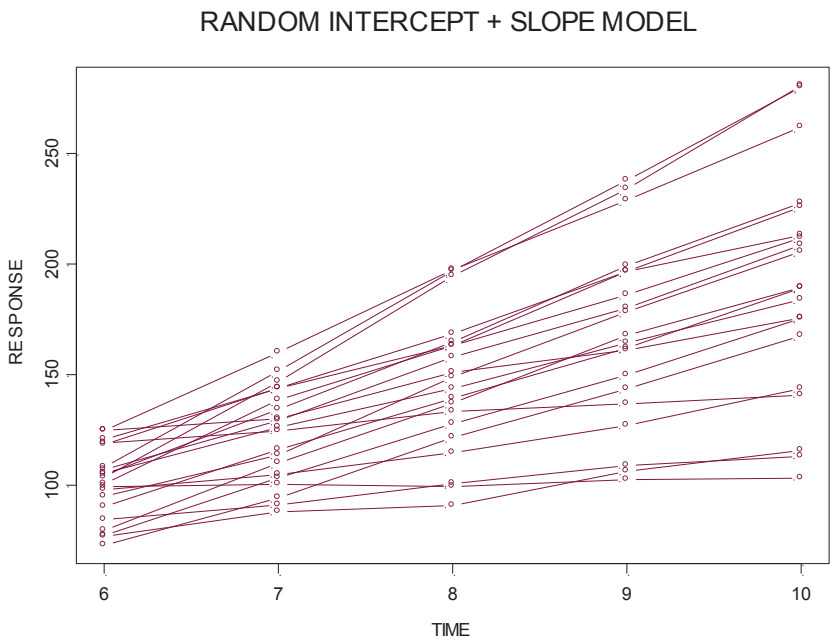
$$\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{G})$$





Random intercept & random intercept + slope model





- RI model: $Y_{ij} = \beta_0 + \beta_1 \,\mathsf{time}_{ij} + b_{0i} + \varepsilon_{ij}$
- RI + RS model: $Y_{ij} = \beta_0 + \beta_1 \,\mathsf{time}_{ij} + b_{0i} + b_{1i}\,\mathsf{time}_{ij} + \varepsilon_{ij}$

3.1.3 Generalizations

Generalized Linear Mixed (effects) Model (GLMM)

- ▷ Two-stage definition:
  - ○ Each subject has his/her longitudinal profile, depending on covariates
  - ○ Given random effects, the response follows a GLM distribution
- ▷ LMM = special case of a GLMM, response has Gaussian distribution & identity link function
- ▷ GLMM popular for analyzing longitudinal studies with a discrete response
- ▷ Interpretation of regression coefficients:
  - ○ LMM: population-averaged & subject-specific
  - ○ GLMM: subject-specific

Multilevel models

- When there are more than 2 levels
- Split-up of the random effects structure into the different levels
- Same interpretation of regression coefficients as for GLMM

Joint modeling

- When there are more than 1 outcome measured repeatedly in time, one speaks of a multivariate mixed (effects) model

- The term **joint modeling** often refers to when a mixed model is combined with a survival outcome, to either:
  - better deal with time-dependent covariates, or
  - deal with Missing-Not-At-Random mechanisms, and then
  - it is an example of a shared parameter model

End of Session 3

Session 4
Missing data processes

- Missing data are common in all (longitudinal) studies, and there can be several reasons for encountering missing data
- In a medical context:
  - participant is not traceable anymore
  - individual is too sick to participate
  - participant is withdrawn from the study
  - participant refuses to (further) respond
  - ...
- As seen before (review papers) many medical papers lack proper treatment of the data in the presence of missing data
- STROBE guidelines (checklist observational studies) pay attention to missing data

STROBE guidelines

▷ Statistical methods 12

(a) Describe all statistical methods, including those used to control for confounding

(b) Describe any methods used to examine subgroups and interactions

(c) Explain how missing data were addressed

(d) Cohort study – If applicable, explain how loss to follow-up was addressed
Case-control study – If applicable, explain how matching of cases and controls was addressed
Cross-sectional study – If applicable, describe analytical methods taking account of sampling strategy

(e) Describe any sensitivity analyses

▷ Participants 13

(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed

(b) Give reasons for non-participation at each stage

(c) Consider use of a flow diagram

▷ Descriptive data 14

(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders

(b) Indicate number of participants with missing data for each variable of interest

(c) Cohort study – Summarise follow-up time (e.g., average and total amount)

Methods used to deal with missing data

- Plenty of approaches have been suggested to deal with missing data:
  - ▷ Complete case analysis
  - ▷ Last-observation-carried forward method
  - ▷ Missing indicator method
  - ▷ Multiple imputation: assuming MVN or using chained equations approach (MICE)
  - ▷ Likelihood approach - Bayesian approach
  - ▷ Weighing methods
- Their appropriateness depends on the missing data mechanism

4.1 Missing data processes

- The appropriateness of the methods that deal with missing data depends on the missing data process
- Rubin (1976) and Little & Rubin (1987) defined the well-known taxonomy:
  - Missing completely at random (MCAR): The missing values generating mechanism works completely independent of the (observed and unobserved) data
  - Missing at random (MAR): The missing data mechanism depends only on the observed data
  - Missing not at random (MNAR): The missing data mechanism depends also on the unobserved data
- This taxonomy applies to both missing responses & missing covariates, but also to a combination of responses and covariates.
- On next 2 slides the formal definition of MCAR, MAR and MNAR is given for missing responses. A similar thing can be done for missing covariates

Formal missing data notation

- Note:
  - $\boldsymbol{Y}$ **represents**: responses with possibly missing values
  - $\boldsymbol{X}$ **represents**: covariates without missing values
- Planned data (omit i): $\boldsymbol{Y} = (Y_1, \ldots, Y_m)^T$, with distribution $p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\theta})$
- Some data are missing: $R_j = 0/1$ if Y_j is non-missing/missing

  $\Rightarrow \mathbf{R} = (R_1, \ldots, R_m)^T$ with distribution $p(\boldsymbol{r} \mid \boldsymbol{X}, \boldsymbol{\psi})$
- Observed data: $\boldsymbol{Y}^{obs}$, missing data: $\boldsymbol{Y}^{mis}$
- Total observed data: $(\boldsymbol{Y}^{obs}, \mathbf{R})$
- Full data $(\boldsymbol{Y} = \{\boldsymbol{Y}^{obs}, \boldsymbol{Y}^{mis}\}, \mathbf{R})$ with model:

  $$p(\boldsymbol{y}, \boldsymbol{r} \mid \boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\psi}) = p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\theta}) p(\boldsymbol{r} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\psi})$$
- Full data model: $p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\psi}) = \sum_{\boldsymbol{r}} p(\boldsymbol{Y}, \boldsymbol{r} \mid \boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\psi})$

Taxonomy: formal description

Due to Rubin (1976) and Little & Rubin (1987):

- Missing completely at random (MCAR):

$$p(\boldsymbol{r} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\psi}) = p(\boldsymbol{r} \mid \boldsymbol{X}, \boldsymbol{\psi})$$

- Missing at random (MAR):

$$p(\boldsymbol{r} \mid \boldsymbol{y}^{obs}, \boldsymbol{y}^{mis}, \boldsymbol{X}, \boldsymbol{\psi}) = p(\boldsymbol{r} \mid \boldsymbol{y}^{obs}, \boldsymbol{X}, \boldsymbol{\psi})$$

- Missing not at random (MNAR):

$$p(\boldsymbol{r} \mid \boldsymbol{y}^{obs}, \boldsymbol{y}^{mis}, \boldsymbol{X}, \boldsymbol{\psi}) \neq p(\boldsymbol{r} \mid \boldsymbol{y}^{obs}, \boldsymbol{y}^{mis*}, \boldsymbol{X}, \boldsymbol{\psi}),$$

  with $\boldsymbol{y}^{mis*}$ different from $\boldsymbol{y}^{mis}$, but also missing

- Note: one cannot distinguish MAR from MNAR based on the observed data

Taxonomy in longitudinal setting

▷ This implies in a longitudinal setting with:

- time points $\{t_1, t_2, \ldots, t_k, \ldots\}$ and corresponding responses $\{Y_1, Y_2, \ldots, Y_k, \ldots\}$
- (time-dependent) covariates $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_k, \ldots\}$

for **responses**:

- **MCAR**: probability that response Y_k is missing does not depend on any other responses, nor on any measured or unmeasured covariates
- **MAR**: probability that response Y_k is missing may depend on $\{Y_1, Y_2, \ldots, Y_{k-1}\}$ but not on $\{Y_k, Y_{k+1}, \ldots\}$. Further, it may depend on measured covariates $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_k, \ldots\}$ but not on unmeasured covariates
- **MNAR**: probability that response Y_k is missing depends on other responses and/or measured and/or unmeasured covariates

for **covariate** at time t_k, say X_{jk}:

- **MCAR**: probability that X_{jk} is missing does not depend on any responses nor on other (measured or unmeasured) covariates
- **MAR**: probability that X_{jk} is missing does not depend on X_{jk} (nor on unmeasured covariates) but may depend on $\{Y_1, Y_2, \ldots, Y_k\}$ and on $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_k\}$ (excluding X_{jk})
- **MNAR**: probability that X_{jk} is missing may depend on X_{jk} and other responses & other measured and unmeasured covariates

▷ **Note**: in this course, we assume the MAR missing data mechanism

Hidden assumptions

- Taxonomy assumes:
  - ▷ Correct model assumptions, e.g. that covariance matrix is correctly specified in a longitudinal study
  - ▷ All necessary covariates are included in the correct shape
- Under model misspecification, MCAR may look like MAR, MAR may look like MNAR
- Note: In basically every study the missing data mechanism is a combination of MCAR, MAR, MNAR
- **Note**: apart from MAR we also assume in this course ignorability: parameters of main model are different from the missing data model

Section 4.2
Statistical methods to deal with missing data

Bayesian Methods for Missing Covariates in Longitudinal Studies

IBC2022 in Riga – July 2022

Statistical methods to deal with missing data

Rubin (2004; actually in the 1970s):

- One imputed value cannot be correct in general.
  ➡ Need to represent missing values by a **number of imputations**.
- To find sensible values to fill in, we need some kind of model.

➡ Missing data has a **distribution** that depends on assumptions that have been made about the model.

➡ Impute missing values in $\mathbf{x}_k$ from the "predictive distribution" of the missing values given the observed values:

$$p(\mathbf{x}_k^{mis} \mid \mathbf{y}, \mathbf{X}_{-k}, \mathbf{x}_k^{obs}) \qquad \text{where } \mathbf{x}_k = (\mathbf{x}_k^{obs}, \mathbf{x}_k^{mis})$$

Statistical methods to deal with missing data

There is no unique (best) way to deal with missing data in a statistical analysis!

But, generally accepted as poor approaches:

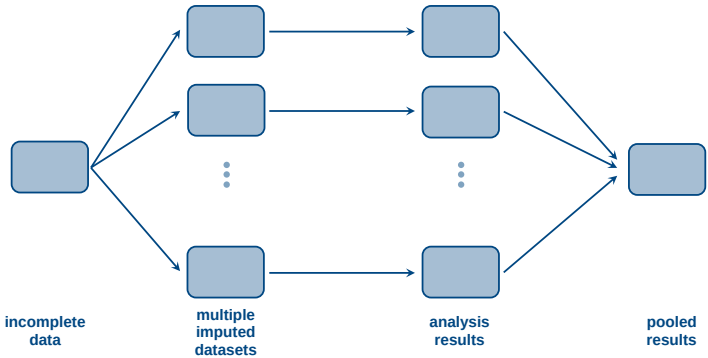
- **Complete case analysis**: applicable when MCAR, but see also White & Carlin (SIM, 2010)
- **LOCF**: for longitudinal data; was quite popular for years, even with FDA; now generally discouraged
- **Single imputation**

Better ways are:

- **Multiple Imputation** (MI)
- **Likelihood & Bayesian approach**

The focus of this course is on the Bayesian approach

Multiple Imputation: Three Steps



1. **Imputation:** impute multiple times ➡ multiple completed datasets
2. **Analysis:** analyse each of the datasets
3. **Pooling:** combine results, taking into account additional uncertainty

Multiple Imputation: Imputation Step

Univariate Missingness

y	x_1	x_2	x_3	…
✓	NA	✓	✓	…
✓	✓	✓	✓	…
✓	NA	✓	✓	…
✓	✓	✓	✓	…
⋮	⋮	⋮	⋮	⋱

Analysis model of interest

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_d x_{id} + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2_\varepsilon)$$

Impute missing values in $\mathbf{x}_1$ from

$$p(\mathbf{x}_1^{mis} \mid \mathbf{y}, \mathbf{x}_1^{obs}, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_d)$$

Notation:

y^{mis}	x_1^{mis}	x_2^{mis}	x_3^{mis}	…
✓	NA	✓	✓	…
⋮	⋮	⋮	⋮	⋮
✓	NA	✓	✓	…

y^{obs}	x_1^{obs}	x_2^{obs}	x_3^{obs}	…
✓	✓	✓	✓	…
⋮	⋮	⋮	⋮	⋮
✓	✓	✓	✓	…

Multiple Imputation: Imputation Step

In practice:

- Fit a (linear) regression model for $\mathbf{x}_1$ on cases with $\mathbf{x}_1$ observed:

$$x_{i1}^{obs} = \gamma_0 + \gamma_1 y_i^{obs} + \gamma_3 x_{i2}^{obs} + \ldots + \gamma_d x_{id}^{obs} + \xi_i, \qquad \xi_i \sim N(0, \sigma_\xi^2),$$

- sample m ($\approx 10 - 20$) imputed values for x_{i1}^{mis} from

$$p(x_{i1}^{mis} \mid y_i^{mis}, x_{i2}^{mis}, \ldots, x_{id}^{mis}, \hat{\boldsymbol{\gamma}}, \hat{\sigma}_\xi) \sim N(\eta_i^{mis}, \hat{\sigma}_\xi^2)$$

  with

$$\eta_i^{mis} = \hat{\gamma}_0 + \hat{\gamma}_1 y_i^{mis} + \hat{\gamma}_3 x_{i2}^{mis} + \ldots + \hat{\gamma}_d x_{id}^{mis}$$

- Create m completed datasets.

Multiple Imputation: Imputation Step

This **implies MAR**:

We assume that missing observations $\mathbf{x}_k^{mis}$ are from the same distribution,

$$p(\mathbf{x}_k^{mis} \mid \mathbf{y}^{mis}, \mathbf{X}_{-k}^{mis}, \gamma),$$

as the available observations $\mathbf{y}^{obs}$,

$$p(\mathbf{x}_k^{obs} \mid \mathbf{y}^{obs}, \mathbf{X}_{-k}^{obs}, \gamma),$$

conditional on the other variables, $\mathbf{y}$ and $\mathbf{X}_{-k}$.

Multiple Imputation: Analysis & Pooling

- **Analyse** each of the m datasets
  ➡ parameter estimates $\tilde{\boldsymbol{\beta}}^1, \ldots, \tilde{\boldsymbol{\beta}}^m$

- **Pool** the $\tilde{\boldsymbol{\beta}}$ using **Rubin's Rules**:
  - overall estimate:

$$\bar{\boldsymbol{\beta}} = \frac{1}{m} \sum_{k=1}^{m} \tilde{\beta}^k$$

  - variance

$$T = \overline{W} + B + \frac{1}{m} B$$

with $\overline{W} = \frac{1}{m} \sum \widehat{\text{var}}(\tilde{\boldsymbol{\beta}}^k)$ (average within imputation variance) and $B = \frac{1}{m-1} \sum (\tilde{\boldsymbol{\beta}}^k - \bar{\boldsymbol{\beta}})^2$ (between imputation variance)

Multiple Imputation: General Case

Multiple Imputation

- is the most (?) popular approach to deal with **missingness in covariates**,
- can also handle missing values in the **response**
- can be applied to **any statistical model** with incomplete data.

- Additional (**"auxiliary"**) covariates, i.e., variables not especially predictive for the response, may be used to improve the imputation.
- MI is (in general) based on the **MAR and ignorability** assumptions (extensions to MNAR are possible).

Multiple Imputation for Multivariate Missingness

From the previous slide(s):
The (incomplete) covariates must be assumed stochastic.

➡ When there are missing values in multiple variables (multivariate missingness)

$$p(\mathbf{X}^{mis} \mid \mathbf{X}^{obs}, \mathbf{y}, \gamma)$$ where $\mathbf{X}^{mis}$: incomplete variables; $\mathbf{X}^{obs}$: compl. observed variables

is **multivariate**.

- $p(\mathbf{X}^{mis} \mid \mathbf{y}, \mathbf{X}^{obs}, \gamma)$ is difficult to specify when covariates are of mixed type (continuous, categorical)
- Approximation via
  - multivariate normal or multinomial distribution (“joint model MI”)
  - **Multivariate Imputation by Chained Equations (MICE)**
    aka Fully Conditional Specification (FCS)

Multiple Imputation: MICE

- Direct extension of approach for univariate missingness.
- Based on the idea of the **Gibbs sampler**:
  iterative sampling from full-conditional distributions
  - $p(\mathbf{x}_1^{mis} \mid \mathbf{x}_1^{obs}, \mathbf{x}_2, \ldots, \mathbf{x}_d, \mathbf{y}, \gamma)$
  - $p(\mathbf{x}_2^{mis} \mid \mathbf{x}_1, \mathbf{x}_2^{obs}, \ldots, \mathbf{x}_d, \mathbf{y}, \gamma)$
  - …
  - $p(\mathbf{x}_d^{mis} \mid \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d^{obs}, \mathbf{y}, \gamma)$
  - $p(\mathbf{y}^{mis} \mid \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d, \mathbf{y}^{obs}, \gamma)$

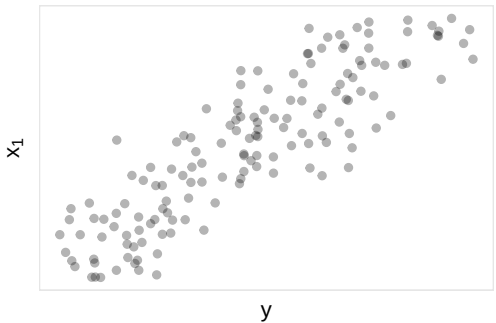
- Start by randomly drawing initial values from the observed data.
- Usually converges fast ($\approx$ 20 iterations).

MICE: Potential Issues

In practice, the full-conditionals are specified using regression models.
➡ This typically implies linear associations between variables.

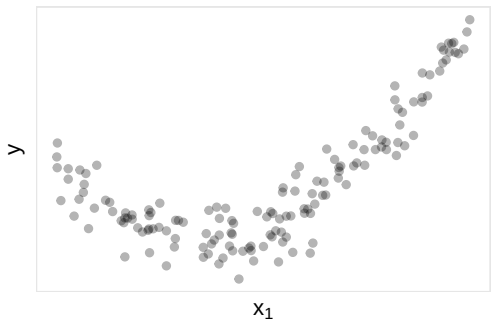
Imputation Model:
Linear association:

$$\mathbf{x}_1 = \gamma_0 + \gamma_1 \mathbf{y} + \gamma_2 \mathbf{x}_2 + \gamma_3 \mathbf{x}_3 + \dots$$



Analysis Model:
Quadratic association:

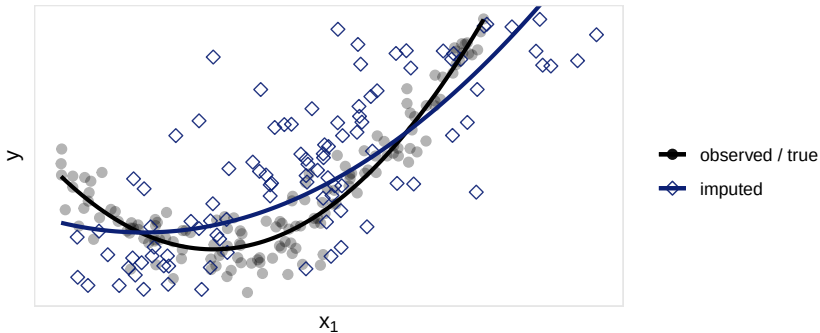
$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_1^2 + \beta_3 \mathbf{x}_2 + \dots$$



MICE: Potential Issues

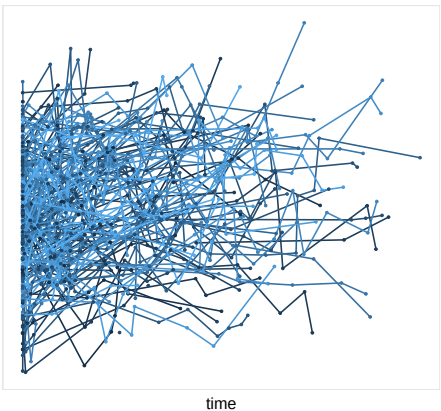
- true association: **non-linear**
- imputation assumption: **linear**

} ➡ bias!



MICE: Potential Issues

Example: Chronic Hepatitis C



(Unbalanced) longitudinal data

➡ **long format**

id	*logFIB4*	time	race	Albumin	…
1	✓	✓	NA	✓	…
1	✓	✓	NA	NA	…
1	✓	✓	NA	NA	…
1	✓	✓	NA	✓	…
2	✓	✓	✓	✓	…
2	✓	✓	✓	NA	…
⋮	⋮	⋮	⋮	⋱	

Imputation of baseline covariates conditional on longitudinal variables.

MICE: Potential Issues

Specification of $p(\mathbf{X}^{mis} \mid \mathbf{y}, \mathbf{X}^{obs}, \gamma)$ is not straightforward when there are **non-linear associations**, e.g.,

- non-linear functions of incomplete covariates
- interactions with incomplete covariates
- analysis models assuming non-linear associations (e.g., Cox model)

and in **multi-level settings**.

The (simple regression) models used in MICE are likely

- incompatible (with each other),
- uncongenial (with the analysis model), and, thus,

➡ do not result in a good approximation of $p(\mathbf{X}^{mis} \mid \mathbf{y}, \mathbf{X}^{obs}, \gamma)$.

Compatibility & Congeniality

Compatibility[1]

$p(A \mid B)$ and $p(B \mid A)$ are **compatible** if there exists a joint distribution $p(A, B)$ with $p(A \mid B)$ and $p(B \mid A)$ as its conditional distributions.

Congeniality[2]

The imputation model(s) are compatible with the analysis model.

[1] Arnold, B. C., & Press, S. J. (1989). Compatible conditional distributions. *Journal of the American Statistical Association*, 84(405), 152-156. DOI: 10.2307/2289858

[2] Xie, X., & Meng, X. L. (2017). Dissecting multiple imputation from a multi-phase inference perspective: what happens when God's, imputer's and analyst's models are uncongenial?. *Statistica Sinica*, 1485-1545. DOI: 10.5705/ss.2014.067

Likelihood Approach

For **missing responses**, there are three established approaches that differ in how the probability of missingness is modelled, which are examples of a **likelihood approach**.

- **Selection Model**:

$$p(\mathbf{y}, \mathbf{r} \mid \mathbf{X}, \theta, \psi) = p(\mathbf{y} \mid \mathbf{X}, \theta)\, p(\mathbf{r} \mid \mathbf{y}, \mathbf{X}, \psi)$$

- **Pattern Mixture Model**:

$$p(\mathbf{y}, \mathbf{r} \mid \mathbf{X}, \alpha, \phi) = p(\mathbf{y} \mid \mathbf{r}, \mathbf{X}, \alpha)\, p(\mathbf{r} \mid \mathbf{X}, \phi)$$

- **Shared Parameter Models**:

$$\begin{aligned} p(\mathbf{y}, \mathbf{r} \mid \mathbf{X}, \eta, \nu) &= \int p(\mathbf{y}, \mathbf{r} \mid \mathbf{X}, \mathbf{b}, \eta, \nu)\, dF(\mathbf{b}) \\ &= \int p(\mathbf{y} \mid \mathbf{X}, \mathbf{b}, \eta)\, p(\mathbf{r} \mid \mathbf{X}, \mathbf{b}, \nu)\, dF(\mathbf{b}) \end{aligned}$$

Likelihood Approach

- **Selection Model**: missing data model can be ignored if
  - **MAR**: $p(\mathbf{r} \mid \mathbf{y}^{obs}, \mathbf{y}^{mis}, \mathbf{X}, \psi) = p(\mathbf{r} \mid \mathbf{y}^{obs}, \mathbf{X}, \psi)$
  - **Ignorability assumption**: θ and ψ have no parameters in common

$$\begin{aligned} p(\mathbf{y}^{obs}, \mathbf{r} \mid \mathbf{X}, \theta, \psi) &= \int p(\mathbf{y}^{obs}, \mathbf{y}^{mis} \mid \mathbf{X}, \theta)\, p(\mathbf{r} \mid \mathbf{y}^{obs}, \mathbf{y}^{mis}, \mathbf{X}, \psi)\, d\mathbf{y}^{mis} \\ &= p(\mathbf{y}^{obs} \mid \mathbf{X}, \theta)\, p(\mathbf{r} \mid \mathbf{y}^{obs}, \mathbf{X}, \psi) \end{aligned}$$

    ➡ straightforward for missingness in **y**

- **Pattern Mixture Model**: provide a more elegant manner to allow for MNAR, see Daniels & Hogan (2008)
- **Shared Parameter Model**: one way to describe MNAR dropout process

Likelihood Approach for Incomplete Covariates

The extension to incomplete covariates requires integrating out the missing values $\mathbf{x}^{mis}$.

- Typically no simple solution.
- Complex integrals for multivariate missingness.

➡ The Bayesian approach avoids solving those integrals via data augmentation.

Focus of this course

- Use of the Bayesian approach, which is based on combining a likelihood with prior information (see next session).
- Use of selection models and shared parameter models.

End of Session 4

Session 5
Bayesian methods for longitudinal studies

- General introduction to Bayesian methodology
- Computational aspects
- Software
- Bayesian approach to deal with missing data: a quick intro

Central idea of Bayesian approach

Combine data (likelihood) with Your prior knowledge (prior probability) to update information on the parameter to result in a revised probability associated with the parameter (posterior probability)

5.1 A quick introduction to the Bayesian approach

- Suppose data $\boldsymbol{y} = \{y_1, \ldots, y_n\}$ have been observed following a distribution $p(\boldsymbol{y} \mid \boldsymbol{\theta})$
- Suppose also that we have some external information on the parameters $\boldsymbol{\theta}$ expressed by a distribution $p(\boldsymbol{\theta})$
- Then we could wonder how to combine the information on $\boldsymbol{\theta}$ coming from the observed data and the prior information
- The answer is given by Bayes Theorem
- Bayes Theorem stipulates that one should standardize the product of the likelihood of the data $L(\boldsymbol{\theta} \mid \boldsymbol{y}) \equiv p(\boldsymbol{y} \mid \boldsymbol{\theta})$ and the prior $p(\boldsymbol{\theta})$ to produce the posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{y})$
- The posterior distribution contains all available information on $\boldsymbol{\theta}$

Bayes theorem – Formal definition

- Bayes Theorem combines prior and data information
  - ▷ Notation:
    - ○ Unknown parameters $\boldsymbol{\theta}$ & data: $\boldsymbol{y}$
    - ○ Likelihood $L(\boldsymbol{\theta} \mid \boldsymbol{y})$: plausibility of $\boldsymbol{\theta}$ given data $\boldsymbol{y}$
    - ○ Prior $p(\boldsymbol{\theta})$: prior density of $\boldsymbol{\theta}$ values (information on $\boldsymbol{\theta}$ independent of $\boldsymbol{y}$)
    - ○ Posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y})$: posterior density of $\boldsymbol{\theta}$ values as a result of combining prior and data information
  - ▷ General Bayes Theorem:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{L(\boldsymbol{\theta} \mid \boldsymbol{y})p(\boldsymbol{\theta})}{p(\boldsymbol{y})} = \frac{L(\boldsymbol{\theta} \mid \boldsymbol{y})p(\boldsymbol{\theta})}{\int L(\boldsymbol{\theta}|\boldsymbol{y})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

- **Thus**:
  - ▷ Prior and posterior attach (Bayesian) probabilities to the different (intervals of) $\boldsymbol{\theta}$ values
  - ▷ Thereby, unknown parameters get a distribution, quite different from classical statistics

5.2 Priors: some reflections

- The prior $p(\boldsymbol{\theta})$ reflects the information that we have about $\boldsymbol{\theta}$ not based on the observed data
- There are several types of priors:
  - **Informative prior**: prior that contains genuine information on the parameter(s), is also called a subjective prior
  - **Non-informative prior**: better called vague prior has the ambition to bring in no or little information into the computations. This is also called an objective prior
  - **Proper/improper prior**: a proper prior is a classical distribution (on the parameters), while an improper prior has infinite AUC
  - **Conjugate prior**: a prior combined with likelihood gives a posterior of the same type as the prior. Examples: beta prior combined with binomial likelihood, a gamma prior combined with Poisson likelihood, a Gaussian prior combined with Gaussian likelihood (σ known)
  - The prior for a multivariate parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^T$ is often split up into priors of its components, e.g. $p(\boldsymbol{\theta}) = p(\theta_1) \times \ldots \times p(\theta_d)$ is the product of independent priors on the elements of $\boldsymbol{\theta}$.

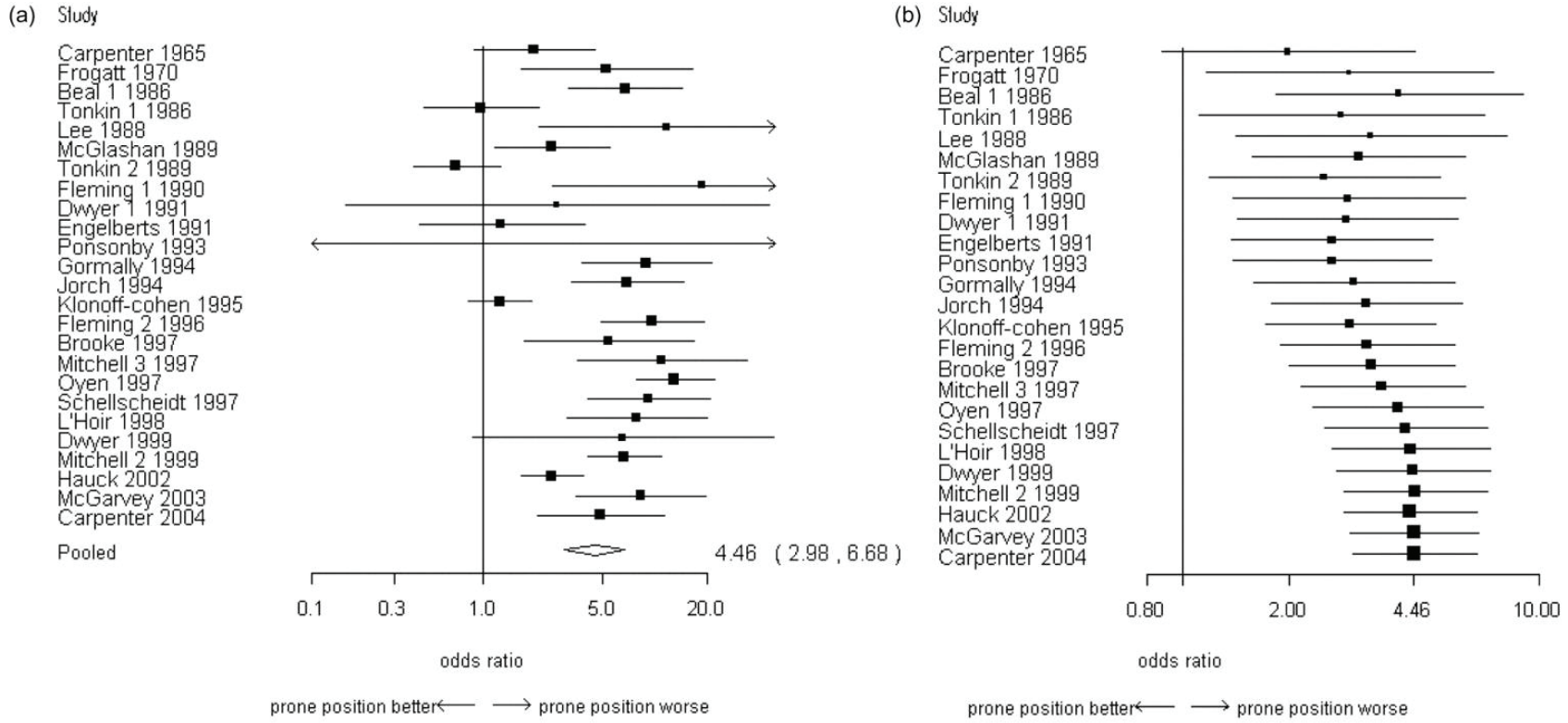
Illustration: SIDS meta-analysis

- **Question**: Should baby sleep in **front** or **back** position to avoid S(udden) I(nfant) D(eath) S(yndrome)?

?

▷ Meta-analysis by Gilbert et al. (2005)

Meta-analysis to see advantage of putting baby in front (prone) or back — Cumulative meta-analysis



- Conclusion from **cumulative meta-analysis**:
  - ▷ Advice to put infants to sleep on the front for nearly a half century was contrary to evidence available from 1970 that this was likely to be harmful.
  - ▷ A systematic review of preventable risk factors for SIDS from 1970 would have led to earlier recognition of the risks of sleeping on the front and might have prevented over 10 000 infant deaths in the UK and at least 50 000 in Europe, the USA, and Australasia.

- **Now suppose the following**:
  - ▷ My name is **Tonkin** and **I am preparing my paper on SIDS** in **1985**
  - ▷ I do a classical analysis of the obtained data, which are:

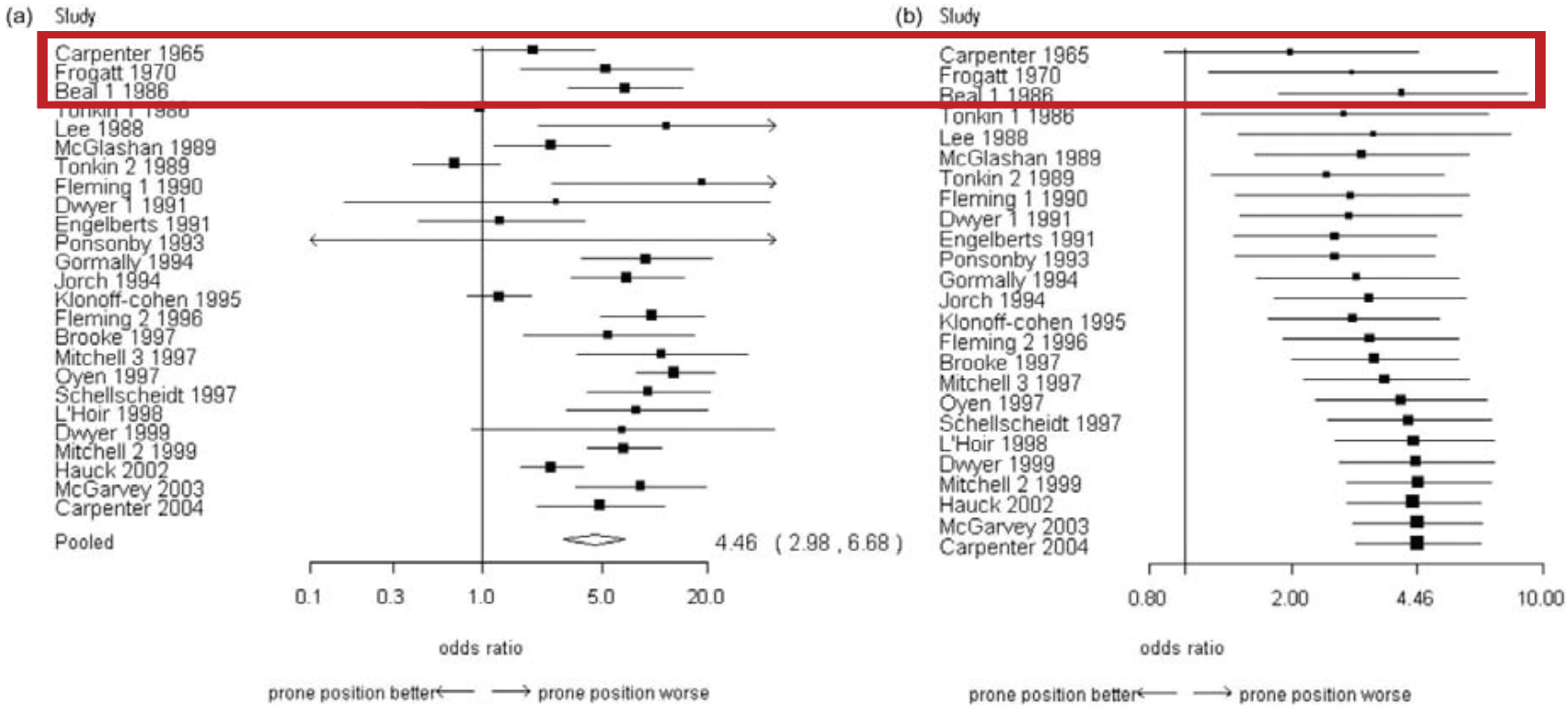
	SIDS=No	SIDS=Yes	
Front	575	51	626
Back	79	8	87
	654	59	713

  - ▷ Using a standard statistical program, I obtain for OR $(= \theta)$:

    $\widehat{OR} = 0.88$ with 95% CI = [0.40, 1.90] and P = 0.74

- **Suppose in a second step**:
  - ▷ I picked up the result of the cumulative meta-analysis after the first 3 studies
  - ▷ Roughly I observe from the figure $OR_0 = 4.20$ with 95% CI $= [1.9, 9.3]$
  - ▷ **What can I do with this prior information?** ... Some possibilities:
    1. Ignore the prior information
    2. Combine the prior information with the current data
    3. Combine discounted prior information with the current data
  - ▷ I have already done case 1
  - ▷ But how to do cases 2 and 3?
  - ▷ We (can) use **Bayesian approach**

▷ Meta-analysis by Gilbert et al. (2005)



Information from first three studies highlighted

▷ Three Bayesian analyses (details next slides) provide estimates of OR for SIDS in front position

	PRIOR INFORMATION on ODDS RATIO		
	Nothing	**Cum MA**	**Reduced**
Prior	1 $[0, 10^6]$	4.20 [1.9, 9.3]	4.20 [1.58, 11.19]
Data	0.88 [0.40, 1.90]	0.88 [0.40, 1.90]	0.88 [0.40, 1.90]
Posterior	0.88 [0.38, 1.81]	2.09 [1.11, 3.63]	1.74 [0.86, 3.39]

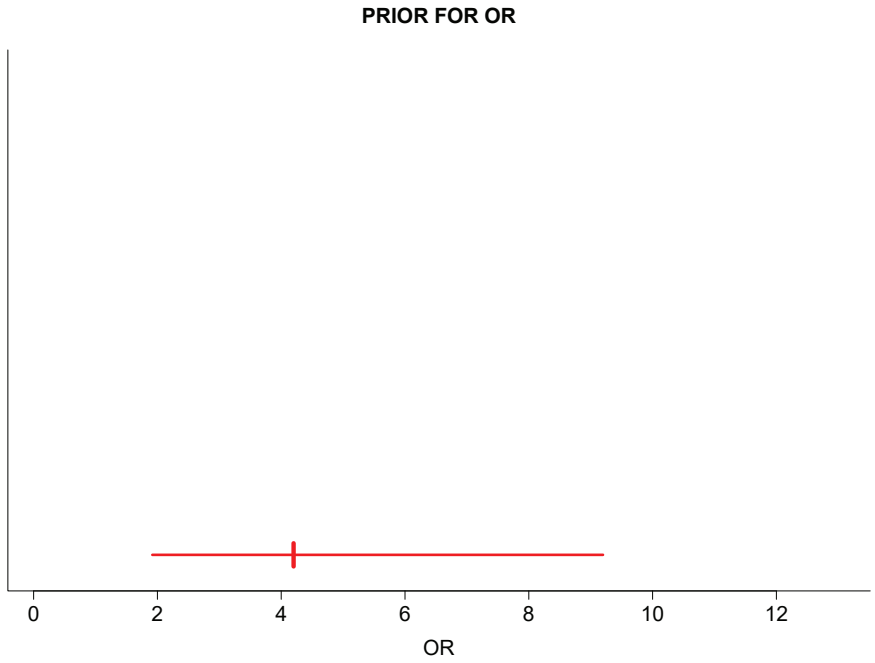
▷ Findings of Bayesian analyses:

- When results from previous 3 studies are taken into account, risk in front position for SIDS $\approx$ as in cumulative MA
- When nothing is assumed, then result of (my) Tonkin study is obtained
- When prior information on risk is reduced, then OR is moved towards 1 but still larger than 1

How to specify prior information on θ?

- Information from Carpenter (1965) & Frogatt (1970) & Beal (1986):
  $\widehat{OR} = 4.20$ with 95% CI $= [1.90, 9.30]$

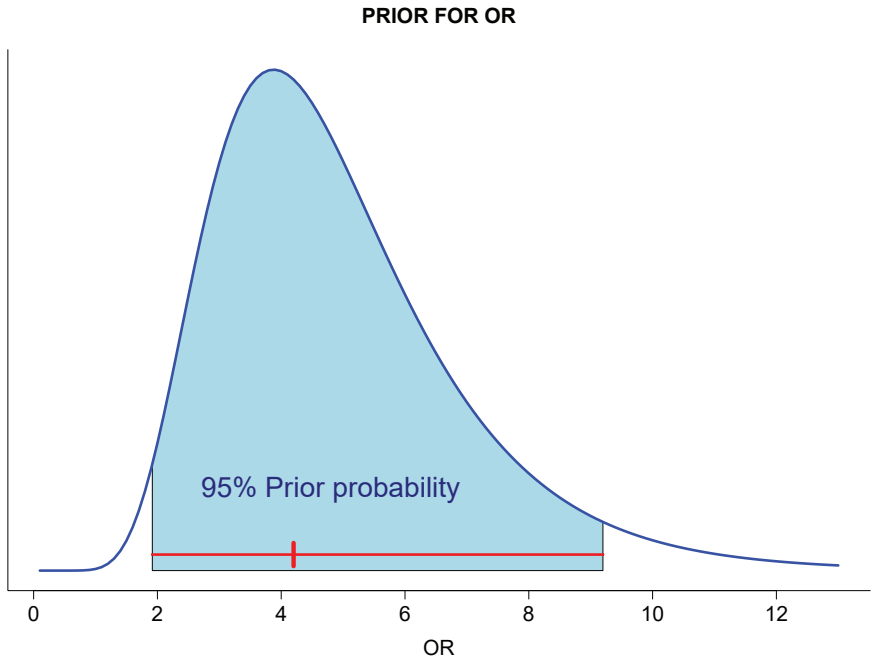
Information from first 3 studies on SIDS



How to specify prior information on θ?

- Information from Carpenter (1965) & Frogatt (1970) & Beal (1986):
  $\widehat{OR} = 4.20$ with 95% CI $= [1.90, 9.30]$

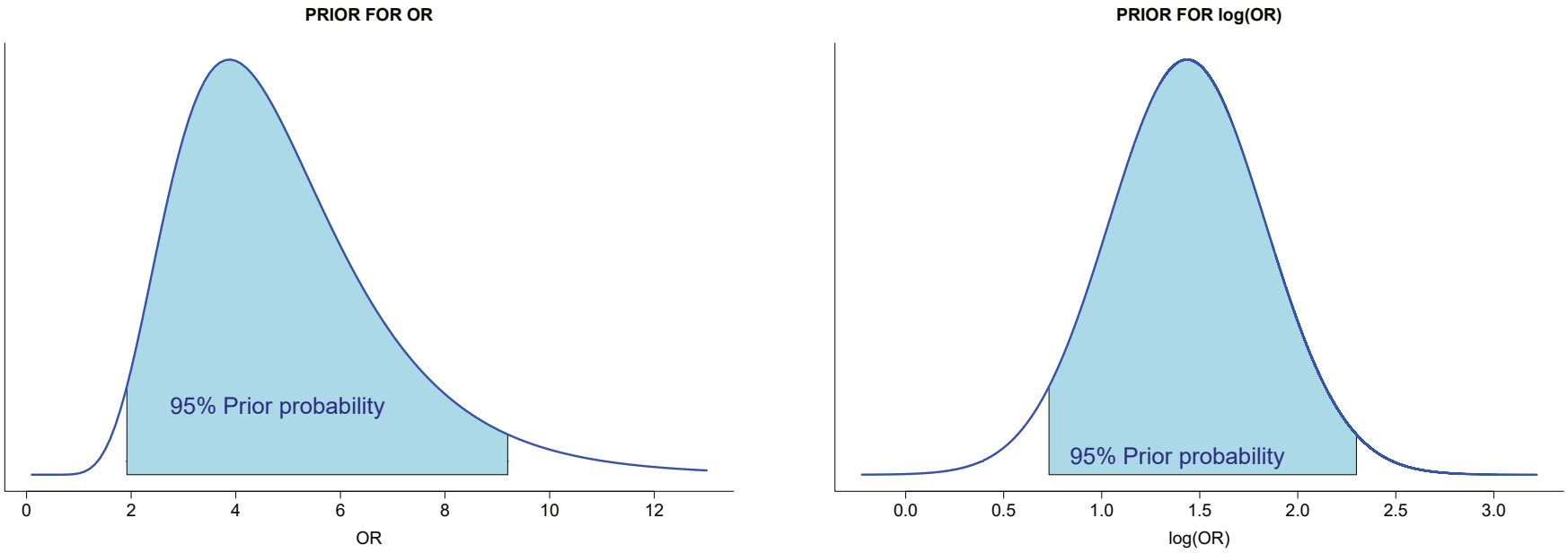
Turn information into a prior distribution



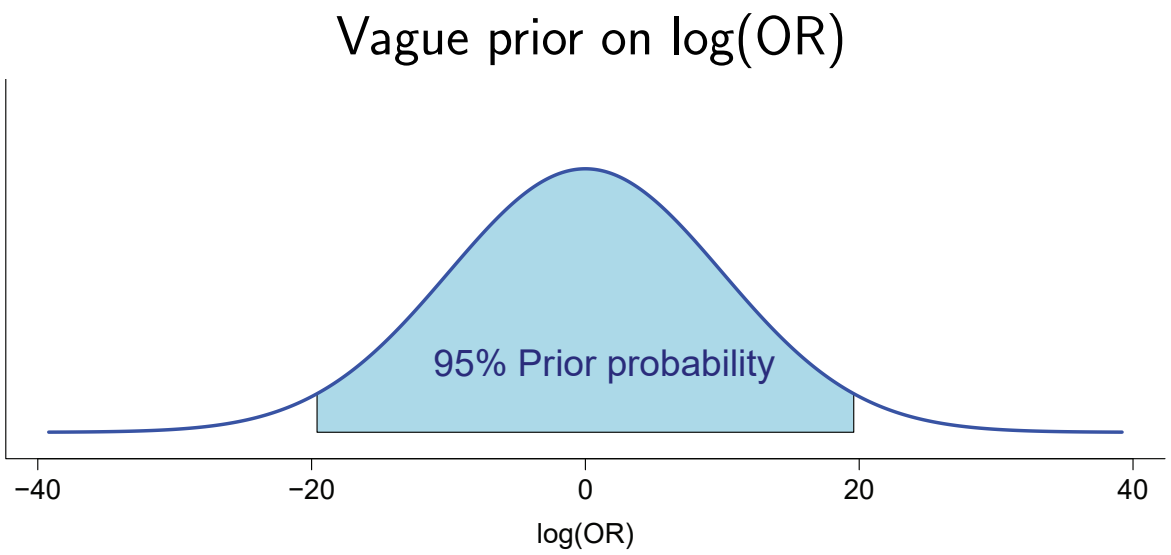
How to specify prior information on θ?

- Prior information can also be specified on log(OR)-scale
- ... and back transformed on OR-scale

Prior on OR and log(OR) from first 3 studies on SIDS



- One can also specify that there is no prior information
- ... as one does in a classical/frequentist approach
- Then one speaks of a non-informative prior, but better is to call it a vague prior

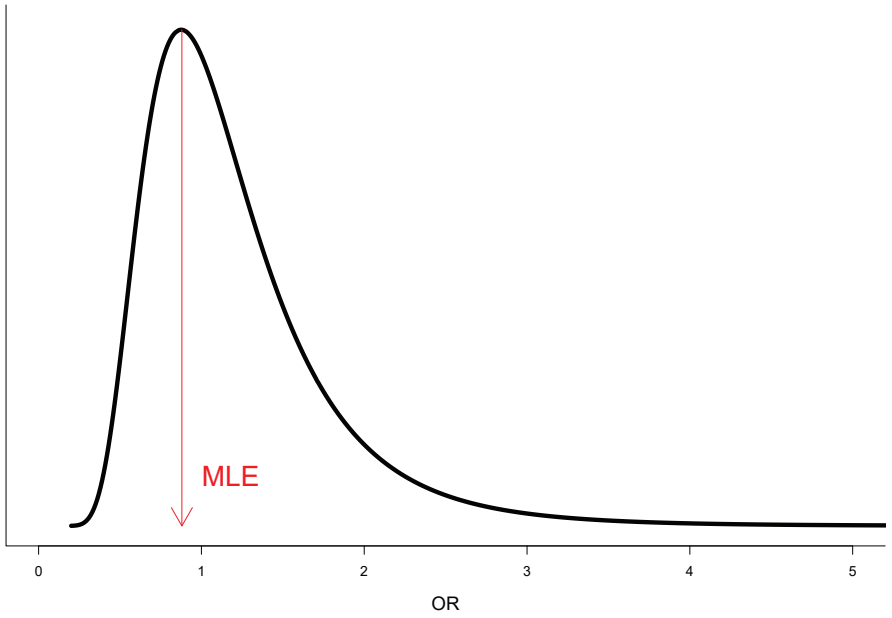


Note: This prior is in fact too "vague",
since one puts (very small) prior probability on too large and too small ORs

What information do the data bring on θ?

- Study results from Tonkin (1986):
  $\widehat{OR} = 0.88$ with 95% CI = [0.40, 1.90] and P = 0.74

Likelihood function for θ



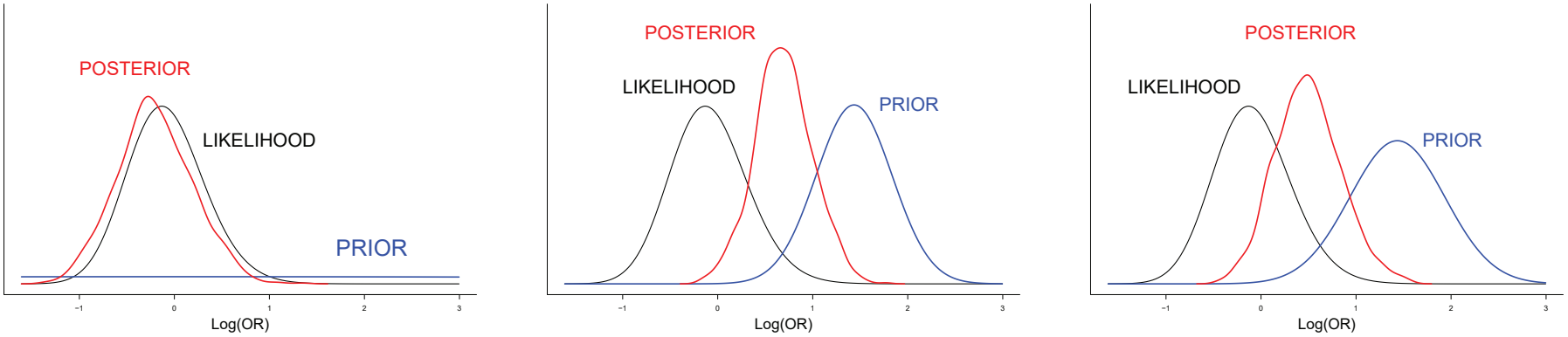
How to combine information on θ?

- Take prior on θ
- And likelihood on θ
- Take product of two functions for each value of θ
- Standardize

$\Rightarrow$ Bayes Theorem

Prior, likelihood and posterior on log(OR)

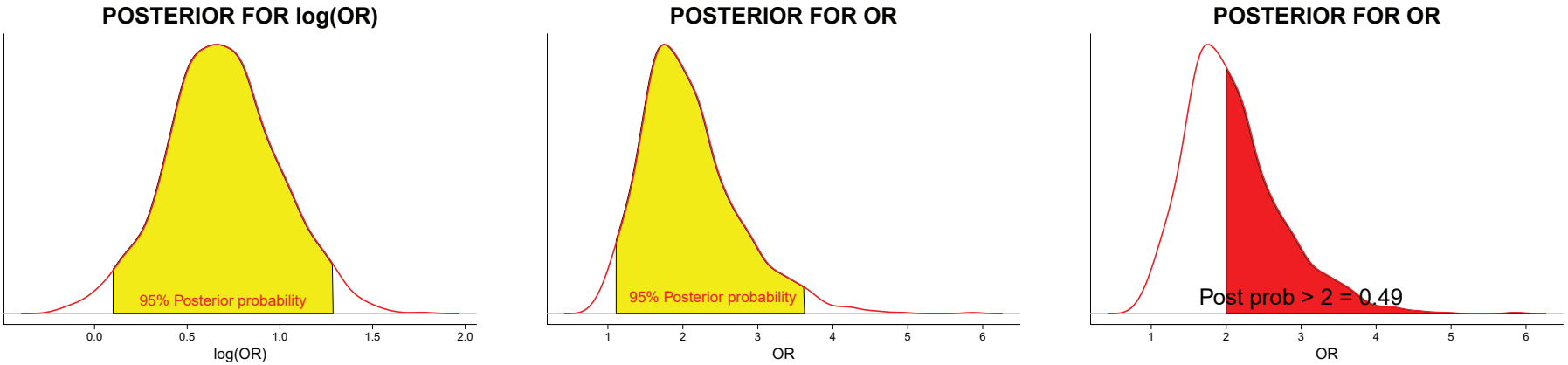
Vague prior **Prior from 1st 3 studies** **Discounted prior**



- ▷ Note:
  - ○ Prior, likelihood and posterior provide for each value of log(OR) plausible and implausible values
  - ○ Posterior combines information from prior and likelihood
  - ○ Posterior is a compromise between prior and likelihood
- ▷ Figures can also be produced for odds-ratio

- **Posterior provides all information what researcher needs**

Posterior on OR and log(OR)



- Also most probably value of OR, posterior mean & median & SD
- And ... Bayesian 95% confidence interval + ...

Asthma RCT: A Bayesian longitudinal analysis

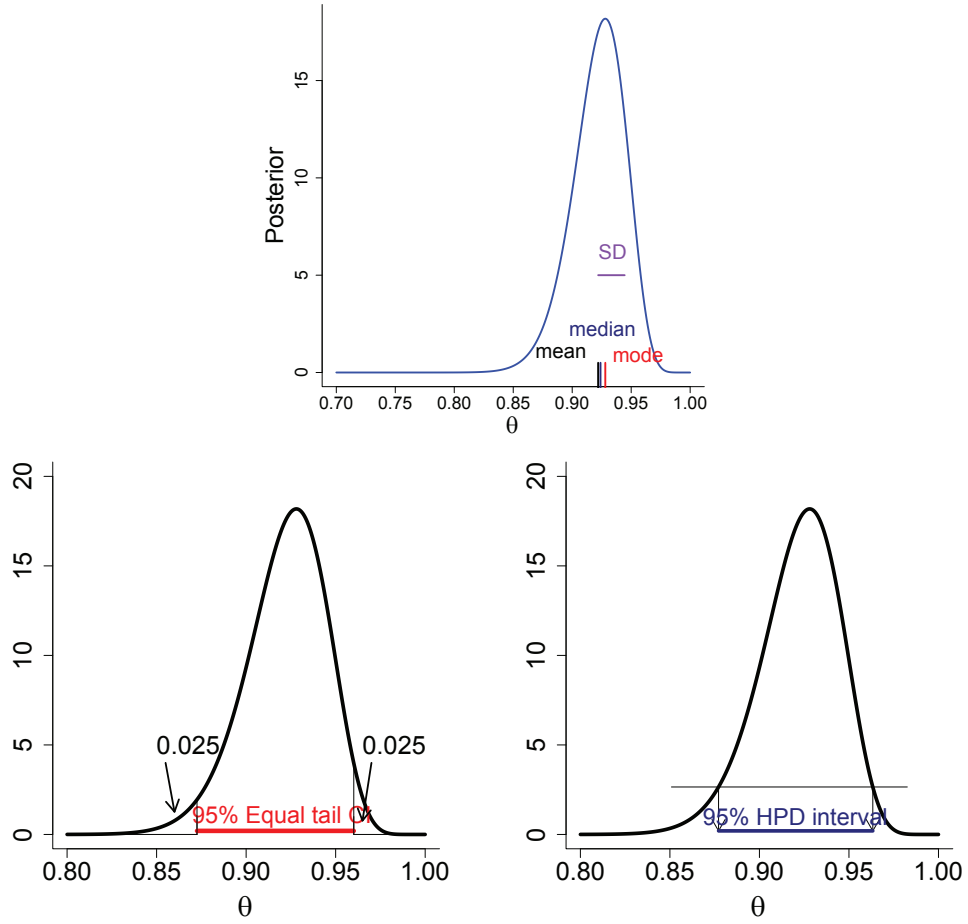
- RCT on children with mild to moderate asthma (N = 35), with outcome = (log2 transformed) PC20, measured after 1, 2 and 3 months
- Covariates: visit & treatment
- Statistical model: compound symmetry longitudinal Gaussian main model
- **Vague priors**:
  - Regression coefficients: N$(0, 10^6)$, called locally uniform priors
  - Variance σ^2: IG(0.001, 0.001), $\approx$ uniform prior on σ
  - Correlation ρ: U[-1,1], uniform prior on interval [-1,1]
- **Informative priors** based on historical study:
  - β_{treat} (L vs I) is around -0.5 with 95% CI $\approx$ [-1.5, 0.5]: inf prior $\beta_{treat} \sim$ N$(-0.5, 0.5^2)$
  - σ^2 is around 2.5 with 95% CI $\approx$ [0.5, 7.5]: inf prior $\sigma^2 \sim$ Gamma$(3, 1)$
  - ρ is around 0.5 with 95% CI $\approx$ [0.3, 0.7]: inf prior $\rho \sim$ Beta$(10, 10)$

Posterior summary measures

- **Point estimates** (univariate, for multivariate the same except for median)
  - Posterior mean: $\overline{\theta} = \int \theta\, p(\theta \mid \boldsymbol{y}) d\theta$
  - Posterior median: $0.5 = \int_{\overline{\theta}_M} p(\theta \mid \boldsymbol{y}) d\theta$
  - Posterior mode: $\widehat{\theta}_M = \arg\max_\theta p(\theta \mid \boldsymbol{y})$
  - Posterior variance: $\overline{\sigma}^2 = \int (\theta - \overline{\theta})^2\, p(\theta \mid \boldsymbol{y}) d\theta$
  - Posterior SD: $\overline{\sigma}$
- **Interval estimates**
  - Univariate: 95% equal tail credible interval or 95% highest posterior density interval
  - Multivariate: 95% HPD region
- **Posterior predictive distribution (PPD)**
  - PPD = distribution of a future observation $\widetilde{y}$ taking into account uncertainty of $\boldsymbol{\theta}$

$$p(\widetilde{y} \mid \boldsymbol{y}) = \int p(\widetilde{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \boldsymbol{y})\, d\boldsymbol{\theta}$$

Posterior summary measures: example



Top: posterior point estimates, Bottom: posterior interval estimates

Bayesian versus frequentist approach

- **Frequentist approach**:
  - θ fixed and inference based on sampling variability
  - Many tests are based on asymptotic arguments
  - Maximization is key tool
  - Hierarchical models: random effects are integrated out
- **Bayesian approach**:
  - Condition on observed data (data fixed), uncertainty about θ (θ stochastic)
  - No asymptotic arguments are needed, all inference depends on posterior
  - Integration is key tool
  - Hierarchical models: random effects are estimated together with fixed effects
- But: frequentist and Bayesian approach may give the same numerical output

5.3 Computational aspects

- To apply Bayes Theorem, one needs to compute a (often complex) integral
- This is only possible in simplest cases
- **About 250 years** one has been looking how to do that in practice
- The solution lies in bypassing the integral by sampling the posterior, and thereby approximating the posterior summary measures
- This is what is called Markov chain Monte Carlo sampling
- Most famous general software is WinBUGS

5.3.1 The Markov Chain Monte Carlo (MCMC) algorithm

- The Markov Chain Monte Carlo algorithm gave a new (great) impetus to the Bayesian paradigm
- Two types of MCMC algorithms:
  - The Gibbs sampler
  - The Metropolis(-Hastings) sampler
- MCMC produces a dependent sample from the posterior distribution
- MCMC enables to fit complex models, but may be quite slow

The Gibbs sampler

Starting position $\boldsymbol{\theta}^0 = (\theta_1^0, \ldots, \theta_d^0)^T$

Iteration $(k+1)$:

1. Sample $\theta_1^{(k+1)}$ from $p(\theta_1 \mid \theta_2^k, \ldots, \theta_{(d-1)}^k, \theta_d^k, \boldsymbol{y})$

2. Sample $\theta_2^{(k+1)}$ from $p(\theta_2 \mid \theta_1^{(k+1)}, \theta_3^k, \ldots, \theta_d^k, \boldsymbol{y})$

⋮

d. Sample $\theta_d^{(k+1)}$ from $p(\theta_d \mid \theta_1^{(k+1)}, \ldots, \theta_{(d-1)}^{(k+1)}, \boldsymbol{y})$

$p(\theta_j \mid \theta_1, \ldots, \theta_{(j-1)}, \theta_{(j+1)}, \ldots, \theta_d, \boldsymbol{y})$: full conditional distribution

Result of Gibbs sampling:

- Chain of vectors: $\boldsymbol{\theta}^k = (\theta_1^k, \ldots, \theta_d^k)^T, k = 1, 2, \ldots$
  - Consists of **dependent** elements
  - **Markov property**: $p(\boldsymbol{\theta}^{(k+1)} \mid \boldsymbol{\theta}^k, \boldsymbol{\theta}^{(k-1)}, \boldsymbol{\theta}^{(k-2)}, \ldots, \boldsymbol{y}) = p(\boldsymbol{\theta}^{(k+1)} \mid \boldsymbol{\theta}^k, \boldsymbol{y})$
- Chain depends on starting value + initial portion/burn-in part is **discarded**
- Under mild conditions: sample from the posterior distribution

$\Rightarrow$ From k_0 on: summary measures calculated from the chain consistently estimate the true posterior measures

The Gibbs sampler vs MICE

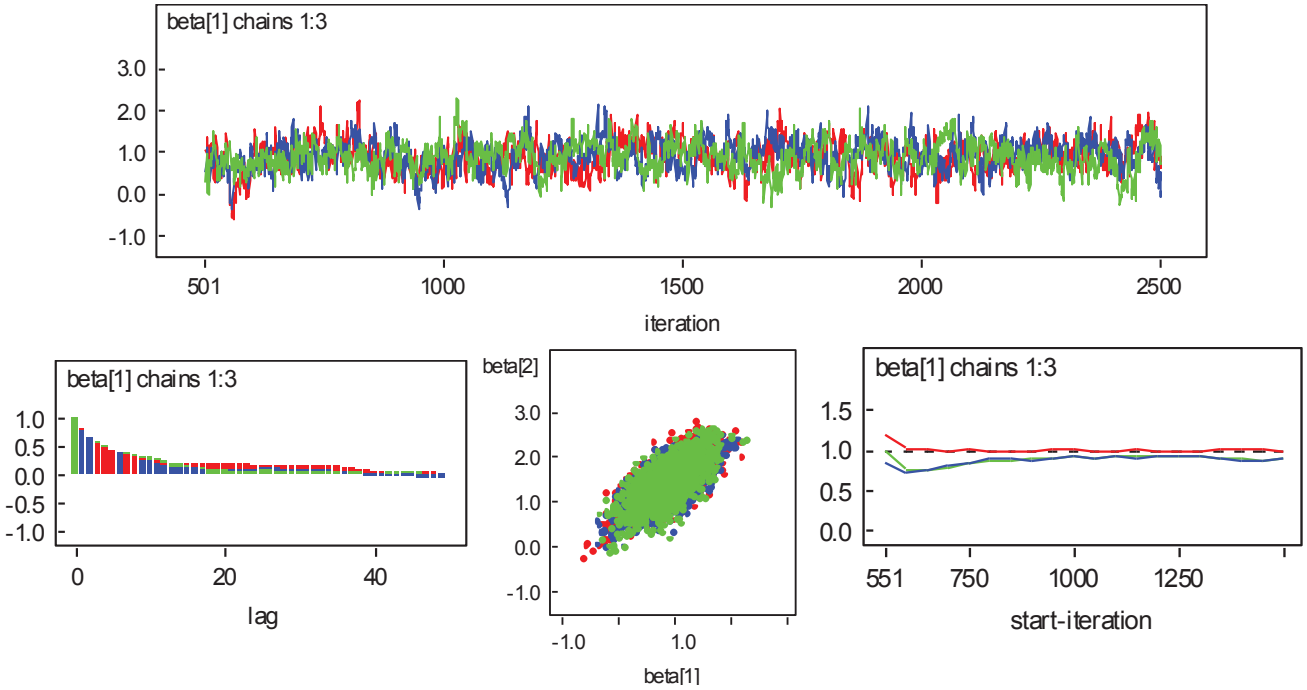
- **Gibbs sampler**:
  - We start with a joint (posterior) distribution known up to a constant
  - From the joint expression we derive the full conditionals
  - The full conditionals are sampled to obtain after K iterations a sample from the posterior
- **MICE**:
  - We start with fexible conditional distributions
  - The conditional distributions are sampled and then provide imputed values for the missing data
  - But, it is not guaranteed that the conditional distributions correspond to a joint distribution of the response(s) and covariates

Checking convergence of Gibbs sampler

- ▷ Checking for convergence
  - ○ Trace plot: visual inspection of stationarity of Markov chain
  - ○ Formal diagnostics: Geweke (single chain) & BGR-diagnostic (multiple chains)
- ▷ Computing the Monte Carlo standard error
  - ○ Sampling error of summary measures: MCerror $< 5\%$ of posterior SD
  - ○ Effective sample size (ESS)

Asthma RCT: Checking convergence

Saturated model + unstructured covariance matrix



Top: trace plot from identifiable model

Bottom, left: autocorrelation plot, Bottom, middle: cross-correlation plot, Bottom, right: Gelman-Rubin diagnostic plot

5.4 Software

- WinBUGS: Windows version of Bayesian inference Using Gibbs Sampling (BUGS)
  - Start: 1989 in MRC Biostatistics at Cambridge with BUGS
  - Final version = 1.4.3, freely available
- OpenBUGS = open source version of WinBUGS, freely available
- BUGS-related software: JAGS, NIMBLE, MultiBUGS
- Stan: different programming language, but becoming increasingly popular
- INLA and R-INLA: based on (sophisticated) Gaussian quadrature methods
- Numerous dedicated Bayesian R packages

5.5 Bayesian approach to deal with missing data: a quick intro

▷ Bayesian approach requires a likelihood

$\Rightarrow$ based on selection/pattern mixture/shared parameter model + priors on all model parameters

▷ We focus here on selection models, but have also a practical example of a shared parameter model

▷ Selection model for missing response: we need to model only the observed data

- MAR: $p(\boldsymbol{r} \mid \boldsymbol{y}^{obs}, \boldsymbol{y}^{mis}, \boldsymbol{X}, \boldsymbol{\psi}) = p(\boldsymbol{r} \mid \boldsymbol{y}^{obs}, \boldsymbol{X}, \boldsymbol{\psi})$

& Ignorability assumption: $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ have no parameters in common + $p(\boldsymbol{\theta}, \boldsymbol{\psi}) = p(\boldsymbol{\theta})p(\boldsymbol{\psi})$

▷ In this course we need to extend this to allow for also missing values in covariates

▷ Note: in the Bayesian approach the missing values are considered as parameters and thus can be given an informative prior

A linear regression analysis with missing data

- Cross-sectional study (Boonen et al., 1996)
- 245 healthy elderly women in a geriatric hospital
- Aim: Find determinants for osteoporosis
- Average age women = 75 yrs with a range of 70-90 yrs
- Marker for osteoporosis = tbbmc (in kg) measured for 234 women
- Multiple linear regression model: regressing tbbmc on age (yrs), weight (kg), length (cm), strength
- Two analyses:
  - Without missing data
  - Missing data created in tbbmc, weight and strength
- Analysis is done using OpenBUGS, see osteomultipleregression missing values.odc

Generated missing data

Subject #	age[]	length[]	weight[]	strength[]	tbbmc[]
1	71.00	157.00	67.00	96.25	NA
2	73.00	163.00	71.00	85.25	NA
3	74.00	160.00	66.00	87.75	NA
4	72.00	158.00	73.00	NA	2.183
5	73.00	153.00	72.00	NA	2.010
6	76.00	152.00	50.00	NA	1.579
7	73.00	161.00	61.00	NA	1.911
8	79.00	163.00	NA	NA	1.681
9	72.00	151.00	**NA**	54.00	NA
10	71.00	159.00	NA	76.00	2.138
11	74.00	168.00	NA	79.25	2.535
12	73.00	152.00	NA	58.50	1.480

Regression model imputing missing data

```
model{

# main model
for (i in 1:N){
    tbbmc[i] ~dnorm(mu[i],tau)
    x[i,1]<-1; x[i,2]<-age[i]; x[i,3]<-weight[i]; x[i,4]<-length[i];
    x[i,5]<- strength[i]
    mu[i] <- inprod(beta[],x[i,])
}
    sigma <- pow(tau,-0.5)
    c <- 1.0E-6
    for (r in 1:5) { mu.beta[r] <- 0.0}
    for (r in 1:5) { for (s in 1:5){
    prec.beta[r,s] <- equals(r,s)*c
    }}
    beta[1:5] ~ dmnorm(mu.beta[], prec.beta[,])
    tau ~ dgamma(1.0E-3,1.0E-3)
```

```
# imputation models

for (i in 1:N){
    strength[i] ~dnorm(mu_s[i],tau_s)
    mu_s[i] <- gamma[1]+gamma[2]*age[i]+gamma[3]*weight[i]+gamma[4]*length[i]

    weight[i] ~dnorm(mu_w[i],tau_w)
    mu_w[i] <- nu[1]+nu[2]*age[i]+nu[3]*length[i]
}
     tau_s ~ dgamma(1.0E-3,1.0E-3)
     tau_w ~ dgamma(1.0E-3,1.0E-3)

     for (r in 1:4) { mu.gamma[r] <- 0.0}
     for (r in 1:3) { mu.nu[r] <- 0.0}

     for (r in 1:4) { for (s in 1:4){
     prec.gamma[r,s] <- equals(r,s)*c
     }}
     for (r in 1:3) { for (s in 1:3){
     prec.nu[r,s] <- equals(r,s)*c
     }}

    gamma[1:4] ~ dmnorm(mu.gamma[], prec.gamma[,])
    nu[1:3] ~ dmnorm(mu.nu[], prec.nu[,])
  }
```

- Red part: analysis model regressing tbbmc on age, length, weight and strength
- Blue part: 2 Gaussian regression imputation models
  1. Regressing strength on age, length and weight
  2. Regressing weight on age and length

- The BUGS program imputes the missing data from
  $f(\text{tbbmc, weight, strength} \mid \text{age, length}) =$
  $f(\text{tbbmc} \mid \text{strength, weight, age, length}) \times f(\text{strength} \mid \text{weight, age, length}) \times$
  $f(\text{weight} \mid \text{age, length})$
- Each of the models (analysis & imputation) generates values for the missing data from their respective PPD
- Except for weight[9] the 95%CI of imputed values includes observed value, see also osteomultipleregression.odc
- **Note**: a sensitivity analysis is required to see how the imputations change when:
  - the chosen imputation models for the above split up are varied
  - the split up is done differently, i.e. instead of first imputing strength given weight, and then weight one could change the order

Some remarks

- To address missing covariate data, only a model for the covariates with missing data needs to be specified
- Auxiliary variables can be included in the imputation models to further improve the predictions
- Bayesian approach provides an estimate of posterior distribution of regression coefficients averaged over missing part in covariates and response
- Since missing data in covariates and response are treated as parameters, one could also provide informative priors to missing data as a result of an elicitation process
- Finally, in Ibrahim, Chen, Lipsitz & Herring (JASA, 2005):

..., fully Bayesian methods are perhaps the most powerful and most general methods for missing covariate data.

End of Session 5

Session 6
Bayesian Analysis of Incomplete Data with JointAI

Bayesian Methods for Missing Covariates in Longitudinal Studies

IBC 2022, Riga

Statistical methods to deal with missing data

Remember:

- Rubin: impute missing values in $\mathbf{x}_k$ from $p(\mathbf{x}_k^{mis} \mid \mathbf{x}_k^{obs}, \mathbf{X}_{-k}^{obs}, \mathbf{y})$.
- Direct specification is not feasible.
- MICE approximates $p(\mathbf{X}^{mis} \mid \mathbf{y}, \mathbf{X}^{obs}, \boldsymbol{\theta})$ using the idea of Gibbs sampling.
- Issues with non-linear associations & multi-level data
- Likelihood difficult for missing covariates due to integration

Bayesian approach

- Aim: Better approximation of $p(\mathbf{X}^{mis} \mid \mathbf{X}^{obs}, \boldsymbol{\theta})$
- Avoid integrals by data augmentation (sampling)

Focus here on missing values in covariates and ignorable missingness.

Getting the Correct Distribution

- We need $p(\mathbf{X}^{mis} \mid \mathbf{y}, \mathbf{X}^{obs}, \boldsymbol{\theta})$
- We assume $p(\mathbf{y} \mid \mathbf{X}^{mis}, \mathbf{X}^{obs}, \boldsymbol{\theta}_{y|x})$ (analysis model)

Bayes Theorem:

$$
\begin{aligned}
p(\mathbf{X}^{mis} \mid \mathbf{y}, \mathbf{X}^{obs}, \boldsymbol{\theta}) &= \frac{p(\mathbf{y}, \mathbf{X}^{obs}, \boldsymbol{\theta} \mid \mathbf{X}^{mis})\, p(\mathbf{X}^{mis})}{p(\mathbf{y}, \mathbf{X}^{obs}, \boldsymbol{\theta})} \\
&\propto \underbrace{p(\mathbf{y}, \mathbf{X}^{obs}, \mathbf{X}^{mis}, \boldsymbol{\theta})}_{\text{joint distribution}}
\end{aligned}
$$

The joint distribution does not have a closed form.

Factorization of the Joint Distribution

$$\underbrace{p(\mathbf{y}, \mathbf{X}^{obs}, \mathbf{X}^{mis}, \boldsymbol{\theta}_{y|x}, \boldsymbol{\theta}_x)}_{\text{joint distribution}} = \underbrace{p(\mathbf{y} \mid \mathbf{X}^{obs}, \mathbf{X}^{mis}, \boldsymbol{\theta}_{y|x})}_{\text{analysis model}} \underbrace{p(\mathbf{X}^{obs}, \mathbf{X}^{mis} \mid \boldsymbol{\theta}_x)}_{\text{covariate model(s)}} \underbrace{p(\boldsymbol{\theta}_{y|x}, \boldsymbol{\theta}_x)}_{\text{priors}}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_{y|x}, \boldsymbol{\theta}_x)$

Covariate Models

- multiple variables
- mixed type (continuous, categorical)
- multi-level

➡ **no closed-form** distribution

➡ Specification as a **sequence of** (univariate) **conditional distributions**

Covariate Models

Example: $p(\mathbf{X} \mid \boldsymbol{\theta}_x) = p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3(t), \mathbf{x}_4(t) \mid \boldsymbol{\theta}_x)$

- $\mathbf{x}_1$: time-constant, incomplete
- $\mathbf{x}_2$: time-constant, complete
- $\mathbf{x}_3(t)$: time-varying, incomplete
- $\mathbf{x}_4(t)$: time-varying, complete

$$
\begin{aligned}
p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3(t), \mathbf{x}_4(t) \mid \boldsymbol{\theta}_x) = {} & p(\mathbf{x}_3(t) \mid \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_4(t), \boldsymbol{\theta}_x) && \text{e.g., GLMM} \\
& p(\mathbf{x}_4(t) \mid \mathbf{x}_1, \mathbf{x}_2, \boldsymbol{\theta}_x) && \text{e.g., GLMM} \\
& p(\mathbf{x}_1 \mid \mathbf{x}_2, \boldsymbol{\theta}_x) && \text{e.g., GLM} \\
& p(\mathbf{x}_2 \mid \boldsymbol{\theta}_x) && \text{(can be omitted)}
\end{aligned}
$$

covariate models $\neq$ imputation models!

Imputation Models

$$\underbrace{p(\mathbf{x}_1 \mid \text{everything else})}_{\text{imputation model}} \propto \underbrace{p(\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}_{y|x}, \boldsymbol{\theta}_x)}_{\text{joint distribution}} = \underbrace{p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}_{y|x})}_{\text{analysis model}} \underbrace{p(\mathbf{X} \mid \boldsymbol{\theta}_x)}_{\text{covariate model(s)}} \underbrace{p(\boldsymbol{\theta}_{y|x}, \boldsymbol{\theta}_x)}_{\text{priors}}$$

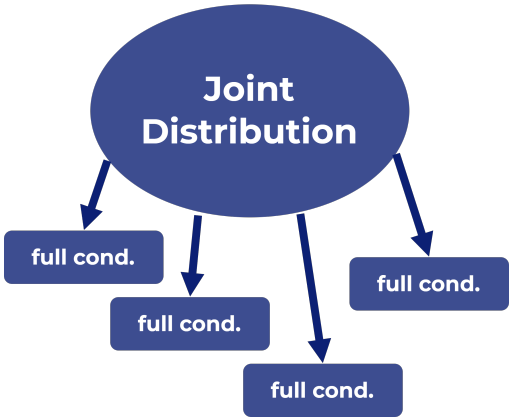
Imputation Models typically do not have a closed form.

➡ Estimation via **MCMC sampling**, with the help of

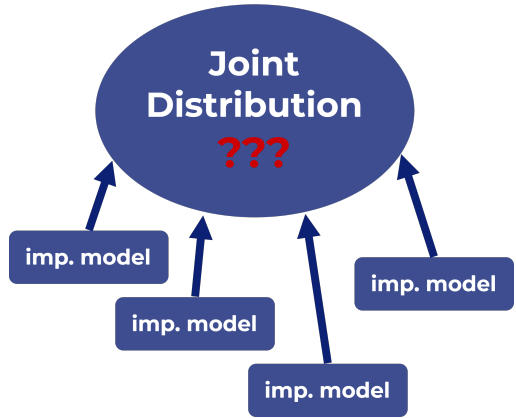
- Gibbs sampler
- Metropolis-Hastings (or the like)

Fully Bayesian vs MICE

Fully Bayesian



MICE / FCS



Advantages of the Bayesian Approach

$$p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) = \underbrace{p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}_{y|x})}_{\text{analysis model}}\, p(\mathbf{x} \mid \boldsymbol{\theta}_x)\, p(\boldsymbol{\theta}_{y|x}, \boldsymbol{\theta}_x) \qquad \text{with } \boldsymbol{\theta} = (\boldsymbol{\theta}_{y|x}, \boldsymbol{\theta}_x)$$

- specification of the joint distribution
  ➡ assures **compatibility** of imputation models
- use of the analysis model in the specification of the joint distribution
  ➡ **parameters of interest** $\boldsymbol{\theta}_{y|x}$ are estimated directly
  ➡ **non-linear associations** are taken into account
  ➡ assures **congeniality**
- response is not in any linear predictor
  ➡ no problem to use **complex outcomes**

In practice: Package JointAI

Specification of the joint distribution:

User

analysis model

JointAI

covariate models

priors

Estimation:

JointAI

pre-processing

rjags/JAGS

MCMC sampling

JointAI

post-processing

JointAI: Model Types

Generalized Linear Models

- `lm_imp()`
- `glm_imp()`
- `clm_imp()`
- `mlogit_imp()`
- `betareg_imp()`
- `lognorm_imp()`

Mixed Models

- `lme_imp()`
- `glme_imp()`
- `clmm_imp()`
- `mlogitmm_imp()`
- `betamm_imp()`
- `lognormm_imp()`

Survival Models

- `coxph_imp()`
- `survreg_imp()`
- `JM_imp()`

Multi-level Models

- **lme4** or **nlme** type specification
- nested and crossed random effects
- 2, 3, 4, … levels of grouping, i.e., `... + (time | id) + (1 | group) + (1 | center ) + (1 | country) + ...`
- multi-level structure also for survival models
- uses hierarchical centering, i.e., $\mathbf{b} \sim N(\mathbf{X}\beta, \mathbf{D})$

Survival with Time-varying Covariates

- proportional hazards model with time-varying covariates (using `+ (1 | id)` and `timevar = "<...>"`)
- baseline hazard: B-spline with `df_basehaz` degrees of freedom
- joint model for longitudinal and survival data (using `formula = list(<...>)` and `timevar = "<...>"`)
- association structure (`assoc_type`)
  - underlying value (`underl.value`)
  - observed/imputed value (`obs.value`)
- multivariate joint model:
  full / block-diagonal / independent random effects

Cumulative Logit Models

- invert the OR (`rev`)

$$\log\left(\frac{P(y_i > k)}{P(y_i \leq k)}\right) = \gamma_k + \eta_i \quad \text{vs} \quad \log\left(\frac{P(y_i \leq k)}{P(y_i > k)}\right) = \gamma_k + \eta_i$$

- partial proportional odds (`nonprop`)

$$\log\left(\frac{P(y_i > k)}{P(y_i \leq k)}\right) = \gamma_k + \eta_i \quad \text{vs} \quad \log\left(\frac{P(y_i \leq k)}{P(y_i > k)}\right) = \gamma_k + \eta_i + \eta_{ki}$$

JointAI: Model Specification

Specification like standard complete data functions:

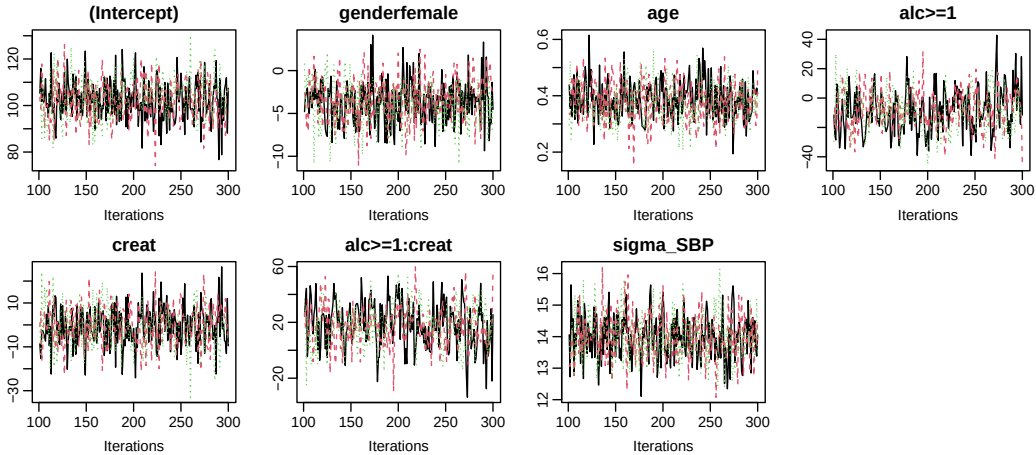
```
library("JointAI")
lm1 <- lm_imp(SBP ~ gender + age + alc * creat, data = NHANES,
              n.iter = 200, seed = 2022)
```

First thing after fitting a model: check convergence!

```
traceplot(lm1)
```

(The NHANES data is a dataset contained in JointAI.)

JointAI: Traceplot



JointAI: Model Summary

```
summary(lm1)
```

```
## Bayesian linear model fitted with JointAI
##
## Call:
## lm_imp(formula = SBP ~ gender + age + alc * creat, data = NHANES,
##     n.iter = 200, seed = 2022)
##
##
## Posterior summary:
##                  Mean      SD    2.5%   97.5% tail-prob. GR-crit MCE/SD
## (Intercept)  102.575  8.7395  85.842 120.251      0.000    1.04 0.0408
## genderfemale  -3.622  2.5565  -8.563   1.488      0.137    1.02 0.0408
## age            0.379  0.0702   0.249   0.519      0.000    1.03 0.0408
## alc>=1        -8.626 13.6774 -33.890  17.652      0.523    1.00 0.0618
## creat         -0.794  9.1653 -19.334  16.431      0.910    1.01 0.0408
## alc>=1:creat  18.209 15.3472 -10.440  47.830      0.247    1.00 0.0607
##
## [...]
```

JointAI: Model Summary

```
## [...]
##
## Posterior summary of residual std. deviation:
##           Mean    SD 2.5% 97.5% GR-crit MCE/SD
## sigma_SBP   14 0.721 12.6  15.4   0.998 0.0408
##
##
## MCMC settings:
## Iterations = 101:300
## Sample size per chain = 200
## Thinning interval = 1
## Number of chains = 3
##
## Number of observations: 186
```

JointAI: Which Models were Fitted?

```
list_models(lm1)
```

```
## Linear model for "SBP"
##    family: gaussian
##    link: identity
## * Predictor variables:
##   (Intercept), genderfemale, age, alc>=1, creat, alc>=1:creat
## * Regression coefficients:
##   beta[1:6] (normal prior(s) with mean 0 and precision 1e-04)
## * Precision of  "SBP" :
##   tau_SBP (Gamma prior with shape parameter 0.01 and rate parameter 0.01)
##
## Binomial model for "alc"
##    family: binomial
##    link: logit
## * Reference category: "<1"
## * Predictor variables:
##   (Intercept), genderfemale, age, creat
## * Regression coefficients:
##   alpha[1:4] (normal prior(s) with mean 0 and precision 1e-04)
##
## [...]
```

JointAI: Which Models were Fitted?

```
## [...]
##
## Linear model for "creat"
##    family: gaussian
##    link: identity
## * Predictor variables:
##   (Intercept), genderfemale, age
## * Regression coefficients:
##   alpha[5:7] (normal prior(s) with mean 0 and precision 1e-04)
## * Precision of  "creat" :
##   tau_creat (Gamma prior with shape parameter 0.01 and rate parameter 0.01)
```

JointAI: Covariate Model Types

```
lm1$models
```

```
##                      SBP                      alc                    creat
## "glm_gaussian_identity"    "glm_binomial_logit"                     "lm"
```

Change the model types:

```
lm2 <- lm_imp(SBP ~ gender + age + alc + creat, data = NHANES,
              n.iter = 200, models = c(creat = "lognorm"), seed = 2022)
lm2$models
```

```
##                      SBP                      alc                    creat
## "glm_gaussian_identity"    "glm_binomial_logit"                "lognorm"
```

JointAI: Covariate Model Types

Generalized linear models:

- `glm_<family>_<link>`
  - `glm_gaussian_identity` (alias: `lm`)
  - `glm_binomial_logit` (alias: `glm_logit`)
  - `glm_poisson_log`
  - …
- `lognorm`
- `beta`
- `clm`
- `mlogit`

Mixed models:

- `glmm_<family>_<link>`
  - `glmm_gaussian_identity` (alias: `lmm`)
  - `glmm_binomial_logit` (alias: `lmm_logit`)
  - `glmm_gamma_inverse`
  - …
- `glmm_lognorm`
- `glmm_beta`
- `clmm`
- `mlogitmm`

JointAI: Model Specification

- auxiliary variables (`auxvars`)
- (ridge) shrinkage (`shrinkage`)
- truncation of continuous distributions (`trunc`)
- setting reference categories (`refcats`)
- change hyper-parameters (`hyperpars`)
- parameters to be monitored (`monitor_params`)
- export/customize the JAGS model (`modelname`, `modeldir`, `keep_model`, `overwrite`)

JointAI: MCMC settings

- adaptive phase (`n.adapt`)
- sampling phase (`n.iter`)
- number of chains (`n.chains`)
- thinning (`thin`)
- initial values (`inits`)
- seed value (`seed`)

- continue sampling (`add_samples()`)
- parallel sampling:

```
library("future")
plan(multisession, workers = 8)
## fit JointAI model ...

## to re-set to sequential evaluation:
plan(sequential)
```

JointAI: Plots

- data distribution (`plot_all()`)
- missing data pattern (`md_pattern()`)

- MCMC chains (`traceplot()`)
- posterior density (`densplot()`)

- Monte Carlo Error (`plot(MC_error(<object>))`)

- imputed vs observed data (`plot_imp_distr()`)

JointAI: Model Infos

- monitor any node
  `monitor_params`
- see the JAGS model
  (`<object>$jagsmodel`)
- print model info
  (`list_models()`,
  `<object>$models`)
- access the JAGS model object
  (`<object>$model`)
- list all monitored parameters
  (`parameters()`)
- list all regression coefficients
  (`<object>$coef_list`)

- Computational infos:
  `<object>$comp_info`
  - start time
  - duration
  - package version
  - R version
  - parallel setting

JointAI: Output Subset & Prediction

- Subset selection
  - iterations (`start`, `end`, `thin`)
  - nodes (`subset`)
  - chains (`exclude_chains`)
  - sub-models (`outcome`)

- create "newdata" for effect plots (`predDF()`)
- prediction (`predict()`)
  - GLM(M) type models: linear predictor or outcome scale
    for now: assumption random effects = 0
  - prop. hazard models: (log) hazard, (-log) survival
  - joint model for longitudinal & survival data: not yet possible

- export imputed values
  (`monitor_params = c(imps = TRUE)`, `get_MIdat()`)

Session 7
Summary & Extensions

Bayesian Methods for Missing Covariates in Longitudinal Studies

IBC2022 in Riga – July 2022

Recap

Frequentist methods for longitudinal data

- (non-likelihood)
- likelihood methods ➡ mixed effects model

Missing data processes

- mechanism of missingness: MCAR / MAR / MNAR
- Multiple Imputation / MICE
  - Potential Issues with MICE with non-linear associations and in multi-level settings

Recap

Bayesian methods for longitudinal studies

- Bayes theorem
- Priors
- Computational aspects
  - MCMC, Gibbs, Metropolis-Hastings
  - Convergence
- Software
- Intro: Bayesian approach to deal with missing data

Recap

Bayesian analysis of incomplete data with JointAI

- Theoretical aspects
  - Specification of the joint distribution as the product of analysis model, covariate models, priors
  - Convenient specification of the (multivariate) distribution of the incomplete covariates via a product of (univariate) conditional distributions
  - Difference Bayesian approach vs MICE
- Implementation in the R package JointAI

Extension to MNAR

So far, we assumed **ignorable missingness**, i.e., M(C)AR and $p(\boldsymbol{\theta}, \psi) = p(\boldsymbol{\theta})p(\psi)$.

In general, the observed data include the missing data indicator **r** and we need a model for **r**, e.g.:

Selection Model

$$p(\mathbf{y}, \mathbf{x}, \mathbf{r}, \boldsymbol{\theta}_{y|x}, \boldsymbol{\theta}_x, \psi) = \underbrace{p(\mathbf{r} \mid \mathbf{y}, \mathbf{x}, \psi)}_{\text{missingness model}} \underbrace{p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}_{y|x})}_{\text{analysis model}} \underbrace{p(\mathbf{x} \mid \boldsymbol{\theta}_x)}_{\text{covariate models}} \underbrace{p(\boldsymbol{\theta}_{y|x})\, p(\boldsymbol{\theta}_x)\, p(\psi)}_{\text{priors}}$$

- $\boldsymbol{\theta}_{y|x}$ applies to complete and incomplete cases
  ➡ complete data inference (contrary to a pattern mixture model)

Extension to MNAR

- $\mathbf{r}$ has the same dimension as $\mathbf{x}^{mis}$ (where $\mathbf{x} = (\mathbf{x}^{obs}, \mathbf{x}^{mis})$)
  ➡ sequential specification

$$p(\mathbf{r} \mid \mathbf{y}, \mathbf{x}, \psi) = p(\mathbf{r}_1 \mid \mathbf{r}_2, \ldots, \mathbf{r}_q, \mathbf{y}, \mathbf{x}, \psi)\; p(\mathbf{r}_2 \mid \mathbf{r}_3, \ldots, \mathbf{r}_q \mathbf{y}, \psi)\; \ldots\; p(\mathbf{r}_q \mid \mathbf{y}, \mathbf{x}, \psi)$$

Extension to MNAR: JointAI

Example:

Create the missing data indicator variables:

```
HCVbase$DM_NA <- is.na(HCVbase$DM)
HCVbase$Creatinin_NA <- is.na(HCVbase$Creatinin)
```

Specify the selection model:

```
lm_MNAR <- lm_imp(list(Albumin ~ Age0 + Sex + DM + Creatinin,
                       DM_NA ~ Creatinin_NA + Albumin + Age0 + Sex + DM + Creatinin,
                       Creatinin_NA ~ Albumin + Age0 + Sex + DM + Creatinin),
                  models = c(Creatinin = "lognorm"),
                  n.iter = 500, data = HCVbase, seed = 2022, n.chains = 8)
```

Extension to MNAR

Huang et al. (2005)[1] propose

- the use of probit models instead of logit models for $\mathbf{r}$ and
- choosing hyper-parameters for $p(\psi)$ using a procedure based on (standard) multiple imputation

to improve convergence and mixing of the Gibbs sampler.

[1] Lan Huang, Ming-Hui Chen and Joseph G. Ibrahim. *Bayesian Analysis for Generalized Linear Models with Nonignorably Missing Covariates*. Biometrics (2005), 61, 767 – 780. DOI: j.1541-0420.2005.00338.x

Multivariate Mixed Models in JointAI

```
fmla <- list(logFIB4 ~ Age0 + Sex + Creatinin + ns(time, df = 3) +
               (ns(time, df = 3) | id) + (1 | center),
             Albumin ~ Age0 + Sex + DM + ns(time, df = 2) + (ns(time, df = 2) | id),
             Creatinin ~ Age0 + Sex + ns(time, df = 3) + (ns(time, df = 3)) | id)
```

```
mvlme <- lmer_imp(fmla, data = datHCV, n.iter = 200, rd_vcov = "full",
                  models = c(Creatinin = "glmm_lognorm"))
```

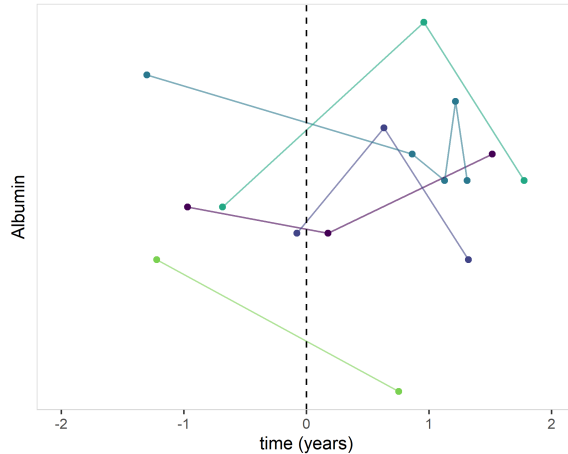
Can handle

- mixed type outcomes,
- outcome specific random effect structures & levels, and
- (outcome specific) random effects variance-covariance matrix structures (`rd_vcov`).

Custom Models

Example

- **Aim:** Fit a PH model with baseline albumin as covariate.
- **Problem:** Some patients do not have albumin observed at baseline.
- **Idea:** Impute from a mixed model on repeated albumin within a relevant time window, but **jointly** with the PH model!



Custom Models

Data setup:
Add baseline albumin (`Albu0`) as covariate to the data:

```
HCVbase$Albu0 <- HCVbase$Albumin
datHCV <- merge(datHCV, subset(HCVbase, select = c("id", "Albu0")))
```

Setup run (without creating MCMC samples):

```
fmla <- list(
  Surv(etime, event) ~ Age0 + Sex + DM + Albu0,
  Albumin ~ Age0 + Sex + ns(time, df = 2) + (ns(time, df = 2) | id)
)

setup <- coxph_imp(fmla, data = datHCV,  no_model = "time",
                   n.adapt = 0, n.iter = 0, scale_vars = FALSE)
```

Custom Models

Relevant part of the default JAGS model (`setup$jagsmodel`):

```
## [...]
##   # Normal mixed effects model for Albumin ---------------------------------------
##   for (i in 1:985) {
##     M_lvlone[i, 1] ~ dnorm(mu_Albumin[i], tau_Albumin)
##     mu_Albumin[i] <- b_Albumin_id[group_id[i], 1] +
##                      b_Albumin_id[group_id[i], 2] * M_lvlone[i, 2] +
##                      b_Albumin_id[group_id[i], 3] * M_lvlone[i, 3]
##   }
##
## [...]
##
##   # Normal model for Albu0 -------------------------------------------------------
##   for (ii in 1:250) {
##     M_id[ii, 3] ~ dnorm(mu_Albu0[ii], tau_Albu0)
##     mu_Albu0[ii] <- M_id[ii, 5] * alpha[1] + M_id[ii, 6] * alpha[2] +
##                     M_id[ii, 7] * alpha[3] + M_id[ii, 8] * alpha[4]
##   }
## [...]
```

Custom Models

Specify the custom model to overwrite the default model for `Albu0`:

```
custom_model <- "for (ii in imp_Albu0) {
    M_id[ii, 3] <- M_lvlone[rows_Albumin[ii], 1]
}"
```

Elements that need to be added to the `data_list`:

```
# elements of Albu0 that need to be imputed
imp_Albu0 <- which(is.na(setup$data_list$M_id[, "Albu0"]))

# position of the corresponding values in the long-format matrix
rows_Albumin <- which(setup$data_list$M_lvlone[, "time"] == 0)
```

Custom Models

```
cox_fit <- update(setup,
                  custom = list(Albu0 = custom_model),
                  append_data_list = list(imp_Albu0 = imp_Albu0,
                                          rows_Albumin = rows_Albumin),
                  n.iter = 500, n.chains = 8, seed = 2022)
```

The model for `Albu0` is now our custom model (`cox_fit$jagsmodel`):

```
## [...]
##
## for (ii in imp_Albu0) {
##     M_id[ii, 3] <- M_lvlone[rows_Albumin[ii], 1]
## }
##
## [...]
```

Thank you for your attention!

Nicole Erler
Department of Biostatistics
Erasmus Medical Center
n.erler@erasmusmc.nl
N_Erler NErler nerler.com

Emmanuel Lesaffre
I-Biostat
KU Leuven
emmanuel.lesaffre@kuleuven.be