

# Chapter 5

## Mixed Models for Discrete Data

## 5.1 Generalized Linear Mixed Models

---

- The previous chapter focused on the framework of Generalized Estimating Equations
  - ▷ this can be seen as the extension of the marginal models for continuous data of Chapter 2 to the setting of categorical longitudinal responses
- In this chapter we will see the analogue of linear mixed models for categorical data



**Generalized Linear Mixed Models (GLMMs)**

## 5.1 Generalized Linear Mixed Models (cont'd)

---

- The intuitive idea behind GLMMs is the same as in linear mixed models, i.e.,
  - ▷ the correlation between the repeated categorical measurements is induced by unobserved random effects
  - ▷ in other words: the categorical longitudinal measurements of a subject are correlated because all of them share the *same* unobserved random effect (**conditional independence assumption**)

## 5.1 Generalized Linear Mixed Models (cont'd)

---

Graphical representation of the conditional independence assumption



## 5.1 Generalized Linear Mixed Models (cont'd)

---

- Similarly to Chapter 4, we will focus on clustered dichotomous/binary data
  - ▷ nonetheless, the same ideas and issues also apply to other categorical responses (e.g., Poisson, ordinal data, multinomial data, etc.)
- Suppose we have a binary outcome  $y_{ij}$

$$y_{ij} = \begin{cases} 1, & \text{if subject } i \text{ has a positive response at measurement } j \\ 0, & \text{if subject } i \text{ has a negative response at measurement } j \end{cases}$$

## 5.1 Generalized Linear Mixed Models (cont'd)

---

- The generic mixed model for  $y_{ij}$  is a *Mixed-Effects Logistic Regression* and has the form:

$$\begin{cases} \log \frac{\pi_{ij}}{1 - \pi_{ij}} = x_{ij}^{\top} \beta + z_{ij}^{\top} b_i \\ b_i \sim \mathcal{N}(0, D) \end{cases}$$

where

- ▷  $\pi_{ij} = \Pr(y_{ij} = 1)$  the probability of a positive response
- ▷  $x_{ij}$  a vector of fixed-effects covariates, with corresponding regression coefficients  $\beta$
- ▷  $z_{ij}$  a vector of random-effects covariates, with corresponding regression coefficients  $b_i$

## 5.1 Generalized Linear Mixed Models (cont'd)

---

- More formally, we have the following three-part specification
  1. Conditional on the random effects  $b_i$ , the responses  $y_{ij}$  are independent and have a Bernoulli distribution with mean  $E(y_{ij} | b_i) = \pi_{ij}$  and variance  $\text{var}(y_{ij} | b_i) = \pi_{ij}(1 - \pi_{ij})$
  2. The conditional mean of  $y_{ij}$  depends upon fixed and random effects via the following expression:

$$\log \frac{\pi_{ij}}{1 - \pi_{ij}} = x_{ij}^\top \beta + z_{ij}^\top b_i$$

3. The random effects follow a multivariate normal distribution with mean zero and variance-covariance matrix  $D$

## 5.1 Generalized Linear Mixed Models (cont'd)

---

- Notes: On the definition of GLMMs
  - ▷ The three-part specification of GLMMs corresponds to a full specification of the distribution of the outcome  $y_{ij}$  – this is in contrast to the GEE approach, which is a semi-parametric method
  - ▷ The mean and correlation structures are simultaneously defined using random effects
    - ⇒ As we will see next, this has direct and important implications with respect to the interpretation of the parameters



## 5.2 Interpretation

---

- Example: In the AIDS dataset, a very low CD4 count (less than  $150 \text{ cells/mm}^3$ ) is an indicator for opportunistic infections
  - ▷ In the following analysis we dichotomize the CD4 cell counts from the AIDS dataset using this threshold
  - ▷ We fit a mixed effects logistic regression with
    - \* *fixed effects*: time, treatment and their interaction
    - \* *random effects*: random intercepts

## 5.2 Interpretation (cont'd)

- The model has the form:

$$\log \frac{\pi_{ij}}{1 - \pi_{ij}} = \beta_0 + \beta_1 \text{Time}_{ij} + \beta_2 \text{ddI}_i + \beta_3 \{\text{Time}_{ij} \times \text{ddI}_i\} + b_i, \quad b_i \sim \mathcal{N}(0, \sigma_b^2)$$

	Value	Std.Err.	z-value	p-value
$\beta_0$	6.250	0.899	6.954	< 0.001
$\beta_1$	0.149	0.044	3.392	0.001
$\beta_2$	-0.811	0.731	-1.109	0.267
$\beta_3$	-0.029	0.059	-0.494	0.622
$\sigma_b$	6.019			

## 5.2 Interpretation (cont'd)

---

- Interpretation of fixed effects

- ▷ At baseline for group ddC the log odds of a low CD4 cell count are on average  $\beta_0 = 6.25$

- \* 95% heterogeneity interval (**not** confidence interval):  
 $(\beta_0 - 1.96\sigma_b ; \beta_0 + 1.96\sigma_b) = (-5.55 ; 18.05)$

- ▷ We translate the log odds to the probability scale: The probability of low CD4 cell count is  $\exp(\beta_0)/\{1 + \exp(\beta_0)\} = 0.99807$

- \* 95% heterogeneity interval:  
 $(1/[1 + \exp\{-(\beta_0 - 1.96\sigma_b)\}] ; 1/[1 + \exp\{-(\beta_0 + 1.96\sigma_b)\}]) = (0.00389 ; 1)$

## 5.2 Interpretation (cont'd)

---

- When we compare the middle point of the transformed heterogeneity interval with the transformed intercept an **important** observation is made:

▷  $\exp(\beta_0) / \{1 + \exp(\beta_0)\} = 0.99807$

▷ mean of transformed interval = 0.50194

**When we transform the fixed effects to the probability scale, they do not correspond to the average probability**

## 5.2 Interpretation (cont'd)

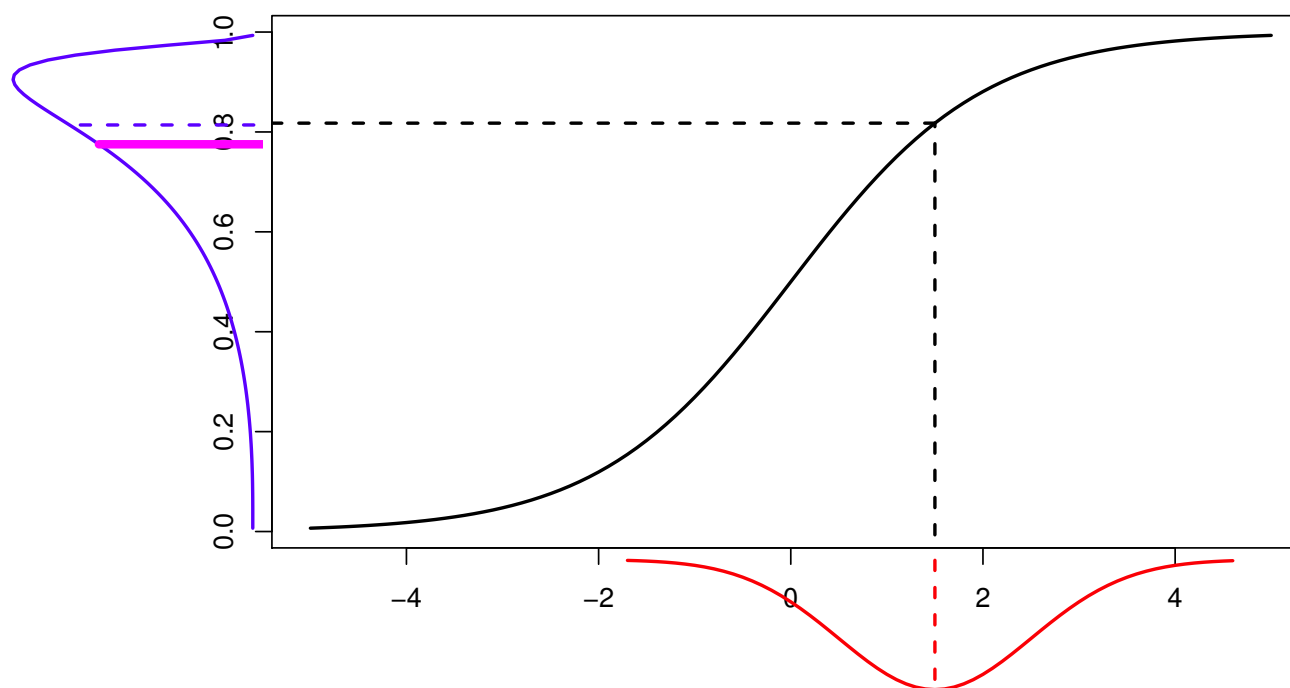
---

- Let's explain this issue graphically ...

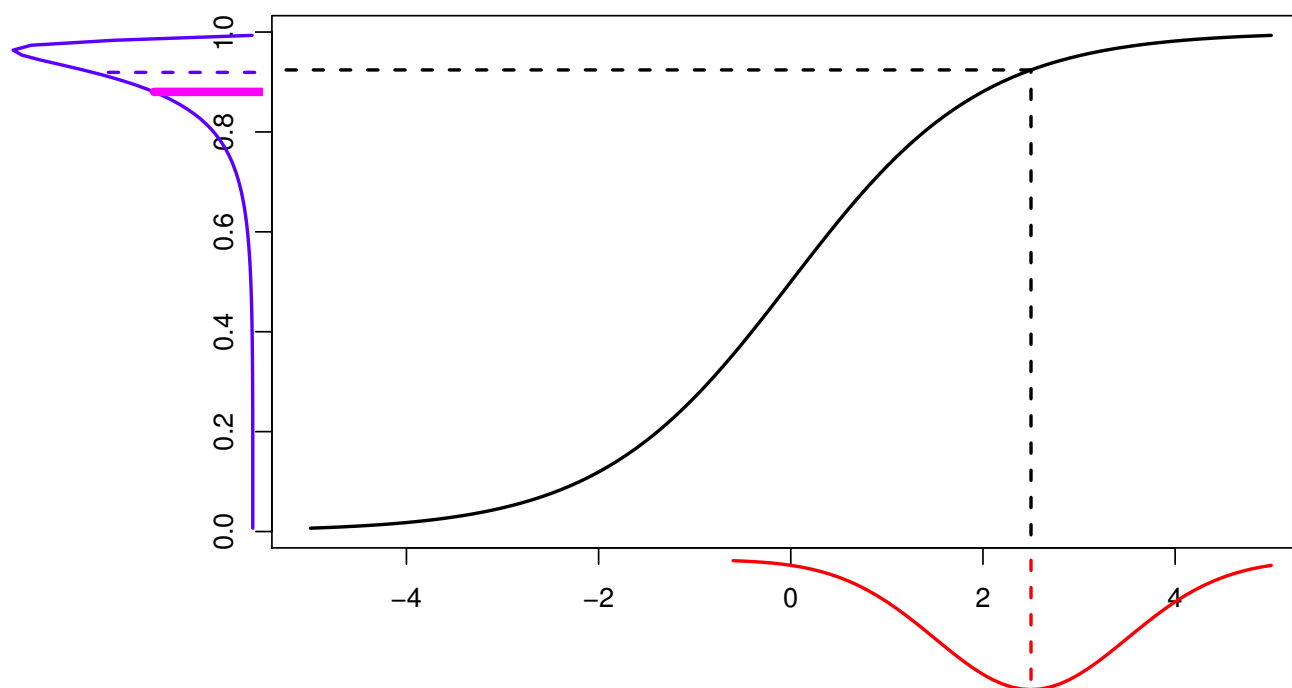
## 5.2 Interpretation (cont'd)



## 5.2 Interpretation (cont'd)



## 5.2 Interpretation (cont'd)



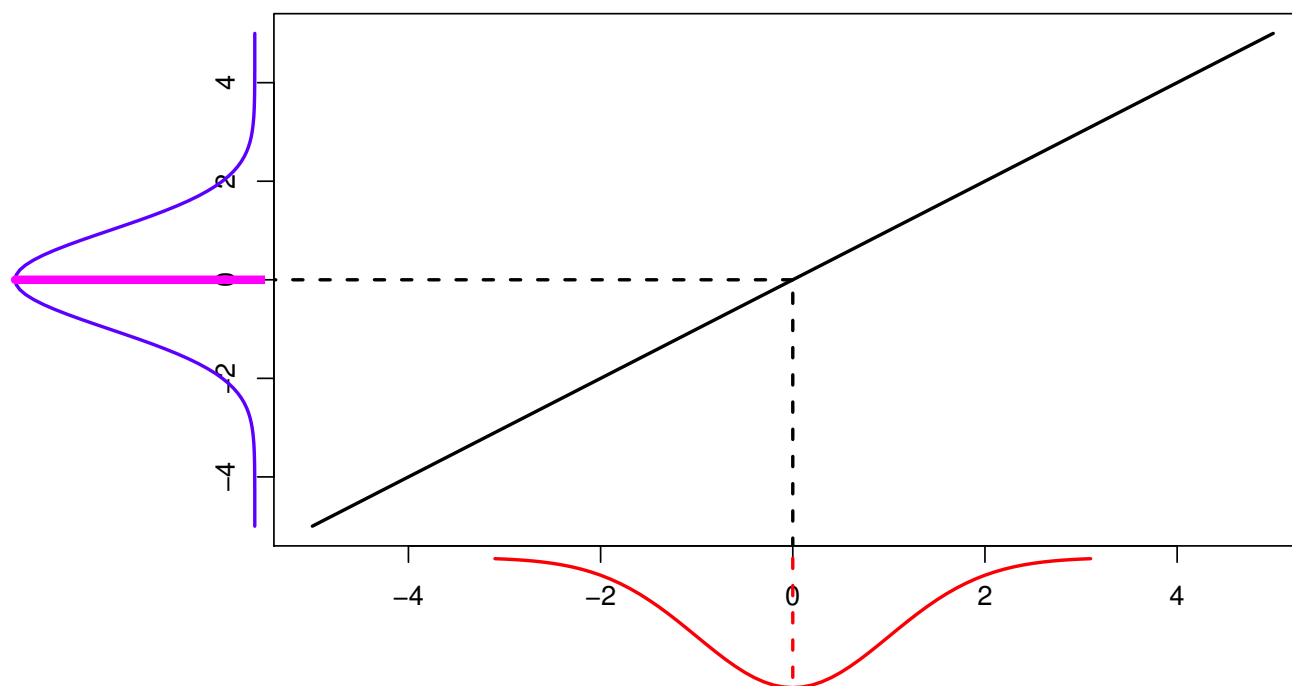


## 5.2 Interpretation (cont'd)

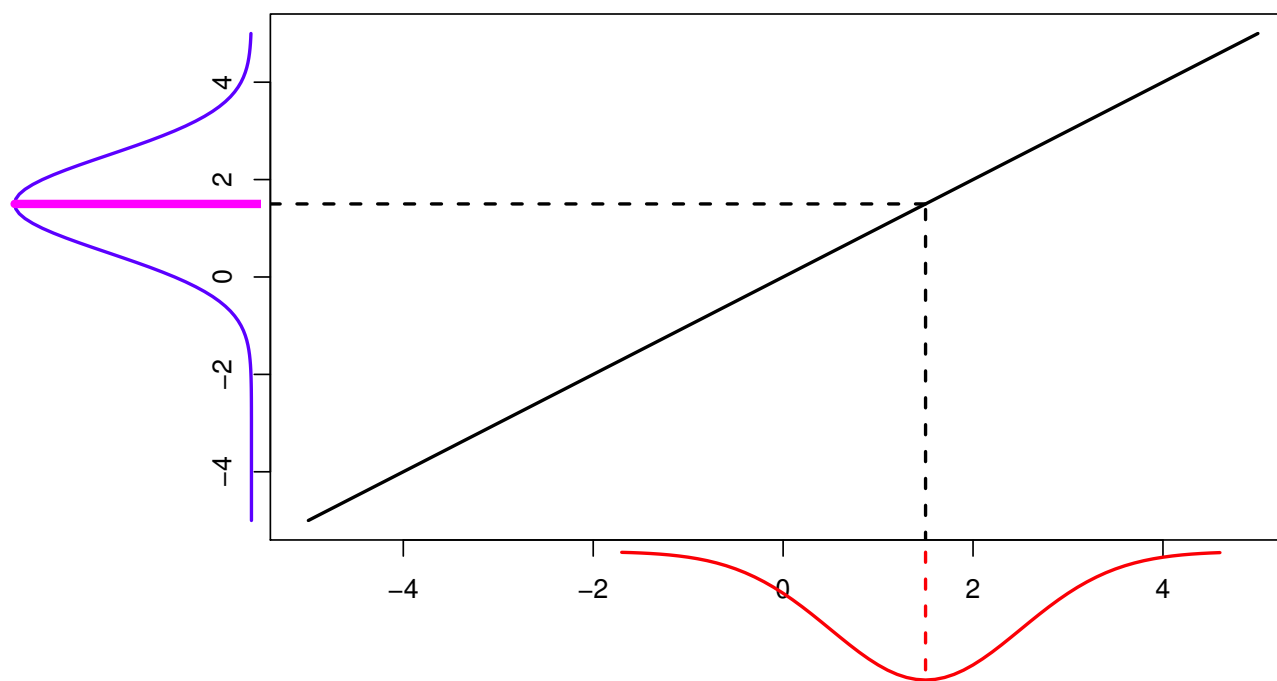
---

- We did not have this problem in the case of the linear mixed model because we did not have a link function
  - ▷ or to put it more precisely, the link function was the identity  $g(x) = x$
- Let's see graphically again why for linear mixed models we do not have the same problem . . .

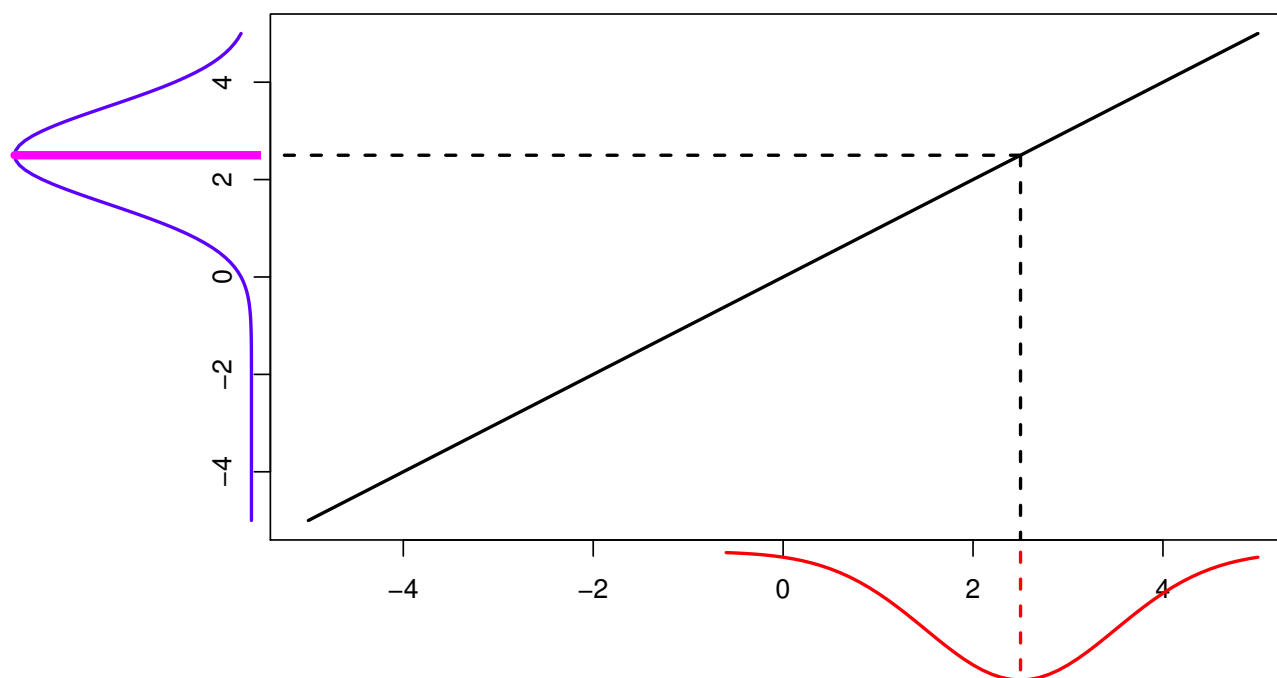
## 5.2 Interpretation (cont'd)



## 5.2 Interpretation (cont'd)



## 5.2 Interpretation (cont'd)



## 5.2 Interpretation (cont'd)

---

- The same complications also hold for the other fixed-effects coefficients of the logistic regression model
  - ▷ e.g.,  $\beta_1$  does **not** have the interpretation of the *average* odds ratio for a month increase in follow-up
- Let's see why
  - ▷ say that we compare two patients at different follow-up times who both took ddC, Patient *i* at month  $m$  and Patient *i'* at month  $m + 1$
  - ▷ the equation of the model for Patient *i* is:

$$\log \frac{\pi_{ij}}{1 - \pi_{ij}} = \beta_0 + \beta_1 \{\text{Time}_{ij} = m\} + b_i$$

## 5.2 Interpretation (cont'd)

---

▷ the equation of the model for Patient  $i'$  is:

$$\log \frac{\pi_{i'j}}{1 - \pi_{i'j}} = \beta_0 + \beta_1 \{\text{Time}_{ij} = m + 1\} + b_{i'}$$

▷ hence, the corresponding odds ratio is:

$$\text{log odds ratio: } \log \frac{\pi_{i'j}}{1 - \pi_{i'j}} - \log \frac{\pi_{ij}}{1 - \pi_{ij}} = \beta_1 + (b_{i'} - b_i) \Rightarrow$$

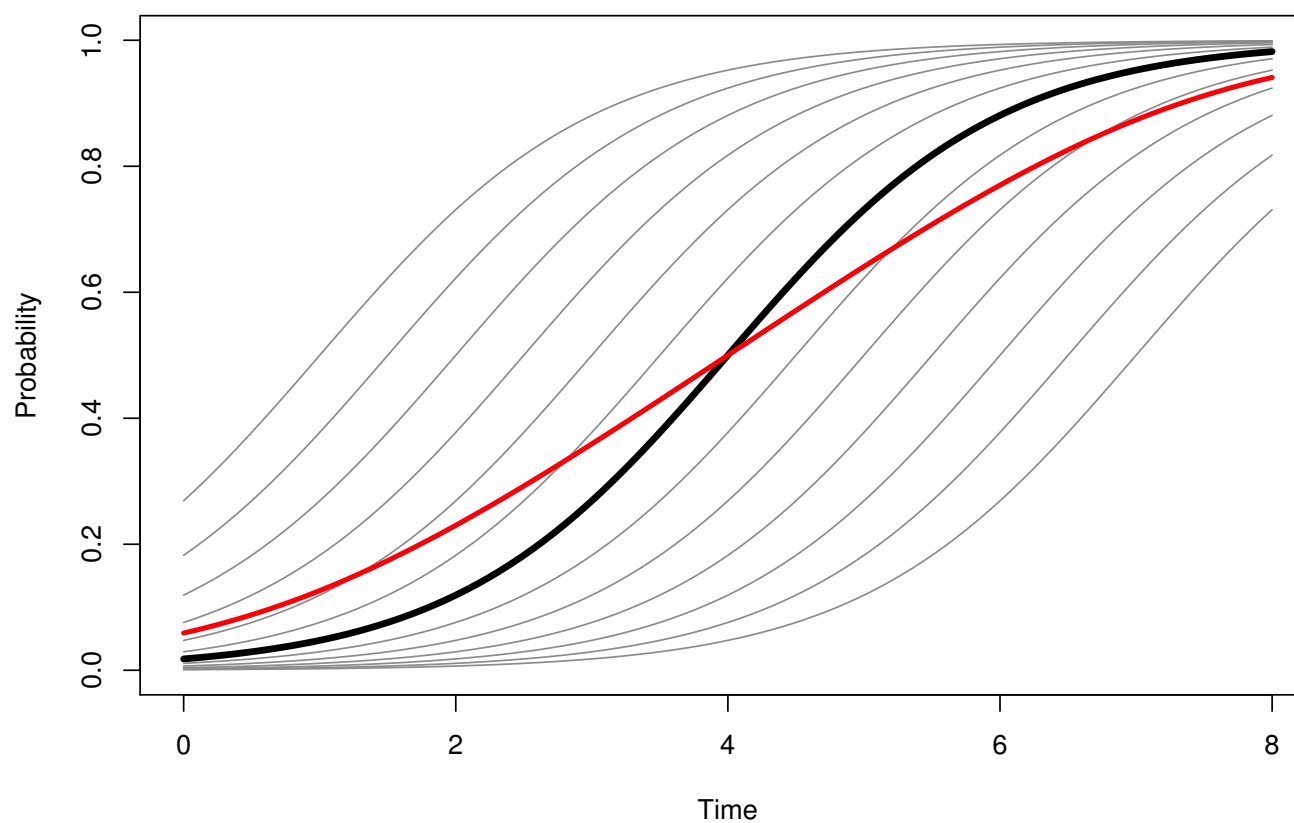
$$\text{odds ratio: } \frac{\pi_{i'j}/(1 - \pi_{i'j})}{\pi_{ij}/(1 - \pi_{ij})} = \exp\{\beta_1 + (b_{i'} - b_i)\} \neq \exp(\beta_1)$$

## 5.2 Interpretation (cont'd)

---

- Hence, the interpretation of  $\beta_1$  is not the log odds for unit increase of Time for all subjects, but rather for subjects with *the same random-effect value*
- To illustrate this again graphically, we depict the relationship between time and the probability of low CD4 cell counts
  - ▷ the grey lines depict 13 random subjects with increasing random effects
  - ▷ the black line corresponds to the subject with  $b_i = 0$  (i.e., the mean individual)
    - ⇒ This line is actually  $1/[1 + \exp\{-(\beta_0 + \beta_1 \text{Time}_{ij})\}]$
  - ▷ the red line that crosses the 13 lines denotes the average longitudinal evolution of the probability of low CD4 cells counts across subjects

## 5.2 Interpretation (cont'd)





## 5.2 Interpretation (cont'd)

---

- To summarize:
  - ▷ The fixed-effects regression coefficients are interpreted in terms of the effects of covariates on changes in an *individual's* transformed mean response, while holding the remaining covariates fixed
  - ▷ Because the components of the fixed effects  $\beta$ , have interpretations that depend upon holding  $b_i$  (the  $i$ -th subject's random effects) fixed, they are often referred to as *subject-specific* regression coefficients
  - ▷ As a result, GLMMs are most useful when the main scientific objective is to make inferences about individuals rather than population averages
  - ▷ Population averages are the targets of inference in marginal models (i.e., GEE)

## 5.2 Interpretation (cont'd)

---

Hence, contrary to the marginal and mixed effects model for continuous data (Chapters 2 & 3), the regression coefficients from marginal models for discrete data **do not** have the same interpretation as the corresponding coefficients from mixed effects models

## 5.2 Interpretation (cont'd)

---

- **Nonetheless**, for the special case of random intercepts, there is a closed-form expression to obtain the marginal regression coefficients from the subject-specific ones, i.e.,

$$\beta^M = \frac{\beta^{SS}}{\sqrt{1 + 0.346\sigma_b^2}}$$

where

- ▷  $\beta^M$  denotes the marginal coefficients
- ▷  $\beta^{SS}$  denotes the subject-specific coefficients
- ▷  $\sigma_b^2$  denotes the variance of the random intercepts

## 5.2 Interpretation (cont'd)

- **Example:** We continue on the previous example from the AIDS dataset (see pp.309) and we compute the corresponding marginal regression coefficients

	Subject-specific				Marginal	
	Value	Std.Err.	z-value	p-value	Value	Std.Err.
$\beta_0$	6.250	0.899	6.954	0.000	1.699	0.244
$\beta_1$	0.149	0.044	3.392	0.001	0.040	0.012
$\beta_2$	-0.811	0.731	-1.109	0.267	-0.220	0.199
$\beta_3$	-0.029	0.059	-0.494	0.622	-0.008	0.016
$\sigma_b$	6.019					

## 5.2 Interpretation (cont'd)

---

- We observe considerable differences between the two sets of parameters
  - ▷ the subject-specific odds ratio for a unit increase in time for a specific ddC patients is 0.54 (95% CI: 0.52; 0.56),
  - ▷ whereas the corresponding marginal odds ratio averaged over all ddC patients equals 0.51 (95% CI: 0.5; 0.52)
  - ▷ note that the lower limit of the 95% CI for the subject-specific odds ratio equals the upper limit of the 95% CI for the marginal odds ratio  
⇒ *the confidence intervals do not overlap*

## 5.2 Interpretation (cont'd)

---

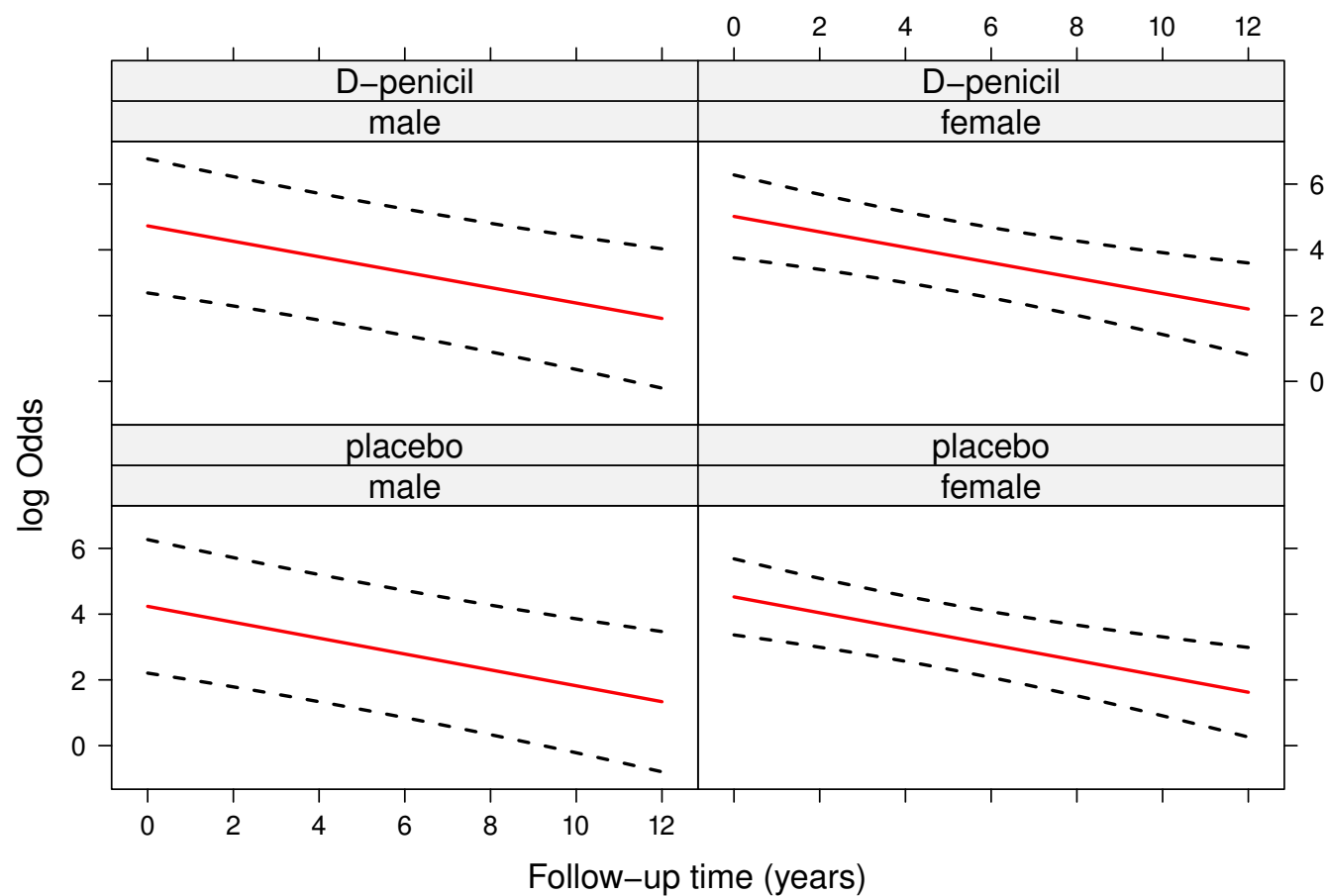
- As we have previously seen, effect plots can be used to effectively communicate complex models
  - ▷ especially in GLMMs, these plots also can be used to depict the marginal average evolutions (i.e., even if the fixed effects coefficients have a subject-specific interpretation, we can still calculate the marginal means)
- **Example:** In the PBC dataset we are interested in the probability of having excess serum cholesterol levels
  - ▷ we include the main effects of time, drug, age & sex
  - ▷ the interaction effect between time and drug, and the interaction effect between age and sex

## 5.2 Interpretation (cont'd)

---

- In the following figure we depict the marginal odds ratio as a function for time, separately for each combination of randomized treatment and sex

## 5.2 Interpretation (cont'd)





## 5.2 Interpretation (cont'd)

---

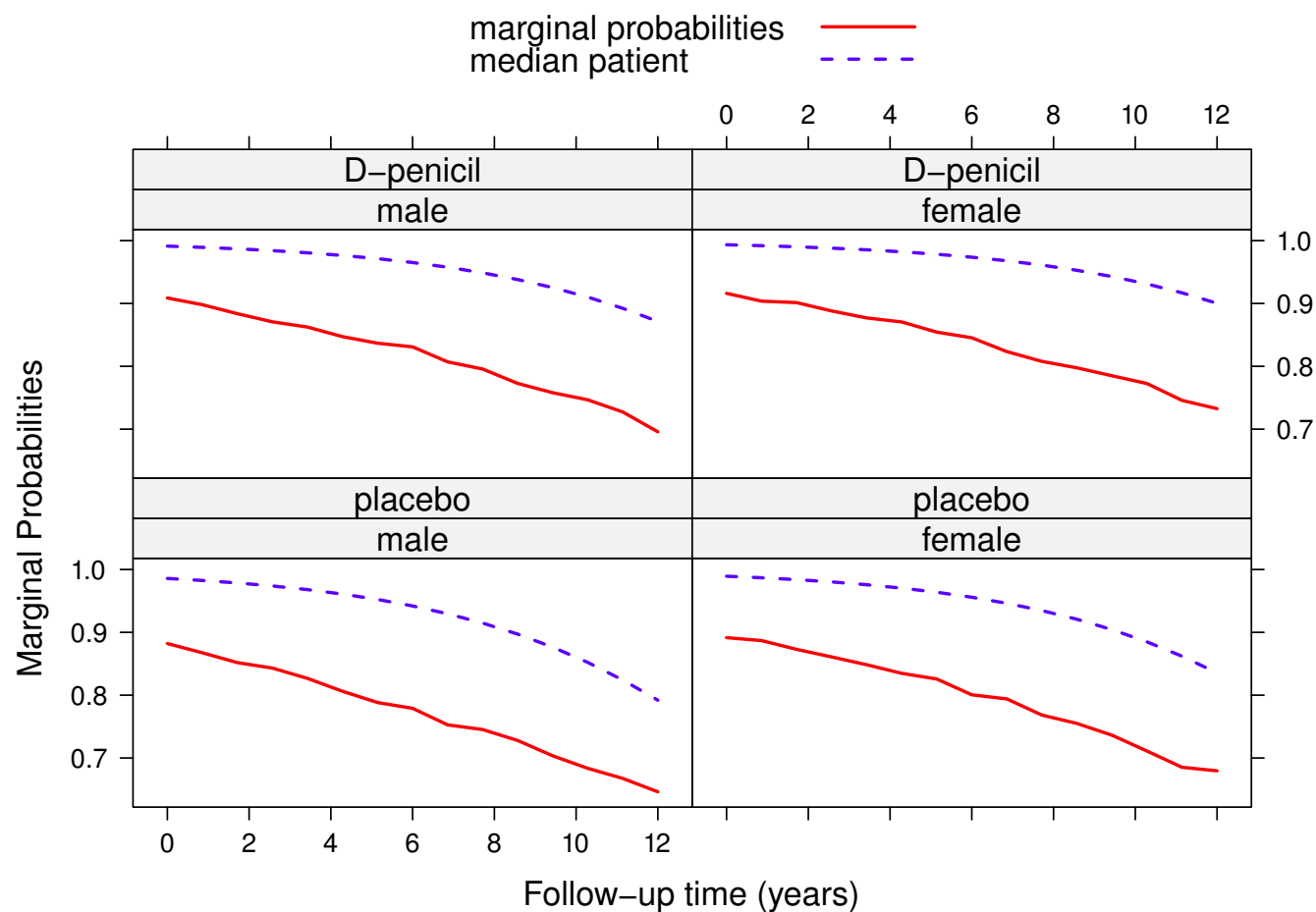
- In the following figure we depict
  - ▷ the marginal probabilities, and
  - ▷ the probabilities of the median patientas a function for time, separately for each combination of randomized treatment and sex

## 5.2 Interpretation (cont'd)

---

- The marginal probabilities are obtained using a Monte Carlo sampling procedure
  - ▷ for each combination of follow-up time, randomized treatment and sex we generate 3000 patients with random effect values coming from the normal distribution  $\mathcal{N}(0, \hat{\sigma}_b^2)$ , where  $\hat{\sigma}_b^2$  denotes the estimated variance of the random effects from the model
  - ▷ for each of these 3000 patients we calculate their probability of having an abnormal serum cholesterol value
  - ▷ we take as an estimate the mean of the 3000 probabilities

## 5.2 Interpretation (cont'd)



## 5.2 Interpretation (cont'd)

---

- Calculation of 95% confidence intervals for the estimated marginal probabilities is not a straightforward task

## 5.3 Estimation

---

- The estimation of GLMMs is based on the same principles as in marginal and mixed models for continuous data
  - ▷ i.e., we have a full specification of the distribution of the data (contrary to GEE), and hence we can use *maximum likelihood*
- Nevertheless, there is an important complication in GLMMs

**The fitting of GLMMs is a computationally challenging task!**

## 5.3 Estimation (cont'd)

---

- Even though the nature of this problem is of rather computational/technical nature, we will need to discuss it in more detail . . .
- What is the problem?
  - ▷ The log-likelihood expression for GLMMs has the same form as in linear mixed models (see pp.153)

$$\ell(\theta) = \sum_{i=1}^n \log \int p(y_i | b_i; \theta) p(b_i; \theta) db_i$$

where  $\theta$  are the parameters of the model

## 5.3 Estimation (cont'd)

---

- In linear mixed effects models both terms in the integrand

- ▷  $p(y_i \mid b_i; \theta)$

- ▷  $p(b_i; \theta)$

are densities of (multivariate) normal distributions, and also because  $y_i$  and  $b_i$  are linearly related

**In linear mixed effects models the integral** in the log-likelihood expression **has a closed-form solution** (i.e., we can compute it on paper)

## 5.3 Estimation (cont'd)

---

- In GLMMs the two terms of the integrand denote densities of different distributions – e.g., in mixed effects logistic regression
  - ▷  $p(y_i | b_i; \theta) \Rightarrow$  Bernoulli distribution
  - ▷  $p(b_i; \theta) \Rightarrow$  multivariate normal distribution

The implication is that

**In GLMMs the same integral does not have a closed-form solution**



## 5.3 Estimation (cont'd)

---

- To overcome this problem two general types of solutions have been proposed in the literature
  - ▷ *Approximation of the integrand*: this entails approximating the product inside the integral (i.e.,  $\{p(y_i | b_i; \theta)p(y_i | b_i; \theta)\}$ ) by a multivariate normal distribution for which the integral has a closed-form solution
    - \* Penalized Quasi Likelihood (PQL)
    - \* Laplace approximation
  - ▷ *Approximation of the integral*: this entails approximating the whole integral (i.e.,  $\int p(y_i | b_i; \theta)p(y_i | b_i; \theta)db_i$ ) by a sum
    - \* Gaussian Quadrature & adaptive Gaussian Quadrature
    - \* Monte Carlo & MCMC (Bayesian approach)

## 5.3 Estimation (cont'd)

---

**From the two alternatives, methods that rely on approximation of the integral have been shown to be superior**

- Though they are (much) more computationally demanding – they have a parameter that controls the accuracy of the approximation:
  - ▷ in Gaussian quadrature rules it is the number of quadrature points (*adaptive Gaussian quadrature with 1 point is equivalent to the Laplace approximation*)
  - ▷ in Monte Carlo/MCMC approaches it is the number of samples

## 5.3 Estimation (cont'd)

---

- **Example:** We continue on the AIDS example, but we now treat the time variable as a factor (i.e., categorical) – the model has the form:

$$\log \frac{\pi_{ij}}{1 - \pi_{ij}} = \beta_0 + \beta_1 \{\text{Time}_{ij} = 2\} + \beta_2 \{\text{Time}_{ij} = 6\} + \beta_3 \{\text{Time}_{ij} = 12\} + \beta_4 \{\text{Time}_{ij} = 18\} + b_i$$

where

- ▷  $\pi_{ij} = \Pr(\text{CD4}_{ij} < 150)$
- ▷  $\{\text{Time}_{ij} = 2\}$  denotes the dummy variable for month 2,  $\{\text{Time}_{ij} = 6\}$  the dummy variable for month 6, and so on

## 5.3 Estimation (cont'd)

---

- We have fitted this model using
  - ▷ PQL
  - ▷ Laplace approximation (adaptive Gaussian quadrature with 1 point)
  - ▷ adaptive Gaussian quadrature with 7, 11, 15 and 21 points
- The following figure depicts the estimated fixed effect coefficients under each approximation with corresponding 95% CIs

## 5.3 Estimation (cont'd)



## 5.3 Estimation (cont'd)

---

- We observe considerable differences between
  - ▷ PQL & Laplace (approximation of the integrand), and
  - ▷ adaptive Gaussian quadrature (approximation of the integral)
- In general, PQL and Laplace will work better as the data get more 'continuous', i.e.,
  - ▷ in Bernoulli data as the number of repeated measurements increases *considerably*
  - ▷ in Binomial data as the number of trials increases
  - ▷ in Poisson data as the rate increases

## 5.3 Estimation (cont'd)

---

- Estimation of the random effects proceeds in a similar manner as in linear mixed models (see pp.165–172)
  - ▷ based on a fitted mixed model, estimates for the random effects are based on the posterior distribution:

$$p(b_i | y_i; \theta) = \frac{p(y_i | b_i; \theta) p(b_i; \theta)}{p(y_i; \theta)}$$

$$\propto p(y_i | b_i; \theta) p(b_i; \theta),$$

in which  $\theta$  is replaced by its MLE  $\hat{\theta}$

## 5.3 Estimation (cont'd)

---

- This is a whole distribution
  - ▷ to obtain estimates for the random effects we typically use measures of location from this posterior distribution (e.g., mean or mode)
  - ▷ as an estimate of the dispersion of the random effect we use the variance of the local curvature around the mode of the posterior distribution
- Contrary to linear mixed models in which this distribution has a closed-form, in GLMMs for categorical responses this is not the case
  - ▷ calculation of the above mentioned measures of location and dispersion is achieved using numerical algorithms



## 5.4 GLMMs in R

---

- R>** In R there are two main packages to fit GLMMs, namely **lme4** and **MCMCglmm** – in this course we will primarily use **lme4**
- The function that fits GLMMs in **lme4** is `glmer()` – this has similar syntax as the `lmer()` function that fits linear mixed models, namely
    - ▷ `formula`: a formula specifying the response vector, the fixed- and random-effects structure
    - ▷ `data`: a data frame containing all the variables
    - ▷ `family`: a description of the error distribution and link function to be used in the model
    - ▷ `nAGQ`: the number of quadrature points

## 5.4 GLMMs in R (cont'd)

---

**R>** The following code fits a mixed effects logistic regression for abnormal serum cholesterol from the PBC dataset with random intercepts and 15 quadrature points for the adaptive Gauss-Hermite rule

```
glmmFit <- glmer(serCholD ~ year * drug + (1 | id),  
                 family = binomial(), data = pbc2, nAGQ = 15)  
  
summary(glmmFit)
```

## 5.4 GLMMs in R (cont'd)

---

**R>** With **MCMCglmm** the same model can be fitted with the code

```
prior <- list(R = list(V = 1, fix = 1),  
             G = list(G1 = list(V = 1e-03, nu = -2)))  
  
glmmFit_mcmc <- MCMCglmm(serCholD ~ year * drug , random = ~ id,  
                        data = pbc2, family = "categorical",  
                        prior = prior, nitt = 200000, thin = 20,  
                        burnin = 5000)  
  
summary(glmmFit_mcmc)
```

## 5.4 GLMMs in R (cont'd)

---

- R>** In the first part of the code we define the `prior` for the variance of the random effects – these options correspond to a non-informative prior that would be equivalent to standard maximum likelihood
- R>** Next in `MCMCglmm()` we have the arguments
- ▷ `fixed`: a formula specifying the response vector and the fixed-effects structure
  - ▷ `random`: a formula specifying the random-effects structure
  - ▷ `data`: a data frame containing all the variables
  - ▷ `family`: a character vector specifying the family
  - ▷ `prior`: the list of prior specifications
  - ▷ `nitt`, `thin`, `burnin` the total number of iterations, the amount of thinning and the number of burn-in iterations

## 5.5 Model Building

---

- Model building for GLMMs proceeds in the same manner as for linear mixed models, i.e.,
  - ▷ we start with an elaborate specification of the fixed-effects structure that contains all the variables we wish to study, and potential nonlinear and interactions terms
  - ▷ following we build-up the random-effects structure, starting from random intercepts, next including also random slopes, quadratic slopes, etc.
    - \* in each step we perform likelihood ratio tests to see whether including the additional random effect improves the fit of the model
  - ▷ having chosen the random-effects structure, we return to the fixed effects and check whether the specification can be simplified
    - \* again we first start by testing the complex terms (i.e., interactions and nonlinear terms), and then we continue to drop explanatory variables, if required

## 5.5 Model Building (cont'd)

---

- **Nevertheless**, quite often, and especially for dichotomous data, extending the random-effects structure may lead to numerical/computational problems
  - ▷ this is because dichotomous data contain the least amount of information
- Hence, for dichotomous data and when we have few to moderate number of repeated measurements per subject, we often can only fit random intercepts models

## 5.6 Hypothesis Testing

---

- Having fitted a GLMM with maximum likelihood, testing of either the fixed- or random-effects structure proceeds in a similar manner as in linear mixed models
- **Important difference:** in GLMMs we do not have REML we always work with full maximum likelihood
  - ▷ when we want to test the random-effects, the fixed-effects structure is also allowed to be different (though comparing nested models is a requirement for using the standard tests)

## 5.6 Hypothesis Testing (cont'd)

---

- **Example:** In the PBC dataset and for the dichotomous longitudinal outcome excess serum cholesterol levels (defined as before as above the threshold of 210 mg/dL), we fit a model that postulates
  - ▷ *fixed effects:*
    - \* main effects of time, treatment, and sex
    - \* interaction effects between time and treatment, and between drug and sex
  - ▷ *random effects:* random intercepts

We are interested in testing whether the model can be simplified by dropping the interaction terms



## 5.6 Hypothesis Testing (cont'd)

- The models under the two hypotheses are:

$$\left\{ \begin{array}{l} H_0 : \log \frac{\pi_{ij}}{1 - \pi_{ij}} = \beta_0 + \beta_1 \text{Time}_{ij} + \beta_2 \text{D-penicil}_i + \beta_3 \text{Female}_i + b_i \\ H_a : \log \frac{\pi_{ij}}{1 - \pi_{ij}} = \beta_0 + \beta_1 \text{Time}_{ij} + \beta_2 \text{D-penicil}_i + \beta_3 \text{Female}_i + \\ \quad \beta_4 \{ \text{Time}_{ij} \times \text{D-penicil}_i \} + \beta_5 \{ \text{Female}_i \times \text{D-penicil}_i \} + b_i \end{array} \right.$$

where  $\pi_{ij} = \Pr(\text{serChol}_{ij} > 210)$

## 5.6 Hypothesis Testing (cont'd)

- With respect to coefficients:

$$\begin{cases} H_0 : \beta_4 = \beta_5 = 0 \\ H_a : \text{at least one different from 0} \end{cases}$$

	df	logLik	AIC	BIC	LRT	p-value
$H_0$	5	-353.57	717.13	742.26		
$H_a$	7	-353.31	720.62	755.79	0.51	0.7736

- The results suggest that the interaction terms do not seem to improve the fit of the model

## 5.6 Hypothesis Testing (cont'd)

---

- Similarly to previous chapters, when we want to test non-nested models we can use information criteria, i.e., the AIC or the BIC

## 5.7 Review of Key Points

---

- GLMMs are the analogue of linear mixed models for categorical data
  - ▷ we include random effects in the linear predictor to account for the correlations in the outcomes belonging to the same groupe/cluster
  
- Features of GLMMs
  - ▷ these models provide a complete specification of the distribution of the grouped/longitudinal outcome – contrary to GEE, which is a semi-parametric method
  - ▷ interpretation of parameters is conditional on the random effects – contrary to GEE, which provide coefficients with a marginal interpretation

## 5.7 Review of Key Points (cont'd)

---

- Features of GLMMs
  - ▷ estimation of GLMMs is more complex, and requires careful choice of numerical algorithms
  - ▷ **they provide valid inferences under MAR – contrary to GEE, which only provide valid inferences under MCAR**
- Model building and hypothesis testing works in the same way as in the previous models we have seen