Chapter 5

Mixed Models for Discrete Data

5.1 Generalized Linear Mixed Models

- The previous chapter focused on 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

$$\Downarrow$$

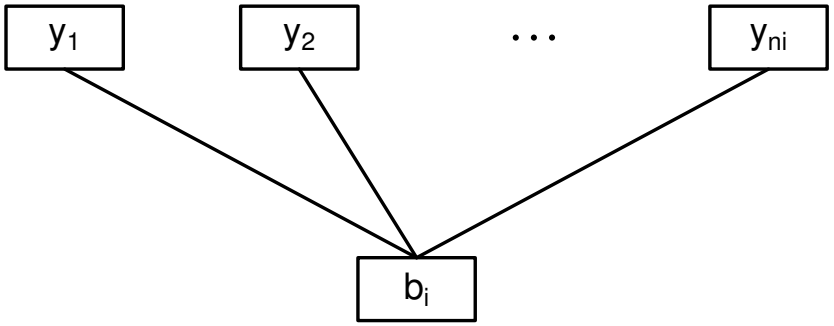
Generalized Linear Mixed Models (GLMMs)

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 the 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 grouped 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:

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

where

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

5.1 Generalized Linear Mixed Models (cont'd)

- Hence, 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 $E(y_{ij} \mid b_i) = \pi_{ij}$ and variance $\text{var}(y_{ij} \mid b_i) = \pi_{ij}(1 - \pi_{ij})$

  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 in contrast to the GEE approach, which is a semi-parametric method

  ▷ The mean and correlation structures are simultaneously defined using random effects
  $\Rightarrow$ As we will see next, this has direct and important implications with respect to the interpretation of parameters

5.2 Interpretation

- Example: In the AIDS dataset, a very low CD4 count (less than 150 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:

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

	Value	Std.Err.	z-value	p-value
β_0	6.250	0.899	6.954	< 0.001
β_1	0.149	0.044	3.392	0.001
β_2	-0.811	0.731	-1.109	0.267
β_3	-0.029	0.059	-0.494	0.622
σ_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:

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

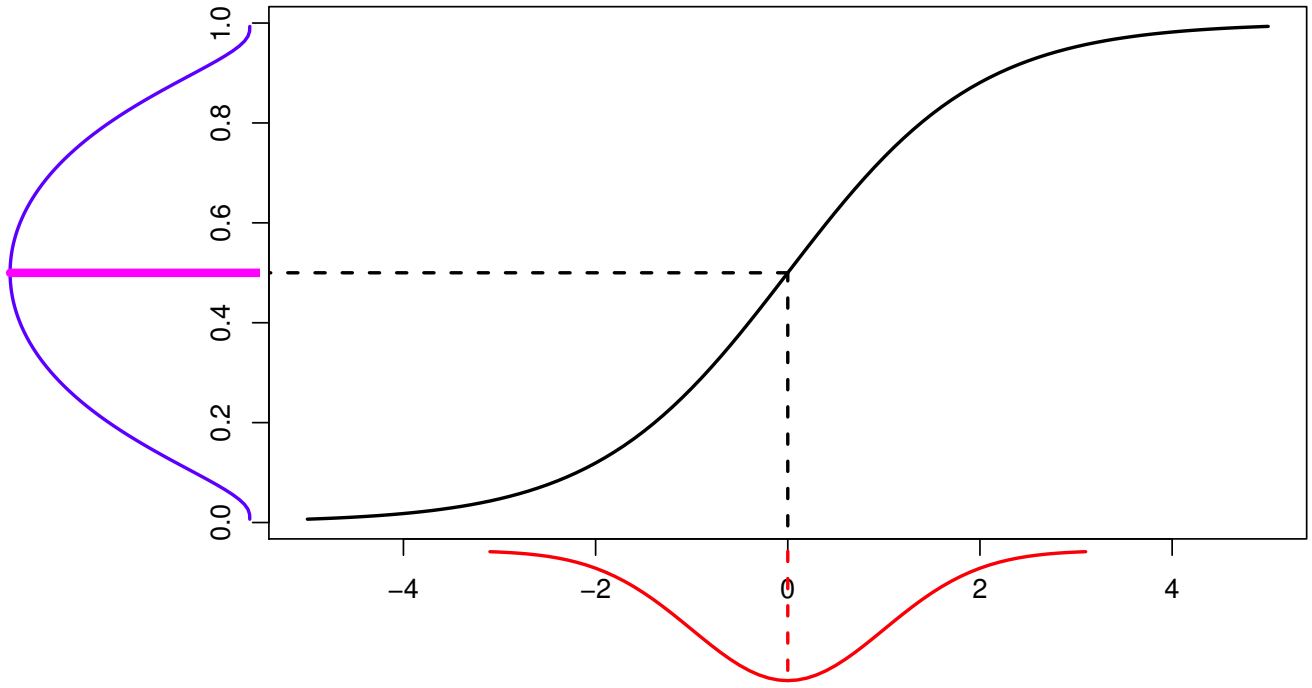
  ▷ mean of transformed interval $= 0.50194$

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

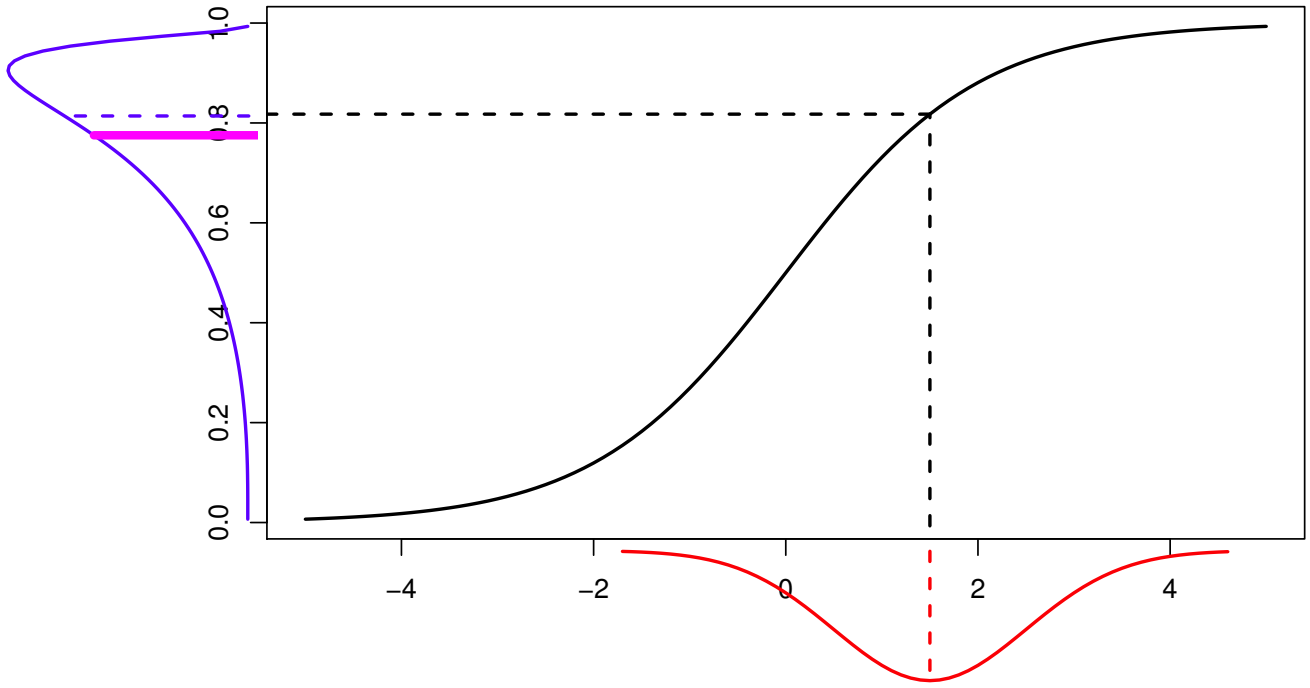
5.2 Interpretation (cont'd)

- Let's explain this issue graphically ...

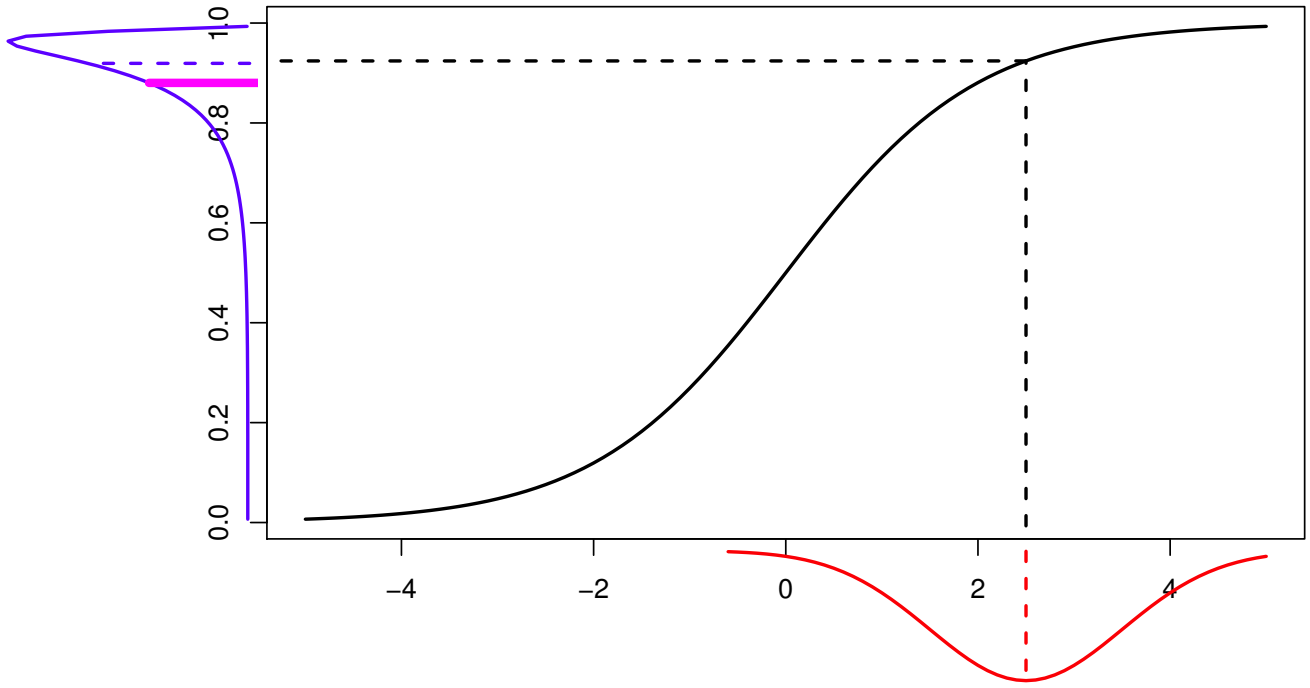
5.2 Interpretation (cont'd)



5.2 Interpretation (cont'd)



5.2 Interpretation (cont'd)

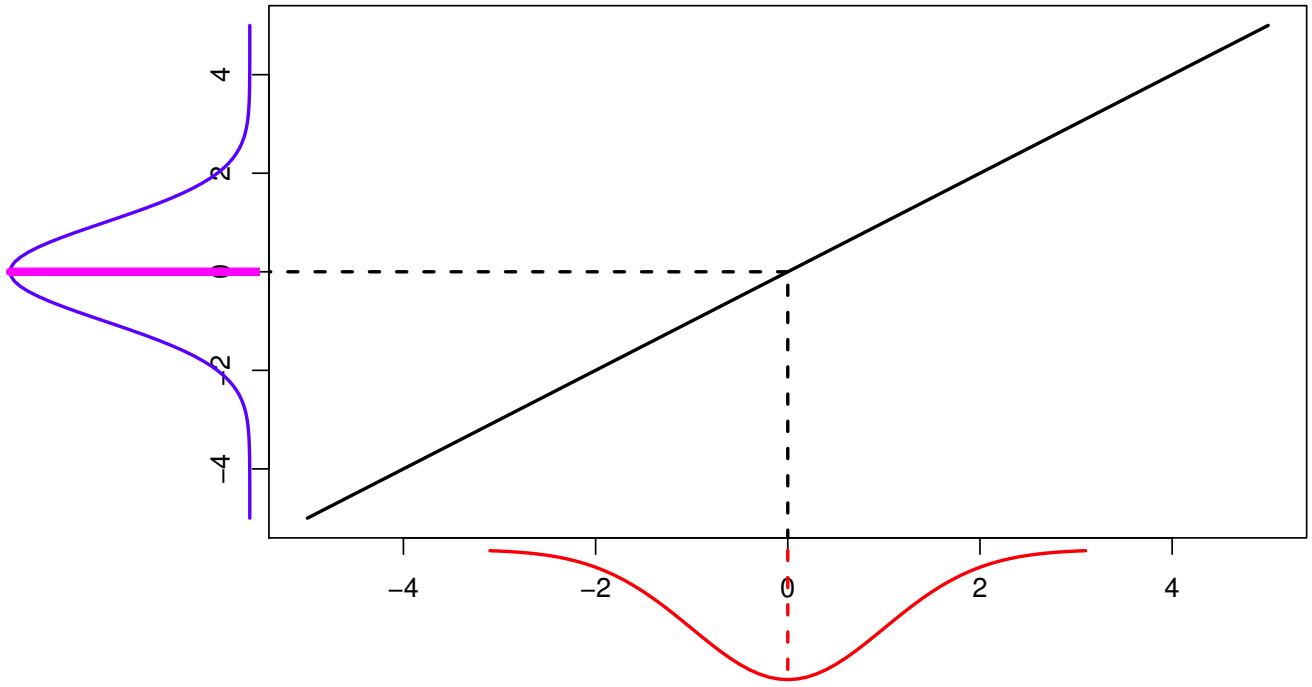


5.2 Interpretation (cont'd)

- We did not have this problem in the case of the linear mixed model because we did not have a link function
  - ▷ or put 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)

- The same complications also hold for the other fixed-effects coefficients of the logistic regression model
  - ▷ e.g., β_1 does **not** have the interpretation of the 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:

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

5.2 Interpretation (cont'd)

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

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

▷ hence, the corresponding odds ratio is:

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

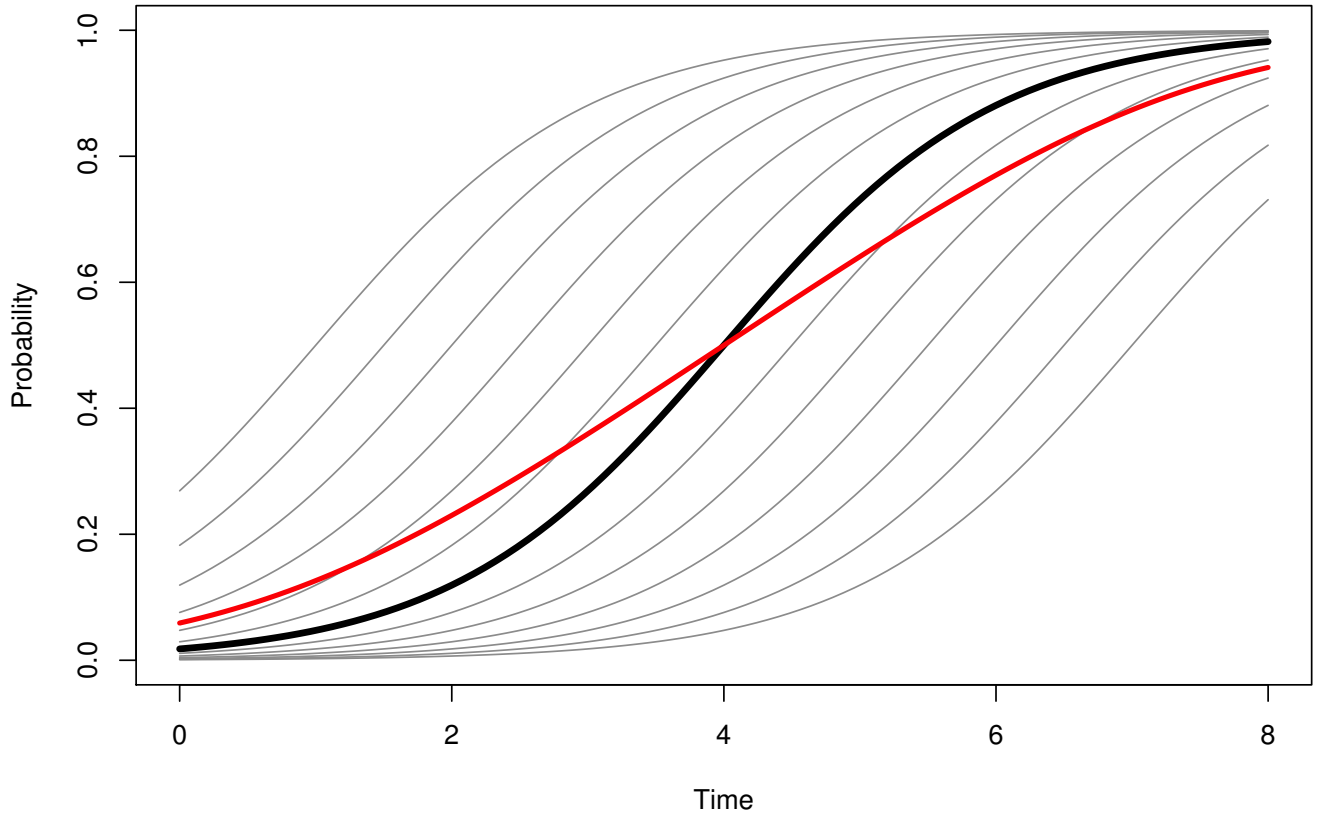
$$\text{odds ratio:} \quad \frac{\pi_{i'j}/(1-\pi_{i'j})}{\pi_{ij}/(1-\pi_{ij})} = \exp\{\beta_1 + (b_{i'} - b_i)\} \neq \exp(\beta_1)$$

5.2 Interpretation (cont'd)

- Hence, the interpretation of β_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) $\Rightarrow$ This line is actually $1/[1 + \exp\{-(\beta_0 + \beta_1 \texttt{Time}_{ij})\}]$
  - ▷ 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 β, 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

  - ▷ β^M denotes the marginal coefficients
  - ▷ β^{SS} denotes the subject-specific coefficients
  - ▷ σ_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.285) and we compute the corresponding marginal regression coefficients

	Subject-specific				Marginal	
	Value	Std.Err.	z-value	p-value	Value	Std.Err.
β_0	6.250	0.899	6.954	0.000	1.699	0.244
β_1	0.149	0.044	3.392	0.001	0.040	0.012
β_2	−0.811	0.731	−1.109	0.267	−0.220	0.199
β_3	−0.029	0.059	−0.494	0.622	−0.008	0.016
σ_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
  $\Rightarrow$ *the confidence intervals do not overlap*

5.3 Estimation

- ***

5.3 Estimation (cont'd)

- ***

5.4 GLMMs in R

- ***

5.4 GLMMs in R (cont'd)

- ***

5.5 Model Building

- ***

5.5 Model Building (cont'd)

- ***

5.6 Hypothesis Testing

- ***

5.6 Hypothesis Testing (cont'd)

- ***

5.7 Review of Key Points

- ***

5.7 Review of Key Points (cont'd)

- ***

5.7 Review of Key Points (cont'd)

- ***