

## O Modelo Linear Misto

- ① Modelo de Efeitos Fixos: apresenta somente fatores fixos, exceto o termo do erro experimental.
- ② Modelo de Efeitos Mistos: apresenta tanto fatores fixos como aleatórios, além do erro experimental.

### Modelo Linear Misto: Ideia

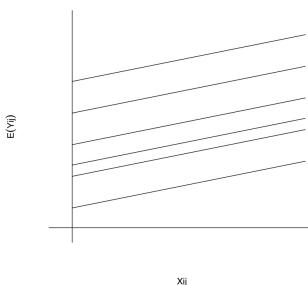
- Os parâmetros da regressão variam de indivíduo para indivíduo explicando as fontes de heterogeneidade da população.
- Cada indivíduo tem a sua própria trajetória média e um subconjunto dos parâmetros de regressão são tomados como aleatórios.
- Efeitos fixos são compartilhados por todos os indivíduos e os aleatórios são específicos de cada um.

### Modelo Linear Misto: Intercepto aleatório

$$Y_{ij} = \beta_{0i} + \beta_1 t_{ij} + \varepsilon_{ij}$$

$$= (\beta_0 + b_{0i}) + \beta_1 t_{ij} + \varepsilon_{ij}$$

Intercepto Aleatório

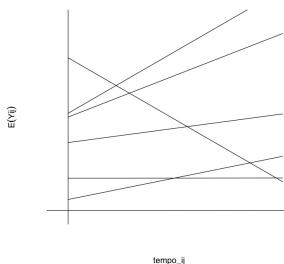


### Modelo Linear Misto: Intercepto e Inclinação Aleatórios

$$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + \varepsilon_{ij}$$

$$= (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) t_{ij} + \varepsilon_{ij}$$

Intercepto e Inclinação Aleatórios



## Modelo Linear Misto

### • Características:

- ① Características populacionais  $\beta$  (fixos);
- ② Características individuais  $\beta_i$  ou  $b_i$  (aleatórios).

### • Efeito:

- ① Média:  $E(Y_i) = X_i\beta$
- ② Estrutura de Covariância: Efeito aleatório induz  $Var(Y_i)$ .
- ③ Separa a variação entre indivíduos daquela intra indivíduos.
- ④ Permite obter estimativa de trajetórias individuais no tempo.

### Modelo Linear Misto - Simetria Composta

Modelo com intercepto aleatório:

$$Y_{ij} = \beta_{0i} + \beta_1 t_{ij} + \varepsilon_{ij}$$

- $\beta_{0i} \sim N(\beta_0, \sigma_{\beta_0}^2)$ .
- $\varepsilon_{ij} \sim N(0, \sigma^2)$ .
- $\beta_{0i}$  e  $\varepsilon_{ij}$  são independentes.

$$\text{① } Var(Y_{ij}) = \sigma^2 + \sigma_{\beta_0}^2.$$

$$\text{② } Cov(Y_{ij}, Y_{ij'}) = \sigma_{\beta_0}^2$$

### Modelo Linear Misto - Inclinação aleatória

Modelo com intercepto e inclinação aleatórios:

$$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + \varepsilon_{ij}$$

- $\beta_{0i} \sim N(\beta_0, \sigma_{\beta_0}^2)$ ,  $\beta_{1i} \sim N(\beta_1, \sigma_{\beta_1}^2)$ ,  $Cov(\beta_{0i}, \beta_{1i}) = \sigma_{\beta_{01}}$ .
- $\varepsilon_{ij} \sim N(0, \sigma^2)$ .
- $\beta' = (\beta_{0i}, \beta_{1i})$  e  $\varepsilon_{ij}$  são independentes.

$$\text{① } Var(Y_{ij}) = \sigma_{\beta_0}^2 + \sigma_{\beta_1}^2 t_{ij}^2 + 2t_{ij}\sigma_{\beta_{01}} + \sigma^2.$$

$$\text{② } Cov(Y_{ij}, Y_{ij'}) = \sigma_{\beta_0}^2 + t_{ij}t_{ij'}\sigma_{\beta_1}^2 + (t_{ij} + t_{ij'})\sigma_{\beta_{01}}.$$

## Vantagens

- Predizer trajetórias individuais (ex: intercepto aleatório)

$$Y_{ij} = X_{ij}\beta + b_i + \varepsilon_{ij}$$

Resposta Média populacional:

$$E(Y_{ij}) = X_{ij}\beta$$

Resposta média para o  $i$ -ésimo indivíduo (trajetória):

$$E(Y_{ij}|b_i) = X_{ij}\beta + b_i.$$

- Flexibilidade em acomodar estruturas não balanceadas.

## Forma Geral do Modelo Misto

A forma geral do modelo é

$$Y_i = X_i\beta + Z_i b_i + \varepsilon_i,$$

em que:

- $\beta$  é um vetor  $p \times 1$  de efeitos fixos;
- $b_i$  é um vetor  $q \times 1$  de efeitos aleatórios;
- $b_i \sim N_q(0, D)$  e  $\varepsilon_{ij} \sim N(0, \sigma^2)$ , sendo  $b_i$  e  $\varepsilon_{ij}$  independentes.
- $q \leq p \Rightarrow Z_i$  é um subconjunto de  $X_i$ .

## Característica do Modelo

### 1 Média Populacional ou Marginal

$$E(Y_i) = X_i\beta.$$

### 2 Média condicional ou específica por indivíduo

$$E(Y_i|b_i) = X_i\beta + Z_i b_i.$$

### 3 Covariância Marginal

$$\text{Var}(Y_i) = Z_i \text{Var}(b_i) Z_i' + \sigma^2 I_{n_i}.$$

### 4 Podemos assumir que $\varepsilon_i \sim N(0, R_i)$ mas o usual é tomar $R_i = \sigma^2 I_{n_i}$ e interpretá-lo como covariância condicional. Ou seja,

$$\text{Var}(Y_i|b_i) = R_i = \sigma^2 I_{n_i}.$$

## Como os Efeitos Aleatórios Capturam Correlação

### Dados os efeitos aleatórios, as medidas de cada indivíduo são independentes (suposição de independência condicional)

$$p(y_i|b_i) = \prod_{j=1}^{n_i} p(y_{ij}|b_i)$$

### Marginalmente (integrando nos efeitos aleatórios, as medidas de cada indivíduo são correlacionadas)

$$p(y_i) = \int p(y_i|b_i)p(b_i)db_i \implies y_i \sim N_{n_i}(X_i\beta, Z_i \text{Var}(b_i) Z_i' + \sigma^2 I_{n_i})$$

## Formulação em dois Estágios do Modelo Linear Misto

No primeiro estágio as medidas longitudinais no  $i$ -ésimo indivíduo são modeladas como:

$$Y_i = Z_i \beta_i + \varepsilon_i,$$

em que  $Z_i$  covariáveis intra-indivíduo (ou tempo dependente) e  $\varepsilon_i \sim N(0, \sigma^2 I_{n_i})$ .

No segundo estágio, temos  $\beta_i$  aleatório (variando de indivíduo para indivíduo) tal que:

$$E(\beta_i) = A_i \beta,$$

em que  $A_i$  ( $q \times p$ ) contém somente covariáveis que variam entre indivíduos (não dependente do tempo) e

$$\text{Var}(\beta_i) = D.$$

Os dois componentes do modelo podem ser combinados para produzir um modelo de efeitos mistos para  $Y_i$ , embora existam restrições.

Reescreva os efeitos  $\beta_i$  como

$$\beta_i = A_i \beta + b_i,$$

em que  $b_i$  tem distribuição normal multivariada com média zero e covariância  $D$ .

Desta forma,

$$\begin{aligned} Y_i &= Z_i(A_i \beta + b_i) + \varepsilon_i \\ &= Z_i(A_i \beta) + Z_i b_i + \varepsilon_i \\ &= X_i \beta + Z_i b_i + \varepsilon_i \end{aligned}$$

Ou seja, sob a restrição que

$$X_i = Z_i A_i$$

obtém-se o modelo de efeitos aleatórios.

Ponderando sobre os efeitos aleatórios  $b_i$ , temos

$$E(Y_i) = (Z_i A_i) \beta = X_i \beta,$$

e

$$\text{Cov}(Y_i) = Z_i D Z_i' + \sigma^2 I_{n_i}.$$

Note que  $Z_i$  aparece tanto no modelo para a média marginal como no modelo para a covariância marginal.

Embora este modelo seja bastante similar àquele apresentado anteriormente, há uma importante diferença.

O modelo de dois estágios impõe uma restrição na escolha da matriz de desenho dos efeitos fixos, que requer a estrutura  $X_i = Z_i A_i$ , em que  $A_i$  contém apenas covariáveis invariantes no tempo e  $Z_i$  contém apenas covariáveis variantes no tempo.

Isso implica que qualquer covariável tempo-dependente deve ser especificada como efeito aleatório, o que é uma restrição desnecessária e, em algumas situações, inconveniente!

Por outro lado, uma estrutura simples para a covariância impõe uma estrutura bastante simples para a média!

Vimos que uma covariância simetria composta é obtida do modelo com interceptos aleatórios,

$$Y_i = Z_i \beta_i + \varepsilon_i,$$

sendo  $Z_i$  um vetor  $n_i \times 1$  de uns.

Tal modelo impede qualquer dependência da resposta no tempo, isto é,

$$E(Y_i) = (Z_i A_i) \beta$$

não pode depender do tempo, pois o tempo, como variável intra-indivíduo, não foi incluído em  $Z_i$  no primeiro estágio.

## Inferência Para o Modelo Misto

Considere o modelo

$$Y_i = X_i\beta + Z_ib_i + \varepsilon_i,$$

em que,  $b_i \sim N_q(0, D(\alpha))$  e  $\varepsilon_{ij} \sim N(0, \sigma^2)$ ,  $b_i$  e  $\varepsilon_{ij}$  independentes.

Tem-se:  $p$  efeitos fixos e  $\frac{q(q+1)}{2} + 1$  efeitos aleatórios.

Inferência Estatística para  $\theta = (\beta, \alpha, \sigma^2)$ :

- ➊ Máxima Verossimilhança.
- ➋ Máxima Verossimilhança Restrita.

## Função de Verossimilhança

A função de verossimilhança é dada por:

$$\begin{aligned} L(\theta|y) &= \prod_{i=1}^N p(y_i|\theta) \\ &= \prod_{i=1}^N \int p(y_i, b_i|\theta) db_i \\ &= \prod_{i=1}^N \int p(y_i|b_i, \theta)p(b_i|\theta) db_i \end{aligned}$$

em que,  $p(y_i|b_i, \theta) \sim N_n(X_i\beta + Z_i b_i, \sigma^2 I_n)$  e  $p(b_i|\theta) \sim N_q(0, D)$ .

## Observações

➊ O EMV É obtido usando verossimilhança perfilada e iterações via algoritmo EM e/ou Newton-Raphson. Detalhes numéricos podem ser encontrados em Pinheiro e Bates (2000), Cap. 2.

➋ O EMVR também pode ser obtido através de

$$I^*(\theta) = I(\theta) + \text{termo}.$$

➌ A função `lme` do R fornece EMVR e EMV usando um enfoque híbrido (EM + Newton-Raphson). Esta função é de autoria de Pinheiro e Bates.

➍ EMV e EMVR têm assintoticamente as propriedades usuais de um estimador de máxima verossimilhança (consistência e normalidade).

## Avaliação dos Componentes de Variância

➊ Número de componentes é igual a  $\frac{q(q+1)}{2} + 1$  em que  $q$  é o número de efeitos aleatórios no modelo.

➋ Muitas situações envolvem  $q = 2$  (intercepto e inclinação aleatórios) e portanto:

$$\frac{2(2+1)}{2} + 1 = 4,$$

que permite termos heterogeneidade de variâncias e covariâncias pois ficam em função do tempo.

➌ A escolha da "melhor" estrutura de variância-covariância pode ser realizada utilizando o teste da RMVR. Estes testes, usualmente, são na fronteira do espaço de parâmetros. Neste caso, a estatística da RMVR não tem, sob  $H_0$ , uma distribuição qui-quadrado.

### Dist. da Estatística da RMVR sob $H_0$

➍ A distribuição neste caso é uma mistura (50:50) de dist. qui-quadrado. Ou seja, por exemplo, para  $H_0 : \sigma_{\beta_1} = 0$

$$RMVR \sim 0.5\chi_q + 0.5\chi_{q+1}$$

➎ Exemplo

Modelo completo:  $q = 2$  (intercepto e inclinação aleatórios)

Modelo restrito:  $q = 1$  (somente intercepto aleatório)

Teste usual (errado): nível de significância: 5,99

Teste correto:

$$RMVR \sim 0.5\chi_1 + 0.5\chi_2$$

nível é 5,14 (Tabela, Apênd. C, Fitzmaurice et al, 2004).

➏ Proposta ad hoc: para testar a 0,05, use o nível de 0,10.

## Predição dos Efeitos Aleatórios

Objetivo: predizer perfis individuais ou identificar indivíduos acima ou abaixo do perfil médio.

Obs.: não dizemos estimar os efeitos pois os mesmos são aleatórios. Dizemos predizer os efeitos aleatórios.

Deseja-se:

$$\hat{Y}_i = \hat{E}(Y_i|b_i) = X_i\hat{\beta} + Z_i\hat{b}_i$$

e para tal necessita-se de  $\hat{b}_i$ , o chamado Estimador BLUP ("Best Linear Unbiased Predictor") de  $b_i$ .

No modelo linear misto,

- ➊  $Y_i$  e  $b_i$  tem uma distribuição conjunta normal multivariada.
- ➋ Usando conhecidas propriedades da normal multivariada, temos que

$$E(b_i|Y_i, \hat{\beta}) = DZ'_i Var(Y_i)^{-1}(Y_i - X_i\hat{\beta})$$

- ➌ Usando os estimadores MVR dos componentes de variância,

$$\hat{b}_i = \hat{D}Z'_i \widehat{Var}(Y_i)^{-1}(Y_i - X_i\hat{\beta})$$

o BLUP de  $b_i$ .

Desta forma obtemos:

$$\begin{aligned} \hat{Y}_i &= X_i\hat{\beta} + Z_i\hat{b}_i \\ &= X_i\hat{\beta} + Z_i\hat{D}Z'_i \widehat{Var}(Y_i)^{-1}(Y_i - X_i\hat{\beta}) \\ &= X_i\hat{\beta} + (Z_i\hat{D}Z'_i + \hat{R}_i - \hat{R}_i)\widehat{Var}(Y_i)^{-1}(Y_i - X_i\hat{\beta}) \\ &= (\hat{R}_i \widehat{Var}(Y_i)^{-1})X_i\hat{\beta} + (I_{n_i} - \hat{R}_i \widehat{Var}(Y_i)^{-1})Y_i \end{aligned}$$

em que  $R_i = Var(Y_i|b_i) = Var(\varepsilon)$ .

Interpretação: média ponderada entre a média populacional  $X_i\hat{\beta}$  e o  $i$ -ésimo perfil observado. Isto significa que o perfil predito é "encolhido" na direção da média populacional.

A quantidade de "encolhimento" depende da magnitude de  $R_i$  e  $Var(Y_i)$ .

- ➊  $R_i$ : variação intra-indivíduo;
- ➋  $Var(Y_i)$ : variação total (entre e intra-indivíduo).

Quando  $R_i$  é relativamente grande, e a variabilidade intra indivíduo é maior que a variabilidade entre indivíduo, mais peso é atribuído a  $X_i\hat{\beta}$ , a média populacional estimada, do que à resposta individual observada.

Por outro lado, quando a variabilidade entre indivíduos é grande em relação à variabilidade intra indivíduos, mais peso é dado à resposta observada  $Y_i$ .

Finalmente, o grau de "encolhimento" em direção à média populacional também depende de  $n_i$ .

Em geral, há maior encolhimento em direção à curva media populacional quando  $n_i$  é pequeno.

Intuitivamente, isso faz sentido já que menos peso deve ser dado à trajetória observada do indivíduo quando menos dados estão disponíveis.

## Respostas Longitudinal Não-Gaussiana

Até o momento, vimos situações em que a variável resposta era contínua. Mais especificamente, consideramos o caso em que a resposta era normal.

Vamos considerar agora outras escalas para a variável resposta. Por exemplo, podemos assumir respostas binárias ou contagens.

Abordaremos os modelos:

- Modelos Lineares Generalizados Mistos.
- Modelos Marginais: GEE.

## Revisão: Modelos Lineares Generalizados

Os Modelos Lineares Generalizados (GLM) são uma classe unificada de modelos de regressão.

1 Considere  $Y_1, \dots, Y_N$  uma amostra aleatória de respostas univariadas (desenho transversal).

2 Um vetor de  $p$ -covariáveis associados a cada resposta  $Y_i$ . Ou seja,

$$X_i = \begin{pmatrix} X_{i0} \\ X_{i1} \\ \vdots \\ X_{ip} \end{pmatrix},$$

em que  $X_{i0} = 1$ .

3 O MLG é definido por três componentes:

- Distribuição de  $Y_i$ .
- Componente Sistemático (preditor linear).
- Função de Ligação.

$$\eta_i = X'_i \beta = \beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}$$

## GLM - Família Exponencial

A distribuição de  $Y_i$  pertence à família exponencial que inclui os principais modelos estatísticos: normal, binomial, poisson, exponencial, etc.

Ou seja,  $Y_i$  tem densidade  $f(Y_i|\theta, \phi)$  pertencente à família exponencial:

$$f(y_i|\theta_i, \phi) = \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right\},$$

em que  $\theta_i$  é o parâmetro natural,  $\phi$  é o parâmetro de escala e  $b(\cdot)$  e  $c(\cdot)$  são funções específicas.

## Modelos Lineares Generalizados

•  $b(\cdot)$  é a função geradora de momentos

- $\mu = E(Y) = b'(\theta)$  e
- $Var(Y) = \phi b''(\theta)$

• Em geral, média e variância são relacionadas.

$$Var(Y) = \phi b''(b'^{-1}(\mu)) = \phi \nu(\mu)$$

• A função  $\nu(\mu)$  é chamada de função de variância.

•  $b'^{-1}$  que relaciona  $\theta$  com  $\mu$  é chamada de função de ligação.

## Exemplo: Modelo Normal ( $\mu, \sigma^2$ )

- $\theta = \mu$
- $\phi = \sigma^2$
- $b(\theta) = \theta^2/2$
- Média:  $\mu = \theta$  e  $\nu(\mu) = 1$
- Observe que no modelo normal, média e variância não são relacionadas

$$\phi \nu(\mu) = \sigma^2$$

- Função de ligação natural:  $\theta = \mu$ .

## Exemplo: Modelo Bernoulli ( $\pi$ )

- $\theta = \log(\pi/(1-\pi))$
- $\phi = 1$
- $b(\theta) = -\log(1-\pi) = \log(1 + \exp(\theta))$
- Média:  $\mu = \pi = \frac{\exp(\theta)}{1+\exp(\theta)}$  e  $\nu(\mu) = \pi(1-\pi) = \frac{\exp(\theta)}{(1+\exp(\theta))^2}$
- Observe que no modelo bernoulli, média e variância são relacionadas

$$\phi \nu(\mu) = \mu(1-\mu)$$

- Função de ligação natural:  $\theta = \log\left(\frac{\mu}{1-\mu}\right)$ .

## Função de Ligação Natural ou Canônica

A função de ligação natural ou canônica é dada por

$$g(\mu_i) = \theta_i = \eta_i = X_i \beta = \beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}.$$

Por exemplo,

- Gaussiano:  $g(\mu_i) = \eta_i$  (identidade)
- Bernoulli:  $g(\mu_i) = \text{logit}(\eta_i)$ .
- Poisson:  $g(\mu_i) = \log(\eta_i)$

## Inferência por Máxima Verossimilhança

- Função de log-verossimilhança  $\log L(\cdot) = l(\cdot)$

$$L(\beta) = \prod_{i=1}^N f(y_i|\theta_i, \phi) = \prod_{i=1}^N \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right\}$$

- Equações escore: derivada de  $l(\cdot)$ .

- Inferência baseada na teoria assintótica de Máxima Verossimilhança.

## Relembrando: Modelo Normal

- Modelo Marginal

$$Y_{ij} = X'_{ij}\beta + \varepsilon_{ij}$$

e

$$E(Y_{ij}|X_{ij}) = X'_{ij}\beta.$$

- Modelo Condisional

$$Y_{ij} = X'_{ij}\beta + Z'_{ij}b_i + \varepsilon_{ij}$$

em que:

- $(\beta)_{p \times 1}$ : efeitos fixos;
- $(b_i)_{q \times 1}$ : efeitos aleatórios.

e,

$$b_i \sim N_q(0, D) \text{ e } \varepsilon_{ij} \sim N(0, \sigma^2)$$

Sendo  $b_i$  e  $\varepsilon_{ij}$  independentes.

- Média Condisional ou Específica por Indivíduo

$$E(Y_{ij}|b_i, X_{ij}) = X'_{ij}\beta + Z'_{ij}b_i.$$

e a Covariância Marginal

$$\text{Var}(Y_{ij}) = Z_i D Z'_i + \sigma^2 I_{n_i}.$$

- Podemos assumir que  $\varepsilon_{ij} \sim N(0, R_i)$  mas o usual é tomar  $R_i = \sigma^2 I_{n_i}$  e interpretá-lo como covariância condicional. Ou seja,

$$\text{Var}(Y_{ij}|b_i) = R_i = \sigma^2 I_{n_i}.$$

## Modelo Bernoulli

- Modelamos

$$\mu_{ij} = E(Y_{ij}|X_{ij}) \text{ (modelo marginal)}$$

$$\mu_{ij} = E(Y_{ij}|b_i, X_{ij}) \text{ (modelo condicional).}$$

- Função de ligação: logit (mais comum)

$$\text{logit}(\mu_{ij}) = X'_{ij}\beta \quad \text{Modelo Marginal}$$

$$\text{logit}(\mu_{ij}) = X'_{ij}\beta + Z'_{ij}b_i; \quad \text{Modelo Condisional}$$

## Modelo Poisson

- Modelamos

$$\mu_{ij} = E(Y_{ij}|X_{ij}) \text{ (modelo marginal)}$$

$$\mu_{ij} = E(Y_{ij}|b_i, X_{ij}) \text{ (modelo condicional).}$$

- Função de ligação: logarítmica (mais comum)

$$\log(\mu_{ij}) = X'_{ij}\beta \quad \text{Modelo Marginal}$$

$$\log(\mu_{ij}) = X'_{ij}\beta + Z'_{ij}b_i \quad \text{Modelo Condisional}$$

## Modelos Lineares Generalizados Longitudinais

- Fácil transferência entre modelos (marginal e condicional) para resposta gaussiana.

- Transferência difícil entre modelos quando a resposta não é gaussiana.

- Modelos Marginais

- Especificação completa: o ajuste por MV pode ser complicado.
- Alternativa Não-Verosimilhança: MQG, GEE, etc.

- Modelos Condicionais: ajuste complicado.

## Modelos Marginais: GEE

Equações de Estimação Generalizadas

$$\sum_{i=1}^N D'_i V_i^{-1} (Y_i - \mu_i) = 0,$$

em que

- $D_i = \partial \mu_i / \partial \beta$  e  $\mu_i = g^{-1}(X_i \beta)$ , o inverso da função de ligação  $g(\cdot)$ .
- $\text{Var}(Y_i) \approx V_i = \phi A_i^{1/2}(\beta) R_i(\alpha) A_i^{1/2}(\beta)$  em que  $A_i$  é uma matriz diagonal formada por  $\text{Var}(Y_{ij})$ ,  $R_i$  é matriz de correlação de trabalho e  $\phi$  é um parâmetro de dispersão/escala.
- $\text{Var}(\hat{\beta})$  é estimada pela variância robusta (estimador sanduíche).

## Formas de Correlação de Trabalho

- independência,  $R_i(\alpha) = I_{n_i}$ :

⇒ dados longitudinais não correlacionados.

- simetria composta, especifica que  $R_i(\alpha) = \rho \mathbf{1}_{n_i} \mathbf{1}'_{n_i} + (1 - \rho) \mathbf{I}_{n_i}$ :

⇒ mesma correlação para qualquer par de tempo.

- AR-1, para a qual  $R_i(\alpha) = \rho^{|j-i|}$ :

⇒ válida para medidas igualmente espaçadas no tempo;

- não estruturada;

⇒ estima todas as  $n_i(n_i - 1)/2$  correlações de  $R$ .

## Variância do Estimador

- Naive ou "baseada no modelo"

$$\widehat{\text{Var}}(\hat{\beta}) = \left( \sum_{i=1}^N \hat{D}'_i V_i(\hat{\alpha})^{-1} \hat{D}_i \right)^{-1}.$$

- Robusta ou "empírica"

$$\widehat{\text{Var}}(\hat{\beta}) = M_0^{-1} M_1 M_0^{-1},$$

em que

$$M_0 = \sum_{i=1}^N \hat{D}'_i V_i(\hat{\alpha})^{-1} \hat{D}_i,$$

$$M_1 = \sum_{i=1}^N \hat{D}'_i V_i(\hat{\alpha})^{-1} (\mathbf{y}_i - \hat{\mu}_i)(\mathbf{y}_i - \hat{\mu}_i)' V_i(\hat{\alpha})^{-1} \hat{D}_i.$$

## Estimando a Correlação de Trabalho

- Liang e Zeger (1986) utilizaram estimativas de momentos para os parâmetros da matriz de correlação de trabalho.

- Ou seja, utilizar estimadores baseados nos resíduos para as quantidades envolvidas em  $R_i$ .

- Resíduos de Pearson:

$$e_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\nu}_{ij}}},$$

em que  $\nu_{ij} = \mu_{ij}(1 - \mu_{ij})$  para resposta binária e  $\nu_{ij} = \mu_{ij}$ , para contagem.

## Estimadores de Momentos usando Resíduos

- **Independência:**  $\text{Corr}(Y_{ij}, Y_{ij'}) = 0$ .
- **Simetria Composta:**  $\text{Corr}(Y_{ij}, Y_{ij'}) = \alpha$ .

$$\hat{\alpha} = \frac{\sum_{i=1}^N \sum_{j>j'} e_{ij} e_{ij'}}{\sum_{i=1}^N n_i(n_i - 1)/2 - p}.$$

- **Não estruturada:**  $\text{Corr}(Y_{ij}, Y_{ij'}) = \alpha_{jj'}$ .

$$\hat{\alpha}_{jj'} = \frac{\sum_{i=1}^N e_{ij} e_{ij'}}{\sum_{i=1}^N n_i - p}.$$

Podemos estimar  $\phi$  por

$$\hat{\phi} = \sum_{i=1}^N \sum_{j=1}^{n_i} e_{ij}^2 / (\sum_{i=1}^N n_i - p).$$

## Ajustando GEE

- 1 Encontrar a estimativa inicial para  $\beta$  por exemplo via MLE assumindo independência (passo 0).
- 2 Encontre os resíduos e estime  $\alpha$  e  $\phi$ .
- 3 Atualize a estimativa de  $\beta$  (passo  $j + 1$ ):  
$$\hat{\beta}^{j+1} = \hat{\beta}^j + \left\{ \sum_{i=1}^N D'_i(\hat{\beta}^j) V_i^{-1}(\hat{\beta}^j) D_i(\hat{\beta}^j) \right\} \left\{ \sum_{i=1}^N D'_i(\hat{\beta}^j) V_i^{-1}(\hat{\beta}^j) (Y_i - \mu_i(\hat{\beta}^j)) \right\}.$$
- 4 Faça iterações em (2)-(3) até a convergência.
- 5 Estime  $\text{Var}(\hat{\beta})$  usando o estimador sanduíche.

## Modelos Lineares Generalizados Mistos

A ideia intuitiva por trás dos modelos lineares generalizados mistos (GLMM) é a mesma dos modelos lineares mistos, isto é:

- A correlação entre as medidas repetidas é induzida por efeitos aleatórios não observados.
- Em outras palavras: as medidas longitudinais de um indivíduo são correlacionadas porque todas elas compartilham os mesmos efeitos aleatórios não observados (suposição de independência condicional).

Vamos considerar agora o caso de respostas não gaussianas.

### 1 Modelos Lineares Generalizados

- Resposta na família exponencial: normal, gama, exponencial, Bernoulli, Poisson, etc.
- Preditor Linear:  $X_i'\beta$ .
- Função de Ligação:  $g(\mu_i) = X_i'\beta$ .

### 2 Modelos Lineares Generalizados Mistos

Preditor Linear:

$$\eta_i = X_i'\beta + Z_i b_i.$$

### 3 Função de Ligação

$$g(E(Y_{ij}|b_i)) = X_{ij}'\beta + Z_{ij}'b_i$$

em que:

- $(\beta)_{p \times 1}$ : efeitos fixos;
- $(b_i)_{q \times 1}$ : efeitos aleatórios.

Assumimos  $b_i \sim N_q(0, D)$ .

## Exemplo: Resposta Binária

Por exemplo, suponha que temos uma variável binária  $Y_{ij}$

$$Y_{ij} = \begin{cases} 1, & \text{se o indivíduo } i \text{ tem uma resposta positiva no tempo } j \\ 0, & \text{se o indivíduo } i \text{ tem uma resposta negativa no tempo } j \end{cases}$$

O modelo misto genérico para  $Y_{ij}$  é uma Regressão Logística de Efeitos Mistos e tem a forma

$$\log \left( \frac{\pi_{ij}}{1 - \pi_{ij}} \right) = X_{ij}'\beta + Z_{ij}'b_i$$
$$b_i \sim N(0, D)$$

em que  $\pi_{ii} = P(Y_{ii} = 1)$  é a probabilidade de uma resposta positiva. Mais formalmente, temos a seguinte especificação:

- 1 Condisional nos efeitos aleatórios  $b_i$ , as respostas  $Y_{ij}$  são independentes e têm uma distribuição de Bernoulli com média  $E(Y_{ij}|b_i) = \pi_{ij}$  e variância  $\text{Var}(Y_{ij}|b_i) = \pi_{ij}(1 - \pi_{ij})$ .
- 2 A média condicional de  $Y_{ij}$  depende de efeitos fixos e aleatórios via

$$\log \left( \frac{\pi_{ij}}{1 - \pi_{ij}} \right) = X_{ij}'\beta + Z_{ij}'b_i$$

- 3 Os efeitos aleatórios seguem uma distribuição normal multivariada com média zero e matriz de variância-covariância  $D$ .

Notas sobre a definição do modelos:

- A especificação do modelo misto corresponde à especificação completa da distribuição da resposta  $Y_{ij}$ , o que contrasta com o enfoque GEE, que é um método semi-paramétrico.
- As estruturas de média e correlação são simultaneamente definidas pelos efeitos aleatórios.
- Isso tem implicações diretas e importantes sobre a interpretação dos parâmetros.

## Modelo Generalizado Misto Longitudinal

- Os coeficientes de regressão fixos são interpretados em termos dos efeitos das covariáveis em mudanças na resposta média transformada *individual*, mantidas constantes as demais covariáveis.
- Como os componentes dos efeitos fixos  $\beta$ , têm interpretações que dependem de manter  $b_i$  fixado, eles são frequentemente chamados de *coeficientes de regressão específicos do indivíduo*.
- Como resultado, os modelos mistos são mais úteis quando o principal objetivo é fazer inferências sobre os indivíduos ao invés da população.
- As médias populacionais são os objetivos de inferência nos modelos marginais (i.e. GEE).

# Estimação

A estimação do modelo linear generalizado misto é baseada nos mesmos princípios vistos nos modelos marginais e mistos para dados contínuos.

Isto é, nós temos uma especificação completa da distribuição dos dados (diferente do GEE) e, portanto, podemos usar *máxima verossimilhança*.

Contudo, há uma importante complicação: o ajuste de um modelo linear generalizado misto é uma tarefa computacionalmente desafiadora!

A função de verossimilhança é dada por

$$\begin{aligned} L(\theta) = L(\theta|y) &= \prod_{i=1}^N p(y_i|\theta) \\ &= \prod_{i=1}^N \int p(y_i; b_i|\theta) db_i \\ &= \prod_{i=1}^N \int p(y_i|b_i, \theta) p(b_i|\theta) db_i \end{aligned}$$

em que,  $p(y_i|b_i, \theta) \sim \text{Bernoulli-logit}/\text{Poisson-log}$ , etc e  $p(b_i|\theta) \sim N_q(0, D)$ .

A função de log-verossimilhança tem a mesma forma daquela do modelo linear misto:

$$l(\theta) = \sum_{i=1}^N \log \int p(y_i|b_i, \theta) p(b_i|\theta) db_i.$$

No modelo linear misto os dois termos do integrando

- $p(y_i|b_i, \theta)$
- $p(b_i|\theta)$

são densidades da distribuição normal (multivariada). A integral da log-verossimilhança tem uma solução de forma fechada.

Já nos modelos lineares generalizados mistos os dois termos do integrando denotam densidades de diferentes distribuições.

Por exemplo, no caso logístico

- $p(y_i|b_i, \theta) \implies$  distribuição Bernoulli
- $p(b_i|\theta) \implies$  distribuição normal multivariada

A implicação é que a mesma integral não tem uma solução de forma fechada.

Para contornar o problema, dois tipos gerais de solução foram propostos na literatura:

- *Aproximação do integrando*: envolve aproximar o produto dentro da integral (i.e.  $p(y_i|b_i, \theta)p(b_i|\theta)$ ) por uma distribuição normal multivariada para a qual a integral tem forma fechada:
  - Quase-verossimilhança Penalizada (PQL)
  - Aproximação de Laplace
- *Aproximação da integral*: envolve aproximar toda a integral (i.e.  $\int p(y_i|b_i, \theta)p(b_i|\theta)db_i$ ) por uma soma:
  - Quadratura Gaussiana e Quadratura Gaussiana Adaptativa
  - Monte Carlo e MCMC (enfoque Bayesiana)

Dentre as duas alternativas, os métodos que se baseiam na aproximação da integral são considerados superiores.

Eles são computacionalmente caros e possuem um parâmetro que controla a acurácia da aproximação:

- Para quadratura gaussiana é o número de pontos de quadratura (quadratura gaussiana adaptativa é equivalente à aproximação de Laplace).
- Nos enfoques Monte Carlo/MCMC é o número de amostras.

A estimativa dos efeitos aleatórios procede de forma similar àquela dos modelos mistos.

Baseado no modelo misto ajustado, as estimativas dos efeitos aleatórios são baseadas na distribuição *a posteriori*

$$\begin{aligned} p(b_i|Y_i; \theta) &= \frac{p(Y_i|b_i; \theta)p(b_i|\theta)}{p(Y_i|\theta)} \\ &\approx p(Y_i|b_i; \theta)p(b_i|\theta) \end{aligned}$$

em que  $\theta$  é substituído pelo seu EMV  $\hat{\theta}$ .

- Para obter estimativas dos efeitos aleatórios nós usamos medidas de locação dessa distribuição *a posteriori* (e.g. média ou moda).
- Para estimativa da dispersão dos efeitos aleatórios nós usamos a variância da curvatura local em torno da moda da distribuição *a posteriori*.
- Diferente do modelo linear misto esta distribuição não tem forma fechada. Assim, o cálculo das medidas de locação de dispersão envolve o uso de algoritmos numéricos.

## Interpretação dos Parâmetros

Como discutido anteriormente:

- O vetor  $\beta$  no GEE tem interpretação populacional. Ou seja, a mesma interpretação dos modelos transversais.
- O vetor  $\beta$  no modelo GLMM tem interpretação condicional sob o nível dos efeitos aleatórios. Ou seja, interpretação específica para cada indivíduo.
- Portanto, as estimativas dos modelos são diferentes!

A seguir, aprofundaremos as conexões entre os dois modelos.

## Coxeções entre modelos de efeitos aleatórios e modelos marginais

No modelo misto modelamos a média condicional,  $\mu_{ij} = E(Y_{ij}|b_i)$ .

Invertendo a função de ligação, obtemos

$$E(Y_{ij}|b_i) = g^{-1}(X'_{ij}\beta + Z'_{ij}b_i).$$

Marginalmente, ponderando sobre os efeitos aleatórios, a média é

$$E(Y_{ij}) = E[E(Y_{ij}|b_i)] = \int g^{-1}(X'_{ij}\beta + Z'_{ij}b_i)f(b_i, D)db_i,$$

em que  $f(b_i, D)$  é a  $N_q(0, D)$ , densidade dos efeitos aleatórios.

## Conexões entre modelos de efeitos aleatórios e modelos marginais

Para a função de ligação identidade,

$$E(Y_{ij}) = \int (X'_{ij}\beta + Z'_{ij}b_i)f(b_i, D)db_i = X'_{ij}\beta.$$

O modelo marginal tem a mesma forma e efeitos  $\beta$ . Isto não é verdade para outras ligações.

Por exemplo, para o modelo logístico

$$E(Y_{ij}) = E\left[\frac{\exp(X'_{ij}\beta + Z'_{ij}b_i)}{1 + \exp(X'_{ij}\beta + Z'_{ij}b_i)}\right],$$

Esta esperança não tem a forma  $\exp(X_{ij})/[1 + \exp(X_{ij})]$ , exceto quando  $b_i$  tem uma distribuição degenerada ( $\sigma_b = 0$ ).

Zeger et. al. (1988) mostraram que, para o modelo condicional,

$$\text{logit}(\mu_{ij}) \approx a(D)X'_{ij}\beta,$$

em que  $a(D) = |c^2DZ'_{ij}Z'_{ij} + 1|^{-q/2}$  e  $c = \frac{16\sqrt{3}}{15\pi}$ .

No caso particular em que  $q = 1$  (intercepto aleatório), temos a relação aproximada:

$$\beta_M \approx \frac{\beta_{EA}}{\sqrt{1 + \frac{16\sqrt{3}}{15\pi}\sigma_b^2}}.$$

Assim, se  $\text{Var}(b_i) = \sigma_b > 0$ , o efeito marginal  $\beta_M$  é menor que o efeito condicional  $\beta_{EA}$ .

Além disso, a discrepância entre  $\beta_{EA}$  e  $\beta_M$  aumenta quando  $\sigma_b$  cresce.

Por exemplo, se  $\sigma_b^2 = 3,5$ , então  $\beta_{EA} \approx 1,49\beta_M$ ; se  $\sigma_b^2 = 9$ , então  $\beta_{EA} \approx 2,03\beta_M$ .

A figura a seguir ilustra por que o efeito marginal é menor que o efeito condicional.

## Dados Ausentes em Estudos Longitudinais

O problema de dados ausentes em estudos longitudinais é muito mais grave que nos estudos transversais, pois a não-resposta pode ocorrer em qualquer ocasião.

Em áreas como a saúde, dados ausentes são a regra e não exceção!

Tipos:

- *intermitentes*: há uma ou mais perdas pontuais;
- *dropout*: há perda completa da informação a partir de um certo instante de tempo.

Consideraremos inicialmente o caso de perda apenas na variável resposta.

### Implicações para Análise

Dados ausentes têm três implicações gerais para a análise:

- i) Acarreta complicações para os métodos de análise que requerem dados balanceados;
- ii) Perda de informação com redução na precisão com que mudanças na resposta média podem ser estimadas;
- iii) Podem introduzir vícios e levar a inferências enganosas.

## Hierarquia de Mecanismos de Dados Ausentes (Rubin, 1976)

Um indivíduo tem um vetor de respostas  $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})'$ , com distribuição governada pelo vetor de parâmetros  $\theta$ .

Seja  $\mathbf{R}_i$  um vetor  $n_i \times 1$  de indicadoras da resposta ser observada  $\mathbf{R}_i = (R_{i1}, R_{i2}, \dots, R_{in_i})'$ , com  $R_{ij} = 1$  se  $Y_{ij}$  é observado e  $R_{ij} = 0$  se  $Y_{ij}$  é dado ausente.

A distribuição de  $\mathbf{R}$ ,  $P(\mathbf{R}|\mathbf{Y}, \psi)$ , pode depender de  $\mathbf{Y}$  assim como de parâmetros desconhecidos  $\psi$ .

Dado  $\mathbf{R}_i$ , temos a partição  $\mathbf{Y}_i = (\mathbf{Y}_{i,obs}, \mathbf{Y}_{i,mis})$ , correspondendo às respostas observadas e aos dados ausentes, respectivamente.

- *Missing Completely at Random* (MCAR): quando a não resposta é independente de dados observados ou não observados, isto é:

$$P(\mathbf{R}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \psi) = P(\mathbf{R}|\psi).$$

Ex: erros administrativos que ocorrem ao acaso, tais como acidentes em laboratório, perda de formulário, etc.

- *Missing at Random* (MAR): quando a probabilidade de não resposta é independente de  $\mathbf{Y}_{mis}$ :

$$P(\mathbf{R}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \psi) = P(\mathbf{R}|\mathbf{Y}_{obs}, \psi).$$

Ex: valores ausentes em indivíduos mais velhos, indivíduos de certa região, ou tempo de calendário.

- *Missing Not at Random* (MNAR): quando a probabilidade de não resposta depende de dados não observados  $\mathbf{Y}_{mis}$ :

$$P(\mathbf{R}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \psi) = P(\mathbf{R}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \psi).$$

Ex: não-resposta em certas questões (orientação sexual, renda, etc...), ou condição clínica (não-resposta se uma condição está presente, a qual não pode ser avaliada de forma precisa).

Compreender o mecanismo de não-resposta é fundamental para fazermos inferências corretas.

### Ignorabilidade: Perda na Resposta

Segundo Rubin (1987), sob uma condição de *ignorabilidade*, não precisamos considerar um modelo para  $\mathbf{R}$  envolvendo os parâmetros de perturbação  $\psi$  ao se fazer inferência sobre  $\theta$ .

Porque os dados observados consistem não apenas de  $\mathbf{Y}_{obs}$  mas também de  $\mathbf{R}$ , A distribuição de probabilidade dos dados observados é dada por:

$$\begin{aligned} P(\mathbf{R}, \mathbf{Y}_{obs}|\theta, \psi) &= \int P(\mathbf{R}, \mathbf{Y}|\theta, \psi)d\mathbf{Y}_{mis} \\ &= \int P(\mathbf{R}|\mathbf{Y}, \psi)P(\mathbf{Y}|\theta)d\mathbf{Y}_{mis} \end{aligned} \quad (1)$$

Sob MAR (1) se torna:

$$\begin{aligned} P(\mathbf{R}, \mathbf{Y}_{obs}|\theta, \psi) &= P(\mathbf{R}|\mathbf{Y}_{obs}, \psi) \int P(\mathbf{Y}|\theta)d\mathbf{Y}_{mis} \\ &= P(\mathbf{R}|\mathbf{Y}_{obs}, \psi)P(\mathbf{Y}_{obs}|\theta). \end{aligned} \quad (2)$$

Quando os dois parâmetros  $\psi$  e  $\theta$  são distintos, inferências de máxima verossimilhança sobre  $\theta$  não serão afetadas por  $\psi$  ou  $P(\mathbf{R}|\mathbf{Y}_{obs}, \psi)$ .

Nesse caso, o mecanismo de não resposta pode ser seguramente ignorado.

## Ignorabilidade: Perda na Resposta

A função de verossimilhança, ignorando o mecanismo de geração da não resposta, é dada por:

$$L(\theta | \mathbf{Y}_{obs}) \propto P(\mathbf{Y}_{obs} | \theta). \quad (3)$$

- Métodos baseados em momentos, tais como o GEE, produzem estimativas inconsistentes sob o mecanismo MAR.
- Modelos multivariados com estrutura de correlação mal especificada também são inválidos sob MAR.
- Inferência baseada em verossimilhança é válida sob perda MAR.
- Modelos multivariados com estrutura de correlação corretamente especificada também são válidos nesta situação.
- Quando os dados são MNAR, praticamente todos os métodos padrão de análise de dados longitudinais são inválidos.
- No caso MNAR é necessário modelar explicitamente a distribuição conjunta  $P(\mathbf{Y}, \mathbf{R})$ .

## Ignorabilidade: Perda na Covariável

No caso de perda na covariável podemos definir os mecanismos de não resposta de forma análoga ao discutido anteriormente.

De forma geral, quando a perda é MNAR na covariável, os métodos usuais de análise ainda retornarão inferências válidas.

O problema ocorre quando a perda é MAR dependendo da resposta. Neste caso, não modelar o mecanismo de perda retornará inferências viesadas, tanto para métodos como GEE quanto para métodos baseados e verossimilhança.

## Métodos para Tratar Dados Ausentes

Três métodos comumente usados para lidar com dados ausentes em estudos longitudinais são:

- Métodos de imputação;
- Métodos baseados em verossimilhança; e
- Métodos de ponderação.

Destes, a imputação múltipla é o método mais comumente usado.

## Imputação Múltipla: Rubin (1987)

Consiste basicamente de três passos:

- Imputação:** Para cada valor ausente são gerados  $M (M \geq 2)$  valores;
- Análise:** Cada conjunto de dados completado é analisado por métodos tradicionais para dados completos;
- Combinação:** Finalmente, os resultados das  $M$  análises são combinados numa análise final permitindo que a incerteza associada à imputação seja considerada.

Seja  $\hat{\beta}_i$  e  $\bar{U}_i$  as estimativas pontuais e de variância para o  $i$ -ésimo conjunto de dados imputado ( $i = 1, 2, \dots, M$ ).

Então a estimativa pontual para  $\beta$  das múltiplas imputações é a média das  $M$  estimativas dos dados completos:

$$\bar{\beta} = \frac{1}{M} \sum_{i=1}^M \hat{\beta}_i.$$

## Imputação Múltipla: Rubin (1987)

Seja  $\bar{U}$  a variância entre-imputações, que é a média das  $M$  estimativas de dados completos:

$$\bar{U} = \frac{1}{M} \sum_{i=1}^M \hat{U}_i,$$

e  $B$  a variância intra-imputações:

$$B = \frac{1}{M-1} \sum_{i=1}^M (\hat{\beta}_i - \bar{\beta})^2.$$

Então, a variância estimada associada com  $\bar{\beta}$  é a variância total:

$$T = \bar{U} + \left(1 + \frac{1}{M}\right) B.$$

A estatística  $(\beta - \bar{\beta}) T^{-1/2}$  é aproximadamente distribuída com distribuição  $t$  com  $v_M$  graus de liberdade, em que

$$v_M = (M-1) \left\{ 1 + \frac{\bar{U}}{(1+M^{-1})B} \right\}^2 \quad (4)$$

Na prática não mais de 10 imputações são geralmente necessárias.

## Simulação de Monte Carlo

### Desenho da Simulação

Caso homocedástico:

$$Y_{ij} = \beta_0 + \beta_1 T_j + \beta_2 G_i + \beta_3 (G_i \times T_j) + b_{0i} + \varepsilon_{ij}. \quad (5)$$

Caso heterocedástico:

$$Y_{ij} = \beta_0 + \beta_1 T_j + \beta_2 G_i + \beta_3 (G_i \times T_j) + b_{0i} + b_{1i} T_j + \varepsilon_{ij}. \quad (6)$$

- $T_j$  (Tempo) = {0, 1, 2, 3, 4}, e
- $G_i$  (Grupo) = {0, 1}, com  $P(G_i = 1) = 0.5$ .

Fixados  $\beta_0 = 25$ ,  $\beta_1 = -1$ ,  $\beta_2 = 0$  e  $\beta_3 = -1$ .

Médias populacionais:

- Grupo 0: 25, 24, 23, 22, 21; e
- Grupo 1: 25, 23, 21, 19, 17.

Componentes de Variância:

- $\varepsilon_{ij} \sim N(0, 2^2)$
- $b_i \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}; \begin{pmatrix} 4 & 0.25 \\ 0.25 & -0.10 \end{pmatrix} \right)$

A matriz de variância-covariância para o caso homocedástico foi

$$V(\mathbf{Y}) = \begin{bmatrix} 8.00 & 4.00 & 4.00 & 4.00 & 4.00 \\ 4.00 & 8.00 & 4.00 & 4.00 & 4.00 \\ 4.00 & 4.00 & 8.00 & 4.00 & 4.00 \\ 4.00 & 4.00 & 4.00 & 8.00 & 4.00 \\ 4.00 & 4.00 & 4.00 & 4.00 & 8.00 \end{bmatrix};$$

ou, em termos de correlação,

$$Cor(\mathbf{Y}) = \begin{bmatrix} 1.00 & 0.50 & 0.50 & 0.50 & 0.50 \\ 0.50 & 1.00 & 0.50 & 0.50 & 0.50 \\ 0.50 & 0.50 & 1.00 & 0.50 & 0.50 \\ 0.50 & 0.50 & 0.50 & 1.00 & 0.50 \\ 0.50 & 0.50 & 0.50 & 0.50 & 1.00 \end{bmatrix}.$$

## Desenho da Simulação

Enquanto para o caso heterocedástico tivemos

$$V(\mathbf{Y}) = \begin{bmatrix} 8.00 & 3.90 & 3.80 & 3.70 & 3.60 \\ 3.90 & 8.05 & 4.20 & 4.35 & 4.50 \\ 3.80 & 4.20 & 8.60 & 5.00 & 5.40 \\ 3.70 & 4.35 & 5.00 & 9.65 & 6.30 \\ 3.60 & 4.50 & 5.40 & 6.30 & 11.20 \end{bmatrix};$$

ou,

$$\text{Cor}(\mathbf{Y}) = \begin{bmatrix} 1.00 & 0.49 & 0.46 & 0.42 & 0.38 \\ 0.49 & 1.00 & 0.50 & 0.49 & 0.47 \\ 0.46 & 0.50 & 1.00 & 0.55 & 0.55 \\ 0.42 & 0.49 & 0.55 & 1.00 & 0.61 \\ 0.38 & 0.47 & 0.55 & 0.61 & 1.00 \end{bmatrix}.$$

## Geração da Não Resposta

- MAR: Se o valor da variável dependente foi menor que 23, então o indivíduo saía da estudo no próximo período de tempo com probabilidade de 80%.

Valores foram escolhidos de forma a produzir em média de 42% de dados ausentes.

Os modelos GEE ajustados:

- Independente (IN);
- Simetria composta (SC);
- Não estruturada (NE); e
- Auto regressiva de ordem 1 (AR).

## Modelo Normal: Caso Homocedástico

O modelo correto para análise assume variabilidade constante entre os tempos.

- Os valores são as médias de 5.000 repetições do processo de geração e perda de dados segundo o mecanismo MAR;
- O tamanho de cada banco criado foi  $n = 100$ , totalizando 500 observações;
- A imputação múltipla foi conduzida para  $M = 5$  bancos utilizando um modelo normal, pacote `norm` do R. Detalhes podem ser obtidos em Schafer (1997).

**Tabela 1:** Imputação Modelo Normal: Estimativa (erro padrão)

		$\beta_0$ (i)	$\beta_1$ (t)	$\beta_2$ (g)	$\beta_3$ (g $\times$ t)
	Valor real	25	-1	0	-1
COMP	GEE-IN	24.993 (0.353)	-0.999 (0.089)	0.010 (0.502)	-1.002 (0.126)
	GEE-SC	24.993 (0.353)	-0.999 (0.089)	0.010 (0.502)	-1.002 (0.126)
	GEE-NE	24.993 (0.351)	-0.999 (0.089)	0.008 (0.499)	-1.002 (0.127)
	GEE-AR	24.991 (0.370)	-0.998 (0.097)	0.013 (0.526)	-1.003 (0.137)
MAR	GEE-IN	24.983 (0.358)	-0.551 (0.152)	-0.027 (0.512)	-0.899 (0.254)
	GEE-SC	24.980 (0.368)	-1.065 (0.133)	0.022 (0.527)	-1.019 (0.217)
	GEE-NE	24.927 (0.363)	-0.706 (0.143)	0.011 (0.519)	-0.976 (0.235)
	GEE-AR	24.986 (0.389)	-1.294 (0.162)	0.021 (0.554)	-1.065 (0.255)
IMP	GEE-IN	24.976 (0.356)	-1.004 (0.091)	0.006 (0.505)	-0.998 (0.129)
	GEE-SC	24.976 (0.356)	-1.004 (0.091)	0.006 (0.505)	-0.998 (0.129)
	GEE-NE*	25.300 (0.348)	-1.060 (0.096)	0.095 (0.496)	-1.020 (0.135)
	GEE-AR	24.974 (0.370)	-1.010 (0.099)	0.004 (0.526)	-0.997 (0.140)

\* Por problemas de convergência os valores apresentados são a mediana

## Resultados: Caso Heterocedástico

O modelo correto para análise deveria incluir uma estrutura de covariância não constante.

- Os valores são as médias de 5.000 repetições do processo de geração e perda de dados segundo o mecanismo MAR;
- O tamanho de cada banco criado foi  $n = 500$ , totalizando 2.500 observações;
- A imputação múltipla foi conduzida para  $M = 5$  bancos utilizando um modelo normal, pacote `norm` do R.

## Modelo Normal: Caso Heterocedástico

**Tabela 2:** Imputação Modelo Normal: Estimativa (erro padrão)

		$\beta_0$ (i)	$\beta_1$ (t)	$\beta_2$ (g)	$\beta_3$ (g $\times$ t)
	Valor real	25	-1	0	-1
COMP	GEE-IN	25.001 (0.160)	-1.001 (0.051)	-0.002 (0.226)	-0.999 (0.072)
	GEE-SC	25.001 (0.160)	-1.001 (0.051)	-0.002 (0.226)	-0.999 (0.072)
	GEE-NE	25.001 (0.160)	-1.001 (0.051)	-0.003 (0.227)	-0.999 (0.074)
	GEE-AR	25.001 (0.166)	-1.001 (0.053)	-0.003 (0.236)	-0.999 (0.076)
MAR	GEE-IN	24.928 (0.162)	-0.455 (0.080)	-0.042 (0.230)	-0.884 (0.133)
	GEE-SC	24.934 (0.166)	-0.970 (0.074)	0.010 (0.237)	-1.015 (0.119)
	GEE-NE	24.902 (0.164)	-0.635 (0.076)	-0.008 (0.233)	-0.957 (0.124)
	GEE-AR	24.984 (0.175)	-1.216 (0.083)	0.007 (0.248)	-1.083 (0.128)
IMP	GEE-IN	24.986 (0.160)	-0.986 (0.050)	-0.007 (0.226)	-0.992 (0.071)
	GEE-SC	24.986 (0.160)	-0.986 (0.050)	-0.007 (0.226)	-0.992 (0.071)
	GEE-NE	25.069 (0.161)	-1.009 (0.053)	0.012 (0.229)	-0.998 (0.075)
	GEE-AR	24.994 (0.166)	-0.989 (0.053)	-0.006 (0.236)	-0.982 (0.074)

## Considerações

### Sobre o modelo GEE:

- dados ausentes podem apresentar grande impacto na estimação de quantidades de interesse;
- o impacto além do vício das estimativas também está na precisão destas;
- diferente do que ocorre com os dados completos a escolha da matriz de correlação de trabalho tem fundamental importância na estimativa final.

A imputação múltipla é uma ferramenta adequada para obtenção de estimativas não viesadas.

## CE075 Análise de Dados Longitudinais: Lista 2

**Questão 1.** Considere um estudo aleatorizado com o objetivo de comparar dois tratamentos orais (A e B) para micose de unha. Os pacientes foram aleatorizados com relação ao grau de onicólise (desprendimento da unha) no baseline (semana 0), e nas semanas 4, 8, 12, 24, 36 e 48. A variável de onicólise resposta é binária (nenhuma ou amena versus moderada ou severa) e foi avaliada em 294 pacientes compreendendo um total de 1908 medidas. O principal objetivo da análise é comparar os efeitos dos tratamentos A e B nas mudanças das probabilidades da resposta ao longo da duração do estudo. Os dados estão no objeto `toenail.dta`.

- a) Considere um **modelo marginal** para as log odds de onicólise moderada ou severa. Usando o **GEE**, ajuste um modelo que assume tendências lineares para as log odds ao longo do tempo, com intercepto comum para os dois tratamentos mas com inclinações distintas:

$$\text{logit}\{E(Y_{ij})\} = \beta_1 + \beta_2 \text{Month}_{ij} + \beta_3 \text{Treatment}_i \times \text{Month}_{ij}.$$

- b) Qual é a interpretação de  $\beta_2$  neste modelo?  
c) Qual é a interpretação de  $\beta_3$  neste modelo?  
d) Quais conclusões você tira da análise? Forneça resultados que embasem as suas conclusões.

- e) Considere agora um **modelo linear generalizado misto** com interceptos aleatórios. Usando máxima verossimilhança, ajuste um modelo com tendências lineares para as log odds no tempo e com inclinações que dependem do grupo tratamento:

$$\text{logit}\{E(Y_{ij}|b_i)\} = (\beta_1 + b_i) + \beta_2 \text{Month}_{ij} + \beta_3 \text{Treatment}_i \times \text{Month}_{ij}.$$

em que, dado  $b_i$ ,  $Y_{ij}$  segue uma distribuição de Bernoulli. Assuma que  $b_i \sim N(0, \sigma_b^2)$ .

- f) Qual é a estimativa de  $\sigma_b^2$ ? Dê uma interpretação da magnitude da variância estimada.  
g) Qual é a interpretação da estimativa de  $\beta_2$ ?  
h) Qual é a interpretação da estimativa de  $\beta_3$ ?  
i) Compare e contraste as estimativas de  $\beta_3$  dos modelos marginal e misto. Por que elas diferem?  
j) Repita a análise sequencialmente aumentando o número de pontos de quadratura usado no ajuste. Compare as estimativas e erros padrões dos parâmetros do modelo quando o número de pontos de quadratura é 2, 5, 10, 20, 30, e 50. Os resultados dependem do número de pontos de quadratura?

**Questão 2.** Análise dados de cirurgia cardíaca. Dados do prof. Antônio Luiz Ribeiro (FM, UFMG). Descrição: O paciente é submetido à cirurgia cardíaca com o auxílio da circulação extracorpórea (CEC) em que o sangue heparinizado entra em contato com superfícies estranhas (oxigenador e tubos do circuito extracorpóreo). Desta forma, apresenta-se ativação de vários sistemas orgânicos do corpo. A cirurgia cardíaca com CEC provoca alterações inflamatórias no organismo conhecidas como síndrome da resposta inflamatória sistêmica (SIRS). As respostas são as dosagens de 4 citocinas (*i*, *t*, *mc*, *mip*) (marcadores de inflamação) no sangue da artéria radial ou linha arterial sanguínea da CEC nos seguintes 6 momentos: