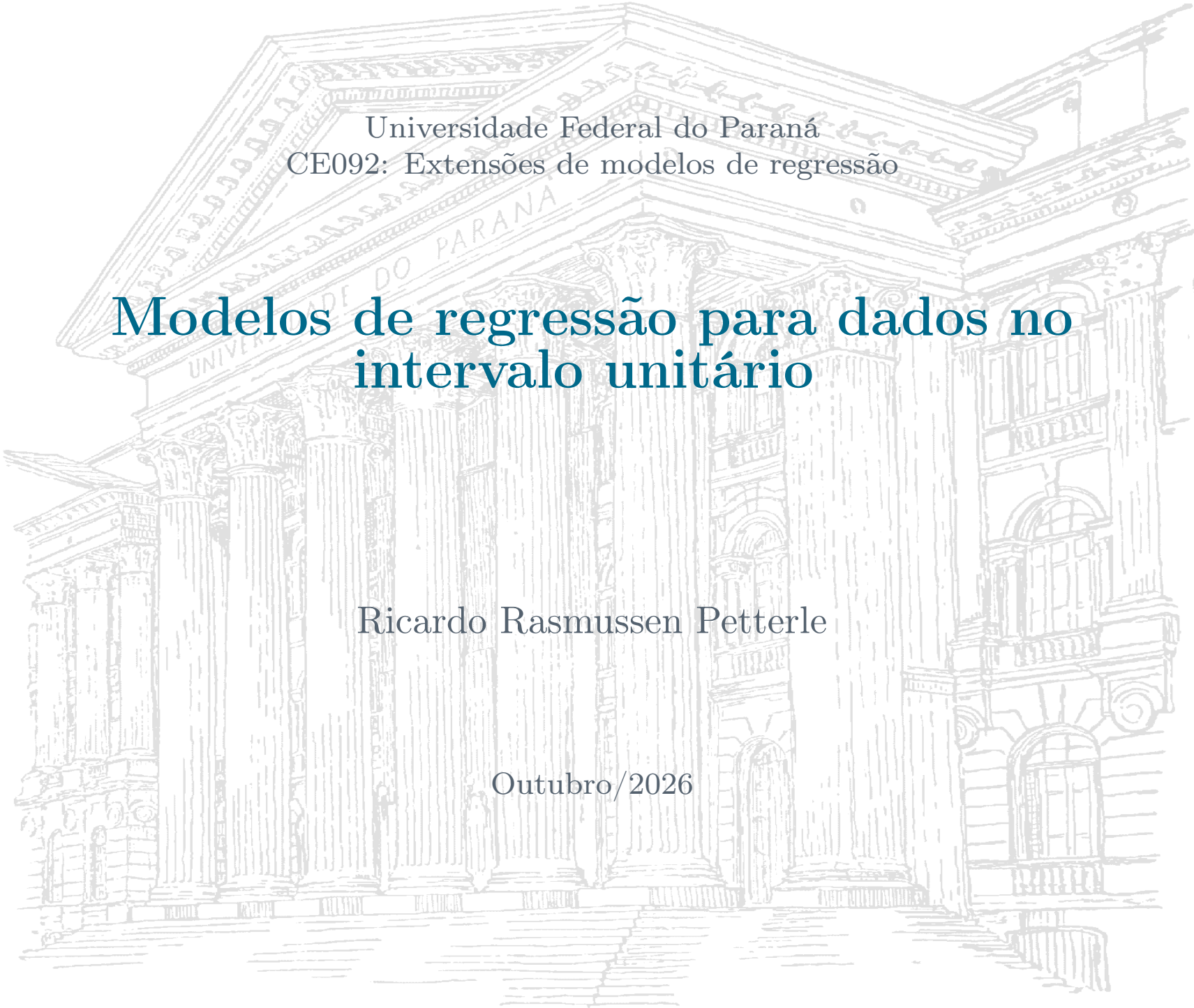


Universidade Federal do Paraná
CE092: Extensões de modelos de regressão

Modelos de regressão para dados no intervalo unitário

Ricardo Rasmussen Petterle

Outubro/2026

Retomando o exemplo da aula anterior

- No conjunto de dados do **índice de qualidade da água**, ajustamos modelos com

$$y \sim \texttt{location} + \texttt{quarter}.$$

  - Cada usina foi observada em diferentes **locais** e **trimestres**;
  - Há observações provenientes da **mesma unidade amostral**;
  - Ignoramos essa estrutura para comparar as **diferentes distribuições**.

- Podemos realmente considerar **independentes** as observações provenientes da mesma usina?

Por que precisamos de efeitos aleatórios?

- Em dados **longitudinais** ou **agrupados**, várias observações podem pertencer à mesma unidade amostral.
- Exemplos:
  - **medidas repetidas** avaliadas no mesmo paciente;
  - alunos pertencentes à mesma classe;
  - percentual de gordura corporal em diferentes regiões do corpo;
- Essa **associação** entre **observações** da mesma unidade é chamada de **dependência intraunidade**.

Estrutura dos dados

- Considere $i = 1, \ldots, n$ **unidades amostrais** e $j = 1, \ldots, n_i$ medidas avaliadas dentro da i-ésima **unidade amostral**;
- Considere y_{ij} observações da variável aleatória Y_{ij};
  - i identifica a **unidade amostral**;
  - j identifica a **observação dentro da unidade** i;
  - n_i é o número de observações realizadas na unidade i.
- Como ficam os indices no conjunto de dados do **índice de qualidade da água**?

Especificação do modelo misto

- Condicionalmente aos efeitos aleatórios, consideramos

$$Y_{ij} \mid \mathbf{u}_{\mu,i} \sim \mathcal{D}(\mu_{ij}, \phi_{ij}),$$

- Com estruturas de regressão

$$\begin{aligned} \text{logit}(\mu_{ij}) &= \mathbf{x}_{\mu,ij}^\top \boldsymbol{\beta}_\mu + \mathbf{z}_{\mu,ij}^\top \mathbf{u}_{\mu,i}, \\ \log(\phi_{ij}) &= \mathbf{x}_{\phi,ij}^\top \boldsymbol{\beta}_\phi, \\ \mathbf{u}_{\mu,i} &\sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Sigma}). \end{aligned}$$

- Note que μ_{ij} pode representar:
  - a **média**;
  - um **quantil** específico;
  - ou a **moda**.
- Para a distribuição Vasicek, utiliza-se ligação **logit** para ϕ_{ij}, pois $\phi_{ij} \in (0, 1)$.

Interpretando o preditor linear misto

- O preditor para μ_{ij} pode ser decomposto em

$$\text{logit}(\mu_{ij}) = \underbrace{\mathbf{x}_{\mu,ij}^\top \boldsymbol{\beta}_\mu}_{\text{efeitos fixos}} + \underbrace{\mathbf{z}_{\mu,ij}^\top \mathbf{u}_{\mu,i}}_{\text{efeitos aleatórios}} .$$

- No caso mais simples, com apenas um intercepto aleatório,

$$\text{logit}(\mu_{ij}) = \mathbf{x}_{\mu,ij}^\top \boldsymbol{\beta}_\mu + u_{\mu,i}, \qquad u_{\mu,i} \sim \mathcal{N}(0, \sigma^2).$$

  - Os efeitos fixos descrevem a tendência média da população.
  - $u_{\mu,i}$ representa o desvio específico da unidade i.
  - σ^2 mede a heterogeneidade entre as unidades.

Verossimilhança condicional

Seja

$$\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^\top,$$

e, para o modelo misto contínuo considerado nesta etapa,

$$\boldsymbol{\theta} = \left(\boldsymbol{\beta}_\mu^\top, \boldsymbol{\beta}_\phi^\top, \boldsymbol{\sigma}^\top\right)^\top,$$

onde $\boldsymbol{\sigma}$ contém os parâmetros da estrutura de covariância dos efeitos aleatórios.

Condicionalmente a $\mathbf{u}_{\mu,i}$,

$$\mathcal{L}_i\left(\boldsymbol{\theta}, \mathbf{u}_{\mu,i} \mid \mathbf{y}_i\right) = \prod_{j=1}^{n_i} f_c\left(y_{ij}; \mu_{ij}, \phi_{ij}, \tau\right).$$

Verossimilhança marginal

Para obter a verossimilhança marginal, integramos os efeitos aleatórios:

$$\mathcal{L}_i(\boldsymbol{\theta} \mid \mathbf{y}_i) = \int_{\mathbb{R}^q} \mathcal{L}_i\left(\boldsymbol{\theta}, \mathbf{u}_{\mu,i} \mid \mathbf{y}_i\right) f\left(\mathbf{u}_{\mu,i} \mid \boldsymbol{\Sigma}\right) \, d\mathbf{u}_{\mu,i}.$$

Assumindo independência entre as N unidades,

$$\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y}) = \prod_{i=1}^{N} \mathcal{L}_i(\boldsymbol{\theta} \mid \mathbf{y}_i).$$

Problema computacional

Em geral, a integral sobre os efeitos aleatórios não possui solução analítica.

A aproximação de Laplace

Defina a log-verossimilhança completa da unidade i por

$$\ell_{c,i}(\boldsymbol{\theta}, \mathbf{u}_{\mu,i}) = \sum_{j=1}^{n_i} \log f_c(y_{ij}; \mu_{ij}, \phi_{ij}, \tau) + \log f(\mathbf{u}_{\mu,i} \mid \boldsymbol{\Sigma}).$$

A moda condicional dos efeitos aleatórios é

$$\widehat{\mathbf{u}}_{\mu,i}(\boldsymbol{\theta}) = \arg\max_{\mathbf{u}_{\mu,i}} \ell_{c,i}(\boldsymbol{\theta}, \mathbf{u}_{\mu,i}).$$

A aproximação de Laplace usa uma expansão quadrática de $\ell_{c,i}$ em torno desse modo.

Aproximação da log-verossimilhança marginal

Se

$$\mathbf{H}_i(\boldsymbol{\theta}) = -\left.\frac{\partial^2 \ell_{c,i}(\boldsymbol{\theta}, \mathbf{u}_{\mu,i})}{\partial \mathbf{u}_{\mu,i} \partial \mathbf{u}_{\mu,i}^\top}\right|_{\mathbf{u}_{\mu,i}=\widehat{\mathbf{u}}_{\mu,i}},$$

então

$$\ell_i^*(\boldsymbol{\theta}) = \ell_{c,i}\left(\boldsymbol{\theta}, \widehat{\mathbf{u}}_{\mu,i}\right) + \frac{q}{2}\log(2\pi) - \frac{1}{2}\log\left|\mathbf{H}_i(\boldsymbol{\theta})\right|.$$

Logo,

$$\ell^*(\boldsymbol{\theta} \mid \mathbf{y}) = \sum_{i=1}^{N} \ell_i^*(\boldsymbol{\theta}), \qquad \widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell^*(\boldsymbol{\theta} \mid \mathbf{y}).$$

Como isso é feito no `unitregTMB`?

A implementação utiliza o `Template Model Builder (TMB)`.

1. A log-verossimilhança completa é especificada em `C++`.
2. Para cada valor de θ, o TMB determina internamente $\widehat{\mathbf{u}}_{\mu,i}$.
3. Os efeitos aleatórios são integrados por **aproximação de Laplace**.
4. As derivadas da função objetivo são obtidas por **diferenciação automática**.
5. A log-verossimilhança marginal aproximada é maximizada em relação a θ.

Erros-padrão e transformações não lineares são obtidos com `TMB::sdreport()`.

Do modelo contínuo ao modelo aumentado

Nesta primeira parte, usamos o componente contínuo

$$f_c(y_{ij}; \mu_{ij}, \phi_{ij}, \tau), \qquad 0 < y_{ij} < 1.$$

Quando houver zeros e/ou uns exatos, o mesmo framework é estendido substituindo f_c pela distribuição aumentada

$$f_{\text{aug}}(y_{ij}; \mu_{ij}, \phi_{ij}, p_{0ij}, p_{1ij}, \tau).$$

Nesse caso, o vetor de parâmetros passa a incluir também

$$\beta_{p_0} \qquad \text{e} \qquad \beta_{p_1}.$$

Exemplo 4: Voltando ao exemplo do IQA

Para o conjunto `wqi`, podemos incluir um intercepto aleatório para a usina:

$$\text{logit}(\mu_{ij}) = \mathbf{x}_{\mu,ij}^{\top}\boldsymbol{\beta}_{\mu} + u_{\mu,i}, \qquad u_{\mu,i} \sim \mathcal{N}(0, \sigma^2).$$

Em R:

```
fit_ug_mixed <- unitregTMB(
  y ~ location + quarter + (1 | id),
  family = unitgamma(model_for = "mean"),
  data = wqi
)
```

Agora a heterogeneidade entre usinas é incorporada diretamente ao modelo.

Exemplo 5: Percentual de gordura corporal

- Estudo realizado no Serviço de Endocrinologia e Metabologia do HC-UFPR.
- 298 indivíduos $\times$ 5 medidas repetidas = **1490 observações**.
- **Variável resposta**: percentual de gordura corporal medido em cinco regiões.

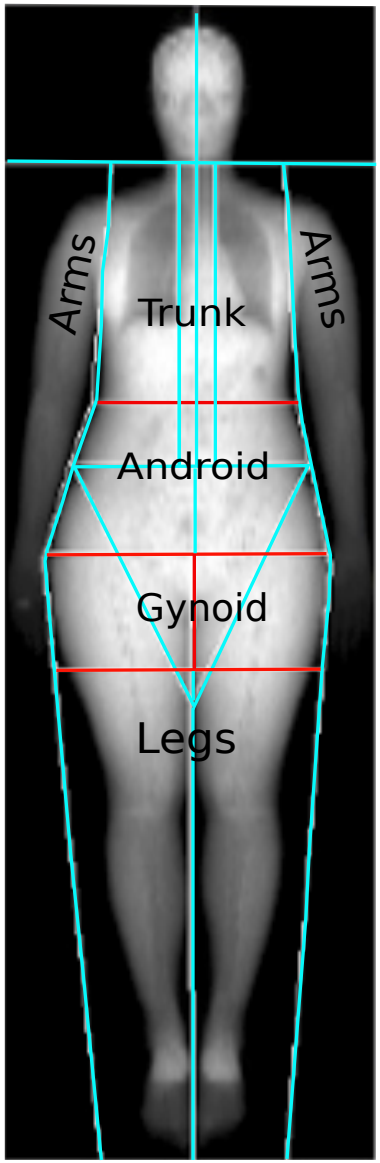
Regiões corporais

1. *Braços*
2. Pernas
3. Tronco
4. Androide
5. Ginoide

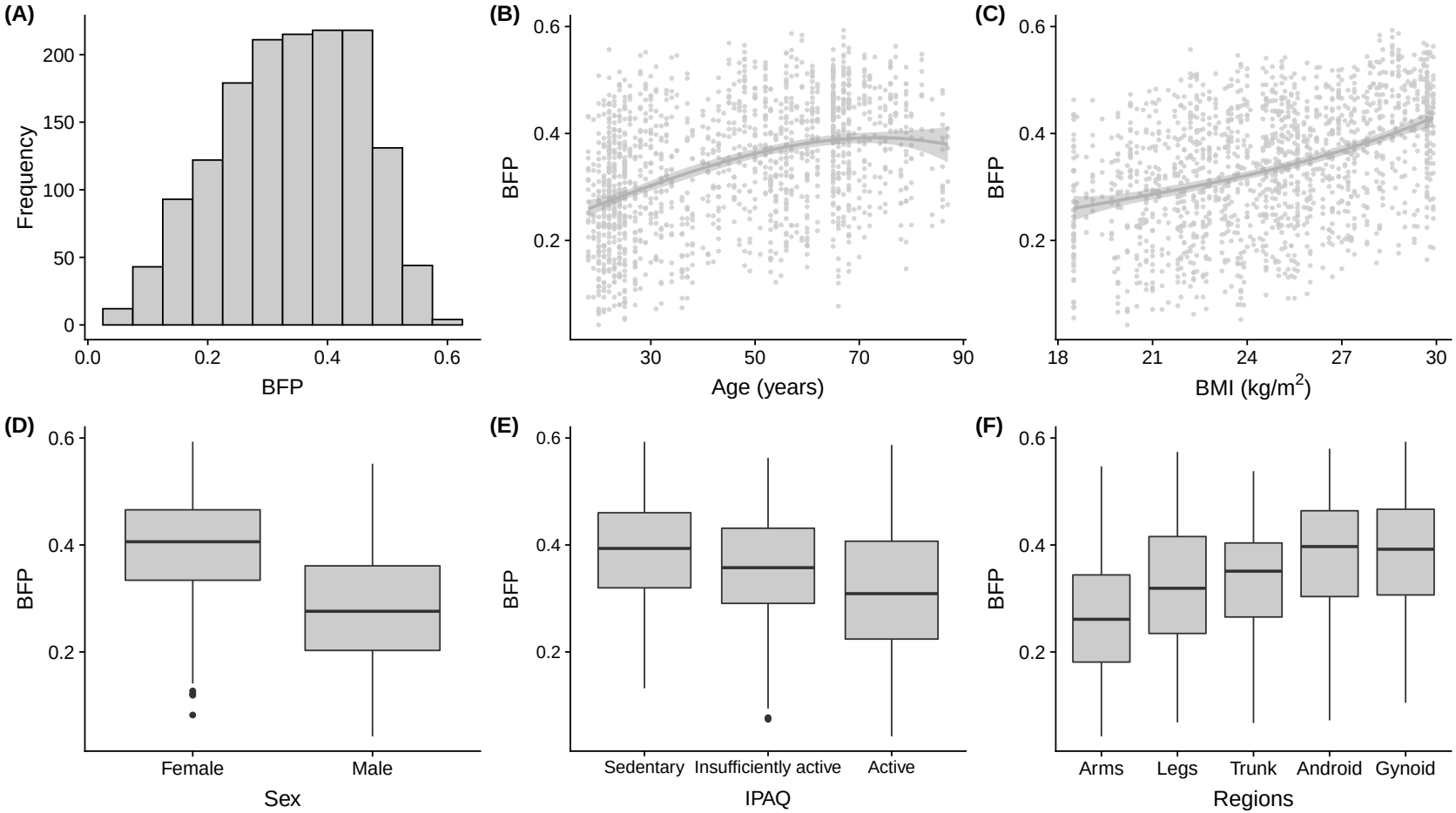
Covariáveis

1. Idade (anos)
2. IMC (kg/m^2)
3. Sexo (*feminino*, masculino)
4. IPAQ (*sedentário*, insuficientemente ativo, ativo)

Exemplo 5: Percentual de gordura corporal



Exemplo 5: Percentual de gordura corporal



Exemplo 5: Percentual de gordura corporal

- Seja $Y_{ij} \mid u_{i1}, u_{i2} \sim \mathrm{UG}(\mu_{ij}, \phi_{ij})$ o percentual de gordura corporal para $i = 1, \ldots, 298$ **indivíduos** e $j = 1, \ldots, 5$ **regiões corporais**.

- **Estrutura da média**:

$$\begin{aligned} \text{logit}(\mu_{ij}) = \beta_0 &+ \beta_1 \, \texttt{idade}_i + \beta_2 \, \texttt{IMC}_i + \beta_3 \, \texttt{sexo}_i \\ &+ \beta_4 \, \texttt{IPAQ}_i + \beta_{5j} \, \texttt{região}_j + u_{i1} \end{aligned} \tag{1}$$

- **Estrutura da precisão**:

$$\log(\phi_{ij}) = \gamma_0 + \gamma_1 \, \texttt{IMC}_i + \gamma_{2j} \, \texttt{região}_j + u_{i2}$$

- **Efeitos aleatórios**:

$$u_{i1} \sim \mathcal{N}(0, \tau_1^2), \qquad u_{i2} \sim \mathcal{N}(0, \tau_2^2)$$

Exemplo 5: Percentual de gordura corporal

Parâmetro	Unit-gama[1]	Beta[1]	Unit-gama[2]	Beta[2]
	Est(EP)	Est(EP)	Est(EP)	Est(EP)
β_0: Intercepto	−2.806(0.122)*	−2.815(0.122)*	−2.179(0.127)*	−2.187(0.127)*
β_1: Idade	0.002(0.001)*	0.002(0.001)*	−0.003(0.010)	−0.003(0.010)
β_2: IMC	0.084(0.005)*	0.084(0.005)*	0.066(0.006)*	0.066(0.006)*
β_3: Sexo-masculino	−0.644(0.030)*	−0.650(0.030)*	−0.757(0.030)*	−0.760(0.030)*
β_{42}: IPAQ-IA	−0.087(0.043)*	−0.088(0.043)*	−0.065(0.044)	−0.065(0.044)
β_{43}: IPAQ-A	−0.179(0.042)*	−0.182(0.042)*	−0.158(0.042)*	−0.159(0.042)*
β_{52}: Região-pernas	0.288(0.019)*	0.291(0.019)*	0.279(0.017)*	0.277(0.016)*
β_{53}: Região-tronco	0.336(0.019)*	0.344(0.018)*	0.342(0.023)*	0.358(0.023)*
β_{54}: Região-androide	0.523(0.018)*	0.537(0.018)*	0.538(0.025)*	0.558(0.025)*
β_{55}: Região-ginoide	0.570(0.018)*	0.580(0.018)*	0.557(0.016)*	0.562(0.016)*
ϕ	56.716(2.330)*	96.136(3.945)*	–	–
τ_1	0.226(0.011)*	0.226(0.011)*	0.244(0.011)*	0.244(0.011)*
γ_0: Intercepto	–	–	2.730(0.413)*	3.152(0.418)*
γ_1: IMC	–	–	0.048(0.016)*	0.052(0.016)*
γ_{22}: Região-pernas	–	–	1.382(0.131)*	1.448(0.132)*
γ_{23}: Região-tronco	–	–	−0.509(0.122)*	−0.497(0.124)*
γ_{24}: Região-androide	–	–	−0.936(0.123)*	−0.875(0.125)*
γ_{25}: Região-ginoide	–	–	2.801(0.420)*	2.838(0.397)*
τ_2	–	–	0.173(0.104)	0.186(0.099)
LogLik	**2208.64**	2198.67	**2305.82**	2303.41
AIC	**−4393.28**	−4373.34	**−4575.64**	−4570.82
BIC	**−4329.60**	−4309.66	**−4480.12**	−4475.30

Nota: EP = erro-padrão; * indica valor-p < 0.05; [1] modelos com precisão constante; [2] modelos com precisão variável.

Exemplo 5: Percentual de gordura corporal

- A inclusão de uma estrutura de **precisão variável** melhorou substancialmente o ajuste do modelo:
  - `LogLik` : 2208.64 $\rightarrow$ **2305.82**.
- **Modelo com precisão variável (melhor ajuste):**
  - Efeitos não significativos: `Idade` (β_1), `IPAQ-IA` (β_{42}) e o efeito aleatório associado à precisão (τ_2).
  - Estrutura da precisão:

Maior precisão ($\downarrow$ variância)	**Menor precisão** ($\uparrow$ variância)
`IMC` ($\gamma_1 = 0.05$)	`tronco` ($\gamma_{23} = -0.51$)
`pernas` ($\gamma_{22} = 1.38$)	`androide` ($\gamma_{24} = -0.94$)
`ginoide` ($\gamma_{25} = 2.80$)	

Exemplo 7: Doenca periodontal

- Estudo clínico realizado para avaliar a condição da **doença periodontal** em indivíduos com diabetes tipo 2.
- Foram avaliados **seis sítios** em cada um dos **28 dentes** (**oito** molares, **oito** pré-molares, **quatro** caninos e **oito** incisivos).
- Nível de inserção clínica (NIC ≥ 3 mm).
- A proporção de sítios dentários com doença é dada por:

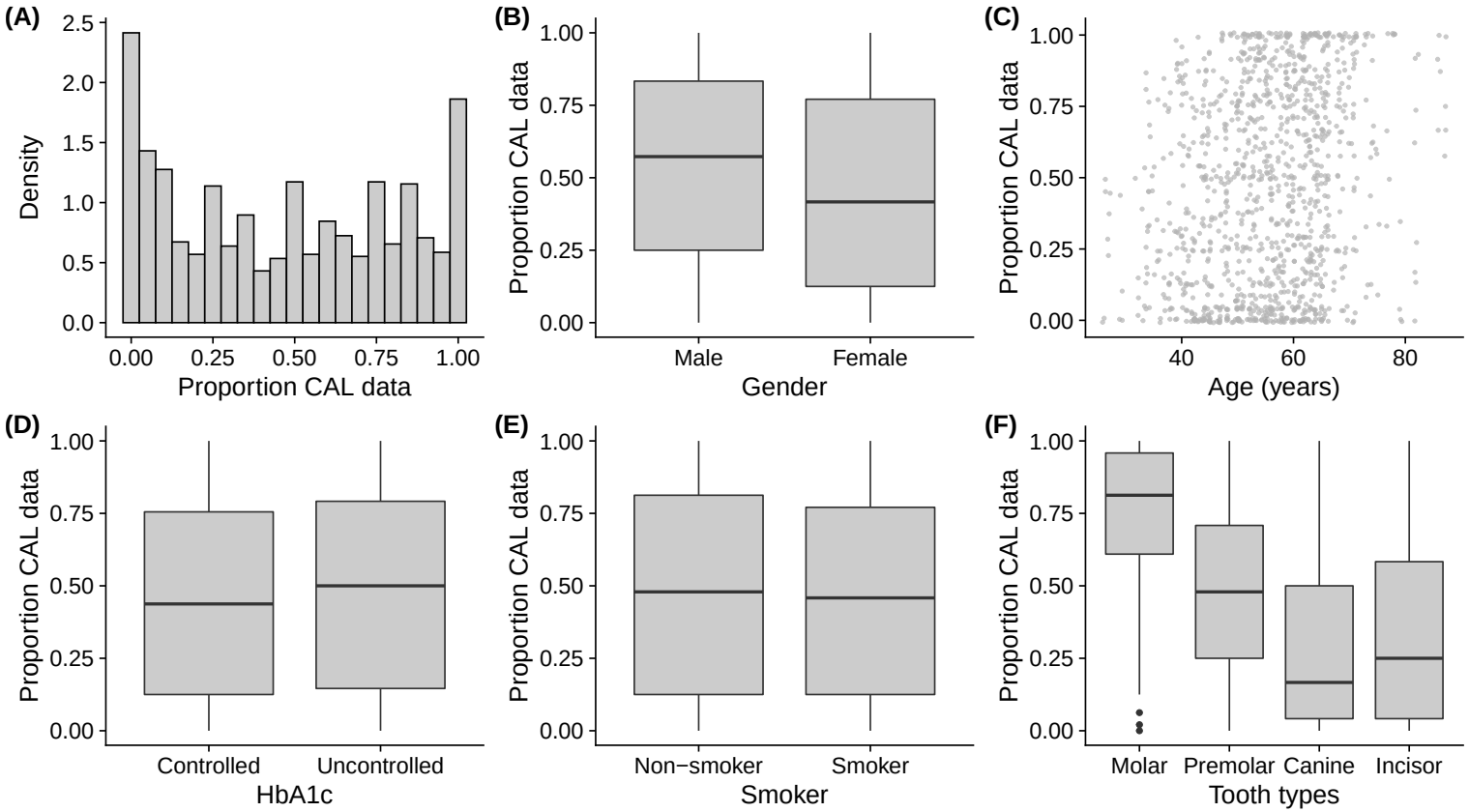
$$\frac{\text{número de sítios com doença}}{\text{número total de sítios}}.$$

- Por exemplo, para **molares** e **caninos**, existem **48** e **24** sítios, respectivamente.
- A **variável resposta** assume valores no intervalo $[0, 1]$.

Exemplo 7: Doenca periodontal

- O conjunto de dados contém 1160 observações (290 indivíduos $\times$ 4 tipos de dentes).
- **Objetivo principal**:
  - Investigar o efeito das **covariáveis** sobre os **quantis** da variável resposta, considerando a dependência entre observações de um mesmo indivíduo.
- Os **quatro tipos de dentes** correspondem às **medidas agrupadas** por indivíduo.
- **Covariáveis**:
  1. `Sexo`: (*masculino*, feminino).
  2. `Idade`: idade do paciente (em anos).
  3. `HbA1c`: (*controlado*, $< 7\%$, não controlado, $\geq 7\%$).
  4. `Tabagismo`: (*não fumante*, fumante).

Exemplo 7: Doenca periodontal



- Os valores zero e um representam, respectivamente, **ausência de doença** $(114, 9.8\%)$ e **comprometimento de todos os sítios** $(94, 8.1\%)$.

Exemplo 7: Doenca periodontal

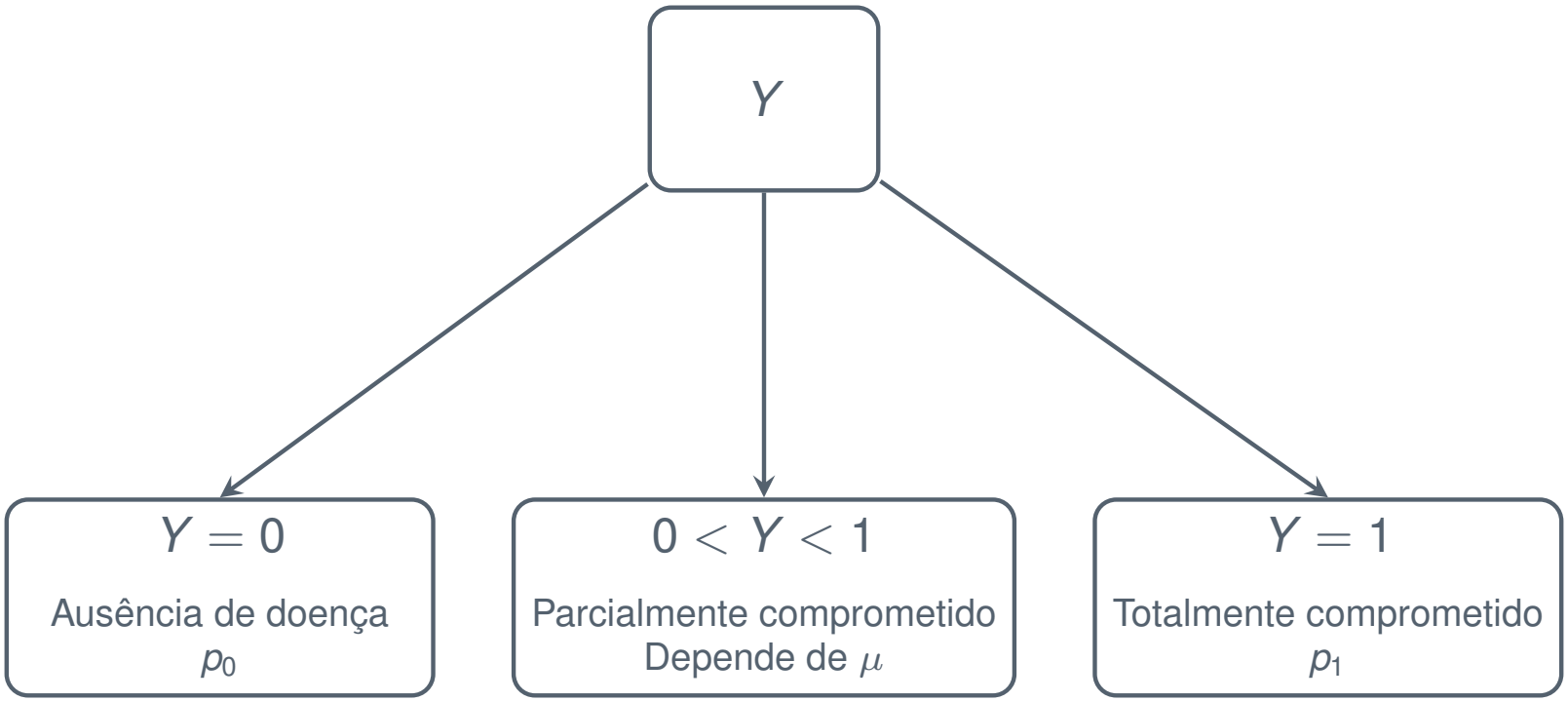
- Seja Y_{ij} a proporção de sítios dentários com doença para $i = 1, \ldots, 290$ **indivíduos** e $j = 1, \ldots, 4$ **tipos de dentes**.
- O modelo hierárquico é especificado por:

$$
\begin{aligned}
Y_{ij}|\mathbf{u}_i &\sim \text{ZOAUW}(\mu_{ij}, \phi, p_{0ij}, p_{1ij}, \tau) \\
\text{logit}(\mu_{ij}) &= \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{u}_i, \\
\text{logit}(p_{0ij}) &= \mathbf{x}_{ij}^\top \boldsymbol{\gamma}, \\
\text{logit}(p_{1ij}) &= \mathbf{x}_{ij}^\top \boldsymbol{\delta}.
\end{aligned}
$$

- $\mathbf{u}_i \sim \mathcal{N}(0, \sigma^2)$.
- $\mathbf{x}_{ij}^\top = (1, \texttt{sexo}_i, \texttt{idade}_i, \texttt{HbA1c}_i, \texttt{fumante}_i, \texttt{molar}_{ij}, \texttt{pré-molar}_{ij}, \texttt{incisivo}_{ij})$.

Exemplo 7: Doenca periodontal

- Estrutura do modelo para a variável resposta Y.



Exemplo 7: Doenca periodontal

- **Configuração geral**
  - **Níveis quantílicos:** $\tau = 0.1,\ 0.25,\ 0.5,\ 0.75,\ 0.9$.
  - **Modelos mistos quantílicos aumentados:**
    1. Unit-Weibull
    2. Kumaraswamy

- **Estruturas de regressão**
  - **Modelo 1:**
    - $\mu_{ij} \sim$ `sexo` + `idade` + `HbA1c` + `fumante` + `dente` + (1|indivíduo)
    - ϕ, p_0 e $p_1 \sim$ constante
  - **Modelo 2:**
    - $\mu_{ij} \sim$ `sexo` + `idade` + `HbA1c` + `fumante` + `dente` + (1|indivíduo)
    - p_{0ij} e $p_{1ij} \sim$ todas as covariáveis
    - $\phi \sim$ constante

Exemplo 7: Doenca periodontal

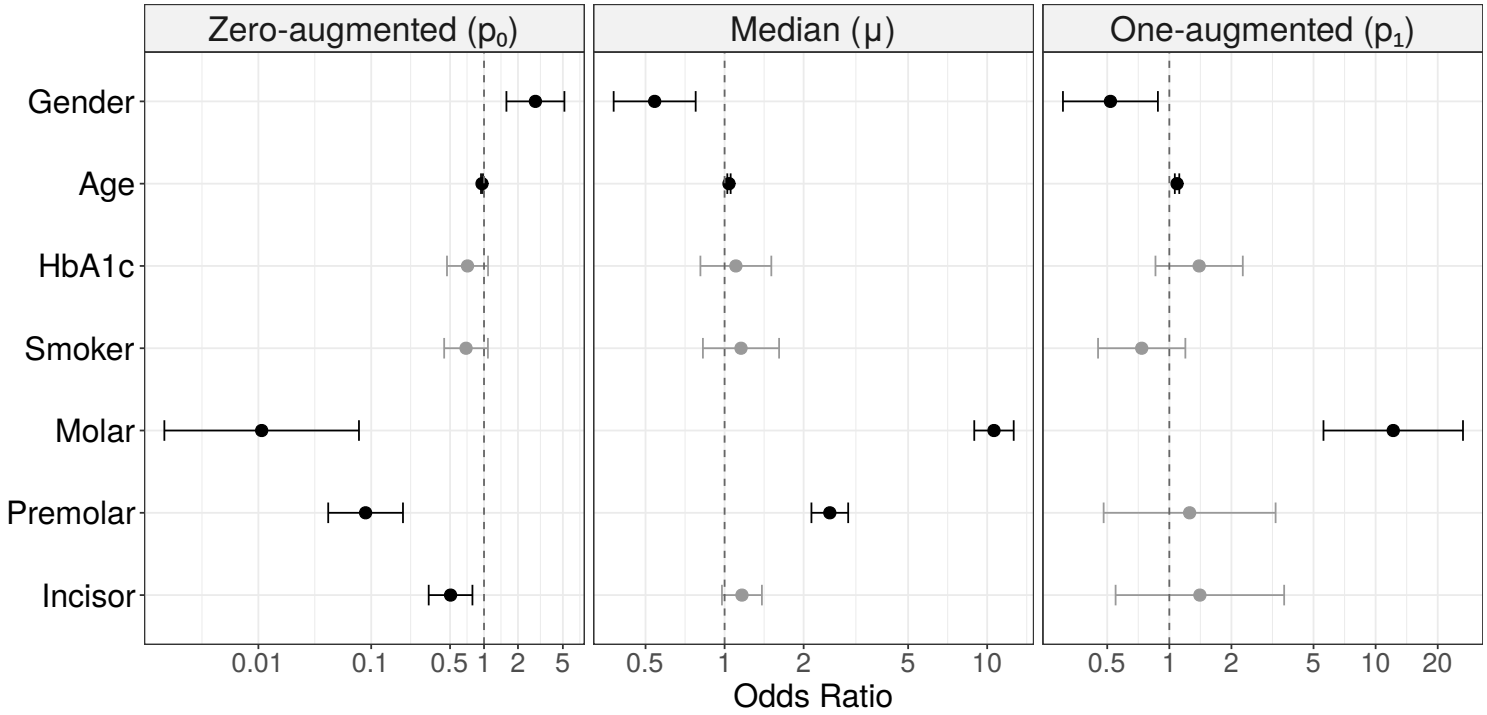
Seleção de modelos:

- Critério de Informação de Akaike (AIC).

Modelo	Distribuição	Nível quantílico				
		0.1	0.25	0.5	0.75	0.9
1	Unit-Weibull	794.75	779.16	769.17	768.89	776.55
	Kumaraswamy	970.77	948.77	925.89	912.78	912.67
2	Unit-Weibull	**539.13**	**523.54**	**513.55**	**513.27**	**520.93**
	Kumaraswamy	715.15	693.15	670.27	657.16	657.05

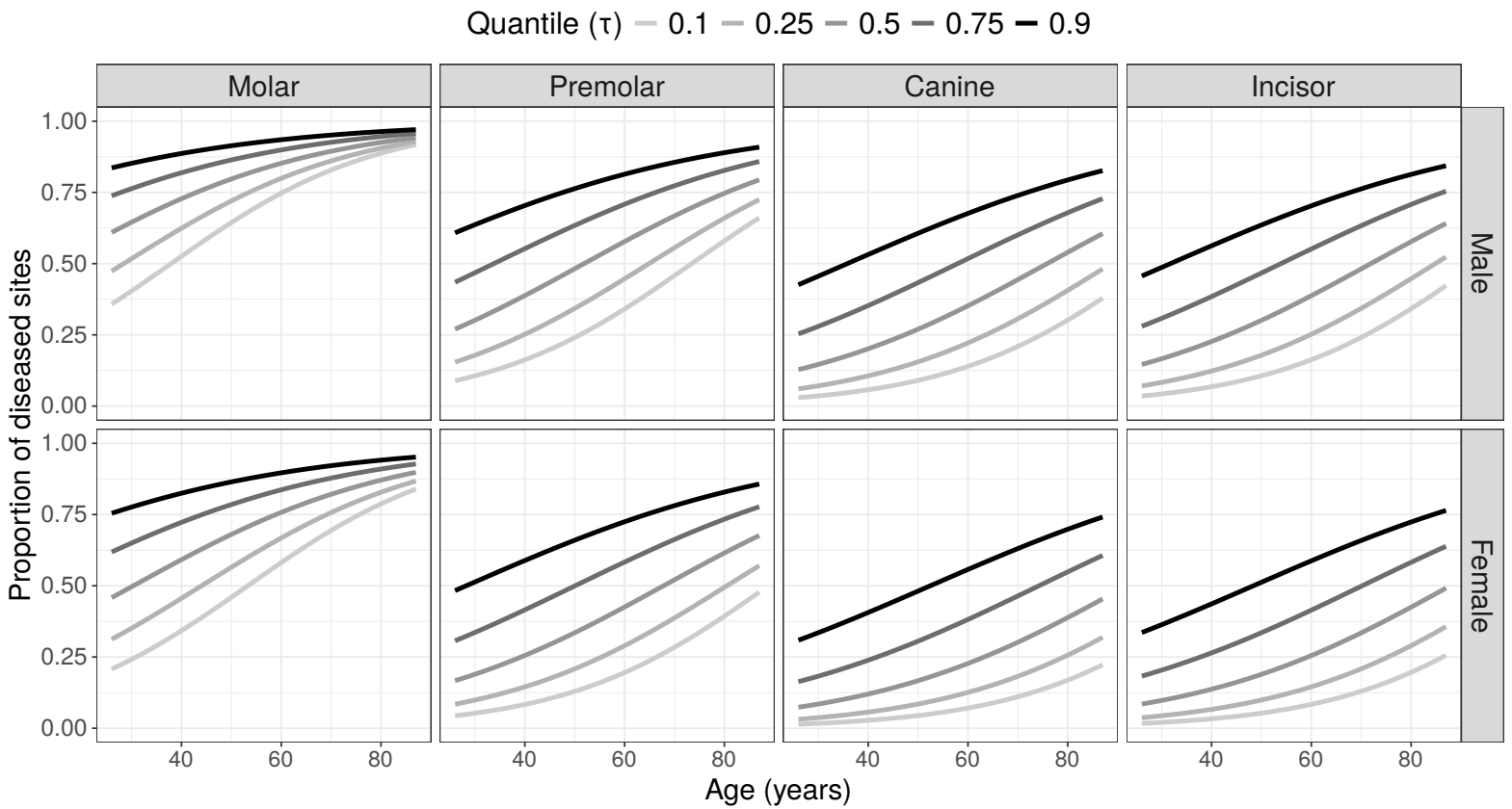
- O modelo misto ZOAUW (**Modelo 2**) apresentou o **melhor ajuste** para todos os níveis quantílicos.

Exemplo 7: Doenca periodontal



- As **mulheres** apresentam chances **2,86 vezes maiores** de estarem **livres de doença periodontal** em comparação aos **homens**.
- Os **molares** apresentam chances medianas aproximadamente **10,6 vezes maiores** do que os **caninos** (categoria de referência).

Exemplo 7: Doenca periodontal



Contato

- Nome: Ricardo Rasmussen Petterle.
- Estatístico.
- Setor de Ciências da Saúde.
- Universidade Federal do Paraná.
- E-mail: ricardopetterle@ufpr.br

- Obrigado!