Métodos computacionais para inferência estatística

20º SINAPE

Modelo Beta Misto

Paulo Justiniano Ribeiro Jr

Wagner Hugo Bonat

Elias Teixeira Krainski

Walmes Marques Zeviani

LEG:Laboratório de Estatística e Geoinformação / *UFPR* / Brasil

http://www.leg.ufpr.br
e-mail:paulojus@ufpr.br

20º SINAPE
20-21 de Julho 2012

Modelo BetaMisto e Motivação

- Motivação: modelos de regressão para respostas em $(0, 1)$
  - IQVT
  - IQA
  - outras potenciais

- Efeitos aleatórios:
  - variabilidade extra
  - estruturas de dependência (grupos, med. repetidas, longitudinais, espaciais, hieráquicos)

Regressão Beta

- Modelos Beta para obs independentes (Paolino, 2001 ; Kieschnick, 2003 ; Ferrari, 2004)
- *a la* GLM
- Extensões:
  - Simas, 2010 modela média e precisão por covariáveis,
  - preditores não-lineares.
  - avaliação do ajuste (verossimilhança) (Espinheira, 2008, 2008a; Rocha, 2010)
  - correções de viés (Vasconcellos, 2005; Ospina, 2006 e Simas, 2010)
  - Bayesiana (Branscum, 2007) em dados de distância genética entre vírus
  - pacote **betareg** (Cribari-Neto, 2010)
  - em séries temporais (McKenzie,2007; Grunwald, 1993; Rocha, 2010)
  - Da-Silva, 2011: modelo beta Bayesiano dinâmico para séries temporais (taxa de desemprego mensal Brasileira)

Objetivo

- Modelos de regressão beta com efeitos aleatórios:
  - superdispersão,
  - medidas repetidas,
  - estrutura longitudinal,
  - sub-parcelas,
  - dentre outras.
- Estimabilidade
- Comparação de modelos e inferência sobre parâmetros
- Métodos numéricos (Laplace) e computacionalmente intensivos (*data-cloning*)
- implementação computacional (avaliar e disponibilizar métodos e algorítimos)
- aplicações

Objetivo

- Modelos de regressão beta com efeitos aleatórios:
  - superdispersão,
  - medidas repetidas,
  - estrutura longitudinal,
  - sub-parcelas,
  - dentre outras.
- Estimabilidade
- Comparação de modelos e inferência sobre parâmetros
- Métodos numéricos (Laplace) e computacionalmente intensivos (*data-cloning*)
- implementação computacional (avaliar e disponibilizar métodos e algorítimos)
- aplicações

Objetivo

- Modelos de regressão beta com efeitos aleatórios:
  - superdispersão,
  - medidas repetidas,
  - estrutura longitudinal,
  - sub-parcelas,
  - dentre outras.
- Estimabilidade
- Comparação de modelos e inferência sobre parâmetros
- Métodos numéricos (Laplace) e computacionalmente intensivos (*data-cloning*)
- implementação computacional (avaliar e disponibilizar métodos e algorítimos)
- aplicações

Objetivo

- Modelos de regressão beta com efeitos aleatórios:
  - superdispersão,
  - medidas repetidas,
  - estrutura longitudinal,
  - sub-parcelas,
  - dentre outras.
- Estimabilidade
- Comparação de modelos e inferência sobre parâmetros
- Métodos numéricos (Laplace) e computacionalmente intensivos (*data-cloning*)
- implementação computacional (avaliar e disponibilizar métodos e algorítimos)
- aplicações

Objetivo

- Modelos de regressão beta com efeitos aleatórios:
  - superdispersão,
  - medidas repetidas,
  - estrutura longitudinal,
  - sub-parcelas,
  - dentre outras.
- Estimabilidade
- Comparação de modelos e inferência sobre parâmetros
- Métodos numéricos (Laplace) e computacionalmente intensivos (*data-cloning*)
- implementação computacional (avaliar e disponibilizar métodos e algorítimos)
- aplicações

Objetivo

- Modelos de regressão beta com efeitos aleatórios:
  - superdispersão,
  - medidas repetidas,
  - estrutura longitudinal,
  - sub-parcelas,
  - dentre outras.
- Estimabilidade
- Comparação de modelos e inferência sobre parâmetros
- Métodos numéricos (Laplace) e computacionalmente intensivos (*data-cloning*)
- implementação computacional (avaliar e disponibilizar métodos e algorítimos)
- aplicações

Beta

- $Y \sim B(\mu, \phi)$, parametrização em termos de média e precisão:

$$f(y|\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-y)\phi-1}, \quad 0 < y < 1, \tag{1}$$

  - $0 < \mu < 1$, $\phi > 0$ (precisão) e $\Gamma(.)$ é a função gama.
  - $E(Y) = \mu$ e $V(Y) = \frac{\mu(1-\mu)}{(1+\phi)}$.

- Regressão Beta:

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} = \eta_i \tag{2}$$

  - parâmetros $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$ $(k \times 1)$
  - covariáveis $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})^T$
  - η_i é o preditor linear, ligação *logit* $g(.) : (0,1) \to \Re$; $g(\mu) = \log \mu/(1-\mu)$,
  - outras possíveis ligações: *probit*, a complemento log-log e Cauchy (Cribari-Neto,2010).

Beta misto

- Modelo com efeitos aleatórios
- (motivação: parcimônia, estruturas de dependência, etc)

$$f_i(y_{ij}|\mathbf{b_i}, \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu_{ij}\phi)\Gamma((1-\mu_{ij})\phi)} y_{ij}^{\mu_{ij}\phi-1}(1-y_{ij})^{(1-y_{ij})\phi-1} \quad (3)$$

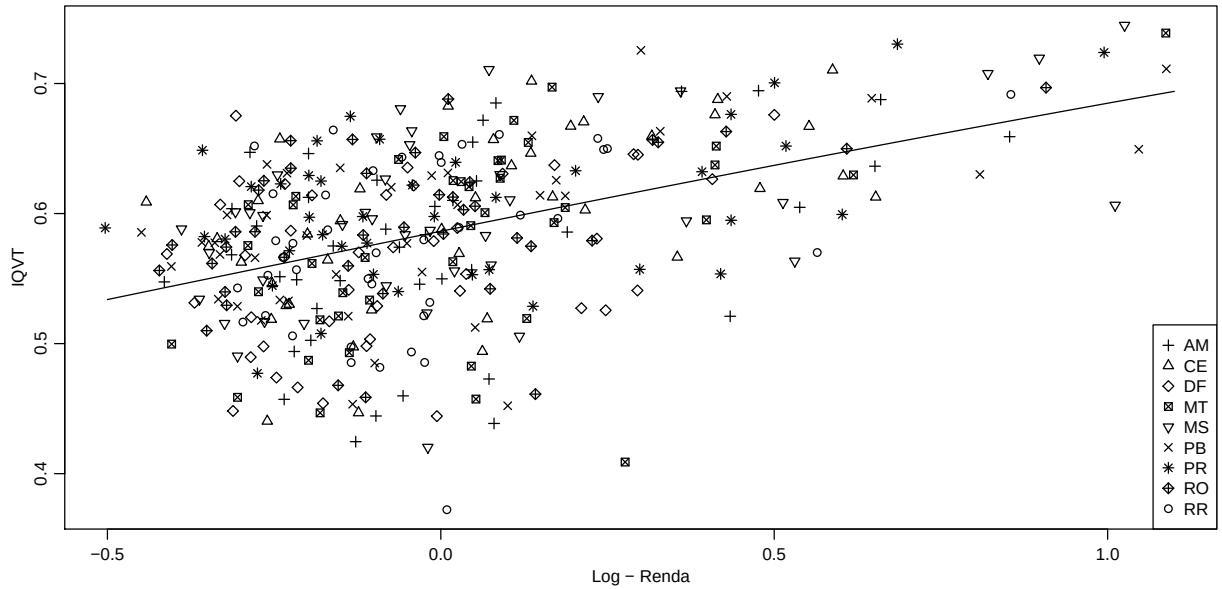
  - $g(\mu_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b_i}$
  - $g(.)$, $\mathbf{x}_{ij}$, $\boldsymbol{\beta}$ como anterior
  - $\mathbf{z}_{ij}$ vetor de covariáveis (ef. al.) de dimensão q
  - $f(\mathbf{b_i}|\Sigma) \sim N(\mathbf{0}, \Sigma)$

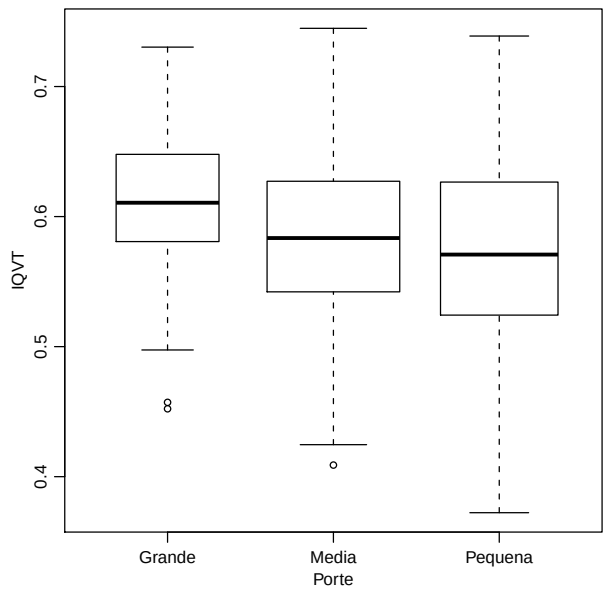
- Independência condicional em $\mathbf{b_i}$

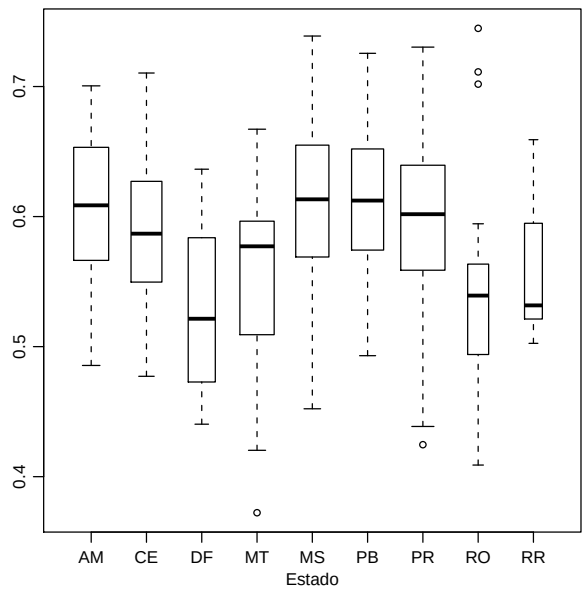
IQVT

- 25 indicadores de 8 áreas temáticas:
  habitação, saúde, educação, saúde integral e segurança no trabalho, desenvolvimento de competências, atribuição de valor ao trabalho e orientação a paetivipação e desempenho,
- Pesquisa SESI por amostragem
- 8 estados + DF
- Respostas: Índice por indústria
- (outras informações : aspectos de qualidade de vida, gastos com benefícios sociais, etc)
- 2 covariáveis
  - renda média (log, centrada)
  - porte: grande (500+), média (100-500) e pequena (20 a 99)

Descritiva


IQVT
0.4
0.5
0.6
0.7
−0.5
0.0
0.5
1.0
Log − Renda
AM
CE
DF
MT
MS
PB
PR
RO
RR


IQVT
0.4
0.5
0.6
0.7
Grande
Media
Pequena
Porte


0.4
0.5
0.6
0.7
AM
CE
DF
MT
MS
PB
PR
RO
RR
Estado

O Modelo

- $Y_{ij} \sim B(\mu_{ij}, \phi)$;
- $g(\mu_{ij}) = (\beta_0 + b_{i1}) + \beta_1 Media + \beta_2 Pequena + (\beta_3 + b_{i2}) Renda$;
- $b_{ij} \sim NMV(\mathbf{0}, \Sigma)$ onde $\Sigma = \begin{bmatrix} 1/\tau_1^2 & \rho \\ \rho & 1/\tau_2^2 \end{bmatrix}$ para $j = 1, 2$.
- Estado como EA
- Ajustes de modelos e submodelos

Verossimilhança (Marginal)

- $$f_i(\mathbf{y}_i|\boldsymbol{\beta}, \Sigma, \phi) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\mathbf{b}_i, \boldsymbol{\beta}, \phi) f(\mathbf{b}_i|\Sigma) d\mathbf{b}_i, \tag{4}$$

  segue que a verossimilhança para $\boldsymbol{\beta}$, Σ e ϕ é dada por

$$L(\boldsymbol{\beta}, \Sigma, \phi) = \prod_{i=1}^{N} f_i(\mathbf{y}_i|\boldsymbol{\beta}, \Sigma, \phi) \tag{5}$$

- Estimação por métodos numéricos envolve integrações
- N integrais, $N \times q$ integrais

Métodos Numéricos e computacionais

- Inferência - I
  - Soluções analíticas
  - Soluções por aproximações numéricas
  - Laplace, Quadraturas (Gaussiana, adaptativas, etc)
  - Soluções por algorítmos de simulação Monte-Carlo
  - Monte Carlo, Quasi Monte-Carlo
  - Simulações via cadeias de Markov
  - MCMC (verossimilhança ou Bayesiana
  - Data-cloning
- Inferência - II
  - Aproximações quadráticas (Hessiano ou data-cloning) I_E
  - Verossimilhanças perfilhadas I_O
  - Crítico para veriâncias/precisão dos ef. aleatórios

Data Cloning

- observações y_{ij} com $i = 1, \ldots, N$ blocos e $j = 1, \ldots, n_i$ repetições em cada bloco, são clonadas K – *vezes* por blocos, ou seja, N blocos passam a ser $N \times K$ blocos.
- Dados clonados y_{ij}^K
- verossimilhança clonada $L(\boldsymbol{\beta}, \Sigma, \phi)^K$
- mesmo máximo
- K vezes a matriz de informação de Fisher

Algoritmo DC

- $$\pi_K(\boldsymbol{\beta}, \Sigma, \phi | y_{ij}) = \frac{[\int f_i(\mathbf{y}_i|\boldsymbol{\beta}, \Sigma, \phi) f(\mathbf{b}_i|\Sigma) d\mathbf{b}_i]^K \pi(\boldsymbol{\beta})\pi(\Sigma)\pi(\phi)}{C(K; y_{ij})} \quad (6)$$

  onde

  $$C(K; y_{ij}) = \int [\int f_i(\mathbf{y}_i|\boldsymbol{\beta}, \Sigma, \phi) f(\mathbf{b}_i|\Sigma) d\mathbf{b}_i]^K \pi(\boldsymbol{\beta})\pi(\Sigma)\pi(\phi) d\boldsymbol{\beta} d\Sigma d\phi \quad (7)$$

- priori torna-se irrelavante

Predição

- $$f_i(\mathbf{b}_i|\mathbf{y}_i, \beta, \Sigma, \phi) = \frac{f_i(\mathbf{y}_i|\mathbf{b}_i, \beta, \phi) f(\mathbf{b}_i|\Sigma)}{\int f_i(\mathbf{y}_i|\mathbf{b}_i, \beta, \phi) f(\mathbf{b}_i|\Sigma) d\mathbf{b}_i} \quad (8)$$

- Bayes Empírico

Ajustes

Tabela: Estimativas pontuais, logaritmo da verossimilhança maximizada e critério de informação de Akaike.

	Model.1	Model.2	Model.3	Model.4	Model.5
b0	0.3479	0.4451	0.4338	0.3962	0.3965
b1		-0.1050	-0.0878	-0.0723	-0.0724
b2		-0.1608	-0.1443	-0.1326	-0.1329
b3			0.4184	0.4703	0.4697
Phi	53.9700	56.7966	72.8577	94.1938	94.1905
Tau.U				62.3648	62.3464
Tau.V					51480.4778
Rho					0.8509
logLik	463.9274	473.2354	518.6716	553.5231	553.5252
AIC	-923.8548	-938.4708	-1027.3433	-1095.0462	-1091.0504

VM e DC

Tabela: Estimativas pontuais e desvio padrão via Verossimilhança Marginal e dClone.

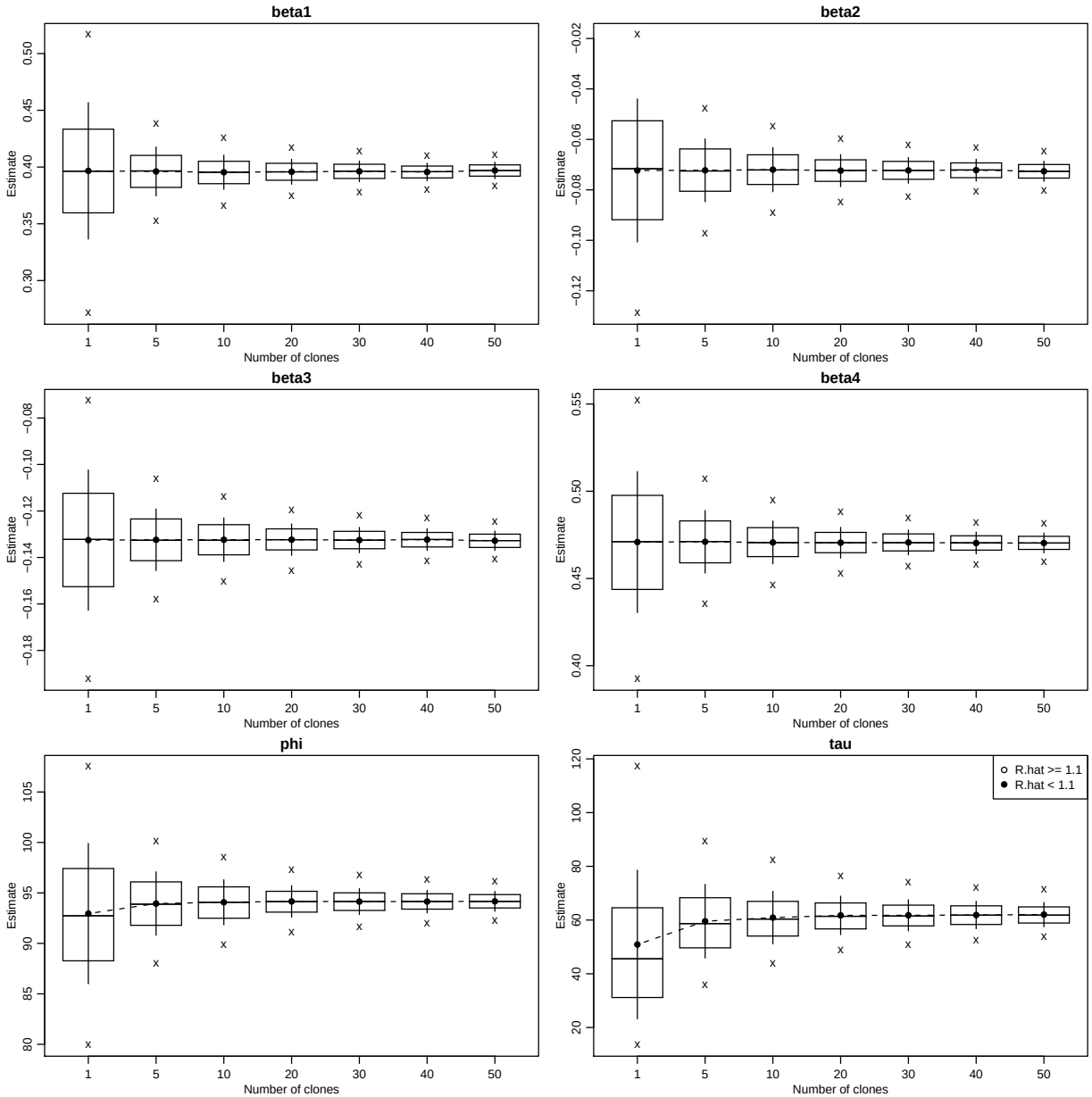
	Pt.Marginal	SD.Marginal	Pt.dclone	SD.dclone
b1	0.3962	0.0474	0.3970	0.0512
b2	-0.0723	0.0269	-0.0726	0.0283
b3	-0.1326	0.0288	-0.1328	0.0296
b4	0.4703	0.0393	0.4704	0.0402
phi	94.1938	7.0256	94.1683	6.9767
tau.U	62.3648	31.8706	62.0308	32.0805

ICs

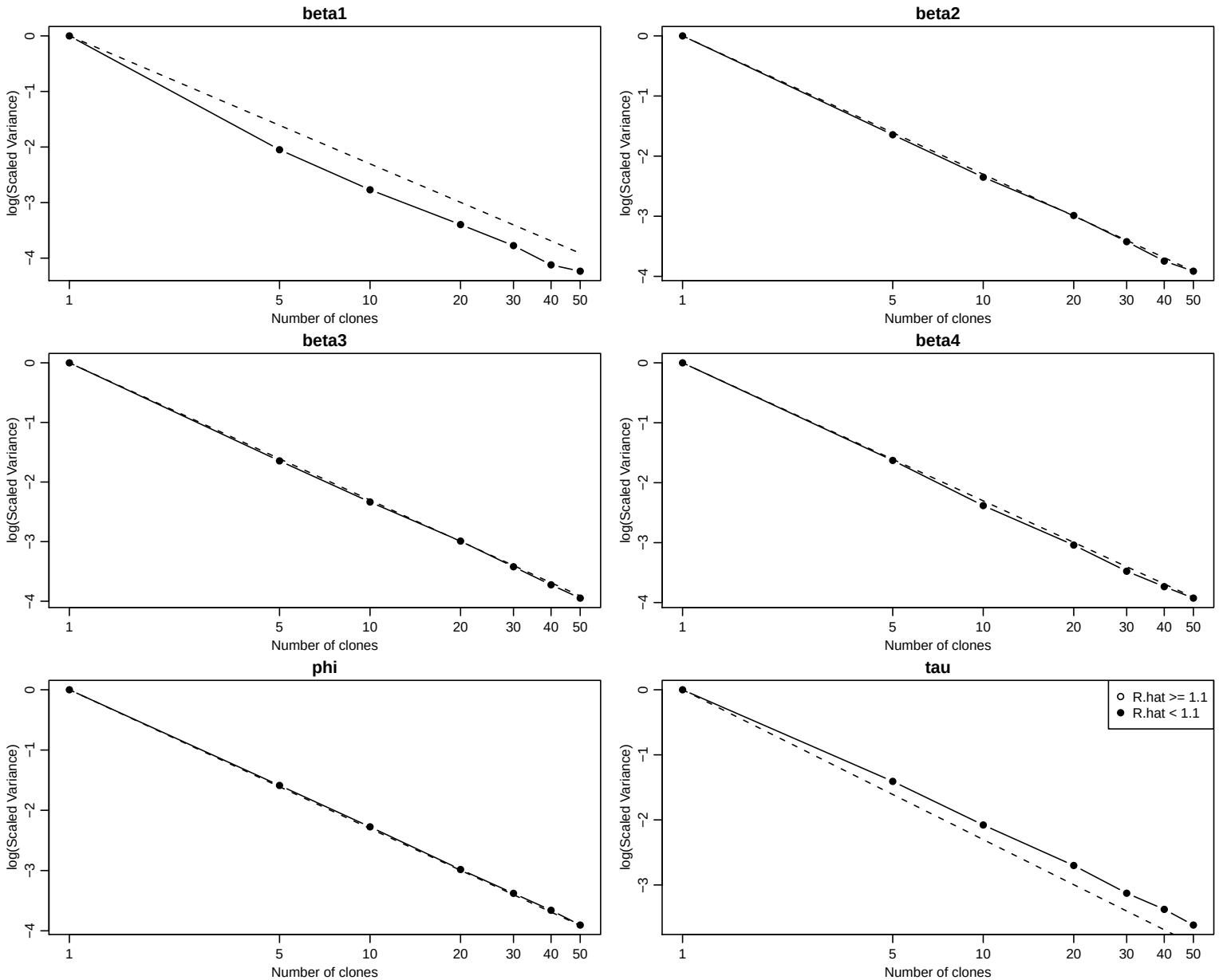
Tabela: Intervalos de confiança assintótico e baseado em perfil de verossimilhança.

	2.5 %	97.5 %	2.5 %	97.5 %
beta1	0.2967	0.4973	0.2918	0.4978
beta2	-0.1281	-0.0171	-0.1275	-0.0172
beta3	-0.1909	-0.0747	-0.1910	-0.0741
beta4	0.3916	0.5491	0.3931	0.5480
phi	80.4943	107.8424	81.0877	108.6460
tau	-0.8458	124.9074	19.7383	156.4794

Estimabilidade I



Estimabilidade II



Comentários

- 1, 5, 10, 20, 30, 40 e 50 clones
- 3 cadeias com 5000 amostras retidas ao final
- Efeito de prioris
- precisão: priori Gama$(0.1, 0.001)$
- diagnóstico de estimabilidade : a taxa de queda da variância é $\frac{1}{k}$

Predição

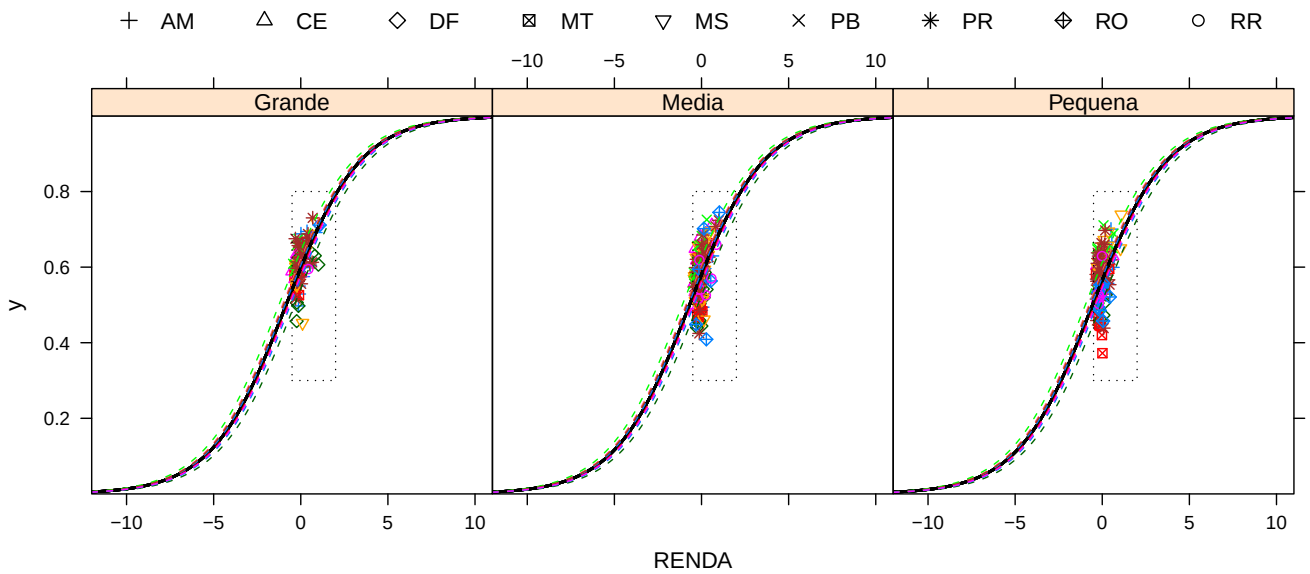
Tabela: Valores preditos por estado, porte e renda. Entre parenteses diferença em percentual em relação a média.

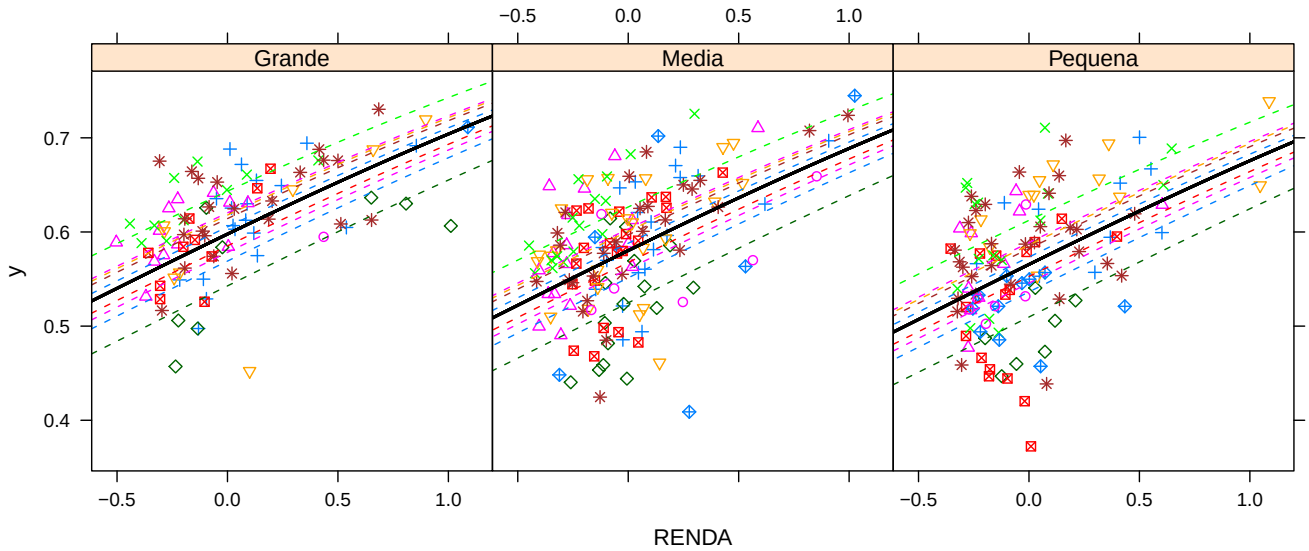
ESTADO	R$ 500,00		
	Grande	Média	Pequena
AM	52.91 (1.52)	51.11 (1.58)	49.6 (1.63)
CE	54.48 (4.52)	52.68 (4.7)	51.17 (4.85)
DF	46.5 (-10.77)	44.71 (-11.13)	43.23 (-11.43)
MT	50.82 (-2.49)	49.01 (-2.58)	47.51 (-2.65)
MS	54.22 (4.04)	52.42 (4.2)	50.92 (4.33)
PB	56.91 (9.2)	55.13 (9.58)	53.64 (9.9)
PR	53.83 (3.29)	52.03 (3.42)	50.52 (3.52)
RO	49.17 (-5.66)	47.36 (-5.86)	45.86 (-6.03)
RR	50.11 (-3.85)	48.31 (-3.99)	46.8 (-4.1)
ESTADO	R$ 2.500,00		
	Grande	Média	Pequena
AM	70.55 (0.95)	69.02 (1)	67.72 (1.04)
CE	71.84 (2.8)	70.35 (2.95)	69.08 (3.07)
DF	64.95 (-7.06)	63.29 (-7.39)	61.88 (-7.68)
MT	68.78 (-1.58)	67.21 (-1.66)	65.87 (-1.73)
MS	71.63 (2.51)	70.14 (2.64)	68.86 (2.75)
PB	73.79 (5.6)	72.37 (5.9)	71.15 (6.16)
PR	71.31 (2.04)	69.81 (2.15)	68.52 (2.24)
RO	67.34 (-3.64)	65.73 (-3.82)	64.36 (-3.97)

Interpretações

- Acima da média: PB, MS, PR, AM, CE
- destaque positivo: PB 9.9% maior que a média nacional para pequeno porte e renda baixa
- destaque negativo: DF até 11.43% menor que a média geral.
- Diferenças/efeitos dimimuem com aumento da renda (porte como o estado perdem importância)
- baixa renda: políticas de apoio
- mais alta renda: menos dependente dos benefícios, renda principal mantenedora de qualidade

Predições


AM
CE
DF
MT
MS
PB
PR
RO
RR
Grande
Media
Pequena
y
RENDA


Grande
Media
Pequena
y
RENDA

IQA

Questão central: avaliar impacto da UH na qualidade da água (licenciamento)
Dados:

- posição: Montante, Reservatório e Jusante
- tempo: coletas trimestrais
- usina: 16 usinas

Questões:

- sem casualização, replicações
  Abordagem: estudo observacional - modelo
- combinações dos fatores: muitos parâmetros
  Abordagem: Efeitos aleatórios

IQA

Questão central: avaliar impacto da UH na qualidade da água (licenciamento)
Dados:

- posição: Montante, Reservatório e Jusante
- tempo: coletas trimestrais
- usina: 16 usinas

Questões:

- sem casualização, replicações
  Abordagem: estudo observacional - modelo
- combinações dos fatores: muitos parâmetros
  Abordagem: Efeitos aleatórios

Modelo

$$Y_{ijt} \sim \text{Beta}(\mu_{ijt}, \phi)$$

$$g(\mu_{ijt}) = \beta_0 + \beta_{1,i} + \beta_{2,t} + b_j + b_{j,t}$$

$$b_j \sim N(0, \tau_U^2)$$

$$b_{jt} \sim N(0, \tau_t^2)$$

Comentários

- Modelo 6: dificuldades com integração numérica (MC, GH) devido a dimensão
- Laplace também apresetna problemas incluindo tempo computacional
- Data-clone é mais flexível
- Estimativas pontuais próximas mas diferenças nos erros-padrão (assimetria?)
- uso de perfil de verossimilhança
- pseudo verossimilhanças

- $M \longrightarrow R : 5,39\%$
- $M \longrightarrow J : 3.55\%$
- sugere comportamento cíclico
- estrutura temporal? tempo contínuo?
- combinar diferentes anos

Comentários

- Modelo 6: dificuldades com integração numérica (MC, GH) devido a dimensão
- Laplace também apresetna problemas incluindo tempo computacional
- Data-clone é mais flexível
- Estimativas pontuais próximas mas diferenças nos erros-padrão (assimetria?)
- uso de perfil de verossimilhança
- pseudo verossimilhanças

- $M \longrightarrow R : 5,39\%$
- $M \longrightarrow J : 3.55\%$
- sugere comportamento cíclico
- estrutura temporal? tempo contínuo?
- combinar diferentes anos