Inferência em Modelos Hierárquicos Generalizados sob Planos Amostrais Informativos

Romy Rodriguez Ravines
(romy@dme.ufrj.br)

Instituto de Matemática
Universidade Federal do Rio de Janeiro

Dissertação de Mestrado
Março, 2003

Orientador: Fernando Moura (fmoura@im.ufrj.br)

Objetivo

Implementar e aplicar a proposta de Pfeffermann, Moura e Silva [Multilevel Modelling Newsletter, v.14, n.1(2002): 8-17], sobre o uso das distribuições amostrais na realização de inferência sobre modelos de superpopulação hierárquicos para dados não normais na presença de desenhos amostrais informativos.

Palavras-Chave

- Distribuição Amostral *(Sampling Model)*
- Modelo de Superpopulação
- Desenho Amostral Informativo

Distribuição Amostral (D.A.)

- Segundo Pfeffermann, Krieger e Rinott [*Statistica Sinica* **8**(1998): 1087-1114], sempre é possível aproximar a distribuição paramétrica dos dados de uma amostra
- Supor que na população $y_i \sim f_p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})$ e usar o Teorema de Bayes para obter a Distribuição (marginal) Amostral de y_i:

$$f_s(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{Pr(I_i = 1 \mid y_i, \boldsymbol{x}_i, \boldsymbol{\phi}) f_p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})}{Pr(I_i = 1 \mid \boldsymbol{x}_i, \boldsymbol{\phi})}$$

  onde $I_i = 1$ indica que o elemento $i \in s$.
- A distribuição amostral é um caso especial da familia de distribuições ponderadas [Bayarri & DeGroot, 1994]:

$$p(y \mid \theta) = \frac{w(y) g(y \mid \theta)}{E_\theta[w(y)]}$$

- Tem-se que

$$Pr(I_i = 1 \mid y_i, \boldsymbol{x}_i, \boldsymbol{\phi}) = \int Pr(I_i = 1 \mid y_i, \boldsymbol{x}_i, \boldsymbol{\phi}, \pi_i) f_p(\pi_i \mid y_i, \boldsymbol{x}_i, \boldsymbol{\phi}) d\pi_i$$
$$= \quad E_p[\pi_i \mid y_i, \boldsymbol{x}_i, \boldsymbol{\phi}]$$

- Então,

$$f_s(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{E_p[\pi_i \mid y_i, \boldsymbol{x}_i, \boldsymbol{\phi}] f_p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})}{E[\pi_i \mid \boldsymbol{x}_i, \boldsymbol{\phi}]} \tag{1}$$

- Por exemplo, seja a fdp Gama com parâmetro de forma α e média μ_i tal que

$$f_p(y_i) \propto y_i^{\alpha-1} \exp(-\alpha y_i / \mu_i),$$

e seja $E_p(\pi_i \mid y_i) \propto y_i$. Então,

$$f_s(y_i) \propto y_i^{(\alpha+1)-1} \exp(-\alpha y_i / \mu_i)$$

Distribuição Amostral na Família Exponencial

- **Proposição**

  Se a fdp da população de y_i é

  $$f_p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_i) = a_i(\boldsymbol{\theta}_i) \exp\left[\sum_{k=1}^{K} \theta_{ki} b_{ki}(y_i) + c_i(y_i)\right]$$

  e as probabilidades de inclusão na amostra obedecem

  $$E_p[\pi_i \mid y_i, \boldsymbol{x}_i] = r_i \exp\left[\sum_{k=1}^{K} d_{ki} b_{ki}(y_i)\right]$$

  então a fdp da amostra pertence também à família exponencial com parâmetros $\theta^*_{ki} = \theta_{ki} + d_{ki}$.

- Se $\theta_{ki} = (\phi_{0k} + \boldsymbol{x}_i'\boldsymbol{\phi}_k)$ e $d_{ki} = (\psi_{0k} + \boldsymbol{x}_i'\boldsymbol{\psi}_k)$ então a fdp amostral pertence à mesma familia com ϕ_{0k} e $\boldsymbol{\phi}_k$ substituídas por $(\phi_{0k} + \psi_{0k})$ e $(\boldsymbol{\phi}_k + \boldsymbol{\psi}_k)$ respectivamente.

Distribuição Amostral nos Modelos Hierárquicos

- Assume-se que o efeito do plano amostral é independente em cada nível. Consequentemente, a equação (1) é utilizada em forma independente em cada nível.
- Devem ser conhecidos os valores esperados das probabilidades de seleção dos elementos em cada nível da hierarquia: $I_i, I_{j|i}, I_{z|j,i} \ldots$
- Pfeffermann, Moura e Silva (2002) desenvolveram a aplicação das Distribuições Amostrais para o Modelo Linear Normal Hierárquico.

D.A. nos Modelos Hierárquicos Generalizados

- Por exemplo, um modelo de superpopulação de dois níveis:

$$\begin{aligned} y_{ij} &\sim FamExp(\theta_{ij}), && j = 1, \ldots, n_i; \\ \eta_{ij} &= g(\theta_{ij}) = \beta_{0i} + \boldsymbol{x}'_{ij}\boldsymbol{\beta}; \\ \beta_{0i} &\sim Normal(\gamma_0 + \boldsymbol{z}'_i\boldsymbol{\gamma}, \sigma^2), && i = 1, \ldots, n. \end{aligned}$$

- Desenho informativo no primeiro nível: $E[\pi_{j|i}] = g_1(y_{ij}) = \exp[A_1 y_{ij} + h_1(\boldsymbol{x}_i)]$
- Desenho informativo no segundo nível: $E[\pi_i] = g_2(\beta_{0i}) = \exp[A_2\beta_{0i} + h_2(\boldsymbol{z}_i)]$
- Com um desenho informativo nos dois níveis tem-se:

$$\begin{aligned} f_s(y_{ij} \mid \beta_{0i}, \boldsymbol{\beta}, \boldsymbol{x}_{ij}, A_1, h_1) &\propto E[\pi_{j|i}]\, f_p(y_{ij} \mid \beta_{0i}, \boldsymbol{\beta}, \boldsymbol{x}_{ij}), && j = 1, \ldots, n_i; \\ \eta_{ij} &= g(\beta_{0i}, \boldsymbol{x}_{ij}, \boldsymbol{\beta}, A_1, h_1); \\ f_s(\beta_{0i} \mid \gamma_0, \boldsymbol{\gamma}, \boldsymbol{z}_i, \sigma^2, A_2, h_2) &\propto E[\pi_i]\, f_p(\beta_{0i} \mid \gamma_0, \boldsymbol{\gamma}, \boldsymbol{z}_i, \sigma^2), && i = 1, \ldots, n. \end{aligned}$$

Simulação

- Modelo de superpopulação:

$$
\begin{aligned}
(y_{ij} \mid \theta_{ij}) &\sim \text{Bernoulli}(\theta_{ij}) \\
\text{logit}\,(\theta_{ij}) &= \beta_{0i} + \boldsymbol{x}'_{ij}\boldsymbol{\beta} \\
(\beta_{0i} \mid \boldsymbol{z}_i, \boldsymbol{\gamma}, \sigma^2_\mu) &\sim N(\boldsymbol{z}'_i\boldsymbol{\gamma}, \sigma^2_\mu)
\end{aligned}
$$

  onde $i =$ escola e $j =$ aluno.

- Dados gerados: 500 populações e 2000 amostras (4 desenhos amostrais diferentes).
- Cada amostra foi utilizada para ajustar três modelos diferentes:
  - ignorando o desenho amostral (IG)
  - usando as distribuições amostrais (SM)
  - incorporando as variáveis do desenho (DV).

- **Geração das Populações**:

Intercepto da Escola: $\beta_{0i} = \gamma_0 + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \mu_i,$
$\mu_i \sim N(0, \sigma^2_\mu)$

Tamanho da Escola: $\log M_i = \alpha_0 + \alpha_1 \beta_{0i} + \varsigma_i$
$\varsigma_i \sim N(0, \sigma^2_M)$

Resposta do Aluno: $y_{ij} \sim \text{Bernoulli}(\theta_{ij})$
$\text{logit}\,(\theta_{ij}) = \beta_{0j} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij}$

Estrato do Aluno: $p_{ij} = \eta_0 + \eta_1 y_{ij} + \zeta_{ij};$
$\zeta_{ij} \sim N(0, \sigma^2_p)$

$$O_{ij} = \begin{cases} 1 & \text{se} \quad p_{ij} < 1,76, \\ 2 & \text{se} \quad 1,76 \leq p_{ij} < 1,97, \\ 3 & \text{se} \quad p_{ij} \geq 1,97. \end{cases}$$

- **Seleção das amostras**:

Tabela 1: Desenhos Amostrais Utilizados

		Seleção de Escolas	
		Aleatória Simples (AAS)	Proporcional ao Tamanho (PPT)
Seleção de Alunos	Aleatória Simples (AAS)	**AAS-AAS**	**PPT-AAS**
	Estratificada (EST)	**AAS-EST**	**PPT-EST**

Tabela 2: Classificação dos Desenhos Amostrais

	Desenho Não Informativo	**Desenho Informativo**
Escolas	AAS	PPT
Alunos	AAS	EST

- **Distribuição Amostral de β_{0i}**

$$f_s(\beta_{0i} \mid \boldsymbol{z}_i, \boldsymbol{\gamma}, \sigma^2_\mu) = \frac{E_p(\pi_i \mid \beta_{0i}, \boldsymbol{z}_i, \boldsymbol{\gamma}, \sigma^2_\mu) f_p(\beta_{0i} \mid \boldsymbol{z}_i, \boldsymbol{\gamma}, \sigma^2_\mu)}{E_p(\pi_i \mid \boldsymbol{z}_i, \boldsymbol{\gamma}, \sigma^2_\mu)}$$

$$f_s(\beta_{0i} \mid \boldsymbol{z}_i, \boldsymbol{\gamma}, \sigma^2_\mu) = \frac{\exp\left[\alpha_0 + \alpha_1 \beta_{0i} + \sigma^2_M/2\right] f_p(\beta_{0i} \mid \boldsymbol{z}_i, \boldsymbol{\gamma}, \sigma^2_\mu)}{\exp[\alpha_0 + \alpha_1 \boldsymbol{z}_i \boldsymbol{\alpha} + (\alpha_1^2 \sigma^2_\mu + \sigma^2_M)/2]}$$

$$= \frac{1}{\sqrt{2\pi}\sigma_\mu} \exp\left[\alpha_1 \beta_{0i} + \frac{\sigma^2_M}{2} - \frac{(\beta_{0i} - \boldsymbol{z}_i'\boldsymbol{\gamma})^2}{2\sigma^2_\mu}\right]$$

$$= \frac{1}{\sqrt{2\pi}\sigma_\mu} \exp\left[-\frac{1}{2\sigma^2_\mu}\left(\beta_{0i} - \boldsymbol{z}_i'\boldsymbol{\gamma} - \alpha_1 \sigma^2_\mu\right)^2\right]$$

Logo, a distribuição amostral de β_{0i} é

$$(\beta_{0i} \mid \boldsymbol{z}_i, \boldsymbol{\gamma}, \sigma^2_\mu) \sim N(\boldsymbol{z}_i'\boldsymbol{\gamma} + \alpha_1 \sigma^2_\mu, \sigma^2_\mu)$$

- **Distribuição Amostral de y_{ij}**

$$f_s(y_{ij} \mid \boldsymbol{x}_{ij}, \beta_{0i}, \boldsymbol{\beta}) = \frac{E_p(\pi_{j|i} \mid y_{ij}, \boldsymbol{x}_{ij}, \beta_{0i}, \boldsymbol{\beta}) f_p(y_{ij} \mid \boldsymbol{x}_{ij}, \beta_{0i}, \boldsymbol{\beta})}{E_p(\pi_{j|i} \mid \boldsymbol{x}_{ij}, \beta_{0i}, \boldsymbol{\beta})}$$

$$\begin{aligned}
E_p(\pi_{j|i} \mid y_{ij}, \boldsymbol{x}_{ij}, \beta_{0i}, \boldsymbol{\beta}) &= (q_1^i - q_2^i)\Phi(\delta_1 - \delta_2 y_{ij}) + (q_2^i - q_3^i)\Phi(\delta_3 - \delta_2 y_{ij}) + q_3^i, \\
E_p(\pi_{j|i} \mid \boldsymbol{x}_{ij}, \beta_{0i}, \boldsymbol{\beta}) &= \Big[(q_1^i - q_2^i)\Phi(\delta_1) + (q_2^i - q_3^i)\Phi(\delta_3) + q_3^i\Big] Pr(y_{ij} = 0) + \\
&\quad \Big[(q_1^i - q_2^i)\Phi(\delta_1 - \delta_2) + (q_2^i - q_3^i)\Phi(\delta_3 - \delta_2) + q_3^i\Big] Pr(y_{ij} = 1)
\end{aligned}$$

Logo, a distribuição amostral de y_{ij} é Bernoulli(θ_{ij}^s) onde

$$\theta_{ij}^s = \frac{1}{1 + \dfrac{(q_1^i - q_2^i)\Phi(\delta_1) + (q_2^i - q_3^i)\Phi(\delta_3) + q_3^i}{\big[(q_1^i - q_2^i)\Phi(\delta_1 - \delta_2) + (q_2^i - q_3^i)\Phi(\delta_3 - \delta_2) + q_3^i\big] \exp\big(\beta_{0i} + \boldsymbol{x}_{ij}'\boldsymbol{\beta}\big)}}$$

Simulação: Amostras PPT-EST

- Desenho Amostral Informativo no nível das escolas (PPT) e Informativo ao nível de Alunos (EST).
- O modelo a ser ajustado com dados da amostra é

$$(y_{ij} \mid \theta_{ij}^s) \sim \text{Bernoulli}(\theta_{ij}^s)$$

$$logit\,(\theta_{ij}^s) = \log(\frac{B_2}{B_1}) + (\beta_{0i} + \boldsymbol{x}_{ij}'\boldsymbol{\beta})$$

$$(\beta_{0i} \mid \boldsymbol{z}_i, \boldsymbol{\gamma}, \sigma_\mu^2) \sim N(\boldsymbol{z}_i'\boldsymbol{\gamma} + \alpha_1\sigma_\mu^2, \sigma_\mu^2)$$

onde:

$$B_1 = (q_1^i - q_2^i)\Phi(\delta_1) + (q_2^i - q_3^i)\Phi(\delta_3) + q_3^i$$

$$B_2 = (q_1^i - q_2^i)\Phi(\delta_1 - \delta_2) + (q_2^i - q_3^i)\Phi(\delta_3 - \delta_2) + q_3^i$$

Tabela 3: PPT-EST: Média das distribuições a posterioris e EQM

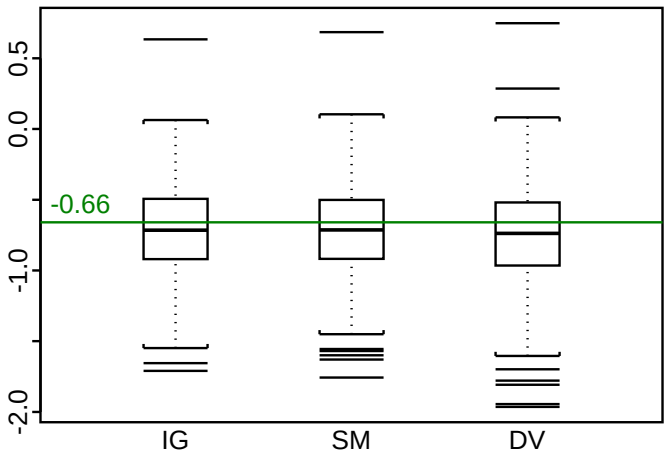
Parâmetro		Média			EQM		
		IG	SM	DV	IG	SM	DV
β_1	-0,66	-0,71	-0,72	-0,74	0,097	0,099	0,127
β_2	-0,95	-1,02	-1,02	-1,07	0,094	0,093	0,130
β_3	-2,10	-2,21	-2,22	-2,33	0,165	0,166	0,274
β_4	-0,43	-0,45	-0,46	-0,47	0,077	0,077	0,100
γ_0	2,65	2,94	2,79	3,99	0,435	0,369	2,989
γ_1	-0,28	-0,30	-0,31	-0,20	0,262	0,236	0,307
γ_2	-0,56	-0,53	-0,59	-0,35	0,358	0,319	0,458
σ^2_μ	0,75	0,73	0,95	0,87	0,148	0,214	0,222

Tabela 4: PPT-EST: Cobertura dos intervalos de 95% de credibilidade

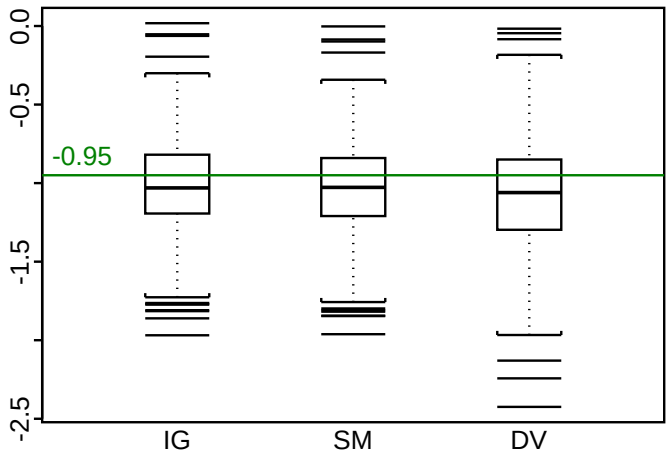
	Modelo				Modelo		
	IG	SM	DV		IG	SM	DV
β_1	93,4	93,0	91,8	γ_0	93,0	94,2	68,0
β_2	92,4	92,4	90,8	γ_1	96,4	95,6	96,6
β_3	92,8	92,0	89,2	γ_2	94,8	96,0	94,4
β_4	94,4	93,8	93,8	σ^2_μ	94,4	94,4	95,0

IG= Modelo Ignorando o Desenho, SM= Modelo Com Distribuições Amostrais e DV= Modelo com Variáveis do Desenho.

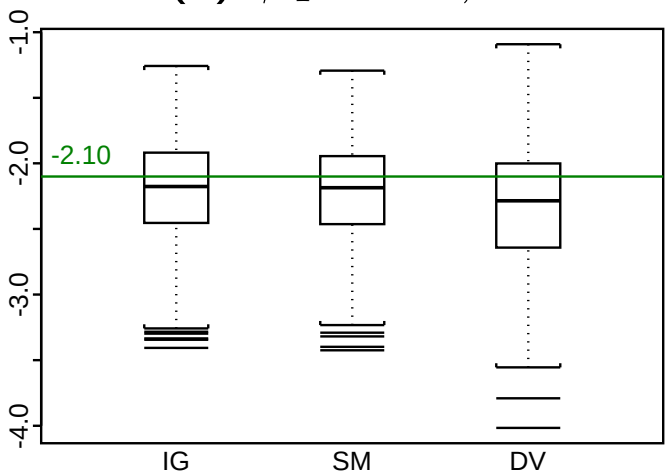
Figura 1: PPT-EST: MÉDIAS A POSTERIORI - 1º nível



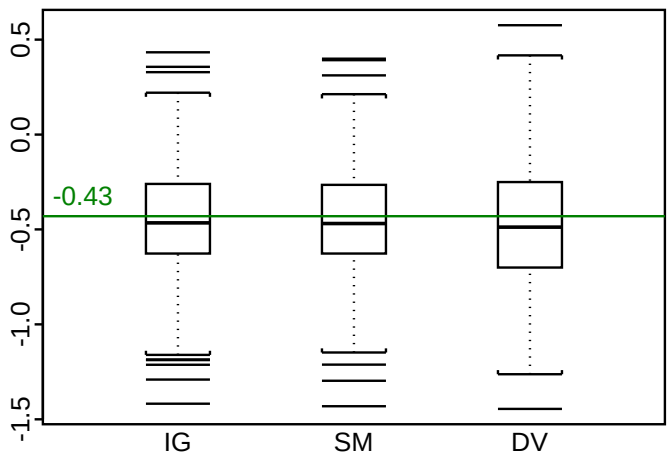
(a) $\beta_1 = -0,66$



(b) $\beta_2 = -0,95$

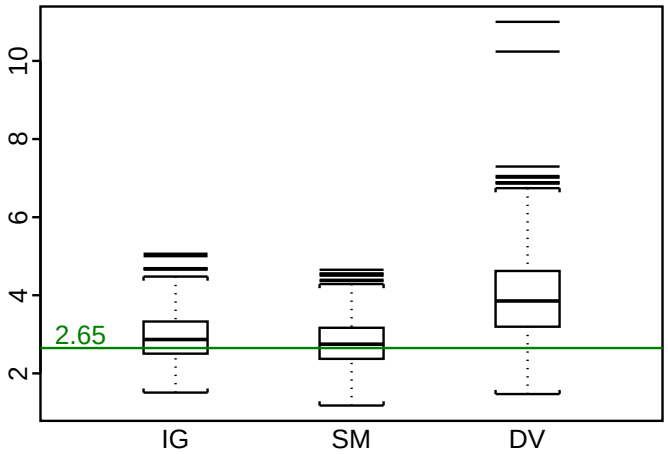


(c) $\beta_3 = -2,10$

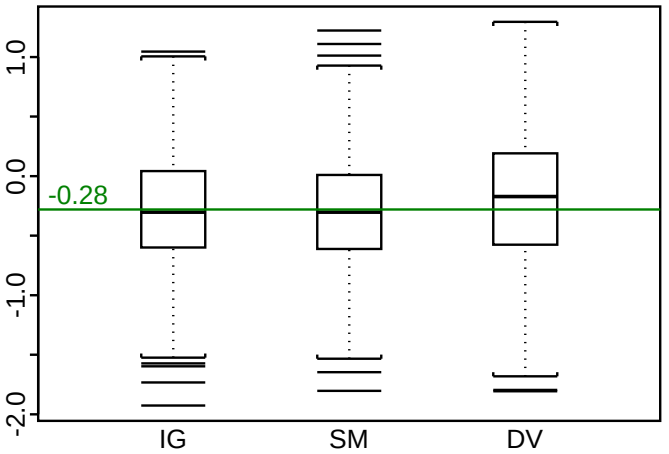


(d) $\beta_4 = -0,43$

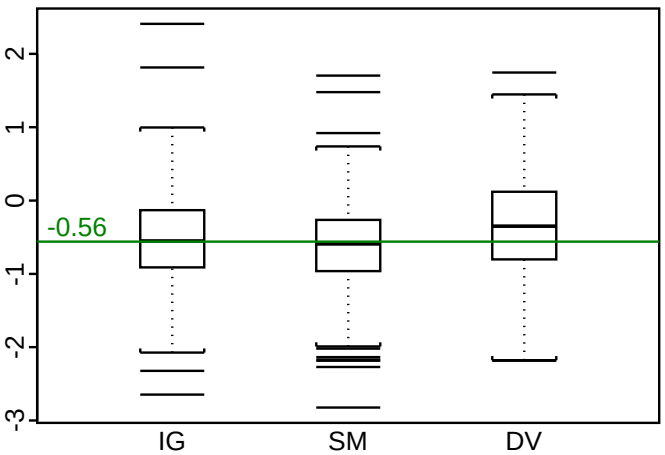
Figura 2: PPT-EST: MÉDIAS A POSTERIORI - 2º nível



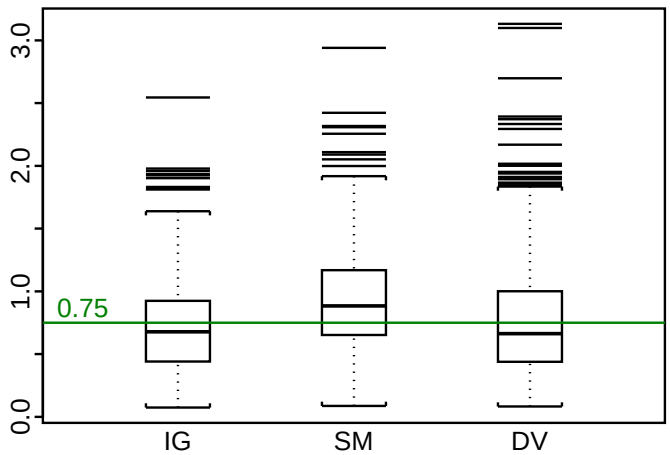
(a) $\gamma_0 = 2,65$



(b) $\gamma_1 = -0,28$



(c) $\gamma_2 = -0,56$



(d) $\sigma^2_\mu = 0,75$

Simulação: Resultados Principais

- ✘ A principal desvantagem do SM é a necessidade de assumir relações adicionais ao modelo de superpopulação.
- ✘ O modelo DV tem os maiores EQM e as menores coberturas.
- ✔ O modelo SM tem os melhores resultados em relação aos parâmetros do 2º nível (Escolas).
- ✔ O modelo SM tem resultados similares ao modelo IG em relação aos parâmetros do 1º nível (Alunos).

Aplicação

- Os dados correspondem à ENAHO-2000.IV
- Amostragem Complexa e Questionário Grande

ENAHO: Unidades e Tipo de amostragem

Unidade	Area	
	Urbana	Rural
Primeira (UP)	CCPP (+ 2000 hab.) PPT	CCPP (500-2000 hab.) ou Grupos de 4 AER PPT
Segunda (US)	Conglomerado PPT	Conglomerado ou 1 AER PPT
Terceira (UT)	Domicílio AAS	Domicílio AAS

- **Modelo de Pobreza**:
  - O objetivo é relacionar algu,mas variáveis sócio-econômicas e demográficas à probabilidade de uma familia ser pobre.
  - Ajusta-se um modelo logístico onde a variável resposta é uma indicadora que assume valor 1 se a familia for *pobre* e 0 se *não for*, no momento da pesquisa.
  - As covariáveis consideradas são:
    - Características do domicílio:
      Material do piso (1=Terra, 0=Outro) e
      Serviço de Saneamento (1=Rede Pública, 0=Outro).
    - Número de Membros da Família.
    - Características do Chefe de Família:
      Sexo (1=Mulher, 0=Homem), Idade, Anos de estudo.
  - Também considera-se algumas variáveis relacionadas com os conglomerados, z_j:
    - Localização geográfica (1=Lima, 0=Outro)
    - Tipo de Localização (1=Urbana, 0=Rural).

Tabela 5: Aplicação: Média e Desvio Padrão a posteriori dos parâmetros

Parâmetros	IG1	SM1	DV1	IG2	SM2	DV2
γ_0	-0,875	-1,060	-1,311	-0,177	-0,366	-0,706
Area	0,074	0,067	0,356	-0,901	-0,878	-0,532
Lima	0,407	0,359	0,207	0,165	0,126	-0,072
σ^2_μ	0,894	0,844	0,854	1,078	1,008	1,007
Piso	1,123	1,121	1,107	–	–	–
Saneamento	-1,035	-1,005	-0,980	–	–	–
Membros	0,455	0,451	0,454	0,432	0,427	0,432
Sexo	-0,107	-0,107	-0,111	-0,148	-0,150	-0,148
Idade	-0,031	-0,030	-0,031	-0,038	-0,038	-0,038
Estudo	-0,171	-0,169	-0,170	-0,220	-0,219	-0,218
Deviance	2256,0	2268,0	2255,0	2349,0	2368,0	2351,0
Sensibilidade	0,6423	0,6411	0,6441	0,6283	0,6228	0,6275
Especificidade	0,7752	0,7732	0,7749	0,7650	0,7638	0,6275
% de acertos	0,7237	0,7220	0,7243	0,7121	0,7108	0,7113

IG1 e IG2 = Modelos Ignorando o Desenho, SM1 e SM2 = Modelos Com Distribuições Amostrais e DV1 e DV2 = Modelos com as Variáveis do Desenho.

- **Aplicação: Discussão de resultados**:
  - A amostra ENAHO-2000.IV é resultado do uso de uma amostragem complexa, porém a inferência sobre parâmetros ao nível de família está livre da influência do plano amostral.
  - No caso da pobreza, o uso do tamanho do conglomerado pode influenciar na presença de famílias pobres na amostra pois é uma variável associada ao tamanho das cidades e conseqüentemente ao desenvolvimento
  - A combinação de co-variáveis presentes no modelo influencia no efeito que o plano amostral tem sobre a estimação dos parâmetros
  - Deve-se avaliar a especificação da esperança condicional

Considerações Finais

- Os desenhos amostrais podem ser informativos.
- A inferencia sobre o modelo de superpopulação deve levar em conta a informação do plano amostral utilizado.
- Existem várias propostas: métodos clássicos e métodos bayesianos
- Em relação as Distribuições Amostrais:
  - Identificabilidade
  - Especificação das Esperanças Condicionais
  - Uso das aproximações de Taylor.
  - Poder Preditivo
  - Tempo computacional
  - Plano não informativo

Trabalhos Futuros

- Aplicar o *Sampling Model* (SM) a outras distribuições, como a Poisson.
- Explorar o uso das aproximações de Taylor para as distribuições amostrais.
- Avaliar o uso de Indicadoras não independentes.
- Trabalhar com a verossimilhança observada completa.
- Estudar outras variáveis da ENAHO-2000.IV e outras pesquisas com amostragem complexa.

Principais Referências

- Pfeffermann, D. and Moura, F. e Silva, P.
  Fitting Multi-level modelling under informative probability sampling.
  Multilevel Modelling Newsletter, 14(1): 8-17, 2002

- Pfeffermann, D. and Krieger, A. e Rinott, Y.
  Parametric distributions of complex survey data under informative probability sampling.
  Statística Sinica, 8:1087–1114, 1998.

- Qin, J. and Leung, D. e Shao, J.
  Estimation with Survey Data Under Nonignorable Nonresponse or Informative Sampling
  Journal of the American Statistical Assosiation, 97(457): 193–200,2002.