Universidade Federal do Rio de Janeiro
Instituto de Matemática

Giuseppe de Abreu Antonaci

Comparação de métodos para estimação de índices de pobreza em pequenas áreas

Rio de Janeiro
2012

Comparação de métodos para estimação de índices de pobreza em pequenas áreas

Giuseppe de Abreu Antonaci

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Estatística, Instituto de Matemática, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Estatística.

Orientador: Fernando Antonio da Silva Moura

Co-orientador: Pedro Luis do Nascimento Silva

Rio de Janeiro

2012

Antonaci, Giuseppe de Abreu
A634c Comparação de métodos para estimação de índices de pobreza em pequenas áreas /Giuseppe de Abreu Antonaci. – Rio de Janeiro: IM/UFRJ, 2012.
xxvi, 72 p. ; 30 cm.

Orientador: Fernando Antonio da Silva Moura.

Dissertação (mestrado) – UFRJ/IM. Programa de Pós-graduação em Estatística, 2012.
Referências: p.89-91.

1.Modelos hierárquicos - Teses 2. Estimação de parâmetro 3. Pobreza -Métodos estatísticos I. Moura, Fernando Antonio da Silva II.Universidade Federal do Rio de Janeiro. Instituto de Matemática III.Título.

Comparação de métodos para estimação de índices de pobreza em pequenas áreas

Giuseppe de Abreu Antonaci

Orientador: Fernando A. S. Moura

Co-orientador: Pedro L. N. Silva

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Estatística, Instituto de Matemática, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Estatística.

Aprovada em:

Prof. Fernando Antonio da Silva Moura, Ph.D., UFRJ

Prof. Helio dos Santos Migon, Ph.D., UFRJ

Prof. Julio da Motta Singer, Ph.D., USP

Rio de Janeiro

2012

Agradecimentos

Aos meus pais, por todo o suporte que eles sempre deram aos meus estudos e tudo o mais que me ensinaram, de maneira voluntária ou involuntária.

A minha namorada, Gabi, que me apoiou nos momentos mais difíceis e pela sua compreensão nesta última etapa do mestrado, quando meus finais de semana livres se tornaram cada vez mais raros.

Aos meus orientadores, Fernando Moura e Pedro Silva, tanto pela sugestão de um tema interessante e de grande utilidade, quanto pelo suporte e sugestões que me deram neste último ano.

Aos membros da banca, prof. Helio Migon e prof. Julio Singer, pelo seu esforço em avaliar meu trabalho e pelas sugestões feitas.

Aos professores do programa de pós-graduação da UFRJ, pela disposição em compartilhar o conhecimento.

Aos meus colegas de turma do mestrado, cuja companhia tornou esta jornada muito mais divertida. Principalmente ao William, cuja disposição em discutir todos os detalhes dos mais complicados teoremas e de destrinchar todas as possibilidades de solução dos mais variados problemas foram de grande auxílio para aprender tudo que este curso me ofereceu.

Ao IBGE e aos meus chefes Marcos Paulo e Sonia, que por meio da minha liberação para cursar este mestrado demonstraram sua confiança na minha capacidade de terminá-lo e retornar ao trabalho um funcionário mais capaz.

In God we trust,
all others must bring data.
- *William E Deming*

Stay a while and listen.
- *Deckard Cain*

Resumo

Comparação de métodos para estimação de índices de pobreza em pequenas áreas

Giuseppe de Abreu Antonaci

Orientador: Fernando A. S. Moura

Co-orientador: Pedro L. N. Silva

Neste trabalho são apresentados e comparados três métodos de estimação em pequenas áreas. O método ELL (Elbers, Lanjouw e Lanjouw, 2003), proposto pelo Banco Mundial e utilizado pelo IBGE (Instituto Brasileiro de Geografia e Estatística) na construção do mapa de pobreza de 2008, e os métodos EB (Empirical Bayes) e HB (Hierarquical Bayes) apresentados por Rao (2003).

Primeiro o índice de pobreza FGT (Foster, Greer e Thorbecke, 1984) é apresentado detalhadamente, assim como os três métodos de estimação em pequenas áreas citados.

Posteriormente, estes métodos foram comparados em dois estudos utilizando dados simulados por meio de modelos de superpopulação. Um dos estudos é somente uma primeira avaliação dos três métodos e utiliza uma população bem regrada e amostra aleatória simples. No segundo estudo a população busca simular características da população brasileira e o deenho da amostra utilizado é comparável aos aplicados nas pesquisas do IBGE.

Por fim, os dados do Censo Demográfico de 2000 e das PNAD (Pesquisa Nacional por Amostragem de Domicílios) de 2001, 2002 e 2003 são utilizados juntamente com os métodos estudados para estimar os índices de pobreza para os municípios do estado de Minas Gerais e os respectivos erros de estimação e uma comparação empírica dos métodos é realizada.

Palavras-chave: modelos hierárquicos, estimação em pequenas áreas, Censo, FGT, ELL, PNAD, IBGE, pobreza.

Abstract

A comparison of methods for small area estimation of poverty indexes

Giuseppe de Abreu Antonaci

Supervisor : Fernando A. S. Moura

Co-supervisor : Pedro L. N. Silva

In this work we present and compare three methods of small area estimation. The ELL method (Elbers, Lanjouw e Lanjouw, 2003), proposed by the World Bank and used by IBGE (Instituto Brasileiro de Geografia e Estatística) to make the poverty map for 2008, and the methods EB (empirical Bayes) and HB (Hierarquical Bayes) presented by Rao (2003).

First we detail the FGT poverty index (Foster, Greer e Thorbecke, 1984) as well as the three small area estimation methods previously cited.

After that, these methods were compared in two studies using simulated data from an superpopulation hierarquical model. One of the studies is just a first evaluation of the three methods and uses a well behaved population model and simple random sample. In the second simulation study the population model tries to emulate key characteristics of the Brazilian population and the sample design used is comparable to those used by IBGE.

Finally, these methods are applied to the data from the Census for 2000 and from PNAD (National Household Sample Survey) for 2001, 2002 and 2003 and used to estimate the poverty indexes for the municipalities of Minas Gerais and their estimation errors. This result is used to conduct an empirical comparison of the methods.

Keywords: hierarchical models, small area estimation, FGT, ELL, Census, PNAD, IBGE, poverty

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

AIC	Critério de Informação Akaike (*Akaike Information Criterion*)
EB	Bayesiano Empírico (*Empirical Bayes*)
ELL	Elbers, Lanjouw, Lanjouw
FGT	Foster, Greer, Thorbecke
HB	Bayesiano Hierárquico (*Hierarquical Bayes*)
IBGE	Fundação Instituto Brasileiro de Geografia e Estatística
MCMC	Monte Carlo via Cadeias de Markov (*Monte Carlo Markov Chain*)
PNAD	Pesquisa Nacional por Amostra de Domicílios

Sumário

1 Introdução

Este trabalho é motivado pela publicação do mapa de pobreza brasileiro pelo IBGE (2008) utilizando a metodologia ELL (Elbers, Lanjouw e Lanjouw, 2002). Existem várias críticas a diferentes aspectos desta metodologia (Haslett et al., 2010; Rao e Molina, 2010; Tarozzi e Deaton, 2009), o que motivou a comparação deste método com outros dois propostos por Rao (2003), o método EB (*Empirical Bayes*) e o método HB (*Hierarquical Bayes*).

O método EB já foi implementado para estimação de índices de pobreza por Rao e Molina (2010) utilizando dados semelhantes ao estudo realizado no capítulo 3 desta dissertação. A contribuição dada por este trabalho é aplicar o método EB para uma população e amostra mais parecida com as obtidas através das pesquisas do IBGE além da aplicação do método HB. Além disso, é realizada uma comparação entre os três métodos citados considerando dados reais destas pesquisas.

No capítulo 1 é feita uma introdução sobre os índices de pobreza e uma atenção maior é dada aos índices FGT (Foster, Greer e Thorbecke, 1984), que serão utilizados neste trabalho. No capítulo 2 os métodos que serão estudados é detalhada. No capítulo 3 é realizada uma primeira comparação dos métodos utilizando uma população simples, com áreas do mesmo tamanho, e desta população retirada uma amostra estratificada simples de todas as áreas. No capítulo 4 novamente os métodos são comparados, no entanto, a população é mais complexa, com áreas de tamanhos desiguais, e a amostra é estratificada com seleção em dois estágios. No capítulo 5 os métodos são aplicados a dados de pesquisas do IBGE e os resultados encontrados por cada método são comparados. No capítulo 6 são apresentadas as conclusões gerais sobre a comparação dos métodos e ressaltando as vantagens e desvantagens de cada um. No capítulo 7 são citados alguns aspectos que não foi possível estudar nesta dissertação assim como outras abordagens que acrescentariam aos resultados apresentados.

1.1 Índices de pobreza

A medição de pobreza é um assunto amplamente discutido desde o início do século passado e vários índices diferentes já foram propostos para resumir as informações sócio-

econômicas disponíveis sobre a população. Uma lista extensa, mas certamente não exaustiva, inclui os Índices de Gini (1955), Theil (1967), Atkinson (1970), Hoover (1936) e FGT (Foster, Greer e Thorbecke, 1984).

A escolha dos índices FGT como foco deste estudo se deve a estes terem sido os índices utilizados pelo IBGE para a estimação dos mapas de pobreza para o Brasil (IBGE, 2008). Ainda assim, os métodos que serão apresentados aqui comportam estimar outros índices e as conclusões obtidas podem ser facilmente estendidas para eles.

1.1.1 Índice FGT

Para uma área m e definida a linha de pobreza z o índice FGT do tipo α é definido como:

$$F_m(\alpha, z) = \frac{1}{N_m} \sum_{j=1}^{N_m} F_{mj}(\alpha, z) \tag{1.1}$$

$$F_{mj}(\alpha, z) = \left(\frac{z - R_{mj}}{z} \right)^{\alpha} I(R_{mj} < z) \tag{1.2}$$

onde $\alpha = 0, 1, 2$, $m = 1, \ldots, M$, N_m é o número de unidades na área m, R_{mj} é uma medida quantitativa pertinente do bem-estar da unidade j da área m, z é a linha de pobreza, ou seja, um valor na mesma escala de R_{mj} que determina se uma unidade é considerada pobre ou não, e $I(R_{mj} < z)$ é uma função indicadora que assume o valor 1 se $R_{mj} < z$ ou 0 em caso contrário.

Para os diferentes valores de α são definidos os índices:

Incidência de pobreza (FGT$(\alpha = 0)$) para cada domicílio, é tão somente uma variável indicadora se este está acima ou abaixo da linha de pobreza, na média para uma área mede a proporção de domicílios que estão abaixo da linha de pobreza;

Hiato de pobreza (FGT$(\alpha = 1)$) para cada domicílio, mede quanto abaixo da linha de pobreza ele está como uma proporção de z. Por exemplo, para um domicílios com $R_{mj} = 0,7z$ se obteria $F_{mj}(1, z) = 0,3$. Para a área mede quanto, em média, seria necessário aumentar a Renda dos domicílios para que estes não estivessem abaixo da linha de pobreza;

Severidade de pobreza (FGT($\alpha = 2$)) tem uma interpretação semelhante de quando $\alpha = 1$ mas dando mais peso aos domicílios mais pobres, sendo portanto mais uma medida do grau de concentração de Renda dos domicílios abaixo da linha de pobreza.

As opções mais comuns para a variável R são medidas padronizadas da renda ou do consumo das unidades pesquisadas. O domicílio será considerado como a unidade básica e desta maneira N_m se refere ao número de domicílios na área m. Como variável de bem estar será utilizada a renda monetária domiciliar *per capita* mensal, daqui por diante referida somente como Renda. Portanto, R_{mj} será a Renda para o domicílio j na área m.

A discussão sobre a escolha de qual linha de pobreza utilizar é extensa e muito importante do ponto de vista sócio-econômico mas foge ao escopo desta dissertação. Em todos os cálculos feitos a linha de pobreza utilizada foi de 60% do valor da mediana da Renda, valor utilizado por vários artigos na área (Osier, 2009; Rao e Molina, 2010).

2 Estimação em Pequenas Áreas

Atualmente pesquisas amostrais são amplamente utilizadas pelos institutos de pesquisas oficiais para obter informação acerca da população. No entanto, estas pesquisas são limitadas na abrangência e precisão das informações disponíveis e é cada vez mais comum que os usuários demandem informações que estas pesquisas não podem fornecer de maneira confiável através dos estimadores usuais.

Tome como exemplo o problema que é tratado nesta dissertação, a estimação índices de pobreza para os municípios de Minas Gerais utilizando os dados da Pesquisa Nacional por Amostra de Domicílios do IBGE (2001, 2002, 2003b). A amostra atual da PNAD permite divulgar estimadores precisos para a Região Metropolitana de Belo Horizonte (que contém 34 municípios) e para o estado de Minas Gerais (contendo 853 municípios). Sem ter como diferenciar a renda ou índices de pobreza entre os municípios da região metropolitana ou entre os municípios do estado não há como fazer um investimento adequado de recursos públicos.

Foi da necessidade de obter estimativas para áreas onde a amostra existente não nos dá informação confiável, ou até mesmo para áreas onde não existe amostra, que a teoria de estimação em pequenas áreas ou domínios se desenvolveu. A principal ideia dos métodos de estimação em pequenas áreas que serão apresentados é que eles usam modelos para "tomar emprestado" informação de outras áreas semelhantes.

2.1 Modelagem e fonte dos dados

Em todos os métodos estudados para a estimação de índices de pobreza o primeiro passo é ajustar um modelo que explique o comportamento de uma variável de bem estar social em função de informações conhecidas dos indivíduos da população. As maneiras de estimar os parâmetros deste modelo variam de acordo com os métodos, mas como eles compartilham algumas definições sobre o modelo e o tratamento dos dados, estes serão apresentados primeiro. Uma representação gráfica da explicação a seguir pode ser vista na figura 2.1.

Considere uma população alvo residente em M áreas com N_m unidades em cada área, onde $m = 1, \ldots, M$ é o índice das áreas e $j = 1, \ldots, N_m$ é o índice das unidades e

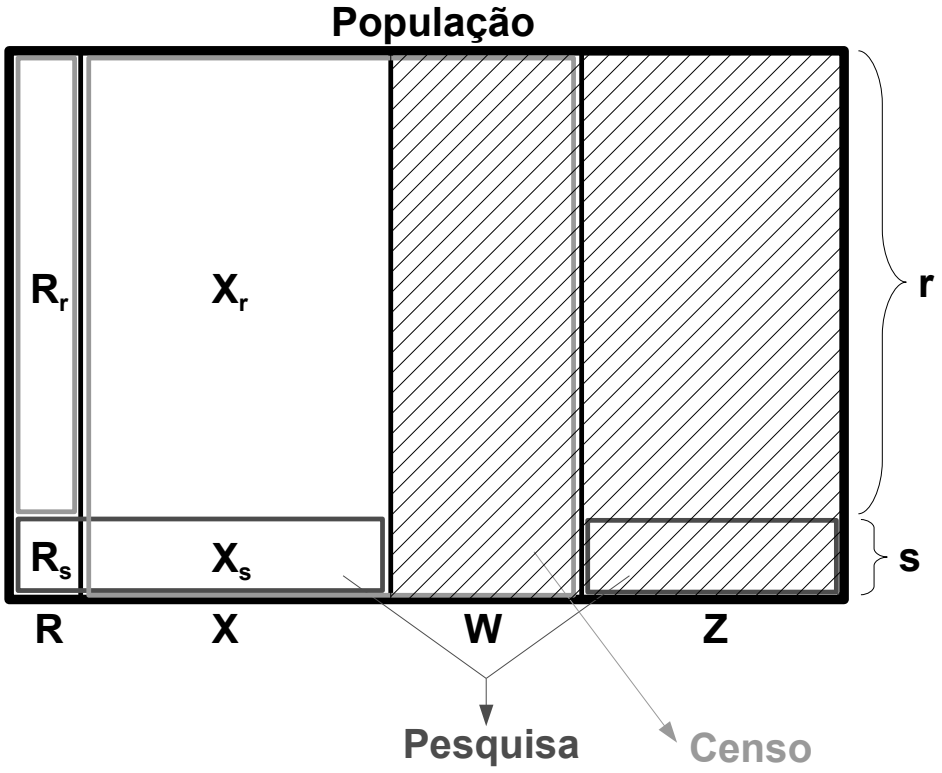


Figura 2.1: Divisão das informações da população nas pesquisas

$N = \sum_{m=1}^{M} N_m$. Seja R a variável que mede o bem estar, necessária para o cálculo dos índices de pobreza.

Usualmente, esta variável é observada somente em uma pequena parcela da população através de uma pesquisa amostral. Sejam s o subgrupo da população que respondeu a esta pesquisa e r seu complemento. Defina como $\boldsymbol{R}_s$ a partição do vetor $\boldsymbol{R}$ que foi observada na amostra e $\boldsymbol{R}_r$ a partição do vetor $\boldsymbol{R}$ não observada na amostra. De maneira independente defina partições do vetor $\boldsymbol{R}$ correspondentes as M áreas. Defina também as interseções $s_m = s \cap m$ e $r_m = r \cap m$ assim como as respectivas interseções das variáveis, ou seja, $\boldsymbol{R}_{s_m}$ e $\boldsymbol{R}_{r_m}$. Por fim defina como R_{mj} o valor observado da variável R para o indivíduo j da área m e n_m o número de unidades na área m que foram observadas na pesquisa amostral.

Suponha que existe informação sobre toda a população, obtida através de um censo populacional, mas que a variável R não faz parte do Censo ou não é adequadamente mensurada para o cálculo de índices de pobreza. Seja $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$ a matriz contendo as informações das variáveis comuns ao censo e à pesquisa amostral, e defina as partições dela assim como as descritas para o vetor $\boldsymbol{R}$.

É importante ressaltar que é inevitável, ao utilizar dois conjuntos de dados de fontes distintas, que algumas variáveis de cada conjunto não sejam comparáveis. Na figura 2.1 representamos por W o conjunto de variáveis que constam no Censo mas não na pesquisa e por Z o conjunto de variáveis que constam na pesquisa mas não no Censo. Estas variáveis serão descartadas da análise, uma vez que não há como utilizá-las para obter informação acerca da variável de bem estar.

Seguindo a notação descrita, tem-se como valores observados $\boldsymbol{R}_s$ e $\boldsymbol{X}' = (\boldsymbol{X}'_s, \boldsymbol{X}'_r)$, onde $\boldsymbol{X}_s$ é observado tanto na pesquisa quanto no censo. O primeiro passo é avaliar a relação entre $\boldsymbol{R}$ e $\boldsymbol{X}$ através de um modelo hierárquico com intercepto aleatórios da forma:

$$\mathrm{g}(R_{mj}) = \beta_0 + \beta_1 X_{1,mj} + \cdots + \beta_p X_{p,mj} + u_m + e_{mj} \tag{2.1}$$

onde u_m são os erros relativos às áreas e e_{mj} são os erros individuais. Ao considerar a renda como variável de bem estar é usual que $\mathrm{g}(\cdot) = \log(\cdot)$. Neste caso é também usual supor que $u_m \sim \mathrm{N}(0, \sigma_u^2)$ e $e_{mj} \sim \mathrm{N}(0, \sigma_e^2)$. Seja também $\boldsymbol{Y} = \mathrm{g}(\boldsymbol{R}) = \log(\boldsymbol{R})$.

Ressalta-se que não é o foco deste trabalho testar qual o melhor modelo para o ajuste da renda e sim a comparação dos métodos de estimação. Baseados no modelo escolhido em IBGE (2008) para Minas Gerais, onde todas as variáveis utilizadas são lineares e poucas interações foram consideradas significativas, adotou-se um modelo puramente linear para realizar a comparação dos métodos.

Definindo $\boldsymbol{\theta} = (\beta_0, \beta_1, \ldots, \beta_p, \sigma_u, \sigma_e) = (\boldsymbol{\beta}, \sigma_u, \sigma_e)$, é possível utilizar a amostra s para obter uma estimativa $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$. Esta, por sua vez, pode ser utilizada para estimar valores para $\boldsymbol{R}_r$ e, por meio deste, obter estimativas mais precisas de $F_m(\alpha, z)$ para as áreas desejadas. As formas de estimação de $\boldsymbol{\theta}$ e como as informações auxiliares são utilizadas para estimar os índices de pobreza variam entre os métodos estudados e serão apresentadas com detalhes posteriormente.

2.2 Método Direto

A estimação dos índices de pobreza com base somente nos dados da amostra, sem o auxílio de outras fontes de dados ou de modelos, é denominado método direto. Foi utilizado neste caso o estimador de Horvitz-Thompson (Cochran, 1977, pág. 259). Apesar de não ser um estimador de pequenas áreas, é pertinente acrescentar esta pequena descrição,

uma vez que este estimador é oportunamente comparado com os estimadores de pequenas áreas.

Seja w_{mj} peso da unidade j da área m, ou seja, o inverso da probabilidade de seleção da unidade j da área m.

O estimador direto do índice de pobreza FGT para a área m é:

$$\hat{F}_m(\alpha, z) = \frac{1}{N_m} \sum_{j \in s_m} w_{mj} \left(\frac{z - R_{mj}}{z} \right)^{\alpha} I(R_{mj} < z).$$

Ressalta-se que este estimador só existe para áreas onde há amostra.

2.3 Método do Banco Mundial (ELL)

Desenvolvido por Elbers, Lanjouw e Lanjouw (2002) e promovido pelo Banco Mundial, este método foi utilizado em diversos países para a estimação e construção de mapas de pobreza, inclusive no Brasil. Tarozzi e Deaton (2009) e Rao e Molina (2010) apresentaram críticas ao método, algumas das quais serão citadas a seguir.

2.3.1 Estimação dos parâmetros do modelo

A estimação dos parâmetros do modelo hierárquico apresentado em (2.1) só utiliza os dados observados na amostra da pesquisa, ou seja, $\boldsymbol{R}_s$ e $\boldsymbol{X}_s$. Os passos propostos para estimação dos parâmetros segundo os autores do método ELL são:

1. Primeiro estime $\boldsymbol{\beta}$ através de mínimos quadrados ordinários (MQO) para regressão linear, obtendo $\hat{\boldsymbol{\beta}}_{\text{MQO}}$;

2. Obtenha então estimativas dos resíduos $\hat{r}_{mj} = Y_{mj} - \boldsymbol{X}'_{mj}\hat{\boldsymbol{\beta}}_{\text{MQO}}$;

3. Decomponha os resíduos em $\hat{r}_{mj} = \hat{u}_m + \hat{e}_{mj}$ onde $\hat{u}_m = \bar{r}_m = \frac{1}{n_m} \sum_{j \in s_m} \hat{r}_{mj}$ e $\hat{e}_{mj} = (\hat{r}_{mj} - \bar{r}_m)$;

4. Estime σ_u^2 como:

$$\hat{\sigma}_u^2 = \max \left\{ \frac{\sum_{m \in s} \lambda_m (\bar{r}_m - \bar{r})^2}{\sum_{m \in s} \lambda_m (1 - \lambda_m)} - \frac{\sum_{m \in s} \lambda_m (1 - \lambda_m) \hat{\tau}_m^2}{\sum_{m \in s} \lambda_m (1 - \lambda_m)}; 0 \right\}$$

   com $\hat{\tau}_m^2 = \frac{1}{n_m(n_m - 1)} \sum_{m \in s} (\hat{e}_{mj} - \bar{e}_m)^2$, $\lambda_m = \frac{w_m}{\sum_{m \in s} w_m}$, $\bar{r} = \frac{1}{M} \sum_{m \in s} \bar{r}_m$, w_m é o peso amostral da área m e $\bar{e}_m = \frac{1}{n_m} \sum_{j \in s_m} \hat{e}_{mj}$;

5. Estime σ_e^2 através de

$$\hat{\sigma}_e^2 = \frac{1}{\sum_{m \in s} n_m - p} \sum_{m \in s} \sum_{j \in s_m} \hat{r}_{mj}^2 - \hat{\sigma}_u^2;$$

6. Agora usando $\hat{\sigma}_e^2$ e $\hat{\sigma}_u^2$ estime novamente os valores de $\boldsymbol{\beta}$ por mínimos quadrados ponderados:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{ELL}} &= \left(\sum_{m \in s} \boldsymbol{X}'_{s_m} \boldsymbol{W}_{s_m} \boldsymbol{V}_{s_m}^{-1} \boldsymbol{X}_{s_m} \right)^{-1} \left(\sum_{m \in s} \boldsymbol{X}'_{s_m} \boldsymbol{W}_{s_m} \boldsymbol{V}_{s_m}^{-1} \boldsymbol{Y}_{s_m} \right) \\
\boldsymbol{W}_{s_m} &= \text{diag}(w_{m1}, \ldots, w_{mn_m}) \\
\boldsymbol{V}_{s_m} &= \hat{\sigma}_u^2 \left(\mathbf{1}_{n_m} \mathbf{1}'_{n_m} \right) + \hat{\sigma}_e^2 \mathbb{I}_{n_m}
\end{aligned}$$

onde $\mathbf{1}_{n_m}$ é um vetor com dimensão n_m e com todos os elementos iguais a 1 e $\mathbb{I}_{n_m}$ é a matriz identidade de tamanho n_m;

7. Estime a variância de $\hat{\boldsymbol{\beta}}_{\text{ELL}}$ através de:

$$\hat{\text{V}} \left(\hat{\boldsymbol{\beta}}_{\text{ELL}} \right) = \boldsymbol{D} \left(\sum_{m \in s} \boldsymbol{X}'_{s_m} \boldsymbol{W}_{s_m} \boldsymbol{V}_{s_m}^{-1} \boldsymbol{W}_{s_m} \boldsymbol{X}_{s_m} \right)^{-1} \boldsymbol{D}$$

onde $\boldsymbol{D} = \left(\sum_{m \in s} \boldsymbol{X}'_{s_m} \boldsymbol{W}_{s_m} \boldsymbol{V}_{s_m}^{-1} \boldsymbol{X}_{s_m} \right)^{-1}$.

A crítica mais óbvia a este método é que os estimadores de variância se baseiam em resíduos obtidos usando $\hat{\boldsymbol{\beta}}_{\text{MQO}}$ enquanto os resíduos ao final do processo serão obtidos usando $\hat{\boldsymbol{\beta}}_{\text{ELL}}$. A correção mais simples seria reestimar alternadamente os parâmetros fixos e de variância do modelo até que as estimativas convergissem. Outras críticas ao método são apresentadas em Haslett et al. (2010).

2.3.2 Estimação dos índices de pobreza e de suas variâncias

Para estimar o índice de pobreza para a área m deseja-se utilizar a esperança do índice dados os parâmetros do modelo hierárquico e os dados obtidos no Censo, ou seja:

$$\text{E} \left[F_m(\alpha, z) | \boldsymbol{\beta}, \sigma_u, \sigma_e, \boldsymbol{X} \right]$$

Note que, no caso do método ELL, a estimativa do índice de pobreza não depende diretamente da amostra da pesquisa. No entanto, como os parâmetros do modelo hierárquico são desconhecidos eles serão substituídos pelas suas estimativas pontuais, estas baseadas

nos dados da pesquisa. Logo a estimativa do índice de pobreza para a área m é obtida por:

$$\mathrm{E}\left[F_m(\alpha, z)|\hat{\boldsymbol{\beta}}, \hat{\sigma}_u, \hat{\sigma}_e, \boldsymbol{X}\right]$$

ou

$$\int \left\{ \frac{1}{N_m} \sum_{j=1}^{N_m} \left(\frac{z - R_{mj}}{z} \right)^{\alpha} I(R_{mj} < z) \right\} dF(R_{mj}|\hat{\boldsymbol{\beta}}, \hat{\sigma}_u, \hat{\sigma}_e, \boldsymbol{X}) \tag{2.2}$$

Para o índice FGT, é possível reescrever a expressão (2.2) como:

$$\frac{1}{N_m} \sum_{j=1}^{N_m} \left\{ \int \left(\frac{z - R_{mj}}{z} \right)^{\alpha} I(R_{mj} < z) dF(R_{mj}|\hat{\boldsymbol{\beta}}, \hat{\sigma}_u, \hat{\sigma}_e, \boldsymbol{X}) \right\} \tag{2.3}$$

A integral apresentada em (2.3) não tem forma fechada sendo necessário utilizar métodos numéricos para aproximá-la. Optou-se por utilizar o método de replicação descrito a seguir:

1. Gere $l = 1, \ldots, L$ réplicas de Y_{mj} de acordo com a fórmula em (2.1), a saber

$$\hat{Y}_{mj}^{(l)} = \hat{\beta}_0^{(l)} + \hat{\beta}_1^{(l)} x_{1,mj} + \cdots + \hat{\beta}_p^{(l)} x_{p,mj} + \hat{u}_m^{(l)} + \hat{e}_{mj}^{(l)}$$

   $m = 1, \ldots, M$, $j = 1, \ldots, N_m$ e

$$\hat{\boldsymbol{\beta}}^{(l)} \sim N\left(\hat{\boldsymbol{\beta}}_{\mathrm{ELL}}, \hat{\mathrm{V}}\left(\hat{\boldsymbol{\beta}}_{\mathrm{ELL}}\right)\right), \qquad \hat{u}_m^{(l)} \sim N(0, \hat{\sigma}_u^2), \qquad \hat{e}_{mj}^{(l)} \sim N(0, \hat{\sigma}_e^2)$$

2. Para cada réplica l calcule $F_m^{(l)}(\alpha, z)$ como em (1.1) e (1.2), onde $R_{mj}^{(l)} = \exp(Y_{mj}^{(l)})$

3. Obtenha a estimativa do índice de pobreza para a área m fazendo a média entre as l estimativas obtidas no passo 2, isto é:

$$\hat{F}_m^{\mathrm{ELL}}(\alpha, z) = \frac{1}{L} \sum_{l=1}^{L} F_m^{(l)}(\alpha, z)$$

Desta maneira obtém-se estimativas pontuais para os índices de pobreza para as áreas.

Elbers et al. (2002) também apresentam uma maneira de estimar a variância do estimador pontual através de

$$\hat{V}\left(\hat{F}_m^{\mathrm{ELL}}(\alpha, z)\right) = \frac{1 + L^{-1}}{L - 1} \sum_{l=1}^{L} \left(F_m^{(l)}(\alpha, z) - \hat{F}_m^{\mathrm{ELL}}(\alpha, z)\right)^2$$

Uma crítica usual ao método ELL é como ele gera valores para $\hat{u}_m^{(l)}$. Para auxiliar a explicação, é fornecido um pequeno exemplo dos resíduos do ajuste de um modelo hierárquico.

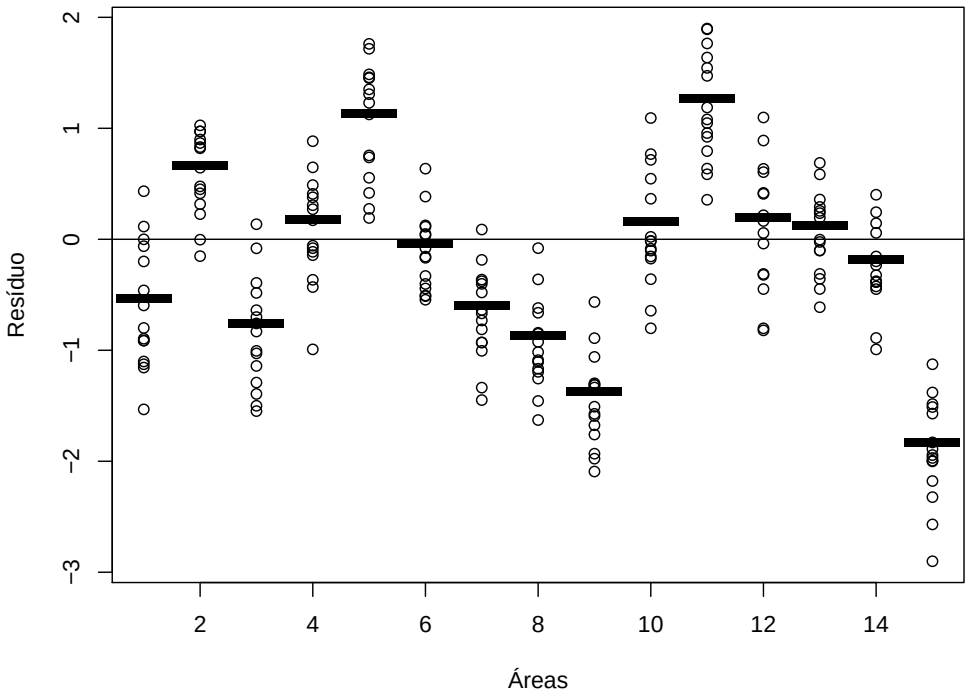


Figura 2.2: Resíduos por área

Observe na figura 2.2 onde são mostrados os resíduos de todas as observações separados por áreas e as médias dos resíduos por área. O valor de $\hat{\sigma}_e$ é obtido pela dispersão dos resíduos em torno das médias das áreas e o valor de $\hat{\sigma}_u$ é obtido pela dispersão das médias dos resíduos por área em torno do zero.

Ao gerar o erro de grupo para a área 15 o método ELL usa a distribuição $u_{15}^{(l)} \sim N(0, \hat{\sigma}_u^2)$. No entanto, como se vê na figura 2.2, existe evidências de que esta área está abaixo da média. No entanto, esta informação é ignorada pelo método e todos os erros de área são estimados como se nada se soubesse sobre aquela área.

2.4 Método Bayesiano empírico (EB)

A base teórica deste método é apresentada de forma geral por (Rao, 2003, Cap. 9), e mais focada na estimação de índices de pobreza em Rao e Molina (2010).

2.4.1 Estimação dos parâmetros do modelo

A teoria apresentada em Rao e Molina (2010) não foca a estimação dos parâmetros do modelo, sugerindo que seja utilizado um dos métodos já aceitos pela comunidade científica para o ajuste do modelo hierárquico. Rao e Molina (2010) sugerem o estimador de máxima

verossimilhança (ML) ou o estimador de máxima verossimilhança residual ou restrito (REML). Este estimador está na biblioteca *lme4* (Bates e Maechler, 2010) do programa estatístico *R* (R Development Core Team, 2010). Considerou-se também os estimadores de mínimos quadrados iterativos generalizados (IGLS) e mínimos quadrados ponderados iterativos generalizados (PWIGLS) apresentados por Pffefermann et al. (1998).

2.4.2 Estimação dos índices de pobreza

Considere um índice genérico υ que seja função de uma variável populacional $\boldsymbol{Y}$, ou seja, $\upsilon = h(\boldsymbol{Y})$. Seja $\hat{\upsilon}$ um estimador de υ que depende somente de $\boldsymbol{Y}_s$, a parte conhecida de $\boldsymbol{Y}$. Logo, o erro quadrático médio de $\hat{\upsilon}$ será:

$$EQM(\hat{\upsilon}) = E_{\boldsymbol{Y}}\{(\hat{\upsilon} - \upsilon)^2\} \tag{2.4}$$

onde $E_{\boldsymbol{Y}}$ denota a esperança com relação ao vetor de dados populacionais. O melhor preditor de υ é a função de $\boldsymbol{Y}_s$ que minimiza (2.4). Considere $\upsilon^* = E_{\boldsymbol{Y}_r}(\upsilon|\boldsymbol{Y}_s)$, onde a esperança é tomada com respeito à distribuição condicional de $\boldsymbol{Y}_r|\boldsymbol{Y}_s$. Então

$$EQM(\hat{\upsilon}) = E_{\boldsymbol{Y}}\{(\hat{\upsilon} - \upsilon^*)^2\} + 2E_{\boldsymbol{Y}}\{(\hat{\upsilon} - \upsilon^*)(\upsilon^* - \upsilon)\} + E_{\boldsymbol{Y}}\{(\upsilon^* - \upsilon)^2\}$$

Observe que o último termo não depende de $\hat{\upsilon}$ e que o segundo termo é nulo pois:

$$\begin{aligned} E_{\boldsymbol{Y}}\{(\hat{\upsilon} - \upsilon^*)(\upsilon^* - \upsilon)\} &= E_{\boldsymbol{Y}_s}\big\{E_{\boldsymbol{Y}_r}\big[(\hat{\upsilon} - \upsilon^*)(\upsilon^* - \upsilon)|\boldsymbol{Y}_s\big]\big\} \\ &= E_{\boldsymbol{Y}_s}\big\{(\hat{\upsilon} - \upsilon^*)(\upsilon^* - E_{\boldsymbol{Y}_r}[\upsilon|\boldsymbol{Y}_s])\big\} \\ &= 0 \end{aligned}$$

Logo, deseja-se que $\hat{\upsilon}$ minimize o valor de $E_{\boldsymbol{Y}}\{(\hat{\upsilon} - \upsilon^*)^2\}$. Como esta quantidade é não negativa ela será mínima se $\hat{\upsilon} = \upsilon^*$. Logo,

$$\hat{\upsilon}^B = \upsilon^* = E_{\boldsymbol{Y}_r}(\upsilon|\boldsymbol{Y}_s) \tag{2.5}$$

é o melhor preditor. É fácil notar também que $\hat{\upsilon}^B$ é não viesado pois

$$E_{\boldsymbol{Y}_s}(\hat{\upsilon}^B) = E_{\boldsymbol{Y}_s}\big\{E_{\boldsymbol{Y}_r}(\upsilon|\boldsymbol{Y}_s)\big\} = E_{\boldsymbol{Y}}(\upsilon).$$

A demonstração pode ser vista no apêndice A.1.

Usualmente, $\hat{\upsilon}^B$ depende de $\boldsymbol{\theta}$, um vetor desconhecido de parâmetros do modelo. Neste caso, um estimador Bayesiano empírico pode ser obtido calculando a esperança em (2.5) considerando $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, onde $\hat{\boldsymbol{\theta}}$ é um estimador apropriado de $\boldsymbol{\theta}$.

Considerando a transformação definida na seção 2.1 pode-se reescrever o índice FGT para a área m como

$$h(\boldsymbol{Y}_m) = F_m(\alpha, z) = \frac{1}{N_m} \sum_{j=1}^{N_m} \left(\frac{z - g^{-1}(Y_{mj})}{z} \right)^{\alpha} I(g^{-1}(Y_{mj}) < z) \tag{2.6}$$

Note que $h(\boldsymbol{Y}_m)$ pode ser definida de forma geral como função de $\boldsymbol{Y}_m$, independentemente do índice de pobreza a ser utilizado. No entanto, como o índice FGT é separável por adição, podemos escrever

$$h(Y_{mj}) = \left(\frac{z - g^{-1}(Y_{mj})}{z} \right)^{\alpha} I(g^{-1}(Y_{mj}) < z)$$

e neste caso,

$$h(\boldsymbol{Y}_m) = \frac{1}{N_m} \sum_{j=1}^{N_m} h(Y_{mj})$$

Tomando $\upsilon = F_m(\alpha, z)$ obtém-se por (2.5) que

$$\hat{F}_m^B(\alpha, z) = E_{\boldsymbol{Y}_r}\big(F_m(\alpha, z)|\boldsymbol{Y}_s\big)$$

Particionando $F_m(\alpha, z)$ tem-se

$$\hat{F}_m^B(\alpha, z) = E_{\boldsymbol{Y}_r} \left(\frac{1}{N_m} \left\{ \sum_{j \in s_m} F_{mj}(\alpha, z) + \sum_{j \in r_m} F_{mj}(\alpha, z) \right\} |\boldsymbol{Y}_s \right)$$

$$\hat{F}_m^B(\alpha, z) = \frac{1}{N_m} \left\{ \sum_{j \in s_m} F_{mj}(\alpha, z) + \sum_{j \in r_m} E_{\boldsymbol{Y}_r}\big(F_{mj}(\alpha, z)|\boldsymbol{Y}_s\big) \right\}$$

$$\hat{F}_m^B(\alpha, z) = \frac{1}{N_m} \left\{ \sum_{j \in s_m} F_{mj}(\alpha, z) + \sum_{j \in r_m} \hat{F}_{mj}^B(\alpha, z) \right\}$$

onde

$$\hat{F}_{mj}^B(\alpha, z) = E_{\boldsymbol{Y}_r}\big(h(Y_{mj})|\boldsymbol{Y}_s\big) = \int h(Y_{mj}) dF(Y_{mj}|\boldsymbol{Y}_s), \ j \in r_m \tag{2.7}$$

Normalmente a integral em (2.7) é intratável devido à complexidade de $h(Y_{mj})$. No entanto, a densidade de Y_{mj} dado $\boldsymbol{Y}_s$, é fácil de ser determinada uma vez que, por hipótese, $\boldsymbol{Y}$ tem distribuição normal (veja (West e Harrison, 1997, pág. 637)).

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_s \\ \boldsymbol{Y}_r \end{pmatrix} \sim N \left[\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_s \\ \boldsymbol{\mu}_r \end{pmatrix}, \mathbf{V} = \begin{pmatrix} \mathbf{V}_s & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_r \end{pmatrix} \right]$$

logo a distribuição condicional de $\boldsymbol{Y}_r$ dado $\boldsymbol{Y}_s$ será

$$\boldsymbol{Y}_r|\boldsymbol{Y}_s \sim N(\boldsymbol{\mu}_{r|s}, \mathbf{V}_{r|s}) \tag{2.8}$$

onde

$$\boldsymbol{\mu}_{r|s} = \boldsymbol{\mu}_r + \mathbf{V}_{rs}\mathbf{V}_s^{-1}(\boldsymbol{Y}_s - \boldsymbol{\mu}_s) \quad \text{e} \quad \mathbf{V}_{r|s} = \mathbf{V}_r - \mathbf{V}_{rs}\mathbf{V}_s^{-1}\mathbf{V}_{sr}$$

A partir desta distribuição, $\hat{F}_{mj}^{B}(\alpha, z)$ pode ser estimado simulando L populações do vetor $\boldsymbol{Y}_r$, geradas através de (2.8). Seja $Y_{mj}^{(l)}$, o l-ésimo valor simulado para a observação j da área m. Então aproxima-se

$$\hat{F}_{mj}^{B}(\alpha, z) = \frac{1}{L}\sum_{l=1}^{L} h(Y_{mj}^{(l)}) \tag{2.9}$$

Neste caso a distribuição de $\boldsymbol{Y}_r|\boldsymbol{Y}_s$ vai depender de $\boldsymbol{\theta}$, o vetor de parâmetros do modelo apresentado em (2.1). Substituindo $\boldsymbol{\theta}$ por $\hat{\boldsymbol{\theta}}$, o estimador final é chamado de estimador Bayesiano empírico (ou EB) e denota-se $\hat{F}_{mj}^{EB}(\alpha, z)$. Finalmente tem-se que

$$\hat{F}_{m}^{EB}(\alpha, z) = \frac{1}{N_m}\left\{\sum_{j\in s_m} F_{mj}(\alpha, z) + \sum_{j\in r_m} \hat{F}_{mj}^{EB}(\alpha, z)\right\} \tag{2.10}$$

2.4.3 Estimação do erro da estimava do índice de pobreza

Rao e Molina (2010) sugerem a utilização de *bootstrap* paramétrico para a estimação do erro quadrático médio dos estimadores dos índices de pobreza. Este *bootstrap* paramétrico para populações finitas é apresentado em González-Manteiga et al. (2008) e a seguir é descrito os passos para a sua implementação.

1. Ajuste o modelo 2.1 ao dados da amostra, $\boldsymbol{Y}_s$ e $\boldsymbol{X}_s$, e obtenha estimativas $\hat{\boldsymbol{\beta}}, \hat{\sigma}_e^2, \hat{\sigma}_u^2$ de $\boldsymbol{\beta}, \sigma_e^2, \sigma_u^2$ usando um método de estimação adequado;

2. Construa um modelo de superpopulação $\xi^* : Y_{mj} = \boldsymbol{X}_{mj}{}'\hat{\boldsymbol{\beta}} + u_m^* + e_{mj}^*,\ m = 1, \ldots, M,\ j = 1, \ldots, N_m$ onde $u_m^* \overset{iid}{\sim} N(0, \hat{\sigma}_u^2)$ e $e_{mj}^* \overset{iid}{\sim} N(0, \hat{\sigma}_e^2)$;

3. Gere réplicas $\boldsymbol{Y}^b$ de populações *bootstrap* de acordo com as distribuições de u_m^* e e_{mj}^*;

4. Calcule os parâmetros populacionais $F_m^b(\alpha, z) = h(\boldsymbol{Y}^b)$;

5. Para cada população b tire uma amostra com o mesmo desenho da amostra s e calcule $\hat{F}_m^{\mathrm{EB},b}(\alpha, z)$ da maneira descrita na subseção 2.4.2;

6. Uma estimativa para o $\mathrm{EQM}(\hat{F}_m^{\mathrm{EB}}(\alpha, z))$ é dada por

$$\mathrm{eqm}(\hat{F}_m^{\mathrm{EB}}(\alpha, z)) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{F}_m^{\mathrm{EB},b}(\alpha, z) - F_m^b(\alpha, z)\right)^2$$

É importante ressaltar que a consistência do estimador do erro quadrático médio depende da consistência dos estimadores de $\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2$.

2.5 Método Bayesiano Hierárquico (HB)

O método Bayesiano Hierárquico ou Bayes completo é apresentado em (Rao, 2003, Cap. 10) e propõe o uso de uma distribuição a priori $f(\boldsymbol{\theta})$ para os parâmetros do modelo. Analogamente ao apresentado para o EB, deseja-se estimar o índice de pobreza a partir da distribuição de $\boldsymbol{Y}_r|\boldsymbol{Y}_s$, mas neste caso será usada uma distribuição a priori $f(\boldsymbol{\theta})$ e não se substituirá o valor desconhecido de $\boldsymbol{\theta}$ por um estimador pontual.

2.5.1 Estimação dos parâmetros do modelo

Pode-se fazer uma definição um pouco diferente dos parâmetros do modelo, que será mais útil para estimar $\boldsymbol{Y}_r$ ao utilizar a inferência Bayesiana. Primeiramente reescreva o modelo apresentado em (2.1) como:

$$\begin{aligned} g(R_{mj}) &= \beta_{0,m} + \beta_1 x_{1,mj} + \cdots + \beta_p x_{p,mj} + e_{mj} \\ \beta_{0,m} &= \beta_0 + u_m \end{aligned}$$

e defina $\boldsymbol{\theta}_2' = (\beta_{0,1}, \ldots, \beta_{0,M}, \beta_0, \beta_1, \ldots, \beta_p, \sigma_e, \sigma_u)$.

Considerando as hipóteses usuais de independência

$$\begin{aligned} Y_{mj}|\boldsymbol{\theta}_2 &\perp Y_{nl}|\boldsymbol{\theta}_2 && \text{se } m \neq n \text{ ou } j \neq l \\ \beta_{0,m}|\beta_0, \sigma_u &\perp \beta_{0,n}|\beta_0, \sigma_u && \text{se } m \neq n \\ (\beta_1, \ldots, \beta_p, \sigma_e) &\perp (\beta_0, \sigma_u) \end{aligned} \tag{2.11}$$

segue-se que

$$\begin{aligned} f(\boldsymbol{Y}_s|\boldsymbol{\theta}_2) &= \prod_{m=1}^{M} \prod_{j \in s_m} f(Y_{mj}|\boldsymbol{\theta}_2) \\ f(\beta_{0,1}, \ldots, \beta_{0,M}|\beta_0, \sigma_u) &= \prod_{m=1}^{M} f(\beta_{0,m}|\beta_0, \sigma_u) \\ f(\beta_1, \ldots, \beta_p, \sigma_e) &= f(\beta_1, \ldots, \beta_p, \sigma_e|\beta_0, \sigma_u) \end{aligned}$$

onde

$$Y_{mj}|\boldsymbol{\theta}_2 \sim N(\beta_{0,m} + \beta_1 x_{1,mj} + \cdots + \beta_p x_{p,mj}, \sigma_e^2)$$
$$\beta_{0,m}|\beta_0, \sigma_u \sim N(\beta_0, \sigma_u^2)$$

A distribuição a posteriori de $\boldsymbol{\theta}_2$ será portanto

$$f(\boldsymbol{\theta}_2|\boldsymbol{Y}_s) \propto f(\boldsymbol{Y}_s|\boldsymbol{\theta}_2) f(\beta_{0,1}, \ldots, \beta_{0,M}|\beta_0, \sigma_u) f(\beta_1, \ldots, \beta_p, \sigma_e|\beta_0, \sigma_u) f(\beta_0, \sigma_u)$$

2.5.2 Estimação dos índices de pobreza

Como na subseção 2.4.2, o objetivo é estimar

$$E_{\boldsymbol{Y}_r}\big(F_m(\alpha, z)|\boldsymbol{Y}_s\big)$$

É necessário primeiro determinar a distribuição de $\boldsymbol{Y}_r|\boldsymbol{Y}_s$, mas diferentemente do que foi feito na seção anterior, $\boldsymbol{\theta}_2$ não será substituído por uma estimativa pontual. De acordo com a hipótese apresentada na expressão (2.11) tem-se que $f(\boldsymbol{Y}_r|\boldsymbol{\theta}_2) = f(\boldsymbol{Y}_r|\boldsymbol{\theta}_2, \boldsymbol{Y}_s)$. Portanto,

$$\begin{aligned} f(\boldsymbol{Y}_r|\boldsymbol{Y}_s) &= \int_{\boldsymbol{\Theta}_2} f(\boldsymbol{Y}_r|\boldsymbol{\theta}_2, \boldsymbol{Y}_s) f(\boldsymbol{\theta}_2|\boldsymbol{Y}_s) d\boldsymbol{\theta}_2 \\ &= \int_{\boldsymbol{\Theta}_2} f(\boldsymbol{Y}_r, \boldsymbol{\theta}_2|\boldsymbol{Y}_s) d\boldsymbol{\theta}_2 \end{aligned}$$

Utilizando a definição da função $h(\cdot)$ de (2.6) tem-se que

$$h(\boldsymbol{Y}) = E_{\boldsymbol{Y}_r}\left(\frac{1}{N_m}\left\{\sum_{j\in s_m} F_{mj}(\alpha, z) + \sum_{j\in r_m} F_{mj}(\alpha, z)\right\}|\boldsymbol{Y}_s\right)$$
$$h(\boldsymbol{Y}) = \frac{1}{N_m}\left\{\sum_{j\in s_m} F_{mj}(\alpha, z) + \sum_{j\in r_m} E_{\boldsymbol{Y}_r}\big(F_{mj}(\alpha, z)|\boldsymbol{Y}_s\big)\right\} \tag{2.12}$$

onde

$$E_{\boldsymbol{Y}_r}\big(F_{mj}(\alpha, z)|\boldsymbol{Y}_s\big) = \int_{Y_{mj}} \int_{\boldsymbol{\Theta}_2} F_{mj}(\alpha, z) f(Y_{mj}, \boldsymbol{\theta}_2|\boldsymbol{Y}_s) d\boldsymbol{\theta}_2 dY_{mj}$$

Dada a impossibilidade de calcular as integrais acima de maneira analítica, métodos numéricos para estimação são necessários. Neste caso, será utilizado Monte Carlo via cadeias de Markov (Gamerman e Lopes, 2006).

Para obter uma estimativa pontual do índice de pobreza, retira-se uma amostra $h^{(1)}, \ldots, h^{(L)}$ onde $h^{(l)} \sim h(\boldsymbol{Y})$, a distribuição a posteriori apresentada em (2.12), e calcula-se

$$\hat{F}_m^{HB}(\alpha, z) = \frac{1}{L}\sum_{l=1}^{L} h^{(l)} \tag{2.13}$$

2.5.3 Estimação do erro da estimativa do índice de pobreza

No caso do método Bayesiano uma estimativa da variância dos índices de pobreza pode ser obtida da amostra da distribuição a posteriori do índice. Neste caso, $\hat{\text{Var}}\left(\hat{F}_m^{HB}(\alpha, z)\right)$ será o estimador de $\text{Var}\left(\hat{F}_m^{HB}(\alpha, z)\right)$ onde

$$\hat{\text{Var}}\left(\hat{F}_m^{HB}(\alpha, z)\right) = \frac{1}{L}\sum_{l=1}^{L}\left(h^{(l)} - \bar{h}\right)^2$$

e $\bar{h} = \frac{1}{L}\sum_{l=1}^{L} h^{(l)}$.

Uma vantagem deste método é não ser necessário supor normalidade para a distribuição a posteriori do índice de pobreza a fim de se obter um intervalo de confiança. Pode-se utilizar a mesma amostra $h^{(1)}, \ldots, h^{(L)}$ obtida de (2.12) e estimar os quantis de $2,5\%$ e $97,5\%$. Assim tem-se um intervalo que contém o valor populacional do índice de pobreza com 95% de probabilidade.

3 Comparação simulando população simples

Neste capítulo serão comparados os vários estimadores utilizando dados gerados de um modelo de superpopulação com parâmetros conhecidos. Na primeira seção, este modelo é apresentado e explica-se a razão pela qual decidiu-se pelos valores utilizados para os parâmetros. Na segunda seção descreve-se a amostra que foi retirada das populações geradas. Nas seções subsequentes são implementados os estimadores apresentados no capítulo 2, a saber ELL, EB e HB, e posteriormente é feita uma comparação entre eles.

Para a comparação entre os estimadores utilizando o modelo de superpopulação foram geradas $I = 1000$ populações distintas e de cada uma delas retirada uma amostra. Para cada população e sua respectiva amostra foram calculados os índices de pobreza para cada área m usando toda a população ($F_m(\alpha, z)$) e, usando apenas a amostra, os estimadores diretos ($\hat{F}_m(\alpha, z)$) e os estimadores de pequenas áreas apresentados ($\hat{F}_m^{\mathrm{ELL}}(\alpha, z)$, $\hat{F}_m^{\mathrm{EB}}(\alpha, z)$ e $\hat{F}_m^{\mathrm{HB}}(\alpha, z)$).

3.1 População simulada

A população foi simulada segundo um modelo hierárquico de dois níveis da mesma forma que aquele apresentado em (2.1) mas usando somente duas variáveis explicativas:

$$Y_{mj} = \beta_0 + \beta_1 X_{1,mj} + \beta_2 X_{2,mj} + u_m + e_{mj} \tag{3.1}$$

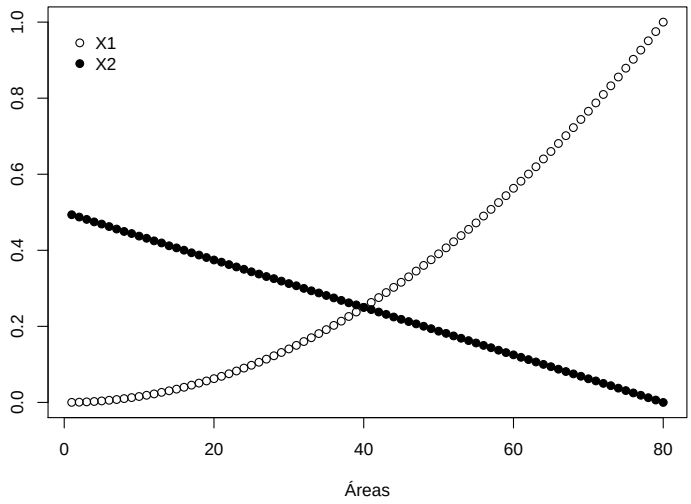
Foram simulados dados para 80 áreas com 250 unidades em cada área, indexadas por m e j respectivamente. As variáveis explicativas também foram simuladas de acordo com uma distribuição de Bernoulli onde os parâmetros das distribuições variam de acordo com o índice da área:

$$X_{1,mj} \sim Ber((m/80)^2) \qquad\qquad X_{2,mj} \sim Ber\left(\frac{(1 - m/80)}{2}\right)$$

Tem-se, por conta disso, que a média por área de X_1 cresce quadraticamente e a de X_2 decresce linearmente com o índice da área. Pode-se ver este comportamento na figura 3.1 onde é apresentada a expectativas dos valores de X_1 e X_2 por área.

Estudos anteriores mostram uma grande disparidade na Incidência de pobreza por município no Brasil. Existem municípios onde menos de 5% dos domicílios são considerados pobres enquanto há outros onde mais da metade dos domicílios estão abaixo da linha

Figura 3.1: Média das variáveis explicativas, X_1 e X_2, por área



de pobreza. Os valores de $\boldsymbol{\beta}' = (\beta_0, \beta_1, \beta_2) = (3, 1, -1.2)$ foram escolhidos levando em consideração esse comportamento de maneira que os índices de pobreza para o modelo de superpopulação representassem essa realidade. Na figura 3.2 são apresentadas as médias dos índices de pobreza por área.

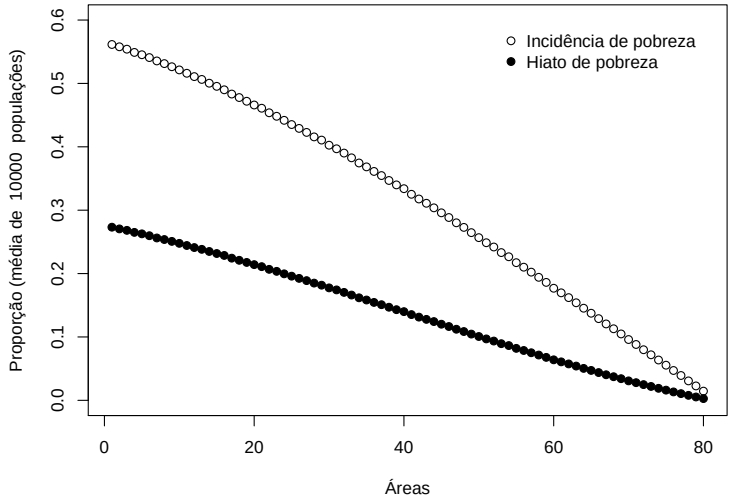


Figura 3.2: Média dos índices de pobreza por área

Como a variável que se pretende utilizar para a modelagem é o logaritmo da Renda que, em geral, apresenta distribuição aproximadamente normal foi escolhida esta distribuição para os erros, nomeadamente

$$u_m \sim N(0, \sigma_u^2) \qquad e_{mj} \sim N(0, \sigma_e^2)$$

$$\sigma_u = 0.06 \qquad \sigma_e = 0.65$$

Para escolher esses valores para os parâmetros de variância foi ajustado um modelo com efeito aleatório aos dados do Censo 2000 de Minas Gerais. Como não era o objetivo no momento realizar um estudo das variáveis disponíveis para análise foi decidido simplesmente ajustar o modelo hierárquico com as 5 variáveis do Censo com maior poder de explicação do logaritmo da renda domiciliar *per capita* e considerando o município como a variável que define os grupos do modelo.

Depois foi realizada uma análise deste modelo ajustado e verificou-se qual a proporção da variância total dos dados era explicada pelos efeitos fixos, pela variância entre os municípios e pela variância das unidades. Os valores escolhidos para o modelo de superpopulação visam representar estas mesmas proporções nas populações simuladas.

3.2 Definição da amostra realizada

A amostra utilizada neste caso foi uma amostra aleatória estratificada em um estágio. Cada área foi considerada um estrato e dentro de cada área foram selecionadas 5 unidades usando amostragem aleatória simples sem reposição. De maneira análoga às definições da seção 2.1, s_m é a amostra de unidades da área m e $s = \cup_{m=1}^{80} s_m$.

3.3 Estimação dos parâmetros do modelo

A partir dos dados da amostra foram testados 3 métodos para estimar os parâmetros do modelo hierárquico apresentado em (3.1), máxima verossimilhança restrita (REML) implementado pelo pacote *lme4*, mínimos quadrados iterativos generalizados (IGLS) como apresentados por Pffefermann et al. (1998) e Monte Carlo via cadeias de Markov (MCMC) com amostrador de Gibbs como implementado pelo OpenBUGS (Lunn et al., 2009).

No caso do estimador Bayesiano a estimativa pontual apresentada é a média de uma amostra de tamanho 500 da distribuição a posteriori estimada dos parâmetros.

Verifica-se nas tabelas 3.1 e 3.2 que os estimadores dos parâmetros fixos foram eficientes em todos os métodos analisados. Os estimadores de variância por IGLS apresentaram resultado muito ruim, não sendo este método uma escolha muito adequada para realizar a estimação pontual.

No caso de σ_u, todos os estimadores apresentam algum problema. A razão mais provável para este problema na estimação de σ_u é sua pouca influência na variância total

Tabela 3.1: Média das estimativas das amostras das 1000 populações simuladas

	Real	REML	IGLS	MCMC
β_0	3.00	3.0018	3.0018	3.0067
β_1	1.00	0.9981	0.9982	0.995
β_2	-1.20	-1.1989	-1.199	-1.206
σ_u	0.06	0.0093	0.008	0.0918
σ_e	0.65	0.6502	0.4192	0.6438

Tabela 3.2: Erro quadrático médio relativo das estimativas das amostras para 1000 populações simuladas (x100)

	REML	IGLS	MCMC
β_0	0.073	0.073	0.0747
β_1	0.537	0.5367	0.5424
β_2	0.4827	0.4825	0.4846
σ_u	7.2973	4.7017	3.0718
σ_e	0.0913	8.3716	0.0924

dos dados, o que acaba sendo difícil de identificar quando se tem uma amostra pequena. Na verdade, é reconhecida na literatura de modelos hierárquicos a dificuldade em estimar efeitos de área (e.g. Pffefermann et al., 1998; Draper e Browne, 2000; Corrêa, 2008).

Por conta da ineficiência do IGLS ao estimar tanto σ_u como σ_e o REML foi o método escolhido para ser utilizado no ELL e no EB.

3.4 Estimação pontual dos índices de pobreza

Para cada uma das 1000 populações finitas simuladas a partir do modelo de superpopulação apresentado em (3.1) foi calculado o valor populacional dos índices de pobreza para cada uma das 80 áreas. Em seguida foi retirada uma amostra de cada população simulada seguindo o desenho amostral apresentado na seção 3.2.

A partir de cada amostra foram estimados os índices de pobreza para as 80 áreas de acordo com os métodos apresentados no capítulo 2 e também por meio do estimador direto. Na figura 3.3 são apresentadas as médias dos valores de Incidência de pobreza obtidos das amostras de cada população simulada por áreas e na figura 3.4 as médias dos

valores de Hiato de pobreza obtidos das amostras de cada população simulada.

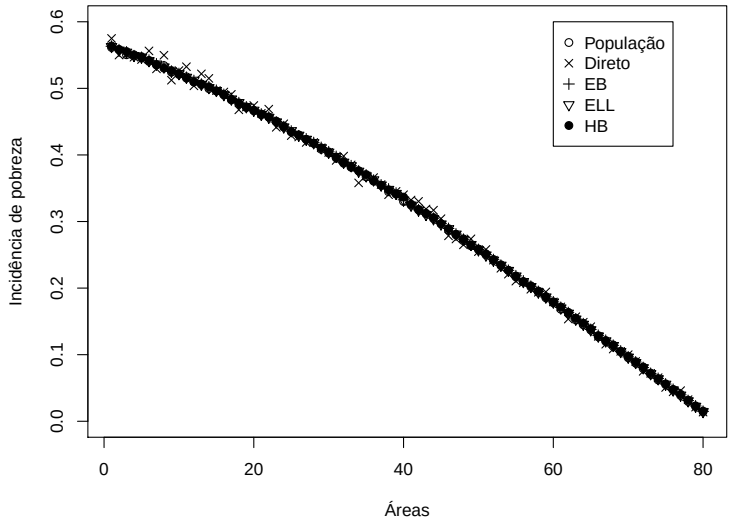


Figura 3.3: Incidência de pobreza por índice das áreas

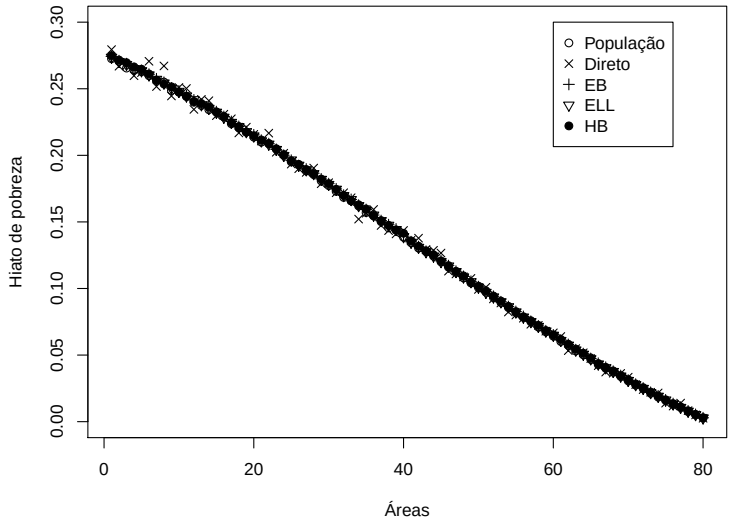


Figura 3.4: Hiato de pobreza por índice das áreas

Em ambos os casos observa-se que as médias dos estimadores foram praticamente iguais às médias dos valores populacionais dos índices de pobreza. Portanto, tem-se evidência de que com um grande número de simulações os estimadores apresentados tendem a se igualar com os valores reais dos índices de pobreza.

3.5 Erros de estimação

A importância do resultado apresentado pelas figuras 3.3 e 3.4 é inegável, mas em uma situação real existe somente uma realização da população finita e uma amostra observada. De maior importância é analisar a precisão dos estimadores, o que será feito através do erro quadrático médio das estimativas.

3.5.1 Erros verdadeiros da estimação

Estima-se o valor do erro quadrático médio verdadeiro dos estimadores através da expressão (3.2). Obviamente tanto o estimador do índice de pobreza quanto o seu valor populacional são calculados com os dados da i-ésima população, mas esta dependência de i não fica evidenciada na fórmula. Os resultados são apresentados nas figuras 3.5 e 3.6.

$$\hat{\mathrm{EQM}}_m = \frac{1}{I}\sum_{i=1}^{I}\Big(\hat{F}_m(\alpha, z) - F_m(\alpha, z)\Big)^2 \tag{3.2}$$

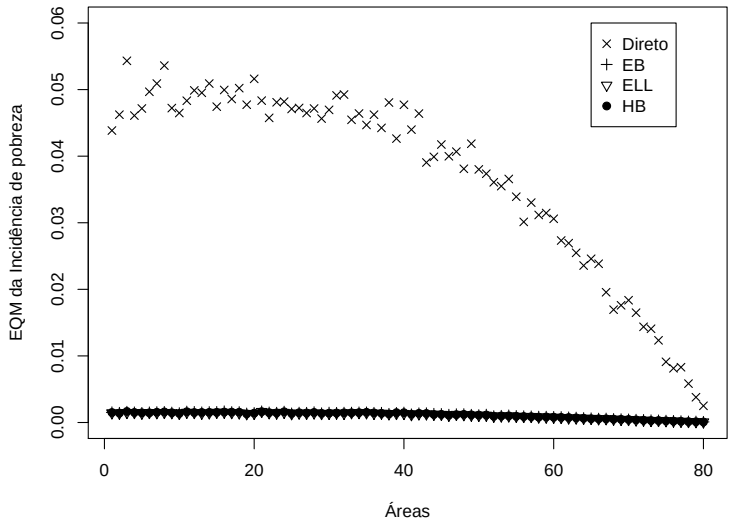


Figura 3.5: Erro quadrático médio da Incidência de pobreza por índice das áreas

Como já era esperado o estimador direto da amostra é menos preciso que os estimadores de pequenas áreas para ambos os índices, chegando a apresentar erro quadrático médio 20 vezes maior para as áreas mais pobres, que são exatamente as que se tem maior interesse em estimar com boa precisão.

Para verificar os erros quadráticos médios dos estimadores de pequenas áreas o erro do estimador direto foi ignorado e os gráficos acima são apresentados novamente nas figuras 3.7 e 3.8.

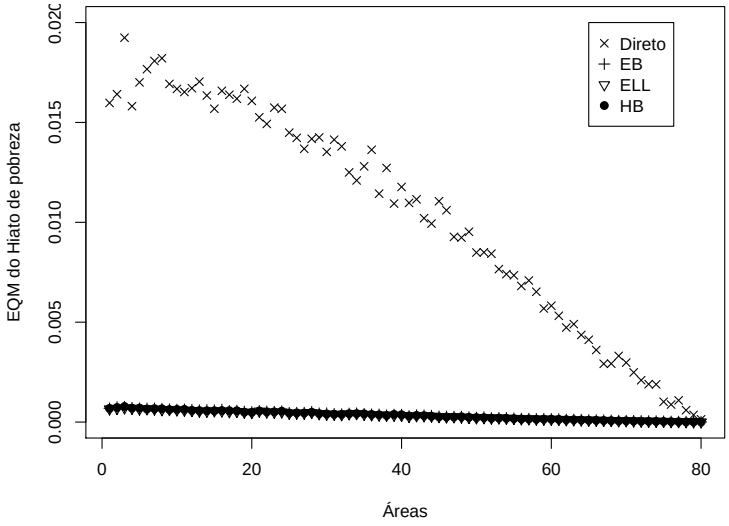


Figura 3.6: Erro quadrático médio do Hiato de pobreza por índice das áreas

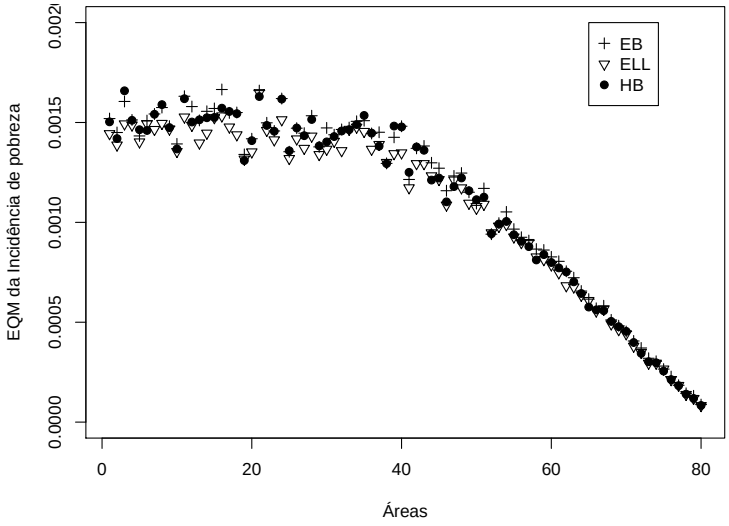


Figura 3.7: Erro quadrático médio da Incidência de pobreza por índice das áreas

Observa-se que não há diferença relevante entre os erros dos estimadores de pequenas áreas, de maneira que para esta simulação pode-se considerá-los igualmente eficientes.

3.5.2 Estimação do erros de estimação

É importante lembrar que em uma aplicação real os valores dos índices calculados com base na população não estão disponíveis e o erro quadrático médio apresentado anteriormente não é calculável. Portanto, é essencial verificar se os métodos de estimação do erro apresentados são eficientes e se trarão resultados confiáveis quando utilizados em dados reais.

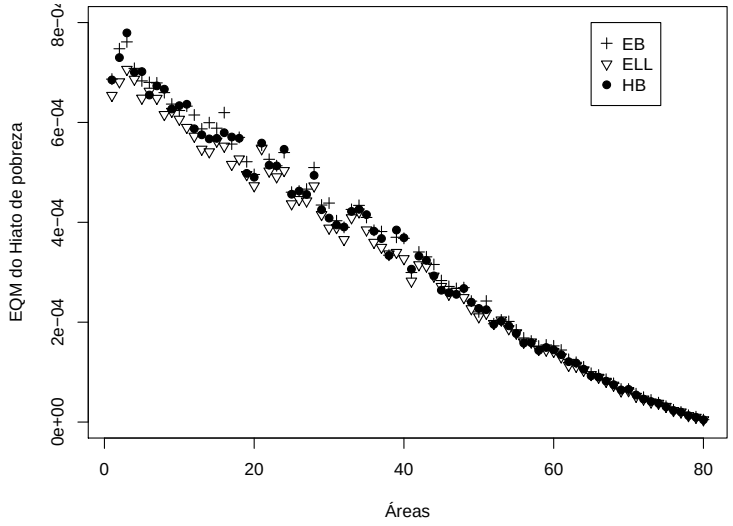


Figura 3.8: Erro quadrático médio do Hiato de pobreza por índice das áreas

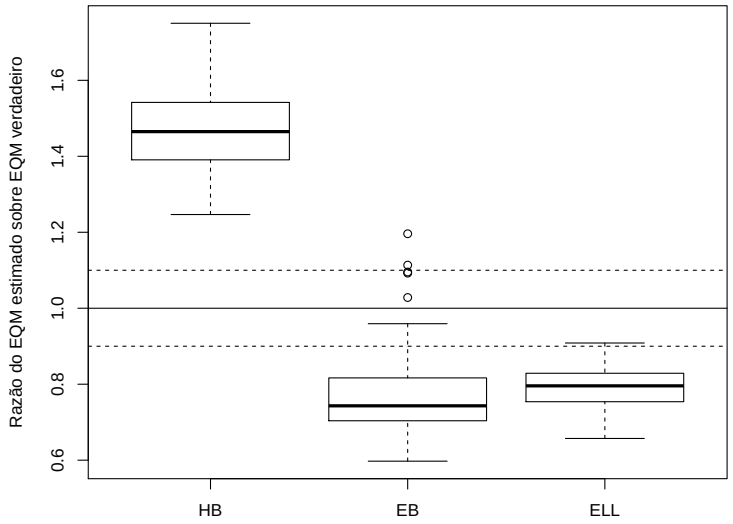


Figura 3.9: Diagrama em caixa da razão da média do EQM estimado sobre o EQM verdadeiro da Incidência de pobreza

Nas figuras 3.9 e 3.10 são apresentados *boxplots* da razão entre a média do EQM estimado sobre o EQM verdadeiro para cada um dos índices de pobreza e para cada método utilizado.

O ideal seria que as razões estivessem próximas de 1, indicando que os EQMs verdadeiros foram bem estimados. No entanto, o método HB superestimou o EQM enquanto os métodos ELL e EB o subestimaram. Note que este desvio da estimação do EQM condiz com o vício encontrado para a estimação de σ_u, (tabela 3.1, pág. 48).

Para verificar se foi realmente o vício de estimação do σ_u que influenciou as estimativas do erro quadrático médio repetiu-se a estimação do EQM no método EB, apresentado

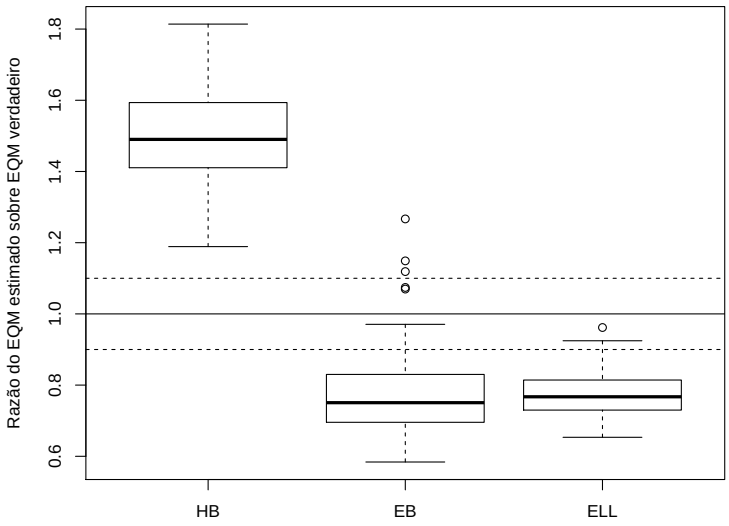


Figura 3.10: Diagrama em caixa da razão da média do EQM estimado sobre o EQM verdadeiro do Hiato de pobreza

na página 40. No entanto, o modelo de superpopulação ξ^* foi construído utilizando $u_m^* \sim N(0, \sigma_u^2)$, o valor verdadeiro da variância de grupo. Desta maneira se exclui a influência que a má estimação de σ_u pode ter na estimação do erro.

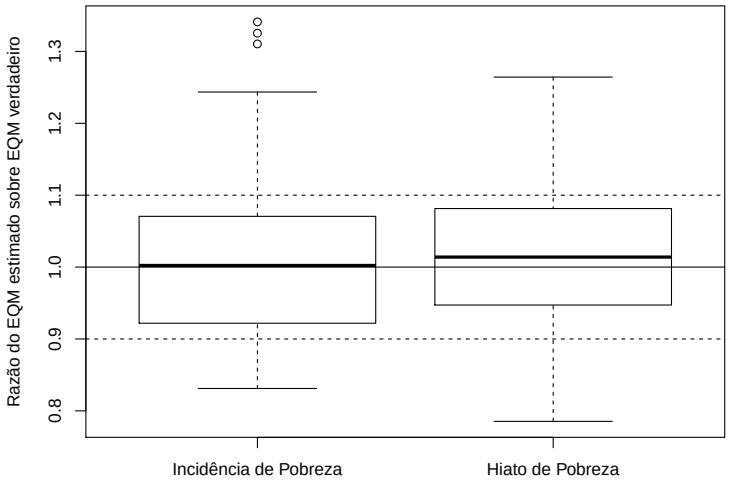


Figura 3.11: EQM verdadeiro por EQM estimado com σ_u verdadeiro

Observando a figura 3.11 percebe-se que tanto o erro quadrático médio da Incidência de pobreza quanto do Hiato de pobreza foram mais bem estimados. No entanto, muitos valores das estimativas têm diferença de mais de 10%, as linhas tracejadas, para os valores verdadeiros do EQM.

3.6 Conclusão

Para o modelo de superpopulação e método de amostragem considerados neste capítulo, os 3 métodos foram igualmente eficientes ao estimar os valores dos índices de pobreza para as áreas. Os problemas de estimação de σ_u não pareceram afetar significativamente o desempenho das estimativas pontuais.

Era esperado que os métodos EB e HB apresentassem melhor resultado do que o ELL por conta da falha deste último ao considerar as informações da amostra, como descrito na subseção 2.3.2. Esta vantagem não foi verificada, mas possivelmente porque a amostra para cada área era muito pequena.

Todos os métodos apresentaram problemas ao estimar o erro das estimativas pontuais. Verificou-se que esta deficiência ao estimar o erro é em grande parte resultado da má estimação de σ_u. Isto só reforça a importância de ao trabalhar com dados reais verificar se o modelo foi bem ajustado e se os parâmetros foram bem estimados.

4 Comparação simulando população complexa

O objetivo neste capítulo é realizar testes semelhantes aos que já foram apresentados no capítulo 3 mas utilizando uma população e um plano amostral mais parecidos com os utilizados pelas pesquisas do IBGE.

Na primeira seção apresentamos uma análise mais profunda dos dados da população obtidos pelo Censo Demográfico 2000, focando principalmente nas características que foram incorporadas à simulação. Na segunda seção é apresentado o plano amostral que será considerado e discutidas algumas de suas características. Na terceira seção são apresentados os estimadores pontuais associados aos métodos de estimação em pequenas áreas. Na quarta seção comparamos os erros de estimação de cada método assim como a eficiência em estimar estes erros.

Para a comparação entre os estimadores foi utilizado o modelo de superpopulação gerando-se $I = 3000$ populações distintas e de cada uma delas retirou-se uma amostra, conforme o plano descrito na seção 4.2. Aumentou-se o número de populações simuladas devido ao aumento de complexidade da população e as restrições impostas ao desenho da amostra.

4.1 População simulada

Neste caso a população também foi simulada segundo um modelo hierárquico de dois níveis, mas com três variáveis explicativas. Nesta população simulada consideramos áreas como classificadas em rurais e urbanas e o número de unidades em cada área é aleatório com distribuição Gamma. O modelo se superpopulação considerado foi:

$$Y_{mj} = \beta_0 + \beta_1 X_{1,mj} + \beta_2 X_{2,mj} + \beta_3 X_{3,mj} + u_m + e_{mj} \tag{4.1}$$

Desta maneira $\boldsymbol{X}_3$ é uma variável indicadora que vale 1 se o indivíduo pertence a uma área urbana ou 0 se pertence a uma área rural e β_3 representa a diferença no logaritmo da renda por conta desta situação. Foram simulados dados para $m = 1, \ldots, 160$ áreas com $j = 1, \ldots, N_m$ unidades em cada área. As variáveis explicativas foram simuladas com distribuições semelhantes às da população simples.

$$X_{1,mj} \sim Ber((m/160)^2) \qquad X_{2,mj} \sim Ber\left(\frac{(1-m/160)}{2}\right)$$

Os parâmetros utilizados nos modelos tem os mesmos valores utilizados anteriormente com a adição de um parâmetro para a situação da área, ou seja

$$\beta_0 = 3, \qquad \beta_1 = 1, \qquad \beta_2 = -1,2, \qquad \beta_3 = 0,4$$

Devido à adição de mais um parâmetro fixo, as variâncias do modelo hierárquico ajustado sofreram pequenas alterações. Para manter as proporções devidas dos erros de áreas e dos erros individuais os valores dos parâmetros foram definidos em:

$$u_m \sim N(0, \sigma_u^2) \qquad e_{mj} \sim N(0, \sigma_e^2)$$
$$\sigma_u^2 = 0.1 \qquad \sigma_e^2 = 0.65$$

4.1.1 Áreas rurais e urbanas

A classificação das áreas em rurais e urbanas é parte da construção deste novo modelo de superpopulação apresentado em (4.1) devido a sua relação com a renda domiciliar e com o tamanho dos setores censitários. A seguir será mostrado como estas duas variáveis se comportam de maneira distinta em cada tipo de setor e como esta diferença será tratada.

A renda domiciliar *per capita* é muito diferente entre as áreas rurais e urbanas. Sendo Renda a variável principal deste estudo, quaisquer características que auxiliem a predição de seu comportamento são importantes.

Nas figuras 4.1 e 4.2 estão representadas as distribuições do logaritmo da renda para os setores urbanos e rurais. A partir da análise das médias para os setores rurais e urbanos, o valor de β_3 foi escolhido de maneira e representar no modelo esta mesma diferença de maneira proporcional.

O tamanho das áreas varia muito também conforme esta característica, mas esse comportamento deriva diretamente de regras estabelecidas pelo IBGE. Como a densidade demográfica em áreas urbanas é maior que em áreas rurais, para que os setores nas áreas rurais não fiquem muito grandes é definido que estes tenham em média 150 domicílios por setor. Em contrapartida, os setores nas áreas urbanas tem 250 domicílios em média.

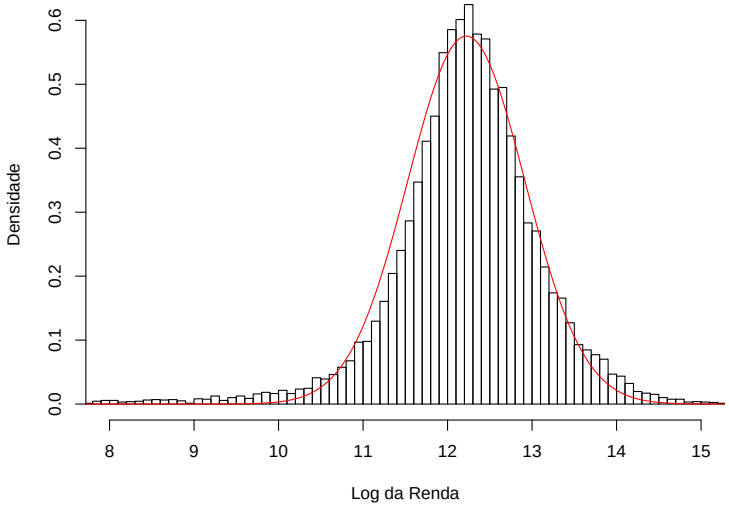


Figura 4.1: Histograma do log da renda domiciliar nos setores urbanos

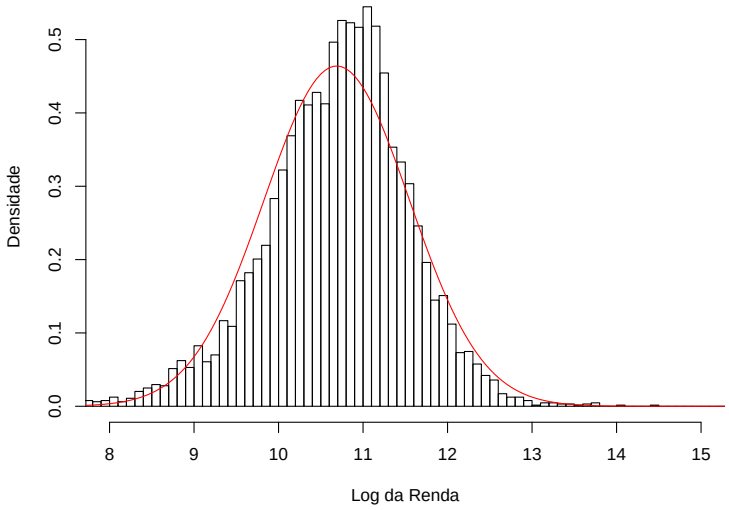


Figura 4.2: Histograma do log da renda domiciliar nos setores rurais

Pelos histogramas apresentados nas figuras 4.3 e 4.4 pode-se verificar que o tamanho dos setores é diferente para os setores urbanos e rurais, com médias 250 e 115 respectivamente. Optou-se por utilizar a distribuição Gamma para gerar os tamanhos dos setores, ainda que o ajuste não seja muito bom no caso dos setores urbanos. Nas simulações foram utilizadas como distribuição do tamanho dos setores:

$$N_{\text{Urb}} \sim Gamma\left(\frac{250^2}{5800}, \frac{250}{5800}\right) \qquad N_{\text{Rur}} \sim Gamma\left(\frac{115^2}{4130}, \frac{115}{4130}\right)$$

Os parâmetros foram escritos da maneira apresentada para deixar evidente os valores das médias e variâncias das distribuições.

Como a amostra da PNAD, que será utilizada na aplicação com dados reais, seleciona

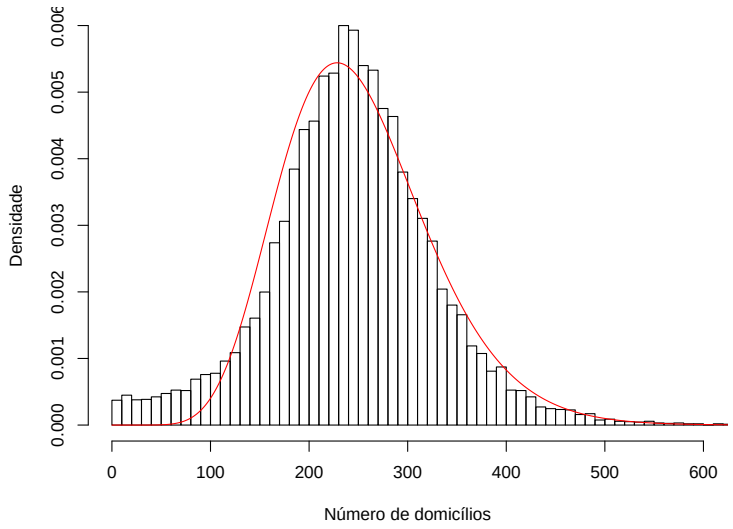


Figura 4.3: Histograma do tamanho dos setores urbanos

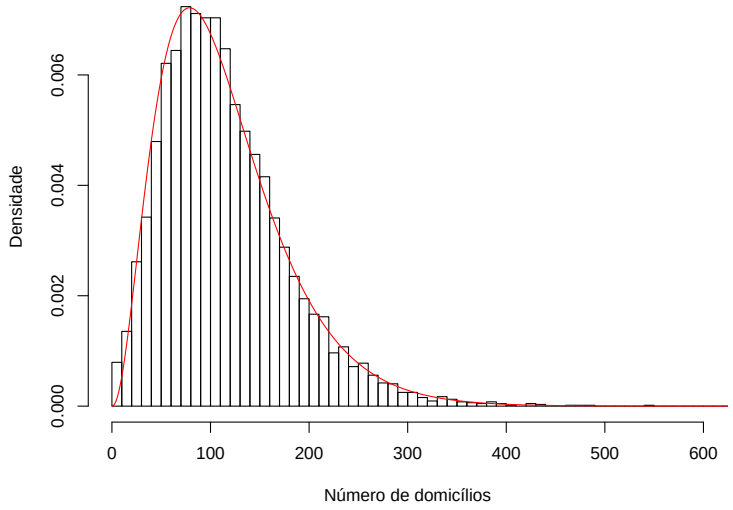


Figura 4.4: Histograma do tamanho dos setores rurais

os setores com probabilidade proporcional ao tamanho, foi considerado importante refletir o comportamento do tamanho dos setores nas populações simuladas.

4.2 Definição do plano amostral

As pesquisas domiciliares do IBGE têm um desenho amostral complexo, incluindo estratificação, conglomeração e seleção de setores com probabilidade desigual. Várias destas características são necessárias quando se planeja uma amostra probabilística para um país tão extenso quanto o Brasil, mas trazem complicações para calcular as estimativas pontuais e suas variâncias.

Com o intuito de testar quão bem os métodos apresentados estimam os indicadores de pobreza quando usada uma amostra complexa, o desenho amostral utilizado nesta simulação foi alterado de maneira a se assemelhar àqueles utilizados pelo IBGE.

Foi realizada uma estratificação das áreas em duas etapas. Primeiro foram criados dois estratos, um com as áreas rurais e outro com as áreas urbanas. Depois as áreas urbanas foram divididos em 3 e as áreas rurais em 2 estratos de renda. Foi utilizada como variável de estratificação a renda domiciliar média por área e os estratos foram definidos usando o pacote *stratification* (Baillargeon e Rivest, 2010) do R que usa o método de Lavallée e Hidiroglou (1988).

Foi retirada uma amostra de 40 áreas e o número de áreas amostradas em cada estrato foi decidido por alocação ótima de Neyman. Dentro de cada área amostrada foram selecionadas 25 unidades. No total foi obtida uma amostra de 1000 unidades de uma população de aproximadamente 33 mil unidades.

4.2.1 Estimação em amostras complexas

É importante, em um desenho amostral como este, levar em consideração não só a probabilidade de seleção das unidades amostradas como também os vários níveis de seleção antes de se chegar à unidade básica.

Cochran (1977) e Pessoa e Silva (1998) mostram que desconsiderar os pesos de uma amostra complexa pode viciar as estimativas pontuais dos parâmetros e ignorar as características de estratificação e conglomeração da amostra pode viciar as estimativas de variância destes estimadores.

No entanto, no caso de estimação de parâmetros de modelos hierárquicos, não há um consenso sobre como considerar os pesos ou outras características do plano amostral. É possível, inclusive, que para dois modelos diferentes que usem os mesmos dados um deles seja influenciado pelo plano amostral e o outro não.

Decidiu-se por testar o método de estimação dos parâmetros do modelo proposto por citePfeffermann1998, que incorpora a probabilidade de seleção no primeiro e segundo estágio de seleção, e compará-lo com o métodos de máxima verosimilhança restrita que assume equiprobabilidade das observações. Também testou-se o método de pseudo máxima verossimilhança que leva em consideração a probabilidade de seleção das unidades.

4.3 Estimação dos parâmetros do modelo

Os parâmetros do modelo hierárquico foram estimados utilizando três métodos clássicos. Primeiro, o de máxima verossimilhança restrita (REML), que não considera o plano amostral, implementado no pacote *lme4* (Bates e Maechler, 2010); segundo, o de pseudo máxima verossimilhança (PML), que considera os pesos das unidades; terceiro, o método de mínimos quadrados iterativos generalizados com pesos probabilísticos (PWIGLS), uma extensão do IGLS, também apresentado por Pffefermann et al. (1998), que considera a probabilidade de seleção da área e a probabilidade de seleção da unidade dentro da área.

Para o estimador Bayesiano foi usado novamente MCMC com amostrador de Gibbs implementado no OpenBUGS (Lunn et al., 2009). Neste caso o estimador pontual foi calculado por meio da a média de uma amostra de 1000 observações da distribuição a posteriori dos parâmetros.

Tabela 4.1: Média das estimativas para as amostras das 1000 populações simuladas

	Real	REML	PWIGLS	PML	MCMC
β_0	3,00	3,0109	3,0058	3,0050	3,0117
β_1	1,00	1,0005	1,0010	1,0006	0,9993
β_2	-1,20	-1,2032	-1,2025	-1,2025	-1,2030
β_3	0,40	0,3961	0,3938	0,3978	0,3968
σ_u	0,10	**0,0824**	**0,0770**	**0,0170**	**0,0933**
σ_e	0,65	0,6513	0,6500	0,6492	0,6521

Tabela 4.2: Erro quadrático médio relativo das estimativas para as amostras das 1000 populações simuladas (x100)

	REML	PWIGLS	PML	MCMC
β_0	0,1276	0,1800	0,1441	0,1282
β_1	0,2086	0,2818	0,2407	0,2069
β_2	0,2408	0,3283	0,2701	0,2400
β_3	0,9802	1,4208	1,1398	0,9784
σ_u	**2,9086**	**2,3047**	**6,9219**	0,8519
σ_e	0,0346	0,0510	0,0416	0,0340

Verifica-se pelas tabelas 4.1 e 4.2 novamente uma maior dificuldade em estimar os

valores de σ_u. Entre os três estimadores clássicos o PML foi o pior e será descartado da análise. Entre o PWIGLS e o REML o primeiro teve um vício maior na estimação de σ_u mas um erro quadrático médio menor. Para os outros parâmetros, o REML teve um erro menor mas não o suficiente para justificar descartar o PWIGLS.

Antes de tomar uma decisão final entre o REML e o PWIGLS decidiu-se testar os dois métodos alterando o número de áreas na população e na amostra. Foram testadas populações com 160, 320, 480, 640 e 800 áreas e amostras com 40, 80, 120, 160, 200, 240, 280, 320, 360 e 400 áreas, ressaltando que a amostra nunca ultrapassou metade das áreas da população. As tabelas completas podem ser vistas no apêndice B.1.

Para cada um dos 30 pares de tamanho de população e amostra foram geradas 1000 populações segundo o modelo em (4.1) e foi seguido o mesmo processo amostral já descrito. Comparou-se as estimativas de σ_u e o REML teve um vício menor que o PWIGLS em todos os casos e um erro quadrático médio menor em 23 dos 30 casos.

De fato, uma característica importante da amostra apresentada na seção 4.2 é ser informativa somente no primeiro estágio de seleção, a seleção das áreas. A seleção das unidades dentro das áreas é por amostragem aleatória simples, logo não há variação nos pesos das unidades de uma mesma área. No próprio artigo onde Pfeffermann apresenta este estimador ele comenta que o ganho para amostras informativas somente no primeiro nível é mínimo e Pfeffermann, Silva e Moura (2006) apresentam resultados semelhantes.

Baseando-se nos resultados apresentados, conclui-se que o REML é mais eficiente nos casos onde a amostra realizada é informativa somente no primeiro nível de seleção. Por ser este o caso nesta simulação optou-se por utilizar este estimador para os parâmetros do modelo hierárquico.

4.4 Estimação pontual dos índices de pobreza

Para cada uma das 3000 populações e amostras simuladas foram calculados os estimadores de pobreza segundo os métodos apresentados. Assim como foi observado na seção 3.4 aqui também as médias das estimativas fornecidas pelos métodos é praticamente igual à média populacional dos índices de pobreza para as áreas, não sendo possível distingui-las nas figuras 4.5 e 4.6.

Pode-se notar uma clara diferença entre a tendência dos índices de pobreza para as

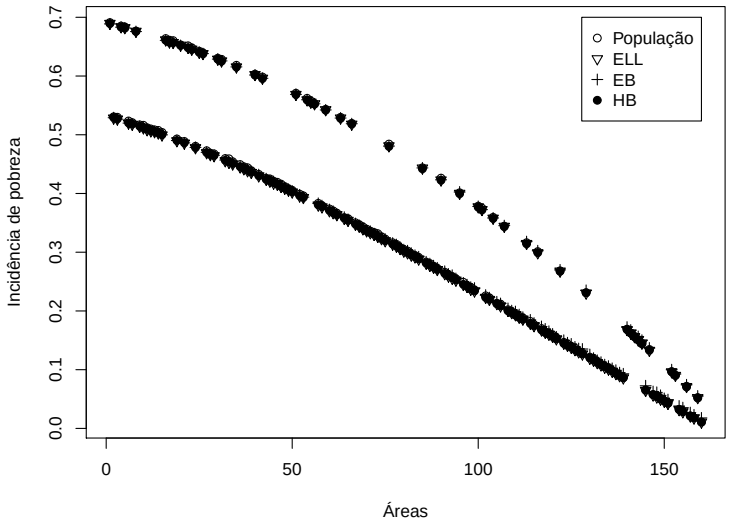


Figura 4.5: Incidência de pobreza por índice das áreas

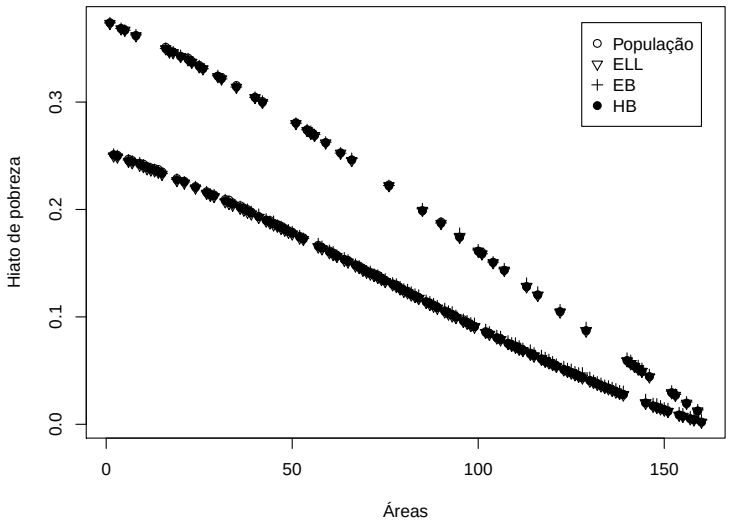


Figura 4.6: Hiato de pobreza por índice das áreas

áreas rurais e urbanas, decorrentes da adição de β_3 ao modelo. Como esperado, as estimativas dos índices de pobreza para as áreas urbanas são menores do que as estimativas para as áreas rurais.

4.5 Erros de estimação

Ao analisar os erros de estimação é importante lembrar que somente 40 das 160 áreas foram amostradas em cada simulação. Como a amostra não foi mantida fixa entre cada réplica da simulação, para uma área m qualquer tanto foi possível estimar seu índice de pobreza usando diretamente os dados da Renda para esta área, quanto foi necessário

utilizar o estimador sintético, que se baseia somente nos parâmetros do modelo. Como é esperado que o erro de estimação seja menor quando a área é amostrada, são analisados separadamente o EQM para quando a área faz parte da amostra e o EQM para quando a área não faz parte da amostra.

Definindo s_i a amostra de áreas da simulação i, $i = 1, \ldots, 3000$, o EQM quando a área está na amostra é definido como

$$\hat{\text{EQM}}_m^D = \frac{1}{\sum_{m \in s_i} 1} \sum_{m \in s_i} \left(\hat{F}_m(\alpha, z) - F_m(\alpha, z) \right)^2$$

e o EQM quando a área está fora da amostra é

$$\hat{\text{EQM}}_m^F = \frac{1}{\sum_{m \notin s_i} 1} \sum_{m \notin s_i} \left(\hat{F}_m(\alpha, z) - F_m(\alpha, z) \right)^2 .$$

4.5.1 Erros verdadeiros da estimação

Nas figuras 4.7 e 4.8 observamos os valores do erros quadráticos médios para a Incidência de pobreza e Hiato de pobreza, respectivamente, quando as áreas estão na amostra.

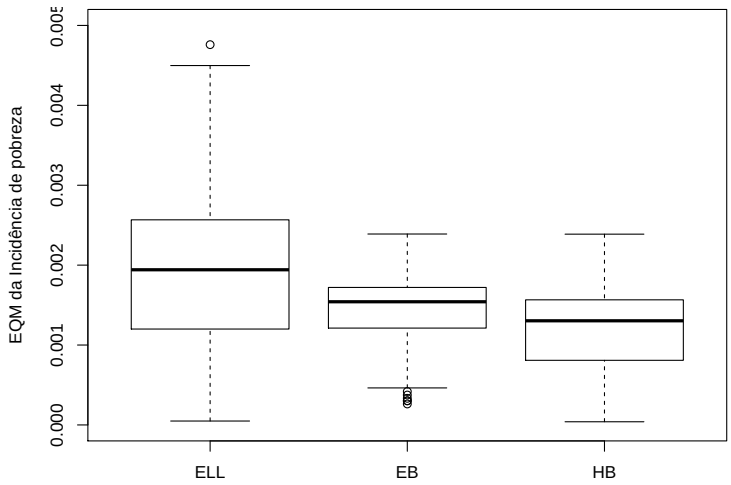


Figura 4.7: EQM da Incidência de pobreza para as áreas na amostra

É fácil perceber que, no geral, o EQM do método HB foi menor que para os outros métodos e que o método EB foi melhor que o método ELL. Mais precisamente, o EQM da Incidência de pobreza pelo método EB é 71% do EQM pelo método ELL enquanto o EQM pelo método HB é 59% do EQM pelo método ELL. Para o Hiato de pobreza o EQM pelo método EB é 76% e pelo método HB é 57% do EQM pelo método ELL.

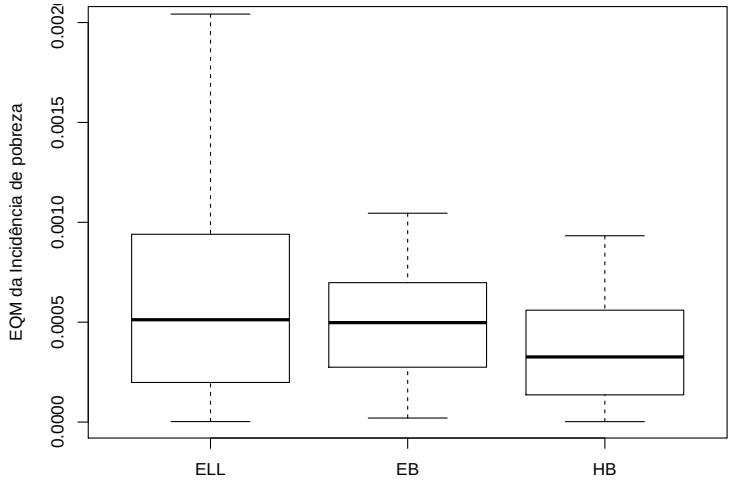


Figura 4.8: EQM do Hiato de pobreza para as áreas fora da amostra

No entanto, o mesmo comportamento não é observado quando se compara o EQM para as áreas fora da amostra. Nas figuras 4.9 e 4.10 não é possível notar nenhuma diferença entre os erros quadráticos médios dos métodos avaliados. Este resultado é porque, independente do método, são utilizados estimadores sintéticos baseados todos no mesmo modelo para as áreas onde não há amostra.

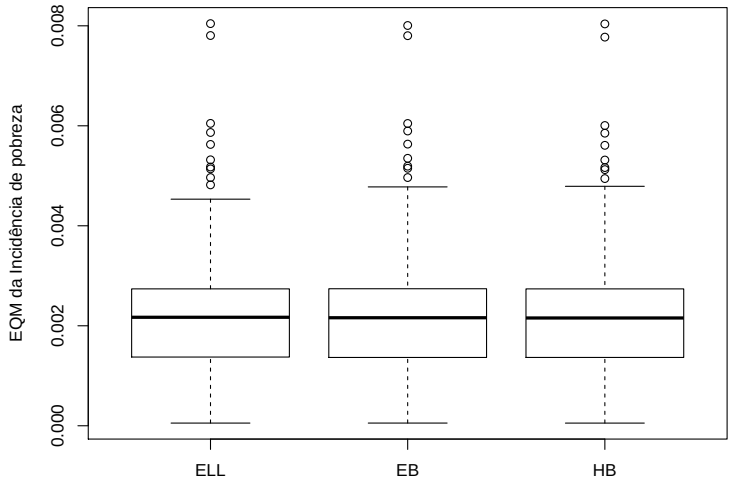


Figura 4.9: EQM da Incidência de pobreza para as áreas na amostra

4.5.2 Estimação dos erros de estimação

Com relação à estimação do erro o resultado encontrado foi bem diferente daquele apresentado no capítulo 3. Para analisar quão bem foi estimado o erro quadrático médio

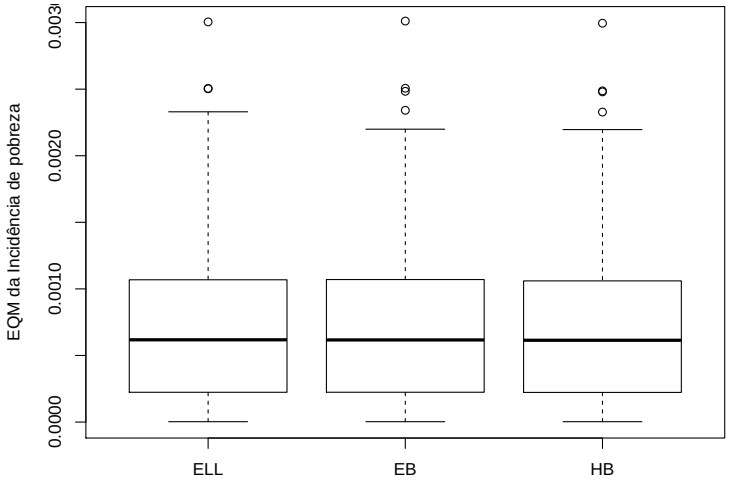


Figura 4.10: EQM do Hiato de pobreza para as áreas fora da amostra

calculamos a razão entre a média das estimativas do EQM e seu valor verdadeiro por área.

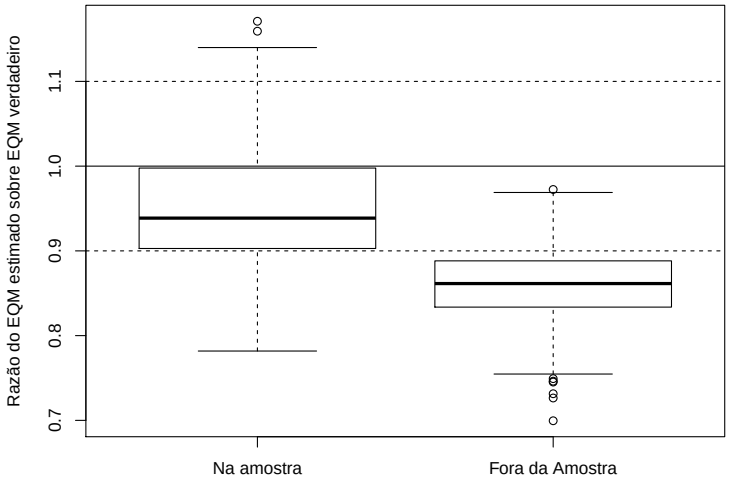


Figura 4.11: Diagrama em caixa da razão do EQM da Incidência de pobreza para áreas na amostra e fora da amostra, respectivamente, segundo o método ELL

Considerando primeiro o método ELL, são apresentadas nas figuras 4.11 e 4.12 as razões das estimativas do EQM pelo seu valor verdadeiro. O ideal seria que todas as estimativas estivessem perto de 1, de preferência contidas na faixa de 10% de erro para mais ou para menos.

Para ambos os índices de pobreza verifica-se que o EQM para as áreas amostradas foi estimado um pouco abaixo do EQM verdadeiro. No entanto, para as áreas onde não havia amostra o erro de estimação foi severamente subestimado, o que pode levar a conclusões

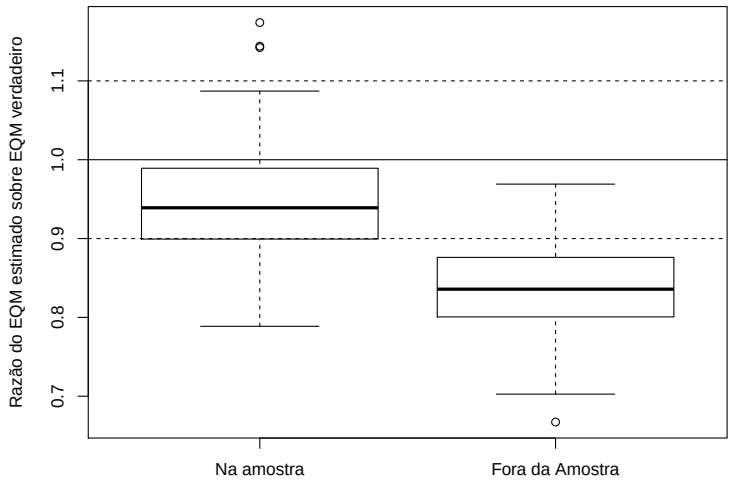


Figura 4.12: Diagrama em caixa da razão do EQM do Hiato de pobreza para áreas na amostra e fora da amostra, respectivamente, segundo o método ELL

erradas sobre a precisão dos estimadores.

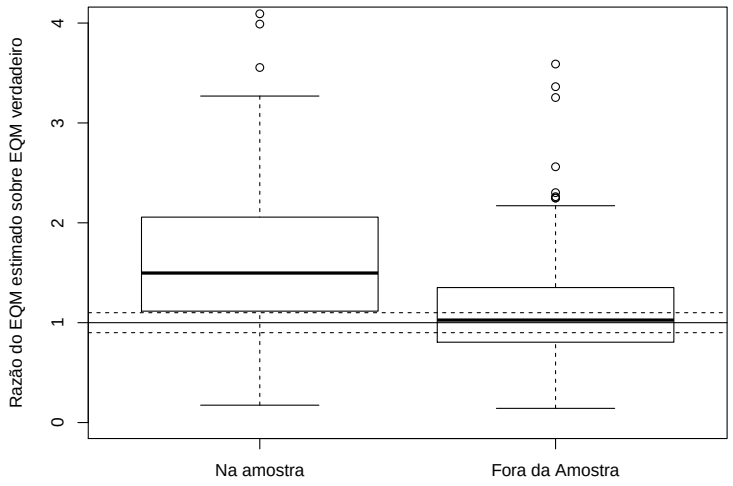


Figura 4.13: Diagrama em caixa da razão do EQM da Incidência de pobreza para áreas na amostra e fora da amostra, respectivamente, segundo o método EB

Considerando o método EB, o resultado ficou aquém do esperado. Percebe-se pelas figuras 4.13 e 4.14 que quando as áreas estão na amostra, o EQM foi muito superestimado, mas ainda assim, pode-se perceber uma grande variação na razão entre o EQM estimado e o verdadeiro.

O método apresentado na subseção 2.4.3 foi implementado utilizando $B = 500$ populações simuladas. É possível que esta estimativa do EQM possa ser melhorada au-

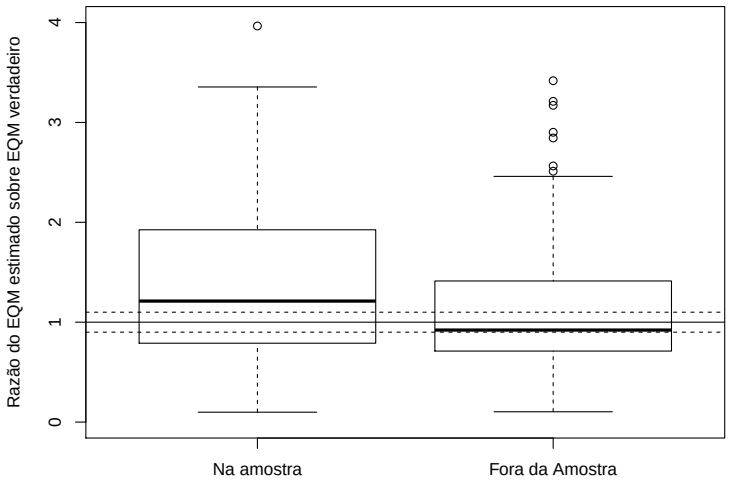


Figura 4.14: Diagrama em caixa da razão do EQM do Hiato de pobreza para áreas na amostra e fora da amostra, respectivamente, segundo o método EB

mentando o valor de B, mas é importante lembrar que estimação por *bootstrap* é computacionalmente muito custosa e o número de réplicas para estimar adequadamente o erro quadrático médio pode ser impraticável. No caso da simulação, onde foi feito um conjunto de $B = 500$ simulações para cada uma das $I = 3000$ populações, esta observação é ainda mais pertinente.

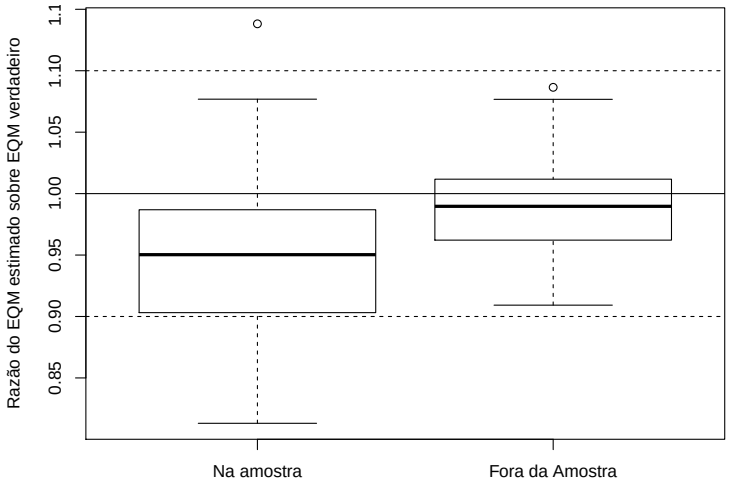


Figura 4.15: Diagrama em caixa da razão do EQM da Incidência de pobreza para áreas na amostra e fora da amostra, respectivamente, segundo o método HB

Considerando o método HB, pode-se perceber pelas figuras 4.15 e 4.16 que a estimação do EQM foi a mais eficiente entre os três métodos. Existe uma pequena subestimação

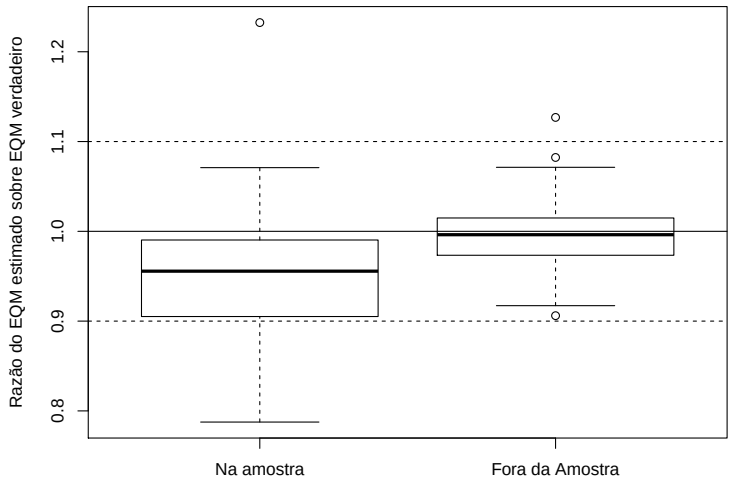


Figura 4.16: Diagrama em caixa da razão do EQM do Hiato de pobreza para áreas na amostra e fora da amostra, respectivamente, segundo o método HB

Tabela 4.3: Cobertura real do intervalo nominal de 95% por índices e métodos

	Inc. de pob.		Hiato de pob.	
	Na amostra	Fora da amostra	Na amostra	Fora da amostra
ELL	91,25	89,43	90,79	88,31
EB	97,12	94,79	94,69	94,06
HB	92,48	92,77	93,83	93,7

no caso das áreas que foram amostradas, mas ainda assim 75% das estimativas ficaram a menos de 10% do valores reais. Já no caso das áreas que ficaram fora da amostra, praticamente todos os EQMs foram estimados próximos do seus valores verdadeiros.

Por fim foi analisada a cobertura de intervalos de confiança nominais de 95% para cada um dos métodos. No caso do ELL e do EB os intervalos foram construídos assumindo normalidade e no caso do HB foram tomados os quantis 2,5% e 97,5% da amostra da distribuição a posteriori dos índices de pobreza.

Os resultados apresentados na tabela 4.3 não são surpreendentes. O método ELL já tinha apresentado uma subestimação dos erros, principalmente quando as áreas não foram amostradas e o HB também já tinha mostrado uma pequena subestimação. No caso do EB, apesar de a cobertura ser mais próxima de 95%, vale lembrar que o erro foi bastante superestimado, o que nos deixa com um intervalo de confiança grande e pouca informação sobre qual pode ser o verdadeiro valor do índice.

4.6 Conclusão

No caso tratado neste capítulo, o método HB foi definitivamente o melhor entre os três. Não só a estimava do índice de pobreza teve um erro quadrático médio menor mas esse erro foi mais bem estimado por este método. É possível que a estimativa do erro do método EB possa ser melhorada, no entanto, com o número de simulações *bootstrap* utilizadas (500) esse método já é aproximadamente 10 vezes mais demorado que a estimação do HB.

De fato, quando utilizado em dados reais não há a necessidade de estimar erros para 3000 populações simuladas, como foi o caso neste exemplo. No entanto, uma pesquisa pode ter em torno de 100 mil domicílios e um censo, dezenas de milhões. Neste caso, esta diferença no tempo de computação pode facilmente se transformar em dias de diferença para encontrar o resultado.

5 Comparação utilizando dados reais

Neste capítulo, os métodos estudados até então, serão aplicados a dados reais produzidos pelo IBGE. Utilizaremos o Censo Demográfico 2000 e as Pesquisas Nacionais por Amostra de Domicílios de 2001, 2002 e 2003. Aplicar estes métodos aos dados de todo o Brasil é uma tarefa maior do que a necessária para comparar a eficiência dos métodos e por conta de limites de tempo e computacionais optou-se por utilizar dados somente de um estado da federação.

Neste caso, a escolha mais indicada para realizar este teste é o estado de Minas Gerais, que tem uma grande extensão territorial e possui grande variação sócio-econômica dentro de sua fronteira. Os municípios do sul do estado, que fazem fronteira com Rio de Janeiro e São Paulo, são influenciados pela economia ativa destes estados e são reconhecidamente mais ricos, enquanto municípios do norte do estado fazem fronteira com o interior da Bahia e são reconhecidamente mais pobres. Por conta desta diversidade, o estado de Minas Gerais é geralmente escolhido para estudos, já tendo sido chamado como "O Brasil dentro do Brasil" (Elbers et al., 2008).

Serão apresentados mais detalhes sobre o Censo e a PNAD, como características das variáveis que compõem as duas pesquisas e seus desenhos amostrais. Depois, serão apresentadas as estimativas para os parâmetros do modelo hierárquico ajustado aos dados, os índices de pobreza estimados para cada município e os erros estimados desses índices. No final os resultados encontrados em cada método são comparados para avaliar o mais eficiente.

5.1 Conjuntos de dados utilizados

Nesta seção serão apresentados alguns detalhes do Censo e da PNAD. Não é intenção explicar todos os detalhes acerca destas pesquisas, sendo apresentadas somente as características julgadas importantes para o estudo aqui apresentado.

Para maiores detalhes sobre estas pesquisas, recomenda-se ao leitor interessado consultar as notas técnicas do Censo Demográfico 2000, publicadas pelo IBGE (2003a), e as sínteses de indicadores das PNAD onde também constam as notas técnicas das pesquisas, também publicadas pelo IBGE (2001, 2002, 2003b).

5.1.1 Censo Demográfico 2000

A população alvo do Censo é muito abrangente, incluindo alguns tipos de domicílios que não são úteis para o estudo aqui apresentado. Podemos citar como exemplos, domicílios coletivos, como quartéis, prisões e hospícios e domicílios improvisados, como fábricas, grutas, espaço sob pontes ou viadutos. Na verdade, foram tomados como fonte de dados neste estudo, somente os domicílios particulares permanentes ocupados.

Entre os domicílios investigados, uma outra diferença importante são as variáveis disponíveis. O Censo possui dois questionários aplicados aos domicílios; um deles, chamado de questionário universo, é mais curto e aplicado a uma grande parcela da população. O outro, chamado de questionário da amostra, investiga as mesmas variáveis do questionário universo além de outras informações acerca do domicílio e seus moradores e é aplicado a uma parcela menor da população.

A parcela que respondeu o questionário da amostra foi selecionada por amostragem sistemática dentro de cada setor. Em municípios com menos de 15 mil habitantes, um domicílio a cada cinco respondeu o questionário da amostra, gerando uma fração amostral de aproximadamente 20%. Em municípios com mais de 15 mil habitantes, um domicílio em cada dez respondeu ao questionário maior, gerando uma fração amostral de aproximadamente 10%.

Para este estudo, o questionário universo é muito simples e as variáveis de que ele dispõe são muito poucas para que se possa ajustar um bom modelo que explique a renda domiciliar *per capita*. Desta maneira optou-se por utilizar os dados do questionário da amostra do Censo 2000, através do qual são obtidas informações principalmente sobre o domicílio e a educação e renda de seus moradores.

5.1.2 Pesquisa Nacional por Amostra de Domicílios

A população alvo da PNAD se limita desde o princípio a moradores em domicílios particulares permanentes ocupados ou em unidades domiciliares em domicílios coletivos. Como os domicílios coletivos são raros na amostra da PNAD e suas características não permitem uma análise apropriada da relação entre renda e as variáveis auxiliares, utilizamos somente os dados referentes aos domicílios particulares permanentes ocupados.

Os moradores entrevistados pela PNAD respondem um questionário extenso com perguntas sobre o domicílio, renda, educação, fecundidade, migração e outros assuntos que

variam de ano para ano. Como é necessário comparar a informação adquirida através da PNAD com a que foi obtida no Censo, os dados do questionário da PNAD usados se limitam às partes sobre o domicílio, renda e educação.

A amostra da PNAD é mais complexa que a do Censo. Dentro de um estado, primeiro os municípios são divididos em autorrepresentativos, com probabilidade 1 de pertencer à amostra, e os outros municípios são divididos em estratos. Estes outros municípios são então selecionados com probabilidade proporcional à população residente. No segundo estágio setores censitários são selecionados dentro dos municípios, também com probabilidade proporcional à população residente nos setores. No último estágio, os domicílios são selecionados com equiprobabilidade dentro dos setores censitários selecionados.

5.1.3 Comparação das variáveis explicativas

O primeiro passo para aplicar um dos métodos de estimação em pequenas áreas com os dados da PNAD e do Censo foi determinar quais variáveis poderiam ser usadas para explicar a renda e fazer a ligação entre as duas bases de dados.

A primeira etapa desta comparação foi temática. Analisando os questionários de ambas as pesquisas, registrou-se quais perguntas eram semelhantes em sua formulação e nas possibilidades de resposta e quais simplesmente não existiam em ambas as pesquisas e não poderiam ser utilizadas no teste.

Mesmo entre aquelas que constavam em ambos os questionários, nem todas as perguntas podiam ser diretamente comparadas. Em algumas foi necessário agrupar categorias das respostas, como por exemplo na pergunta sobre escoamento sanitário. Enquanto a PNAD faz diferença entre fossa séptica com ou sem ligação com a rede de esgoto, o Censo somente questionava sobre a existência de fossa séptica.

Em outros casos a informação desejada não era diretamente obtida em nenhuma das duas pesquisas, por exemplo, o curso mais elevado concluído pelo entrevistado. Em ambos os casos, essa informação foi obtida através de filtros e análises de algumas perguntas do questionário.

O segundo passo foi realizar uma análise estatística para comparar as variáveis. Apesar da comparação temática realizada, as variáveis foram observadas através de questionários diferentes e sequências de perguntas diferentes. Por isso, foi comparada a distribuição de cada variável da PNAD com a da sua correspondente no Censo e foi feito um teste de

aderência. Como é esperada uma alteração nas variáveis entre os anos foi utilizado um valor-p de corte de 10%.

Para a PNAD 2001 as variáveis escolhidas foram: tipo do domicílio[1], número de cômodos, número de dormitórios, condição do domicílio[2], método de abastecimento de água, método de despejo de esgoto, método de descarte do lixo, existência de máquina de lavar, existência de televisão e total de moradores; com relação ao chefe da família, sexo, raça, grau de instrução, tipo de trabalho[3] e horas trabalhadas por semana.

5.2 Estimação dos parâmetros do modelo

A análise e a escolha adequada do modelo é parte fundamental do estudo. A partir dos resultados apresentados na seção anterior tem-se uma lista de variáveis comparáveis, candidatas a fazerem parte do modelo hierárquico para explicar o logaritmo da renda domiciliar *per capita*.

Como o estudo pretende estimar os índices de pobreza por município, é natural que este seja o nível escolhido para o modelo hierárquico. Ainda assim foram testados modelos usando distrito ou setor censitário no lugar de município. O ajuste do modelo com município teve o menor valor do critério AIC, confirmando a sua escolha como variável que define os grupos de domicílios no modelo hierárquico.

Cada variável explicativa foi testada através de um método *stepwise* partindo de um modelo completo e comparando o critério AIC. No caso da PNAD 2001 o método de abastecimento de água foi a única variável excluída do modelo neste passo.

Após verificar a significância das variáveis, foram testados os vários níveis das variáveis categóricas e sua significância para o modelo. A junção das categorias não significativas teve um viés temático sempre quando possível. Por exemplo, no caso do destino do lixo, apesar da categoria "jogado em rio, lago ou mar" não ser estatisticamente significante foi dada preferência a incorporá-la a "jogado em terreno baldio ou logradouro" ao invés de uni-lo a categoria base que é "coletado por serviço de limpeza".

Decidido qual seria o modelo final, os dados do censo foram adaptados, separando as variáveis pertinentes e realizando as mesmas agregações de categorias. Por fim, foram

[1]Casa, cômodo, apartamento

[2]Próprio, alugado, cedido, etc.

[3]Empregado, empregador, conta-própria, etc.

estimados os parâmetros do modelo pela maneira clássica, utilizando o pacote *lme4* e pela maneira Bayesiana utilizando o *OpenBUGS*.

Para o ajuste do modelo Bayesiano foram utilizados como valores iniciais dos coeficientes de regressão o valor 0 e para os parâmetros de precisão o valor 1. A convergência da cadeia pode ser verificada visualmente e confirmada pela linha horizontal que indica o valor estimado para o mesmo parâmetro pelo método clássico. Em todos os casos, a convergência aconteceu antes da milésima iteração. Com relação à independência das amostras observou-se que a autocorrelação de todos os parâmetros é insignificante a partir do *lag* 400.

5.3 Estimação dos índices de pobreza

Partindo para a estimação dos índices de pobreza no método EB foram utilizadas $L = 500$ populações *bootstrap* e para o ELL e o HB foram utilizadas $L = 1000$ populações *bootstrap*. Optou-se por utilizar um número menor para o EB uma vez que com esta simulação só é estimado o índice pontual, e a estimação do erro dependerá de outra simulação *bootstrap*.

Percebe-se nas figuras 5.1 e 5.2 que as estimativas dos índices de pobreza seguem a tendência daquelas obtidas diretamente da amostra do Censo. No entanto o efeito de *shrinkage* é evidente em todos os casos. A única exceção é no caso das estimativas obtidas pelo método HB para os municípios com amostra que acompanha muito bem os extremos do índice.

O questionário sobre rendimentos da PNAD é extenso, perguntando sobre vários tipos de ocupação e trabalhos, sendo muito mais efetivo em capturar informação acerca de trabalhos temporários ou trabalhos informais realizados no próprio domicílio que muitas vezes são esquecidos ou considerados de pouca importância pelo entrevistado.

É importante lembrar que, no caso do índice FGT, um pequeno acréscimo à renda em domicílios que estão acima da linha da pobreza não altera em nada o valor do índice, enquanto acréscimos em domicílios que estão abaixo da linha podem tirar esse domicílio da condição de pobreza ou alterar consideravelmente o valor monetário necessário para tirá-lo da pobreza.

O que é importante perceber nas figuras é que os vários métodos de estimação são

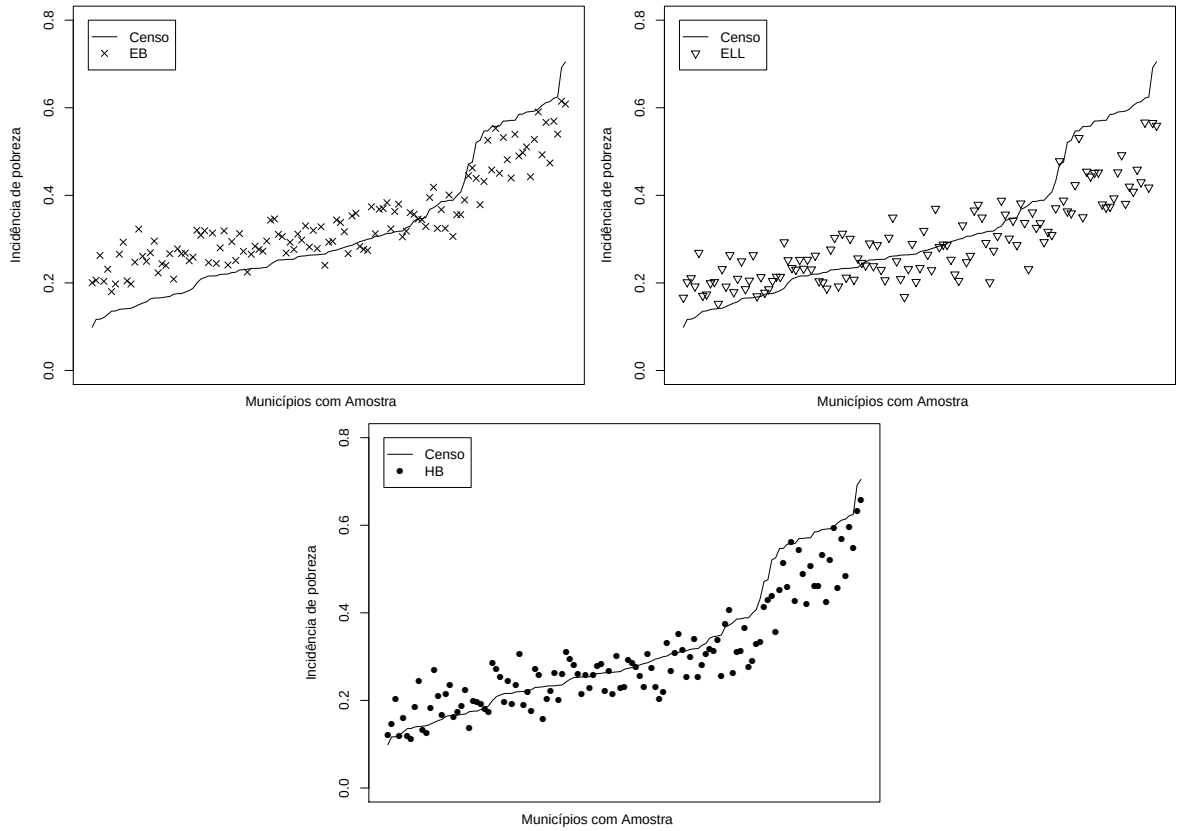


Figura 5.1: Incidência de pobreza para os municípios com amostra

próximos uns dos outros e seguem uma tendência aproximada à do Censo.

O mesmo comportamento pode ser observado nas figuras 5.3 e 5.4, lembrando sempre das diferenças na captura da renda pela PNAD.

5.4 Estimação dos erros de estimação

Partindo para a comparação dos erros de estimação, primeiro a raiz quadrada do erro quadrático médio estimado por cada método foi comparada.

Na figura 5.5 observa-se a raiz do EQM da Incidência de Pobreza para os municípios que tinham amostra e os que não tinham, respectivamente. No caso dos municípios com amostra, o método HB apresenta o menor erro estimado, e partindo das simulações realizadas anteriormente, há razões para crer que estas estimativas são boas.

Note, em particular, o único *outlier* no diagrama em caixa do HB para os municípios com amostra. Este ponto representa o município de Belo Horizonte, com uma amostra de 1665 domicílios na PNAD. Mesmo um estimador direto da amostra já teria um erro pequeno e não é esperado que a utilização de modelos prejudique o estimador. Ou seja,

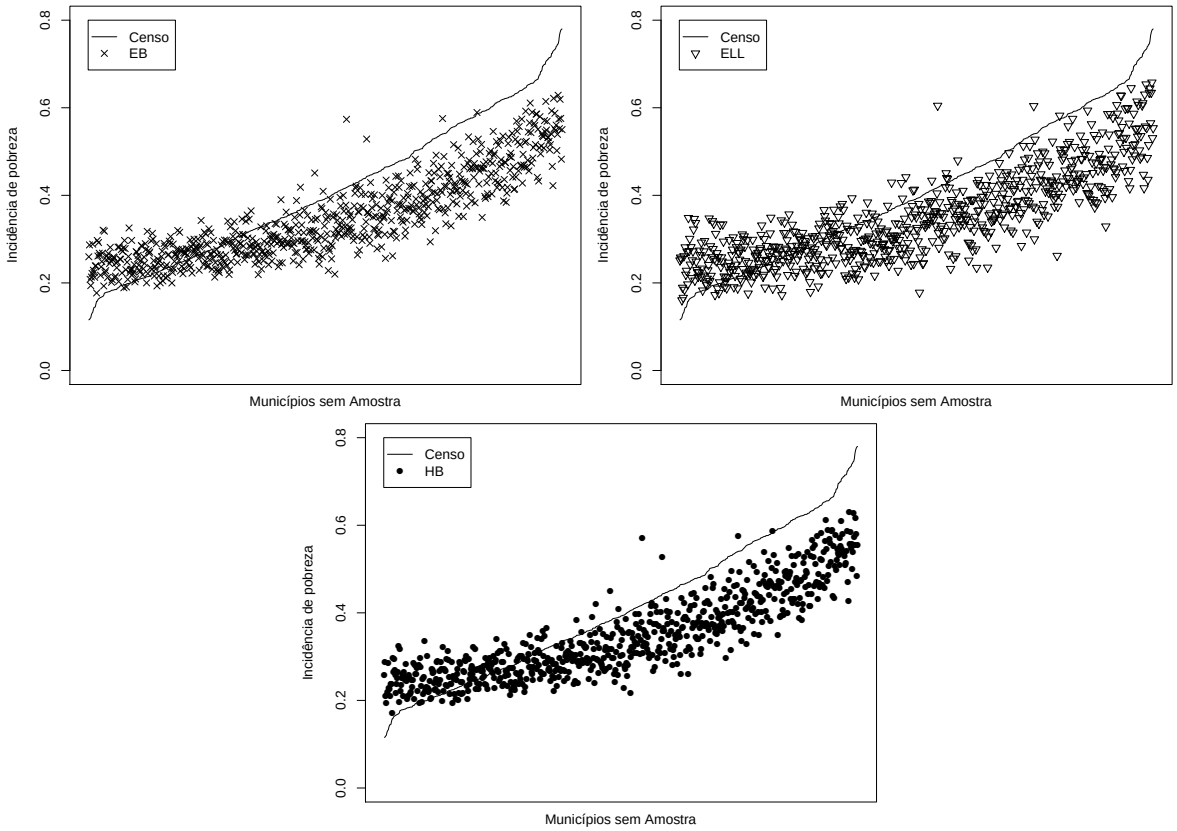


Figura 5.2: Incidência de pobreza para os municípios sem amostra

este ponto, com um EQM relativo de 3,2%, era esperado e o que parece estranho é porque um comportamento semelhante não é observado nos outros métodos.

Na verdade, como o ELL não leva em consideração parte da informação vinda da amostra e trata municípios que têm amostra como aqueles que não têm, este comportamento não era realmente inesperado. Verifica-se na figura 5.5 que a raiz do EQM pelo método ELL é muito parecida entre os municípios com ou sem amostra.

No caso do EB, novamente se observa um comportamento do EQM estranho, com os erros para os municípios com amostra maiores que os erros para os municípios sem amostra e ambos fora do padrão dos outros métodos.

A figura 5.6 apresenta a mesma informação da figura 5.5 mas para o Hiato de Pobreza. A análise aqui é a mesma e os mesmos padrões de comportamento dos erros são observados.

Para dar uma ideia da ordem de grandeza dos erros com relação às estimativas de pobreza, estes índices são apresentados em conjunto nas figuras 5.7, 5.8, 5.9, 5.10, 5.11 e 5.12.

O mais importante para se notar nestes gráficos é o intervalo de confiança apresentado.

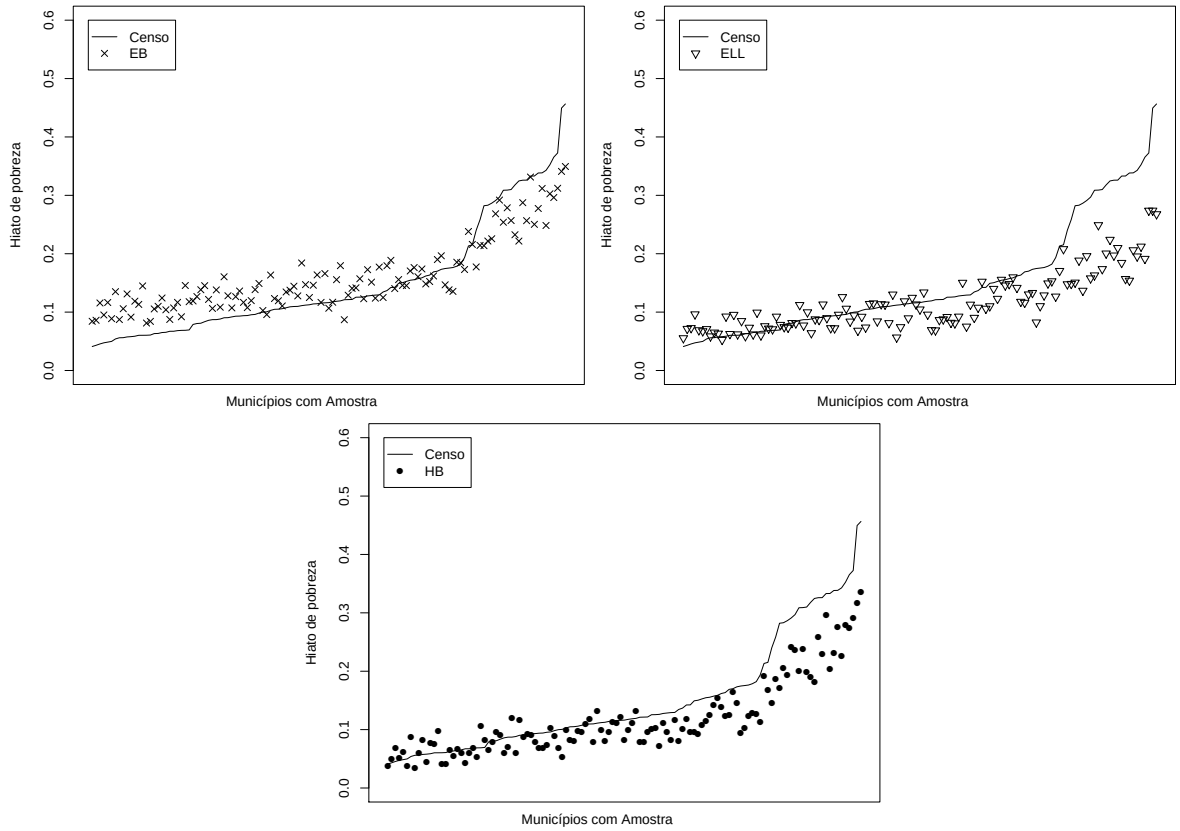


Figura 5.3: Hiato de pobreza para os municípios com amostra

Ainda que para o HB existam alguns pontos onde este intervalo é bem pequeno, ele é em geral grande o suficiente para que uma comparação entre os vários municípios não seja definitiva.

É possível perceber novamente o comportamento errático das estimativas do erro pelo método EB e como em alguns casos o intervalo de confiança fica fora dos valores possíveis do parâmetro. Obviamente o mais correto seria limitar o intervalo inferiormente com o valor zero, mas optou-se por deixar o valor assim para ressaltar que este é um problema possível na aproximação normal mas não no caso do HB, onde são utilizados quantis estimados.

Outro comportamento que também é possível perceber nestas figuras é a falta de variabilidade no EQM estimado pelo método ELL. Todos os municípios tem um intervalo de confiança muito próximo, independente de ter ou não amostra da PNAD ou do número de observações provenientes do Censo. Esse comportamento é ainda mais ressaltado se for comparado com o do método HB, onde é claro que os municípios têm intervalos de confiança muito diferentes, influenciados pela informação disponível para cada um deles.

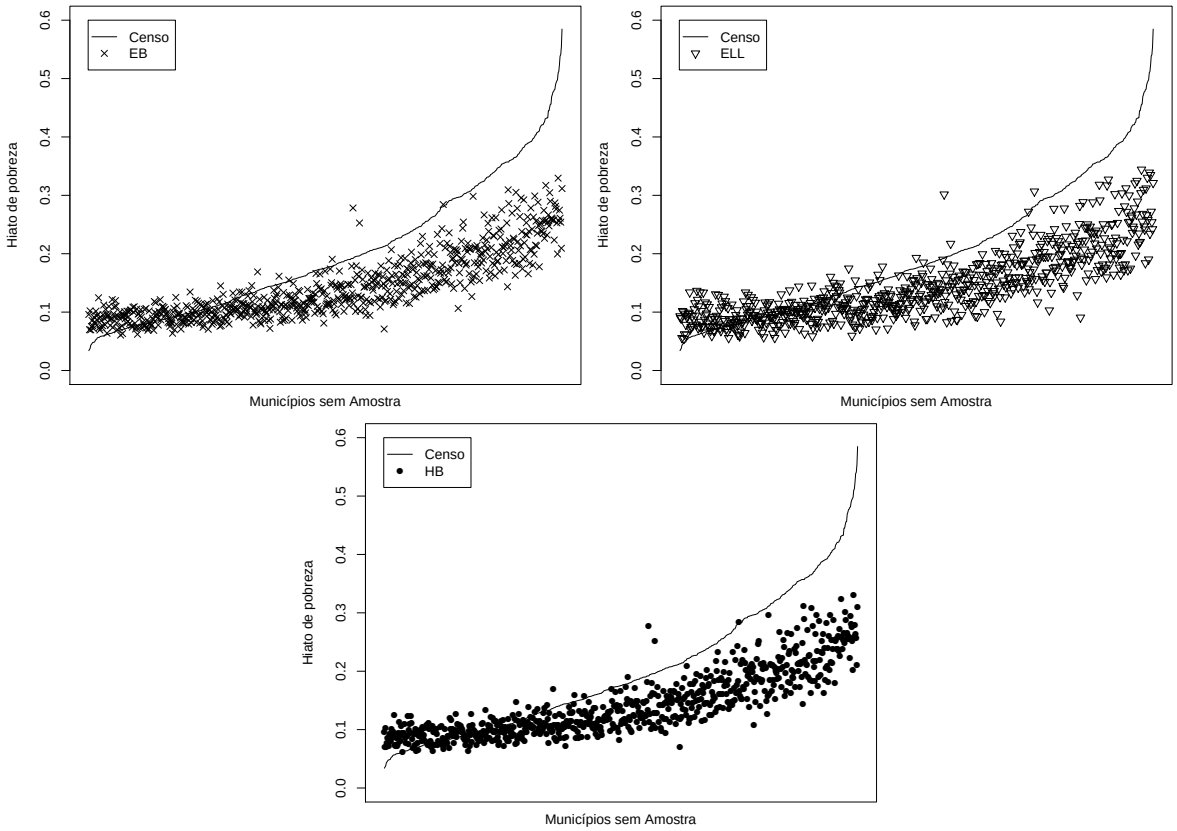


Figura 5.4: Hiato de pobreza para os municípios sem amostra

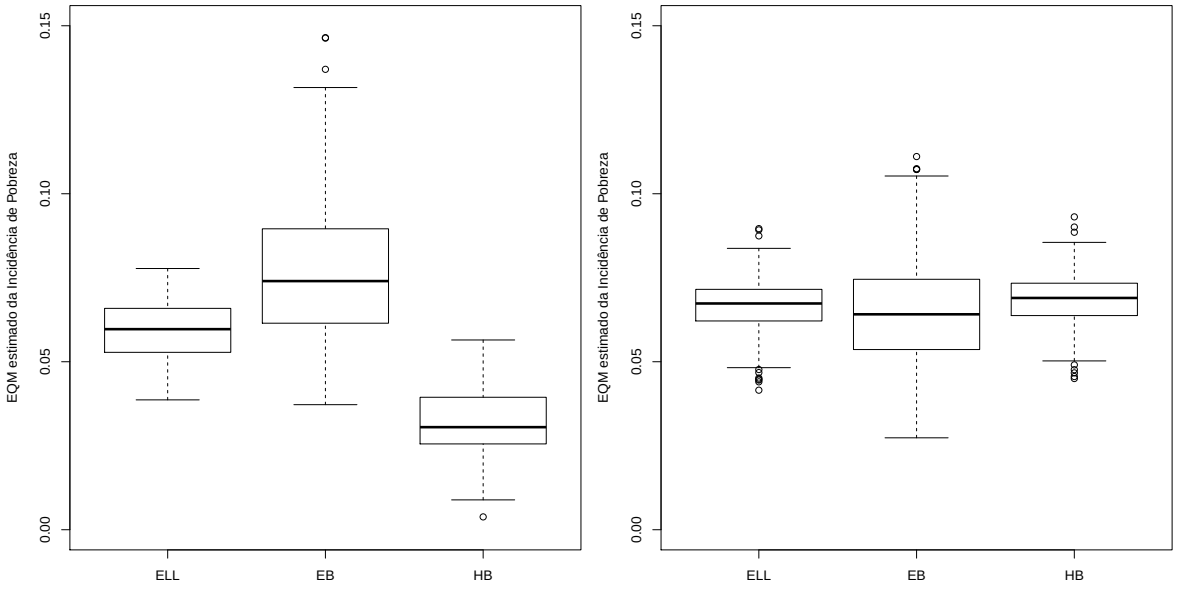


Figura 5.5: Diagrama em caixa da raiz do EQM da Incidência de Pobreza para municípios na amostra e fora da amostra da PNAD

Novamente o método HB se mostrou o mais eficiente. Apesar de não haver ganho para os municípios sem amostra, é evidente que o intervalo de confiança para os municípios

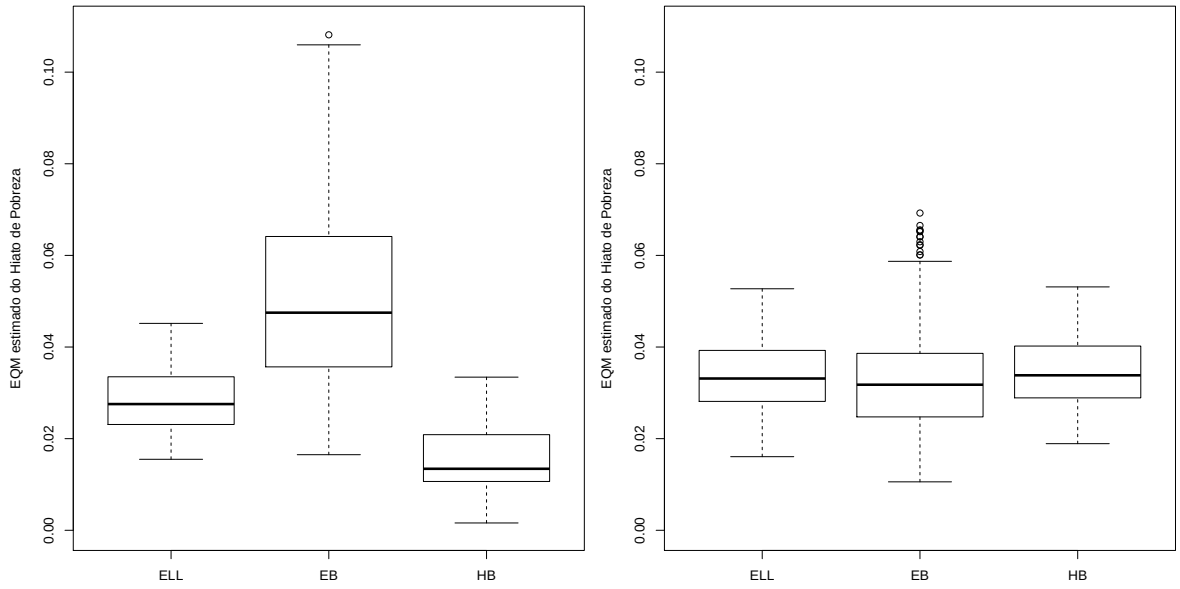


Figura 5.6: Diagrama em caixa da raiz do EQM do Hiato de Pobreza na amostra e fora da amostra

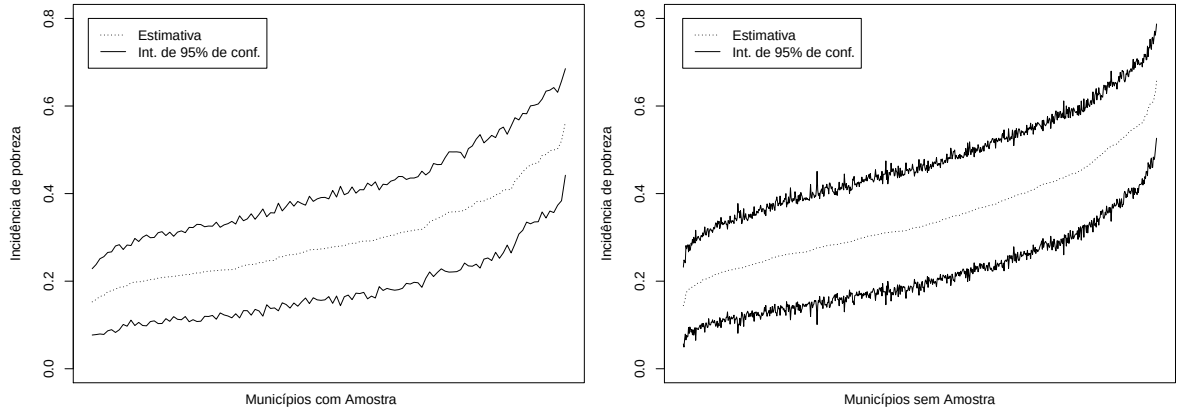


Figura 5.7: Estimativa e intervalo de confiança de 95% da Incidência de pobreza pelo método ELL

com amostra é melhor que dos outros dois métodos.

5.5 Conclusão

Todos os resultados apresentados neste capítulo tiveram como base a PNAD 2001, apesar de os mesmos resultados estarem disponíveis para a PNAD 2002 e 2003. Optou-se por não detalhar os passos ou mostrar os gráficos para essas duas bases de dados devido à similaridade dos resultados. Obviamente, existe uma diferença pequena nas estimativas dos índices de pobreza, mas é esperado que a população sofra alterações com o decorrer

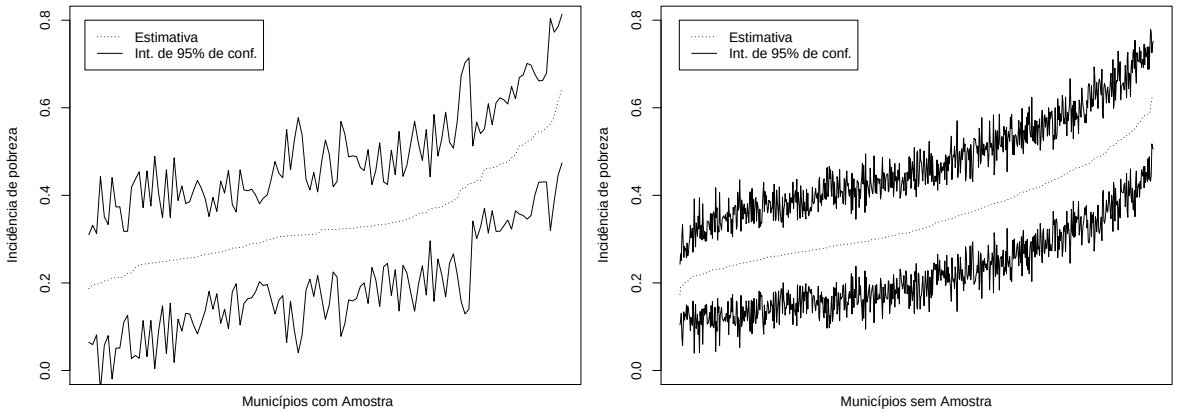


Figura 5.8: Estimativa e intervalo de confiança de 95% da Incidência de Pobreza pelo método EB

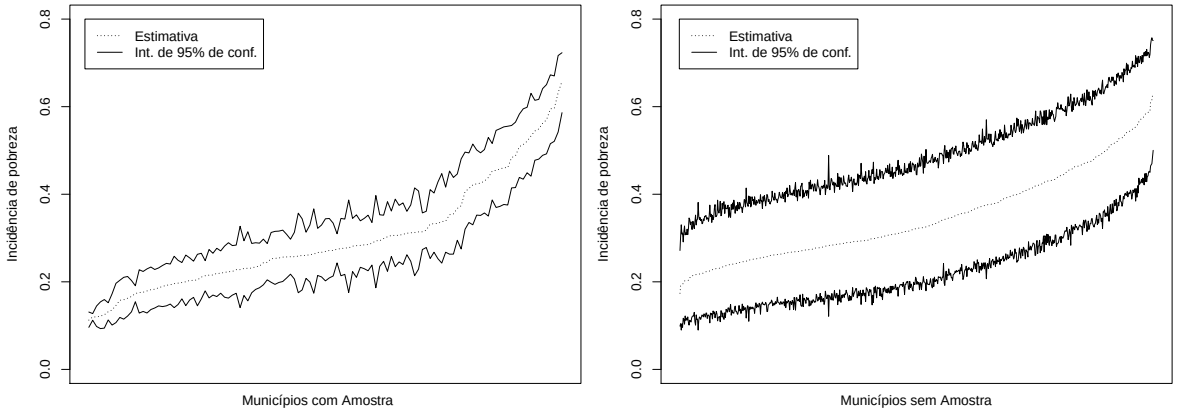


Figura 5.9: Estimativa e intervalo de confiança de 95% da Incidência de Pobreza pelo método HB

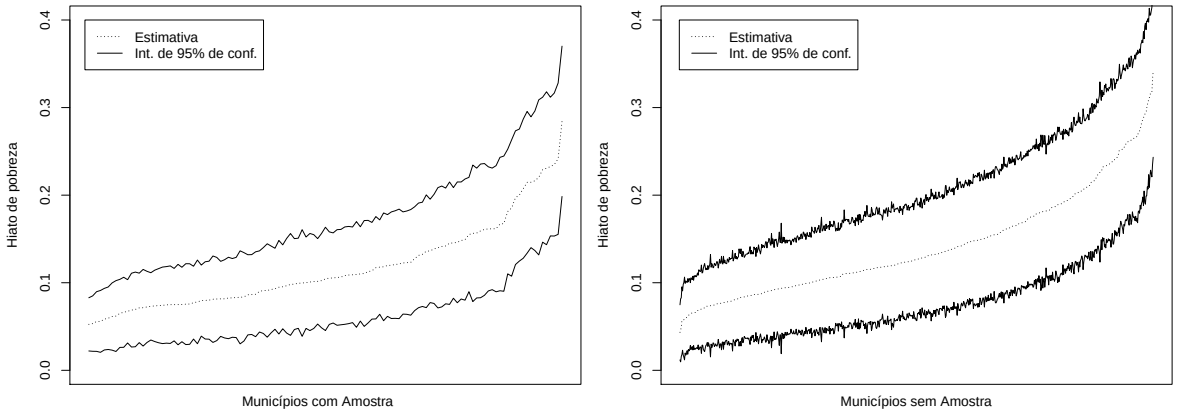


Figura 5.10: Estimativa e intervalo de confiança de 95% da Hiato de pobreza pelo método ELL

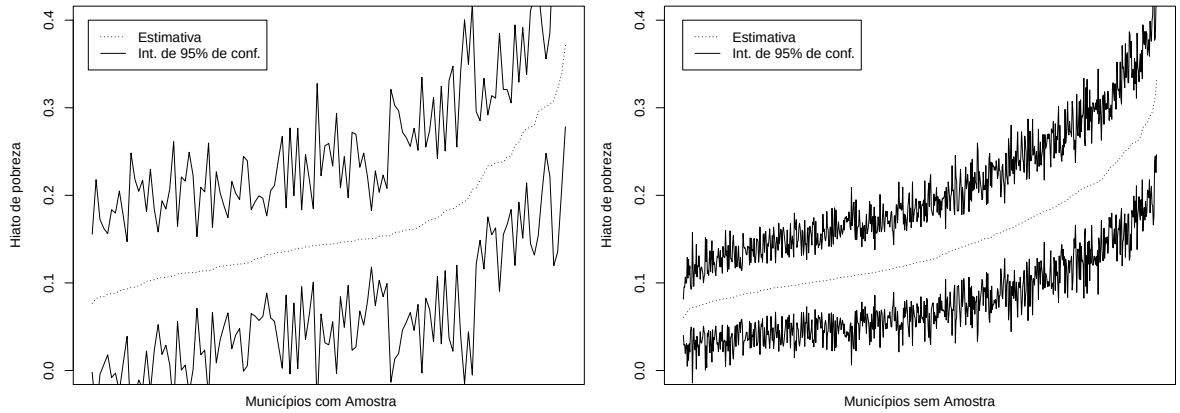


Figura 5.11: Estimativa e intervalo de confiança de 95% da Hiato de Pobreza pelo método EB

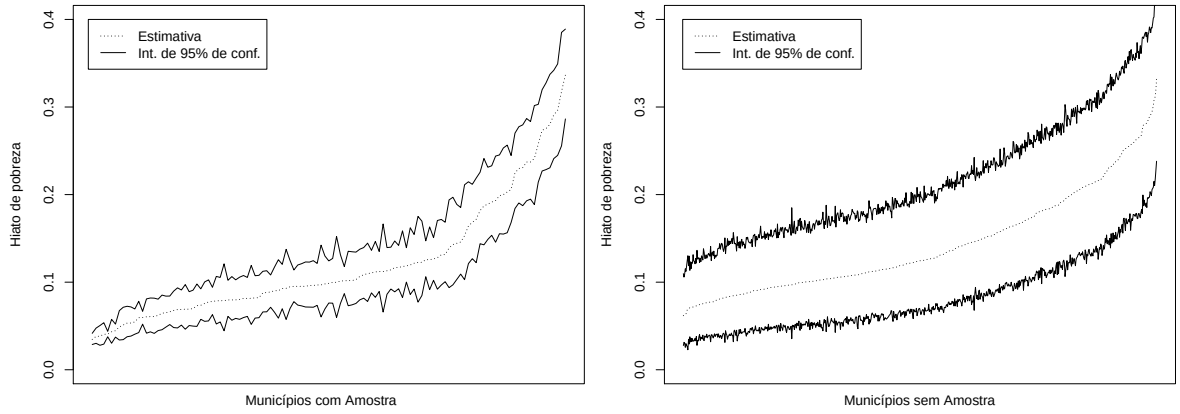


Figura 5.12: Estimativa e intervalo de confiança de 95% da Hiato de Pobreza pelo método HB

dos anos. Já as estimativas dos EQMs são muito semelhantes, não levando a conclusões diferentes daquelas obtidas com base na PNAD 2001, apresentadas a seguir.

Com relação aos estimadores pontuais não se observou grandes diferenças entre os 3 métodos analisados. Certamente os métodos não estimaram os mesmos valores para os índices de cada município, mas ainda assim forneceram valores muito próximos. Infelizmente a precisão destas estimativas não pode ser confirmada por valores externos ou pelos dados do Censo devido à diferença dos métodos de coleta dos dados de renda.

Com relação ao erro de estimação, aqueles encontrados pelo método EB utilizando o *bootstrap* foram piores do que o esperado, ainda mais considerando que as estimativas dos parâmetros do modelo foram bem estimadas. Para o método *bootstrap* para a estimação do erro foram utilizadas 1000 replicações. É possível que este não tenha sido um valor adequado e sejam necessárias mais réplicas para obter uma estimativa correta.

No entanto, as estimativas do erro para os municípios com amostra não parecem estar sequer próximas do que seria esperado, o que indica que se o método for eficiente ainda seriam necessárias muito mais réplicas, e a execução destas 1000 já levaram 4 dias sendo executadas em 10 núcleos. Comparando este tempo de execução com o do método HB que foi de apenas 1 dia em 1 núcleo ou do ELL que demorou somente 3 horas, também em 1 núcleo, é importante levantar a questão se é mesmo razoável realizar esta estimação por *bootstrap*.

Independentemente disto, entre os 3 métodos o HB apresentou os melhores resultados e os testes realizados nas seções anteriores e a análise feita dos erros leva a crer que as estimativas apresentadas são confiáveis.

6 Conclusão

Quanto à comparação dos métodos, é inegável a vantagem que o método Bayesiano Hierárquico (HB) apresentou em todos os testes realizados. Desde a simulação mais simples, onde ele foi tão bom quanto os outros dois métodos, quanto na simulação complexa, onde ele obteve o menor erro e a estimação mais precisa deste erro, até a aplicação em dados reais onde os índices e os erros estimados parecem os mais confiáveis, por conta do erro estimado para Belo Horizonte e do comportamente diferenciado para municípios com e sem amostra.

Com relação ao método Bayesiano Empírico (EB), apesar de este ter apresentado resultados próximos àqueles encontrados no HB com relação aos estimadores pontuais, a ineficiência do método de estimação do erro foi realmente uma falha inesperada. Infelizmente isso traz enormes prejuízos ao método, uma vez que o conhecimento do erro das estimativas é parte essencial do estudo.

No caso do método ELL, apesar de ser aquele computacionalmente mais eficiente, o ganho apresentado pelo HB compensa a demora na estimação. Além disso, é importante lembrar que o método ELL foi aplicado com alguns aprimoramentos, o mais significativo a estimação dos parâmetros do modelo hierárquico que foi feita utilizando máxima verossimilhança ao invés do método de estimação sugerido pelos autores.

Infelizmente, independentemente do método, há um problema nos resultados apresentados que não pode ser resolvido por nenhum método estatístico, ou seja os erros das estimativas. Note que mesmo para o método HB, onde foram obtidos os menores erros, não é possível fazer uma separação clara entre os índices de pobreza dos municípios. Considerando que o objetivo final destas estimativas é indicar quais os municípios mais pobres para que estes recebam os recursos necessários ao combate da pobreza, nenhum dos método propostos cumpre o objetivo. O fator mais importante para esta limitação seria o pouco espalhamento da amostra, não sendo possível obter informações diretas de renda para aproximadamente 80% dos municípios.

Existe uma concepção errada do poder da estimação em pequenas áreas e dos limites que ela possui. Apesar de ser uma ferramenta poderosa, sempre há um limite inerente aos dados e as informações disponíveis e por vezes algumas concessões têm que ser feitas. Neste caso, é inevitável a necessidade de agrupar alguns municípios para fazer a comparação

desejada.

Obviamente o método HB permite que este agrupamento seja menor do que seria necessário para outros métodos e certamente muito menor do que o necessário para utilizar estimadores diretos. Mas ainda assim, para uma publicação oficial, se faz necessário melhorar os erros das estimativas obtidas ou rever as áreas para as quais se pretende fornecer estimativas.

7 Trabalhos Futuros

Aproveitando que o método HB se apresentou como o mais eficiente e que a PNAD é uma pesquisa que se repete anualmente, é possível adicionar uma componente temporal aos parâmetros do modelo e utilizar modelos dinâmicos para melhorar a precisão das estimativas.

Apesar de não ser possível adicionar componentes de correlação espacial com base nas amostras apresentadas aqui a amostra da PNAD foi modificada recentemente e está muito mais espalhada, com mais municípios com amostra. Seria possível, em um estudo posterior com essas novas bases de dados, testar a inclusão de componentes espaciais.

Outra modificação possível seria realizar este estudo sem supor normalidade dos erros, usando maneiras alternativas de parametrização do modelo ou tentando modelar diretamente os índices de pobreza através de uma função logit.

Referências

Atkinson, A. B. (1970) On the measurement of inequality. *Journal of Economic Theory*, **2**, 244–263.

Baillargeon, S. e Rivest, L.-P. (2010) *stratification: Univariate Stratification of Survey Populations*. URL: `http://CRAN.R-project.org/package=stratification`. R package version 2.0-2.

Bates, D. e Maechler, M. (2010) *lme4: Linear mixed-effects models using S4 classes*. URL: `http://CRAN.R-project.org/package=lme4`. R package version 0.999375-34.

Cochran, W. G. (1977) *Sampling Techniques*. Probability and Mathematical Statistics-Applied. John Wiley & Sons, Inc., 3ª edn.

Corrêa, S. T. (2008) *Bias corrections in multilevel modelling of survey data with applications to small area estimation*. Tese de Doutorado, University of Southampton.

Draper, D. e Browne, W. J. (2000) Implementation and performance issues in the bayesian and likelihood fitting of multilevel models. *Computational Statistics*, **15**, 391–420.

Elbers, C., Lanjouw, J. O. e Lanjouw, P. (2002) Micro-level estimation of welfare. *Policy Research Working Paper 1*, The World Bank.

— (2003) Micro-level estimation of poverty and inequality. *Econometrica*, **71**, 355–364.

Elbers, C., Leite, P. G. e Lanjouw, P. (2008) Brazil within brazil: Testing the poverty map methodology in minas gerais. *Policy Research Working Paper 4513*, The World Bank.

Foster, J., Greer, J. e Thorbecke, E. (1984) A class of decomposable poverty measures. *Econometrica*, **52**, 761–766.

Gamerman, D. e Lopes, H. F. (2006) *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Texts in Statistical Science. Champan and Hall, 2ª edn.

Gini, C. (1955) *Variabilità e Mutabilità*. Rome: Libreria Eredi Virgilio Veschi. Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T).

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D. e Santamaria, L. (2008) Bootstrap mean squared error of a small-area eblup. *Journal of Statistical Computation and Simulation*, **78**, 443–462.

Haslett, S., Isidro, M. e Jones, G. (2010) Comparison of survey regression techniques in the context of small area estimation of poverty. *Survey Methodology*, **36**, 157–170.

Hoover, jr., E. M. (1936) The measurement of industrial localization. *The Review of Economics and Statistics*, **18**, 162–171.

IBGE (2001) *Pesquisa Nacional por Amostra de Domicílios 2001 – Síntese de Indicadores*. Instituto Brasileiro de Geografia e Estatística. URL: `http://www.ibge.gov.br/home/estatistica/populacao/trabalhoerendimento/pnad2001/default.shtm`.

— (2002) *Pesquisa Nacional por Amostra de Domicílios 2002 – Síntese de Indicadores*. Instituto Brasileiro de Geografia e Estatística. URL: `http://www.ibge.gov.br/home/estatistica/populacao/trabalhoerendimento/pnad2002/default.shtm`.

— (2003a) *Metodologia do Censo Demográfico 2000*, vol. 25 de *Série Relatórios Metodológicos*. Instituto Brasileiro de Geografia e Estatística.

— (2003b) *Pesquisa Nacional por Amostra de Domicílios 2003 – Síntese de Indicadores*. Instituto Brasileiro de Geografia e Estatística. URL: `http://www.ibge.gov.br/home/estatistica/populacao/trabalhoerendimento/pnad2003/default.shtm`.

— (2008) Mapa de pobreza e desigualdade: municípios brasileiros. *Relatório técnico*, Instituto Brasileiro de Geografia e Estatística.

Lavallée, P. e Hidiroglou, M. A. (1988) On the stratification of skewed populations. *Survey Methodology*, **14**, 33–43.

Lunn, D., Spiegelhalter, D. e Thomas, A. (2009) The bugs project: Evolution, critique, and future directions. *Statistics in Medicine*, **28**, 3049–3067.

Osier, G. (2009) Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Survey Research Methods*, **3**, 167–195.

Pessoa, D. e Silva, P. (1998) *Análise de Dados Amostrais Complexos*. Caxambu: ABE - Associação Brasileira de Estatística. 13o SINAPE.

Pfeffermann, D., Silva, P. L. d. N. e Moura, F. d. S. (2006) Multi-level modelling under informative sampling. *Biometrika*, **93**, 943–959.

Pffefermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H. e Rasbash, J. (1998) Weighting for unequal selection probabilities in multilevel model. *Journal of the Royal Statistical Society series B*, **60**, 23–40.

R Development Core Team (2010) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: `http://www.R-project.org`. ISBN 3-900051-07-0.

Rao, J. N. K. (2003) *Small Area Estimation*. Wiley Series in Survey Methodology. John Wiley & Sons, Inc.

Rao, J. N. K. e Molina, I. (2010) Small area estimation of poverty indicators. *The Canadian Journal of Statistics*, **38**, 369–385.

Tarozzi, A. e Deaton, A. (2009) Using census and survey data to estimate poverty and inequality for small areas. *The Review of Economics and Statistics*, **91**, 773–792.

Theil, H. (1967) *Economics and information theory*. North-Holland Pub. Co.

West, M. e Harrison, J. (1997) *Bayesian Forecasting and Dynamic Models*. Springer Series in Statistics. Springer, 2ª edn.

ANEXOS

A Detalhes das provas

A.1 Vício do estimador $\hat{v}^B$

Tem-se que

$$\hat{v}^B = E_{\boldsymbol{Y}_r}(v|\boldsymbol{Y}_s)$$

$$v = h(\boldsymbol{Y})$$

$$\boldsymbol{Y}' = (\boldsymbol{Y}_s', \boldsymbol{Y}_r')$$

Logo,

$$\begin{aligned}
E_{\boldsymbol{Y}_s}(\hat{v}^B) &= E_{\boldsymbol{Y}_s}\left[E_{\boldsymbol{Y}_r}(v|\boldsymbol{Y}_s)\right] \\
&= E_{\boldsymbol{Y}_s}\left[E_{\boldsymbol{Y}_r}(h(\boldsymbol{Y})|\boldsymbol{Y}_s)\right] \\
&= \int \left\{ \int h(\boldsymbol{Y}) f(\boldsymbol{Y}_r|\boldsymbol{Y}_s) d\boldsymbol{Y}_r \right\} f(\boldsymbol{Y}_s) d\boldsymbol{Y}_s \\
&= \int \int h(\boldsymbol{Y}) f(\boldsymbol{Y}_s, \boldsymbol{Y}_r) d\boldsymbol{Y}_r d\boldsymbol{Y}_s \\
&= \int h(\boldsymbol{Y}) f(\boldsymbol{Y}) d\boldsymbol{Y} \\
&= E_{\boldsymbol{Y}}[h(\boldsymbol{Y})] \\
&= E_{\boldsymbol{Y}}(v)
\end{aligned}$$

B Tabelas e figuras adicionais

B.1 Tabelas de comparação entre o ML e o PWIGLS para vários tamanhos de populações e amostras

Tabela B.1: Média das estimativas de σ_u para as amostras das 1000 populações simuladas

M	m	ML	PWIGLS	M	m	ML	PWIGLS
160	40	-0,1629	-0,2409	640	160	-0,0102	-0,0511
160	80	-0,0674	-0,1569	640	200	-0,0164	-0,045
320	40	-0,1376	-0,2034	640	240	-0,0105	-0,0355
320	80	-0,056	-0,1142	640	280	-0,0121	-0,0401
320	120	-0,0322	-0,086	640	320	-0,0141	-0,0382
320	160	-0,0252	-0,0723	800	40	-0,1802	-0,2299
480	40	-0,1755	-0,2369	800	80	-0,058	-0,1212
480	80	-0,0299	-0,0972	800	120	-0,0328	-0,0776
480	120	-0,0357	-0,0943	800	160	-0,0119	-0,0469
480	160	-0,0173	-0,06	800	200	-0,0142	-0,0406
480	200	-0,0134	-0,0388	800	240	-0,015	-0,0343
480	240	-0,0137	-0,0405	800	280	-0,0059	-0,026
640	40	-0,1565	-0,2067	800	320	-0,0118	-0,0278
640	80	-0,0641	-0,1185	800	360	-0,0066	-0,0256
640	120	-0,0168	-0,0659	800	400	-0,0086	-0,024

Tabela B.2: Erro quadrático médio relativo de σ_u para as amostas de 1000 populações simuladas (x100)

M	m	ML	PWIGLS	M	m	ML	PWIGLS
160	40	0,516	0,4729	640	160	0,1626	0,1988
160	80	0,3096	0,358	640	200	0,1469	0,1825
320	40	0,5187	0,4587	640	240	0,1265	0,1569
320	80	0,3201	0,32	640	280	0,1157	0,1563
320	120	0,2053	0,2425	640	320	0,1111	0,1503
320	160	0,1727	0,2222	800	40	0,5326	0,4623
480	40	0,5326	0,4672	800	80	0,3037	0,3177
480	80	0,2868	0,3011	800	120	0,2267	0,242
480	120	0,2238	0,2564	800	160	0,1702	0,1924
480	160	0,1802	0,2088	800	200	0,1442	0,1761
480	200	0,1438	0,1768	800	240	0,1288	0,1531
480	240	0,1268	0,162	800	280	0,1121	0,143
640	40	0,5219	0,4548	800	320	0,1069	0,1405
640	80	0,3402	0,3252	800	360	0,105	0,1351
640	120	0,2083	0,2321	800	400	0,0964	0,1255