UFRJ

Inferência em Modelos Hierárquicos Generalizados sob Planos Amostrais Informativos

Romy Elena Rodríguez Ravines

Dissertação de Mestrado submetida ao Programa de Pós-graduação em Estatística do Instituto de Matemática da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários para obtenção do grau de Mestre em Ciências Estatísticas.

Orientador: Prof. Fernando Moura

Rio de Janeiro

Março de 2003

Inferência em Modelos Hierárquicos Generalizados sob Planos Amostrais Informativos

Romy Elena Rodríguez Ravines

Orientador: Prof. Fernando Moura

Dissertação de Mestrado submetida ao Programa de Pós-graduação em Estatística do Instituto de Matemática da Universidade Federal do Rio de Janeiro - UFRJ, como parte dos requisitos necessários para obtenção do grau de Mestre em Ciências Estatísticas.

Aprovada por :

_____

Presidente, Prof. Fernando Moura

_____

Prof. Dani Gamerman

_____

Prof. Heleno Bolfarine

Rio de Janeiro

Março de 2003

Ravines, Romy Elena Rodriguez

Inferência em Modelos Hierárquicos Generalizados sob Planos Amostrais Informativos/ Romy Elena Rodriguez Ravines.- Rio de Janeiro: UFRJ/IM, 2003.

xiii, 116f.: il.; 31cm.

Orientador: Fernando Moura

Dissertação (mestrado) - UFRJ/IM/ Programa de Pós-graduação em Estatística, 2003.

Referências Bibliográficas: f.95-99.

1. Amostragem Informativa. 2. Modelos Hierárquicos. 3. Inferência Analítica. I. Moura, Fernando. II. Universidade Federal do Rio de Janeiro, Instituto de Matemática. III. Título.

Agradecimentos

Esta dissertação só foi possível graças ao apoio financeiro do CNPq, a valiosa colaboração e atenção dedicada a este trabalho do meu Orientador Fernando Moura, o incentivo de todos os Professores do DME, em particular, Dani Gamerman e Hélio Migon, e o apoio incondicional de minha família e de meus caros amigos do RJ.

A todos, muito obrigada.

RESUMO

Inferência em Modelos Hierárquicos Generalizados sob Planos Amostrais Informativos

Romy Elena Rodríguez Ravines

Orientador: Prof. Fernando Moura

Resumo da Dissertação de Mestrado submetida ao Programa de Pós-graduação em Estatística do Instituto de Matemática da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários para obtenção do grau de Mestre em Ciências Estatísticas.

Os desenhos amostrais complexos são muito utilizados nas pesquisas sociais, nas quais os dados têm uma estrutura hierárquica intrínseca. Os modelos multi-níveis ou hierárquicos são os mais apropriados para descreverem essas estruturas, porém eles geralmente são ajustados independentemente do mecanismo utilizado para a obtenção das amostras.

Uma importante distinção relacionada com o efeito dos desenhos amostrais complexos na inferência é entre desenhos informativos e não informativos. Realizar inferência analítica ignorando o desenho amostral quando de fato ele é informativo tem conseqüências importantes. Neste trabalho estende-se a proposta de Pfeffermann, D., Moura, F.A.S. e Silva, P.L.N. [*Multilevel Modelling Newsletter*, v.14, n.1 (2002) : 8-17], sobre o uso das distribuições amostrais em modelos hierárquicos normais na presença de desenhos amostrais informativos, para modelos hierárquicos generalizados. Os resultados de um estudo de simulação em 500 populações e 2000 amostras e de uma aplicação a dados reais também são apresentados.

Palavras-chave: Modelo de superpopulação, Inferência analítica, Desenho amostral informativo, Distribuição Amostral.

ABSTRACT

Inference in Generalized Hierarchical Models under Informative Probability Sampling

Romy Elena Rodríguez Ravines

Orientador: Prof. Fernando Moura

Abstract da Dissertação de Mestrado submetida ao Programa de Pós-graduação em Estatística do Instituto de Matemática da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários para obtenção do grau de Mestre em Ciências Estatísticas.

Complex sample designs are often used in social science researches, in which the data have an intrinsic hierarchical structure. The hierarchical models are suitable to describe these structures, however they often are fitted independently of the sample design.

An important distinction relating to the effect of complex sample designs on the inference is between informative and noninformative designs. Carry out analytic inference ignoring the sample design when in fact, is informative, has important consequences. In this research, the work of Pfeffermann, D., Moura, F.A.S. and Silva, P.L.N. [*Multilevel Modelling Newsletter*, v.14, n.1 (2002) : 8-17], about the use of sampling distributions in normal hierarchical models under informative sampling designs, is extended to generalized hierarchical models. The results of a simulation study with 500 populations and 2000 samples and an application in a real data set are also presented.

Key-words: Superpopulation model, Analytic Inference, Informative Probability Sampling, Sampling Distribution.

RESUMEN

Inferencia en Modelos Hierárquicos Generalizados bajo Diseños Muestrales Informativos

Romy Elena Rodríguez Ravines

Orientador: Prof. Fernando Moura

Resumen da Dissertação de Mestrado submetida ao Programa de Pós-graduação em Estatística do Instituto de Matemática da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários para obtenção do grau de Mestre em Ciências Estatísticas.

Los diseños muestrales complejos son usados frecuentemente en investigaciones sociales, debido a que las poblaciones en estudio tienen una estrutura hierárquica intrínseca. Los modelos hierárquicos son apropiados para describir esas estructuras, sin embargo, generalmente son ajustados independentemente del mecanismo utilizado para obtener las muestras.

Una importante distinción relacionada con el efecto de los diseños muestrales complejos en la inferencia es entre diseños informativos y no informativos. Realizar inferencia analítica ignorando el diseño muestral cuando de hecho es informativo, tiene consecuencias importantes. En este trabajo se extiende la propuesta de Pfeffermann, D., Moura, F.A.S. y Silva, P.L.N. [*Multilevel Modelling Newsletter*, v.14, n.1 (2002) : 8-17], sobre el uso de las distribuciones muestrales en modelos hierárquicos normales en la presencia de disenhos muestrales informativos, para modelos hierárquicos generalizados. También se presentan los resultados de un estudio de simulación en 500 poblaciones y 2000 muestras e de una aplicación a datos reales, .

Palabras-clave: Modelo de superpopulación, Inferencia Analítica, Disenho Muestral Informativo, Distribuciones Muestrales.

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

Capítulo 1

INTRODUÇÃO

Segundo Pfeffermann, Krieger, e Rinott (1998), os dados amostrais podem ser considerados como o resultado de dois processos aleatórios: o processo que gera a população finita ou modelo de superpopulação e o processo de seleção da amostra ou mecanismo de seleção de amostras. A maioria das pesquisas por amostragem utilizam mecanismos de seleção complexos onde as unidades da população são selecionadas em vários estágios e (ou) com probabilidades de seleção distintas em algumas ou em todas as etapas do processo de amostragem.

Freqüentemente, dados de pesquisas por amostragem são utilizados para fazer inferência sobre os parâmetros do modelo de superpopulação, entretanto esta estimação é feita ignorando-se o mecanismo através do qual os dados foram obtidos. Com isso, as unidades da amostra são analisadas como se fossem independentes e identicamente distribuídas, o que nem sempre é verdadeiro, pois o mecanismo de seleção da amostra pode ser do tipo informativo, i.e., ser um desenho onde as probabilidades de seleção dos elementos da população estão correlacionadas com as variáveis respostas.

Realizar inferência estatística sem considerar o desenho amostral quando de fato ele é informativo tem conseqüências importantes sob o ponto de vista freqüentista como também Bayesiano. Do ponto de vista freqüentista, como descutem Pfeffermann et al. (1998) e Corrêa (2001), uma analise como essa pode acarretar a produção de estimativas viciadas para os parâmetros do modelo de interesse (bem como para a precisão destas estimativas), levando a uma visão distorcida do fenômeno em estudo. Do ponto de vista Bayesiano, segundo Gelman, Carlin, Stern, e Rubin (1995), mesmo

com verossimilhanças e dados fixos, a distribuição à posteriori muda de acordo com diferentes desenhos não ignoráveis da coleta de dados.

Na literatura existem algumas metodologias propostas para o tratamento de dados amostrais dessa natureza (obtidos com desenhos informativos). Na abordagem clássica, a maioria delas limita-se a obtenção de estimativas pontuais. Segundo Duarte (1999), existe bastante literatura sobre a estimação de medidas descritivas que incorporem o desenho amostral usado na obtenção dos dados, mas, existe pouca literatura sobre modelagem de dados de pesquisa por amostragem, e ainda há pouca literatura sobre como incorporar o desenho amostral na análise de modelos lineares. Já o trabalho de Gelman et al. (1995), resume claramente como sob o paradigma Bayesiano este problema pode naturalmente ser considerado, modificando-se a verossimilhança.

Por outro lado, sabe-se que os desenhos amostrais complexos são utilizados com maior freqüência nas pesquisas sociais onde os dados têm uma estrutura hierárquica intrínseca. O exemplo mais conhecido deste tipo de dados se encontra na área de educação onde estudantes agrupam-se em turmas, turmas em escolas, escolas em distritos escolares e assim por diante. O estudo da estrutura hierárquica deste tipo de população é de grande interesse para os pesquisadores. Os modelos multi-níveis têm utilidade incontestada nas ciências sociais, (Draper, 1995), porém eles geralmente são ajustados independentemente do mecanismo utilizado para a obtenção das amostras.

Os trabalhos de Pfeffermann, Skinner, Holmes, Goldstein, e Rasbash (1998) e Pfeffermann, Moura, e Silva (2002) são alternativas para a realização de inferência sobre os parâmetros de modelos hierárquicos a partir de amostras obtidas com desenhos informativos. O primeiro propõe um procedimento de ponderação das unidades da amostra para corrigir vícios de estimação e o segundo propõe a utilização da "distribuição amostral" utilizando modelos para as probabilidades de inclusão para cada nível hierárquico. Ambos trabalhos só foram desenvolvidos para dados normais (variável resposta normal)

O objetivo principal desta dissertação é implementar e aplicar a proposta de Pfeffermann et al. (2002) sobre o uso das distribuições amostrais em modelos hierárquicos normais na presença de desenhos amostrais informativos, para modelos hierárquicos generalizados. Com o objetivo de ser avaliadas as correções propostas, realizou-se um estudo de simulação em 500 populações e 2000 amostras obtidas com 4 diferentes desenhos amostrais. Uma aplicação em dados da "Encuesta Nacional de Hogares (ENAHO) - 2000 IV", pesquisa realizada pelo "Instituto Nacional de Estadística e Informática"(INEI) do Peru entre outubro e dezembro do ano 2000, é apresentada com detalhes.

Esta dissertação está dividida em 7 Capítulos. No Capítulo 2 são apresentadas a notação e algumas definições importantes. No Capítulo 3 são considerados os procedimentos que podem ser adotados no ajuste de modelos lineares sob desenhos amostrais informativos, tanto do ponto de vista freqüentista quanto do Bayesiano. No Capítulo 4 apresentam-se métodos de estimação de modelos lineares hierárquicos para amostras informativas. O experimento de simulação é apresentado no Capítulo 5. A aplicação de alguns dos procedimentos mencionados no Capítulo 4 na modelagem de um indicador de estado de pobreza, é considerada no Capítulo 6. Finalmente, o Capítulo 7 contém as considerações finais e as sugestões de trabalhos futuros.

Capítulo 2

DESENHOS AMOSTRAIS INFORMATIVOS

Neste Capítulo são apresentadas a notação e algumas definições importantes a serem utilizadas ao longo desta dissertação

2.1 Notação

Considere uma *População Finita* (P) de tamanho N, na qual temos interesse numa característica $\mathcal{Y}$. Seja $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)'$ o vetor $N \times 1$ de *dados completos* correspondentes aos valores da característica $\mathcal{Y}$ das N unidades da população.

Define-se como *amostra* (s) de tamanho n, um subconjunto de n unidades selecionadas de P, sendo $\boldsymbol{y}_s = (y_1, y_2, \ldots, y_n)'$ o vetor $n \times 1$ que representa o conjunto de *dados observados* ou medidos da característica $\mathcal{Y}$ para as n unidades que pertencem à amostra s. Os dados não observados (voluntária ou involuntariamente) definem o conjunto de *dados faltantes*, $\boldsymbol{y}_{\bar{s}}$.

A forma (ou mecanismo) como os dados são selecionados constitui o *desenho amostral*. Seja $\boldsymbol{I}$ o vetor indicador de seleção, i.e, $\boldsymbol{I} = (I_1, \ldots, I_N)$, onde $I_i = 1$ se $i \in s$ e $I_i = 0$ se $i \notin s$. Portanto, o conjunto de dados observados pode ser representado por $s = \{i : I_i = 1\}$ e o conjunto de dados não observados, ou dados faltantes, por $\bar{s} = \{i : I_i = 0\}$. As *probabilidades de seleção*, i.e, as probabilidades dos indivíduos da população de serem incluídos na amostra s são representadas por $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$.

As variáveis utilizadas na seleção da amostra, por exemplo, variáveis indicadoras de estratos ou conglomerados que determinam o grupo a que pertence a unidade e

variáveis quantitativas como as medidas de tamanho, são denominadas *variáveis do desenho*. Estas variáveis podem fazer parte ou não do conjunto de *covariáveis* $\boldsymbol{x}$ a serem incluídas nos modelos. No caso dos modelos hierárquicos de dois níveis, $\boldsymbol{x}$ representa as covariáveis do 1° nível e $\boldsymbol{z}$ representa as covariáveis do 2° nível.

2.2 Desenho Amostral Informativo

O desenho amostral pode ser informativo ou não informativo. Após propor um modelo, deve-se analisar se as probabilidades de seleção, $\boldsymbol{\pi}$, dos elementos da população estão relacionadas com as variáveis respostas, $\boldsymbol{y}$, condicionadas às covariáveis, $\boldsymbol{x}$, do modelo. Se essa relação existe, então, o desenho amostral é *informativo*. Se o desenho amostral é informativo, a distribuição dos valores da amostra, $f_s(y_i \mid \boldsymbol{x}, \boldsymbol{\theta})$, é diferente da sua distribuição na população, $f_p(y_i \mid \boldsymbol{x}, \boldsymbol{\theta})$.

Quando $f_s(y_i \mid \boldsymbol{x}, \boldsymbol{\theta})$ e $f_p(y_i \mid \boldsymbol{x}, \boldsymbol{\theta})$ são iguais, os dados não observados, $\boldsymbol{y}_{\bar{s}}$, não fornecem informação adicional ao modelo proposto e o desenho amostral é *ignorável* ou não informativo. Segundo Binder e Roberts (2001) o que é ignorável do ponto de vista Bayesiano pode não ser ignorável do ponto de vista freqüentista. A classificação de um desenho amostral em informativo ou ignorável depende das informações disponíveis sobre o desenho, as variáveis de interesse e o modelo proposto.

Exemplos de amostragem informativa podem ser encontrados em estudos ecológicos, sociais, da saúde pública e em pesquisas onde as unidades são selecionadas com probabilidades proporcionais a seus valores, intencionalmente ou não.

Na amostragem estratificada e na amostragem por conglomerados, por exemplo, as unidades da amostra final são selecionadas com probabilidades desiguais. Se essas probabilidades estão correlacionadas com as variáveis resposta, o desenho torna-se informativo e o modelo apropriado para se ajustar aos dados amostrais é diferente do modelo para se ajustar aos dados populacionais (Pfeffermann et al., 2002).

Outro exemplo prático é a Não Resposta Não Ignorável. A não resposta é um

fenômeno comum nas pesquisas por amostragem. A Não Resposta Não Ignorável acontece quando o mecanismo de não resposta depende dos valores da variável não respondida, e que, segundo Qin, Leung, e Shao (2002), é o tipo de não resposta mais difícil de ser controlado. Neste caso, se a amostra planejada é considerada a população de interesse e o mecanismo de não resposta é considerado como o mecanismo de seleção, então o desenho amostral é informativo para os valores observados.

2.3 Superpopulação

O processo de inferência estatística a partir de uma amostra compreende um conjunto de princípios e procedimentos que podem envolver, por exemplo, o conhecimento de algum processo aleatório que possa ter gerado o verdadeiro valor desconhecido da característica de interesse $\mathcal{Y}$ para cada unidade da população. Esse processo é representado por um modelo que é utilizado como base para a realização de inferências. Esta abordagem é denominada de *modelos de superpopulação*.

O termo *Superpopulação* refere-se então, ao modelo $f_p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})$ que especifica a distribuição conjunta dos valores da variável de interesse $\boldsymbol{y}$ na população, isto é, a distribuição conjunta de $\boldsymbol{y} = (y_1, \ldots, y_N)$. A *inferência analítica* em dados amostrais refere-se à inferência sobre os parâmetros do modelo de superpopulação ($\boldsymbol{\theta}$).

O principal problema abordado nesta dissertação é a realização de inferência analítica a partir de dados obtidos por amostragem complexas, particularmente nos casos quando a amostragem é informativa.

Capítulo 3

MODELOS LINEARES SOB DESENHOS AMOSTRAIS INFORMATIVOS

O ajuste de modelos lineares, em particular o modelo de regressão, é uma atividade realizada freqüentemente e de forma quase imediata graças ao desenvolvimento de muitos pacotes computacionais. Entretanto, as hipóteses estatísticas necessárias para a aplicação dos métodos de estimação, como por exemplo, de que os dados foram obtidos através de uma amostragem aleatória simples com reposição, são freqüentemente violados. Ajustar modelos lineares ignorando o desenho amostral pode levar a interpretações distorcidas do fenômeno em estudo (Corrêa, 2001).

Neste Capítulo apresentam-se alguns dos métodos propostos na literatura para a realização de inferência sobre os parâmetros de modelos lineares sob desenhos amostrais informativos. Na Seção 3.1 mencionam-se alguns dos métodos de estimação sob o ponto de vista clássico. O procedimento Bayesiano é apresentado na Seção 3.2.

3.1 Inferência Clássica

Esta Seção contem um resumo dos principais trabalhos de estatísticos clássicos sobre métodos de estimação dos coeficientes de modelos de regressão para amostras complexas. O texto completo está baseado nos Capítulos 2 e 3 de Duarte (1999), onde encontram-se detalhes sobre cada um dos métodos abaixo mencionados.

Os primeiros autores que abordaram o problema de estimação de coeficientes de regressão em amostras complexas foram Kish e Frankel (1974). Eles usaram o método de Linearização de Taylor, o método das Replicações Repetidas Balancea-

das e o método das Replicações Repetidas Jackknife, considerados métodos indiretos, para a obtenção de estimativas de coeficientes de regressão com dados de diversos desenhos amostrais complexos. Nenhum dos métodos mostrou ser melhor ou pior que os outros, porém, nesse trabalho mostraram-se que as estimativas obtidas para os erros padrões com modelos que ignoram o desenho amostral subestimam os erros padrões verdadeiros.

Nathan e Holt (1974) consideram a incorporação das variáveis de desenho como informação auxiliar no modelo linear. Os autores estimaram os parâmetros do modelo de superpopulação e verificaram que o estimador usual de Mínimos Quadrados Ordinarios não é apropriado para inferências sobre modelos de regressão e que se deve levar em conta a informação da amostra usada. Eles propuseram dois estimadores alternativos conhecidos como estimadores de "Pearson" e de "Pearson-ajustado" respectivamente.

Fuller (1975) estudou a forma analítica de estimação de modelos de regressão com amostras obtidas de forma aleatória simples sem reposição. Pfeffermann e Nathan (1979, 1981) propuseram um método de estimação no caso em que diferentes grupos da população apresentam diferentes relações de regressão, mas apenas uma parte dos grupos pode ser incluída na amostra. Os autores trataram os coeficientes de regressão de cada grupo como variáveis aleatórias não correlacionadas e o parâmetro populacional foi definido como sendo uma média ponderada desses coeficientes de regressão separados.

Pfeffermann e Holmes (1985) complementaram o estudo de Nathan e Holt (1974) e verificaram que o estimador de Pearson é sensível com respeito à especificação correta das relações entre as variáveis de regressão e as variáveis de desenho. Assim, sugerem que a distribuição das probabilidades sob o desenho não pode ser ignorada no processo de inferência. Os autores propuseram o uso de dois estimadores diferentes, ambos baseados no desenho: (a) Estimadores ponderados pelas probabilidades de inclusão

na amostra e (b) Estimadores ponderados pelas probabilidades ajustadas.

Pfeffermann e Holmes (1985) observaram que a modelagem da relação entre as variáveis de regressão e as de desenho faz surgir uma grande e possivelmente mais eficiente família de estimadores, que utilizam tanto a modelagem usual quanto as informações sobre o desenho amostral. O estimador de Máxima Pseudo-Verossimilhança é um exemplo.

Godambe e Thompson (1986) utilizaram os inversos das probabilidades de seleção dos indivíduos como pesos nas equações de Pseudo-Verossimilhança e concluíram que o estimador obtido é um estimador ótimo e que o estimador da variância é um estimador consistente. Este procedimento é simples e atualmente encontra-se implementado em vários pacotes estatísticos.

Silva (1996) investigou o aproveitamento de informações populacionais auxiliares para a estimação de modelos paramétricos empregando o método de Máxima Pseudo-Verossimilhança. Duarte (1999) estendeu o estudo de simulação de Silva (1996) para avaliar o desempenho dos estimadores de variância de diferentes estimadores de um modelo de regressão linear com dados provenientes de uma amostragem aleatória simples e de uma amostragem estratificada.

Em relação à classe de modelos lineares generalizados, Liang e Zeger (1986) propuseram estimar os coeficientes a partir de uma equação de quase-verossimilhança e demonstraram que os estimadores obtidos são consistentes e assintoticamente normais. Rotnitzky e Jewell (1990) consideraram o problema de realizar testes de hipóteses sobre os coeficientes de regressão de modelos na familia exponencial com observações em conglomerados. O ajuste de modelos de regressão em epidemiologia para amostras complexas foi estudado por Binder (1992).

3.2 Inferência Bayesiana

Sugden (1985) argumenta que se os dados não selecionados são considerados dados faltantes e se todas as variáveis usadas na construção do desenho são conhecidas para todas as unidades da população, seria possível considerar que o desenho amostral não faz parte da inferência. Porém, comenta que, geralmente os dados são analisados por pesquisadores ou estatísticos ("analistas") enquanto que a seleção da amostra é realizada por outras pessoas ("amostristas"). Conseqüentemente, os analistas não dispõem de toda a informação da população utilizada no desenho amostral. Portanto, para eles o desenho amostral não é ignorável e as probabilidades de seleção, normalmente reportadas como parte dos dados, carregam toda a informação sobre o desenho amostral. Sugden (1985) fornece vários argumentos com os quais o desenho pode ser considerado ignorável mas, em geral, a inferência com dados de pesquisa por amostragem depende do desenho sempre que apenas uma parte da informação sobre o desenho esteja disponível.

Rubin (1985) argumenta que, embora as probabilidades de inclusão sejam utilizadas pelos freqüentistas para obter estimadores não viciados elas são geralmente consideradas irrelevantes na inferência Bayesiana. Apesar desta posição Rubin (1985) afirma que as probabilidades de inclusão têm um papel importante dentro de inferência Bayesiana aplicada, mas não de uma forma simples. Ele propõe que a modelagem da variável de interesse, $\boldsymbol{y}$, seja condicionada nas probabilidades de inclusão, $\boldsymbol{\pi}$, e não condicionada em todas as variáveis do desenho, $\boldsymbol{v}$, pois $\boldsymbol{\pi} = f(\boldsymbol{v})$ constitui um resumo adequado de $\boldsymbol{v}$. Dado que modelar em função de $\boldsymbol{\pi}$ é mais simples do que modelar em $\boldsymbol{v}$, o Bayesiano que se concentrar em modelos com $\boldsymbol{\pi}$ possivelmente será mais calibrado do que o Bayesiano que constrói modelos com todas as $\boldsymbol{v}$. Rubin (1985) conclui que as probabilidades de inclusão podem ter um papel importante na análise Bayesiana de dados.

Gelman et al. (1995) dedicaram um Capítulo do seu livro ao papel do desenho

na análise Bayesiana ressaltando a importância de sua incorporação dentro da modelagem a ser realizada. Os autores afirmam que se a inferência Bayesiana é utilizada estritamente para analisar a distribuição a posteriori dos parâmetros com um modelo fixo, então, para todo desenho ignorável, o processo de seleção dos dados é irrelevante na inferência desses parâmetros. Porém, não se pode esquecer que:

1. o analista de dados sempre deve utilizar todas as informações relevantes e a forma como os dados foram observados pode ser *informativa*;

2. fazer análise de sensibilidade é parte da inferência Bayesiana e os desenhos *ignoráveis* são apropriados para produzir dados para os quais as inferências são pouco sensíveis à escolha do modelo;

3. pensar no desenho e nos dados pode ajudar na estrutura da inferência sobre modelos e previsões sobre a população finita, e o mais importante;

4. mesmo com verossimilhanças fixas, $p(\boldsymbol{y} \mid \boldsymbol{x}, \theta)$, a distribuição à posteriori muda de acordo com diferentes desenhos não ignoráveis.

Baseados nas observações anteriores, Gelman et al. (1995) afirmam que é necessário, trabalhar com uma estrutura formal para modelar a forma de escolha da amostra, i.e, incluir o desenho amostral durante a realização de inferência estatística.

3.2.1 Verossimilhança Completa e Verossimilhança Observada

Usando a notação apresentada na Seção 2.1, tem-se que a forma natural de modelar $\boldsymbol{y}$ levando em conta o desenho amostral é *expandir o espaço amostral* incluindo, além dos dados de interesse, $\boldsymbol{y}$, a variável indicadora, $\boldsymbol{I}$, cujo elemento I_i toma valor 1 se o elemento y_i foi observado (pertence à amostra). Logo, o espaço amostral, no caso de dados obtidos sob desenhos informativos, é o produto do espaço amostral usual para $\boldsymbol{y}$ e o espaço amostral para $\boldsymbol{I}$.

Considerar o desenho ou plano amostral na estimação de modelos implica a modelagem conjunta de $\boldsymbol{y}$ e $\boldsymbol{I}$. É útil dividir a distribuição conjunta $p(\boldsymbol{y}, \boldsymbol{I} \mid \boldsymbol{\theta}, \boldsymbol{\phi})$ em duas partes: (1) O modelo de superpopulação $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ e (2) O modelo para o vetor de inclusão $Pr(\boldsymbol{I} \mid \boldsymbol{y}, \boldsymbol{\phi})$.

Fazendo as seguintes hipóteses:

1. $\boldsymbol{I}$ é conhecido para toda a população.

2. As covariáveis $\boldsymbol{x}$ (quando são utilizadas) são conhecidas para toda a população.

3. O mecanismo de seleção ou medição não muda a distribuição dos dados, i.e., a distribuição dos dados completos $\boldsymbol{y}$ não são afetados pelo indicador de seleção $\boldsymbol{I}$ (hipótese de estabilidade),

e denotando a informação disponível por $(\boldsymbol{y}_{obs}, \boldsymbol{I})$ ou, na presença de covariáveis $\boldsymbol{x}$, $(\boldsymbol{y}_{obs}, \boldsymbol{I}, \boldsymbol{x})$, a modelagem pode ser realizada utilizando-se as seguintes distribuições:

1. Verossimilhança dos dados completos: Dada uma população P, os dados consistem em $(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{I})$ e

$$p(\boldsymbol{y}, \boldsymbol{I} \mid \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\phi}) = p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) Pr(\boldsymbol{I} \mid \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\phi}) \tag{3.1}$$

2. Verossimilhança dos dados observados: Dada uma amostra s, os dados disponíveis consistem em $(\boldsymbol{y}_s, \boldsymbol{x}, \boldsymbol{I})$. A distribuição conjunta de $\boldsymbol{y}_s$ e $\boldsymbol{I}$, dado $\boldsymbol{x}$, é obtida integrando-se (3.1),

$$\begin{aligned} p(\boldsymbol{y}_s, \boldsymbol{I} \mid \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\phi}) &= \int p(\boldsymbol{y}, \boldsymbol{I} \mid \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\phi}) d\boldsymbol{y}_{\bar{s}} \\ &= \int Pr(\boldsymbol{I} \mid \boldsymbol{y}_s, \boldsymbol{y}_{\bar{s}}, \boldsymbol{x}, \boldsymbol{\phi}) p(\boldsymbol{y}_s, \boldsymbol{y}_{\bar{s}} \mid \boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{y}_{\bar{s}} \end{aligned} \tag{3.2}$$

A equação (3.2) não impõe restrições sobre o mecanismo de seleção da amostra.

3. Distribuição à posteriori conjunta de $(\boldsymbol{\theta}, \boldsymbol{\phi})$:

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \boldsymbol{x}, \boldsymbol{y}_s, \boldsymbol{I}) &\propto p(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \boldsymbol{x}) p(\boldsymbol{y}_s, \boldsymbol{I} \mid \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\phi}) \\ &= p(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \boldsymbol{x}) \int p(\boldsymbol{y}, \boldsymbol{I} \mid \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\phi}) d\boldsymbol{y}_{\bar{s}} \\ &= p(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \boldsymbol{x}) \int p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) Pr(\boldsymbol{I} \mid \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\phi}) d\boldsymbol{y}_{\bar{s}} \end{aligned} \quad (3.3)$$

4. Distribuição à posteriori de $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{y}_s, \boldsymbol{I}) = p(\boldsymbol{\theta} \mid \boldsymbol{x}) \int \int p(\boldsymbol{\phi} \mid \boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) Pr(\boldsymbol{I} \mid \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\phi}) d\boldsymbol{y}_{\bar{s}} d\boldsymbol{\phi} \quad (3.4)$$

Sendo a última a distribuição de maior interesse pois na prática $\boldsymbol{\phi}$ geralmente carece de interesse científico.

3.2.2 Ignorabilidade

Ignorar o desenho amostral significa não se considerar $Pr(\boldsymbol{I} \mid \boldsymbol{y}_s, \boldsymbol{y}_{\bar{s}}, \boldsymbol{x}, \boldsymbol{\phi})$ na equação (3.2). Este procedimento é adequado quando as probabilidades de seleção não dependem de $\boldsymbol{y}$, i.e., $Pr(\boldsymbol{I} \mid \boldsymbol{y}_s, \boldsymbol{y}_{\bar{s}}, \boldsymbol{x}) = Pr(\boldsymbol{I} \mid \boldsymbol{x})$. Neste caso (3.2) torna-se:

$$p(\boldsymbol{y}_s, \boldsymbol{I} \mid \boldsymbol{x}, \boldsymbol{\theta}) = \int p(\boldsymbol{y}_s, \boldsymbol{y}_{\bar{s}} \mid \boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{y}_{\bar{s}} \quad (3.5)$$

Na inferência Bayesiana as seguintes duas condições são suficientes e necessárias para assegurar que o desenho é ignorável:

1. Faltantes ao acaso (*Missing at Random*): Dado $\boldsymbol{\phi}$, $Pr(I \mid .)$ depende só de $\boldsymbol{x}$ e $\boldsymbol{y}_s$, i.e.,

$$Pr(\boldsymbol{I} \mid \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\phi}) = Pr(\boldsymbol{I} \mid \boldsymbol{x}, \boldsymbol{y}_s, \boldsymbol{\phi})$$

2. Parâmetros diferentes: Os parâmetros do processo dos dados faltantes são independentes, dados os valores das covariáveis $\boldsymbol{x}$, dos parâmetros do processo gerador dos dados, i.e.,

$$p(\boldsymbol{\phi} \mid \boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{\phi} \mid \boldsymbol{x})$$

Então quando essas duas condições são satisfeitas, o desenho é dito ignorável e $p(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{y}_s) = p(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{y}_s, \boldsymbol{I})$.

Na realidade, dizer que um desenho é ignorável não significa que ele não fornece informações úteis, mas sim que as probabilidades de seleção não fornecem informação adicional daquela já fornecida pelas variáveis do desenho, as quais podem fazer parte ou não do vetor $\boldsymbol{x}$.

A maioria dos desenhos estatísticos são ignoráveis. Nestes casos só é necessário o conhecimento das distribuições $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})$ e $p(\theta)$ para a realização apropriada da inferência sobre $\boldsymbol{\theta}$. Exemplos desta classe de desenhos são:

- Amostragem Aleatória Simples
- Experimentos Completamente Aleatorizados

Gelman et al. (1995) afirmam que o conceito de ignorabilidade não fornece uma boa justificativa para se afirmar que utilizando-se dados e modelos fixos, o desenho amostral sempre é irrelevante para inferência Bayesiana. Porém, destacam que se deve considerar o desenho amostral na análise pois com uma função de verossimilhança fixa $p(\boldsymbol{y} \mid \boldsymbol{\theta})$, e dados fixos $\boldsymbol{y}$, a distribuição à posteriori depende dos mecanismos de seleção *não ignoráveis*.

Sob desenhos amostrais não ignoráveis ou informativos, é possível incluir covariáveis apropriadas no modelo para tornar o desenho ignorável. Esta regra não é exclusividade da inferência Bayesiana. Aumentar covariáveis no modelo parece ser uma solução adequada e simples, contudo alguns dos seguintes problemas podem ser enfrentados:

1. O número de parâmetros do modelo pode crescer demasiadamente devido ao aumento de muitas variáveis no modelo, isto é fácil de acontecer quando, por exemplo, a população está dividida em muitos conglomerados ou estratos.

2. Os novos parâmetros do modelo podem não ter uma interpretação válida ou não ser de interesse científico.

3. O modelo pode-se tornar sensível a pequenas modificações.

Há muitos cenários nos quais o mecanismo de seleção dos dados é conhecido mas não ignorável. Dois exemplos importantes são os dados censurados e os dados truncados. Outro exemplo é o caso dos dados de pesquisa por amostragem onde as probabilidades de seleção (π_i) são conhecidas só para os elementos pertencentes à amostra.

3.2.3 Exemplo

Nesta Seção reproduz-se o exemplo de Gelman et al. (1995). Este exemplo ilustra o caso em que se faz inferência a partir de dados de uma amostra, sob um plano amostral informativo, onde as probabilidades de seleção são conhecidas só para os indivíduos na amostra.

Considere uma pesquisa entre adultos, onde as mulheres têm π_1 de probabilidade de serem amostradas e os homens, π_2. Assuma que π_1, π_2 e N são conhecidos mas $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$, as quantidades de mulheres e homens na população, são desconhecidas. O mecanismo de seleção é não ignorável pois a variável sexo, $\boldsymbol{x} = (x_1, \ldots, x_N)$ onde $x_i =$ 1 ou 2, não é observada para todos os elementos da população. Mas, condicionado em $\boldsymbol{\lambda}$ o desenho é ignorável.

Seja $\boldsymbol{y} = (y_1, \ldots, y_N)$ a variável de interesse, com distribuição normal condicionada em $\boldsymbol{x}$. Então os parâmetros a serem modelados são $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \mu_1, \mu_2, \sigma_1, \sigma_2)$ e

os dados observados são $(\boldsymbol{y}_s, \boldsymbol{x}_s), n_1$ e n_2, onde n_1 e n_2 são o número de mulheres e de homens na amostra.

O mecanismo de seleção está representado na seguinte distribuição:

$$\begin{aligned} Pr(\boldsymbol{I} \mid \boldsymbol{\lambda}) &= \prod_{i=1}^{N} \pi_{x_i}^{I_i}(1-\pi_{x_i})^{1-I_i} \\ &= \pi_1^{n_1}(1-\pi_1)^{(\lambda_1 N - n_1)}\pi_2^{n_2}(1-\pi_2)^{(\lambda_1 N - n_2)} \\ &\propto (1-\pi_1)^{\lambda_1 N}(1-\pi_2)^{\lambda_2 N} \qquad , \end{aligned}$$

As distribuições das variáveis $\boldsymbol{y}$ e $\boldsymbol{x}$ na população são

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) \sim \prod_{i=1}^{N} p(y_i \mid \mu_{x_i}, \sigma_{x_i}^2), \qquad p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \binom{N}{\lambda_1 N}^{-1},$$

e a distribuição a posteriori de $\boldsymbol{\theta}$, condicionada na amostra disponível, é:

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{y}_s, \boldsymbol{x}_s, I) &\propto p(\boldsymbol{\theta}) p(\boldsymbol{y}_s, \boldsymbol{x}_s, \boldsymbol{I} \mid \boldsymbol{\theta}) \\ &= p(\boldsymbol{\theta}) \sum_{\boldsymbol{x}_{\bar{s}}} \int p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{I} \mid \boldsymbol{\theta}) dy_{\bar{s}} \\ &\propto p(\boldsymbol{\theta}) \sum_{\boldsymbol{x}_{\bar{s}}} \int \binom{N}{\lambda_1 N}^{-1} \left[\prod_{i=1}^{N} p(y_i \mid \mu_{x_i}, \sigma_{x_i}^2)\right] (1-\pi_1)^{\lambda_1 N}(1-\pi_2)^{\lambda_2 N} dy_{\bar{s}} \\ &= p(\boldsymbol{\theta}) \binom{N}{\lambda_1 N}^{-1} \binom{N-n}{\lambda_1 N - n_1} (1-\pi_1)^{\lambda_1 N}(1-\pi_2)^{\lambda_2 N} \prod_{i=1}^{n_1+n_2} p(y_{s,i} \mid \mu_{x_{s,i}}, \sigma_{x_{s,i}}^2) \\ &\propto p(\boldsymbol{\theta}) \binom{\lambda_1 N}{n_1} \binom{\lambda_2 N}{n_2} (1-\pi_1)^{\lambda_1 N}(1-\pi_2)^{\lambda_2 N} \prod_{i=1}^{n_1+n_2} p(y_{s,i} \mid \mu_{x_{s,i}}, \sigma_{x_{s,i}}^2). \end{aligned}$$

3.3 Aproximação da Distribuição Amostral

Em Pfeffermann et al. (1998), afirma-se que em geral é sempre possível aproximar a distribuição paramétrica dos dados de uma amostra e a partir dela, fazer inferência sobre a distribuição da população de origem, explorando a relação existente entre ambas distribuições.

Os autores fazem uso do teorema de Bayes para obter a distribuição (marginal) amostral de y_i , condicionando a distribuição dos elementos observados ao fato de terem sido incluídos na amostra, i.e.,

$$f_s(y_i \mid \boldsymbol{\theta}, \boldsymbol{\phi}) = p(y_i \mid \boldsymbol{\theta}, \boldsymbol{\phi}, I_i = 1) = \frac{Pr(I_i = 1 \mid y_i, \boldsymbol{\phi}) f_p(y_i \mid \boldsymbol{\theta})}{Pr(I_i = 1 \mid \boldsymbol{\phi})}, \tag{3.6}$$

onde $I_i = 1$ indica que o elemento $i \in s$ e $\boldsymbol{\phi}$ representa os parâmetros relativos ao mecanismo de seleção.

No caso em que a distribuição populacional depende de variáveis concomitantes, a densidade amostral marginal de y_i é definida por

$$f_s(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{Pr(I_i = 1 \mid y_i, \boldsymbol{x}_i, \boldsymbol{\phi}) f_p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})}{Pr(I_i = 1 \mid \boldsymbol{x}_i, \boldsymbol{\phi})}. \tag{3.7}$$

Segundo Pfeffermann et al. (1998) esta densidade pode ser vista como um caso especial da familia de distribuições ponderadas ("*weighted distributions*") definida por Rao (1965). As distribuições ponderadas surgem quando a probabilidade (ou densidade) de uma potencial observação y é "distorcida", i.e., a probabilidade (ou densidade) $g(y \mid \theta)$ é multiplicada por alguma função (não-negativa) de ponderação $w(y)$, a qual pode envolver alguns parâmetros desconhecidos. Na equação (3.7), os dados observados constituem uma amostra aleatória da seguinte versão ponderada de $g(y \mid \theta)$:

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

onde a esperança do denominador é a constante de normalização de $p(y \mid \theta)$. Bayarri e DeGroot (1992) apresentaram um resumo do estudo realizado por eles sobre essa familia de distribuições.

Pfeffermann et al. (1998) recomendam e justificam o uso da distribuição amostral pois demonstraram, através de resultados teóricos e de simulação, que sob certas condições, as observações de uma amostra proveniente de uma população de observações independentes são assintoticamente independentes. Logo, podem ser uti-

lizados procedimentos padrões de estimação eficiente, o que, segundo os autores, é a principal vantagem deste método.

No mesmo artigo apresenta-se uma expressão alternativa a (3.6). Tem-se que mesmo quando $\pi_i = Pr(I_i = 1 \mid \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\phi}) \neq Pr(I_i = 1 \mid y_i, \boldsymbol{x}_i, \boldsymbol{\phi})$, cumpre-se a seguinte relação:

$$\begin{aligned} Pr(I_i = 1 \mid y_i, \boldsymbol{x}_i, \boldsymbol{\phi}) &= \int Pr(I_i = 1 \mid y_i, \boldsymbol{x}_i, \boldsymbol{\phi}, \pi_i) f_p(\pi_i \mid y_i, \boldsymbol{x}_i, \boldsymbol{\phi}) d\pi_i \\ &= \quad E_p[\pi_i \mid y_i, \boldsymbol{x}_i, \boldsymbol{\phi}], \end{aligned} \tag{3.8}$$

pois $Pr(I_i = 1 \mid y_i, \boldsymbol{x}_i, \boldsymbol{\phi}, \pi_i) = \pi_i$. Então, substituindo (3.8) em (3.7) tem-se

$$f_s(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{\phi}) = f(y_i \mid \boldsymbol{x}_i, I_i = 1, \boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{E_p[\pi_i \mid y_i, \boldsymbol{x}_i, \boldsymbol{\phi}] f_p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})}{E_p[\pi_i \mid \boldsymbol{x}_i, \boldsymbol{\phi}]}. \tag{3.9}$$

A partir de (3.9) os autores afirmam que para qualquer fdp populacional dada, a correspondente fdp amostral é totalmente determinada pela esperança condicional $E_p[\pi_i \mid y_i, \boldsymbol{x}_i, \boldsymbol{\phi}]$.

Os autores desta proposta lembram que sob as amostragens padrões, as observações na amostra não são independentes. Porém, eles estabeleceram algumas condições sob as quais, observações que são independentes na população, são assintoticamente independentes na amostra. Então considerando independência assintótica, a distribuição conjunta dos dados observados é:

$$f_s(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \prod_{i \in s} \frac{E_p[\pi_i \mid y_i, \boldsymbol{x}_i, \boldsymbol{\phi}] f_p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})}{E[\pi_i \mid \boldsymbol{x}_i, \boldsymbol{\phi}]}. \tag{3.10}$$

Um dos resultados apresentado no artigo afirma que sob algumas condições de regularidade, os valores esperados $E_p[\pi_i \mid y_i, \boldsymbol{x}_i, \boldsymbol{\phi}]$ podem ser aproximados por polinômios de baixa ordem em y_i e $\boldsymbol{x}_i$, ou por exponenciais de tais polinômios, via a expansão de séries de Taylor. Assim, para o primeiro caso tem-se:

$$E_p[\pi_i \mid y_i, \boldsymbol{x}_i] \approx \sum_{j=0}^{J} A_j y_i^j + h(\boldsymbol{x}_i), \tag{3.11}$$

onde $h(\boldsymbol{x}_i) = \sum_{p=1}^{m} \sum_{k=1}^{K(p)} B_{kp} x_{ip}^k$ e $\{A_j\}$ e $\{B_{kp}\}$ são parâmetros desconhecidos a serem estimados a partir dos dados observados. Substituindo (3.11) em (3.9) e assumindo a existência de $E^{(j)} = E_p[Y_i^j \mid \boldsymbol{x}_i]$, a função de distribuição amostral pode ser aproximada por:

$$f_s(y_i \mid \boldsymbol{x}_i) \approx \frac{\sum_{j=1}^{J}(A_j E^{(j)}) f_p^{(j)}(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) + [A_0 + h(\boldsymbol{x}_i)] f_p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})}{\sum_{j=1}^{J}(A_j E^{(j)}) + [A_0 + h(\boldsymbol{x}_i)]}, \qquad (3.12)$$

onde $f_p^{(j)}(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) = y_i^j f_p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})/E^{(j)}$. Observa-se em (3.12) que a função de distribuição amostral é agora uma mistura das densidades $f_p^{(j)}(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}), j = 1, \ldots, J$.

3.4 Comentários

Como exposto em Smith (2001) o conjunto de resultados demonstra convincentemente que uma análise de dados de pesquisa por amostragem complexa baseada em suposições da amostragem aleatória simples, não é apropriada.

Entre as características comuns dos métodos propostos na Inferência Clássica tem-se que todos requerem a hipótese de normalidade assintótica implicando a necessidade de contar com tamanhos de amostra grandes e impossibilidade de se utilizar procedimentos da inferência clássica tais como gráfico de resíduos e testes estatísticos. Além disso, cada método está desenvolvido analiticamente para casos particulares de desenhos amostrais, geralmente para amostragem aleatória simples e amostragem estratificada.

Alguns dos métodos propostos requerem o conhecimento de informações detalhadas sobre os elementos da amostra, como estratos e conglomerados aos quais pertencem e suas probabilidades de inclusão na amostra. Outros requerem informações auxiliares sobre a população. Contudo, uma vantagem do método de Pseudo-Verossimilhança é sua simplicidade o que permitiu sua disponibilidade em pacotes estatísticos comerciais.

Uma observação importante a ser feita é que os trabalhos mencionados na Seção

3.1 tratam do problema do ajuste de modelos lineares com dados de pesquisa por amostragem complexa, que têm por objetivo principal fornecer estimadores não viciados. Mas, nenhum desses métodos trata explicitamente do problema dos desenhos amostrais informativos, que é uma das possíveis conseqüências do uso de amostragem complexa.

O tratamento dos desenhos amostrais informativos dentro da Inferência Bayesiana é análogo ao tratamento do problema de não resposta não ignorável (Qin et al., 2002). Como foi mencionado no Capítulo 1, os dados amostrais podem ser considerados como resultado de dois processos aleatórios. O primeiro processo, ou modelo de superpopulação, gera a população. Porém, os dados não são completamente observados neste primeiro processo. Condicionado às observações do primeiro processo, o segundo processo (o mecanismo de seleção de amostras) gera um subconjunto de dados que são completamente observados. Qin et al. (2002) afirmam que os dados de pesquisa com não resposta são um exemplo desse tipo de dados onde o segundo processo corresponde ao mecanismo de resposta.

Na pratica não é comum conhecer as probabilidades de seleção nem as variáveis do desenho de todos os elementos da população, estas são conhecidas só para os elementos da amostra. Como Rubin (1985) afirma, a inferência envolvendo valores não observados de $\boldsymbol{y}$, i.e. quando $I_i = 0$, deve se sustentar em hipóteses que não são diretamente verificáveis a partir dos dados observados.

A equação (3.2), usando o resultado de independência assintótica obtido por Pfeffermann et al. (1998), pode ser escrita como:

$$p(\boldsymbol{y}_s, \boldsymbol{I} \mid \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \prod_{i \in s} \Big[\frac{E(\pi_i \mid y_i, \boldsymbol{x}_i, \boldsymbol{\phi}) f_p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})}{Pr(i \in s \mid \boldsymbol{x}_i, \boldsymbol{\phi})} \Big] \prod_{i \in s} Pr(i \in s \mid \boldsymbol{x}_i, \boldsymbol{\phi}) \prod_{i \notin s} \big(1 - Pr(i \in s \mid \boldsymbol{x}_i, \boldsymbol{\phi})\big), \quad (3.13)$$

onde o termo entre corchetes ([]) corresponde à distribuição amostral definida em (3.9).

Métodos numéricos são necessários para aproximar as distribuições (3.3) e (3.4). Existem algumas propostas para trabalhar com verossimilhanças do tipo (3.13), especificamente para o tratamento do problema de dados faltantes (*missing data*). Uma das mais recentes é de Qin et al. (2002) que propôs o uso de um modelo semi-paramétrico, assumindo um modelo paramétrico para o mecanismo de resposta ($Pr(i \in s)$) mas um modelo não paramétrico para a distribuição de y ($f_p(y_i \mid \boldsymbol{x}_i)$). Outra proposta, envolvendo o algoritmo EM, foi apresentada por Ibrahim, Chen, e Lipsitz (2001).

Comparando as expressões (3.10) e (3.13) observa-se que a distribuição conjunta amostral é uma das parcelas da verossimilhança dos dados observados. Segundo Pfeffermann et al. (1998) a parcela faltante em (3.13) não é operacional dado que o produto $\prod_{i \notin s}\big(1 - Pr(i \in s \mid \boldsymbol{x}_i, \boldsymbol{\phi})\big)$ depende de valores $\boldsymbol{x}_i$ que geralmente não fazem parte dos dados disponíveis para o analista.

A distribuição amostral pode depender de muitos mais parâmetros que a distribuição da população, porém, permite usar a inferência Bayesiana de forma natural e os resultados estarão contidos nas distribuições a posteriori, assim, supera-se os métodos clássicos que se limitam à estimação pontual.

Capítulo 4

MODELOS HIERÁRQUICOS SOB DESENHOS AMOSTRAIS INFORMATIVOS

4.1 Introdução

Grande parte das populações investigadas nas ciências sociais para serem respondidas perguntas científicas e/ou para tomar decisões têm uma estrutura hierárquica. Economia, Educação e Saúde Pública são apenas algumas áreas onde os exemplos surgem naturalmente. Draper (1995) argumenta que o uso de Modelos Hierárquicos (MHs) têm três vantagens claras sobre outros métodos utilizados na análise de dados de ciências sociais. Primeiro, os MHs fornecem um ambiente natural onde expressar e comparar teorias sobre possíveis relações estruturais entre variáveis de cada nível. Segundo, o ajuste de MHs produz avaliações de incerteza melhor calibradas na presença de correlações intraclasses positivas típicas das ciências sociais. Finalmente, os MHs oferecem *framework* explícito para expressar a permutabilidade das unidades, permitindo combinar informação sobre unidades de diferentes níveis (por exemplo, alunos e escolas) para a obtenção de previsões acuradas e bem calibradas.

Enquanto a modelagem linear hierárquica (MLH) é extensamente aplicada, os pesquisadores percebem que os dados disponíveis quase sempre são provenientes de pesquisas por amostragem complexa e de grande escala. Os procedimentos de seleção das amostras geralmente são de várias etapas, com probabilidades desiguais de seleção, de conglomerados, etc. Então, a amostra disponível é o produto do modelo hierárquico subjacente e o procedimento de seleção da amostra. Ignorar o efeito de seleção pode, do ponto de vista freqüêntista, causar viés tanto nas estimativas

pontuais quanto nas variâncias das mesmas (Zhang & Mike, 2000).

No Capítulo anterior descreveu-se o problema de estimação de modelos lineares, especificamente de modelos de regressão com dados de pesquisa com amostragem informativa. Neste Capítulo descrevem-se algumas alternativas disponíveis na literatura que tratam do problema de estimação de parâmetros de modelos de superpopulação do tipo hierárquico ou multinível, que no caso de uma variável resposta normal pode ser representado por:

$$y_{ij} \mid \boldsymbol{\beta}, \boldsymbol{x}_{ij}, \sigma^2 \sim N(\boldsymbol{x}'_{ij}\boldsymbol{\beta}, \sigma^2) \tag{4.1}$$

$$\beta_{0i} \mid \boldsymbol{z}_i, \boldsymbol{\gamma}, \sigma^2_{\beta_0} \sim N(\boldsymbol{z}'_i\boldsymbol{\gamma}, \sigma^2_{\beta_0}) \tag{4.2}$$

O procedimento de ponderação MQGIPP e o método da Distribuição Amostral são as duas propostas para modelos lineares hierárquicos normais, que se apresentam nas Seções 4.3 e 4.4. A extensão do uso das distribuições amostrais para modelos lineares hierárquicos generalizados, na presença de desenhos amostrais informativos, é apresentada na Seção 4.5.

4.2 Modelos Multinível e Amostragem Complexa

Do ponto de vista freqüêntista, diferentes abordagens tem sido propostas para evitar o problema da presença de viés nas estimativas pontuais de parâmetros de um modelo hierárquico com dados obtidos com uma amostragem complexa. Uma delas é utilizar um modelo de regressão, onde os dados das unidades do segundo nível (p.e. escolas) são combinados com dados do primeiro nível (p.e. alunos) e é ajustado um modelo de somente um nível. Este método usa o seguinte estimador de mínimos quadrados ponderados para estimar $\boldsymbol{\gamma}$:

$$\hat{\boldsymbol{\gamma}}_{MQP} = \left(\boldsymbol{z}'_s\boldsymbol{x}'_s\boldsymbol{\pi}_s^{-1}\boldsymbol{x}_s\boldsymbol{z}_s\right)^{-1}\boldsymbol{z}'_s\boldsymbol{x}'_s\boldsymbol{\pi}_s^{-1}\boldsymbol{y}_s,$$

onde $\boldsymbol{\pi}_s = diag(\pi_1, \ldots, \pi_n)$ e π_i é a probabilidade da unidade i de pertencer à amostra. Este procedimento ignora a estrutura hierárquica da população.

A análise de regressão multinível ordinaria é uma segunda abordagem. Ela leva em conta a estrutura hierárquica dos erros mas ignora o desenho amostral. Um argumento, equívoco, desta abordagem é que a natureza multinível dos MHs modela diretamente o desenho amostral de vários estágios usado para selecionar a amostra (Zhang & Mike, 2000).

Um caso particular de interesse, onde a análise de regressão multinível ordinaria pode ser utilizada é aquele onde todas as variáveis do desenho são incorporadas como covariáveis do modelo, i.e, fazem parte de $\boldsymbol{x}_s$ ou $\boldsymbol{z}_s$. Supondo que $\boldsymbol{x}_s$ e $\boldsymbol{z}_s$ representam as variáveis utilizadas na seleção de s, então, o conhecimento do mecanismo de seleção é redundante para $\boldsymbol{y}_s$ dado ($\boldsymbol{x}_s$,$\boldsymbol{z}_s$), portanto pode ser ignorado.

Segundo Zhang e Mike (2000), situações onde as variáveis do desenho coincidem com as covariáveis não são raras na área de educação. Por exemplo, freqüêntemente o tipo de escola (pública ou privada) e a etnia do aluno são variáveis de estratificação utilizadas respectivamente na primeira e segunda etapa do processo de seleção das amostras. Porém, não se deve esquecer que na prática a informação sobre as variáveis do desenho limita-se à amostra disponível e que, nesta condição, o desenho ainda pode ser informativo para a realização de inferência sobre os parâmetros.

Outra abordagem utilizada para a estimação de MHs é denominada Análise de Regressão Multinível Ordinária Ponderada. Este método incorpora pesos amostrais durante a estimação. Para o caso da amostragem por conglomerados, Pfeffermann e LaVange (1989) propuseram o seguinte estimador:

$$\hat{\boldsymbol{\gamma}}_{MQGP} = \sum_{c}^{m} \frac{1}{\pi_c} \Big[\boldsymbol{x}_c^{*'} \boldsymbol{w}_c \boldsymbol{x}_c^{*} - \boldsymbol{x}_c^{*'} \boldsymbol{w}_c \boldsymbol{x}_c Q_{c,w}^{-1} \boldsymbol{x}_c' \boldsymbol{w}_c \boldsymbol{x}_c^{*} \Big] R$$

onde $Q_c = \boldsymbol{x}_c' \boldsymbol{x}_c + \sigma^2 \Delta^{-1}$. π_c é a probabilidade de inclusão do conglomerado c, $\boldsymbol{w}_c = diag(w_{c1}, \ldots, w_{cn_c})$ é a matriz de pesos amostrais correspondente as unidades dentro do conglomerado c, $Q_{c,w} = \boldsymbol{x}_c' \boldsymbol{w}_c \boldsymbol{x}_c + \sigma^2 \Delta^{-1}$, $\boldsymbol{x}_c^{*} = \boldsymbol{x}_c \boldsymbol{z}_c$ e $R = \sum_{c=1}^{m} \frac{1}{\pi_c} \Big[\boldsymbol{x}_c^{*'} \boldsymbol{w}_c \boldsymbol{y}_c - \boldsymbol{x}_c^{*'} \boldsymbol{w}_c \boldsymbol{x}_c Q_{c,w}^{-1} \boldsymbol{x}_c' \boldsymbol{w}_c \boldsymbol{y}_c \Big]$. Nesta abordagem somente é considerada uma parte de informação do desenho (pesos amostrais). Sugden e Smith (1984) expõem alguns casos

onde o conhecimento dos pesos amostrais é suficiente para assumir a ignorabilidade. Porém deve ser destacado que o conhecimento das probabilidades de inclusão para todas as unidades não é suficiente para ignorar o desenho amostral.

A análise de Regressão Multinível baseada no desenho é um aperfeiçoamento do método anterior, o procedimento de ponderação MQGIPP é um exemplo desta abordagem.

4.3 Procedimento de Ponderação MQGIPP

Na Seção anterior foram descritos alguns dos métodos propostos para o ajuste de modelos multinível com dados de pesquisa por amostragen complexa mas que não tratam dos desenhos amostrais informativos em forma particular. Pfeffermann, Skinner, Holmes, Goldstein, e Rasbash (1998), propuseram um procedimento de ponderação para a estimação dos parâmetros de modelos lineares hierárquicos com o objetivo de corrigir vícios na estimação dos parâmetros sob desenhos amostrais informativos. Esse procedimento é uma adaptação do método dos Mínimos Quadrados Generalizados Iterativo (MQGIPP), por analogia ao método de máxima pseudo-verossimilhança.

A idéia básica do procedimento de ponderação MQGIPP é que a seleção da amostra não acarretaria vícios na estimação se os valores das variáveis de interesse fossem observados para todas as unidades da população (como em um censo). O procedimento consiste em usar as probabilidades de inclusão na amostra como ponderadores dos valores observados e logo obter estimadores consistentes e aproximadamente não viciados das estimativas "censais", os principais passos desse procedimento são

1. Supor que todas as unidades da população foram observadas e escrever o "modelo completo".

2. Escrever as equações necessárias para utilizar o método dos Mínimos Quadrados

Generalizados Iterativo no "modelo completo".

3. Nas equações resultantes, substituir todos as valores (somatórios) populacionais pelos respectivos valores amostrais observados, ponderados pelos respectivos inversos das probabilidades de inclusão na amostra

4. Aplicar o método dos Mínimos Quadrados Generalizados Iterativo no "modelo censal".

Uma aplicação do MQGIPP foi realizada por Corrêa (2001) onde ajustou um modelo linear normal de dois níveis para relacionar um indicador do estado nutricional de adultos com outras variáveis determinantes da qualidade de vida da população das regiões Nordeste e Sudeste do Brasil a partir de dados da amostra da Pesquisa sobre Padrões de Vida - PPV, desenvolvida pelo Instituto Brasileiro de Geografia e Estatística (IBGE) nos anos 1996-1997. Corrêa (2001) comparou o procedimento de ponderação proposto por (Pfeffermann et al., 1998) com três tipos de ajustes [1] disponíveis no pacote computacional MlwiN 1.10 [2](Rasbash et al. (2000)). Corrêa (2001) concluiu que os valores obtidos com o MQGIPP foram idênticos aos obtidos com a opção de pesos amostrais padronizados do MlwiN 1.10. Contudo os desvios padrões deste último ajuste apresentaram vício. As outras duas alternativas forneceram estimativas discrepantes.

Entretanto, em Pfeffermann et al. (2002), os autores afirmam que o MQGIPP tem quatro importantes limitações:

1. As variâncias dos estimadores ponderados são geralmente maiores que as variâncias dos correspondentes estimadores não ponderados.

---

[1] com pesos padronizados, com pesos não padronizados e sem pesos

[2] o procedimento do MlwiN considera que os pesos são independentes dos efeitos aleatórios de cada nível

2. A inferência é restrita principalmente à estimação pontual. Não é possível determinar a distribuição exata dos estimadores pontuais ponderados.

3. O uso dos "pesos amostrais" não permite em geral condicionar sob as probabilidades de seleção das unidades de segundo ou mais alto nível ou nas variáveis independentes do modelo.

4. Não é claro como fazer previsão dos efeitos de segundo e mais alto nível.

4.4 A Distribuição Amostral no Modelo Linear Hierárquico Normal

Pfeffermann, Moura, e Silva (2002), desenvolveram uma proposta sobre o uso das distribuições amostrais propostas por Pfeffermann, Krieger, e Rinott (1998) para modelos lineares hierárquicos normais. A idéia utilizada foi obter o modelo hierárquico amostral como função do modelo populacional e das probabilidades de inclusão de primeira ordem das unidades na amostra e ajustar-lo utilizando técnicas usuais de estimação.

Os autores formularam um modelo de superpopulação hierárquico linear normal de dois níveis e fizeram estudo de simulação com 400 populações e 1600 amostras. As amostras foram obtidas com desenhos amostrais de dois estágios e os modelos foram ajustados com o método de estimação MCMC. Eles compararam os resultados da estimação usando as distribuições amostrais (SM) com o método de ponderação MQGIPP [3] e concluíram que os viés estimados com os dois métodos é geralmente muito menor do que os viés estimados com o modelo que ignora o desenho (IG), sendo que os viés observados destes dois modelos (MQGIPP e SM) foram similares sob todos os planos amostrais avaliados, exceto para as componentes de variância. Entretanto, o uso do modelo amostral permitiu obter melhores coberturas dos intervalos de confiança. Na simulação, o uso das distribuições amostrais (SM) produz

[3] no artigo o método é denominado *Probability Weighting* (PW)

percentagens de cobertura quase perfeitos para todos os parâmetros sob todos os planos amostrais, o que não aconteceu com o método MQGIPP e o modelo IG. Os autores atribuem a má performance do MQGIPP ao tamanho amostral utilizado, eles afirmam que neste caso, a aproximação normal não é válida para a obtenção dos intervalos de confiança.

Pfeffermann et al. (2002) afirmam que o MQGIPP tem duas vantagens sobre o uso das Distribuições Amostrais: o modelo populacional não requer modificação e necessita menor esforço computacional. Porém, este método apresenta sérias limitações, já mencionadas na Seção 4.3. O uso das distribuições amostrais (SM) é mais flexível e a sua principal vantagem é a boa cobertura dos intervalos de credibilidade. Contudo este método tem algumas desvantagens:

- Requer a especificação das esperanças condicionais das probabilidades de seleção em cada um dos níveis do modelo;

- A robustez do uso de distribuições amostrais a má especificação ainda não foi avaliada.

4.5 A Distribuição Amostral no Modelo Linear Hierárquico Generalizado

Nos trabalhos citados nas Seções 4.3 e 4.4 apresentam-se resultados teóricos e práticos sobre a realização de inferência analítica a partir de amostras sob desenhos amostrais informativos, para modelos onde a variável de interesse tem distribuição normal. Dos trabalhos mencionados no Capítulo anterior, tanto Gelman et al. (1995), sob o ponto de vista Bayesiano como Pfeffermann et al. (1998), do ponto de vista freqüêntista, não particularizam os seus resultados ao caso normal, mas não explicitam a sua extensão para modelos hierárquicos. Nesta Seção apresentam-se alguns resultados teóricos sobre a realização de inferência analítica a partir de amostras sob desenhos informativos em modelos lineares hierárquicos generalizados.

4.5.1 A Distribuição Amostral na Família Exponencial

Antes de apresentar a forma de obter e utilizar as distribuições amostrais em modelos hierárquicos, apresenta-se a proposição de invariância do artigo do Pfeffermann et al. (1998) para as distribuições amostrais de variáveis cuja distribuição populacional pertence à família exponencial. Segundo esta proposição, se a esperança da probabilidade de seleção dos elementos tem uma forma particular definida [4], a Distribuição Amostral pertence também à família exponencial. Esta proposição é anunciada a seguir:

Seja a fdp da população pertencente à familia exponencial, i.e.,

$$f_p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_i) = a_i(\boldsymbol{\theta}_i) \exp\left[\sum_{k=1}^{K} \theta_{ki} b_{ki}(y_i) + c_i(y_i)\right] \tag{4.3}$$

onde $\boldsymbol{\theta}_i = (\theta_{1i}, \ldots, \theta_{Ki})'$ toma valores no espaço de parâmetros $\Theta \subset \boldsymbol{R}^K$, e $b_{ki}(.)$ e $c_i(.)$ são funções conhecidas.

Supondo que as probabilidades de inclusão na amostra têm média

$$E_p(\pi_i \mid y_i, \boldsymbol{x}_i) = r_i \exp\left[\sum_{k=1}^{K} d_{ki} b_{ki}(y_i)\right] \tag{4.4}$$

onde r_i e $\{d_{ki}\}$ são constantes que podem depender de $\boldsymbol{x}_i$, mas não de y_i. A seguinte proposição fornece uma "propriedade de invariância de distribuição".

Proposição 4.5.1 *Se a fdp da população de y_i pertence à família exponencial definida por (4.3) e as probabilidades de inclusão na amostra obedecem (4.4), então a fdp da amostra pertence também à família exponencial com parâmetros $\theta^*_{ki} = \theta_{ki} + d_{ki}$.*

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

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

[4] Ver Proposição

e seja a esperança das probabilidades de seleção $E_p(\pi_i \mid y_i) \propto y_i$. Então, a distribuição amostral de y_i é outra vez Gama com parâmetro de forma $(\alpha + 1)$ e com média $\mu_i(\alpha + 1)/\alpha$.

O resultado estabelecido na Proposição 4.5.1 é parecido com o resultado familiar da identificação de distribuições a priori conjugadas na Inferência Bayesiana. Interessante é, segundo Pfeffermann et al. (1998) , que Cox e Hinkley (1974) chamaram à família de distribuições à priori para as quais a distribuição à posteriori pertence a mesma familia de distribuições fechadas por amostragem (*closed under sampling*), termo apropriado para este contexto.

A dependência do $\boldsymbol{x}_i$ nas equações (4.3) e (4.4) opera de uma forma muito geral através de θ_{ki} e d_{ki} respectivamente. Esta dependência pode ser mais explícita para a classe de modelos de regressão de $\boldsymbol{y}$ sobre $\boldsymbol{x}$ se as seguintes relações lineares são assumidas:

$$\theta_{ki} = \phi_{0k} + \boldsymbol{x}_i'\boldsymbol{\phi}_k; \qquad d_{ki} = \Psi_{0k} + \boldsymbol{x}_i'\boldsymbol{\Psi}_k. \tag{4.5}$$

Corolario 4.5.1 *Sob as condições da Proposição 4.5.1 e os supostos (4.5), a fdp amostral pertence à mesma familia restrita com ϕ_{0k} e $\boldsymbol{\phi}_k$ substituídas por $(\phi_{0k} + \Psi_{0k})$ e $(\boldsymbol{\phi}_k + \boldsymbol{\Psi})$ respectivamente. Em particular, se as funções d_{ki} não dependem de $\boldsymbol{x}_i$, i.e., $\Psi_k = 0$, os coeficientes de $\boldsymbol{x}_i$ na parametrização natural da pdf amostral são os mesmos para da fdp populacional.*

Lembrando que a distribuição amostral é um caso particular das distribuições ponderadas, a Proposição 4.5.1 é importante também porque garante que possam ser utilizadas as prioris e métodos de aproximação usuais para o MCMC, o que não acontece com uma classe particular das distribuições ponderadas e que segundo Bayarri e DeGroot (1992), o uso das prioris "de rotina" (prioris impróprias ou prioris conjugadas) pode ser inadequado.

4.5.2 A Distribuição Amostral em Modelos Hierárquicos

Para a utilização do método da Distribuição Amostral em Modelos Hierárquicos é importante considerar a seguinte hipótese: o efeito do plano amostral é independente em cada nível da hierarquia. Logo, para estabelecer as distribuições amostrais, necessitam-se conhecer os valores esperados das probabilidades de seleção dos elementos em cada nível da hierarquia, i.e, a variável indicadora I que denota se o indivíduo pertence à amostra, é fatorada em tantas indicadoras $I_i, I_{j|i}, I_{z|j,i} \ldots$ quantas hierarquias o modelo possuir. Os valores esperados necessários para a especificação das distribuições amostrais são calculados independentemente para cada variável indicadora.

Por exemplo, no caso de 2 níveis, utilizam-se duas variáveis indicadoras, I_i que é igual a 1 se a unidade i do segundo nível for selecionado e $I_{j|i}$ que é igual a 1 se a unidade j do primeiro nível for selecionada, dado que a unidade i do segundo nível foi selecionada. Neste caso, as esperanças necessárias são $E[\pi_{j|i} \mid \zeta_1]$ e $E[\pi_i \mid \zeta_2]$, onde $\pi_{j|i}$ é a probabilidade de $I_{j|i}$ ser igual a 1 e π_i é a probabilidade de I_i ser igual a 1. ζ_1 e ζ_2 representam os parâmetros e variáveis das quais dependem as probabilidades de seleção em cada um dois níveis.

A hipótese de efeitos do desenho independentes é suficiente para a utilização do teorema de Bayes em cada nível do modelo, analogamente a (3.6), e em conseqüência, obter as distribuição amostral de cada variável dependente em função exclusivamente da esperança condicional das probabilidades de seleção associadas a seu nível.

Analogamente ao caso da Distribuição Normal, desenvolvido por Pfeffermann et al. (2002), a necessidade de se assumir uma relação entre as probabilidades de seleção e as variáveis dependentes de cada nível é a principal desvantagem deste método.

4.5.3 Em Modelos Lineares Hierárquicos Generalizados

Os Modelos Lineares Generalizados são uma extensão dos Modelos Lineares Clássicos onde os componentes de $\boldsymbol{y}$ são variáveis aleatórias independentes com distribuição normal e variância constante (McCullagh & Nelder, 1989). Os componentes de um Modelo Linear Generalizado são três:

1. O componente aleatório: formado pelos dados observados que são variáveis aleatórias, y, independentes com média μ e variância σ^2. A distribuição de y pertence à família exponencial, i.e:

$$f_p(y \mid \boldsymbol{x}, \boldsymbol{\theta}) = a(\boldsymbol{\theta}) \exp \left[\sum_{k=1}^{K} \theta_k b_k(y) + c(y) \right];$$

2. O componente sistemático: formado pelas covariáveis, $\boldsymbol{x} = (x_1, \ldots, x_p)$, que produzem um preditor linear, η, dado por

$$\eta = \boldsymbol{x}'\boldsymbol{\beta} = \sum_{l=1}^{p} x_l \beta_l;$$

3. A função de ligação: dada por uma função, g, que relaciona o componente aleatório com o componente sistemático tal que $\eta = g(\mu)$, ou seja, a função de ligação descreve a relação entre o preditor linear η e o valor esperado μ dos dados y.

Por exemplo, na distribuição Bernoulli, $f_p(y \mid \theta) = \theta^y(1-\theta)^{1-y}$ onde $0 < \theta < 1$ e $\mu = E_p[y] = \theta$, é comum utilizar a função Logit como função de ligação, assim $\eta = \log[\mu/(1-\mu)]$.

Neste trabalho considera-se os Modelos Lineares Hierárquicos Generalizados como uma classe de modelos estatísticos onde a variável resposta, no primeiro nível, tem uma distribuição que pertence à familia exponencial e servem para modelar dados provenientes de uma população de interesse que tem uma estrutura hierárquica

intrínseca. Um caso particular desta classe de modelos é o chamado na literatura clássica de Modelo de Intercepto Aleatório, que pode ser representado por:

$$\begin{aligned} y_{ij} &\sim FamExp(\theta_{ij}), && j = 1, \ldots, n_i; \\ \eta_{ij} &= g(\theta_{ij}) = \beta_{0i} + \beta_1 x_{1ij} + \ldots + \beta_p x_{pij}; \\ \beta_{0i} &\sim Normal(\boldsymbol{z}'\boldsymbol{\gamma}, \sigma^2), && i = 1, \ldots, n. \end{aligned} \tag{4.6}$$

Então, para o desenho amostral ser informativo no primeiro nível, as probabilidades de seleção das unidades, j, devem estar relacionadas com as variáveis y_{ij}. Para o desenho ser informativo no segundo nível, as probabilidades de seleção das unidades do segundo nível, i, devem estar associadas aos interceptos β_{0i}. O desenho amostral pode ser informativo nos dois níveis ou somente num deles.

A extensão do método da Distribuição Amostral consiste em propor uma equação que represente a relação entre as probabilidades de seleção com as variáveis respostas respectivas de cada nível e a partir delas obter as esperanças condicionais necessárias para a determinação das distribuições amostrais. No caso do Modelo de Intercepto Aleatório, se o desenho for informativo nos dois níveis, as distribuições amostrais de y_{ij} e de β_{0j} devem ser obtidas. Este aspecto da modelagem é ilustrado detalhadamente no Capítulo seguinte.

4.5.4 Exemplos

Para ilustrar o uso do método da Distribuição Amostral em Modelos Lineares Hierárquicos onde a distribuição da variável resposta pertence à família exponencial, apresentam-se alguns exemplos:

- Seja a fdp de y_i, Gama com parâmetro de forma α e média μ_i tal que

$$\begin{aligned} f_p(y_i) &\propto y_i^{\alpha-1} \exp\{-\alpha y_i/\mu_i\}, \text{ e} \\ \log(\mu_i) &= \beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta}. \end{aligned}$$

Seja a esperança das probabilidades de seleção $E_p(\pi_i \mid y_i) \propto y_i$. Então, para valores dados $\boldsymbol{x}_i$ a distribuição amostral de y_i é outra vez Gama com:

$$E_s(y_i \mid \boldsymbol{x}_i) = \exp\Big\{\beta_0 + \log\big[(\alpha+1)/\alpha\big] + \boldsymbol{x}_i'\boldsymbol{\beta}\Big\}$$

onde os parâmetros $\boldsymbol{\beta}$ das fdps populacionais e amostrais são iguais.

- Seja y_i uma variável categórica que toma valores $0, 1, \ldots, K-1$. Seja $\boldsymbol{x}_i$ um conjunto de covariáveis e suponha que $Pr(y_i = k \mid \boldsymbol{x})$ possa ser modelado usando a regressão logística tal que

$$Pr(y_i = k \mid \boldsymbol{x}_i) = \frac{\exp[\alpha_k + \boldsymbol{x}_i'\boldsymbol{\beta}_k]}{\sum_{j=0}^{K-1}\exp[\alpha_j + \boldsymbol{x}_i'\boldsymbol{\beta}_j]},$$

onde $\alpha_0 = 0$ e $\boldsymbol{\beta}_0 = \mathbf{0}$ por unicidade, e seja

$$Pr(I_i = 1 \mid y_i = k, \boldsymbol{x}_i) = E[\pi_i \mid y_i = k, \boldsymbol{x}_i] = P_k, \qquad k = 0, \ldots, K-1.$$

A fdp amostral é então,

$$\begin{aligned} Pr(y_i = k \mid \boldsymbol{x}_i, I_i = 1) &= \frac{P_k \exp[\alpha_k + \boldsymbol{x}_k'\boldsymbol{\beta}_k]}{\sum_{j=0}^{K-1} P_j \exp[\alpha_j + \boldsymbol{x}_i'\boldsymbol{\beta}_j]} \\ &= \frac{\exp[\alpha_k^* + \boldsymbol{x}_i'\boldsymbol{\beta}_k]}{\sum_{j=0}^{K-1}\exp[\alpha_j^* + \boldsymbol{x}_i'\boldsymbol{\beta}_j]} \end{aligned}$$

onde $\alpha_k^* = \big[\log(P_k/P_0) + \alpha_k\big]$, logo $\alpha_0^* = 0$. Portanto, a fdp amostral é também logística com os mesmos coeficientes de inclinação, mas com interceptos diferentes.

- Dado que a Distribuição Normal pertence à familia exponencial, a distribuição amostral de uma variável resposta normal,i.e, $y_i \sim Normal(\theta, \sigma_y^2)$, onde $\theta = \boldsymbol{x}_i'\boldsymbol{\beta}$, também pode ser obtida usando a Proposição 4.5.1:

$$\begin{aligned} f_p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}, \sigma_y^2) &= \frac{1}{\sqrt{2\pi}\sigma_y}\exp\Big[-\frac{1}{2\sigma_y^2}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2\Big] \\ &= \exp\Big[-\frac{1}{2\sigma_y^2}\big[y_i^2 - 2y_{0i}\boldsymbol{x}_i'\boldsymbol{\beta} + (\boldsymbol{x}_i'\boldsymbol{\beta})^2 - \frac{1}{2}\log 2\pi\sigma_y^2\big]\Big] \\ &= a_i(\boldsymbol{x}_i'\boldsymbol{\beta}) \times \exp\Big[\boldsymbol{x}_i'\boldsymbol{\beta} \times b_j(y_i) + c_i(y_i)\Big] \end{aligned}$$

onde:

$$a_i(\boldsymbol{x}_i'\boldsymbol{\beta}) = \exp\Big[-\frac{1}{2}\log 2\pi\sigma_y^2 - \frac{(\boldsymbol{x}_i'\boldsymbol{\beta})^2}{\sigma_y^2}\Big]$$
$$b_i(y_i) = \frac{y_i}{\sigma_y^2}$$
$$c_i(y_i) = -\frac{y_i^2}{2\sigma_y^2}$$

Supondo que as unidades i são selecionadas com amostragem proporcional ao tamanho M_i e que $M_i \mid y_i, \boldsymbol{\alpha}, \sigma_M^2 \sim logN(\alpha_0 + \alpha_1 y_i, \sigma_M^2)$ tem-se:

$$\begin{aligned} E_p[\pi_i \mid y_i, \boldsymbol{x}_i, \boldsymbol{\beta}, \sigma_y^2] &= \exp\Big[\alpha_0 + \alpha_1 y_i + \frac{\sigma_M^2}{2}\Big] \\ &= \exp\Big[\frac{\alpha_0 + \sigma_M^2}{2}\Big]\exp\Big[\alpha_1 y_i\Big] \\ &= \exp\Big[\frac{\alpha_0 + \sigma_M^2}{2}\Big]\exp\Big[\alpha_1\sigma_y^2\frac{y_i}{\sigma_y^2}\Big] \\ &= r \times \exp\Big[d \times b_i(y_i)\Big] \end{aligned}$$

onde:

$$r = \exp\Big[\frac{\alpha_0 + \sigma_M^2}{2}\Big]$$
$$d = \alpha_1\sigma_y^2$$

Logo, pela Proposição 4.5.1, na amostra,

$$y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}, \sigma_y^2 \sim N(\boldsymbol{x}_i'\boldsymbol{\beta} + \alpha_1\sigma_y^2, \sigma_y^2)$$

- No caso y_i seja Poisson com parâmetro θ, a fpd é

$$\begin{aligned} f_p(y_i \mid \theta) &= \frac{\exp[-\theta]\theta^{y_i}}{y_i!} \\ &= \exp[-\theta] \times \exp[y_i\log\theta - \log y_i!] \\ &= a(\theta) \times \exp[b(y_i)\log\theta + c(y_i)] \end{aligned} \tag{4.7}$$

onde

$$\begin{aligned} a(\theta) &= \exp[-\theta] \\ b(y_i) &= y_i \\ c(y_i) &= \log y_i! \end{aligned}$$

Supondo, analogamente ao exemplo anterior, que M_i é uma variável de tamanho que define as probabilidades de seleção das unidades i, sendo que $M_i \mid y_i, \boldsymbol{\alpha}, \sigma_M^2 \sim logN(\alpha_0 + \alpha_1 y_i, \sigma_M^2)$, tem-se

$$E_p[\pi_i \mid y_i, \theta] = r \times \exp\Big[d \times b_i(y_i)\Big]$$

onde:

$$\begin{aligned} r &= \exp\Big[\frac{\alpha_0 + \sigma_M^2}{2}\Big] \\ d &= \alpha_1 \end{aligned}$$

Logo, pela Proposição 4.5.1, na amostra,

$$y_i \mid \theta \sim Poisson(\theta + \alpha_1)$$

Os exemplos acima apresentados, ilustram como a distribuição amostral de variáveis aleatórias, y_i, cuja distribuição populacional pertence à família exponencial é obtida após a especificação das esperanças condicionais $E[\pi_i \mid y_i, .]$. Nos modelos hierárquicos, esse procedimento deve ser feito em cada nível em forma independente. Por exemplo, seja o modelo de superpopulação de dois níveis hierárquicos tal que a variável resposta, no primeiro nível, tem distribuição Poisson como em (4.7) com $\theta = \exp[\beta_{0i} + \beta_1 x_{1ij} + \ldots + \beta_p x_{pij}]$ e β_{0i}, no segundo nível, tem distribuição Normal como em (4.6), i.e.,

$$\begin{aligned} y_{ij} &\sim Poisson(\theta) && j = 1, \ldots, n_i; \\ \log\theta &= \beta_{0i} + \beta_1 x_{1ij} + \ldots + \beta_p x_{pij}; \\ \beta_{0i} &\sim Normal(\boldsymbol{z}'\boldsymbol{\gamma}, \sigma^2), && i = 1, \ldots, n. \end{aligned}$$

Supondo um desenho amostral em dois estágios com seleção aleatória simples das unidades i do segundo nível e com seleção Proporcional ao Tamanho das unidades j do primeiro nível, onde o tamanho está definido por M_{ij} com distribuição $LogN(\alpha_0 + \alpha_1 y_{ij}, \sigma^2_M)$, então o desenho amostral é informativo só no primeiro nível e o modelo a ser ajustado com os dados da amostra é

$$\begin{aligned} y_{ij} &\sim Poisson(\theta + \alpha_1) && j = 1, \ldots, n_i; \\ \log\theta &= \beta_{0i} + \beta_1 x_{1ij} + \ldots + \beta_p x_{pij}; \\ \beta_{0i} &\sim Normal(\boldsymbol{z}'\boldsymbol{\gamma}, \sigma^2), && i = 1, \ldots, n. \end{aligned}$$

Agora, se o desenho é em dois estágios, com probabilidade proporcional ao tamanho em ambos estágios e se

$$\begin{aligned} M_{2i} &\sim LogN(\delta_0 + \delta_1\beta_{0i}, \sigma^2_{M_2}), \text{ e} \\ M_{1ij} &\sim LogN(\alpha_0 + \alpha_1 y_{ij}, \sigma^2_{M_1}). \end{aligned}$$

são os tamanhos utilizados para selecionar unidades no segundo e primeiro nível, respectivamente, então, o desenho amostral é informativo nos dois níveis e o modelo a ser ajustado com os dados da amostra observada é

$$\begin{aligned} y_{ij} &\sim Poisson(\theta + \alpha_1) && j = 1, \ldots, n_i; && (4.8) \\ \log\theta &= \beta_{0i} + \beta_1 x_{1ij} + \ldots + \beta_p x_{pij}; \\ \beta_{0i} &\sim Normal(\boldsymbol{z}'\boldsymbol{\gamma} + \delta_1\sigma^2, \sigma^2), && i = 1, \ldots, n. && (4.9) \end{aligned}$$

Em (4.8) e (4.9) observa-se a presença de mais parâmetros nas distribuições amostrais do que nas distribuições populacionais. Os novos parâmetros (α_1 e δ_1) fazem parte das distribuições das variáveis do desenho, (M_1 e M_2). Este fato deve ser levado em conta no momento do ajuste do modelo para não ter problemas com a identificabilidade. Neste caso em particular, deve-se incluir na verossimilhança, os valores observados dos tamanhos das unidades selecionadas e devem ser modeladas

com as suas respectivas distribuições amostrais. Mesmo quando o modelo fica mais complexo do que o modelo que ignora o desenho, o método da Distribuição Amostral tem a vantagem de trabalhar só com os valores observados das unidades da amostra.

Capítulo 5

SIMULAÇÃO

Neste Capítulo apresenta-se um experimento de simulação utilizando a Distribuição Amostral num caso particular dos Modelos Lineares Hierárquicos Generalizados. No experimento, geram-se dados de escolas e alunos com estrutura hierárquica para testar a relevância de incluir o mecanismo de seleção dos dados nos modelos hierárquicos sob diferentes desenhos amostrais. A simulação realizada nesta dissertação é uma extensão do trabalho para dados normais de Pfeffermann et al. (2002) que baseou-se num conjunto de dados educacionais de alunos e escolas do município do Estado do Rio de Janeiro coletado em 1996 (BEES).

O modelo de superpopulação escolhido foi:

$$y_{ij} \mid \theta_{ij} \sim Bernoulli(\theta_{ij}) \tag{5.1}$$

$$logit\,(\theta_{ij}) = \beta_{0i} + \boldsymbol{x}'_{ij}\boldsymbol{\beta} \tag{5.2}$$

$$\beta_{0i} \mid \boldsymbol{z}_i, \boldsymbol{\gamma}, \sigma^2_\mu \sim N(\boldsymbol{z}'_i\boldsymbol{\gamma}, \sigma^2_\mu), \tag{5.3}$$

onde y_{ij} representa o nível de proficiência do aluno j da escola i. y_{ij} toma valor 1 se o nível for bom ou 0 se o nível for ruim. $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2_\mu)$ é o vetor de parâmetros para os quais temos interesse em realizar inferência. Foram geradas 500 populações, cada uma com 392 escolas [1]. Além das variáveis do modelo de superpopulação foram geradas duas variáveis de desenho, i.e., informações utilizadas na seleção de amostras. De cada população foram selecionadas 4 amostras por meio de 4 desenhos

[1] As populações foram geradas com o pacote R versão 1.4.1, a rotina utilizada encontra-se no Apêndice B.1

amostrais diferentes[2]. Cada amostra foi utilizada para ajustar três modelos diferentes: o primeiro, ignorando o desenho amostral (IG), o segundo, utilizando as distribuições amostrais (SM) e o terceiro, incorporando as variáveis do desenho (DV).

As estimativas foram obtidas por meio do pacote WinBUGS versão 1.3 (Spiegelhalter, Thomas, & Best, 2000) onde, para cada parâmetro do modelo foram geradas duas cadeias de 10 000 valores sendo que as 5 000 primeiras foram descartadas. Todas as distribuições à priori consideradas foram próprias, mas pouco informativas, i.e., com variâncias grandes em relação aos valores médios esperados de cada parâmetro. Especificamente utilizaram-se distribuições de Pareto como prioris para as variâncias e distribuições normais com média zero para os outros parâmetros. Em todos os casos a convergência das cadeias foi verificada com o teste de Gelman-Rubin disponível no pacote WinBUGS.

Nas seções seguintes descrevem-se os passos da geração das 500 populações e os tipos de desenhos utilizados na seleção das amostra. Apresentam-se também as comparações dos resultados obtidos com os modelos ajustados a cada conjunto de amostras.

5.1 Geração dos dados das Populações

5.1.1 Geração do Intercepto da Escola β_{0i}

O intercepto aleatório foi gerado independentemente para cada escola segundo a equação (5.4)

$$\beta_{0i} = \boldsymbol{z}_i'\boldsymbol{\gamma} + \mu_i = \gamma_0 + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \mu_i, \qquad \mu_i \sim N(0, \sigma_\mu^2),\ i = 1, \ldots, N \tag{5.4}$$

com $\boldsymbol{\gamma}' = (\gamma_0, \gamma_1, \gamma_2) = (2,65; -0,28; -0,56)$, $\sigma_\mu^2 = 0.75$ e $N = 392$.

[2] As amostras foram selecionadas com o pacote SAS versão 8.0 (SAS Institute Inc. (1999)), as rotinas utilizadas encontram-se no Apêndice B.2

z_{1i} e z_{2i} foram as variáveis indicadoras de localização da escola utilizadas por Pfeffermann et al. (2002), assim

$$z_{ki} = \begin{cases} 1 & \text{se a escola pertence à região } k, \\ 0 & \text{caso contrário.} \end{cases} \tag{5.5}$$

5.1.2 Geração do Tamanho da Escola M_i

Nesta etapa foi gerado o número de alunos de cada escola, variável que foi utilizada na seleção de amostras de escolas com probabilidade proporcional ao tamanho. Na prática, considerar que o tamanho da escola está relacionado com a variável resposta significa, por exemplo, supor que um aluno de uma escola pequena (com poucos alunos) tem um ensino quase personalizado e portanto, tem proficiência escolar melhor. Entretanto, nas escolas muito grandes os professores não tem tempo para uma atenção personalizada de todos os seus alunos o que pode aumentar as possibilidades de um aluno apresentar uma pior performance escolar.

O tamanho M_i, i.e. o número total de alunos de cada escola, foi gerado segundo a equação (5.6)

$$\log M_i = \alpha_0 + \alpha_1 \beta_{0i} + \varsigma_i \quad ; \qquad \varsigma_i \sim N(0, \sigma_M^2), \tag{5.6}$$

onde $\alpha_0 = 3,99$, $\alpha_1 = 0,52$ e $\sigma_M^2 = 0,18$.

A equação (5.6) implica que

$$\log M_i \mid \beta_{0i}, \boldsymbol{\alpha}, \sigma_M^2 \sim N(\alpha_0 + \alpha_1 \beta_{0i}, \sigma_M^2), \text{ e que} \tag{5.7}$$

$$M_i \mid \beta_{0i}, \boldsymbol{\alpha}, \sigma_M^2 \sim logN(\alpha_0 + \alpha_1 \beta_{0i}, \sigma_M^2), \tag{5.8}$$

5.1.3 Geração da Resposta do Aluno y_{ij}

Antes de gerar a variável resposta foi necessário gerar as covariáveis $\boldsymbol{x}_{ij}$. Todas as covariáveis de alunos $\boldsymbol{x}_{ij} = (x_{1ij}, x_{2ij}, x_{3ij}, x_{4ij})$ são de natureza dicotômica e foram

selecionadas aleatoriamente com reposição das observações originais do BEES. Assim, $x_{1ij} = 1$ se o aluno fosse do sexo feminino, $x_{2ij} = 1$ se tivesse 15 ou 16 anos de idade, $x_{3ij} = 1$ se tivesse 17 ou mais anos e $x_{4ij} = 1$ se pelo menos um dos pais do aluno tivesse educação universitária.

A partir da geração das covariáveis $\boldsymbol{x}_{ij}$, as respostas foram geradas segundo a equação (5.9),

$$\begin{aligned} logit\,(\theta_{ij}) &= \beta_{0j} + \boldsymbol{x}'_{ij}\boldsymbol{\beta}, \\ &= \beta_{0j} + \beta_1 x_{1ij} + \beta_2 x_{1ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} \\ y_{ij} &\sim Bernoulli(\theta_{ij}) \end{aligned} \tag{5.9}$$

com $\beta_1 = -0,66, \beta_2 = -0,95, \beta_3 = -2,10$ e $\beta_4 = -0,43$.

5.1.4 Geração do Estrato do Aluno O_{ij}

Após da geração de resposta y_{ij}, para cada aluno j da escola i foi construída a variável p_{ij} tal que

$$p_{ij} = \eta_0 + \eta_1 y_{ij} + \zeta_{ij}; \qquad \zeta_{ij} \sim N(0, \sigma_p^2) \tag{5.10}$$

com $\eta_0 = 1,67$, $\eta_1 = 0,29$ e $\sigma_p^2 = 0,24^2$. Supôs-se a existência de três (03) estratos (k=1,2,3) onde cada aluno foi alocado segundo o valor de p_{ij} de acordo com a seguinte regra:

$$O_{ij} = \begin{cases} 1 & \text{se} \quad p_{ij} < 1,76, \\ 2 & \text{se} \quad 1,76 \leq p_{ij} < 1,97, \\ 3 & \text{se} \quad p_{ij} \geq 1,97, \end{cases} \tag{5.11}$$

onde $O_{ij} = k$ indica o que aluno j da escola i pertence ao estrato k.

Esta variável foi construída para ser utilizada na seleção de alunos através de uma

amostragem estratificada. Nota-se que, de (5.11):

$$\begin{aligned} Pr(O_{ij} = 1) &= Pr(p_{ij} < 1,76), \\ Pr(O_{ij} = 2) &= Pr(p_{ij} < 1,97) - Pr(p_{ij} \leq 1,76), \\ Pr(O_{ij} = 3) &= Pr(p_{ij} \geq 1,97). \end{aligned}$$

A partir de (5.10) tem-se que, $p_{ij} \sim N(\eta_0 + \eta_1 y_{ij}, \sigma_p^2)$, portanto:

$$\begin{aligned} Pr(O_{ij} = 1) &= \Phi\big(\delta_1 - \delta_2 y_{ij}\big), \\ Pr(O_{ij} = 2) &= \Phi\big(\delta_3 - \delta_2 y_{ij}\big) - \Phi\big(\delta_1 - \delta_2 y_{ij}\big) \\ Pr(O_{ij} = 3) &= 1 - \Phi\big(\delta_3 - \delta_2 y_{ij}\big). \end{aligned} \tag{5.12}$$

onde $\delta_1 = \left(\frac{1,76-\eta_0}{\sigma_p}\right)$, $\delta_2 = \frac{\eta_1}{\sigma_p}$, $\delta_3 = \left(\frac{1,97-\eta_0}{\sigma_p}\right)$.

5.2 Obtenção das Amostras

A seleção de cada amostra foi realizada em duas etapas, na primeira houve uma seleção de 40 escolas e na segunda realizou-se uma seleção de 10 alunos dentro de cada escola selecionada na primeira etapa. Para a obtenção das 4 amostras de cada população foram utilizados os 4 desenhos amostrais diferentes. Esses desenhos foram o resultado da combinação de 2 formas diferentes de seleção de escolas com 2 formas diferentes de seleção de alunos apresentadas na Tabela 5.1. Como foi mencionado no Capítulo anterior, a seleção Aleatória Simples é sempre não informativa. Entretanto, dado (5.6), a seleção com probabilidade proporcional ao tamanho (PPT) é um desenho informativo para escolas, e, no caso dos alunos, de (5.10) nota-se que a amostragem estratificada (EST) é também um desenho amostral informativo. Neste último caso as amostras de alunos estiveram constituídas por 4 alunos do estrato 1, 4 do estrato 2 e 2 de estrato 3. De (5.1) e (5.10) conclui-se que o vetor $\boldsymbol{\phi}$ está formado por $(\boldsymbol{\alpha}, \boldsymbol{\delta})$.

Cada desenho implica probabilidades diferentes de seleção dos elementos (escolas ou alunos) da população. O cálculo destas probabilidades será abordado nas Seções seguintes.

Tabela 5.1: Classificação dos Desenhos Amostrais

	Desenho Não Informativo	Desenho Informativo
Escolas	Aleatória Simples (AAS)	Proporcional ao Tamanho (PPT)
Alunos	Aleatória Simples (AAS)	Estratificada (EST)

Tabela 5.2: Desenhos Amostrais Utilizados

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

5.3 Análise das amostras AAS-EST

Usando amostragem aleatória simples (AAS) de escolas e amostragem estratificada (EST) de alunos dentro da escola i (selecionada no primeiro estágio), a probabilidade do aluno j ser selecionado é:

$$\begin{aligned} Pr(I_{j|i} = 1 \mid y_{ij}, \boldsymbol{\eta}, \sigma_p, \boldsymbol{q}_i) &= \sum_{k=1}^{3} q_k^i Pr\big(O_{ij} = k \mid y_{ij}, \boldsymbol{\eta}, \sigma_p\big) \\ &= q_1^i \Phi\big(\delta_1 - \delta_2 y_{ij}\big) + q_2^i \Big[\Phi\big(\delta_3 - \delta_2 y_{ij}\big) - \Phi\big(\delta_1 - \delta_2 y_{ij}\big)\Big] \\ &\quad + q_3^i \Big[1 - \Phi\big(\delta_3 - \delta_2 y_{ij}\big)\Big] \\ &= (q_1^i - q_2^i)\Phi\big(\delta_1 - \delta_2 y_{ij}\big) + (q_2^i - q_3^i)\Phi\big(\delta_3 - \delta_2 y_{ij}\big) + q_3^i, \end{aligned} \tag{5.13}$$

onde q_k^i é a fração de amostragem do estrato k da escola i.

Observando a expressão (5.13) conclui-se que o desenho amostral é informativo

pois a probabilidade de seleção $\pi_{ij} = Pr(I_{j|i} = 1)$ depende diretamente da variável resposta y_{ij}, portanto, é necessário levar em conta este efeito durante a realização da inferência.

Nesta situação pode-se considerar a inclusão de variáveis indicadoras do estrato como covariáveis do aluno para tornar o desenho ignorável e a inferência seria feita da maneira usual, mas, esta alternativa nem sempre é a mais prática (Pfeffermann et al. (2002)).

Podemos também, seguindo a proposta de Pfeffermann et al. (2002), achar a distribuição amostral do y_{ij} e fazer inferência a partir dela. A distribuição amostral de y_{ij} é Bernoulli de parâmetro [3]:

$$\theta_{ij}^{s} = \frac{1}{1 + \dfrac{(q_1^i - q_2^i)\Phi(\delta_1) + (q_2^i - q_3^i)\Phi(\delta_3) + q_3^i}{\left[(q_1^i - q_2^i)\Phi(\delta_1 - \delta_2) + (q_2^i - q_3^i)\Phi(\delta_3 - \delta_2) + q_3^i\right]\exp\left(\beta_{0i} + \boldsymbol{x}_{ij}'\boldsymbol{\beta}\right)}}. \quad (5.14)$$

Neste caso, o estrato a que pertence o aluno é também uma informação relevante no ajuste do modelo pois serviu para a determinação das probabilidades de seleção, e portanto é recomendável a sua inclusão na inferência. Porém, tem-se somente a informação do estratos dos alunos na amostra, logo, deve-se utilizar a distribuição amostral de O_{ij} que é dada por [4]

$$\begin{aligned}
Pr(O_{ij} = 1) &= \frac{q_1^i}{\sum_{k=1}^{3} q_k^i Pr\left(O_{ij} = k \mid y_{ij}, \boldsymbol{\eta}, \sigma_p\right)} \times \Phi\left(\delta_1 - \delta_2 y_{ij}\right), \\
Pr(O_{ij} = 2) &= \frac{q_2^i}{\sum_{k=1}^{3} q_k^i Pr\left(O_{ij} = k \mid y_{ij}, \boldsymbol{\eta}, \sigma_p\right)} \times \left[\Phi\left(\delta_3 - \delta_2 y_{ij}\right) - \Phi\left(\delta_1 - \delta_2 y_{ij}\right)\right] \\
Pr(O_{ij} = 3) &= \frac{q_3^i}{\sum_{k=1}^{3} q_k^i Pr\left(O_{ij} = k \mid y_{ij}, \boldsymbol{\eta}, \sigma_p\right)} \times \left[1 - \Phi\left(\delta_3 - \delta_2 y_{ij}\right)\right],
\end{aligned}$$

onde $\delta_1 = \left(\frac{1.76 - \eta_0}{\sigma_p}\right)$, $\delta_2 = \frac{\eta_1}{\sigma_p}$, $\delta_3 = \left(\frac{1.97 - \eta_0}{\sigma_p}\right)$.

---

[3] veja-se a demonstração completa no Apêndice A.4

[4] Veja a demonstração completa no Apêndice A.3

A verossimilhança é então:

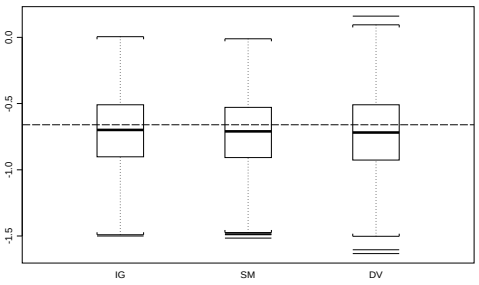
$$
\begin{aligned}
f(\boldsymbol{y}, \boldsymbol{O} \mid \{I_{ij}=1\}, \boldsymbol{x}_{ij}, \boldsymbol{z}_i, \{\beta_{0i}\}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \sigma_p^2, \sigma_\mu^2) = \\
&= \prod_{i=1}^{n} \prod_{j=1}^{m_j} Pr(O_{ij} \mid I_{ij}=1, y_{ij}, \boldsymbol{\eta}, \sigma_p^2) f(y_{ij} \mid I_{ij}=1, \boldsymbol{x}_{ij}, \beta_{0i}, \boldsymbol{\beta}) \\
&= \prod_{i=1}^{n} \prod_{j=1}^{m_j} Pr(O_{ij} \mid I_{j|i}=1, y_{ij}, \boldsymbol{\eta}, \sigma_p^2) f(y_{ij} \mid I_{j|i}=1, \boldsymbol{x}_{ij}, \beta_{0i}, \boldsymbol{\beta}) \\
&= \prod_{i=1}^{n} \prod_{j=1}^{m_j} Pr_s(O_{ij} \mid y_{ij}, \boldsymbol{\eta}, \sigma_p^2) f_s(y_{ij} \mid \boldsymbol{x}_{ij}, \beta_{0i}, \boldsymbol{\beta}) \quad .
\end{aligned}
$$

A distribuição conjunta a partir da qual foram obtidas as distribuições a posteriores condicionais completas é dada por:

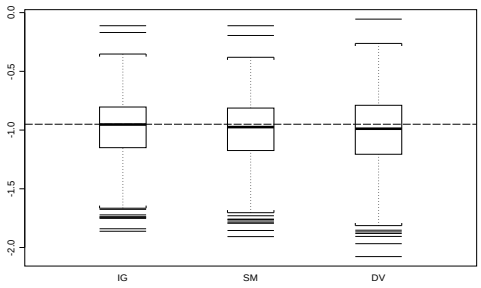
$$
\begin{aligned}
f(\boldsymbol{y}, \boldsymbol{O}, \{\beta_{0i}\}, \boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_\mu^2 \mid I_{ij}=1, \boldsymbol{x}_{ij}, \boldsymbol{z}_i) = \\
&\times \prod_{i=1}^{n} \prod_{j=1}^{m_j} Pr_s(O_{ij} \mid y_{ij}, \boldsymbol{\eta}, \sigma_p^2) f_s(y_{ij} \mid \boldsymbol{x}_{ij}, \beta_{0i}, \boldsymbol{\beta}) \\
&\qquad f_p(\beta_{0i} \mid \boldsymbol{z}_i, \boldsymbol{\gamma}, \sigma_\mu^2) p(\boldsymbol{\eta}) p(\boldsymbol{\beta}) p(\boldsymbol{\gamma}) p(\sigma_\mu^2),
\end{aligned}
$$

onde $p(\boldsymbol{\eta}), p(\boldsymbol{\beta}), p(\boldsymbol{\gamma})$ e $p(\sigma_\mu^2)$ denotam as prioris para $\boldsymbol{\eta}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ e σ_μ^2 respectivamente.

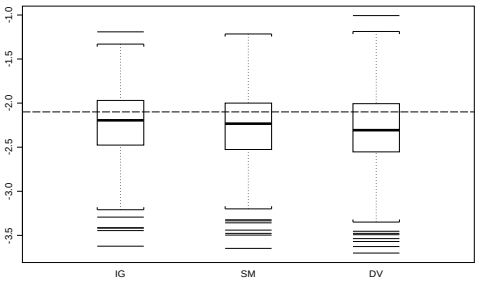
Com o objetivo de comparar a performance do modelo com as distribuições amostrais (SM), para cada amostra foram ajustados também, o modelo que ignora o desenho amostral (IG), i.e., o modelo idêntico ao modelo de superpopulação, e o modelo que inclui as variáveis do desenho (DV) como covariáveis, i.e., foram incluídas duas variáveis indicadoras do estrato a que pertence o aluno. Na Figura 5.1 representam-se as médias a posteriori de cada modelo utilizado (IG, SM e DV), obtidas com as 500 amostras. Observa-se que, exceto para γ_1, as medianas das estimativas com o modelo DV ficam mais afastadas dos valores utilizados na geração dos dados (representados pela linha horizontal) do que as medianas dos outros dois modelos, em particular, observa-se uma péssima performance na estimação de γ_0. Já entre os box-plots dos modelos IG e SM não se observam diferenças muito significativas.



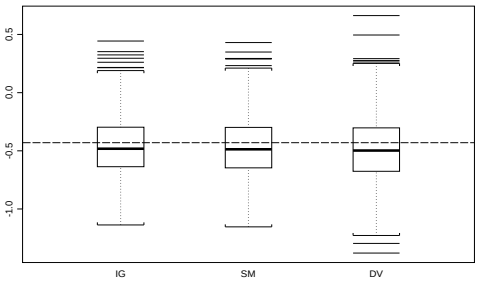
(a) $\beta_1 = -0,66$



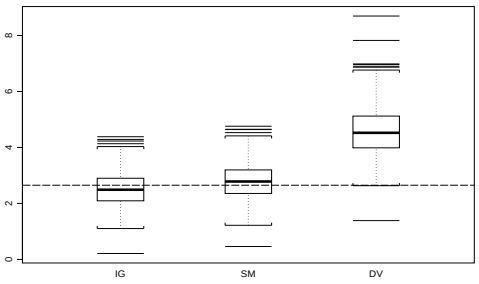
(b) $\beta_2 = -0,95$



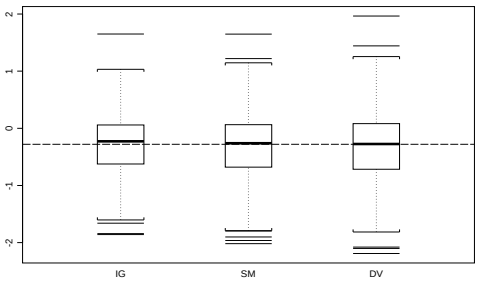
(c) $\beta_3 = -2,10$



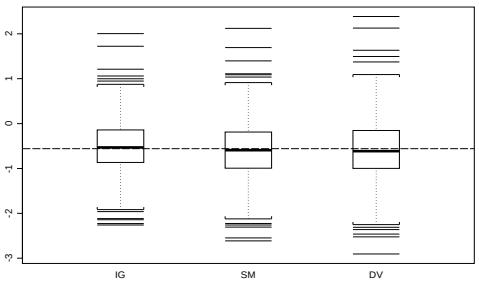
(d) $\beta_4 = -0,43$



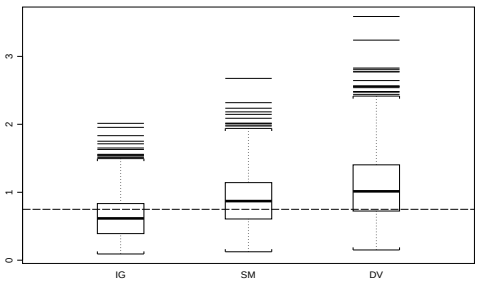
(e) $\gamma_0 = 2,65$



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



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



(h) $\sigma_\mu^2 = 0,75$

Figura 5.1: AAS-EST: Box-Plots das médias a posteriori das 500 amostras

Tabela 5.3: AAS-EST: Média das distribuições a posterioris e Erro Quadrático Médio (EQM)

Parâmetro		Média			EQM		
		IG[1]	SM[2]	DV[3]	IG[1]	SM[2]	DV[3]
β_1	-0,66	-0,71	-0,72	-0,73	0,087	0,091	0,110
β_2	-0,95	-0,98	-1,00	-1,02	0,079	0,084	0,107
β_3	-2,10	-2,23	-2,27	-2,31	0,169	0,189	0,231
β_4	-0,43	-0,46	-0,47	-0,48	0,068	0,072	0,090
γ_0	2,65	2,52	2,80	4,59	0,343	0,393	4,484
γ_1	-0,28	-0,26	-0,30	-0,29	0,273	0,328	0,388
γ_2	-0,56	-0,50	-0,58	-0,59	0,322	0,387	0,458
σ^2_μ	0,75	0,66	0,92	1,21	0,121	0,193	0,410

Nota:[1]Ignorando o desenho, [2] Usando Distribuições Amostrais, [3] Incluindo as Variáveis do Desenho

Na Tabela 5.3 apresentam-se um resumo das médias das distribuições a posteriori de cada um dos modelos ajustados. Observa-se que, en média, as estimativas pontuais [5] dos parâmetros ao nível de alunos ($\boldsymbol{\beta}$) são muito parecidas nos três modelos, mas, o maiores Erros Quadráticos Médios (EQM) [6] correspondem ao modelo DV. Em relação aos parâmetros do segundo nível ($\boldsymbol{\gamma}, \sigma^2_\mu$), as melhores estimativas, em média, correspondem ao modelo IG. O modelo SM tem uma performance pior, com respeito ao EQM, do que o modelo IG e melhor em relação ao modelo DV. Em particular, observa-se que o EQM do modelo DV para γ_0 é elevado.

Comparando as porcentagens de cobertura dos intervalos de 95% de credibilidade,

[5] considerando perda quadrática

[6] O Erro Quadrático Médio é dado por

$$EQM(\beta_j) = \frac{1}{500}\sum_{i=1}^{500}(\beta_j^i - \beta_j)^2$$

apresentados na Tabela 5.4, pode-se concluir que os três métodos têm a mesma performance em relação aos parâmetros do primeiro nível, $\boldsymbol{\beta}$. Para $\boldsymbol{\gamma}$, as coberturas do modelo SM são todas maiores do que as coberturas do modelo IG. Porém, para σ^2_μ, a maior cobertura é a do modelo IG. Observa-se também que o modelo DV tem uma cobertura muito baixa para γ_0 e σ^2_μ.

Tabela 5.4: AAS-EST: Porcentagem de Cobertura dos intervalos de 95% de credibilidade

	Modelo				Modelo		
Parâmetro	IG[1]	SM[2]	DV[3]	Parâmetro	IG[1]	SM[2]	DV[3]
β_1	91,2	91,4	91,8	γ_0	94,8	95,4	31,2
β_2	93,0	93,0	92,0	γ_1	95,0	96,0	95,8
β_3	92,0	91,4	91,6	γ_2	95,0	95,4	95,8
β_4	94,2	94,0	93,4	σ^2_μ	96,0	93,0	88,6

Nota:[1]Ignorando o desenho, [2] Usando Distribuições Amostrais, [3] Incluindo as Variáveis do Desenho

5.4 Análise das amostras PPT-AAS

No caso da amostragem de escolas com Probabilidade Proporcional ao Tamanho (PPT), a probabilidade de selecionar uma escola (π_i) de tamanho M_i numa amostra de tamanho n é dada por:

$$\pi_i = n\frac{M_i}{\sum_{i=1}^{N} M_i} = n\frac{M_i}{M.}, \tag{5.15}$$

onde $M. = \sum_{i=1}^{N} M_i$ é o total de alunos de todas as escolas da população. Após da seleção das escolas, a probabilidade de seleção do aluno j pertencente à escola i é

dada por:

$$\begin{aligned}\pi_{ij} &= \frac{n_i}{M_i} \times \frac{nM_i}{\sum_{i=1}^{392} M_i} \\ &= \frac{10}{M_i} \times \frac{40M_i}{M.} \\ &= \frac{400}{M.}.\end{aligned} \tag{5.16}$$

A expressão (5.16) é a probabilidade final (π_{ij}) de seleção de um aluno para um desenho amostral de duas etapas, onde numa primeira, as escolas são selecionadas com probabilidade proporcional ao tamanho e na segunda etapa, alunos dentro das escolas selecionadas, são selecionados de forma aleatória simples. As probabilidades π_{ij} não dependem diretamente da variável resposta y_{ij} e ainda, são iguais para todos os alunos pois, supondo conhecido o tamanho da população de alunos ($M.$) , esta probabilidade não depende de nenhuma variável. Porém, sabe-se que os alunos de escolas diferentes foram selecionados com probabilidades diferentes pois estas probabilidades dependem do tamanho da escola, e o tamanho da escola está diretamente relacionado com o intercepto do modelo hierárquico, então a relação tamanho-intercepto deve ser levada em conta para a realização da inferência. Este caso é um exemplo da necessidade de fazer à análise do efeito do desenho em cada nível do modelo a ser ajustado. Assim, seguindo (Pfeffermann et al., 1998), tem-se

- Ao nível de alunos: $E[\pi_{j|i} \mid y_{ij}, .] = 10/M_i$ e $E[\pi_{j|i} \mid .] = 10/M_i$, logo $f_s(y_{ij} \mid .) = f_p(y_{ij} \mid .)$, portanto, a distribuição amostral de y_{ij} é a mesma que a distribuição populacional.

- Ao nível de escolas: supondo conhecido o número total de alunos na população (i.e. $N\bar{M}$ conhecido), $E[\pi_i \mid \beta_{0i}, .] = \dfrac{nE[M_i \mid \beta_{0i}, .]}{N\bar{M}}$ e $E[\pi_i \mid .] = \dfrac{nE[M_i \mid .]}{N\bar{M}}$, logo $f_s(\beta_{0i} \mid .) \neq f_p(\beta_{0i} \mid .)$, daqui que é necessário achar a distribuição amostral de β_{0i}.

Seguindo Pfeffermann et al. (2002), a distribuição amostral do β_{0i} é $N(\boldsymbol{z}_i'\boldsymbol{\gamma} + \alpha_1\sigma_\mu^2, \sigma_\mu^2)$[7], diferindo da distribuição populacional somente na média.

Neste caso, o tamanho da escola, M, é também uma informação relevante no ajuste do modelo, pois serve para a determinação das probabilidades de seleção e está relacionado com a distribuição amostral de β_{0i}. Portanto é necessária a sua inclusão no modelo. Como, somente a informação do tamanho das escolas da amostra está disponível, deve-se utilizar a distribuição amostral de M_i: $logN(\alpha_0 + \alpha_1\beta_{0i} + \sigma_M^2, \sigma_M^2)$ [8], que difere da distribuição populacional apenas na média.

A verossimilhança é então:

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{M} \mid \{I_{ij}=1\}, \boldsymbol{x}_{ij}, \boldsymbol{\beta}, \{\beta_{0i}\}, \boldsymbol{z}_i, \boldsymbol{\gamma}, \sigma_\mu^2, \sigma_M^2) = \\ = \prod_{i=1}^{n}\prod_{j=1}^{m_j} f(y_{ij} \mid I_{ij}=1, \boldsymbol{x}_{ij}, \beta_{0i}, \boldsymbol{\beta}) f(M_i \mid I_{ij}=1, \beta_{0i}, \boldsymbol{\alpha}, \sigma_M^2) \\ = \prod_{i=1}^{n}\prod_{j=1}^{m_j} f(y_{ij} \mid \boldsymbol{x}_{ij}, \beta_{0i}, \boldsymbol{\beta}) f(M_i \mid I_i=1, \beta_{0i}, \boldsymbol{\alpha}, \sigma_M^2) \\ = \prod_{i=1}^{n}\prod_{j=1}^{m_j} f_p(y_{ij} \mid \boldsymbol{x}_{ij}, \beta_{0i}, \boldsymbol{\beta}) f_s(M_i \mid \beta_{0i}, \boldsymbol{\alpha}, \sigma_M^2) \quad . \end{aligned}$$

A distribuição conjunta a partir da qual foram obtidas as distribuições a posteriores condicionais completas necessárias para a implementação do método MCMC é dada por:

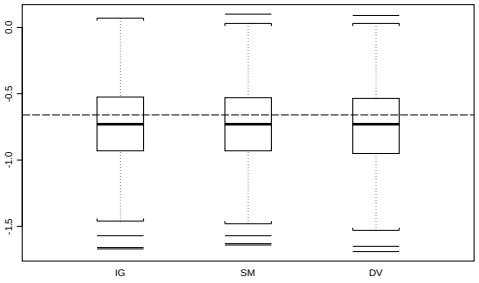
$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{M}, \{\beta_{0i}\}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \sigma_\mu^2, \sigma_M^2 \mid I_{ij}=1, \boldsymbol{x}_{ij}, \boldsymbol{z}_i) = \\ \prod_{i=1}^{n}\prod_{j=1}^{m_j} f_p(y_{ij} \mid \boldsymbol{x}_{ij}, \beta_{0i}, \boldsymbol{\beta}) f_s(M_i \mid \beta_{0i}, \boldsymbol{\alpha}, \sigma_M^2) \\ \times f_s(\beta_{0i} \mid \boldsymbol{z}_i, \boldsymbol{\gamma}, \sigma_\mu^2) p(\boldsymbol{\beta}) p(\boldsymbol{\alpha}) p(\boldsymbol{\gamma}) p(\sigma_M^2) p(\sigma_\mu^2), \end{aligned}$$

onde $p(\boldsymbol{\beta}), p(\boldsymbol{\alpha}), p(\boldsymbol{\gamma})$, $p(\sigma_M^2)$ e $p(\sigma_\mu^2)$ denotam as prioris para $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$, σ_M^2 e σ_μ^2 respectivamente.

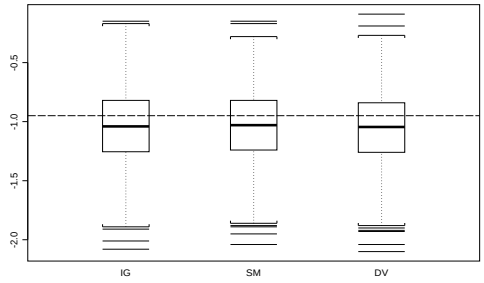
---

[7] Veja a demonstração completa no Apêndice A.2

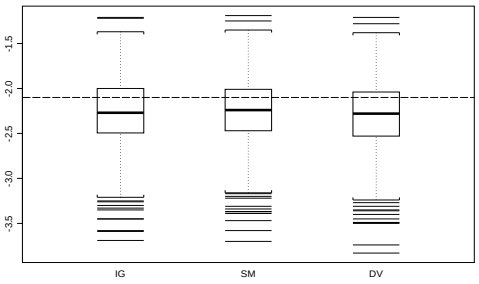
[8] Veja a demonstração completa no Apêndice A.1



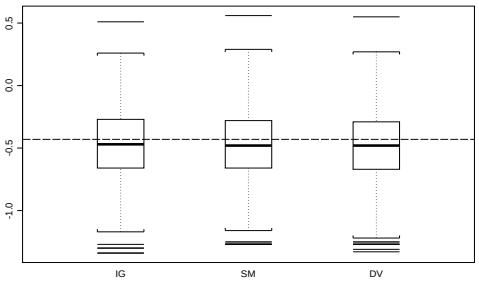
(a) $\beta_1 = -0,66$



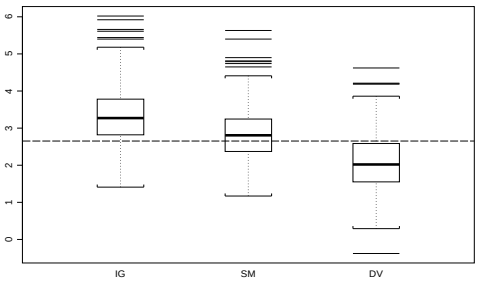
(b) $\beta_2 = -0,95$



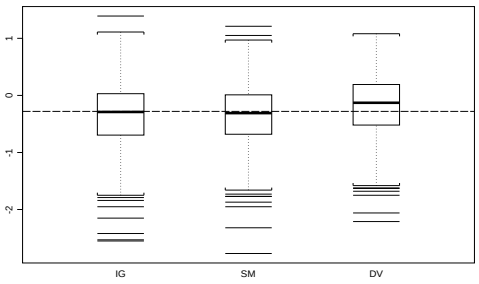
(c) $\beta_3 = -2,10$



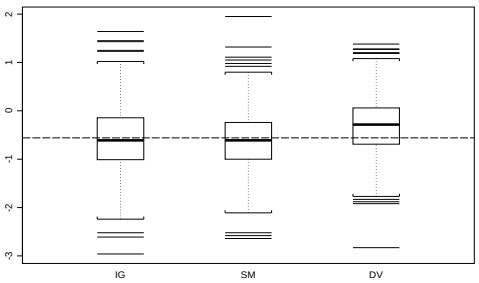
(d) $\beta_4 = -0,43$



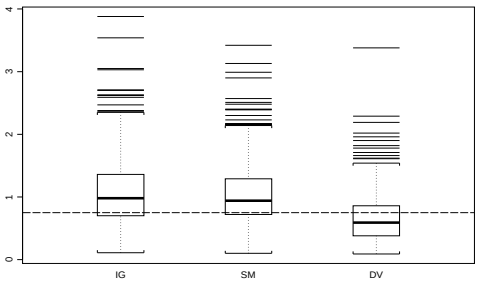
(e) $\gamma_0 = 2,65$



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



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



(h) $\sigma_\mu^2 = 0,75$

Figura 5.2: PPT-AAS: Box-Plots das médias a posteriori das 500 amostras

Além do modelo usando as distribuições amostrais (SM) foram ajustados os modelos IG e DV. A Figura 5.2 contem as representações das 500 médias a posteriori obtidas com cada modelo. Como era esperado, os três modelos têm box-plots parecidos para os parâmetros $\boldsymbol{\beta}$, este fato deve-se que ao nível de alunos todos os modelos são idênticos, contudo observa-se também que as medianas do modelo SM ficam mais próximas da linha horizontal (i.e. do valor utilizado na geração da superpopulação). Este resultado é também observado na Tabela 5.5 pois os EQM do modelo SM são os menores.

Em relação aos parâmetros do segundo nível $\boldsymbol{\gamma}$, os box-plots da Figura 5.2 mostram que a estimação com o modelo SM foi a melhor. Observa-se que as três medianas estão muito próximas da linha horizontal o que não acontece com os outros dois modelos. Além disso, segundo os EQM apresentados na Tabela 5.5, as estimativas do modelo SM foram as mais precisas. Já para σ_μ^2 os resultados indicam que o estimador SM é melhor do que o estimador obtida pelo modelo IG, porém o modelo DV exibe o menor EQM, resultado explicado pelo maior número de covariáveis presentes no modelo, i.e., uma parte da variância é atribuída à variável *Tamanho*.

As coberturas dos intervalos de 95% de credibilidade são apresentadas na Tabela 5.6. Observa-se que para γ_0, o modelo SM apresenta uma cobertura de 12 pontos percentuais maior do que a cobertura do modelo IG, para os demais parâmetros, as coberturas são similares. Em relação ao modelo DV, observam-se coberturas menores para todos os parâmetros, exceto para σ_μ, este resultado pode-se dever à presença de uma covariável a mais no modelo.

5.5 Análise das amostras PPT-EST

O desenho amostral PPT-EST, neste experimento, é informativo nos dois níveis. Em cada nível do modelo tem-se:

- $E_p[\pi_{j|i} \mid O_{ij}, y_{ij}, \boldsymbol{\eta}] = q_j^i$ e $E_p[\pi_{j|i} \mid y_{ij}, \boldsymbol{\eta}] = \sum_{k=1}^{3} q_k^i Pr\big(O_{ij} = k \mid y_{ij}, \boldsymbol{\eta}, \sigma_p\big)$,

Tabela 5.5: PPT-AAS: Média das distribuições a posterioris e Erro Quadrático Médio (EQM)

Parâmetro		Média			EQM		
		IG[1]	SM[2]	DV[3]	IG[1]	SM[2]	DV[3]
β_1	-0,66	-0,74	-0,73	-0,74	0,098	0,094	0,098
β_2	-0,95	-1,05	-1,05	-1,06	0,109	0,105	0,110
β_3	-2,10	-2,27	-2,25	-2,30	0,177	0,170	0,194
β_4	-0,43	-0,48	-0,47	-0,48	0,090	0,088	0,091
γ_0	2,65	3,33	2,86	2,09	0,925	0,466	0,868
γ_1	-0,28	-0,34	-0,33	-0,19	0,362	0,294	0,305
γ_2	-0,56	-0,60	-0,61	-0,32	0,453	0,369	0,429
σ^2_μ	0,75	1,09	1,05	0,66	0,408	0,338	0,158

Nota:[1]Ignorando o desenho, [2] Usando Distribuições Amostrais, [3] Incluindo as Variáveis do Desenho

Tabela 5.6: PPT-AAS: Porcentagem de Cobertura dos intervalos de 95% de credibilidade

Parâmetro	Modelo			Parâmetro	Modelo		
	IG[1]	SM[2]	DV[3]		IG[1]	SM[2]	DV[3]
β_1	93,6	93,8	92,4	γ_0	82,6	95,0	85,4
β_2	92,6	92,0	91,8	γ_1	94,8	94,6	95,2
β_3	94,2	93,2	93,0	γ_2	96,0	95,0	92,8
β_4	95,0	95,2	94,6	σ^2_μ	89,0	89,8	96,4

Nota:[1]Ignorando o desenho, [2] Usando Distribuições Amostrais, [3] Incluindo as Variáveis do Desenho

logo $f_s(y_{ij} \mid .) \neq f_p(y_{ij} \mid .)$.

- $E[\pi_i \mid \beta_{0i}, .] = \dfrac{nE[M_i \mid \beta_{0i}, .]}{N\bar{M}}$ e $E[\pi_i \mid .] = \dfrac{nE[M_i \mid .]}{N\bar{M}}$, logo $f_s(\beta_{0i} \mid .) \neq f_p(\beta_{0i} \mid .)$.

Como nos casos anteriores, incluem-se na modelagem, as variáveis: Estrato do aluno, O_{ij}, e Tamanho da escola, M_i, por serem parte do desenho amostral e estarem associadas a y_{ij} e β_{0i}, respectivamente. Usando a proposta de Pfeffermann et al. (1998), a verossimilhança é dada por:

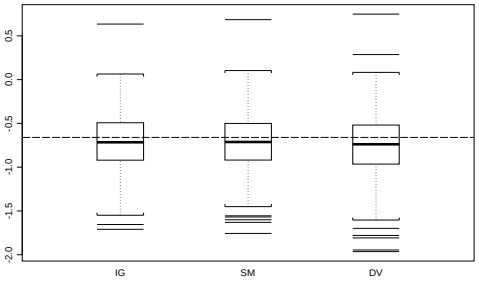
$$\begin{aligned}
f(\boldsymbol{y}, \boldsymbol{O}, \boldsymbol{M} \mid \{I_{ij}=1\}, \boldsymbol{x}_{ij}, \boldsymbol{z}_i, \{\beta_{0i}\}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \sigma_p^2, \sigma_\mu^2, \boldsymbol{\alpha}, \sigma_M^2) = \\
= \prod_{i=1}^{n}\prod_{j=1}^{m_j} Pr(O_{ij} \mid I_{ij}=1, y_{ij}, \boldsymbol{\eta}, \sigma_p^2) f(y_{ij} \mid I_{ij}=1, \boldsymbol{x}_{ij}, \beta_{0i}, \boldsymbol{\beta}) \\
\times f(M_i \mid I_{ij}=1, \beta_{0i}, \boldsymbol{\alpha}, \sigma_M^2) \\
= \prod_{i=1}^{n}\prod_{j=1}^{m_j} Pr(O_{ij} \mid I_{j|i}=1, y_{ij}, \boldsymbol{\eta}, \sigma_p^2) f(y_{ij} \mid I_{j|i}=1, \boldsymbol{x}_{ij}, \beta_{0i}, \boldsymbol{\beta}) \\
\times f(M_i \mid I_i=1, \beta_{0i}, \boldsymbol{\alpha}, \sigma_M^2) \\
= \prod_{i=1}^{n}\prod_{j=1}^{m_j} Pr_s(O_{ij} \mid y_{ij}, \boldsymbol{\eta}, \sigma_p^2) f_s(y_{ij} \mid \boldsymbol{x}_{ij}, \beta_{0i}, \boldsymbol{\beta}) f_s(M_i \mid \beta_{0i}, \boldsymbol{\alpha}, \sigma_M^2).
\end{aligned}$$

Sendo que, a distribuição conjunta a partir da qual foram obtidas as distribuições a posteriores condicionais completas é dada por:

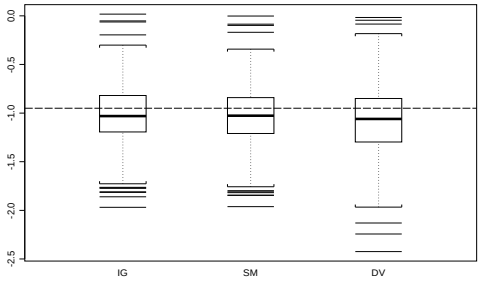
$$\begin{aligned}
f(\boldsymbol{y}, \boldsymbol{O}, \boldsymbol{M}, \{\beta_{0i}\}, \boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \sigma_M^2, \sigma_\mu^2 \mid I_{ij}=1, \boldsymbol{x}_{ij}, \boldsymbol{z}_i) = \\
\prod_{i=1}^{n}\prod_{j=1}^{m_j} Pr_s(O_{ij} \mid y_{ij}, \boldsymbol{\eta}, \sigma_p^2) f_s(y_{ij} \mid \boldsymbol{x}_{ij}, \beta_{0i}, \boldsymbol{\beta}) f_s(M_i \mid \beta_{0i}, \boldsymbol{\alpha}, \sigma_M^2) \\
f_s(\beta_{0i} \mid \boldsymbol{z}_i, \boldsymbol{\gamma}, \sigma_\mu^2) p(\boldsymbol{\eta}) p(\boldsymbol{\beta}) p(\boldsymbol{\alpha}) p(\boldsymbol{\gamma}) p(\sigma_M^2) p(\sigma_\mu^2),
\end{aligned}$$

onde $p(\boldsymbol{\eta}), p(\boldsymbol{\beta}), p(\boldsymbol{\alpha}), p(\boldsymbol{\gamma})$, $p(\sigma_M^2)$ e $p(\sigma_\mu^2)$ denotam as prioris para $\boldsymbol{\eta}$, $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$, σ_M^2 e σ_μ^2 respectivamente.

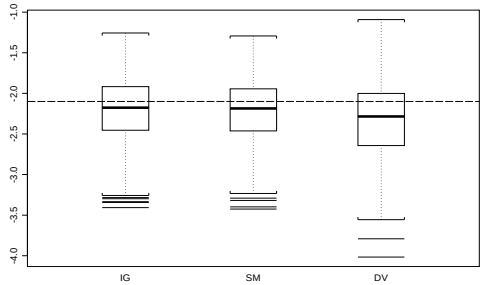
Como nos casos anteriores, além do modelo usando as distribuições amostrais (SM) foram ajustados mais dois modelos: o primeiro é idêntico ao modelo populacional, i.e., ignorando o desenho amostral (IG) e o segundo, incluindo O e M como covariáveis (DV). Na Figura 5.3 representam-se as médias a posteriori obtidas com as 500 amostras, observa-se que os box-plots correspondentes as modelos IG e SM



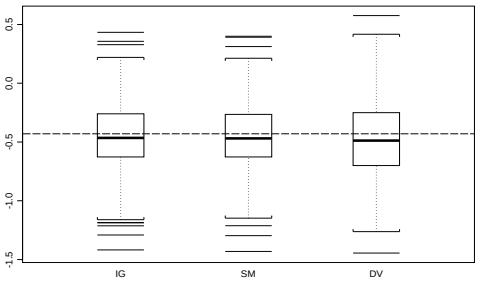
(a) $\beta_1 = -0,66$



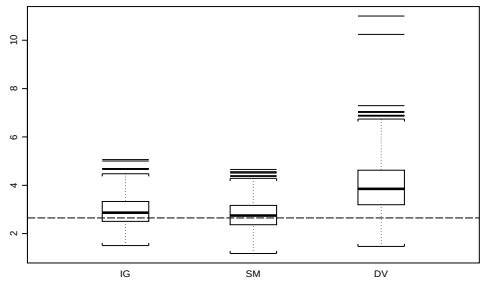
(b) $\beta_2 = -0,95$



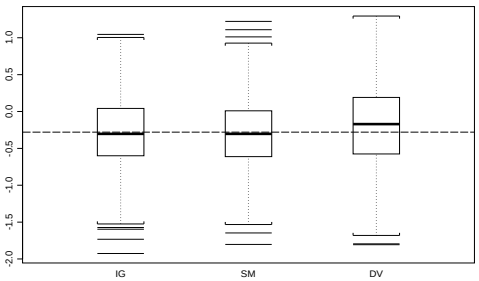
(c) $\beta_3 = -2,10$



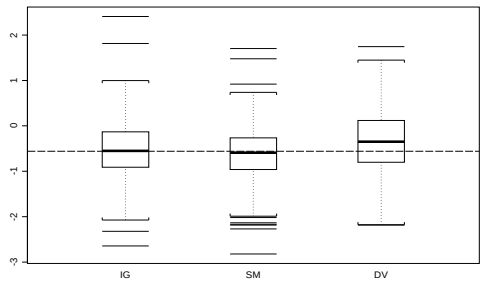
(d) $\beta_4 = -0,43$



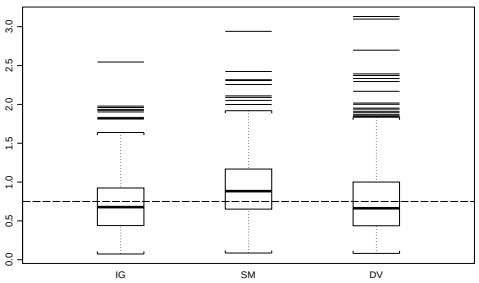
(e) $\gamma_0 = 2,65$



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



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



(h) $\sigma_\mu^2 = 0,75$

Figura 5.3: PPT-EST: Box-Plots das médias a posteriori das 500 amostras

são similares e tem as suas medianas próximas aos valores utilizados na geração do modelo de superpopulação. Já os box-plots do modelo DV, em particular, os de γ_0 e β_3, têm as suas medianas afastadas dos valores reais e um maior número de valores extremos.

Na Tabela 5.7 apresentam-se as médias dos valores esperados das distribuições a posteriori das 500 amostras. Em relação ao nível de alunos, $\boldsymbol{\beta}$, observa-se que, em média, os três modelos forneceram valores similares, porém, o EQM do modelo DV é até 65% maior do que o EQM do modelo SM. Entretanto, as diferenças entre o modelo IG e o modelo SM não são significativas. Em relação ao nível de escolas, $\boldsymbol{\gamma}$,o modelo DV têm as piores médias e EQM e o modelo SM tem menores EQM do que o modelo IG. Já para σ^2_μ o melhor resultado é do modelo IG.

Tabela 5.7: PPT-EST: Média das distribuições a posterioris e Erro Quadrático Médio (EQM)

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

Nota:[1]Ignorando o desenho, [2] Usando Distribuições Amostrais, [3] Incluindo as Variáveis do Desenho

Na Tabela 5.8 observa-se que a porcentagem de cobertura dos intervalos de 95% de credibilidade dos modelos IG e SM é maior do que 90% para todos os parâmetros. Sendo que, para $\boldsymbol{\beta}$ as coberturas do modelo IG foram melhores com respeito ao modelo

SM, enquanto para $\boldsymbol{\gamma}$ e σ_μ as coberturas do modelo SM foram melhores com respeito ao modelo IG. Já o modelo DV têm coberturas menores em comparação aos outros dois modelos, em particular, tem uma cobertura muito baixa (68.0%) para γ_0.

Tabela 5.8: PPT-EST: Porcentagem de Cobertura dos intervalos de 95% de credibilidade

	Modelo				Modelo		
Parâmetro	IG[1]	SM[2]	DV[3]	Parâmetro	IG[1]	SM[2]	DV[3]
β_1	93,4	93,0	91,8	γ_0	93,0	94,2	68,0
β_2	92,4	92,4	90,8	γ_1	96,4	95,6	96,6
β_3	92,8	92,0	89,2	γ_2	94,8	96,0	94,4
β_4	94,4	93,8	93,8	σ^2_μ	94,4	94,4	95,0

Nota:[1]Ignorando o desenho, [2] Usando Distribuições Amostrais, [3] Incluindo as Variáveis do Desenho

5.6 Análise das amostras AAS-AAS

A amostragem aleatória simples é um plano amostral não informativo pois a probabilidade de seleção dos elementos da população não está associada à resposta y_{ij} [9]. Nesta simulação foi utilizado um desenho com amostragem aleatória simples nas duas etapas (AAS-AAS) onde a probabilidade final de selecionar alunos de diferentes escolas é diferente, porém, o desenho foi considerado ignorável para o ajuste do modelo hierárquico. A razão de tal consideração é simples, a avaliação do efeito do desenho amostral é feita em cada nível hierarquico do modelo.

Neste caso, as distribuições amostrais de y_{ij} e β_{0i} são calculadas assim:

- Ao nível de alunos: tem-se $E[\pi_{j|i} \mid y_{ij}, .] = 10/M_i$ e $E[\pi_{j|i} \mid .] = 10/M_i$, logo $f_s(y_{ij} \mid .) = f_p(y_{ij} \mid .)$.

[9] Lembre-se que dita probabilidade depende só do número de elementos na população e do número de elementos na amostra

- Ao nível de escolas: tem-se $E[\pi_i \mid \beta_{0i}, .] = 40/\sum M_i$ e $E[\pi_i \mid .] = 40/\sum M_i$, logo, $f_s(\beta_{0i} \mid .) = f_p(\beta_{0i} \mid .)$

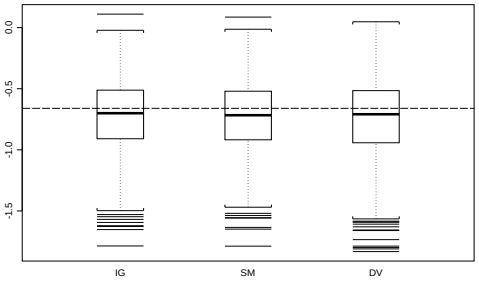
Em conseqüência, o modelo amostral AAS-AAS é idêntico ao modelo da população dado por (5.1), (5.2) e (5.3), e a distribuição conjunta a partir da qual são obtidas as distribuições condicionais completas é dado por:

$$f(\boldsymbol{y}, \{\beta_{0i}\}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2_\mu \mid \boldsymbol{x}_{ij}, \boldsymbol{z}_i,) = \prod_{i=1}^{n} \prod_{j=1}^{m_j} f_p(y_{ij} \mid \boldsymbol{x}_{ij}, \beta_{0i}, \boldsymbol{\beta}) f_p(\beta_{0i} \mid \boldsymbol{z}_i, \boldsymbol{\gamma}, \sigma^2_\mu) p(\boldsymbol{\beta}) p(\boldsymbol{\gamma}) p(\sigma^2_\mu),$$

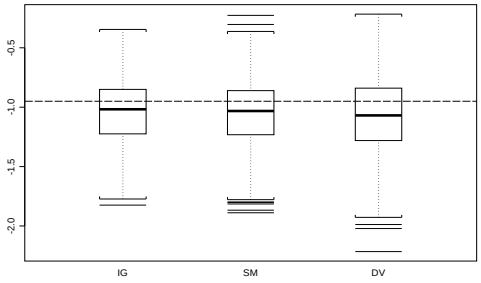
onde $p(\boldsymbol{\beta}), p(\boldsymbol{\gamma})$ e $p(\sigma^2_\mu)$ denotam as prioris para $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ e σ^2_μ respectivamente.

Como no caso dos outros planos amostrais, além do modelo IG, foram ajustados os modelos SM e DV supondo amostragem informativa nos dois níveis. O objetivo deste experimento é observar as conseqüências de supor amostragem informativa quando de fato não é. As médias a posteriori das 500 amostras estão representadas na Figura 5.4. Pode-se observar que as medianas do modelo IG estão mais próximas das linhas horizontais e que os box-plots para os parâmetros do segundo nível exibem muitos valores extremos para os três modelos. Em geral, as estimativas com o modelo SM são mais parecidas com o modelo IG do que as estimativas com o modelo DV. Resultado que pode ser confirmado com a Tabela 5.9, onde observa-se que, em média, as estimativas pontuais do modelo IG são as melhores em termos do viés. No primeiro nível, as médias do modelo IG e SM são quase idênticas. Já no segundo nível, o modelo exibe os menores EQM para $\boldsymbol{\gamma}$, porém, a estimativa de σ^2_μ é a pior, tendo um viés absoluto 36% maior do que o viés do modelo IG.

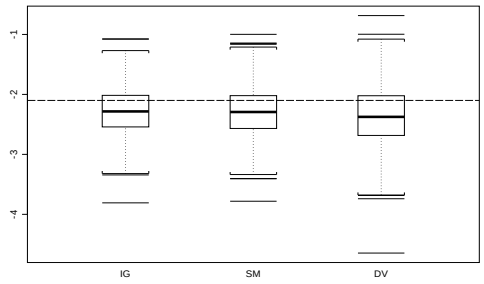
Na Tabela 5.10 observa-se que as coberturas dos intervalos de credibilidade confirmam a melhor performance do modelo IG, e que a principal conseqüência do ajuste do modelo SM é a perda de eficiência na estimação de σ^2_μ.



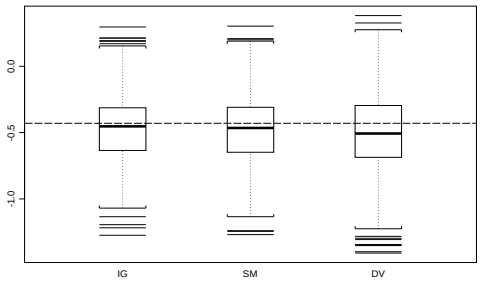
(a) $\beta_1 = -0,66$



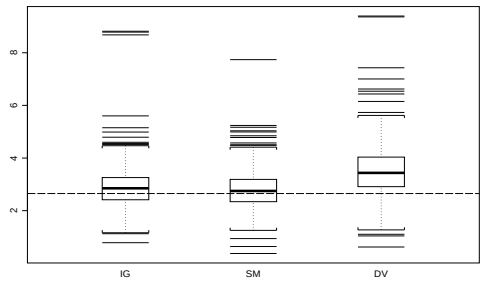
(b) $\beta_2 = -0,95$



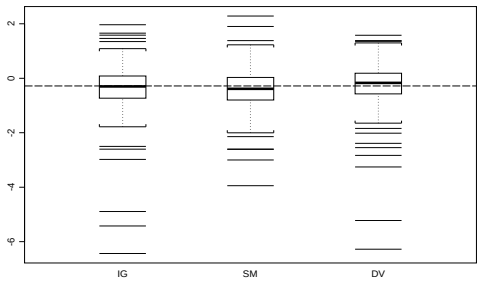
(c) $\beta_3 = -2,10$



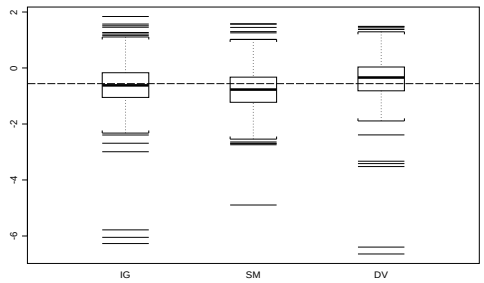
(d) $\beta_4 = -0,43$



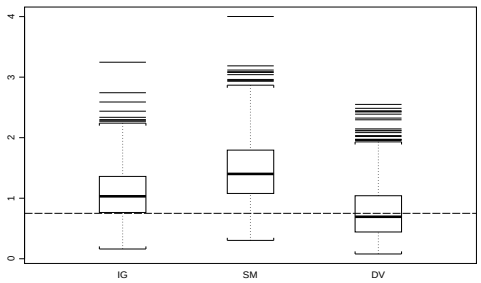
(e) $\gamma_0 = 2,65$



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



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



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

Figura 5.4: AAS-AAS: Box-Plots das médias a posteriori das 500 amostras

Tabela 5.9: AAS-AAS: Média das distribuições a posterioris e Erro Quadrático Médio (EQM)

Parâmetro		Média			EQM		
		IG[1]	SM[2]	DV[3]	IG[1]	SM[2]	DV[3]
β_1	-0,66	-0,72	-0,73	-0,74	0,095	0,093	0,132
β_2	-0,95	-1,04	-1,04	-1,07	0,088	0,091	0,127
β_3	-2,10	-2,29	-2,30	-2,37	0,203	0,214	0,321
β_4	-0,43	-0,46	-0,48	-0,50	0,068	0,072	0,102
γ_0	2,65	2,89	2,78	3,51	0,710	0,562	1,732
γ_1	-0,28	-0,33	-0,39	-0,21	0,560	0,437	0,510
γ_2	-0,56	-0,63	-0,75	-0,39	0,660	0,569	0,656
σ^2_μ	0,75	1,09	1,48	0,80	0,328	0,860	0,228

Nota:[1]Ignorando o desenho, [2] Usando Distribuições Amostrais, [3] Incluindo as Variáveis do Desenho

Tabela 5.10: AAS-AAS: Porcentagem de Cobertura dos intervalos de 95% de credibilidade

Parâmetro	Modelo			Parâmetro	Modelo		
	IG[1]	SM[2]	DV[3]		IG[1]	SM[2]	DV[3]
β_1	93,2	92,2	91,8	γ_0	94,4	95,4	82,8
β_2	94,2	93,4	92,2	γ_1	95,8	96,0	95,0
β_3	92,0	91,0	89,6	γ_2	95,2	95,0	93,2
β_4	94,4	94,2	95,0	σ^2_μ	89,6	66,2	94,8

Nota:[1]Ignorando o desenho, [2] Usando Distribuições Amostrais, [3] Incluindo as Variáveis do Desenho

5.7 Bondade de Ajuste e Seleção de Modelos

O objetivo desta Seção é apresentar os resultados de um exercício realizado com umas das amostras de cada tipo de plano amostral utilizado no experimento de simulação. A finalidade do exercício foi em primeiro lugar, calcular e comparar a bondade de

ajuste de cada um dos modelos em avaliação: Ignorando o desenho (IG), usando as distribuições amostrais (SM) e incluindo as variáveis de desenho como covariáveis (DV), e em segundo lugar, selecionar um modelo para cada amostra utilizando um critério de seleção convencional.

As medidas de bondade de ajuste ou de poder preditivo do modelo utilizadas neste exercício foram: Sensibilidade, Especifidade, Porcentagem de Acertos (Pac) e Porcentagem de Acertos Individuais (Pacpi). Quanto maior o valor da cada medida, melhor a performance do modelo. Para a seleção de modelos, utilizaram-se dois critérios: o Deviance e o DIC. A definição e forma de cálculo de cada uma dessas medidas é explicada em detalhe no Apêndice C.

5.7.1 Amostra AAS-EST

As médias e os erros padrões das distribuições a posteriori de cada parâmetro encontram-se na Tabela 5.11. Em relação aos parâmetros do 1º nível ($\boldsymbol{\beta}$) observa-se que as estimativas do modelo DV têm os maiores desvios absolutos respeito a média verdadeira e que para β_1, β_2 e β_3 os menores desvios correspondem as médias a posteriori do modelo SM. Este resultado indica que, para esta amostra, a distribuição amostral produz estimativas mais acuradas dos parâmetros referidos. A situação é diferente em relação aos parâmetros do segundo nível ($\boldsymbol{\gamma}, \sigma^2_\mu$) pois todos os modelos forneceram médias a posterioris com altos desvios absolutos, em particular, o modelo DV onde os desvios em relação a média de todos os parâmetros superaram ao 100%. Contudo, as médias a posteriori do modelo SM foram as mais parecidas com as do modelo IG. As estimativas pontuais ruins para os parâmetros do 2° nível podem ser atribuídas a um efeito da amostra escolhida pois, como se observa na Tabela 5.3, no experimento de simulação, os resultados da estimação destes parâmetros foram satisfatórios.

As distribuições das medidas de sensibilidade e especificidade encontram-se representadas nas Figuras 5.5 e 5.6 respectivamente. Em relação à sensibilidade, embora

Tabela 5.11: AAS-EST: Médias e Erro Padrão a Posteriori

Parâmetro	Na População	Média a Posteriori			Erro Padrão		
		IG[1]	SM[2]	DV[3]	IG[1]	SM[2]	DV[3]
β_1	-0,66	-0,62	-0,64	-0,72	0,28	0,28	0,30
β_2	-0,95	-0,77	-0,82	-0,84	0,27	0,27	0,30
β_3	-2,10	-2,15	-2,21	-2,45	0,37	0,37	0,43
β_4	-0,43	-0,17	-0,17	-0,13	0,26	0,27	0,30
γ_0	2,65	2,73	3,08	5,43	0,60	0,65	0,89
γ_1	-0,28	-0,75	-0,82	-0,98	0,60	0,67	0,71
γ_2	-0,56	-1,30	-1,41	-1,68	0,64	0,70	0,76
σ^2_μ	0,75	1,30	1,87	2,17	0,61	0,82	0,95

Nota:[1]Ignorando o desenho, [2] Usando Distribuições Amostrais, [3] Incluindo as Variáveis do Desenho

em média os três modelos tenham valores similares, a análise detalhada da Figura 5.5 indica que o modelo SM teve uma melhor performance geral pois tem o menor número de simulações com sensibilidade menor do que 0,60 e apresenta valores acima de 0,75. Já o modelo IG apresenta o maior número de simulações com sensibilidade baixa (menos de 0,60). No caso da especificidade, não é claro qual é o modelo com melhor performance, dado que os três modelos têm distribuições quase simétricas e as médias são muito parecidas.

No caso da Porcentagem de acertos, na Figura 5.7 observa-se que, em média, os três modelos tiveram a mesma performance, porém o modelo SM apresenta o maior número de simulações com Pac alto (acima de 0,62).

Na Tabela 5.12 apresentam-se os valores calculados para as medidas Deviance e DIC, o critério de seleção em ambos casos indica que a modelo SM é preferível ao modelo IG, e que o modelo DV é preferível ao SM. Este resultado é esperado desde que os critérios utilizados avaliam os modelos pelos desvios dos valores replicados em relação aos valores observados. Porém, estes critérios não devem ser utilizados

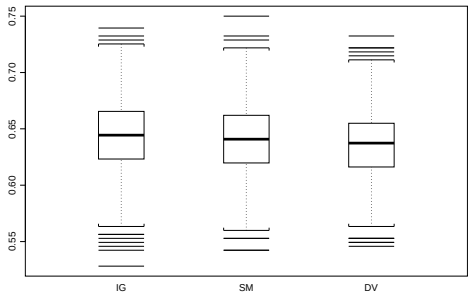


Figura 5.5: Distribuição da medida de sensibilidade da amostra AAS-EST

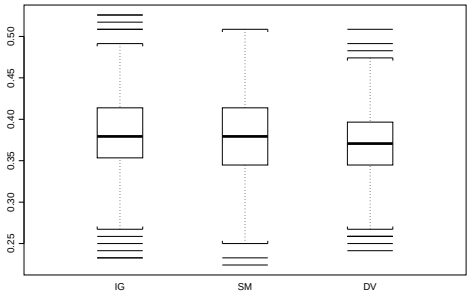


Figura 5.6: Distribuição da medida de especificidade da amostra AAS-EST

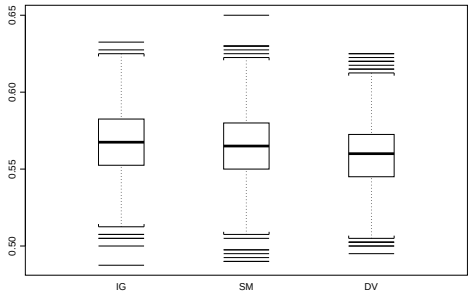


Figura 5.7: Porcentagem de acertos da amostra AAS-EST

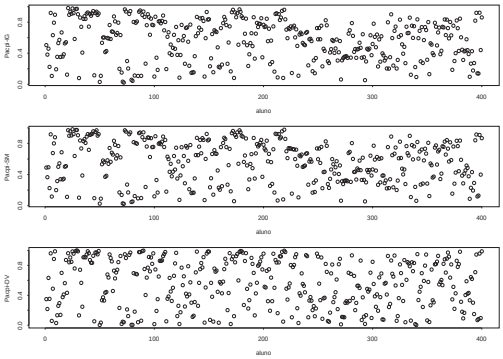


Figura 5.8: Porcentagem de acertos individuais da amostra AAS-EST

isoladamente de outros resultados pois, como neste caso, o modelo DV apresentou as piores estimativas pontuais de todos os parâmetros do modelo de superpopulação.

5.7.2 Amostra PPT-AAS

No caso da amostra PPT-AAS, como se pode observar na Tabela 5.13, o efeito do uso das distribuições amostrais (SM) é ilustrado pelas médias a posteriori obtidas para γ_0, γ_2 e σ^2_μ onde o desvio absoluto em relação a média obtido é até 7 vezes menor que o desvio absoluto obtido com o modelo IG. O resultado mais surpreendente desta amostra corresponde às estimativas pontuais (médias a posteriori) fornecidas pelo

Tabela 5.12: AAS-EST: Deviance e DIC

Modelo	$E[d(y_{rep}, y_{obs})]$	$\bar{D}$	DIC
IG[1]	299,9	416,671	449,431
SM[2]	295,5	416,096	451,560
DV[3]	241,2	346,950	383,725

Nota:[1]Ignorando o desenho, [2] Usando Distribuições Amostrais,[3] Incluindo as Variáveis do Desenho. $\bar{D}$ e DIC referem-se só a contribuição de y

modelo DV, todas elas têm desvios acima do 34% e no caso de γ_1 e γ_2 o sinal não corresponde ao utilizado no modelo de superpopulação. Além dos menores desvios em relação a média, o modelo SM têm as distribuições a posteriori com os menores erros padrões.

Ao nível dos alunos, onde a amostragem foi AAS, as médias e erros padrões a posterioris dos parâmetros $\boldsymbol{\beta}$ são similares para os três modelos. Este fato era esperado, pois a distribuição da variável $\boldsymbol{y}$ foi a mesma em todos os casos, mas, é um exemplo onde o uso das distribuições amostrais no segundo nível não afetou a inferência sobre os parâmetros do primeiro nível.

Das Figuras 5.9 e 5.10 conclui-se que a porcentagem de acertos de valores positivos (iguais a 1) e de valores negativos (iguais a 0) têm uma distribuição similar para os três modelos. Sendo que a menor sensibilidade média e a maior especificidade média correspondem ao SM. A maior porcentagem de acertos médio é do modelo IG, contudo, o modelo DV apresenta alguns valores acima de 0,75. Neste caso, o SM tem Pacs baixos (veja-se a Figura 5.11).

As medidas acima mencionadas sobre o poder preditivo dos modelos nesta amostra indicam que os três tiveram performance similar durante as replicações dos dados. Os resultado apresentados na Tabela 5.14 levam à mesma conclusão. Observa-se que analogamente ao caso da amostra Aas-Est, o menor $E[d(y_{rep}, y_{obs})]$ corresponde ao

Tabela 5.13: PPT-AAS: Médias e Erro Padrão a Posteriori

	Na	Média a Posteriori			Erro Padrão		
Parâmetro	População	IG[1]	SM[2]	DV[3]	IG[1]	SM[2]	DV[3]
β_1	-0,66	-0,72	-0,74	-0,74	0,29	0,29	0,28
β_2	-0,95	-0,74	-0,75	-0,77	0,27	0,27	0,28
β_3	-2,10	-2,25	-2,15	-2,17	0,40	0,37	0,37
β_4	-0,43	-0,55	-0,62	-0,61	0,27	0,26	0,27
γ_0	2,65	2,94	2,87	0,98	0,64	0,59	0,72
γ_1	-0,28	-0,34	-0,70	0,33	0,62	0,52	0,56
γ_2	-0,56	-1,65	-0,86	0,05	0,66	0,55	0,58
σ^2_μ	0,75	0,95	0,87	0,49	0,48	0,45	0,34

Nota:[1]Ignorando o desenho, [2] Usando Distribuições Amostrais, [3] Incluindo as Variáveis do Desenho

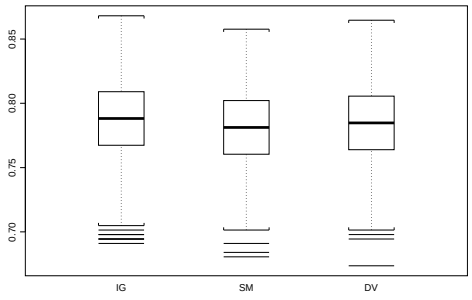


Figura 5.9: Distribuição da medida de sensibilidade da amostra PPT-AAS

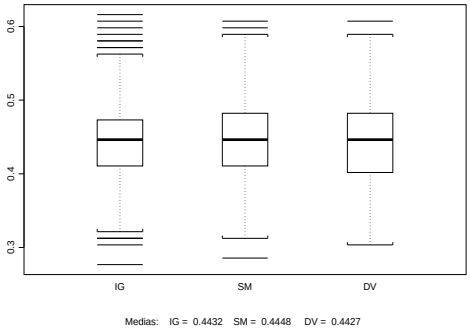


Figura 5.10: Distribuição da medida de especificidade da amostra PPT-AAS

modelo DV, porém este critério não é adequado para selecionar o modelo devido à má performance dele na inferência sobre os parâmetros do modelo de superpopulação.

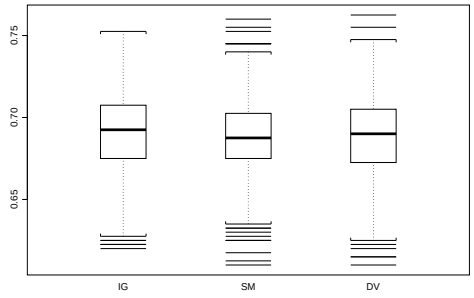


Figura 5.11: Porcentagem de acertos da amostra PPT-AAS

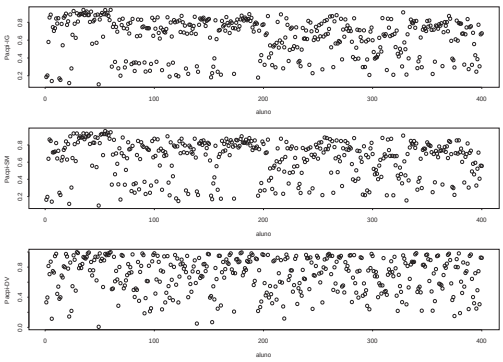


Figura 5.12: Porcentagem de acertos individuais da amostra PPT-AAS

Tabela 5.14: PPT-AAS: Deviance e DIC

Modelo	$E[d(y_{rep}, y_{obs})]$	$\bar{D}$	DIC
IG[1]	273,4	389,965	418,823
SM[2]	273,6	389,765	410,683
DV[3]	272,2	389,175	411,772

Nota:[1]Ignorando o desenho, [2] Usando Distribuições Amostrais,[3] Incluindo as Variáveis do Desenho. $\bar{D}$ e DIC referem-se só a contribuição de y

5.7.3 Amostra PPT-EST

Na Tabela 5.15 encontram-se a média e o erro padrão das distribuições a posteriori de cada parâmetro obtidas no ajuste dos modelos para a amostra PPT-EST. Observa-se que o modelo DV apresenta os maiores desvios absolutos em relação a média verdadeira nos 4 parâmetros do 1º nível (β_1, β_2, β_3, β_4). Entretanto, para β_1 e β_2, o menor desvio corresponde ao modelo SM. Em relação aos parâmetros do segundo nível, tem-se que para γ_0 o menor desvio corresponde ao valor estimado com SM, sendo que para γ_1 e γ_2 todos os modelos apresentam um desvio absoluto superior ao 50%. Em relação ao erro padrão, os modelos IG e SM apresentam valores similares, entretanto, os valores do erro padrão do modelo DV de todos os parâmetros, exceto σ^2_μ,

é maior do que o modelo SM. Os resultados apresentados na Tabela 5.15 indicam, em forma geral, que a pior performance em relação a estimação pontual dos parâmetros e os erros padrão corresponde ao modelo DV, sendo que o modelo SM, neste aspecto, tem resultados melhores ou parecidos com o modelo IG.

Tabela 5.15: PPT-EST: Médias e Erro Padrão a Posteriori

	Na	Média a Posteriori			Erro Padrão		
Parâmetro	População	IG[1]	SM[2]	DV[3]	IG[1]	SM[2]	DV[3]
β_1	-0,66	-0,56	-0,62	-0,62	0,26	0,25	0,28
β_2	-0,95	-0,07	-0,09	-0,02	0,24	0,24	0,28
β_3	-2,10	-2,42	-2,47	-2,69	0,36	0,35	0,40
β_4	-0,43	-0,42	-0,41	-0,34	0,25	0,24	0,26
γ_0	2,65	2,52	2,53	3,63	0,51	0,53	0,80
γ_1	-0,28	-0,79	-0,82	-0,44	0,48	0,49	0,55
γ_2	-0,56	-1,28	-1,15	-0,99	0,49	0,57	0,58
σ^2_μ	0,75	0,33	0,53	0,36	0,23	0,28	0,27

Nota:[1]Ignorando o desenho, [2] Usando Distribuições Amostrais, [3] Incluindo as Variáveis do Desenho

As Figuras 5.13 e 5.14 apresentam a distribuição das medidas de sensibilidade e especificidade respectivamente. Observa-se que as medidas dos modelos IG e SM tem distribuições parecidas, porém o melhor poder preditivo, segundo estas medidas, corresponde ao modelo DV. Em relação ao modelo IG, em média, o SM tem uma sensibilidade superior em 0,62% e uma especificidade superior em 1,04%. Já o modelo DV tem uma sensibilidade superior em 6,06% e uma especificidade superior em 17,38%.

A porcentagem de acertos (Pac) e a porcentagem de acertos individuais (Pacpi) são apresentadas nas Figuras 5.15 e 5.16 respectivamente. Analogamente aos casos de sensibilidade e especificidade, as melhores medidas correspondem ao modelo DV,

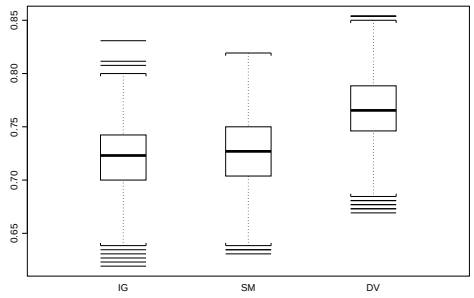


Figura 5.13: Distribuição da medida de sensibilidade da amostra PPT-EST

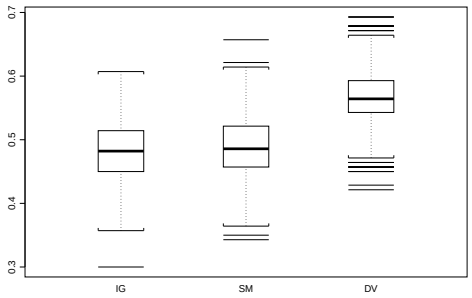


Figura 5.14: Distribuição da medida de especificidade da amostra PPT-EST

cuja Pac é, em média, 9,05% superior à Pac do IG. O Pac do modelo SM é superior ao do modelo IG em 0,74%. Definindo a Porcentagem de "Uns" como medida $T(y, \zeta)$ [10], o p-valor Bayesiano do modelo IG é $0,526$, do modelo SM é $0,542$ e do modelo DV é $0,513$. Em relação ao Pacpi, os três modelo apresentam resultados similares.

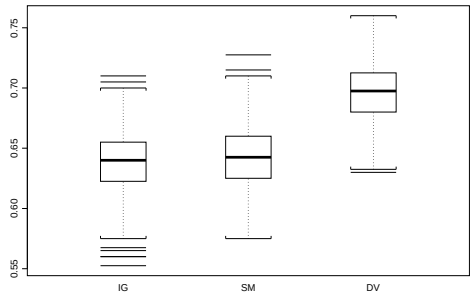


Figura 5.15: Porcentagem de Acertos da amostra PPT-EST

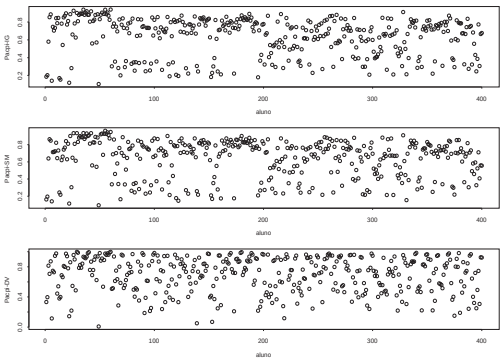


Figura 5.16: Porcentagem de Acertos Individuais da amostra PPT-EST

As medidas de sensibilidade, especificidade e Porcentagem de acertos, em forma conjunta, indicam que o modelo DV apresenta o melhor poder de discriminação entre os três modelos avaliados. Em outras palavras, na amostra utilizada, com o modelo

[10] Veja definição no apêndice C

DV é possível classificar bem as observações positivas ($y = 1$) e ao mesmo tempo, classificar bem as observações negativas ($y = 0$). Esta conclusão é reforçada pela Tabela 5.16 onde observa-se que o menor *Deviance* e o menor *DIC* correspondem ao modelo DV. Observa-se, também, que os valores do modelo SM são menores do que os do modelo IG.

Tabela 5.16: PPT-EST: Deviance e DIC

Modelo	$E[d(y_{rep}, y_{obs})]$	$\bar{D}$	DIC
IG[1]	317,8	442,459	462,668
SM[2]	315,0	440,198	459,691
DV[3]	265,2	380,826	402,929

Nota:[1]Ignorando o desenho, [2] Usando Distribuições Amostrais,[3] Incluindo as Variáveis do Desenho. $\bar{D}$ e DIC referem-se só a contribuição de y

Em conclusão, o modelo DV que contem as variáveis do desenho como covariáveis tem a melhor performance, segundo todas as medidas, em relação ao poder preditivo da variável resposta. Porém, se o interesse principal da pesquisa é a inferência sobre os parâmetros do modelo de superpopulação o modelo SM é o melhor. Além disso, se a inclusão das variáveis do desenho no modelo carece de interesse científico, o melhor modelo é o SM.

5.7.4 Amostra AAS-AAS

O objetivo principal de investigar a amostra AAS-AAS é a avaliação das conseqüências de ajustar o modelo com as distribuições amostrais num conjunto de dados provenientes de uma amostragem não informativa. Na Tabela 5.17 apresentam-se a média e o erro padrão das distribuições a posteriori de cada parâmetro dos modelos na amostra AAS-AAS. Os resultados demonstram que neste caso, o uso do modelo SM não prejudicou a estimação dos parâmetros. Observa-se que os modelos IG e SM

fornecem as mesmas (ou muito parecidas) médias e erros padrões para $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$. A maior diferença entre ambos modelos deve-se a σ_μ^2 onde o modelo SM fornece uma estimativa com um viés duas vezes maior ao obtido com o modelo IG. Já o modelo DV tem os maiores erros padrões, exceto para σ_μ.

Tabela 5.17: AAS-AAS: Médias e Erro Padrão a Posteriori

	Na	Média a Posteriori			Erro Padrão		
Parâmetro	População	IG[1]	SM[2]	DV[3]	IG[1]	SM[2]	DV[3]
β_1	-0,66	-0,95	-1,05	-1,17	0,31	0,32	0,35
β_2	-0,95	-1,78	-1,82	-1,65	0,31	0,28	0,32
β_3	-2,10	-2,43	-2,48	-2,63	0,42	0,42	0,48
β_4	-0,43	-0,00	-0,02	0,20	0,28	0,28	0,31
γ_0	2,65	4,31	4,39	3,96	0,77	0,77	0,94
γ_1	-0,28	-1,16	-1,21	-0,64	0,70	0,70	0,70
γ_2	-0,56	-2,12	-2,50	-1,31	0,77	0,77	0,78
σ_μ^2	0,75	1,02	1,40	0,67	0,54	0,62	0,45

Nota:[1]Ignorando o desenho, [2] Usando Distribuições Amostrais, [3] Incluindo as Variáveis do Desenho

As medidas de sensibilidade e especificidade são apresentadas nas Figuras 5.17 e 5.18 respectivamente. Observa-se claramente que o modelo DV tem as maiores medidas o que significa que este modelo tem melhor poder preditivo que os modelos IG e SM. Observa-se também que as médias e medianas das medidas de sensibilidade e especificidade dos modelos IG e SM são similares, porém, o modelo SM têm maior número de medidas extremamente baixas.

A porcentagem de acertos (Pac) e a porcentagem de acertos individuais (Pacpi) são apresentadas nas Figuras 5.19 e 5.20 respectivamente. Analogamente aos casos de sensibilidade e especificidade, as melhores medidas correspondem ao modelo DV, cuja Pac é, em média, 5,05% superior à Pac do IG. Em relação ao Pacpi, os três

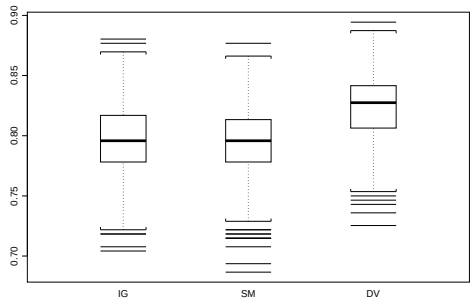


Figura 5.17: Distribuição da medida de sensibilidade da amostra AAS-AAS

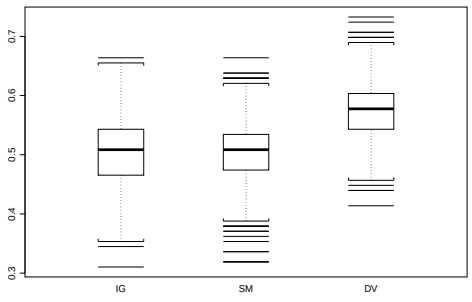


Figura 5.18: Distribuição da medida de especificidade da amostra AAS-AAS

modelos apresentam resultados similares.

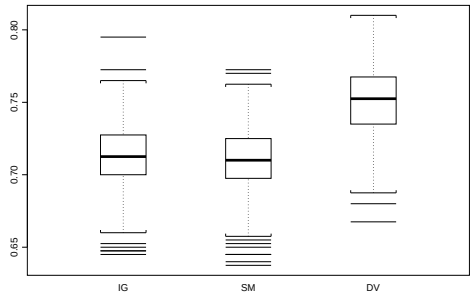


Figura 5.19: Porcentagem de Acertos da amostra AAS-AAS

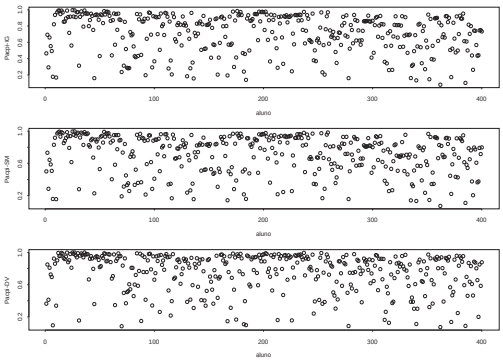


Figura 5.20: Porcentagem de Acertos Individuais da amostra AAS-AAS

Na Tabela 5.18 observa-se que o menor *Deviance* e o menor *DIC* correspondem ao modelo DV e que os valores do modelo SM são similares aos do modelo IG.

5.8 Discussão

O experimento de simulação para verificar coberturas dos intervalos de credibilidade e o exercício empírico para avaliar o poder preditivo dos modelos, permitiram obter às seguintes conclusões, válidas no caso da inferência em modelos hierárquicos logísticos:

Tabela 5.18: AAS-AAS: Deviance e DIC

Modelo	$E[d(y_{rep}, y_{obs})]$	$\bar{D}$	DIC
IG[1]	253,0	362,522	390,916
SM[2]	253,0	362,743	388,088
DV[3]	217,0	322,263	347,414

Nota:[1]Ignorando o desenho, [2] Usando Distribuições Amostrais,[3] Incluindo as Variáveis do Desenho. $\bar{D}$ e DIC referem-se só a contribuição de y

- O uso dos modelos com distribuições amostrais (SM) deu resultados satisfatórios na estimação dos parâmetros do segundo nível, em todos os planos amostrais utilizados, este fato permite afirmar que, ante uma amostragem proporcional ao tamanho (PPT) e o conhecimento de uma distribuição adequada para o *tamanho* a inferência sobre os parâmetros do modelo de superpopulação é mais eficiente com o uso das distribuições amostrais. Além disso, independente do tipo de amostragem utilizado nas unidades do primeiro nível, com o modelo SM, a estimação dos parâmetros do segundo nível continua sendo eficiente, principalmente na cobertura dos intervalos de credibilidade.

- No caso da amostragem estratificada (EST), o uso da distribuição amostral não teve um ganho significativo em eficiência em relação ao modelo que ignora o desenho amostral. Este resultado deve-se a que as frações de amostragem geradas em cada estrato forma muito similares dentro das escolas, este fato que ocasionou que as diferenças $(q_1^i - q_2^i)$ e $(q_3^i - q_2^i)$ presentes nos parâmetros das distribuições amostrais de y_{ij}, ficarem perto de zero. Portanto, nesta simulação, os parâmetros da distribuições amostrais foram muito similares aos parâmetros das distribuições na população.

- A inclusão das variáveis do desenho como covariáveis teve um bom desempenho

segundo os indicadores do poder preditivo, mas a sua utilização não é recomendada no caso que o objetivo principal do modelo seja estimar os valores dos parâmetros, pois os resultados demonstraram que este modelo tem problemas com alguns parâmetros, principalmente interceptos e variâncias.

- O uso das distribuições amostrais (SM) em dados com amostragem não informativa, como no caso da amostragem aleatória simples (AAS) teve como principal conseqüência a sobre-estimação da variância do segundo nível e uma baixa cobertura dos intervalos de credibilidade.

Capítulo 6

APLICAÇÃO

O propósito deste Capítulo é apresentar e comparar alguns dos métodos tratados no Capítulo 4 sobre o ajuste de modelos hierárquicos a dados reais obtidos sob desenhos amostrais complexos. Os dados utilizados correspondem à "Encuesta Nacional de Hogares" (ENAHO-2000.IV) realizada no Peru entre outubro e dezembro do ano 2000 pelo Instituto Nacional de Estadística e Informática (INEI).

Como caso ilustrativo, relaciona-se a situação de pobreza ou não de famílias peruanas com alguns fatores sócio-econômicos e demográficos que as caracterizam.

6.1 ENAHO: Aspectos Principais

6.1.1 Objetivos

Os objetivos gerais da ENAHO-2000.IV são:

- Gerar indicadores que permitam conhecer a evolução da pobreza e das condições de vida das famílias.
- Efetuar diagnósticos sobre as condições de vida e pobreza da população.
- Servir de fonte de informação para pesquisadores.

6.1.2 Desenho amostral

Estrutura do cadastro de domicílios

O Peru está dividido em 24 "departamentos", cada "departamento" se divide em provincias e cada provincia em distritos. Cada distrito é formado por "centros poblados"(CCPP). Os CCPP com mais de 2000 habitantes formam a área urbana e os CCPP com menos de 2000 habitantes formam a área rural.

No cadastro, cada CCPP urbano está dividido em zonas que têm aproximadamente 50 quarteirões e cada zona está formada por 4 ou 5 setores, ou conglomerados urbanos, de 150 domicílios aproximadamente. Cada CCPP rural com 500 a menos de 2000 habitantes está dividido em zonas e estas zonas em setores ou conglomerados rurais. Os CCPP rurais com menos de 500 habitantes estão agrupados em áreas (AER) com 100 domicílios aproximadamente. Estas áreas também são chamadas de conglomerados rurais.

Unidades de amostragem

O processo de seleção da amostra é em três etapas tanto na área urbana como na área rural. As unidades de amostragem em cada etapa são apresentadas na Tabela 6.1.

Tabela 6.1: Unidades de amostragem da ENAHO 2000.IV

Unidade	Tipo de área	
	Urbana	Rural
Primaria (UP)	CCPP (+ 2000 hab.)	CCPP (500-2000 hab.) ou Grupos de 4 AER
Secundária (US)	Conglomerado	Conglomerado ou 1 AER
Terciária (UT)	Domicílio	Domicílio

Nota: CCPP = "Centro Poblado", AER = Area de Cadastramento Rural

Mecanismo de seleção

Antes da seleção das unidades para a amostra, os CCPP são classificados em três estratos:

1. Grandes Cidades : CCPP com mais de 100 000 habitantes (áreas metropolitanas)

2. Resto Urbano: CCPP com mais de 2000 e menos de 100 000 habitantes (áreas urbanas medianas e pequenas)

3. Rural : CCPP com menos de 2000 habitantes.

No caso das Grandes Cidades, não houve uma seleção de unidades primarias (CCPP) pois todas foram incluídas na amostra. Na segunda etapa, para selecionar conglomerados, foi utilizada a amostragem proporcional ao tamanho (PPT), considerando o número de domicílios particulares como tamanho de cada conglomerado e na terceira etapa utilizou-se a seleção sistemática simples ao acaso.

A seleção no restante do pais foi realizada de maneira similar exceto porque na primeira houve uma seleção de CCPP com PPT (Número total de domicílios). A Tabela 6.2 apresenta um resumo do mecanismo de seleção da amostra total.

Tabela 6.2: Mecanismo de seleção da ENAHO 2000.IV

	Estrato		
Etapa	Grandes Cidades	Urbana	Rural
Primeira	Todos	PPT [a]	PPT
Segunda	PPT	PPT	PPT
Terceira	Sistemático	Sistemático	Sistemático

[a] Probabilidade Proporcional ao Tamanho

Tamanho da amostra

O tamanho da amostra final é 4083 domicílios dos quais 2560 pertenciam a área urbana. Estes domicílios correspondem a 835 conglomerados selecionados.

Tabela 6.3: Tamanho da amostra da ENAHO 2000.IV

		Tipo de área	
	Total	Urbana	Rural
Domicílios	4083	2560	1523
Conglomerados	835	695	140

Probabilidade de Seleção de cada vivenda

A probabilidade de seleção final de cada domicílio da amostra é calculada da seguinte forma

$$p_{hij} = \underbrace{\left[n_h \frac{M_{hi}}{M_h}\right]}_{1^o etapa} \times \underbrace{\left[g_{hi} \frac{M_{hij}}{M_{hi}}\right]}_{2^o etapa} \times \underbrace{\left[\frac{m_{hij}}{M'_{hij}}\right]}_{3^o etapa} \tag{6.1}$$

onde:

- p_{hij}: Probabilidade de seleção dos domicílios na $j-$ésima US dentro da $i-$ésima UP no $h-$ésimo estrato.
- n_h: Número de UP selecionadas no $h-$ésimo estrato.
- M_h: Número total de domicílios no $h-$ésimo estrato.
- M_{hi}: Total de domicílios na $i-$ésima UP selecionada no $h-$ésimo estrato.
- g_{hi}: Número de US selecionadas na $i-$ésima UP do $h-$ésimo estrato.

- M_{hij}: Total de domicílios selecionados na $j-$ésima US dentro da $i-$ésima UP no $h-$ésimo estrato.

- m_{hij}: Número de domicílios selecionados na $j-$ésima US dentro da $i-$ésima UP selecionada no $h-$ésimo estrato.

- M'_{hij}: Total de domicílios na $j-$ésima US selecionada dentro da $i-$ésima UP no $h-$ésimo estrato.

O peso inicial de cada domicílio é o inverso da probabilidade final de seleção, i.e.,

$$\begin{aligned} w_{hij} &= \frac{1}{p_{hij}} \\ &= \frac{M_h \times M'_{hij}}{n_h \times g_{hi} \times M_{hij} \times m_{hij}} \end{aligned} \tag{6.2}$$

onde w_{hij} é o peso inicial para os domicílios selecionados na $j-$ésima US dentro da $i-$ésima UP selecionada no $h-$ésimo estrato.

Os pesos finais w'_{hij} são os pesos iniciais w_{hij} ajustados considerando a magnitude da "não resposta" segundo a equação (6.3).

$$w'_{hij} = w_{hij} \times \frac{m'_{hij}}{m''_{hij}} \tag{6.3}$$

onde:

- m'_{hij}: Total de domicílios selecionados na $j-$ésima US selecionada dentro da i-ésima UP selecionada no $h-$ésimo estrato (i.e, o número de entrevistas realizadas mais o número de não respostas)

- m''_{hij}: Total de domicílios entrevistados na j-ésima US selecionada dentro da $i-$ésima UP no $h-$ésimo estrato .

Os pesquisadores são recomendados pelo INEI a incluir os pesos finais w'_{hij} durante a utilização da base de dados da ENAHO-2000.IV.

6.2 Modelo Probabilístico de Pobreza

O objetivo deste modelo é determinar as variáveis mais associadas à pobreza, i.e., não pretende identificar variáveis causais de pobreza mas sim variáveis com alta correlação. Considera-se a *família* como a unidade econômica relevante e como variável resposta à indicadora que toma valor 1 se a familia for *pobre* e 0 se *não for*. A classificação da familia segundo o seu estado de pobreza foi realizada pelo INEI utilizando o método da Linha de Pobreza. Esta variável é fornecida como parte dos dados.

Por estar num contexto social é formulado um modelo hierárquico de intercepto aleatório (com dois níveis). A variável resposta do modelo y_{ij} é igual a 1 se a família estivesse em estado de pobreza na época da pesquisa e as covariáveis, $\boldsymbol{x}_{ij}$, são algumas características sócio-econômicas e demográficas das famílias:

- Características do domicílio
  - Material do piso (1=Terra, 0=Outro)
  - Serviço de Saneamento (1=Rede Pública, 0=Outro)
- Número de Membros da Família
- Características do Chefe de Família
  - Sexo (1=Mulher, 0=Homem)
  - Idade
  - Anos de estudo

Consideraram-se também duas variáveis relacionadas com os conglomerados, $\boldsymbol{z}_j$:

- Localização geográfica (1=Lima, 0=Outro)
- Tipo de Localização (1=Urbana, 0=Rural)

A formulação matemática do modelo, sem levar em conta o desenho amostral, é a seguinte:

$$\begin{aligned} y_{ij} \mid \theta_{ij}, &\sim Bernoulli(\theta_{ij}) \\ logit\,(\theta_{ij}) &= \beta_{0i} + \boldsymbol{x}'_{ij}\boldsymbol{\beta} \\ \beta_{0i} \mid \boldsymbol{z}_j, \boldsymbol{\gamma}, \sigma^2_\mu &\sim N(\boldsymbol{z}'_j\boldsymbol{\gamma}, \sigma^2_\mu) \end{aligned} \tag{6.4}$$

Como foi descrito na Seção 6.1.2, a amostra da ENAHO-2000.IV é resultado da utilização de um plano amostral complexo. Nas duas primeiras etapas utiliza-se uma amostragem PPT onde o tamanho está definido como o número de conglomerados e o número de domicílios particulares, respectivamente. Dado que nas áreas urbanas existem mais domicílios contíguos do que nas áreas rurais e que a proporção de famílias pobres nas áreas urbanas é sempre menor que nas áreas rurais, avaliou-se a relação do tamanho do conglomerado (número de domicílios) com a probabilidade da familia ser pobre. Em outras palavras, ajustaram-se modelos sob a hipótese de que a amostra ENAHO-2000.IV é informativa e compararam-se os resultados com os obtidos no ajuste do modelo (6.4).

Embora este trabalho não tenha como objetivo a análise da pobreza, senão o de avaliar um método de estimação de modelos hierárquicos, considerou-se importante verificar se o plano amostral estaria trazendo informação para o modelo onde a variável resposta é o estado de pobreza pois é uma das variáveis sociais mais utilizadas para avaliar e formular políticas governamentais cuja a finalidade é o combate a pobreza. A pesquisa em relação a pobreza é incentivada por muitos programas e organismos nacionais e internacionais. Como é estabelecido num dos objetivos de ENAHO-2000.IV, essa base de dados é utilizada freqüentemente por pesquisadores da área social que formulam modelos para determinar variáveis e relações que definam as características da população pobre ou em risco de entrar nesse estado. Contudo, as pesquisas até agora publicadas pelo INEI são baseadas em estimações de regressões lineares simples ou em algumas aplicações de técnicas multivariadas e mesmo quando

as estimativas pontuais, como por exemplo, a Proporção de Pobres do pais, são obtidas utilizando os pesos fornecidos com os dados. Todos os modelos são ajustados ignorando o plano amostral da ENAHO-2000.IV ou, em alguns casos, os pesos são utilizados para repetir o dado observado tantas vezes quanto seu peso indicar, assim, trabalha-se com uma "amostra" do tamanho da população, em conseqüência, alguns dos parâmetros dos modelos alcançam a significancia estatística somente devido ao elevado número de observações.

É importante lembrar também que pesquisas como a ENAHO são desenhadas sob algumas restrições administrativas e de custos e que geralmente são usadas para estudar vários assuntos simultaneamente, por exemplo, o questionário da ENAHO tem mais de 5 seções pois investigaram-se temas como vivenda, educação, saúde, emprego, aceso a programas sociais, etc. A determinação do tamanho da amostra realiza-se em função da precisão desejada para uma ou poucas das variáveis, este fato deixa a possibilidade de que o desenho seja informativo para algumas das tantas variáveis estudadas na mesma pesquisa.

6.2.1 Modelos propostos

O objetivo agora é propor modelos adequado para representar a estrutura hierárquica intrínseca da população que leve em conta a informação do desenho amostral. O primeiro passo foi revisar a disponibilidade de dados. Além das covariáveis enumeradas na Seção anterior, a base de dados contém variáveis indicadoras do estrato e conglomerado que pertence cada família. Contém também o peso associado a cada família e que por meio de algumas operações aritméticas fornece o tamanho real de cada conglomerado presente na amostra. Observa-se que não existem variáveis para identificar os CCPP, portanto os modelos propostos consideram só dois estágios da amostragem coincidentes com os dois níveis do modelo. É claro que o tipo de CCPP está relacionado com a probabilidade de seleção de domicílios (primeira etapa da

amostragem), por esta razão a variável AREA (1=Urbana, 0=Rural) esteve presente em todos os modelos formulados.

Os modelos foram divididos em dois grupos, o primeiro deles tem os conglomerados como unidades do segundo nível e as famílias como unidades do primeiro nível. O segundo grupo de modelos tem os "departamentos" como unidades do segundo nível e as famílias como unidades do primeiro nível. A razão da formulação deste segundo grupo de modelos é que a determinação do tamanho da amostra, como é relatado em Instituto Nacional de Estadística e Informática (2001), é realizada por departamentos. O segundo grupo de modelos ilustra o caso em que os níveis do modelo hierárquico não coincidem com os estágios da amostragem.

Os modelos ajustados foram:

1. Modelo Hierárquico Logístico (MHLOG) de dois níveis com 6 covariáveis de famílias e 2 de conglomerados,

2. MHLOG usando as distribuições amostrais do tamanho e do intercepto ao nível de conglomerados,

3. MHLOG incluindo o tamanho do conglomerado como covariável do 2º nível,

4. MHLOG incluindo o estrato do conglomerado como covariável,

5. MHLOG com 4 covariáveis de famílias (sem Material do piso nem Serviço de Saneamento devido a sua pouca variabilidade num mesmo conglomerado e a sua relação com o Tipo de Localização),

6. MHLOG com 4 covariáveis de famílias e a distribuição amostral do tamanho e do intercepto do conglomerado,

7. MHLOG com 4 covariáveis de famílias com tamanho do conglomerado como covariável

8. MHLOG de dois níveis com 8 covariáveis de famílias e nenhuma de departamento,

9. MHLOG usando as distribuições amostrais do tamanho e do intercepto ao nível de departamentos,

10. MHLOG incluindo o tamanho do departamento como covariável do 2º nível,

Em forma análoga ao experimento de simulação, para a determinação das distribuições amostrais assumiu-se que os tamanhos de conglomerados e "departamentos" seguem a distribuição Lognormal. O modelo usando a distribuição Multinomial para determinar a distribuição amostral do estrato não foi possível de se ajustar devido ao desconhecimento das frações de amostragem de cada estrato.

6.2.2 Comparação de Resultados

Na Tabela 6.4 apresentam-se os resultados dos ajustes dos 9 modelos formulados [1]. Todos os modelos têm as familias como unidades do primeiro nível, enquanto que os modelos I e II têm os conglomerados e o modelo III tem os departamentos, como unidades de segundo nível. Observa-se que o grupo do modelo III tem os maiores valores de D e os menores valores de sensibilidade, especificidade e porcentagem de acertos, o que indica que esses modelos tem menor poder preditivo que os modelos I e II. Este resultado pode ser atribuido ao fato de que os níveis do modelo não correspondem aos estágios da amostragem e o efeito dela não está bem representada.

Ao comparar os modelos I e II, observa-se que as medidas de seleção de modelos e do poder preditivo são melhores para o modelo I. Este resultado indica que a presença das variáveis Tipo de Piso e de Serviço Sanitário não prejudica a performance preditiva do modelo. Entre os modelos I, todos eles fornecem médias e erros padrões

[1] As estimativas da distribuições a posteriori foram obtidas no WinBugs 1.4. As medidas do poder preditivo foram calculadas no pacote R.

Tabela 6.4: Comparação das médias e erros padrões a posteriori para modelos hierárquicos ajustados no WinBUGS (método MCMC)

	Modelo I			Modelo II			Modelo III		
Parâmetros	IG	SM	DV	IG	SM	DV	IG	SM	DV
γ_0	-0,875	-1,060	-1,311	-0,177	-0,366	-0,706	-0,852	-0,919	-0,888
	(0,136)	(0,141)	(0,182)	(0,115)	(0,132)	(0,163)	(0,171)	(0,224)	(0,185)
Area	0,074	0,067	0,356	-0,901	-0,878	-0,532	0,217	0,220	0,213
	(0,166)	(0,153)	(0,176)	(0,139)	(0,141)	(0,157)	(0,120)	(0,118)	(0,121)
Lima	0,407	0,359	0,207	0,165	0,126	-0,072	0,129	0,143	0,124
	(0,179)	(0,167)	(0,183)	(0,164)	(0,173)	(0,178)	(0,218)	(0,221)	(0,226)
σ^2_μ	0,894	0,844	0,854	1,078	1,008	1,007	0,411	0,362	0,377
	(0,152)	(0,138)	(0,150)	(0,162)	(0,164)	(0,157)	(0,152)	(0,144)	(0,133)
Piso	1,123	1,121	1,107	–	–	–	1,028	1,032	1,028
	(0,111)	(0,109)	(0,113)	–	–	–	(0,099)	(0,100)	(0,099)
Saneamento	-1,035	-1,005	-0,980	–	–	–	-0,900	-0,910	-0,906
	(0,144)	(0,137)	(0,143)	–	–	–	(0,123)	(0,122)	(0,125)
Membros	0,455	0,451	0,454	0,432	0,427	0,432	0,408	0,408	0,407
	(0,025)	(0,025)	(0,024)	(0,024)	(0,024)	(0,024)	(0,022)	(0,021)	(0,022)
Sexo	-0,107	-0,107	-0,111	-0,148	-0,150	-0,148	-0,122	-0,124	-0,125
	(0,122)	(0,124)	(0,126)	(0,122)	(0,127)	(0,124)	(0,110)	(0,115)	(0,114)
Idade	-0,031	-0,030	-0,031	-0,038	-0,038	-0,038	-0,029	-0,028	-0,029
	(0,003)	(0,004)	(0,004)	(0,003)	(0,003)	(0,004)	(0,003)	(0,003)	(0,003)
Estudo	-0,171	-0,169	-0,170	-0,220	-0,219	-0,218	-0,170	-0,170	-0,170
	(0,015)	(0,015)	(0,015)	(0,014)	(0,014)	(0,014)	(0,013)	(0,013)	(0,012)
$\bar{D}$	3184,3	3195,6	3185,7	3305,8	3323,3	3311,3	3465,8	3466,5	3466,2
DIC	3449,4	3448,2	3443,9	3610,0	3612,7	3602,2	3495,7	3496,0	3496,2
D	2256,0	2268,0	2255,0	2349,0	2368,0	2351,0	2480,0	2478,0	2479,0
Sensibilidade	0,6423	0,6411	0,6441	0,6283	0,6228	0,6275	0,6090	0,6090	0,6076
Especificidade	0,7752	0,7732	0,7749	0,7650	0,7638	0,6275	0,7526	0,7519	0,7520
% de acertos	0,7237	0,7220	0,7243	0,7121	0,7108	0,7113	0,6970	0,6966	0,6961

Nota: Para os modelos I e II, as unidades do 2º nível são os conglomerados. Para o modelo III, as unidades do 2º nível são os departamentos, IG = Ignorando o desenho, SM = usando distribuições amostrais, DV incluindo os tamanhos das unidades do 2º nível como covariáveis

a posteriori muito similares para todos os parâmetros do primeiro nível (família) dado a que não existem diferenças entre as formulações dos modelos a este nível.

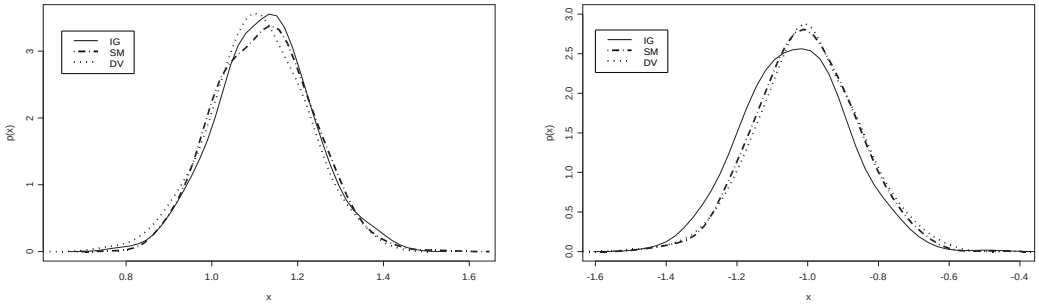
Contudo, observa-se que no caso dos parâmetros do segundo nível, os erros padrões a posteriori do modelo usando a distribuição amostral do tamanho do conglomerado e do intercepto (SM) são os menores, porém, as medidas do poder preditivo indicam que a melhor performance foi dos modelos incluindo as variáveis do desenho como covariáveis.

Na Figura 6.1 apresentam-se as distribuições a posteriori de alguns dos parâmetros dos modelos IG, SM e DV (correspondentes às colunas 2-4 da Tabela 6.4). Observa-se claramente que para os parâmetros associados às características da familia, os três modelos fornecem estimativas parecidas. Já para os parâmetros associados aos conglomerados, as densidades a posteriori estão centralizadas em pontos diferentes, existindo mais proximidade entre as densidades dos modelos IG e SM. Pode-se observar também que os desvios padrões das densidades do modelo DV são um pouco maiores do que os desvios dos outros dois modelos.

De forma complementar, foram ajustados alguns modelos no pacote estatístico MlwiN que permite a incorporação dos pesos amostrais na estimação de modelos hierárquicos. O MlwiN utiliza o método IGLS. Os resultados são apresentados na Tabela 6.5. Observa-se que em relação as estimativas pontuais, não existem diferenças significativas entre os modelos e que, como era esperado, o uso dos pesos tem como conseqüência principal, o aumento dos desvios padrões.

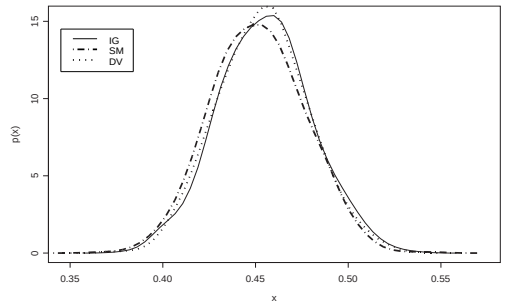
6.3 Discussão

A amostra ENAHO-2000.IV é resultado do uso de uma amostragem complexa, muito comum nas pesquisas sociais. Dado que no último estágio os domicílios são selecionados ao acaso, sistematicamente, a inferência sobre parâmetros ao nível de família e/ou domicílios está livre da influência do plano amostral. Já no caso da inferência a níveis agregados, como conglomerados ou departamentos, é recomendável uma análise sobre a natureza, informativa ou não, do desenho amostral em relação à variável de inte-

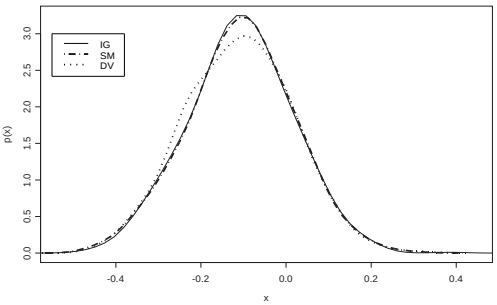


(a) Piso do domicílio

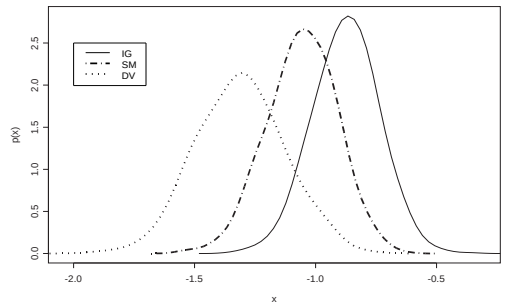
(b) Serviço de Saneamento



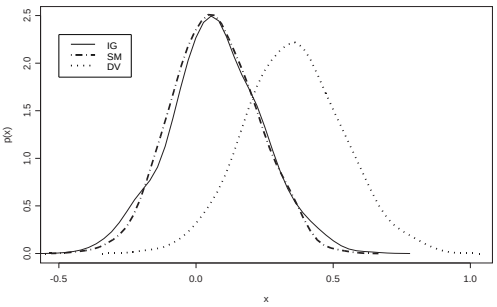
(c) Tamanho da família



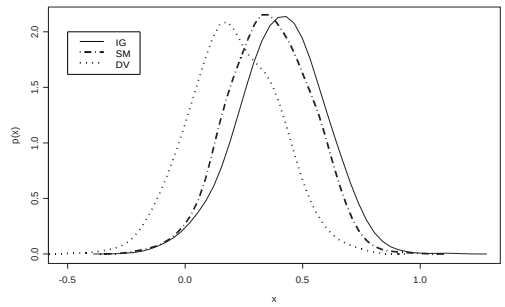
(d) Sexo do Chefe



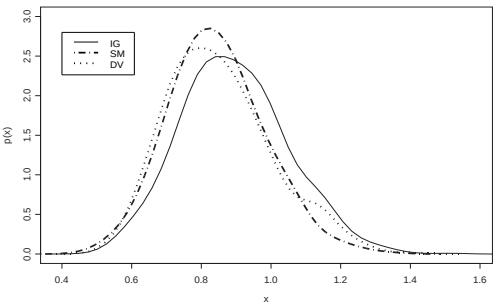
(e) γ_0



(f) Area Urbana-Rural

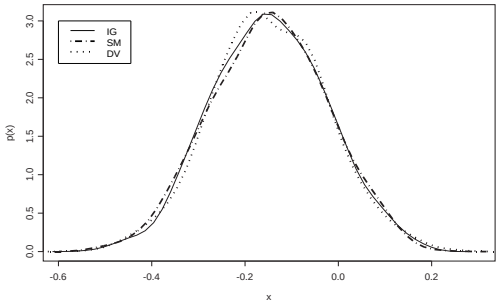


(g) Lima

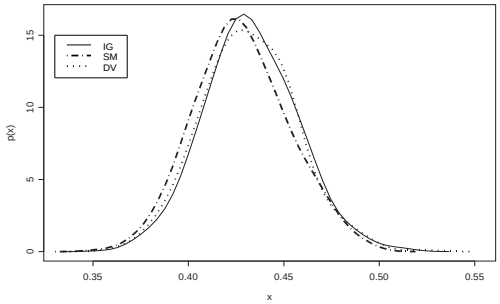


(h) σ_μ^2

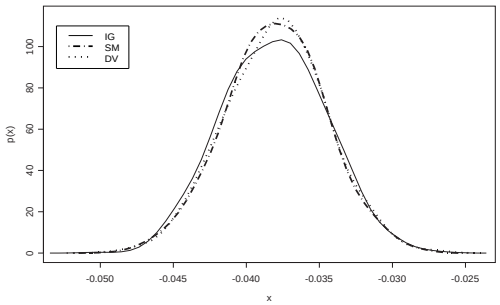
Figura 6.1: Densidades a posteriori dos parâmetros do Modelo I da Tabela 6.4



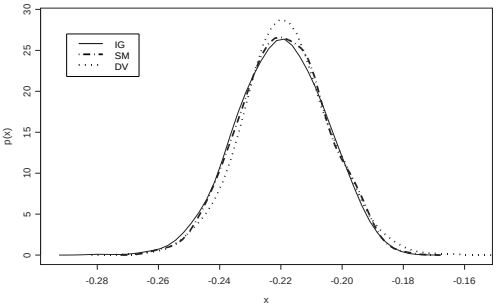
(a) Sexo do Chefe



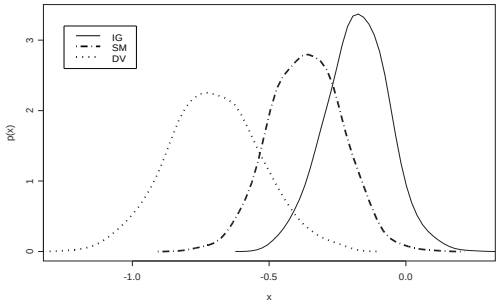
(b) Tamanho da família



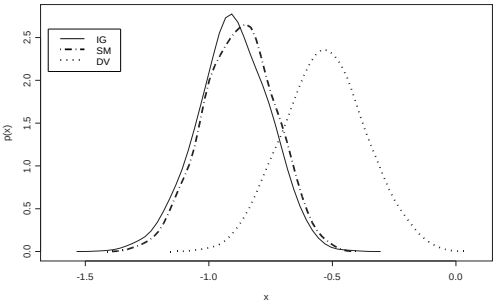
(c) Idade do Chefe



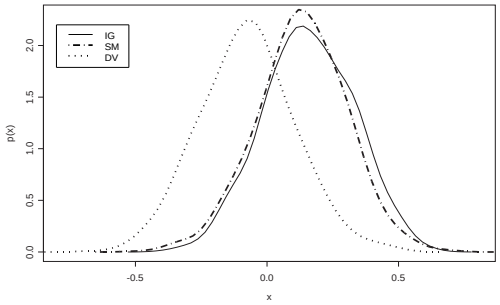
(d) Estudos do Chefe



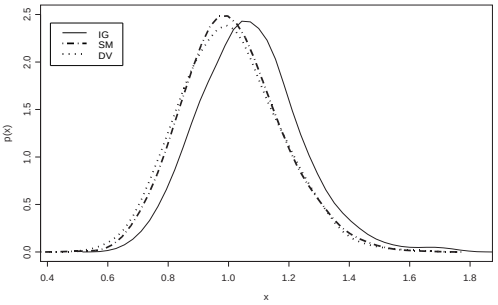
(e) γ_0



(f) Area Urbana-Rural



(g) Lima



(h) σ_μ^2

Figura 6.2: Densidades a posteriori dos parâmetros do Modelo II da Tabela 6.4

resse. No caso da pobreza, o uso do tamanho do conglomerado pode influenciar na presença de famílias pobres na amostra pois é uma variável associada ao tamanho das cidades e conseqüentemente ao desenvolvimento e tipo de oportunidades de emprego e programas sociais disponíveis para as famílias.

Os resultados da Tabela 6.4 indicam os resultados mais satisfatórios para o modelo com todas as covariáveis de família (Modelo I). Indicam também que o uso da distribuição amostral (SM) tem como conseqüência a redução do desvio padrão das densidades a posteriori. Contudo, ao fazer uma interpretação dos parâmetros, consideramos importante destacar que a presença das variáveis: tipo de piso e de serviço de saneamento influencia tanto no sinal quanto na significancia estatística da variável que indica se o conglomerado está localizado na área urbana ou rural. Influencia também na significancia estatística da variável que indica se o conglomerado está localizado em Lima ou não. Interessante é observar que as variáveis mencionadas (tipo de piso, serviço de saneamento, área urbana ou rural e localização em Lima ou não) estão associados pelo mesmo motivo que justificou a suspeita de influência do tamanho do conglomerado na inferência. Esta parece ser a explicação ao fato de se ter obtido quase os mesmos resultados com os modelos IG e SM (ver Figura 6.2).

Em conclusão, tem-se que a combinação de covariáveis presentes no modelo influência no efeito que o plano amostral tem sobre a estimação dos parâmetros. Se as covariáveis levam em conta ou representam o efeito das variáveis do desenho, o modelo IG é aconselhado. No caso da ENAHO-2000.IV, a interpretação dos parâmetros obtidos ajudou na escolha do melhor modelo, pois embora as medidas de seleção de modelos e de poder preditivo indicam que a melhor performance é dos modelos I, a análise dos sinais e a significancia estatística dos parâmetros indicam que os modelos II são melhores. Entre eles, o modelo SM não apresenta melhor performance do que o IG. Se a inclusão das variáveis AREA e LIMA não fosse desejada, então o modelo SM com as 6 covariáveis de família deve ser utilizado.

E importante lembrar que as conclusões anteriores são válidas para os modelos formulados, onde o tamanho foi representado por uma distribuição Lognormal. Existe ainda, a possibilidade de que a relação escolhida entre o intercepto e os tamanhos não seja a mais adequada e portanto o SM não tenha captado o efeito do plano amostral.

Tabela 6.5: Comparação das médias e erros padrões das estimativas para modelos ajustados com o MlwiN (Método IGLS)

	Modelo	MHLOG I		MHLOGII		MHLOGIII
	Linear	s.p.	p.p.	s.p.	p.p.	s.p.
Conglomerado						
β_0	0,029	-0,054	-0,033	-0,085	0,007	-0,301
	(0,239)	(0,254)	(0,325)	(0,260)	(0,333)	(0,264)
σ^2_μ	–	0,534	0,576	0,515	0,571	0,513
	–	(0,093)	(0,132)	(0,091)	(0,127)	(0,092)
Piso	1,018	0,948	0,952	0,987	0,993	0,929
	(0,092)	(0,100)	(0,136)	(0,101)	(0,137)	(0,100)
Saneamento	-0,721	-0,808	-0,948	-0,880	-0,979	-0,715
	(0,096)	(0,109)	(0,168)	(0,126)	(0,194)	(0,112)
Membros	0,386	0,390	0,398	0,390	0,395	0,391
	(0,021)	(0,021)	(0,030)	(0,021)	(0,030)	(0,021)
Sexo	-0,074	-0,088	-0,345	-0,097	-0,349	-0,072
	(0,109)	(0,113)	(0,145)	(0,114)	(0,147)	(0,114)
Idade	-0,028	-0,026	-0,026	-0,027	-0,027	-0,027
	(0,003)	(0,003)	(0,004)	(0,003)	(0,004)	(0,003)
Estudos	-0,158	-0,147	-0,123	-0,152	-0,129	-0,146
	(0,013)	(0,013)	(0,017)	(0,013)	(0,017)	(0,013)
Departamento						
β_0		-0,034	-0,11	-0,118	-0,186	-0,099
		(0,249)	(0,278)	(0,253)	(0,280)	(0,251)
σ^2_μ		0,355	0,417	0,341	0,406	0,342
		(0,078)	(0,087)	(0,076)	(0,089)	(0,076)
Piso		0,934	0,886	0,966	0,916	0,945
		(0,097)	(0,116)	(0,098)	(0,112)	(0,097)
Saneamento		-0,728	-0,753	-0,837	-0,854	-0,740
		(0,105)	(0,148)	(0,121)	(0,161)	(0,105)
Membros		0,380	0,366	0,382	0,367	0,381
		(0,021)	(0,022)	(0,021)	(0,022)	(0,021)
Sexo		-0,095	-0,173	-0,116	-0,188	-0,096
		(0,111)	(0,100)	(0,111)	(0,101)	(0,110)
Idade		-0,027	-0,024	-0,027	-0,024	-0,027
		(0,003)	(0,003)	(0,003)	(0,003)	(0,003)
Estudo		-0,152	-0,132	-0,156	-0,136	-0,153
		(0,013)	(0,018)	(0,013)	(0,018)	(0,013)

Nota: MHLOG I = Modelo Hierárquico Logístico de Intercepto Aleatório, MHLOG II = MHLOG I incluindo AREA e LIMA como covariáveis, MHLOG III = MHLOG I incluindo TAMANHO como covariável, s.p. = sem pesos, p.p. = com pesos padronizados

Capítulo 7

CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho tratou do problema da realização de inferência sobre modelos hierárquicos com dados provenientes de amostras complexas, em particular sob desenhos amostrais informativos. Avaliou-se o uso das distribuições amostrais propostas por Pfeffermann et al. (2002) no caso de variáveis resposta tipo 0-1. Foi realizado um experimento de simulação para verificar a eficiência do método na estimação dos parâmetros do modelo de superpopulação. Compararam-se os resultados do modelo de distribuições amostrais (SM) com o modelo que ignora o desenho amostral (IG). Aplicou-se o método em dados reais provenentes da 'Encuesta Nacional de Hogares" (ENAHO-2000.IV).

Por meio do experimento de simulaçao pode-se concluir que o método da distribuição amostral (SM) mostrou melhor performance na inferência de parâmetros quando a amostragem foi Proporcional ao Tamanho (PPT) do que o modelo que ignora o desenho (IG). Com o modelo SM, os erros quadráticos médios das médias das distribuições a posteriori foram menores e a cobertura do intervalos de credibilidade foram maiores. Quando a amostragem foi Estratificada (EST) os resultados não foram os mesmos. O modelo SM mostrou resultados similares ao modelo IG. Este fato deve ser melhor investigado.

A realização do experimento de simulação e a aplicação do método em dados reais permitiu observar os seguintes aspectos relacionados com o uso da distribuição amostral:

- Identificabilidade: Para a obtenção das distribuições amostrais, é necessário

supor uma relação entre as variáveis do desenho e a variável resposta. Essa relação implica a incorporação de novos parâmetros às distribuições de interesse, os quais não podem ser estimados sem a inclusão de todas as relações durante a estimação. Em conseqüência, as rotinas a serem utilizadas crescem bastante em relação as rotinas do modelo IG, e sem o conhecimento das variáveis do desenho, por exemplo, o tamanho e as frações de amostragem, a estimação de todos os parâmetros das distribuições amostrais não seria possível.

- Especificação das esperanças condicionais: o modelo deve possuir robustez a má especificação das esperanças condicionais pois a distribuição amostral fica completamente determinada após a determinação delas. A má determinação das esperanças condicionais implica a má determinação do modelo completo. Em particular, deve-se realizar um cálculo cuidadoso quando os níveis do modelo não coincidem com os estágios da amostragem.

- Poder Preditivo: ao avaliar o poder preditivo dos modelos através de um exercício empírico, o modelo que inclui as variáveis do desenho como covariáveis (DV) teve a melhor performance, segundo todas as medidas utilizadas. Porém, se o interesse principal da pesquisa é a inferência sobre os parâmetros do modelo de superpopulação ou se a inclusão das variáveis do desenho no modelo carece de interesse científico, o modelo SM deve ser utilizado.

- Tempo computacional: a estimação com o modelo SM foi até 50% mais lenta do que com o modelo IG. Este resultado deve-se ao aumento na complexidade do modelo.

Com a aplicação, em particular, observou-se a importância de determinar se o desenho é informativo ou não, pois uma amostra complexa, como no caso da ENAHO-2000.IV, não necessariamente é informativa. Decidir qual é a relação apropriada entre

as variáveis do desenho não é uma tarefa fácil, em particular, a relação entre tamanhos e interceptos, pois estes últimos não são observáveis.

Trabalhos futuros

- A proposta de usar as distribuições amostrais é interessante do ponto de vista teórico e prático. Porém, a aplicação das distribuições amostrais em modelos hierárquicos até agora só foi realizadas em distribuições Normais e Bernoulli. Uma extensão trivial a outras distribuições, como a Poisson, pode ser realizada.

- A utilização da expansão de Taylor para aproximar as esperanças condicionais deve ser avaliada pois pode ajudar a tornar os modelos robustos a má especificação das mesmas.

- Pesquisas futuras podem trabalhar com a verossimilhança observada completa, como é a proposta Bayesiana por ser a mais natural para representar a relação do plano amostral com a variáveis de interesse.

- É importante ressaltar que a base de dados da ENAHO-2000.IV foi utilizada só no o ajuste de modelos lineares hierárquicos para a Pobreza usando poucas variáveis independentes. A idéia de estudar se o desenho é informativo para outras variáves é bastante pertinente pois a ENAHO-2000.IV é só uma das pesquisas sociais trimestrais realizadas no Perú entre 1995-2001. O estudo ao longo do tempo é de interesse científico e político. Além disso, pesquisas na área social são geralmente financiadas por organismos internacionais e atualmente existem bases de dados similares em vários países da América Latina.

REFERÊNCIAS BIBLIOGRÁFICAS

Bayarri, M., & DeGroot, M. (1992). A "BAD" view of weighted distributions and selection models. Em J. Bernardo, J. Berger, A. Dawid, & A. Smith (Eds.), *Bayesian Statistics. vol. 4* (pp. 17–33).

Binder, D. (1992). Fitting Cox's proportional hazards model from survey data. *Biometrika*, *79*, 139–147.

Binder, D. A., & Roberts, G. R. (2001). Can informative designs be ignorable? *Survey Research Methods Section Newsletter*, 1–3.

Corrêa, S. (2001). *Modelos lineares hierárquicos em pesquisas por amostragem - relacionando o Índice de massa corporal às variáveis da pesquisa sobre padrões de vida.* Dissertação de mestrado, IBGE - ENCE, RJ, Brasil.

Da Costa, L. (2000). *Uso de modelos hierárquicos para o mapeamento da desnutrição infantil no Brasil.* Dissertação de mestrado, IM - UFRJ, RJ, Brasil.

Draper, D. (1995). Inference and hierarquical modelling in the social science. *Journal of Educational and Behavioral Statistics*, *20*, 115–147, 233–239.

Duarte, R. (1999). *Ajuste de modelos lineares usando estimadores de regressão para amostras complexas.* Dissertação de mestrado, IME - USP, SP, Brasil.

Fuller, W. (1975). Regression analysis for sample survey. *Sankhyã: The Indian Journal of Statistics, Series C*, *37*, 117–132.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (1995). *Bayesian data analysis.* London: Chapman and Hall.

Godambe, V., & Thompson, M. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review*, *54*(2).

Ibrahim, J., Chen, M., & Lipsitz, S. (2001). Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika*, *88*(2), 551–564.

Instituto Nacional de Estadística e Informática. (2001). *Encuesta Nacional de Hogares 2000 - 4º trimestre* [Banco de Microdados online (Disponível em http://www.inei.gob.pe)]. Lima: INEI.

Kish, L., & Frankel, M. (1974). Inference from complex samples. *Journal of the Royal Statistical Society, Series B*, *36*, 1–37.

Laud, P., & Ibrahim, J. (1995). Predictive model selection. *Journal of the Royal Statistical Society, Serie B*, *57*, 247–262.

Liang, K., & Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 13–22.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models* (2º ed.). London: Chapman and Hall.

Nathan, G., & Holt, D. (1974). Inference from complex samples. *Journal of the Royal Statistical Society, Series B*, *42*(3).

Pfeffermann, D., & Holmes, D. (1985). Robustness considerations in the choice of a method of inference for regression analysis of survey data. *Journal of the Royal Statistical Society, Series A*, *148*, 268–278.

Pfeffermann, D., Krieger, A., & Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statística Sinica*, *8*, 1087–1114.

Pfeffermann, D., & LaVange, L. (1989). Regression models for stratified multi-stage cluster samples. *Analysis of Complex Survey*, 237–260.

Pfeffermann, D., Moura, F., & Silva, P. (2002). Fitting multi-level modelling under informative probability sampling. *Multi-level Modelling Newsletter*, *14*(1), 8–17.

Pfeffermann, D., & Nathan, G. (1979). Analysis of data from complex samples. Em *Proceedings of the 41° session of the ISI. xlvii, livro 3* (pp. 21–42). Viena.

Pfeffermann, D., & Nathan, G. (1981). Regression analysis of data from a cluster sample. *Journal of the American Statistical Assosiation*, *76*(375).

Pfeffermann, D., Skinner, C., Holmes, D., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B*, *60*, 23–40,41–56(discussion).

Qin, J., Leung, D., & Shao, J. (2002). Estimation with survey data under nonignorable nonresponse or informative sampling. *Journal of the American Statistical Assosiation*, *97*(457), 193–200.

Rao, C. (1965). On discrete distributions arising out of methods of ascertainment. Em S. Atkinson, A.C. amd Fienberg (Ed.), *Classical and contagious discrete distributions* (pp. 320–332).

Rasbash, J., Browne, W., Healy, M., Cameron, B., & Charlton, C. (2000). *Mlwin version 1.10.* Cambridge: Multilevel Models project. Institute of Education. (Disponível em http://www.ioe.ac.uk/mlwin/)

Rotnitzky, A., & Jewell, N. (1990). Hypotesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, *77*, 485–497.

Rubin, D. (1985). The use of propensity scores in applied Bayesian inference. Em J. Bernardo, M. DeGroot, D. Lindley, & A. Smith (Eds.), *Bayesian Statistics. vol. 2.* (pp. 463–472).

SAS Institute Inc. (1999). *Sas onlinedoc®, version 8.* Cary, NC: SAS Institute Inc.

Silva, P. (1996). *Utilizing auxiliary information for estimation and analysis in sample surveys.* Tese de doutorado, University of Southampton, Department of Social Statistics, Southampton.

Smith, T. (2001). Biometrika centenary: Sample surveys. *Biometrika*, *88*(1), 167–194.

Spiegelhalter, D., Thomas, A., & Best, N. (2000). *Winbugs version 1.3. user manual.* Cambridge: Medical Research Council Biostatistics Unit. (Disponível em http://www.mrc-bsu.cam.ac.uk/bugs)

Spiegelhalter, D. J., Best, N. G., & Carlin, B. P. (1998). *Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models* (Tech. Rep.). Cambridge, U.K. (Disponível em http://www.mrc-bsu.cam.ac.uk/Publications/preslid.shtml)

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Linde, A. Van der. (2001). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, *64*, 583–639.

Sugden, R. (1985). A Bayesian view of ignorable designs in survey sampling inference. Em J. Bernardo, M. DeGroot, D. Lindley, & A. Smith (Eds.), *Bayesian Statistics. vol. 2* (pp. 751–754).

Sugden, R. A., & Smith, T. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, *74*, 495–506.

Waller, L., Carlin, B., Xia, H., & Gelfand, A. (1997). Hierarquical spatio-temporal mappings of disease rates. *Journal of the American Statistical Association*, *92*(438), 607–617.

Zhang, F., & Mike, C. (2000). *Multilevel linear regression analysis of complex survey data.* Proceedings of the Survey Research Methods Section. (Disponível em http://www.amstat.org/sections/srms/proceedings/papers/2000_029.pdf)

Apêndice A

DISTRIBUIÇÕES AMOSTRAIS

A.1 Distribuição Amostral de M_i

De (5.7) tem-se que $M_i \mid \beta_{0i}, \boldsymbol{\alpha}, \sigma^2_M \sim logN(\alpha_0 + \alpha_1\beta_{0i}, \sigma^2_M)$ então,

$$f_p(M_i \mid \beta_{0i}, \boldsymbol{\alpha}, \sigma^2_M) = \frac{1}{\sqrt{2\pi}\sigma_M M_i} \exp\left[-\frac{1}{2\sigma^2_M}(\log M_i - \alpha_0 - \alpha_1\beta_{0i})^2\right] \tag{A.1}$$

e

$$E_p[M_i \mid \beta_{0i}, \boldsymbol{\alpha}, \sigma^2_M] = \exp\left[\alpha_0 + \alpha_1\beta_{0i} + \frac{\sigma^2_M}{2}\right]. \tag{A.2}$$

Usando a proposta de Pfeffermann et al. (1998), a distribuição amostral de M é dada por

$$f_s(M_i \mid \beta_{0i}, \boldsymbol{\alpha}, \sigma^2_M) = \frac{E_p[\pi_i \mid M_i, \beta_{0i}, \boldsymbol{\alpha}, \sigma^2_M] f_p(M_i \mid \beta_{0i}, \boldsymbol{\alpha}, \sigma^2_M)}{E_p[\pi_i \mid \beta_{0i}, \boldsymbol{\alpha}, \sigma^2_M]}, \tag{A.3}$$

onde

$$\pi_i = \frac{n \times M_i}{\sum_{i=1}^N M_i} = \frac{nM_i}{M.},$$

logo, supondo $M. = \sum_{i=1}^N M_i$ conhecido,

$$E_p[\pi_i \mid M_i, M., \beta_{0i}, \boldsymbol{\alpha}, \sigma^2_M] = \frac{n \times M_i}{\sum_{i=1}^N M_i} \tag{A.4}$$

$$\begin{aligned} E_p[\pi_i \mid M., \beta_{0i}, \boldsymbol{\alpha}, \sigma^2_M] &= \frac{n \times E[M_i \mid \beta_{0i}, \boldsymbol{\alpha}, \sigma^2_M]}{\sum_{i=1}^N M_i} \\ &= \frac{n}{\sum_{i=1}^N M_i} \exp\left[\alpha_0 + \alpha_1\beta_{0i} + \frac{\sigma^2_M}{2}\right] \end{aligned} \tag{A.5}$$

De (A.1), (A.4), (A.5) em (A.3) tem-se

$$f_s(M_i \mid \beta_{0i}, \boldsymbol{\alpha}, \sigma_M^2) = \frac{n\frac{M_i}{\sum_{i=1}^N M_i}\frac{1}{\sqrt{2\pi}\sigma_M M_i}\exp\left[-\frac{1}{2\sigma_M^2}(\log M_i - \alpha_0 - \alpha_1\beta_{0i})^2\right]}{\frac{n}{\sum_{i=1}^N M_i}\exp\left[\alpha_0 + \alpha_1\beta_{0i} + \frac{\sigma_M^2}{2}\right]}$$
$$= \frac{1}{\sqrt{2\pi}\sigma_M M_i}\exp\left[-\frac{1}{2\sigma_M^2}(\log M_i - \alpha_0 - \alpha_1\beta_{0i} - \sigma_M^2)^2\right].$$

Logo, na amostra,

$$M_i \mid \beta_{0i}, \boldsymbol{\alpha}, \sigma_M^2 \sim logN(\alpha_0 + \alpha_1\beta_{0i} + \sigma_M^2, \sigma_M^2). \tag{A.6}$$

A.2 Distribuição Amostral de β_{0i}

De (5.4) tem-se $\beta_{0i} \sim N(\boldsymbol{z}_i'\boldsymbol{\gamma}, \sigma_\mu^2)$, e seguindo Pfeffermann et al. (1998), a distribuição amostral de β_{0i} é dada por

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

Usando(A.5),

$$f_s(\beta_{0i} \mid \boldsymbol{z}_i, \boldsymbol{\gamma}, \sigma_\mu^2) = \frac{\exp\left[\alpha_0 + \alpha_1\beta_{0i} + \sigma_M^2/2\right] f_p(\beta_{0i} \mid \boldsymbol{z}_i, \boldsymbol{\gamma}, \sigma_\mu^2)}{\exp[\alpha_0 + \alpha_1\boldsymbol{z}_i\boldsymbol{\alpha} + (\alpha_1^2\sigma_\mu^2 + \sigma_M^2)/2]}$$
$$= \frac{1}{\sqrt{2\pi}\sigma_\mu}\exp\left[\alpha_1\beta_{0i} + \frac{\sigma_M^2}{2} - \frac{(\beta_{0i} - \boldsymbol{z}_i'\boldsymbol{\gamma})^2}{2\sigma_\mu^2}\right]$$
$$= \frac{1}{\sqrt{2\pi}\sigma_\mu}\exp\left[-\frac{1}{2\sigma_\mu^2}\big(\beta_{0i} - \boldsymbol{z}_i'\boldsymbol{\gamma}\big)^2 + \alpha_1\big(\beta_{0i} - \boldsymbol{z}_i'\boldsymbol{\gamma}\big) - \frac{\alpha_1^2\sigma_\mu^2}{2}\right]$$
$$= \frac{1}{\sqrt{2\pi}\sigma_\mu}\exp\left\{-\frac{1}{2\sigma_\mu^2}\Big[\big(\beta_{0i} - \boldsymbol{z}_i'\boldsymbol{\gamma}\big)^2 - 2\alpha_1\sigma_\mu^2\big(\beta_{0i} - \boldsymbol{z}_i'\boldsymbol{\gamma}\big) + \alpha_1^2\sigma_\mu^4\Big]\right\}$$
$$= \frac{1}{\sqrt{2\pi}\sigma_\mu}\exp\left[-\frac{1}{2\sigma_\mu^2}\big(\beta_{0i} - \boldsymbol{z}_i'\boldsymbol{\gamma} - \alpha_1\sigma_\mu^2\big)^2\right].$$

Logo, na amostra

$$\beta_{0i} \mid \boldsymbol{z}_i, \boldsymbol{\gamma}, \sigma_\mu^2 \sim N(\boldsymbol{z}_i'\boldsymbol{\gamma} + \alpha_1\sigma_\mu^2, \sigma_\mu^2). \tag{A.7}$$

A.3 Distribuição Amostral de O_{ij}

A distribuição populacional do estrato está dada por (5.12), tem-se também que

$$E_p[\pi_{j|i} \mid O_{ij}, y_{ij}, \boldsymbol{\eta}] = q_j^i,$$

i.e, a fração de amostragem do estrato a que pertence o aluno, e

$$E_p[\pi_{j|i} \mid y_{ij}, \boldsymbol{\eta}] = \sum_{k=1}^{3} q_k^i Pr\big(O_{ij} = k \mid y_{ij}, \boldsymbol{\eta}, \sigma_p\big).$$

Logo, a distribuição amostral de O_{ij} está dada por

$$Pr_s(O_{ij} = 1) = \frac{q_1^i}{\sum_{k=1}^{3} q_k^i Pr\big(O_{ij} = k \mid y_{ij}, \boldsymbol{\eta}, \sigma_p\big)} \times \Phi\big(\delta_1 - \delta_2 y_{ij}\big),$$

$$Pr_s(O_{ij} = 2) = \frac{q_2^i}{\sum_{k=1}^{3} q_k^i Pr\big(O_{ij} = k \mid y_{ij}, \boldsymbol{\eta}, \sigma_p\big)} \times \Big[\Phi\big(\delta_3 - \delta_2 y_{ij}\big) - \Phi\big(\delta_1 - \delta_2 y_{ij}\big)\Big]$$

$$Pr_s(O_{ij} = 3) = \frac{q_3^i}{\sum_{k=1}^{3} q_k^i Pr\big(O_{ij} = k \mid y_{ij}, \boldsymbol{\eta}, \sigma_p\big)} \times \Big[1 - \Phi\big(\delta_3 - \delta_2 y_{ij}\big)\Big],$$

onde $\delta_1 = \left(\frac{1.76-\eta_0}{\sigma_p}\right)$, $\delta_2 = \frac{\eta_1}{\sigma_p}$, $\delta_3 = \left(\frac{1.97-\eta_0}{\sigma_p}\right)$.

A.4 Distribuição Amostral de y_{ij}

De (5.9) $y_{ij} \sim Bernoulli(\theta_{ij})$, uma vez mais, seguindo Pfeffermann et al. (1998), a distribuição amostral de y_{ij} é dada por

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

Usando (5.13),

$$E_p[\pi_{j|i} \mid y_{ij}, \boldsymbol{x}_{ij}, \beta_{0i}, \boldsymbol{\beta}] = (q_1^i - q_2^i)\Phi\big(\delta_1 - \delta_2 y_{ij}\big) + (q_2^i - q_3^i)\Phi\big(\delta_3 - \delta_2 y_{ij}\big) + q_3^i,$$

e

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

Por outro lado, de (5.1) e (5.2) tem-se

$$\begin{aligned} f_p(y_{ij} \mid \theta_{ij}) &= \theta_{ij}^{y_{ij}}(1-\theta_{ij})^{1-y_{ij}} \\ \log\Big(\frac{\theta_{ij}}{1-\theta_{ij}}\Big) &= \beta_{0i} + \boldsymbol{x}'_{ij}\boldsymbol{\beta} \\ \theta_{ij} &= \frac{\exp\big(\beta_{0i} + \boldsymbol{x}'_{ij}\boldsymbol{\beta}\big)}{1+\exp\big(\beta_{0i} + \boldsymbol{x}'_{ij}\boldsymbol{\beta}\big)}, \end{aligned}$$

logo,

$$f_p(y_{ij} \mid \boldsymbol{x}_{ij}, \beta_{0i}, \boldsymbol{\beta}) = \frac{\exp\big[y_{ij}(\beta_{0i} + \boldsymbol{x}'_{ij}\boldsymbol{\beta})\big]}{1+\exp\big(\beta_{0i} + \boldsymbol{x}'_{ij}\boldsymbol{\beta}\big)}, \tag{A.8}$$

daí,

$$\begin{aligned} Pr(y_{ij}=0 \mid \boldsymbol{x}_{ij}, \beta_{0i}, \boldsymbol{\beta}) &= \frac{1}{1+\exp\big(\beta_{0i} + \boldsymbol{x}'_{ij}\boldsymbol{\beta}\big)} \\ Pr(y_{ij}=1 \mid \boldsymbol{x}_{ij}, \beta_{0i}, \boldsymbol{\beta}) &= \frac{\exp\big(\beta_{0i} + \boldsymbol{x}'_{ij}\boldsymbol{\beta}\big)}{1+\exp\big(\beta_{0i} + \boldsymbol{x}'_{ij}\boldsymbol{\beta}\big)} \end{aligned}$$

$$\begin{aligned} f_s(y_{ij} \mid \boldsymbol{x}_{ij}, \beta_{0i}, \boldsymbol{\beta}) = \big[(q_1^i - q_2^i)\Phi\big(\delta_1 - \delta_2 y_{ij}\big) + (q_2^i - q_3^i)\Phi\big(\delta_3 - \delta_2 y_{ij}\big) + q_3^i\big] \\ \times \exp\big[y_{ij}(\beta_{0i} + \boldsymbol{x}'_{ij}\boldsymbol{\beta})\big] \times \Big[\big[(q_1^i - q_2^i)\Phi(\delta_1) + (q_2^i - q_3^i)\Phi(\delta_3) + q_3^i\big] \\ + \big[(q_1^i - q_2^i)\Phi(\delta_1 - \delta_2) + (q_2^i - q_3^i)\Phi(\delta_3 - \delta_2) + q_3^i\big]\exp\big(\beta_{0i} + \boldsymbol{x}'_{ij}\boldsymbol{\beta}\big)\Big]^{-1}. \end{aligned} \tag{A.9}$$

Logo, na amostra $y_{ij} \sim Bernoulli(\theta_{ij}^s)$ onde

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

Apêndice B

ROTINAS COMPUTACIONAIS

B.1 Geração das populações no *R*

```
#-------------------------------------------------------------------------
# ROTINA R PARA A GERAÇÃO DAS POPULAÇÕES DE ESCOLAS E ALUNOS
        Esta rotina cria 05 populações independentemente e salva os
        dados em formato txt.
#-------------------------------------------------------------------------

options(type="decimal",digits=4,object.size=5e8)

escola <- matrix(scan("d:/dissertacao/escola.txt"),392,5,byrow=T)
aluno <- matrix(scan("d:/dissertacao/aluno.txt"),14831,5,byrow=T)
POBESCOLA <- matrix(NA,1,8); POBALUNO  <- matrix(NA,1,12)

for(k in 1:5){
#*************************************************************************
#   Passo 01: GERAÇÃO DAS COVARIAVEIS DAS ESCOLAS
#*************************************************************************
    M <- 392
    ESCOLA <- matrix(NA,M,8)
    B0J <- rep(NA,M)
    ESCOLA[,1] <- k
    ESCOLA[,4] <- escola[,4]
    ESCOLA[,5] <- escola[,5]
#*************************************************************************
#   Passo 02: GERAÇÃO DOS INTERCEPTOS DAS ESCOLAS
#*************************************************************************
    for(j in 1:M)
    {
        ESCOLA[j,2] <- j
        U0J <- rnorm(1,0,sqrt(0.75))
        B0J[j] <- (2.65 - 0.28*ESCOLA[j,4] - 0.56*ESCOLA[j,5] + U0J)
    }
#*************************************************************************
#   Passo 03: GERAÇÃO DOS TAMANHOS DAS ESCOLAS
#*************************************************************************
        meanlogMj <- rep(NA,M)
        LOGMJ <- rep(NA,M)
        for(j in 1:M)
```

```
        {
            meanlogMj[j] <- (3.99 + 0.52*B0J[j])
            LOGMJ[j] <- rnorm(1,meanlogMj[j],sqrt(0.18))
        }
        MJ <- round(exp(LOGMJ))
        N <- sum(MJ)
        ESCOLA[,3] <- MJ
#**********************************************************************
#   Passo 04: GERAÇÃO DAS COVARIAVEIS DOS ALUNOS
#**********************************************************************
    ALUNO <- matrix(NA,N,12)
    ALUNO[,1] <- k
    aux <- 0
    for(j in 1:M)
    {
        for(i in 1:MJ[j])
        {
            ALUNO[i+aux,2] <- j
            ALUNO[i+aux,3] <- i
        }
        aux <- aux + MJ[j]
    }
            ALUNO[,5] <- sample(aluno[,2], size=N, replace=T)
            ALUNO[,6] <- sample(aluno[,3], size=N, replace=T)
            ALUNO[,7] <- sample(aluno[,4], size=N, replace=T)
            ALUNO[,8] <- sample(aluno[,5], size=N, replace=T)
#**********************************************************************
#   Passo 05: GERAÇÃO DAS RESPOSTAS DOS ALUNOS
#**********************************************************************
    # Repetindo os interceptos das escolas
        B0J1 <- rep(NA,N)
        aux <- 0
        for(j in 1:M)
        {
            for(i in 1:MJ[j])
            {
                B0J1[i+aux] <- B0J[j]
            }
            aux <- aux + MJ[j]
        }
    # Gerando os logits, proporções e respostas

        LOGIT <- rep(NA,N)
        PI <- rep(NA,N)
        LOGIT <- B0J1 - 0.66*ALUNO[,5] - 0.95*ALUNO[,6] - 2.1*ALUNO[,7] - 0.43*ALUNO[,8]
        PI <- (exp(LOGIT))/(1+exp(LOGIT))
        ALUNO[,4] <- rbinom(N,1,PI)
#**********************************************************************
#   Passo 06: GERAÇÃO DAS ESTRATOS DOS ALUNOS
```

```
#*********************************************************************
        E1J <- rep(0,M)
        E2J <- rep(0,M)
        E3J <- rep(0,M)
while(min(E1J)<4 || min(E2J)<4 || min(E3J)<2) {
        E1J <- rep(0,M)
        E2J <- rep(0,M)
        E3J <- rep(0,M)
    PROB1EST <- rep(NA,N)
    PROB1EST <- (1.67 + 0.29*ALUNO[,4] + rnorm(N,0,0.24))
    for(i in 1:N)
        {
        if(PROB1EST[i] <  1.76)
            {ALUNO[i,9] <- 1
            ALUNO[i,10:12] <- c(1,0,0)}
            else {
                    if(PROB1EST[i] >= 1.76 && PROB1EST[i] < 1.97)
                        {ALUNO[i,9] <- 2
                        ALUNO[i,10:12] <- c(0,1,0)}
                        else {
                            if(PROB1EST[i] >= 1.97)
                                {ALUNO[i,9] <- 3
                                ALUNO[i,10:12] <- c(0,0,1)}
                                }
                    }
        }
        aux <- 0
        for(j in 1:M)
        {
            for(i in 1:MJ[j])
            {
                E1J[j] <- E1J[j] + ALUNO[(i+aux),10]
                E2J[j] <- E2J[j] + ALUNO[(i+aux),11]
                E3J[j] <- E3J[j] + ALUNO[(i+aux),12]
            }
            aux <- aux + MJ[j]
        }
}

    n1 <-4 ; n2 <- 4 ; n3 <- 2
    ESCOLA[,6] <- n1/E1J
    ESCOLA[,7] <- n2/E2J
    ESCOLA[,8] <- n3/E3J
    POBESCOLA <- rbind(POBESCOLA,ESCOLA)
    POBALUNO  <- rbind(POBALUNO,ALUNO)
}
#*********************************************************************
#   Passo 07: EXPORTAÇÃO DOS DADOS
#*********************************************************************
```

```
    write(t(POBESCOLA), file="d:/Dissertacao/Populacao/ESCOLA01.txt",ncolumns=8)
    write(t(POBALUNO), file="d:/Dissertacao/Populacao/ALUNO01.txt",ncolumns=12)
#*********************************************************************
#   Passo 08 (opcional): NOMES DAS VARIAVEIS GERADAS
#*********************************************************************
    escola.lab_c("N°Escola","Tamanho","Regiao1","Regiao2","Es1", "Es2", "Es3")
    dimnames(ESCOLA) <- list(NULL, escola.lab)
    aluno.lab_c("N°Escola","N°Aluno","Y","X1","X2","X3","X4","Estrato","Es1","Es2","Es3")
    dimnames(ALUNO) <- list(NULL, aluno.lab)
```

B.2 Obtenção de amostras no *SAS*

```
/* ------------------------------------------------------
    AMOSTRA.sas
    Esta macro seleciona amostras com 04 planos amostrais
    ------------------------------------------------------
*/
\%macro AMOSTRA(popesc,amoesc,popalu,amoalu,nesc,nalu);

    /* Parte I: Gera amostras AASAAS e AASEST */

    * Selecionando amostra de escolas usando AAS;
    PROC SURVEYSELECT
        DATA=&popesc
        METHOD=srs
        SAMPSIZE=&nesc
        OUT=AasAas.&amoesc;
        ID popula escola tamanho reg1 reg2 f1 f2 f3;
        STRATA popula;
    RUN;

    * Preparando arquivo para selecionar amostras de alunos AAS;
        * Juntando arquivos;
        DATA poptemp1;
            MERGE &popalu AasAas.&amoesc;
            BY popula escola;
            RUN;
        * Selecionando linhas;
        DATA amostra1;
            SET poptemp1;
            IF tamanho >= 0;
            RUN;

    * Selecionando amostra de alunos usando AAS;
    PROC SURVEYSELECT
        DATA=amostra1
        METHOD=srs
        SAMPSIZE=&nalu
        OUT=AasAas.&amoalu;
```

```
    STRATA popula escola;
    ID popula escola aluno Y X1 X2 X3 X4 estrato est1 est2 est3;
    RUN;

* Preparando arquivo para selecionar amostras de alunos EST;
PROC SORT data=amostra1;
    BY popula escola estrato;
    RUN;

* Selecionando amostra de alunos usando EST;
    * (amostra simples em cada estrato);
PROC SURVEYSELECT
    DATA=amostra1
    METHOD=srs
    SAMPSIZE=(4 4 2 ... 4 4 2)
    OUT=AasEst.&amoalu;
    STRATA popula escola estrato;
    ID popula escola aluno Y X1 X2 X3 X4 estrato est1 est2 est3;
    RUN;

/* Parte II: Gera amostras PPTAAS e PPTEST */

* Selecionando amostra de escolas usando PPT;
PROC SURVEYSELECT
    DATA=&popesc
    METHOD=pps_sampford
    SAMPSIZE=&nesc
    OUT=PptAas.&amoesc;
    SIZE tamanho;
    ID popula escola tamanho reg1 reg2 f1 f2 f3;
    STRATA popula;
    RUN;

* Preparando arquivo para selecionar amostras de alunos AAS;
    * Juntando arquivos;
    DATA poptemp2;
        MERGE &popalu PptAas.&amoesc;
        BY popula escola;
        RUN;
    * Selecionando linhas;
    DATA amostra2;
        SET poptemp2;
        IF tamanho >= 0;
        RUN;

* Selecionando amostra de alunos usando AAS;
PROC SURVEYSELECT
    DATA=amostra2
    METHOD=srs
```

```
        SAMPSIZE=&nalu
        OUT=PptAas.&amoalu;
        STRATA popula escola;
        ID popula escola aluno Y X1 X2 X3 X4 estrato est1 est2 est3;
        RUN;

    * Preparando arquivo para selecionar amostras de alunos EST;
    PROC SORT data=amostra2;
        BY popula escola estrato;
        RUN;

    * Selecionando amostra de alunos usando EST;
        * (amostra simples em cada estrato);
    PROC SURVEYSELECT
        DATA=amostra2
        METHOD=srs
        SAMPSIZE=( 4 4 2 ... 4 4 2)
        OUT=PptEst.&amoalu;
        STRATA popula escola estrato;
        ID popula escola aluno Y X1 X2 X3 X4 estrato est1 est2 est3;
        RUN;
\%mend AMOSTRA;
/*************************************************************************
 Argumentos:
    popesc  : Arquivo com dados da população de escolas
    amoesc  : Arquivo para guardar dados da amostra de escolas
    popalu  : Arquivo com dados da população de alunos
    amoalu  : Arquivo para guardar dados da amostra de alunos
    nesc    : Tamanho da amostra de escolas
    nalu    : Tamanho da amostra de alunos por escola
/*************************************************************************
 Libraries:
    pops    : Pasta com populações
    AasAas  : Pasta com amostras AASAAS
    AasEst  : Pasta com amostras AASEST
    PptAas  : Pasta com amostras PPTAAS
    PptEst  : Pasta com amostras PPTEST
 *************************************************************************/
```

B.3 Rotina do *WinBUGS*

```
.............................................................
        AMOSTRAGEM PPT-EST (informativa nos 2 níveis)
        Modelando Y, beta0, M e O com as distribuições amostrais,
        Ultima modificação: 22/01/2003
.............................................................
    model
        {
        # Prioris
```

```
 beta1 ~ dnorm(0,0.01)
 beta2 ~ dnorm(0,0.01)
 beta3 ~ dnorm(0,0.01)
 beta4 ~ dnorm(0,0.01)
 gama0 ~ dnorm(0,0.01)
 gama1 ~ dnorm(0,0.01)
 gama2 ~ dnorm(0,0.01)
 taubeta0 ~ dpar(1,0.01)
 s2beta0 <- 1 /taubeta0
alpha0 ~ dnorm(0,0.01)
alpha1 ~ dnorm(0,0.01)
tautam ~ dpar(1,0.01)
 s2tam <- 1 / tautam
delta1 ~ dnorm(0,0.01)
 delta2 ~ dnorm(0,0.01)
 delta3 ~ dnorm(0,0.01)
 phi1 <- phi(delta1)
 phi2 <- phi(delta3)
 phi3 <- phi(delta1 - delta2)
 phi4 <- phi(delta3 - delta2)
 eta0   <- 1.76 - delta1*0.24
 eta1   <- delta2*0.24

for( i in  1:40)
{
    # Distribuição amostral do tamanho (Mi)
        meantam[i]<- alpha0 + alpha1*BETA0[i] + s2tam
        TAMANHO[i] ~ dlnorm(meantam[i],tautam)

    # Distribuição amostral do intercepto (beta0i)
        mbeta0[i]<- gama0 + gama1*REG1[i] + gama2*REG2[i]
                          + alpha1*s2beta0
        BETA0[i] ~ dnorm(mbeta0[i],taubeta0)

    for( j in (n[i]+1):n[i+1])
 {
    # Distribuição amostral do estrato (Oij)
        Op[j,1] <- phi(delta1 - delta2*Y[j])
        Op[j,2] <- phi(delta3 - delta2*Y[j]) - Op[j,1]
        Op[j,3] <- 1- Op[j,1] - Op[j,2]
        den1[j] <- f[i,1]*Op[j,1] + f[i,2]*Op[j,2] + f[i,3]*Op[j,3]
        Os[j,1] <- (f[i,1]/den1[j]) * Op[j,1]
        Os[j,2] <- (f[i,2]/den1[j]) * Op[j,2]
        Os[j,3] <- (f[i,3]/den1[j]) * Op[j,3]
        ESTRATO[j,1:3] ~ dmulti(Os[j,1:3],1)

    # Distribuição amostral da resposta (Yij)
        p0[j] <-  exp(BETA0[i] + beta1*X1[j] + beta2*X2[j]
                               + beta3*X3[j] + beta4*X4[j])
```

```
                e1[j] <- (f[i,1]-f[i,2])*phi1 + (f[i,2]-f[i,3])*phi2 + f[i,3]
                e2[j] <- (f[i,1]-f[i,2])*phi3 + (f[i,2]-f[i,3])*phi4 + f[i,3]

                ts[j] <- 1/(1 + e1[j]/(e2[j]*p0[j]))

                Y[j] ~ dbern(ts[j])

            # Deviance
                YHAT[j] ~ dbern(ts[j])
                L[j] <- (Y[j] + 0.5)*(log(Y[j] + 0.5)-log(YHAT[j] + 0.5))+
                         (1.5 - Y[j])*(log(1.5 - Y[j])-log(1.5 - YHAT[j]))
            }
        }
            D <- 2*sum(L[])
    }
.............................................................
```

Apêndice C

MEDIDAS DE BONDADE DE AJUSTE E SELEÇÃO DE MODELOS

Medidas de Bondade de Ajuste

Nesta seção descrevem-se as medidas de Bondade de Ajuste utilizadas para avaliar a performance dos modelos nas Seções 5.7 e 6.2. Dado que os dados y_{obs} só tomam valor 0 ou 1, seguiu-se o trabalho de Da Costa (2000) que realizou a avaliação e comparação de modelos hierárquicos para o mapeamento da desnutrição infantil no Brasil, e definiu a variável y_{ki} sendo igual a 1 se a criança i do estado k for considerada desnutrida e sendo igual a zero caso contrário.

Sensibilidade e Especificidade

A *Sensibilidade* indica a proporção de indivíduos para os quais o modelo prevê o valor "1" corretamente, i.e. $y_{j,rep} = y_{j,obs} = 1$ (verdadeiros positivos). A *Especificidade* indica a proporção de indivíduos para os quais o modelo prevê o valor "0" corretamente, i.e. $y_{j,rep} = y_{j,obs} = 0$ (verdadeiros negativos).

As medidas de Sensibilidade e Especificidade de cada modelo foram obtidas mediante a seguinte aproximação:

1. Simulou-se o vetor de parâmetros de modelo da respectiva distribuição a posteriori

2. Calculou-se o vector de π_j com o vector de parâmetros simulado no passo anterior

3. Gerou-se o valor $Y_{j,rep}$ com distribuição Bernoulli de parâmetro π_j

4. Construiu-se a seguinte tabela:

Amostra		y_{obs}		Total
		1	0	
y_{rep}	1	n_{11}	n_{12}	$n_{1.}$
	0	n_{21}	n_{22}	$n_{2.}$
Total		$n_{.1}$	$n_{.2}$	$n_{..}$

5. A sensibilidade foi estimada por $n_{11}/n_{.1}$ e a especificidade por $n_{22}/n_{.2}$

Os 5 passos anteriores foram repetidos 1000 vezes a fim de obter 1000 simulações de Sensibilidade e Especificidade.

Percentual de acertos na amostra preditiva

Devido a natureza dicotômica de y_{obs}, pode-se calcular o número de vezes que o modelo faz uma boa predição (replicação), i.e, as vezes em que $y_{j,rep} = y_{j,obs}$, e utilizar a Proporção de Acertos na Amostra Preditiva (Pac) como medida de discriminação do modelo (Da Costa (2000)).

Utilizando o mesmo algoritmo para a obtenção da Sensibilidade e Especificidade, a partir da tabela 4, tem-se que

$$Pac = \frac{n_{11} + n_{22}}{n_{..}} \tag{C.1}$$

Outra medida que pode ser utilizada no caso das variáveis 0-1 é a Proporção de Acertos por Indivíduo (Pacpi). Para cada indivíduo (aluno ou família) o Pacpi é:

$$Pacpi_j = \frac{1}{L}\sum_{l=1}^{L} I(y^l_{j,rep} = y_{j,obs}), j = 1, \ldots, n \tag{C.2}$$

onde L é o número de replicações e n é o número de indivíduos na amostra. Como no caso das outras medidas de Bondade de Ajuste, foram obtidas 1000 simulações do Pacpi para cada indivíduo.

O ideal é que o Pacpi esteja perto de 100%, caso contrário estaria indicando que em grande parte das L simulações, o valor replicado, $y^l_{j,rep}$, é diferente do valor real $y_{j,obs}$, o que significa que o indivíduo não segue um padrão similar aos outros indivíduos com o mesmo perfil das covariáveis.

Percentual de "uns" na amostra preditiva

O p-valor Bayesiano é definido por Gelman et al. (1995) como a probabilidade de que os dados replicados sejam mais extremos do que os dados observados, quando medidos por $T(y, \zeta)$, uma medida escalar que resume os parâmetros do modelo, ζ, e os dados.

Neste trabalho utilizamos o "número de uns" na amostra como estatística T e o p-valor estimado foi o número de vezes em que $T(y^l_{rep}) \geq T(y_{obs})$, dividido pelo número de amostras do vetor y_{rep}, que neste caso foi L=1000.

Seleção de Modelos

Critério do *Deviance*

Waller, Carlin, Xia, e Gelfand (1997) propuseram trabalhar no espaço preditivo, estendendo os métodos Laud e Ibrahim (1995), para selecionar modelos não regulares.

A distribuição básica necessária é

$$f(y_{rep} \mid y_{obs}) = \int f(y_{rep} \mid \zeta) f(\zeta \mid y_{obs}) d\zeta, \tag{C.3}$$

onde ζ denota todos os parâmetros do modelo e y_{rep} é o vetor de valores replicados

do vetor y_{obs}. Para o modelo M_i, (C.3) é

$$f(y_{rep} \mid y_{obs}, M_i) = \int f(y_{rep} \mid \zeta^{(i)}, M_i) f(\zeta^{(i)} \mid y_{obs}, M_i) d\zeta^{(i)}, \tag{C.4}$$

O procedimento para selecionar modelos proposto por Waller et al. (1997) consiste em:

1. Identificar uma função de discrepância $d(y_{rep}, y_{obs})$,

2. Calcular

$$E[d(y_{rep}, y_{obs}) \mid y_{obs}, M_i] \tag{C.5}$$

3. Selecionar o modelo que minimiza (C.5).

Para verossimilhanças Gaussianas, Laud e Ibrahim (1995) propuseram

$$d(y_{rep}, y_{obs}) = (y_{rep} - y_{obs})^t (y_{rep} - y_{obs}). \tag{C.6}$$

Para verossimilhanças não Gaussianas, Waller et al. (1997) propuseram o uso do *Deviance*.

Em conseqüência, no caso da distribuição Bernoulli, utiliza-se a seguinte função de discrepância:

$$d(y_{rep}, y_{obs}) = 2 \sum_j \Big\{ y_{j,obs} \log\big(\frac{y_{j,obs}}{y_{j,rep}}\big) + (1 - y_{j,obs}) \log\big(\frac{1 - y_{j,obs}}{1 - y_{j,rep}}\big) \Big\}, \tag{C.7}$$

e com a finalidade de evitar o problemas de cálculo devido aos "zeros", faz-se a seguinte correção em (C.7):

$$d(y_{rep}, y_{obs}) = 2 \sum_j \Big\{ \big(y_{j,obs} + 0.5\big) \log\big(\frac{y_{j,obs} + 0.5}{y_{j,rep} + 0.5}\big) + \big(1.5 - y_{j,obs}\big) \log\big(\frac{1.5 - y_{j,obs}}{1.5 - y_{j,rep}}\big) \Big\} \tag{C.8}$$

O valor do $E[d(y_{rep}, y_{obs})]$ foi aproximado pelo método de Monte Carlo. As rotinas de estimação dos modelos foram implementadas no WinBUGS 1.4. Geraram-se

cadeias de 10 000 iterações de cada parâmetro do modelo com o que foi atingida a convergência. Em seguida, foram realizadas 1000 iterações incluindo o cálculo de $d(y_{rep}, y_{obs})$, segundo (C.8),a média delas foi a aproximação utilizada para $E[d(y_{rep}, y_{obs})]$.

DIC

O *Deviance Information Criterion* (DIC) pode ser utilizado para avaliar a complexidade de um modelo e para comparar modelos diferentes . Detalhes sobre o DIC encontram-se em Spiegelhalter, Best, Carlin, e Linde (2001) e Spiegelhalter, Best, e Carlin (1998).

O DIC é dado por $DIC = \bar{D} + p_D = D(\bar{\zeta}) + 2p_D$, onde

1. $\bar{D}$ é a média a posteriori do deviance. O *deviance* é definido como $D = -2\log\big[f(y \mid \zeta)\big]$.

2. $D(\bar{\zeta})$ é uma estimativa pontual do deviance obtido ao substituir as médias a posteriori de ζ no deviance, assim, $D(\bar{\zeta}) = -2\log\big[f(y \mid \bar{\zeta})\big]$.

3. p_D é o "número efetivo de parâmetros" dado por $p_D = \bar{D} - D(\bar{\zeta})$.

O cálculo do DIC vem incorporado na versão 1.4 do WinBUGS. O menor DIC indica o modelo que fará melhores predições a curto prazo, no mesmo sentido que o AIC. i.e., indica o modelo que "replica" melhor o conjunto de dados.

Uma observação importante é que os DICs só são comparáveis em cima de modelos com exatamente os mesmos dados observados, mas não há nenhuma necessidade de eles serem aninhados. Daqui que o DIC não pode ser utilizado para comparar as performances dos modelos IG e SM, mas sim, os modelos SM e os que contem as variáveis do desenho.