Universidade Federal de Pernambuco
Centro de Ciências Exatas e da Natureza
Departamento de Estatística

Pós-graduação em Estatística

Inferência sob planos amostrais de cadastro duplo

Hemílio Fernandes Campos Coêlho

Tese de Doutorado

Recife
23 de fevereiro de 2011

Universidade Federal de Pernambuco
Centro de Ciências Exatas e da Natureza
Departamento de Estatística

Hemílio Fernandes Campos Coêlho

Inferência sob planos amostrais de cadastro duplo

Trabalho apresentado ao Programa de Pós-graduação em Estatística do Departamento de Estatística da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Doutor em Estatística.

Orientador: Cristiano Ferraz

Recife
23 de fevereiro de 2011

Catalogação na fonte
Bibliotecária Jane Souto Maior, CRB4-571

Coelho, Hemílio Fernandes Campos
Inferência sob planos amostrais de cadastro duplo / Hemílio Fernandes Campos Coelho - Recife: O Autor, 2011.
xv, 144 p. : il., fig., tab.

Orientador: Cristiano Ferraz.
Tese (doutorado) Universidade Federal de Pernambuco. CCEN. Estatística, 2011.

Inclui bibliografia e apêndices.

1. Estatística aplicada. 2. Amostragem. 3. Inferência. I. Ferraz, Cristiano (orientador). II. Título.

310 CDD (22. ed.) MEI2011 – 024

Universidade Federal de Pernambuco
Pós-Graduação em Estatística

23 de fevereiro de 2011

Nós recomendamos que a tese de doutorado de autoria de

Hemílio Fernandes Campos Coêlho

intitulada

"Inferência sob planos amostrais de cadastro duplo"

seja aceita como cumprimento parcial dos requerimentos para o grau de Doutor em Estatística.

Coordenador da Pós-Graduação em Estatística

Banca Examinadora:

Cristiano Ferraz orientador

Getúlio José Amorim Amaral

Júlio da Motta Singer (USP)

Marcel de Toledo Vieira (UFJF)

Fernando Antônio da Silva Moura (UFRJ)

Este documento será anexado à versão final da dissertação.

Dedico este trabalho à minha esposa,
Merciany Rodrigues Ferreira.

Agradecimentos

A realização deste trabalho é uma prova da Fé em Deus. Uma Fé que nos momentos mais difíceis surge como uma força maior que nos move pra frente sem deixar aumentar a dor nem sangrar as feridas. Fé no Deus que nos mostra que para Ele nada é impossível, e que nos dá condições na vida para enfrentar todas as adversidades, sejam elas quais forem. Portanto, meu primeiro agradecimento é para Deus, por ter me dado uma missão importante neste mundo, que sempre estará condicionada pelas minhas ações.

Aos meus pais, por toda a força e incentivo dada desde a infância, em especial à minha mãe, Wilma, uma das pernambucanas mais fantásticas que existe. Uma guerreira que incondicionalmente luta no dia-a-dia pelo melhor para seus filhos, e que sempre esteve presente nos momentos em que mais precisei de seu apoio para seguir firme nos estudos.

À minha esposa Merciany Rodrigues, por ser meu ágape. Ágape em grego significa **o amor incondicional**. Nos últimos dez anos, desde a graduação até o fim do Doutorado, ela esteve ao meu lado como amiga, namorada, noiva e esposa, sendo a minha fortaleza e ágape. Nos momentos em que apenas a emoção transbordava em mim, ela vinha e se tornava a razão para o meu equilíbrio e perseverança.

Aos meus irmãos, pelos constantes momentos engraçados na vida de irmãos e também por toda a força dada nessa jornada.

Aos meus sogros, Izaías e Mercedes, pelo exemplo de Fé e pelos valorosos conselhos.

À minha avó, Maria José, que carinhosamente chamo de “Mãe Bezinha”, pelas valorosas lições de vida.

Ao meu tio Jadiel Sobreira (in memorian), que mesmo não estando neste mundo há quase seis anos, ainda transmite importantes lições para toda a família pelo exemplo que sempre será em todos os aspectos. Agradeço também a todos os primos que acompanharam minha jornada.

Aos meus tios maternos, em especial Nilma e Lindoval, pela constante alegria e apoio.

Ao professor Cristiano Ferraz, que foi mais que um orientador. O professor Cristiano é um verdadeiro amigo e um exemplo de profissional, pois confiou plenamente na minha capacidade de desenvolver este trabalho e sempre me incentivou a ter perseverança e jamais desistir diante dos difíceis desafios que surgem na vida.

Aos professores Marcel Vieira e Damião Nóbrega, pelas importantes contribuições dadas na qualificação do Doutorado. Ao professor Damião, um grande agradecimento por ter me recebido na UFRN durante dois dias e ter contribuído de forma extraordinária para o desenvolvimento desta tese.

Aos estatísticos Jeffey Bailey e Chadd Crouse, do National Agricultural Statistics Service of the United States Departament of Agriculture (NASS/USDA), pelo material fornecido sobre

o trabalho desenvolvido com cadastro duplo nos Estados Unidos.

Aos grandes amigos da Estatística - UFPE: Abraão David, Carlos Renato, Claudyvan Paiva, Emmanuelle Araújo, Leila Rameh, Luiz Medeiros, Marcelo Rodrigo, Oscar Raposo, Raul Siqueira, Romero Filho, Robson Florêncio, Tadeu Rodrigues, Vanessa Santos, Lucemberg Pedrosa, Camilla Rocha, Hélio, Lucianna, Mariana Batista e Natália Pires, por todos os bons momentos vivenciados.

Aos grandes amigos da pós-graduação em Estatística, em especial Artur Lemonte, Basílides, Carlos Raphael, Daniel Neyra, Fábio Bayer, Fábio Fajardo, Nataly Monroy, Jane, Jeremias Leão, Juliana Pires, Luz Marina, Maria Lídia, Rejane Brito, Rita Lima, Sílvio Fernando, Tarciana Liberal, Tatiene Souza, Themis Abensur, e Vínicius Quintas. Cada um fez a diferença nos momentos de estudo, alegria, preocupação, perseverança.

Ao mestrando Oscar Lopez da Universidade Nacional da Colômbia, pela disponibilidade e atenção nas discussões sobre o tema desta tese.

Aos amigos professores que tive a honra de conhecer durante o prazeroso período em que trabalhei na UFPI, no saudoso Campus Ministro Reis Velloso, cidade de Parnaíba, Piauí: Fuad Hazime, Patrícia Hazime, André Perinotto, Cíntia Martins, Marcelo Filgueiras, Erik Rodarte, Roberto Ramos, Alexandro Marinho, Marcelo Rêgo, Gustavo Portela, Anna Carolina, Fábio Motta, Marcelo Coerltjens, Patrícia Coerltjens, Carla Eiras, France Keiko, André Luiz, Jand Venes, Valmária, Cleide e Fernando.

Aos professores do Departamento de Estatística da UFPE, em especial Maria Cristina Falcão Raposo, Sylvio José Pereira dos Santos, Klaus Leite Pinto Vasconcellos, Francisco Cribari Neto e Cláudia Regina Oliveira de Paiva Lima, pelo apoio e importantes orientações.

Aos amigos professores do Departamento de Estatística da UFPB, em especial Ulisses Umbelino, Eufrásio Neto e José Carlos, pelos valorosos conselhos e apoio para a finalização do Doutorado.

À secretária da pós, Valéria Bittencourt, uma profissional da mais alta competência e uma grande amiga que tive a honra de conhecer na pós.

Ao secretário da graduação, Lódino Neto, pela amizade e pelas constantes conversas e apostas sobre o futebol pernambucano.

Aos funcionários do Departamento em Estatística, em especial Jimmy, Maurício e Cícero.

Aos meus alunos.

À CAPES, pelo apoio financeiro nos dois anos iniciais do doutorado.

Resumo

A abordagem de cadastro duplo envolve um levantamento amostral onde dois cadastros são utilizados com o propósito de fornecer maior cobertura e identificar elementos de uma única população-alvo. Tal abordagem tem sido vastamente utilizada na literatura em situações onde um único cadastro não consegue fornecer cobertura completa da população-alvo, ou ainda, quando há diferenças de custo de amostragem realizada em cada um dos cadastros disponíveis. Esta tese estuda estratégias de inferência assistida por modelos aplicadas à abordagem de cadastro duplo, propondo estimadores do tipo regressão generalizado. Variáveis auxiliares são consideradas disponíveis em ambos os cadastros e utilizadas no processo de estimação para aumentar a precisão de estimativas. As propriedades de consistência, centralidade assintótica e erro quadrático médio assintótico dos novos estimadores propostos são apresentadas e uma simulação é realizada para analisar sua eficiência relativa a estimadores já propostos na literatura, sob a situação em que um plano de amostragem aleatória simples é aplicado em cada um dos cadastros. Os estimadores propostos têm potencialidade de aplicação direta em diversas áreas, como a de pesquisa agropecuária.

Palavras-chave: Cadastro duplo, estimador regressão, consistência, centralidade assintótica.

Abstract

A dual frame sampling approach is defined as a survey sample design with two sample frames providing coverage for the same target population. Under this context, we propose several generalized regression estimators are proposed and we proved that they are asymptotically design-unbiased, design-consistent and have an asymptotic mean square error under some conditions. Their performance is compared with estimators found in literature and investigated for several dual frame scenarios in a Monte Carlo study for simple random sampling design. The proposed estimators are motivated by their potential application to many areas, such as agricultural research, health and ecology.

key words: Dual frame, generalized regression estimator, design consistency, agricultural research, asymptotic design unbiased, asymptotic mean square error.

“Enquanto eu puder trabalhar eu trabalho, porque o trabalho me distrai, não é sacrifício. Cada problema que aparece é um esforço para resolver que me agrada muito e que ainda consigo fazer.”

–OSCAR NIEMEYER, ao completar 103 anos em 2010.

“Provai e vede como o Senhor é bom. Feliz quem encontra Nele o seu refúgio.”

–SALMO 34.

LISTA DE TABELAS

LISTA DE FIGURAS

SUMÁRIO

CAPÍTULO 1

Introdução

1 A abordagem de cadastro duplo

A abordagem de cadastro duplo é definida como um levantamento onde dois cadastros, denotados por A e B, são utilizados simultaneamente para identificar elementos de uma população-alvo. Nesta abordagem, amostras aleatórias independentes, denotadas por S_A e S_B, são obtidas de cada cadastro, respectivamente, através de planos amostrais com probabilidades $p(S_A)$ e $p(S_B)$ possivelmente diferentes. Hartley (1962) provavelmente foi o primeiro a apresentar uma estratégia de estimação que trouxe vantagens na relação custo-benefício em favor da abordagem de cadastro duplo. Motivados por esta contribuição inicial, vários autores contribuíram para o desenvolvimento e para a melhoria de estratégias de estimação sob essa abordagem, como Bankier, Lepkowski & Groves (1986), Lund (1968), Fuller & Burmeister (1972), Skinner (1991), Skinner & Rao (1996), Bosecker & Ford (1976), Haines & Pollock (1998), Carfagna (2004), Alpizar-Jara, Pollock & Haines (2005), Lohr (2007) e Lu (2007). No contexto de estratégias de estimação assistida por modelos, a lista de autores parece ser mais reduzida. Singh & Wu (2003) propuseram estimadores do tipo regressão modificados, baseados em métodos de calibração, e Coelho (2007) apresentou estratégias de estimação assistida por modelos baseadas em estimadores do tipo razão.

Existem inúmeras situações em que é possível utilizar a abordagem de cadastro duplo. Por exemplo, quando um dos cadastros não tem um grau de cobertura desejável da população-alvo, porém existe um segundo cadastro, que em conjunto com o primeiro, fornece cobertura completa. Esta situação é representada pela figura 1.1.

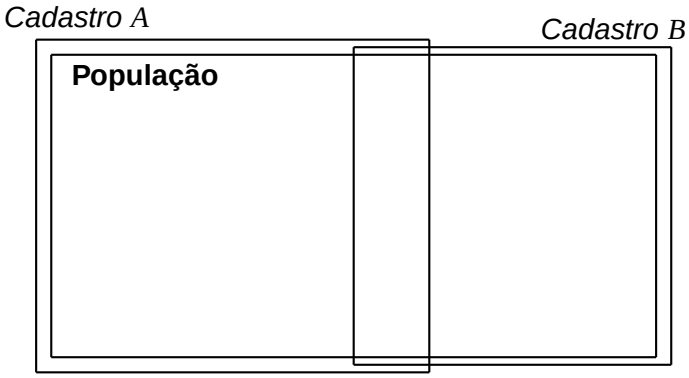


Figura 1.1: Cadastro duplo fornecendo cobertura completa para a população-alvo

Outro exemplo está na situação em que um cadastro cobre por completo a população-alvo, mas tem alto custo de seleção da amostra. Nesse caso, se houver um outro cadastro disponível e de abrangência menor que o primeiro, é possível utilizar ambos os cadastros de modo a tornar vantajosa a relação custo-benefício de seleção de elementos. Esta situação é representada pela figura 1.2.

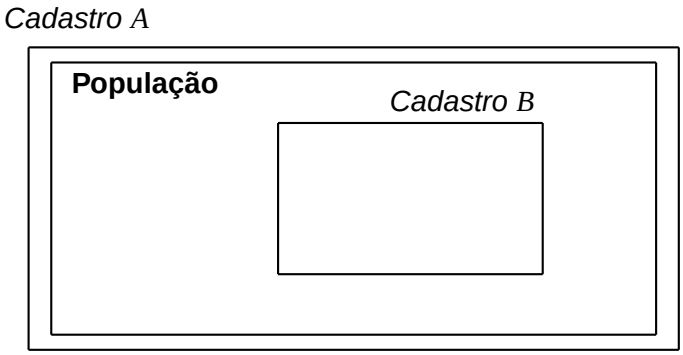


Figura 1.2: Cadastro duplo para melhorar a relação custo-benefício na seleção de elementos da população-alvo

Denote por $\mathscr{A}$ o conjunto de elementos pertencentes ao cadastro A e $\mathscr{B}$ o conjunto de elementos pertencentes ao cadastro B. O uso de dois cadastros induz a existência de no máximo três domínios disjuntos, a, b e ab, tais que $a = \mathscr{A} \cap \mathscr{B}^c$, $b = \mathscr{A}^c \cap \mathscr{B}$ e $ab = \mathscr{A} \cap \mathscr{B}$. Quando $\mathscr{B} \subset \mathscr{A}$ apenas dois domínios são gerados, como ilustra a figura 1.3.(b).

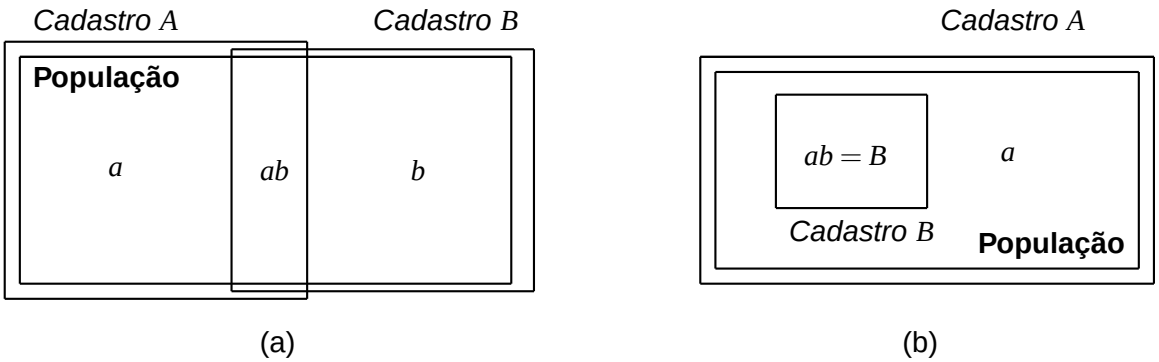


Figura 1.3: Domínios a, b e ab induzidos pela abordagem de cadastro duplo

A abordagem de cadastro duplo é um caso particular da abordagem de cadastros múltiplos em que F cadastros são utilizados ($F \geq 2$). Em geral, tal abordagem implica a existência de $2^F - 1$ domínios possíveis de serem identificados.

2 Aplicações

Situações reais em que a abordagem de cadastro duplo pode ser aplicada são encontradas na literatura, abrangendo diversas áreas de aplicação. A seguir, serão apresentados estudos que utilizaram a abordagem com sucesso ou que ainda estudam as suas potencialidades.

2.1 Agricultura

1. Pesquisa Agropecuária Nacional - Canadá

Armstrong (1979) estudou o uso da abordagem de cadastro duplo na Divisão de Agropecuária e Estatística do Canadá na condução da Pesquisa Agropecuária Nacional, baseada em um cadastro do tipo área. Devido a problemas neste tipo de cadastro, como a vulnerabilidade a altas taxas de não-resposta e a não-atualização ao longo dos anos, era necessário verificar se outras técnicas poderiam ser mais eficientes na produção de estimativas agropecuárias. Para isto, uma das províncias ligadas à produção agrícola, a de New Brunswick no ano de 1978 foi escolhida para a realização de um estudo piloto através da abordagem de cadastro duplo. Foram considerados dois cadastros:

- Um cadastro do tipo *área* (listagem de segmentos de área cobrindo um determinado território), que fornece total cobertura da população-alvo;
- e um cadastro do tipo *específico* (listagem de elementos pertencentes a população da província).

A complexidade encontrada no uso da abordagem neste cenário estava na forma de tratamento do cadastro do tipo área. Por ser um cadastro que contemplava grandes áreas, garantia a total cobertura da população-alvo. Neste cadastro, as grandes fazendas (que possuíam a maioria das informações de interesse) identificadas em censos anteriores sempre eram incluídas na pesquisa com probabilidade 1, para que e na amostra obtida não se corresse o risco das informações destas fazendas gerarem superestimativas dos parâmetros de interesse. Além disso, uma eventual exclusão de uma destas grandes fazendas poderia comprometer seriamente a qualidade das estimativas de interesse. Os demais elementos do cadastro de área eram selecionados com base em planos amostrais complexos, como o de conglomerado em dois estágios.

No estudo, o chamado estimador *screening* foi empregado, e tem este nome porque para ser utilizado, é necessário identificar quais elementos da amostra obtida do cadastro de área estão no cadastro de lista. Após a identificação, os elementos em duplicata são desprezados. Considerando A como sendo o cadastro do tipo área e B o cadastro do tipo específico, tem-se que $\mathscr{B} \subset \mathscr{A}$ (ver figura 1.3.(b)). Considerando uma amostragem aleatória simples em ambos os cadastros, o estimador *screening* para o total populacional $t_y = t_{ya} + t_{yB}$ é dado por:

$$\hat{t}_y = N_a \overline{y}_a + N_B \overline{y}_B,$$

onde t_{ya} é o total populacional do domínio a, t_{yB} é o total populacional do cadastro B, N_a é o tamanho populacional do domínio a, N_B é o tamanho populacional do cadastro B, $\overline{y}_a$ é a média amostral do domínio a e $\overline{y}_B$ é a média amostral do cadastro B. A variância do estimador *screening* é dada por

$$\text{Var}(\hat{t}_y) = (N_a)^2 \text{Var}(\overline{y}_a) + (N_B)^2 \text{Var}(\overline{y}_B)$$

onde $\text{Var}(\overline{y}_a)$ e $\text{Var}(\overline{y}_B)$ são as variâncias dos estimadores para a média no domínio a e cadastro B, respectivamente. Foram obtidos resultados satisfatórios na produção de estimativas agropecuárias em New Brunswick quando estes foram comparados aos da pesquisa realizada apenas com o cadastro de área. Esta experiência foi uma das primeiras a apresentar potencial de aplicação da abordagem de cadastro duplo em pesquisas agropecuárias.

2. National Agricultural Statistics Service (NASS-USA)

As pesquisas do setor agropecuário nos Estados Unidos são desenvolvidas pelo *United States Department of Agriculture* (USDA) através do *National Agritcultural Statistics Service* (NASS), que fornece oficialmente toda a informação estatística referente à agricultura. O NASS utiliza métodos de pesquisa baseados em múltiplos cadastros. Desde 1950, abordagens como a de cadastro duplo fornecem grande contribuição para as pesquisas agropecuárias do país.

O programa desenvolvido pelo NASS, intitulado *Multiple Frame Agricultural Survey*, foi iniciado em 1986 em 27 estados daquele país e conduzido nos meses de junho, julho, agosto, setembro, janeiro e março, com o intuito de obter estimativas referentes a áreas produtivas, número de unidades agrícolas em atividade, área total ocupada por unidades agrícolas, produção total das unidades agrícolas, dentre outros. São considerados os seguintes tipos de cadastros:

- Cadastro de área;
- Cadastros específicos (listagens de elementos da população-alvo).

O cadastro do tipo área é resultante da combinação dos cadastros de área construídos para cada estado, independentemente. A cada 10 ou 20 anos o cadastro é atualizado. Já os cadastros específicos são regularmente atualizados, e utilizados em combinação com o cadastro de área. Existem aproximadamente 2 milhões de unidades agrícolas nos Estados Unidos, e os cadastros específicos, quando construídos em grande escala, conseguem uma cobertura de aproximadamente 1,1 milhões de unidades agrícolas.

As informações coletadas para composição do cadastro específico são divididas em duas partes. A primeira é referente aos dados para identificar, localizar e contactar fazendeiros e empresários do setor: nome, endereço, telefone, estado, distrito, município, razão social, número de identificação de empregador, etc. A segunda parte contém dados sobre a fazenda e/ou negócio. Uma vez construído, o cadastro é constantemente atualizado através de pesquisas feitas ao longo do ano, e através do censo agropecuário. O cadastro específico é construído com múltiplos propósitos, dentre os quais é possível citar os seguintes:

i. identificação e classificação de fazendeiros e/ou empresários do setor que estão diretamente ligados a determinados produtos de interesse;

ii. estratificação para identificar grupos de fazendeiros e produtores que possuem determinado produto em comum, para que se possa coletar amostras em cada grupo;

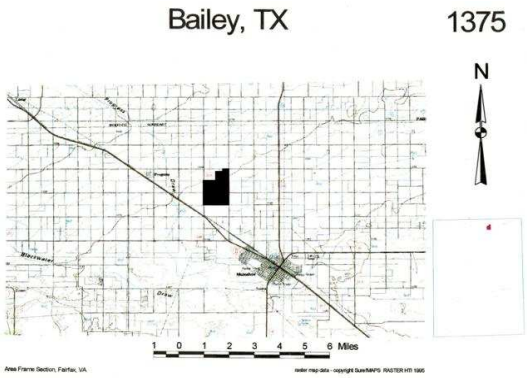
iii. pesquisas específicas, ou seja a coleta de dados é feita de forma a atender uma população-alvo em particular.

iv. possibilidade de se atribuir probabilidades de seleção para cada elemento do cadastro

O cadastro do tipo área do NASS é construído a partir de toda a área dos Estados Unidos, que é dividida em segmentos de terra através de delimitações, associando fazendas, produção, animais, etc. com a área pertencente a cada segmento. Imagens de satélite, mapas topográficos, sistemas de informação geográfico e fotos aéreas também são utilizados para construção dos cadastros de área. A vantagem desse cadastro está na cobertura completa de

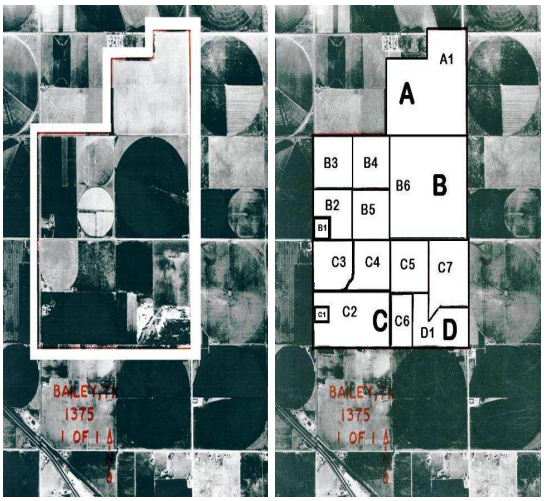
todas as unidades agrícolas em atividade. As figuras (1.4) e (1.5) ilustram um exemplo de segmento de área selecionado pelo NASS. Com o passar dos anos, novas tecnologias ajudaram a construção dos cadastros do tipo área.

Figura 1.4: Exemplo de seleção de segmento de área



Fonte: National Agritcultural Statistics Service (NASS)

Figura 1.5: Segmento escolhido e subdivisão do segmento em estratos



Fonte: National Agritcultural Statistics Service (NASS)

3. Pesquisas Agropecuárias do Departamento Administrativo Nacional de Estatística (Colômbia)

As pesquisas do setor agropecuário na Colômbia são desenvolvidas pelo *Departamento Administrativo Nacional de Estadística* (DANE) em função de seu papel como coordenador do *Sistema Estadístico Nacional* (SEN) daquele país. O DANE trabalha atualmente com o objetivo de fortalecer e consolidar o SEN através das chamadas "estatísticas estratégicas", na consolidação das informações obtidas através das pesquisas agropecuárias e adoção de instrumentos de coleta e iniciativas mais eficientes no sentido de melhorar a qualidade da informação, sua disponibilidade e acessabilidade. Tudo isto é focado na necessidade de obtenção de informações estatísticas que atendam à demanda e aos múltiplos propósitos das pesquisas implementadas na região.

A abordagem de cadastro duplo não está sendo utilizada no momento nas pesquisas colombianas. Porém está sendo considerada como possível alternativa para melhorar a relação custo-benefício. López (2010) talvez seja o primeiro a apresentar um estudo sobre a abordagem naquele país, pois utilizou os dados provenientes da chamada *Encuesta Nacional Agropecuria* (ENA), pesquisa realizada anualmente para analisar a eficiência de estimadores propostos na literatura diante dos múltiplos propósitos das pesquisas agropecuárias realizadas na Colômbia, como:

- Estimar a área cultivada para os principais produtos de caráter permanente ou temporário;
- Determinar a implantação de políticas voltadas para o meio-ambiente;
- Estabelecer as situações em que será necessária uma intervenção técnica;
- Determinar o uso e aproveitamento do solo;
- Conhecer a tecnologia utilizada nos sistemas produtivos dos principais produtos agrícolas (provisório ou permanentes)
- Manejo tecnológico do gado e controle da produção de leite.

2.2 Ecologia

1. Estimação de ninhos de águias norte-americanas

Haines & Pollock (1998), interessados no monitoramento de animais silvestres, consideraram a abordagem de cadastro duplo para estimar o total de ninhos bem-sucedidos (ou seja, que contém pelo menos um casal com um filhote) da ave símbolo e uma das mais conhecidas dos Estados Unidos, a águia norte-americana, que se encontra ameaçada de extinção. O monitoramento da espécie é importante pois a quantidade de ninhos serve como indicador para a qualidade do ecossistema, e até mesmo para a qualidade da água da região de estudo, uma vez que a espécie constrói ninhos em regiões com grande quantidade de água de boa qualidade. A águia norte-americana é muito sensível a toxinas e a presença destas no habitat da espécie é prejudicial dependendo do tipo de presa que capturam. Por outro lado, os filhotes que ainda não estão prontos para deixar os ninhos são bastante prejudicados. Em regiões onde são encontrados ninhos vazios, análises foram feitas e infelizmente nestes ninhos foram encontrados resíduos destas toxinas. Ovos mais frágeis e baixa produção de ovos são exemplos de informações obtidas a partir da análise dos ninhos desta espécie de ave.

O conhecimento de informações sobre a qualidade do ecossistema utilizando essa espécie de águia é baseado necesseriamente em várias variáveis, como mortalidade, quantidade de determinado tipo de toxina e número de ovos por ninho por exemplo. A necessidade de uso da abordagem de cadastro duplo neste caso foi importante porque dois tipos de cadastros estavam disponíveis:

- Um cadastro *específico*, que continha informações como a localização dos ninhos, por exemplo. Apesar de bastante eficiente, este tipo de cadastro pode ser desatualizado com facilidade, principalmente se forem levadas em consideração variáveis como tempo, uma vez que novos ninhos podem vir a ser construídos em épocas distintas.
- Um cadastro de área, que continha informações sobre os limites geográficos das regiões da espécie, possuindo a vantagem de fornecer completa cobertura da população de interesse e a desvantagem de proporcionar alto custo de seleção de elementos para amostras a serem coletadas.

Com o intuito de minimizar as deficiências individuais destes cadastros, a abordagem de cadastro duplo foi utilizada com o objetivo de estimar o total de ninhos de águia, e apresentou resultados satisfatórios. Neste estudo, sempre foi assumido que o cadastro de área estava completo, e o que era sempre atualizado era o cadastro tipo específico. Além disso, o uso desta abordagem revelou outra regra importante no comportamento da espécie: a desconsideração da teoria que estabelecia a permanência de um casal de aves em um ninho construído. Ficou comprovado que um casal de aves mantêm vários ninhos em seu território, revelando que a

variável *número de territórios ocupados* se tornou mais efetiva do que o número de ninhos bem-sucedidos.

2.3 Saúde

1. Pesquisa Nacional de Idosos (EUA)

Choudhry, Park & Li (2002) consideraram a abordagem de cadastro duplo a partir de informações da Pesquisa Nacional de Idosos realizada em 2000, conduzida para obter estimativas da população de idosos. A pesquisa foca, em especial, grupos populacionais de interesse do Departamento de Assistência ao Idoso (*Departament of Veteran Affairs* - DVA) dos Estados Unidos, de modo a estruturar o planejamento de programas destinados à serviços de assistência geral para idosos.

O DVA tinha interesse em obter estimativas dos grupos populacionais em questão (mulheres, afro-americanos, e latino-americanos). A população-alvo considerada na pesquisa foi todo o conjunto de idosos residentes nos Estados Unidos e em Porto Rico, e com base em critérios determinados pelo DVA, a população é então estratificada com base nos grupos populacionais de interesse. Existiam dois cadastros disponíveis:

- Um cadastro específico, com números de telefones de idosos identificados de acordo com Casady & Lepkowski (1991); Casady & Lepkowski (1993) e Potter *et al.* (1991).

- Um cadastro específico constituído pela combinação de duas fontes administrativas, a primeira com informações de idosos cadastrados nos serviços de assistência à saúde e a segunda com informações do registro nacional de seguridade social e pensões.

O primeiro cadastro, apesar de conter a quase totalidade das informações, em geral apresentava grande taxa de não-resposta ou de não-conclusão de uma entrevista, e alto custo para obtenção das informações, uma vez que se faz necessário um tempo elevado ao telefone para realização das entrevistas. Já o segundo cadastro, apesar de incompleto, não apresentava as deficiências do primeiro. Com isso, a ideia dos autores foi combinar ambos os cadastros de modo a alcançar uma maior precisão nas estimativas de interesse, com custo o mais reduzido possível. Ao utilizar a abordagem de cadastro duplo, foi verificada uma redução significativa do custo quando se comparou a abordagem com um plano amostral aplicado apenas no primeiro cadastro.

2. Pesquisa Nacional de Imunização (EUA)

Shin, Molinari & Wolter (2008) consideraram a abordagem de cadastro duplo na Pesquisa Nacional de Imunização (National Immunization Survey - NIS) realizada pelo Centro de Pesquisas de Opiniões Nacionais (National Opinion Research Center - NORC), de modo a fornecer informações para os centros de Controle e Prevensão de Doenças (Center for Disease Control and Prevention - CDC). Este tipo de pesquisa monitora a campanha de vacinação em crianças de 19 a 35 meses de vida. A cada ano, esta pesquisa realiza uma série de aproximadamente 24000 entrevistas em domícilios selecionados. Nesta pesquisa dois cadastros foram utilizados:

- Um cadastro de telefones registrados em um ou mais bancos de telefones do país;
- Um cadastro específico de domícilios que continham pelo menos uma criança.

De todos os telefones cadastrados, menos de 25% destes eram associados a domícilios e aproximadamente 68% dos telefones em funcionamento estavam presentes neste cadastro. Os autores encontraram evidência de que a aplicação da abordagem de cadastro duplo na pesquisa NIS representa uma alternativa de baixo custo quando comparada à pesquisa feita apenas no cadastro A, principalmente em situações em que é necessário utilizar um plano de amostragem estratificada com base no custo de seleção de unidades amostrais em cada estrato.

3 Contribuições desta tese

Quando uma ou mais variáveis auxiliares estão disponíveis em uma abordagem com um único cadastro, seu uso pode melhorar consideravelmente as estimativas dos parâmetros, através de estimadores do tipo regressão. Coelho (2007) mostrou que, como era de se esperar, o uso de estimadores do tipo regressão também pode melhorar consideravelmente estimativas em uma abordagem de cadastro duplo.

O objetivo desta tese de doutorado é apresentar estratégias de inferência utilizando formas gerais de estimadores do tipo regressão, adaptados à abordagem de cadastro duplo e compará-las de modo a informar qual destas estratégias tem melhor desempenho em situações onde a informação de uma ou mais variáveis auxiliares está disponível. Estas estratégias permitem propor diferentes tipos de estimadores de regressão, dentre os quais os propostos por Coelho (2007) são casos particulares. A investigação conduzida contempla a proposição de novos estimadores, bem como o estudo de suas propriedades estatísticas, com respeito a variância, consistência com respeito ao plano amostral, centralidade assintótica e erro quadrático médio assintótico, e distribuição assintótica. Um estudo de simulação foi realizado para avaliar os estimadores sob o caso específico de amostragem aleatória simples empregada nos dois cadastros, através do método de simulação de Monte Carlo.

4 Organização da tese

A presente tese está dividida em sete capítulos. Neste primeiro capítulo foi apresentada uma introdução sobre a abordagem de cadastro duplo. O capítulo 2 apresenta uma revisão das principais estratégias de estimação sob a abordagem de cadastro duplo já propostas na literatura, as quais não utilizam a informação de variáveis auxiliares no processo de estimação. O capítulo 3 apresenta, como contribuição original, estimadores do tipo regressão para a abordagem de cadastro duplo e suas propriedades de variância. O capítulo 4 apresenta, ainda como contribuição original desta tese, o estudo da consistência, centralidade assintótica, erro quadrático médio assintótico e distribuição assintótica dos estimadores propostos. O capítulo 5 apresenta casos particulares dos estimadores apresentados no capítulo 3 e suas propriedades de variância sob os planos de amostragem aleatória simples e estratificada. O capítulo 6 apresenta a avaliação computacional dos estimadores propostos, através do método de simulação de Monte Carlo. O ambiente de programação, análise de dados e gráficos `R` em sua versão 2.12.0 foi a plataforma computacional escolhida para a avaliação computacional. Por fim, no capítulo 7 são apresentadas as considerações finais deste trabalho.

CAPÍTULO 2

Inferência em uma abordagem de cadastro duplo

1 Estratégia de Hartley(1962)

Hartley (1962) avaliou a eficiência de uso da abordagem de cadastro duplo em um estudo realizado no *Statistical Laboratory* da *Iowa State University* sobre efeitos da industrialização na agricultura, sob a perspectiva de um plano de amostragem aleatória simples aplicado em cada cadastro. Os responsáveis pela pesquisa eram do Departamento de Economia e Sociologia da mesma universidade. Os cadastros utilizados nessa pesquisa foram os seguintes:

- Um cadastro do tipo área rural, utilizado para seleção de produtores rurais;
- Um cadastro do tipo específico, constituído por empregados de um empresa automotiva e que também eram produtores rurais.

O uso simultâneo destes cadastros para a cobertura da população-alvo forneceu uma experiência bem sucedida no que diz respeito ao custo-benefício na seleção de elementos. De modo a facilitar a compreensão do leitor a respeito do processo de inferência sob a abordagem de cadastro duplo, a descrição do estimador de Hartley (1962) será realizada de forma detalhada. Considere que dois cadastros A e B estejam disponíveis e que juntos forneçam completa cobertura para a população-alvo, ilustrado pela figura 1.3(a). Para a implementação da abordagem, duas condições são necessárias:

1. Todos os elementos da população-alvo devem ser identificados pela união dos cadastros;
2. Pode-se verificar se qualquer elemento de um cadastro pertence ou não ao outro.

Denotando por $U = \mathscr{A} \cup \mathscr{B}$ o conjunto de elementos da população-alvo e considerando a notação de domínios apresentada na figura 1.3(a), Hartley (1962) faz uma descrição dos possíveis cenários gerados por um levantamento amostral realizado sob a abordagem de cadastro duplo. Estes cenários dependem basicamente da disponibilidade de informações populacionais, como mostra a tabela (2.1) a seguir.

Tabela 2.1: Situações possíveis para uso da abordagem de Cadastro Duplo

Tipo de Informação Disponível	**CENÁRIOS**			
	1	**2**	**3**	**4**
Tamanho dos domínios e dos cadastros	Tamanhos dos domínios e cadastros conhecidos	Tamanhos dos domínios e cadastros conhecidos	Apenas os tamanhos dos cadastros conhecidos	Apenas as magnitudes relativas dos cadastros conhecidas
Possibilidade de alocação da amostra	Alocação de amostra aos domínios	Alocação de amostra aos cadastros	Alocação de amostra aos cadastros	Alocação de amostra aos cadastros

É possível notar que $U = a \cup b \cup ab$, como mostra a figura 1.3(a). Os cenários 1 e 2 ilustram a situação em que os tamanhos dos domínios são conhecidos, o que implica que os tamanhos populacionais nos cadastros também o são. O cenário 1 permite que a amostra seja alocada a cada domínio, o que ilustra o caso de um plano sob estratificação, enquanto que o cenário 2 admite a alocação da amostra apenas aos cadastros. Os cenários 3 e 4 só permitem que a alocação da amostra seja feita nos cadastros. No cenário 3, apenas os tamanhos populacionais nos cadastros são conhecidos, enquanto o cenário 4 apresenta como única informação disponível o tamanho relativo dos cadastros, fornecendo informações bastante restritivas, tornando possível apenas a estimação de médias populacionais. Assim, todo o desenvolvimento nesta tese será referente aos cenários 2 e 3. A tabela (2.2) fornece a notação utilizada para cálculos feitos a partir das amostras obtidas dos cadastros A e B. Nela, as quantidades $n_{ab(A)}$, $\hat{t}_{yab(A)}$ e $\tilde{y}_{ab(A)}$ são funções de elementos na amostra que pertencem ao domínio ab e que foram selecionados do cadastro A. As quantidades $n_{ab(B)}$, $\hat{t}^B_{yab}$, e $\tilde{y}^B_{ab}$ são definidas analogamente e estão relacionadas ao cadastro B.

Tabela 2.2: Notação utilizada na estratégia de Hartley

Quantidades de Interesse	CADASTRO		Domínio		
	A	B	a	b	ab
População	A	B	A	B	$A \cap B$
Tamanho da População	N_A	N_B	N_a	N_b	N_{ab}
Total populacional	t_{yA}	t_{yB}	t_{ya}	t_{yb}	t_{yab}
Média populacional	μ_{yA}	μ_{yB}	μ_{ya}	μ_{yb}	μ_{yab}
Amostra	S_A	S_B	S_A	S_B	S_{ab}
Tamanho da amostra	n_A	n_B	n_a	n_b	$n_{ab(A)}$ $n_{ab(B)}$
Estimador do Total	$\widehat{t}_{yA}$	$\widehat{t}_{yB}$	$\widehat{t}_{ya}$	$\widehat{t}_{yb}$	$\hat{t}^A_{yab}$ $\hat{t}^B_{yab}$
Média amostral	$\overline{y}_A$	$\overline{y}_B$	$\tilde{y}_a$	$\tilde{y}_b$	$\tilde{y}^A_{ab}$ $\tilde{y}^B_{ab}$

Considerando a implementação de um plano de amostragem aleatória simples em cada cadastro, por exemplo, a ideia de Hartley pode ser representada pela figura a seguir.

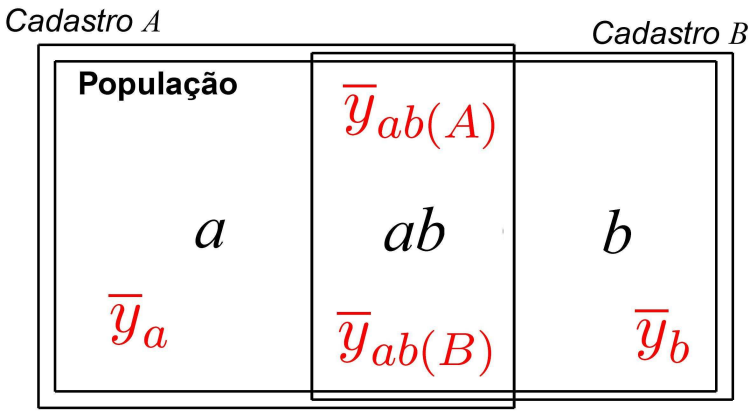


Figura 2.1: Quantidades amostrais geradas pela estratégia de Hartley

Nesta tese, a menos que seja dito o contrário, y_k será sempre considerado como o valor da variável de interesse y associada ao elemento k na população-alvo. O método de estimação desenvolvido por Hartley (1962), que considera o cenário 2, é proposto com base em uma variável y^*_k definida da seguinte forma:

$$y^*_k = \begin{cases} y_k, & \text{se} \quad k \in a \text{ ou } k \in b \\ py_k, & \text{se} \quad k \in ab \text{ e } k \in \mathscr{A} \\ (1-p)y_k, & \text{se} \quad k \in ab \text{ e } k \in \mathscr{B}, \end{cases}$$

onde p é uma constante de ponderação para identificação dos elementos de cada cadastro na população, tal que $0 \leq p \leq 1$. Dessa forma , é possível reescrever o total populacional, t_y, da

seguinte maneira:

$$\begin{aligned} t_y &= t_{ya} + t_{yb} + t_{yab} \\ &= t_{ya} + t_{yb} + (p + 1 - p)t_{yab} \\ &= t_{ya} + pt_{yab} + t_{yb} + (1-p)t_{yab} \\ &= \sum_{k \in \mathscr{A}} y^*_{kA} + \sum_{k \in \mathscr{B}} y^*_{kB} \\ &= t^*_{yA} + t^*_{yB}. \end{aligned} \tag{2.1}$$

Observando a estrutura da expressão (2.1), Hartley enfatiza o fato de que t_y pode ser estimado utilizando estimadores para cada um dos domínios a, b e ab. É possível também representar t_y ou a média populacional, μ, da seguinte forma:

$$\begin{aligned} t_y &= t_{yA} + t_{yB} - t_{yab} \\ &= t_{yA} + t_{yB} - \left(pt_{yab} + (1-p)t_{yab}\right) \end{aligned} \tag{2.2}$$

$$\mu = N^{-1} t_y = N^{-1}\left(t_{yA} + t_{yB} - t_{yab}\right), \tag{2.3}$$

onde $N = N_A + N_B - N_{ab}$. Nesta tese, o estimador de Hartley será motivado pelas expressões (2.2) e (2.3). Dessa forma, considerando $\hat{t}_{yA}$, $\hat{t}_{yB}$ e $\hat{t}_{yab}$ como estimadores do total para os cadastros A, B e para o domínio ab, respectivamente, e denotando por $\overline{y}_H$ o estimador de Hartley para μ, tem-se que

$$\overline{y}_H = N^{-1}\left(\hat{t}_{yA} + \hat{t}_{yB} - \hat{t}_{yab}\right). \tag{2.4}$$

Quando um plano de amostragem aleatória simples é empregado em cada cadastro (caso originalmente considerado em Hartley (1962)),

$$\hat{t}_{yA} = N_A \overline{y}_A\,, \qquad \hat{t}_{yB} = N_B \overline{y}_B \quad \text{e} \quad \hat{t}_{yab} = pN_{ab}\overline{y}_{ab(A)} + (1-p)N_{ab}\overline{y}_{ab(B)}.$$

Ao substituir estas quantidades em (2.4), a expressão do estimador de Hartley é

$$\overline{y}_H = N^{-1}\left[N_A\overline{y}_A + N_B\overline{y}_B - \left(pN_{ab}\overline{y}_{ab(A)} + (1-p)N_{ab}\overline{y}_{ab(B)}\right)\right]$$

$$= \quad w_A\overline{y}_A + w_B\overline{y}_B - \left(pw_{ab}\overline{y}_{ab(A)} + (1-p)w_{ab}\overline{y}_{ab(B)}\right), \tag{2.5}$$

onde $w_A = N_A/N$, $w_B = N_B/N$ e $w_{ab} = N_{ab}/N$. Em uma situação onde estimadores de Horvitz-Thompson são utilizados para estimar as quantidades populacionais nos cadastros e domínios, tem-se que

$$\hat{t}_{yA} = \sum_{k\in S_A}\frac{y_k}{\pi_k}\ , \qquad \hat{t}_{yB} = \sum_{k\in S_B}\frac{y_k}{\pi_k} \quad \text{e} \quad \hat{t}_{yab} = p\sum_{k\in S_A\cap ab}\frac{y_k}{\pi_k} + (1-p)\sum_{k\in S_B\cap ab}\frac{y_k}{\pi_k}.$$

Substituindo estes valores em (2.4), obtém-se

$$\begin{aligned}
\overline{y}_H &= N^{-1}\left[\sum_{k\in S_A}\frac{y_k}{\pi_k^A} + \sum_{k\in S_B}\frac{y_k}{\pi_k^B} - \left(p\sum_{\substack{k\in ab\\ k\in S_A}}\frac{y_k}{\pi_k^A} + (1-p)\sum_{\substack{k\in ab\\ k\in S_B}}\frac{y_k}{\pi_k^B}\right)\right] \\
&= N^{-1}\left[N_A\sum_{k\in S_A}\frac{y_k}{N_A\pi_k} + N_B\sum_{k\in S_B}\frac{y_k}{N_B\pi_k} - \left(pN_{ab}\sum_{\substack{k\in ab\\ k\in S_A}}\frac{y_k}{N_{ab}\pi_k} + (1-p)\sum_{\substack{k\in ab\\ k\in S_B}}\frac{y_k}{N_{ab}\pi_k}\right)\right] \\
&= w_A\overline{y}_{HT(A)} + w_B\overline{y}_{HT(B)} - \left(pw_{ab}\overline{y}^A_{HT(ab)} + (1-p)w_{ab}\overline{y}^B_{HT(ab)}\right),
\end{aligned} \tag{2.6}$$

onde

- π_k^A é a probabilidade de inclusão do elemento k do cadastro A na amostra;
- π_k^B é a probabilidade de inclusão do elemento k do cadastro B na amostra;
- $\overline{y}_{HT(A)}$ é o estimador de Horvitz-Thompson para a média populacional do cadastro A;
- $\overline{y}_{HT(B)}$ é o estimador de Horvitz-Thompson para a média populacional do cadastro B;
- $\overline{y}^A_{HT(ab)}$ é o estimador de Horvitz-Thompson para a média populacional do domínio ab, no cadastro A e
- $\overline{y}^B_{HT(ab)}$ é o estimador de Horvitz-Thompson para a média populacional do domínio ab, no cadastro B.

Nota-se que o estimador de Hartley para a média populacional μ é uma média ponderada entre os estimadores de Horvitz-Thompson para a média em cada cadastro e também para a interseção entre eles.

As variância aproximadas de $\overline{y}_H$ e dos outros estimadores que serão apresentados nesta tese podem ser expressas em termos matriciais na forma

$$\text{AVar}\left(\hat{\theta}\right) = \mathbf{d}^T \Sigma \mathbf{d}, \tag{2.7}$$

onde $\mathbf{d}^T$ é vetor coluna e Σ é a matriz de variâncias e covariâncias do estimador $\hat{\theta}$, diagonal em blocos, pois planos amostrais são aplicados independentemente em cada cadastro. Os elementos de Σ são representados pelas variâncias dos estimadores referentes a cadastros e domínios para cada estimador. Para o estimador de Hartley, tem-se que

$$\mathbf{d} = \begin{bmatrix} w_A \\ pw_{ab} \\ w_B \\ (1-p)w_{ab} \end{bmatrix} \quad \text{e} \quad \Sigma_H = \begin{pmatrix} \Sigma_A & \mathbf{0} \\ \mathbf{0} & \Sigma_B \end{pmatrix} = \begin{pmatrix} \sigma_A^2 & \sigma_{A;ab(A)} & 0 & 0 \\ \sigma_{A;ab(A)} & \sigma_{ab(A)}^2 & 0 & 0 \\ 0 & 0 & \sigma_B^2 & \sigma_{B;ab(B)} \\ 0 & 0 & \sigma_{B;ab(B)} & \sigma_{ab(B)}^2 \end{pmatrix}.$$

Os termos σ_A^2 e σ_B^2 representam as variâncias dos estimadores de Horvitz-Thompson para a média populacional nos cadastros A e B. Por exemplo,

$$\sigma_A^2 = \frac{1}{N_A^2} \sum_{k \in \mathscr{A}} \sum_{l \in \mathscr{A}} \Delta_{kl}^A \frac{y_k}{\pi_k^A} \frac{y_l}{\pi_l^A},$$

com $\Delta_{kl}^A = \pi_{kl}^A - \pi_k^A \pi_l^A$, sendo π_{kl}^A a probabilidade de inclusão de segunda ordem dos elementos k e l do cadastro A. $\sigma_{ab(A)}^2$ e $\sigma_{ab(B)}^2$ representam as variâncias dos estimadores para a média populacional do domínio ab nos cadastros A e B respectivamente, e $\sigma_{A;ab(A)}$ $\sigma_{B;ab(B)}$ representam as covariâncias entre os estimadores calculados dentro dos cadastros A e B respectivamente. Por exemplo,

$$\sigma_{A;ab(A)} = \text{Cov}\left(\overline{y}_{HT(A)}; \overline{y}_{HT(ab)}^A\right) = \frac{1}{N_A} \frac{1}{N_{ab}} \sum_{k \in \mathscr{A}} \sum_{l \in ab} \Delta_{kl}^A \frac{y_k}{\pi_k^A} \frac{y_l}{\pi_l^A}.$$

Σ_H é uma matriz diagonal em blocos, pois planos amostrais são aplicados independentemente em cada cadastro. Hartley mostrou ainda que o valor de p que minimiza a expressão (2.7) é dado por

$$p = \frac{N_b \left(f^B\right)^{-1} w_B^{ab} \sigma_{ab(B)}^2}{N_a \left(f^A\right)^{-1} w_A^{ab} \sigma_{ab(A)}^2 + N_b \left(f^B\right)^{-1} w_B^{ab} \sigma_{ab(B)}^2}, \tag{2.8}$$

onde $w_A^{ab} = N_{ab}/N_A$, $w_B^{ab} = N_{ab}/N_B$, $f_A = n_A/N_A$ e $f_B = n_B/N_B$. Quando o valor de ao menos uma das covariâncias apresentadas for muito alto, o valor ótimo de p (2.8) pode ficar fora do intervalo $[0,1]$. Um estimador não-viesado para Var$(\overline{y}_H)$ pode ser obtido ao substituir as

quantidades populacionais pelos respectivos estimadores de Horvitz-Thompson.

Hartley (1962) mostrou que a abordagem de cadastro duplo é mais eficiente do que a abordagem de selecionar uma amostra de um único cadastro, para o caso em que $B \subset A$. O critério escolhido para a comparação foi a razão

$$\text{Var}\left(\overline{y}_H\right)/\text{Var}\left(\overline{y}'\right)$$

onde $\overline{y}'$ é o estimador para a média populacional do cadastro A e $\overline{y}_H$ é o estimador de Hartley para μ sob a abordagem de cadastro duplo. Considerando c_A e c_B como os custos de seleção de elementos dos cadastros A e B respectivamente, e σ_B^2 e σ_a^2 como as variâncias dos estimadores obtidos no cadastro B e domínio a respectivamente, as tabelas a seguir apresentam a redução da variância do estimador de Hartley em função de c_B/c_A e N_B/N_A, para valores fixos de σ_B^2/σ_a^2.

Tabela 2.3: Valores da razão $\text{Var}\left(\overline{y}_H\right)/\text{Var}\left(\overline{y}'\right)$ quando $\sigma_B^2/\sigma_a^2 = 16$

c_B/c_A	N_B/N						
	0.5	0.6	0.7	0.8	0.9	0.95	1
0.01	0.096	0.076	0.059	0.045	0.031	0.024	0.010
0.05	0.154	0.134	0.118	0.102	0.086	0.075	0.050
0.10	0.206	0.188	0.174	0.160	0.143	0.131	0.100
0.20	0.288	0.278	0.269	0.261	0.248	0.237	0.200
0.30	0.359	0.356	0.355	0.353	0.347	0.338	0.300
0.40	0.423	0.428	0.435	0.440	0.441	0.436	0.400
0.50	0.483	0.496	0.510	0.524	0.533	0.532	0.500
1.00	0.096	0.784	0.836	0.889	0.944	0.972	1.000

Tabela 2.4: Valores da razão $\text{Var}\left(\overline{y}_H\right)/\text{Var}\left(\overline{y}'\right)$ quando $\sigma_B^2/\sigma_a^2 = 4$

c_B/c_A	N_B/N						
	0.5	0.6	0.7	0.8	0.9	0.95	1
0.01	0.259	0.201	0.152	0.108	0.066	0.044	0.010
0.05	0.340	0.284	0.234	0.186	0.137	0.107	0.050
0.10	0.404	0.352	0.304	0.257	0.205	0.172	0.100
0.20	0.500	0.456	0.415	0.372	0.322	0.287	0.200
0.30	0.576	0.540	0.507	0.472	0.426	0.393	0.300
0.40	0.640	0.613	0.588	0.561	0.523	0.493	0.400
0.50	0.696	0.678	0.661	0.642	0.614	0.589	0.500
1.00	0.900	0.914	0.932	0.953	0.976	0.988	1.000

Por exemplo, para $\sigma_B^2/\sigma_a^2 = 4$, na situação em que $c_B/c_A = 0.30$ e $N_B/N = 0.7$, observa-se que o uso da abordagem de cadastro duplo comparada à abordagem de uso de apenas um cadastro apresenta um valor da razão igual a 0.507.

Exemplo de aplicação da abordagem de Hartley

Nesta seção será apresentado um exemplo de estimação através da estratégia de Hartley. Considere uma população com $N = 169587$ elementos gerados aleatoriamente. De modo a exemplificar o uso da abordagem, foram considerados cadastros A e B, de acordo com a situação descrita pela figura 1.3(a), da seguinte forma:

1. Cadastro A: constituído do elemento 1 ao elemento 84793, com $N_A = 84793$;
2. Cadastro B: constituído do elemento 74793 ao elemento 169587, com $N_B = 84794$.
3. Domínio ab: constituído por $N_{ab} = 10000$ elementos, comuns aos cadastros gerados.

O parâmetro de interesse é a média populacional $\mu = 21.1$. Através do plano de amostragem aleatória simples, duas amostras independentes, de tamanho 1000 (cadastro A) e 800 (cadastro B) foram selecionadas, obtendo-se

$$\overline{y}_A = 22.56\,, \qquad \overline{y}_a = 21.35\,, \qquad \overline{y}_{ab(A)} = 31.91$$

$$\overline{y}_B = 22.57\,, \qquad \overline{y}_b = 24.04\,, \qquad \overline{y}_{ab(B)} = 12.67$$

$$w_{ab} = 10000/169587 = 0.0589$$

Logo, basta substituir estas quantidades na forma do estimador de Hartley, bem como na forma do valor ótimo de p, para se obter um estimador para a média populacional.

2 Estratégia de Fuller & Burmeister(1972)

A estratégia de Hartley é utilizada nas situações em que N_{ab} é conhecido (cenário 2, tabela 2.1.). Quando isto não acontece, o seguinte estimador proposto por Fuller & Burmeister (1972) pode ser utilizado. De uma forma geral, utilizando estimadores do tipo Horvitz-Thompson, tem-se

$$\begin{aligned}
\overline{y}_{FB} &= N^{-1}\left\{\hat{t}_{ya} + \hat{t}_{yb} + \beta_1\hat{t}^A_{yab} + (1-\beta_1)\hat{t}^B_{yab} + \beta_2\left(\hat{N}^A_{ab} - \hat{N}^B_{ab}\right)\right\} \\
&= \hat{w}_a\overline{y}_{HT(a)} + \hat{w}_b\overline{y}_{HT(b)} + \beta_1\hat{w}_{ab(A)}\overline{y}_{HTab(A)} + \hat{w}_{ab(B)}(1-\beta_1)\overline{y}_{HTab(B)} + \\
&+ N^{-1}\beta_2\left(\hat{N}^A_{ab} - \hat{N}^B_{ab}\right), \qquad (2.9)
\end{aligned}$$

em que

- $\hat{w}_a = \frac{\hat{N}_a}{N}$, $\hat{w}_b = \frac{\hat{N}_b}{N}$, $\hat{w}_{ab(A)} = \frac{\hat{N}^A_{ab}}{N}$ e $\hat{w}_{ab(B)} = \frac{\hat{N}^B_{ab}}{N}$, em que, por exemplo, $\hat{N}_a = \sum_{k \in S_a} (1/\pi^A_k)$;
- $\hat{t}_{ya}$ é o estimador de Horvitz-Thompson para o total populacional do domínio a;
- $\hat{t}_{yb}$ é o estimador de Horvitz-Thompson para o total populacional do domínio b;
- $\hat{t}^A_{yab}$ é o estimador de Horvitz-Thompson para o total populacional do domínio ab no cadastro A;
- $\hat{t}^B_{yab}$ é o estimador de Horvitz-Thompson para o total populacional do domínio ab no cadastro B;
- $\hat{N}^A_{ab}$ e $\hat{N}^B_{ab}$ são estimadores para o tamanho populacional do domínio ab em cada cadastro, respectivamente.
- β_1 e β_2 são constantes de ponderação, escolhidas como os valores que minimizam a variância do estimador de Fuller & Burmeister, que será apresentada mais adiante.

O estimador de Fuller & Burmeister (1972) é função de estimadores da forma $\hat{Q}_= \hat{w}_d \overline{y}_d$. Através do método de linearização de Taylor é possível obter um pseudo-estimador para que seja possível o cálculo da variância aproximada de $\hat{Q}$. Tem-se que

$$Q_L = w\mu + \sum_{i=1}^{2} b_i(\hat{\theta}_i - \theta_i), \text{ onde:}$$

$$\begin{aligned} Q &= w_d\mu_d, \qquad \hat{\theta}_1 = \hat{w}_d \qquad \hat{\theta}_2 = \hat{\mu}_d \\ b_1 &= \left.\frac{\partial \hat{Q}}{\partial \hat{w}}\right|_{(\hat{w},\hat{\mu})=(w,\mu)} = \mu \quad \text{e} \\ b_2 &= \left.\frac{\partial \hat{Q}}{\partial \hat{\mu}}\right|_{(\hat{w},\hat{\mu})=(w,\mu)} = w. \end{aligned}$$

Assim,

$$\begin{aligned} \hat{Q}_{(L)} &= Q + \mu_d(\hat{w}_d - w_d) + w_d(\hat{\mu}_d - \mu_d) \\ &= Q + \mu_d\hat{w}_d - \mu w_d + w_d\hat{\mu}_d - w_d\mu_d \\ &= \mu_d\hat{w}_d + w_d\hat{\mu}_d - w_d\mu_d, \end{aligned}$$

onde $w_d = N_d/N$ e μ_d é a média populacional do domínio d. O resultado obtido implica um pseudo-estimador para $\overline{y}_{FB}$, dado por

$$\begin{aligned}\overline{y}_{FB(L)} &= (\mu_a\hat{w}_a + w_a\hat{\mu}_a - w_a\mu_a) + (\mu_b\hat{w}_b + w_b\hat{\mu}_b - w_b\mu_b) \\ &+ \beta_1(\mu_{ab(A)}\hat{w}_{ab(A)} + w_{ab(A)}\hat{\mu}_{ab(A)} - w_{ab(A)}\mu_{ab(A)}) \\ &+ (1-\beta_1)\left(\mu_{ab(B)}\hat{w}_{ab(B)} + w_{ab(B)}\hat{\mu}_{ab(B)} - w_{ab(B)}\mu_{ab(B)}\right) + N^{-1}\beta_2\left(\hat{N}^A_{ab} - \hat{N}^B_{ab}\right),\end{aligned}$$

Considerando a expressão (2.7), a variância do estimador de Fuller & Burmeister é dada em função de

$$\mathbf{d} = \begin{bmatrix} w_a \\ w_{ab(A)}\beta_1 \\ w_b \\ w_{ab(B)}(1-\beta_1) \\ N^{-1}\beta_2 \\ N^{-1}\beta_2 \end{bmatrix} \tag{2.10}$$

e

$$\Sigma_{FB} = \begin{pmatrix} \sigma^2_a & \sigma_{a;ab(A)} & 0 & 0 & 0 & 0 \\ \sigma_{a;ab(A)} & \sigma^2_{ab(A)} & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2_b & \sigma_{b;ab(B)} & 0 & 0 \\ 0 & 0 & \sigma_{b;ab(B)} & \sigma^2_{ab(B)} & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_{N_{A(ab)}} & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2_{N_{B(ab)}} \end{pmatrix} \tag{2.11}$$

Na matriz Σ_{FB}, por exemplo, $\sigma^2_{A(a)} = \text{Var}(\overline{y}_{HT(a)})$, e os demais termos são definidos como anteriormente. Sob um plano de amostragem aleatória simples, por exemplo, o estimador de Fuller & Burmeister (1972) é dado por

$$\overline{y}_{FB} = \hat{w}_a\overline{y}_a + \hat{w}_b\overline{y}_b + \hat{w}_{ab,s}\overline{y}_{ab}, \tag{2.12}$$

onde

$$\overline{y}_{ab} = \left[n_{ab(A)}(1-f_B) + n_{ab(B)}(1-f_A)\right]^{-1}\left[(1-f_B)\,n_{ab(B)}\overline{y}_{ab(A)} + (1-f_A)\overline{y}_{ab(B)}\right],$$

$$w_{ab,s} = \hat{N}_{ab,s}/N,\ f_A = n_A/N_A \text{ e } f_B = n_B/N_B$$

Na expressão, $\hat{N}_{ab,s}$ é definida como a menor raiz da seguinte equação quadrática em x

$$\begin{aligned}&(n_A g_B + n_B g_A)\, x^2 - \left(n_A N_B g_B + n_B N_A g_A + n_{ab(A)} N_A g_B + n_{ab(B)} N_B g_A\right) x \\ &+ \left(n_{ab(A)} g_B + n_{ab(B)} g_A\right) N_A N_B = 0.\end{aligned} \tag{2.13}$$

Fuller & Burmeister provaram que a variância de $\hat{N}_{ab,s}$ é dada por

$$\text{Var}\left(\hat{N}_{ab,s}\right) = \frac{N_a N_b N_{ab} g_A g_B}{n_A N_b g_B + n_B N_a g_A} + \mathbf{O}(1) \quad \text{e} \quad E(\hat{N}_{ab,s}) = N_{ab} + \mathbf{O}\left(\frac{1}{n}\right),$$

onde $\mathbf{O}(.)$ indica ordem de magnitude, como usado em análise real[1].

3 Estratégia de Bankier, Lepkowski & Groves(1986)

A estratégia de estimação sob a abordagem de cadastro duplo adotada por Bankier, Lepkowski & Groves (1986), descrita aqui simplesmente por estratégia BLG, atribui pesos, às unidades amostrais de A e B. A ideia dos autores foi adotar uma abordagem que envolvesse apenas quantidades amostrais referentes aos cadastros A e B, como mostra a figura (2.2) a seguir.

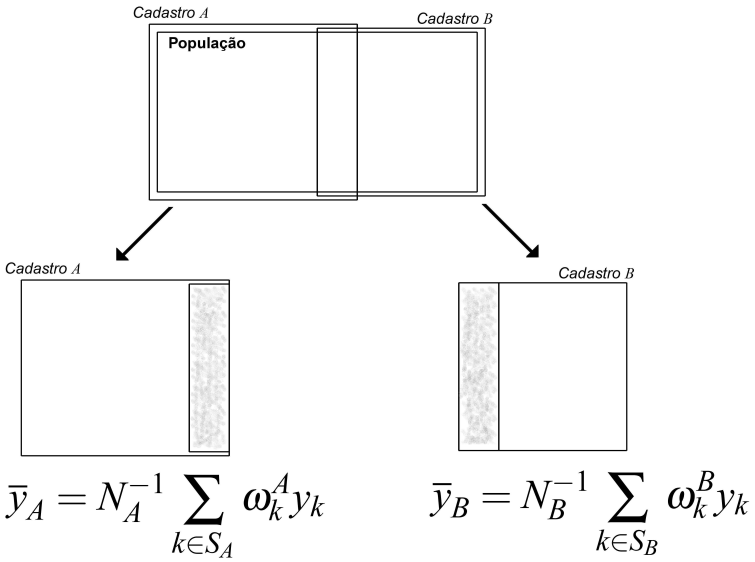


Figura 2.2: Quantidades amostrais geradas pela estratégia BLG

[1]Para um maior aprofundamento, o leitor pode consultar Fuller (1996). Uma breve revisão sobre o tema é apresentada no apêndice A.

Denotando por ω_k^C o peso associado a um cadastro C, tem-se que

$$\omega_k^C = \begin{cases} \dfrac{1}{\pi_k^A}, & k \in a \\ \dfrac{1}{\pi_k^B}, & k \in b \\ \dfrac{1}{\pi_k^A + \pi_k^B}, & k \in ab \end{cases}$$

Quando há o interesse em estimar por exemplo uma média populacional denotada por μ, tem-se que o estimador BLG é dado por

$$\begin{aligned} \overline{y}_{BLG} &= N^{-1}\left(\sum_{k\in S_A}\omega_k^A y_k + \sum_{k\in S_B}\omega_k^B y_k\right) = w_A\frac{\sum\limits_{k\in S_A}\omega_k^A y_k}{N_A} + w_B\frac{\sum\limits_{k\in S_B}\omega_k^B y_k}{N_B} \\ &= N^{-1}\left(\sum_{k\in a}\omega_k y_k + \sum_{k\in b}\omega_k y_k + \sum_{k\in ab}\omega_k y_k\right) = w_a\overline{y}_a + w_b\overline{y}_b + w_{ab}\overline{y}_{ab} \end{aligned} \tag{2.14}$$

com $w_A = N_A/N$, $w_B = N_B/N$, $w_a = N_a/N$, $w_b = N_b/N$ e $w_{ab} = N_{ab}/N$. A variância de $\overline{y}_{BLG}$ é obtida a partir da expressão (2.7), onde

$$\mathbf{d} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \text{e} \qquad \Sigma_{BLG} = \begin{pmatrix} \sigma_{yA}^2 & 0 \\ 0 & \sigma_{yB}^2 \end{pmatrix}$$

Na matriz Σ_{BLG}, $\sigma_{yA}^2 = \text{Var}\left(\sum\limits_{k\in\mathscr{A}}\omega_k^A y_k\right)$ e $\sigma_{yB}^2 = \text{Var}\left(\sum\limits_{k\in\mathscr{B}}\omega_k^B y_k\right)$. É possível obter um estimador não-viesado para a variância do estimador de Bankier, substituindo em Σ_{BLG} as quantidades populacionais pelos seus respectivos estimadores de Horvitz-Thompson.

A partir de (2.14) Bankier, Lepkowski & Groves (1986) propuseram ainda um estimador iterativo chamado *raking ratio estimator*, dado por

$$\overline{y}_{BLG(r)} = N^{-1}\left(\sum_{k\in S_A}\omega_{ka}^{(r)} y_k + \sum_{k\in S_B}\omega_{kb}^{(r)} y_k + \sum_{k\in ab}\omega_{kab}^{(r)} y_k\right). \tag{2.15}$$

A ideia é que se façam r ajustes à constante ω_k, onde

$$\omega_{ka}^{(r)} = \frac{N_a}{\hat{N}^{A(r-1)}}\omega_{ka}^{(r-1)}, \qquad \omega_{kb}^{(r)} = \omega_{kb}^{(r-1)}, \qquad \omega_{kab}^{(r)} = \frac{N_a}{\hat{N}^{A(r-1)}}\omega_{kab}^{(r-1)} \quad \text{para } r = 1,3,5,\ldots$$

$$\omega_{ka}^{(r)} = \omega_{ka}^{(r-1)}, \qquad \omega_{kb}^{(r)} = \frac{N_b}{\hat{N}_\nu^{B(r-1)}}\omega_{kb}^{(r-1)}, \qquad \omega_{kab}^{(r)} = \frac{N_b}{\hat{N}_\nu^{B(r-1)}}\omega_{kab}^{(r-1)} \quad \text{para } r = 2,4,6,\ldots$$

$$\omega_{ka}^{(0)} = \frac{1}{\pi_k^A}, \qquad \omega_{kb}^{(0)} = \frac{1}{\pi_k^B}, \qquad \omega_{kab}^{(0)} = \frac{1}{\left(\pi_k^A + \pi_k^B\right)},$$

$$\hat{N}_{A(r)} = \omega_{ka}^{(r)}N_a + \omega_{kab}^{(r)}N_{ab} \qquad \text{e} \qquad \hat{N}_{B(r)} = \omega_{kb}^{(r)}N_b + \omega_{kab}^{(r)}n_{ab}.$$

O estimador $\overline{y}_{BLG(r)}$ é obtido fazendo $\overline{y}_{BLG(0)} = \overline{y}_{BLG}$ e aplicando r ajustes com respeito aos domínios, alternadamente. Como para um determinado valor de r as expressões para a variância do estimador $\overline{y}_{BLG(r)}$ são bastante complexas (Bankier, Lepkowski & Groves (1986) ; Brackstone & Rao (1979)), Skinner (1991) provou que quando o número de ajustes é suficientemente grande, ou seja, quando $r \to \infty$, tem-se que

$$\overline{y}_{BLG(\infty)} = N^{-1}\left\{\left(N_A - \tilde{N}_{ab}\right)\overline{y}_a + \tilde{N}_{ab}\overline{y}_{ab} + \left(N_B - \tilde{N}_{ab}\right)\overline{y}_b\right\}, \tag{2.16}$$

Logo, a variância aproximada do estimador raking é dada por

$$\begin{aligned}\text{AVar}\left(\overline{y}_{BLG(\infty)}\right) &= \left(\frac{1}{\pi_k^A}\right)^{-1}N_a\sigma_a^2 + \left(\frac{1}{\pi_k^B}\right)^{-1}N_b\sigma_b^2 + \left(\pi_k^A + \pi_k^B\right)^{-1}N_{ab}\sigma_{ab}^2 \\ &+ \left(\mu_a + \mu_b - \mu_{ab}\right)^2\frac{N_{ab}N_aN_b}{n_AN_b + n_BN_a}\left(1 + \lambda^2\right),\end{aligned} \tag{2.17}$$

onde μ_a, μ_b e μ_{ab} são as médias populacionais de cada domínio e σ_a^2, σ_b^2 e σ_{ab}^2 são as variâncias populacionais de cada domínio, e

$$\lambda^2 = \frac{N_aN_b(N_{ab})^2\left[\left(\pi_k^A\right)^2N_A - \left(\pi_k^B\right)^2N_B\right]^2}{N_AN_B\pi_k^A\pi_k^B\left(\pi_k^A + \pi_k^B\right)^2\left((N_{ab})^2 - N_AN_B\right)^2}.$$

4 Estratégia de Máxima Pseudo-Verossimilhança

Skinner & Rao (1996) apresentaram o estimador de máxima pseudo-verossimilhança (*pseudo maximum likelihood* - PML), considerando a situação descrita pelo cenário 3. Esta estratégia é considerada pois o estimador de Fuller & Burmeister não pode ser aplicado diretamente em planos amostrais complexos, por não ser consistente para o parâmetro de interesse. Dessa forma, a estratégia de máxima pseudo-verossimilhança considera uma modificação no estimador de Fuller & Burmeister, de modo a validar a propriedade de consistência. De um modo geral, utilizando novamente estimadores do tipo Horvitz-Thompson para obter estimadores nos domínios, tem-se que:

$$\begin{aligned}\overline{y}_{PML} &= N^{-1}\left[\left(N_A-\hat{N}_{ab,PML}\right)\overline{y}_{HT(a)}+\hat{N}_{ab,PML}\overline{y}_{ab(*)}+\left(N_B-\hat{N}_{ab,PML}\right)\overline{y}_{HT(b)}\right]\\ &= w_{a(*)}\overline{y}_{HT(a)}+w_{ab(*)}\overline{y}_{ab(*)}+w_{b(*)}\overline{y}_{HT(b)} \qquad (2.18)\end{aligned}$$

onde

$$\begin{aligned}\overline{y}_{ab(*)} &= \frac{\left[\pi^A\hat{N}^A_{ab}\overline{y}_{ab(A)}+\pi^B\hat{N}^B_{ab}\overline{y}_{ab(B)}\right]}{\left[\pi^A\hat{N}^A_{ab}+\pi^B\hat{N}^B_{ab}\right]}\\ w_{a(*)} &= \left(N_A-\hat{N}_{ab,PML}\right)/N \qquad w_{b(*)} = \left(N_B-\hat{N}_{ab,PML}\right)/N\\ w_{ab(*)} &= \hat{N}_{ab,PML}/N\\ \hat{N}^A_{ab} &= \sum_{k\in ab}\frac{1}{\pi^A} \qquad \hat{N}^B_{ab} = \sum_{k\in ab}\frac{1}{\pi^B}.\end{aligned}$$

Com $\pi^A=\pi^A_k$ e $\pi^B=\pi^B_k$, para todo $k\in\mathscr{A}$ e $k\in\mathscr{B}$ respectivamente. $\hat{N}_{ab,PML}$ é a menor raiz da equação quadrática

$$ax^2+bx+c=0, \qquad (2.19)$$

com

$$\begin{aligned}a &= n_A+n_B\,,\\ b &= n_AN_B+n_BN_A+n_A\hat{N}^A_{ab}+n_B\hat{N}^B_{ab};,\end{aligned}$$

$$c \quad = \quad n_A \hat{N}_{ab}^A N_B + n_B \hat{N}_{ab}^B N_A.$$

Como $b^2 - 4ac = \left(n_A N_B + n_B N_A + n_A \hat{N}_{ab}^A + n_B \hat{N}_{ab}^B\right)^2 + 4 n_A n_B \left(N_A - \hat{N}_{ab}^A\right)\left(N_B - \hat{N}_{ab}^B\right)$, (2.19) não possui raízes complexas, pois $\hat{N}_{ab}^A \leq N_A$ e $\hat{N}_{ab}^B \leq N_B$. Além disso, Skinner & Rao (1996) mostraram que

$$\text{AVar}\left(\hat{N}_{ab,PML}\right) = \left(\frac{N^2}{n}\right)\left[\alpha^2 \sigma_{ab(A)}^2 + (1-\alpha)^2 \sigma_{ab(B)}^2\right],$$

onde $N = N_A + N_B - N_{ab}$, $N = n_A + n_B - n_{ab}$, $n_{ab} = n_{ab(A)} + n_{ab(B)}$ e α é uma constante tal que $0 \leq \alpha \leq 1$, cujo valor que minimiza $\text{AVar}\left(\hat{N}_{ab,PML}\right)$ é dado por

$$\alpha = \frac{\sigma_{ab(B)}^2}{\sigma_{ab(A)}^2 + \sigma_{ab(B)}^2}.$$

Para cálculo da variância do estimador PML utilizando o método de linearização de Taylor, são necessários desenvolvimentos apresentados anteriormente para o estimador de Fuller & Burmeister. Além destes desenvolvimentos é necessário utilizar o método de linearização de Taylor em $\overline{y}_{ab(*)}$. Logo,

$$\begin{aligned}
\overline{y}_{ab*} &\approx \frac{\pi^A N_{ab}\mu_{ab} + \pi^B N_{ab}\mu_{ab}}{\pi^A \hat{N}_{ab}^A + \pi^B \hat{N}_{ab}^B} + \frac{1}{\pi^A N_{ab} + \pi^B N_{ab}} \left\{ \pi^A \hat{N}_{ab}^A \overline{y}_{ab(A)} + \pi^B \hat{N}_{ab}^B \overline{y}_{ab(B)} \right. \\
&- \left. \left(\frac{\pi^A N_{ab}\mu_{ab} + \pi^B N_{ab}\mu_{ab}}{\pi^A N_{ab} + \pi^B N_{ab}}\right)\left(\pi^A \hat{N}_{ab}^A + \pi^B \hat{N}_{ab}^B\right) \right\} \\
&= \mu_{ab} + \frac{(N_{ab})^{-1}}{\pi^A + \pi^B} \left\{ \pi^A \hat{N}_{ab}^A \overline{y}_{ab(A)} + \pi^B \hat{N}_{ab}^B \overline{y}_{ab(B)} - \mu_{ab}\left(\pi^A \hat{N}_{ab}^A + \pi^B \hat{N}_{ab}^B\right) \right\} \\
&\approx \mu_{ab} + \frac{(N_{ab})^{-1}}{\pi^A + \pi^B} \left\{ \pi^A \left(\hat{N}_{ab}^A \mu_{ab} + N_{ab}\overline{y}_{ab(A)} - N_{ab}\mu_{ab}\right) + \pi^B \left(\hat{N}_{ab}^B \mu_{ab} + N_{ab}\overline{y}_{ab(B)} - N_{ab}\mu_{ab}\right) - \right. \\
&- \left. \mu_{ab}\pi^A \hat{N}_{ab}^A - \mu_{ab}\hat{N}_{ab}^B \pi^B \right\} \\
&= \mu_{ab} + \frac{(N_{ab})^{-1}}{\pi^A + \pi^B} \left\{ N_{ab}\pi^A \overline{y}_{ab(A)} - \pi^A N_{ab}\mu_{ab} + N_{ab}\pi^B \overline{y}_{ab(B)} - \pi^B N_{ab}\mu_{ab} \right\} \\
&= \mu_{ab} + \left(\frac{\pi^A}{\pi^A + \pi^B}\right)\overline{y}_{ab(A)} + \left(\frac{\pi^B}{\pi^A + \pi^B}\right)\overline{y}_{ab(B)} - \left(\frac{\pi^A + \pi^B}{\pi^A + \pi^B}\right)\mu_{ab} \\
&= \left(\frac{\pi^A}{\pi^A + \pi^B}\right)\overline{y}_{ab(A)} + \left(\frac{\pi^B}{\pi^A + \pi^B}\right)\overline{y}_{ab(B)}
\end{aligned}$$

Logo, um pseudo-estimador para o estimador PML é dado por

$$\begin{aligned}\overline{y}_{PML(L)} &\approx w_a\overline{y}_a + w_{a*}\mu_a - w_a\mu_a + w_b\overline{y}_b + w_{b*}\mu_b - w_b\mu_b + \\ &+ w_{ab}\left\{\left(\frac{\pi^A}{\pi^A+\pi^B}\right)\overline{y}_{ab(A)} + \left(\frac{\pi^B}{\pi^A+\pi^B}\right)\overline{y}_{ab(B)}\right\} \\ &+ w_{ab*}\mu_{ab} - w_{ab}\mu_{ab}. \end{aligned} \tag{2.20}$$

Dessa forma, baseando-se na expressão (2.7), a variância aproximada do estimador PML é dada em função das quantidades

$\mathbf{d}^T = [w_a \;\; w_b \;\; w_{ab} \;\; w_{ab} \;\; \mu_a/N \;\; \mu_b/N \;\; \mu_{ab}/N]$ e

$$\Sigma_{PML} = \begin{pmatrix} \sigma^2_{yA} & \sigma_{yA;yab(A)} & 0 & 0 & 0 & 0 & 0 \\ \sigma_{yA;yab(A)} & \sigma^2_{yab(A)} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2_{yB} & \sigma_{yB;yab(B)} & 0 & 0 & 0 \\ 0 & 0 & \sigma_{yB;yab(B)} & \sigma^2_{yab(B)} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_{N_{a,PML}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2_{N_{b,PML}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2_{N_{ab,PML}} \end{pmatrix} \tag{2.21}$$

Foram revisadas quatro estratégias de estimação sob a abordagem de cadastro duplo, sob os cenários 2 e 3. A partir da forma dos estimadores e de suas respectivas variâncias, é possível tirar as seguintes conclusões:

1. A partir da forma da variância do estimador $\overline{y}_{FB}$, quando $\beta_1 = \gamma$ e $\beta_2 = (\alpha - \gamma)\,\mu_{ab} + (1-\alpha)\mu^a + \alpha\mu^b$, onde $\alpha = N_a N_{ab}/(N_a N_b + N_b N_a)$ e $\gamma = N_a N^B/(N_a N_b + N_b N_a)$, $\mathbf{d}^T$ será equivalente à forma apresentada para o estimador PML;
2. Sob a condição comentada no item 1, tem-se que $\text{Var}(\overline{y}_{PML}) \geq \text{Var}(\overline{y}_{FB})$;
3. Considerando as condições 1 e 2 e o estimador $\overline{y}_{FB}$, quando $\beta_2 = 0$, tem-se que $\overline{y}_{FB} \approx \overline{y}_H$, e portanto $\text{Var}(\overline{y}_{FB}) \geq \text{Var}(\overline{y}_H)$. Dessa forma, tem-se que $\text{Var}(\overline{y}_{PML}) \geq \text{Var}(\overline{y}_H)$;
4. Caso os valores de β_1 e β_2 sejam escolhidos de outra forma, o estimador sob a estratégia PML não será necessariamente mais eficiente que o estimador sob a estratégia de Hartley.

As estratégias revisadas não utilizam informações de possíveis variáveis auxiliares no processo de estimação. A estratégia apresentada a seguir representa uma forma de fazer uso de informações auxiliares.

5 Estratégia de Singh & Wu (2003)

No contexto de estimadores do tipo regressão, Singh & Wu (2003) apresentaram uma estratégia de estimação sob abordagem de cadastro duplo baseada em um estimador intuitivo, dado por

$$\overline{y} = w_a\overline{y}_{Ra} + w_b\overline{y}_{Rb} + \frac{1}{2}\left(w_{ab}\overline{y}_{Rab(A)} + w_{ab}\overline{y}_{Rab(B)}\right), \tag{2.22}$$

onde $\overline{y}_{Ra}$, $\overline{y}_{Rb}$, $\overline{y}_{Rab(A)}$ e $\overline{y}_{Rab(B)}$ são os estimadores regressão para as médias dos domínios a, b e ab e que serão apresentados no próximo capítulo, $w_a = N_a/N$, $w_b = N_b/N$, $w_{ab} = N_{ab}/N$ e $N = N_A + N_B - N_{ab}$. Este estimador foi proposto para situações em que pode ocorrer $n_{ab(A)} = n_{ab(B)} = 0$, não sendo necessário o cálculo de valores ótimos para minimização da sua variância. Esta estratégia não envolve obtenção de estimativas por mínimos quadrados, pois calcula os coeficientes de regressão através de métodos de calibração, e portanto não será explorada no desenvolvimento desta tese. A partir deste estimador, foi proposto o estimador regressão de calibração (RC), dado por

$$\overline{y}_{RC} = w_a\overline{y}_{RCa} + w_b\overline{y}_{RCb} + \zeta_A\left(w_{ab}\overline{y}_{RCab(A)}\right) + \zeta_B\left(w_{ab}\overline{y}_{RCab(B)}\right), \tag{2.23}$$

onde ζ_A e ζ_B são coeficientes de ajuste.

CAPÍTULO 3

Estimadores do tipo regressão em uma abordagem de cadastro duplo

1 Introdução

O estimador regressão para o caso em que q variáveis auxiliares estão disponíveis é visto em várias referências. Para um maior aprofundamento, o leitor pode consultar Särndal, Swensson & Wretman (2003). Neste tipo de situação, admite-se que o vetor de informações auxiliares, associado ao elemento k da população, é conhecido e denotado

$$\mathbf{x}_k = \left(x_{1k}, x_{2k}, \ldots, x_{qk}\right).$$

Considere, por exemplo, a estimação de um total populacional

$$t_y = \sum_{k \in U} y_k. \tag{3.1}$$

Defina as seguintes quantidades:

$$\check{y}_k = \frac{y_k}{\pi_k}, \qquad \hat{t}_{y\pi} = \sum_{k \in S} \check{y}_k \qquad \hat{\mathbf{T}} = \sum_{k \in S} \frac{\mathbf{x}_k \mathbf{x}_k^T}{\sigma_k^2 \pi_k} \qquad \hat{\mathbf{t}} = \sum_{k \in S} \frac{\mathbf{x}_k \check{y}_k}{\sigma_k^2},$$

$$\check{x}_{jk} = \frac{x_{jk}}{\pi_{jk}} \; (j = 1, 2, \ldots, q), \qquad \hat{\mathbf{t}}_{x\pi} = \left(\hat{t}_{x_1\pi}, \hat{t}_{x_2\pi}, \ldots, \hat{t}_{x_q\pi}\right), \quad \text{onde } \hat{t}_{x_j\pi} = \sum_{k \in S} \check{x}_{jk}$$

O estimador regressão generalizado para t_y, utilizando q variáveis auxiliares pode ser expresso da seguinte forma:

$$\begin{aligned}
\hat{t}_{yR} &= \hat{t}_{y\pi} + \sum_{j=1}^{q} \hat{\beta}_{vj} \left(t_{x_j} - \hat{t}_{x_j\pi} \right) = \hat{t}_{y\pi} + \left(\mathbf{t}_x - \hat{\mathbf{t}}_x \right)^T \underset{\sim}{\hat{\beta}} \\
&= \sum_{k\in S} \check{y}_k + \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi} \right)^T \left(\hat{\mathbf{T}} \right)^{-1} \sum_{k\in S} \sigma_k^{-2} \mathbf{x}_k \check{y}_k \\
&= \sum_{k\in S} \left[1 + \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi} \right)^T \left(\hat{\mathbf{T}} \right)^{-1} \sigma_k^{-2} \mathbf{x}_k \right] \check{y}_k = \sum_{k\in S} g_{ks} \check{y}_k \\
&= f\left(\hat{t}_{y\pi}, \hat{\mathbf{t}}_{x\pi}, \hat{\mathbf{T}}, \hat{\mathbf{t}} \right), \qquad (3.2)
\end{aligned}$$

onde $f(.)$ é uma função não-linear dos estimadores $\hat{t}_{y\pi}$, $\hat{\mathbf{t}}_{x\pi}$, $\hat{\mathbf{T}}$ e $\hat{\mathbf{t}}$. A variância aproximada deste estimador (Särndal, 2003) é dada por

$$\text{AVar}(\hat{t}_{yR}) = \sum_{k\in U} \sum_{l\in U} \Delta_k \frac{E_k}{\pi_k} \frac{E_l}{\pi_l}, \qquad \text{onde} \quad E_k = y_k - \mathbf{x}_k^T \underset{\sim}{\beta} \qquad (3.3)$$

A seguir, serão apresentadas estratégias de estimação para a abordagem de cadastro duplo que fazem uso de estimadores do tipo regressão, e que foram inspiradas nas estratégias apresentadas no capítulo 2. Tais estratégias constituem a contribuição original desta tese.

2 Estimador regressão sob a estratégia de Hartley

Considere a situação de uso de uma abordagem de cadastro duplo sob o cenário 2, e que as seguintes condições são satisfeitas:

- O vetor de q informações auxiliares, denotado por $\mathbf{x}_{kA} = \left(x_{1k}, x_{2k}, \ldots, x_{qk} \right)$ é conhecido para todo $k \in \mathscr{A}$;
- O vetor de m informações auxiliares, denotado por $\mathbf{x}_{kB} = \left(x_{1k}, x_{2k}, \ldots, x_{mk} \right)$ é conhecido para todo $k \in \mathscr{B}$. É importante notar que as informações de variáveis auxiliares disponíveis em um cadastro não precisam necessariamente coincidir com as informações auxiliares disponíveis no outro.

Nos capítulos anteriores foi apresentada a seguinte partição da média populacional

$$t_y = \sum_{k\in\mathscr{A}} y_k + \sum_{k\in\mathscr{B}} y_k - \sum_{k\in ab} y_k.$$

Dessa forma, o seguinte estimador regressão para a média populacional é proposto e denominado estimador regressão sob a estratégia de Hartley.

Estimador regressão sob a estratégia de Hartley

$$\begin{aligned}
\overline{y}_{RH} &= N^{-1}\left\{\hat{t}_{yR}^{A}+\hat{t}_{yR}^{B}-p\hat{t}_{yR}^{ab(A)}-(1-p)\hat{t}_{yR}^{ab(B)}\right\} \\
&= N^{-1}\left\{N_A\overline{y}_{RA}+N_B\overline{y}_{RB}-N_{ab}p\overline{y}_{Rab(A)}-(1-p)N_{ab}\overline{y}_{Rab(B)}\right\} \\
&= w_A\overline{y}_{RA}+w_B\overline{y}_{RB}-w_{ab}\left(p\overline{y}_{Rab(A)}+(1-p)\overline{y}_{Rab(B)}\right) \\
&= w_A\frac{f\left(\hat{t}_{y\pi}^{A},\hat{\mathbf{t}}_{x\pi}^{A},\hat{\mathbf{T}}^{A},\hat{\mathbf{t}}^{A}\right)}{N_A}+w_B\frac{f\left(\hat{t}_{y\pi}^{B},\hat{\mathbf{t}}_{x\pi}^{B},\hat{\mathbf{T}}^{B},\hat{\mathbf{t}}^{B}\right)}{N_B} \\
&- w_{ab}\left\{p\frac{f\left(\hat{t}_{y\pi}^{ab(A)},\hat{\mathbf{t}}_{x\pi}^{ab(A)},\hat{\mathbf{T}}^{ab(A)},\hat{\mathbf{t}}^{ab(A)}\right)}{N_{ab}}+(1-p)\frac{f\left(\hat{t}_{y\pi}^{ab(B)},\hat{\mathbf{t}}_{x\pi}^{ab(B)},\hat{\mathbf{T}}^{ab(B)},\hat{\mathbf{t}}^{ab(B)}\right)}{N_{ab}}\right\}, \quad (3.4)
\end{aligned}$$

onde $\hat{t}_{yR}^{A}$ e $\hat{t}_{yR}^{B}$ são os estimadores do tipo regressão para o total dos cadastros A e B e $\hat{t}_{yR}^{ab(A)}$ e $\hat{t}_{yR}^{ab(B)}$ são os estimadores do tipo regressão para o total do domínio ab. $\overline{y}_{RA}$ e $\overline{y}_{RB}$ são estimadores do tipo regressão para a média dos cadastros A e B e $\overline{y}_{Rab(A)}$ e $\overline{y}_{Rab(B)}$ são estimadores do tipo regressão para a média do domínio ab. $\hat{t}_{y\pi}^{ab(A)}$, $\hat{\mathbf{t}}_{x\pi}^{ab(A)}$, $\hat{\mathbf{T}}^{ab(A)}$ e $\hat{\mathbf{t}}^{ab(A)}$ são definidas de forma análoga às quantidades apresentadas em (3.2). Da mesma forma que a vista para a estratégia de Hartley sem considerar variáveis auxiliares, $0 \leq p \leq 1$.

Resultado 1 *A variância aproximada do estimador $\overline{y}_{RH}$ é função de variâncias aproximadas e covariâncias aproximadas para cada estimador regressão obtido nos cadastros e domínios:*

$$\begin{aligned}
AVar(\overline{y}_{RH}) &= (w_A)^2 AVar(\overline{y}_{RA})+(w_B)^2 AVar(\overline{y}_{RB})+(w_{ab})^2p^2 AVar\left(\overline{y}_{Rab(A)}\right) \\
&+ (w_{ab})^2p^2 AVar\left(\overline{y}_{Rab(A)}\right)+(w_{ab})^2(1-p)^2 AVar\left(\overline{y}_{Rab(B)}\right) \\
&- 2pw_Aw_{ab}\mathsf{A}Cov\left(\overline{y}_{RA}\,;\,\overline{y}_{Rab(A)}\right)-2(1-p)w_Bw_{ab}\mathsf{A}Cov\left(\overline{y}_{RB}\,;\,\overline{y}_{Rab(B)}\right) \quad (3.5)
\end{aligned}$$

A partir da expressão (2.7), a variância aproximada de $\overline{y}_{RH}$ é função de

$$\mathbf{d}=\begin{bmatrix} w_A \\ pw_{ab} \\ w_B \\ (1-p)w_{ab} \end{bmatrix} \quad \text{e} \quad \Sigma_{RH}=\begin{pmatrix} \sigma_A^2 & \sigma_{A;ab(A)} & 0 & 0 \\ \sigma_{A;ab(A)} & \sigma_{ab(A)}^2 & 0 & 0 \\ 0 & 0 & \sigma_B^2 & \sigma_{B;ab(B)} \\ 0 & 0 & \sigma_{B;ab(B)} & \sigma_{ab(B)}^2 \end{pmatrix}.$$

Os elementos de Σ_{RH} representam respectivamente as variâncias e covariâncias com base em estimadores regressão para cadastros e domínios respectivamente, como apresentado por Särndal, Swensson & Wretman (2003). Por exemplo,

$$\sigma_A^2 = \text{AVar}(\overline{y}_{RA}) = \sum_{k\in\mathscr{A}}\sum_{l\in\mathscr{A}} \Delta_{kl}^A \frac{E_k^A}{\pi_k^A}\frac{E_l^A}{\pi_l^A}, \qquad \text{onde} \quad E_k^A = y_k - \mathbf{x}_{kA}^T \underset{\sim}{\beta}^A$$

$$\sigma_B^2 = \text{AVar}(\overline{y}_{RB}) = \sum_{k\in\mathscr{B}}\sum_{l\in\mathscr{B}} \Delta_{kl}^B \frac{E_k^B}{\pi_k^B}\frac{E_l^B}{\pi_l^B}, \qquad \text{onde} \quad E_k^B = y_k - \mathbf{x}_{kB}^T \underset{\sim}{\beta}^B$$

3 Estimador regressão sob a estratégia de Fuller & Burmeister

Na situação em que os tamanhos populacionais dos domínios são desconhecidos (cenário 3), a estratégia de Fuller & Burmeister (1972) é considerada como alternativa à abordagem de Hartley (1962). Neste caso, o estimador regressão para a abordagem de cadastro duplo é denominado estimador regressão sob a estratégia de Fuller & Burmeister, e dado por

$$\overline{y}_{RFB} \;=\; w_a\overline{y}_{Ra} + w_b\overline{y}_{Rb} + \beta_1 w_{ab}\overline{y}_{Rab(A)} + w_{ab}\,(1-\beta_1)\,\overline{y}_{Rab(B)} + N^{-1}\beta_2\left(\hat{N}_{ab}^A - \hat{N}_{ab}^B\right), \qquad (3.6)$$

onde β_1 e β_2 são constantes de ponderação escolhidas de modo a minimizar a variância aproximada do estimador $\overline{y}_{RFB}$.

Resultado 2 *Admitindo a expressão (2.7), a expressão em termos matriciais da variância do estimador regressão sob a estratégia de Fuller & Burmeister é dada em função das quantidades*

$$\mathbf{d} = \begin{bmatrix} w_a \\ w_a p_1 \\ w_b \\ w_b(1-p_1) \\ N^{-1}p_2 \\ N^{-1}p_2 \end{bmatrix} \quad e \quad \Sigma_{RFB} = \begin{pmatrix} \sigma^2_{A(a)} & \sigma_{A(a);A(ab)} & 0 & 0 & 0 & 0 \\ \sigma_{A(a);A(ab)} & \sigma^2_{A(ab)} & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2_{B(b)} & \sigma_{B(b);B(ab)} & 0 & 0 \\ 0 & 0 & \sigma_{B(b);B(ab)} & \sigma^2_{B(ab)} & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_{\hat{N}_{A(ab)}} & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2_{\hat{N}_{B(ab)}} \end{pmatrix}.$$

Os elementos de Σ_{RFB} representam as variâncias aproximadas dos estimadores do tipo regressão para quantidades nos cadastros e domínios, os quais estão presentes no estimador sob a estratégia. Da mesma forma que a vista para o estimador regressão sob a estratégia de Hartley (1962), um estimador para a variância deste estimador pode ser obtido substituindo-se as respectivas quantidades populacionais pelos seus respectivos estimadores de Horvitz-Thompson.

4 Estimador regressão sob a estratégia BLG

Esta estratégia de estimação no contexto de estimadores regressão é inspirada em Bankier, Lepkowski & Groves (1986), e também utiliza a mesma ponderação nos elementos dos cadastros apresentada anteriormente, e dada por

$$\omega_k = \begin{cases} \dfrac{1}{\pi_k^A}, & k \in a \\ \dfrac{1}{\pi_k^B}, & k \in b \\ \dfrac{1}{\pi_k^{A\cup B}} = \dfrac{1}{\pi_k^A + \pi_k^B}, & k \in ab \end{cases} \tag{3.7}$$

Dessa forma, é natural conceber um estimador regressão generalizado a partir desta estratégia. Considerando o problema de estimar, por exemplo, uma média populacional dada por

$$\mu = N^{-1}\left(t_{yA} + t_{yB}\right), \qquad \text{onde } N = N_A + N_B,$$

tem-se que o estimador regressão BLG é dado por

$$\begin{aligned}
\overline{y}_{RBLG} &= N^{-1}\left(\sum_{k\in S_A} g^*_{ks(A)}\omega_k y_k + \sum_{k\in S_B} g^*_{ks(B)}\omega_k y_k\right) \\
&= N^{-1}\left(\sum_{k\in S_a} g_{ksa}\omega_k y_k + \sum_{k\in S_b} g_{ksb}\omega_k y_k + \sum_{k\in S_{ab(A)}} g_{ksab(A)}\omega_k y_k + \sum_{k\in S_{ab(B)}} g_{ksab(B)}\omega_k y_k\right) \\
&= N^{-1}\left(\sum_{k\in S_a} g_{ksa}\frac{y_k}{\pi_k^A} + \sum_{k\in S_b} g_{ksb}\frac{y_k}{\pi_k^B} + \sum_{k\in S_{ab(A)}} g_{ksab(A)}\frac{y_k}{\pi_k^{A\cup B}} + \sum_{k\in S_{ab(B)}} g_{ksab(B)}\frac{y_k}{\pi_k^{A\cup B}}\right) \\
&= N^{-1}\left(\sum_{k\in S_a} g_{ksa}\frac{y_k}{\pi^A} + \sum_{k\in S_b} g_{ksb}\frac{y_k}{\pi_k^B} + \sum_{k\in S_{ab}} g_{ksab}\frac{y_k}{\pi_k^{A\cup B}}\right) \\
&= w_a\overline{y}_{Ra} + w_b\overline{y}_{Rb} + w_{ab}\overline{y}_{Rab}\,,
\end{aligned} \tag{3.8}$$

onde $w_a = \frac{N_a}{N}$, $w_b = \frac{N_b}{N}$ e $w_{ab} = \frac{N_{ab}}{N}$. S_a, S_b e S_{ab} representam os elementos da amostra nos domínios a, b e ab. Dessa forma, a variância aproximada do estimador $\overline{y}_{RBLG}$ é dada de modo mais simples que a apresentada para a estratégia de Hartley:

$$\begin{aligned}
\mathsf{AVar}(\overline{y}_{RBLG}) &= (w_a)^2 \mathsf{AVar}(\overline{y}_{Ra}) + (w_b)^2 \mathsf{AVar}(\overline{y}_{Rb}) + (w_{ab(A)})^2 \mathsf{AVar}\left(\overline{y}_{Rab(A)}\right) \\
&+ (w_{ab(B)})^2 \mathsf{AVar}\left(\overline{y}_{Rab(B)}\right) + 2 w_a w_{ab(A)} \mathsf{ACov}\left(\overline{y}_{Ra}\,;\, \overline{y}_{Rab(A)}\right) \\
&+ 2 w_b w_{ab(B)} \mathsf{ACov}\left(\overline{y}_{Rb}\,;\, \overline{y}_{Rab(B)}\right) \qquad (3.9)
\end{aligned}$$

Resultado 3 *Admitindo a expressão (2.7), a variância do estimador BLG expressa em termos matriciais é dada em função de*

$$\mathbf{d} = \begin{bmatrix} w_a \\ w_b \\ w_{ab(A)} \\ w_{ab(B)} \end{bmatrix} \quad e \quad \Sigma_{RBLG} = \begin{pmatrix} \sigma_a^2 & \sigma_{a;ab(A)} & 0 & 0 \\ \sigma_{a;ab(A)} & \sigma_{ab(A)}^2 & 0 & 0 \\ 0 & 0 & \sigma_b^2 & \sigma_{b;ab(B)} \\ 0 & 0 & \sigma_{b;ab(B)} & \sigma_{ab(B)}^2 \end{pmatrix},$$

Na expressão da variância, $\sigma_a^2 = \mathsf{AVar}(\overline{y}_{Ra})$, $\sigma_{a;ab(A)} = \mathsf{ACov}\left(\overline{y}_{Ra};\overline{y}_{Rab(A)}\right)$ são estimadores regressão para os domínios apresentadas na expressão (3.9). Os elementos de Σ_{RBLG} representam as variâncias e covariâncias aproximadas dos estimadores regressão para os domínios, respectivamente. É possível obter um estimador para esta matriz. Basta substituir as respectivas quantidades populacionais pelos respectivos estimadores de Horvitz-Thompson. Ainda, com base na estratégia, é possível também propor uma versão *raking* do estimador regressão. Neste caso, considerando novamente a mesma estrutura proposta por Skinner (1991), com

$$\begin{aligned}
\omega_{ka}^{(r)} &= \frac{N^A}{\hat{N}^{A(r-1)}} \omega_{ka}^{(r-1)}, \qquad \omega_{kb}^{(r)} = \omega_{kb}^{(r-1)} \qquad \omega_{kab}^{(r)} = \frac{N^A}{\hat{N}^{A(r-1)}} \omega_{kab}^{(r-1)} \quad \text{para } r = 1,3,5,\ldots \\
\omega_{ka}^{(r)} &= \omega_{ka}^{(r-1)}, \qquad \omega_{kb}^{(r)} = \frac{N^B}{\hat{N}^{B(r-1)}} \omega_{kb}^{(r-1)} \qquad \omega_{kab}^{(r)} = \frac{N^B}{\hat{N}^{B(r-1)}} \omega_{kab}^{(r-1)} \quad \text{para } r = 2,4,6,\ldots \\
\omega_{ka}^{(0)} &= \frac{1}{\pi_k^A}, \qquad \omega_{kb}^{(0)} = \frac{1}{\pi_k^B}, \qquad \omega_{kab}^{(0)} = \frac{1}{\left(\pi_k^A + \pi_k^B\right)} \\
\hat{N}^{A(r)} &= \omega_{ka}^{(r)} n^a + \omega_{kab}^{(r)} n_{ab} \\
\hat{N}^{B(r)} &= \omega_{kb}^{(r)} n^b + \omega_{kab}^{(r)} n_{ab}
\end{aligned}$$

tem-se que

$$\overline{y}_{RBLG(r)} = N^{-1}\left(\sum_{k\in S_a} g_{ksa}\omega_{ka}^{(r)}y_k + \sum_{k\in S_b} g_{ksa}\omega_{kb}^{(r)}y_k + \sum_{k\in S_{ab}} g_{ksa}\omega_{kab}^{(r)}y_k\right), \tag{3.10}$$

No caso em que $r \longrightarrow \infty$, tem-se que

$$\overline{y}_{RBLG(\infty)} = \hat{w}_{a*}\overline{y}_{Ra*} + \hat{w}_{b*}\overline{y}_{Rb*} + \hat{w}_{ab*}\overline{y}_{Rab*}, \tag{3.11}$$

onde

$$\hat{w}_{a*} = N_A - \tilde{N}_{ab}\,, \qquad \hat{w}_{b*} = N_B - \tilde{N}_{ab}\,, \qquad \hat{w}_{ab*} = \tilde{N}_{ab}$$

$$\overline{y}_{Ra*} = \frac{\sum_{k\in S_a} g_{ksa}y_k}{n_a}\,, \qquad \overline{y}_{Rb*} = \frac{\sum_{k\in S_b} g_{ksb}y_k}{n_b}\,, \qquad \overline{y}_{Rab*} = \frac{\sum_{k\in S_{ab}} g_{ksab}y_k}{n_{ab}} \quad \text{e}$$

$\tilde{N}_{ab}$ é a menor raiz da equação quadrática em x,

$$\left(n_{ab(A)} + n_{ab(B)}\right)x^2 - \left\{\left(n_{ab(A)} + n_{ab(B)}\right)\left(N_A + N_B\right) + \left[\left(\frac{1}{\pi_k^A}\right) + \left(\frac{1}{\pi_k^B}\right)\right]n_a n_b\right\}x + \left(n_{ab(A)} + n_{ab(B)}\right)N_A N_B = 0.$$

Resultado 4 *A variância aproximada do estimador raking-regressão para a abordagem de cadastro duplo é dada por:*

$$\begin{aligned} AVar\left(\overline{y}_{RBLG(\infty)}\right) &= (w_{a*})^2\, AVar(\overline{y}_{Ra*}) + (w_{b*})^2\, AVar(\overline{y}_{Rb*}) + (w_{ab*})^2\, AVar(\overline{y}_{Rab*}) \\ &+ (\mu_a + \mu_b - \mu_{ab})^2\,\frac{N_{ab}N_aN_b}{n_AN_b + n_BN_a}\left(1+\lambda^2\right), \end{aligned} \tag{3.12}$$

com

$$\lambda^2 = \frac{N_aN_bN_{ab}^2\left[\left(\pi_k^A\right)^2N_A - \left(\pi_k^B\right)^2N_B\right]^2}{N_AN_B\pi_k^A\pi_k^B\left(\pi_k^A + \pi_k^B\right)^2\left(N_{ab}^2 - N_AN_B\right)^2}.$$

A partir de (2.7), esta variância é dada em função de

$$\mathbf{d} = \begin{bmatrix} w_a^* \\ w_b^* \\ w_{ab}^* \\ 1 \end{bmatrix} \quad e \quad \Sigma_{RBLG(\infty)} = \begin{bmatrix} \sigma_a^2 & 0 & 0 & 0 \\ 0 & \sigma_b^2 & 0 & 0 \\ 0 & 0 & \sigma_{ab}^2 & 0 \\ 0 & 0 & 0 & \sigma_{\hat{N}_{ab}}^2 \end{bmatrix}$$

Na matriz $\Sigma_{RBLG(\infty)}$, os valores de σ_a^2, σ_b^2 e σ_{ab}^2 são obtidos da expressão (3.12). Por exemplo, $\sigma_a^2 = \text{AVar}(\overline{y}_{Ra*})$.

5 Estimador regressão sob a estratégia de Máxima Pseudo-Verossimilhança

Considere a situação de uso de uma abordagem de cadastro duplo sob o cenário 3, e que as informações sobre as variáveis auxiliares são conhecidas, conforme descrito anteriormente. Tem-se que o estimador regressão para a estratégia de máxima pseudo-verossimilhança é dado por

$$\begin{aligned} \overline{y}_{RPML} &= \frac{\left(N_A - \hat{N}_{ab,PML}\right)\overline{y}_{Ra} + \hat{N}_{ab,PML}\overline{y}_{ab*} + \left(N_B - \hat{N}_{ab,PML}\right)\overline{y}_{Rb}}{N} \\ \overline{y}_{RPML} &= \hat{w}_{a*}\overline{y}_{Ra} + \hat{w}_{ab*}\overline{y}_{ab*} + \hat{w}_{b*}\overline{y}_{Rb} \end{aligned} \tag{3.13}$$

$$\overline{y}_{ab*} = \frac{\left[\pi^A \hat{N}_{ab}^A \overline{y}_{Rab(A)} + \pi^B \hat{N}_{ab}^B \overline{y}_R^{ab(B)}\right]}{\left[\pi^A \hat{N}_{ab}^A + \pi^B \hat{N}_{ab}^B\right]}$$

As demais quantidades deste estimador são as mesmas apresentadas para a expressão (2.18).

Resultado 5 *Admitindo a expressão (2.7), a variância aproximada do estimador regressão sob a estratégia PML é dada em função de*

$$\mathbf{d}^T = \begin{bmatrix} w_a & w_{ab}\left(\frac{\pi^A}{\pi^A + \pi^B}\right) & w_b & w_{ab}\left(\frac{\pi^B}{\pi^A + \pi^B}\right) & \frac{\mu_a}{N} & \frac{\mu_b}{N} & \frac{\mu_{ab}}{N} \end{bmatrix} \quad e$$

$$\Sigma_{RPML} = \begin{pmatrix} \sigma_a^2 & \sigma_{(a;ab(A))} & 0 & 0 & 0 & 0 & 0 \\ \sigma_{(a;ab(A))} & \sigma_{ab(A)}^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_b^2 & \sigma_{(b;ab(B))} & 0 & 0 & 0 \\ 0 & 0 & \sigma_{(b;ab(B))} & \sigma_{ab(B)}^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_{\hat{N}_{PML}^a}^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_{\hat{N}_{PML}^b}^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma_{\hat{N}_{PML}^{ab}}^2 \end{pmatrix}$$

Prova: Através de procedimentos já descritos anteriormente, tem-se que um pseudo-estimador para o estimador PML é dado por

$$\begin{aligned} \overline{y}_{RPML_L} &\approx w_a\overline{y}_{Ra} + \hat{w}_{a*}\mu_a - w_a\mu_a + w_b\overline{y}_{Rb} + \hat{w}_{b*}\mu_b - w_b\mu_b \\ &+ w_{ab}\left\{\left(\frac{\pi^A}{\pi^A+\pi^B}\right)\overline{y}_{Rab(A)} + \left(\frac{\pi^B}{\pi^A+\pi^B}\right)\overline{y}_{Rab(B)}\right\} \\ &+ \hat{w}_{ab*}\mu_{ab} - w_{ab}\mu_{ab} \end{aligned} \tag{3.14}$$

Além dos procedimentos anteriores, é preciso linearizar o estimador $\overline{y}_{ab*}$. Tem-se que

$$\begin{aligned} \overline{y}_{ab*} &\approx \frac{\pi^A N_{ab}\mu^{ab} + \pi^B N_{(B)}\mu_{ab}}{\pi^A\hat{N}_{ab(A)} + \pi^B\hat{N}_{ab(B)}} + \frac{1}{\pi^A N_{ab} + \pi^B N_{ab}}\left\{\pi^A\hat{N}_{ab}^A\overline{y}_{Rab(A)} + \pi^B\hat{N}_{ab}^B\overline{y}_{Rab(B)}\right. \\ &- \left.\left(\frac{\pi^A N_{ab}\mu_{ab} + \pi^B N_{ab}\mu_{ab}}{\pi^A N_{ab} + \pi^B N_{ab}}\right)\left(\pi^A\hat{N}_{ab(A)} + \pi^B\hat{N}_{ab(B)}\right)\right\} \\ &= \mu_{ab} + \frac{(N_{ab})^{-1}}{\pi^A+\pi^B}\left\{\pi^A\hat{N}_{ab(A)}\overline{y}_{Rab(A)} + \pi^B\hat{N}_{ab(B)}\overline{y}_{Rab(B)} - \mu_{ab}\left(\pi^A\hat{N}_{ab(A)} + \pi^B\hat{N}_{ab(B)}\right)\right\} \\ &\approx \mu_{ab} + \frac{(N_{ab})^{-1}}{\pi^A+\pi^B}\left\{\pi^A\left(\hat{N}_{ab(A)}\mu^{ab} + N_{ab}\overline{y}_{Rab(A)} - N_{ab}\mu_{ab}\right) + \pi^B\left(\hat{N}_{ab(B)}\mu_{ab} + N_{ab}\overline{y}_{Rab(B)} - N_{ab}\mu_{ab}\right)\right. \\ &- \left.\mu_{ab}\pi^A\hat{N}_{ab(A)} - \mu_{ab}\hat{N}_{ab(B)}\pi^B\right\} \\ &= \mu_{ab} + \frac{(N_{ab})^{-1}}{\pi^A+\pi^B}\left\{N_{ab}\pi^A\overline{y}_{Rab(A)} - \pi^A N_{ab}\mu_{ab} + N_{ab}\pi^B\overline{y}_{Rab(B)} - \pi^B N_{ab}\mu_{ab}\right\} \\ &= \mu_{ab} + \left(\frac{\pi^A}{\pi^A+\pi^B}\right)\overline{y}_{Rab(A)} + \left(\frac{\pi^B}{\pi^A+\pi^B}\right)\overline{y}_{Rab(B)} - \left(\frac{\pi^A+\pi^B}{\pi^A+\pi^B}\right)\mu_{ab} \end{aligned}$$

$$= \left(\frac{\pi^A}{\pi^A+\pi^B}\right)\overline{y}_{Rab(A)} + \left(\frac{\pi^B}{\pi^A+\pi^B}\right)\overline{y}_{Rab(B)}.$$

Logo, obtendo a variância no estimador $\overline{y}_{PML_L}$, temos o resultado 5. ■

Sobre os estimadores propostos originalmente nesta tese é possível observar o seguinte:

1. Na forma da variância do estimador regressão sob a estratégia PML, se $p = \pi^A/(\pi^A+\pi^B)$ e $1-p = \pi^B/(\pi^A+\pi^B)$, a variância do estimador regressão sob a estratégia de Hartley tende a se aproximar da variância do estimador regressão sob a estratégia PML, desde que $\sigma^2_{\hat{N}_{a,PML}}$, $\sigma^2_{\hat{N}_{b,PML}}$ e $\sigma^2_{\hat{N}_{ab,PML}}$ sejam negligíveis. Se esta condição não ocorrer, e $w_{a*} \approx w_a$, $w_{b*} \approx w_b$ e $w_{ab*} \approx w_{ab}$, a variância do estimador regressão sob a estratégia PML tende a se aproximar da variância do estimador regressão sob a estratégia BLG;

2. As variâncias dos estimadores regressão generalizados sob as estratégias PML e Fuller & Burmeister terão o mesmo vetor $\mathbf{d}^T$ quando $\beta_1 = \theta$ e $\beta_2 = (\phi-\theta)\,\mu_{ab} + (1-\phi)\mu_a + \phi\mu_b$, onde $\phi = n_A N_{ab}/\left(n_A N_b + n_B N_a\right)$ e $\theta = n_A N_B/\left(n_A N_B + n_B N_A\right)$;

3. A partir da condição 2, AVar$(\overline{y}_{RPML}) \geq$ AVar$(\overline{y}_{FB})$;

4. Dadas as condições 2 e 3, quando $\beta_2 = 0$, tem-se que $\overline{y}_{RFB} = \overline{y}_{RH}$, e portanto, analogamente aos estimadores já conhecidos, conclui-se que *A* Var$(\overline{y}_{RFB}) \geq$ AVar$(\overline{y}_{RH})$;

5. É possível notar que para as estratégias propostas, AVar$(\overline{y}_{RPML}) \geq$ AVar$(\overline{y}_{RH})$.

CAPÍTULO 4

Propriedades Estatísticas dos Estimadores

1 Introdução

Conceitos sobre sequências de populações finitas são importantes para investigar o comportamento assintótico dos estimadores propostos nesta tese. Neste capítulo, tais conceitos são apresentados no contexto da abordagem de cadastro duplo. Para uma revisão da teoria considerando um cadastro apenas, o leitor pode consultar Särndal, Swensson & Wretman (2003), Park (2002) e Fuller (2009).

Considere uma sequência infinita de elementos identificáveis por $k = 1,2,3,\ldots$, associada a uma sequência infinita de valores $y_1, y_2, y_3, \ldots$, onde y_k é o valor associado ao elemento k. Uma sequência de populações $U_1, U_2, U_3, \ldots$ é definida de modo que U_ν contém os N_ν primeiros elementos da sequência infinita, isto é, $U_\nu = \{1,2,\ldots,N_\nu\}$. Dessa forma,

$$U_1 \subset U_2 \subset U_3 \subset \cdots \qquad \text{e} \qquad N_1 < N_2 < N_3 < \cdots$$

θ_ν é o valor de um parâmetro de interesse na população U_ν, que é função do vetor $\mathbf{Y}_\nu = \{y_1, y_2, \ldots, y_{N_\nu}\}$. Considere a situação em que $U_\nu = \mathscr{A}_\nu \cup \mathscr{B}_\nu$, onde $\mathscr{A}_\nu$ e $\mathscr{B}_\nu$ representam os conjuntos de elementos da sequência pertencentes aos cadastros A e B respectivamente. Neste caso, $N_\nu = N_{A\nu} + N_{B\nu} - N_{ab\nu}$. O cenário que define a relação entre as sequências $\mathscr{A}_\nu$ e $\mathscr{B}$ é descrito pela figura (4.1) a seguir para $\nu = 2$. Outros cenários podem ser verificados no apêndice D.

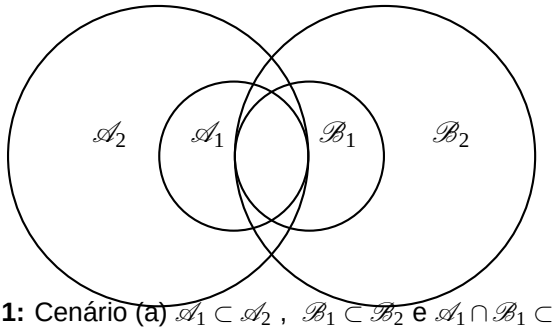


Figura 4.1: Cenário (a) $\mathscr{A}_1 \subset \mathscr{A}_2$, $\mathscr{B}_1 \subset \mathscr{B}_2$ e $\mathscr{A}_1 \cap \mathscr{B}_1 \subset \mathscr{A}_2 \cap \mathscr{B}_2$

A figura ilustra populações com sequências para cada cadastro encaixadas. Isto também acontece com a união e a interseção entre os cadastros. Todos os estimadores apresentados nesta tese são baseados neste cenário.

São considerados planos amostrais probabilísticos denotados por $p_{A\nu}(.)$ e $p_{B\nu}(.)$ para $\mathscr{A}_\nu$ e $\mathscr{B}_\nu$, que associam probabilidades $p_{A\nu}(S_{A\nu})$ e $p_{B\nu}(S_{B\nu})$ para as amostras $S_{A\nu}$ e $S_{B\nu}$ em cada cadastro. Os tamanhos das amostras obtidas de $\mathscr{A}_\nu$ e $\mathscr{B}_\nu$ são denotados por $n_{A\nu}$ e $n_{B\nu}$, e será admitido que $n_{A1} < n_{A2} < \ldots$ e $n_{B1} < n_{B2} < \ldots$. As probabilidades de inclusão de primeira ordem dos elementos de $\mathscr{A}_\nu$ e $\mathscr{B}_\nu$ são denotadas por $\pi^A_{\nu k}$ e $\pi^B_{\nu k}$. As probabilidades de inclusão de segunda ordem dos elementos de $\mathscr{A}_\nu$ e $\mathscr{B}_\nu$ são denotadas por $\pi^A_{\nu kl}$ e $\pi^B_{\nu kl}$. Dessa forma, tem-se uma sequência de populações finitas sob a abordagem de cadastro duplo indexadas por ν, tal que, quando $\nu \to \infty$, observa-se que $n_{A\nu} \to \infty$, $N_{A\nu} \to \infty$, $n_{B\nu} \to \infty$ e $N_{B\nu} \to \infty$ simultaneamente.

Considere θ_ν como um parâmetro de interesse. Sob a abordagem de cadastro duplo, é possível escrever $\theta_\nu = \theta_{A\nu} + \theta_{B\nu} - \theta_{ab\nu}$, onde

- $\theta_{A\nu}$ é um parâmetro de interesse em $\mathscr{A}_\nu$;
- $\theta_{B\nu}$ é um parâmetro de interesse em $\mathscr{B}_\nu$;
- $\theta_{ab\nu}$ é um parâmetro de interesse em em $\mathscr{A}_\nu \cap \mathscr{B}_\nu$.

Por exemplo, quando θ_ν é uma média populacional,

$$\begin{aligned}
\theta_\nu &= \mu_\nu = N_\nu^{-1} \sum_{k \in U_\nu} y_k = N_\nu^{-1} \left(\sum_{k \in \mathscr{A}_\nu} y_k + \sum_{k \in \mathscr{B}_\nu} y_k - \sum_{k \in ab_\nu} y_k \right) \\
&= N_\nu^{-1} \left(N_{A\nu} \mu_{A\nu} + N_{B\nu} \mu_{B\nu} - N_{ab} \mu_{ab\nu} \right) \\
&= w_{A\nu} \mu_{A\nu} + w_{B\nu} \mu_{B\nu} - w_{ab\nu} \mu_{ab\nu} \\
&= \theta_{A\nu} + \theta_{B\nu} - \theta_{ab\nu} \qquad (4.1)
\end{aligned}$$

onde $w_{Av} = N_{Av}/N_v$, $w_{Bv} = N_{Bv}/N_v$ e $w_{abv} = N_{abv}/N_v$. Um estimador centrado para $\hat{\theta}_v$ é dado por $\hat{\theta}_v = \hat{\theta}_{Av} + \hat{\theta}_{Bv} - \hat{\theta}_{abv}$. Para o caso da média populacional,

$$\begin{aligned}
\hat{\theta}_{Av} &= w_A \sum_{k \in S_{Av}} \left(y_k/\pi^A_{vk}\right) \\
\hat{\theta}_{Bv} &= w_B \sum_{k \in S_{Bv}} \left(y_k/\pi^B_{vk}\right) \\
\hat{\theta}_{abv} &= w_{ab} \sum_{k \in S_{Av} \cap ab} \left(y_k/\pi^A_{vk}\right) + w_{ab} \sum_{k \in S_{Bv} \cap ab} \left(y_k/\pi^B_{vk}\right)
\end{aligned}$$

2 Centralidade assintótica e consistência do estimador sob a abordagem de cadastro duplo

A partir da sequência de populações descrita anteriormente, é possível apresentar as seguintes definições:

Definição 1 *Um estimador $\hat{\theta}_v$ é **assintoticamente centrado** para θ_v se*

$$\lim_{v \to \infty} \left[E\left(\hat{\theta}_v - \theta_v\right)\right] = 0,$$

Sob a abordagem de cadastro duplo, tem-se a condição quando

$$\begin{aligned}
\lim_{v \to \infty} \left[E\left(\hat{\theta}_{Av} - \theta_{Av}\right)\right] &= 0, \\
\lim_{v \to \infty} \left[E\left(\hat{\theta}_{Bv} - \theta_{Bv}\right)\right] &= 0, \\
\lim_{v \to \infty} \left[E\left(\hat{\theta}_{abv} - \theta_{abv}\right)\right] &= 0.
\end{aligned}$$

Definição 2 *Um estimador $\hat{\theta}_v$ é **consistente de acordo com o plano** para θ_v se, para qualquer constante $\varepsilon > 0$,*

$$\lim_{v \to \infty} P\left(\left|\hat{\theta}_v - \theta_v\right| > \varepsilon\right) = 0$$

Sob a abordagem de cadastro duplo, tem-se a condição quando

$$\begin{aligned}
\lim_{v \to \infty} P\left(\left|\hat{\theta}_{Av} - \theta_{Av}\right| > \varepsilon\right) &= 0, \\
\lim_{v \to \infty} P\left(\left|\hat{\theta}_{Bv} - \theta_{Bv}\right| > \varepsilon\right) &= 0, \\
\lim_{v \to \infty} P\left(\left|\hat{\theta}_{abv} - \theta_{abv}\right| > \varepsilon\right) &= 0.
\end{aligned}$$

Condições necessárias para verificação de consistência do estimador de Horvitz-Thompson

Para o caso em que é considerado apenas um cadastro, Isaki & Fuller (1982) provaram que as condições a seguir são suficientes para garantir a propriedade de consistência do estimador de Horvitz-Thompson. Estas condições serão necessárias para o desenvolvimento dos resultados nesta tese.

C1. A forma do estimador de Horvitz-Thompson é dada por

$$\overline{y}_{HT} = \frac{1}{N_\nu} \sum_{k \in S_\nu} \frac{y_k}{\pi_{\nu k}},$$

C2. As probabilidades de inclusão de primeira e segunda ordem são dadas respectivamente por:

$$\pi_{\nu k} = P\left(k \in S_\nu\right) = n_\nu \left(\sum_{k \in S_\nu} \omega_k \right)^{-1} \omega_k, \qquad k = 1, \ldots, N_\nu$$
$$\pi_{\nu kl} = P\left(k \in S_\nu \, , \, l \in S_\nu\right), \qquad k, l = 1, \ldots, N_\nu$$

C3. Seja

$$g_{\nu kl} = \begin{cases} \pi_{\nu k}\pi_{\nu l} - \pi_{\nu kl}, & \text{se } \pi_{\nu k}\pi_{\nu l} - \pi_{\nu kl} > 0 \\ 0, & c.c. \end{cases}$$

C4. A sequência $\{y_k, \omega_k\}$ é tal que

i. $\dfrac{1}{N_\nu^2} \sum_{k \neq l = 1} g_{\nu kl}^r = \mathbf{O}\left(n_\nu^{-2r\delta}\right)$

ii. $\dfrac{1}{N_\nu} \sum_{i \in U_\nu} \left[\dfrac{N_\nu}{n_\nu} \left(y_k - \overline{y}\right) \right]^{2s} < M < \infty$

C5. Para $\delta > 0$ e para todo N_ν e μ_ν é a média populacional da variável de interesse Y, as probabilidades de inclusão satisfazem:

$$N_\nu^{-2} \sum_{k \neq l} \sum_{\in U_\nu} g_{\nu kl}^r = \mathbf{O}\left(n_\nu^{-2r\delta}\right),$$

onde para $r = 1$, $\left| \dfrac{y_k}{\pi_{\nu k}} \right|$ é limitado. Se $s = 1$, então $n_\nu^2 g_{\nu kl}$ é limitado para todo k, l.

Para o caso em que o parâmetro de interesse é uma média populacional, estas condições condições garantem que

$$\overline{y}_{HT} - \mu_\nu = \mathbf{O}_p\left(n_\nu^{-1/2}\right).$$

onde $\delta = 1/2$ e $\mathbf{O}_p(.)$ indica ordem em probabilidade, como usado em teoria da probabilidade[2]. Sobre as amostras obtidas de cada cadastro e seus estimadores, ainda é necessário estabelecer as seguintes condições:

A1. Denotando por $S_{A\nu}$ e $S_{B\nu}$ as amostras obtidas de A e B, respectivamente, tem-se que, para qualquer variável de interesse y,

1. $\max\limits_{k\in S_{A\nu}} y_k = \mathbf{o}_p\left\{\left(n_{A\nu}\right)^{1/2}\right\}$ e
2. $\max\limits_{k\in S_{B\nu}} y_k = \mathbf{o}_p\left\{\left(n_\nu^B\right)^{1/2}\right\}$,

Esta condição é satisfeita para qualquer plano amostral em que $\pi_k^A \neq 0$ e $\pi_k^B \neq 0$. O uso simultâneo dos dois cadastros induz à obtenção do vetor que representa a união entre as amostras obtidas de cada cadastro, dado por

$$\mathbf{S}_\nu = S_{A\nu} \cup S_{B\nu} = S_{a\nu} \cup S_{b\nu} \cup S_{ab\nu}, \qquad \text{onde } S_{ab\nu} = S_{A\nu} \cap S_{B\nu}$$

Tem-se que $n_{A\nu} = n_{a\nu} + n_{ab\nu}$, $n_{B\nu} = n_{b\nu} + n_{ab\nu}$. O conjunto S_ν é apenas uma das possíveis amostras obtidas através do uso simultâneo de dois cadastros. De modo geral, $\mathbf{S}_\nu$ é um conjunto aleatório definido da seguinte forma:

$$\mathbf{S}_\nu = \mathbf{S}_\nu^A \cup \mathbf{S}_\nu^B = \left\{k \in \mathscr{A}_\nu, l \in \mathscr{B}_\nu |\ I_{\nu k}^A = 1,\ I_{\nu l}^B = 1,\ 1 \leq k \leq N_\nu^A\ ;\ 1 \leq l \leq N_\nu^B\right\}.$$

$\mathbf{S}_\nu$ é uma das $\binom{N_\nu^A}{n_\nu^A} \times \binom{N_\nu^B}{n_\nu^B}$ amostras resultantes do processo de seleção utilizando cadastro duplo .

A2. Assintoticamente, o estimador de Horvitz-Thompson possui distribuição normal para qualquer variável de interesse y e satisfaz a condição **A1**, conforme descrito por Wu & Rao (2006).

A3. Sobre o tamanho populacional dos domínios, supondo que $N_a > 0$ e $N_b > 0$ e admitindo que $\nu \to \infty$, tem-se que

[2]Para um maior aprofundamento, o leitor pode consultar Sen, Singer & Lima (2010) e Fuller (2009). Uma breve revisão sobre o tema é apresentado no apêndice A. A prova das propriedades para o estimador de Hartley podem ser consultadas no apêndice B, página 106.

$$\frac{n_{A\nu}}{n_{A\nu}+n_\nu^B} \to c_1 \in (0,1) \qquad \frac{N_{a\nu}}{N_{A\nu}} \to c_2 \in (0,1) \qquad \frac{N_\nu^b}{N_\nu^B} \to c_3 \in (0,1)$$

$$\frac{N_{a\nu}}{N_\nu} \to w_a \in (0,1) \qquad \frac{N_\nu^b}{N_\nu} \to w_b \in (0,1)$$

É possível estender os resultados para um contexto mais amplo, quando F cadastros estão disponíveis. Neste caso, a sequência $U_1, U_2, U_3, \ldots$ é considerada de tal forma que

$$U_\nu = U_\nu^1 \cup U_\nu^2 \cup U_\nu^3 \cdots \cup U_\nu^F = \bigcup_{f=1}^{F} U_\nu^f, \qquad \text{onde } f = 1, \ldots, F.$$

Lohr (2007, 2010) apresenta resultados para $F = 3$. Tem-se ainda que

$$N_\nu = \sum_{i=1}^{F} N_\nu^{U^i} - \sum_{i<j} N_\nu^{U_\nu^i \cap U_\nu^j} + \sum_{i<j<k} N_\nu^{U_\nu^i \cap U_\nu^j \cap U_\nu^k} + \cdots + (-1)^{F+1} N_\nu^{\bigcap_{i=1}^{F} U_\nu^i}.$$

Logo,

$$\left(\bigcup_{f=1}^{F} U_1^f\right) \subset \left(\bigcup_{f=1}^{F} U_2^f\right) \subset \left(\bigcup_{f=1}^{F} U_3^f\right) \subset \cdots \tag{4.2}$$

3 Centralidade assintótica e consistência do estimador regressão generalizado

No contexto de populações finitas, ao considerar a situação em que é de interesse estimar uma média populacional $\mu_\nu = N_\nu^{-1} \sum_{k \in U_\nu} y_k$, o estimador regressão generalizado é dado por

$$\hat{\mu}_{yR\nu} = N_\nu^{-1} \hat{t}_{yR} = N_\nu^{-1} \left[\hat{t}_{y\pi\nu} + \sum_{j=1}^{q} \hat{\beta}_{\nu j} \left(t_{x_j \nu} - \hat{t}_{x_j \pi \nu} \right) \right],$$

onde $\hat{t}_{yR}$ é o estimador regressão para o total populacional. Tem-se ainda que o estimador original de Horvitz-Thompson é obtido na situação particular em que $\hat{\beta}_{\nu j} = 0$, $\forall j$. Seguindo a descrição de Robinson & Särndal (1983), sob a situação em que é considerado um modelo de superpopulação, $\hat{\beta}_{\nu j}$ é função dos vetores aleatórios $\mathbf{Y}_\nu$ e $\mathbf{I}_\nu$, onde $\mathbf{Y}_\nu = \{Y_1, Y_2, \ldots, Y_{N_\nu}\}$ e $\mathbf{I}_\nu = \{I_{k1}, I_{k2}, \ldots, I_{kN_\nu}\}$, onde $I_{\nu k}$ é uma variável indicadora de inclusão do elemento k na amostra, onde

$$I_{\nu k} = \begin{cases} 1, & \text{se } k \in S_\nu \\ 0, & \text{c.c.} \end{cases}$$

Robinson & Särndal (1983) mostraram que o estimador regressão generalizado é consistente de acordo com o plano amostral e assintoticamente centrado, considerando válidas as condições a seguir:

A1. $\limsup_{\nu\to\infty}\left\{\dfrac{\sum_{k\in U_\nu} x_{kj}^2}{N_\nu}\right\}<\infty \quad \text{para} \quad j=0,1,\ldots,q$

A2. $\limsup_{\nu\to\infty}\left\{\dfrac{\sum_{k\in U_\nu} Y_k^2}{N_\nu}\right\}<\infty$, com probabilidade 1.

A3. $\limsup_{\nu\to\infty}\left\{E\left(\sum_{j=0}^{q}\hat{\beta}_{\nu j}^2\middle|\mathbf{Y}_\nu\right)\right\}<\infty$, com probabilidade 1.

A4. $\liminf_{\nu\to\infty}\left\{N_\nu\left[\min_{1\le k\le N_\nu}\pi_{\nu k}\right]\right\}=\infty.$

A5. $\lim_{\nu\to\infty}\left\{\max_{1\le k\neq l\le N_\nu}\left[\dfrac{\pi_{\nu kl}}{\pi_{\nu k}\pi_{\nu l}}-1\right]\right\}=0.$

A partir destas considerações, são apresentados os seguintes resultados para o estimadores regressão propostos no capítulo 3:

Resultado 6 *O estimador $\bar{y}_{RH}$ é assintoticamente centrado e consistente de acordo com o plano amostral sob as condições **A1**-**A5**.*

Resultado 7 *O estimador $\bar{y}_{RFB}$ é assintoticamente centrado e consistente de acordo com o plano amostral sob as condições **A1**-**A5**.*

Resultado 8 *Os estimadores $\bar{y}_{RBLG}$ e $\bar{y}_{RBLG(\infty)}$ são assintoticamente centrados e consistentes de acordo com o plano amostral sob as condições **A1**-**A5**.*

Resultado 9 *O estimador $\bar{y}_{RPML}$ é assintoticamente centrado e consistente de acordo com o plano amostral sob as condições **A1**-**A5**.*

Prova: Apêndice **B**.

4 Erro quadrático médio assintótico dos estimadores regressão sob a abordagem de cadastro duplo

A seguir serão apresentadas condições necessárias para a minimização do erro quadrático médio, definido como $E\left\{(\overline{y}_{R\nu}-\overline{y}_{\nu})^2\,|\mathbf{Y}\right\}$, dos estimadores regressão propostos. Para o caso em que apenas um cadastro é considerado,

$$\mu_k = E(Y_k) \qquad \sigma_k^2 = \text{Var}(Y_k)$$

Dessa forma, em cada cadastro, adicionalmente são assumidas as seguintes condições sobre a estrutura das observações da variável de interesse:

A6. Os Y_k's são não-correlacionados e

$$\limsup_{\nu\to\infty}\frac{\sum\limits_{k\in U_\nu}\mu_k^2}{N}<\infty, \quad \text{e} \quad \limsup_{\nu\to\infty}\frac{\sum\limits_{k\in U_\nu^d}\mu_k^2}{N_\nu}<\infty\,; \quad \sigma_k^2<\infty$$

A7. Para um conjunto de q variáveis auxiliares no cadastro, existem constantes β_j tal que para todo ν suficientemente grande,

$$n_\nu\sum_{j=0}^{q}E\left[\left(\hat{\beta}_{\nu j}-\beta_j\right)^2\right]<\infty,$$

A8. $\liminf\limits_{\nu\to\infty}\dfrac{N_\nu}{n_\nu}\left\{\min\limits_{1\le k\le N_\nu}\pi_{\nu k}\right\}>0$

A9. $\limsup\limits_{\nu\to\infty}\dfrac{(N_\nu)^2}{n_\nu}\left\{\max\limits_{1\le k\ne\le N_\nu}|\pi_{\nu kl}-\pi_{\nu k}\pi_{\nu l}|\right\}<\infty$

A10. Existe uma constante γ tal que

$$\mu_k=\gamma\pi_{\nu k}+\sum_{j=0}^{q}\beta_j x_{jk}\,, \quad 1\le k\le N_\nu$$

A partir destas condições para cadastros, as quais também podem ser admitidas para os domínios, serão apresentadas expressões para os EQM's dos estimadores propostos para cada uma das estratégias de estimadores regressão propostas nesta tese.

Resultado 10 *Dada uma sequência de de populações finitas U_ν, tem-se que para cada um dos cadastros, sob as condições **A1**, **A2**, **A5**, **A6**−**A9** o erro quadrático médio assintótico do estimador regressão é dado por*

$$n_\nu E\left(\overline{y}_{R\nu}-\overline{y}_\nu\right)^2 = D_\nu + W_\nu + Q_\nu < \infty, \quad \textit{quando} \quad \nu \to \infty$$

onde

$$\begin{aligned}
D_\nu &= \frac{n_\nu}{(N_\nu)^2}\sum_{k\in U_\nu}\sigma_k^2\left(\frac{1}{\pi_{\nu k}-1}\right),\\
W_\nu &= \frac{n_\nu}{(N_\nu)^2}\sum_{k\in U_\nu}\sum_{\substack{l\in U_\nu\\ k\neq l}}\sigma_k^2\left(\pi_{kl}-\pi_{\nu k}\pi_l\right)\left(\frac{g_k}{\pi_{\nu k}}-\frac{g_l}{\pi_l}\right)^2, \quad \text{com} \quad g_k = \mu_k - \sum_{j=0}^{q}\beta_j x_{kj}\\
Q_\nu &= \frac{n_\nu}{(N_\nu)^2}E\left\{\left[\sum_{j=0}^{q}\left(\hat{\beta}_{\nu j}-\beta_j\right)\right]\frac{1}{N_\nu}\sum_{k\in U_\nu}x_{kj}\left(\frac{I_{\nu k}}{\pi_{\nu k}}-1\right)\right\}^2 +\\
&+ 2E\left\{\left[\frac{\sqrt{n_\nu}}{N_\nu}\sum_{k\in U_\nu}(y_k-\mu_k)\left(\frac{I_{\nu k}}{\pi_{\nu k}}-1\right)\right]\left[\frac{\sqrt{n_\nu}}{N_\nu}\sum_{j=0}^{q}\left(\hat{\beta}_{\nu j}-\beta_j\right)\frac{1}{N_\nu}\sum_{k\in U_\nu}x_{kj}\left(\frac{I_{\nu k}}{\pi_{\nu k}}-1\right)\right]\right\}\\
&+ 2E\left\{\left[\frac{\sqrt{n_\nu}}{N_\nu}\sum_{k\in U_\nu}(y_k-\mu_k)\left(\frac{I_{\nu k}}{\pi_{\nu k}}-1\right)\right]\frac{\sqrt{n_\nu}}{N_\nu}\sum_{k\in U_\nu}g_k\left(\frac{I_{\nu k}}{\pi_{\nu k}}-1\right)\right\}\\
&+ 2E\left\{\left[\frac{\sqrt{n_\nu}}{N_\nu}\sum_{j=0}^{q}\left(\hat{\beta}_{\nu j}-\beta_j\right)\frac{1}{N_\nu}\sum_{k\in U_\nu}x_{kj}\left(\frac{I_{\nu k}}{\pi_{\nu k}}-1\right)\right]\frac{\sqrt{n_\nu}}{N_\nu}\sum_{k\in U_\nu}g_k\left(\frac{I_{\nu k}}{\pi_{\nu k}}-1\right)\right\}
\end{aligned}$$

com $Q_\nu \longrightarrow 0$, quando $\nu \to \infty$.

Prova: Seja

$$\sqrt{n}\left(\overline{y}_{R\nu}-\overline{y}_\nu\right) = m_1 + m_2 + m_3,$$

onde

$$m_1 = \frac{\sqrt{n_\nu}}{N_\nu}\sum_{k\in U}(y_k-\mu_k)\left(\frac{I_{\nu k}}{\pi_{\nu k}}-1\right)$$

$$\begin{aligned} m_2 &= \frac{\sqrt{n_\nu}}{N_\nu}\sum_{j=0}^{q}\left(\hat{\beta}_{\nu j}-\beta_j\right)\frac{1}{N}\sum_{k\in U}x_{kj}\left(\frac{I_{\nu k}}{\pi_{\nu k}}-1\right) \\ m_3 &= \frac{\sqrt{n_\nu}}{N_\nu}\sum_{k\in U}g_k\left(\frac{I_{\nu k}}{\pi_{\nu k}}-1\right) \end{aligned}$$

Logo,

$$n_\nu E\left(\overline{y}_{R\nu}-\overline{y}_\nu\right)^2 = E(m_1^2)+E(m_2^2)+E(m_3^2)+E(m_1m_2)+E(m_1m_3)+E(m_2m_3)$$

É possível notar que,

$$E(m_2) = E\{E(m_1^2|\mathbf{I})\} = E\left\{\frac{n_\nu}{N_\nu}\sum_{k\in U}\sigma_k^2\left(\frac{I_{\nu k}}{\pi_{\nu k}}-1\right)^2\right\} = D_\nu.$$

Sob **A6** e **A8** tem-se que

$$\limsup_{\nu\to\infty} D_\nu \leq \limsup_{\nu\to\infty}\left(\frac{n_\nu}{N_\nu \min\pi_{\nu k}}\right)\frac{1}{N_\nu}\sum_{k\in U_\nu}\sigma_k^2 < \infty \tag{4.3}$$

Ainda, sob **A1**, **A4**, **A5** e **A7**,

$$E(m_2^2) \leq E\left\{\frac{n_\nu}{(N_\nu)^2}\sum_{j=0}^{q}E\left[\left(\hat{\beta}_{\nu j}-\beta_j\right)^2|\mathbf{I}\right]\sum_{j=0}^{q}\left(\hat{\beta}_{\nu j}-\beta_j\right)\frac{1}{N_\nu}\sum_{k\in U_\nu}x_{kj}\left(\frac{I_{\nu k}}{\pi_{\nu k}}-1\right)\right\} \longrightarrow 0, \tag{4.4}$$

quando $\nu\to\infty$.

$$\begin{aligned} E(m_3^2) &= \frac{n_\nu}{(N_\nu)^2}\sum_{k\in U_\nu}g_k^2\left(\frac{I_{\nu k}}{\pi_{\nu k}}-1\right) \\ &+ \frac{n_\nu}{(N_\nu)^2}\sum_{k\in U_\nu}\sum_{\substack{l\in U_\nu \\ k\neq l}}g_kg_l\left(\frac{I_{\nu kl}}{\pi_{\nu k}\pi_{\nu l}}-1\right) = W_\nu, \end{aligned} \tag{4.5}$$

onde

$$\begin{aligned} \limsup_{\nu\to\infty} W_\nu &\leq \limsup_{\nu\to\infty}\left\{\left(\frac{N_\nu^2}{n_\nu}\max_\nu|\pi_{\nu kl}-\pi_{\nu k}\pi_{\nu l}|\right)\left(\frac{n_\nu^2}{N_\nu^2\max\{\pi_{\nu k}^2\}}\right)\frac{8}{N_\nu}\left(\sum_{k\in U_\nu}\mu_k^2+2\sum_{j=0}^{q}\beta_j^2\sum_{k\in U_\nu}x_{kj}^2\right)\right\} \\ &< \infty. \end{aligned} \tag{4.6}$$

Ainda, sob **A1**, **A7**, **A9** e **A10**:

$$\begin{aligned} E(m_1 m_2) &\leq \left[E(m_1)^2 E(m_2)^2\right]^{1/2} \to 0, \\ E(m_1 m_3) &= E\left[E(m_1|\mathbf{I}_\nu) m_2\right]^{1/2} = 0, \\ E(m_2 m_3) &\leq \left[E(m_2)^2 E(m_3)^2\right]^{1/2} \to 0, \end{aligned}$$

o que leva ao resultado 2. ■

Todas as condições apresentadas são utilizadas sob a suposição de que n_ν cresça com mesma taxa de N_ν. Se isso não acontecer, tem-se que $D_\nu \longrightarrow 0$ e $W_\nu \longrightarrow 0$, o que ilustra a consistência do estimador, mas que não leva a um resultado interessante por não apresentar uma forma geral para o EQM.

Considerando **C15**, $\pi_{\nu k}$ precisa ser constante para todo ν, se $\gamma \neq 0$. Sob a perspectiva de um plano amostral de amostragem aleatória simples, por exemplo, D_ν é mínimo quando

$$\pi_{\nu k} = \frac{n\sigma_k}{\sum\limits_{k \in U_\nu} \sigma_k}.$$

Dessa forma, tem-se que

$$EQM(\overline{y}_{R\nu}) = \frac{\left[(N_\nu)^{-1} \sum\limits_{k \in U_\nu} \sigma_k\right]^2 - \left(\frac{n_\nu}{(N_\nu)^2}\right) \sum\limits_{k \in U_\nu} \sigma_k^2}{n_\nu}$$

A seguir, serão apresentadas expressões para o erro quadrático médio assintótico de cada um dos estimadores regressão propostos.

4.1 Estratégia de Hartley

Sob a estratégia de Hartley, o erro quadrático médio do estimador $\overline{y}_{RH}$ é dado pela expressão:

$$\begin{aligned} \text{EQM}(\overline{y}_{RH}) = E(\overline{y}_{RH} - \mu)^2 &= (w_A)^2 E(\overline{y}_{RA} - \mu_A)^2 + (w_B)^2 E(\overline{y}_{RB} - \mu_B)^2 \\ &+ 2w_A w_B E\left\{(\overline{y}_{Ra} - \mu_A)(\overline{y}_{Rb} - \mu_B)\right\} \\ &- 2w_A w_{ab} pE\left[(\overline{y}_{Ra} - \mu_A)\left(\overline{y}_{Rab(A)} - \mu_{ab}\right)\right] - \end{aligned}$$

$$
\begin{aligned}
&- \quad 2w_A w_{ab}(1-p)E\left[(\overline{y}_{Ra}-\mu_A)\left(\overline{y}_{Rab(B)}-\mu_{ab}\right)\right] \\
&- \quad 2w_B w_{ab}pE\left[(\overline{y}_{Rb}-\mu_B)\left(\overline{y}_{Rab(A)}-\mu_{ab}\right)\right] \\
&- \quad 2w_B w_{ab}(1-p)E\left[(\overline{y}_{Rb}-\mu_B)\left(\overline{y}_{Rab(B)}-\mu_{ab}\right)\right] \\
&+ \quad p^2\left(w_{ab}\right)^2 E\left(\overline{y}_{Rab(A)}-\mu_{ab}\right)^2 \\
&+ \quad 2p(1-p)\left(w_{ab}\right)^2 E\left[\left(\overline{y}_{Rab(A)}-\mu_{ab}\right)\left(\overline{y}_{Rab(B)}-\mu_{ab}\right)\right] \\
&+ \quad \left(w_{ab}\right)^2(1-p)^2 E\left[\left(\overline{y}_{Rab(B)}-\mu_{ab}\right)^2\right] \qquad (4.7)
\end{aligned}
$$

Quando $v \to \infty$, tem-se que

$$
\begin{aligned}
E\left\{(\overline{y}_{RAv}-\mu_{Av})(\overline{y}_{RBv}-\mu_{Bv})\right\} &= 0; \\
E\left\{\left(\overline{y}_{Rab(A)v}-\mu_{abv}\right)\left(\overline{y}_{Rab(B)v}-\mu_{abv}\right)\right\} &= 0; \\
E\left\{(\overline{y}_{RAv}-\mu_{Av})\left(\overline{y}_{Rab(B)v}-\mu_{abv}\right)\right\} &= 0; \\
E\left\{\left(\overline{y}_{Rab(A)v}-\mu_{abv}\right)\left(\overline{y}_{Rab(B)v}-\mu_{abv}\right)\right\} &= 0,
\end{aligned}
$$

pois planos independentes são aplicados em A e B. Além disso, tem-se que,

$$
\begin{aligned}
E\left\{(\overline{y}_{RAv}-\mu_{Av})\left(\overline{y}_{Rab(A)v}-\mu_{vab}\right)\right\} &\longrightarrow C_{(A,ab(A))}; \\
E\left\{(\overline{y}_{RBv}-\mu_{Bv})\left(\overline{y}_{Rab(B)v}-\mu_{ab}\right)\right\} &\longrightarrow C_{(B,ab(B))}; \\
E\left\{(\overline{y}_{RAv}-\mu_{Av})^2\right\} &= D_v^A + W_v^A < \infty; \\
E\left\{(\overline{y}_{RBv}-\mu_{Bv})^2\right\} &= D_v^B + W_v^B < \infty; \\
E\left\{\left(\overline{y}_{Rab(A)}-\mu_{ab}\right)^2\right\} &= D_v^{ab(A)} + W_v^{ab(A)} < \infty; \\
E\left\{\left(\overline{y}_{Rab(B)v}-\mu_{ab}\right)^2\right\} &= D_v^{ab(B)} + W_v^{ab(B)} < \infty,
\end{aligned}
$$

$C_{(s_1,s_2)}$ indica a covariância entre as quantidades provenientes dos conjuntos s_1 e s_2, conforme descrito por Särndal (1992). Logo,

$$\begin{aligned}
\text{EQM}(\overline{y}_{RHv}) = E(\overline{y}_{RHv} - \mu_v)^2 &= (w_{vA})^2 \left(D_v^A + W_v^A\right) + (w_{vB})^2 \left(D_v^B + W_v^B\right) \\
&- 2w_{vA}w_{vab}pC_{(A,ab(A))} - 2w_{vB}w_{vab}(1-p)C_{(A,ab(A))} \\
&+ p^2 \left(D_v^{ab(A)} + W_v^{ab(A)}\right) \\
&+ (w_{vab})^2 (1-p)^2 E \left(D_v^{ab(B)} + W_v^{ab(B)}\right) \\
&\leq (w_{vA})^2 \left(D_v^A + W_v^A\right) + (w_{vB})^2 \left(D_v^B + W_v^B\right) \\
&+ p^2 \left(D_v^{ab(A)} + W_v^{ab(A)}\right) \\
&+ (w_{vab})^2 (1-p)^2 E \left(D_v^{ab(B)} + W_v^{ab(B)}\right) \qquad (4.8)
\end{aligned}$$

Resultado 11 *Admitindo que $w_A = \tau_A$, $w_B = \tau_B$ e $w_{vab} = \tau_{ab}$, tem-se então que o erro quadrático médio assintótico do estimador regressão sob a estratégia de Hartley é dado por*

$$\begin{aligned}
EQM(\overline{y}_{RHv}) &\leq \tau_A^2 EQM(\overline{y}_{RAv}) + \tau_B^2 EQM(\overline{y}_{RBv}) \\
&+ p^2 EQM(\overline{y}_{Rab(A)v}) + (1-p)^2 \tau_{ab}^2 EQM(\overline{y}_{Rab(B)v}) < \infty.
\end{aligned}$$

4.2 Estratégia de Fuller & Burmeister

Sob a estratégia em que a magnitude dos domínios é desconhecida tem-se(apêndice B),

$$\begin{aligned}
\overline{y}_{RFB} - \mu_v &\approx w_a(\overline{y}_{Rav} - \mu_{va}) + w_{vb}(\overline{y}_{Rbv} - \mu_b) + p_1 w_{vab}\left(\overline{y}_{vRab(A)} - \mu_{vab}\right) + (1-p_1)w_{vab}\left(\overline{y}_{Rab(B)v} - \mu_{vab}\right) \\
&+ N^{-1}\left(\hat{N}_a\mu_{va} + \hat{N}_b\mu_{vb} + p_1\hat{N}_{ab(A)}\mu_{vab} + (1-p_1)\hat{N}_{vab(B)}\mu_{vab} + p_2\hat{N}_{ab(A)} - p_2\hat{N}_{ab(B)}\right)
\end{aligned}$$

$$
\begin{aligned}
&= w_a\left(\overline{y}_{Rav}-\mu_{va}\right)+p_1 w_{vab}\left(\overline{y}_{Rab(A)}-\mu_{vab}\right)\\
&+ w_b\left(\overline{y}_{Rbv}-\mu_{vb}\right)+(1-p_1)w_{vab}\left(\overline{y}_{Rab(B)v}-\mu_{vab}\right)+\mathbf{O}(1).
\end{aligned}
$$

Dessa forma,

$$
\begin{aligned}
\left(\overline{y}_{RFBv}-\mu_v\right)^2 &\approx \left(w_{va}\right)^2\left(\overline{y}_{Rav}-\mu_{va}\right)^2+2p_1 w_{va}w_{vab}\left(\overline{y}_{Rav}-\mu_{va}\right)\left(\overline{y}_{Rab(A)v}-\mu_{vab}\right)\\
&+ p_1^2\left(w_{vab}\right)^2\left(\overline{y}_{Rab(A)v}-\mu_{vab}\right)^2+2w_{va}w_{vb}\left(\overline{y}_{Rav}-\mu_{va}\right)\left(\overline{y}_{Rbv}-\mu_{vb}\right)\\
&+ 2(1-p_1)w_{va}w_{vab}\left(\overline{y}_{Rav}-\mu_{va}\right)\left(\overline{y}_{Rab(B)v}-\mu_{vab}\right)\\
&+ 2p_1 w_{vb}w_{vab}\left(\overline{y}_{Rbv}-\mu_{vb}\right)\left(\overline{y}_{Rab(A)v}-\mu_{vab}\right)\\
&+ 2p_1(1-p_1)\left(w_{vab}\right)^2\left(\overline{y}_{Rab(A)v}-\mu_{vab}\right)\left(\overline{y}_{Rab(B)v}-\mu^{vab}\right)\\
&+ \left(w_{vb}\right)^2\left(\overline{y}_{Rbv}-\mu_{vb}\right)^2+2(1-p_1)w_{vb}w_{vab}\left(\overline{y}_{Rbv}-\mu_{vb}\right)\left(\overline{y}_{Rab(B)v}-\mu_{vab}\right)\\
&+ (1-p_1)^2\left(w_{vab}\right)^2\left(\overline{y}_{Rab(B)v}-\mu_{vab}\right)^2+\mathbf{O}\left(N_v^{-1}\right)
\end{aligned}
$$

Calculando o valor esperado, ou seja, $E\left(\overline{y}_{RFBv}-\mu_v\right)^2$ e admitindo $v\to\infty$, tem-se

$$
\begin{aligned}
E\left(\overline{y}_{Rav}-\mu_{va}\right)^2 &= D_v^a+W_v^a<\infty\\
E\left(\overline{y}_{Rbv}-\mu_{vb}\right)^2 &= D_v^b+W_v^b<\infty\\
E\left(\overline{y}_{Rab(A)v}-\mu_{vab}\right)^2 &= D_v^{ab(A)}+W_v^{ab(A)}<\infty\\
E\left(\overline{y}_{Rab(B)v}-\mu_{vab}\right)^2 &= D_v^{ab(B)}+W_v^{ab(B)}<\infty\\
E\left\{\left(\overline{y}_{Rav}-\mu_{va}\right)\left(\overline{y}_{Rb}-\mu^b\right)\right\} &= 0\\
E\left\{\left(\overline{y}_{Rab(A)}-\mu_{vab}\right)\left(\overline{y}_{Rab(B)v}-\mu_{vab}\right)\right\} &= 0\\
E\left\{\left(\overline{y}_{Rav}-\mu_{va}\right)\left(\overline{y}_{Rab(B)v}-\mu_{vab}\right)\right\} &= 0
\end{aligned}
$$

$$
\begin{aligned}
E\left\{\left(\overline{y}_{Rbv}-\mu_{vb}\right)\left(\overline{y}_{Rab(A)v}-\mu_{vab}\right)\right\} &= 0 \\
E\left\{\left(\overline{y}_{Rav}-\mu_{va}\right)\left(\overline{y}_{Rab(A)v}-\mu_{vab}\right)\right\} &= C_{(a;ab(A))}<\infty \\
E\left\{\left(\overline{y}_{Rbv}-\mu_{bv}\right)\left(\overline{y}_{Rab(B)v}-\mu_{vab}\right)\right\} &= C_{(b;ab(A))}<\infty
\end{aligned}
$$

Resultado 12 *O erro quadrático médio assintótico do estimador regressão sob a estratégia de Fuller & Burmeister é dado por*

$$
\begin{aligned}
EQM\left(\overline{y}_{RFB}\right) &\approx \left(w_{va}\right)^2\left(D_v^a+W_v^a\right)^2+p_1^2\left(w_{vab}\right)^2\left(D_v^{ab(A)}+W_v^{ab(A)}\right) \\
&+ \left(w_{vb}\right)^2\left(D_v^b+W_v^b\right)+\left(1-p_1\right)^2\left(w_{vab}\right)^2\left(D_v^{ab(B)}+W_v^{ab(B)}\right) \\
&+ 2p_1w_{va}w_{vab}C_{(a;ab(A))}+2(1-p_1)w_{vb}w_{vab}C_{b;ab(B)}
\end{aligned}
$$

4.3 Estratégia BLG

Sob a estratégia adotada por Bankier, Lepkowski & Groves (1986), é possível notar que a forma do estimador apresentada em (3.8) é inicialmente escrita da seguinte forma:

$$
\overline{y}_{RBLGv} = \frac{\sum\limits_{k\in S_{Av}} g^*_{ks(A)}\omega_k y_k+\sum\limits_{k\in S_{Bv}} g^*_{ks(B)}\omega_k y_k}{N} = w_{Av}\overline{y}_{RAv}+w_{Bv}\overline{y}_{Rv},
$$

onde

$$
\omega_k=\begin{cases}
\dfrac{1}{\pi^A_{vk}}, & k\in a \\
\dfrac{1}{\pi^B_{vk}}, & k\in b \\
\dfrac{1}{\pi^{A\cup B}_{vk}}=\dfrac{1}{\pi^A_{vk}+\pi^B_{vk}}, & k\in ab
\end{cases}
$$

Dessa forma, tem-se que

$$
\overline{y}_{RBLGv}-\mu = w_v^A\left(\overline{y}^A_{Rv}-\mu^A_v\right)+w_v^B\left(\overline{y}^B_{Rv}-\mu^B_v\right)
$$

$$\begin{aligned}(\overline{y}_{RBLGv}-\mu_v)^2 &= \left(w_v^A\right)^2\left(\overline{y}_{Rv}^A-\mu_v^A\right)^2+\left(w_v^B\right)^2\left(\overline{y}_{Rv}^B-\mu_v^B\right)^2 \\ &+ 2w_v^Aw_v^B\left(\overline{y}_{Rv}^A-\mu_v^A\right)\left(\overline{y}_{Rv}^B-\mu_v^B\right)\end{aligned}$$

$$E\left(\overline{y}_{RBLGv}-\mu_v\right)^2 = \left(w_v^A\right)^2E\left(\overline{y}_{Rv}^A-\mu_v^A\right)^2+\left(w_v^B\right)^2E\left(\overline{y}_{Rv}^B-\mu_v^B\right)^2 \quad (4.9)$$

Resultado 13 *Para $w_v^A\to\tau_A$ e $w_v^B\to\tau_B$, tem-se que o erro quadrático médio do estimador regressão sob a estratégia BLG é dado por*

$$EQM(\overline{y}_{RBLG}) = \tau_A^2EQM(\overline{y}_{Rv}^A)+\tau_B^2EQM(\overline{y}_{Rv}^B)$$

A partir da estratégia de Hartley, é possível notar que $\text{EQM}(\overline{y}_{RBLG})<\text{EQM}(\overline{y}_{RHv})$. Os EQM's serão aproximadamente iguais quando os EQM's dos estimadores referentes aos domínios são aproximadamente nulos.

Para a versão *raking* o estimador regressão apresentado em (3.11) também temos resultado similar, basta notar que existem constantes α^A, α^B que satisfazem:

$$\begin{aligned}\omega_{ka}^{(\infty)} &= \frac{1}{\pi_{vk}^A}\left[\prod_{r\ \text{par}}\left(\frac{N_{Av}}{\hat{N}^{A(r-1)}}\right)\right]=\frac{1}{\pi_{vk}^A}\alpha^A \\ \omega_{kb}^{(\infty)} &= \frac{1}{\pi_{vk}^B}\left[\prod_{r\ \text{ímpar}}\left(\frac{N_v^B}{\hat{N}^{B(r-1)}}\right)\right]=\frac{1}{\pi_{vk}^B}\alpha^B \\ \omega_{kab}^{(\infty)} &= \frac{1}{\pi_{vk}^A+\pi_{vk}^B}\left[\prod_{r\ \text{par}}\left(\frac{N_{Av}}{\hat{N}^{A(r-1)}}\right)\prod_{r\ \text{ímpar}}\left(\frac{N_v^B}{\hat{N}^{B(r-1)}}\right)\right]=\left(\frac{1}{\pi_{vk}^A+\pi_{vk}^B}\right)\alpha^A\alpha^B\end{aligned} \quad (4.10)$$

4.4 Estratégia de Máxima Pseudo-Verossimilhança

Com base na estratégia de Máxima Pseudo-Verossimilhança, é possível notar que

$$\begin{aligned}\overline{y}_{RPML} &= w_{va}\overline{y}_{Rv}^a+w_{vb}\overline{y}_{Rv}^b+w_{vab}\left(\frac{\pi_{vk}^A}{\pi_{vk}^A+\pi_{vk}^B}\right)\overline{y}_{Rab(A)}+w_{vab}\left(\frac{\pi_{vk}^B}{\pi_{vk}^A+\pi_{vk}^B}\right)\overline{y}_{Rv}^{ab(B)} \\ &+ \hat{w}^a\mu_{va}+\hat{w}^b\mu_{vb}+\hat{w}^{ab}\mu_{vab}-w_{va}\mu_{va}-w_{vb}\mu_{vb}-w_{vab}\mu_{vab}\end{aligned}$$

Logo,

$$\begin{aligned}\overline{y}_{RPMLv} - \mu_v &\approx w_{va}(\overline{y}^a_{Rv} - \mu_{va}) + w_{vb}\left(\overline{y}^b_{Rv} - \mu_{vb}\right) \\ &+ w_{vab}\left(\frac{\pi^A_{vk}}{\pi^A_{vk}+\pi^B_{vk}}\right)\overline{y}_{Rab(A)v} + w_{vab}\left(\frac{\pi^B_{vk}}{\pi^A_{vk}+\pi^B_{vk}}\right)\overline{y}_{Rab(B)v} + O(1)\end{aligned}$$

No caso em que $\left(\frac{\pi^A_{vk}}{\pi^A_{vk}+\pi^B_{vk}}\right) = p_1$ e $\left(\frac{\pi^B_{vk}}{\pi^A_{vk}+\pi^B_{vk}}\right) = (1-p_1)$, onde p_1 é uma constante de ponderação, tal que $0 \leq p_1 \leq 1$, o estimador $\overline{y}_{RPML}$ assume forma similar à do estimador $\overline{y}_{RFB}$. Dessa forma, $EQM(\overline{y}_{RPMLv})$ terá forma similar à $EQM(\overline{y}_{RFBv})$.

5 Distribuições assintóticas dos estimadores propostos

Skinner & Rao (1996), mostraram que é possível representar os estimadores $\overline{y}_H$, $\overline{y}_{FB}$ e $\overline{y}_{PML}$ como funções do vetor

$$\hat{\eta} = \left(\begin{array}{ccccccc} \overline{y}_{HTa} & \overline{y}_{HTab(A)} & \hat{N}^{ab(A)}/N & \overline{y}_{HTb} & \overline{y}_{HTab(B)} & \hat{N}^{ab(B)}/N \end{array}\right),$$

que é estimador de

$$\eta = \left(\begin{array}{cccccc} \mu_a & \mu_{ab} & N_{ab}/N & \mu_b & \mu_{ab} & N_{ab}/N \end{array}\right).$$

O estimador $\overline{y}_{BLG}$ também pode ser representado de forma análoga, como função de

$$\hat{\eta} = \left(\begin{array}{cc} \overline{y}_{HT(A)} & \overline{y}_{HT(B)} \end{array}\right)$$

sendo mais simples que a forma apresentada para os outros estimadores. Para todas as estratégias, $\hat{\eta}$ é consistente para η. Logo,

$$n^{1/2}(\hat{\eta} - \eta) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma)$$

onde $n = n_A + n_B - n_{ab}$. Σ é uma matriz bloco diagonal da forma

$$\Sigma = \left(\begin{array}{cc} \Sigma_A & \mathbf{0} \\ \mathbf{0} & \Sigma_B \end{array}\right)$$

onde

$$\Sigma_A = \begin{pmatrix} \sigma_a^2 & \sigma_{a;ab(A)} & 0 \\ \sigma_{a;ab(A)} & \sigma_{ab(A)}^2 & 0 \\ 0 & 0 & \sigma_{N_{ab}^A} \end{pmatrix} \qquad \Sigma_B = \begin{pmatrix} \sigma_b^2 & \sigma_{b;ab(B)} & 0 \\ \sigma_{b;ab(B)} & \sigma_{ab(B)}^2 & 0 \\ 0 & 0 & \sigma_{N_{ab}^B} \end{pmatrix}$$

Os estimadores do tipo regressão propostos também podem ser escritos como funções de um vetor denotado por

$$\hat{\varphi} = \begin{pmatrix} \overline{y}_{Ra} & \overline{y}_{Rab(A)} & \hat{N}^{ab(A)}/N & \overline{y}_{Rb} & \overline{y}_{Rv}^{ab(B)} & \hat{N}^{ab(B)}/N \end{pmatrix},$$

Foram apresentadas condições no capítulo 3 que tornam $\hat{\varphi}$ consistente para φ. Logo,

$$n^{1/2}(\hat{\varphi} - \varphi) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma_*).$$

onde, de acordo com a estratégia utilizada, Σ_* é dada por Σ_{RH}, Σ_{RFB}, Σ_{RBLG} ou Σ_{RPML}. Todos os estimadores propostos podem ser dados na forma $\hat{\theta} = f(\hat{\varphi})$, onde $f(.)$ é uma função contínua, não linear e diferenciável em relação aos componentes de $\hat{\varphi}$. Por exemplo, sob a estratégia de Hartley, $f(\hat{\phi}) = \overline{y}_{RH}$. Logo, pelo método método delta, tem-se que

$$n^{1/2}(f(\hat{\varphi}) - f(\varphi)) \xrightarrow{d} \mathcal{N}\left(0, [f'(\varphi)]^T \Sigma_* [f'(\varphi)]\right). \tag{4.11}$$

onde $f'(\varphi)$ é vetor de derivadas parciais. Ao substituir quantidades populacionais pelas suas respectivas quantidades amostrais, é razoável considerar que um intervalo de confiança aproximado para $f(\varphi)$ é dado por

$$f(\hat{\varphi}) \pm z_{1-\alpha/2} \sqrt{[f'(\varphi)]^T \hat{\Sigma}_* [f'(\varphi)]/n^{1/2}} \tag{4.12}$$

CAPÍTULO 5

Inferência para casos particulares do estimador regressão generalizado

Nos capítulos anteriores foram apresentados vários estimadores do tipo regressão para cadastro duplo. Neste capítulo são ilustrados casos particulares destes estimadores, considerando que um plano de amostragem aleatória simples é aplicado em cada um dos cadastros.

1 Estimadores do tipo razão sob a estratégia de Hartley

Considerando-se o estimador regressão sob a estratégia de Hartley apresentado pela expressão (3.4), e considerando o cenário em que o estimador regressão reduz-se ao caso de estimadores do tipo razão, é possível obter os estimadores apresentados a seguir.

1.1 Estimador razão RH1

Sob a situação particular em que uma variável auxiliar está disponível em ambos os cadastros, e considerando que um plano de amostragem aleatória simples é empregado em cada cadastro, o estimador RH1 do tipo razão para μ, sob a estratégia de Hartley, é considerado para situações em que o total das variáveis auxiliares é conhecido para os domínios. Tem-se que

$$\overline{y}_{RH1} = w_a \frac{\tilde{y}_a}{\tilde{x}_a}\mu_{X_a} + w_b \frac{\tilde{y}_b}{\tilde{x}_b}\mu_{X_b} + w_{ab} p \frac{\tilde{y}_{ab(A)}}{\tilde{x}_{ab(A)}}\mu_{X_{ab}} + w_{ab}(1-p)\frac{\tilde{y}_{ab(B)}}{\tilde{x}_{ab(B)}}\mu_{X_{ab}}, \tag{5.1}$$

onde

- $w_a = N_a/N$;

- $w_b = N_b/N$;
- $w_{ab} = N_{ab}/N$;
- μ_{X_b} é a média populacional da variável auxiliar no domínio b;
- μ_{X_a} é a média populacional da variável auxiliar no domínio a;
- μ_{X_b} é a média populacional da variável auxiliar no domínio b;
- $\tilde{y}_a$ é a média amostral da variável de interesse no domínio a;
- $\tilde{y}_b$ é a média amostral da variável de interesse no domínio b;
- $\tilde{y}_{ab(A)}$ é a média amostral das informações da variável de interesse no cadastro A e que pertencem ao domínio ab;
- $\tilde{y}_{ab(B)}$ é a média amostral das informações da variável de interesse no cadastro B e que pertencem ao domínio ab.
- $\tilde{x}_a$ é a média amostral da variável auxiliar no domínio a;
- $\tilde{x}_b$ é a média amostral da variável auxiliar no domínio b;
- $\tilde{x}_{ab(A)}$ é a média amostral das informações da variável auxiliar no cadastro A e que pertencem ao domínio ab;
- $\tilde{x}_{ab(B)}$ é a média amostral das informações da variável auxiliar no cadastro B e que pertencem ao domínio ab.

$\tilde{y}_a$, $\tilde{y}_b$, e $\tilde{y}_{ab(B)}$ representam as médias amostrais dos valores pertencentes aos domínios a, b e ab e identificados nas amostras de tamanho n_A e n_B obtidas de cada um dos cadastros. A variância aproximada deste estimador, expressa em termos matriciais, é dada por $\text{AVar}(\bar{y}_{RH1}) = \mathbf{d}^T \Sigma_{RH1} \mathbf{d}$, onde

$$\mathbf{d} = \begin{bmatrix} 1 \\ p \\ -R_a \\ -pR_{ab(A)} \\ 1 \\ 1-p \\ -R_b \\ -(1-p)R_{ab(B)} \end{bmatrix} \quad \text{e} \quad \Sigma_{RH1} = \begin{pmatrix} \Sigma_A & \mathbf{0} \\ \mathbf{0} & \Sigma_B \end{pmatrix},$$

onde

$$\Sigma_A = \begin{pmatrix} \Sigma_{y(a;ab(A))} & \Sigma_{xy(a;ab(A))} \\ \Sigma_{xy(a;ab(A))} & \Sigma_{x(a;ab(A))} \end{pmatrix} = \begin{pmatrix} \sigma^2_{ya} & \sigma_{(ya;yab(A))} & \sigma_{(ya;xa)} & \sigma_{(ya;xab(A))} \\ \sigma_{(ya;yab(A))} & \sigma^2_{yab(A)} & \sigma_{(xa;yab(A))} & \sigma_{(yab(A);xab(A))} \\ \sigma_{(ya;xa)} & \sigma_{(xa;yab(A))} & \sigma^2_{xa} & \sigma_{(xa;xab(A))} \\ \sigma_{(ya;xab(A))} & \sigma_{(xab(A);yab(A))} & \sigma_{(xa;xab(A))} & \sigma^2_{xab(A)} \end{pmatrix}$$

$$\Sigma_B = \begin{pmatrix} \Sigma_{y(b;ab(B))} & \Sigma_{xy(b;ab(B))} \\ \Sigma_{xy(b;ab(B))} & \Sigma_{x(b;ab(B))} \end{pmatrix} = \begin{pmatrix} \sigma^2_{yb} & \sigma_{(yb;yab(B))} & \sigma_{(yb;xb)} & \sigma_{(yb;xab(B))} \\ \sigma_{(yb;yab(B))} & \sigma^2_{yab(B)} & \sigma_{(xa;yab(B))} & \sigma_{(yab(B);xab(B))} \\ \sigma_{(yb;xb)} & \sigma_{(xb;yab(B))} & \sigma^2_{xb} & \sigma_{(xb;xab(B))} \\ \sigma_{(yb;xab(B))} & \sigma_{(xab(B);yab(B))} & \sigma_{(xb;xab(B))} & \sigma^2_{xab(B)} \end{pmatrix}$$

A forma da variância aproximada deste estimador é dada por

$$\begin{aligned} A\text{Var}(\bar{y}_{RH1}) &= N^{-2}\sigma^2_{RH1}, \\ \text{em que } \sigma^2_{RH1} &= (N_A)^2\left(\frac{1-f_A}{n_A}\right)w_a\sigma^2_{ya} + p^2(N_A)^2\left(\frac{1-f_A}{n_A}\right)w_{ab(A)}\sigma^2_{yab(A)} \\ &+ R^2_a(N_A)^2\left(\frac{1-f_A}{n_A}\right)w_a\sigma^2_{xa} + p^2R^2_{ab(A)}(N_A)^2\left(\frac{1-f_A}{n_A}\right)w_{ab(A)}\sigma^2_{xab(A)} \\ &- 2R^2_a\left(\frac{N_a}{n_a}\right)^2 n_A(1-f_A)\rho_{xya}\sigma_{ya}\sigma_{xa} \\ &- 2p^2R_{ab(A)}\left(\frac{N_{ab}}{n_{ab(A)}}\right)^2 n_A(1-f_A)\rho_{xyab(A)}\sigma_{yab(A)}\sigma_{xab(A)} \\ &+ \left(N^B\right)^2\left(\frac{1-f_B}{n_B}\right)w_b\sigma^2_{yb} + (1-p)^2\left(N^B\right)^2\left(\frac{1-f_B}{n_B}\right)w_{ab(B)}\sigma^2_{yab(B)} \\ &+ R^2_b\left(N^B\right)^2\left(\frac{1-f_B}{n_B}\right)w_b\sigma^2_{xb} + (1-p)^2R^2_{ab(B)}(N_A)^2\left(\frac{1-f_B}{n_B}\right)w_{ab(B)}\sigma^2_{xab(B)} \\ &- 2R^2_b\left(\frac{N_b}{n_b}\right)^2 n_B(1-f_B)\rho_{xyb}\sigma_{yab}\sigma_{xb} \\ &- 2(1-p)^2R_{ab(B)}\left(\frac{N_{ab}}{n_{ab(B)}}\right)^2 n_A(1-f_A)\rho_{xyab(B)}\sigma_{yab(B)}\sigma_{xab(B)}, \end{aligned}$$

onde

- $w_{ab(A)} = N_{ab}/N_A, \quad w_{ab(B)} = N_{ab}/N_B$;
- $R_a = \dfrac{\mu_{ya}}{\mu_{xa}}, \quad R_{ab(A)} = \dfrac{\mu_{yab(A)}}{\mu_{xab(A)}}$;
- $R_b = \dfrac{\mu_{yb}}{\mu_{xb}}, \quad R_{ab(B)} = \dfrac{\mu_{yab(B)}}{\mu_{xab(B)}}$;
- σ^2_{ya}, σ^2_{yb}, $\sigma^2_{yab(A)}$, $\sigma^2_{yab(B)}$ são as variâncias populacionais referentes aos domínios;
- ρ_{xya}, $\rho_{xyab(A)}$, ρ_{xyb} e $\rho_{xyab(B)}$ representam os coeficientes de correlação linear entre as variáveis auxiliares e as variáveis de interesse nos domínios.

Ao substituir as quantidades populacionais pelas suas respectivas quantidades amostrais, tem-se um estimador para a variância do estimador $RH1$. Na situação em que apenas um cadastro contém informação auxiliar disponível, A por exemplo, este estimador reduz-se à forma apresentada e analisada por Coelho (2007), onde

$$\overline{y}_{RH1} = w_a \frac{\tilde{y}_a}{\tilde{x}_a}\mu_{X_a} + w_{ab} p \frac{\tilde{y}_{ab(A)}}{\tilde{x}_{ab(A)}}\mu_{X_{ab}} + w_b \tilde{y}_b + w_{ab}(1-p)\tilde{y}_{ab(B)}. \tag{5.2}$$

1.2 Estimador razão RH2

O estimador $RH2$ do tipo razão para μ, sob a estratégia de Hartley, é considerado para a situação em que existe informação apenas para o total da variável auxiliar nos cadastros. Considerando novamente o caso em que um plano de amostragem aleatória simples é utilizado, tem-se que

$$\overline{y}_{RH2} = \left(\frac{w_a\tilde{y}_a + p w_{ab}\tilde{y}_{ab(A)}}{w_a\tilde{x}_a + p w_{ab}\tilde{x}_{ab(A)}}\right)\mu_{X_A} + \left(\frac{w_b\tilde{y}_b + (1-p) w_{ab}\tilde{y}_{ab(B)}}{w_b\tilde{x}_b + (1-p) w_{ab}\tilde{x}_{ab(A)}}\right)\mu_{X_B}, \tag{5.3}$$

onde w_a, w_b, w_{ab}, $\tilde{y}_a$, $\tilde{y}_b$, $\tilde{y}_{ab(A)}$, $\tilde{y}_{ab(B)}$, $\tilde{x}_a$, $\tilde{x}_b$, $\tilde{x}_{ab(A)}$ e $\tilde{x}_{ab(B)}$ são definidos de maneira análoga à apresentada para o estimador $RH1$. μ_{X_A} e μ_{X_B} são as médias populacionais da variável auxiliar nos cadastros A e B. A variância aproximada deste estimador, expressa em termos matriciais, é dada por

$$\text{AVar}\,(\overline{y}_{RH2}) = \mathbf{d}^T \Sigma_{RH2} \mathbf{d},$$

onde

$$\mathbf{d}^T = \left[\ \frac{\mu_{X_A}}{\mu^*_{x(A)}}\ ,\ \frac{p\mu_{X_A}}{\mu^*_{x(A)}}\ ,\ -\frac{G_A\mu_{X_A}}{\mu^*_{x(A)}}\ ,\ -\frac{pG_A\mu_{X_A}}{\mu^*_{x(A)}}\ ,\ \frac{\mu_{X_B}}{\mu^*_{x(B)}}\ ,\ \frac{(1-p)\mu_{X_B}}{\mu^*_{x(B)}}\ ,\ -\frac{G_B\mu_{X_B}}{\mu^*_{x(B)}}\ ,\ -\frac{G_B(1-p)\mu_{X_B}}{\mu^*_{x(B)}}.\ \right]$$

Na expressão da variância aproximada, $\Sigma_{RH2} = \Sigma_{RH1}$. Tem-se ainda que

$$\begin{aligned} \mu^*_{y(A)} &= \frac{N_a\mu_{ya} + pN_{ab}\mu_{yab(A)}}{N_A} & \mu^*_{x(A)} &= \frac{N_a\mu_{xa} + pN_{ab}\mu_{xab(A)}}{N_A} \\ \mu^*_{y(B)} &= \frac{N_b\mu_{yb} + pN_{ab}\mu_{yab(B)}}{N_B} & \mu^*_{x(B)} &= \frac{N_b\mu_{xb} + pN_{ab}\mu_{xab(B)}}{N_B} \\ G_A &= \frac{\mu^*_{y(A)}}{\mu^*_{x(A)}} & G_B &= \frac{\mu^*_{y(B)}}{\mu^*_{x(B)}} \end{aligned}$$

Ao substituir as quantidades populacionais pelas suas respectivas quantidades amostrais, tem-se um estimador para a variância do estimador $RH2$. Considerando novamente a situação em que apenas o cadastro A possui informação auxiliar disponível, este estimador reduz-se à outra forma de estimador razão apresentada por Coelho (2007), onde

$$\overline{y}_{RH2} = \left(\frac{w_a\tilde{y}_a + pw_{ab}\tilde{y}_{ab(A)}}{w_a\tilde{x}_a + pw_{ab}\tilde{x}_{ab(A)}}\right)\mu_{X_A} + w_b\tilde{y}_b + (1-p)w_{ab}\tilde{y}_{ab(B)}, \tag{5.4}$$

2 Estimadores do tipo razão sob a estratégia de Fuller & Burmeister

Os estimadores do tipo razão sob a estratégia de Hartley são propostos para o caso em que N_a, N_b e N_{ab} são conhecidos. Quando isto não acontecer, e admitindo o uso de um plano de amostragem aleatória simples nos cadastros, e considerando o cenário em que o estimador regressão sob a abordagem de Fuller & Burmeister reduz-se ao estimador do tipo razão, é possível obter os estimadores apresentados a seguir.

2.1 Estimador razão FB1

O estimador $RFB1$ do tipo razão para μ, sob a estratégia de Fuller & Burmeister, é dado por

$$\overline{y}_{RFB1} = \hat{w}_a\frac{\tilde{y}_a}{\tilde{x}_a}\mu_{X_a} + \hat{w}_b\frac{\tilde{y}_b}{\tilde{x}_b}\mu_{X_b} + \hat{w}_{ab}\beta_1\frac{\tilde{y}_{ab(A)}}{\tilde{x}_{ab(A)}}\mu_{X_{ab}} + \hat{w}_{ab}(1-\beta_1)\frac{\tilde{y}_{ab(B)}}{\tilde{x}_{ab(B)}}\mu_{X_{ab}}, \tag{5.5}$$

onde

- $\hat{w}_a = (N_A - \hat{N}_{ab,s})/N = \hat{N}_a/N$, $\hat{w}_b = (N_B - \hat{N}_{ab,s})/N = \hat{N}_b/N$ e $\hat{w}_{ab} = \hat{N}_{ab,s}/N$ são estimadores de w_a, w_b e w_{ab}. $\hat{N}_{ab,s}$ é definido como a menor raiz de (2.13), e os demais termos são definidos de maneira análoga à apresentada para o estimador razão sob a estratégia de Hartley.

Aplicando o método de linearização de Taylor em cada um dos termos do estimador apresentado em (5.5), é possível obter um pseudo-estimador de $\bar{y}_{RFB1}$, para que seja possível a obtenção de uma variância aproximada do estimador. Tem-se

$$A\text{Var}(\bar{y}_{RFB1}) = N^{-2}\mathbf{d}^T\Sigma_{RFB1}\mathbf{d} + O(1), \tag{5.6}$$

onde

$$\mathbf{d} = \begin{bmatrix} 1 \\ \beta_1 \\ -R_a \\ -\beta_1 R_{ab(A)} \\ 1 \\ 1-\beta_1 \\ -R_b \\ -(1-\beta_1)R_{ab(B)} \\ 1 \\ 1 \end{bmatrix}$$

com $R_a = \mu_{ya}/\mu_{xa}$, $R_a = \mu_{yab(A)}/\mu_{xab(A)}$, $R_b = \mu_{yb}/\mu_{xb}$ e $R_{ab(B)} = \mu_{yab(B)}/\mu_{xab(B)}$. A matriz de variâncias e covariâncias do estimador é dada por

$$\Sigma_{RFB1} = \begin{pmatrix} \Sigma_{y(a;ab(A))} & \Sigma_{xy(a;ab(A))} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \Sigma_{xy(a;ab(A))} & \Sigma_{x(a;ab(A))} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Sigma_{y(b;ab(B))} & \Sigma_{xy(b;ab(B))} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Sigma_{xy(b;ab(B))} & \Sigma_{x(b;ab(B))} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \Sigma_{N^{ab}_{(A,B)}} \end{pmatrix}$$

as submatrizes de Σ_{RFB1} com exceção de $\Sigma_{N^{ab}_{(A,B)}}$ são iguais às submatrizes de Σ_{RH1}. Tem-se ainda que

$$\Sigma_{N^{ab}_{(A,B)}} = \begin{pmatrix} \text{Var}\left(\hat{N}^A_{ab}\right) & 0 \\ 0 & \text{Var}\left(\hat{N}^B_{ab}\right) \end{pmatrix}$$

Ao substituir as quantidades populacionais pelas suas respectivas quantidades amostrais, tem-se um estimador para a variância do estimador $RFB1$. Considerando novamente a situação em

que apenas o cadastro A possui informação auxiliar disponível, este estimador reduz-se à outra forma de estimador razão apresentada e analisada por Coelho (2007), onde

$$\overline{y}_{RFB1} = \hat{w}_a \frac{\tilde{y}_a}{\tilde{x}_a} \mu_{X_a} + \hat{w}_{ab} \beta_1 \frac{\tilde{y}_{ab(A)}}{\tilde{x}_{ab(A)}} \mu_{X_{ab}} + \hat{w}_b \tilde{y}_b + \hat{w}_{ab}(1-\beta_1)\tilde{y}_{ab(B)}. \tag{5.7}$$

2.2 Estimador razão RFB2

O estimador $RFB2$ do tipo razão para μ, sob a estratégia de Fuller & Burmeister, é considerado para a situação em que existe informação apenas para o total da variável auxiliar nos cadastros. Considerando-se novamente o caso em que um plano de amostragem aleatória simples é utilizado em ambos os cadastros, tem-se que

$$\overline{y}_{RFB2} = \left(\frac{\hat{w}_a \tilde{y}_a + \beta_1 \hat{w}_{ab} \tilde{y}_{ab(A)}}{\hat{w}_a \tilde{x}_a + \beta_1 \hat{w}_{ab} \tilde{x}_{ab(A)}} \right) \mu_{X_A} + \left(\frac{\hat{w}_b \tilde{y}_b + (1-\beta_1)\hat{w}_{ab} \tilde{y}_{ab(B)}}{w_b \tilde{x}_b + (1-\beta_1)\hat{w}_{ab} \tilde{x}_{ab(A)}} \right) \mu_{X_B}, \tag{5.8}$$

onde μ_{X_A} e μ_{X_B} são as médias populacionais para as únicas variáveis auxiliares nos cadastros A e B, e as demais quantidades são definidas de forma análoga às quantidades apresentadas para o estimador $RFB1$. Sob a situação em que apenas uma variável está disponível para um dos cadastro, A por exemplo, tem-se que

$$\overline{y}_{RFB2} = \left(\frac{\hat{w}_a \tilde{y}_a + \beta_1 \hat{w}_{ab} \tilde{y}_{ab(A)}}{\hat{w}_a \tilde{x}_a + \beta_1 \hat{w}_{ab} \tilde{x}_{ab(A)}} \right) \mu_{X_A} + \hat{w}_b \tilde{y}_b + \hat{w}_{ab}(1-p)\tilde{y}_{ab(B)}, \tag{5.9}$$

com variância aproximada igual a

$$\mathrm{AVar}(\overline{y}_{RFB2}) = N_v^{-2} \left[t_{X_A}^2 \mathrm{AVar}(\hat{G}_A) + \mathrm{AVar}(\hat{G}_B) \right],$$

onde t_{X_A} representa o total populacional da variável auxiliar no cadastro A. Tem-se que

$$\begin{aligned} A\mathrm{Var}(\hat{G}_A) &= \left(\frac{1}{t_{x_A}^*} \right)^2 \left\{ \left(\mu_{ya}^2 + \beta_1^2 \mu_{yab(A)} + G_A^2 \mu_{xa}^2 + p^2 G_A^2 \mu_{xab(A)}^2 \right) \frac{N_a N_b N_{ab} g_A g_B}{n_A N_b g_B + n_B N_a g_A} \right. \\ &+ N_a^2 \left(\frac{1-f_A}{n_A P_a} \right) \sigma_{ya}^2 + \beta_1^2 N_{ab}^2 \left(\frac{1-f_{\mathscr{A}}}{n_{\mathscr{A}} P_{ab(A)}} \right) \sigma_{yab(A)}^2 + G_A^2 N_a^2 \left(\frac{1-f_A}{n_A P_a} \right) \sigma_{xa}^2 \\ &+ \left. \beta_1^2 G_A^2 N_{ab}^2 \left(\frac{1-f_A}{n_A P_{ab(A)}} \right) \sigma_{xab(A)}^2 - 2 G_A N_a^2 \frac{n_A}{n_a^2} \left(1 - \frac{n_A}{N_A} \right) \rho_{xya} \sigma_{xa} \sigma_{ya} \right. \end{aligned}$$

$$- \quad 2\beta_1^2 G_A N_{ab}^2 \frac{n_{\mathscr{A}}}{n_{ab(A)}^2}\left(1-\frac{n_{\mathscr{A}}}{N_{\mathscr{A}}}\right)\rho_{xyab(A)}\sigma_{xab(A)}\sigma_{yab(A)}\Bigg\},$$

onde $G_{\mathscr{A}} = \dfrac{t_{y\mathscr{A}}^*}{t_{x\mathscr{A}}^*}$, com $t_{y\mathscr{A}}^* = N_a\mu_{ya} + pN_{ab}\mu_{yab(A)}$, $t_{x\mathscr{A}}^* = N_a\mu_{xa} + pN_{ab}\mu_{xab(A)}$ e σ_{xa}^2, $\sigma_{xab(A)}^2$, σ_{ya}^2 e $\sigma_{yab(A)}^2$ representam as variâncias das variáveis nos domínios. Tem-se ainda que

$$\begin{aligned} A\text{Var}(\hat{G}_B) &= G_B + \left\{(1-\beta_1)^2(\mu_{yab(B)}^2) + (\mu_{yb}^2)\right\}\frac{N_aN_bN_{ab}g_Ag_B}{n_AN_bg_B + n_BN_ag_A} \\ &+ (1-\beta_1)^2N_{ab}^2\left(\frac{1-f_B}{n_BP_{ab(B)}}\right)\sigma_{yab(B)}^2 + N_b^2\left(\frac{1-f_{\mathscr{B}}}{n_{\mathscr{B}}P_b}\right)\sigma_{yb}^2 \end{aligned}$$

onde $G_{\mathscr{B}} = N_b\mu_{yb} + (1-p)N_{ab}\mu_{yab(B)}$, $P_a = n_a/n_A$, $P_{ab(A)} = n_ab(A)/n_A$, $P_b = n)b/n_B$ e $P_{ab(B)} = n_{ab(B)}/n_B$. A derivação da variância do estimador $RFB2$, bem como a obtenção do valor de β_1 que minimiza sua variância pode ser consultada em Coelho (2007).

3 Estimadores do tipo razão sob a estratégia BLG

Sob a estratégia adotada por Bankier, Lepkowski & Groves (1986), e considerando que existe informação de uma variável auxiliar em ambos os cadastros, o estimador $RBLG$ do tipo razão para μ, sob a estratégia BLG é dado por

$$\bar{y}_{RBLG} = \frac{\dfrac{\sum\limits_{k\in S_A}\omega_k y_k}{\sum\limits_{k\in S_A}\omega_k x_k}t_{X_A} + \dfrac{\sum\limits_{k\in S_B}\omega_k y_k}{\sum\limits_{k\in S_B}\omega_k x_k}t_{X_B}}{N} \tag{5.10}$$

em que $\tilde{y}_a$, $\tilde{x}_a$, $\tilde{y}_b$ e $\tilde{x}_b$ representam as médias amostrais das variáveis de interesse e das variáveis auxiliares em cada cadastro, respectivamente. Através do método de linearização de Taylor, tem-se que a variância aproximada do estimador de Bankier é dada por

$$\begin{aligned} A\text{Var}(\bar{y}_{RBLG}) &= (w_A)^2\,\text{Var}(\tilde{y}_a) + \left(w_A\frac{\mu_{Y_A}}{\mu_{X_A}}\right)^2\text{Var}(\tilde{x}_a) \\ &+ (w_B)^2\,\text{Var}(\tilde{y}_b) + \left(w_B\frac{\mu_{Y_B}}{\mu_{X_B}}\right)^2\text{Var}(\tilde{x}_b) \\ &- 2(w_A)^2\frac{\mu_{Y_A}}{\mu_{X_A}}\text{Cov}(\tilde{y}_a,\tilde{x}_a) - 2(w_B)^2\frac{\mu_{Y_B}}{\mu_{X_B}}\text{Cov}(\tilde{y}_b,\tilde{x}_b). \end{aligned} \tag{5.11}$$

Em termos matriciais, a variância aproximada do estimador RBLG é dada por

$$A\text{Var}\left(\overline{y}_{RBLG}\right) = \mathbf{d}^T \Sigma_{RBLG} \mathbf{d},$$

em que

$$\mathbf{d} = \begin{bmatrix} w_A \\ w_A \dfrac{\mu_{Y_A}}{\mu_{X_A}} \\ w_B \\ w_B \dfrac{\mu_{Y_B}}{\mu_{X_B}} \end{bmatrix} \quad \text{e} \quad \Sigma_{RBLG} = \begin{pmatrix} \sigma_{yA}^2 & \sigma_{xyA} & 0 & 0 \\ \sigma_{xyA} & \sigma_{xA}^2 & 0 & 0 \\ 0 & 0 & \sigma_{yB}^2 & \sigma_{xyB} \\ 0 & 0 & \sigma_{xyB} & \sigma_{xB}^2 \end{pmatrix}$$

Da mesma forma que a vista para o estimador regressão, é possível propor uma versão *raking* para este estimador, da seguinte forma:

$$\begin{aligned} \overline{y}_{RBLG} &= \frac{\dfrac{\sum\limits_{k\in S_A} \omega_k^{(r)} y_k}{\sum\limits_{k\in S_A} \omega_k^{(r)} x_k} t_{X_A} + \dfrac{\sum\limits_{k\in S_B} \omega_k^{(r)} y_k}{\sum\limits_{k\in S_B} \omega_k^{(r)} x_k} t_{X_B}}{N} \\ &= w_A \frac{\tilde{y}_{(r)A}}{\tilde{x}_{(r)A}} \mu_{X_A} + w_B \frac{\tilde{y}_{(r)B}}{\tilde{x}_{(r)B}} \mu_{X_B}, \end{aligned} \tag{5.12}$$

em que a variância aproximada é similar à apresentada em (5.11).

4 Estimadores do tipo razão sob a estratégia de Máxima-Pseudo Verossimilhança

Sob esta estratégia, a proposta de estimador razão se torna simples. Sob a situação em que é de interesse a estimação de uma média populacional sob a abordagem de cadastro duplo, tem-se que

$$\overline{y}_{RPML} = N^{-1} \left\{ \frac{\left(N_A - \hat{N}_{ab,PML}\right)\tilde{y}_a}{\left(N_A - \hat{N}_{ab,PML}\right)\tilde{x}_a} t_{X_a} + \frac{\hat{N}_{ab,PML}\tilde{y}_{ab(*)}}{\hat{N}_{ab,PML}\tilde{x}_{ab(*)}} t_{X_{ab}} + \frac{\left(N_B - \hat{N}_{ab,PML}\right)\tilde{y}_b}{\left(N_B - \hat{N}_{ab,PML}\right)\tilde{x}_b} t_{X_b} \right\}$$

$$= \quad N^{-1}\left\{\frac{\tilde{y}_a}{\tilde{x}_a}t_{X_a}+\frac{\tilde{y}_{ab(*)}}{\tilde{x}_{ab(*)}}t_{X_{ab}}+\frac{\tilde{y}_b}{\tilde{x}_b}t_{X_b}\right\}$$

Como visto anteriormente, através do método de linearização de Taylor, um pseudo-estimador para $\tilde{y}_{ab(*)}$ é dado por

$$\begin{aligned}\tilde{y}_{ab(*)} &\approx \left(\frac{\pi^A}{\pi^A+\pi^B}\right)\tilde{y}_{ab(A)}+\left(\frac{\pi^B}{\pi^A+\pi^B}\right)\tilde{y}_{ab(B)}\\ &= p_A\tilde{y}_{ab(A)}+p_B\tilde{y}_{ab(B)}=p_A\tilde{y}_{ab(A)}+(1-p_A)\tilde{y}_{ab(B)}=p_A\left(\tilde{y}_{ab(A)}-\tilde{y}_{ab(B)}\right)+\tilde{y}_{ab(B)}=g(\tilde{y}_{ab(A)},\tilde{y}_{ab(B)})\end{aligned}$$

Dessa forma, tem-se então que

$$\frac{g(\tilde{y}_{ab(A)},\tilde{y}_{ab(B)})}{g(\tilde{x}_{ab(A)},\tilde{x}_{ab(B)})} \approx \frac{g(\mu_{yab},\mu_{yab})}{g(\mu_{xab},\mu_{xab})}+\frac{1}{g(\mu_{xab},\mu_{xab})}\left(g(\tilde{y}_{ab(A)},\tilde{y}_{ab(B)})-\frac{g(\mu_{yab},\mu_{yab})}{g(\mu_{xab},\mu_{xab})}g(\tilde{x}_{ab(A)},\tilde{x}_{ab(B)})\right)$$

É possível notar que $g(\mu_{yab},\mu_{yab})=\mu_{yab}$ e $g(\mu_{xab},\mu_{xab})=\mu_{xab}$. Dessa forma,

$$\frac{\tilde{y}_{ab(*)}}{\tilde{x}_{ab(*)}}=\frac{g(\tilde{y}_{ab(A)},\tilde{y}_{ab(B)})}{g(\tilde{x}_{ab(A)},\tilde{x}_{ab(B)})} \approx \frac{\mu_{yab}}{\mu_{xab}}+\frac{1}{\mu_{xab}}\left(g(\tilde{y}_{ab(A)},\tilde{y}_{ab(B)})-\frac{\mu_{yab}}{\mu_{xab}}g(\tilde{x}_{ab(A)},\tilde{x}_{ab(B)})\right)$$

Logo,

$$\begin{aligned}\overline{y}_{RPML} &\approx N^{-1}\left\{\left[\frac{\mu_{ya}}{\mu_{xa}}+\frac{1}{\mu_{xa}}\left(\tilde{y}_a-\frac{\mu_{ya}}{\mu_{xa}}\tilde{x}_a\right)\right]t_{X_a}+\left[\frac{\mu_{yb}}{\mu_{xb}}+\frac{1}{\mu_x^b}\left(\tilde{y}_b-\frac{\mu_{yb}}{\mu_{xb}}\tilde{x}_b\right)\right]t_{X_b}\right.\\ &+ \left.\left[\frac{\mu_{yab}}{\mu_{xab}}+\frac{1}{\mu_{xab}}\left(g(\tilde{y}_{ab(A)},\tilde{y}_{ab(B)})-\frac{\mu_{yab}}{\mu_{xab}}g(\tilde{x}_{ab(A)},\tilde{x}_{ab(B)})\right)\right]t_{X_{ab}}\right\}\end{aligned}$$

$$\begin{aligned}A\mathrm{Var}(\overline{y}_{RPML}) &= \left(\frac{t_{X_a}}{\mu_{xa}}\right)^2\mathrm{Var}(\tilde{y}_a)+\left(\frac{\mu_{ya}}{\mu_{xa}}t_{X_a}\right)^2\mathrm{Var}(\tilde{x}_a)-2\mu_{ya}\left(\frac{t_{X_a}}{\mu_{xa}}\right)^2\mathrm{Cov}(\tilde{y}_a;\tilde{x}_a)\\ &+ \left(\frac{t_{X_b}}{\mu_x^b}\right)^2\mathrm{Var}(\tilde{y}_b)+\left(\frac{\mu_{yb}}{\mu_{xb}}t_{X_b}\right)^2\mathrm{Var}(\tilde{x}_b)-2\mu_{ya}\left(\frac{t_{X_b}}{\mu_x^b}\right)^2\mathrm{Cov}(\tilde{y}_b;\tilde{x}_b)\\ &+ p_A^2\left(\frac{t_{X_{ab}}}{\mu_{xab}}\right)^2\mathrm{Var}\left(\tilde{y}_{ab(A)}\right)+p_A^2\left(\frac{\mu_{yab}}{\mu_{xab}}\right)^2\mathrm{Var}\left(\tilde{x}_{ab(A)}\right)-2p_A^2\mu_{yab}\left(\frac{t_{X_{ab}}}{\mu_{xab}}\right)^2\mathrm{Cov}\left(\tilde{y}_{ab(A)};\tilde{x}_{ab(A)}\right)\end{aligned}$$

$$
\begin{aligned}
&+ \quad p_B^2 \left(\frac{t_{X_{ab}}}{\mu_{xab}}\right)^2 \text{Var}\left(\tilde{y}_{ab(B)}\right) + p_B^2 \left(\frac{\mu_{yab}}{\mu_{xab}}\right)^2 \text{Var}\left(\tilde{x}_{ab(B)}\right) - 2p_B^2 \mu_{yab} \left(\frac{t_{X_{ab}}}{\mu_{xab}}\right)^2 \text{Cov}\left(\tilde{y}_{ab(B)}; \tilde{x}_{ab(B)}\right) \\
&= \quad \mathbf{d}^T \Sigma_{RPML} \mathbf{d},
\end{aligned}
$$

em que $\mathbf{d}^T$ e Σ_{RPML} são dados por

$$\mathbf{d}^T = [\mathbf{d}_1^T \quad \mathbf{d}_2^T]$$

$$\mathbf{d}_1^T = \left[\begin{array}{ccccc} \frac{t_{X_a}}{\mu_{xa}} & \frac{\mu_{ya}}{\mu_{xa}} t_{X_a} & \sqrt{\mu_{ya}} \frac{t_{X_a}}{\mu_{xa}} & p_A \frac{t_{X_{ab}}}{\mu_{xab}} & p_A \frac{\mu_{yab}}{\mu_{xab}} \quad p_A \sqrt{\mu_{yab}} \frac{t_{X_{ab}}}{\mu_{xab}} \end{array} \right]$$

$$\mathbf{d}_2^T = \left[\begin{array}{ccccc} \frac{t_{X_b}}{\mu_x^b} & \frac{\mu_{yb}}{\mu_{xb}} t_{X_b} & \sqrt{\mu_y^b} \frac{t_{X_b}}{\mu_x^b} & p_B \frac{t_{X_{ab}}}{\mu_{xab}} & p_B \frac{\mu_{yab}}{\mu_{xab}} \quad p_B \sqrt{\mu_{yab}} \frac{t_{X_{ab}}}{\mu_{xab}} \end{array} \right]$$

$$\Sigma_{RPML} = \begin{pmatrix} \Sigma_A & \mathbf{0} \\ \mathbf{0} & \Sigma_B \end{pmatrix}$$

em que

$$\Sigma_A = \begin{pmatrix} \sigma_{ya}^2 & \sigma_{y(a;ab(A))} & \sigma_{xya} & \sigma_{xy(a;ab(A))} \\ \sigma_{y(a;ab(A))} & \sigma_{yab}^2 & \sigma_{xy(a;ab(A))} & \sigma_{xy(a;ab(A))} \\ \sigma_{xya} & \sigma_{xy(a;ab(A))} & \sigma_{xa}^2 & \sigma_{x(a;ab(A))} \\ \sigma_{xy(a;ab(A))} & \sigma_{xa}^2 & \sigma_{x(a;ab(A))} & \sigma_{xab(A)}^2 \end{pmatrix}$$

$$\Sigma_B = \begin{pmatrix} \sigma_{yb}^2 & \sigma_{y(b;ab(B))} & \sigma_{xyb} & \sigma_{xy(b;ab(B))} \\ \sigma_{y(b;ab(B))} & \sigma_{yab}^2 & \sigma_{xy(b;ab(B))} & \sigma_{xy(b;ab(B))} \\ \sigma_{xyb} & \sigma_{xy(b;ab(B))} & \sigma_{xb}^2 & \sigma_{x(b;ab(B))} \\ \sigma_{xy(b;ab(B))} & \sigma_{xb}^2 & \sigma_{x(b;ab(B))} & \sigma_{xab(B)}^2 \end{pmatrix}$$

Sob um plano de amostragem aleatória simples, é possível observar que a variância aproximada da estratégia PML coincide com a variância do estimador razão $R1$ de Hartley. Como visto, a informação de uma variável auxiliar pode ser alocada no processo de estimação sob cada estratégia proposta. Qual das estratégias será a melhor vai depender do cenário em questão. Se os cadastros estiverem sob o cenário 2 (N_A, N_B, N_a, N_b e N_{ab} conhecidos), os estimadores de Hartley e Bankier podem ser utilizados, sendo a estratégia de Bankier uma estratégia mais simples de implementação, caso exista grande quantidade de informação no

domínio ab. Já se o interesse for obter uma maior precisão, a estratégia de Hartley deve ser considerada, uma vez que a constante p é escolhida de modo a minimizar a variância do estimador utilizado. Caso a situação seja a do cenário 3 (N_A, N_B, N_a, N_b e N_{ab} desconhecidos), os estimadores de Fuller & Burmeister e PML podem ser utilizados, sendo a estratégia de Fuller & Burmeister mais eficiente que a PML pela mesma justificativa apresentada para a estratégia de Hartley. Para ambos os cenários, as estratégias de Bankier e PML surgem como alternativas mais direcionadas à agilidade de obtenção de resultados do que eficiência.

CAPÍTULO 6

Avaliação Numérica

1 Discussão metodológica

Os desempenhos dos estimadores propostos nesta tese foram avaliados através do método de simulação de Monte Carlo. Admitindo-se que em ambos os cadastros exista informação de duas variáveis auxiliares disponíveis, o cenário considerado foi que a relação entre a variável de interesse e as variáveis auxiliares é descrita pelo modelo

$$\xi: \; y_k = \beta_0 + \beta_1 x_{k1} + \cdots + \beta_q x_{kq} + \varepsilon_k = \beta_0 + \sum_{r=1}^{q} x_{kr} + \varepsilon_k = \mathbf{x}_k^T \underset{\sim}{\beta} + \varepsilon_k, \tag{6.1}$$

onde foi fixado $q = 2$ em ambos os cadastros. Os estimadores apresentados nos capítulos 2 e os estimadores regressão propostos no capítulo 3 foram avaliados de acordo com o plano de amostragem aleatória simples aplicado em cada cadastro.

Foi admitido que $N_A + N_B - N_{ab} = 4500$, conforme descrição na tabela (6.1) a seguir. Para cada tamanho populacional, foram consideradas amostras de tamanho $n_A = n_B = 125$, 250, 500, 750, 1000, 1250.

Tabela 6.1: Tamanhos populacionais utilizados para a, b e ab

POPULAÇÃO	N_a	N_b	N_{ab}	$N_A = N_B$
1	2000	2000	500	2500
2	1750	1750	1000	2750
3	1500	1500	1500	3000

Para a avaliação dos estimadores foram considerados o viés relativo, desvio padrão e

erro quadrático médio dos estimadores para a média populacional. O viés de um estimador $\hat{\theta}$ é dado por

$$B(\hat{\theta}) = E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta$$

e o erro quadrático médio por

$$EQM(\hat{\theta}) = \text{Var}(\hat{\theta}) + \{B(\hat{\theta})\}^2.$$

O viés relativo é dado por

$$VR(\hat{\theta}) = \frac{E(\hat{\theta} - \theta)}{\theta}$$

onde θ é o parâmetro de interesse e $\hat{\theta}$ é estimador de θ. O esquema de geração de dados foi conduzido da seguinte forma:

Passo 1. Gerar pares (i, j) referentes às observações das variáveis auxiliares para o domínio a, onde $i, j = 1, \ldots, N_a$, considerando uma distribuição $\mathcal{N}(\mu_{xaq}, \sigma^2_{xaq})$, com $q = 1, 2$ e $\mu_{ax1} = 4$ e $\mu_{ax2} = 8$. Foram ainda considerados dois cenários para a variância das variáveis auxiliares:

1. $\sigma^2_{xa1} = 1$ e $\sigma^2_{xa2} = 1$;
2. $\sigma^2_{xa1} = 10$ e $\sigma^2_{xa2} = 4$.

Passo 2. Gerar pares (i, j) para as variáveis auxiliares no domínio ab, onde $i, j = 1, \ldots, N_{ab}$, a partir de valores de uma distribuição $\mathcal{N}(\mu_{xabq}, \sigma^2_{xabq})$, q=1,2. Analogamente ao item anterior, tem-se que $\mu_{xab1} = 2$, $\mu_{xab1} = 1$, e dois cenários para as variâncias das variáveis auxiliares neste domínio:

1. $\sigma^2_{xab1} = 1$ e $\sigma^2_{xab2} = 1$;
2. $\sigma^2_{xab1} = 10$ e $\sigma^2_{xab2} = 4$.

Passo 3. Gerar N_a observações da variável de interesse y para o domínio a, onde $i = 1, \ldots, N_a$, seguindo distribuição $\mathcal{N}(\beta_0 + \beta^a_1 x_{1kA} + \beta^a_2 x_{2kA}, \sigma^2_k)$, onde $\beta^a_i = 1$ e $\sigma^2_k = \sigma^2_{ya} = 1$;

Passo 4. Gerar N_{ab} observações da variável de interesse para o domínio ab, a partir de uma distribuição $\mathcal{N}(\beta_0 + \beta^{ab}_1 x_{1kA} + \beta^{ab}_2 x_{2kA}, \sigma^2_k)$, onde $\beta^{ab}_i = 1$ e $\sigma^2_k = \sigma^2_{yab} = 1$.

Passo 5. Repetir os passos 1, 2, 3 e 4 para o domínio b e cadastro B. No passos 1, os cenários para as variâncias foram os seguintes:

1. $\sigma^2_{xb1} = 1$, $\sigma^2_{xb2} = 1$;

2. $\sigma^2_{xb1} = 8$ e $\sigma^2_{xb2} = 12$.

Passo 6. Na estrutura de populações geradas nos passos de 1 a 5 para os cadastros A e B, coletar uma amostra de tamanho especificado através do plano de amostragem aleatória simples e obter estimativas para a média populacional utilizando os estimadores dos capítulos 2 e 3;

Passo 7. Obter estimativas de desvio padrão (DP), Viés Relativo (VR), Erro Quadrático Médio (EQM). Além disso, para cada estimador $\hat{\theta}$ é construído um intervalo do tipo

$$\hat{\theta} \pm 1.96\sqrt{\text{Var}(\hat{\theta})}.$$

Este intervalo foi considerado para cálculo da proporção de vezes em que o intervalo de confiança dos estimadores propostos cobre o verdadeiro valor do parâmetro para cada caso. Esta proporção será chamada de *Proporção de Cobertura do Intervalo* (PCI).

As populações geradas nos passos de 1 a 5 permaneceram fixas no estudo de simulação. Utilizando o método de Monte Carlo, os passos 6 e 7 foram repetidos $r = 1000$ vezes, para cada tamanho de amostra especificado. Os resultados são apresentados nas tabelas 6.2 a 6.13 e pelas figuras 6.1 a 6.15. Além dos estimadores para a variável de interesse, foram considerados também no estudo de simulação os estimadores para N_{ab} sob as estratégias de Fuller & Burmeister (1972) e Skinner & Rao (1996), denotados por $\hat{N}_{ab,s}$ e $\hat{N}_{ab,PML}$.

2 Resultados da Simulação

2.1 Desempenho dos estimadores propostos sob o plano de amostragem aleatória simples

O desempenho dos estimadores propostos sob o plano de amostragem aleatória simples, sob os cenários 2 e 3, é apresentado a seguir pelas tabelas (6.2) a (6.13) e figuras (6.1) a (6.6).

De modo geral, é possível observar que os estimadores já propostos na literatura tiveram desempenho similar tanto em relação à proporção de cobertura do intervalo de confiança quanto na redução da variância à medida em que o tamanho de amostra nos cadastros aumenta. Observando-se os cenários tem-se que, sob o cenário 2, o estimador de Hartley apresentou evidência de melhor desempenho em relação ao estimador BLG, como mostram os valores de desvio padrão e EQM. Sob o cenário 3, tem-se evidência de melhor desempenho do estimador PML em relação ao estimador de Fuller & Burmeister, quando observados também os valores de desvio padrão e EQM. Isto foi verificado para todos os valores populacionais de N_{ab}, o que fornece evidência de que em situações onde a variabilidade é idêntica para

todos os cadastros, basta que o cenário em questão seja determinado e utilizado o estimador apropriado para este cenário.

Comparados aos estimadores de Hartley, Fuller, Bankier e PML quando o tamanho de amostra nos cadastros aumenta, as versões propostas no contexto de estimadores regressão apresentaram considerável redução em relação à variância e erro quadrático médio, apresentando um viés relativo maior. Além disso, há evidência de que os estimadores regressão propostos são mais confiáveis para se estimar o parâmetro populacional de interesse, como mostra as PCI's para cada estimador, bem próximas dos níveis nominais de um intervalo de confiança calculado com $z_{1-\alpha/2} = 1.96$. Este desempenho foi o mesmo para todos os valores de N_{ab} apresentados.

Os resultados mostraram também que cada um dos estimadores regressão propostos apresentou viés relativo maior que sua respectiva versão simples apresentada no capítulo 2. Porém, os resultados forneceram evidência de que esse viés tende a diminuir com o aumento do tamanho da amostra em ambos os cadastros.

Na situação em que as variâncias em domínios e cadastros são diferentes, os estimadores regressão considerados sob as estratégias de Hartley (cenário 2) e Fuller & Burmeister (cenário 3) ainda continuaram a apresentar considerável redução da variância em relação aos estimadores já propostos quando o tamanho da amostra aumenta nos cadastros. Os estimadores de Bankier (cenário 2) e PML (cenário 3) apresentaram redução nas PCI's quando o tamanho de N_{ab} aumenta. Isto fornece evidência de que estes estimadores são sensíveis ao aumento da quantidade de informações do domínio ab, pois à medida em que N_{ab} aumenta, é maior a chance de selecionar mais elementos deste domínio, e constatou-se que os estimadores mencionados apresentaram problemas devido a esse fato. Ainda, é possível notar que se for considerada a quantidade $w_{ab} = N_{ab}/N$, onde $N = N_A + N_B - N_{ab}$ (total populacional), tem-se que os estimadores foram avaliados nas situações em que $w^{ab} = 500/4500 = 0.11$, $w_{ab} = 1000/4500 = 0.22$ e $w_{ab} = 1500/4500 = 0.33$. Como mostram as tabelas (6.8.) a (6.12), a PCI para estes estimadores diminuiu quando w_{ab} ficou acima de 0.10 (10%) do total populacional.

A distribuição empírica de cada um dos estimadores propostos foi obtida, como mostram as figuras (6.7) a (6.9), e fornece evidência de que a distribuição assintótica de todos os estimadores propostos é centrada no real valor do parâmetro, que no estudo de simulação foi a média populacional. Para todos os casos, há evidência de que os estimadores obtidos seguem distribuição normal, como apresentado no capítulo 4.

Os gráficos referentes aos desvios padrões dos estimadores revelam que sob o cenário 2, os estimadores regressão sob as estratégias de Hartley e Bankier apresentaram desempenho similar na redução da variância, com pequena vantagem para o estimador regressão sob a estratégia BLG. Sob o cenário 3, o estimador regressão sob a estratégia de Fuller & Burmeister apresentou melhor desempenho para todos os casos. Desconsiderando-se os cenários,

os estimadores regressão sob as estratégias PML e Fuller & Burmeister foram os que apresentaram menor variância em comparação aos demais estimadores, bem como apresentaram também maior redução com o aumento do tamanho da amostra em cada cadastro.

2.2 Estudo sobre a distribuição de $\hat{N}_{ab,s}$ e $\hat{N}_{ab,PML}$

Os estimadores de N_{ab} propostos por Fuller & Burmeister (1972) e Skinner & Rao (1996), e denotados por $\hat{N}_{ab,s}$ e $\hat{N}_{ab,PML}$, são definidos como a solução de equações quadráticas e utilizados nos casos em que não é possível identificar duplicatas nas amostras obtidas de A e B. Além disso, não estão em função dos valores das variáveis utilizadas no processo de estimação. As figuras (6.10) a (6.12) ilustram a distribuição empírica de $\hat{N}_{ab,s}$ e as figuras (6.13) a (6.15) ilustram a distribuição empírica de $\hat{N}_{ab,PML}$. Para cada uma das populações geradas na tabela (6.10), foram obtidas 1000 estimativas de N_{ab} de acordo com as estratégias mencionadas e construídos gráficos de densidade estimada.

É possível verificar que, para $N_{ab} = 500$ por exemplo, à medida em que os tamanhos de amostra aumentam, a distribuição empírica das estimativas fornece evidência de que as distribuições de $\hat{N}_{ab,s}$ e $\hat{N}_{ab,PML}$ convergem para uma distribuição centrada no real valor de N_{ab}, o que sugere distribuição normal. Além disso, verifica-se também que há evidência de redução da variância do estimador para tamanhos de amostra elevados, como o observado para $n_A = n_B = 1250$ em ambos os casos. Para os demais valores de referência para N_{ab} analisados no estudo de simulação, o comportamento observado foi o mesmo, o que evidencia que $\hat{N}_{ab,s}$ e $\hat{N}_{ab,PML}$ são estimadores que garantem uma boa aproximação para N_{ab}, e essa aproximação se torna ainda melhor à medida em que os tamanhos de amostra nos cadastros aumentam.

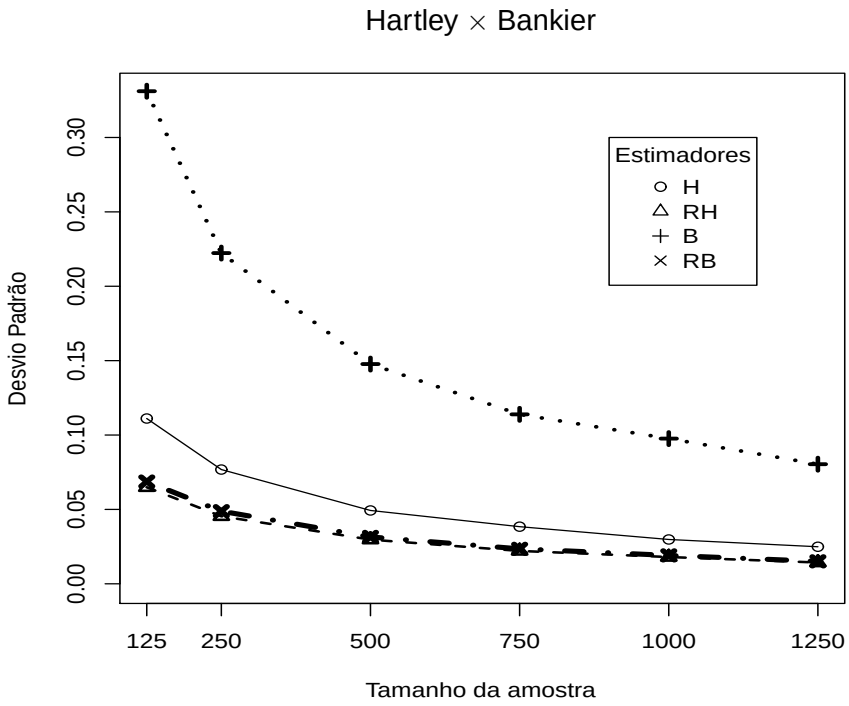
Tabela 6.2: Avaliação das estratégias de Hartley, Fuller, Bankier e PML para $N = 4500$, $N_{ab} = 500$ e desvios unitários.

Cenário	Estimador	$n_A = n_B$	Estatística			
			Desvio Padrão	100×VR	EQM	PCI
2	$\overline{y}_H$	125	0.1111	-0.0191	0.0123	0.953
		250	0.0768	0.0386	0.0059	0.940
		500	0.0493	0.0059	0.0024	0.955
		750	0.0384	0.0189	0.0014	0.952
		1000	0.0298	0.0201	0.0008	0.952
		1250	0.0249	0.0255	0.0006	0.957
	$\overline{y}_{BLG}$	125	0.3311	-0.0453	0.0042	0.948
		250	0.2223	0.0915	0.0496	0.946
		500	0.1477	0.0740	0.0219	0.947
		750	0.1139	-0.0208	0.0219	0.944
		1000	0.0976	0.0446	0.0095	0.959
		1250	0.0804	0.0034	0.0064	0.948
3	$\overline{y}_{FB}$	125	0.1108	-0.0224	0.0122	0.953
		250	0.0861	0.0374	0.0059	0.942
		500	0.0692	0.0058	0.0024	0.953
		750	0.0393	0.0019	0.0014	0.952
		1000	0.0303	0.0019	0.0008	0.951
		1250	0.0278	0.0025	0.0006	0.956
	$\overline{y}_{PML}$	125	0.1108	-0.0002	0.0117	0.950
		250	0.0767	0.0003	0.0045	0.948
		500	0.0491	0.0005	0.0020	0.951
		750	0.0384	0.0001	0.0010	0.951
		1000	0.0298	0.0001	0.0006	0.952
		1250	0.0248	0.0025	0.0005	0.953

Tabela 6.3: Avaliação das estratégias de estimador regressão para $N = 4500$, $N_{ab} = 500$ e desvios unitários.

Cenário	Estimador	$n_A = n_B$	Estatística			
			Desvio Padrão	100×VR	EQM	PCI
2	$\overline{y}_{RH}$	125	0.0649	0.0034	0.0042	0.948
		250	0.0454	0.0044	0.0021	0.939
		500	0.0297	0.0240	0.0009	0.956
		750	0.0221	0.0192	0.0004	0.949
		1000	0.0180	0.0204	0.0003	0.950
		1250	0.0143	0.0194	0.0002	0.947
	$\overline{y}_{RBLG}$	125	0.0687	0.0511	0.0047	0.956
		250	0.0488	0.0629	0.0024	0.939
		500	0.0315	0.0400	0.0010	0.954
		750	0.0235	0.0342	0.0005	0.940
		1000	0.0191	0.0375	0.0004	0.943
		1250	0.0151	0.0392	0.0002	0.942
3	$\overline{y}_{RFB}$	125	0.1108	-0.0224	0.0123	0.953
		250	0.0442	0.0420	0.0020	0.944
		500	0.0287	0.0230	0.0008	0.952
		750	0.0215	0.0189	0.0005	0.940
		1000	0.0176	0.0184	0.0003	0.956
		1250	0.0138	0.0203	0.0001	0.950
	$\overline{y}_{RPML}$	125	0.0821	0.0218	0.0067	0.952
		250	0.0544	0.0394	0.0029	0.951
		500	0.0358	0.0258	0.0013	0.949
		750	0.0280	0.0142	0.0008	0.955
		1000	0.0229	0.0116	0.0005	0.943
		1250	0.0184	0.0053	0.0003	0.947

Figura 6.1: Desvio Padrão dos estimadores propostos para $N = 4500$, $N_{ab} = 500$ e desvios unitários.



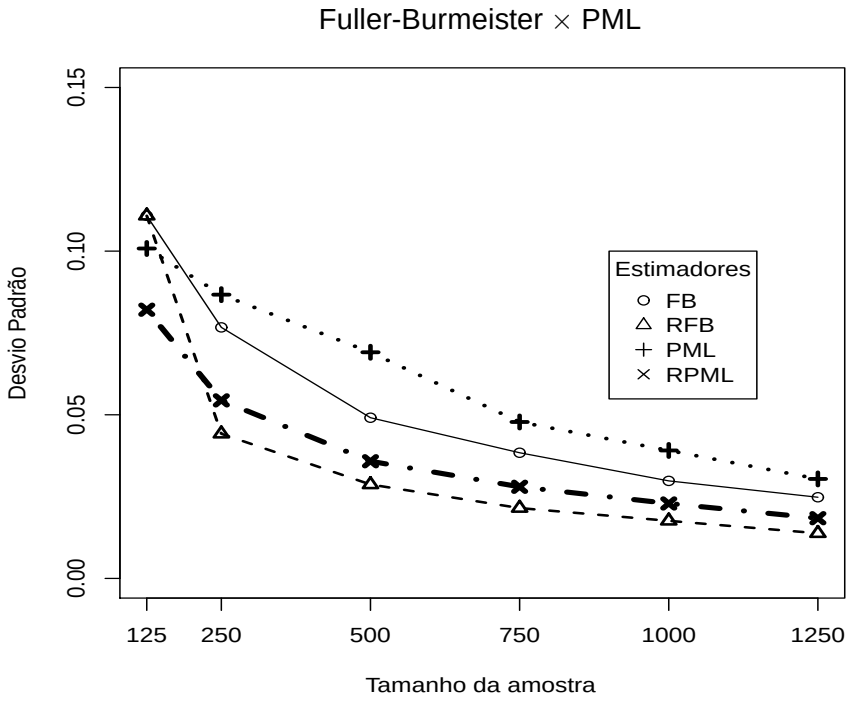


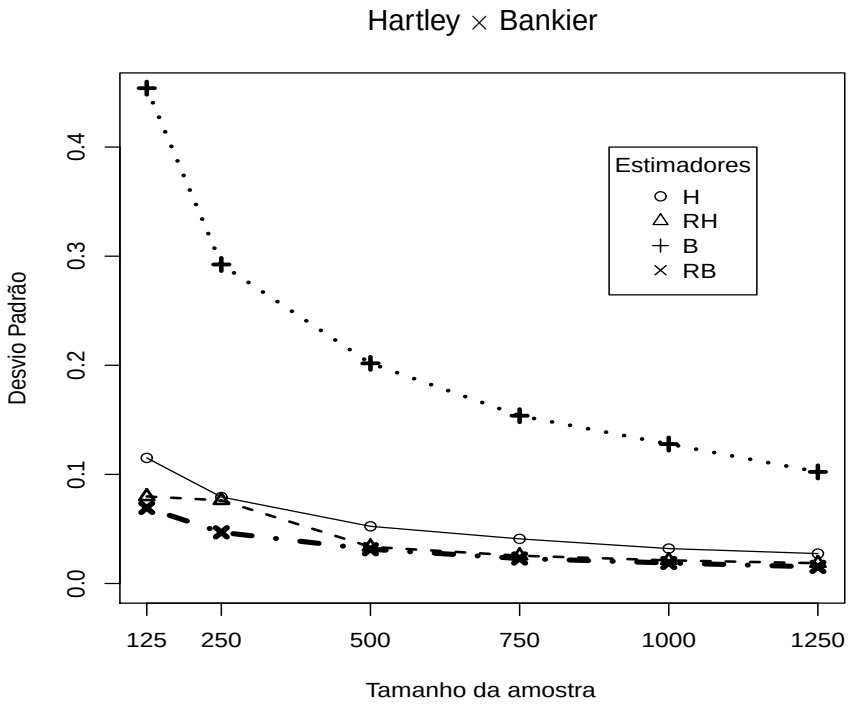
Tabela 6.4: Avaliação das estratégias de Hartley, Fuller, Bankier e PML para $N = 4500$, $N_{ab} = 1000$ e desvios unitários.

Cenário	Estimador	$n_A = n_B$	Estatística			
			Desvio Padrão	100×VR	EQM	PCI
2	$\overline{y}_H$	125	0.1151	0.0559	0.0133	0.945
		250	0.0793	0.0470	0.0063	0.947
		500	0.0524	0.0414	0.0028	0.949
		750	0.0409	0.0385	0.0017	0.950
		1000	0.0320	0.0219	0.0010	0.950
		1250	0.0274	0.0267	0.0007	0.949
	$\overline{y}_{BLG}$	125	0.4540	0.0938	0.2062	0.951
		250	0.2924	0.0149	0.0855	0.951
		500	0.2018	0.0967	0.0409	0.947
		750	0.1538	0.0647	0.0237	0.940
		1000	0.1278	0.0578	0.0163	0.942
		1250	0.1023	0.0679	0.0105	0.942
3	$\overline{y}_{FB}$	125	0.1154	0.0570	0.0133	0.943
		250	0.0792	0.0112	0.0063	0.946
		500	0.0525	0.0405	0.0028	0.950
		750	0.0410	0.0384	0.0017	0.950
		1000	0.0320	0.0219	0.0010	0.949
		1250	0.0274	0.0266	0.0008	0.947
	$\overline{y}_{PML}$	125	0.1125	0.0112	0.0126	0.940
		250	0.0761	0.0123	0.0058	0.949
		500	0.0518	0.0590	0.0026	0.953
		750	0.0399	0.0189	0.0016	0.955
		1000	0.0321	0.0002	0.0010	0.949
		1250	0.0274	0.0003	0.0008	0.947

Tabela 6.5: Avaliação das estratégias de estimador regressão para $N = 4500$, $N_{ab} = 1000$ e desvios unitários.

Cenário	Estimador	$n_A = n_B$	Estatística			
			Desvio Padrão	100×VR	EQM	PCI
2	$\overline{y}_{RH}$	125	0.0798	0.0138	0.0056	0.948
		250	0.0762	0.0118	0.0058	0.953
		500	0.0336	0.0656	0.0011	0.941
		750	0.0255	0.0121	0.0006	0.950
		1000	0.0212	0.0034	0.0004	0.945
		1250	0.0186	0.0032	0.0003	0.946
	$\overline{y}_{RBLG}$	125	0.0693	0.1228	0.0051	0.938
		250	0.0469	0.1304	0.0025	0.934
		500	0.0314	0.1258	0.0012	0.921
		750	0.0229	0.1197	0.0008	0.908
		1000	0.0188	0.1128	0.0006	0.910
		1250	0.0152	0.1143	0.0004	0.900
3	$\overline{y}_{RFB}$	125	0.0674	0.0188	0.0045	0.946
		250	0.0446	0.0155	0.0019	0.946
		500	0.0308	0.0045	0.0009	0.944
		750	0.0234	0.0059	0.0005	0.949
		1000	0.0189	0.0044	0.0004	0.939
		1250	0.0162	0.0029	0.0003	0.947
	$\overline{y}_{RPML}$	125	0.0817	0.1120	0.0069	0.950
		250	0.0539	0.0790	0.0030	0.946
		500	0.0357	0.0008	0.0014	0.942
		750	0.0267	0.1006	0.0008	0.919
		1000	0.0210	0.1033	0.0006	0.913
		1250	0.0177	0.0991	0.0004	0.915

Figura 6.2: Desvio Padrão das estratégias de estimador regressão para $N = 4500$, $N_{ab} = 1000$ e desvios unitários.



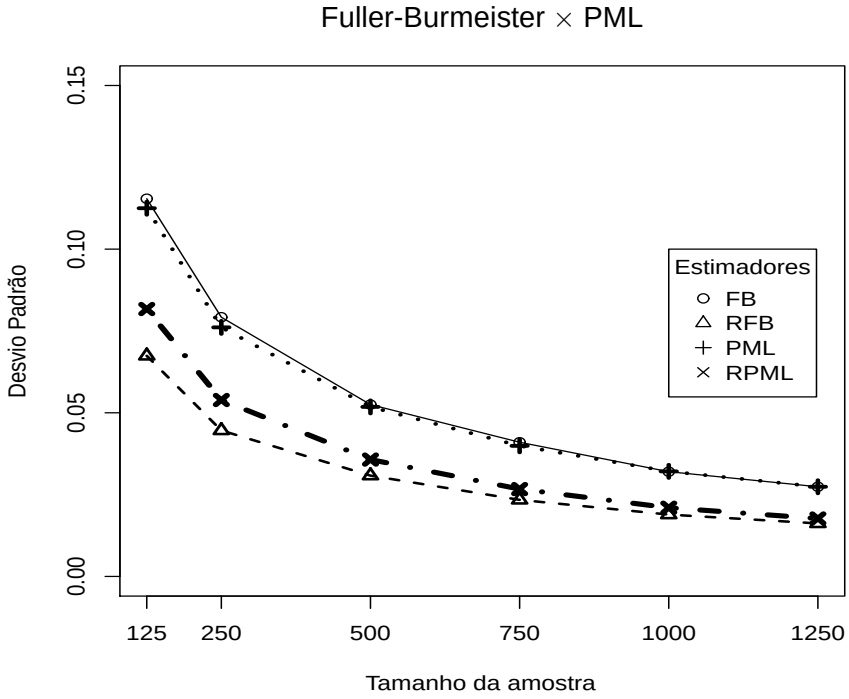


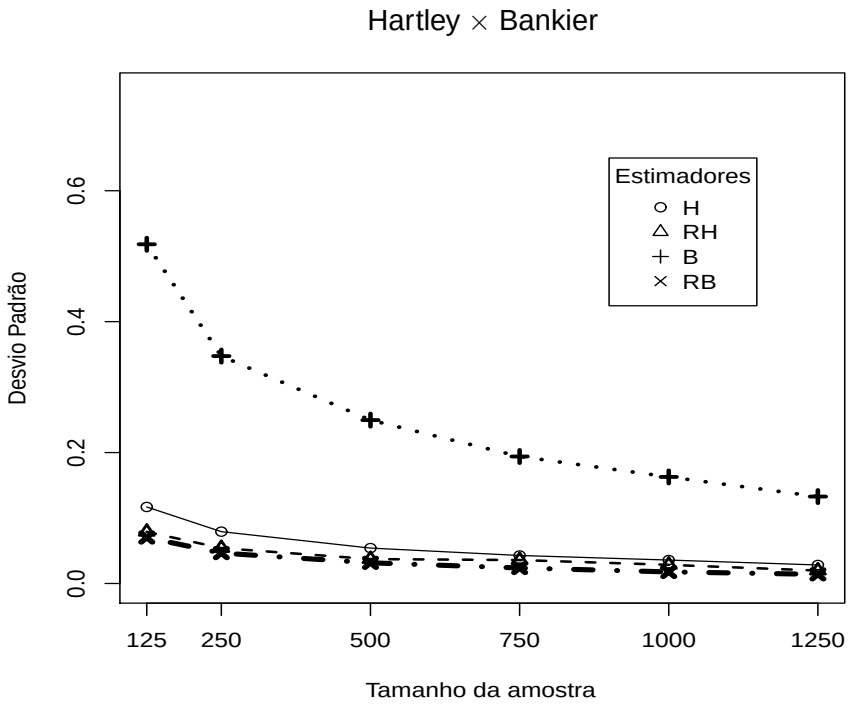
Tabela 6.6: Avaliação das estratégias de Hartley, Fuller, Bankier e PML para $N = 4500$, $N_{ab} = 1500$ e desvios unitários.

Cenário	Estimador	$n_A = n_B$	Estatística			
			Desvio Padrão	100×VR	EQM	PCI
2	$\overline{y}_H$	125	0.1169	0.0430	0.0137	0.952
		250	0.0792	0.0341	0.0063	0.952
		500	0.0541	0.0056	0.0029	0.950
		750	0.0428	0.0014	0.0018	0.953
		1000	0.0358	0.0012	0.0013	0.950
		1250	0.0283	0.0012	0.0008	0.950
	$\overline{y}_{BLG}$	125	0.5182	0.1671	0.2688	0.949
		250	0.3474	0.0975	0.1207	0.954
		500	0.2496	0.0448	0.0623	0.949
		750	0.1940	0.0676	0.0376	0.958
		1000	0.1628	0.0228	0.0265	0.956
		1250	0.1328	0.0149	0.0176	0.950
3	$\overline{y}_{FB}$	125	0.1164	0.0388	0.0135	0.946
		250	0.0792	0.0324	0.0062	0.953
		500	0.0541	0.0499	0.0029	0.947
		750	0.0427	0.0116	0.0018	0.953
		1000	0.0358	0.0133	0.0013	0.950
		1250	0.0284	0.0014	0.0008	0.950
	$\overline{y}_{PML}$	125	0.1194	0.0458	0.0195	0.946
		250	0.0892	0.0344	0.0082	0.951
		500	0.0642	0.0399	0.0059	0.949
		750	0.0527	0.0316	0.0019	0.950
		1000	0.0378	0.0233	0.0017	0.950
		1250	0.0289	0.0230	0.0009	0.939

Tabela 6.7: Avaliação das estratégias de estimador regressão para $N = 4500$, $N_{ab} = 1500$ e desvios unitários.

Cenário	Estimador	$n_A = n_B$	Estatística			
			Desvio Padrão	100×VR	EQM	PCI
2	$\overline{y}_{RH}$	125	0.0789	0.0261	0.0062	0.955
		250	0.0541	0.0153	0.0029	0.948
		500	0.0373	0.0158	0.0013	0.953
		750	0.0358	0.0023	0.0008	0.954
		1000	0.0285	0.0134	0.0013	0.950
		1250	0.0197	0.0008	0.0004	0.952
	$\overline{y}_{RBLG}$	125	0.0703	0.1547	0.0053	0.940
		250	0.0465	0.1221	0.0024	0.943
		500	0.0316	0.1296	0.0012	0.904
		750	0.0237	0.1289	0.0008	0.901
		1000	0.0177	0.1327	0.0006	0.931
		1250	0.0140	0.1410	0.0005	0.899
3	$\overline{y}_{RFB}$	125	0.0684	0.0189	0.0047	0.941
		250	0.0461	0.0083	0.0021	0.951
		500	0.0311	0.0150	0.0009	0.949
		750	0.0240	0.0035	0.0006	0.947
		1000	0.0235	0.0133	0.0013	0.950
		1250	0.0165	0.0046	0.0003	0.943
	$\overline{y}_{RPML}$	125	0.0843	0.1165	0.0073	0.946
		250	0.0535	0.1326	0.0031	0.938
		500	0.0372	0.1387	0.0016	0.943
		750	0.0282	0.1271	0.0010	0.919
		1000	0.0209	0.1325	0.0007	0.937
		1250	0.0160	0.1255	0.0005	0.899

Figura 6.3: Desvio Padrão das estratégias de estimador regressão para $N = 4500$, $N_{ab} = 1500$ e desvios unitários.



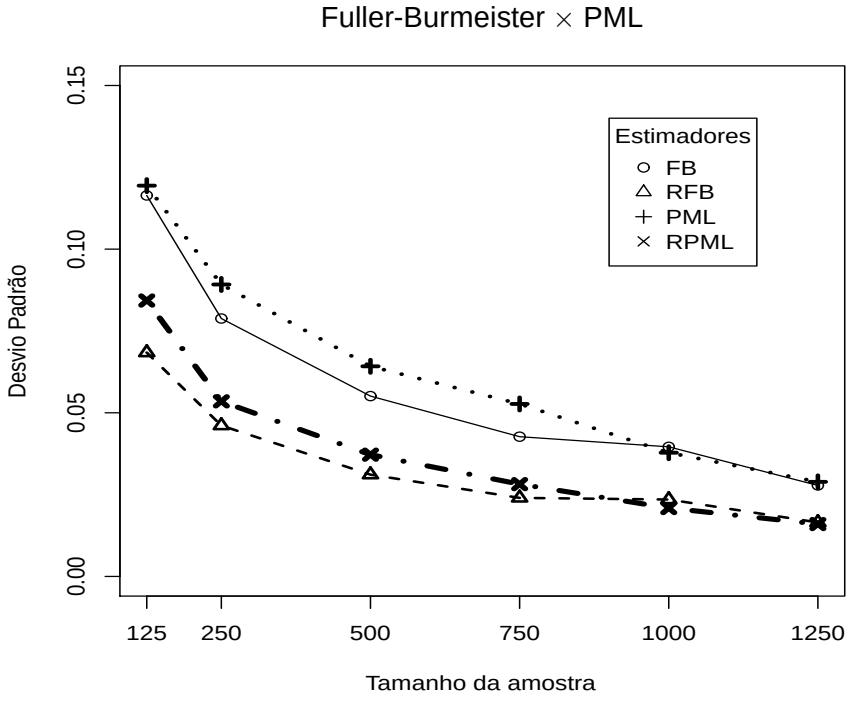


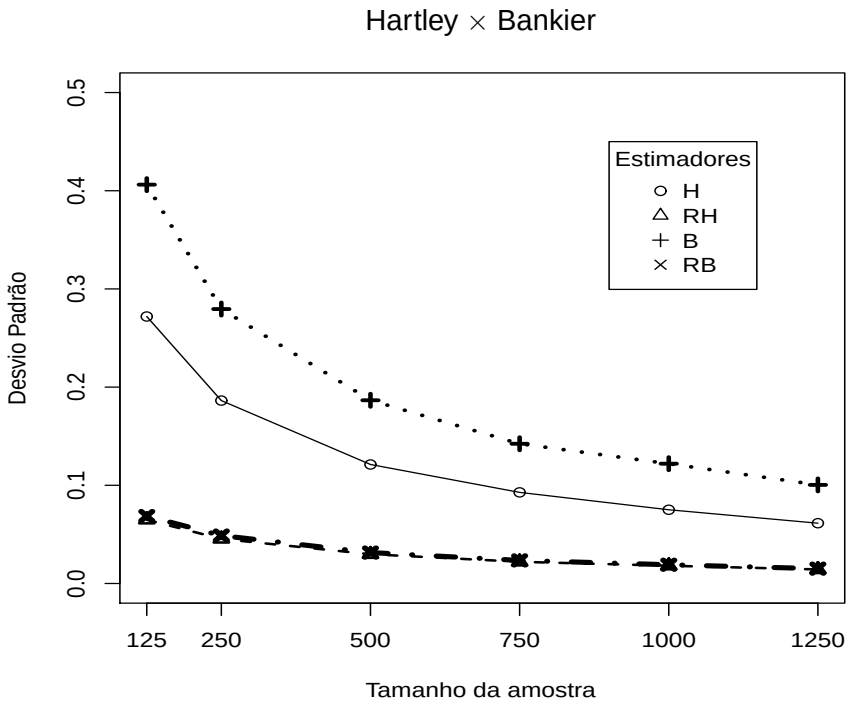
Tabela 6.8: Avaliação das estratégias de Hartley, Fuller, Bankier e PML para $N = 4500$, $N_{ab} = 500$ e desvios diferentes.

Cenário	Estimador	$n_A = n_B$	Estatística			
			Desvio Padrão	100×VR	EQM	PCI
2	$\overline{y}_H$	125	0.2719	0.1317	0.0742	0.948
		250	0.1863	0.0351	0.0347	0.946
		500	0.1212	0.0268	0.0147	0.953
		750	0.0928	0.0190	0.0100	0.937
		1000	0.0751	0.0236	0.0056	0.953
		1250	0.0614	0.0274	0.0037	0.961
	$\overline{y}_{BLG}$	125	0.4062	0.1653	0.1655	0.950
		250	0.2795	0.0856	0.0782	0.945
		500	0.1866	0.0409	0.0349	0.944
		750	0.1423	0.0215	0.0203	0.946
		1000	0.1220	0.0481	0.0149	0.957
		1250	0.1004	0.0191	0.0101	0.953
3	$\overline{y}_{FB}$	125	0.2714	0.1389	0.0740	0.940
		250	0.1859	0.0322	0.0346	0.947
		500	0.1209	0.0273	0.0146	0.955
		750	0.0928	0.0185	0.0086	0.939
		1000	0.0751	0.0232	0.0056	0.953
		1250	0.0614	0.0272	0.0037	0.961
	$\overline{y}_{PML}$	125	0.2818	0.1400	0.0741	0.947
		250	0.1944	0.0325	0.0350	0.947
		500	0.1218	0.0276	0.0147	0.960
		750	0.1032	0.0190	0.0090	0.940
		1000	0.0842	0.0233	0.0056	0.949
		1250	0.0745	0.0235	0.0038	0.959

Tabela 6.9: Avaliação das estratégias de estimador regressão propostos para $N = 4500$, $N_{ab} = 500$ e desvios diferentes.

Cenário	Estimador	$n_A = n_B$	Estatística			
			Desvio Padrão	100×VR	EQM	PCI
2	$\overline{y}_{RH}$	125	0.0649	0.029	0.0042	0.948
		250	0.0454	0.0396	0.0021	0.939
		500	0.0298	0.0193	0.0009	0.956
		750	0.0221	0.0144	0.0004	0.948
		1000	0.0181	0.0156	0.0003	0.953
		1250	0.0143	0.0146	0.0002	0.949
	$\overline{y}_{RBLG}$	125	0.0687	0.1056	0.0049	0.945
		250	0.0488	0.0398	0.0025	0.935
		500	0.0315	0.1166	0.0012	0.927
		750	0.0235	0.1224	0.0008	0.902
		1000	0.0191	0.1911	0.0006	0.880
		1250	0.0151	0.1518	0.0005	0.811
3	$\overline{y}_{RFB}$	125	0.0617	0.0275	0.0038	0.951
		250	0.0442	0.0371	0.0020	0.945
		500	0.0287	0.0182	0.0008	0.953
		750	0.0215	0.0135	0.0005	0.940
		1000	0.0176	0.0136	0.0003	0.956
		1250	0.0138	0.0155	0.0002	0.954
	$\overline{y}_{RPML}$	125	0.0815	0.1680	0.0071	0.938
		250	0.0550	0.1861	0.0036	0.936
		500	0.0360	0.1728	0.0018	0.911
		750	0.0281	0.1509	0.0012	0.881
		1000	0.0232	0.1586	0.0010	0.863
		1250	0.0185	0.1518	0.0007	0.818

Figura 6.4: Desvio Padrão das estratégias de estimador regressão para $N = 4500$, $N_{ab} = 500$ e desvios diferentes.



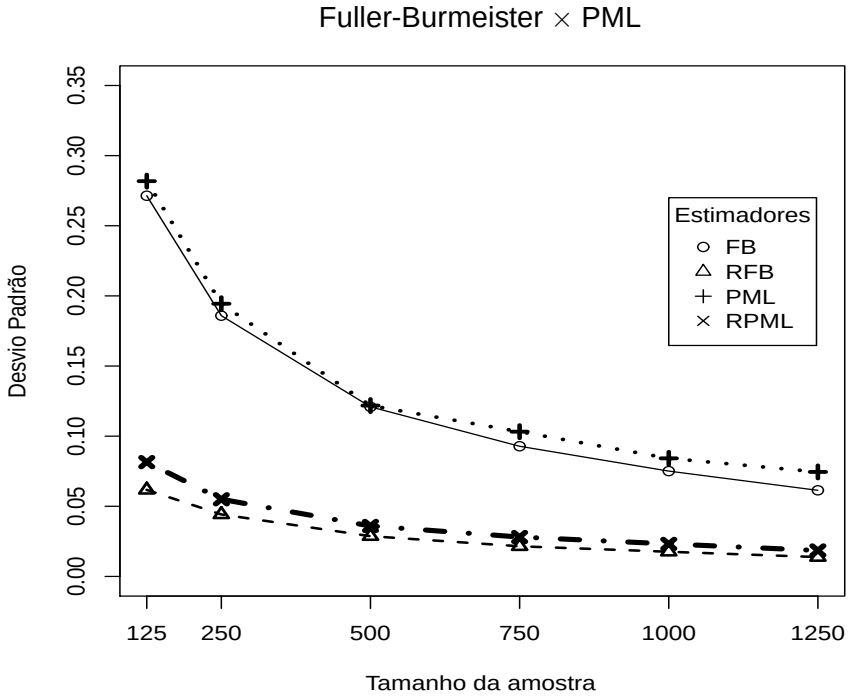


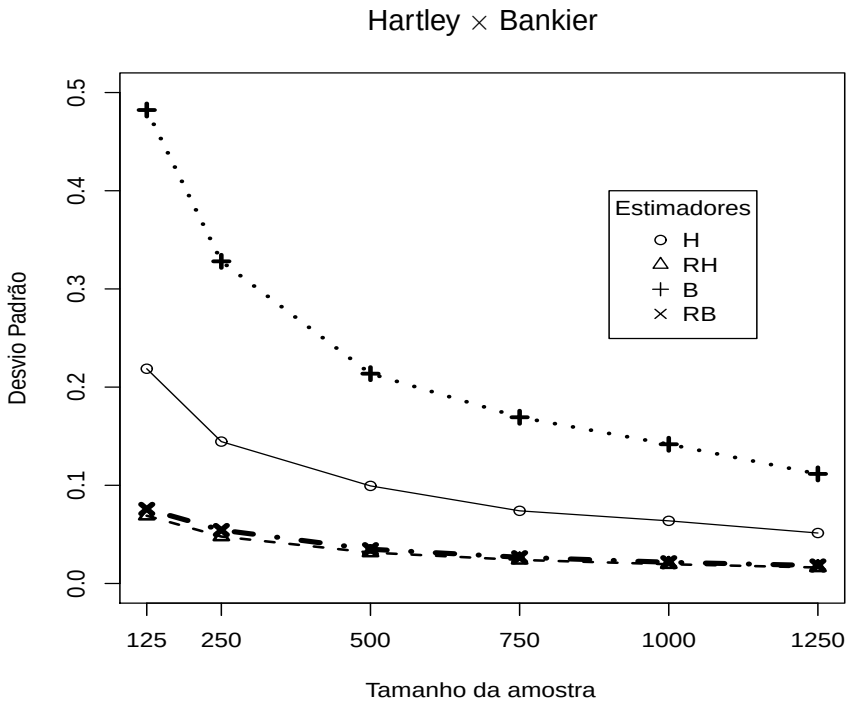
Tabela 6.10: Avaliação das estratégias de Hartley, Fuller, Bankier e PML para $N = 4500$, $N_{ab} = 1000$ e desvios diferentes.

Cenário	Estimador	$n_A = n_B$	Estatística			
			Desvio Padrão	100×VR	EQM	PCI
2	$\overline{y}_H$	125	0.2187	0.0714	0.0479	0.954
		250	0.1445	0.0417	0.0209	0.954
		500	0.0994	0.0919	0.0098	0.953
		750	0.0740	0.0853	0.5485	0.955
		1000	0.0638	0.0207	0.0041	0.951
		1250	0.0514	0.0667	0.0027	0.955
	$\overline{y}_{BLG}$	125	0.4822	0.2200	0.2332	0.955
		250	0.3281	0.1280	0.1079	0.948
		500	0.2137	0.0159	0.0456	0.957
		750	0.1693	0.0261	0.0287	0.950
		1000	0.1418	0.0726	0.0202	0.949
		1250	0.1116	0.0228	0.0124	0.954
3	$\overline{y}_{FB}$	125	0.2184	0.0676	0.0478	0.953
		250	0.1446	0.0430	0.0209	0.953
		500	0.0994	0.0097	0.0100	0.953
		750	0.0741	0.0086	0.0054	0.955
		1000	0.0638	0.0207	0.0041	0.951
		1250	0.0514	0.0668	0.0027	0.955
	$\overline{y}_{PML}$	125	0.2203	0.0678	0.0490	0.953
		250	0.1354	0.0425	0.0210	0.953
		500	0.1042	0.0101	0.0115	0.953
		750	0.0839	0.0088	0.0060	0.955
		1000	0.0745	0.0210	0.0043	0.951
		1250	0.0516	0.0670	0.0030	0.955

Tabela 6.11: Avaliação das estratégias de estimador regressão propostos para $N = 4500$, $N_{ab} = 1000$ e desvios diferentes.

Cenário	Estimador	$n_A = n_B$	Estatística			
			Desvio Padrão	100×VR	EQM	PCI
2	$\overline{y}_{RH}$	125	0.0690	0.0311	0.0047	0.945
		250	0.0478	0.0419	0.0023	0.947
		500	0.0315	0.0338	0.0010	0.956
		750	0.0240	0.0330	0.0005	0.940
		1000	0.0196	0.0350	0.0004	0.945
		1250	0.0163	0.0340	0.0003	0.939
	$\overline{y}_{RBLG}$	125	0.0759	0.0539	0.0057	0.887
		250	0.0544	0.0174	0.0029	0.852
		500	0.0349	0.0294	0.0012	0.856
		750	0.0267	0.0369	0.0007	0.906
		1000	0.0216	0.0292	0.0005	0.909
		1250	0.0183	0.0286	0.0003	0.912
3	$\overline{y}_{RFB}$	125	0.0647	0.0261	0.0042	0.955
		250	0.0461	0.0432	0.0022	0.952
		500	0.0296	0.0361	0.0010	0.949
		750	0.0227	0.0299	0.0005	0.945
		1000	0.0186	0.0341	0.0004	0.948
		1250	0.0155	0.0344	0.0002	0.940
	$\overline{y}_{RPML}$	125	0.0933	0.0923	0.0088	0.806
		250	0.0640	0.1247	0.0043	0.901
		500	0.0398	0.1113	0.0018	0.912
		750	0.0325	0.0993	0.0012	0.901
		1000	0.0259	0.1149	0.0010	0.909
		1250	0.0205	0.1000	0.0005	0.888

Figura 6.5: Desvio Padrão das estratégias de estimador regressão para $N = 4500$, $N_{ab} = 1000$ e desvios diferentes.



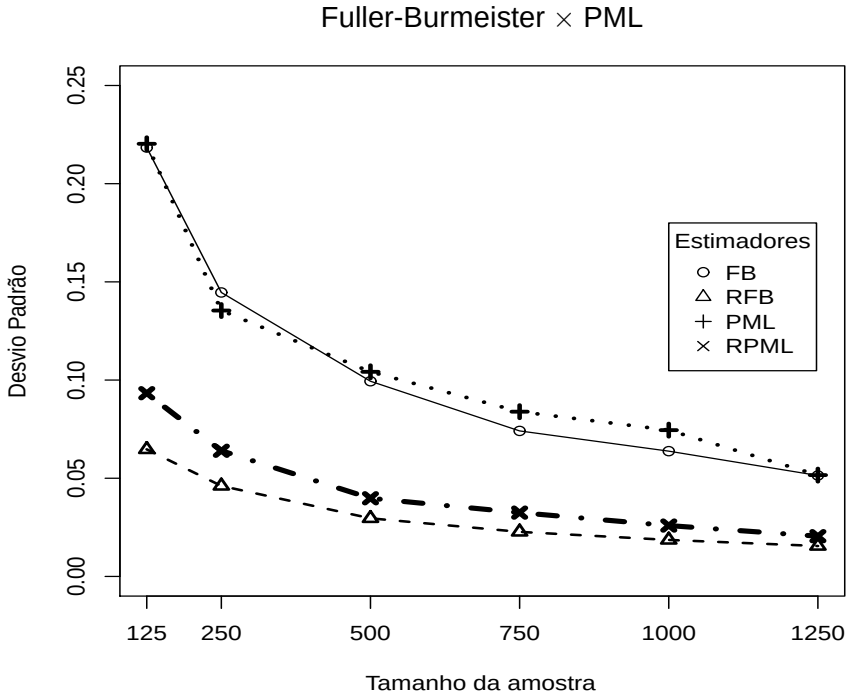


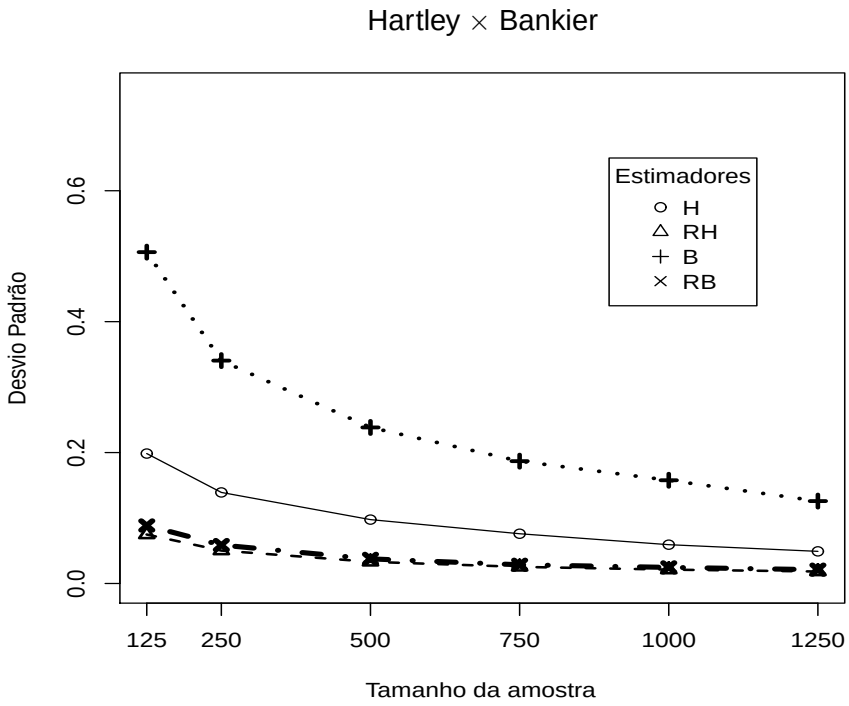
Tabela 6.12: Avaliação das estratégias de Hartley, Fuller, Bankier e PML para $N = 4500$, $N_{ab} = 1500$ e desvios diferentes.

Cenário	Estimador	$n_A = n_B$	Estatística			
			Desvio Padrão	100×VR	EQM	PCI
2	$\overline{y}_H$	125	0.1985	0.0724	0.0395	0.948
		250	0.1391	0.0109	0.0194	0.952
		500	0.0976	0.0023	0.0095	0.951
		750	0.0760	0.0031	0.0058	0.952
		1000	0.0593	0.0014	0.0035	0.955
		1250	0.0491	0.0085	0.0024	0.955
	$\overline{y}_{BLG}$	125	0.5062	0.0260	0.2562	0.945
		250	0.3405	0.0549	0.1159	0.943
		500	0.2384	0.0088	0.0568	0.944
		750	0.1869	0.0479	0.0349	0.942
		1000	0.1577	0.0174	0.0249	0.948
		1250	0.1259	0.0448	0.0159	0.941
3	$\overline{y}_{FB}$	125	0.1983	0.0694	0.0394	0.949
		250	0.1391	0.0117	0.0193	0.951
		500	0.0977	0.0029	0.0095	0.951
		750	0.0760	0.0031	0.0058	0.951
		1000	0.0593	0.0014	0.0035	0.954
		1250	0.0491	0.0083	0.0024	0.957
	$\overline{y}_{PML}$	125	0.1900	0.0702	0.0400	0.939
		250	0.1404	0.0120	0.0198	0.941
		500	0.0880	0.0032	0.0095	0.941
		750	0.0776	0.0035	0.0056	0.939
		1000	0.0590	0.0019	0.0037	0.950
		1250	0.0479	0.0081	0.0028	0.957

Tabela 6.13: Avaliação das estratégias de estimador regressão propostos para $N = 4500$, $N_{ab} = 1500$ e desvios diferentes.

Cenário	Estimador	$n_A = n_B$	Estatística			
			Desvio Padrão	100×VR	EQM	PCI
2	$\overline{y}_{RH}$	125	0.0748	0.0303	0.0056	0.954
		250	0.0503	0.0264	0.0025	0.954
		500	0.0331	0.0158	0.0011	0.956
		750	0.0254	0.0087	0.0006	0.952
		1000	0.0212	0.0375	0.0005	0.940
		1250	0.0183	0.0123	0.0003	0.944
	$\overline{y}_{RBLG}$	125	0.0886	0.5346	0.0113	0.788
		250	0.0587	0.5340	0.0069	0.721
		500	0.0376	0.5283	0.0048	0.670
		750	0.0286	0.5191	0.0041	0.763
		1000	0.0246	0.5499	0.0043	0.799
		1250	0.0211	0.5249	0.0038	0.709
3	$\overline{y}_{RFB}$	125	0.0692	0.0231	0.0048	0.952
		250	0.0458	0.0177	0.0021	0.951
		500	0.0298	0.0142	0.0010	0.945
		750	0.0226	0.0114	0.0005	0.951
		1000	0.0196	0.0332	0.0004	0.945
		1250	0.0167	0.0134	0.0003	0.954
	$\overline{y}_{RPML}$	125	0.1097	0.4907	0.0149	0.786
		250	0.0716	0.5027	0.0082	0.785
		500	0.0495	0.4918	0.0054	0.797
		750	0.0379	0.4948	0.0044	0.701
		1000	0.0324	0.4768	0.0038	0.731
		1250	0.0262	0.4950	0.0037	0.751

Figura 6.6: Desvio Padrão das estratégias de estimador regressão para $N = 4500$, $N_{ab} = 1500$ e desvios diferentes.



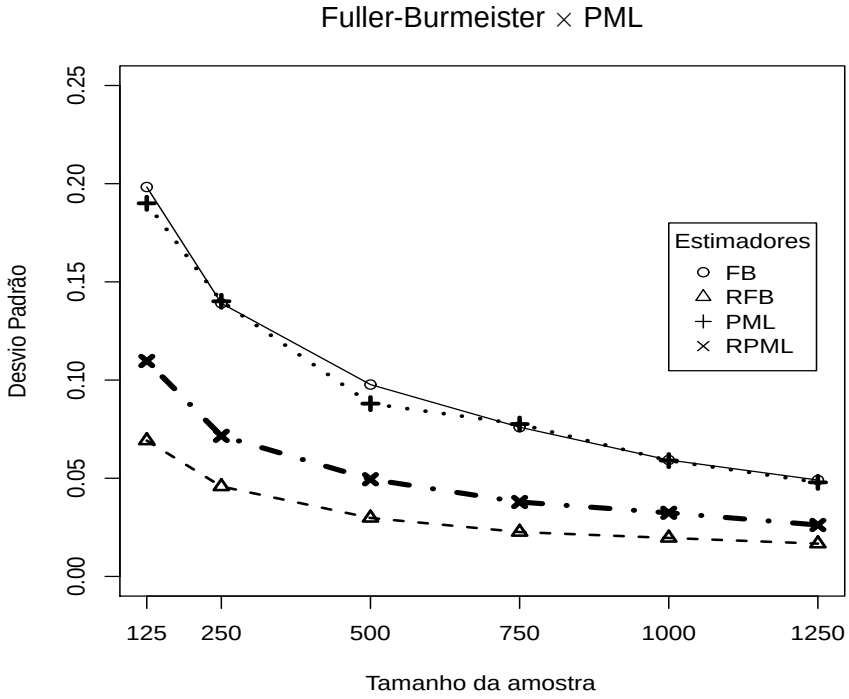


Figura 6.7: Densidade estimada dos estimadores propostos para $N = 4500$ e $N_{ab} = 500$

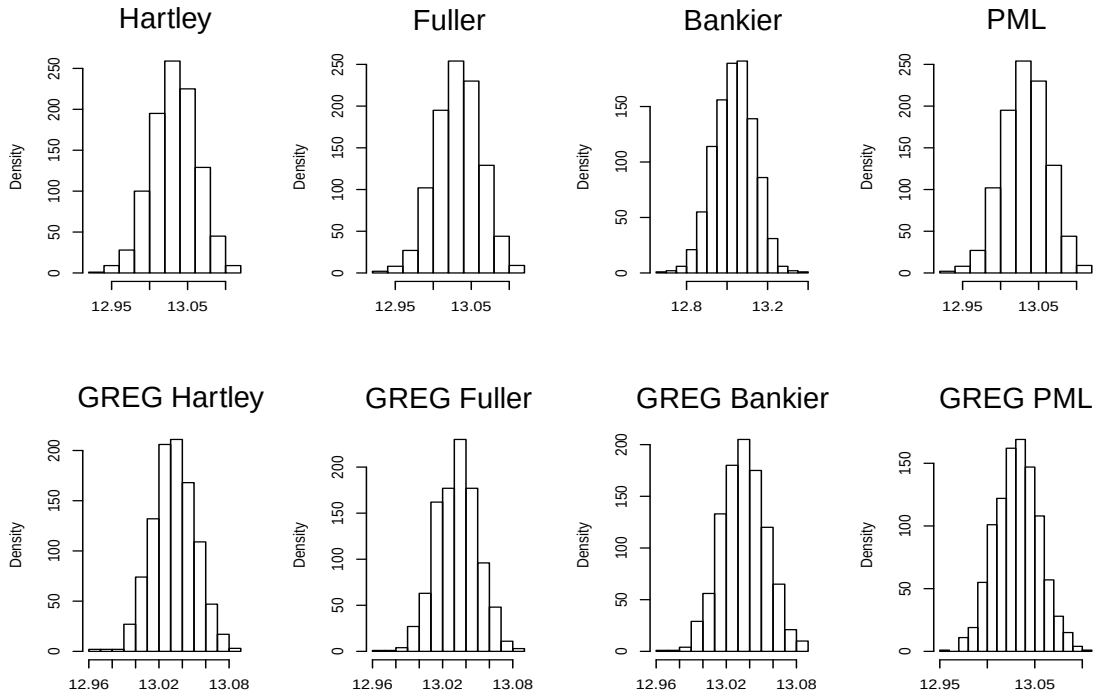


Figura 6.8: Densidade estimada dos estimadores propostos para $N = 4500$ e $N_{ab} = 1000$

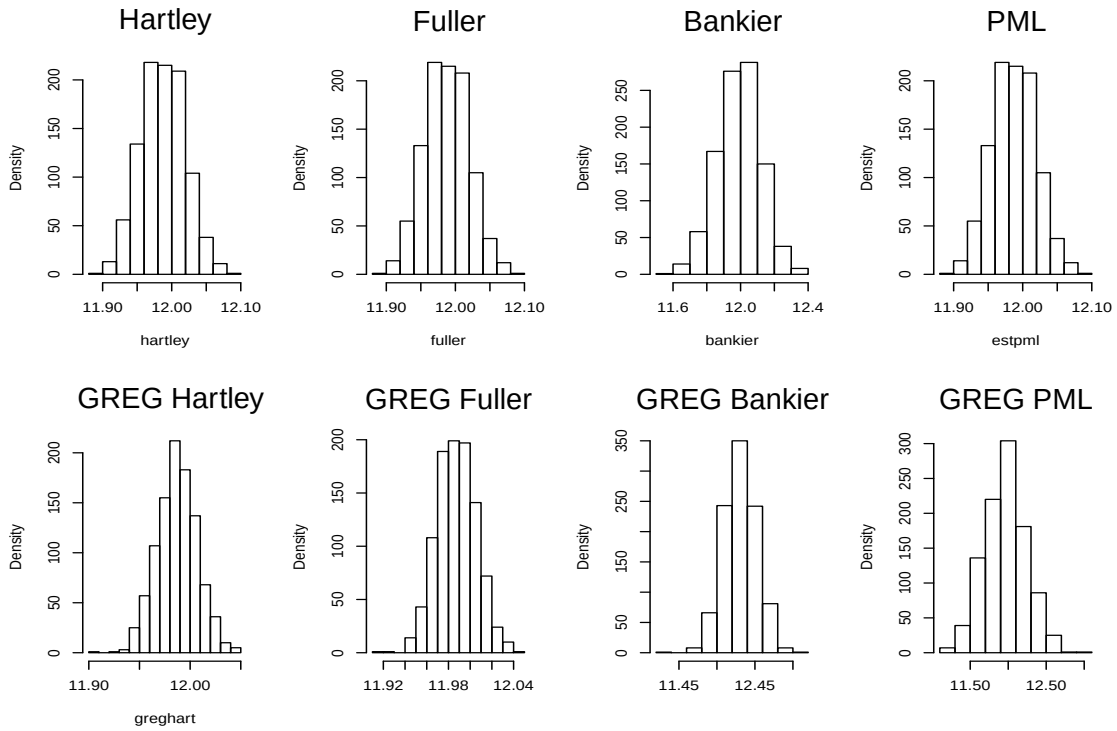


Figura 6.9: Densidade estimada dos estimadores propostos para $N = 4500$ e $N_{ab} = 1500$

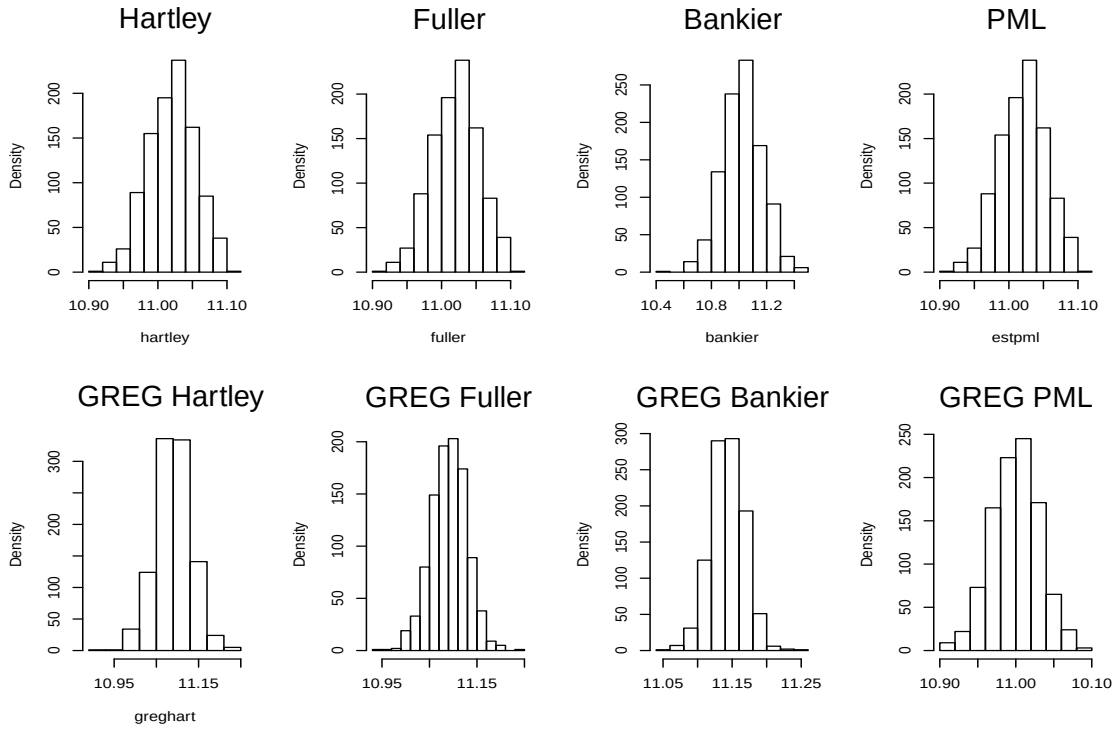


Figura 6.10: Densidade Estimada de $\hat{N}_{ab,s}$, com valor de referência $N_{ab} = 500$

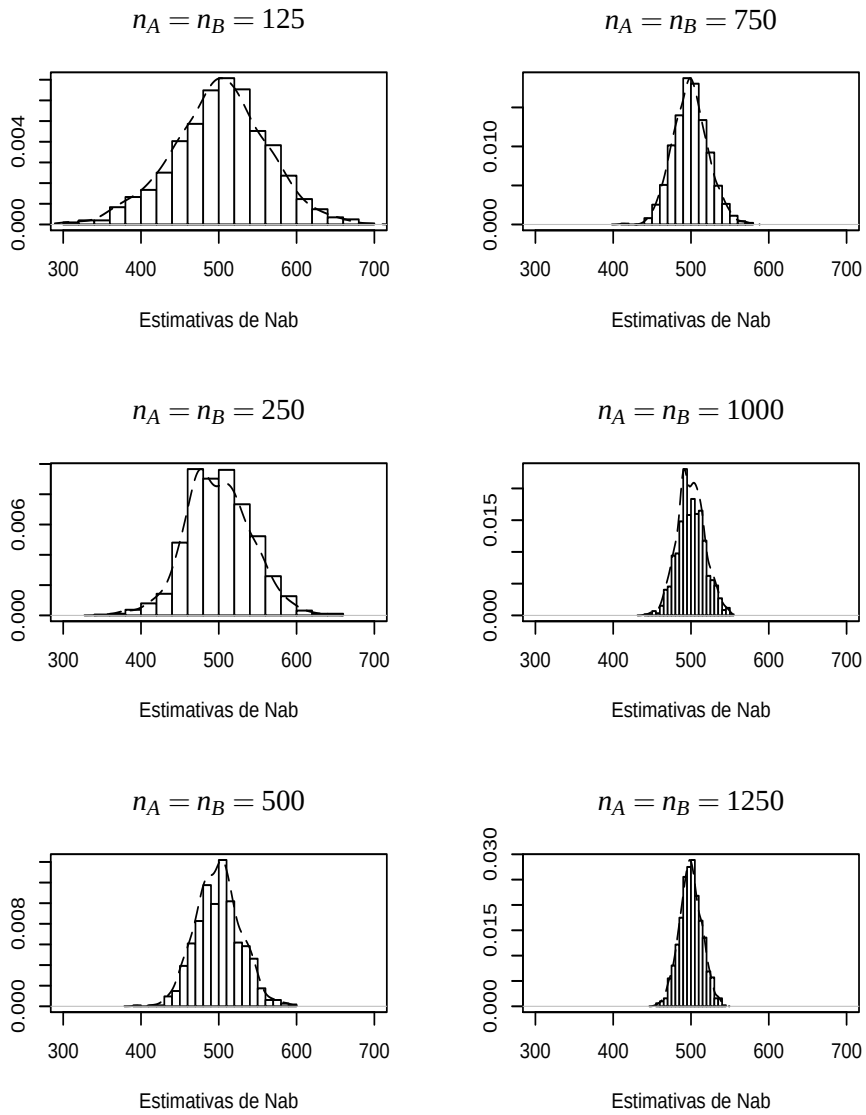


Figura 6.11: Densidade Estimada de $\hat{N}_{ab,s}$, com valor de referência $N_{ab} = 1000$

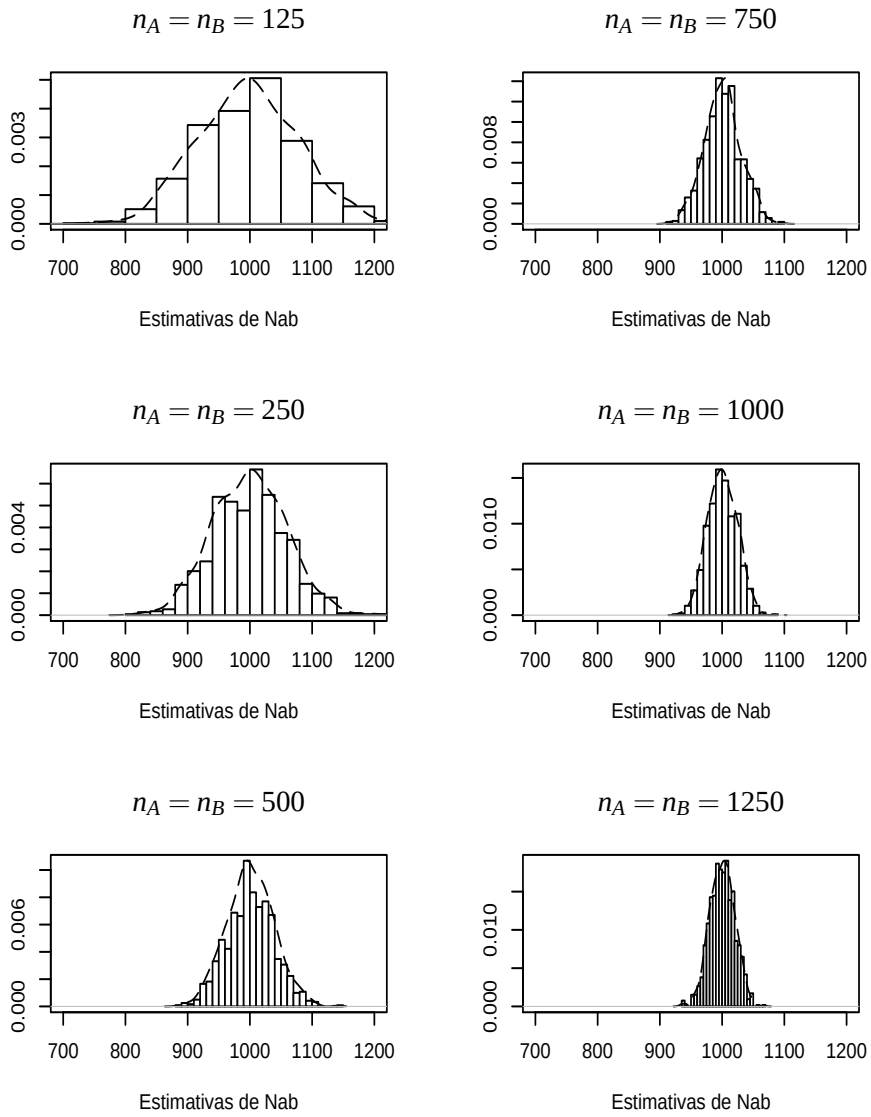


Figura 6.12: Densidade Estimada de $\hat{N}_{ab,s}$, com valor de referência $N_{ab} = 1500$

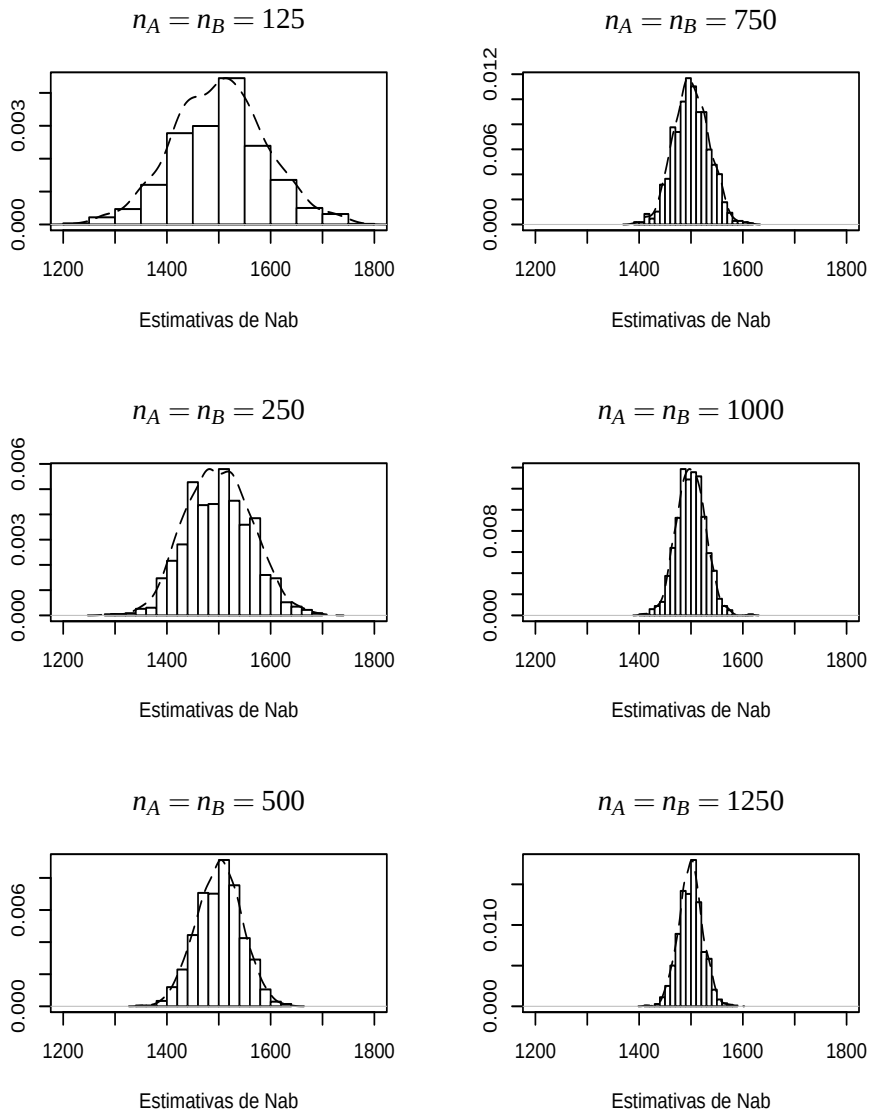


Figura 6.13: Densidade Estimada de $\hat{N}_{ab,PML}$, com valor de referência $N_{ab} = 500$

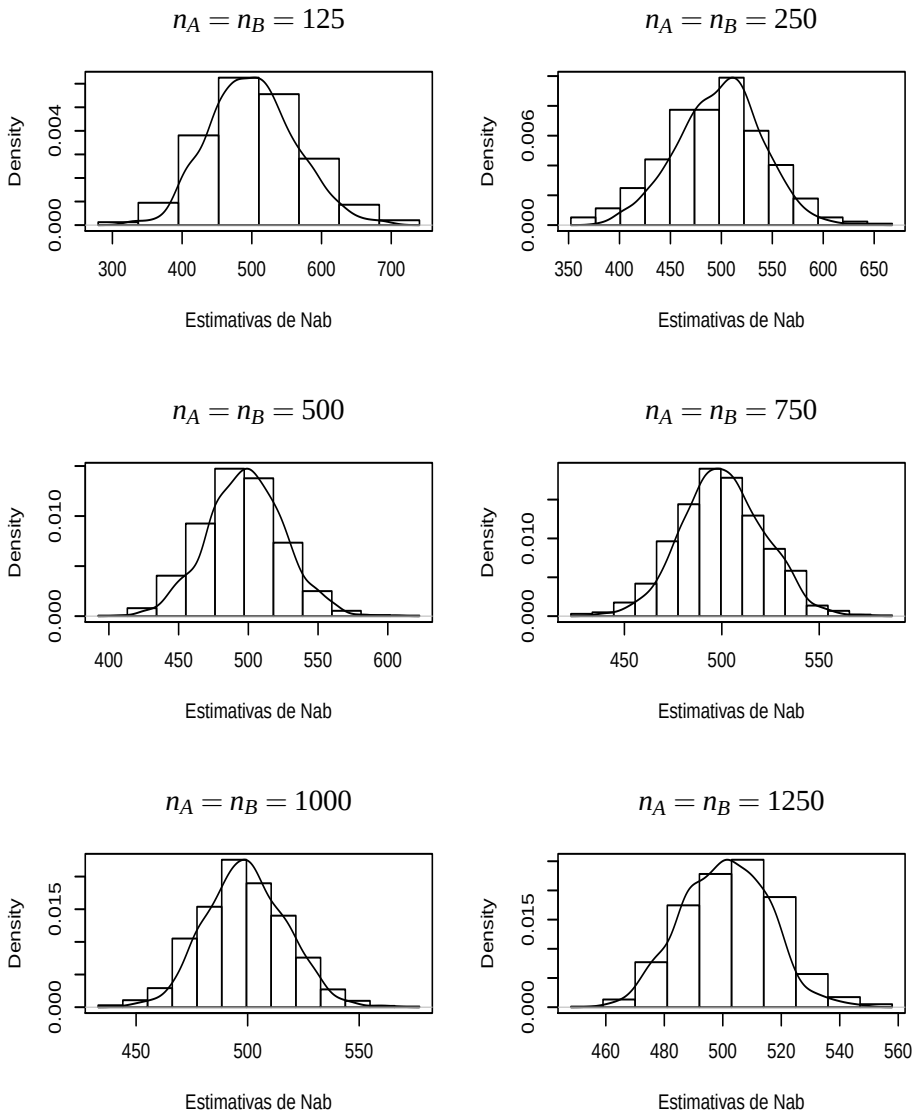


Figura 6.14: Densidade Estimada de $\hat{N}_{ab,PML}$, com valor de referência $N_{ab} = 1000$

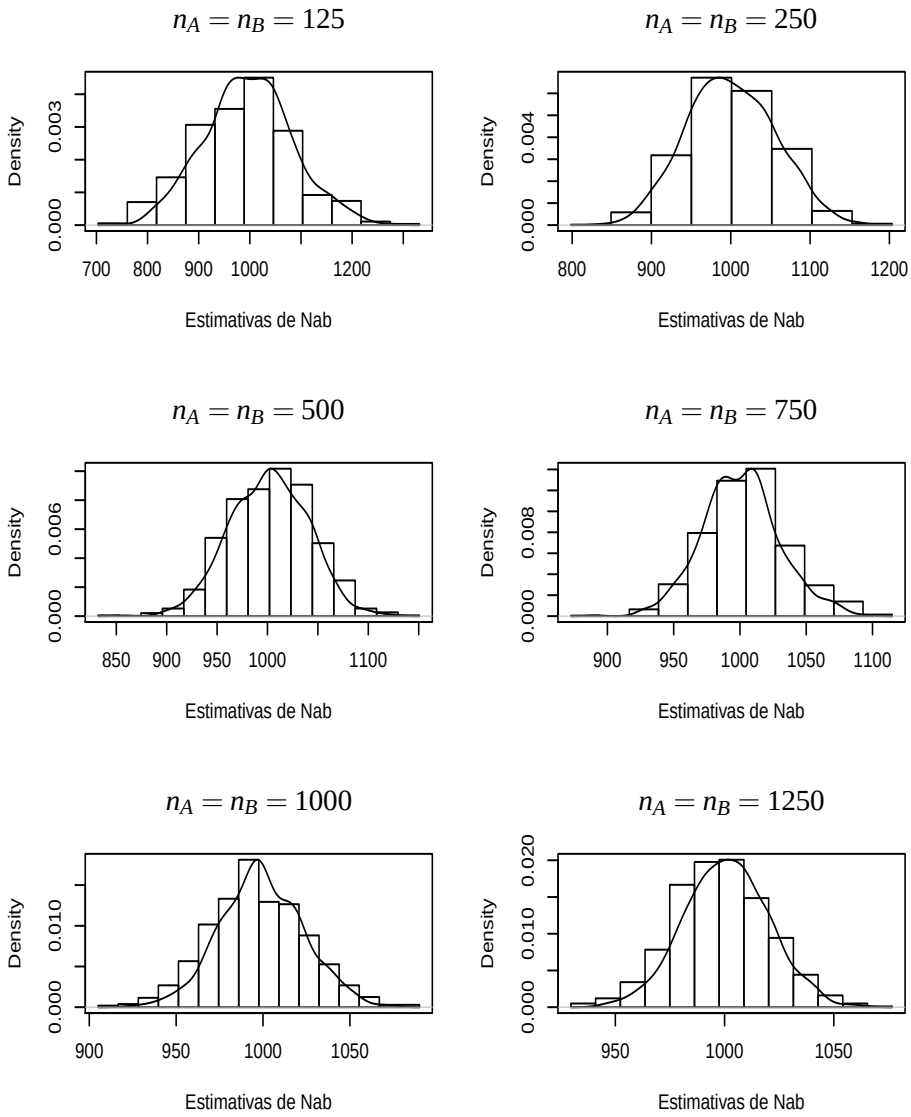
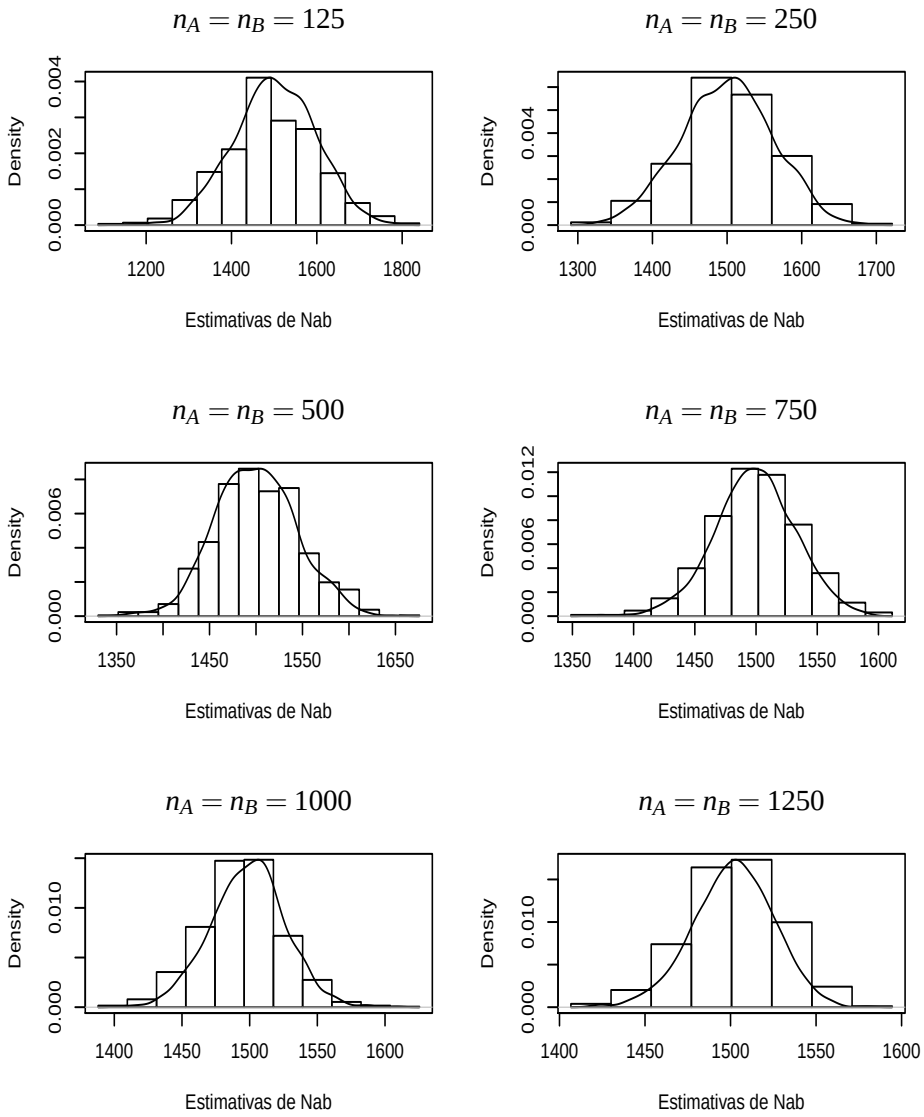


Figura 6.15: Densidade Estimada de $\hat{N}_{ab,PML}$, com valor de referência $N_{ab} = 1500$



CAPÍTULO 7

Considerações Finais

No desenvolvimento desta tese, foram apresentados quatro novos estimadores do tipo regressão generalizados para uso na abordagem de cadastro duplo. Estes estimadores foram propostos a partir das estratégias apresentadas por Hartley (1962), Fuller & Burmeister (1972), Bankier, Lepkowski & Groves (1986) e Skinner & Rao (1996). A forma da variância aproximada de todos os estimadores propostos foi derivada e apresentada em termos matriciais na forma $\mathbf{d}^T\Sigma\mathbf{d}$, onde $\mathbf{d}$ e Σ são dadas de acordo com cada uma das estratégias apresentadas. Além disso, as propriedades de consistência, centralidade assintótica e expressões para o erro quadrático médio destes estimadores foram apresentadas de forma rigorosa com o intuito de motivar o uso de variáveis auxiliares na abordagem de cadastro duplo, uma vez que é esperado que informações auxiliares forneçam uma melhora no processo de estimação de uma parâmetro populacional. Além disso, foi provado através do método delta que, assintoticamente, os estimadores propostos possuem distribuição normal, e que forneceu suporte para cálculo das variâncias aproximadas dos estimadores propostos.

Os resultados obtidos nesta tese generalizam todos os resultados para o estimador do tipo razão propostos por Coelho (2007), que considerou informação de apenas uma variável auxiliar em apenas um dos cadastros, além de fornecer novas formas de estimador do tipo razão sob as estratégias propostas por Bankier, Lepkowski & Groves (1986) e Skinner & Rao (1996).

O desempenho dos novos estimadores regressão para a abordagem de cadastro duplo foi avaliado através de um estudo de simulação sob os planos de amostragem aleatória simples, comparando-os com os estimadores já disponíveis na literatura. Os resultados das simulações mostraram que os estimadores regressão propostos apresentaram melhor desempenho em comparação aos estimadores propostos na literatura, apresentando considerável redução de

variância e erro quadrático médio, como era esperado.

Os estimadores propostos sob as versões de Hartley (1962) e Fuller & Burmeister (1972) apresentaram menor sensibilidade à heterogeneidade das observações dentro dos cadastros e domínios e ao aumento do tamanho populacional no domínio N_{ab}, quando comparados aos estimadores de Bankier, Lepkowski & Groves (1986) e Skinner & Rao (1996), que só apresentaram desempenho equivalente quando $N_{ab}/N \approx 0.10$, valor este que pode servir como uma recomendação exigida para uso destes estimadores quando se desejar facilidade de implementação do método.

Os estimadores regressão propostos sob as estratégias de Hartley (1962) e Fuller (2009) são mais eficientes do que os demais estimadores apresentados sob o cenário 2. Porém, para tamanhos de amostras elevados na situação descrita pelo cenário 2, é verificada vantagem de uso do estimador RBLG.

Os estimadores regressão sob as estratégias de Fuller & Burmeister e PML apresentaram melhor desempenho que os demais estimadores apresentados sob o cenário 3. O estimador regressão sob a estratégia de Fuller & Burmeister foi o que teve maior destaque por apresentar maior redução de sua variância quando o tamanho da amostra aumenta.

De um modo geral, ao desconsiderar os cenários os estimadores regressão propostos sob as estratégias de Fuller & Burmeister e PML foram os estimadores que apresentaram melhor desempenho, por apresentar maior redução da variância quando o tamanho da amostra aumenta em ambos os cadastros. Dessa forma, em um cenário de aplicação onde o tamanho populacional dos domínios é conhecido (cenário 2), recomenda-se o uso do estimador de Bankier, por ser mais eficiente que a estratégia de Hartley, além de ser relativamente simples de ser implementado. Se a situação envolver o não conhecimento do tamanho populacional dos domínios (cenário 3), recomenda-se o uso do estimador de Fuller & Burmeister.

Esta tese dá abertura para várias pesquisas sobre estratégias de inferência assistida por modelos sob a abordagem de cadastro duplo. Dentre as propostas de pesquisas futuras, por exemplo, podemos citar as seguintes:

- Apresentar os estimadores propostos sob o contexto da estratificação, considerando-se estimadores regressão nas versões separado e combinado, de modo a generalizar os resultados apresentados por Coelho & Ferraz (2007);
- Estender toda a teoria desenvolvida nesta tese para o estimador regressão logístico generalizado, utilizado para estimação em populações assistida por modelos variáveis dicotômicas, discutido por Poveda (2007);
- Apresentar os estimadores propostos para diversos planos amostrais e realizar estudos de alocação ótima de amostras aos cadastros sob estes planos;
- Analisar o desempenho dos estimadores propostos sob planos amostrais complexos;

- Analisar o desempenho dos estimadores sob técnicas de imputação de dados para tratamento da não-resposta.

Apêndice A

Resumo de conceitos importantes

A seguir, será apresenta uma revisão de alguns conceitos utilizados ao longo desta tese.

1 Ordens de Magnitude de sequências de números reais ($\mathbf{O}$ e $\mathbf{o}$)

Sejam a_n e b_n sequências de números reais. Dizemos que

1. $a_n = \mathbf{O}(b_n)$ se existirem $M > 0$ e $n_0 = n(M)$ tal que $\left|\frac{a_n}{b_n}\right| \leq M$, $\forall\, n \geq n_0$;
2. $a_n = \mathbf{o}(b_n)$ se para todo $\varepsilon > 0$ existir um número positivo $n_0 = n_0(\varepsilon)$ tal que $\left|\frac{a_n}{b_n}\right| < \varepsilon$, $\forall\, n \geq n_0$

1.1 Teorema 1

Sejam a_n, b_n, c_n e d_n sequências de números reais. Então:

1. Se $a_n = \mathbf{o}(b_n)$, então $a_n = \mathbf{O}(b_n)$;
2. Se $a_n = \mathbf{O}(b_n)$ e $c_n = \mathbf{O}(d_n)$, então:
   (a) $a_n c_n = \mathbf{O}(b_n d_n)$;
   (b) $|a_n|^s = \mathbf{O}(|b_n|^s)$, para todo $s > 0$;
   (c) $a_n + c_n = \mathbf{O}(\max\{|b_n|, |d_n|\})$
3. Se $a_n = \mathbf{o}(b_n)$ e $c_n = \mathbf{o}(d_n)$, então:
   (a) $a_n c_n = \mathbf{o}(b_n d_n)$;
   (b) $|a_n|^s = \mathbf{o}(|b_n|^s)$, para todo $s > 0$;
   (c) $a_n + c_n = \mathbf{o}(\max\{|b_n|, |d_n|\})$

2 Desigualdade de Markov

Seja X uma variável aleatória qualquer. Então para todo $t > 0$, temos:

$$P(|X| \geq \varepsilon) \leq \frac{E(|X|)^t}{\varepsilon^t}$$

3 Convergência Estocástica

3.1 Ordens de Magnitude de sequências de números reais ($\mathbf{O}_p(.)$ e $\mathbf{o}_p(.)$)

Sejam $\{X_n\}_{n\geq 1}$ sequência de variáveis aleatórias e $\{b_n\}_{n\geq 1}$ uma sequência de números reais (ou variáveis aleatórias). Dessa forma, temos que:

1. $X_n = \mathbf{O}_p(b_n)$ se para todo $\varepsilon > 0$ existir M tal que $M = M(\varepsilon)$ e um número inteiro positivo n_0 onde $n_0 = n_0(\varepsilon)$ tais que

$$P\left(\left|\frac{X_n}{b_n}\right| \geq M\right) \leq \varepsilon, \ \forall n \geq n_0$$

   ou seja, a sequência $\left\{\frac{X_n}{b_n}\right\}_{n\geq 1}$ é limitada em probabilidade para todo n suficientemente grande.

2. $X_n = \mathbf{o}_p(b_n)$ se para todo $\varepsilon > 0$ e para todo $\lambda > 0$ existir n_0 tal que $n_0 = n_0(\varepsilon, \lambda)$ tais que

$$P\left(\left|\frac{X_n}{b_n}\right| \geq \varepsilon\right) < \lambda, \ \forall n \geq n_0,$$

   ou seja, $X_n = \mathbf{o}_p(b_n)$ se para todo $\varepsilon > 0$, $P\left(\left|\frac{X_n}{b_n}\right| \geq \varepsilon\right) \to 0$, quando $n \to \infty$.

3.2 Teorema 2

Sejam $\{X_n\}_{n\geq 1}$, $\{Y_n\}_{n\geq 1}$ sequência de variáveis aleatórias e $\{a_n\}_{n\geq 1}$ $\{b_n\}_{n\geq 1}$ sequências de números reais (ou variáveis aleatórias). Temos então que:

1. Se $X_n = \mathbf{o}_p(a_n)$, então $X_n = \mathbf{O}_p(a_n)$;
2. Se $X_n = \mathbf{O}_p(a_n)$ e $Y_n = \mathbf{O}_p(b_n)$, então:
   (a) $X_n Y_n = \mathbf{O}_p(a_n b_n)$;
   (b) $|X_n|^s = \mathbf{O}_p(|a_n|^s)$, para todo $s > 0$;
   (c) $X_n + Y_n = \mathbf{O}_p(\max\{|b_n|, |d_n|\})$

3. Se $X_n = \mathbf{o}_p(a_n)$ e $Y_n = \mathbf{o}_p(b_n)$, então:

   (a) $X_n Y_n = \mathbf{o}_p(a_n b_n)$;

   (b) $|X_n|^s = \mathbf{o}_p(|a_n|^s)$, para todo $s > 0$;

   (c) $X_n + Y_n = \mathbf{o}_p(\max\{|b_n|, |d_n|\})$

   (d) Se $X_n = \mathbf{O}_p(a_n)$ e $Y_n = \mathbf{o}_p(b_n)$, então: $X_n Y_n = \mathbf{o}_p(a_n b_n)$

4 Tipos de convergência estocástica

1. Se $X_n \xrightarrow{q.c} X$, então $X_n \xrightarrow{P} X$;

2. Sejam $\{X_n; n \geq 1\}$ e $\{Y_n; n \geq 1\}$ sequências de variáveis aleatórias. Se $|X_n - Y_n| \xrightarrow{P} 0$ e $Y_n \xrightarrow{d} X$, então $X_n \xrightarrow{d} X$;

3. Se $X_n \xrightarrow{P} X$, então $X_n \xrightarrow{d} X$;

4. Seja c uma constante. Se Se $X_n \xrightarrow{d} c$, então $X_n \xrightarrow{P} c$;

5. Se $X_n \xrightarrow{d} X$, então $X_n \xrightarrow{P} 0$, então, $X_n Y_n \xrightarrow{P} 0$;

6. **Teorema de Slutsky:**

   - Se $X_n \xrightarrow{d} X$ e $Y_n \xrightarrow{P} 0$, então $X_n Y_n \xrightarrow{P} 0$;
   - Seja c uma constante. Se $X_n \xrightarrow{d} X$ e $Y_n \xrightarrow{P} c$, então $X_n + Y_n \xrightarrow{d} X + c$, $X_n Y_n \xrightarrow{d} Xc$, $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$;

7. Seja $g : \mathbb{R} \to \mathbb{R}$ uma função contínua. Então,

   - Se $X_n \xrightarrow{q.c} X$, então $g(X_n) \xrightarrow{q.c.} g(X)$;
   - Se $X_n \xrightarrow{P} X$, então $g(X_n) \xrightarrow{P} g(X)$;
   - Se $X_n \xrightarrow{d} X$, então $g(X_n) \xrightarrow{d} g(X)$;

8. (**Método Delta**) Suponha que $\sqrt{n}(X_n - \mu) \xrightarrow{d} N(0, \sigma^2)$. Se g é uma função diferenciável em μ, então:

$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} N\left(0, \sigma^2 \left[g'(\mu)\right]^2\right).$$

Apêndice B

Demonstrações de resultados propostos

A seguir serão apresentadas as demonstrações dos resultados enunciados nesta tese de doutorado.

Consistência do estimador de Hartley

Resultado: O estimador $\overline{y}^{A\cup B}$ é consistente, ou seja:

$$P\left(\left|\overline{y}_H^{A\cup B}-\overline{y}_\nu\right|>\varepsilon|\mathscr{F}_\nu\right)\longrightarrow 0,\ \text{quando } \nu\to\infty$$

Prova: A prova deste resultado é construtiva, e depende das condições a seguir:

C1. Dado $N_\nu = N_{A\nu}+N_{B\nu}-N_{ab\nu}$, seja $\theta_\nu^{A\cup B}$ o parâmetro de interesse. Então, existe sequência de estimadores de $\theta_\nu^{A\cup B}$, denotada por $\left\{\hat{\theta}_\nu;\ \nu\geq 1\right\}$ que satisfaz a seguinte propriedade: $\forall\varepsilon>0$, existe $\delta_\varepsilon\in(0,\infty)$ tal que

$$P\left(\left|\hat{\theta}_\nu^{A\cup B}-\theta_\nu^{A\cup B}\right|>\delta_\varepsilon\right)<\varepsilon$$

C2. No cadastro A, $\forall N_{A\nu}$, $\forall k\in\mathscr{A}_\nu$, $\min\limits_{k\in\mathscr{A}_\nu}\pi_{\nu k}^A\geq\pi_\nu^*>0$, onde $N_{A\nu}\pi_\nu^*\to\infty$ e existe $s>0$ tal que $(N_{A\nu})^{1/2+s}\left(\pi_{\nu k}^A\right)\to\infty$ e

$$\max_{k\in U_\nu^A}\sum_{l\in U_\nu^A:k\neq l}\left(\pi_{\nu kl}-\pi_{\nu k}^A\pi_{\nu l}^A\right)^2=O\left((N_{A\nu})^{-2s}\right),\ N_{A\nu}\to\infty.$$

C3. As observações da variável de interesse no cadastro A, $\{y_k\}_{k\in U_\nu^A}$ satisfazem:

$$\limsup_{N_{A\nu}\to\infty}\sum_{k\in S_\nu^A}y_k^2<\infty,$$

C4. As mesmas condições **C2** e **C3** são admitidas ao se considerar as correspondentes quantidades no cadastro B.

Breidt & Opsomer (2008) mostraram que a condição suficiente para garantiar a consistência de acordo com o plano é a convergência em média quadrática, ou seja, basta provar que

$$\overline{y}^{A\cup B} \xrightarrow{(2)} \mu_v^{A\cup B}$$

Logo,

$$\begin{aligned}
\overline{y}_{Hv} - \mu_v &= w_{Av}\overline{y}_{HTv}^{A} + w_{Bv}\overline{y}_{HTv}^{B} - w_{abv}\overline{y}_{HTv}^{A\cap B} \\
&- \left(w_{Av}\mu_{Av} + w_{Bv}\mu_{Bv} - w_{abv}\mu_{(A\cap B)v}\right) \\
&= w_{Av}\left(\overline{y}_{Av} - \mu_{Av}\right) + w_{Bv}\left(\overline{y}_{Bv} - \mu_{Bv}\right) - w_{abv}\left(\overline{y}^{A\cap B} - \mu_{Av}\right) \\
&= w_{Av}\left(\overline{y}_{Av} - \mu_{Av}\right) + w_{Bv}\left(\overline{y}_{Bv} - \mu_{Bv}\right) - w_{abv}\left(\overline{y}_{(A\cap B)} - \mu_{(A\cap B)v}\right)
\end{aligned}$$

$$\begin{aligned}
\left(\overline{y}_{Hv} - \mu_v\right)^2 &= \left(w_{Av}\right)^2\left(\overline{y}_{Av} - \mu_{Av}\right)^2 + \left(w_{Bv}\right)^2\left(\overline{y}_{Bv} - \mu_{Bv}\right)^2 \\
&+ 2w_{Av}w_{Bv}\left(\overline{y}_{Av} - \mu_{Av}\right)\left(\overline{y}_{Bv} - \mu_{Bv}\right) \\
&- 2w_{Av}w_{abv}\left(\overline{y}_{(A\cap B)} - \mu_{(A\cap B)v}\right)\left(\overline{y}_{Av} - \mu_{Av}\right) \\
&- 2w_{Bv}w_{abv}\left(\overline{y}_{(A\cap B)} - \mu_{(A\cap B)v}\right)\left(\overline{y}_{Bv} - \mu_{Bv}\right)
\end{aligned}$$

$$\begin{aligned}
E\left(\overline{y}_{Hv} - \mu_v\right)^2 &= \left(w_{Av}\right)^2 E\left(\overline{y}_{Av} - \mu_{Av}\right)^2 + \left(w_{Bv}\right)^2 E\left(\overline{y}_{Bv} - \mu_{Bv}\right)^2 \\
&+ 2w_{Av}w_{Bv}E\left(\overline{y}_{Av} - \mu_{Av}\right)E\left(\overline{y}_{Bv} - \mu_{Bv}\right) \\
&- 2w_{Av}w_{abv}E\left[\left(\overline{y}_{(A\cap B)v} - \mu_{(A\cap B)v}\right)\left(\overline{y}_{Av} - \mu_{Av}\right)\right] \\
&- 2w_{Bv}w_{abv}E\left[\left(\overline{y}_{(A\cap B)v} - \mu_{(A\cap B)v}\right)\left(\overline{y}_{Bv} - \mu_{Bv}\right)\right]
\end{aligned}$$

Para as quantidades referentes ao cadastro A, tem-se que

$$\begin{aligned}
E\left(\left(\overline{y}_{Av} - \mu_{Av}\right)\right)^2 &= E\left(\overline{y}_{Av} - \mu_{Av}\right)^2 = \frac{1}{\left(N_{Av}\right)^2}\sum_{k,l\in U_v^A}\left(\pi_{vkl} - \pi_{vk}\pi_{vl}\right)\frac{y_k y_l}{\pi_{vk}\pi_{vl}} \\
&\leq \frac{1}{N_{Av}\pi_v^*}\sum_{k\in U_v^A}\frac{y_k^2}{N_{Av}} + \frac{1}{\left(N_{Av}\right)^2\left(\pi_v^*\right)^2}\left(\sum_{k,l\in U_v^A:\, k\neq l}\left(\pi_{vkl} - \pi_{vk}^2\pi_{vl}^2\right)^2\right)^{1/2}\left(\sum_{k,l\in U_v^A:\, k\neq l} y_k^2 y_l^2\right)^{1/2} \\
&\leq \frac{1}{N_{Av}\pi_v^*}\sum_{k\in U_v^A}\frac{y_k^2}{N_{Av}}
\end{aligned}$$

$$+ \quad \frac{\pi_{N_{Av}}^2}{(N_{Av})^{1/2+s}} \left(\frac{N_{Av} \max\limits_{k \in \mathscr{A}_v} \sum\limits_{l \in \mathscr{A}_v : k \neq l} (\pi_{vkl} - \pi_{vk}\pi_{vl})^2}{(N_{Av})^{1-2s}} \right)^{1/2} \left\{ \left(\sum_{k \in U_v^A} \frac{y_k^2}{N_{Av}} \right)^2 \right\}^{1/2},$$

que converge para zero quando $v \to \infty$. Analogamente, $E\left[(\overline{y}_{Bv} - \mu_{Bv})\right]^2 \longrightarrow 0, \ \ v \to \infty$. Admitindo que

$$E\left[(\overline{y}_{Av} - \mu_{Av}) \left(\overline{y}_{(A \cap B)} - \mu_{(A \cap B)v} \right) \right] < \infty$$

$$E\left[(\overline{y}_{Bv} - \mu_{Bv}) \left(\overline{y}_{(A \cap B)} - \mu_{(A \cap B)v} \right) \right] < \infty$$

é possível notar que uma vez que N_{Av} e N_{Bv} convergem para infinito com magnitudes diferentes, então a quantidade $N_v = N_{Av} + N_{Bv} - N_{(A \cap B)v}$ (onde $N_{(A \cap B)v}$ é finita) converge para infinito a uma taxa maior que as de N_{Av} e N_{Bv}. Logo, $w_{Av} \to 0$, $w_{Bv} \to 0$ e $w_{abv} \to 0$, quando $v \to \infty$. Logo,

$$E\left(\overline{y}^{A \cup B} - \mu_v\right)^2 \to 0, \text{ quando } \ v \to \infty$$

Pela desigualdade de Markov, tem-se que

$$P\left(\left|\overline{y}_v^{A \cup B} - \overline{y}_N\right| > \varepsilon\right) \leq \frac{E\left(\overline{y}^{A \cup B} - \overline{y}_N\right)^r}{\varepsilon^r}$$

Quando $r = 2$ e $v \to \infty$, temos o resultado. ■

Resultados 6, 7, 8 e 9 - página 46

(i) Estimador $\overline{y}_{RH}$: **Prova:** Para quantidades nos cadastros e nos domínios, admite-se as condições apresentadas por Isaki & Fuller (1982). É possível notar que

$$\begin{aligned} \overline{y}_{RH(v)} - \mu_v &= w_{Av}\overline{y}_{RAv} + w_{Bv}\overline{y}_{RBv} - w_{abv}\left(p\overline{y}_{Rab(A)v} + (1-p)\overline{y}_{Rab(B)v} \right) \\ &- w_{Av}\mu_{Av} + w_{Bv}\mu_{Bv} - w_{abv}\left(p\mu_{ab(A)v} + (1-p)\mu_{ab(B)v} \right) \end{aligned}$$

$$\begin{aligned} \overline{y}_{RH(v)} - \mu_v &= w_{Av}\left(\overline{y}_{RAv} - \mu_{Av}\right) + w_B\left(\overline{y}_{RBv} - \mu_{Bv}\right) \\ &- w_{A \cap B}\left\{ p\left(\overline{y}_{Rab(A)v} - \mu_{ab(A)v} \right) + (1-p)\left(\overline{y}_{Rab(B)v} - \mu_{ab(B)v} \right) \right\} \end{aligned}$$

Tem-se que para A:

$$\overline{y}_R^A - \mu_{Av} = \frac{1}{N_{Av}} \sum_{k \in \mathscr{A}_v} y_k \left(\frac{I_{vk}}{\pi_{vk}} - 1 \right) - \frac{1}{N_{Av}} \sum_{j=0}^{q} \sum_{k \in \mathscr{A}_v} \hat{\beta}_{vj}^A x_{kj} \left(\frac{I_{vk}}{\pi_{vk}} - 1 \right)$$

Através do método de linearização de Taylor aplicado no segundo termo da expressão acima, tem-se que:

$$\overline{y}_{RA} - \mu_{Av} \approx \frac{1}{N_{Av}} \sum_{k \in \mathscr{A}_v} y_k \left(\frac{I_{vk}}{\pi_{vk}} - 1 \right) - \sum_{j=0}^{q} \beta_j^A \left(\overline{x}_{jt}^A - \overline{x}_{N_{Av}} \right)$$

A variância aproximada de $\overline{y}_{RAv} - mu_{Av}$ é dada por:

$$\begin{aligned} A\text{Var}\left(\overline{y}_{RAv} - \mu_{Av}\right) &= \text{Var}\left\{ \sum_{k \in \mathscr{A}_v} y_k \left(\frac{I_{vk}}{\pi_{vk}} - 1 \right) \right\} + \text{Var}\left\{ \sum_{j=0}^{q} \beta_j^A \left(\overline{x}_{jv}^A - \overline{x}_{N_{Av}} \right) \right\} \\ &- \text{Cov}\left[\frac{1}{N_{Av}} \sum_{k \in \mathscr{A}_v} y_k \left(\frac{I_{vk}}{\pi_{vk}} - 1 \right) \; ; \; \sum_{j=0}^{q} \beta_j^A \left(\overline{x}_{jv}^A - \overline{x}_{N_{Av}} \right) \right] \end{aligned}$$

Logo,

$$\begin{aligned} A\text{Var}\left(\overline{y}_{RAv} - \mu_{Av}\right) &\leq \text{Var}\left\{ \sum_{k \in \mathscr{A}_v} y_k \left(\frac{I_{vk}}{\pi_{vk}} - 1 \right) \right\} + \text{Var}\left\{ \sum_{j=0}^{q} \beta_{vj}^A \left(\overline{x}_{jt}^A - \overline{x}_{N_{Av}} \right) \right\} \\ &= \text{Var}\left\{ \sum_{k \in \mathscr{A}_v} y_k \left(\frac{I_{vk}}{\pi_{vk}} - 1 \right) \right\} + \sum_{j=0}^{q} \left(\beta_{vj}^A \right)^2 \text{Var}\left(\overline{x}_{jt}^A - \overline{x}_{N_{Av}} \right) \\ &+ \sum_{j \neq j'} \beta_j^A \beta_{j'}^A \text{Cov}\left[\left(\overline{x}_{vj}^A - \overline{x}_{N_{Av}} \right) \; ; \; \left(\overline{x}_{vj'}^A - \overline{x}_{N_{Av}} \right) \right] \\ &= \alpha_1 + \alpha_2 + \alpha_3 \end{aligned}$$

Pelos resultados obtidos por Isaki & Fuller (1982), tem-se que

1. $\alpha_1 = \text{Var}\left\{ \sum_{k \in \mathscr{A}_v} y_k \left(\frac{I_{vk}}{\pi_{vk}} - 1 \right) \right\} = O\left((n_v^A)^{-2\delta} \right)$;

2. $\alpha_2 = \sum_{j=0}^{q} \left(\beta_{vj}^A \right)^2 \text{Var}\left(\overline{x}_{jt}^A - \overline{x}_{N_{Av}} \right) = O\left((n_v^A)^{-2\delta} \right)$

Além disso, tem-se que

$$\begin{aligned}
\alpha_3 &= \sum_{j\neq j'} \beta_j^A \beta_{j'}^A \text{Cov}\left[\left(\overline{x}_{vj}^A - \overline{x}_{N_{Av}}\right) ; \left(\overline{x}_{vj'}^A - \overline{x}_{N_{Av}}\right)\right] \\
&\leq \sum_{j\neq j'} \beta_{vj}^A \beta_{vj'}^A \left\{\text{Cov}\left[\left(\overline{x}_{vj}^A - \overline{x}_{N_{Av}}\right) ; \left(\overline{x}_{vj'}^A - \overline{x}_{N_{Av}}\right)\right]\right\}^2 \\
&\leq \sum_{j\neq j'} \beta_{vj}^A \beta_{vj'}^A \text{Var}\left[\left(\overline{x}_{vj}^A - \overline{x}_{N_{Av}}\right)\right] \text{Var}\left[\left(\overline{x}_{vj'}^A - \overline{x}_{N_{Av}}\right)\right] = O\left((n_v^A)^{-4\delta}\right)
\end{aligned}$$

Portanto, $A\text{Var}\left(\overline{y}_{RAv} - \mu_{Av}\right) = \alpha_1 + \alpha_2 + \alpha_3 = O\left(\max\left\{n_{t(A)}^{-2\delta} \, , \, (n_v^A)^{-4\delta}\right\}\right) = O\left((n_v^A)^{-2\delta}\right)$

Analogamente, é possível mostrar que

- $A\text{Var}\left(\overline{y}_{RBv} - \mu_{Bv}\right) = O\left((n_v^B)^{-2\delta}\right)$
- $A\text{Var}\left(\overline{y}_{Rab(A)v} - \mu_{ab(A)v}\right) = O\left[\left(n_v^{ab(A)}\right)^{-2\delta}\right]$
- $A\text{Var}\left(\overline{y}_{Rab(B)v} - \mu_{ab(B)v}\right) = O\left[\left(n_v^{ab(B)}\right)^{-2\delta}\right]$

É possível mostrar ainda que

$$\begin{aligned}
A\text{Var}\left(\overline{y}_{RH(v)} - \mu_v\right) &= (w_{Av})^2 \left[A\text{Var}\left(\overline{y}_{RAv} - \mu_{Av}\right)\right] + (w_{Bv})^2 \left[A\text{Var}\left(\overline{y}_{RBv} - \mu_{Bv}\right)\right] \\
&+ (w_{abv})^2 p^2 \left[A\text{Var}\left(\overline{y}_{Rab(A)v} - \mu_{ab(A)v}\right)\right] + w_{abv}^2 (1-p)^2 \left[A\text{Var}\left(\overline{y}_{Rab(B)v} - \mu_{ab(B)v}\right)\right] \\
&+ 2p w_{Av} w_{abv} A\text{Cov}\left[\left(\overline{y}_{RAv} - \mu_{Av}\right) ; \left(\overline{y}_{Rab(A)v} - \mu_{ab(A)v}\right)\right] \\
&+ 2(1-p) w_{Bv} w_{abv} A\text{Cov}\left[\left(\overline{y}_{RB} - \mu_{Bv}\right) ; \left(\overline{y}_{Rab(B)v} - \mu_{ab(B)v}\right)\right] \\
&\leq (w_{Av})^2 \left[A\text{Var}\left(\overline{y}_{RAv} - \mu_{Av}\right)\right] + (w_{Bv})^2 \left[A\text{Var}\left(\overline{y}_{RBv} - \mu_{Bv}\right)\right] \\
&+ (w_{abv})^2 p^2 \left[A\text{Var}\left(\overline{y}_{Rab(A)v} - \mu_{ab(A)v}\right)\right] + w_{abv}^2 (1-p)^2 \left[A\text{Var}\left(\overline{y}_{Rab(B)v} - \mu_{ab(B)v}\right)\right] \\
&+ p w_{Av} w_{abv} \left\{A\text{Cov}\left[\left(\overline{y}_{RAv} - \mu_{Av}\right) ; \left(\overline{y}_{Rab(A)v} - \mu_{ab(A)v}\right)\right]\right\}^2 \\
&+ (1-p) w_{Bv} w_{abv} \left\{A\text{Cov}\left[\left(\overline{y}_{RB} - \mu_{Bv}\right) ; \left(\overline{y}_{Rab(B)v} - \mu_{ab(B)v}\right)\right]\right\}^2 \\
&\leq (w_{Av})^2 \left[A\text{Var}\left(\overline{y}_{RAv} - \mu_{Av}\right)\right] + w_B^2 \left[A\text{Var}\left(\overline{y}_{RBv} - \mu_{Bv}\right)\right] \\
&+ (w_{abv})^2 p^2 \left[A\text{Var}\left(\overline{y}_{Rab(A)v} - \mu_{ab(A)v}\right)\right] + w_{abv}^2 (1-p)^2 \left[A\text{Var}\left(\overline{y}_{Rab(B)v} - \mu_{ab(B)v}\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&+ \quad pw_{Av}w_{abv}AVar\left[(\overline{y}_{RAv}-\mu_{Av})\right]AVar\left[\left(\overline{y}_{Rab(A)v}-\mu_{ab(A)v}\right)\right] \\
&+ \quad (1-p)w_{Bv}w_{abv}AVar\left[(\overline{y}_{RBv}-\mu_{Bv})\right]AVar\left[\left(\overline{y}_{Rab(B)v}-\mu_{ab(B)v}\right)\right] \\
&= \quad \mathbf{O}\left\{\max\left((n_{Av})^{-2\delta}\,,\,(n_{Bv})^{-2\delta}\right)\right\}.
\end{aligned}
$$

Com este resultado e conforme condições descritas para a situação de apenas um cadastro, apresentadas por Isaki & Fuller (1982), tem-se finalmente que

$$\overline{y}_{RH(v)}-\mu_v=\mathbf{O}_p\left(\max\left\{(n_{Av})^{-\delta};(n_{Bv})^{-\delta}\right\}\right).$$

Robinson & Särndal (1983) mostraram que os estimadores regressão para cadastros e domínios são assintoticamente centrados, ou seja:

$$\lim_{v\to\infty}\left\{E(\overline{y}_R^C|\mathbf{Y}_v)-\mu_v\right\}=0 \quad \text{e} \quad \lim_{v\to\infty}\left\{E(\overline{y}_R^d|\mathbf{Y}_v)-\mu_v\right\}=0$$

Estes resultados são garantidos pela desigualdade de Markov, que fornece condições suficientes para mostrar que

$$\lim_{v\to\infty}E\left\{\overline{y}^C_{GREG(v)}-\overline{y}_v|\mathbf{Y}_v\right\}=0,$$

da seguinte forma: seja

$$\overline{y}_R^C-\mu_v=\frac{1}{N_v^C}\sum_{k\in U_v^C}y_k\left(\frac{I_{vk}}{\pi_{vk}}-1\right)-\sum_{j=1}^{q}\hat{\beta}_{jv}\frac{1}{N_v}\sum_{k\in U_v^C}x_{kj}\left(\frac{I_{vk}}{\pi_{vk}}-1\right)$$

Logo,

$$
\begin{aligned}
E\left\{\overline{y}_R^C-\mu_v|\mathbf{Y}_v\right\} &\leq \left\{E\left[\left(\frac{1}{N_v^C}\sum_{k\in U_v^C}y_k\left(\frac{I_{vk}}{\pi_{vk}}-1\right)\right)^2\middle|\mathbf{Y}_v\right]\right\}^{1/2} \\
&+ \left\{E\left[\sum_{j=0}^{q}\hat{\beta}_{vj}^2|\mathbf{Y}_v\right]\sum_{j=0}^{q}E\left\{\left[\sum_{k\in U_v^C}x_{kj}\left(\frac{I_{vk}}{\pi_{vk}}-1\right)\right]^2\right\}\right\}^{1/2} \\
&= Q_1^{1/2}+Q_2^{1/2}.
\end{aligned}
$$

Tem-se que

$$\begin{aligned} Q_1 &= E\left[\left(\frac{1}{N_\nu^C}\sum_{k\in U_\nu^C} y_k\left(\frac{I_{\nu k}}{\pi_{\nu k}}-1\right)\right)^2 \Big| \mathbf{Y}_\nu\right] \\ &= \frac{1}{N_\nu^2}E\left\{\sum_{k\in U_\nu^C} y_k^2\left(\frac{I_{\nu k}}{\pi_{\nu k}}-1\right)^2 + 2\sum_k\sum_{\substack{l\\ k\neq l}} y_k y_l\left(\frac{I_{\nu k}I_{\nu l}}{\pi_{\nu k}\pi_{\nu l}}-\frac{I_{\nu k}}{\pi_{\nu k}}-\frac{I_{\nu l}}{\pi_{\nu l}}-1\right)\right\} \\ &= \frac{1}{N_\nu^2}\sum_{k\in U_\nu^C} y_k^2\left(\frac{1}{\pi_{\nu k}}-1\right)^2 + \frac{2}{N_\nu^2}\sum_k\sum_{\substack{l\\ k\neq l}} y_k y_l\left(\frac{\pi_{\nu kl}}{\pi_{\nu k}\pi_{\nu l}}-1\right). \end{aligned}$$

Dessa forma,

$$\begin{aligned} \frac{1}{N_\nu^2}\sum_{k\in U_\nu^C} y_k^2\left(\frac{1}{\pi_{\nu k}}-1\right)^2 &\leq \frac{N_\nu^{-1}\sum_{k\in U_\nu^C} y_k^2}{N_\nu \min_{1\leq k\leq N_\nu}\pi_{\nu k}} \longrightarrow 0, \quad \text{quando } \nu\to\infty \\ \frac{2}{N_\nu^2}\sum_k\sum_{\substack{l\\ k\neq l}} y_k y_l\left(\frac{\pi_{\nu kl}}{\pi_{\nu k}\pi_{\nu l}}-1\right) &\leq \max_{1\leq k\neq l\leq N_\nu}\left|\frac{\pi_{\nu kl}}{\pi_{\nu k}\pi_{\nu l}}-1\right|\left(\frac{\sum_{k\in U_\nu^C}|y_k|}{N_\nu}\right)^2 \\ &\leq \max_{1\leq k\neq l\leq N_\nu}\left|\frac{\pi_{\nu kl}}{\pi_{\nu k}\pi_{\nu l}}-1\right|\frac{\sum_{k\in U_\nu^C} y_k^2}{N_\nu} \longrightarrow 0, \quad \text{quando } \nu\to\infty \end{aligned}$$

Analogamente, $Q_2 \to 0$, quando $\nu\to\infty$, pois

$$E\left\{\left[\sum_{k\in U_\nu^C} x_{kj}\left(\frac{I_{\nu k}}{\pi_{\nu k}}-1\right)\right]^2\right\} \longrightarrow 0, \quad \text{quando } \nu\to\infty,$$

o que é garantido pelas condições **C6**, **C8**, **C9** e **C10**. Portanto, o estimador $\overline{y}_{RH(\nu)}$, dado em função de estimadores regressão para cadastros e domínios é assintoticamente centrado e consistente. ■

(ii) Estimador $\overline{y}_{RFB}$: É possível notar que através do método de linearização de Taylor, tem-se que

$$\begin{aligned}
\overline{y}_{RFB(v)} &\approx (\hat{w}_{av}\mu_a + w_{av}\overline{y}_{RAv} - w_{av}\mu_a) + (\hat{w}_{bv}\mu_b + w_{bv}\overline{y}_{RBv} - w_{bv}\mu_b) \\
&+ p_1\left(\hat{w}_{ab(A)v}\mu_{abv} + w_{abv}\overline{y}_{Rab(A)v} - w_{abv}\mu_{abv}\right) + (1-p_1)\left(\hat{w}_{abv}\mu_{abv} + w_{abv}\mu_{abv}\right) \\
&+ N_v^{-1}p_2\left(\hat{N}_{ab(A)v} - \hat{N}_{ab(B)v}\right) \\
&= w_{av}\overline{y}_{Rav} + w_{bv}\overline{y}_{Rbv} + p_1 w^{abv}\overline{y}_{Rab(A)v} + (1-p_1)w_{abv}\overline{y}_{Rab(B)v} \\
&+ \hat{w}_{av}\mu_{av} + \hat{w}_{bv}\mu_{bv} + p_1\hat{w}_{abv}\mu_{abv} + (1-p_1)\hat{w}_{abv}\mu_{abv} + N_v^{-1}p_2\left(\hat{N}_{ab(A)} - \hat{N}_{ab(B)}\right) \\
&- w_{av}\mu_{av} - w_{bv}\mu_{bv} - w_{abv}\mu_{abv}
\end{aligned}$$

Admitindo que $\mu_v = w_{av}\mu_{av} + w_{bv}\mu_{bv} + pw_{abv}\mu_{abv} + (1-p)w_{abv}\mu_{abv}$, tem-se que

$$\begin{aligned}
\overline{y}_{RFB(v)} - \mu_v &\approx w_a\left(\overline{y}_{Rav} - \mu_{av}\right) + w_b\left(\overline{y}_{Rbv} - \mu_{bv}\right) + p_1 w_{abv}\left(\overline{y}_{Rab(A)v} - \mu_{abv}\right) + (1-p_1)w_{abv}\left(\overline{y}_{Rab(B)v} - \mu_{abv}\right) \\
&+ N_v^{-1}\left(\hat{N}_a\mu_{av} + \hat{N}_b\mu_{bv} + p_1\hat{N}_{ab}^A\mu_{abv} + (1-p_1)\hat{N}_{ab}^B\mu_{abv} + p_2\hat{N}_{ab}^A - p_2\hat{N}_{ab}^B\right) \\
&= w_{av}\left(\overline{y}_{RAv} - \mu_{av}\right) + p_1 w_{abv}\left(\overline{y}_{Rab(A)v} - \mu_{abv}\right) \\
&+ w_{bv}\left(\overline{y}_{RBv} - \mu_{bv}\right) + (1-p_1)w_{abv}\left(\overline{y}_{Rab(B)v} - \mu_{abv}\right) + \mathbf{O}(1).
\end{aligned}$$

Logo, a variância aproximada dessa diferença é dada por

$$\begin{aligned}
A\text{Var}\left(\overline{y}_{RFB(v)} - \mu_v\right) &= (w_{av})^2 A\text{Var}\left(\overline{y}_{Rav} - \mu_{av}\right) + (w_{bv})^2 A\text{Var}\left(\overline{y}_{Rbv} - \mu_{bv}\right) \\
&+ (p_1 w_{abv})^2 A\text{Var}\left(\overline{y}_{Rab(A)v} - \mu_{abv}\right) + (1-p_1)^2(w_{abv})^2 A\text{Var}\left(\overline{y}_{Rab(B)v} - \mu_{abv}\right) \\
&+ 2p_1 w_{av}w_{abv} A\text{Cov}\left\{\left(\overline{y}_{RAv} - \mu_{av}\right)\left(\overline{y}_{Rab(A)v} - \mu_{abv}\right)\right\}
\end{aligned}$$

$$
\begin{aligned}
&+ \quad 2(1-p_1)w_{bv}w_{abv}A\text{Cov}\left\{(\overline{y}_{RBv}-\mu_{bv})\left(\overline{y}_{Rab(B)v}-\mu_{abv}\right)\right\} \\
&+ \quad \mathbf{O}(N_v^{-1}) \\
&\leq \quad (w_{av})^2 A\text{Var}(\overline{y}_{RAv}-\mu_{av})+(w_{bv})^2 A\text{Var}(\overline{y}_{RBv}-\mu_{bv}) \\
&+ \quad (p_1 w_{abv})^2 A\text{Var}\left(\overline{y}_{Rab(A)v}-\mu_{abv}\right)+(1-p_1)^2(w_{abv})^2 A\text{Var}\left(\overline{y}_{Rab(B)v}-\mu_{abv}\right) \\
&+ \quad 2p_1 w_{av}w_{abv}A\text{Var}(\overline{y}_{RAv}-\mu_{av})A\text{Var}\left(\overline{y}_{Rab(A)v}-\mu_{abv}\right) \\
&+ \quad 2(1-p_1)w_{bv}w_{abv}A\text{Var}(\overline{y}_{RBv}-\mu_{bv})A\text{Var}\left(\overline{y}_{Rab(B)v}-\mu_{abv}\right)+\mathbf{O}(N_v^{-1}) \\
&= \quad \mathbf{O}\left\{\max\left[(n_{av})^{-2\delta}\,,\,(n_{bv})^{-2\delta}\,,\,(n_v^{ab(A)})^{-2\delta}\,,\,(n_v^{ab(B)})^{-2\delta}\right]\right\}=\mathbf{O}\left\{(n_v^*)^{-2\delta}\right\}
\end{aligned}
$$

Além disso, para cadastros e domínios individualmente vale também os resultados apresentados Robinson & Särndal (1983), da mesma forma que a feita para o estimador $\overline{y}_{RH(v)}$. Portanto, sob a abordagem de cadastro duplo, o estimador $\overline{y}_{RFB(v)}$, função de estimadores regressão para cadastros e domínios também é assintoticamente centrado e consistente.

(iii) Estimador $\overline{y}_{RBLG}$:

Da mesma forma que as apresentadas para o teorema 3, são admitidas para os domínios as condições apresentadas por Isaki & Fuller (1982). Como a estratégia de Bankier é uma adaptação da estratégia de Horvitz-Thompson a partir da identificação dos elementos pertencentes aos domínios a, b e ab, tem-se que

$$
\begin{aligned}
\overline{y}_{RBa(v)} &= (w_{av})\overline{y}_{Rav}+(w_{bv})\overline{y}_{Rbv}+w_{abv}\overline{y}_{Rabv} \\
\mu_v &= \frac{t_{yA}+t_{yB}}{N_v}=\frac{t_{ya}+t_{yab(A)}+t_{yb}+t_{yab(B)}}{N}=\frac{t_{ya}+t_{yb}+t_{yab}}{N}, \quad \text{onde} \quad t_{yab}=t_{yab(A)}+t_{yab(B)} \\
\mu_v &= (w_{av})\mu_{av}+(w_{bv})\mu_{bv}+w_{abv}\mu_{abv}
\end{aligned}
$$

Dessa forma, tem-se que

$$
\begin{aligned}
\overline{y}_{RBa(v)} - \mu_v &= (w_{av})\,(\overline{y}_{Rav} - \mu_{av}) + (w_{bv})\,(\overline{y}_{RBv} - \mu_{bv}) + w_{abv}\,(\overline{y}_{Rabv} - \mu_{abv}) \\
\\
AVar\left(\overline{y}_{RBa(v)} - \mu_v\right) &= (w_{av})^2 AVar\,(\overline{y}_{Rav} - \mu_{av}) + (w_{bv})^2 AVar\,(\overline{y}_{Rbv} - \mu_{bv}) \\
&+ (w_{abv})^2 AVar\,(\overline{y}_{Rabv} - \mu_{abv}) \\
\\
&+ 2(w_{av})w_{abv} ACov\left[(\overline{y}_{Rav} - \mu_{av})\ ;\ (\overline{y}_{Rabv} - \mu_{abv})\right] \\
\\
&+ 2(w_{bv})w_{abv} ACov\left[(\overline{y}_{Rbv} - \mu_{bv})\ ;\ (\overline{y}_{Rabv} - \mu_{abv})\right] \\
\\
&\leq (w_{av})^2 AVar\,(\overline{y}_{Rav} - \mu_{av}) + (w_{bv})^2 AVar\,(\overline{y}_{Rbv} - \mu_{bv}) \\
&+ (w_{abv})^2 AVar\,(\overline{y}_{Rabv} - \mu_{abv}) \\
\\
&+ 2(w_{av})w_{abv} \left\{ACov\left[(\overline{y}_{Rav} - \mu_{av})\ ;\ (\overline{y}_{Rabv} - \mu_{abv})\right]\right\}^2 \\
\\
&+ 2(w_{bv})w_{abv} \left\{ACov\left[(\overline{y}_{Rbv} - \mu_{bv})\ ;\ (\overline{y}_{Rabv} - \mu_{abv})\right]\right\}^2 \\
\\
&\leq (w_{av})^2 AVar\,(\overline{y}_{Rav} - \mu_{av}) + (w_{bv})^2 AVar\,(\overline{y}_{Rbv} - \mu_{bv}) \\
&+ (w_{abv})^2 AVar\,(\overline{y}_{Rabv} - \mu_{abv}) \\
\\
&+ 2(w_{av})w_{abv} \left\{AVar\,(\overline{y}_{Rav} - \mu_{av})\, AVar\,(\overline{y}_{Rabv} - \mu_{abv})\right\} \\
\\
&+ 2(w_{bv})w_{abv} \left\{AVar\,(\overline{y}_{Rbv} - \mu_{bv})\, AVar\,(\overline{y}_{Rabv} - \mu_{abv})\right\} \\
\\
&= \mathbf{O}\left(n_{a(v)}^{-2\delta}\right) + \mathbf{O}\left((n_{bv})^{-2\delta}\right) + \mathbf{O}\left((n_v^{ab})^{-2\delta}\right) + \mathbf{O}\left((n_{av})^{-2\delta}(n_v^{ab})^{-2\delta}\right) \\
\\
&+ \mathbf{O}\left((n_{bv})^{-2\delta}(n_v^{ab})^{-2\delta}\right) = \mathbf{O}\left(\max\left\{(n_{av})^{-2\delta}, (n_{bv})^{-2\delta}, (n_v^{ab})^{-2\delta}\right\}\right),
\end{aligned}
$$

Adaptando os resultados apresentados por Isaki & Fuller (1982) e Robinson & Särndal (1983) para as quantidades referentes à cadastros e domínios, da mesma forma que a feita para o estimador $\overline{y}_{RH(v)}$, tem-se que o estimador $\overline{y}_{RBa(v)}$ é assintoticamente centrado e consistente.

(iv) Estimador $\overline{y}_{RBLG|(\infty)}$:

Inicialmente, sejam $\omega_{ka}^{(\infty)}$, $\omega_{kb}^{(\infty)}$, $\omega_{kab}^{(\infty)}$ dados da mesma forma que a apresentada por Skinner (1991):

$$\begin{aligned}
\omega_{ka}^{(\infty)} &= \frac{1}{\pi_{vk}^{A}}\left[\prod_{r \text{ par}}\left(\frac{N_{Av}}{\hat{N}^{A(r-1)}}\right)\right] = \frac{1}{\pi_{vk}^{A}}\alpha^{A} \\
\omega_{kb}^{(\infty)} &= \frac{1}{\pi_{vk}^{B}}\left[\prod_{r \text{ ímpar}}\left(\frac{N_{Bv}}{\hat{N}^{B(r-1)}}\right)\right] = \frac{1}{\pi_{vk}^{B}}\alpha^{B} \\
\omega_{kab}^{(\infty)} &= \frac{1}{\pi_{vk}^{A}+\pi_{vk}^{B}}\left[\prod_{r \text{ par}}\left(\frac{N_{Av}}{\hat{N}^{A(r-1)}}\right)\prod_{r \text{ ímpar}}\left(\frac{N_{Bv}}{\hat{N}^{B(r-1)}}\right)\right] = \left(\frac{1}{\pi_{vk}^{A}+\pi_{vk}^{B}}\right)\alpha^{A}\alpha^{B}
\end{aligned}$$

α^{A} e α^{B} convergem para valores fixos quanto $r \to \infty$. Dessa forma, basta substituir as quantidades apresentadas em (3.10) e temos o resultado, da mesma forma que a feita para o estimador $\overline{y}_{RBa(v)}$.

(v) Estimador $\overline{y}_{RPML}$:

É possível notar que

$$\begin{aligned}
\overline{y}_{RPML(v)} &\approx w_{av}\overline{y}_{Ra(v)} + w_{a(*)}\mu_{av} - w_{av}\mu_{av} + w_{bv}\overline{y}_{Rb(v)} + w_{abv(*)}\mu_{Bv} - w_v^b\mu_{Bv} \\
&+ w_{abv}\left\{\left(\frac{\pi_{vk}^A}{\pi_{vk}^A+\pi_{vk}^B}\right)\overline{y}_{Rab(A)v} + \left(\frac{\pi_{vk}^B}{\pi_{vk}^A+\pi_{vk}^B}\right)\overline{y}_{Rab(B)v}\right\} \\
&+ w^{ab(*)}\mu_{abv} - w_{abv}\mu_{abv} \\
&= w_{av}\overline{y}_{Ra(v)} + w_{bv}\overline{y}_{Rb(v)} \\
&+ w_{abv}\left\{\left(\frac{\pi_{vk}^A}{\pi_{vk}^A+\pi_{vk}^B}\right)\overline{y}_{Rab(A)v} + \left(\frac{\pi_{vk}^B}{\pi_{vk}^A+\pi_{vk}^B}\right)\overline{y}_{Rab(B)v}\right\} + O(1)
\end{aligned}$$

Dessa forma, admitindo que $\mu_v = w_{av}\mu_{av} + w_{bv}\mu_b + w_{abv}\mu_{abv}$, e que

$$\begin{aligned}
w_{abv}\left(\frac{\pi_{vk}^A}{\pi_{vk}^A+\pi_{vk}^B}\right)\overline{y}_{Rab(A)v} + w_{abv}\left(\frac{\pi_{vk}^B}{\pi_{vk}^A+\pi_{vk}^B}\right)\overline{y}_{Rab(B)v} &= \frac{w_{abv}}{\pi_{vk}^A+\pi_{vk}^B}\left(\pi_{vk}^A\overline{y}_{Rab(A)v} + \pi_{vk}^B\overline{y}_{Rab(B)v}\right) \\
&= w_v^{ab(*)}\overline{y}^{ab(*)},
\end{aligned}$$

tem-se que

$$\overline{y}_{RPML(v)} - \mu_v \approx w_{av}\left(\overline{y}_{Ra(v)} - \mu_{av}\right) + w_{bv}\left(\overline{y}_{Rb(v)} - \mu_{bv}\right) + w_{abv(*)}\left(\overline{y}_{Rab(v)} - \mu_{abv}\right) + O(1)$$

Dessa forma, a variância aproximada dessa diferença é dada por

$$\begin{aligned}
A\text{Var}\left(\overline{y}_{RPML(v)} - \mu_v\right) &= (w_{av})^2 A\text{Var}\left(\overline{y}_{Ra(v)} - \mu_{av}\right) + (w_{bv})^2 A\text{Var}\left(\overline{y}_{Rb(v)} - \mu_{bv}\right) \\
&+ \left(w_{abv(*)}\right) A\text{Var}\left(\overline{y}_{Rab(v)} - \mu_{abv}\right) + O(1) \\
&= \mathbf{O}\left(\max\left\{(n_{av})^{-2\delta}, (n_{bv})^{-2\delta}, (n_v^{ab(A)})^{-2\delta}, (n_v^{ab(B)})^{-2\delta}\right\}\right) = \mathbf{O}\left((n_v^*)^{-2\delta}\right),
\end{aligned}$$

de acordo com os resultados obtidos Isaki & Fuller (1982) e Robinson & Särndal (1983) e apresentados anteriormente para os outros estimadores propostos.

Apêndice C

Programas utilizados

As avaliações computacionais realizadas ao longo desta tese foram feitas utilizando o ambiente de programação, análise de dados e gráficos R em sua versão 2.12.0, e que se encontra gratuitamente disponível em http://www.r-project.org. A opção por esta linguagem foi devido ao compromisso entre a flexibilidade oferecida por algumas linguagens compiladas e a conveniência dos tradicionais pacotes estatísticos, além da facilidade de implementação dos estimadores propostos. A seguir será apresentado o programa utilizado para a avaliação dos estimadores sob o plano de amostragem aleatória simples.

```
################################################################################
#UNIVERSIDADE FEDERAL DE PERNAMBUCO
#CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
#PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA
#DOUTORADO EM ESTATÍSTICA
################################################################################
#INFERÊNCIA SOB PLANOS AMOSTRAIS DE CADASTROS DUPLOS
################################################################################
#HEMÍLIO FERNANDES CAMPOS COELHO
################################################################################

################################################################################
#O PROGRAMA A SEGUIR FOI UTILIZADO PARA A APLICAÇÃO APRESENTADA NO CAPÍTULO 2
################################################################################

################################################################################
#Função para criar uma população artificial de tamanho N baseada em dados de uma
#amostra de tamanho n.
################################################################################
# name -> Nome que a variável receberá
################################################################################

artipop<-function(N,y,name)
{
y<-as.vector(unlist(y))
n<-length(y)
nv<-log(N/n)/(log(2))
nvr<-floor(nv+1)
```

```
for(k in 1:nvr){
y<-append(y,y, after=n)}
var<-data.frame(y[1:N])
names(var)<-c(name)
return(var)
}

################################################################################
#Função para cálculo do p-otimo do estimador de Hartley caso normal.
################################################################################
# N_A -> Tamanho da população do cadastro A
# NB -> Tamanho da população do cadastro B
# Nab -> Tamanho da população do domínio "ab"
# yabC -> Variável resposta na amostra para o domínio ab do cadastro C
################################################################################

pH=function(N_A,NB,Nab,yabA,yabB)
{
var1=var(yabA)
var2=var(yabB)
fAinv=N_A/nA
fBinv=NB/nB
wabA=Nab/N_A
wabB=Nab/NB
a=(2*NB*fBinv*wabB*var2)/(2*N_A*fAinv*wabA*var1+2*NB*fBinv*wabB*var2)
return(a)
}

################################################################################
#Função que calcula o estimador de Hartley
################################################################################
# p -> Função "pH" aplicada anteriormente
# Na -> Tamanho da população do domínio "a"
# Nb -> Tamanho da população do domínio "b"
# yd -> Variável de interesse para o domínio d
# ydC -> Variável de interesse para o domínio d, cadastro C
################################################################################

esthart=function(p,Na,Nb,ya,yb,yabA,yabB)
{
est=Na*mean(ya)+Nb*mean(yb)+Nab*p*mean(yabA)
+Nab*(1-p)*mean(yabB)
return(est)
}


############################ CONSTRUÇÃO DO BANCO DE DADOS ###################################
library(TeachingSampling)
#Cadastro de Hortaliças (Cadastro B)
cadhor<- read.table("marco hortalizas.tab", header=TRUE, sep="", na.strings="NA", dec=".", strip.white=TRUE)

#Dados de área cultivada, área colhida e produção (variáveis de interesse)
acult<- read.table("asemh.tab", header=TRUE, sep="", na.strings="NA", dec=".", strip.white=TRUE)
acolhi<- read.table("acosh.tab", header=TRUE, sep="", na.strings="NA", dec=".", strip.white=TRUE)
```

```
producao<- read.table("prodh.tab", header=TRUE, sep="", na.strings="NA", dec=".", strip.white=TRUE)

#Cadastro do DANE (Cadastro A)
cadENA <- read.table("universo1.tab", header=TRUE, sep="", na.strings="NA", dec=".", strip.white=TRUE)

#Criação de um cadastro geral
N=nrow(cadENA)
Nab= 10000
Na=trunc(nrow(cadENA)/2)-Nab
Nb=trunc(nrow(cadENA)/2+1)-Nab
N_A=Na+Nab
N_B=Nb+Nab
domA=append(matrix(1,Na),matrix(0,Nab))
domB=append(matrix(0,Nab),matrix(1,Nb))
cult= artipop(N,acult,"area cultivada")
colhi=artipop(N,acolhi,"area colhida")
prod=artipop(N,producao,"produção")
cadgeral=data.frame(cadENA,cult,colhi,prod)

####CADASTROS
cadA=data.frame(cadgeral[1:N_A,],domA)
cadB=data.frame(cadgeral[(N_A+1-Nab):(N_A+N_B-Nab),],domB)
domab=subset(cadA,domA==0)


####PARÂMETROS DE INTERESSE
media_tot=(sum(cadA\$produção)+sum(cadB\$produção)-sum(domab\$produção))/N


####COLETA DAS INFORMAÇÕES DA AMOSTRA
nA=1000
nB=800
selA=S.SI(N_A,nA)
amoA=cadA[selA,]
nabA=nrow(domabA)
selB=S.SI(N_B,nB)
amoB=cadB[selB,]

####DOMÍNIOS
doma=subset(amoA,domA==1)
domabA=subset(amoA,domA==0)
domb=subset(amoB,domB==1)
domabB=subset(amoB,domB==0)


#####CÁLCULO DO ESTIMADOR DE HARTLEY
phart=pH(N_A,N_B,Nab,domabA\$produção,domabB\$produção)
hartley=esthart(phart,Na,Nb,doma\$produção,domb\$produção,domabA\$produção,domabB$produção)/N
mean(doma\$produção)
mean(domb\$produção)
mean(domabA\$produção)
mean(domabB\$produção)
```

```
################################################################################
# SIMULAÇÃO:
#O PROGRAMA A SEGUIR AVALIA OS ESTIMADORES DE HARTLEY, FULLER, BANKIER,
#PML, E OS ESTIMADORES PROPOSTOS ATRAVÉS DO USO DE ESTIMADORES GREG
################################################################################

################################################################################
#Função para cálculo do p-otimo do estimador de Hartley caso normal.
################################################################################
# nA -> Tamanho da amostra obtida do cadastro A
# N_A -> Tamanho da população do cadastro A
# nB -> Tamanho da amostra obtida do cadastro B
# NB -> Tamanho da população do cadastro B
# sA -> Amostra obtida do cadastro A
# sB -> Amostra obtida do cadastro B
# Nab -> Tamanho da população do domínio "ab"
# dab -> Indicador numérido do domínio "ab"
################################################################################

pH=function(nA,N_A,nB,NB,sA,sB,Nab,dab)
{
domabA=subset(sA,sA[,2]==dab)
doma=subset(sA,sA[,2]==(1-dab))
domabB=subset(sB,sB[,2]==dab)
domb=subset(sB,sB[,2]==(1-dab))
var1=var(domabA[,1])
var2=var(domabB[,1])
fAinv=N_A/nA
fBinv=NB/nB
wabA=Nab/N_A
wabB=Nab/NB
a=(2*NB*fBinv*wabB*var2)/(2*N_A*fAinv*wabA*var1+2*NB*fBinv*wabB*var2)
return(a)
}

################################################################################
#Função que calcula o estimador de Hartley
################################################################################
# p -> Função "pH" aplicada anteriormente
# nA -> Tamanho da amostra obtida do cadastro A
# N_A -> Tamanho da população do cadastro A
# nB -> Tamanho da amostra obtida do cadastro B
# NB -> Tamanho da população do cadastro B
# sA -> Amostra obtida do cadastro A
# sB -> Amostra obtida do cadastro B
# Nab -> Tamanho da população do domínio "ab"
# dab -> Indicador numérido do domínio "ab"
################################################################################
```

```
esthart=function(p,n_A,nB,N_A,NB,sA,sB,Nab,dab)
{
domabA=subset(sA,sA[,2]==dab)
doma=subset(sA,sA[,2]==(1-dab))
domabB=subset(sB,sB[,2]==dab)
domb=subset(sB,sB[,2]==(1-dab))
Na=N_A-Nab
Nb=NB-Nab
est=Na*mean(doma[,4])+Nb*mean(domb[,4])+Nab*p*mean(domabA[,4])
+Nab*(1-p)*mean(domabB[,4])
return(est)
}


#################################################################################
#Função que estima Nab para o estimador de Fuller-Burmeister
#################################################################################
# N_A -> Tamanho da população do cadastro A
# NB -> Tamanho da população do cadastro B
# sA -> Amostra obtida do cadastro A
# sB -> Amostra obtida do cadastro B
# dA -> Variável indicadora de domínios (1 se pertence a "a"; 0 se pertence a "ab")
# dB -> Variável indicadora de domínios (1 se pertence a "b"; 0 se pertence a "ab")
# dab -> Indicador numérido do domínio "ab"
#################################################################################

Nabful<- function(N_A,NB,sA,sB,dA,dB,dab)
{
domabA<-subset(sA,dA==dab)
domabB<-subset(sB,dB==dab)
nabA<-length(domabA[,1])
nabB<-length(domabB[,1])
nA<-length(sA[,1])
nB<-length(sB[,1])
gA<-(N_A - nA)/(N_A - 1)
gB<-(NB - nB)/(NB - 1)
z3 <-(nA*gB)+(nB*gA)
z2 <-nA*NB*gB+nB*N_A*gA+nabA*N_A*gB+nabB*NB*gA
z1 <-(nabA*gB+nabB*gA)*N_A*NB
if(missing(z1) || is.null(z1) || is.na(z1) || !is.numeric(z1)){z1<-0}
if(missing(z2) || is.null(z2) || is.na(z2) || !is.numeric(z2)){z2<-0}
if(missing(z3) || is.null(z3) || is.na(z3) || !is.numeric(z3)){z3<-0}
roots<-polyroot(c(z1,-z2,z3))
NabFB <- min(Re(roots))
return(NabFB)
}

#################################################################################
#Função que calcula o estimador de Fuller & Burmeister
#################################################################################
#Precisa da função "Nabful"
#################################################################################
# nA -> Tamanho da amostra obtida do cadastro A
# N_A -> Tamanho da população do cadastro A
```

```
# nB -> Tamanho da amostra obtida do cadastro B
# NB -> Tamanho da população do cadastro B
# sA -> Amostra obtida do cadastro A
# sB -> Amostra obtida do cadastro B
# Nabf -> Tamanho da população do domínio "ab", estimada pela função "Nabful"
# dab -> Indicador numérido do domínio "ab"
################################################################################

estful=function(n_A,nB,N_A,NB,sA,sB,Nabf,dab)
{
domabA=subset(sA,sA[,2]==dab)
doma=subset(sA,sA[,2]==(1-dab))
domabB=subset(sB,sB[,2]==dab)
domb=subset(sB,sB[,2]==(1-dab))
nabA=length(domabA[,1])
nabB=length(domabB[,1])
fA=n_A/N_A
fB=nB/NB
w=nabA*(1-fB)/(nabA*(1-fB)+nabB*(1-fA))
yabest=w*mean(domabA[,4])+(1-w)*mean(domabB[,4])
Naest=(N_A-Nabf)
Nbest=(NB-Nabf)
estful=Naest*mean(doma[,4])+Nbest*mean(domb[,4])+Nabf*yabest
return(estful)
}

################################################################################
#Função que calcula o estimador de Bankier
################################################################################
# nA -> Tamanho da amostra obtida do cadastro A
# N_A -> Tamanho da população do cadastro A
# nB -> Tamanho da amostra obtida do cadastro B
# NB -> Tamanho da população do cadastro B
# sA -> Amostra obtida do cadastro A
# sB -> Amostra obtida do cadastro B
# dab -> Indicador numérido do domínio "ab"
################################################################################

estbank=function(n_A,nB,N_A,NB,sA,sB,dab)
{
domabA=subset(sA,sA[,2]==dab)
doma=subset(sA,sA[,2]==(1-dab))
domabB=subset(sB,sB[,2]==dab)
domb=subset(sB,sB[,2]==(1-dab))
fA=n_A/N_A
fB=nB/NB
somapondA=(1/fA)*sum(doma[,4])
somapondabA=(1/(fA+fB))*sum(domabA[,4])
soma1=somapondA+somapondabA
somapondB=(1/fB)*sum(domb[,4])
somapondabB=(1/(fA+fB))*sum(domabB[,4])
soma2=somapondB+somapondabB
est=soma1+soma2
return(est)
```

```
}


#############################################################################################
#Função que calcula o estimador de Horvitz-Thompson para o caso do estimador Bankier Combinado
#############################################################################################
#Esta função é uma alteração da função "E.SI" do pacote "TeachingSampling"
#############################################################################################
E.SImod=function(N_A,NB, n_A,nB,nd,y)
{
y <- as.data.frame(y)
val1=n_A/N_A
val2=nB/NB
pik <- rep(1/(val1+val2),nd)
dk <- 1/pik
ty <- sum(y * dk)
return(ty)
}


#############################################################################
#Função que calcula o estimador de Nab para a estratégia PML
#############################################################################
# nA -> Tamanho da amostra obtida do cadastro A
# N_A -> Tamanho da população do cadastro A
# nB -> Tamanho da amostra obtida do cadastro B
# NB -> Tamanho da população do cadastro B
# sA -> Amostra obtida do cadastro A
# sB -> Amostra obtida do cadastro B
# dab -> Indicador numérido do domínio "ab"
#############################################################################

NabPML = function(n_A,nB,N_A,NB,sA,sB,dab)
{
domabA=subset(sA,sA[,2]==dab)
doma=subset(sA,sA[,2]==(1-dab))
domabB=subset(sB,sB[,2]==dab)
domb=subset(sB,sB[,2]==(1-dab))
fA=n_A/N_A
fB=nB/NB
vetor1=rep(1,length(domabA[,4]))
vetor2=rep(1,length(domabB[,4]))
nabA=sum(vetor1)
nabB=sum(vetor2)
NabA=(1/fA)*nabA
NabB=(1/fB)*nabB
z1=z2=z3=0
z1 = n_A*NabA*NB+nB*NabB*N_A
z2 = n_A*NB+nB*N_A+n_A*NabA+nB*NabB
z3 = n_A+nB
if(missing(z1) || is.null(z1) || is.na(z1) || !is.numeric(z1)){z1<-0}
if(missing(z2) || is.null(z2) || is.na(z2) || !is.numeric(z2)){z2<-0}
if(missing(z3) || is.null(z3) || is.na(z3) || !is.numeric(z3)){z3<-0}
roots<-polyroot(c(z1,-z2,z3))
```

```
NabPML = min(Re(roots))
return(NabPML)
}

##############################################################################
#Função que calcula o estimador para o total via PML
##############################################################################
# nA -> Tamanho da amostra obtida do cadastro A
# N_A -> Tamanho da população do cadastro A
# nB -> Tamanho da amostra obtida do cadastro B
# NB -> Tamanho da população do cadastro B
# sA -> Amostra obtida do cadastro A
# sB -> Amostra obtida do cadastro B
# dab -> Indicador numérido do domínio "ab"
##############################################################################

estPML = function(n_A,nB,N_A,NB,Nab,sA,sB,dab)
{
fA=n_A/N_A
fB=nB/NB
domabA=subset(sA,sA[,2]==dab)
doma=subset(sA,sA[,2]==(1-dab))
domabB=subset(sB,sB[,2]==dab)
domb=subset(sB,sB[,2]==(1-dab))
vetor1=rep(1,length(domabA[,1]))
vetor2=rep(1,length(domabB[,1]))
NabA=(1/fA)*sum(vetor1)
NabB=(1/fB)*sum(vetor2)
valor1=fA*NabA*mean(domabA[,4])+fB*NabB*mean(domabB[,4])
valor2=fA*NabA+fB*NabB
muabest=valor1/valor2
est=(N_A-Nab)*mean(doma[,4])+(NB-Nab)*mean(domb[,4])+Nab*muabest
return(est)
}

##############################################################################################
#Função que disponibiliza o vetor de resíduos e Função que calcula a covariância estimada entre
#estimadores GREG
##############################################################################################
# n -> Tamanho da amostra no cadastro de interesse
# N -> Tamanho da população no cadastro de interesse
# y -> Vetor com as informações da variável de interesse
# x -> Vetor ou matriz com informações das variáveis auxiliares
# tx -> Vetor com os totais para as variáveis auxiliares
# ck -> Por default, igual a 1. Vetor de pesos referente à variância do modelo
##############################################################################################

residuo=function (N, n, y, x, tx, b, b0 = FALSE)
{
y <- as.data.frame(y)
x <- as.matrix(x)
pik <- rep(n/N, n)
dk <- 1/pik
if (b0 == TRUE) {
```

```
x <- as.matrix(cbind(1, x))
}
for (k in 1:dim(y)[2]) {
e <- y[, k] - (x %*% as.matrix(b[, k]))
 }
return(e)
}

covarg=function(n,N,e1,e2)
{
a=n/(N^2)
b=n/N
c=(n-1)/(N-1)
V1=1-a/(b*c)
covar=V1*sum(e1%*%t(e2))
return(covar)
}

#############################################################################################################
#Função que calcula o p-ótimo para o estimador regressão com base na estratégia de Hartley
#############################################################################################################
#-> Precisa da função "covarg"
#-> Precisa da função "residuo"
#############################################################################################################
# N -> Total populacional referente a toda população
# N_A -> Total populacional referente a A
# NB -> Total populacional referente a B
# nA -> Tamanho da amostra referente ao cadastro A
# nB -> Tamanho da amostra referente ao cadastro B
# Nab -> Total populacional referente ao domínio ab
# nabA -> Tamanho da amastro em ab, coletada de A
# nabB -> Tamanho da amastro em ab, coletada de B
# eA -> Vetor de resíduos para o estimador regressão referente a A
# eB -> Vetor de resíduos para o estimador regressão referente a B
# eabA -> Vetor de resíduos para o estimador regressão referente a ab(cadastro A)
# eabB -> Vetor de resíduos para o estimador regressãoo referente a ab(cadastro B)
# vabA ->  Variância estimada do estimador regressão para ab (cadastro A)
# vabB ->  Variância estimada do estimador regressão para ab (cadastro B)
# covabA-> Covariância estimada entre o estimador para A e o estimador para ab(cadastro A)
# covabB-> Covariância estimada entre o estimador para B e o estimador para ab(cadastro B)
#############################################################################################################

preghart=function(N,nA,nB, N_A,NB,nabA,nabB,Nab,eA,eB,eabA,eabB)
{
wab=Nab/N
wA=N_A/N
wB=NB/N
q1=(Nab^2)*(1-(nabA/Nab))*var(eabA)/(nabA)
q2=(Nab^2)*(1-(nabB/Nab))*var(eabB)/(nabB)
q3=covarg(nA,N_A,eA,eabA)
q4=covarg(nB,NB,eB,eabB)
num=wA*q3-wB*q4
den=wab*(q1-q2)
p=num/den
```

```
return(p)
}


#################################################################################################
#Função que calcula o estimador greg para um domínio
#################################################################################################
#obs: Desempenha o mesmo papel que a função GREG.SI, apenas foi modificada para cálculo correto
#dentro dos domínios
#################################################################################################

GREG.SIESPd=function (N, n, nd, y, x, tx, b, b0 = FALSE)
{
y <- as.data.frame(y)
x <- as.matrix(x)
pik <- rep(n/N, nd)
dk <- 1/pik
if (b0 == TRUE) {
x <- as.matrix(cbind(1, x))
}
Total <- matrix(NA, nrow = 3, ncol = dim(y)[2])
rownames(Total) = c("Estimation", "Variance", "CVE")
colnames(Total) <- names(y)
for (k in 1:dim(y)[2]) {
xHT <- t(x) %*% dk
yHT <- sum(y[, k] * dk)
ty <- yHT + (tx - t(xHT)) %*% as.matrix(b[, k])
e <- y[, k] - (x %*% as.matrix(b[, k]))
Vty <- (N^2) * (1 - (n/N)) * var(e)/(n)
CVe <- 100 * sqrt(Vty)/ty
Total[, k] <- c(ty, Vty, CVe)
}
return(Total)
}


#################################################################################################
#Função que calcula o estimador regressão de acordo com a estratégia de Hartley
#################################################################################################
#Precisa da função "preghart"
#################################################################################################
# p -> Valor de "p" obtido pela função "preghart"
# amo1   -> Amostra de A
# amo2   -> Amostra de B
# pop1   -> População do cadastro A
# pop2   -> População do cadastro B
# N_A   -> Tamanho da população do cadastro A
# nA -> Tamanho da amostra obtida do cadastro A
# NB -> Tamanho da população do cadastro B
# nB -> Tamanho da amostra obtida do cadastro B
# Nab -> Tamanho da população do domínio "ab"
# dab -> Identificador do domínio ab
# name1 -> Identificador para a existência de intercepto no modelo para a pop. de A
# ("TRUE" ou "FALSE")
```

```
# name2 -> Identificador para a existência de intercepto no modelo para a pop. de B
# ("TRUE" ou "FALSE")
##############################################################################################

GREG.SIH=function(p,pop1,pop2,amo1,amo2,N_A, nA, NB,nB,Nab,dab,name1,name2)
{
amo1 <- as.data.frame(amo1)
amo2 <- as.data.frame(amo2)
popxA=data.frame(pop1[,1],pop1[,3])
txA=c(length(popxA[,1]),sum(pop1[,1]),sum(pop1[,3]))
popa=subset(pop1,pop1[,2]==(1-dab))
popxa=data.frame(popa[,1],popa[,3])
txa=c(length(popa[,1]),sum(popa[,1]),sum(popa[,3]))
popxB=data.frame(pop2[,1],pop2[,3])
txB=c(length(popxB[,1]),sum(pop2[,1]),sum(pop2[,3]))
popb=subset(pop2,pop2[,2]==(1-dab))
popxb=data.frame(popb[,1],popb[,3])
txb=c(length(popb[,1]),sum(popb[,1]),sum(popb[,3]))
xA=data.frame(amo1[,1],amo1[,3])
xB=data.frame(amo2[,1],amo2[,3])
Na=length(popa[,1])
Nb=length(popb[,1])
fA=nA/N_A
fB=nB/NB
pikA=rep(fA,nA)
pikB=rep(fB,nB)
popabA=subset(pop1,pop1[,2]==dab)
popabB=subset(pop2,pop2[,2]==dab)
popxabA=data.frame(popabA[,1],popabA[,3])
popxabB=data.frame(popabB[,1],popabB[,3])
domabA=subset(amo1,amo1[,2]==dab)
xabA=data.frame(domabA[,1],domabA[,3])
txabA=c(Nab,sum(popabA[,1]),sum(popabA[,3]))
doma=subset(amo1,amo1[,2]==(1-dab))
xa=data.frame(doma[,1],doma[,3])
domabB=subset(amo2,amo2[,2]==dab)
xabB=data.frame(domabB[,1],domabB[,3])
txabB=c(Nab,sum(popabB[,1]),sum(popabB[,3]))
domb=subset(amo2,amo2[,2]==(1-dab))
xb=data.frame(domb[,1],domb[,3])
pika=rep(fA,length(doma[,1]))
pikb=rep(fB,length(domb[,1]))
pikabA=rep(fA,length(domabA[,1]))
pikabB=rep(fB,length(domabB[,1]))
na=length(doma[,1])
nb=length(domb[,1])
ba=E.Beta(doma[,4],xa,pika,1,b0=name1)
bb=E.Beta(domb[,4],xb,pikb,1,b0=name2)
babA=E.Beta(domabA[,4],xabA,pikabA,1,b0=name1)
babB=E.Beta(domabB[,4],xabB,pikabB,1,b0=name2)
somapondA=GREG.SIESPd(N_A,nA,length(doma[,4]),doma[,4],xa,txa,ba,b0=name1)
somapondabA=GREG.SIESPd(N_A,nA,length(domabA[,1]),domabA[,4],xabA,txabA,babA,b0=name2)
somapondB=GREG.SIESPd(NB,nB,length(domb[,4]),domb[,4],xb,txb,bb,b0=name2)
somapondabB=GREG.SIESPd(NB,nB,length(domabB[,1]),domabB[,4],xabB,txabB,babB,b0=name2)
```

```
esthart=somapondA[1]+p*somapondabA[1]+somapondB[1] + (1-p)*somapondabB[1]
return(esthart[1])
}


##############################################################################################
#Função que calcula o estimador regressão de acordo com a estratégia de Bankier
##############################################################################################
# Desempenha mesmo papel que a função GREG.SI. Aqui, a função foi modificada para poder
# receber os pesos conforme determina a estratégia de Bankier. Precisa da função "GREG.SIESPab"
##############################################################################################
# amo1   -> Amostra de A
# amo2   -> Amostra de B
# pop1   -> População do cadastro A
# pop2   -> População do cadastro B
# N_A   -> Tamanho da população do cadastro A
# nA -> Tamanho da amostra obtida do cadastro A
# NB -> Tamanho da população do cadastro B
# nB -> Tamanho da amostra obtida do cadastro B
# Nab -> Tamanho da população do domínio "ab"
# dab -> Identificador do domínio ab
# name1 -> Identificador para a existência de intercepto no modelo para a pop. de A ("TRUE" ou "FALSE")
# name2 -> Identificador para a existência de intercepto no modelo para a pop. de B ("TRUE" ou "FALSE")
##############################################################################################
#-> Utiliza a primeira função a ser apresentada, a "GREG.SIESPab" (Função que calcula o
# estimador greg para o domínio ab sob a estratégia de bankier)
#-> Precisa da função "GREG.SIESPd" para os domínios
##############################################################################################

GREG.SIESPab=function (N, n, N1,n1, nab,y, x, tx, b, b0 = FALSE)
{
y <- as.data.frame(y)
x <- as.matrix(x)
val1=n1/N1
val2=n/N
pik <- rep(1/(val1+val2), nab)
dk <- 1/pik
if (b0 == TRUE) {
x <- as.matrix(cbind(1, x))
}
Total <- matrix(NA, nrow = 3, ncol = dim(y)[2])
rownames(Total) = c("Estimation", "Variance", "CVE")
colnames(Total) <- names(y)
for (k in 1:dim(y)[2]) {
xHT <- t(x) %*% dk
yHT <- sum(y[, k] * dk)
ty <- yHT + (tx - t(xHT)) %*% as.matrix(b[, k])
e <- y[, k] - (x %*% as.matrix(b[, k]))
Vty <- (N^2) * (1 - (n/N)) * var(e)/(n)
CVe <- 100 * sqrt(Vty)/ty
Total[, k] <- c(ty, Vty, CVe)
}
return(Total)
}
```

```
GREG.SIBA=function(pop1,pop2,amo1,amo2,N_A, nA, NB, nB,Nab,dab,name1,name2)
{
amo1 <- as.data.frame(amo1)
amo2 <- as.data.frame(amo2)
popxA=data.frame(pop1[,1],pop1[,3])
txA=c(length(popxA[,1]),sum(pop1[,1]),sum(pop1[,3]))
popa=subset(pop1,pop1[,2]==(1-dab))
popxa=data.frame(popa[,1],popa[,3])
txa=c(length(popa[,1]),sum(popa[,1]),sum(popa[,3]))
popxB=data.frame(pop2[,1],pop2[,3])
txB=c(length(popxB[,1]),sum(pop2[,1]),sum(pop2[,3]))
popb=subset(pop2,pop2[,2]==(1-dab))
popxb=data.frame(popb[,1],popb[,3])
txb=c(length(popb[,1]),sum(popb[,1]),sum(popb[,3]))
xA=data.frame(amo1[,1],amo1[,3])
xB=data.frame(amo2[,1],amo2[,3])
Na=length(popa[,1])
Nb=length(popb[,1])
fA=nA/N_A
fB=nB/NB
pikA=rep(fA,nA)
pikB=rep(fB,nB)
popabA=subset(pop1,pop1[,2]==dab)
popabB=subset(pop2,pop2[,2]==dab)
popxabA=data.frame(popabA[,1],popabA[,3])
popxabB=data.frame(popabB[,1],popabB[,3])
domabA=subset(amo1,amo1[,2]==dab)
xabA=data.frame(domabA[,1],domabA[,3])
txabA=c(Nab,sum(popabA[,1]),sum(popabA[,3]))
doma=subset(amo1,amo1[,2]==(1-dab))
xa=data.frame(doma[,1],doma[,3])
domabB=subset(amo2,amo2[,2]==dab)
xabB=data.frame(domabB[,1],domabB[,3])
txabB=c(Nab,sum(popabB[,1]),sum(popabB[,3]))
domb=subset(amo2,amo2[,2]==(1-dab))
xb=data.frame(domb[,1],domb[,3])
pika=rep(fA,length(doma[,1]))
pikb=rep(fB,length(domb[,1]))
pikabA=rep(1/(fA+fB),length(domabA[,1]))
pikabB=rep(1/(fA+fB),length(domabB[,1]))
na=length(doma[,1])
nb=length(domb[,1])
ba=E.Beta(doma[,4],xa,pika,1,b0=name1)
bb=E.Beta(domb[,4],xb,pikb,1,b0=name2)
babA=E.Beta(domabA[,4],xabA,pikabA,1,b0=name1)
babB=E.Beta(domabB[,4],xabB,pikabB,1,b0=name2)
somapondA=GREG.SIESPd(N_A,nA,length(doma[,4]),doma[,4],xa,txa,ba,b0=name1)
somapondabA=GREG.SIESPab(N_A,nA,NB,nB,length(domabA[,1]),domabA[,4],xabA,txabA,babA,b0=name2)
soma1=somapondA[1]+somapondabA[1]
somapondB=GREG.SIESPd(NB,nB,length(domb[,4]),domb[,4],xb,txb,bb,b0=name2)
somapondabB=GREG.SIESPab(N_A,nA,NB,nB,length(domabB[,1]),domabB[,4],xabB,txabB,babB,b0=name2)
soma2=somapondB[1]+somapondabB[1]
est=soma1+soma2
```

```
return(est[1])
}


#############################################################################################################
#Função que calcula o estimador regressão de acordo com a estratégia de Fuller
#############################################################################################################
#-> Precisa da função "GREG.SIESPd" para os domínios
#############################################################################################################
# amo1   -> Amostra de A
# amo2   -> Amostra de B
# pop1   -> População do cadastro A
# pop2   -> População do cadastro B
# N_A   -> Tamanho da população do cadastro A
# nA -> Tamanho da amostra obtida do cadastro A
# NB -> Tamanho da população do cadastro B
# nB -> Tamanho da amostra obtida do cadastro B
# Nabf -> Tamanho da população do domínio "ab"
# dab -> Identificador do domínio ab
# name1 -> Identificador para a existência de intercepto no modelo para a pop. de A
# ("TRUE" ou "FALSE")
# name2 -> Identificador para a existência de intercepto no modelo para a pop. de B
# ("TRUE" ou "FALSE")
#############################################################################################################

GREG.SIFUL=function(pop1,pop2,amo1,amo2,N_A, nA, NB,nB,Nabf,dab,name1,name2)
{
amo1 <- as.data.frame(amo1)
amo2 <- as.data.frame(amo2)
popxA=data.frame(pop1[,1],pop1[,3])
txA=c(length(popxA[,1]),sum(pop1[,1]),sum(pop1[,3]))
popa=subset(pop1,pop1[,2]==(1-dab))
popxa=data.frame(popa[,1],popa[,3])
txa=c(length(popa[,1]),sum(popa[,1]),sum(popa[,3]))
popxB=data.frame(pop2[,1],pop2[,3])
txB=c(length(popxB[,1]),sum(pop2[,1]),sum(pop2[,3]))
popb=subset(pop2,pop2[,2]==(1-dab))
popxb=data.frame(popb[,1],popb[,3])
txb=c(length(popb[,1]),sum(popb[,1]),sum(popb[,3]))
xA=data.frame(amo1[,1],amo1[,3])
xB=data.frame(amo2[,1],amo2[,3])
Na=length(popa[,1])
Nb=length(popb[,1])
fA=nA/N_A
fB=nB/NB
pikA=rep(fA,nA)
pikB=rep(fB,nB)
popabA=subset(pop1,pop1[,2]==dab)
popabB=subset(pop2,pop2[,2]==dab)
popxabA=data.frame(popabA[,1],popabA[,3])
popxabB=data.frame(popabB[,1],popabB[,3])
domabA=subset(amo1,amo1[,2]==dab)
xabA=data.frame(domabA[,1],domabA[,3])
txabA=c(length(popabA[,1]),sum(popabA[,1]),sum(popabA[,3]))
```

```
doma=subset(amo1,amo1[,2]==(1-dab))
xa=data.frame(doma[,1],doma[,3])
domabB=subset(amo2,amo2[,2]==dab)
xabB=data.frame(domabB[,1],domabB[,3])
txabB=c(length(popabB[,1]),sum(popabB[,1]),sum(popabB[,3]))
domb=subset(amo2,amo2[,2]==(1-dab))
xb=data.frame(domb[,1],domb[,3])
pika=rep(fA,length(doma[,1]))
pikb=rep(fB,length(domb[,1]))
pikabA=rep(fA,length(domabA[,1]))
pikabB=rep(fB,length(domabB[,1]))
na=length(doma[,1])
nb=length(domb[,1])
nabA=length(domabA[,1])
nabB=length(domabB[,1])
ba=E.Beta(doma[,4],xa,pika,1,b0=name1)
bb=E.Beta(domb[,4],xb,pikb,1,b0=name2)
babA=E.Beta(domabA[,4],xabA,pikabA,1,b0=name1)
babB=E.Beta(domabB[,4],xabB,pikabB,1,b0=name2)
somapondA=GREG.SIESPd(N_A,nA,length(doma[,4]),doma[,4],xa,txa,ba,b0=name1)
somapondabA=GREG.SIESPd(N_A,nA,length(domabA[,1]),domabA[,4],xabA,txabA,babA,b0=name2)
somapondB=GREG.SIESPd(NB,nB,length(domb[,4]),domb[,4],xb,txb,bb,b0=name2)
somapondabB=GREG.SIESPd(NB,nB,length(domabB[,1]),domabB[,4],xabB,txabB,babB,b0=name2)
w=(nabA*(1-fB)/(nabA*(1-fB)+nabB*(1-fA)))
estful=somapondA[1]+somapondB[1]+w*somapondabA[1]+(1-w)*somapondabB[1]
return(estful)
}
```

```
#############################################################################################
#Função que calcula o estimador regressão de acordo com a estratégia PML
#############################################################################################
#-> Precisa da função "GREG.SIESPd" para os domínios
#-> Precisa da função "NabPML" para estimar Nab
#############################################################################################
# amo1   -> Amostra de A
# amo2   -> Amostra de B
# pop1   -> População do cadastro A
# pop2   -> População do cadastro B
# N_A   -> Tamanho da população do cadastro A
# nA -> Tamanho da amostra obtida do cadastro A
# NB -> Tamanho da população do cadastro B
# nB -> Tamanho da amostra obtida do cadastro B
# Nabpml -> Estimador PML para Nab
# dab -> Identificador do domínio ab
# name1 -> Identificador para a existência de intercepto no modelo para a pop. de A
# ("TRUE" ou "FALSE")
# name2 -> Identificador para a existência de intercepto no modelo para a pop. de B
# ("TRUE" ou "FALSE")
```

```
##################################################################################################


GREG.SIPML=function(pop1,pop2,amo1,amo2,N_A, nA, NB, nB,Nabpml,dab,name1,name2)
{
amo1 <- as.data.frame(amo1)
amo2 <- as.data.frame(amo2)
popxA=data.frame(pop1[,1],pop1[,3])
txA=c(length(popxA[,1]),sum(pop1[,1]),sum(pop1[,3]))
popa=subset(pop1,pop1[,2]==(1-dab))
popxa=data.frame(popa[,1],popa[,3])
txa=c((N_A-Nabpml),sum(popa[,1]),sum(popa[,3]))
popxB=data.frame(pop2[,1],pop2[,3])
txB=c(length(popxB[,1]),sum(pop2[,1]),sum(pop2[,3]))
popb=subset(pop2,pop2[,2]==(1-dab))
popxb=data.frame(popb[,1],popb[,3])
txb=c((NB-Nabpml),sum(popb[,1]),sum(popb[,3]))
xA=data.frame(amo1[,1],amo1[,3])
xB=data.frame(amo2[,1],amo2[,3])
Na=length(popa[,1])
Nb=length(popb[,1])
fA=nA/N_A
fB=nB/NB
pikA=rep(fA,nA)
pikAu=pikA[1]
pikB=rep(fB,nB)
pikBu=pikB[1]
popabA=subset(pop1,pop1[,2]==dab)
popabB=subset(pop2,pop2[,2]==dab)
popxabA=data.frame(popabA[,1],popabA[,3])
popxabB=data.frame(popabB[,1],popabB[,3])
domabA=subset(amo1,amo1[,2]==dab)
xabA=data.frame(domabA[,1],domabA[,3])
doma=subset(amo1,amo1[,2]==(1-dab))
xa=data.frame(doma[,1],doma[,3])
domabB=subset(amo2,amo2[,2]==dab)
xabB=data.frame(domabB[,1],domabB[,3])
domb=subset(amo2,amo2[,2]==(1-dab))
xb=data.frame(domb[,1],domb[,3])
pika=rep(fA,length(doma[,1]))
pikb=rep(fB,length(domb[,1]))
pikabA=rep(1/pikAu,length(domabA[,1]))
pikabB=rep(1/pikBu,length(domabB[,1]))
NabestabA=sum(pikabA)
NabestabB=sum(pikabB)
txabA=c(NabestabA,sum(popabA[,1]),sum(popabA[,3]))
txabB=c(NabestabB,sum(popabB[,1]),sum(popabB[,3]))
na=length(doma[,1])
nb=length(domb[,1])
ba=E.Beta(doma[,4],xa,pika,1,b0=name1)
bb=E.Beta(domb[,4],xb,pikb,1,b0=name2)
babA=E.Beta(domabA[,4],xabA,pikabA,1,b0=name1)
babB=E.Beta(domabB[,4],xabB,pikabB,1,b0=name2)
somapondA=GREG.SIESPd(N_A,nA,length(doma[,4]),doma[,4],xa,txa,ba,b0=name1)
```

```
somapondabA=GREG.SIESPd(N_A,nA,length(domabA[,1]),domabA[,4],xabA,txabA,babA,b0=name2)
somapondB=GREG.SIESPd(NB,nB,length(domb[,4]),domb[,4],xb,txb,bb,b0=name2)
somapondabB=GREG.SIESPd(NB,nB,length(domabB[,1]),domabB[,4],xabB,txabB,babB,b0=name2)
tgregabpml=(pikAu*somapondabA[1]+pikBu*somapondabB[1])/(pikAu*NabestabA+pikBu*NabestabB)
pmlgreg=somapondA[1]+tgregabpml+somapondB[1]
return(pmlgreg)
}


#######################INÍCIO DO PROGRAMA###########################
library(TeachingSampling)
RNGkind("Marsaglia-Multicarry")
seed=.Random.seed
set.seed(seed)
####################################################################
##ARMAZENANDO VALORES POPULACIONAIS E
##TAMANHOS DE AMOSTRA
####################################################################
N=c(2000,1750,1500,1250,1000,750)
Nab=c(500,1000,1500,2000,2500,3000)
nA=c(125,250,500,750,1000,1250)
nB=c(125,250,500,750,1000,1250)
####################################################################
####################################################################


#######INÍCIO DO LOOP PARA OS TAMANHOS DE "N"
for(i in 1:length(N))
{
Total=matrix(NA,nrow=2,ncol=3)
colnames(Total)=c("NC","Nab","Soma")
Total[2,]=c(N[i],Nab[i],N[i]+Nab[i])
print(Total[2,])
print("@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@")

#####Carregando o banco de dados
xA11=4+rnorm(N[i])
xA21=8+rnorm(N[i])
xA31=2+rnorm(Nab[i])
xA32=1+rnorm(Nab[i])
xA1=append(xA11,xA31)
xA2=append(xA21,xA32)
yAae=2+xA1[1:N[i]]+xA2[1:N[i]]+rnorm(N[i])
yabA=2+xA31+xA32+rnorm(Nab[i])
yA=append(yAae,yabA)
domA=c(append(matrix(1,N[i]),matrix(0,Nab[i])))
N_A=length(domA)
dadosA=data.frame(XA1=xA1,DA=domA,XA2=xA2,YA=yA)
xB11=4+rnorm(N[i])
xB21=8+rnorm(N[i])
xB31=xA31
xB32=xA32
xB1=append(xB11,xB31)
xB2=append(xB21,xB32)
```

```
yBe=2+xB1[1:N[i]]+xB2[1:N[i]]+rnorm(N[i])
yB=append(yBe,yabA)
domB=c(append(matrix(1,N[i]),matrix(0,Nab[i])))
NB=length(domB)
dadosB=data.frame(XB1=xB1,DB=domB,XB2=xB2,YB=yB)
a=sum(yA)
b=sum(yB)
c=sum(yabA)
soma=a + b - c
total=length(yA)+length(yB)-length(yabA)
par=soma/total
par=round(par,2)


#######Iniciando o loop para os tamanhos de amostra
for(j in 1:length(nA))
{
print("-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#")
Total=matrix(NA,nrow=2,ncol=2)
colnames(Total)=c("nA","nB")
Total[2,]=c(nA[j],nB[j])
print(Total[2,])
print("----------------------------------------------------------------")

########Início do loop de Monte Carlo
nrep=100
hartley=matrix(0,nrep)
greghart=matrix(0,nrep)
Nab_fuller=c()
fuller=matrix(0,nrep)
gregfuller=matrix(0,nrep)
bankier=matrix(0,nrep)
gregbankier=matrix(0,nrep)
Nab_PML=matrix(0,nrep)
estpml=matrix(0,nrep)
gregestpml=matrix(0,nrep)
for(r in 1:nrep)
{
selA=S.SI(N_A,nA[j])
amoA=dadosA[selA,]
domabA=subset(amoA,amoA[,2]==0)
nabA=length(domabA[,4])
###nA=length(amoA[,1])
selB=S.SI(NB,nB[j])
amoB=dadosB[selB,]
###nB=length(amoB[,1])
domabB=subset(amoB,amoB[,2]==0)
nabB=length(domabB[,4])
phart=pH(nA[j],N_A,nB[j],NB,amoA,amoB,Nab[i],0)
hartley[r]=esthart(phart,nA[j],nB[j],N_A,NB,amoA,amoB,Nab[i],0)/total
fuller[r]=estful(nA[j],nB[j],N_A,NB,amoA,amoB,Nab_fuller[i],0)/total
bankier[r]=estbank(nA[j],nB[j],N_A,NB,amoA,amoB,0)/total
Nab_PML[r]=NabPML(nA[j],nB[j],N_A,NB,amoA,amoB,0)
estpml[r]=estPML(nA[j],nB[j],N_A,NB,Nab[i],amoA,amoB,0)/total
```

```
dataA=data.frame(amoA[,1],amoA[,3])
pikA=rep(nA[j]/N_A,nA[j])
bA=E.Beta(amoA[,4],dataA,pikA,1,b0=TRUE)
resA=residuo(N_A, nA[j], amoA[,4], dataA, totalA, bA, b0 = TRUE)
dataB=data.frame(amoB[,1],amoB[,3])
totalB=c(sum(amoB[,1]),sum(amoB[,3]))
pikB=rep(nB[j]/NB,nB[j])
bB=E.Beta(amoB[,4],dataB,pikB,1,b0=TRUE)
resB=residuo(NB, nB[j], amoB[,4], dataB, totalB, bB, b0 = TRUE)
covg=covarg(nA[j]+nB[j],N_A+NB,resA,resB)
domabA=subset(amoA,amoA[,2]==0)
domabB=subset(amoB,amoB[,2]==0)
nabA=length(domabA[,1])
nabB=length(domabB[,1])
dataabA=data.frame(domabA[,1],domabA[,3])
dataabB=data.frame(domabB[,1],domabB[,3])
babA=E.Beta(domabA[,4],dataabA,rep(nA[j]/N_A,nabA),1,b0=TRUE)
babB=E.Beta(domabB[,4],dataabB,rep(nB[j]/NB,nabB),1,b0=TRUE)
dadosabA=subset(dadosA,dadosA[,2]==0)
#Nab=length(dadosabA[,1])
dadosabB=subset(dadosB,dadosB[,2]==0)
totalabA=c(sum(dadosabA[,1]),sum(dadosabA[,3]))
totalabB=c(sum(dadosabB[,1]),sum(dadosabB[,3]))
resabA=residuo(Nab[i],nabA,domabA[,4],dataabA,totalabA,babA,b0=TRUE)
resabB=residuo(Nab[i],nabB,domabB[,4],dataabB,totalabB,babB,b0=TRUE)
pgh=preghart(N_A+NB-Nab[i],nA[j],nB[j], N_A,NB,nabA,nabB,Nab[i],resA,resB,resabA,resabB)
greghart[r]=GREG.SIH(pgh,dadosA,dadosB,amoA,amoB,N_A, nA[j], NB,nB[j],Nab[i],0
,"TRUE","TRUE")/total
gregbankier[r]=GREG.SIBA(dadosA,dadosB,amoA,amoB,N_A, nA[j], NB,nB[j],Nab[i],0
,"TRUE","TRUE")/total
gregfuller[r]=GREG.SIFUL(dadosA,dadosB,amoA,amoB,N_A, nA[j], NB,nB[j],Nab[i],0
,"TRUE","TRUE")/total
gregestpml[r]=GREG.SIPML(dadosA,dadosB,amoA,amoB,N_A, nA[j], NB,nB[j],Nab_PML[r],
0,"TRUE","TRUE")/total


}
print(c("ESTIMADORES NORMAIS","valor do parâmetro:",par))
Total=matrix(NA,nrow=4,ncol=4)
colnames(Total)=c("Média","Desvio Padrão", "V.R.", "EQMR")
Total[1,]=c(mean(hartley),sd(hartley),(mean(hartley)-par)/par,var(hartley)
+(mean(hartley)-par)^2)
Total[2,]=c(mean(fuller),sd(fuller),(mean(fuller)-par)/par,var(fuller)
+(mean(fuller)-par)^2)
Total[3,]=c(mean(bankier),sd(bankier),(mean(bankier)-par)/par,var(bankier)
+(mean(bankier)-par)^2)
Total[4,]=c(mean(estpml),sd(estpml),(mean(estpml)-par)/par,var(estpml)
+(mean(estpml)-par)^2)
print(Nab_fuller)

print(Total)
print(c("ESTIMADORES GREG","valor do parâmetro:",par))
Total=matrix(NA,nrow=4,ncol=4)
colnames(Total)=c("Média","Desvio Padrão", "V.R.", "EQMR")
```

```
Total[1,]=c(mean(greghart),sd(greghart),(mean(greghart)-par)/par,var(greghart)
+(mean(greghart)-par)^2)
Total[2,]=c(mean(gregfuller),sd(gregfuller),(mean(gregfuller)-par)/par,var(gregfuller)
+(mean(gregfuller)-par)^2)
Total[3,]=c(mean(gregbankier),sd(gregbankier),(mean(gregbankier)-par)/par,var(gregbankier)
+(mean(gregbankier)-par)^2)
Total[4,]=c(mean(gregestpml),sd(gregestpml),(mean(gregestpml)-par)/par,var(gregestpml)
+(mean(gregestpml)-par)^2)
M=par(mfrow=c(2,4))
M[1]=hist(hartley,main=c("Hartley",N[i],Nab[i],nA[j],nB[j]))
M[2]=hist(fuller,main=c("Fuller"))
M[3]=hist(bankier,main=c("Bankier") )
M[4]=hist(estpml,main=c("PML"))
M[5]=hist(greghart,main=c("Hartley-GREG"))
M[6]=hist(gregfuller,main=c("Fuller-GREG"))
M[7]=hist(gregbankier,main=c("Bankier-GREG"))
M[8]=hist(gregestpml,main=c("PML-GREG"))
x11()
print(Total)
print(c("fim","para","indsamp",j))
j=j+1
}


final=matrix(NA,nrow=1,ncol=3)
final[1,]=c("FIM","PASSO",i)
print(final)
print("----------------------------------------------------------")
i=i+1
print("@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@")
}
############FIM DO LOOP PARA OS TAMANHOS DE N


##############################FIM DO PROGRAMA###########################
```

Apêndice D

Cenários de sequências de populações finitas sob a abordagem de cadastro duplo

É possível apresentar exemplos das estruturas das populações finitas que podem ser consideradas sob a abordagem de cadastro duplo. Estes exemplos estão apresentados nas figuras a seguir para $\nu = 2$.

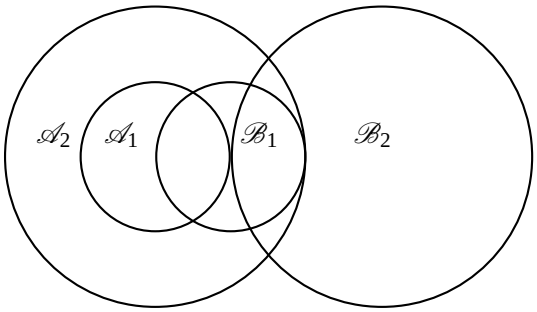


Figura 1: Cenário (b) $\mathscr{A}_1 \subset \mathscr{A}_2$ e $\mathscr{B}_1 \subset \mathscr{A}_2$, e $\mathscr{A}_1 \cup \mathscr{B}_1 \subset \mathscr{A}_2$

O cenário (b) ilustra a união entre os primeiros elementos de cada sequência encaixados no segundo elemento da sequência referente ao cadastro A.

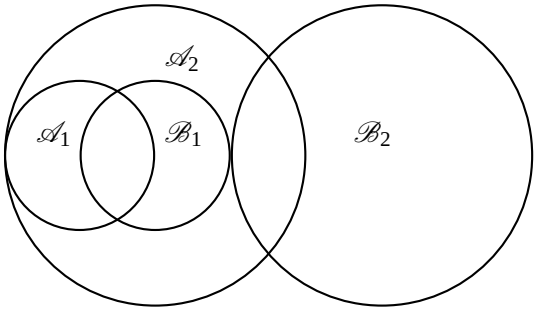


Figura 2: Cenário (c) $\mathscr{A}_1 \subset \mathscr{A}_2$ e $\mathscr{A}_1 \cup \mathscr{B}_1 \subset \mathscr{A}_2$, com $(\mathscr{A}_1 \cup \mathscr{B}_1) \cap \mathscr{B}_2 = \emptyset$

O cenário (c) ilustra situação equivalente ao cenário (b), mas com o detalhe de que a união dos cadastros não contemplar elementos da interseção entre os elementos das sequências.

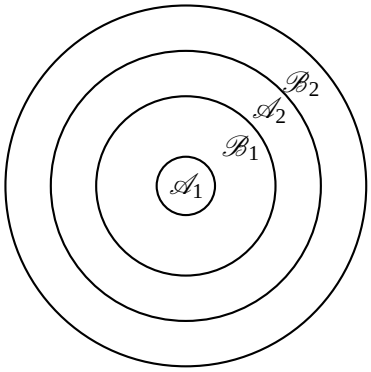


Figura 3: Cenário (d) $\mathscr{A}_1 \subset \mathscr{A}_2 \subset \mathscr{B}_1 \subset \mathscr{B}_2$

No cenário (d) é possível observar a situação em que o cadastro A está contido no cadastro B, e vice-versa.

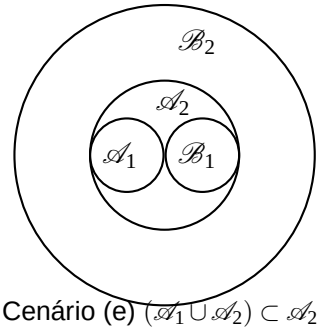


Figura 4: Cenário (e) $(\mathscr{A}_1 \cup \mathscr{A}_2) \subset \mathscr{A}_2$ e $\mathscr{A}_2 \subset \mathscr{B}_2$

O cenário (e) apresenta como característica a não existência de interseção entre os primeiros elementos das sequências referentes a A e B. Porém, a união destes elementos é embutida nos demais elementos das sequências.

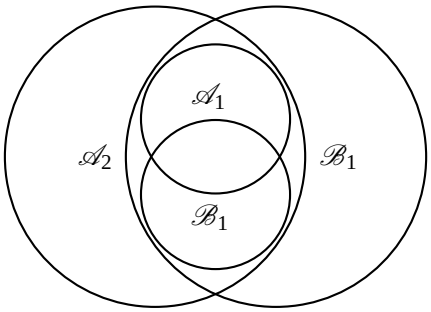


Figura 5: Cenário (f) $(\mathscr{A}_1 \cup \mathscr{B}_1) \subset \mathscr{A}_2 \cap \mathscr{B}_2$

O cenários (f) e (g) apresentam situação similar ao apresentado anteriormente, apenas com o fato de que a interseção entre os primeiros elementos das sequências existe.

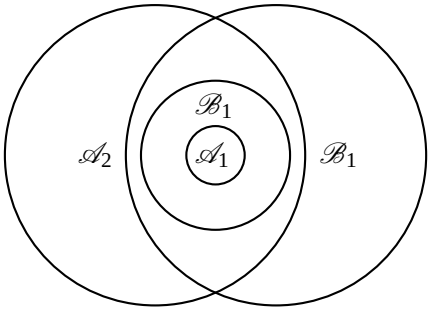


Figura 6: Cenário (g) $\mathscr{A}_1 \subset \mathscr{B}_1$ e $(\mathscr{A}_1 \cup \mathscr{B}_1) \subset \mathscr{A}_2 \cap \mathscr{B}_2$

REFERÊNCIAS

ALPIZAR-JARA, R.; POLLOCK, K. H.; HAINES, D. E. **Mark-recapture estimator for dual frame population size of prominent nesting structures: the effect of uncertain detection probability**. *Environmental and Ecological Statistics*, 2005.

ARMSTRONG, B. **Test of multiple frame sampling techniques for agricultural surveys: New brunswick, 1978**. *Proceedings of the Section on Survey Research Methods of the American Statistical Association, pp. 295-300*, 1979.

BANKIER; LEPKOWSKI, J. M.; GROVES, R. M. **Estimators based on several stratified samples with applications to multiple frame surveys**. *Journal of the American Statistical Association*, 1986.

BOSECKER, R. R.; FORD, B. L. **Multiple frame estimation with stratified overlap domain**. *Proceedings of the Social Statistics Section, American Statistical Association*, 1976.

BRACKSTONE, G. J.; RAO, J. N. K. **An investigation of raking ratio estimation**. Sankhya: The Indian Journal of Statistics, 16, Supp., 47-55, 1979.

BREIDT, F. J.; OPSOMER, J. D. **Endougenous post-stratification in surveys: Classifying with a sample-fitted model**. *Annals of Statistics*, 2008.

CARFAGNA, E. **List frames, area frames and administrative data, are complementary or in competition?** *invited paper to the Meeting MEXSAI Conference organized by Eurostat, FAO, OCSE, UN/ECE, NASS/USDA, JRC, ISI, Istat, SAGARPA*, 2004.

CASADY, R. J.; LEPKOWSKI, J. M. **Optimal allocation for stratified telephone survey designs**. *Proceedings of the Section on Survey Research Methods of the American Statistical Association, pp. 111-116*, 1991.

CASADY, R. J.; LEPKOWSKI, J. M. **Stratified telephone survey designs**. *Survey Methodology, Vol. 19, pp. 103-113*, 1993.

CHOUDHRY, G. H.; PARK, I.; LI, W. T. **Dual frame sample design for a national sample of veterans**. *Joint Statistical Meetings, Section of Survey Research Methods*, 2002.

COELHO, H. F. C. ***A Abordagem de Cadastro Duplo: Estimação Assistida por Modelos com Aplicações em Pesquisas Agropecuárias***. Dissertação (Mestrado) — Universidade Federal de Pernambuco, Departamento de Estatística, 2007.

COELHO, H. F. C.; FERRAZ, C. **Ratio type estimators for stratified dual frame surveys.** *Proceedings of the 56th Session of the ISI*, 2007.

FULLER, W. A. **Introduction to statistical time series**. Wiley-Interscience, 1996.

FULLER, W. A. **Sampling statistics**. *Wiley series in Survey Methodology*, 2009.

FULLER, W. A.; BURMEISTER, L. F. **Estimators for samples selected from two overlapping frames**. *Proceedings of the Social Statistics Section, American Statistical Association, 245-249*, 1972.

HAINES, D. E.; POLLOCK, K. H. **Estimating the number of active and successful bald eagle nests: an application of the dual frame method**. *Environmental and Ecological Statistics*, 1998.

HARTLEY, H. O. **Multiple frame surveys**. *Proceedings of the Social Statistics Association-ASA*, 1962.

ISAKI, C. T.; FULLER, W. A. **Survey design under the regression superpopulation model**. *Journal of American Statistical Association*, 1982.

LOHR, S. L. **Inference from dual frame surveys**. *Journal of American Statistical Association*, 2007.

LóPEZ, O. F. M. ***Comparación de la eficiencia de estimadores para marcos duales***. Dissertação (Mestrado) — Universidad Nacional de Colombia, Faculdad de Ciencias-Departamento de Estadística, 2010.

LU, Y. ***Longitudinal Estimation in Dual Frame Surveys***. Tese (Doutorado) — Arizona State University, 2007.

LUND, R. E. **Estimators in multiple frame surveys**. *Proceedings ot the Social Statistics Section, American Statistical Association*, 1968.

PARK, M. ***Regression Estimation of the Mean in Survey Sampling***. Tese (Doutorado) — Iowa State University, 2002.

POTTER, F. J. *et al.* **List-assisted rdd surveys**. *Proceedings of the Section on Survey Research Methods of the American Statistical Association, pp. 117-122*, 1991.

POVEDA, L. M. R. ***Estimação em populações assistidas por modelos para variáveis dicotômicas***. Dissertação (Mestrado) — Universidade Federal de Pernambuco, Departamento de Estatística, 2007.

ROBINSON, P. M.; SäRNDAL, C. E. **Asymptotic properties of the generalized regression estimator in probability sampling**. *Sankhya: The Indian Journal of Statistics, 45, Series B, 240-248*, 1983.

SEN, P. K.; SINGER, J. M.; LIMA, A. C. Pedroso de. **From finite sample to asymptotic methods in statistics**. Cambridge Series in Statistical and Probabilistic Mathematics (No. 29), 2010.

SHIN, H. C.; MOLINARI, N. A.; WOLTER, K. **A dual frame design for the national immunization survey**. *Section on Survey Research Methods - AAPOR*, 2008.

SINGH, A. C.; WU, S. **An extension of generalized estimator to dual frame surveys**. *Join*

Statistical Meetings - Section on Survey Research Methods, 2003.

SKINNER, C. J. **On the efficiency of raking ratio estimation for multiple frame surveys**. *Journal of the American Statistical Association*, 1991.

SKINNER, C. J.; RAO, J. N. K. **Estimation in dual frame surveys with complex designs**. *Journal of the American Statistical Association*, 1996.

SäRNDAL, C. E. **Model assisted survey sampling**. *New York: Springer*, 1992.

SäRNDAL, C. E.; SWENSSON, B.; WRETMAN, J. **Model assisted survey sampling**. *Springer Series in Statistics*, 2003.

WU, C.; RAO, J. N. K. **Pseudo-empirical likelihood ratio confidence intervals for complex surveys**. *The Canadian Journal of Statistics, 34, 359-375*, 2006.