UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA
DOUTORADO EM ESTATÍSTICA

JOÃO EUDES MIQUÉIAS MACIEL TORRES

ESTIMAÇÃO ASSISTIDA POR MODELOS LINEARES GENERALIZADOS EM PLANOS AMOSTRAIS DE CADASTROS MÚLTIPLOS

RECIFE
2024

João Eudes Miquéias Maciel Torres

ESTIMAÇÃO ASSISTIDA POR MODELOS LINEARES GENERALIZADOS EM PLANOS AMOSTRAIS DE CADASTROS MÚLTIPLOS

Tese apresentada ao Curso de Doutorado em Estatística do Programa de Pós-Graduação em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal de Pernambuco, como requisito parcial à obtenção do título de doutor em Estatística. Área de Concentração: Ciência Exatas e da Terra

Orientador: Cristiano Ferraz
Coorientador: Raydonal Ospina Martinez

Recife

2024

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Torres, João Eudes Miqueias Maciel.
Estimação assistida por modelos lineares generalizados em planos amostrais de cadastros múltiplos / Joao Eudes Miqueias Maciel Torres. - Recife, 2024.
74f.: il.

Tese (Doutorado) - Universidade Federal de Pernambuco, Centro de Ciências Exatas e da Natureza, Programa de Pós-Graduação em Estatística, 2024.
Orientação: Cristiano Ferraz.
Coorientação: Raydonal Ospina Martinez.

1. Múltiplos cadastros; 2. Estimador de multiplicidade; 3. Estimador ótimo; 4. Modelo linear generalizado; 5. Amostragem de área; 6. Pesquisa agropecuária. I. Ferraz, Cristiano. II. Martinez, Raydonal Ospina. III. Título.

UFPE-Biblioteca Central

Folha de aprovação: Inserir a folha de aprovação enviada pela Secretaria do curso de Pós-Graduação. A folha deve conter a **data de aprovação**, estar **sem assinaturas** e em formato **PDF**.

AGRADECIMENTOS

Em primeiro lugar, a Deus. Reconheço o Seu amor infinito em tudo o que Ele tem me proporcionado. Por muitas vezes, Ele me deu alívio, e em Seus braços pude descansar. A Ele, toda honra, glória, majestade e poder, pelos séculos dos séculos. Amém

À minha esposa, Assíria Torres, por seu amor, incentivo e companheirismo. Nos momentos mais difíceis, você esteve ao meu lado em oração.

À minha família: tive em vocês todo o apoio e suporte de que precisei, infalivelmente. Mãe, seu amor e cuidado me deram forças para continuar. Pai, você é meu maior exemplo.

A todos os amigos com quem tive o prazer de compartilhar grande parte dos meus dias na pós-graduação e que, de agora em diante, farão parte da minha vida.

Aos meus orientadores, Cristiano Ferraz e Raydonal Ospina, pela ajuda, apoio e compreensão, por me ensinarem e me pressionarem sempre que necessário, e por todos os ensinamentos, não apenas no âmbito acadêmico.

À banca examinadora, pelas valiosas sugestões, que, sem dúvida, enriqueceram a qualidade deste trabalho.

Aos professores do Programa de Pós-Graduação em Estatística (PPGE) da UFPE, pelo ensino de excelência que proporcionaram. Com muitos, venho aprendendo desde a graduação.

À secretaria do PPGE pelo excelente atendimento, pela solução de dúvidas e pelos conselhos fornecidos, que contribuíram para a conclusão do Doutorado.

Aos amigos na Diretoria de Avaliação Institucional (DAI) da UFPE, por toda ajuda, apoio e suporte que me deram nas etapas finais deste trabalho. Em especial, à Priscila Gonçalves, por todo incentivo e suporte.

RESUMO

Esta tese considera o problema de estimação de parâmetros de populações finitas usando um estimador assistido por modelo linear generalizado (GEREG), quando a amostra é selecionada a partir de múltiplos cadastros sobrepostos. O GEREG considera a disponibilidade de variáveis auxiliares relacionadas à variável de interesse através de um modelo linear generalizado adequado. Nas situações em que a distribuição empírica da variável de interesse pode ser considerada como membro da família exponencial, espera-se que o GEREG apresente um melhor desempenho estatístico do que o estimador de regressão geral usual (GREG). Esta tese estende o GEREG para um plano amostral de múltiplos cadastros, utilizando a abordagem de estimação por multiplicidade. Sua forma geral, bem como propriedades estatísticas são introduzidas. É apresentado um estudo de Monte Carlo, comparando o GEREG com concorrentes, para estimar totais e proporções populacionais, considerando um plano amostral de múltiplos cadastros. A aplicação dos estimadores GEREG em pesquisas agropecuárias que utilizam um cadastro de área de segmentos quadrados conjuntamente com cadastros de lista também é avaliada por meio de simulação. Os resultados obtidos corroboram que o estimador GEREG tende a demonstrar melhor desempenho em relação ao estimador GREG quando modelos lineares generalizados são adequados para descrever a distribuição da variável de interesse.

Palavras-chave: Múltiplos cadastros; Estimador de multiplicidade; Estimador ótimo; Modelo linear generalizado; Amostragem de área; Pesquisa agropecuária.

ABSTRACT

This dissertation considers the problem of estimating finite population parameters using a generalized linear model assisted estimator (GEREG), when the sample is selected from multiple overlapping frames. The GEREG considers the availability of auxiliary variables related to the variable of interest through a suitable generalized linear model. In situations where the empirical distribution of the variable of interest can be regarded as a member of the exponential family, the GEREG is expected to show a better statistical performance than the usual general regression estimator (GREG). This dissertation extends the GEREG for a multiple frame sampling design, using the multiplicity estimator approach. Its general form, as well as statistical properties are introduced. A Monte Carlo study, comparing the GEREG with competitors, for estimating population totals and proportions is presented, considering a multiple frame. The application of GEREG estimators in agricultural survey that uses an area frame of square segment together with list frames is also evaluated through simulation. The results obtained corroborate that the GEREG estimator tends to demonstrate better performance in relation to the GREG estimator when generalized linear models are suitable to describe the distribution of the variable of interest.

Keywords: Multiple frame; Multiplicity estimator; Optimum estimator; Generalized linear model; Area sampling; Agricultural survey.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

GEREG	Estimador de Regressão Linear Generalizado
GREG	Estimador de Regressão Linear Geral
HT	Horvitz-Thompson
LGREG	Estimador de Regressão Linear Generalizado Logístico
MLGREG	Estimador de Regressão Linear Generalizado Logístico Multinomial
MS	Multiplicidade Simples
MT	Mato Grosso
NASS	National Agricultural Statistics Service
R.EQM	Raiz do Erro Quadrático Médio
USDA	United States Department of Agriculture

SUMÁRIO

1 INTRODUÇÃO

Um objetivo comum em pesquisas amostrais é realizar inferência estatística válida em uma população finita de maneira mais rápida e econômica em comparação aos estudos censitários. Frequentemente, informações já conhecidas sobre a população, obtidas de outros estudos amostrais ou censitários e dados administrativos, podem ser incorporadas ao estudo amostral para melhorar sua eficiência. Tais informações, conhecidas na literatura como *informações auxiliares*, podem ser úteis tanto para aprimorar o processo de seleção da amostra quanto para aumentar a qualidade das estimativas.

Na fase de estimação, as informações auxiliares podem ser empregadas na concepção de modelos preditivos a partir de dados amostrais. O termo *estimação assistida por modelos* refere-se a abordagens de estimação baseadas na aleatorização que utilizam modelos preditivos para modificar a estrutura do estimador, preservando propriedades desejáveis como centralidade e consistência, e melhorando sua precisão. Breidt & Opsomer (2017) oferecem uma revisão abrangente das abordagens de estimação assistida por modelos, que utilizam técnicas de predição paramétricas e não paramétricas. Além disso, eles sistematizaram uma maneira de derivar propriedades assintóticas desejáveis dos estimadores a partir das propriedades do estimador de diferença.

No contexto da estimação assistida por modelos paramétricos, Rondon, Vanegas & Ferraz (2012) propuseram o Estimador Assistido por Modelo Linear Generalizado (GEREG), que considera a disponibilidade de variáveis auxiliares relacionadas à variável de interesse por meio de um modelo linear generalizado adequado. Quando a distribuição empírica da variável de interesse pode ser considerada membro da família exponencial e as variáveis auxiliares disponíveis estão linearmente relacionadas à variável de interesse por meio de uma função de ligação, espera-se que o GEREG apresente um desempenho estatístico superior ao do estimador usual de Horvitz-Thompson ou do Estimador de Regressão Geral (GREG).

Explorando alguns exemplos, nos casos em que a variável de interesse tem distribuição normal e a função de ligação é canônica, o GEREG corresponde ao estimador GREG usual. Nos casos em que a variável de interesse é discreta, proveniente de um processo de contagem, a utilização de um GEREG com uma função de ligação logarítmica levaria a uma estimativa do total populacional, bem como de taxas e proporções, assistida por um modelo de Poisson. Tal estimativa aproveitaria melhor a relação entre a variável de interesse e as variáveis auxiliares

disponíveis, em comparação com uma estimativa baseada em um modelo linear normal

O caso em que a variável resposta é binária é outro exemplo. Com a distribuição de Bernoulli, o uso da função *logit* como ligação leva ao Estimador Assistido por Modelo Linear Generalizado Logístico (LGREG), originalmente proposto por Lehtonen & Veijanen (1998). Estimar proporções populacionais usando o LGREG tem a vantagem de garantir que a estimativa esteja no suporte coerente, assumindo valores no intervalo (0,1), além de aumentar a precisão em comparação com estimadores concorrentes, como o de Horvitz-Thompson e o GREG. Lehtonen & Veijanen (1998) também generalizaram seu estimador para acomodar o caso logístico multinomial, introduzindo o chamado Estimador Assistido por Modelo Linear Generalizado Logístico Multinomial (MLGREG).

O tema geral de estimação assistida por modelos em planos amostrais de múltiplos cadastros foi abordado por alguns autores, geralmente enfatizando o caso especifico de cadastro duplo. Singh & Wu (2003) utilizaram variáveis auxiliares para propor um estimador inspirado na ideia de Singh (1996) de estimadores do tipo regressão modificada, em conjunto com uma abordagem de calibração, para cadastros duplos.

Coêlho (2007, 2011) aplicou os estimadores de regressão e razão no contexto de cadastro duplo sob a estratégia estimação de domínios de Hartley (1962). Ele também considerou adaptações de estimadores de Fuller & Burmeister (1972) e Skinner & Rao (1996) para o caso de planos amostrais de cadastros duplos.

Ranalli *et al.* (2016) estenderam a abordagem de calibração, originalmente proposta por Deville & Särndal (1992) para o contexto de cadastro único, para planos amostrais de cadastro duplo. Molina *et al.* (2015) propuseram o uso de estimadores MLGREG em planos de cadastro duplo, também a partir de uma abordagem de domínios. Ainda, Rueda *et al.* (2018) desenvolveram estimadores para dados categóricos ordinais utilizando a estratégia de multiplicidade, com aplicação em planos amostrais de múltiplos cadastros.

Mais recentemente, Yan, Ye & Zhang (2021) desenvolveram estimadores de regressão não paramétrica para pesquisas de cadastro duplo utilizado a estratégia de domínios. Por sua vez, Rueda *et al.* (2021) propuseram estimadores para cadastro duplo baseados no método de verossimilhança empírica populacional proposto por Chen & Kim (2014), bem como uma extensão dessa abordagem para múltiplos cadastros aplicando a estratégia de multiplicidade.

Planos amostrais de múltiplos cadastros são adotados quando dois ou mais cadastros são utilizados simultaneamente a fim de fornecer cobertura completa da população-alvo ou quando um cadastro que cobre completamente a população-alvo, mas apresenta alto custo de

operação, é complementado por outro, com cobertura incompleta e um menor custo de utilização. Um exemplo deste segundo caso são pesquisas agropecuárias que utilizam um cadastro de área em conjunto com um cadastro de lista, pois apesar do primeiro fornecer cobertura completa para a população de interesse, tem alto custo operacional se comparado com cadastros do tipo lista que, por sua vez, muitas vezes não garantem um bom nível de cobertura (CARFAGNA, 2001).

Neste trabalho, o problema de estimação de parâmetros em populações finitas com base em planos amostrais de múltiplos cadastros será abordado considerando a incorporação de informações auxiliares disponíveis a nível de cadastro, por meio de estimadores assistidos por modelos lineares generalizados (GEREG). O desenvolvimento teórico apresentado foi motivado pelo potencial de aplicação dos estimadores GEREG em pesquisas amostrais no setor agropecuário, onde a combinação eficiente de múltiplos cadastros pode melhorar a precisão das estimativas e reduzir os custos operacionais.

A contribuição desta tese é estender a forma e os principais resultados dos GEREG para serem utilizados em planos amostrais baseados em dois ou mais cadastros. Os resultados estendidos envolvem uma proposta de estimador para cadastro duplo, baseada na abordagem de Hartley (1962, 1974), e uma proposta de estimador para múltiplos cadastros, baseada na abordagem de multiplicidade (MECATTI, 2007).

Os desempenhos dos estimadores propostos foram avaliados por meio de estudos de simulação e no contexto de pesquisas agropecuárias. Para este fim, populações fictícias foram criadas a partir dos resultados do Censo Agropecuário do Brasil, realizado no ano de 2017. Os estudos de simulação foram desenvolvidos seguindo a proposta de Ferraz, Mecatti & Torres (2022).

Esta tese está dividida em sete capítulos, incluindo esta introdução. No capítulo 2, são discutidos conceitos essenciais relacionados à estimação em populações finitas assistida por modelos lineares generalizados. O capítulo 3 oferece uma breve revisão das estratégias de estimação em planos amostrais que envolvem múltiplos cadastros. São apresentadas as abordagens de inferência por domínios, com maior enfoque em planos cadastro duplo, e por multiplicidade. No capítulo 4, define os estimadores GEREG sob a estratégia de domínios e multiplicidades simples, juntamente como seus valores esperados aproximados e as respectivas variâncias aproximadas. No capítulo 5, são demonstradas as condições sob as quais os estimadores propostos podem ser expressos em termos dos modelos (MLG) ajustados por meio das amostras selecionadas em cada cadastro. No capítulo 6, são conduzidos estudos de simulação e, por fim, no capítulo 7, são apresentadas considerações finais e perspectivas para trabalhos futuros.

2 INFERÊNCIA ASSISTIDA POR MODELOS

Estudos amostrais visam a estimação de parametros populacionais dentre os quais se destacam proporções médias e totais. Considere y_k como o valor de uma variável de interesse associado ao elemento $k \in U$. Totais populacionais correspondem a somas de todos os valores observados nos elementos de uma população. Eles podem ser interpretados em diversos contextos. Quando y_k assume valores quantitativos como, por exemplo, a área plantada com milho pelo produtor $k \in U$, o total populacional, expresso como $Y = \sum_{k \in U} y_k$, corresponde ao total de área plantada com milho no território definido pela população-alvo. Quando y_k é definido como uma variável indicadora de presença de um atributo qualitativo - por exemplo, se o produtor k utiliza máquinas para a colheita de sua produção - Y corresponde ao total de elementos com o atributo qualitativo na população, ou seja, o número total de produtores que utilizam máquinas de colheita.

Uma vez conhecido o número de elementos que compõem a população, estimar totais, médias ou proporções são problemas equivalentes. Nesta tese, o foco está na estimação de parâmetros definidos como proporções e totais populacionais.

2.1 ESTIMAÇÃO ASSISTIDA POR MODELO LINEAR

Em inferência assistida por modelos, a estimação de parâmetros leva em consideração a relação entre a variável de interesse e as variáveis auxiliares disponíveis, descrita através de um modelo geral de regressão. Quanto melhor a adequação do modelo formulado entre a variável de interesse e as variáveis auxiliares, maior será o ganho de precisão do estimador assistido pelo modelo considerado. Quando o objetivo é estimar o total Y, os valores não observados na amostra, inclusive por não-resposta, podem ser preditos pela equação:

$$\mathrm{E}(y_k|\boldsymbol{x}_k) = \mu_k = h(\boldsymbol{\beta};\boldsymbol{x}_k), \;\; k = 1,\ldots,N, \tag{2.1}$$

em que $\boldsymbol{\beta}$ é um vetor de parâmetros desconhecidos e $\boldsymbol{x}_k = (x_{1k},\ldots,x_{Jk})$ é um vetor de informação auxiliar para o k-ésimo elemento da população-alvo. Para o modelo de regressão linear geral, com $\hat{\boldsymbol{\beta}}$ denotando o valor estimado de $\boldsymbol{\beta}$ com base na amostra, a estrutura de $h(.)$ é dada por

$$\hat{\mu}_k = \hat{y}_k = \boldsymbol{x}_k^T \hat{\boldsymbol{\beta}}.$$

Na estimação do vetor de parâmetros faz-se a suposição que um modelo linear descreve adequadamente a dispersão entre a variável de interesse e as variáveis auxiliares. Definindo $\text{Var}(y_k) = \sigma_k$, o estimador para $\boldsymbol{\beta}$ pode ser derivado usando os princípios de estimação de parâmetros complexos - quando o parâmetro pode ser escrito em função de totais populacionais, e seu estimador é obtido substituindo os totais pelos respectivos estimadores de Horvitz-Thompson (SÄRNDAL; SWENSSON; WRETMAN, 2003):

$$\hat{\boldsymbol{\beta}} = \left(\sum_{k \in S} \frac{\boldsymbol{x}_k \boldsymbol{x}_k^T}{\pi_k \sigma_k} \right)^{-1} \left(\sum_{k \in S} \frac{\boldsymbol{x}_k \boldsymbol{y}_k^T}{\pi_k \sigma_k} \right),$$

em que π_k é a probabilidade de inclusão do elemento k na amostra S.

O estimador do tipo regressão é obtido baseado na diferença entre os valores estimados e os valores observados para a variável de interesse na amostra. Uma das formas de expressar o estimador regressão para o total Y é fornecida por:

$$\hat{Y}_{GREG} = \sum_{k \in U} \hat{y}_k + \sum_{k \in S} \frac{(y_k - \hat{y}_k)}{\pi_k}. \tag{2.2}$$

O estimador (2.2) é a soma das estimativas do modelo para variável de interesse na população U mais um termo de ajuste com base na diferença entre os valores estimados e os valores observados na amostra. A variância do estimador de regressão, derivada por linearização de Taylor, é expressa por

$$\text{Var}_p\left(\hat{Y}_{GREG}\right) \approx \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) \frac{(y_k - \hat{y}_k)}{\pi_k} \frac{(y_l - \hat{y}_l)}{\pi_l},$$

em que π_{kl} é a probabilidade de inclusão de segunda ordem para os elementos k e l da população na amostra S. Um estimador não viesado da variância é dado por

$$\widehat{\text{Var}}_p\left(\hat{Y}_{GREG}\right) = \sum_{k \in U} \sum_{l \in U} \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_{kl}} \frac{(y_k - \hat{y}_k)}{\pi_k} \frac{(y_l - \hat{y}_l)}{\pi_l}.$$

2.2 ESTIMAÇÃO ASSISTIDA POR MODELO LINEAR GENERALIZADO

Modelos lineares generalizados expandem a relação dada em (2.1) para casos em que a variável resposta tem distribuição pertencente à família exponencial (MCCULLAGH; NELDER, 1989). Nesse caso, o valor esperado da variável dependente é relacionado de maneira biunívoca com a componente sistemática $h(\boldsymbol{\beta}; \boldsymbol{x}_k)$ por meio de uma função de ligação. Considerando um modelo linear nos parâmetros, os valores não observados na amostra podem ser estimados por

$$\mathrm{E}(y_k|\boldsymbol{x}_k) = \mu_k = g^{-1}\left(\boldsymbol{x}_k^T\boldsymbol{\beta}\right), \;\; k = 1,\ldots,N, \tag{2.3}$$

onde $g(\cdot)$ é uma função contínua duplamente diferenciável. Para populações finitas, os parâmetros do modelo devem ser estimados, fazendo-se distinção entre as estimativas obtidas para a amostra $\hat{\boldsymbol{\beta}}^S$ e para a população $\hat{\boldsymbol{\beta}}^U$ por meio da maximização da função de log-verossimilhança amostral e populacional, respectivamente.

A influência do plano amostral é inserida no processo de estimação por meio da ponderação da função de log-verossimilhança pelas respectivas probabilidades de inclusão de primeira ordem, i.e.,

$$\hat{\boldsymbol{\beta}}^S = \arg\max_{\boldsymbol{\beta}} \sum_{k\in S} l_k^{\pi}(\mu_k;\boldsymbol{\beta}),$$

em que $l_k^{\pi}(\mu_k;\boldsymbol{\beta}) = \frac{1}{\pi_k} l_k(\mu_k;\boldsymbol{\beta})$ é conhecida como função de *pseudo log-verossimilhança* de y_k. Essa função pode ser entendida como uma estimativa para a função de log-verossimilhança populacional através da amostra. Quando o modelo populacional é corretamente especificado, é possível mostrar que $\hat{\boldsymbol{\beta}}^S$ é consistente para $\boldsymbol{\beta}$ (LUMLEY; SCOTT, 2017).

2.2.1 Família exponencial de distribuições

A família exponencial abrange muitas distribuições como: gama, normal inversa, exponencial, Poisson, Bernoulli e multinomial. Essas distribuições permitem analisar variáveis resposta que tenham comportamento assimétrico, variáveis resposta com domínio restrito a um intervalo no conjunto dos reais, como o intervalo (0,1), e variáveis resposta discretas. A classe de distribuições da família exponencial é caracterizada por distribuições cujas funções de probabilidade é da forma

$$f(y;\theta,\phi) = \exp\{\phi[y\theta - b(\theta)] + c(y,\phi)\}, \tag{2.4}$$

em que $b(\cdot)$ e $c(\cdot)$ são funções conhecidas, θ o parâmetro canônico e $\phi^{-1} > 0$ é o parâmetro de dispersão.

Essa classe de distribuições apresenta boas propriedades estatísticas, como a facilidade de se calcular momentos por meio das derivadas de $b(\theta)$ (função geradora de cumulantes)

em relação ao parâmetro canônico θ. Por exemplo, o valor esperado e variância são dados por:

$$\mathrm{E}(y) = \mu = b'(\theta) \quad \text{e} \quad \mathrm{Var}(y) = \phi^{-1} b''(\theta) = \phi^{-1} V(\mu),$$

em que $V(\mu) = \partial\mu/\partial\theta$ é conhecida como função de variância. Os modelos lineares generalizados utilizados nessa tese, seguem a estrutura utilizada por Rondon, Vanegas & Ferraz (2012). Assim, os modelos ξ propostos para estimar os valores das variáveis de interesse não observados na amostra consideram a seguinte estrutura:

$$\begin{cases} \mathrm{E}_\xi(y_k) = \mu_k, \\ \mathrm{Var}_\xi(y_k) = \phi_k^{-1} V(\mu_k), \\ g(\mu_k) = \eta_k = \left(\boldsymbol{x}_k^T \boldsymbol{\beta}\right) \end{cases} \tag{2.5}$$

onde $\mathrm{E}_\xi(\cdot)$ e $\mathrm{Var}_\xi(\cdot)$ representam o valor esperado e a variância com respeito ao modelo ξ e $g(\cdot)$ uma função de ligação.

Uma escolha particular de função de ligação está relacionada com o parâmetro canônico da família exponencial. Se o preditor linear modela diretamente o parâmetro canônico, ou seja, $\theta_k = \eta_k = g(\mu_k)$, então a função $g(\cdot)$ é chamada de ligação canônica. O capítulo 5 apresenta alguns resultados que possibilitam a simplificação do estimador do tipo regressão quando o modelo especificado utiliza uma função de ligação canônica.

2.2.2 Estimador de regressão linear generalizado (GEREG)

Considere $\hat{\mu}_k^S = g^{-1}\left(\boldsymbol{x}_k^T \hat{\boldsymbol{\beta}}^S\right)$ um estimador para $\hat{y}_k$ baseado no modelo amostral com estrutura na forma (2.5). O estimador GEREG para o total Y pode ser expresso por:

$$\hat{Y}_{GEREG} = \sum_{k\in U} \hat{\mu}_k^S + \sum_{k\in S} \frac{(y_k - \hat{\mu}_k^S)}{\pi_k}. \tag{2.6}$$

Sob a suposição que $\hat{\mu}_k^S \approx \hat{\mu}_k^U$, pode-se mostrar que o viés do estimador (2.6) é aproximadamente nulo. Considere $E_k = y_k - \hat{\mu}_k^U$ o k-ésimo resíduo do modelo proposto com base na população U e $\mathrm{E}_p(\cdot)$ o valor esperado com respeito ao plano amostral $p(\cdot)$, então,

$$\mathrm{E}_p(\hat{Y}_{GEREG}) \approx \sum_{k\in U} \hat{\mu}_k^S + \mathrm{E}_p\left(\sum_{k\in S} \frac{E_k}{\pi_k}\right) = Y,$$

em que,

$$\begin{aligned}
\mathrm{E}_p\left(\sum_{k\in S}\frac{E_k}{\pi_k}\right) &= \mathrm{E}_p\left(\sum_{k\in S}\frac{y_k}{\pi_k}\right) - \mathrm{E}_p\left(\sum_{k\in S}\frac{\hat{\mu}_k^U}{\pi_k}\right) \\
&= \sum_{k\in U}\frac{y_k \mathrm{E}_p(I_k)}{\pi_k} - \sum_{k\in U}\frac{\hat{\mu}_k^U \mathrm{E}_p(I_k)}{\pi_k} \\
&= \sum_{k\in U} y_k - \sum_{k\in U}\hat{\mu}_k^U = Y - \sum_{k\in U}\hat{\mu}_k^U,
\end{aligned}$$

com $\mathrm{E}_p(I_k) = \pi_k$ a probabilidade de inclusão de primeira ordem do elemento $k \in U$. De maneira análoga, pode-se obter uma expressão aproximada para a variância de $\hat{Y}_{GEREG}$:

$$\mathrm{Var}_p\left(\hat{Y}_{GEREG}\right) \approx \mathrm{Var}_p\left(\sum_{k\in S}\frac{E_k}{\pi_k}\right) = \sum_{k\in U}\sum_{l\in U}\left(\pi_{kl} - \pi_k\pi_l\right)\frac{E_k}{\pi_k}\frac{E_l}{\pi_l}, \tag{2.7}$$

onde π_{kl} é a probabilidade de inclusão de segunda ordem para os elementos k e l da população. A equação (2.7) fornece uma aproximação da variância do estimador (2.6) obtida através da fórmula da variância do estimador de Horvitz-Thompson aplicada aos resíduos do modelo proposto. Com base em (2.7), define-se um estimador para a variância de $\hat{Y}_{GEREG}$ expresso por:

$$\widehat{\mathrm{Var}}_p\left(\hat{Y}_{GEREG}\right) = \sum_{k\in S}\sum_{l\in S}\frac{\left(\pi_{kl} - \pi_k\pi_l\right)}{\pi_{kl}}\frac{e_k}{\pi_k}\frac{e_l}{\pi_l}, \tag{2.8}$$

em que $e_k = y_k - \hat{\mu}_k^S$ é o resíduo amostral do elemento k.

Os resultados aproximados que foram apresentados dependem do quão próximo é o modelo $\hat{\mu}_k^S$, estimado por meio da amostra, do modelo populacional $\hat{\mu}_k^U$ estimado a partir da população U e isso está diretamente relacionado com o quão próximas são as estimativas $\hat{\boldsymbol{\beta}}^S$ e $\hat{\boldsymbol{\beta}}^U$.

2.2.3 Estimador Generalizado Logístico (LGREG)

Um caso particular do estimador GEREG considera o modelo logístico quando a variável de interesse é dicotômica. O estimador de regressão generalizado logístico (LGREG), proposto por Lehtonen & Veijanen (1998), utiliza como função de ligação do modelo a função *logit*. No contexto em que a variável resposta é dicotômica, ou seja, $Y_k = 1$ ou $Y_k = 0$ para $k \in U$, a aplicação do modelo logístico é mais adequado no sentido de garantir que o valor predito pertença ao intervalo unitário; em (2.3) tem-se que $g^{-1}\left(\boldsymbol{x}_k^T\boldsymbol{\beta}\right) \in (0,1)$. Denotando como $\mu_k = \mathrm{P}(Y_k = 1|\boldsymbol{x}_k)$, a função de ligação logit é dada por:

$$g(\mu_k) = \log\left(\frac{\mu_k}{1-\mu_k}\right)$$

e os valores populacionais podem ser obtidos por

$$\mu_k = g^{-1}\left(\boldsymbol{x}_k^T\boldsymbol{\beta}\right) = \frac{\exp\left(\boldsymbol{x}_k^T\boldsymbol{\beta}\right)}{1+\exp\left(\boldsymbol{x}_k^T\boldsymbol{\beta}\right)}, \tag{2.9}$$

com $\mathrm{E}(Y_k|\boldsymbol{x}_k) = \mathrm{P}(Y_k = 1|\boldsymbol{x}_k)$. No contexto de populações finitas, o modelo populacional (2.9) pode ser estimado com base em uma amostra. O estimador LGREG tem a mesma estrutura do estimador em (2.6) e pode ser expresso por:

$$\hat{Y}_{LGREG} = \sum_{k\in U}\hat{\mu}_k^S + \sum_{k\in S}\frac{y_k - \hat{\mu}_k^S}{\pi_k} = \sum_{k\in S}\frac{g_{ks}y_k}{\pi_k},$$

onde

$$\hat{\mu}_k^S = \frac{\exp\left(\boldsymbol{x}_k^T\hat{\boldsymbol{\beta}}^S\right)}{1+\exp\left(\boldsymbol{x}_k^T\hat{\boldsymbol{\beta}}^S\right)},$$

e

$$g_{ks} = 1 + \frac{\sum_{k\in U}\hat{\mu}_k^S - \sum_{k\in S}\frac{\hat{\mu}_k^S}{\pi_k}}{\hat{Y}_\pi},$$

em que, $\hat{Y}_\pi$ é o estimador de Horvitz-Thompson para o total Y, g_{ks} um fator de calibração para π_k e $\hat{\boldsymbol{\beta}}^S$ é a estimativa de máxima pseudo log-verossimilhança para a amostra S obtida através de

$$\hat{\boldsymbol{\beta}}^S = \arg\max_{\boldsymbol{\beta}}\sum_{k\in S}\frac{1}{\pi_k}\left[y_k\log\mu_k + (1-y_k)\log(1-\mu_k)\right]. \tag{2.10}$$

Para maximizar a expressão (2.10) é necessário aplicar métodos numéricos de otimização, como o de Newton-Raphson e scoring de Fisher. Um estimador para a variância do LGREG obtido a partir da aplicação direta da expressão (2.8) é

$$\widehat{\mathrm{Var}}_p\left(\hat{Y}_{LGREG}\right) = \sum_{k\in S}\sum_{l\in S}\frac{(\pi_{kl}-\pi_k\pi_l)}{\pi_{kl}}\frac{e_k}{\pi_k}\frac{e_l}{\pi_l},$$

em que $e_k = y_k - \hat{\mu}_k^S$. Outro estimador para a variância do LGREG que utiliza os fatores de calibração g_{ks} em sua forma é dado por

$$\widehat{\mathrm{Var}}_p\left(\hat{Y}_{LGREG}\right) = \sum_{k\in S}\sum_{l\in S}\frac{(\pi_{kl}-\pi_k\pi_l)}{\pi_{kl}}\frac{g_{ks}e_k}{\pi_k}\frac{g_{ls}e_l}{\pi_l}.$$

Nos dois casos, os estimadores da variância do LGREG são aproximações baseadas em grandes amostras (SÄRNDAL; SWENSSON; WRETMAN, 2003, p. 234 - 238).

2.2.4 Estimador Generalizado Logístico Multinomial (MLGREG)

Uma generalização para o caso em que a variável resposta é dicotômica ocorre quando y_k pode assumir valores correspondentes a m classes distintas. Lehtonen & Veijanen (1998) propuseram um estimador baseado no LGREG para o contexto de variáveis respostas multinomiais. Assumindo a codificação das classes nos valores discretos $1,\ldots,m$, o objetivo é estimar a distribuição de frequência de y para a população U. Para estimar a distribuição de frequência, define-se m variáveis indicadoras tais que

$$z_{ik} = \begin{cases} 1, & \text{se } y_k = i \\ 0, & \text{caso contrário} \end{cases} \tag{2.11}$$

para $i = 1,\ldots,m$. Considerando $\boldsymbol{x}_k$ um vetor de informação auxiliar para o k-ésimo elemento da população-alvo, o problema de estimar as proporções pode ser abordado utilizando o modelo logístico usual. O problema consiste agora em estimar as proporções $P_i = N^{-1}\sum_{k\in U} z_{ik} =^{-1} Z_i$, com Z_i denotando o total da variável indicadora z_i, como definido em (2.11), para $i = 1,\ldots,m$. Com base no modelo logístico para variáveis dicotômicas descrito pela equação (2.9), um estimador para $\mu_{ik} = \mathrm{P}(z_{ik} = 1|\boldsymbol{x}_k) = \mathrm{P}(Y_k = i|\boldsymbol{x}_k)$ é dado por:

$$\hat{\mu}_{ik}^S = \frac{\exp\left(\boldsymbol{x}_k^T\hat{\boldsymbol{\beta}}_i^S\right)}{1+\sum_{r=1,\ldots,m}\exp\left(\boldsymbol{x}_k^T\hat{\boldsymbol{\beta}}_i^S\right)},\ i = 1,\ldots,m.$$

Considere a distribuição de frequência populacional $\boldsymbol{Y} = (Z_{k1},\ldots,Z_{km})$. O estimador de regressão logístico multinomial (MLGREG) para $\boldsymbol{Y}$ é dado por $\hat{\boldsymbol{Y}}_{MLGREG} = \left(\hat{Z}_{k1},\ldots,\hat{Z}_{km}\right)$, em que,

$$\hat{Z}_i = \sum_{k\in U}\hat{\mu}_{ik}^S + \sum_{k\in S}\frac{z_{ik}-\hat{\mu}_{ik}^S}{\pi_k} = \sum_{k\in S}\frac{w_{ks}^i z_{ki}}{\pi_k},\ i = 1,\ldots,m, \tag{2.12}$$

com $\pi_k = \mathrm{P}(y_k \in S)$, $k \in U$, e

$$w_{ks}^{i} = 1 + \frac{\sum_{k\in U} \hat{\mu}_{ik}^{S} - \sum_{k\in S} \frac{\hat{\mu}_{ik}^{S}}{\pi_k}}{\hat{Z}_{i\pi}}.$$

onde $\hat{Z}_{i\pi}$ o estimador de Horvitz-Thompson para o total Z_i. As etapas descritas até agora para explicitar o estimador MLGREG são análogas às aplicadas para definir o estimador LGREG com variável resposta dicotômica. Contudo, o processo de estimação do vetor de parâmetros $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_m^T\right)^T$ deve considerar as m classes simultaneamente, sendo inadequado o uso da equação (2.10) com a substituição de y_k por z_{ik}. Dessa forma, o estimador de pseudo máxima verossimilhança da amostra S para $\boldsymbol{\beta}$ é a solução numérica de

$$\hat{\boldsymbol{\beta}}^{S} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{m} \left(\sum_{k\in S} z_{ik} \ln \mu_{ik} \right). \tag{2.13}$$

2.3 ESTIMAÇÃO COM ESTRATIFICAÇÃO

Em muitas pesquisas, a população de interesse pode ser particionada em subpopulações disjuntas que são identificadas a priori. Nesta situação, é possível adotar um plano amostral estratificado por tais subpopulações. Nesse caso, amostras independentes são selecionadas em cada estrato e a amostra final é composta por todos os elementos selecionados. Considere a população $U = \cup_{h=1}^{H} U_h$ e a amostra $S = \cup_{h=1}^{H} S_h$, em que H é o número de estratos e S_h a amostra selecionada através do plano amostral $p_h(\cdot)$ aplicado no estrato h. O total da variável de interesse y pode ser escrito por

$$Y = \sum_{h=1}^{H} \sum_{k\in U_h} y_k = \sum_{h=1}^{H} Y_h.$$

em que Y_h é o total da variável de interesse no estrato h.

O estimador de Y com base em uma amostra estratificada é expresso por

$$\hat{Y} = \sum_{h=1}^{H} \hat{Y}_h,$$

em que $\hat{Y}_h$ é um estimador para Y_h. Se $\hat{Y}_h$ for um estimador não viesado para $h = 1, \ldots, H$, então o estimador acima é não viesado para Y. Além disso, como planos amostrais são aplicados em cada estrato de maneira independente, a variância de $\hat{Y}$ é a soma das variâncias dos estimadores de cada estrato. Assim,

$$\text{Var}_p(\hat{Y}) = \sum_{h=1}^{H} \text{Var}_p(\hat{Y}_h)$$

com estimador não-viesado dado por

$$\widehat{\text{Var}}_p(\hat{Y}) = \sum_{h=1}^{H} \widehat{\text{Var}}_p(\hat{Y}_h),$$

desde que $\widehat{\text{Var}}_p(\hat{Y}_h)$ seja um estimador não viesado para $\text{Var}_p(\hat{Y}_h)$.

Assumindo que informações auxiliares estejam disponíveis para toda população, uma possibilidade é utilizar o estimador GEREG, expresso em (2.6), para estimar os totais $Y_h,\ h = 1,\ldots,H$. A seguir, são revistas duas formas de aplicação de estimadores do tipo regressão no contexto de amostragem estratificada.

2.3.1 Estimador de regressão separado

Um estimador do tipo regressão é considerado separado quando um modelo é ajustado em cada estrato. Esse tipo de abordagem é adequada quando os elementos pertencentes ao mesmo estrato possuem características similares, enquanto elementos pertencentes a diferentes estratos possuem características heterogêneas. Nesse caso, uma suposição razoável é que a relação entre a variável resposta e as variáveis auxiliares não é a mesma em todos os estratos. Consequentemente, modelos são ajustados separadamente para cada estrato a partir das amostras $S_h,\ h = 1,\ldots,H$.

Considerando modelos amostrais com estrutura expressa em (2.5), o valores da variável de interesse são estimados por

$$\hat{y}_k = \hat{\mu}_k^{S_h} = g^{-1}\left(\boldsymbol{x}_k^T \hat{\boldsymbol{\beta}}^{S_h}\right),\ \forall k \in U_h \text{ e } h = 1,\ldots,H,$$

em que $\hat{\boldsymbol{\beta}}^{S_h}$ é a estimativa de pseudo máxima-verossimilhança para $\boldsymbol{\beta}^{U_h}$ em S_h dada por

$$\hat{\boldsymbol{\beta}}^{S_h} = \underset{\boldsymbol{\beta}^{U_h}}{\arg\max} \sum_{k \in S_h} \frac{1}{\pi_k^h}\, l\left(\boldsymbol{\beta}^{U_h}\right)$$

em que π_k^h é a probabilidade de inclusão de primeira ordem do elemento $k \in h$. Assim, o estimador separado para o total Y assistido por modelos lineares generalizados fica definido como

$$\hat{Y}_{GEREG_S} = \sum_{h=1}^{H} \left[\sum_{k \in U_h} \hat{\mu}_k^{S_h} + \sum_{k \in S_h} \frac{\left(y_k - \hat{\mu}_k^{S_h} \right)}{\pi_k^h} \right].$$

A variância aproximada do estimador $\hat{Y}_{GEREG_S}$ obtido a partir de expressão (2.7) é

$$\mathrm{Var}_p \left(\hat{Y}_{GEREG_S} \right) \approx \sum_{h=1}^{H} \left(\sum_{k \in U_h} \sum_{l \in U_h} \left(\pi_{kl}^h - \pi_k^h \pi_l^h \right) \frac{E_k}{\pi_k^h} \frac{E_l}{\pi_l^h} \right).$$

em que $E_k = y_k - \hat{\mu}_k^{U_h}$ e π_{kl}^h é a probabilidade de inclusão de segunda ordem dos elementos $k, l \in h$.

2.3.2 Estimador de regressão combinado

Um estimador do tipo regressão é considerado combinado quando um único modelo é formulado para toda população, supondo que a relação entre a variável de interesse e as variáveis auxiliares seja a mesma em todos os estratos. Nesse caso, a estimação do modelo utiliza os dados das amostras de cada estrato de forma combinada, ou seja, todos os estratos são considerados simultaneamente para a estimação dos parâmetros do modelo. Assim,

$$\hat{\boldsymbol{\beta}}^S = \arg\max_{\boldsymbol{\beta}} \sum_{h=1}^{H} \sum_{k \in S_h} \frac{1}{\pi_k^h} \, l(\boldsymbol{\beta}),$$

em que $l(\boldsymbol{\beta})$ é função de log-verossimilhança de uma distribuição pertencente à família exponencial. O estimador combinado para o total Y assistido por modelos lineares generalizados é dado por

$$\hat{Y}_{GEREG_C} = \sum_{k \in U} \hat{\mu}_k^S + \sum_{k \in S} \frac{\left(y_k - \hat{\mu}_k^S \right)}{\pi_k^h},$$

em que $\hat{y}_k = \hat{\mu}_k^S = g^{-1}(\boldsymbol{x}_k^T \hat{\boldsymbol{\beta}}^S)$, para todo $k \in U$. Esse estimador tem forma igual aquela expressa em (2.6), porém, na etapa de estimação dos parâmetros do modelo, os diferentes planos amostrais aplicados nos estratos são considerados por meio da ponderação da função de log-verossimilhança por π_k^h. Da mesma forma, a variância do estimador $\hat{Y}_{GEREG_C}$ é obtido a partir de expressão (2.7).

3 INFERÊNCIA COM MÚLTIPLOS CADASTROS

Neste capítulo, serão revisadas duas estratégias para realizar inferência em planos amostrais de cadastro duplo: inferência por domínio e inferência por multiplicidade. As principais vantagens e desvantagens de cada uma delas serão discutidas, tendo como objetivo a estimação de um total populacional.

3.1 INFERÊNCIA POR DOMÍNIOS

Hartley (1962) foi provavelmente o primeiro a introduzir uma metodologia inferencial completa para planos amostrais de múltiplos cadastros. Na abordagem de Hartley, é idealizado um plano amostral no qual vários cadastros sobrepostos são usados para fornecer cobertura completa da população de interesse, com as unidades que compõe a população identificadas em pelo menos um dos cadastros.

Em planos de cadastro duplo, o uso simultâneo de dois cadastros, denotados por A e B, cuja interseção parcial não é nula, induz a três domínios de estimação, denotados por a, b e $ab(A) = ab(B)$, respectivamente, que podem ser estrategicamente utilizados para gerar estimativas de parâmetros. Os domínios de estimação formam uma partição da população U de modo que é possível escrever

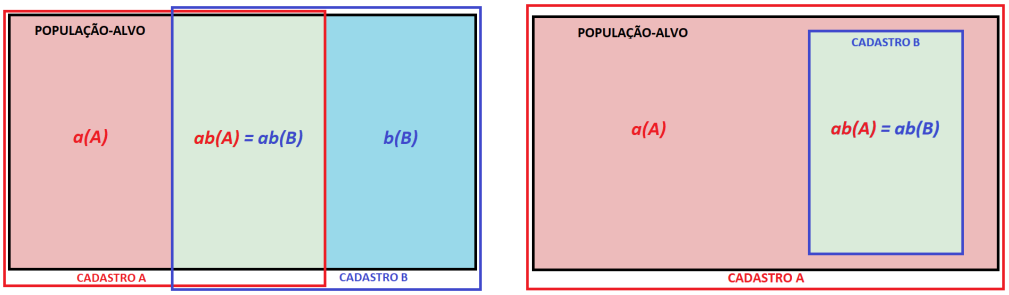
$$U = a \cup b \cup ab(A) = a \cup b \cup ab(B),$$

em que a é o conjunto de elementos identificados apenas pelo cadastro A, b o conjunto de elementos identificados apenas pelo cadastro B, $ab(A)$ o conjunto de elementos da interseção identificados pelo cadastro A, e $ab(B)$ o conjunto de elementos da interseção identificados pelo cadastro B. Na forma acima, é possível obter o total populacional Y a partir da soma de todos os valores observados nos domínios a, b e ab, com $ab(A) = ab(B)$.

Utilizar dois cadastros pode também ser uma solução para reduzir custos em pesquisas. Em alguns casos, um cadastro com cobertura completa tem alto custo de utilização, enquanto outro com cobertura parcial tem custo menor. Assim, utilizando os dois cadastros, é possível obter uma relação custo-benefício mais vantajosa na seleção da amostra, mesmo que o domínio b seja vazio. A figura 1 ilustra duas possíveis sobreposições dos cadastros. No cenário (a), o uso simultâneo de dois cadastros, cuja interseção é não vazia, fornece cobertura completa

para a população de interesse. Já no cenário (b), o cadastro A sozinho oferece cobertura completa, enquanto B é um subconjunto de A. Esse último caso é comum em pesquisas que utilizam um cadastro de área, como Pesquisas Agropecuárias e Ambientais, onde a população de interesse pode ser delimitada geograficamente (FERRAZ; MECATTI; TORRES, 2022; GALLEGO *et al.*, 1995). O caso em que $A \cap B = \emptyset$ não é considerado como cadastro duplo pois cada cadastro pode ser tratado como um estrato, permitindo a adoção de estratégias usuais de estimação sob estratificação.

Figura 1 – Domínios de estimação induzidos por uso simultâneo de dois cadastros



(a) **(b)**

Fonte: De autoria própria

3.1.1 Estratégia de Hartley

Considere duas amostras, S_A e S_B, selecionadas de maneira independente nos cadastros A e B, respectivamente, por meio de planos amostrais possivelmente diferentes. A partição de U nos domínios de estimação permite que a amostra S_A seja escrita como $S_A = S_a \cup S_{ab(A)}$, com S_a representando os elementos de S_A que pertencem ao domínio a; e $S_{ab(A)}$, os elementos de S_A que pertencem ao domínio $ab(A)$. De maneira análoga, S_B pode ser escrito como $S_B = S_b \cup S_{ab(B)}$, com S_b representando os elementos de S_B que pertencem ao domínio b; e $S_{ab(B)}$, os elementos de S_B que pertencem ao domínio $ab(B)$. Neste caso, ficam determinadas para o domínio ab duas amostras, selecionadas a partir de planos amostrais independentes.

Na abordagem por domínio, o total populacional Y é decomposto na soma dos totais para domínios de estimação, Y_a, Y_b e $Y_{ab(A)} = Y_{ab(B)}$. Contudo, é possível obter simultaneamente duas estimativas para o domínio ab, valendo-se das amostras $S_{ab(A)}$ e $S_{ab(B)}$. A estratégia usada por Hartley considera uma combinação linear dos estimadores para o total Y_{ab}:

$$\hat{Y}_{ab} = \alpha \hat{Y}_{ab(A)} + (1-\alpha)\hat{Y}_{ab(B)},$$

em que $\alpha \in [0,1]$ é uma constante e $\hat{Y}_{ab(A)}$ e $\hat{Y}_{ab(B)}$ são estimadores de Horvitz-Thompson (HT)

para o total no domínio *ab* a partir de $S_{ab(A)}$ e $S_{ab(B)}$, respectivamente. Assim, a classe de estimadores proposta por Hatley é dada por:

$$\hat{Y}_H = \hat{Y}_a + \alpha\hat{Y}_{ab(A)} + (1-\alpha)\hat{Y}_{ab(B)} + \hat{Y}_b, \tag{3.1}$$

com variância expressa por:

$$\begin{aligned}\text{Var}_p\left(\hat{Y}_H|\alpha\right) &= \text{Var}_p\left[\hat{Y}_a + \alpha\hat{Y}_{ab(A)}\right] + \text{Var}_p\left[(1-\alpha)\hat{Y}_{ab(B)} + \hat{Y}_b\right] \\ &= \text{Var}_p\left[\hat{Y}_a\right] + \text{Var}_p\left[\hat{Y}_b\right] \\ &+ \alpha^2\text{Var}_p\left[\hat{Y}_{ab(A)}\right] + (1-\alpha)^2\text{Var}_p\left[\hat{Y}_{ab(B)}\right] \\ &+ 2\alpha\text{Cov}_p\left[\hat{Y}_a, \hat{Y}_{ab(A)}\right] + 2(1-\alpha)\text{Cov}_p\left[\hat{Y}_b, \hat{Y}_{ab(B)}\right].\end{aligned}$$

A variância do estimador de Hartley pode ser expressa em termos matriciais considerando o vetor de pesos e de a matriz bloco diagonal de variâncias e covariâncias dos domínios:

$$\text{Var}_p\left(\hat{Y}_H|\alpha\right) = \boldsymbol{\alpha}^T\boldsymbol{\Sigma}\,\boldsymbol{\alpha}, \tag{3.2}$$

em que,

$$\boldsymbol{\alpha} = \begin{bmatrix} 1 \\ \alpha \\ 1 \\ 1-\alpha \end{bmatrix} \quad \text{e} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_a^2 & \sigma_{a;ab(A)} & 0 & 0 \\ \sigma_{a;ab(A)} & \sigma_{ab(A)}^2 & 0 & 0 \\ 0 & 0 & \sigma_b^2 & \sigma_{b;ab(B)} \\ 0 & 0 & \sigma_{b;ab(B)} & \sigma_{ab(B)}^2 \end{bmatrix}. \tag{3.3}$$

Um caso particular do estimador de Hartley, conhecido como estimador *screening*, corresponde a escolher $\alpha = 0$ em (3.1). Uma vantagem do estimador screening é que apenas o domínio *a* precisa ser completamente identificado nas amostras, pois, neste caso, o estimador para Y é:

$$\hat{Y}_{screening} = \hat{Y}_a + \hat{Y}_B, \tag{3.4}$$

em que $\hat{Y}_B$ é o estimador de Horvitz-Thompson a partir do cadastro B. Além disso, a variância do estimador screening também fica simplificada por

$$\text{Var}_p\left(\hat{Y}_{screening}\right) = \text{Var}_p\left(\hat{Y}_a\right) + \text{Var}_p\left(\hat{Y}_B\right), \tag{3.5}$$

e, na forma matricial, por

$$\text{Var}_p\left(\hat{Y}_{screening}\right) = \begin{bmatrix} \sigma_a^2 & 0 \\ 0 & \sigma_B^2 \end{bmatrix}.$$

3.1.2 Estimador ótimo de Hartley

Uma abordagem ótima é adotada quando o valor escolhido de α minimiza a variância na expressão (3.2), ou seja, o valor ótimo de α é

$$\alpha^{ot} = \underset{0\leq\alpha\leq 1}{\arg\min} \ \left\{ \mathrm{Var}_p\left(\hat{Y}_H\right)\right\}.$$

O próprio Hartley em 1974 derivou este resultado de tal forma a minimizar a variância assintótica do estimador (3.1). O valor ótimo pode ser escrito por:

$$\alpha^{ot} = \frac{\mathrm{Var}_p\left[\hat{Y}_{ab(B)}\right] + \mathrm{Cov}_p\left[\hat{Y}_b, \hat{Y}_{ab(B)}\right] - \mathrm{Cov}_p\left[\hat{Y}_a, \hat{Y}_{ab(A)}\right]}{\mathrm{Var}_p\left[\hat{Y}_{ab(A)}\right] + \mathrm{Var}_p\left[\hat{Y}_{ab(B)}\right]}.$$

É comum que os valores das variâncias e covariâncias populacionais sejam desconhecidos e seja necessário estimar esses parâmetros por meio das informações disponíveis nas amostras S_A e S_B. Assim, um estimador natural para α^{ot} é dado por:

$$\hat{\alpha}^{ot} = \frac{\widehat{\mathrm{Var}}_p\left[\hat{Y}_{ab(B)}\right] + \widehat{\mathrm{Cov}}_p\left[\hat{Y}_b, \hat{Y}_{ab(B)}\right] - \widehat{\mathrm{Cov}}_p\left[\hat{Y}_a, \hat{Y}_{ab(A)}\right]}{\widehat{\mathrm{Var}}_p\left[\hat{Y}_{ab(A)}\right] + \widehat{\mathrm{Var}}_p\left[\hat{Y}_{ab(B)}\right]}. \tag{3.6}$$

Desde que $\hat{\alpha}^{ot}$ seja consistente para α^{ot}, o estimador que é assintoticamente ótimo entre todos os estimadores da classe dada por (3.1) é expresso por:

$$\hat{Y}_H^{ot} = \hat{Y}_a + \hat{\alpha}^{ot}\hat{Y}_{ab(A)} + (1-\hat{\alpha}^{ot})\hat{Y}_{ab(B)} + \hat{Y}_b. \tag{3.7}$$

Embora o estimador ótimo de Hartley seja assintoticamente ótimo, quando o valor de $\widehat{\mathrm{Cov}}_p\left[\hat{Y}_b, \hat{Y}_{ab(B)}\right]$ ou $\widehat{\mathrm{Cov}}_p\left[\hat{Y}_a, \hat{Y}_{ab(A)}\right]$ for grande em valor absoluto, as estimativas de α^{ot} fornecidas por (3.6) podem estar fora do intervalo $(0,1)$.

Vários autores consideraram a estratégia de usar domínios para estimar totais após Hartley. No entanto, aplicar a abordagem de estimação por domínios em cenários com mais de dois cadastros pode rapidamente se tornar complicado. Isso ocorre porque, no caso de haver Q cadastros, existem $2^Q - 1$ possíveis domínios que podem ser identificados. Por exemplo, para $Q = 3$, o estimador de Y baseado nos domínios de estimação é expresso na seguinte forma

$$\begin{aligned}\hat{Y}_H = {} & \hat{Y}_a + \hat{Y}_b + \hat{Y}_c + \alpha_1\hat{Y}_{ab(A)} + (1-\alpha_1)\hat{Y}_{ab(B)} \\ & + \alpha_2\hat{Y}_{ac(A)} + (1-\alpha_2)\hat{Y}_{ac(C)} + \alpha_3\hat{Y}_{bc(B)} + (1-\alpha_3)\hat{Y}_{bc(C)} \\ & + \alpha_4\hat{Y}_{abc(A)} + \alpha_5\hat{Y}_{abc(B)} + (1-\alpha_4-\alpha_5)\hat{Y}_{abc(C)},\end{aligned}$$

em que, $0 \leq \alpha_1, \alpha_2, \alpha_3 \leq 1$ e $0 \leq \alpha_4 + \alpha_5 \leq 1$. Nesse caso, além de ser necessário identificar completamente sete domínios de estimação, é preciso escolher os valores de $\alpha_1, \ldots, \alpha_5$ que minimizam a variância de $\hat{Y}$.

3.2 INFERÊNCIA POR MULTIPLICIDADE

Uma abordagem alternativa para a estimação em planos de múltiplos cadastros é baseada na abordagem de multiplicidade, que pode acomodar naturalmente casos nos quais a quantidade de cadastros Q, cobrindo a população-alvo, é maior que dois, embora não utilize nenhum critério ótimo para reduzir a variância do estimador (MECATTI, 2007).

3.2.1 Estimador de multiplicidade simples

A abordagem de multiplicidade simples define m_k como o número de cadastros que identificam e dão acesso ao elemento $k \in U$. Se U_q denota o conjunto de elementos da população no cadastro q, $q \in \{1, \ldots, Q\}$, então $m_k = \sum_{q=1}^{Q} 1_{[k \in U_q]}$, onde $1_{[k \in U_q]}$ é uma variável indicadora assumindo valor 1 se $k \in U_q$, e zero, caso contrário. Em planos de múltiplos cadastros, o total da população pode ser expresso como uma função de fatores de multiplicidade simples:

$$Y = \sum_{q=1}^{Q} \sum_{k \in U_q} \frac{y_k}{m_k}.$$

O estimador de multiplicidade simples (MS) proposto por Mecatti (2007) tem forma:

$$\hat{Y}_{MS} = \sum_{q=1}^{Q} \sum_{k \in S_q} \frac{y_k}{m_k \pi_{k(q)}}, \tag{3.8}$$

em que S_q representa a amostra selecionada do cadastro q e $\pi_{k(q)}$ é a probabilidade de inclusão de primeira ordem do elemento k de U_q por meio do plano amostral $p_q(\cdot)$. Mecatti mostrou que (3.8) é um estimador não viesado para o total Y. Além disso, se amostras são selecionadas de maneira independente em cada cadastro, a variância de $\hat{Y}_{MS}$ é dada por

$$\mathrm{Var}_p\left(\hat{Y}_{MS}\right) = \sum_{q=1}^{Q} \sum_{k \in U_q} \sum_{l \in U_q} \left(\pi_{k(q)}\pi_{l(q)} - \pi_{kl(q)}\right) \left(\frac{y_k}{m_k \pi_{k(q)}} - \frac{y_l}{m_l \pi_{l(q)}} \right)^2,$$

onde $\pi_{kl(q)}$ denota a probabilidade de inclusão de segunda ordem do plano amostral do cadastro q. Um estimador não-viesado da variância é dado por

$$\widehat{\mathrm{Var}}_p\left(\hat{Y}_{MS}\right) = \sum_{q=1}^{Q} \sum_{k<l} \sum_{l \in S_q} \frac{\left(\pi_{k(q)}\pi_{l(q)} - \pi_{kl(q)}\right)}{\pi_{kl(q)}} \left(\frac{y_k}{m_k \pi_{k(q)}} - \frac{y_l}{m_l \pi_{l(q)}} \right)^2. \tag{3.9}$$

3.2.2 Estimador de multiplicidade generalizado

O estimador de multiplicidade simples foi ainda estendido, levando a uma classe de estimadores de multiplicidade generalizada (SINGH; MECATTI, 2011). Estimadores dessa classe tem forma expressa por

$$\hat{Y}_{MG} = \sum_{q=1}^{Q} \sum_{k \in S_q} \frac{y_k \alpha_{q(k)}}{\pi_{k(q)}} \tag{3.10}$$

e variância dada por

$$\mathrm{Var}_p\left(\hat{Y}_{MG}\right) = \sum_{q=1}^{Q} \sum_{k<l} \sum_{\in U_q} \left(\pi_{k(q)}\pi_{l(q)} - \pi_{kl(q)}\right) \left(\frac{y_k \alpha_{q(k)}}{\pi_{k(q)}} - \frac{y_l \alpha_{q(l)}}{\pi_{l(q)}}\right)^2,$$

em que, $\alpha_{q(k)} \in [0,1]$, com $\sum_q \alpha_{q(k)} = 1$, são os fatores de ajuste de multiplicidade correspondentes aos q cadastros para a unidade k. O estimador centrado para variância de $\hat{Y}_{MG}$ é expresso por

$$\widehat{\mathrm{Var}}_p\left(\hat{Y}_{MG}\right) = \sum_{q=1}^{Q} \sum_{k<l} \sum_{\in S_q} \frac{\left(\pi_{k(q)}\pi_{l(q)} - \pi_{kl(q)}\right)}{\pi_{kl(q)}} \left(\frac{y_k \alpha_{q(k)}}{\pi_{k(q)}} - \frac{y_l \alpha_{q(l)}}{\pi_{l(q)}}\right)^2.$$

O estimador de multiplicidade simples é obtido ao substituir $\alpha_{q(k)}$ pelo inverso de m_k em (3.10). De fato, o estimador de regressão generalizado abrange a maioria dos estimadores de múltiplos cadastros disponíveis na literatura, especificando adequadamente um conjunto de parâmetros (RUEDA *et al.*, 2018). No entanto, ainda há necessidade de estudos de seus respectivos desempenhos estatísticos.

As estratégias de estimação baseadas em domínio e multiplicidade têm grandes diferenças que são bastante evidentes quando se trata de pesquisas que usam mais de dois cadastros. A complexidade de identificar e classificar unidades amostrais em domínios cresce rapidamente, mesmo no caso de três cadastros sobrepostos, o que não ocorre com o fator de multiplicidade simples.

4 ESTIMADOR GEREG EM PLANOS DE MÚLTIPLOS CADASTROS

Em planos amostrais de cadastro duplo, é possível que informações auxiliares estejam disponíveis em cada cadastro. Nesse caso, um estimador semelhante ao dado em (2.6) pode ser obtido com modelos ajustados a partir das amostras em cada cadastro. Estimadores do tipo regressão, como expresso em (2.2), no contexto de cadastro duplo foram estudados por Coêlho (2011). Rondon, Vanegas & Ferraz (2012) propuseram o estimador GEREG no contexto de uma pesquisa com uso de um único cadastro. As contribuições originais desta tese são apresentadas a partir deste capítulo. Propomos a extensão dos estimadores assistidos de modelos lineares generalizados, conforme descrito no Capítulo 2, sob as abordagens de estimação de domínios e multiplicidade, descritas no Capítulo 3.

4.1 ESTIMADOR GEREG PARA CADASTRO DUPLO SOB A ABORDAGEM DE DOMÍNIOS

Suponha que para cada elemento $k \in A$ tenha-se a disposição o vetor de variáveis auxiliares $\boldsymbol{x}_{kA}$ e, além disso, para os elementos na amostra $k \in S_A$ sejam conhecidos os valores da variável de interesse y_k. Da mesma forma, para cada $k \in B$, que exista o vetor auxiliar $\boldsymbol{x}_{kB}$. Os valores da variável de interesse também são conhecidos apenas para os elementos na amostra S_B. Além disso, as variáveis auxiliares que compõem os vetores $\boldsymbol{x}_{kA}$ e $\boldsymbol{x}_{kB}$ podem ser diferentes.

É possível ajustar modelos a partir de S_A e S_B, separadamente, e estimar os valores não observados de y em cada cadastro. Da expressão (2.3) tem-se

$$\hat{\mu}_k^{S_A} = g_A^{-1}\left(\boldsymbol{x}_{kA}^T \hat{\boldsymbol{\beta}}_{S_A}^A\right), \;\; k = 1, \ldots, N_A, \tag{4.1}$$

em que, N_A é o tamanho da população U_A, $\hat{\boldsymbol{\beta}}_{S_A}^A$, a estimativa de pseudo máxima-verossimilhança de $\boldsymbol{\beta}^A$ com base em S_A e $g_A(\cdot)$ a função de ligação que melhor descreve a correlação entre as variáveis auxiliares e a variável de interesse no cadastro A. Desse forma, $\hat{\mu}_k^{S_A}$ fica determinado em função da estimação do vetor de parâmetros do modelo e da escolha de uma função de ligação $g_A(\cdot)$.

A estimação do parâmetro desconhecido $\boldsymbol{\beta}^A$ pode ser realizada por meio da maximização da função de pseudo log-verossimilhança em S_A. Considere d_k um fator de expansão definido em função dos domínios de estimação, tal que,

$$d_k = \begin{cases} \frac{1}{\pi_{kA}}, & \text{se } k \in S_a \\ \frac{\alpha}{\pi_{kA}}, & \text{se } k \in S_{ab(A)} \\ \frac{1-\alpha}{\pi_{kB}}, & \text{se } k \in S_{ab(B)} \\ \frac{1}{\pi_{kB}}, & \text{se } k \in S_b \end{cases} . \tag{4.2}$$

com $0 \geq \alpha \geq 1$. A definição de d_k requer que cada elemento nas amostras S_A e S_B seja identificado em um domínio de estimação. Caso essa identificação seja possível, a estimativa de pseudo máxima-verossimilhança de $\boldsymbol{\beta}^A$ é solução da equação

$$\hat{\boldsymbol{\beta}}^A_{S_A} = \arg\max_{\boldsymbol{\beta}^A} \sum_{k \in S_A} d_k l\left(\boldsymbol{\mu}_k; \boldsymbol{\beta}^A\right), \tag{4.3}$$

em que $l\left(\boldsymbol{\mu}_k; \boldsymbol{\beta}^A\right)$ é a função de log-verossimilhança de uma distribuição de probabilidade pertencente à família exponencial com forma (2.4). Uma vez que os valores de $\boldsymbol{\mu}_k$ são determinados por $\boldsymbol{\beta}^A$, como visto em (2.3), a partir de agora adotaremos a notação $l(\boldsymbol{\mu}_k)$ para representar a log-verossimilhança do elemento k avaliada em $\boldsymbol{\beta}^A$.

Com o modelo $\hat{\mu}^{S_A}$, pode-se obter estimativas por meio do estimador GEREG para os totais nos domínios a e $ab(A)$:

$$\hat{Y}_{GEREG_a} = \sum_{k \in a} \hat{\mu}_k^{S_A} + \sum_{k \in S_a} \frac{\left(y_k - \hat{\mu}_k^{S_A}\right)}{\pi_{kA}}$$

e

$$\hat{Y}_{GEREG_{ab(A)}} = \sum_{k \in ab(A)} \hat{\mu}_k^{S_A} + \sum_{k \in S_{ab(A)}} \frac{\left(y_k - \hat{\mu}_k^{S_A}\right)}{\pi_{kA}},$$

em que π_{kA} é probabilidade de inclusão de primeira ordem para o k-ésimo elemento no cadastro A por meio do plano amostral $p_A(\cdot)$. De maneira similar, são obtidas estimativas com auxílio do modelo $\hat{\mu}^{S_B}$ para os domínios b e $ab(B)$:

$$\hat{Y}_{GEREG_b} = \sum_{k \in b} \hat{\mu}_k^{S_B} + \sum_{k \in S_b} \frac{\left(y_k - \hat{\mu}_k^{S_B}\right)}{\pi_k}$$

e

$$\hat{Y}_{GEREG_{ab(B)}} = \sum_{k \in ab(B)} \hat{\mu}_k^{S_B} + \sum_{k \in S_{ab(B)}} \frac{\left(y_k - \hat{\mu}_k^{S_B}\right)}{\pi_{kB}},$$

em que π_{kB} é probabilidade de inclusão de primeira ordem para o k-ésimo elemento no cadastro B por meio do plano amostral $p_B(\cdot)$. Sob a abordagem de domínios, o estimador para o total Y é

dado por:

$$\begin{aligned}\hat{Y}^{H}_{GEREG} &= \hat{Y}_{GEREG_a} + \alpha\hat{Y}_{GEREG_{ab(A)}} + (1-\alpha)\,\hat{Y}_{GEREG_{ab(B)}} + \hat{Y}_{GEREG_b} \\ &= \sum_{k\in a}\hat{\mu}_k^{S_A} + \alpha\sum_{k\in ab(A)}\hat{\mu}_k^{S_A} + (1-\alpha)\sum_{k\in ab(B)}\hat{\mu}_k^{S_B} + \sum_{k\in b}\hat{\mu}_k^{S_B} \\ &+ \sum_{S_A} d_k\left(y_k - \hat{\mu}_k^{S_A}\right) + \sum_{S_B} d_k\left(y_k - \hat{\mu}_k^{S_B}\right),\end{aligned} \tag{4.4}$$

em que d_k é definido em (4.2) e o sobrescrito H indica que o estimador utiliza a estratégia de Hartley.

O estimador (4.4) pode ser visto como a soma dos totais estimados nos domínios por meio dos modelos $\hat{\mu}^{S_A}$ e $\hat{\mu}^{S_B}$, um vez que $0 \leq \alpha \leq 1$, adidos termos de ajuste devido a diferença entre os valores estimados pelos modelos e os valores observados nas amostras S_A e S_B.

4.1.1 Centralidade e variância aproximadas

Admitindo que os modelos $\hat{\mu}^{S_A}$ e $\hat{\mu}^{S_B}$ ajustados a partir das amostras fornecem boas aproximações para o modelos populacionais $\hat{\mu}^{A}$ e $\hat{\mu}^{B}$ tem-se que o viés do estimador $\hat{Y}^{H}_{GEREG}$ é aproximadamente nulo como argumentado a seguir.

Considere $E_k^A = y_k - \hat{\mu}_k^A$ o resíduo do elemento k do modelo proposto com base na população U_A e $E_k^B = y_k - \hat{\mu}_k^B$ o resíduo do elemento k do modelo proposto com base na população U_B. Em condições sob as quais $\hat{\mu}_k^{S_A} \approx \hat{\mu}_k^A$ para $k \in U_A$ e $\hat{\mu}_k^{S_B} \approx \hat{\mu}_k^B$ para $k \in U_B$, tem-se que

$$\begin{aligned}\mathrm{E}_{p_A}\left[\hat{Y}_{GEREG_{ab(A)}}\right] &= \sum_{k\in ab(A)}\hat{\mu}_k^{S_A} + \mathrm{E}_{p_A}\left[\sum_{k\in S_{ab(A)}}\frac{\left(y_k - \hat{\mu}_k^{S_A}\right)}{\pi_{kA}}\right] \\ &\approx \sum_{k\in ab(A)}\hat{\mu}_k^{A} + \mathrm{E}_{p_A}\left(\sum_{k\in S_{ab(A)}}\frac{E_k^A}{\pi_{kA}}\right) = Y_{ab},\end{aligned}$$

em que a última igualdade deve-se a:

$$\begin{aligned}\mathrm{E}_{p_A}\left(\sum_{k\in S_{ab(A)}}\frac{E_k^A}{\pi_{kA}}\right) &= \mathrm{E}_{p_A}\left(\sum_{k\in S_{ab(A)}}\frac{y_k}{\pi_{kA}}\right) - \mathrm{E}_{p_A}\left(\sum_{k\in S_{ab(A)}}\frac{\hat{\mu}_k^A}{\pi_{kA}}\right) \\ &= \sum_{k\in ab(A)}\frac{y_k\mathrm{E}_{p_A}(I_k)}{\pi_{kA}} - \sum_{k\in ab(A)}\frac{\hat{\mu}_k^A\mathrm{E}_{p_A}(I_k)}{\pi_{kA}} \\ &= \sum_{k\in ab(A)} y_k - \sum_{k\in ab(A)}\hat{\mu}_k^A = Y_{ab} - \sum_{k\in ab(A)}\hat{\mu}_k^A,\end{aligned}$$

com $\mathrm{E}_{p_A}(I_k) = \pi_{kA}$ a probabilidade de inclusão de primeira ordem conforme o plano amostral $p_A(\cdot)$ aplicado ao cadastro A. Os valores esperados para os demais domínios são calculados de maneira análoga e, por linearidade do valor esperado, tem-se que

$$\begin{aligned}\mathrm{E}_p\left(\hat{Y}^H_{GEREG}\right) &= \mathrm{E}_{p_A}\left(\hat{Y}_{GEREG_a}\right) + \alpha \mathrm{E}_{p_A}\left(\hat{Y}_{GEREG_{ab(A)}}\right) \\ &\quad + (1-\alpha)\,\mathrm{E}_{p_B}\left(\hat{Y}_{GEREG_{ab(B)}}\right) + \mathrm{E}_{p_B}\left(\hat{Y}_{GEREG_b}\right) \\ &\approx Y_a + \alpha Y_{ab} + (1-\alpha) Y_{ab} + Y_b = Y.\end{aligned}$$

A variância aproximada do estimador $\hat{Y}^H_{GEREG}$ pode ser escrita como função das variâncias e covariâncias aproximadas dos estimadores GEREG obtidos para cada domínio. A partir da expressão (3.2), a variância de $\hat{Y}^H_{GEREG}$ tem forma matricial dada por:

$$\mathrm{Var}_p\left(\hat{Y}^H_{GEREG}\right) = \boldsymbol{a}^T \boldsymbol{\Sigma}_{GEREG}\, \boldsymbol{a}, \tag{4.5}$$

em que os termos de

$$\boldsymbol{\Sigma}_{GEREG} = \begin{bmatrix} \sigma^2_a & \sigma_{a;ab(A)} & 0 & 0 \\ \sigma_{a;ab(A)} & \sigma^2_{ab(A)} & 0 & 0 \\ 0 & 0 & \sigma^2_b & \sigma_{b;ab(B)} \\ 0 & 0 & \sigma_{b;ab(B)} & \sigma^2_{ab(B)} \end{bmatrix} \tag{4.6}$$

são definidos por meio da aproximação expressa em (2.7). Por exemplo, a variância do estimador GEREG no domínio a é dada por

$$\sigma^2_a = \mathrm{Var}_{p_A}\left[\hat{Y}_{GEREG_a}\right] \approx \sum_{k\in U_a}\sum_{l\in U_a} (\pi_{kl} - \pi_k\pi_l)\,\frac{E^A_k}{\pi_k}\frac{E^A_l}{\pi_l}.$$

Na prática as quantidades populacionais da matriz bloco diagonal $\boldsymbol{\Sigma}_{GEREG}$ são desconhecidas, fazendo-se necessário estimar esses valores por meio das amostras. Por exemplo,

$$\begin{aligned}\hat{\sigma}^2_a &= \widehat{\mathrm{Var}}_{p_A}\left[\hat{Y}_{GEREG_a}\right] = \sum_{k\in S_a}\sum_{l\in S_a} \frac{(\pi_{klA} - \pi_{kA}\pi_{lA})}{\pi_{klA}} d_k e^A_k d_l e^A_l \\ \hat{\sigma}_{a;ab(A)} &= \widehat{\mathrm{Cov}}_{p_A}\left[\hat{Y}_{GEREG_a}, \hat{Y}_{GEREG_{ab(A)}}\right] = \sum_{k\in S_a}\sum_{l\in S_{ab(A)}} \frac{(\pi_{klA} - \pi_{kA}\pi_{lA})}{\pi_{klA}} d_k e^A_k d_l e^A_l \\ \hat{\sigma}^2_{ab(A)} &= \widehat{\mathrm{Var}}_{p_A}\left[\hat{Y}_{GEREG_{ab(A)}}\right] = \sum_{k\in S_{ab(A)}}\sum_{l\in S_{ab(A)}} \frac{(\pi_{klA} - \pi_{kA}\pi_{lA})}{\pi_{klA}} d_k e^A_k d_l e^A_l,\end{aligned}$$

em que $e^A_k = y_k - \hat{\mu}^{S_A}_k$, d_k como definido em (4.2) e π_{klA} a probabilidade de inclusão de segunda ordem para o cadastro A. A proximidade entre os modelos $\hat{\mu}^{S_A}$ e $\hat{\mu}^A$ está diretamente relacionada com a proximidade entre estimativas $\hat{\boldsymbol{\beta}}^A_{S_A}$ e $\hat{\boldsymbol{\beta}}^A$.

4.1.2 Mesmo conjunto de variáveis auxiliares

A estimação dos modelos $\hat{\mu}^{S_A}$ e $\hat{\mu}^{S_B}$ é um processo que utiliza conjuntos de informações complementares disponíveis nos cadastros A e B. Em particular, suponha que as componentes dos vetores $\boldsymbol{x}_{kA}$ e $\boldsymbol{x}_{kB}$ sejam iguais, ou seja, para cada elemento $k \in U$ as mesmas variáveis auxiliares estejam disponíveis. Se um dos cadastros identificar um subgrupo com características específicas da população-alvo, é possível que modelos ajustados em cada cadastro aproveitem melhor a correlação entre as variáveis auxiliares e a variável de interesse. Essa abordagem pode ser vista como uma adaptação para o contexto de múltiplos cadastros do estimador de regressão separado descrito na seção (2.3.1). De outro modo, pode-se ajustar um único modelo com base nas amostras S_A e S_B, simultaneamente, que considere a estrutura do plano amostral de cadastro duplo na etapa de estimação dos parâmetros do modelo.

Nesse caso, o ajuste do modelo amostral, $S = S_A \cup S_B$, envolve estimar o vetor de parâmetros $\boldsymbol{\beta}^S$ que maximize a função de pseudo log-verossimilhança da forma

$$l(\boldsymbol{\beta}) = \sum_{k \in S_A} d_k l(\mu_k | \boldsymbol{x}_{kA}) + \sum_{k \in S_B} d_k l(\mu_k | \boldsymbol{x}_{kB}), \tag{4.7}$$

em que $l(\mu_k|\cdot)$ é função de log-verossímilhança de uma distribuição pertencente a família exponencial e μ_k expresso em função do parâmetro $\boldsymbol{\beta}$. A estimativa $\hat{\boldsymbol{\beta}}^S$ obtida pela maximização da função (4.7) leva em consideração a estrutura de um plano amostral de cadastro duplo. Isso a difere da estimativa obtida ao maximizar a função de pseudo log-verossimilhança para amostras com um único cadastro.

A partir da expressão (4.4) e com o modelo amostral $\hat{\mu}^S$, o estimador para o total Y fica simplificado na forma

$$\hat{Y}^H_{GEREG} = \sum_{k \in U} \hat{\mu}^S_k + \sum_{k \in S} d_k \left(y_k - \hat{\mu}^S_k \right), \tag{4.8}$$

em que d_k é definido na expressão (4.2). Esse estimador pode ser visto como uma adaptação para o contexto de múltiplos cadastros do estimador de regressão combinado, como descrito na seção (2.3.2). A variância do estimador, expresso em (4.8), bem com o seu estimador não viesado, são os mesmos do estimador GEREG no contexto de um único cadastro, dados nas expressões (2.7) e (2.8), respectivamente.

4.2 ESTIMADOR MLGREG NO CONTEXTO DE CADASTRO DUPLO

A aplicação do estimador MLGREG para estimar proporções no contexto de cadastro duplo proposta por Molina *et al.* (2015) foi baseada na estratégia de estimação por domínio. Considere o estimador para proporção $P_i = N^{-1}\sum_{k\in U} z_{ki}$ sob a abordagem de Hartley:

$$\hat{P}_i^H = N^{-1}\left[\hat{Z}_{ai} + \alpha\hat{Z}_{ab(A)i} + (1-\alpha)\hat{Z}_{ab(B)i} + \hat{Z}_{bi}\right],$$

em que, $\hat{Z}_{ai}$ é um estimador para a contagem populacional da classe i no domínio a e analogamente para os outros domínios. A proposta de Molina *et al.* (2015) foi utilizar o estimador MLGREG para obter as estimativas por domínios considerando as variáveis auxiliares que estão disponíveis em cada cadastro.

4.2.1 Mesmo conjunto de variáveis auxiliares para os cadastros

Suponha que, para cada elemento $k \in U$ tenha-se à disposição um vetor de variáveis auxiliares $\boldsymbol{x}_k$, o qual é independente de o elemento ser identificado no cadastro A, B ou em ambos. Além disso, para cada elemento $k \in S$ observa-se o valor da variável resposta y_k que pode ser recodificada por meio do vetor de variáveis indicadoras $(z_{k1},\ldots,z_{km})$, conforme definido na expressão (2.11). Os elementos que compõem a amostra S são selecionados a partir dos cadastros A e B de maneira independente. Dessa forma, faz sentido definir a função de pseudo log-verossimilhança, como em (2.13), em função dos domínios de estimação. Assim, pode-se estimar $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^T,\ldots,\boldsymbol{\beta}_m^T\right)^T$ maximizando a função de pseudo log-verossimilhança dada por:

$$l(\boldsymbol{\beta}) = \sum_{i=1,\ldots,m}\left(\sum_{k\in S_A} d_k z_{ki}\ln\mu_{ik} + \sum_{k\in S_B} d_k z_{ki}\ln\mu_{ik}\right), \tag{4.9}$$

em que d_k é definido na expressão (4.2).

Com o modelo amostral $\hat{\mu}_{ik}^S$ segue-se a estimação para os domínios utilizando o estimador MLGREG como definido em (2.12). Assim, o estimador para ao total Z_i do domínio a é dado por

$$\hat{Z}_{ai} = \sum_{k\in a}\hat{\mu}_{ik}^S + \sum_{k\in S_a}\frac{z_{ki}-\hat{\mu}_{ik}^S}{\pi_k}.$$

Da mesma forma são obtidas as estimativas para os outros domínios. Desde que o vetor $\boldsymbol{x}_k$ seja conhecido para todos os elementos da população-alvo, o estimador para a proporção da i-ésima classe assistido por um modelo logístico multinomial pode ser escrito como

$$\begin{aligned}\hat{P}^H_{MLGREGi} &= N^{-1}\left[\sum_{k\in U}\hat{\mu}^S_{ik} + \sum_{k\in S_A} d_k\left(z_{ki}-\hat{\mu}^S_{ik}\right) + \sum_{k\in S_B} d_k\left(z_{ki}-\hat{\mu}^S_{ik}\right)\right] \\ &= N^{-1}\left[\sum_{k\in U}\hat{\mu}^S_{ik} + \sum_{k\in S} d_k\left(z_{ki}-\hat{\mu}^S_{ik}\right)\right].\end{aligned}$$

4.2.2 Diferentes conjuntos de variáveis auxiliares para os cadastros

Na seção (4.2.1) foi feita a suposição que as informações auxiliares estão disponíveis a nível de população. Contudo uma suposição razoável é de que informações auxiliares estejam disponíveis também a nível de cadastro. Para cada $k \in A$ tem-se o vetor auxiliar $\boldsymbol{x}_{kA}$ e para cada $k \in B$ tem-se o vetor auxiliar $\boldsymbol{x}_{kB}$, com as componentes de $\boldsymbol{x}_{kA}$ e $\boldsymbol{x}_{kB}$ diferentes. Nesse caso, pode-se usar a informação auxiliar disponível para ajustar um modelo logístico multinomial em cada cadastro separadamente. Assim, para cada $k \in A$, tem-se

$$\hat{\mu}^{S_A}_{ik} = \frac{\exp\left(\boldsymbol{x}^T_{kA}\hat{\boldsymbol{\beta}}^A_{iS_A}\right)}{\sum_{r=1,\ldots,m}\exp\left(\boldsymbol{x}^T_{kA}\hat{\boldsymbol{\beta}}^A_{iS_A}\right)}, \tag{4.10}$$

onde $\hat{\boldsymbol{\beta}}^A_{S_A} = \left(\hat{\boldsymbol{\beta}}^T_{1S_A},\ldots,\hat{\boldsymbol{\beta}}^T_{mS_A}\right)^T$ é a estimativa para o parâmetro $\boldsymbol{\beta}^A$ obtida por meio da maximização da função de pseudo log-verossimilhança dada por

$$l(\boldsymbol{\beta}^A) = \sum_{i=1,\ldots,m}\left(\sum_{k\in S_A} d_k z_{ki}\ln\mu^A_{ik}\right).$$

De maneira análoga, são obtidos $\hat{\mu}^{S_B}_{ik}$. O estimador para a proporção da i-ésima classe assistido por um modelo logístico multinomial quando as variáveis auxiliares são diferentes entre os cadastros é dado por:

$$\begin{aligned}\hat{P}^H_{MLGREGi} = N^{-1}\Bigg[&\sum_{k\in a}\hat{\mu}^{S_A}_{ik} + \alpha\sum_{k\in ab(A)}\hat{\mu}^{S_A}_{ik} + (1-\alpha)\sum_{k\in ab(B)}\hat{\mu}^{S_B}_{ik} + \sum_{k\in b}\hat{\mu}^{S_B}_{ik} \\ &+ \sum_{k\in S_A} d_k\left(y_k-\hat{\mu}^{S_A}_k\right) + \sum_{k\in S_B} d_k\left(y_k-\hat{\mu}^{S_B}_k\right)\Bigg],\end{aligned} \tag{4.11}$$

em que d_k é definido na expressão (4.2).

4.3 ESTIMADOR GEREG SOB A ABORDAGEM DE MULTIPLICIDADE

Como discutido no capítulo 3, a abordagem de multiplicidade não utiliza como critério a minimização da variância para escolher os pesos atribuídos aos totais estimados em cada domínio. De fato, em planos de cadastro duplo, temos que $\alpha = \frac{1}{2}$ sob a abordagem de multiplicidade. Em contrapartida, não é necessário identificar os domínios de estimação, apenas contabilizar o valor de m_k, equivalente a quantos cadastros o elemento $k \in U$ é identificado. Além disso, se planos amostrais independentes são aplicados em cada cadastro, a variância do estimador do total Y é simplificada como a soma das variâncias em cada cadastro.

Considere d_k^{MS} o fator de expansão definido sob a abordagem de multiplicidade, tal que,

$$d_k^{MS} = \begin{cases} \frac{1}{m_k \pi_{kA}}, & \text{se } k \in S_A, \\ \frac{1}{m_k \pi_{kB}}, & \text{se } k \in S_B, \end{cases} \tag{4.12}$$

em que m_k é o fator de multiplicidade simples. Nesse caso, a estimativa de pseudo máxima-verossimilhança para $\boldsymbol{\beta}^A$ em S_A é dada por

$$\hat{\boldsymbol{\beta}}_{S_A}^A = \arg\max_{\boldsymbol{\beta}^A} \sum_{k \in S_A} d_k^{MS}\, l(\boldsymbol{\mu}_k). \tag{4.13}$$

De maneira análoga é obtida a estimativa $\hat{\boldsymbol{\beta}}_{S_B}^B$. Com os modelos amostrais $\hat{\mu}^{S_A}$ e $\hat{\mu}^{S_B}$, com forma expressa em (4.1), estima-se os valores da variável de interesse não observados nas amostras S_A e S_B, respectivamente. Dessa forma, o estimador sob a abordagem de multiplicidade para o total Y é escrito por

$$\hat{Y}_{GEREG}^{MS} = \sum_{k \in U_A} \frac{\hat{\mu}_k^{S_A}}{m_k} + \sum_{k \in S_A} d_k^{MS}(y_k - \hat{\mu}_k^{S_A}) + \sum_{k \in U_B} \frac{\hat{\mu}_k^{S_B}}{m_k} + \sum_{k \in S_B} d_k^{MS}(y_k - \hat{\mu}_k^{S_B}). \tag{4.14}$$

4.3.1 Centralidade e variância aproximadas

Admitindo que $\hat{\mu}_k^{S_A} \approx \hat{\mu}_k^A$ para $k \in U_A$ e $\hat{\mu}_k^{S_B} \approx \hat{\mu}_k^B$ para $k \in U_B$, então o estimador GEREG sob a abordagem de multiplicidade tem viés aproximadamente nulo. A partir da expressão (4.14), o valor esperado de $\hat{Y}_{GEREG}^{MS}$ sob um plano amostral de cadastro duplo é dado por:

$$\begin{aligned}\mathrm{E}_p\left(\hat{Y}^{MS}_{GEREG}\right) &= \sum_{k\in U_A}\frac{\hat{\mu}_k^{S_A}}{m_k} + \mathrm{E}_{p_A}\left[\sum_{k\in S_A} d_k^{MS}(y_k-\hat{\mu}_k^{S_A})\right] + \sum_{k\in U_B}\frac{\hat{\mu}_k^{S_B}}{m_k} + \mathrm{E}_{p_B}\left[\sum_{k\in S_B} d_k^{MS}(y_k-\hat{\mu}_k^{S_B})\right] \\ &\approx \sum_{k\in U_A}\frac{y_k}{m_k} + \sum_{k\in U_B}\frac{y_k}{m_k} = Y,\end{aligned}$$

em que,

$$\mathrm{E}_{p_A}\left[\sum_{k\in S_A} d_k^{MS}(y_k-\hat{\mu}_k^{S_A})\right] = \sum_{k\in U_A}\frac{y_k\mathrm{E}_{p_A}(I_k)}{m_k\pi_{kA}} - \sum_{k\in U_A}\frac{\hat{\mu}_k^{S_A}\mathrm{E}_{p_A}(I_k)}{m_k\pi_{kA}} \approx \sum_{k\in U_A}\frac{y_k}{m_k} - \sum_{k\in U_A}\frac{\hat{\mu}_k^{A}}{m_k}, \text{ e}$$

$$\mathrm{E}_{p_B}\left[\sum_{k\in S_B} d_k^{MS}(y_k-\hat{\mu}_k^{S_B})\right] = \sum_{k\in U_B}\frac{y_k\mathrm{E}_{p_B}(I_k)}{m_k\pi_{kB}} - \sum_{k\in U_B}\frac{\hat{\mu}_k^{S_B}\mathrm{E}_{p_B}(I_k)}{m_k\pi_{kB}} \approx \sum_{k\in U_B}\frac{y_k}{m_k} - \sum_{k\in U_B}\frac{\hat{\mu}_k^{B}}{m_k}.$$

Se planos amostrais independentes são aplicados em cada cadastro, a partir da equação (2.7), a variância aproximada do estimador $\hat{Y}^{MS}_{GEREG}$ é a soma das variâncias dos estimadores GEREG para as populações U_A e U_B. Assim,

$$\begin{aligned}\mathrm{Var}_p\left(\hat{Y}^{MS}_{GEREG}\right) &= \mathrm{Var}_{p_A}\left(\hat{Y}_{GEREG_A}\right) + \mathrm{Var}_{p_B}\left(\hat{Y}_{GEREG_B}\right) \\ &\approx \sum_{k\in U_A}\sum_{l\in U_A}\left(\pi_{klA}-\pi_{kA}\pi_{lA}\right) d_k^{MS}E_k^A d_l^{MS}E_l^A + \\ &\quad \sum_{k\in U_B}\sum_{l\in U_B}\left(\pi_{klB}-\pi_{kB}\pi_{lB}\right) d_k^{MS}E_k^B d_l^{MS}E_l^B,\end{aligned}$$

com d_k^{MS} como definido em (4.12). Por fim, o estimador da variância de $\hat{Y}^{MS}_{GEREG}$ obtido a partir da equação (2.8) é dado por

$$\begin{aligned}\widehat{\mathrm{Var}}_p\left(\hat{Y}^{MS}_{GEREG}\right) &= \sum_{k\in S_A}\sum_{l\in S_A}\frac{\left(\pi_{klA}-\pi_{kA}\pi_{lA}\right)}{\pi_{klA}}\frac{e_k^A}{m_k\pi_{kA}}\frac{e_l^A}{m_l\pi_{lA}} + \\ &\quad \sum_{k\in S_B}\sum_{l\in S_B}\frac{\left(\pi_{klB}-\pi_{kB}\pi_{lB}\right)}{\pi_{klB}}\frac{e_k^B}{m_k\pi_{kA}}\frac{e_l^B}{m_l\pi_{lB}},\end{aligned}$$

em que m_k é o fator de multiplicidade simples.

4.3.2 Mesmo conjunto de variáveis auxiliares

O estimador $\hat{Y}^{MS}_{GEREG}$, expresso em (4.14), bem como os resultados derivados na seção (4.3.1), consideram que modelos são ajustados com base nas amostras S_A e S_B utilizando as variáveis auxiliares disponíveis em cada cadastro. Supondo que as informações auxiliares disponíveis sejam as mesmas para ambos os cadastros, é possível ajustar um único modelo $\hat{\mu}^S$ com base nas amostras S_A e S_B, simultaneamente, levando em consideração a estrutura do plano amostral de cadastro duplo. Nesse caso, para todo $k \in U$ o vetor de variáveis auxiliares

$\boldsymbol{x}_k$ está disponível e que para $k \in S = S_A \cup S_B$ o valor da variável de interesse y_k é observado. Os parâmetros do modelo são estimados maximizando a função de pseudo log-verossimilhança amostral, ou seja,

$$\hat{\boldsymbol{\beta}}^S = \arg\max_{\boldsymbol{\beta}} \left[\sum_{k \in S_A} d_k^{MS}\, l(\boldsymbol{\mu}_k) + \sum_{k \in S_B} d_k^{MS}\, l(\boldsymbol{\mu}_k) \right],$$

em que $l(\boldsymbol{\mu}_k)$ uma função de log-verossimilhança de uma distribuição de probabilidade pertencente à família exponencial definida em (2.4). Assim, o estimador (4.14) para o total Y pode ser reescrito como

$$\hat{Y}_{GEREG}^{MS} = \sum_{k \in U} \hat{\mu}_k^S + \sum_{k \in S} d_k^{MS}(y_k - \hat{\mu}_k^S).$$

O estimador $\hat{Y}_{GEREG}^{MS}$ é semelhante ao estimador obtido a partir da abordagem de domínios, dado em (4.8), mas a estratégia de estimação por multiplicidade simples fica implícita no fator de expansão d_k^{MS}.

4.4 ESTIMADOR GEREG PARA MÚLTIPLOS CADASTROS

Uma das vantagens que a abordagem de multiplicidade apresenta em relação a abordagem por domínios é a naturalidade em que planos com vários cadastros podem ser utilizados. Por essa razão, os resultados apresentados para planos amostrais de cadastro duplo podem ser facilmente generalizados para Q cadastros. Considere a generalização fator de expansão d_k^{MS} em função dos planos amostrais aplicados em cada cadastro, tal que, $d_k^{MS} = \frac{1}{m_k \pi_{k(q)}}$, se $k \in S_q$. O estimador para o total Y sob abordagem de multiplicidade é dado por:

$$\hat{Y}_{GEREG}^{MS} = \sum_{q=1}^{Q} \left[\sum_{k \in U_q} \frac{\hat{\mu}_k^{S_q}}{m_k} + \sum_{k \in S_q} d_k^{MS} \left(y_k - \hat{\mu}_k^{S_q} \right) \right]. \tag{4.15}$$

4.4.1 Centralidade e variância aproximadas

Sob a condição $\hat{\mu}_k^{S_q} \approx \hat{\mu}_k^q, \forall q \in \{1,\ldots,Q\}$, mostra-se que o estimador (4.15) é aproximadamente não viesado:

$$\begin{aligned}
\mathrm{E}_p\left(\hat{Y}_{GEREG}^{MS}\right) &= \sum_{q=1}^{Q} \mathrm{E}_{p_q}\left[\sum_{k\in U_q}\frac{\hat{\mu}_k^{S_q}}{m_k} + \sum_{k\in S_q} d_k^{MS}\left(y_k - \hat{\mu}_k^{S_q}\right)\right] \\
&= \sum_{q=1}^{Q}\left\{\sum_{k\in U_q}\frac{\hat{\mu}_k^{S_q}}{m_k} + \mathrm{E}_{p_q}\left[\sum_{k\in S_q} d_k^{MS}\left(y_k - \hat{\mu}_k^{S_q}\right)\right]\right\} \\
&\approx \sum_{q=1}^{Q}\sum_{k\in U_q}\frac{y_k}{m_k} = Y,
\end{aligned}$$

em que a última passagem é validada a partir da definição do fator de multiplicidade simples, $\sum_{q=1}^{Q} I_{[k\in U_q]} = m_k$, e

$$\mathrm{E}_{p_q}\left[\sum_{k\in S_q} d_k^{MS}\left(y_k - \hat{\mu}_k^{S_q}\right)\right] = \sum_{k\in U_q}\frac{y_k \mathrm{E}_{p_q}(I_k)}{m_k\pi_{k(q)}} - \sum_{k\in U_q}\frac{\hat{\mu}_k^{S_q}\mathrm{E}_{p_q}(I_k)}{m_k\pi_{k(q)}} \approx \sum_{k\in U_q}\frac{y_k}{m_k} - \sum_{k\in U_q}\frac{\hat{\mu}_k^{q}}{m_k}.$$

Da expressão (2.8), um estimador para a variância de $\hat{Y}_{GEREG}^{MS}$ é dado por:

$$\widehat{\mathrm{Var}}_p\left(\hat{Y}_{GEREG}^{MS}\right) = \sum_{q=1}^{Q}\sum_{k\in S_q}\frac{\left[\pi_{kl(q)} - \pi_{k(q)}\pi_{l(q)}\right]}{\pi_{kl(q)}}\frac{e_k^q}{m_k\pi_{k(q)}}\frac{e_l^q}{m_l\pi_{l(q)}},$$

em que $e_k^q = y_k - \mu_k^{S_q}$.

4.4.2 Expressão alternativa para o estimador GEREG

O estimador $\hat{Y}_{GEREG}^{MS}$ possui fácil interpretação em função das médias aritméticas dos valores estimado da variável resposta. Por exemplo, se o elemento $k \in U$ é identificado através de m_k cadastros, então a variável de interesse y_k pode ser estimada por meio de m_k modelos. Além disso, para os valores observados na amostra $S = S_1 \cup \ldots \cup S_Q$, são calculados os termos de ajustes. Dessa forma, a contribuição do elemento k para o total estimado de Y pode ser escrita por $\hat{y}_k = \hat{\mu}_k^0 + \hat{e}_k^0$, em que,

$$\hat{\mu}_k^0 = \frac{1}{m_k}\sum_{q=1}^{Q}\hat{\mu}_k^{S_q} I_{[k\in U_q]} \quad \text{e} \quad \hat{e}_k^0 = \frac{1}{m_k}\sum_{q=1}^{Q}\left(\frac{y_k - \hat{\mu}_k^{S_q}}{\pi_{k(q)}}\right) I_{[k\in S_q]},$$

com $I_{[k\in U_q]}$ uma variável indicadora de $k \in U_q$ e $I_{[k\in S_q]}$ uma variável indicadora de $k \in S_q$. Dessa forma, o estimador $\hat{Y}_{GEREG}^{MS}$ pode ser reescrito na seguinte forma

$$\hat{Y}_{GEREG}^{MS} = \sum_{k\in U}\left(\hat{\mu}_k^0 + \hat{e}_k^0\right). \tag{4.16}$$

Na expressão acima, para $m_k > 1$, $\hat{\mu}_k^0$ pode ser entendido como a média dos valores estimados para y_k com base em m_k modelos amostrais.

4.5 ESTIMADOR MLGREG PARA CADASTRO TRIPLO

A aplicação do estimador MLGREG no contexto de pesquisas amostrais de cadastro duplo descrita nesse capítulo considerou a estratégia de estimação por domínio. Uma contribuição desta tese é utilizar o estimador MLGREG sob a abordagem de multiplicidade em planos amostrais com dois ou mais cadastros.

No caso em que dois cadastros estão em uso, adotar a estratégia de estimação por multiplicidade simples equivale a fazer $\alpha = \frac{1}{2}$ nos estimadores (4.2.1) e (4.11). Inicialmente, será considerado o cenário onde três cadastros são utilizados simultaneamente para obter cobertura completa da população de interesse.

Suponha que além dos cadastros A e B já considerados nos resultados descritos anteriormente, tenha-se outro cadastro, denotado por C, que identifique algum subgrupo da população-alvo. Ficam definidos através da sobreposição desses três cadastros sete domínios de estimação. Além disso, amostras são selecionadas independentemente em cada cadastro. Considere ainda d_k^{MS} um fator de expansão, tal que,

$$d_k^{MS} = \begin{cases} \frac{1}{m_k \pi_{kA}}, & \text{se } k \in S_A \\ \frac{1}{m_k \pi_{kB}}, & \text{se } k \in S_B \\ \frac{1}{m_k \pi_{kC}}, & \text{se } k \in S_C \end{cases}, \tag{4.17}$$

em que π_{kA}, π_{kB} e π_{kC} são as probabilidades de inclusão de primeira ordem para os planos $P_A(\cdot)$, $P_B(\cdot)$ e $P_C(\cdot)$, respectivamente, aplicados em cada cadastro. Considere o estimador para a proporção da i-ésima classe $P_i = N^{-1} \sum_{k \in U} z_{ki}$ sob a abordagem de multiplicidade simples

$$\hat{P}_{MSi} = N^{-1} \left(\hat{Z}_{Ai}^* + \hat{Z}_{Bi}^* + \hat{Z}_{Ci}^* \right), \tag{4.18}$$

em que o subscrito MS faz alusão abordagem de multiplicidade simples e $\hat{Z}_{Ai}^* = \sum_{k \in S_A} d_k^{MS} z_{ki}$, e analogamente para $\hat{Z}_{Bi}^*$ e $\hat{Z}_{Ci}^*$. Assim como na seção 4.2, propomos utilizar um modelo logístico multinomial para obter as estimativas quando informações complementares estão disponíveis. Inicialmente, consideramos que um mesmo conjunto de informações auxiliares está disponível para toda população.

4.5.1 Mesmo conjunto de variáveis auxiliares para os cadastros

Suponha que para cada elemento $k \in U$ esteja disponível um vetor de variáveis auxiliares x_k, cujas componentes são iguais para os três cadastros. Uma amostra S é selecionada

a partir dos cadastros A, B e C, com $S = S_A \cup S_B \cup S_C$. e, além disso, para cada $k \in S$ observa-se o valor da variável resposta y_k que é recodificada no vetor de variáveis indicadoras $(z_{k1}, \ldots, z_{km})$. A função de pseudo log-verossimilhança que leva em consideração a estrutura do plano amostral com três cadastros pode ser definida da mesma forma que (4.9). Assim, estima-se $\boldsymbol{\beta}$ maximizando a função dada por:

$$l(\boldsymbol{\beta}) = \sum_{i=1,\ldots,m} \left(\sum_{k \in S_A} d_k^{MS} z_{ki} \ln \mu_{ik} + \sum_{k \in S_B} d_k^{MS} z_{ki} \ln \mu_{ik} + \sum_{k \in S_C} d_k^{MS} z_{ki} \ln \mu_{ik} \right). \tag{4.19}$$

Um modelo amostral geral $\hat{\mu}_{ik}^S$ pode ser proposto utilizando a estimativa $\hat{\boldsymbol{\beta}}^S$ obtida a partir da maximização de (4.19). Novamente, o processo de estimação de $\hat{\boldsymbol{\beta}}^S$ leva em consideração a estrutura de um plano amostral que, nesse caso, é de cadastro triplo. Utilizando o fator de expansão dado em (4.17), pode-se escrever as componentes da equação (4.18) em função do fator de multiplicidade e assistido por um modelo logístico multinomial. Assim,

$$\hat{Z}_{Ai}^{LM} = \sum_{k \in A} \frac{\hat{\mu}_{ik}^S}{m_k} + \sum_{k \in S_A} d_k^{MS} \left(z_{ki} - \hat{\mu}_{ik}^S \right),$$

em que o sobrescrito *LM* indica que a estimação foi assistida por um modelo Logístico Multinomial. Da maneira análoga, são obtidas as estimativas Z_{Bi}^{LM} e Z_{Ci}^{LM}. Dessa forma, pode-se definir o estimador expresso por:

$$\hat{P}_{MSi}^{LM} = N^{-1} \left[\sum_{k \in U} \frac{\hat{\mu}_{ik}^S}{m_k} + \sum_{k \in S} d_k^{MS} \left(z_{ki} - \hat{\mu}_{ik}^S \right) \right]$$

como o estimador multiplicidade simples assistido por modelo logístico multinomial para planos amostrais de três cadastros.

4.5.2 Diferentes conjuntos de variáveis auxiliares para os cadastros

Suponha agora que informações auxiliares estão disponíveis a nível de cadastro. Assim, para cada $k \in A$ tem-se o vetor auxiliar $\boldsymbol{x}_{kA}$, para cada $k \in B$ o vetor auxiliar $\boldsymbol{x}_{kB}$ e para cada $k \in C$ o vetor auxiliar $\boldsymbol{x}_{kC}$, com as componentes de $\boldsymbol{x}_{kA}$, $\boldsymbol{x}_{kB}$ e $\boldsymbol{x}_{kC}$ possivelmente diferentes. Nesse caso, modelos semelhantes ao dado em (4.10) podem ser ajustados com base nas amostras selecionadas em cada cadastro. Considere a amostra S_A selecionada a partir do cadastro A. A função de pseudo log-verossimilhança, sob a abordagem de multiplicidade, é dado por:

$$l(\boldsymbol{\beta}_A) = \sum_{i=1,\ldots,m} \left\{ \sum_{k \in S_A} d_k^{MS} z_{ki} \ln \mu_{ik} \right\}. \tag{4.20}$$

O modelo logístico multinomial para o cadastro A obtido por meio de S_A, é expresso por:

$$\hat{\mu}_{ik}^{S_A} = \frac{\exp\left(\boldsymbol{x}_k^T \hat{\boldsymbol{\beta}}_{iS_A}^{A}\right)}{\sum_{r=1,\ldots,m} \exp\left(\boldsymbol{x}_k^T \hat{\boldsymbol{\beta}}_{iS_A}^{A}\right)},$$

onde $\hat{\boldsymbol{\beta}}_{S_A}^{A} = \left(\hat{\boldsymbol{\beta}}_{1S_A}^{T}, \ldots, \hat{\boldsymbol{\beta}}_{mS_A}^{T}\right)^T$ é a estimativa para o parâmetro $\boldsymbol{\beta}^A$ que maximiza a função de pseudo log-verossimilhança (4.20). Escrevendo as componentes da equação (4.18) em função do fator de multiplicidade e do modelo $\hat{\mu}_{ik}^{A}$ tem-se:

$$\hat{Z}_{Ai}^{LM} = \sum_{k \in A} \frac{\hat{\mu}_{ik}^{S_A}}{m_k} + \sum_{k \in S_A} d_k^{MS} \left(z_{ki} - \hat{\mu}_{iS}^{S_A}\right).$$

Da maneira análoga, são obtidas as estimativas para Z_{Bi}^{LM} e Z_{Ci}^{LM}. O estimador para a proporção da i-ésima classe assistido por um modelo logístico multinomial, sob a abordagem de multiplicidade, quando as variáveis auxiliares são diferentes entre os cadastros é dado por:

$$\begin{aligned} \hat{P}_{SMi}^{LM} = N^{-1} \Bigg[& \sum_{k \in A} \frac{\hat{\mu}_{ik}^{S_A}}{m_k} + \sum_{k \in B} \frac{\hat{\mu}_{ik}^{S_B}}{m_k} + \sum_{k \in C} \frac{\hat{\mu}_{ik}^{S_C}}{m_k} + \\ & \sum_{k \in S_A} d_k^{MS} \left(z_{ki} - \hat{\mu}_{ik}^{S_A}\right) + \sum_{k \in S_B} d_k^{MS} \left(z_{ki} - \hat{\mu}_{ik}^{S_B}\right) + \sum_{k \in S_C} d_k^{MS} \left(z_{ki} - \hat{\mu}_{ik}^{S_C}\right) \Bigg]. \end{aligned} \tag{4.21}$$

5 ESTIMADOR GEREG SIMPLIFICADO

Os estimadores de regressão discutidos nos capítulos anteriores consistem na soma dos valores estimados a partir de um modelo amostral para a população U, juntamente com um termo de ajuste que leva em consideração a diferença entre os valores observados na amostra e os valores estimados pelo modelo. Foram demostradas as condições nas quais o termo de ajuste se anula quando um único cadastro está em uso, tanto para o estimador de regressão linear geral (SÄRNDAL; SWENSSON; WRETMAN, 2003, p. 230) quanto para o estimador de regressão linear generalizado foco desse trabalho (RONDON; VANEGAS; FERRAZ, 2012).

Considere o estimador GEREG para planos de cadastro único expresso em (2.6) com o modelo amostral assumindo estrutura igual a (2.5). Uma condição suficiente para que $\sum_S (y_k - \hat{\mu}_k^S)/\pi_k = 0$ é que exista um vetor constante $\boldsymbol{a}^T = (a_1, \ldots, a_J)$ tal que

$$\boldsymbol{H}\boldsymbol{a} = [\ \boldsymbol{h}_1 \quad \cdots \quad \boldsymbol{h}_J\ ]\boldsymbol{a} = \mathbf{1},$$

em que $\boldsymbol{h}_j^T = (h_{1j}, \ldots, h_{Nj})$ com $h_{kj} = \phi_k \frac{\partial \theta_k}{\partial \eta_k} x_{kj}$ para $j = 1, \ldots, J$ e $\mathbf{1}$ é um vetor unitário de dimensão J. Os índices N e J representam o tamanho populacional e o número de variáveis auxiliares, respectivamente. Satisfazem essa condição modelos amostrais especificados com homocedasticidade no parâmetro de dispersão, $\phi_k = \phi$ para $k \in U$, e componente sistemática com intercepto e função de ligação canônica (RONDON; VANEGAS; FERRAZ, 2012).

5.1 ESTIMADOR SIMPLIFICADO

Resultado equivalente ao supracitado pode ser provado no contexto de múltiplos cadastros, onde a influência do plano amostral sobre os estimadores do tipo regressão está na ponderação aplicada a função de pseudo log-verossimilhança.

Resultado 1. *Considere o estimador GEREG para planos amostrais de múltiplos cadastros e com modelos amostrais assumindo a estrutura ξ expressa em (2.5). Se existir vetor constante $\boldsymbol{a}_q^T = (a_{q1}, \ldots, a_{qJ})$ tal que $\boldsymbol{H}_{\boldsymbol{q}}\boldsymbol{a}_{\boldsymbol{q}} = [\ \boldsymbol{h}_{a1} \quad \cdots \quad \boldsymbol{h}_{aJ}\ ]\boldsymbol{a}_{\boldsymbol{q}} = \mathbf{1}$, em que $\boldsymbol{h}_{qj}^T = (h_{1j}^q, \ldots, h_{N_q j}^q)$ com $h_{kj}^q = \phi_k \frac{\partial \theta_k}{\partial \eta_k} x_{kjq}$ para $j = 1, \ldots, J_q$ e $\mathbf{1}$ é um vetor unitário de dimensão J_q, então,*

$$\sum_{k \in S_q} d_k^{MS} \left(y_k - \hat{\mu}_k^{S_q} \right) = 0,$$

com $d_k^{MS} = \frac{1}{m_k \pi_{k(q)}}$, se $k \in S_q$.

Demonstração. As estimativas $\hat{\boldsymbol{\beta}}^{q}_{S_q}$ satisfazem a $\boldsymbol{U}_{S_q}(\hat{\boldsymbol{\beta}}^{q}_{S_q}) = \mathbf{0}$, onde $\boldsymbol{U}_{S_q}$ é a função escore obtida a partir da função de pseudo log-verossimilhança em S_q. Sob a abordagem de multiplicidade simples, tem-se que, $\boldsymbol{U}_{S_q j}(\hat{\boldsymbol{\beta}}^{q}_{S_q}) = \boldsymbol{h}^{T}_{S_q j}\boldsymbol{e}^{d^{MS}} = 0$, em que $\boldsymbol{h}^{T}_{S_q j} = (h^{q}_{1j},\ldots,h^{q}_{n_q j})$ e $\boldsymbol{e}^{d^{MS}} = (e^{d^{MS}_1},\ldots,e^{d^{MS}_{n_q}})$ vetores n_q-dimensionais, com $h^{q}_{kj} = \phi_k \frac{\partial \theta_k}{\partial \eta_k} x_{kj(q)}$ para $j = 1,\ldots,J_q$ e $e^{d^{MS}_k} = d^{MS}_k(y_k - \hat{\mu}^{S_q}_k)$ para $k \in S_q$. Se existe $\boldsymbol{a}_q$ tal que $\sum_{j=1}^{J_q} a_{qj}\boldsymbol{h}_{qj} = \mathbf{1}_{N_q}$, então para S_q, $\sum_{j=1}^{J_q} a_{qj}\boldsymbol{h}_{S_q j} = \mathbf{1}_{n_q}$. Assim, $\sum_{k\in S_q} d^{MS}_k \left(y_k - \hat{\mu}^{S_q}_k\right) = \mathbf{1}^{T}_{n_q}\boldsymbol{e}^{d^{MS}} = \left(\sum_{j=1}^{J_q} a_{qj}\boldsymbol{h}_{S_q j}\right)^{T}\boldsymbol{e}^{d^{MS}} = \sum_{j=1}^{J_q} a_{qj}\boldsymbol{h}^{T}_{S_q j}\boldsymbol{e}^{d^{MS}} = 0$ □

Esse resultado é útil para simplificar os estimadores GEREG quando os modelos aplicados em cada cadastro utilizam a função de ligação canônica, intercepto e homocedasticidade do parâmetro de dispersão.

5.1.1 Estimador simplificado sob a abordagem de domínios

O resultado 1 também pode ser provado para a abordagem de estimação por domínios substituindo d^{MS}_k por um fator de expansão correspondente. Em planos de cadastro duplo, por exemplo, tem-se que $\sum_{k\in S_A} d_k \left(y_k - \hat{\mu}^{S_A}_k\right) = 0$ e $\sum_{k\in S_B} d_k \left(y_k - \hat{\mu}^{S_B}_k\right) = 0$, com d_k expresso em (4.2). Além disso, considerando que em cada cadastro um modelo que satisfaz a condição do resultado 1 esteja especificado, então o estimador GEREG sob a abordagem de domínios é expresso por

$$\hat{Y}^{H^*}_{GEREG} = \sum_{k\in a}\hat{\mu}^{S_A}_k + \sum_{k\in ab}\left[\alpha\hat{\mu}^{S_A}_k + (1-\alpha)\hat{\mu}^{S_B}_k\right] + \sum_{k\in b}\hat{\mu}^{S_B}_k. \tag{5.1}$$

Estimadores simplificados ainda podem ser expressos em função dos valores observados por meio da amostra. Considere $S = S_A \cup S_B$, uma forma alternativa para o estimador GEREG é dada por

$$\hat{Y}^{H^*}_{GEREG} = \sum_{k\in S} y_k + \sum_{k\in a^*}\hat{\mu}^{S_A}_k + \sum_{k\in ab^*}\left[\alpha\hat{\mu}^{S_A}_k + (1-\alpha)\hat{\mu}^{S_B}_k\right] + \sum_{k\in b^*}\hat{\mu}^{S_B}_k,$$

em que a^*, b^* e ab^* são os conjuntos de elementos identificados no domínio a, b e ab, respectivamente, mas não observados na amostra $S = S_A \cup S_B$. Contudo, o estimador escrito dessa forma não é assintoticamente centrado considerando a estrutura do plano amostral.

5.2 ESTIMADOR SIMPLIFICADO PARA MÚLTIPLOS CADASTROS

Como discutido nos capítulos 3 e 4, no contexto de planos amostrais com mais de dois cadastros há um aumento da complexidade de estimação atrelada a utilização da abordagem de domínios. Nesse trabalho, considerou-se apenas a abordagem de estimação de multiplicidade simples em planos amostrais com um maior número de cadastros. Os resultados derivados a seguir utilizam a expressão para o estimador $\hat{Y}^{MS}_{GEREG}$, fornecida em (4.15), como referência.

5.2.1 Estimador simplificado sob a abordagem de multiplicidade

Considere Q cadastros cobrindo completamente a polução de interesse e que planos amostrais independentes sejam aplicados em cada um deles. Suponha ainda que os modelos amostrais aplicados em cada cadastro satisfaçam a condição do resultado 1. Assim, o estimador $\hat{Y}^{MS}_{GEREG}$ fica simplificado por

$$\hat{Y}^{MS^*}_{GEREG} = \sum_{q=1}^{Q} \sum_{k \in U_q} \frac{\hat{\mu}_k^{S_q}}{m_k}, \tag{5.2}$$

com o caso particular de dois cadastros dado por

$$\hat{Y}^{MS}_{GEREG} = \sum_{k \in U_A} \frac{\hat{\mu}_k^{S_A}}{m_k} + \sum_{k \in U_B} \frac{\hat{\mu}_k^{S_B}}{m_k}.$$

Por sua vez, o estimador $\hat{Y}^{MS^*}_{GEREG}$ pode ser escrito considerando os valores que são observados em $S = \bigcup_{q=1}^{Q} S_q$:

$$\hat{Y}^{MS^*}_{GEREG} = \sum_{k \in S} y_k + \sum_{q=1}^{Q} \sum_{k \in U_q^*} \frac{\hat{\mu}_k^{S_q}}{m_k},$$

em que U_q^*, $q = \ldots, Q$, são os elementos da população identificados no cadastro q, mas não observado na amostra S. Novamente, o estimador expresso dessa forma não é consistente considerando o plano amostral.

5.2.2 Expressão alternativa para o estimador simplificado

Na seção (4.4.2) foi apresentada outra maneira de escrever o estimador GEREG. A contribuição do k-ésimo elemento para o total estimado $\hat{Y}$ é dado por $\hat{y}_k = \hat{\mu}_k^0 + \hat{e}_k^0$, onde $\hat{\mu}_k^0$ é a média dos valores estimados para y_k com base em m_k modelos e $\hat{e}_k^0$ um termo de ajuste agregado com base no estrutura do plano amostral. Uma vez que a condição do resultado 1 é satisfeita nos

modelos postulados em todos os cadastros, tem-se que $\sum_{k\in U}\hat{e}_k^0 = 0$, pois

$$\sum_{k\in U}\hat{e}_k^0 = \sum_{k\in U}\left\{\frac{1}{m_k}\sum_{q=1}^{Q}\left(\frac{y_k-\hat{\mu}_k^{S_q}}{\pi_{k(q)}}\right)I_{[k\in S_q]}\right\} = \sum_{q=1}^{Q}\left[\sum_{k\in S_q}\frac{1}{m_k}\left(\frac{y_k-\hat{\mu}_k^{S_q}}{\pi_{k(q)}}\right)\right]$$
$$= \sum_{q=1}^{Q}\left[\sum_{k\in S_q} d_k^{MS}\left(y_k-\hat{\mu}_k^{S_q}\right)\right] = 0.$$

Assim, o estimador GEREG para o total Y sob a abordagem de multiplicidade fica expresso por

$$\hat{Y}_{GEREG}^{MS^*} = \sum_{k\in U}\hat{\mu}_k^0, \tag{5.3}$$

com $\hat{\mu}_k^0 = \frac{1}{m_k}\sum_{q=1}^{Q}\hat{\mu}_k^{S_q}I_{[k\in U_q]}$.

6 AVALIAÇÃO NUMÉRICA

Os desempenhos dos estimadores propostos nos capítulos 4 e 5 foram avaliados numericamente através de simulações de Monte Carlo. As simulações foram divididas em duas partes. Inicialmente, foram geradas populações fictícias a partir de distribuições de probabilidades pertencentes à família exponencial de distribuições. Nesse caso, foram considerados três distribuições: gama, Poisson e Bernoulli. Por fim, como uma proposta de aplicação, os estimadores GEREG foram avaliados no contexto de pesquisas amostrais em agropecuária.

Nesse estudo numérico, foram comparados os desempenhos de estimadores em planos amostrais de três cadastros considerando a estratégia de estimação por multiplicidade simples. Em planos amostrais de cadastro duplo, foram utilizadas as estratégias de estimação por domínios e por multiplicidade. Todas as etapas desse estudo foram realizadas utilizando o ambiente de *software* livre para computação gráfica e estatística R, versão 4.3.2 (R CORE TEAM, 2017).

6.1 PROTOCOLO DE SIMULAÇÃO

O primeiro objetivo em nossa avaliação numérica foi construir a partir de distribuições de probabilidade pertencentes à família exponencial populações fictícias de tamanho 2.000 compostas por uma variável de interesse y e duas variáveis auxiliares x_1 e x_2, correlacionadas com a variável de interesse. Em seguida, foram criados três cadastros com ao menos uma variável auxiliar e que conjuntamente listam todos os valores de y. Uma descrição dos cadastros do tipo listagem é dada a seguir:

- L1: composta por 1500 elementos da população selecionados aleatoriamente e com variável auxiliar x_1;
- L2: composta pelos 500 elementos da população não selecionados para L1 e mais 500 elementos selecionados aleatoriamente a partir de L1, com variável auxiliar x_2; e
- L3: composta por 500 elementos da população com os maiores valores de y e com variáveis auxiliares x_1 e x_2.

A proposta do cadastro L3 é representar um grupo com características específicas da população-alvo. Além disso, a maneira como os três cadastros foram gerados também permite que planos de cadastro duplo sejam aplicados nas comparações, uma vez que L1 e L2 fornecem cobertura total para as populações.

Com o objetivo de garantir uma comparação apropriada dos estimadores entre os diferentes planos amostrais, optou-se por fixar o tamanho amostral total em 400, o que corresponde a uma fração amostral de 20%. Assim, no cenário em que os três cadastros foram considerados, foram selecionadas amostras de tamanho $n_1 = 175$ de L1, $n_2 = 125$ de L2 e $n_3 = 100$ de L3. Para o cenário de cadastro duplo, as amostras foram de tamanho $n_1 = 225$ de L1 e $n_2 = 175$ de L2. Todas as amostras foram selecionadas utilizando uma estratégia de amostragem aleatória simples.

6.2 ESTUDO DE MONTE CARLO

Foram realizados estudos de Monte Carlo com 10.000 réplicas para avaliar a precisão dos estimadores em vários cenários, gerados com base em distribuições de probabilidade pertencentes à família exponencial. Em cada réplica de Monte Carlo, uma amostra aleatória simples foi selecionada utilizando o pacote SAMPLING do R (TILLÉ; MATEI, 2016), respeitando uma fração amostral de 20%, e as estimativas para o total e a variância do estimador do total foram calculadas.

O estimador de Monte Carlo para o valor esperado dos estimadores pode ser expresso por

$$\hat{Y}_{mc} = \hat{\mathrm{E}}(\hat{Y}) = \sum_{i=1}^{R} \frac{\hat{Y}_i}{R} \tag{6.1}$$

e a variância desse estimador é

$$\mathrm{Var}(\hat{Y}_{mc}) = \frac{\sum_{i=1}^{R} (\hat{Y}_i - \hat{Y}_{mc})^2}{R}, \tag{6.2}$$

em que, $\hat{Y}_i$ é a estimativa do total para a i-ésima réplica, R é o número de réplicas e $\mathrm{E}(\cdot)$ é o operador de esperança matemática. Com as duas quantidades supracitadas foi calculado o coeficiente de variação (CV).

De maneira análoga, o estimador para o valor esperado da variância do estimador do total pode ser expresso

$$\hat{v}_{mc} = \hat{\mathrm{E}}\left[\widehat{\mathrm{Var}}(\hat{Y})\right] = \sum_{i=1}^{R} \frac{\widehat{\mathrm{Var}}(\hat{Y}_i)}{R} \tag{6.3}$$

e sua variância é

$$\mathrm{Var}(\hat{v}_{mc}) = \frac{\sum_{i=1}^{R} \left[\widehat{\mathrm{Var}}(\hat{Y}_i) - \mathrm{Var}(\hat{Y}_{mc})\right]^2}{R}, \tag{6.4}$$

em que $\widehat{\text{Var}}(\hat{Y}_i)$ é a estimativa da variância do estimador do total na i-ésima réplica. Os desvios padrão de (6.2) e (6.4) também foram computados. Os vieses relativos para (6.1) e (6.3) foram computados, respectivamente, por

$$\text{Viés R}(\hat{Y}_{mc}) = \left(\hat{Y}_{mc} - Y\right)/Y$$
$$\text{Viés R}(\hat{v}_{mc}) = \left[\hat{v}_{mc} - \text{Var}(\hat{Y}_{mc})\right]/\text{Var}(\hat{Y}_{mc}).$$

Por último, para cada réplica, foi calculado um intervalo de confiança baseado na normalidade para as estimativas do total, dado por $\text{IC}_i(95\%) = [\hat{Y}_i \pm z_{(1-\alpha/2)}\sqrt{\widehat{Var}(\hat{Y}_i)}]$ e o percentual de cobertura desse intervalo foi mensurado.

6.3 CASOS PARTICULARES

A simulação considerou três casos particulares de modelos de regressão generalizados. No primeiro caso, a variável resposta foi gerada seguindo uma distribuição gama. Esse modelo é comumente empregado quando a variável de resposta é contínua e positiva, apresentando uma distribuição empírica com assimetria. Em seguida, a distribuição de Poisson foi escolhida para gerar a variável de interesse, adequada para modelar dados de contagem. Por fim, a distribuição de Bernoulli foi utilizada para simular uma variável resposta dicotômica.

Outro fator incorporado na avaliação numérica dos estimadores foi a força da correlação linear entre os valores estimados pelos modelos e os valores observados para a variável de interesse. Rondon, Vanegas e Ferraz (2012) mostraram que, com um único cadastro e sob amostragem aleatória simples, a eficiência dos estimadores GEREG em relação ao estimador de HT é dada por

$$\frac{\text{Var}_p(\hat{Y}_{GEREG})}{\text{Var}_p(\hat{Y}_{HT})} \approx 1 - R_U^2,$$

em que R_U^2 é o coeficiente de correlação linear de Pearson calculado entre y_k e $\hat{\mu}_k$, com $k \in U$. Para o contexto de múltiplos cadastros, o R_U^2 foi calculado utilizando os valores preditos $\hat{\mu}_k$ a partir um modelo populacional completo. Em seguida, foram postulados modelos com as variáveis resposta e auxiliares disponíveis em cada cadastro, com os quais foram calculados os coeficientes de correlação linear de Pearson R_{L1}^2, R_{L2}^2 e R_{L3}^2.

Os estimadores de multiplicidade simples $\hat{Y}_{MS}$, como expresso em (3.8), serviram como referência para a comparação de desempenho. Além disso, o estimador ótimo de Hartley $\hat{Y}_H^{ot}$, descrito em (3.7), foi incluído nas avaliações numéricas nas quais apenas dois cadastros

foram utilizados. O valor de $\hat{\alpha}^{ot}$ também foi utilizado para avaliar o desempenho dos estimadores assistido por modelo sob a abordagem de domínios. Embora α^{ot} seja o valor que minimiza a variância do estimador Hartley, é importante ressaltar que esse não é necessariamente a escolha de α que minimiza as variâncias dos estimadores assistidos por modelos lineares generalizados. Como notação, o sobrescrito H faz alusão a estimadores que foram computados sob a estratégia de estimação por domínios, com $\alpha = \hat{\alpha}^{ot}$, e o sobrescrito MS indica estimadores computados sob a abordagem de multiplicidade simples.

Por fim, em todos os cenários avaliados foi incluído o estimador $\hat{Y}_{Greg}$ assistido por modelo de regressão normal e com função de ligação identidade. Esse estimador difere do estimador GREG usual por aplicar no processo de estimação dos parâmetros a maximização de uma função de pseudo log-verossimilhança que considera a estrutura do plano amostral de múltiplos cadastros, vide as equações (4.3) e (4.13).

6.3.1 Distribuição Gama

Com o objetivo de aplicar o estimador GEREG com base no modelo de regressão gama, foram geradas variáveis dependentes utilizando da relação entre o preditor linear e o parâmetro canônico dessa distribuição. Considerando que a função de ligação canônica é $g(\hat{\mu}_k) = -1/\hat{\mu}_k$, a variável dependente foi gerada a partir de realizações de uma distribuição gama com parâmetro de forma igual a $\hat{\mu}_k$, com

$$\hat{\mu}_k = (\beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k})^{-1},$$

em que $x_{1k} \sim \chi^2_{(1)}$ e $x_{2k} = \exp(-x_{1k})$. A Tabela 1 mostra os valores dos parâmetros escolhidos de modo que o R^2_U entre y_k e $\hat{\mu}_k$ seja 0.5 e 0.8. Nota-se ainda que os valores da correlação de Pearson calculados em cada cadastro foram, em geral, menores do que o R^2_U estipulado, principalmente os valores de R^2_{L3}.

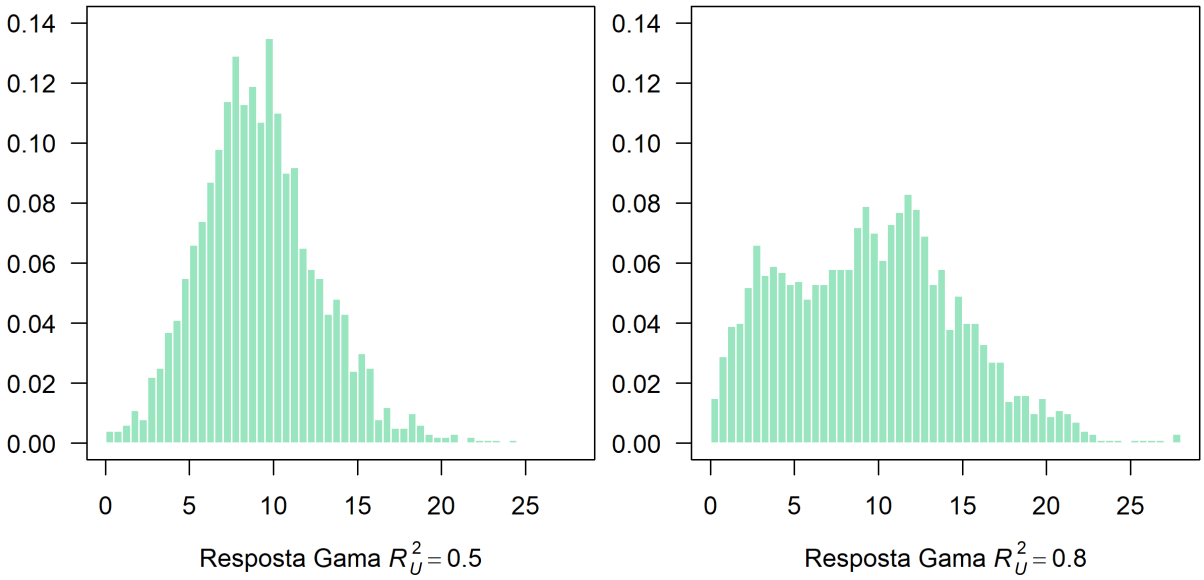
Tabela 1 – Valores dos parâmetros usados para gerar respostas gama

$R^2_U = 0.5$		$R^2_U = 0.8$	
$\beta_0 = 0.01$	$R^2_{L1} = 0.41$	$\beta_0 = 0.01$	$R^2_{L1} = 0.80$
$\beta_1 = 0.05$	$R^2_{L2} = 0.29$	$\beta_1 = 0.105$	$R^2_{L2} = 0.73$
$\beta_2 = 0.1$	$R^2_{L3} = 0.10$	$\beta_2 = 0.06$	$R^2_{L3} = 0.21$

Com os valores dos parâmetros da Tabela 1 foram geradas duas populações com variáveis resposta $y_k \sim \text{Gama}(\hat{\mu}_k)$ e variáveis auxiliares x_{1k} e x_{2k}, conforme o protocolo descrito na seção (6.1). A Figura 2 mostra a distribuição empírica da variáveis respostas geradas na

mesma escala permitindo a comparação visual direta. Além disso, as variáveis respostas são levemente assimétricas, com coeficientes de assimetria de 0.45 e 0.30 para $R^2_U = 0.5$ e $R^2_U = 0.8$, respectivamente.

Figura 2 – Histogramas das variáveis respostas geradas a partir da distribuição de gama



Fonte: De autoria própria

A Tabela 2 apresenta os desempenhos de alguns dos estimadores descritos nesta tese para o total Y. Nesse cenário, os estimadores GEREG assistidos por um modelo de regressão gama com função de ligação canônica são representados por $\hat{Y}_{Gama}$ e $\hat{Y}_{Gama^*}$, este último sendo o estimador simplificado descrito no capítulo 5.

Tabela 2 – Resultados para o estimador do total com variável resposta Gama

Estimador	$R^2_U = 0.5$					$R^2_U = 0.8$				
	CV (%)	Viés R (%)	DP	R.EQM	Cobertura (%)	CV (%)	Viés R (%)	DP	R.EQM	Cobertura (%)
três cadastros										
$\hat{Y}_{MS}$	1.84	0.004	338.1	338.1	94.9	2.29	0.026	435.9	436.0	94.6
$\hat{Y}^{MS}_{Greg}$	1.49	0.084	273.9	**274.3**	94.6	1.65	−0.082	313.4	313.8	94.0
$\hat{Y}^{MS}_{Gama}$	1.53	0.114	280.0	280.8	94.7	1.55	0.096	295.4	**296.0**	94.5
$\hat{Y}^{MS}_{Gama^*}$	1.53	0.114	280.0	280.8	94.7	1.55	0.096	295.4	**296.0**	94.5
dois cadastros										
$\hat{Y}_{MS}$	2.26	−0.004	414.1	414.1	95.1	2.98	0.013	566.8	566.8	94.5
$\hat{Y}^{MS}_{Greg}$	1.68	0.044	308.6	**308.7**	94.8	1.84	−0.081	349.3	349.6	94.2
$\hat{Y}^{MS}_{Gama}$	1.72	0.061	314.6	314.8	94.8	1.64	0.053	312.9	**313.1**	94.6
$\hat{Y}^{MS}_{Gama^*}$	1.72	0.061	314.6	314.8	94.8	1.64	0.053	312.9	**313.1**	94.6
$\hat{Y}^{ot}_{H}$	1.83	0.004	336.0	336.0	94.8	2.64	0.014	503.2	503.2	96.9
$\hat{Y}^{H}_{Greg}$	1.69	0.026	309.1	309.1	93.6	1.83	−0.109	348.2	348.8	94.3
$\hat{Y}^{H}_{Gama}$	1.72	0.045	315.1	315.2	93.6	1.64	0.032	313.1	**313.2**	94.5
$\hat{Y}^{H}_{Gama^*}$	1.72	0.045	315.1	315.2	93.6	1.64	0.032	313.1	**313.2**	94.5

Fonte: De autoria própria

Para $R^2_U = 0.5$, o estimador $\hat{Y}^{MS}_{Greg}$ baseado no modelo normal apresentou o melhor desempenho em termos da raiz do erro quadrático médio (R.EQM), tanto para dois quanto para três cadastros. Para três cadastros, a eficiência relativa observada entre o estimador $\hat{Y}^{MS}_{Greg}$

e o estimador $\hat{Y}_{MS}$ foi de 0.66. Já para dois cadastros, a eficiência relativa foi de 0.55. O estimador ótimo de Hartley $\hat{Y}_H^{ot}$ teve um desempenho superior em relação ao estimador por multiplicidade simples $\hat{Y}_{MS}$, como esperado, uma vez que $\hat{Y}_H^{ot}$ utiliza um critério ótimo na escolha de α. Entretanto, esse comportamento não foi evidente para os estimadores assistidos por modelos que utilizaram a estratégia de estimação de Hartley, com $\alpha = \alpha^{ot}$. As eficiências relativas dos estimadores assistidos por modelos foram mais próximas de 1 quando comparados com $\hat{Y}_H^{ot}$. Mesmo assim, houve ganhos de precisão para esses estimadores por utilizarem as informações auxiliares disponíveis, especialmente no contexto de três cadastros.

Para a população com $R_U^2 = 0.8$, os estimadores $\hat{Y}_{Gama}$ e $\hat{Y}_{Gama^*}$ apresentaram os menores R.EQM's. A eficiência relativa desses estimadores, em comparação com o estimador $\hat{Y}_{MS}$, foi de 0.46 no contexto de três cadastros e 0.30 no contexto de dois cadastros. Além disso, não houve diferenças significativas entre as estimativas obtidas por meio das estratégias de estimação de Hartey e multiplicidade simples.

O resultado 1 pode ser corroborado, uma vez que os desempenhos apresentados pelos estimadores $\hat{Y}_{Gama}$ e $\hat{Y}_{Gama^*}$ foram numericamente iguais. Além disso, os resultados sugerem que a adição de um cadastro pode melhorar o desempenho de todos os estimadores avaliados. A centralidade assintótica dos estimadores GEREG também fica evidenciada por meio do viés relativo, uma vez que, ao aumentar o número de cadastros, o tamanho da amostra selecionada em cada cadastro é reduzido.

A Tabela 3 apresenta os desempenhos dos estimadores da variância. Os coeficientes de variação (CV), expressos em percentual, sugerem que os estimadores das variâncias dos estimadores do tipo regressão são consistentes em torno de $\hat{v}_{mc}$. Para $R_U^2 = 0.5$, observa-se um viés relativo próximo a -10% para os estimadores assistidos por modelos sob a estratégia de estimação por domínios, com $\alpha = \alpha^{ot}$. Essa subestimação influenciou os percentuais de cobertura observados para os intervalos de confiança, que ficaram abaixo de 95%, conforme indicado na Tabela 2. Para $R_U^2 = 0.5$, os estimadores das variâncias apresentaram vieses negativos, com exceção de $\widehat{\text{Var}}(\hat{Y}_H^{ot})$, que apresentou uma superestimação de 23.7%. Além disso, quando o R_U^2 passou de 0.5 para 0.8, houve aumento nas raízes dos erros quadráticos médios de todos os estimadores da variância avaliados.

Tabela 3 – Resultados para o estimador da variância do total estimado com variável resposta Gama

Estimador	$R^2_U = 0.5$					$R^2_U = 0.8$				
	$\hat{v}_{mc}$	CV (%)	Viés R (%)	DP	R.EQM	$\hat{v}_{mc}$	CV (%)	Viés R (%)	DP	R.EQM
					três cadastros					
$\widehat{\text{Var}}(\hat{Y}_{MS})$	113549.0	5.92	−0.649	6725.1	6766.0	187888.7	6.05	−1.134	11364.9	11567.3
$\widehat{\text{Var}}(\hat{Y}^{MS}_{Greg})$	74424.0	7.86	−0.797	5853.5	5883.9	94132.3	9.32	−4.194	8776.5	9695.7
$\widehat{\text{Var}}(\hat{Y}^{MS}_{Gama})$	77903.0	7.87	−0.630	6130.5	6150.4	85070.4	8.59	−2.505	7309.9	7629.7
$\widehat{\text{Var}}(\hat{Y}^{MS}_{Gama^*})$	77903.0	7.87	−0.630	6130.5	6150.4	85070.4	8.59	−2.505	7309.9	7629.7
					dois cadastros					
$\widehat{\text{Var}}(\hat{Y}_{MS})$	171729.8	7.67	0.137	13180.9	13183.0	313327.2	7.05	−2.461	22096.6	23468.2
$\widehat{\text{Var}}(\hat{Y}^{MS}_{Greg})$	95462.0	9.26	0.259	8842.9	8846.4	117026.0	10.6	−4.062	12354.7	13311.4
$\widehat{\text{Var}}(\hat{Y}^{MS}_{Gama})$	99162.4	9.18	0.163	9107.5	9109.0	95738.7	9.71	−2.207	9293.2	9541.1
$\widehat{\text{Var}}(\hat{Y}^{MS}_{Gama^*})$	99162.4	9.18	0.163	9107.5	9109.0	95738.7	9.71	−2.207	9293.2	9541.1
$\widehat{\text{Var}}(\hat{Y}^{ot}_{H})$	111688.2	17.5	−1.09	19570.7	19609.4	313327.2	7.05	23.758	22096.6	64079.5
$\widehat{\text{Var}}(\hat{Y}^{H}_{Greg})$	85864.9	11.0	−10.12	9440.1	13517.3	117026.0	10.5	−3.480	12354.7	13055.4
$\widehat{\text{Var}}(\hat{Y}^{H}_{Gama})$	89202.5	10.9	−10.13	9741.5	14001.1	95738.7	9.71	−2.372	9293.2	9579.9
$\widehat{\text{Var}}(\hat{Y}^{H}_{Gama^*})$	89202.5	10.9	−10.13	9741.5	14001.1	95738.7	9.71	−2.372	9293.2	9579.9

Fonte: De autoria própria

6.3.2 Distribuição Poisson

O desempenho dos estimadores propostos também foi avaliado quando a variável de interesse é uma contagem. Para tanto, variáveis dependentes foram geradas a partir da distribuição Poisson. Nesse caso, o preditor linear do modelo utiliza como função de ligação canônica $g(\hat{\mu}_k) = \log \hat{\mu}_k$. Assim, as variáveis dependentes foram geradas a partir a distribuição de Poisson, com parâmetro $\hat{\mu}_k$ determinado por

$$\hat{\mu}_k = \exp\left(\beta_0 + \beta 1 x_{1k} + \beta_2 x_{2k}\right),$$

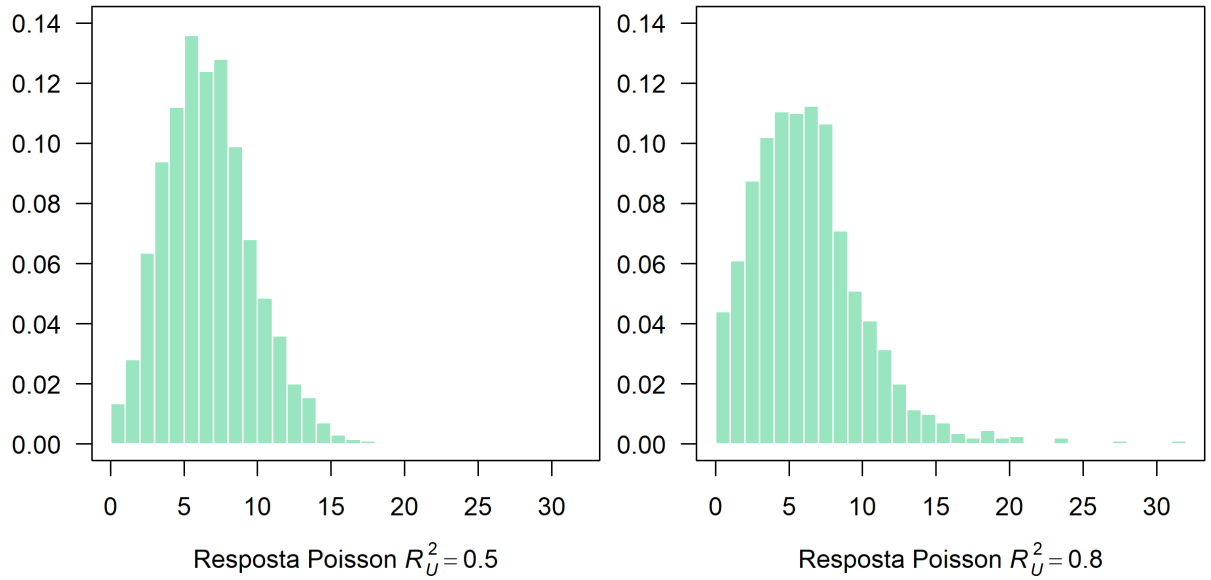
em que, $x_{1k} \sim \chi^2_{(1)}$ e $x_{2k} \sim \chi^2_{(2)}$. Os valores do parâmetro de forma $\hat{\mu}_k$ aumenta exponencialmente a medida que o preditor linear aumenta. A Tabela 4 apresenta os valores dos parâmetros usados para gerar populações com R^2_U iguais a 0.5 e 0.8. Diferente do cenário com variáveis respostas com distribuição gama, os valores de R^2_{L3} foram próximos dos respectivos valores de R^2_U.

Tabela 4 – Valores dos parâmetros usados para gerar respostas poisson

$R^2_U = 0.5$		$R^2_U = 0.8$	
$\beta_0 = 2.00$	$R^2_{L1} = 0.41$	$\beta_0 = 2.00$	$R^2_{L1} = 0.67$
$\beta_1 = 0.09$	$R^2_{L2} = 0.34$	$\beta_1 = 0.18$	$R^2_{L2} = 0.49$
$\beta_2 = -0.08$	$R^2_{L3} = 0.45$	$\beta_2 = -0.18$	$R^2_{L3} = 0.82$

A Figura 3 apresenta os histogramas para variáveis respostas $y_k \sim \text{Poisson}(\hat{\mu}_k)$, geradas a partir dos parâmetros apresentados na Tabela 4. Para a população com $R^2_U = 0.5$, a variável resposta exibe uma leve assimetria, com um coeficiente de assimetria de 0.52. Em contraste, para a população com $R^2_U = 0.8$, a variável resposta apresenta uma forte assimetria, com coeficiente de assimetria de 2.90. Além disso, observam-se algumas realizações distante do centro de massa da distribuição empírica.

Figura 3 – Histogramas da variáveis respostas geradas a partir da distribuição de Poisson



Fonte: De autoria própria

Os estimadores GEREG utilizados neste cenário, denotados por $\hat{Y}_{Poisson}$ e $\hat{Y}_{Poisson^*}$, foram assistidos por um modelo de regressão poisson com função de ligação canônica, sendo o último um estimador simplificado. A Tabela 5 apresenta os resultados da avaliação numérica de alguns estimadores para o total. Novamente, os resultados numericamente iguais dos estimadores $\hat{Y}_{Poisson}$ e $\hat{Y}_{Poisson^*}$ corroboram com o resultado 1. Além disso, todos os estimadores apresentaram melhores desempenhos, em termos da raiz do erro quadrático médio, ao considerar o plano amostral com três cadastros.

Tabela 5 – Resultados para o estimador do total com variável resposta Poisson

Estimador	$R_U^2 = 0.5$					$R_U^2 = 0.8$				
	CV (%)	Viés R (%)	DP	R.EQM	Cobertura (%)	CV (%)	Viés R (%)	DP	R.EQM	Cobertura (%)
					três cadastros					
$\hat{Y}_{MS}$	1.81	−0.007	258.8	258.8	95.0	2.44	0.014	335.5	335.5	94.5
$\hat{Y}^{MS}_{Greg}$	1.70	0.034	242.4	**242.4**	94.5	2.25	−0.093	308.9	309.1	94.0
$\hat{Y}^{MS}_{Poisson}$	1.70	0.063	242.5	242.6	94.4	2.14	−0.035	294.9	**295.0**	94.2
$\hat{Y}^{MS}_{Poisson^*}$	1.70	0.063	242.5	242.6	94.4	2.14	−0.035	294.9	**295.0**	94.2
					dois cadastros					
$\hat{Y}_{MS}$	2.53	−0.002	361.1	361.1	94.5	3.48	−0.021	478.5	478.6	94.3
$\hat{Y}^{MS}_{Greg}$	1.90	−0.032	271.1	**271.1**	94.8	2.56	−0.166	352.0	352.7	93.8
$\hat{Y}^{MS}_{Poisson}$	1.90	0.016	271.6	271.6	94.7	2.46	−0.028	338.1	**338.2**	94.0
$\hat{Y}^{MS}_{Poisson^*}$	1.90	0.016	271.6	271.6	94.7	2.46	−0.028	338.1	**338.2**	94.0
$\hat{Y}^{ot}_{H}$	2.07	−0.036	295.4	295.4	94.9	3.16	−0.119	434.3	434.6	93.9
$\hat{Y}^{H}_{Greg}$	1.90	−0.045	271.1	**271.1**	93.4	2.57	−0.234	352.7	354.1	92.6
$\hat{Y}^{H}_{Poisson}$	1.90	0.002	271.5	271.5	93.4	2.46	−0.094	338.6	338.9	92.8
$\hat{Y}^{H}_{Poisson^*}$	1.90	0.002	271.5	271.5	93.4	2.46	−0.094	338.6	338.9	92.8

Fonte: De autoria própria

Para $R_U^2 = 0.5$, os estimadores assistidos por modelos de regressão normal e poisson apresentaram resultados muito próximos, com ligeira vantagem para o estimador $\hat{Y}_{Greg}$. Considerando três cadastros, a variância do estimador $\hat{Y}^{MS}_{Poisson}$ foi reduzida em 12,4% em relação à

variância de $\hat{Y}_{MS}$. Já com apenas dois cadastros, essa redução foi de 43.4%. Para $R^2_U = 0.8$, os estimadores assistidos por modelo de regressão de poisson demonstraram maior precisão na raiz do erro quadrático médio. A eficiência dos estimadores $\hat{Y}^{MS}_{Poisson}$ e $\hat{Y}^{MS}_{Poisson^*}$ em relação a $\hat{Y}_{MS}$ foi de 0.77 quando utilizados três cadastros, e de 0.50, quando utilizados dois cadastros.

O estimador ótimo de Hartley apresentou melhor desempenho em relação a $\hat{Y}_{MS}$ apenas no contexto de cadastro duplo. A adição de um terceiro cadastro foi suficiente para que a estratégia de estimação por multiplicidade simples produzisse estimativas mais precisas se comparado com $\hat{Y}^{ot}_{H}$. Além disso, não houve diferenças significativas nos desempenhos dos estimadores assistidos por modelos ao adotar a estratégia de estimação por domínios, com $\alpha = \alpha^{ot}$ ou a estratégia de multiplicidade, no contexto de cadastro duplo.

Os resultados da avaliação numérica para os estimadores das variâncias são apresentados na Tabela 6. De modo geral, os vieses relativos foram negativos, indicando que, em média, houve uma subestimação das variâncias dos estimadores. Em particular, esses vieses foram superiores a 11% para os estimadores do total assistidos por modelos sob a abordagem de domínios, afetando o percentual de cobertura os intervalos de confiança apresentados na Tabela 5. Os coeficientes de variação inferiores a 10% para a população com $R^2_U = 0.5$ sugerem que os estimadores das variâncias dos estimadores do tipo regressão são consistentes em torno de $\hat{v}_{mc}$. No entanto, a distribuição empírica da variável resposta para população com $R^2_U = 0.8$ apresentou algumas observações distantes do centro de massa, o que pode ter inflacionado os valores do coeficiente de variação neste caso.

Tabela 6 – Resultados para o estimador da variância do total estimado com variável resposta Poisson

Estimador	$R^2_U = 0.5$					$R^2_U = 0.8$				
	$\hat{v}_{mc}$	CV (%)	Viés R (%)	DP	R.EQM	$\hat{v}_{mc}$	CV (%)	Viés R (%)	DP	R.EQM
três cadastros										
$\widehat{\text{Var}}(\hat{Y}_{MS})$	67035.0	6.24	0.078	4180.6	4180.9	111029.0	18.5	−1.353	20548.4	20604.8
$\widehat{\text{Var}}(\hat{Y}^{MS}_{Greg})$	56886.9	7.55	−3.166	4294.0	4679.6	91633.7	14.5	−3.946	13263.8	13787.7
$\widehat{\text{Var}}(\hat{Y}^{MS}_{Poisson})$	56659.5	7.51	−3.612	4257.5	4757.6	83046.9	11.3	−4.521	9414.6	10202.8
$\widehat{\text{Var}}(\hat{Y}^{MS}_{Poisson^*})$	56659.5	7.51	−3.612	4257.5	4757.6	83046.9	11.3	−4.521	9414.6	10202.8
dois cadastros										
$\widehat{\text{Var}}(\hat{Y}_{MS})$	127316.7	7.95	−2.377	10123.8	10587.7	225333.2	25.5	−1.603	57492.1	57609.2
$\widehat{\text{Var}}(\hat{Y}^{MS}_{Greg})$	72363.5	7.89	−1.517	5709.9	5817.7	118415.4	17.5	−4.411	20731.1	21439.3
$\widehat{\text{Var}}(\hat{Y}^{MS}_{Poisson})$	72405.5	7.94	−1.815	5750.5	5904.3	107698.9	14.6	−5.815	15723.5	17071.9
$\widehat{\text{Var}}(\hat{Y}^{MS}_{Poisson^*})$	72405.5	7.94	−1.815	5750.5	5904.3	107698.9	14.6	−5.815	15723.5	17071.9
$\widehat{\text{Var}}(\hat{Y}^{ot}_{H})$	85251.7	12.6	−2.285	10742.7	10926.1	185082.4	29.6	−1.868	54716.7	54830.0
$\widehat{\text{Var}}(\hat{Y}^{H}_{Greg})$	65352.6	9.60	−11.048	6273.0	10258.2	108632.4	19.2	−12.65	20876.3	26141.4
$\widehat{\text{Var}}(\hat{Y}^{H}_{Poisson})$	65363.7	9.66	−11.332	6312.0	10470.1	98343.3	16.4	−14.24	16118.4	22941.9
$\widehat{\text{Var}}(\hat{Y}^{H}_{Poisson^*})$	65363.7	9.66	−11.332	6312.0	10470.1	98343.3	16.4	−14.24	16118.4	22941.9

Fonte: De autoria própria

6.3.3 Distribuição Bernoulli

No último cenário, foi considerada uma variável de resposta binária gerada a partir da distribuição Bernoulli. Nesse caso, a função $logit(\cdot)$ é a ligação canônica do modelo, ou seja, $g(\hat{\mu}_k) = \log\left(\frac{\hat{\mu}_k}{1-\hat{\mu}_k}\right)$. Assim, as variáveis resposta foram geradas a partir de realizações de uma distribuição Bernoulli com a probabilidade de $y_k = 1$ determinada por

$$\hat{\mu}_k = \frac{\exp(\beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k})}{1 + \exp(\beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k})},$$

em que $x_{1k} \sim U_{[-3,\,3]}$ e $x_{2k} = \exp(-x_{1k})$. É comum se referir ao evento $y_k = 1$ como a observação de um *sucesso* e $y_k = 0$ como a observação de *fracasso*. O intento dessa avaliação numérica foi estimar o total de sucessos $Y = \sum_{k \in U} y_k$ com $y_k \sim \text{Bernoulli}(\hat{\mu}_k)$.

Os valores dos parâmetros usados para gerar duas populações com diferentes R^2_U estão apresentados na Tabela 7. Para ambas populações, o proporção de sucessos foi próxima de 0.5, o que sugere um equilíbrio entre as probabilidades de sucesso e fracasso. Nota-se ainda que os valores de R^2_{L3} não ficaram próximos dos respectivos valores estabelecidos para R^2_U.

Tabela 7 – Valores dos parâmetros usados para gerar respostas bernoulli

$R^2_U = 0.5$		$R^2_U = 0.8$	
$\beta_0 = 0.01$	$R^2_{L1} = 0.47$	$\beta_0 = 0.02$	$R^2_{L1} = 0.77$
$\beta_1 = 0.6$	$R^2_{L2} = 0.48$	$\beta_1 = 1.65$	$R^2_{L2} = 0.76$
$\beta_2 = -0.05$	$R^2_{L3} = 0.19$	$\beta_2 = -0.20$	$R^2_{L3} = 0.36$
$p = 0.48$		$p = 0.48$	

Neste cenário, foram considerados três estimadores assistidos por modelos lineares generalizados para a distribuição de bernoulli. O estimador denotado por $\hat{Y}_{Logit}$ é baseado em um modelo de regressão logístico proposto por Molina *et. al.* (2015) com uma generalização do LGREG para planos amostrais de cadastro duplo. Por fim, considerou-se o modelo de regressão probito, para o qual

$$\hat{\mu}_k = \Phi(\beta_0 + \beta 1 x_{1k} + \beta_2 x_{2k}),$$

em que $\Phi(\cdot)$ é a função de distribuição acumulada da normal padrão (CORDEIRO; DEMÉTRIO, 2008, p. 31). O estimador assistido pelo modelo probito foi denotado por $\hat{Y}_{Probit}$.

Tanto o modelo logístico quanto o probito estimam a chance de $y_k = 1$. Nesse caso, ao invés que usar diretamente os valores ajustados $\hat{\mu}_k$, foi aplicada uma regra de predição a estes com base na proporção de sucessos amostral. Para esta avaliação numérica, os valores preditos

foram dados por:

$$\hat{y}_k = \begin{cases} 1, & \text{se } \hat{\mu}_k \geq p_{s_q}^{(r)} \\ 0, & \text{se } \hat{\mu}_k < p_{s_q}^{(r)} \end{cases}, \; k \in U_q,$$

em que $p_{s_q}^{(r)}$ é a proporção observada de sucessos para a amostra selecionada no cadastro q na r-ésima réplica de Monte Carlo.

A Tabela 8 apresenta os resultados dos estimadores para $Y = \sum_{k \in U} y_k$. Considerando a população com $R_U^2 = 0.5$, os estimadores do tipo regressão não apresentaram desempenhos melhores do que $\hat{Y}_H^{ot}$ e $\hat{Y}_{MS}$ no contexto de dois cadastros. Além disso, a adição de um terceiro cadastro diminuiu a precisão de todos os estimadores avaliados. Esse panorama mudou com a população $R_U^2 = 0.8$. Os estimadores $\hat{Y}_{Logit}$ e $\hat{Y}_{Probit}$ apresentaram resultados similares e foram os mais precisos dentre todos os estimadores avaliados, principalmente ao utilizar um terceiro cadastro. A eficiência de $\hat{Y}_{Logit}$ em relação a $\hat{Y}_{MS}$ foi de 0.42 com a utilização de três cadastros e de 0.49 com o aplicação de dois cadastros.

Tabela 8 – Resultados para o estimador do total com variável resposta Bernoulli

Estimador	$R_U^2 = 0.5$					$R_U^2 = 0.8$				
	CV (%)	Viés R (%)	DP	R.EQM	Cobertura (%)	CV (%)	Viés R (%)	DP	R.EQM	Cobertura (%)
					três cadastros					
$\hat{Y}_{MS}$	5.58	−0.015	53.9	53.9	94.8	5.39	0.048	51.9	51.9	94.9
$\hat{Y}_{Greg}^{MS}$	5.58	0.119	54.0	54.0	94.4	3.87	0.348	37.5	37.6	94.1
$\hat{Y}_{Logit}^{MS}$	5.53	0.097	53.5	**53.5**	95.0	3.48	0.195	33.6	**33.6**	94.3
$\hat{Y}_{Probit}^{MS}$	5.53	0.103	53.5	53.5	94.8	3.49	0.181	33.7	33.8	94.3
					dois cadastros					
$\hat{Y}_{MS}$	5.20	0.003	50.3	50.3	94.8	5.19	0.039	50.0	50.0	94.9
$\hat{Y}_{Greg}^{MS}$	5.38	0.027	52.1	52.1	94.5	3.85	0.232	37.2	37.3	94.2
$\hat{Y}_{Logit}^{MS}$	5.37	0.018	52.0	52.0	94.5	3.62	0.042	34.9	35.0	94.1
$\hat{Y}_{Probit}^{MS}$	5.37	0.022	51.9	51.9	94.5	3.64	0.035	35.1	35.1	94.3
$\hat{Y}_H^{ot}$	5.03	0.020	48.6	**48.6**	94.7	4.99	0.052	48.2	37.4	94.9
$\hat{Y}_{Greg}^{H}$	5.37	0.036	52.0	52.0	93.1	3.86	0.233	37.3	1397	92.7
$\hat{Y}_{Logit}^{H}$	5.36	0.024	51.9	51.9	93.2	3.62	0.043	34.9	**34.9**	92.7
$\hat{Y}_{Probit}^{H}$	5.36	0.025	51.8	51.8	93.1	3.64	0.034	35.1	35.1	92.9

Fonte: De autoria própria

Os resultados também mostram que os estimadores assistidos por modelos demostraram desempenhos semelhantes se comparadas as estratégias de estimação por domínios, com $\alpha = \alpha^{ot}$, e multiplicidade simples.

Os resultados da avaliação numérica para os estimadores das variâncias são apresentados na Tabela 9. Os vieses relativos para os estimadores das variâncias dos estimadores assistidos por modelo foram negativos, indicando que, em média, houve uma subestimação. Em particular, esses vieses foram superiores a 13% quando a abordagem inferencial de domínios foi aplicada, o que afetou o percentual de cobertura os intervalos de confiança apresentados na Tabela 8. Além disso, os coeficientes de variação foram inferiores a 20%, tanto para a população

com $R^2_U = 0.5$ quanto para a população com $R^2_U = 0.8$. Isso sugerem que os estimadores das variâncias são consistentes em torno de $\hat{v}_{mc}$.

Tabela 9 – Resultados para o estimador da variância do total estimado com variável resposta Bernoulli

Estimador	$R^2_U = 0.5$					$R^2_U = 0.8$				
	$\hat{v}_{mc}$	CV (%)	Viés R (%)	DP	R.EQM	$\hat{v}_{mc}$	CV (%)	Viés R (%)	DP	R.EQM
					três cadastros					
$\widehat{\text{Var}}(\hat{Y}_{MS})$	2870.2	5.34	−1.323	153.4	158.1	2709.9	5.69	0.435	154.1	154.6
$\widehat{\text{Var}}(\hat{Y}^{MS}_{Greg})$	2818.0	10.82	−3.497	304.9	321.5	1324.4	17.5	−5.741	232.0	245.7
$\widehat{\text{Var}}(\hat{Y}^{MS}_{Logit})$	2801.8	10.99	−2.069	308.0	313.7	1109.5	18.3	−1.649	203.4	204.2
$\widehat{\text{Var}}(\hat{Y}^{MS}_{Probit})$	2806.4	10.96	−2.075	307.5	313.2	1124.1	18.2	−1.092	204.2	204.6
					dois cadastros					
$\widehat{\text{Var}}(\hat{Y}_{MS})$	2496.3	3.30	−1.379	82.3	89.4	2505.1	3.20	0.098	80.2	80.2
$\widehat{\text{Var}}(\hat{Y}^{MS}_{Greg})$	2596.2	8.71	−4.287	226.2	254.4	1328.9	13.0	−4.154	172.5	181.9
$\widehat{\text{Var}}(\hat{Y}^{MS}_{Logit})$	2590.8	8.89	−4.111	230.3	255.6	1177.0	14.4	−3.642	169.8	175.5
$\widehat{\text{Var}}(\hat{Y}^{MS}_{Probit})$	2592.7	8.81	−3.774	228.5	250.1	1189.1	14.3	−3.631	169.8	175.7
$\widehat{\text{Var}}(\hat{Y}^{ot}_{H})$	2322.8	3.52	−1.671	81.7	90.7	2332.6	3.39	0.545	79.1	80.1
$\widehat{\text{Var}}(\hat{Y}^{H}_{Greg})$	2334.0	11.27	−13.67	263.1	453.6	1191.1	14.3	−14.42	178.6	268.7
$\widehat{\text{Var}}(\hat{Y}^{H}_{Logit})$	2333.9	11.41	−13.29	266.4	446.1	1062.2	16.7	−12.93	177.2	237.2
$\widehat{\text{Var}}(\hat{Y}^{H}_{Probit})$	2334.4	11.36	−13.15	265.1	441.9	1073.0	16.5	−13.06	177.0	239.4

Fonte: De autoria própria

6.3.4 Discussão dos resultados

Nas simulações de Monte Carlo apresentadas nesta seção, foram analisados vários aspectos considerados relevantes para a estimação assistida por modelos em planos amostrais de múltiplos cadastros, conforme apontado nesta tese. Além disso, foram avaliadas as propriedades de centralidade e variância aproximadas, bem como a condição sob a qual é possível desconsiderar o termo de ajuste dos estimadores GEREG, expressando-os de forma simplificada.

O primeiro aspecto avaliado foi o ganho de precisão dos estimadores ao incluir um cadastro adicional que, embora não seja necessário para garantir cobertura total, contenha informações sobre um subgrupo específico da população de interesse. Os resultados demonstram que, em geral, os estimadores GEREG apresentaram ganhos de precisão, em termos da raiz do erro quadrático médio, quando três cadastros foram utilizados (a exceção foi o cenário com distribuição Bernoulli e $R^2_U = 0.5$, possivelmente devido ao processo preditivo empregado aos modelos logit e probit). Entretanto, observa-se um aumento no viés relativo dos estimadores, uma vez que, ao considerar mais cadastros, os tamanhos das amostras selecionadas em cada cadastro foram reduzidos para respeitar o tamanho final da amostra.

Os resultados da simulação também evidenciam que os estimadores assistidos por modelos demonstram maior precisão ao considerar a natureza da variável resposta durante o ajuste do modelo. Para cada uma das três distribuições consideradas, os estimadores GEREG,

baseados no modelo da própria distribuição, apresentaram melhores desempenhos em comparação com o estimador GREG, baseado no modelo normal e função de ligação identidade. A única exceção foi no cenário com distribuição gama e $R_U^2 = 0.5$, no qual a variável resposta gerada foi levemente assimétrica. Contudo, nos cenários em que a variável de interesse apresenta uma maior assimetria, espera-se que o estimador assistido por um modelo de regressão gama supere a precisão do estimador GREG até para valores moderados de R_U^2.

Ao comparar as estratégias inferenciais de domínio e multiplicidade simples, foram percebidas diferenças de desempenho apenas entre os estimadores $\hat{Y}_H^{ot}$ e $\hat{Y}_{MS}$. Neste estudo, a utilização extrapolada de $\alpha = \alpha^{ot}$ com os estimadores assistidos por modelos sob a abordagem de domínios não resultou na redução das variâncias desses estimadores, se comparados com os estimadores correspondentes sob a abordagem de multiplicidade. Embora não seja o foco desta tese, enfatiza-se que encontrar o valor de α que minimiza a variância do estimador GEREG é um problema recursivo. O valor de α é utilizado para estimar os parâmetros do modelo que, por sua vez, geram um novo estimador cuja variância deve ser minimizada.

Por fim, a condição na qual o termo de ajuste do estimador GEREG se anula (resultado 1 do capítulo 5) também pode ser verificada por meio das simulações envolvendo as distribuições Gama e Poisson, tanto para a abordagem de domínios quanto para a de multiplicidade simples.

6.4 APLICAÇÃO EM AGROPECUÁRIA

Os estimadores propostos também foram avaliados no contexto de pesquisas amostrais em agropecuária. O desenvolvimento da teoria de estimação em planos amostrais de cadastro duplo foi amplamente motivado por aplicações em estudos agrícolas. Por exemplo, desde 1954, o *National Agricultural Statistics Service* (NASS) do *United States Department of Agriculture* (USDA) tem desenvolvido, usado e analisado cadastros de área como principal ferramenta para conduzir pesquisas agropecuária (COTTER; NEALON, 1987). Nesses estudos, um cadastro de área que fornece cobertura total para a população-alvo é complementado por cadastros com menor custo associados à sua utilização.

Nessa proposta de aplicação, foram gerados tanto um cadastro quanto uma população fictícia que possuem elementos inerentes a pesquisas com cadastros de área. Para emular uma população de produtores, foram utilizados como norteadores os resultados do Censo Agropecuário de 2017. Nesse sentido, dados acerca da área colhida com soja (em kg) e tamanho

do rebanho bovino (em cabeças) para os estabelecimentos agropecuários do município de Sorriso, no estado do Mato Grosso, foram consideradas como variáveis de interesse.

Foram considerados três cadastros nesta aplicação: um cadastro de área de segmentos quadrados (GALLEGO; DELINCÉ; CARFAGNA, 1994) e dois cadastros com uma listagem de produtores. Os estimadores do tipo regressão foram aplicados apenas nos cadastros do tipo lista. Para o cadastro de área foi utilizado o estimador baseado na subamostragem dos segmentos por pontos, como descrito em (FERRAZ; MECATTI; TORRES, 2022). Para uma revisão sistemática da aplicação de cadastros de área no âmbito de pesquisas em agropecuária veja (TORRES, 2018).

6.4.1 Cadastros utilizados

Segundo o Censo Agropecuário de 2017, Sorriso é o município do estado de Mato Grosso (MT) com a maior área colhida de soja em quilos. A área territorial desse município é 9.293,63 km², correspondendo a cerca de 42% do território do estado de Sergipe. A Figura 4 mostra a localização geográfica do município de Sorriso em relação ao território do Brasil.

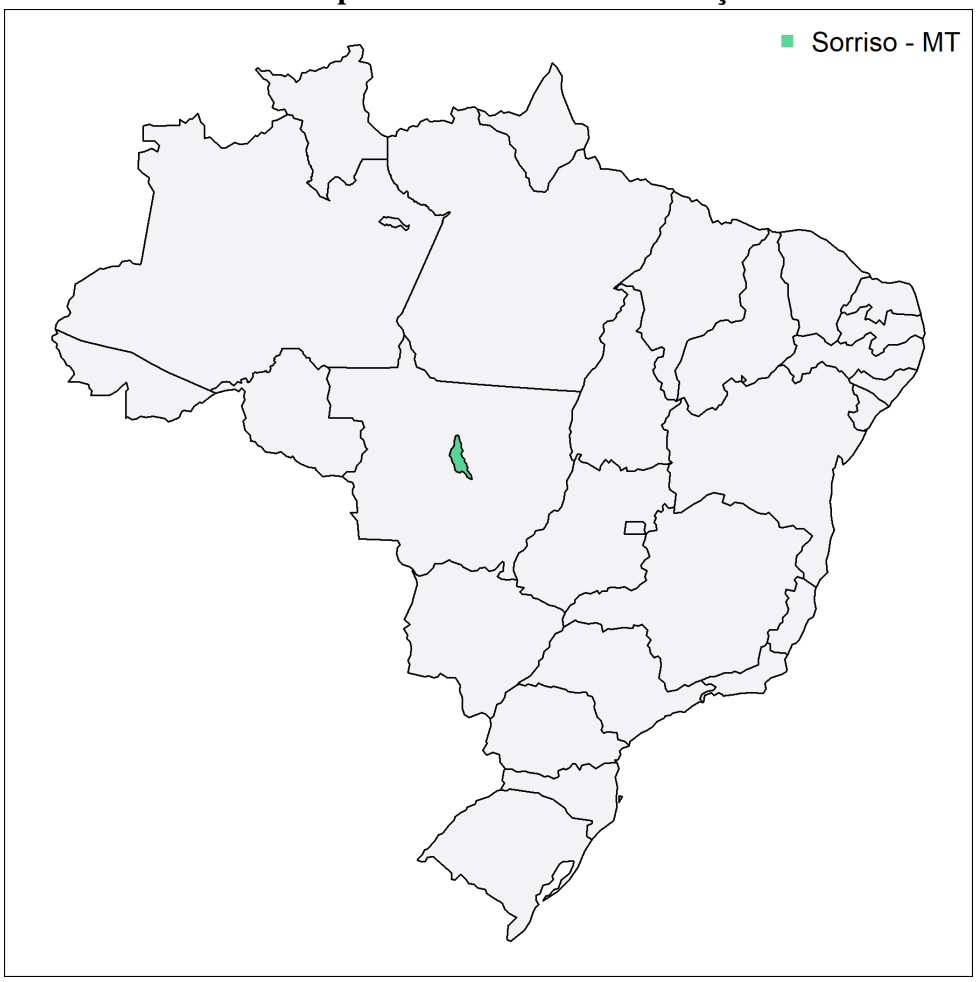
Na construção e uso do cadastro de área foram adotadas práticas consideradas de baixo custo, como o uso de segmentos quadrados com subamostragem de pontos dentro dos segmentos (GALLEGO, 2005). Para cobrir a região de Sorriso, um cadastro de área foi construído com 9.263 segmentos quadrados, cada um com dimensão de 100 hectares (1000m $\times$ 1000m). Nas regiões de fronteira, segmentos cuja área de interseção com o município era menor que 50% foram retirados do cadastro.

Complementares ao cadastro área, foram elaborados dois cadastros do tipo lista. O primeiro, denotado por L1, contém a relação dos produtores com as maiores áreas colhida com soja (em kg) e disponibiliza como variável auxiliar a área com total de lavouras temporárias na propriedade. O segundo, denotado por L2, contém a relação dos produtores com os maiores rebanhos bovinos e utiliza como variável auxiliar a área total de pastagem na propriedade.

6.4.2 Geração de uma população com características espaciais

Para preservar as características espaciais intrínsecas à amostragem de área, foi necessário gerar uma população de campos produtivos georreferenciados. Com base nos resultados do Censo Agropecuário de 2017 para Sorriso, foi emulada uma população artificial de 1.100 estabelecimentos agrícolas que, pelo menos, produziam soja ou possuíam rebanhos bovinos. O passo seguinte foi associar a cada estabelecimento um polígono correspondente à sua área

Figura 4 – Território do município de Sorriso - MT em relação ao território brasileiro



Fonte: De autoria própria

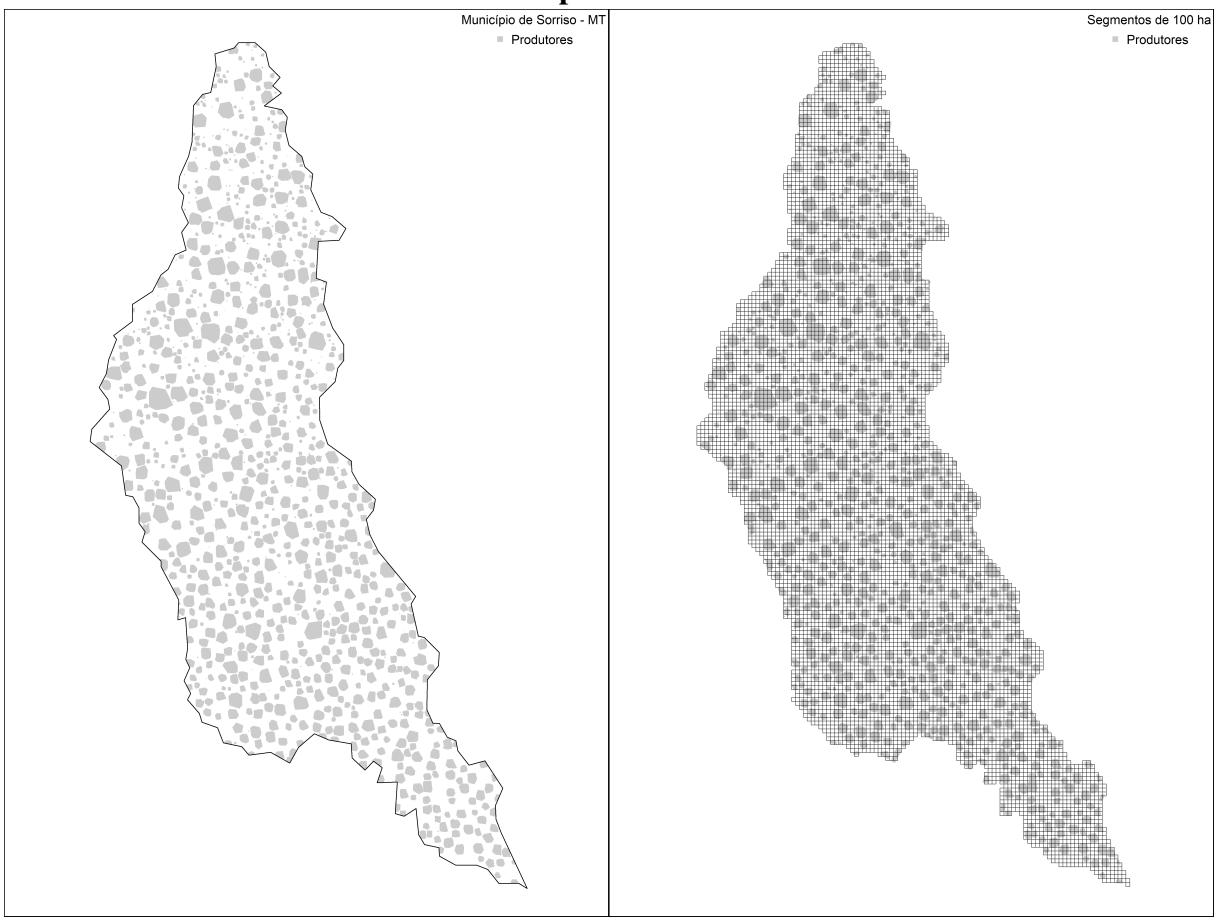
total de exploração, que foi inserido dentro dos limites geográficos de Sorriso sem que houvesse interseção com outros polígonos. A Figura 5 mostra a população de campos resultante desse procedimento, bem como o cadastro de área com segmentos quadrados de 100ha.

Também foram criados os dois cadastros de lista a partir da população artificial de 1.100 estabelecimentos. Os 300 produtores com as maiores áreas colhidas com soja formaram o cadastro L1, e os 250 produtores com os maiores rebanhos bovinos formaram o cadastro L2. Além disso, 99 produtores foram identificados tanto em L1 quanto em L2.

A figura 6 mostra os histogramas das variáveis área colhida com soja (kg) e rebanho bovino (em cabeças) para os produtores identificados no cadastro L1. Em ambos os gráficos, nota-se que alguns valores das variáveis de interesse estão mais afastadas do centro de massa. Além disso, os rebanhos bovinos de muitos produtores foram zero ou próximos de zero. Isso já era esperado uma vez que o cadastro L1 tem foco em grandes produtores de soja.

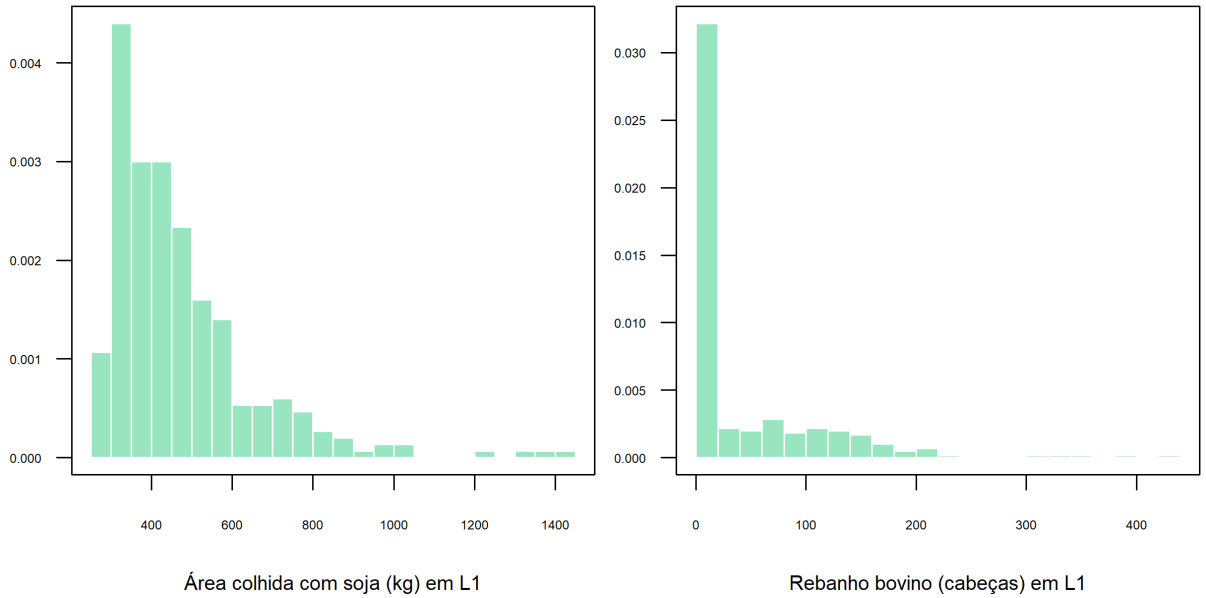
Os histogramas das variáveis de interesse para o cadastro L2 são mostrados na Figura 7. Embora esse cadastro tenha foco nos maiores rebanhos bovinos, muitos produtores

Figura 5 – População emulada de produtores e cadastro de segmentos quadrados de 100ha para Sorriso - MT



Fonte: De autoria própria

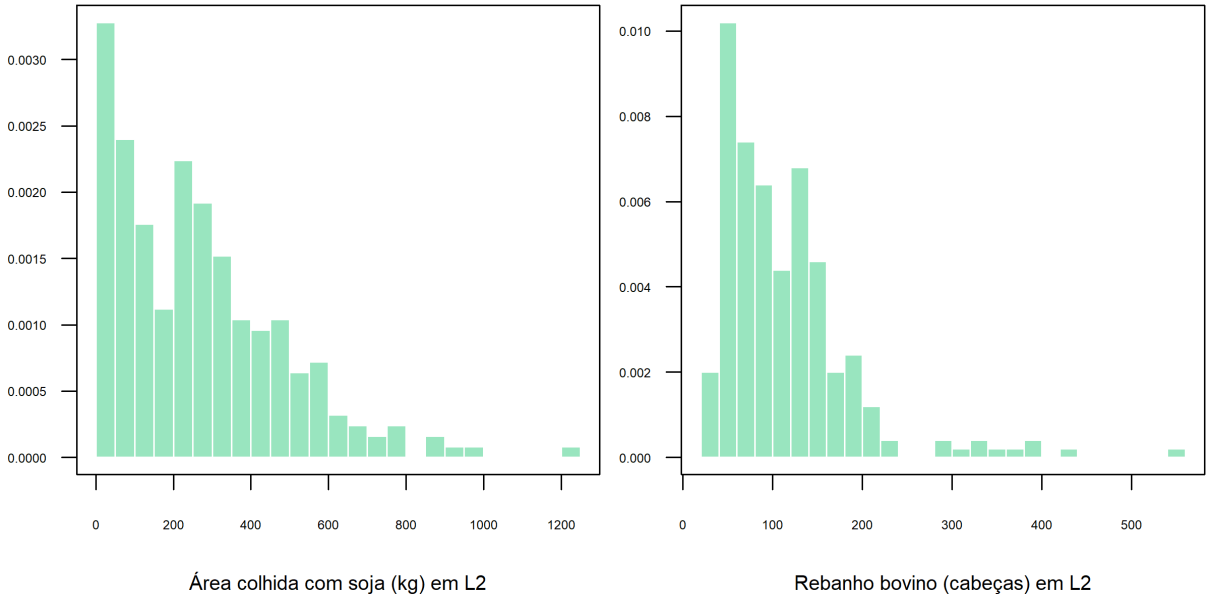
Figura 6 – Histogramas das variáveis área colhida com soja (kg) e rebanho bovino (cabeças) para os produtores identificados no cadastro L1



Fonte: De autoria própria

apresentaram áreas colhidas com soja. Nota-se também valores atípicos na distribuição empírica da variável rebanho bovino em L2.

Figura 7 – Histogramas das variáveis área colhida com soja (kg) e rebanho bovino (cabeças) para os produtores identificados no cadastro L2



Fonte: De autoria própria

6.4.3 Considerações de custo e alocação de amostra

Nesta proposta de aplicação, o custo foi predefinido para ser equivalente ao necessário para entrevistar 400 agricultores utilizando um cadastro do tipo lista. Nesse sentido, é necessário saber quantos produtores podem ser identificados nos segmentos, uma vez que são as unidades amostrais do cadastro de área. Para o cadastro de segmentos de 100ha descrito na seção (6.4.1), o número médio de produtores identificadas por segmento foi de $0,84$. Considerando que o tempo gasto para entrevistar um agricultor por meio de um cadastro de área equivale a realizar pelo menos duas entrevistas através de um cadastro de lista (FERRAZ, 2018), selecionar um segmento tem o custo aproximado de selecionar $1,68$ produtores através de um cadastro de lista.

Como base nos resultados apresentados por Ferraz, Mecatti & Torres (2022), foi considerada uma alocação de 55% do tamanho da amostra no cadastro de área, correspondente a seleção de $n_A = (0.55 \times 400)/1.68 = 131$ segmentos. Para os cadastros de lista a alocação foi proporcional ao tamanho do cadastro, com 25% para L1, correspondendo a $n_{L1} = 100$, e com 20% para L2, correspondendo a $n_{L2} = 80$.

Para avaliar o desempenho dos estimadores GEREG no contexto de pesquisas agropecuárias, foi realizado um estudo de Monte Carlo com 10.000 réplicas. Em cada réplica, um plano de seleção aleatório simples foi aplicado a cada cadastro. Além disso, cinco pontos foram usados para subamostrar os n_A segmentos selecionados nesse primeiro estágio.

6.4.4 Resultados e discussão

Considerando a abordagem de multiplicidade, os estimadores para os totais das variáveis de interesse são dados pela soma dos estimadores dos totais em cada cadastro. Para o cadastro de área foi utilizado um único estimador, denotado por $\hat{Y}_{Area}$. Os estimadores $\hat{Y}(L1)$ e $\hat{Y}(L2)$ são os estimadores por multiplicidade simples, expresso em (3.8), para os cadastros L1 e L2, respectivamente. Dessa forma, o estimador $\hat{Y} = \hat{Y}_{Area} + \hat{Y}(L1) + \hat{Y}(L2)$ não faz uso das informações auxiliares disponíveis e será usado como estimador de referência nos comparativos numéricos.

Os estimadores GEREG consideraram um modelo de regressão gama para estimar o total de área colhida com soja e um modelo de regressão de Poisson para estimar o rebanho bovino total. Os estimadores $\hat{Y}_{Gama}(L1)$, $\hat{Y}_{Poisson}(L1)$ e $\hat{Y}_{Greg}(L1)$ utilizaram a variável auxiliar disponível cadastro L1 (área com lavoura temporária na propriedade) para ajustar os modelos amostrais. Por sua vez, os estimadores $\hat{Y}_{Gama}(L2)$, $\hat{Y}_{Poisson}(L2)$ e $\hat{Y}_{Greg}(L2)$ utilizaram a variável auxiliar disponível cadastro L2 (área de pastagem na propriedade).

A Tabela 10 mostra os desempenhos de algumas combinações de estimadores usados para estimar o total de área colhida com soja. Nesse caso, o uso das informações auxiliares disponíveis no cadastro L2 não resultou em ganho de precisão para os estimadores de regressão, uma vez que esse cadastro lista os produtores com maiores rebanhos bovino. Por outro lado, os estimadores de regressão que utilizaram as informações auxiliares disponíveis no cadastro apresentaram melhores desempenhos se comparados ao estimador $\hat{Y}(L1)$, com eficiência relativa de 0.42 para o estimador $\hat{Y}_{Gama}(L1)$ e de 0.27 para o estimador $\hat{Y}_{Greg}(L1)$. A Figura 8 mostra as densidades empíricas dos estimadores do total de área colhida com soja considerando os cadastros L1 e L2. A distribuição empírica de $\hat{Y}_{Greg}(L1)$ mostra que as estimativas desse estimador são mais concentradas em torno do valor do parâmetro.

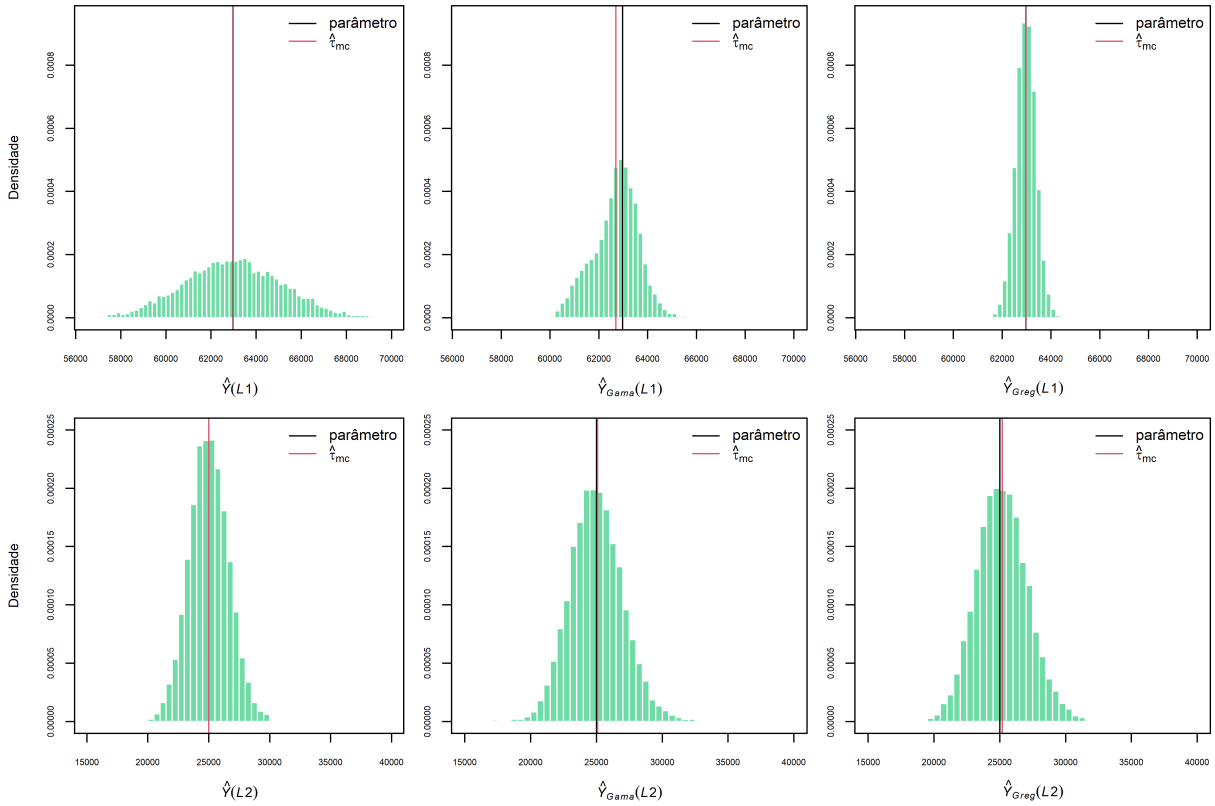
As combinações que apresentaram os melhores desempenhos, em termos da raiz do erro quadrático médio, utilizaram um estimador de regressão com o cadastro L1 e o estimador de multiplicidade simples com o cadastro L2, como, por exemplo, $\hat{Y} = \hat{Y}_{Area} + \hat{Y}_{Greg}(L1) + \hat{Y}(L2)$ e

Tabela 10 – Resultados para o total de área colhida com soja (em kg)

Estimador	CV (%)	Viés R (%)	DP	R.EQM	Cobertura (%)
$\hat{Y}_{Area}+\hat{Y}(L1)+\hat{Y}(L2)$	6.85	−0.358	14625.8	14628.7	96.1
$\hat{Y}_{Area}+\hat{Y}_{Gama}(L1)+\hat{Y}_{Gama}(L2)$	7.61	−0.456	16223.2	16248.1	93.9
$\hat{Y}_{Area}+\hat{Y}_{Gama^*}(L1)+\hat{Y}_{Gama^*}(L2)$	7.61	−0.456	16223.2	16248.1	93.9
$\hat{Y}_{Area}+\hat{Y}_{Gama^*}(L1)+\hat{Y}(L2)$	6.79	−0.491	14488.6	14497.8	95.5
$\hat{Y}_{Area}+\hat{Y}_{Greg}(L1)+\hat{Y}_{Greg}(L2)$	6.79	−0.284	14500.2	14491.4	94.5
$\hat{Y}_{Area}+\hat{Y}_{Greg}(L1)+\hat{Y}(L2)$	6.77	−0.374	14464.2	14456.8	95.4

Fonte: De autoria própria

Figura 8 – Densidades empíricas dos estimadores do total de área colhida com soja para os cadastros L1 e L2



Fonte: De autoria própria

$\hat{Y}=\hat{Y}_{Area}+\hat{Y}_{Gama}(L1)+\hat{Y}(L2)$. Além disso, o desempenho do estimador simplificado $\hat{Y}_{Gama^*}$ foi numericamente igual ao desempenho de $\hat{Y}_{Gama}$.

A Tabela 11 mostra os resultados para algumas combinações de estimadores utilizados para estimar o total de rebanho bovino. Semelhantemente ao observado para a estimação da área colhida com soja, a utilização da informação auxiliar disponível no cadastro L1 (área total com lavouras temporárias) não resultou em ganho de precisão para os estimadores de regressão. Além disso, em cerca de 5% das réplicas de Monte Carlo o estimador $\hat{Y}_{Poisson}(L1)$ apresentou estimativas muito elevadas, o que comprometeu o desempenho das combinações com esse estimador. Por outro lado, a utilização da variável auxiliar disponível no cadastro L2 resultou em estimadores de regressão mais precisos em relação a $\hat{Y}(L2)$. Os estimadores $\hat{Y}_{Poisson}$ e $\hat{Y}_{Greg}(L2)$

apresentaram uma eficiência em relação a $\hat{Y}_{Greg}(L2)$ de 0.76 e 0.74, respectivamente. Além disso, o desempenho do estimador simplificado $\hat{Y}_{Poisson^*}$ foi numericamente igual ao desempenho de $\hat{Y}_{Poisson}$.

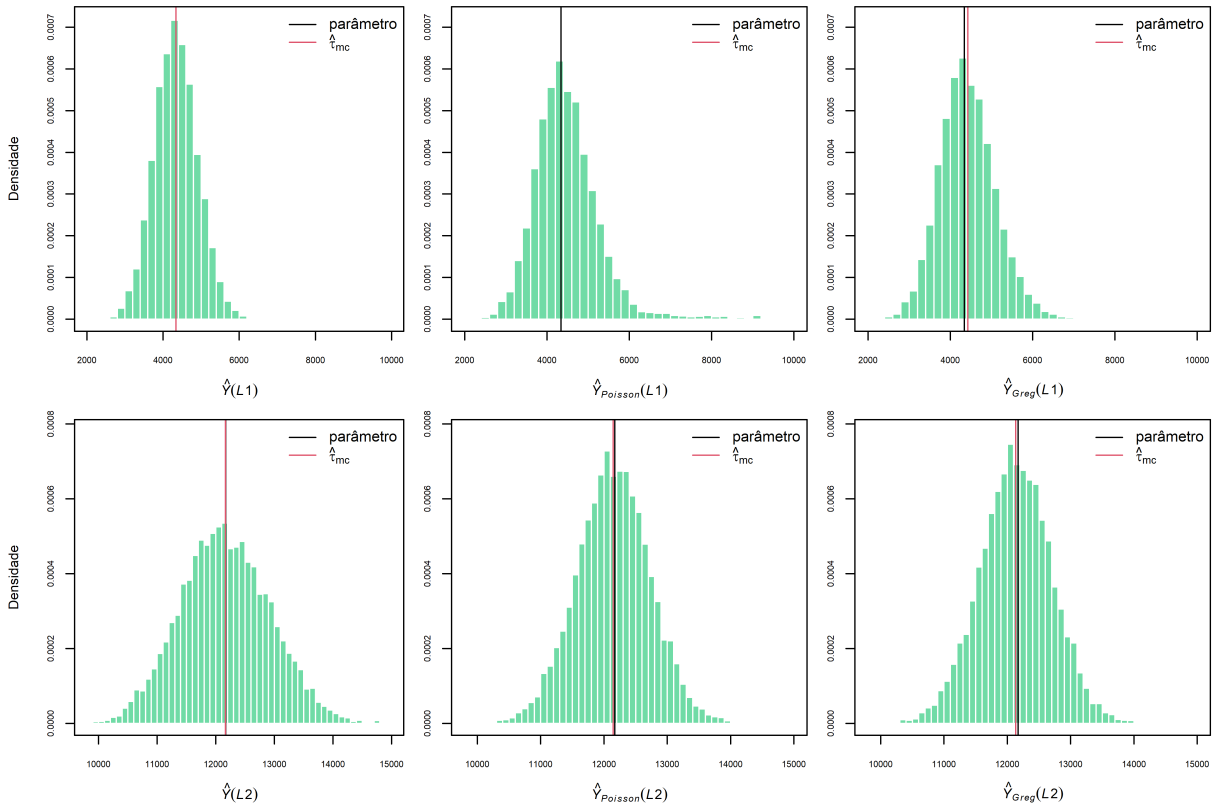
Tabela 11 – Resultados para o total de rebanho bovino (em cabeças)

Estimador	CV (%)	Viés R (%)	DP	R.EQM	Cobertura (%)
$\hat{Y}_{Area}+\hat{Y}(L1)+\hat{Y}(L2)$	13.56	−0.415	4408.1	4404.5	95.2
$\hat{Y}_{Area}+\hat{Y}_{Poisson}(L1)+\hat{Y}_{Poisson}(L2)$	18.87	1.933	6279.2	6308.7	88.2
$\hat{Y}_{Area}+\hat{Y}_{Poisson^*}(L1)+\hat{Y}_{Poisson^*}(L2)$	18.87	1.933	6279.2	6308.7	88.2
$\hat{Y}_{Area}+\hat{Y}(L1)+\hat{Y}_{Poisson^*}(L2)$	13.48	−0.474	4378.6	4370.3	93.1
$\hat{Y}_{Area}+\hat{Y}_{Greg}(L1)+\hat{Y}_{Greg}(L2)$	13.49	−0.255	4391.6	4381.8	90.0
$\hat{Y}_{Area}+\hat{Y}(L1)+\hat{Y}_{Greg}(L2)$	13.48	−0.517	4377.8	4370.2	93.1

Fonte: De autoria própria

A Figura 9 mostra as densidades empíricas dos estimadores do total de rebanho bovino para os cadastros L1 e L2. É possível identificar para $\hat{Y}_{Poisson}(L1)$ estimativas que se afastam do centro de massa distribuição empírica.

Figura 9 – Densidades empíricas dos estimadores do total de rebanho bovino para os cadastros L1 e L2



Fonte: De autoria própria

Em contraste com a estimação do total de área colhida com soja, as combinações que apresentaram os melhores desempenhos, em termos da raiz do erro quadrático médio, utilizaram

o estimador de multiplicidade simples com o cadastro L1 e um estimador de regressão com o cadastro L2. Nesse sentido, a combinação dos estimadores $\hat{Y} = \hat{Y}_{Area} + \hat{Y}_{Greg}(L1) + \hat{Y}(L2)$ e $\hat{Y} = \hat{Y}_{Area} + \hat{Y}_{Gama}(L1) + \hat{Y}(L2)$ apresentaram desempenhos aproximadamente iguais.

Por se tratar um estudo de Monte Carlo, nenhuma análise de diagnóstico foi realizada com os modelos ajustados em cada amostra. A depender das unidades selecionadas, uma observação pode causar efeito de alavanca ou se tornar um ponto aberrante. Além disso, foram reportados problemas de convergência algumas réplicas do experimento. É possível que os estimadores GEREG apresentem melhores desempenhos sob ajustes mais criteriosos.

7 CONSIDERAÇÕES FINAIS

Nesta tese, a utilização dos estimadores assistidos por modelos lineares generalizados foi expandida para o contexto de planos amostrais de múltiplos cadastros. Para tanto, foram apresentadas expressões para estes estimadores a partir das estratégias inferenciais de domínio e de multiplicidade, bem como a derivação das propriedades de centralidade e variância aproximadas. Além disso, a condição para que o termo de ajuste do estimador GEREG se anule foi expandida para o contexto de múltiplos cadastros.

Inicialmente, apresentou-se uma revisão dos principais conceitos acerca da estimação assistida em planos amostrais de cadastro único e uma breve revisão de duas estratégias inferenciais para planos amostrais de múltiplos cadastros. Esses conceitos foram necessários para a proposição dos estimadores, bem como a derivação das propriedade de centralidade e variância aproximadas.

O problema da estimação dos parâmetros dos modelos regressão linear generalizados foi abordado sob a perspectiva de planos amostrais de múltiplos cadastros. Os parâmetros dos modelos foram estimados por meio da maximização de uma função de pseudo log-verossimilhança que considerou o plano de múltiplos cadastros em sua estrutura. Isso diferencia os estimadores GEREG postulados no contexto de planos amostrais de cadastro único dos estimadores GEREG desenvolvidos para planos de múltiplos cadastros.

O desempenho dos estimadores GEREG foi avaliado por meio de um estudo de simulação, no qual foram elaborados cenários com variáveis respostas baseadas em três distribuições pertencentes à família exponencial. Os resultados das simulações mostraram que, ao considerar a natureza da variável de interesse na fase de concepção do modelo, o estimador GEREG demonstra melhor desempenho em comparação com o estimador GREG, apresentando uma redução da variância e do erro quadrático médio. Além disso, foi apresentada uma proposta de aplicação dos estimadores GEREG em pesquisas agropecuárias que utilizam um cadastro de área e dois cadastros de lista.

Esta tese abre caminho para outras pesquisas acerca da estimação assistida por modelos lineares generalizados no âmbito de planos amostrais de múltiplos cadastros. Dentre as possíveis propostas de pesquisas futuras, faz-se menção dos seguintes tópicos:

- Propor outros métodos de determinar o valor de α para os estimadores GEREG no contexto de cadastro duplo com a abordagem de domínios.
- Estudar outras técnicas que sejam factíveis para estimar a variância do estimador GEREG.

- Avaliar os desempenhos dos estimadores GEREG sob a abordagem de multiplicidade simples a medida que o número de cadastros é inflacionado.

REFERÊNCIAS

BREIDT, F. J.; OPSOMER, J. D. Model-assisted survey estimation with modern prediction techniques. 2017.

CARFAGNA, E. Multiple frame sample surveys: advantages, disadvantages and requirements. **Bulletin of the International Statistical Institute, 53rd Session, Proceedings, Book**, v. 1, p. 271–274, 2001.

CHEN, S.; KIM, J. K. Population empirical likelihood for nonparametric inference in survey sampling. **Statistica Sinica**, JSTOR, v. 24, n. 1, p. 335–355, 2014.

COÊLHO, H. F. C. **A abordagem de cadastro duplo (Dual Frame): estimação assistida por modelos Lineares com aplicação em pesquisas agropecuárias**. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2007.

COÊLHO, H. F. C. **Inferência sob planos amostrais de cadastro duplo**. Tese (Doutorado) — Universidade Federal de Pernambuco, 2011.

CORDEIRO, G. M.; DEMÉTRIO, C. G. Modelos lineares generalizados e extensões. **Piracicaba: USP**, p. 31, 2008.

COTTER, J.; NEALON, J. **Area frame design for agricultural surveys**. [S.l.]: US Department of Agriculture, National Agricultural Statistics Service Washington, DC, 1987.

DEVILLE, J.-C.; SÄRNDAL, C.-E. Calibration estimators in survey sampling. **Journal of the American statistical Association**, Taylor & Francis, v. 87, n. 418, p. 376–382, 1992.

FERRAZ, C. **Global Strategy to improve Agricultural and Rural Statistics. Brazil's Master Sampling Frame Experiments**. [S.l.], 2018.

FERRAZ, C.; MECATTI, F.; TORRES, J. Dual frame design in agricultural surveys: reviewing roots and methodological perspectives. **Statistical Methods & Applications**, Springer, p. 1–25, 2022.

FULLER, W. A.; BURMEISTER, L. F. Estimators for samples selected from two overlapping frames. In: **Proceedings of the Social Statistics Section, American Statistical Association**. [S.l.: s.n.], 1972. v. 245249.

GALLEGO, F. J. Stratified sampling of satellite images with a systematic grid of points. **ISPRS Journal of Photogrammetry and Remote Sensing**, Elsevier, v. 59, n. 6, p. 369–376, 2005.

GALLEGO, F. J.; DELINCÉ, J.; CARFAGNA, E. Two stage area frame sampling on square segments for farm surveys. **Survey Methodology**, v. 20, n. 2, p. 107–115, 1994.

GALLEGO, F. J. *et al.* **Sampling frames of square segments**. [S.l.]: Office for Official Publ. of the European Communities, 1995.

HARTLEY, H. O. Multiple frame surveys. In: WASHINGTON, DC. **Proceedings of the social statistics section, American Statistical Association**. [S.l.], 1962. v. 19, n. 6, p. 203–206.

HARTLEY, H. O. Multiple frame methodology and selected applications. **Sankhya**, v. 36, n. 997, p. 118, 1974.

LEHTONEN, R.; VEIJANEN, A. Logistic generalized regression estimators. **Survey Methodology**, Statistics Canada, v. 24, p. 51–56, 1998.

LUMLEY, T.; SCOTT, A. Fitting regression models to survey data. **Statistical Science**, JSTOR, p. 265–278, 2017.

MCCULLAGH, P.; NELDER, J. Binary data. In: **Generalized linear models**. [S.l.]: Springer, 1989. p. 98–148.

MECATTI, F. A single frame multiplicity estimator for multiple frame surveys. **Survey methodology**, v. 33, n. 2, p. 151–157, 2007.

MOLINA, D.; RUEDA, M. d. M.; ARCOS, A.; RANALLI, M. G. Multinomial logistic estimation in dual frame surveys. **SORT**, Institut d'Estadística de Catalunya, v. 39, n. 2, p. 309–336, 2015.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2017. Disponível em: <https://www.R-project.org/>.

RANALLI, M. G.; ARCOS, A.; RUEDA, M. d. M.; TEODORO, A. Calibration estimation in dual-frame surveys. **Statistical Methods & Applications**, Springer, v. 25, p. 321–349, 2016.

RONDON, L. M.; VANEGAS, L. H.; FERRAZ, C. Finite population estimation under generalized linear model assistance. **Computational Statistics & Data Analysis**, Elsevier, v. 56, n. 3, p. 680–697, 2012.

RUEDA, M. d. M.; ARCOS, A.; MOLINA, D.; RANALLI, M. G. Estimation techniques for ordinal data in multiple frame surveys with complex sampling designs. **International Statistical Review**, Wiley Online Library, v. 86, n. 1, p. 51–67, 2018.

RUEDA, M. del M.; RANALLI, M. G.; ARCOS, A.; MOLINA, D. Population empirical likelihood estimation in dual frame surveys. **Statistical Papers**, Springer, v. 62, p. 2473–2490, 2021.

SÄRNDAL, C.-E.; SWENSSON, B.; WRETMAN, J. **Model assisted survey sampling**. [S.l.]: Springer Science & Business Media, 2003.

SINGH, A. Combining information in survey sampling by modified regression. In: **Proceedings of the Section on Survey Research Methods, American Statistical Association**. [S.l.: s.n.], 1996. v. 91, n. 1, p. 120–129.

SINGH, A. C.; MECATTI, F. Generalized multiplicity-adjusted horvitz-thompson estimation as a unified approach to multiple frame surveys. **Journal of official statistics**, v. 27, n. 4, p. 633, 2011.

SKINNER, C. J.; RAO, J. N. Estimation in dual frame surveys with complex designs. **Journal of the American Statistical Association**, Taylor & Francis, v. 91, n. 433, p. 349–356, 1996.

TILLÉ, Y.; MATEI, A. **sampling: Survey Sampling**. [S.l.], 2016. R package version 2.8. Disponível em: <https://CRAN.R-project.org/package=sampling>.

TORRES, J. E. M. M. **Amostragem de área e aplicações em agropecuária**. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2018.

WU, C.; SINGH, A. C. An extension of generalized regression estimator to dual frame surveys. In: JOINT STATISTICAL MEETINGS - SECTION ON SURVEY RESEARCH METHODS. [S.l.], 2003.

YAN, L.; YE, F.; ZHANG, G. Nonparametric regression estimators in dual frame surveys. **Communications in Statistics - Simulation and Computation**, Taylor & Francis, v. 50, n. 3, p. 854–864, 2021. Disponível em: <https://doi.org/10.1080/03610918.2019.1568474>.