ESTIMAÇÃO EM POPULAÇÕES FINITAS ASSISTIDA POR MODELOS PARA VARIÁVEIS DICOTÔMICAS

LUZ MARINA RONDÓN POVEDA

Orientador: Prof. Dr. Cristiano Ferraz
Co-orientadora: Prof. Dra. Carla Almeida Vivacqua

Área de Concentração: Estatística Aplicada

Dissertação submetida como requerimento parcial para obtenção do grau de Mestre em Estatística pela Universidade Federal de Pernambuco

Recife/PE
Dezembro de 2006

Rondón Poveda, Luz Marina

Estimação em populações finitas assistida por modelos para variáveis dicotômicas / Luz Marina Rondón Poveda. – Recife : O Autor, 2006.

x, 130 folhas : il., fig., quadros.

Dissertação (mestrado) – Universidade Federal de Pernambuco. CCEN. Estatística, 2006.

Inclui bibliografia e apêndices.

1. Estatística aplicada – Amostragem. 2. Estimadores de regressão, GREG (Generalized Regression Estimator) e LGREG (Logistic Generalized Regression Estimator) – Estratificação - Estimadores separados e combinados. 3. Pseudo-verossimilhança – Variáveis dicotômicas – Estimação. I. Título.

311.213.2	**CDU (2.ed.)**	**UFPE**
519.52	**CDD (22.ed.)**	**BC2006 – 581**

Universidade Federal de Pernambuco
Pós-Graduação em Estatística

20 de dezembro de 2006
(data)

Nós recomendamos que a dissertação de mestrado de autoria de

Luz Marina Rondón Poveda

intitulada

"Estimação em populações finitas assistidas por modelos para variáveis dicotômicas"

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

Coordenador da Pós-Graduação em Estatística

Prof. Klaus L. P. Vasconcellos
Coordenador da
Pós-graduação em
Estatística, UFPE

Banca Examinadora:

Cristiano Ferraz orientador

Dione Maria Valença (UFRN)

Getúlio José de Amorim Amaral

Este documento será anexado à versão final da dissertação.

Ao grande amor da minha vida, Luis Hernando,

e à minha mãe, Alicia.

Agradecimentos

Quero agradecer ...

A Deus pela minha vida e pelas forças para seguir o caminho que às vezes parecia muito difícil.

Ao meu esposo, Luis Hernando, por me ensinar que é maior a pessoa que se levanta depois de escorregar, enquanto caminhava, que aquela que não se atreveu a caminhar para não escorregar. Por estar sempre com os braços abertos e um bom conselho no momento oportuno, pela compreensão, paciência, atenção, incentivo, ajuda, carinho e apoio incondicional por ele sempre oferecidos. Enfim, por todos os momentos de alegria e amor que me tem dedicado.

Aos meus pais, Noe e Alicia, pela educação, carinho e apoio, em especial a minha mãe, pelo seu imensurável esforço e dedicação.

Aos meus irmãos, Jeisson pelo carinho e Lizbeth pelos momentos de alegria e comprenssão que tem me proporcionado.

Ao meu orientador Cristiano Ferraz, pela oportunidade concedida, confiança, apoio, incentivo, disponibilidade, competência, paciência, e excelente orientação.

Ao Programa de Mestrado em Estatística da Universidade Federal de Pernambuco, pela oportunidade e pelo apoio a mim concedidos, que me permitiram realizar o mestrado neste maravilhoso país, e em especial, aos seus coorde-

nadores, os professores Francisco Cribari Neto e Klaus Vasconcellos.

Aos professores do Programa de Mestrado em Estatística da Universidade Federal de Pernambuco, pela sua contribuição na minha formação pessoal, acadêmica e profissional.

As minhas amigas, Luisa Fernanda e Rossemary, pelo incentivo, carinho e amizade.

Aos meus colegas do mestrado pela convivência nestes dois anos, em especial, Rejane Brito e Hemílio Fernandes, pela amizade, companhia e atenção que me brindaram.

A Themis Abensur, pela convivência, companhia, amizade, as longas conversações e momentos de diversão.

A Valeria Bittencourt, pelo carinho e por ser muito competente em seu trabalho.

Aos professores Yves Tillé e Pierre Duchesne, pela colaboração na disposição de materiais que contribuíram no enriquecimento deste trabalho.

A todas as pessoas que não mencionei e sempre me acompanharam no caminho, estão no meu coração.

À banca de examinadores pelas valiosas sugestões que contribuiram e enriqueceram a qualidade deste trabalho.

À CAPES, pelo apoio financeiro.

Resumo

Neste trabalho é discutida a estimação de proporções em populações finitas assistida por modelos. A teoria envolvendo estimadores de regressão linear generalizados é revista, sob uma abordagem proposta de estimadores assistidos por modelos da família exponencial. O trabalho de Tillé (1998), que deriva o estimador de regressão via probabilidades condicionais de inclusão na amostra, é revisto juntamente com o de Lehtonen e Veijanen (1998), que propõem o estimador de regressão generalizado logístico (LGREG), num contexto de amostra aleatória simples. A aplicação dos estimadores LGREG num cenário de amostragem estratificada é discutida e formas para estimadores LGREG separado e combinado são propostas. As propriedades dos estimadores propostos são investigadas através de um estudo de simulação Monte Carlo, envolvendo os planos de amostragem aleatória simples, de Bernoulli e estratificado.

Palavras-chave: Estimador de regressão generalizado logístico (LGREG), pseudo-verossimilhança, estimador de regressão combinado e separado.

Abstract

In this work, we discuss finite population proportion estimation under a model-assisted approach. The generalized linear regression estimator theory is revisited under a proposed setup of exponential family model-assisted estimators. The work by Tillé (1998), which derives the regression estimator via conditional sample inclusion probabilities is reviewed as well as the work by Lehtonen and Veijanen (1998), which propose the logistic generalized regression estimator (LGREG), under simple random sample. We discuss the application of LGREG estimators under a stratified sample design and propose the forms of a separate and combined LGREG estimators. The statistical properties of all the proposed estimators are investigated through a Monte Carlo simulation study involving simple random sample, Bernoulli sample and stratified sample designs.

Key Words: Logistic generalized regression estimator (LGREG), pseudo-likelihood, combined and separate regression estimator.

Sumário

Lista de Quadros

CAPÍTULO 1

Introdução

A estimação de parâmetros referentes a uma ou mais variáveis de interesse em uma população finita é abordada pela teoria estatística de amostragem. Nesta área, é possível identificar duas etapas no processo de inferência, relacionadas entre si: a de planejamento amostral e a de estimação.

Nesta dissertação, define-se como etapa de planejamento amostral aquela que engloba estudos para identificar o melhor plano e esquema amostral probabilísticos, incluindo a seleção dos indivíduos que comporão a amostra. Ainda nesta etapa são conduzidos estudos que dão suporte à escolha de estimadores a serem utilizados. A etapa de estimação é aquela na qual são obtidas as estimativas dos parâmetros de interesse, através dos estimadores escolhidos, bem como as estimativas das variâncias desses estimadores, a partir da amostra selecionada.

A qualidade estatística da inferência em uma população finita depende da adoção de uma estratégia adequada de amostragem, definida como a escolha de ambos, plano amostral e estimador. Por este motivo, os esforços dos estatísticos envolvidos em levantamentos amostrais concentram-se na procura de planos que minimizem variações amostrais e estimadores que apresentem baixo erro quadrático médio.

A procura por uma boa estratégia de amostragem envolve necessariamente esforços para identificar toda informação possível de se obter a respeito da população sob estudo, na etapa de planejamento amostral. Tais informações dizem respeito a variáveis comumente chamadas na literatura de variáveis

auxiliares (Cochran, 1977; Särndal, Swensson e Wretman, 1992; Lohr, 1999). Variáveis auxiliares podem ser utilizadas para reduzir a variância do estimador de Horvitz-Thompson (Horvitz e Thompson, 1952) quando são empregadas no plano ou esquema amostral. Exemplos que ilustram tal situação incluem o uso de estratificação e de esquemas amostrais com probabilidades de inclusão na amostra proporcionais ao tamanho da variável auxiliar. Uma outra forma de utilizar variáveis auxiliares é incorporá-las à forma do estimador a ser utilizado. Os estimadores assim obtidos são denominados estimadores de regressão generalizados. Nessa dissertação será adotada a abreviação GREG, do inglês *generalized regression estimator*, para referir-se a estes estimadores.

Vários autores apresentam os estimadores de regressão generalizados sob a abordagem de estimação assistida por modelos (Särndal, Swensson e Wretman, 1992, pág.219; Lohr, 1999, pág.372; Särndal, 2001). Através dessa abordagem, um modelo de regressão é utilizado apenas para descrever a relação entre as variáveis de interesse e as auxiliares na população finita.

Teoricamente, quanto maior for a adequação do modelo para descrever a relação entre essas variáveis, maior será a eficiência do GREG em comparação com o estimador de Horvitz-Thompson, que não usa informação auxiliar em sua forma funcional. Uma abordagem menos difundida para derivar estimadores GREG é a apresentada por Tillé (1998), que utiliza probabilidades condicionais de inclusão na amostra.

Em diversas situações é de interesse estimar a proporção de indivíduos da população sob estudo, que possuem determinada característica. Nesse caso, a variável de interesse pode ser vista como uma variável dicotômica assumindo valores 1 (um), quando o indivíduo da população possui a característica e 0 (zero), caso contrário. Apesar de ser possível utilizar estimadores GREG nesse contexto, a relação entre a variável de interesse e possíveis variáveis auxiliares é melhor descrita através de um modelo de regressão logística. O estimador resultante da assistência de tal modelo foi originalmente proposto por Lehtonen e Veijanen (1998), para o caso de uma amostra aleatória simples e denominado estimador de regressão generalizado

logístico, ou, abreviando, LGREG, do inglês, *logistic generalized regression estimator*.

Esta dissertação tem como objetivo geral apresentar uma revisão de literatura envolvendo estimadores do tipo regressão e propor estimadores de regressão assistidos por modelos pertencentes à família exponencial, envolvendo assim modelos lineares e não-lineares. Os estimadores que usam estes modelos no processo de estimação ainda serão chamados neste trabalho de estimadores de regressão generalizados (GREG), por conveniência e adequação, embora que, em livros como Särndal, Swensson e Wretman (1992) e Lohr (1999), estimadores GREG sejam apresentados como sendo assistidos só por modelos lineares. Esta dissertação também visa estudar as propriedades do estimador LGREG e discutir possibilidades de sua aplicação no contexto de planos amostrais estratificados. Os objetivos específicos são: contribuir para a divulgação da abordagem de probabilidades condicionais de inclusão como uma forma alternativa de derivação do estimador GREG; investigar as propriedades estatísticas do estimador LGREG, no caso de amostragem aleatória simples e Bernoulli, através de estudos de simulação Monte Carlo; propor como aplicar e estudar as propriedades estatísticas do LGREG, no caso de uma amostra aleatória estratificada, através de estudos de simulação Monte Carlo.

Os trabalhos desenvolvidos são apresentados ao longo de 8 capítulos. No capítulo 2 são apresentados os conceitos básicos de amostragem e os modelos da família exponencial, que neste trabalho serão usados para assistir a estimação de parâmetros em populações finitas.

No capítulo 3 é proposto o estimador de regressão generalizado (GREG) assistido por modelos pertencentes à família exponencial, apresentando as suas principais propriedades e características, discutindo-se as possíveis aplicações dos GREG's no contexto de estratificação. Além disso, considera-se como caso particular desta classe de estimadores os estimadores assistidos por modelos de regressão lineares.

No capítulo 4 é mostrado que o estimador de regressão pode ser obtido

usando as probabilidades de inclusão condicionais segundo o enfoque desenvolvido por Tillé (1998).

No capítulo 5, é definido o estimador de regressão generalizado logístico (LGREG), suas propriedades e características mais importantes. É apresentada também a estimação de proporções usando os estimadores GREG assistidos por um modelo de regressão linear e o LGREG, por um modelo de regressão logística. Além disso, são discutidas as possíveis aplicações do estimador LGREG no contexto de estratificação.

No capítulo 6, são apresentados estudos de simulação desenvolvidos com o objetivo de avaliar e comparar as propriedades dos estimadores Horvitz-Thompson, GREG e LGREG no caso em que o parâmetro de interesse é uma proporção.

No capítulo 7, ilustra-se a aplicação dos estimadores GREG's usando um subconjunto de dados da Pesquisa Mensal de Emprego (PME), realizada pelo IBGE, no mês de outubro do ano 2005, usando o pacote estatístico SAS. Além disso, no apêndice D, é apresentado um relato de como utilizar o PROC SURVEYLOGISTIC do pacote SAS, no contexto de estimação assistida por modelos. Para terminar, no capítulo 8 são apresentadas as considerações finais deste trabalho.

CAPÍTULO 2

Noções Básicas de Amostragem e Modelos da Família Exponencial

2.1 Noções Básicas de Amostragem

Considere $U = \{1, 2, \ldots, N\}$, o conjunto dos índices que identificam os elementos que compõem a população finita, de tamanho N, e S um subconjunto de U chamado de amostra ($S \subset U$).
A amostra S é considerada ser probabilística se são satisfeitas as seguintes condições:

i) É possível definir o conjunto $\zeta = \{S_1, \ldots, S_T\}$ de todas as amostras possíveis que podem ser selecionadas da população seguindo um plano amostral $p(\cdot)$, chamado de espaço amostral.

ii) O mecanismo de escolha da amostra deve dar uma probabilidade maior que zero para cada elemento da população.

iii) A seleção da amostra deve ser aleatória, ou seja, o processo de seleção das amostras tem que associar a cada amostra possível S uma probabilidade exata de seleção $p(S)$.

iv) É possível identificar para cada uma das amostras que pertencem a ζ a probabilidade de serem selecionadas $p(S)$.

Denote por y uma variável de interesse na população, e y_k o valor dessa variável referente ao indivíduo k. Denote ainda por $\pi_k = P(k \in S)$ e $\pi_{kl} =$

$P(k, l \in S)$ as probabilidades de inclusão de primeira e segunda ordem, respectivamente.

Por simplicidade, considere o objetivo de estimar um parâmetro unidimensional $\theta = \theta(1, \ldots, k, \ldots, N)$ através de um estimador $\hat{\theta} = \hat{\theta}(k \in S)$. O total e a média populacional dados por $t_y = \sum_{k \in U} y_k$, e $\bar{y}_U = N^{-1} t_y$, respectivamente, são exemplos freqüentes de parâmetros de interesse, que acomodam variáveis contínuas e discretas.
Quando a variável de interesse é de tipo dicotômico, por exemplo, é conveniente definir

$$y_k = \begin{cases} 1, & \text{se o atributo está presente no } k\text{-ésimo indivíduo;} \\ 0, & \text{caso contrário.} \end{cases}$$

Dessa forma, t_y representa o total de elementos na população que possuem o atributo de interesse e $\bar{y}_U = P = \frac{t_y}{N}$ a proporção populacional com o atributo desejado.

O estimador de Horvitz-Thompson para t_y é dado pela seguinte expressão

$$\hat{t}_\pi = \sum_{k \in S} \frac{y_k}{\pi_k}.$$

É possível mostrar facilmente que este é um estimador não-viesado, sua variância pode ser expressa por

$$V_p(\hat{t}_\pi) = \sum_{k \in U} \sum_{l \in U} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l},$$

onde $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$ com $\pi_{kl} > 0$ para todo $k, l \in U$, e um estimador não-viesado para $V(\hat{t}_\pi)$ é dado por

$$\hat{V}_p(\hat{t}_\pi) = \sum_{k \in S} \sum_{l \in S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}.$$

Além do estimador de Horvitz-Thompson, nesta dissertação serão estudados outros estimadores. Para avaliar a qualidade de um estimador é necessário conhecer as suas propriedades estatísticas do ponto de vista do plano amostral. Por este motivo, as seguintes propriedades são revisadas:

- A esperança de $\hat{\theta}$, $E_p(\hat{\theta})$ é dada por

$$E_p(\hat{\theta}) = \sum_{S \in \zeta} p(S)\hat{\theta}(S), \tag{2.1}$$

  onde $p(S)$ denota a probabilidade de selecionar a amostra S da população.

- A variância de $\hat{\theta}$ dada por

$$V_p(\hat{\theta}) = \sum_{S \in \zeta} p(S)\{\hat{\theta}(S) - E_p(\hat{\theta})\}^2. \tag{2.2}$$

- O viés é a diferença entre a média da distribuição amostral e o valor verdadeiro do parâmetro, ou seja,

$$B_p(\hat{\theta}) = E_p(\hat{\theta}) - \theta.$$

  Quando $B_p(\hat{\theta}) = 0$, o estimador $\hat{\theta}$ é dito ser um estimador não-viesado para θ.

- O erro quadrático médio é uma medida que pode ser expressa como

$$\begin{aligned} EQM_p(\hat{\theta}) &= \sum_{S \in \zeta} p(S)(\hat{\theta}(S) - \theta)^2 = E_p(\hat{\theta} - \theta)^2 \\ &= V_p(\hat{\theta}) + B_p^2(\hat{\theta}). \end{aligned}$$

Quando é de interesse obter uma estimação intervalar do parâmetro θ e não há informação sobre $V_p(\hat{\theta})$, recorre-se ao estimador $\hat{V}_p(\hat{\theta})$. Além disso, se as condições que atendem a um Teorema Central do Limite como o de Hájek (1960) são satisfeitas é possível construir o seguinte intervalo de confiança:

$$\hat{\theta} \pm z_{1-\alpha/2}\sqrt{\hat{V}_p(\hat{\theta})}, \tag{2.3}$$

sendo $z_{1-\alpha/2}$ uma constante tal que $P(Z > z_{1-\alpha/2}) = \alpha/2$, com $Z \sim N(0,1)$ e $100(1-\alpha)\%$ o nível de confiança desejado para o intervalo.
A qualidade do estimador intervalar (2.3) para θ pode ser medida através da taxa de cobertura, dada pela seguinte expressão

$$TC(\hat{\theta}, \hat{V}(\hat{\theta}), \alpha) = \frac{\sum_{S \in \zeta} Z(S)}{T}, \tag{2.4}$$

em que T é o número total de amostras possíveis que podem ser selecionadas da população e

$$Z(S) = \begin{cases} 1, & \text{se} \quad \theta \in \left(\hat{\theta}(S) \pm z_{1-\alpha/2}\sqrt{\hat{V}_p(\hat{\theta})}\right); \\ 0, & \text{caso contrário.} \end{cases}$$

Uma outra medida de qualidade é o coeficiente de variação de $\hat{V}(\hat{\theta})$, dado por

$$CV(\hat{V}(\hat{\theta})) = 100\frac{\sqrt{V(\hat{V}(\hat{\theta}))}}{E(\hat{V}(\hat{\theta}))}.$$

As vezes é de interesse comparar vários estimadores para o mesmo problema de estimação e sob o mesmo plano amostral. Nesse caso, deve ser considerada uma medida que compare a eficiência obtida com cada estimador, com a intenção de fazer a escolha apropriada. A eficiência relativa de um estimador pode ser medida usando a seguinte expressão

$$eff(\hat{\theta}_1, \hat{\theta}_2) = \frac{V(\hat{\theta}_1)}{V(\hat{\theta}_2)}. \tag{2.5}$$

Se $eff(\hat{\theta}_1, \hat{\theta}_2)$ é inferior, igual ou superior a 1, é dito que $\hat{\theta}_1$ é mais, igualmente ou menos eficiente que $\hat{\theta}_2$, respectivamente. Nesta dissertação, um dos planos utilizados é o de Bernoulli, que será descrito a seguir.

2.1.1 Amostragem de Bernoulli

Um plano amostral BE consiste em uma série de experimentos independentes, um para cada elemento da população. O plano atribui probabilidade igual de seleção, π e de não seleção $(1 - \pi)$, a cada elemento da população. Neste plano, o tamanho da amostra, denotado por n_S, é uma variável aleatória. Sob um plano BE, tem-se que

$$p(S) = \pi^{n_S}(1 - \pi)^{N-n_S},$$

em que $\pi_k = \pi$ e $\pi_{kl} = \pi^2$ são as probabilidades de inclusão de primeira e segunda ordem, respectivamente. Um esquema amostral para selecionar uma amostra seguindo um plano BE é o seguinte:

Passo 1. Considere um valor para π $(0 < \pi < 1)$.

Passo 2. Denote por $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$, uma série de N realizações de uma distribuição uniforme $(0, 1)$.

Passo 3. Se $\varepsilon_k \leq \pi$, então, o elemento k é selecionado para compor a amostra S.

Passo 4. Repetir o procedimento anterior com cada elemento da população.

2.2 Modelos da Família Exponencial

Estes modelos são muito usados na prática (veja McCullagh e Nelder, 1989; Wei, 1998) pois com eles é possível analisar estatisticamente conjuntos de dados com resposta discreta, como nos modelos binomial e poisson, e com resposta contínua restrita ao intervalo $(0, \infty)$, como nos modelos gamma e normal inversa. Além disso, os modelos da família exponencial proporcionam grande flexibilidade para a especificação da relação entre a variável resposta e as variáveis explicativas, pois nestes modelos é assumida a existência de uma função que relaciona a média da variável resposta e o preditor. Os modelos normais lineares e não-lineares fazem parte desta classe de modelos de regressão.

2.2.1 Definição

Sejam $Y_1, \ldots, Y_k, \ldots, Y_n$ um conjunto de variáveis aleatórias independentes cada uma seguindo uma distribuição de probabilidade pertencente à família exponencial. A função de densidade de Y_k (função de probabilidade no caso discreto) pode ser expressa como

$$f(y; \theta_k, \phi_k) = \exp\{\phi_k[y\theta_k - b(\theta_k)] + c(y, \phi_k)\}, \tag{2.6}$$

onde $c(\cdot)$ é uma função conhecida, $\mathrm{E}(Y_k) = \mu_k = b'(\theta_k)$, $\mathrm{Var}(Y_k) = \phi_k^{-1} V_k$, $V_k = \partial\mu_k/\partial\theta_k$ é a função de variância e $\phi_k^{-1} > 0$ é o parâmetro de dispersão. A função de variância determina, de forma biunívoca, a classe correspondente de distribuições. Essa propriedade é muito importante pois permite a

comparação de distribuições através de um teste simples para a função de variância (Jørgensen, 1987). Os modelos da família exponencial são definidos por (2.6) e pela componente sistemática

$$g(\mu_k) = \eta_k = h(\boldsymbol{\beta}; \mathbf{x}_k), \tag{2.7}$$

onde $\boldsymbol{\beta}$ é um vetor de parâmetros desconhecidos, $\mathbf{x}_k = (x_{k1}, \ldots, x_{kJ})$ um vetor de variáveis explicativas para o indivíduo k, $h(\cdot; \mathbf{x}_k)$ uma função contínua, duplamente diferenciável e $g(\cdot)$ uma função monótona e diferenciável, denominada função de ligação. Quando a função $g(\cdot)$ é tal que $\theta_k = \eta_k$ então esta função é chamada de ligação canônica. No Quadro 2.1 apresentam-se algumas das distribuições da família exponencial. Além das distribuições do Quadro 2.1, como exemplos típicos desta classe podem-se citar os modelos logit, probit e loglinear.

Quadro 2.1. Principais distribuições pertencentes à família exponencial.

Distribuição	$b(\theta)$	Ligação Canônica	ϕ	$V(\mu)$
Normal	$\theta^2/2$	μ	$1/\sigma^2$	1
Poisson	e^θ	$\log \mu$	1	μ
Bernoulli	$\log(1+e^\theta)$	$\log\{\mu/(1-\mu)\}$	1	$\mu(1-\mu)$
Gama	$-\log(-\theta)$	$-1/\mu$	$1/(CV)^2$	μ^2
N. Inversa	$-\sqrt{-2\theta}$	$-1/2\mu^2$	ϕ	μ^3

2.2.2 Estimação dos Parâmetros do Modelo

Os modelos da família exponencial podem ser usados para assistir a estimação de parâmetros em populações finitas. Nesse caso, eles são usados apenas para descrever as relações entre as variáveis de interesse e auxiliares, sendo importante identificar as diferenças entre μ_k, $\hat{\mu}_k^U$ e $\hat{\mu}_k^S$. Assim, μ_k refere-se ao parâmetro do modelo formulado, o qual é desconhecido, $\hat{\mu}_k^U$ e $\hat{\mu}_k^S$ são as estimativas de μ_k, baseadas na população U e na amostra S, respectivamente. Da mesma forma, pode-se diferenciar entre $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_U$, $\hat{\boldsymbol{\beta}}_S$ e $\hat{\boldsymbol{\beta}}_S^\pi$, onde $\boldsymbol{\beta}$ é o parâmetro de interesse, $\hat{\boldsymbol{\beta}}_U$ é uma estimativa de $\boldsymbol{\beta}$, baseada em U, ou seja, levando em conta todos os indivíduos da população através de um método de

estimação (quadrados mínimos ordinários, máxima verossimilhança, etc) segundo o modelo formulado. Por outro lado, quando somente está disponível uma amostra para estimar $\boldsymbol{\beta}$, tem-se duas opções: a primeira consiste em aplicar um método de estimação aos dados que compõem a amostra, obtendo $\hat{\boldsymbol{\beta}}_S$ sem levar em conta o plano amostral. A segunda, leva em conta o plano amostral, aplicando o método de estimação ponderado pelas probabilidades de inclusão, obtendo $\hat{\boldsymbol{\beta}}_S^{\pi}$. O Quadro 2.2 resume o descrito no parágrafo anterior.

Quadro 2.2. Estimação de μ_k.

Com informação sobre toda a população	Com informação sobre uma amostra	
	Com ponderação	Sem poderação
$\hat{\mu}_k^U = g^{-1}(h(\hat{\boldsymbol{\beta}}_U; \mathbf{x}_k))$	$\hat{\mu}_k^S = g^{-1}(h(\hat{\boldsymbol{\beta}}_S^{\pi}; \mathbf{x}_k))$	$\hat{\mu}_k^S = g^{-1}(h(\hat{\boldsymbol{\beta}}_S; \mathbf{x}_k))$

O vetor de parâmetros $\boldsymbol{\beta}$ pode ser estimado por $\hat{\boldsymbol{\beta}}_U$, usando o método de máxima-verossimilhança, o qual consiste em maximizar uma função que expresse a chance de observar os dados que compõem a amostra em função dos parâmetros do modelo. Em modelos lineares de resposta normal, o estimador de máxima-verossimilhança corresponde ao estimador de quadrados mínimos. Para o modelo definido na expressão (2.7), o logaritmo da função de verossimilhança considerando todos os indivíduos da população pode ser expresso como

$$L_U(\boldsymbol{\beta}) = \sum_{k \in U} \{\phi_k [y_k \theta(\boldsymbol{\beta}; \mathbf{x}_k) - b(\theta(\boldsymbol{\beta}; \mathbf{x}_k)] + c(y_k, \phi_k)\},$$

o que implica que $\hat{\boldsymbol{\beta}}_U = \arg\max_{\boldsymbol{\beta}} L_U(\boldsymbol{\beta})$ e $\hat{\mu}_k^U = g^{-1}(h(\hat{\boldsymbol{\beta}}_U; \mathbf{x}_k))$ são os estimadores de máxima-verossimilhança de $\boldsymbol{\beta}$ e μ_k, respectivamente.

Para modelos normais lineares o estimador $\hat{\boldsymbol{\beta}}_U$ assume a mesma forma do estimador de quadrados mínimos ponderados que pode ser escrito como

$$\hat{\boldsymbol{\beta}}_U = (\mathbf{X}_U^T \mathbf{W}_U \mathbf{X}_U)^{-1} \mathbf{X}_U^T \mathbf{W}_U \mathbf{Y}_U,$$

em que $\mathbf{X}_U = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^T$, $\mathbf{Y}_U = (y_1, \ldots, y_N)^T$ e a matriz de pesos é dada por $\mathbf{W}_U = \text{diag}\{w_1, \ldots, w_N\}$ com $w_k = \phi_k$.
O logaritmo da função de verossimilhança para a amostra S, considerando os pesos amostrais, é chamado de função de pseudo log-verossimilhança e pode ser expresso como

$$L_S(\boldsymbol{\beta}) = \sum_{k \in S} \frac{1}{\pi_k} \{\phi_k [y_k \theta(\boldsymbol{\beta}; \mathbf{x}_k) - b(\theta(\boldsymbol{\beta}; \mathbf{x}_k))] + c(y_k, \phi_k)\}, \tag{2.8}$$

o que implica que $\hat{\boldsymbol{\beta}}_S^{\pi} = \arg\max_{\boldsymbol{\beta}} L_S(\boldsymbol{\beta})$ e $\hat{\mu}_k^S = g^{-1}(h(\hat{\boldsymbol{\beta}}_S^{\pi}; \mathbf{x}_k))$ são os estimadores de pseudo máxima-verossimilhança (Lehtonen e Pahkinen, 2004, pág. 284) de $\boldsymbol{\beta}$ e μ_k, respectivamente.
Para modelos normais lineares o estimador $\hat{\boldsymbol{\beta}}_S^{\pi}$ pode ser escrito como

$$\hat{\boldsymbol{\beta}}_S^{\pi} = (\mathbf{X}_S^T \mathbf{W}_S \mathbf{X}_S)^{-1} \mathbf{X}_S^T \mathbf{W}_S \mathbf{Y}_S, \tag{2.9}$$

em que $\mathbf{X}_S = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, $\mathbf{Y}_S = (y_1, \ldots, y_n)^T$ e a matriz de pesos é dada por $\mathbf{W}_S = \text{diag}\{w_1, \ldots, w_n\}$ com $w_k = \phi_k / \pi_k$.
Na expressão (2.8) pode-se observar que os estimadores $\hat{\boldsymbol{\beta}}_S^{\pi}$ e $\hat{\boldsymbol{\beta}}_S$ são equivalentes quando $\pi_k = \pi_l$ para todos $k, l \in U$. Ou seja, para planos amostrais como Amostragem Aleatória Simples (com e sem reposição) e Bernoulli tem-se que $\hat{\boldsymbol{\beta}}_S^{\pi}$ e $\hat{\boldsymbol{\beta}}_S$ são equivalentes.

2.2.3 Modelos de Regressão para Variáveis Dicotômicas

Este tipo de modelo de regressão é aplicado em muitos campos do conhecimento como, por exemplo, nas áreas química, médica e biológica, onde o interesse primário da análise de dados, é avaliar a influência de uma ou mais variáveis explicativas sobre a ocorrência ou não de um evento de interesse. Por exemplo, este tipo de modelo pode ser usado pelas autoridades da saúde de alguma região para avaliar e quantificar o efeito da idade, sexo e raça das pessoas na chance de desenvolver algum tipo de doença. Os modelos de regressão dicotômicos lineares e não-lineares podem ser considerados como um caso particular dos modelos da família exponencial onde a variável resposta é assumida como binomial ou Bernoulli. Em particular,

pode-se supor que para cada indivíduo ou unidade experimental k tem-se o vetor $(y_k, x_{k1}, \ldots, x_{kJ})$, em que y_k pode assumir somente um de dois valores possíveis, denotados por conveniência 1 e 0 (1: sucesso; 0: fracasso), e que $\mathbf{x}_k = (x_{k1}, \ldots, x_{kJ})$ seja um conjunto de variáveis observadas para explicar e/ou predizer o valor de y_k. Denota-se a probabilidade de sucesso, condicionada pela informação no vetor $\mathbf{x}_k$, como

$$\pi(\mathbf{x}_k) = P(Y_k = 1|x_{k1}, \ldots, x_{kJ}) = P(Y_k = 1|\mathbf{x}_k),$$

em que

$$g(\pi(\mathbf{x}_k)) = h(\boldsymbol{\beta}; \mathbf{x}_k)$$

é a função de ligação. Entre as possíveis formas de funções de ligação usadas em modelos de regressão para variáveis dicotômicas podem-se citar:

- *Probit:* $g(\pi(\mathbf{x}_k)) = \Phi^{-1}[\pi(\mathbf{x}_k)] = \eta_k$, sendo $\Phi(\cdot)$ a função de distribuição acumulada normal padrão;
- *Logit:* $g(\pi(\mathbf{x}_k)) = \log[\pi(\mathbf{x}_k)/(1 - \pi(\mathbf{x}_k))] = \eta_k$;
- *Complemento log-log:* $g(\pi(\mathbf{x}_k)) = \log[-\log(1 - \pi(\mathbf{x}_k))] = \eta_k$;
- *Aranda-Ordaz:* $g(\pi(\mathbf{x}_k)) = \log\left\{\dfrac{(1 - \pi(\mathbf{x}_k))^{\alpha} - 1}{\alpha}\right\} = \eta_k$, em que α é uma constante.

A função de ligação "logit" dá lugar ao conhecido modelo de regressão logística. Tendo em vista a importância deste modelo nesta dissertação discute-se a seguir possíveis interpretações para os seus parâmetros.

Considere duas variáveis dicotômicas X e Y, codificadas como 0 e 1 (0 Ausência de atributo; 1 Presença de atributo) para o respectivo atributo de interesse, em que Y é assumida como a variável dependente. Além disso, suponha que estas variáveis são observadas com o objetivo de avaliar a possível associação que possa existir entre elas. O Quadro 2.3 resume a distribuição de probabilidades para o fenômeno em estudo, em que $\pi(i) = P(Y = 1|X = i)$, com $i = 0, 1$.

Com o objetivo de quantificar o grau de associação existente entre X e Y, é definida a estatística chamada de razão de chances, em inglês "*odds ratio*" (OR), a qual pode ser expressa na forma abaixo

Quadro 2.3. Distribuição de probabilidades $P(Y = y|X = x)$.

		Y	
		0	1
X	0	$1 - \pi(0)$	$\pi(0)$
	1	$1 - \pi(1)$	$\pi(1)$

$$OR = \frac{\pi(1)(1 - \pi(0))}{(1 - \pi(1))\pi(0)}. \tag{2.10}$$

Suponha, por exemplo, que Y denota a presença ou ausência de câncer pulmonar e X classifica as pessoas entre fumantes e não fumantes. Então, um $OR = 2$ indica que uma pessoa fumante tem duas vezes mais chance de ter câncer pulmonar do que uma pessoa não fumante (exemplo tomado de Hosmer e Lemeshow (1989, pag.40)). A razão de chances (OR) também mede a direção da associação entre as variáveis Y e X. Esta medida está em escala exponencial, portanto, pode tomar valores no intervalo $(0, \infty)$. Observando a expressão (2.12) é possível concluir que um OR igual a 1 indica independência ou ausência de associação. Um OR maior a 1 indica que a variável independente $X = 1$ é um "fator de risco" para $Y = 1$, ou seja, é mais freqüente obter um sucesso no grupo em que $X = 1$ do que no grupo $X = 0$. Quando o OR é menor que 1 a interpretação é análoga e é denominada "fator protetor". Os nomes "fator protetor" e "fator de risco" são devidos ao contexto bioestatístico onde normalmente é usada a razão de chances (OR) como medida de associação.

Quando a variável explicativa é de tipo quantitativo é preciso formular um modelo. O seguinte exemplo considera um modelo de regressão logística com uma variável explicativa contínua

$$\log\left[\frac{\pi(X)}{1 - \pi(X)}\right] = \beta_0 + \beta_1 X,$$

em que

$$\pi(x_k) = P(Y_k = 1|X = x_k) = \frac{\exp(\beta_0 + \beta_1 x_k)}{1 + \exp(\beta_0 + \beta_1 x_k)}. \tag{2.11}$$

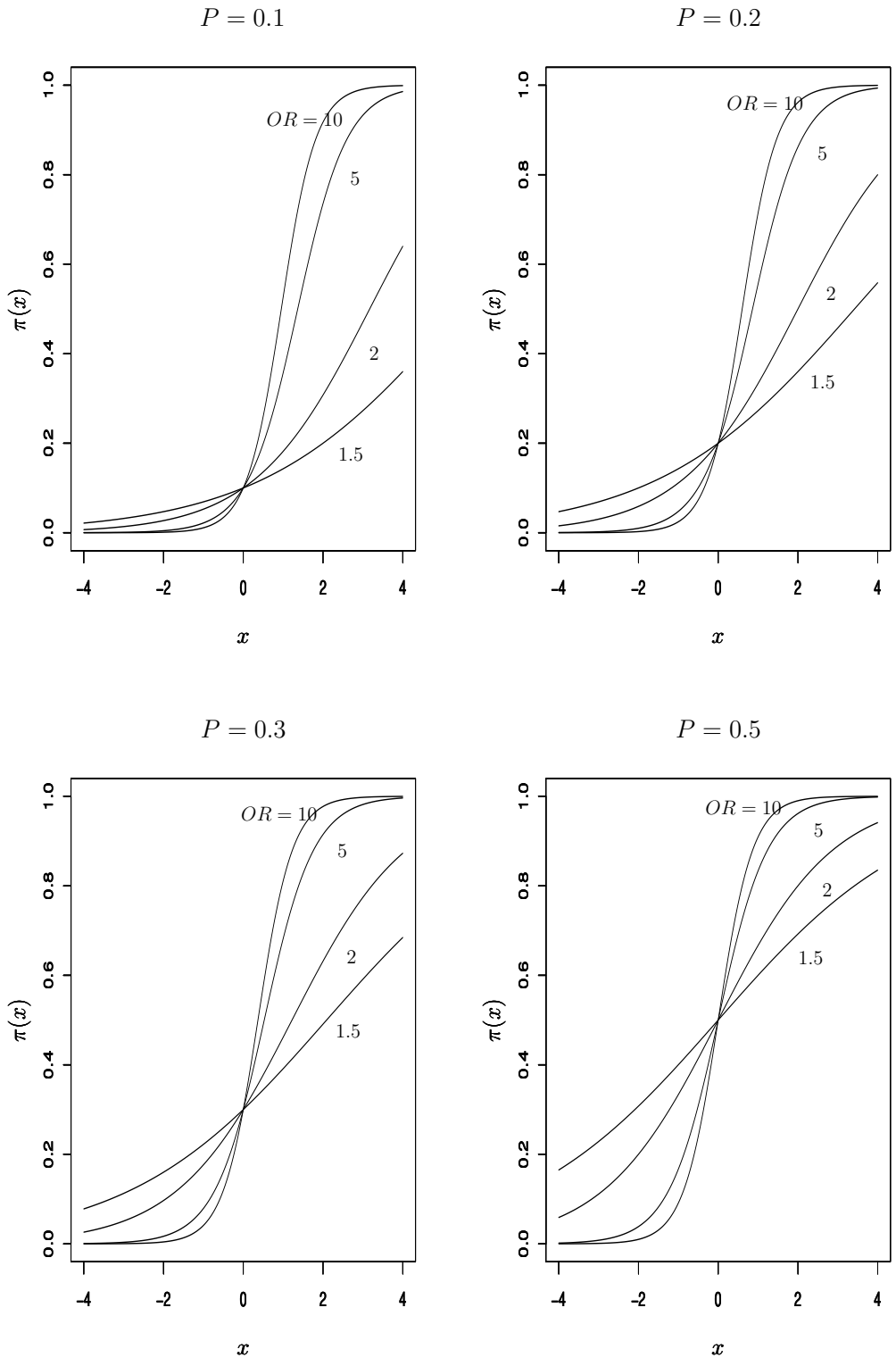
O objetivo é avaliar a associação existente entre X e Y, portanto, é necessário medir o quão freqüente é obtido um sucesso entre os indivíduos que apresentam $X = x + 1$ comparados com os que apresentam $X = x$. Substituindo a equação (2.11) em (2.12), tem-se que

$$OR = \frac{\pi(X+1)[1-\pi(X)]}{[1-\pi(X+1)]\pi(X)} = e^{\beta_1}. \tag{2.12}$$

Baseado neste resultado, é possível ver que um aumento de uma unidade em X faz com que a chance de obter um sucesso aumente (ou diminua) e^{β_1} vezes. Por exemplo, se Y denota a presença ou ausência de osteoporose e X a idade em anos para um grupo de indivíduos, então um $OR = 1.5$ indica que a cada ano que passa estes indivíduos têm uma chance 1.5 vezes maior de sofrer de osteoporose. Daquí para a frente será utilizada a notação tradicional de amostragem, em que não se faz diferença entre letras maiúsculas para variáveis aleatórias e minúsculas para realizações das mesmas.

Na Figura 2.1 é apresentado o comportamento das probabilidades de sucesso $\pi(x)$ em relação à variável explicativa para o modelo (2.11), em que P é a proporção de indivíduos na população com o atributo de interesse, a razão de chances (OR) é o grau de associação entre a variável de interesse (y) e a variável auxiliar (x). Neste caso tem-se que y_k segue uma distribuição de Bernoulli com parâmetro $\pi(x)$ e x segue uma distribuição normal padrão. Nesta figura pode ser observado que quando o grau de associação (OR) entre as variáveis aumenta e, com o aumento P o grau de associação entre as variáveis também aumenta. Quando o grau de associação (OR) entre as variáveis pertence ao intervalo $(0, 1)$ a direção da associação é inversa à apresentada na Figura 2.1. O leitor interessado em saber um pouco mais sobre regressão logística pode consultar, por exemplo, McCullagh e Nelder (1989) e Agresti (1990).

Figura 2.1. Comportamento das probabilidades de sucesso $\pi(x)$ em relação à variável explicativa para o modelo (2.11).



CAPÍTULO 3

Estimador de Regressão Generalizado (GREG)

Este estimador usa informação auxiliar na etapa da estimação, formulando um modelo de regressão entre a variável de interesse e as variáveis auxiliares. A idéia por trás dele é usar o modelo formulado para "estimar" os valores da variável de interesse para os indivíduos que não pertencem à amostra, incrementando desta maneira a eficiência da medição. Quanto maior a adequação do modelo formulado entre a variável de interesse e as variáveis auxiliares, maior será a eficiência do estimador GREG. Tradicionalmente a expressão GREG é utilizada para estimadores assistidos por modelos normais lineares. O estimador de regressão generalizado com base em modelos normais lineares tem sido considerado por vários autores como, por exemplo, Fuller (2002), Holt, Smith, e Winter (1980), Isaki e Fuller (1982), Lohr (1999), Särndal (2001), Särndal, Swensson e Wretman (1992) e Wright (1983). Nesta dissertação a expressão GREG assume um contexto mais amplo, englobando estimadores assistidos por modelos da família exponencial. Essa concepção ampliada de estimadores GREG é parte da contribuição deste trabalho.

Quando o objetivo é estimar o total populacional t_y, é proposto o estimador GREG que pode ser expresso na seguinte forma

$$\hat{t}_{_{GREG}} = \sum_{k \in U} \hat{\mu}_k^S + \sum_{k \in S} \frac{(y_k - \hat{\mu}_k^S)}{\pi_k}, \tag{3.1}$$

onde o modelo formulado pode ser escrito como

$$E(Y_k) = \mu_k = g^{-1}(h(\boldsymbol{\beta}; \mathbf{x}_k)), \quad k = 1, \ldots, N, \tag{3.2}$$

com $\boldsymbol{\beta}$ um vetor de parâmetros desconhecidos, $g(\cdot)$ uma função contínua e duplamente diferenciável e $\mathbf{x}_k = (x_{k1}, \ldots, x_{kJ})$ o vetor de informação auxiliar para o k-ésimo elemento da população. Muitos modelos são possíveis de serem formulados, dependendo da natureza dos dados, da informação auxiliar disponível para o ajuste e da relação entre a variável de interesse e as variáveis auxiliares. Esta característica é muito importante pois proporciona grande flexibilidade para a aplicação do estimador GREG, sendo possível considerar várias alternativas para a componente sistemática bem como para a componente aleatória do modelo assumido.
Supondo que $\hat{\mu}_k^S \approx \hat{\mu}_k^U$, o estimador (3.1) pode ser escrito como

$$\hat{t}_{GREG} \approx \sum_{k \in U} \hat{\mu}_k^U + \sum_{k \in S} \frac{E_k}{\pi_k}, \tag{3.3}$$

em que $E_k = y_k - \hat{\mu}_k^U$. Da equação acima, pode-se avaliar o viés aproximado do $\hat{t}_{GREG}$ da seguinte maneira

$$E_p(\hat{t}_{GREG}) \approx \sum_{k \in U} \hat{\mu}_k^U + E_p\left(\sum_{k \in S} \frac{y_k - \hat{\mu}_k^U}{\pi_k}\right) = t_y.$$

em que

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

com $E_p(I_k) = \pi_k$. Da mesma forma, pode-se usar a expressão (3.3) para obter uma expressão aproximada para a variância de $\hat{t}_{GREG}$, a qual pode ser expressa na forma

$$V_p(\hat{t}_{GREG}) \approx V\left(\sum_{k \in S} \frac{E_k}{\pi_k}\right) = \sum_{k \in U} \sum_{l \in U} \Delta_{kl} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l}, \tag{3.4}$$

com $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$, π_k e π_{kl} as probabilidades de inclusão de primeira e segunda ordem, respectivamente. Ou seja, uma aproximação da variância do estimador $\hat{t}_{GREG}$ é obtida aplicando a fórmula da variância do estimador de Horvitz-Thompson aos resíduos do modelo proposto. A partir da equação (3.4) é possível definir um estimador para a variância de $\hat{t}_{GREG}$ como segue

$$\hat{V}_p(\hat{t}_{GREG}) = \sum_{k\in S}\sum_{l\in S}\frac{\Delta_{kl}}{\pi_{kl}}\frac{e_k}{\pi_k}\frac{e_l}{\pi_l},$$

em que $e_k = y_k - \hat{\mu}_k^S$.

Como um caso particular do estimador de regressão generalizado tem-se o *estimador da razão*. Este estimador é obtido assumindo um modelo de regressão linear entre a variável de interesse e a variável auxiliar, o qual segue uma estrutura da forma

$$\begin{cases} E(Y_k) = \beta x_k; \\ V(Y_k) = \sigma^2 x_k. \end{cases} \tag{3.5}$$

Assumindo este modelo, o estimador GREG pode ser expresso por

$$\begin{aligned}\hat{t}_{GREG1} &= \sum_{k\in U}\hat{\beta}_S^{\pi}x_k = \frac{\sum_{k\in U}x_k}{\sum_{k\in S}\frac{x_k}{\pi_k}}\sum_{k\in S}\frac{y_k}{\pi_k} \\ &= \sum_{k\in S}\frac{g_{ks}y_k}{\pi_k},\end{aligned}$$

que corresponde ao estimador da razão, com

$$\hat{\beta}_S^{\pi} = \frac{\sum_{k\in S}\frac{y_k}{\pi_k}}{\sum_{k\in S}\frac{x_k}{\pi_k}} \tag{3.6}$$

e

$$g_{ks} = \frac{\sum_{k\in U}x_k}{\sum_{k\in S}\frac{x_k}{\pi_k}},$$

onde $\hat{\beta}_S^{\pi}$ também pode ser obtido a partir da expressão (2.9), com $w_k = 1/(\sigma^2 x_k\pi_k)$. Este estimador é muito usado na prática pois é muito fácil de ser aplicado, sendo usado inclusive quando a variável de interesse está categorizada.

Uma aproximação da variância do estimador $\hat{t}_{GREG1}$ pode ser obtida aplicando a expressão (3.4), em que $E_k = y_k - \hat{\beta}_U x_k$, com $\hat{\beta}_U = \frac{t_y}{t_x}$. O estimador da variância do estimador $\hat{t}_{GREG1}$ é expresso por

$$\hat{V}(\hat{t}_{GREG1}) = \sum_{k \in S} \sum_{l \in S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{g_{ks} e_k}{\pi_k} \frac{g_{ls} e_l}{\pi_l},$$

com $e_k = y_k - \hat{\beta}_S^{\pi} x_k$.

O Estimador de regressão generalizado (GREG), como apresentado em (3.1) pode ser interpretado como a soma dos valores preditos pelo modelo considerado para todos indivíduos da população mais um termo de ajuste. É possível formular condições sob as quais o termo de ajuste desaparece, quando a estimação é assistida por modelos normais lineares Särndal, Swensson e Wretman (1992, pag.231) apresentam condicões similares para o caso do estimador de regressão generalizado. A seguir é apresentado um lema que de generaliza os resultados citados acima e que é parte integrante da contribuição desta dissertação.

Lema 1. *Se o estimador de regressão generalizado (GREG) descrito na expressão (3.1) considera um modelo de regressão linear ou não-linear da família exponencial onde tem-se:*

S1. Homogeneidade no parâmetro de dispersão, ou seja, $\phi_k = \phi$ para todo $k \in U$;

S2. Componente sistemática com intercepto, ou seja, existe β_j em $\boldsymbol{\beta}$ tal que $\partial \eta_k / \partial \beta_j = C$ para todo $k \in U$, com C uma constante;

S3. Componente sistemática com ligação canônica, ou seja, $\theta_k = \eta_k$ para todo $k \in U$;

Então o estimador GREG para t_y pode ser escrito como

$$\hat{t}_{GREG} = \sum_{k \in U} \hat{\mu}_k^S = \sum_{k \in U} g^{-1}(h(\hat{\boldsymbol{\beta}}_S^{\pi}; \mathbf{x}_k)).$$

Além disso, o total de y pode ser expresso da seguinte forma

$$t_y = \sum_{k \in U} \hat{\mu}_k^U = \sum_{k \in U} g^{-1}(h(\hat{\boldsymbol{\beta}}_U; \mathbf{x}_k)).$$

A prova deste lema pode ser encontrada no Apêndice A. A aplicação do Lema 1 implica numa simplificação da expressão de $\hat{t}_{GREG}$. O Lema 1 permite concluir que o estimador GREG para o total t_y pode ser expresso de uma maneira mais simples em modelos como, por exemplo:

- Regressão logística linear e não linear com intercepto.
- Regressão linear e não linear homoscedastica com intercepto e ligação identidade.
- Regressão de poisson linear e não linear com intercepto e ligação logaritmo.
- Regressão gama linear e não linear com intercepto e ligação $1/\mu$.
- Regressão normal inversa linear e não linear com intercepto e ligação $1/\mu^2$.

3.1 Estimador de Regressão Generalizado no Contexto de Estratificação

Em muitas pesquisas é comum encontrar populações compostas por subpopulações bem definidas que podem ser identificadas a priori. Quando estas subpopulações são disjuntas, podem dar origem a *estratos*. A estratificação é apresentada em alguns casos de forma evidente e quando ela é usada procura-se que exista homogeneidade nos elementos que pertencem a cada estrato e heterogeneidade entre os estratos. A seleção dos indivíduos em cada estrato é independente, ou seja, pode ser retirada uma amostra seguindo um plano amostral $p(\cdot)$ diferente para cada estrato. A estratificação é um método eficiente e flexível usado com muita freqüência na prática. A seguir serão apresentadas algumas possíveis razões para usar estratificação:

- Às vezes é possível identificar a priori subpopulações para as quais deseja-se obter estimativas com precisões pré-especificadas. Neste caso, cada subpopulação pode ser tratada como uma "população" no processo de inferência.

- A conveniência administrativa pode algumas vezes sugerir estratificação. Por exemplo, se a institução responsável pela pesquisa tem vários escritórios dispersos pela população de interesse, então cada escritório pode encarregar-se da região na qual está localizado recorrendo desta maneira à estratificação, considerando como um estrato a área correspondente a cada escritório.
- É possível ainda que, para algumas subpopulações específicas, o contexto (existência de informações auxiliares, por exemplo) indique um procedimento diferente de estimação. Nestes casos, cada subpopulação específica seria um estrato.

O procedimento de estimação na amostragem estratificada é realizado considerando cada estrato como se fosse uma subpopulação, obtendo as estimativas dos parâmetros de interesse em cada estrato. Uma vez obtidas estas estimativas é feita uma combinação delas para desta maneira, estimar os parâmetros na população total. O processo de estimação em cada estrato pode ser realizado com diferentes métodos. O importante é que as amostras selecionadas em cada estrato sejam independentes, obtendo assim, fórmulas diretas de estimação para os parâmetros populacionais.

Uma das vantagens de usar a amostragem estratificada é que sob certas condições, os estimadores são mais eficientes e com menor variância. Entretanto, existem situações onde a implementação de estratificação tem um custo alto o qual afeta o orçamento e leva a diminuir o tamanho da amostra total. A estratificação também permite planejar estimações para os estratos com um nível de confiança e precisão estabelecidos previamente.

3.1.1 Plano Amostral e Estimação sob Estratificação

Em amostragem estratificada (AE), a população U em estudo é particionada em H estratos de tamanhos $N_1, N_2, \ldots, N_H$, respectivamente, onde

$$U = \bigcup_{h=1}^{H} U_h,$$

em que $U_h = \{k \in U : k \in \text{estrato } h\}$.

Um processo físico de aleatorização é empregado dentro de cada estrato h, independente, para gerar uma amostra S_h de tamanho n_h $(h = 1, 2, \ldots, H)$. A amostra final (de tamanho n) é composta por todos os elementos selecionados, isto é

$$S = \bigcup_{h=1}^{H} S_h,$$

com $n = \sum_{h=1}^{H} n_h$. Denote por p_h o plano amostral implementado pela aleatorização imposta ao estrato h. Como as amostras $S_1, S_2, \ldots, S_H$ foram geradas independentemente, o plano AE atribui probabilidade de seleção da amostra S, dado por

$$p(S) = \prod_{h=1}^{H} p_h(S_h).$$

O número de elementos no estrato h, chamado tamanho do estrato h, é denotado por N_h. Considerando que cada estrato forma uma partição de U, tem-se que $N = \sum_{h=1}^{H} N_h$. Além disso, o total populacional pode ser decomposto como

$$t = \sum_{k \in U} y_k = \sum_{h=1}^{H} t_h = \sum_{h=1}^{H} N_h \bar{y}_{U_h},$$

em que $t_h = \sum_{k \in U_h} y_k$ e $\bar{y}_{U_h}$ são o total e a média do estrato h, respectivamente. Adicionalmente, defina $a_h = N_h/N$ como o peso do estrato h em U. Então, a média populacional pode ser expressa por

$$\bar{y}_U = \sum_{h=1}^{H} a_h \bar{y}_{U_h}.$$

O estimador do tipo Horvitz-Thompson total populacional, sob uma AE, com H estratos, assume a forma

$$\hat{t}_\pi = \sum_{h=1}^{H} \hat{t}_{h\pi},$$

onde $\hat{t}_{h\pi}$ é o estimador de $t_h = \sum_{k \in U_h} y_k$. A sua variância pode ser escrita como

$$V(\hat{t}_\pi) = \sum_{h=1}^{H} V(\hat{t}_{h\pi}).$$

Além disso,

$$\hat{V}(\hat{t}_\pi) = \sum_{h=1}^{H} \hat{V}(\hat{t}_{h\pi}),$$

é um estimador não-viesado para $V(\hat{t}_\pi)$, desde que $\hat{V}(\hat{t}_{h\pi})$ seja um estimador não-viesado para $V(\hat{t}_{h\pi})$, para $h = 1, 2, \ldots, H$.

Uma aplicação importante dos estimadores de regressão, descritos neste trabalho, ocorre quando o plano empregado na seleção dos indivíduos é amostragem estratificada. Neste contexto podem ser identificados dois tipos de estimadores de regressão, os estimadores separado e combinado.

3.1.2 Estimador de Regressão Generalizado Combinado

Os estimadores de regressão são chamados de estimadores de regressão combinados, quando o modelo formulado entre a variável de interesse e as variáveis auxilares é o mesmo para toda a população, sem fazer diferença entre a relação destas variáveis em cada estrato. O estimador de regressão generalizado combinado (GREGC), denotado por $\hat{t}_{GREGC}$, assume a forma dada em (3.1), em que $\hat{\mu}_k^S = g^{-1}(h(\hat{\boldsymbol{\beta}}_S^\pi; \mathbf{x}_k))$ e $\hat{\boldsymbol{\beta}}_S^\pi = \arg\max_{\boldsymbol{\beta}} L_S(\boldsymbol{\beta})$, com

$$L_S(\boldsymbol{\beta}) = \sum_{h=1}^{H} \sum_{k \in S_h} \frac{1}{\pi_k} \{\phi_k [y_k \theta(\boldsymbol{\beta}; \mathbf{x}_k) - b(\theta(\boldsymbol{\beta}; \mathbf{x}_k))] + c(y_k, \phi_k)\}.$$

Uma aproximação da variância de $\hat{t}_{GREGC}$ pode ser expressa como

$$V(\hat{t}_{GREGC}) = \sum_{h=1}^{H} \left[\sum_{k \in U_h} \sum_{l \in U_h} \Delta_{kl} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l} \right], \tag{3.7}$$

em que $E_k = y_k - \hat{\mu}_k^U$, com $\hat{\mu}_k^U = g^{-1}(h(\hat{\boldsymbol{\beta}}_U; \mathbf{x}_k))$. A variância deste tipo de estimador pode estar inflacionada quando os coeficientes de regressão são diferentes de estrato para estrato na população de interesse.

3.1.3 Estimador de Regressão Generalizado Separado

O estimador de regressão separado é aplicado quando é considerado em cada estrato um modelo de regressão diferente, ou seja, quando a relação entre a

variável de interesse e as variáveis auxiliares em cada estrato assumem uma associação diferente, tendo que recorrer à formulação de modelos distintos para estas relações em cada estrato. Os estimadores de regressão separados estão mais sujetos a ser viesados, sendo comparados com os estimadores combinados, na medida em que os tamanhos de amostra para cada estrato sejam pequenos. O estimador de regressão generalizado separado (GREGS), pode ser escrito na seguinte forma

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

em que $\hat{\mu}_k^{S_h} = g^{-1}(h(\hat{\boldsymbol{\beta}}_{S_h}^{\pi}; \mathbf{x}_k))$ e $\hat{\boldsymbol{\beta}}_{S_h}^{\pi} = \arg\max_{\boldsymbol{\beta}} L_{S_h}(\boldsymbol{\beta})$, com

$$L_{S_h}(\boldsymbol{\beta}) = \sum_{k \in S_h} \frac{1}{\pi_k} \{\phi_k [y_k \theta(\boldsymbol{\beta}; \mathbf{x}_k) - b(\theta(\boldsymbol{\beta}; \mathbf{x}_k))] + c(y_k, \phi_k)\}.$$

Uma aproximação da variância do estimador $\hat{t}_{GREGS}$ pode ser obtida usando a expressão (3.7), em que $E_k = y_k - \hat{\mu}_k^{U_h}$, com $\hat{\mu}_k^{U_h} = g^{-1}(h(\hat{\boldsymbol{\beta}}_{U_h}; \mathbf{x}_k))$.

3.2 Estimadores Assistidos por Modelos de Regressão Lineares

Particularmente, para um modelo de regressão linear

$$E(Y_k) = \mu_k = \sum_{j=1}^{J} \hat{\beta}_j x_{kj}, \tag{3.8}$$

tem-se que o estimador GREG pode ser expresso da seguinte forma

$$\hat{t}_{GREG} = \hat{t}_{\pi} + \sum_{j=1}^{J} \hat{\beta}_j^{\pi} (t_{x_j} - \hat{t}_{x_j \pi}), \tag{3.9}$$

onde $\hat{t}_{\pi}$ é o estimador de Horvitz-Thompson para o total de y, $\hat{t}_{x_j \pi}$ é o estimador de Horvitz-Thompson do total da variável auxiliar x_j e $\hat{\beta}_1^{\pi}, \ldots, \hat{\beta}_J^{\pi}$ são os componentes do vetor $\hat{\boldsymbol{\beta}}_S^{\pi}$. Usando o Lema 1, apresentado na seção anterior é possível concluir que, se o modelo formulado em (3.8) considera

intercepto então o estimador $\hat{t}_{GREG}$ é dado por

$$\hat{t}_{GREG} = \sum_{k \in U} \hat{\mu}_k^S.$$

O estimador GREG pode ser expresso de várias formas, sendo a apresentada em (3.9) apenas uma delas. A seguir serão mostradas outras possíveis maneiras de expressar (3.1) para o caso linear. Uma forma de apresentar o estimador GREG é motivada por conseguir expressá-lo como uma soma de valores ponderados. Neste caso, é necessario introduzir as seguintes medidas, as quais permitem expressar $\hat{\boldsymbol{\beta}}_S^{\pi}$ de uma maneira diferente da equação dada em (2.9):

$$\widehat{\mathbf{T}} = \sum_{k \in S} \frac{\mathbf{x}_k \mathbf{x}_k^T}{\sigma_k^2 \pi_k} \qquad \text{e} \qquad \hat{\mathbf{t}} = \sum_{k \in S} \frac{\mathbf{x}_k y_k}{\sigma_k^2 \pi_k},$$

sendo $\hat{\boldsymbol{\beta}}_S^{\pi} = \widehat{\mathbf{T}}^{-1} \hat{\mathbf{t}}$. Além disso, podem ser definidos $\mathbf{t}_x = (t_{x1}, \ldots, t_{xJ})^T$ e $\hat{\mathbf{t}}_{x\pi} = (\hat{t}_{x\pi}, \ldots, \hat{t}_{x\pi})$ vetores dos totais e os estimadores de Horvitz-Thompson das variáveis auxiliares, respectivamente. Então, tomando como base (3.9) e usando as medidas definidas acima, tem-se

$$\begin{aligned}
\hat{t}_{GREG} &= \hat{t}_{\pi} + \sum_{j=1}^{J} \hat{\beta}_j^{\pi} (\hat{t}_{x_j} - \hat{t}_{x_j \pi}) \\
&= \hat{t}_{\pi} + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})^T \hat{\boldsymbol{\beta}}_S^{\pi} \\
&= \sum_{k \in S} \frac{y_k}{\pi_k} + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})^T \widehat{\mathbf{T}}^{-1} \sum_{k \in S} \mathbf{x}_k \frac{y_k}{\sigma_k^2 \pi_k} \\
&= \sum_{k \in S} \left[1 + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})^T \widehat{\mathbf{T}}^{-1} \mathbf{x}_k / \sigma_k^2 \right] \frac{y_k}{\pi_k} \\
&= \sum_{k \in S} g_{ks} \frac{y_k}{\pi_k},
\end{aligned}$$

em que g_{ks} pode ser considerado como um fator de calibração para π_k.

A seguir, são apresentados dois casos particulares do estimador GREG quando o modelo considera somente uma variável auxiliar. Inicialmente, considere um modelo sem intercepto, o estimador assistido por este modelo pode ser denominado $\hat{t}_{GREG1}$ e que corresponde ao *estimador da razão* tratado no começo deste capítulo.

O segundo estimador considerado é o resultado de aplicar um modelo com intercepto e variância constante, o qual segue uma estrutura da forma

$$\begin{cases} E(Y_k) = \alpha + \beta x_k; \\ V(Y_k) = \sigma^2. \end{cases} \tag{3.10}$$

podendo expressar o estimador GREG como

$$\begin{aligned} \hat{t}_{GREG2} &= \sum_{k\in U}(\hat{\alpha}_S^\pi + \hat{\beta}_S^\pi x_k) + \sum_{k\in S}\frac{(y_k - \hat{\alpha}_S^\pi - \hat{\beta}_S^\pi x_k)}{\pi_k} \\ &= N[\tilde{y}_S + \hat{\beta}_S^\pi(\bar{x}_U - \tilde{x}_S)] = \sum_{k\in S}\frac{g_{ks}y_k}{\pi_k}, \end{aligned}$$

em que

$$\hat{\beta}_S^\pi = \frac{\sum_{k\in S}(y_k - \tilde{y}_S)(x_k - \tilde{x}_S)/\pi_k}{\sum_{k\in S}(x_k - \tilde{x}_k)^2/\pi_k}, \qquad \hat{\alpha}_S^\pi = \tilde{y}_S - \hat{\beta}_S^\pi\tilde{x}_S,$$

$$g_{ks} = \frac{N}{\hat{N}}[1 + a_S(x_k - \tilde{x}_S)], \qquad a_S = \frac{(\bar{x}_U - \tilde{x}_S)\hat{N}}{\sum_{k\in S}(x_k - \tilde{x}_S)^2/\pi_k},$$

$$\tilde{y}_S = \frac{1}{\hat{N}}\sum_{k\in S}\frac{y_k}{\pi_k}, \qquad \tilde{x}_S = \frac{1}{\hat{N}}\sum_{k\in S}\frac{x_k}{\pi_k}, \qquad \hat{N} = \sum_{k\in S}\frac{1}{\pi_k}.$$

Este estimador é comumente chamado na literatura de estimador de regressão. Uma aproximação da variância de $\hat{t}_{GREG2}$ pode ser obtida aplicando a expressão (3.4), onde

$$E_k = y_k - \hat{\alpha}_U - \hat{\beta}_U x_k, \tag{3.11}$$

com $\hat{\beta}_U = \frac{S_{xy}}{S_x^2}$ e $\hat{\alpha} = \bar{y}_U - \hat{\beta}_U\bar{x}_U$.

O estimador da variância do estimador $\hat{t}_{GREG2}$ é dado por

$$\hat{V}(\hat{t}_{GREG2}) = \sum_{k\in S}\sum_{l\in S}\frac{\Delta_{kl}}{\pi_{kl}}\frac{g_{ks}e_k}{\pi_k}\frac{g_{ls}e_l}{\pi_l},$$

com $e_k = y_k - \hat{\alpha}_S^\pi - \hat{\beta}_S^\pi x_k$.

Quando o modelo de regressão formulado entre a variável de interesse e as variáveis auxiliares é linear e o plano amostral é estatificado, tem-se os estimadores descritos a seguir.

3.2.1 Estimador de Regressão Combinado

O estimador de regressão combinado assume a seguinte forma

$$\hat{t}_{GREGC} = N[\tilde{y}_S + \hat{\boldsymbol{\beta}}_S^{\pi}(\bar{x}_U - \tilde{x}_S)],$$

onde

$$\tilde{y}_S = \sum_{h=1}^{H} a_h \tilde{y}_{S_h},$$

com $a_h = N_h/N$,

$$\hat{\boldsymbol{\beta}}_S^{\pi} = \frac{\sum_{h=1}^{H}\sum_{k\in S_h}(x_k - \tilde{x}_S)(y_k - \tilde{y}_S)/\pi_k}{\sum_{h=1}^{H}\sum_{k\in S_h}(x_k - \tilde{x}_S)^2/\pi_k},$$

e $\tilde{x}_S$ é definido de forma análoga a $\tilde{y}_S$.

3.2.2 Estimador de Regressão Separado

O estimador de regressão separado pode ser expresso por

$$\hat{t}_{GREGS} = \sum_{h=1}^{H} N_h[\tilde{y}_{S_h} - \hat{\beta}_h(\bar{x}_{U_h} - \tilde{x}_{S_h})],$$

em que

$$\hat{\beta}_h = \frac{\sum_{k\in S_h}(x_k - \tilde{x}_{S_h})(y_k - \tilde{y}_{S_h})/\pi_k}{\sum_{k\in S_h}(x_k - \tilde{x}_{S_h})^2/\pi_k},$$

e

$$\tilde{y}_{S_h} = \frac{\sum_{k\in S_h} y_k/\pi_k}{\sum_{k\in S_h} 1/\pi_k},$$

análogo para $\tilde{x}_{S_h}$.

CAPÍTULO 4

Uma Forma Alternativa de Derivação do Estimador de Regressão

O objetivo deste capítulo é apresentar o método proposto por Tillé (1998), para derivar o estimador de regressão generalizado (GREG), quando o modelo que assiste à estimação é linear, baseado na metodologia da correção do viés condicional (CVC). A inferência condicional tem sido estudada amplamente na área de amostragem, no contexto de obter estimadores não-viesados, ou estimadores com um viés condicional pequeno. Os procedimentos aplicados para obter estimadores não-viesados condicionalmente, recorrem freqüentemente à estimação do viés condicional e à aplicação de um fator de correção ao estimador original. O resultado destes procedimentos é um estimador com menor ou sem viés condicional. Este assunto tem sido discutido por Fuller e Isaki (1981), Deville (1992), Montanari (1997, 1998) e Rao (1994,1997). Além disso, Casady e Valliant (1993) estudaram as propriedades condicionais do estimador usado no caso de pós-estratificação. O método proposto por Tillé usa as probabilidades de inclusão condicionais para construir um estimador com um viés condicional pequeno.

A CVC pode ser aplicada devido à existência da informação auxiliar, estimando a esperança condicional com respeito a uma estatística, denominada estatística auxiliar e denotada por η. A seguir, é apresentado como o estimador obtido através da CVC pode ser mais eficiente do que um estimador incondicional.

Considere-se o estimador $\hat{\theta}$ não-viesado para θ. Se $B(\hat{\theta}|\eta) = E(\hat{\theta}|\eta) - \theta$ é o viés condicional de $\hat{\theta}$ dado que η é conhecida, então o estimador ajustado $\hat{\theta}^*$ pode ser construído assim:

$$\hat{\theta}^* = \hat{\theta} - B(\hat{\theta}|\eta).$$

Neste caso,

$$V(\hat{\theta}^*) = V(\hat{\theta}) + V(B(\hat{\theta}|\eta)) - 2Cov(\hat{\theta}, B(\hat{\theta}|\eta)),$$

onde

$$\begin{aligned} Cov(\hat{\theta}, B(\hat{\theta}|\eta)) &= E((\hat{\theta} - \theta)(E(\hat{\theta}) - \theta)) \\ &= E\{E((\hat{\theta} - \theta)(E(\hat{\theta}) - \theta)|\eta)\} \\ &= V(E(\hat{\theta}|\eta)). \end{aligned}$$

Então, obtém-se

$$V(\hat{\theta}^*) = V(\hat{\theta}) - V(E(\hat{\theta}|\eta)).$$

Ou seja, a variância do estimador $\hat{\theta}^*$ é menor que a variância do estimador $\hat{\theta}$. O problema apresentado usando $\hat{\theta}^*$ é que, ainda que o viés condicional possa ser em geral estimado, o ganho em reduzir a variância pode ser frustrado pela inestabilidade do estimador condicionalmente viesado usado. De maneira geral, nesta seção no lugar de obter $\hat{\theta}^*$ de $\hat{\theta}$ por meio do viés condicional ajustado, a construção do estimador para θ é feita usando a CVC e as probabilidades de inclusão condicionais.

4.1 Estimadores Condicionalmente Não-viesados

Considere $\eta = \eta(x_k, k \in S)$ uma estatística. Como a população é finita, η só pode assumir um número finito de valores, denotados por $(\eta_1, \ldots, \eta_i, \ldots, \eta_l)$. O objetivo é estimar $\bar{y}$ com um viés condicional o menor possível com respeito à estatística η. Então, são definidas as probabilidades condicionais de primeira ordem

$$\pi_{k|\eta} = E(I_k|\eta),\ k \in U,$$

e as probabilidades de segunda ordem

$$\pi_{kl|\eta} = E(I_k I_l|\eta), \;\; k \in U, \; l \in U \text{ com } k \neq l,$$

onde I_k é a variável indicadora de inclusão na amostra, que assume o valor 1 se o k-ésimo elemento pertence à amostra, e 0 caso contrário. Suponha que as probabilidades de inclusão condicionais podem ser calculadas para algum possível valor da estatística η. O estimador construído usando as probabilidades de inclusão condicionais recebe o nome de estimador ponderado condicionalmente (CW). O estimador ponderado condicionalmente simples (SCW), pode ser expresso por

$$\hat{\bar{y}}_{|\eta} = \frac{1}{N} \sum_{k \in S} \frac{y_k}{\pi_{k|\eta}},$$

em que as probabilidades de inclusão condicionais podem ser calculadas para algum possível valor da estatística auxiliar η.

Na teoria de amostragem, uma condição necessária para a existência de um estimador não-viesado de $\bar{y}$ é que $\pi_k > 0$ para todo $k \in U$. Este resultado pode ser adaptado para a existência de um estimador condicionalmente não-viesado, usando como condição necessária $\pi_{k|\eta} > 0$ para todo $k \in U$, e para todos os possíveis valores de η.

Note que, $\pi_{k|\eta}$ pode ser zero até mesmo quando π_k é estritamente positiva. Então, um estimador não-viesado condicionalmente exato raramente existe na prática. Por esta causa, Tillé propõe uma definição de estimadores condicionalmente não-viesados menos exigente.

Definição 1. *O estimador $\hat{\bar{y}}$ de $\bar{y}$ é dito ser virtualmente condicionalmente não-viesado (VCU) com respeito à estatística η se seu viés condicional depende só de quantidades com probabilidades de inclusão condicionais de primeira ordem nulas. Ou seja,*

$$B(\hat{\bar{y}}|\eta) = \sum_{k \in U} y_k \alpha_k(\eta) I[\pi_{k|\eta} = 0]$$

para todo $(y_1, \ldots, y_N) \in I\!R^N$, onde os coeficientes $\alpha_k(\eta)$ podem depender de η.

Exemplo 1. *O viés condicional do estimador SCW pode ser expresso por*

$$\begin{aligned} B(\hat{\bar{y}}_{\pi|\eta}) &= E(\hat{\bar{y}}_{\pi|\eta}) - \bar{y} \\ &= \frac{1}{N} \sum_{\substack{k \in U \\ \pi_{k|\eta} > 0}} E\left(\frac{y_k I_k}{\pi_{k|\eta}} \Big| \eta \right) - \bar{y} \\ &= -\frac{1}{N} \sum_{k \in U} y_k I[\pi_{k|\eta} = 0], \end{aligned}$$

onde $I(\cdot)$ é uma função indicadora dada por

$$I[\pi_{k|\eta} = 0] = \begin{cases} 1 \text{ se } \pi_{k|\eta} = 0 \\ 0 \text{ se } \pi_{k|\eta} > 0 \end{cases}$$

Exemplo 2. *Uma amostra de tamanho $n > 0$ é tomada, sem reposição, seguindo um plano de amostragem aleatória simples, de uma população de tamanho N. Neste caso, se $n = \eta$ tem-se que $\pi_{k|\eta} = \dfrac{n}{N}$ para algum $k \in U$. Então $\pi_{k|\eta} > 0$ para todo $k \in U$ e um estimador não-viesado condicionalmente exato com respeito a n sempre existe.*

Outros estimadores ponderados condicionalmente podem ser derivados usando o estimador incondicional não-viesado e podem ser chamados de estimadores ponderados condicionalmente corrigidos (CCW). Eles são dados por:

$$\hat{\bar{y}}_{c|\eta} = \frac{1}{N} \sum_{k \in S} \frac{y_k}{h_k \pi_{k|\eta}},$$

onde $h_k = E(I[\pi_{k|\eta} > 0]) = P(\pi_{k|\eta} > 0)$. Seu viés condicional pode ser expresso por

$$B(\hat{\bar{y}}_{c|\eta}|\eta) = \frac{1}{N} \sum_{k \in U} y_k \left(\frac{I[\pi_{k|\eta} > 0]}{h_k} - 1 \right).$$

O estimador CCW não é VCU, mas é incondicionalmente viesado, pois

$$B(\hat{\bar{y}}_{c|\eta}) = E(B(\hat{\bar{y}}_{c|\eta}|\eta)) = 0.$$

Os estimadores SCW e CCW não são invariantes por alocação. Ou seja, estes estimadores não incrementam de um valor de C quando todas as unidades y_k são incrementadas por um valor C. Ou seja,

$$\frac{1}{N} \sum_{k \in S} \frac{y_k + C}{\pi_{k|\eta}} = \hat{\bar{y}}_{c|\eta} + \frac{C}{N} \sum_{k \in S} \frac{1}{\pi_{k|\eta}} \neq \hat{\bar{y}}_{c|\eta} + C.$$

Como uma solução para este problema, duas versões do estimador de razão podem ser usadas:

1. O estimador de razão ponderado condicionalmente (SCW), que pode ser expresso por
$$\hat{\bar{y}}_{r|\eta} = \left(\sum_{k\in S}\frac{1}{\pi_{k|\eta}}\right)^{-1}\sum_{k\in S}\frac{y_k}{\pi_{k|\eta}} \tag{4.1}$$
2. O estimador de razão corrigido ponderado condicionalmente (CCW) dado por
$$\hat{\bar{y}}_{cr|\eta} = \left(\sum_{k\in S}\frac{1}{h_k\pi_{k|\eta}}\right)^{-1}\sum_{k\in S}\frac{y_k}{h_k\pi_{k|\eta}}.$$

Um estimador condicionalmente não-viesado raramente existe. Por esta razão, poderia ser necessário admitir um leve viés condicional, o qual leva a concluir que sempre é possível fazer uma correção do estimador CW para que este seja incondicionalmente não-viesado. Entretanto, esta correção faz que o viés condicional seja maior, pelo qual há um incremento do erro quadrático médio (EQM). Por esta causa é preferível usar o estimador dado em (4.1) quando a soma inversa das probabilidades de inclusão $(w_k = 1/\pi_k)$ não é igual a N.

4.2 Probabilidades de Inclusão Condicionais

Na construção do estimador CW é necessário avaliar as probabilidades de inclusão condicionais. Aplicando o teorema de Bayes pode ser observado que

$$\begin{aligned}\pi_{k|\eta} &= E(I_k|\eta=\eta_i)\\ &= P(k\in S|\eta=\eta_i)\\ &= \frac{P(k\in\ S,\eta=\eta_i)}{P(\eta=\eta_i)}\\ &= \pi_k\frac{P(\eta=\eta_i|k\in\ S)}{P(\eta=\eta_i)},\quad i=1,\ldots,I,\end{aligned}$$

onde I é o número de valores que pode assumir η. A distribuição de probabilidade de η pode ser calculada teoricamente do plano amostral $p(\cdot)$, tendo

que:

$$P(\eta = \eta_i) = \sum_{S|\eta=\eta_i} p(S),$$

e

$$P(\eta = \eta_i | k \in S) = \frac{P(\eta = \eta_i \wedge k \in S)}{\pi_k} = \frac{1}{\pi_k} \sum_{\substack{S|\eta=\eta_i \\ S \ni k}} p(S).$$

Em alguns casos não é possível calcular as probabilidades de inclusão condicionais exatas, Sendo necessário usar uma aproximação.

4.3 Estimador de Regressão

Considere que a informação auxiliar disponível é a média populacional $\bar{x}_U$ de uma variável aleatória $\mathbf{x}$, e $\hat{\bar{x}}_{x\pi}$ é o estimador de Horvitz-Thompson de $\bar{x}$. O objetivo é derivar o estimador SCW da média populacional $\bar{y}$ para a variável de interesse y, usando $\eta = \hat{\bar{x}}_{x\pi}$ como a estatística auxiliar. Então, o estimador SCW é dado por

$$\hat{\bar{y}}_{|\hat{\bar{x}}} = \frac{1}{N} \sum_{k \in S} \frac{y_k}{\pi_{k|\hat{\bar{x}}_{x\pi}}},$$

onde $\pi_{k|\hat{\bar{x}}} = E(I_k|\hat{\bar{x}})$. Se o vetor aleatório $\hat{\mathbf{x}}$ assume o valor z, uma aproximação de $\pi_{k|\hat{\bar{x}}}$ usando o teorema de Bayes pode ser expressa por

$$E(I_k|\hat{\mathbf{x}} = z) = P(k \in S|\hat{\mathbf{x}} = z) = \frac{\pi_k P(\hat{\mathbf{x}} = z|k \in S)}{P(\hat{\mathbf{x}} = z)}.$$

Como foi mencionado anteriormente, para derivar a forma final do estimador, é necessário avaliar, pelo menos aproximadamente $\pi_{k|\hat{\bar{x}}}$. Especificamente, é necessário conhecer a distribuição de probabilidade de $\pi_{k|\hat{\bar{x}}}$ incondicional e condicionalmente na presença das unidades amostrais ($k \in S$). Em geral, o cálculo destas probabilidades é muito complexo. Neste caso é necessário usar uma aproximação para construir um estimador SCW aproximado. Assim, é possível calcular a média e a variância de $\hat{\bar{x}}$ condicional e

incondicionalmente na presença de cada unidade na amostra, como segue

$$\bar{x} = E(\hat{\bar{x}}) = \frac{1}{N}\sum_{l\in U} x_l,$$

$$\bar{x}_{|k} = E(\hat{\bar{x}}|k \in S) = \frac{1}{N}\sum_{\substack{l\in U\\ l\neq k}} \frac{x_l\pi_{kl}}{\pi_k\pi_l} + \frac{x_k}{\pi_k N}, \tag{4.2}$$

$$V_x = V(\hat{\bar{x}}) = \frac{1}{N^2}\sum_{k\in U}\frac{x_l^2}{\pi_l}(1-\pi_l) + \frac{1}{N^2}\sum_{l\in U}\sum_{\substack{m\in U\\ m\neq l}} \frac{x_k x_m}{\pi_l\pi_m}(\pi_{lm}-\pi_l\pi_m), \tag{4.3}$$

e

$$\begin{aligned} V_{x|k} = V(\hat{\bar{x}}|k \in S) = \frac{1}{N^2}\sum_{\substack{l\in U\\ l\neq k}} \frac{x_l^2\pi_{kl}}{\pi_k\pi_l^2}\left(1-\frac{\pi_{kl}}{\pi_k}\right) + \\ \frac{1}{N^2}\sum_{\substack{l\in U\\ l\neq k}}\sum_{\substack{m\in U\\ m\neq l\\ m\neq k}} \frac{x_l x_m}{\pi_k\pi_l\pi_m}\left(\pi_{klm}-\frac{\pi_{kl}\pi_{km}}{\pi_k}\right). \end{aligned} \tag{4.4}$$

Note-se que $V(\hat{\bar{x}})$ pode ser escrita como

$$V(\hat{\bar{x}}) = \frac{1}{N}\sum_{k\in U}(\bar{x}_{|k}-\bar{x})x_k.$$

Como exemplo, em um plano de amostragem aleatória simples, tem-se que as expressões (4.2), (4.3) e (4.4) são expressas por

$$\bar{x}_{|k} = \bar{x} + \frac{N-n}{N-1}\frac{x_k-\bar{x}}{n}, \tag{4.5}$$

$$V_x = \frac{N-n}{N-1}\frac{\sigma_x^2}{n}, \tag{4.6}$$

e

$$V_{x|k} = \frac{N(N-n)(n-1)}{(N-2)(N-1)n^2}\left\{\sigma_x^2 - \frac{(x_k-\bar{x})^2}{N-1}\right\}, \tag{4.7}$$

onde

$$\sigma_x^2 = \frac{1}{N}\sum_{k\in U}(x_k-\bar{x})^2.$$

Para amostragem aleatória simples, a normalidade do estimador da média foi provada por Madow (1948) sobre algumas condições e para tamanhos

de amostra grandes. Supondo que $\hat{\bar{x}}$ segue distribuição normal condicional e incondicionalmente na presença das unidades amostrais ($k \in S$), então tem-se que

$$a_k(\hat{\bar{x}}) = \frac{n}{N\pi_{k|\hat{\bar{x}}}} = \frac{P(\hat{\bar{x}})}{P(\hat{\bar{x}}|k \in S)} = \frac{f(\hat{\bar{x}})}{f_k(\hat{\bar{x}})}$$

em que $f(\hat{\bar{x}})$ e $f_k(\hat{\bar{x}})$ são as funções de densidade de uma variável que segue distribuição normal com médias $\bar{x}$ e $\bar{x}_{|k}$, e variâncias V_x e $V_{x|k}$, respectivamente. Desta maneira

$$a_k(\hat{\bar{x}}) = \frac{V_x^{-1/2} \exp\left(-\frac{(\hat{\bar{x}}-\bar{x})^2}{2V_x}\right)}{V_{x|k}^{-1/2} \exp\left(-\frac{(\hat{\bar{x}}-\bar{x}_{|k})^2}{2V_{x|k}}\right)}. \tag{4.8}$$

Então, o estimador SCW é dado por

$$\hat{\bar{y}}_{|\eta} = \frac{1}{n}\sum_{k\in S} a_k(\hat{\bar{x}})y_k.$$

Resultado 1. *Uma aproximação para o estimador SCW de $\bar{y}$ condicionado por $\hat{\bar{x}}$, no caso de AAS e se $\hat{\bar{x}}$ tem uma distribuição normal incondicional e condicionalmente na presença de cada unidade na amostra, é dada por*

$$\bar{y}_{|\hat{\bar{x}}} = \hat{\bar{y}} + (\bar{x} - \hat{\bar{x}})D^* + O_p(n^{-1}). \tag{4.9}$$

em que $D^* = \frac{1}{n\sigma_x^2}\sum_{k\in S}(x_k - \bar{x})y_k$.

A prova do resultado 1 é apresentada no Apêndice B.

É possível observar a semelhança do estimador de regressão com a expressão dada em (4.9). A diferença está em que a forma usual do estimador de regressão é usado $D = \frac{1}{n\hat{\sigma}_x^2}\sum_{k\in S}(x_k - \hat{\bar{x}})y_k$, no lugar de D^*. Ou seja,

$$\hat{\bar{y}}_R = \hat{\bar{y}} + (\bar{x} - \hat{\bar{x}})\frac{1}{n\hat{\sigma}_x^2}\sum_{k\in S}(x_k - \hat{\bar{x}})y_k.$$

Então, usando o resultado 1 é possível introduzir o estimador de regressão como uma aproximação natural do estimador SCW para grandes amostras.

CAPÍTULO 5

Estimador de Regressão Generalizado Logístico (LGREG)

Como foi apresentado no capítulo 3, o estimador GREG para o total t_y pode ser assistido por um modelo de regressão linear. Entretanto, quando a variável de interesse está categorizada, um modelo linear pode não ser razoável. É natural que, no caso em que Y é dicotômica, seja preferível um modelo logístico, pois este é mais apropriado. Neste contexto, é possível definir um estimador como um caso particular do estimador de regressão generalizado GREG, onde a variável de interesse é dicotômica e o modelo formulado é um modelo de regressão logística. Na presença da matriz de informação auxiliar $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_J)$, o estimador de regressão generalizado logístico (LGREG) para o total da variável de interesse y foi proposto por Lehtonen e Veijanen (1998a, 1998b) e pode ser expresso como

$$\hat{t}_{_{LGREG}} = \sum_{k \in U} \hat{\pi}_S(\mathbf{x}_k) + \sum_{k \in S} \frac{y_k - \hat{\pi}_S(\mathbf{x}_k)}{\pi_k} = \sum_{k \in S} \frac{g_{ks} y_k}{\pi_k}, \tag{5.1}$$

onde

$$\hat{\pi}_S(\mathbf{x}_k) = \frac{\exp\left(h(\hat{\boldsymbol{\beta}}_S^{\pi}; \mathbf{x}_k)\right)}{1 + \exp\left(h(\hat{\boldsymbol{\beta}}_S^{\pi}; \mathbf{x}_k)\right)} \tag{5.2}$$

e

$$g_{ks} = 1 + \frac{\sum_{k \in U} \hat{\pi}_S(\mathbf{x}_k) - \sum_{k \in S} \frac{\hat{\pi}_S(\mathbf{x}_k)}{\pi_k}}{\hat{t}_{\pi}}. \tag{5.3}$$

Neste caso, a estimação é assistida pelo seguinte modelo para Y_k

$$\begin{cases} E(Y_k) = \pi(\mathbf{x}_k) \\ V(Y_k) = \pi(\mathbf{x}_k)[1 - \pi(\mathbf{x}_k)] \\ \log\left(\dfrac{\pi(\mathbf{x}_k)}{1 - \pi(\mathbf{x}_k)}\right) = h(\boldsymbol{\beta}; \mathbf{x}_k) \end{cases} \tag{5.4}$$

Do Lema 1, pode-se concluir que, se o modelo formulado em (5.4) considera intercepto então o estimador $\hat{t}_{_{LGREG}}$ assume a seguinte forma

$$\hat{t}_{_{LGREG}} = \sum_{k \in U} \hat{\pi}_S(\mathbf{x}_k).$$

As estimativas do parâmetro $\boldsymbol{\beta}$, e por conseguinte, as estimativas de $\pi(\mathbf{x}_k)$ são obtidas maximizando a função de pseudo log-verossimilhança, que para o caso Bernoulli, adota a seguinte expressão

$$L_S(\boldsymbol{\beta}) = \sum_{k \in S} \frac{1}{\pi_k} \left\{ y_k \log(\pi(\boldsymbol{\beta}; \mathbf{x}_k)) + \log(1 - \pi(\boldsymbol{\beta}; \mathbf{x}_k))(1 - y_k) \right\}.$$

Para esta maximização podem ser usados, por exemplo, o método de Newton-Raphson ou o método scoring de Fisher.

Uma expressão para a variância aproximada do estimador $\hat{t}_{_{LGREG}}$ pode ser obtida aplicando a expressão dada em (3.4), em que

$$E_k = y_k - \hat{\pi}_U(\mathbf{x}_k), \tag{5.5}$$

onde

$$\hat{\pi}_U(\mathbf{x}_k) = \frac{\exp\left(h(\hat{\boldsymbol{\beta}}_U; \mathbf{x}_k)\right)}{1 + \exp\left(h(\hat{\boldsymbol{\beta}}_U; \mathbf{x}_k)\right)},$$

com

$$\hat{\boldsymbol{\beta}}_U \; = \; \arg\max_{\boldsymbol{\beta}} \left\{ \sum_{k \in U} [y_k \log(\pi(\boldsymbol{\beta}; \mathbf{x}_k)) + \log(1 - \pi(\boldsymbol{\beta}; \mathbf{x}_k))(1 - y_k)] \right\}.$$

Para o estimador da variância do estimador $\hat{t}_{_{LGREG}}$ tem-se duas opções, a primeira, denotada por $\hat{V}_1(\hat{t}_{_{LGREG}})$, pode ser expressa na forma abaixo

$$\hat{V}_1(\hat{t}_{_{LGREG}}) = \sum_{k \in S} \sum_{l \in S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{g_{ks} e_k}{\pi_k} \frac{g_{ls} e_l}{\pi_l},$$

com $e_k = y_k - \hat{\pi}_S(\mathbf{x}_k)$ e g_{ks} como na equação (5.3). A segunda assume a seguinte expressão

$$\hat{V}_2(\hat{t}_{LGREG}) = \sum_{k \in S} \sum_{l \in S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k}{\pi_k} \frac{e_l}{\pi_l}.$$

5.1 Estimação de Proporções

A estimação de proporções é um dos importantes objetivos de levantamentos amostrais onde a variável de interesse é dicotômica. Neste contexto, é possível considerar a estimação em presença de informação auxiliar, caso em que podem ser aplicados os estimadores de regressão generalizados (GREG), e os estimadores de regressão generalizados logísticos (LGREG). Entretanto, na prática é muito comum abordar a estimação de proporções assumindo a variável de interesse como se fosse contínua e formulando um modelo de regressão linear entre a variável de interesse e as variáveis auxiliares, o qual pode não ser adequado devido à natureza da variável de interesse. Por exemplo, quando existe somente uma variável auxiliar e o plano adotado para a seleção dos indivíduos que comporão à amostra é uma amostragem aleatória simples, podem ser consideradas três opções para o modelo que assiste a estimação.

5.1.1 GREG Usando um Modelo de Regressão Linear sem Intercepto

Este modelo foi apresentado em (3.5), que adota a seguinte expressão para o estimador da proporção:

$$\hat{P}_{GREG1} = \frac{\hat{t}_{GREG1}}{N} = \left(\frac{\bar{x}_U}{\bar{x}_S} \right) \hat{P}_{HT},$$

em que

$$\hat{P}_{HT} = \frac{\hat{t}_\pi}{N} \tag{5.6}$$

é o estimador de Horvitz-Thompson para a porporção de sucessos P. Nesse caso, a sua variância é dada pela seguinte expressão

$$V(\hat{P}_{GREG1}) = \frac{(N-n) \sum_{k \in U} E_k^2}{nN(N-1)}, \tag{5.7}$$

com $E_k = y_k - \hat{\beta}_U x_k$ e $\hat{\beta}_U = \frac{P}{\bar{x}_U}$. O estimador da variância do estimador $\hat{P}_{GREG1}$ pode ser escrito como

$$\hat{V}(\hat{P}_{GREG1}) = \left(\frac{\bar{x}_U}{\bar{x}_S}\right)^2 \frac{(N-n)\sum_{k\in S} e_k^2}{nN(N-1)},$$

onde $e_k = y_k - \hat{\beta}_S^{\pi} x_k$ e $\hat{\beta}_S^{\pi}$ como em (3.6).

5.1.2 GREG Usando um Modelo de Regressão Linear com Intercepto

Este tipo de modelo foi apresentado em (3.10). A expressão para este estimador é a seguinte

$$\hat{P}_{GREG2} = \frac{\hat{t}_{GREG2}}{N} = \hat{P}_{HT} + \hat{\beta}_S^{\pi}(\bar{x}_U - \bar{x}_S), \tag{5.8}$$

em que

$$\hat{\beta}_S^{\pi} = \frac{\sum_{k\in S} x_k y_k - \hat{P}_{HT}\bar{x}_S}{\sum_{k\in S}(x_k - \bar{x})^2}.$$

Sua variância pode ser expressa como em (5.7), com E_k como na equação (3.11). O estimador da variância do estimador é dado da seguinte forma

$$\hat{V}(\hat{P}_{GREG2}) = \frac{(N-n)\sum_{k\in S}(\tilde{e}_k - \bar{\tilde{e}})^2}{nN(N-1)} \tag{5.9}$$

com $\tilde{e}_k = g_{ks} e_k$, $e_k = y_k - \hat{\alpha}_S^{\pi} - \hat{\beta}_S^{\pi} x_k$, $\hat{\alpha}_S^{\pi} = \hat{P}_{HT} - \hat{\beta}_S^{\pi}\bar{x}_S$, e

$$g_{ks} = 1 + \frac{n(\bar{x}_U - \bar{x}_S)(x_k \bar{x}_S)}{\sum_{k\in S}(x_k - \bar{x}_S)^2}. \tag{5.10}$$

5.1.3 GREG Usando um Modelo de Regressão Logística (LGREG)

Este modelo foi dado em (5.4) considerando intercepto, que adota a seguinte forma

$$\hat{P}_{LGREG} = \frac{\hat{t}_{LGREG}}{N} = \frac{\sum_{k\in U} \hat{\pi}_S(x_k)}{N}. \tag{5.11}$$

A variância do estimador adota a forma dada em (5.7), com E_k como em (5.5). As expressões do estimador da variância do estimador são dadas por

$$\hat{V}_1(\hat{P}_{LGREG}) = \frac{(N-n) g_{ks}^2 \sum_{k\in S} e_k^2}{nN(n-1)}, \tag{5.12}$$

com $e_k = y_k - \hat{\pi}_S(x_k)$ e

$$g_{ks} = \frac{\hat{P}_{LGREG}}{\hat{P}_{HT}}, \tag{5.13}$$

$$\hat{V}_2(\hat{P}_{LGREG}) = \frac{(N-n)\sum_{k\in S} e_k^2}{nN(n-1)}. \tag{5.14}$$

O estimador $\hat{P}_{LGREG}$ descrito em (5.11) pode ser mais adequado que $\hat{P}_{GREG1}$ e $\hat{P}_{GREG2}$ para a estimação de proporções. Entretanto, este estimador requer para sua aplicação conhecimento da variável auxiliar para todos os indivíduos da população, enquanto que, $\hat{P}_{GREG1}$ por exemplo, requer somente conhecimento da variável auxiliar para os indivíduos que compõem a amostra e o total populacional, o que facilita sua aplicação.

5.2 Estimador de Regressão Generalizado Logístico no Contexto de Estratificação

Nesta seção apresenta-se a forma adotada pelo estimador de regressão generalizado logístico, sob amostragem estratificada, usando como referência as expressões apresentadas na seção 2.3.1.

5.2.1 Estimador de Regressão Generalizado Logístico Combinado

O estimador de regressão generalizado logístico combinado (LGREGC) para o total t_y é denotado por $\hat{t}_{LGREGC}$ e, é expresso como na equação (5.1), em que $\hat{\pi}_S(\mathbf{x}_k)$ como em (5.2) e as estimativas do parâmetro $\boldsymbol{\beta}$ são obtidas maximizando a seguinte função de pseudo log-verossilhança

$$L_S(\boldsymbol{\beta}) = \sum_{h=1}^{H} \sum_{k\in S_h} \frac{1}{\pi_k} \left\{ y_k \log(\pi(\boldsymbol{\beta}; \mathbf{x}_k)) + \log(1 - \pi(\boldsymbol{\beta}; \mathbf{x}_k))(1 - y_k) \right\}.$$

5.2.2 Estimador de Regressão Generalizado Logístico Separado

O estimador de regressão generalizado logístico separado (LGREGS), pode ser expresso por

$$\hat{t}_{_{LGREGS}} = \sum_{h=1}^{H} \left[\sum_{k \in U_h} \hat{\pi}_{S_h}(\mathbf{x}_k) + \sum_{k \in S_h} \frac{(y_k - \hat{\pi}_{S_h}(\mathbf{x}_k))}{\pi_k} \right],$$

onde

$$\hat{\pi}_{S_h}(\mathbf{x}_k) = \frac{\exp\left(h(\hat{\boldsymbol{\beta}}^{\pi}_{S_h}; \mathbf{x}_k)\right)}{1 + \exp\left(h(\hat{\boldsymbol{\beta}}^{\pi}_{S_h}; \mathbf{x}_k)\right)},$$

e $\hat{\boldsymbol{\beta}}^{\pi}_{S_h} = \arg\max_{\boldsymbol{\beta}} L_{S_h}(\boldsymbol{\beta})$, com

$$L_{S_h}(\boldsymbol{\beta}) = \sum_{k \in S_h} \frac{1}{\pi_k} \left\{ y_k \log(\pi(\boldsymbol{\beta}; \mathbf{x}_k)) + \log(1 - \pi(\boldsymbol{\beta}; \mathbf{x}_k))(1 - y_k) \right\}.$$

CAPÍTULO 6

Avaliação dos estimadores

6.1 Estudo de Simulação

Considere uma população finita $U = \{1, 2, \ldots, N\}$, da qual é de interesse estimar um parâmetro θ através do estimador $\hat{\theta}$. Quando as propriedades de $\hat{\theta}$ não podem ser descritas analiticamente devido à sua complexidade é possível obter uma aproximação de suas características usando um estudo de simulação de Monte Carlo. Entretanto, sob o plano amostral $p(\cdot)$, o número total de amostras possíveis T pode ser muito amplo. Nesses casos, procede-se a simular somente M $(M < T)$ das amostras possíveis. As medidas de interesse são a esperança de $\hat{\theta}$, $E(\hat{\theta})$, a variância $V(\hat{\theta})$ e a taxa de cobertura $TC(\hat{\theta}, \hat{V}(\hat{\theta}), \alpha)$ cujas expressões foram dadas em (2.1), (2.2) e (2.4), respectivamente. Neste contexto é assumida uma amostra aleatória, ou seja, uma sucesão $\hat{\theta}(S_1), \hat{\theta}(S_2), \ldots, \hat{\theta}(S_M)$ de variáveis aleatórias independentes e identicamente distribuidas com $E_p(\hat{\theta}(S_t)) = E_p(\hat{\theta})$ e $V_p(\hat{\theta}(S_t)) = V_p(\hat{\theta})$, onde $t = 1, \ldots, M$. Então pode-se definir $\bar{\hat{\theta}}$ como

$$\bar{\hat{\theta}} = \sum_{t=1}^{M} \frac{\hat{\theta}(S_t)}{M},$$

em que

$$E_p(\bar{\hat{\theta}}) = E_p(\hat{\theta}) \qquad \text{e} \qquad V_p(\bar{\hat{\theta}}) = \frac{V_p(\hat{\theta})}{M}.$$

A variância de $\hat{\theta}$ pode ser estimada sem tendência usando a seguinte expressão

$$S^2_{\hat{\theta}} = \sum_{t=1}^{M} \frac{(\hat{\theta}(S_t) - \bar{\hat{\theta}})^2}{M-1},$$

onde

$$E_p(S^2_{\hat{\theta}}) = V(\hat{\theta}) \qquad \text{e} \qquad V_p(S^2_{\hat{\theta}}) \longrightarrow 0 \quad \text{quando } M \text{ aumenta.}$$

A taxa de cobertura pode ser estimada por

$$\widehat{TC}_p(\hat{\theta}, \hat{V}(\hat{\theta}), \alpha) = \sum_{t=1}^{M} \frac{Z(S_t)}{M}, \tag{6.1}$$

em que

$$E_p(\widehat{TC}) = TC \qquad \text{e} \qquad V_p(\widehat{TC}) = \frac{TC[1-TC]}{M}.$$

Os resultados numéricos apresentados neste capítulo fornecem uma aproximação às propriedades dos estimadores de Horvitz-Thompsom (HT), de regressão generalizado (GREG) e de regressão generalizado logístico (LGREG), quando são usados para estimar a proporção P de sucessos em uma população finita. No processo de simulação foi considerado um vetor de informação auxiliar composto por uma só variável, a qual foi gerada seguindo distribuição normal padrão. Os modelos formulados entre a variável de interesse e a variável auxiliar foram um modelo de regressão linear com intercepto como em (3.10) e um modelo de regressão logística como foi dado em (5.4) considerando intercepto, para os estimadores $\hat{P}_{_{GREG}}$ e $\hat{P}_{_{LGREG}}$, respectivamente.
As populações de interesse foram geradas com um tamanho de $N = 10000$ indivíduos cada uma, variando os seguintes parâmetros com o objetivo de obter populações com diferentes estruturas e/ou características:

- Grau de associação entre a variável de interesse e a variável auxiliar (OR), com valores: 0.1, 0.2, 0.5, 0.65, 1.5, 2, 5 e 10.

- Proporção de indivíduos na população com o atributo de interesse P com valores: 0.2, 0.3 e 0.5.

Para gerar cada uma das populações seguindo os parâmetros descritos acima foi usado o seguinte procedimento:

Passo 1. Gerar 10000 números aleatórios seguindo uma distribuição normal padrão, denotados por x_k, com $k = 1, 2, \ldots, 10000$.

Passo 2. Calcular a probabilidade de sucesso $\pi(\mathbf{x}_k)$, sendo construída seguindo a grau e a direção da associação com a variável auxiliar dada pela razão de chances (veja expressão 2.12), então

$$\pi(\mathbf{x}_k) = \frac{\exp(\beta_0 + \beta_1 x_k)}{1 + \exp(\beta_0 + \beta_1 x_k)},$$

com

$$\beta_0 = \log\left(\frac{P}{1-P}\right) \tag{6.2}$$

de acordo com o resultado obtido no Apêndice C e

$$\beta_1 = \log(OR).$$

Observe que β_0 não foi considerado diretamente no processo de simulação, pois na prática é de maior interesse conhecer as propriedades e o desempenho dos estimadores para diferentes valores de P do que para diferentes valores de β_0.

Passo 3. Calcular os valores de y_k, conhecendo o valor de $\pi(\mathbf{x}_k)$ para toda a população e gerando 10000 números aleatórios seguindo distribuição uniforme $(0, 1)$, denotados por u_k com $k = 1, 2, \ldots, 10000$, a variável de interesse na população y_k foi gerada seguindo uma distribuição de Bernoulli.

$$y_k = \begin{cases} 1, & \text{se } u_k \leq \pi(x_k); \\ 0, & \text{se } u_k > \pi(x_k). \end{cases}$$

Para determinar as propriedades exatas dos estimadores investigados seria necessário obter as estimativas do parâmetro de interesse para cada uma das amostras possíveis. Entretanto, dado que este número é demasiado grande, foram utilizadas $M = 10000$ amostras no estudo de Monte Carlo. As comparações entre os estimadores foram feitas através das seguintes medidas:

1. Estimação do viés relativo do estimador

$$VR_{\hat{P}} = \frac{|\bar{\hat{P}} - P|}{P}. \tag{6.3}$$

2. Estimação da eficiência do estimador, empregando o estimador $\hat{P}_{HT}$ como referência

$$eff(\hat{P}) = \frac{S^2_{\hat{P}}}{S^2_{\hat{P}_{HT}}}. \tag{6.4}$$

3. Estimação da eficiência do estimador do ponto do vista do erro quadrático médio (EQM)

$$\Lambda(\hat{P}) = \frac{EQM(\hat{P})}{EQM(\hat{P}_{HT})} = \frac{S^2_{\hat{P}} + (\bar{\hat{P}} - P)^2}{S^2_{\hat{P}_{HT}} + (\bar{\hat{P}}_{HT} - P)^2}. \tag{6.5}$$

4. Estimação do viés relativo do estimador da variância do estimador

$$VRV_{\hat{V}(\hat{P})} = \frac{|\bar{\hat{V}}(\hat{P}) - S^2_{\hat{P}}|}{S^2_{\hat{P}}}. \tag{6.6}$$

5. Estimação do coeficiente de variação do estimador de $V(\hat{P})$

$$CV(\hat{P}, \hat{V}(\hat{P})) = 100\frac{\sqrt{S^2_{\hat{V}(\hat{P})}}}{\bar{\hat{V}}(\hat{P})}. \tag{6.7}$$

6. Taxa de cobertura para um intevalo de confiança, calculado como em (6.1).

O desempenho dos estimadores foi observado em planos de amostragem aleatória simples (AAS), amostragem de Bernoulli (BE) e amostragem estratificada (AE).

6.1.1 Amostragem Aleatória Simples

Para este cenário de simulação foram geradas populações para cada combinação do grau de associação e proporção de indivíduos com o atributo de

interesse na população, resultando em 24 populações, de cada uma delas foram selecionadas 10000 (réplicas) amostras de tamanho $n = 50, 100, 250, 500$ e 1000 seguindo um plano AAS. Para cada uma das amostras selecionadas foram calculadas as estimativas dos parâmetros P usando as expressões apresentadas a seguir:

- Estimador de Horvitz-Thompson, $\hat{P}_{HT}$. Para este estimador foi usada a expressão dada em (5.6), e como estimador da variância do estimador

$$\hat{V}(\hat{P}_{HT}) = \frac{(N-n)\hat{P}_{HT}(1-\hat{P}_{HT})}{N(n-1)}. \tag{6.8}$$

- Estimador GREG, $\hat{P}_{GREG}$. Este estimador foi calculado usando a equação (5.8), e o estimador da variância de $\hat{P}_{GREG}$ é expresso como em (5.9).

- Estimador LGREG, $\hat{P}_{LGREG}$. A expressão para calcular este estimador foi apresentada em (5.11), e para o estimador da variância do estimador tem-se duas expressões apresentadas em (5.12) e (5.14).

Uma vez obtidas estas estimações foram calculadas as medidas dadas nas expressões (6.3), (6.4), (6.5), (6.6), (6.7) e (6.1).

6.1.2 Amostragem de Bernoulli

Neste cenário de simulação foram geradas 24 populações seguindo o procedimento descrito na seção anterior. De cada uma destas populações foram selecionadas 10000 (réplicas) amostras de tamanho esperado $n = 50, 100, 250,$ 500 e 1000 seguindo um plano BE. Em cada amostra selecionada foram aplicados os estimadores $\hat{P}_{HT}$, $\hat{P}_{GREG}$ e $\hat{P}_{LGREG}$, usando as expressões apresentadas a seguir:

- Estimador Horvitz-Thompson. Para calcular este estimador foi usada a equação (5.6)que em amostragem de Bernoulli adota a seguinte expressão

$$\hat{P}_{HT} = \frac{n_S}{N\pi}\bar{y}_S,$$

onde n_S é o tamanho da amostra. A variância deste estimador pode ser escrita como

$$V(\hat{P}_{HT}) = \frac{1}{N}\left(\frac{1}{\pi} - 1\right) P,$$

e o estimador da variância de $\hat{P}_{HT}$ é dado por

$$\hat{V}(\hat{P}_{HT}) = \frac{1}{N}\left(\frac{1}{\pi} - 1\right) \hat{P}_{HT}.$$

- Estimador GREG. Este estimador apresentado em (5.8) assume a seguinte forma:

$$\hat{P}_{GREG} = \frac{N\pi}{n_S}\hat{P}_{HT} + \hat{\beta}_S^{\pi}(\bar{x}_U - \bar{x}_S),$$

em que

$$\hat{\beta}_S^{\pi} = \frac{\sum_{k\in S} x_k y_k - n_S \bar{y}_S \bar{x}_S}{\sum_{k\in S}(x_k - \bar{x}_S)^2}.$$

A variância de $\hat{P}_{GREG}$ pode ser expressa como segue

$$V(\hat{P}_{GREG}) = \frac{1}{N^2}\left(\frac{1}{\pi} - 1\right) \sum_{k\in U} E_k^2, \tag{6.9}$$

com E_k como na expressão (3.11). Além disso, o estimador da variância $\hat{P}_{GREG}$ é dado por

$$\hat{V}(\hat{P}_{GREG}) = \frac{1}{N^2\pi}\left(\frac{1}{\pi} - 1\right) \sum_{k\in S} (\tilde{e} - \bar{\tilde{e}})^2,$$

com $\tilde{e}_k = e_k g_{ks}$ e

$$g_{ks} = \frac{N\pi}{n_S}\left[1 + \frac{(x_k - \bar{x}_S)(\bar{x}_U - \bar{x}_S)n_S}{\sum_{k\in S}(x_k - \bar{x}_S)^2}\right].$$

- Estimador LGREG. Usando a expressão (5.11) este estimador pode ser expresso por

$$\hat{P}_{LGREG} = \sum_{k\in U} \frac{\hat{\pi}_S(x_k)}{N}.$$

A variância de $\hat{P}_{GREG}$ é dada por (6.9), com E_k dado na expressão (5.5). Os estimadores da variância do estimador $\hat{P}_{LGREG}$ são apresentados a seguir

$$\hat{V}_1(\hat{P}_{LGREG}) = \frac{g_{ks}^2}{N^2\pi}\left(\frac{1}{\pi} - 1\right) \sum_{k\in S} e_k^2,$$

em que g_{ks} como na equação (5.13), e o segundo é dado por

$$\hat{V}_2(\hat{P}_{_{LGREG}}) = \frac{1}{N^2\pi}\left(\frac{1}{\pi} - 1\right)\sum_{k\in S} e_k^2.$$

Com as estimativas apresentadas acima para cada amostra, procede-se a calcular as medidas dadas nas expressões (6.3), (6.4), (6.5), (6.6), (6.7) e (6.1).

6.1.3 Amostragem Aleatória Estratificada

Sob este plano foram considerados 2 cenários de simulação diferentes, com o intuito de avaliar o desempenho dos estimadores quando é usada a estratificação da população. Mais especificamente os estimadores de regressão separados e combinados. Foram geradas 12 populações de tamanho $N = 10000$, com três estratos cada uma. Os tamanos dos estratos considerados foram $N_1 = 5000$, $N_2 = 3000$ e $N_3 = 2000$. O tamanho das amostras considerados estão dados pelas frações amostrais $f_1 = 1\%$, $f_2 = 5\%$, e $f_3 = 10\%$, tendo desta maneira $n = 100, 500$ e 1000. Os tamanhos das amostras em cada estrato são proporcionais ao tamanho do estrato, obtendo amostras para o estrato 1 de tamanhos $n_1 = 50, 250$ e 500, para o estrato 2 $n_2 = 30, 150$ e 300 e no estrato 3 tamanhos $n_3 = 20, 100$ e 200. As amostras em cada estrato foram selecionadas seguindo um plano AAS. A proporção de indivíduos na população com o atributo de interesse P obedeceu os valores: 0.2, 0.3 e 0.5.

Dois cenários gerais foram considerados, diferindo com respeito aos valores das razões de chances (OR) em cada estrato:

- **Cenário 1:** O grau de associação entre a variável de interesse e a variável auxiliar varia entre estratos. Neste cenário foram consideradas duas formas de estratificação, como sugere o Quadro (6.1).
- **Cenário 2:** A razão de chances (OR) foi mantida a mesma para cada estrato, sendo igual a 0.5 na forma (a) e a 5 na forma (b).

Quadro 6.1. Variação do OR entre estratos para o Cenário 1.

Estrato	Forma	
	(a)	(b)
1	10	2
2	0.2	0.1
3	0.5	5

A geração de cada um dos estratos foi feita independentemente, seguindo os valores acima. A seguir serão apresentadas as fórmulas que adotam os estimadores $\hat{P}_{_{HT}}$, $\hat{P}_{_{GREG}}$ e $\hat{P}_{_{LGREG}}$, quando o plano usado para selecionar os indivíduos que comporão a amostra é um AAE.
O estimador de Horvitz-Thompson para a proporção populacional sob um plano AAE assume a seguinte forma:

$$\hat{P}_{_{HT}} = \sum_{h=1}^{H} a_h \hat{P}_{_{HT_h}},$$

com $\hat{P}_{_{HT_h}} = \bar{y}_{S_h}$ e $a_h = \frac{N_h}{N}$. A variância de $\hat{P}_{_{HT}}$ é dada na seguinte expressão

$$V(\hat{P}_{_{HT}}) = \sum_{h=1}^{H} a_h^2 \frac{N_h(N_h - n_h)}{n_h(N_h - 1)} P_h(1 - P_h).$$

O estimador de $V(\hat{P}_{_{HT}})$ é expresso como segue

$$\hat{V}(\hat{P}_{_{HT}}) = \sum_{h=1}^{H} a_h^2 \frac{(N_h - n_h)}{N_h(n_h - 1)} \hat{P}_{_{HT_h}}(1 - \hat{P}_{_{HT_h}}).$$

A seguir, serão apresentadas as expressões que adotam os estimadores $\hat{P}_{_{GREG}}$ e $\hat{P}_{_{LGREG}}$, quando é formulado um modelo entre a variável de interesse e a variável resposta em cada um dos estratos, ou seja, será apresentado o caso dos estimadores de regressão separados:

- Estimador $\hat{P}_{_{GREGS}}$ pode ser expresso por

$$\hat{P}_{_{GREGS}} = \sum_{h=1}^{H} a_h [\hat{P}_{_{HT}} + \hat{\beta}^{\pi}_{S_h}(\bar{x}_{U_h} - \bar{x}_{S_h})],$$

em que

$$\hat{\beta}^{\pi}_{S_h} = \frac{\sum_{k\in S_h}(x_k - \bar{x}_{S_h})(y_k - \hat{P}_{HT_h})}{\sum_{k\in S_h}(x_k - \bar{x}_{S_h})^2}.$$

A variância de $\hat{P}_{GREGS}$ é dada por

$$V(\hat{P}_{GREGS}) = \sum_{h=1}^{H} a_h^2 \frac{(N_h - n_h)}{n_h N_h (N_h - 1)} \sum_{k\in U_h} E_k^2, \tag{6.10}$$

em que E_k é dado na expressão (3.11). Além disso, o estimador da variância de $\hat{P}_{GREGS}$ é dado por

$$\hat{V}(\hat{P}_{GREGS}) = \sum_{h=1}^{H} a_h^2 \frac{(N_h - n_h)}{n_h N_h (n_h - 1)} \sum_{k\in S_h} (\tilde{e}_k - \bar{\tilde{e}})^2,$$

em que $\tilde{e}_k = g_{ks_h} e_k$ e g_{ks_h} como em (5.10).

- Estimador $\hat{P}_{LGREGS}$, este estimador assume a seguinte forma:

$$\hat{P}_{LGREGS} = \sum_{h=1}^{H} a_h \hat{P}_{LGREG_h}.$$

A variância do estimador é expressa por (6.10), mas com E_k como em (5.5). Os estimadores da variância do estimador $\hat{P}_{LGREGS}$ são dados por

$$\hat{V}_1(\hat{P}_{LGREGS}) = \sum_{h=1}^{H} a_h^2 \frac{(N_h - n_h)}{n_h N_h (n_h - 1)} g_{ks_h}^2 \sum_{k\in S_h} e_k^2,$$

com $g_{ks_h} = \frac{\hat{P}_{LGREG_h}}{\hat{P}_{HT_h}}$, e

$$\hat{V}_2(\hat{P}_{LGREGS}) = \sum_{h=1}^{H} a_h^2 \frac{(N_h - n_h)}{n_h N_h (n_h - 1)} \sum_{k\in S_h} e_k^2.$$

A seguir são apresentadas as fórmulas para os estimadores da proporção P, do tipo combinado, ou seja, onde é formulado o mesmo modelo entre a variável de interesse e variável resposta para toda a população, sem fazer diferença entre estratos.

- O estimador $\hat{P}_{GREGC}$ pode ser expresso da seguinte maneira

$$\hat{P}_{GREGC} = \hat{P}_{HT} + \hat{\beta}_S^{\pi} \left(\bar{x}_U - \sum_{h=1}^{H} a_h \bar{x}_{S_h} \right),$$

em que

$$\hat{\beta}_S^{\pi} = \frac{\sum_{h=1}^{H} \frac{N_h}{n_h} \left[\sum_{k \in S_h} (x_k - \tilde{x}_S)(y_k - \hat{P}_{HT}) \right]}{\sum_{h=1}^{H} \frac{N_h}{n_h} \left[\sum_{k \in S_h} (x_k - \bar{x}_S)^2 \right]},$$

e $\tilde{x}_S = \sum_{h=1}^{H} a_h \bar{x}_{S_h}$. A variância de $\hat{P}_{GREGC}$ é dada por

$$V(\hat{P}_{GREGS}) = \sum_{h=1}^{H} \left[a_h^2 \frac{(N_h - n_h)}{n_h N_h (N_h - 1)} \sum_{k \in U_h} (E_k - \bar{E}_h)^2 \right], \qquad (6.11)$$

em que E_k é dado por (3.11) e $\bar{E}_h = \frac{1}{N_h} \sum_{k \in U_h} E_k$. O estimador da variância do estimador $\hat{P}_{GREGC}$ é dado por

$$\hat{V}(\hat{P}_{GREGC}) = \sum_{h=1}^{H} a_h^2 \frac{(N_h - n_h)}{n_h N_h (n_h - 1)} \sum_{k \in S_h} (\tilde{e}_k - \bar{\tilde{e}}_h)^2,$$

em que $\tilde{e}_k = g_{ks} e_k$, $\bar{\tilde{e}}_h = \frac{1}{n_h} \sum_{k \in S_h} \tilde{e}_k$ e

$$g_{ks} = 1 + \frac{(x_k - \tilde{x}_S)(\bar{x}_U - \tilde{x}_S) N}{\sum_{h=1}^{H} \sum_{k \in S_h} \frac{N_h}{n_h} (x_k - \tilde{x}_S)^2}$$

- O estimador $\hat{P}_{LGREGC}$ assume a seguinte forma:

$$\hat{P}_{LGREGC} = \sum_{h=1}^{H} \sum_{k \in U_h} \frac{\hat{\pi}_S(x_k)}{N},$$

A variância do estimador é expressa por (6.11), onde E_k como foi apresentado em (5.5). Os estimadores da variância do estimador $\hat{P}_{LGREGC}$ são dados por

$$\hat{V}_1(\hat{P}_{LGREGC}) = \sum_{h=1}^{H} a_h^2 \frac{(N_h - n_h)}{n_h N_h (n_h - 1)} g_{ks}^2 \sum_{k \in S_h} (e_k - \bar{e}_h)^2,$$

com g_{ks} dado em (5.13), e

$$\hat{V}_2(\hat{P}_{LGREGC}) = \sum_{h=1}^{H} a_h^2 \frac{(N_h - n_h)}{n_h N_h (n_h - 1)} \sum_{k \in S_h} (e_k - \bar{e}_h)^2.$$

6.2 Resultados

Os resultados do processo de simulação referentes ao estimador $\hat{P}_{_{LGREG}}$ permitem fazer as seguintes considerações:

- O viés relativo do estimador: em todos os cenários considerados (veja os Quadros 6.2, 6.8 e 6.14) o estimador $\hat{P}_{_{LGREG}}$ apresenta um viés relativo que pode ser considerado como desprezível, pois esta medida em todos os casos é menor do que $1\%(0.01)$, ou seja, o estimador $\hat{P}_{_{LGREG}}$ pode ser considerado como assintoticamente não-viesado.

- A eficiência do estimador: o estimador $\hat{P}_{_{LGREG}}$ em todos os cenários de simulação (veja Quadros 6.3, 6.9 e 6.15) apresenta-se mais eficiente que os estimadores $\hat{P}_{_{HT}}$ e $\hat{P}_{_{GREG}}$. ou seja, $eff(\hat{P}_{_{LGREG}}) < eff(\hat{P}_{_{GREG}})$ e $eff(\hat{P}_{_{LGREG}}) < eff(\hat{P}_{_{HT}})$, e é sempre menor que 1. Além disso, o estimador $\hat{P}_{_{LGREG}}$ é mais eficiente quando aumenta o grau de associação, a proporção de indivíduos com o atributo de interesse P e o tamanho da amostra considerado, tendo maior influência nesta mudança o grau de associação.

- A eficiência do estimador do ponto do vista do EQM: nesta medida podem ser observados resultados muito similares (veja Quadros 6.4, 6.10 e 6.16) aos apresentados pela eficiência do estimador ($eff(\hat{P}_{_{LGREG}})$), pois o estimador $\hat{P}_{_{LGREG}}$ é apresentado como assintoticamente não-viesado, o que faz com que $eff(\hat{P}_{_{LGREG}}) \approx \Lambda(\hat{P}_{_{LGREG}})$.

- O viés relativo do estimador da variância do estimador: os estimadores da variância do estimador $\hat{P}_{_{LGREG}}$ (veja Quadros 6.5, 6.11 e 6.17) são apresentados como não-viesados, sendo em quase todos os cenários menor do 5%.

- O coeficiente de variação do estimador: esta medida representa a estabilidade do estimador quando é feita a estimação intervalar do parâmetro P, portanto, quando este coeficiente é pequeno pode-se esperar intervalos de confiança cujas amplitudes variam menos de amostra para amostra, sendo assim mais homogêneos. No caso do estimador

$\hat{P}_{LGREG}$ (veja os Quadros 6.6, 6.12 e 6.18) o coeficiente de variação vai diminuindo a medida que o tamanho da amostra aumenta, e observe também que este valor e menor quando o grau de associação entre as variáveis é mais alto.

- A taxas de cobertura do estimador: para um intervalo de confiança de 95% todos os cenários apresentam taxas maiores que 0.90. Para o estimador $\hat{P}_{LGREG}$ (veja os Quadros 6.7, 6.13 e 6.19), usando os dois estimadores da variância do estimador, pode-se notar a estimação intervalar de P, usando o estimador $\hat{V}_2(\hat{P}_{LGREG})$ é sempre mais estável e homogênea que usando $\hat{V}_1(\hat{P}_{LGREG})$, pois apresenta coeficientes de variação são menores.

6.2.1 Resultados para Amostragem Aleatória Simples

Nos Quadros 6.2 a 6.7 são apresentadas as medidas que permitem fazer a comparação e a avaliação dos estimadores $\hat{P}_{HT}$, $\hat{P}_{GREG}$ e $\hat{P}_{LGREG}$ quando é usado um plano AAS.

No Quadro 6.2 encontram-se os viéses relativos dos estimadores, obtidos nos cenários de simulação. Estes valores mostram os estimadores como não-viesados, ou com um viés relativo desprezível, pois sem importar o tamanho da amostra (n), a proporção de indivíduos na população com a característica de interesse (P), e a força de associação entre as variáveis de interesse e a auxiliar (OR), o valor apresenta-se menor do que 1%.

No Quadro 6.3 é apresentada a eficiência relativa dos estimadores $\hat{P}_{GREG}$ e $\hat{P}_{LGREG}$ com respeito ao estimador $\hat{P}_{HT}$. Os estimadores $\hat{P}_{GREG}$ e $\hat{P}_{LGREG}$ são, em todos os cenários, mais eficientes que o estimador de Horvitz-Thompson, pois todos os valores são menores que 1. Quanto menor é o valor de $eff(\hat{P})$ maior eficiência apresentada pelo estimador. O valor de $eff(\hat{P})$ diminue, na medida que o tamanho da amostra, a proporção de indivíduos com a característica de interesse na população, e o grau de associação aumentam. Outro comportamento que pode ser observado, é que quanto maior é a associação entre as variáveis, sem importar a direção, a eficiência dos estimadores aumenta. Embora, os valores de $eff(\hat{P}_{GREG})$ e $eff(\hat{P}_{LGREG})$ apresentem compor-

tamentos similares, o $eff(\hat{P}_{_{LGREG}})$ é menor em todos os casos, o que leva a concluir que o estimador $\hat{P}_{_{LGREG}}$ é o mais eficiente.

No Quadro 6.4 pode ser observada a eficiência dos estimadores do ponto de vista do erro quadrático médio. A $\Lambda(\hat{P})$ apresenta um comportamento similar a $eff(\hat{P})$. Este comportamento pode ser explicado devido a que o viés relativo do estimador é aproximadamente zero, o que faz com que o erro quadrático médio dos estimadores seja muito próximo da estimativa da variância do estimador, a qual é usada para calcular a eficiência, obtendo desta maneira $eff(\hat{P}) \approx \Lambda(\hat{P})$.

Com respeito a estimação intervalar dos estimadores são apresentadas várias medidas como o viés relativo do estimador da variância, o coeficiente de variação e a taxa de cobertura para um intervalo de 95% de confiança do estimador, apresentadas nos Quadros 6.5 a 6.7. Nestes Quadros são considerados quatro estimadores das variâncias dos estimadores, pois para o estimador $\hat{P}_{_{LGREG}}$ tem-se dois estimadores da variância do estimador. No Quadro 6.5 apresenta-se o viés relativo dos estimadores das variâncias do estimador. Em todos os cenários de simulação esta medida é menor do que $5\%(0.05)$, o que leva a concluir que estes podem ser considerados como estimadores não-viesados.

No Quadro 6.6 pode-se observar os coeficientes de variação dos estimadores. Esta medida é considerada para comparar a estabilidade dos estimadores. Os coeficientes de variação dos estimadores são muito parecidos, mas o estimador $\hat{P}_{_{LGREG2}}$ é mais estável ao ser comparado com o estimador $\hat{P}_{_{LGREG1}}$, pois seus coeficientes de variação são menores do que os de $\hat{P}_{_{LGREG2}}$. O coeficiente de variação começa a diminuir quando o o tamanho da amostra considerado aumenta, mostrando assim mais estáveis os estimadores para tamanhos de amostra grandes.

No Quadro 6.7 apresentam-se as taxas de cobertura para um intervalo de confiança de 95% do estimador. Estas taxas são, em todos os casos, maiores que 0.92. Além disso, as taxas apresentam um comportamento semelhante aos coeficientes de variação, deixando assim, ver a relação entre estas medidas quando é de interesse obter estimações por intervalo.

Quadro 6.2. Viés relativo do estimador de P usando um plano AAS.

n	P	Estimador	Razão de chances (OR)							
			0.1	0.2	0.5	0.65	1.5	2	5	10
50	0.2	$\hat{P}_{HT}$	0.0008	0.0007	0.0011	0.0021	0.0001	0.0007	0.0005	0.0002
		$\hat{P}_{GREG}$	0.0058	0.0074	0.0013	0.0017	0.0017	0.0029	0.0081	0.0076
		$\hat{P}_{LGREG}$	0.0040	0.0005	0.0027	0.0049	0.0013	0.0015	0.0004	0.0002
	0.3	$\hat{P}_{HT}$	0.0006	0.0013	0.0012	0.0008	0.0010	0.0015	0.0006	0.0003
		$\hat{P}_{GREG}$	0.0031	0.0018	0.0001	0.0007	0.0016	0.0033	0.0048	0.0050
		$\hat{P}_{LGREG}$	0.0004	0.0018	0.0020	0.0021	0.0003	0.0012	0.0007	0.0006
	0.5	$\hat{P}_{HT}$	0.0017	0.0008	0.0009	0.0020	0.0002	0.0003	0.0004	0.0007
		$\hat{P}_{GREG}$	0.0009	0.0001	0.0007	0.0015	0.0002	0.0001	0.0001	0.0001
		$\hat{P}_{LGREG}$	0.0011	0.0003	0.0008	0.0016	0.0003	0.0001	0.0004	0.0010
100	0.2	$\hat{P}_{HT}$	0.0025	0.0009	0.0016	0.0013	0.0002	0.0002	0.0010	0.0012
		$\hat{P}_{GREG}$	0.0052	0.0036	0.0025	0.0014	0.0007	0.0024	0.0032	0.0032
		$\hat{P}_{LGREG}$	0.0021	0.0007	0.0008	0.0002	0.0003	0.0006	0.0005	0.0002
	0.3	$\hat{P}_{HT}$	0.0023	0.0022	0.0017	0.0005	0.0005	0.0006	0.0004	0.0002
		$\hat{P}_{GREG}$	0.0034	0.0034	0.0019	0.0005	0.0011	0.0020	0.0021	0.0025
		$\hat{P}_{LGREG}$	0.0017	0.0018	0.0011	0.0007	0.0007	0.0011	0.0002	0.0005
	0.5	$\hat{P}_{HT}$	0.0009	0.0010	0.0008	0.0012	0.0002	0.0003	0.0009	0.0005
		$\hat{P}_{GREG}$	0.0001	0.0003	0.0012	0.0017	0.0006	0.0003	0.0015	0.0013
		$\hat{P}_{LGREG}$	0.0001	0.0004	0.0012	0.0016	0.0004	0.0003	0.0015	0.0012
250	0.2	$\hat{P}_{HT}$	0.0001	0.0002	0.0012	0.0018	0.0016	0.0021	0.0012	0.0011
		$\hat{P}_{GREG}$	0.0016	0.0012	0.0017	0.0019	0.0018	0.0027	0.0023	0.0023
		$\hat{P}_{LGREG}$	0.0002	0.0009	0.0011	0.0016	0.0014	0.0020	0.0008	0.0007
	0.3	$\hat{P}_{HT}$	0.0001	0.0002	0.0001	0.0007	0.0012	0.0011	0.0013	0.0009
		$\hat{P}_{GREG}$	0.0009	0.0010	0.0005	0.0001	0.0013	0.0014	0.0018	0.0016
		$\hat{P}_{LGREG}$	0.0003	0.0003	0.0001	0.0001	0.0011	0.0010	0.0010	0.0006
	0.5	$\hat{P}_{HT}$	0.0003	0.0007	0.0006	0.0004	0.0006	0.0003	0.0007	0.0005
		$\hat{P}_{GREG}$	0.0002	0.0008	0.0002	0.0001	0.0006	0.0003	0.0001	0.0004
		$\hat{P}_{LGREG}$	0.0001	0.0001	0.0005	0.0003	0.0003	0.0003	0.0002	0.0003
500	0.2	$\hat{P}_{HT}$	0.0004	0.0008	0.0009	0.0016	0.0010	0.0009	0.0016	0.0011
		$\hat{P}_{GREG}$	0.0002	0.0005	0.0009	0.0017	0.0008	0.0003	0.0006	0.0001
		$\hat{P}_{LGREG}$	0.0009	0.0012	0.0013	0.0019	0.0010	0.0007	0.0013	0.0008
	0.3	$\hat{P}_{HT}$	0.0006	0.0010	0.0009	0.0004	0.0003	0.0005	0.0009	0.0014
		$\hat{P}_{GREG}$	0.0006	0.0010	0.0009	0.0005	0.0001	0.0002	0.0002	0.0006
		$\hat{P}_{LGREG}$	0.0009	0.0013	0.0011	0.0006	0.0002	0.0004	0.0006	0.0011
	0.5	$\hat{P}_{HT}$	0.0008	0.0006	0.0004	0.0006	0.0003	0.0005	0.0006	0.0007
		$\hat{P}_{GREG}$	0.0011	0.0009	0.0006	0.0007	0.0002	0.0003	0.0004	0.0005
		$\hat{P}_{LGREG}$	0.0011	0.0008	0.0006	0.0007	0.0002	0.0004	0.0004	0.0005
1000	0.2	$\hat{P}_{HT}$	0.0001	0.0002	0.0006	0.0002	0.0005	0.0011	0.0011	0.0009
		$\hat{P}_{GREG}$	0.0007	0.0008	0.0001	0.0002	0.0004	0.0008	0.0005	0.0003
		$\hat{P}_{LGREG}$	0.0002	0.0001	0.0001	0.0003	0.0004	0.0009	0.0008	0.0006
	0.3	$\hat{P}_{HT}$	0.0005	0.0003	0.0003	0.0002	0.0004	0.0005	0.0008	0.0007
		$\hat{P}_{GREG}$	0.0001	0.0008	0.0003	0.0001	0.0003	0.0002	0.0004	0.0003
		$\hat{P}_{LGREG}$	0.0005	0.0001	0.0002	0.0001	0.0003	0.0003	0.0006	0.0005
	0.5	$\hat{P}_{HT}$	0.0008	0.0001	0.0002	0.0002	0.0001	0.0002	0.0005	0.0009
		$\hat{P}_{GREG}$	0.0001	0.0004	0.0001	0.0002	0.0001	0.0001	0.0002	0.0001
		$\hat{P}_{LGREG}$	0.0003	0.0009	0.0001	0.0002	0.0001	0.0001	0.0002	0.0005

Quadro 6.3. Eficiência relativa do estimador de P usando um plano AAS.

n	P	Estimador	Razão de chances (OR)							
			0.1	0.2	0.5	0.65	1.5	2	5	10
50	0.2	$\hat{P}_{GREG}$	0.6469	0.7566	0.9474	0.9914	0.9900	0.9450	0.7712	0.6612
		$\hat{P}_{LGREG}$	0.4826	0.7140	0.9456	0.9901	0.9887	0.9415	0.7381	0.6030
	0.3	$\hat{P}_{GREG}$	0.6297	0.7289	0.9353	0.9789	0.9904	0.9436	0.7454	0.6405
		$\hat{P}_{LGREG}$	0.5738	0.6986	0.9275	0.9743	0.9881	0.9377	0.7197	0.5935
	0.5	$\hat{P}_{GREG}$	0.6168	0.7239	0.9207	0.9810	0.9917	0.9341	0.7294	0.7136
		$\hat{P}_{LGREG}$	0.5709	0.6960	0.9109	0.9745	0.9876	0.9282	0.7083	0.6233
100	0.2	$\hat{P}_{GREG}$	0.6325	0.7393	0.9424	0.9823	0.9912	0.9407	0.7535	0.6443
		$\hat{P}_{LGREG}$	0.5650	0.6970	0.9387	0.9816	0.9911	0.9369	0.7243	0.5883
	0.3	$\hat{P}_{GREG}$	0.6054	0.7146	0.9209	0.9678	0.9827	0.9331	0.7358	0.6278
		$\hat{P}_{LGREG}$	0.5544	0.6863	0.9152	0.9653	0.9824	0.9317	0.7158	0.5847
	0.5	$\hat{P}_{GREG}$	0.5941	0.6943	0.9002	0.9669	0.9794	0.9247	0.7283	0.6148
		$\hat{P}_{LGREG}$	0.5550	0.6723	0.8965	0.9644	0.9782	0.9233	0.7144	0.5750
250	0.2	$\hat{P}_{GREG}$	0.6349	0.7435	0.9324	0.9772	0.9762	0.9281	0.7515	0.6406
		$\hat{P}_{LGREG}$	0.5748	0.7088	0.9293	0.9767	0.9758	0.9232	0.7182	0.5816
	0.3	$\hat{P}_{GREG}$	0.6135	0.7185	0.9208	0.9669	0.9744	0.9221	0.7276	0.6237
		$\hat{P}_{LGREG}$	0.5656	0.6946	0.9180	0.9662	0.9736	0.9185	0.7068	0.5784
	0.5	$\hat{P}_{GREG}$	0.5977	0.7079	0.9096	0.9678	0.9731	0.9190	0.7185	0.6086
		$\hat{P}_{LGREG}$	0.5606	0.6916	0.9078	0.9669	0.9726	0.9178	0.7037	0.5723
500	0.2	$\hat{P}_{GREG}$	0.6319	0.7415	0.9316	0.9769	0.9738	0.9273	0.7486	0.6383
		$\hat{P}_{LGREG}$	0.5672	0.7007	0.9273	0.9763	0.9733	0.9221	0.7185	0.5865
	0.3	$\hat{P}_{GREG}$	0.6019	0.7080	0.9190	0.9684	0.9668	0.9099	0.7209	0.6143
		$\hat{P}_{LGREG}$	0.5515	0.6845	0.9158	0.9677	0.9665	0.9082	0.7013	0.5706
	0.5	$\hat{P}_{GREG}$	0.5916	0.6987	0.9032	0.9635	0.9682	0.9112	0.7086	0.6045
		$\hat{P}_{LGREG}$	0.5583	0.6867	0.9030	0.9634	0.9682	0.9110	0.6965	0.5688
1000	0.2	$\hat{P}_{GREG}$	0.6190	0.7261	0.9228	0.9679	0.9821	0.9337	0.7578	0.6472
		$\hat{P}_{LGREG}$	0.5566	0.6943	0.9204	0.9682	0.9816	0.9296	0.7290	0.5911
	0.3	$\hat{P}_{GREG}$	0.5983	0.7067	0.9083	0.9593	0.9787	0.9254	0.7370	0.6193
		$\hat{P}_{LGREG}$	0.5498	0.6855	0.9077	0.9594	0.9784	0.9238	0.7153	0.5733
	0.5	$\hat{P}_{GREG}$	0.5835	0.6906	0.8936	0.9588	0.9711	0.9131	0.7134	0.6069
		$\hat{P}_{LGREG}$	0.5498	0.6761	0.8931	0.9585	0.9711	0.9126	0.7000	0.5675

Quadro 6.4. Eficiência do ponto de vista do EQM do estimador de P usando um plano AAS.

n	P	Estimador	Razão de chances (OR)							
			0.1	0.2	0.5	0.65	1.5	2	5	10
50	0.2	$\hat{P}_{GREG}$	0.6476	0.7577	0.9228	0.9914	0.9901	0.9451	0.7724	0.6626
		$\hat{P}_{LGREG}$	0.5932	0.7140	0.9204	0.9904	0.9887	0.9415	0.7381	0.6030
	0.3	$\hat{P}_{GREG}$	0.6300	0.7289	0.9353	0.9789	0.9904	0.9438	0.7461	0.6413
		$\hat{P}_{LGREG}$	0.5739	0.6987	0.9275	0.9744	0.9881	0.9377	0.7197	0.5935
	0.5	$\hat{P}_{GREG}$	0.6168	0.7239	0.9207	0.9809	0.9917	0.9341	0.7294	0.6233
		$\hat{P}_{LGREG}$	0.5709	0.6960	0.9109	0.9744	0.9876	0.9283	0.7083	0.5936
100	0.2	$\hat{P}_{GREG}$	0.6336	0.7398	0.9316	0.9823	0.9912	0.9409	0.7539	0.6447
		$\hat{P}_{LGREG}$	0.5650	0.6970	0.9274	0.9815	0.9911	0.9369	0.7242	0.5883
	0.3	$\hat{P}_{GREG}$	0.6059	0.7150	0.9210	0.9678	0.9828	0.9333	0.7360	0.6282
		$\hat{P}_{LGREG}$	0.5544	0.6863	0.9151	0.9653	0.9824	0.9318	0.7158	0.5848
	0.5	$\hat{P}_{GREG}$	0.5940	0.6942	0.9003	0.9670	0.9794	0.9247	0.7285	0.6149
		$\hat{P}_{LGREG}$	0.5549	0.6722	0.8966	0.9645	0.9782	0.9233	0.7145	0.5751
250	0.2	$\hat{P}_{GREG}$	0.6352	0.7437	0.9325	0.9772	0.9763	0.9284	0.7519	0.6411
		$\hat{P}_{LGREG}$	0.5748	0.7088	0.9293	0.9767	0.9758	0.9231	0.7181	0.5815
	0.3	$\hat{P}_{GREG}$	0.6137	0.7186	0.9208	0.9670	0.9744	0.9222	0.7279	0.6240
		$\hat{P}_{LGREG}$	0.5656	0.6946	0.9180	0.9662	0.9736	0.9185	0.7068	0.5784
	0.5	$\hat{P}_{GREG}$	0.5977	0.7079	0.9096	0.9678	0.9731	0.9190	0.7185	0.6086
		$\hat{P}_{LGREG}$	0.5606	0.6916	0.9078	0.9669	0.9726	0.9178	0.7037	0.5723
500	0.2	$\hat{P}_{GREG}$	0.6319	0.7415	0.9426	0.9769	0.9737	0.9272	0.7483	0.6380
		$\hat{P}_{LGREG}$	0.5673	0.7009	0.9386	0.9765	0.9732	0.9221	0.7185	0.5864
	0.3	$\hat{P}_{GREG}$	0.6019	0.7081	0.9190	0.9685	0.9667	0.9099	0.7207	0.6141
		$\hat{P}_{LGREG}$	0.5517	0.6849	0.9159	0.9678	0.9665	0.9082	0.7013	0.5706
	0.5	$\hat{P}_{GREG}$	0.5922	0.6990	0.9033	0.9636	0.9682	0.9111	0.7085	0.6044
		$\hat{P}_{LGREG}$	0.5588	0.6870	0.9031	0.9635	0.9682	0.9110	0.6964	0.5687
1000	0.2	$\hat{P}_{GREG}$	0.6190	0.7261	0.9474	0.9679	0.9820	0.9335	0.7575	0.6470
		$\hat{P}_{LGREG}$	0.5567	0.6943	0.9457	0.9682	0.9815	0.9295	0.7288	0.5911
	0.3	$\hat{P}_{GREG}$	0.5983	0.7067	0.9083	0.9593	0.9787	0.9254	0.7368	0.6191
		$\hat{P}_{LGREG}$	0.5498	0.6856	0.9076	0.9594	0.9784	0.9237	0.7152	0.5732
	0.5	$\hat{P}_{GREG}$	0.5835	0.6906	0.8936	0.9588	0.9710	0.9130	0.7135	0.6069
		$\hat{P}_{LGREG}$	0.5498	0.6761	0.8931	0.9585	0.9711	0.9125	0.7000	0.5675

Quadro 6.5. Viés relativo do estimador da variância do estimador de P usando um plano AAS.

n	P	Estimador	Razão de chances (OR)							
			0.1	0.2	0.5	0.65	1.5	2	5	10
50	0.2	$\hat{P}_{HT}$	0.0011	0.0078	0.0101	0.0006	0.0015	0.0119	0.0036	0.0093
		$\hat{P}_{GREG}$	0.0753	0.0729	0.0654	0.0585	0.0568	0.0691	0.0680	0.0626
		$\hat{P}_{LGREG1}$	0.0056	0.0202	0.0389	0.0302	0.0315	0.0437	0.0228	0.0137
		$\hat{P}_{LGREG2}$	0.0230	0.0532	0.0511	0.0402	0.0391	0.0552	0.0529	0.0541
	0.3	$\hat{P}_{HT}$	0.0085	0.0014	0.0017	0.0011	0.0082	0.0135	0.0033	0.0065
		$\hat{P}_{GREG}$	0.0718	0.0559	0.0549	0.0547	0.0673	0.0770	0.0601	0.0656
		$\hat{P}_{LGREG1}$	0.0203	0.0058	0.0226	0.0258	0.0430	0.0483	0.0165	0.0236
		$\hat{P}_{LGREG2}$	0.0569	0.0333	0.0319	0.0323	0.0476	0.0564	0.0409	0.0569
	0.5	$\hat{P}_{HT}$	0.0211	0.0219	0.0118	0.0066	0.0005	0.0069	0.0244	0.0269
		$\hat{P}_{GREG}$	0.0503	0.0166	0.0473	0.0533	0.0640	0.0560	0.0376	0.0474
		$\hat{P}_{LGREG1}$	0.0030	0.0167	0.0130	0.0248	0.0394	0.0271	0.0043	0.0324
		$\hat{P}_{LGREG2}$	0.0302	0.0179	0.0196	0.0284	0.0414	0.0322	0.0130	0.0215
100	0.2	$\hat{P}_{HT}$	0.0152	0.0144	0.0127	0.0236	0.0155	0.0191	0.0087	0.0006
		$\hat{P}_{GREG}$	0.0488	0.0406	0.0200	0.0067	0.0231	0.0166	0.0346	0.0307
		$\hat{P}_{LGREG1}$	0.0158	0.0083	0.0055	0.0065	0.0131	0.0058	0.0150	0.0068
		$\hat{P}_{LGREG2}$	0.0389	0.0256	0.0115	0.0024	0.0153	0.0100	0.0288	0.0261
	0.3	$\hat{P}_{HT}$	0.0108	0.0210	0.0107	0.0066	0.0064	0.0012	0.0098	0.0100
		$\hat{P}_{GREG}$	0.0355	0.0425	0.0339	0.0297	0.0396	0.0380	0.0275	0.0279
		$\hat{P}_{LGREG1}$	0.0086	0.0144	0.0166	0.0155	0.0295	0.0271	0.0098	0.0115
		$\hat{P}_{LGREG2}$	0.0281	0.0285	0.0216	0.0184	0.0306	0.0301	0.0212	0.0276
	0.5	$\hat{P}_{HT}$	0.0208	0.0015	0.0079	0.0216	0.0064	0.0026	0.0170	0.0062
		$\hat{P}_{GREG}$	0.0370	0.0095	0.0259	0.0461	0.0396	0.0323	0.0289	0.0388
		$\hat{P}_{LGREG1}$	0.0147	0.0118	0.0097	0.0325	0.0282	0.0198	0.0129	0.0196
		$\hat{P}_{LGREG2}$	0.0292	0.0141	0.0136	0.0345	0.0289	0.0218	0.0208	0.0315
250	0.2	$\hat{P}_{HT}$	0.0066	0.0134	0.0166	0.0156	0.0221	0.0321	0.0024	0.0007
		$\hat{P}_{GREG}$	0.0269	0.0326	0.0256	0.0277	0.0323	0.0411	0.0099	0.0132
		$\hat{P}_{LGREG1}$	0.0222	0.0244	0.0205	0.0230	0.0282	0.0357	0.0051	0.0042
		$\hat{P}_{LGREG2}$	0.0303	0.0304	0.0223	0.0241	0.0292	0.0378	0.0015	0.0046
	0.3	$\hat{P}_{HT}$	0.0127	0.0168	0.0199	0.0171	0.0214	0.0130	0.0036	0.0045
		$\hat{P}_{GREG}$	0.0405	0.0333	0.0311	0.0272	0.0338	0.0254	0.0112	0.0259
		$\hat{P}_{LGREG1}$	0.0347	0.0238	0.0255	0.0224	0.0292	0.0183	0.0010	0.0159
		$\hat{P}_{LGREG2}$	0.0410	0.0286	0.0269	0.0231	0.0299	0.0201	0.0065	0.0232
	0.5	$\hat{P}_{HT}$	0.0035	0.0099	0.0176	0.0244	0.0036	0.0109	0.0267	0.0163
		$\hat{P}_{GREG}$	0.0109	0.0335	0.0346	0.0387	0.0191	0.0284	0.0483	0.0414
		$\hat{P}_{LGREG1}$	0.0057	0.0277	0.0292	0.0339	0.0143	0.0225	0.0385	0.0354
		$\hat{P}_{LGREG2}$	0.0104	0.0309	0.0302	0.0344	0.0148	0.0237	0.0423	0.0408

Continuação da Quadro 6.5

n	P	Estimador	Razão de chances (OR)							
			0.1	0.2	0.5	0.65	1.5	2	5	10
500	0.2	$\hat{P}_{HT}$	0.0168	0.0028	0.0243	0.0143	0.0023	0.0008	0.0073	0.0145
		$\hat{P}_{GREG}$	0.0024	0.0159	0.0201	0.0061	0.0015	0.0053	0.0116	0.0210
		$\hat{P}_{LGREG1}$	0.0075	0.0019	0.0249	0.0092	0.0074	0.0024	0.0057	0.0208
		$\hat{P}_{LGREG2}$	0.0024	0.0059	0.0234	0.0082	0.0001	0.0030	0.0081	0.0241
	0.3	$\hat{P}_{HT}$	0.0155	0.0058	0.0050	0.0213	0.0017	0.0038	0.0025	0.0093
		$\hat{P}_{GREG}$	0.0089	0.0067	0.0003	0.0132	0.0007	0.0082	0.0048	0.0124
		$\hat{P}_{LGREG1}$	0.0186	0.0140	0.0038	0.0163	0.0024	0.0102	0.0006	0.0084
		$\hat{P}_{LGREG2}$	0.0143	0.0108	0.0026	0.0156	0.0023	0.0098	0.0025	0.0111
	0.5	$\hat{P}_{HT}$	0.0031	0.0252	0.0100	0.0086	0.0022	0.0071	0.0122	0.0086
		$\hat{P}_{GREG}$	0.0017	0.0389	0.0169	0.0149	0.0090	0.0012	0.0065	0.0073
		$\hat{P}_{LGREG1}$	0.0011	0.0129	0.0150	0.0129	0.0071	0.0033	0.0095	0.0058
		$\hat{P}_{LGREG2}$	0.0041	0.0236	0.0159	0.0134	0.0072	0.0030	0.0082	0.0078
1000	0.2	$\hat{P}_{HT}$	0.0327	0.0138	0.0095	0.0076	0.0253	0.0149	0.0200	0.0158
		$\hat{P}_{GREG}$	0.0295	0.0039	0.0023	0.0044	0.0355	0.0245	0.0350	0.0346
		$\hat{P}_{LGREG1}$	0.0248	0.0044	0.0015	0.0038	0.0351	0.0243	0.0340	0.0316
		$\hat{P}_{LGREG2}$	0.0271	0.0063	0.0022	0.0042	0.0350	0.0243	0.0348	0.0328
	0.3	$\hat{P}_{HT}$	0.0250	0.0221	0.0138	0.0192	0.0049	0.0235	0.0137	0.0065
		$\hat{P}_{GREG}$	0.0238	0.0176	0.0057	0.0157	0.0165	0.0343	0.0363	0.0032
		$\hat{P}_{LGREG1}$	0.0202	0.0162	0.0058	0.0149	0.0159	0.0337	0.0321	0.0017
		$\hat{P}_{LGREG2}$	0.0221	0.0177	0.0063	0.0152	0.0158	0.0338	0.0327	0.0007
	0.5	$\hat{P}_{HT}$	0.0329	0.0200	0.0122	0.0138	0.0044	0.0073	0.0072	0.0003
		$\hat{P}_{GREG}$	0.0189	0.0518	0.0063	0.0131	0.0036	0.0008	0.0038	0.0183
		$\hat{P}_{LGREG1}$	0.0220	0.0037	0.0054	0.0120	0.0029	0.0021	0.0006	0.0122
		$\hat{P}_{LGREG2}$	0.0234	0.0234	0.0058	0.0123	0.0028	0.0021	0.0011	0.0131

Quadro 6.6. Coeficiente de variação do estimador de P usando um plano AAS.

n	P	Estimador	Razão de chances (OR)							
			0.1	0.2	0.5	0.65	1.5	2	5	10
50	0.2	$\hat{P}_{HT}$	11.341	14.381	18.778	19.843	20.516	19.310	14.257	11.381
		$\hat{P}_{GREG}$	16.525	17.104	19.203	20.062	20.741	19.731	16.715	16.035
		$\hat{P}_{LGREG1}$	32.253	28.398	22.185	21.782	22.557	22.928	27.373	32.382
		$\hat{P}_{LGREG2}$	23.324	21.904	19.936	20.314	21.042	20.617	21.024	23.735
	0.3	$\hat{P}_{HT}$	7.344	8.708	11.501	12.088	11.858	11.417	8.803	7.340
		$\hat{P}_{GREG}$	15.742	15.077	13.417	13.076	12.857	13.430	14.906	15.545
		$\hat{P}_{LGREG1}$	31.470	25.515	16.672	15.019	14.847	16.839	24.905	30.664
		$\hat{P}_{LGREG2}$	23.086	18.787	13.996	13.287	13.095	14.051	18.517	22.432
	0.5	$\hat{P}_{HT}$	2.785	2.813	2.782	2.804	2.825	2.852	2.808	3.734
		$\hat{P}_{GREG}$	15.402	14.368	9.317	6.645	6.463	8.972	14.114	13.349
		$\hat{P}_{LGREG1}$	28.571	23.363	13.269	9.276	9.004	12.695	22.860	23.241
		$\hat{P}_{LGREG2}$	21.486	17.393	9.707	6.781	6.577	9.336	16.915	16.112
100	0.2	$\hat{P}_{HT}$	8.284	10.082	13.025	13.800	14.330	13.376	9.966	7.935
		$\hat{P}_{GREG}$	11.473	11.995	13.324	13.935	14.363	13.512	11.574	11.170
		$\hat{P}_{LGREG1}$	23.520	20.064	15.275	14.846	15.538	15.816	19.378	22.674
		$\hat{P}_{LGREG2}$	17.178	15.228	13.741	14.053	14.522	14.059	14.407	16.269
	0.3	$\hat{P}_{HT}$	5.058	6.1175	8.065	8.401	8.198	7.850	5.983	4.955
		$\hat{P}_{GREG}$	11.003	10.634	9.434	9.062	8.770	9.206	10.306	10.767
		$\hat{P}_{LGREG1}$	21.869	18.038	11.509	10.112	10.044	11.711	17.648	21.263
		$\hat{P}_{LGREG2}$	15.925	13.150	9.788	9.166	8.877	9.574	12.617	15.196
	0.5	$\hat{P}_{HT}$	1.423	1.422	1.416	1.425	1.424	1.423	1.406	1.392
		$\hat{P}_{GREG}$	10.792	9.923	6.351	4.293	4.087	5.987	9.862	10.669
		$\hat{P}_{LGREG1}$	19.823	16.168	9.051	5.985	5.780	8.677	16.144	20.025
		$\hat{P}_{LGREG2}$	14.773	11.843	6.561	4.354	4.121	6.130	11.620	14.754
250	0.2	$\hat{P}_{HT}$	5.095	6.282	8.258	8.804	9.131	8.597	6.198	4.948
		$\hat{P}_{GREG}$	7.116	7.407	8.335	8.821	9.163	8.686	7.221	6.944
		$\hat{P}_{LGREG1}$	14.412	12.420	9.317	9.206	9.527	9.756	11.901	13.877
		$\hat{P}_{LGREG2}$	10.533	9.315	8.529	8.854	9.219	8.963	8.900	9.971
	0.3	$\hat{P}_{HT}$	3.082	3.737	5.010	5.241	5.159	4.919	3.741	3.054
		$\hat{P}_{GREG}$	6.810	6.568	5.796	5.571	5.483	5.751	6.466	6.686
		$\hat{P}_{LGREG1}$	13.351	11.064	7.100	6.163	6.041	7.074	10.843	13.110
		$\hat{P}_{LGREG2}$	9.728	8.029	5.974	5.610	5.523	5.927	7.817	9.366
	0.5	$\hat{P}_{HT}$	0.553	0.565	0.571	0.562	0.557	0.563	0.565	0.563
		$\hat{P}_{GREG}$	6.705	6.174	3.900	2.551	2.401	3.696	6.111	6.735
		$\hat{P}_{LGREG1}$	12.269	10.031	5.625	3.615	3.367	5.296	9.898	12.351
		$\hat{P}_{LGREG2}$	9.126	7.294	4.004	2.579	2.405	3.748	7.146	9.180

Continuação da Quadro 6.6

n	P	Estimador	Razão de chances (OR)							
			0.1	0.2	0.5	0.65	1.5	2	5	10
500	0.2	$\hat{P}_{HT}$	3.506	4.343	5.629	6.030	6.279	5.879	4.326	3.459
		$\hat{P}_{GREG}$	4.907	5.107	5.727	6.065	6.253	5.877	5.028	4.852
		$\hat{P}_{LGREG1}$	10.161	8.707	6.408	6.311	6.492	6.670	8.304	9.665
		$\hat{P}_{LGREG2}$	7.315	6.418	5.855	6.086	6.283	6.056	6.189	6.957
	0.3	$\hat{P}_{HT}$	2.102	2.558	3.431	3.575	3.545	3.388	2.586	2.108
		$\hat{P}_{GREG}$	4.750	4.610	3.994	3.833	3.741	3.930	4.485	4.754
		$\hat{P}_{LGREG1}$	9.427	7.839	4.914	4.247	4.098	4.860	7.511	9.078
		$\hat{P}_{LGREG2}$	6.776	5.641	4.106	3.855	3.766	4.045	5.413	6.597
	0.5	$\hat{P}_{HT}$	0.266	0.265	0.271	0.273	0.269	0.267	0.273	0.277
		$\hat{P}_{GREG}$	4.718	4.364	2.717	1.749	1.641	2.569	4.274	4.658
		$\hat{P}_{LGREG1}$	8.588	7.083	3.914	2.459	2.312	3.686	6.858	8.497
		$\hat{P}_{LGREG2}$	6.408	5.139	2.782	1.765	1.640	2.604	4.961	6.341
1000	0.2	$\hat{P}_{HT}$	2.475	3.012	3.943	4.202	4.382	4.069	2.995	2.380
		$\hat{P}_{GREG}$	3.410	3.551	3.992	4.203	4.358	4.068	3.440	3.332
		$\hat{P}_{LGREG1}$	6.977	5.986	4.445	4.330	4.560	4.665	5.809	6.665
		$\hat{P}_{LGREG2}$	5.049	4.454	4.082	4.214	4.375	4.186	4.238	4.763
	0.3	$\hat{P}_{HT}$	1.478	1.791	2.390	2.517	2.442	2.360	1.788	1.437
		$\hat{P}_{GREG}$	3.262	3.143	2.781	2.692	2.557	2.712	3.088	3.232
		$\hat{P}_{LGREG1}$	6.528	5.412	3.398	2.933	2.850	3.401	5.227	6.273
		$\hat{P}_{LGREG2}$	4.657	3.839	2.863	2.707	2.573	2.791	3.720	4.489
	0.5	$\hat{P}_{HT}$	0.131	0.127	0.129	0.128	0.124	0.126	0.128	0.129
		$\hat{P}_{GREG}$	3.188	3.000	1.853	1.178	1.119	1.751	2.903	3.220
		$\hat{P}_{LGREG1}$	5.937	4.931	2.686	1.676	1.605	2.566	4.719	5.912
		$\hat{P}_{LGREG2}$	4.312	3.528	1.894	1.188	1.118	1.769	3.361	4.362

Quadro 6.7. Taxas de cobertura para um intervalo de confiança de 95% do estimador de P usando um plano AAS.

n	P	Estimador	Razão de chances (OR)							
			0.1	0.2	0.5	0.65	1.5	2	5	10
50	0.2	$\hat{P}_{HT}$	0.9236	0.9349	0.9436	0.9213	0.9446	0.9288	0.9370	0.9280
		$\hat{P}_{GREG}$	0.9284	0.9293	0.9281	0.9235	0.9334	0.9281	0.9313	0.9353
		$\hat{P}_{LGREG1}$	0.9365	0.9304	0.9302	0.9301	0.9331	0.9306	0.9326	0.9305
		$\hat{P}_{LGREG2}$	0.9312	0.9270	0.9284	0.9262	0.9364	0.9288	0.9319	0.9292
	0.3	$\hat{P}_{HT}$	0.9599	0.9464	0.9289	0.9523	0.9222	0.9281	0.9439	0.9352
		$\hat{P}_{GREG}$	0.9326	0.9336	0.9332	0.9343	0.9293	0.9305	0.9341	0.9323
		$\hat{P}_{LGREG1}$	0.9294	0.9345	0.9380	0.9380	0.9335	0.9354	0.9324	0.9284
		$\hat{P}_{LGREG2}$	0.9293	0.9364	0.9359	0.9369	0.9331	0.9338	0.9347	0.9292
	0.5	$\hat{P}_{HT}$	0.9392	0.9415	0.9383	0.9382	0.9339	0.9369	0.9385	0.9386
		$\hat{P}_{GREG}$	0.9356	0.9391	0.9375	0.9367	0.9332	0.9375	0.9377	0.9380
		$\hat{P}_{LGREG1}$	0.9334	0.9392	0.9412	0.9393	0.9351	0.9397	0.9408	0.9321
		$\hat{P}_{LGREG2}$	0.9343	0.9398	0.9413	0.9397	0.9358	0.9398	0.9405	0.9354
100	0.2	$\hat{P}_{HT}$	0.9480	0.9545	0.9444	0.9485	0.9531	0.9349	0.9565	0.9452
		$\hat{P}_{GREG}$	0.9392	0.9380	0.9424	0.9451	0.9423	0.9410	0.9403	0.9423
		$\hat{P}_{LGREG1}$	0.9362	0.9377	0.9450	0.9460	0.9431	0.9438	0.9402	0.9397
		$\hat{P}_{LGREG2}$	0.9364	0.9389	0.9441	0.9465	0.9435	0.9424	0.9412	0.9401
	0.3	$\hat{P}_{HT}$	0.9495	0.9367	0.9456	0.9491	0.9490	0.9467	0.9411	0.9496
		$\hat{P}_{GREG}$	0.9444	0.9407	0.9455	0.9460	0.9391	0.9424	0.9460	0.9439
		$\hat{P}_{LGREG1}$	0.9403	0.9420	0.9459	0.9466	0.9419	0.9409	0.9455	0.9407
		$\hat{P}_{LGREG2}$	0.9402	0.9413	0.9460	0.9480	0.9415	0.9422	0.9446	0.9408
	0.5	$\hat{P}_{HT}$	0.9436	0.9422	0.9402	0.9442	0.9430	0.9438	0.9478	0.9472
		$\hat{P}_{GREG}$	0.9423	0.9437	0.9432	0.9434	0.9417	0.9421	0.9445	0.9436
		$\hat{P}_{LGREG1}$	0.9418	0.9436	0.9440	0.9453	0.9428	0.9418	0.9415	0.9395
		$\hat{P}_{LGREG2}$	0.9424	0.9446	0.9444	0.9446	0.9425	0.9426	0.9423	0.9398
250	0.2	$\hat{P}_{HT}$	0.9454	0.9449	0.9488	0.9424	0.9490	0.9409	0.9530	0.9536
		$\hat{P}_{GREG}$	0.9429	0.9421	0.9428	0.9453	0.9438	0.9421	0.9464	0.9459
		$\hat{P}_{LGREG1}$	0.9444	0.9436	0.9427	0.9459	0.9431	0.9443	0.9479	0.9475
		$\hat{P}_{LGREG2}$	0.9416	0.9434	0.9412	0.9455	0.9442	0.9435	0.9477	0.9460
	0.3	$\hat{P}_{HT}$	0.9421	0.9462	0.9495	0.9471	0.9453	0.9445	0.9480	0.9470
		$\hat{P}_{GREG}$	0.9446	0.9449	0.9443	0.9466	0.9436	0.9476	0.9478	0.9490
		$\hat{P}_{LGREG1}$	0.9413	0.9441	0.9447	0.9475	0.9431	0.9478	0.9478	0.9460
		$\hat{P}_{LGREG2}$	0.9426	0.9445	0.9446	0.9458	0.9438	0.9476	0.9477	0.9462
	0.5	$\hat{P}_{HT}$	0.9424	0.9539	0.9428	0.9513	0.9522	0.9510	0.9491	0.9512
		$\hat{P}_{GREG}$	0.9443	0.9447	0.9446	0.9445	0.9456	0.9476	0.9442	0.9450
		$\hat{P}_{LGREG1}$	0.9432	0.9455	0.9441	0.9455	0.9467	0.9481	0.9444	0.9429
		$\hat{P}_{LGREG2}$	0.9443	0.9453	0.9452	0.9450	0.9462	0.9482	0.9445	0.9435

Continuação da Quadro 6.7

n	P	Estimador	Razão de chances (OR)							
			0.1	0.2	0.5	0.65	1.5	2	5	10
500	0.2	$\hat{P}_{HT}$	0.9499	0.9464	0.9507	0.9503	0.9502	0.9528	0.9534	0.9443
		$\hat{P}_{GREG}$	0.9486	0.9486	0.9518	0.9496	0.9488	0.9484	0.9482	0.9455
		$\hat{P}_{LGREG1}$	0.9481	0.9476	0.9523	0.9487	0.9492	0.9492	0.9465	0.9438
		$\hat{P}_{LGREG2}$	0.9489	0.9478	0.9519	0.9488	0.9489	0.9495	0.9467	0.9447
	0.3	$\hat{P}_{HT}$	0.9540	0.9511	0.9473	0.9473	0.9519	0.9507	0.9488	0.9450
		$\hat{P}_{GREG}$	0.9483	0.9509	0.9486	0.9511	0.9496	0.9495	0.9468	0.9468
		$\hat{P}_{LGREG1}$	0.9487	0.9501	0.9497	0.9516	0.9502	0.9500	0.9489	0.9467
		$\hat{P}_{LGREG2}$	0.9496	0.9496	0.9485	0.9515	0.9504	0.9498	0.9503	0.9474
	0.5	$\hat{P}_{HT}$	0.9521	0.9537	0.9443	0.9481	0.9513	0.9516	0.9510	0.9463
		$\hat{P}_{GREG}$	0.9492	0.9478	0.9463	0.9476	0.9484	0.9493	0.9489	0.9476
		$\hat{P}_{LGREG1}$	0.9471	0.9466	0.9476	0.9479	0.9484	0.9482	0.9488	0.9476
		$\hat{P}_{LGREG2}$	0.9473	0.9467	0.9473	0.9482	0.9484	0.9488	0.9489	0.9469
1000	0.2	$\hat{P}_{HT}$	0.9486	0.9488	0.9526	0.9555	0.9479	0.9440	0.9497	0.9428
		$\hat{P}_{GREG}$	0.9443	0.9509	0.9502	0.9510	0.9426	0.9474	0.9455	0.9456
		$\hat{P}_{LGREG1}$	0.9470	0.9514	0.9498	0.9510	0.9438	0.9475	0.9432	0.9458
		$\hat{P}_{LGREG2}$	0.9460	0.9502	0.9497	0.9513	0.9439	0.9480	0.9423	0.9463
	0.3	$\hat{P}_{HT}$	0.9460	0.9443	0.9444	0.9463	0.9499	0.9489	0.9452	0.9493
		$\hat{P}_{GREG}$	0.9467	0.9484	0.9518	0.9472	0.9494	0.9440	0.9449	0.9502
		$\hat{P}_{LGREG1}$	0.9466	0.9468	0.9518	0.9481	0.9483	0.9437	0.9458	0.9499
		$\hat{P}_{LGREG2}$	0.9463	0.9470	0.9513	0.9473	0.9482	0.9436	0.9456	0.9504
	0.5	$\hat{P}_{HT}$	0.9434	0.9474	0.9489	0.9502	0.9488	0.9486	0.9465	0.9493
		$\hat{P}_{GREG}$	0.9458	0.9471	0.9477	0.9498	0.9499	0.9489	0.9485	0.9492
		$\hat{P}_{LGREG1}$	0.9480	0.9483	0.9480	0.9497	0.9503	0.9489	0.9488	0.9506
		$\hat{P}_{LGREG2}$	0.9483	0.9484	0.9482	0.9499	0.9510	0.9490	0.9491	0.9501

6.2.2 Resultados para Amostragem de Bernoulli

Nos Quadros 6.8 a 6.13 encontram-se as medidas que permitem comparar e avaliar os estimadores $\hat{P}_{HT}$, $\hat{P}_{GREG}$ e $\hat{P}_{LGREG}$ quando o plano usado para a seleção da amostra é um BE.

Nos Quadros 6.8, 6.11, 6.12 e 6.13 podem ser observados resultados similares aos obtidos quando o plano adotado para a seleção da amostra foi um AAS, descritos na seção anterior. O que evidencia que os estimadores $\hat{P}_{HT}$, $\hat{P}_{GREG}$ e $\hat{P}_{LGREG}$ sobre estes dois planos apresentam propriedades semelhantes. Isto é, as estratégias de amostragem (plano e estimador) têm desempenhos equivalentes quando é de interesse estimar a proporção de indivíduos na população com uma característica particular, levando em conta o grau de associação entre a variável de interesse e as variáveis auxiliares. Em geral, os estimadores e os estimadores das suas variâncias podem ser considerados como não-viesados. Quando é de interesse a estimação intervalar do parâmetro P, as quatro aproximações consideradas (incluindo os dois estimadores da variância do estimador para $\hat{P}_{LGREG}$) são similares, com respeito ao coeficiente de variação e às taxas de cobertura, obtendo assim, uma boa estimação, pois em todos os casos, esta taxa é maior que 0.92.

Embora, os resultados no Quadro 6.9, possuam comportamentos parecidos aos apresentados em 6.3, pode ser observada uma mudança maior neste cenário, quando P aumenta. Por exemplo, quando $P = 0.5$, a eficiência encontra-se no intervalo $(0.45, 0.50)$, e quando $P = 0.3$, no intervalo $(0.65, 0.70)$. Em todos os casos, o mais eficiente é o estimador $\hat{P}_{LGREG}$, ainda que, os resultados para $\hat{P}_{GREG}$ e $\hat{P}_{LGREG}$ apresentam uma diferença pequena.

No Quadro 6.10 pode ser observada a eficiência dos estimadores do ponto de vista do erro quadrático médio. Sendo observado, um comportamento similar entre a $\Lambda(\hat{P})$ e $eff(\hat{P})$, concluindo de novo que $eff(\hat{P}) \approx \Lambda(\hat{P})$.

Em termos gerais, as propriedades dos estimadores sobre o plano AAS e o plano BE são muito parecidas. A diferença encontra-se na eficiência ($eff(\hat{P})$) dos estimadores, pois neste plano é maior, concluindo então que, quando é usado um plano BE, os estimadores $\hat{P}_{GREG}$ e $\hat{P}_{LGREG}$ apresentam-se mais eficientes.

Quadro 6.8. Viés relativo do estimador de P usando um plano BE.

n	P	Estimador	Razão de chances (OR)							
			0.1	0.2	0.5	0.65	1.5	2	5	10
50	0.2	$\hat{P}_{HT}$	0.0048	0.0049	0.0049	0.0044	0.0042	0.0047	0.0008	0.0046
		$\hat{P}_{GREG}$	0.0066	0.0055	0.0044	0.0040	0.0040	0.0048	0.0028	0.0059
		$\hat{P}_{LGREG}$	0.0038	0.0035	0.0029	0.0026	0.0029	0.0036	0.0007	0.0036
	0.3	$\hat{P}_{HT}$	0.0007	0.0003	0.0045	0.0016	0.0005	0.0012	0.0019	0.0017
		$\hat{P}_{GREG}$	0.0011	0.0011	0.0017	0.0034	0.0030	0.0004	0.0026	0.0024
		$\hat{P}_{LGREG}$	0.0003	0.0009	0.0023	0.0029	0.0035	0.0009	0.0017	0.0011
	0.5	$\hat{P}_{HT}$	0.0011	0.0020	0.0013	0.0021	0.0015	0.0013	0.0013	0.0022
		$\hat{P}_{GREG}$	0.0013	0.0021	0.0015	0.0020	0.0014	0.0011	0.0010	0.0016
		$\hat{P}_{LGREG}$	0.0013	0.0021	0.0015	0.0020	0.0014	0.0011	0.0011	0.0016
100	0.2	$\hat{P}_{HT}$	0.0006	0.0006	0.0002	0.0006	0.0011	0.0014	0.0010	0.0015
		$\hat{P}_{GREG}$	0.0002	0.0007	0.0001	0.0011	0.0016	0.0018	0.0015	0.0010
		$\hat{P}_{LGREG}$	0.0013	0.0014	0.0004	0.0013	0.0018	0.0021	0.0022	0.0019
	0.3	$\hat{P}_{HT}$	0.0006	0.0004	0.0003	0.0031	0.0026	0.0002	0.0015	0.0016
		$\hat{P}_{GREG}$	0.0020	0.0006	0.0005	0.0017	0.0007	0.0001	0.0012	0.0011
		$\hat{P}_{LGREG}$	0.0016	0.0010	0.0004	0.0016	0.0006	0.0002	0.0017	0.0017
	0.5	$\hat{P}_{HT}$	0.0001	0.0007	0.0015	0.0017	0.0020	0.0020	0.0016	0.0009
		$\hat{P}_{GREG}$	0.0006	0.0010	0.0017	0.0019	0.0021	0.0020	0.0015	0.0008
		$\hat{P}_{LGREG}$	0.0005	0.0010	0.0017	0.0019	0.0021	0.0020	0.0015	0.0008
250	0.2	$\hat{P}_{HT}$	0.0031	0.0026	0.0021	0.0021	0.0018	0.0023	0.0007	0.0027
		$\hat{P}_{GREG}$	0.0028	0.0022	0.0017	0.0017	0.0011	0.0015	0.0006	0.0012
		$\hat{P}_{LGREG}$	0.0031	0.0024	0.0018	0.0018	0.0012	0.0016	0.0007	0.0016
	0.3	$\hat{P}_{HT}$	0.0002	0.0020	0.0007	0.0005	0.0002	0.0007	0.0014	0.0015
		$\hat{P}_{GREG}$	0.0006	0.0018	0.0005	0.0010	0.0003	0.0003	0.0002	0.0001
		$\hat{P}_{LGREG}$	0.0008	0.0019	0.0006	0.0011	0.0005	0.0003	0.0004	0.0004
	0.5	$\hat{P}_{HT}$	0.0022	0.0001	0.0007	0.0008	0.0014	0.0014	0.0018	0.0020
		$\hat{P}_{GREG}$	0.0001	0.0009	0.0002	0.0004	0.0008	0.0006	0.0010	0.0010
		$\hat{P}_{LGREG}$	0.0001	0.0008	0.0002	0.0004	0.0008	0.0006	0.0010	0.0010
500	0.2	$\hat{P}_{HT}$	0.0006	0.0006	0.0001	0.0004	0.0006	0.0007	0.0011	0.0006
		$\hat{P}_{GREG}$	0.0002	0.0004	0.0005	0.0001	0.0003	0.0005	0.0012	0.0002
		$\hat{P}_{LGREG}$	0.0004	0.0005	0.0003	0.0001	0.0003	0.0004	0.0011	0.0005
	0.3	$\hat{P}_{HT}$	0.0008	0.0001	0.0003	0.0002	0.0008	0.0003	0.0008	0.0010
		$\hat{P}_{GREG}$	0.0005	0.0002	0.0001	0.0001	0.0012	0.0002	0.0005	0.0006
		$\hat{P}_{LGREG}$	0.0006	0.0002	0.0002	0.0001	0.0012	0.0002	0.0004	0.0005
	0.5	$\hat{P}_{HT}$	0.0003	0.0003	0.0003	0.0002	0.0004	0.0004	0.0003	0.0007
		$\hat{P}_{GREG}$	0.0003	0.0003	0.0002	0.0009	0.0001	0.0001	0.0001	0.0003
		$\hat{P}_{LGREG}$	0.0003	0.0003	0.0002	0.0009	0.0001	0.0001	0.0009	0.0003
1000	0.2	$\hat{P}_{HT}$	0.0004	0.0004	0.0002	0.0001	0.0001	0.0002	0.0005	0.0006
		$\hat{P}_{GREG}$	0.0003	0.0004	0.0004	0.0003	0.0004	0.0005	0.0009	0.0005
		$\hat{P}_{LGREG}$	0.0003	0.0005	0.0004	0.0003	0.0005	0.0004	0.0010	0.0004
	0.3	$\hat{P}_{HT}$	0.0011	0.0001	0.0002	0.0002	0.0002	0.0011	0.0007	0.0003
		$\hat{P}_{GREG}$	0.0002	0.0002	0.0003	0.0003	0.0003	0.0007	0.0001	0.0001
		$\hat{P}_{LGREG}$	0.0001	0.0002	0.0003	0.0003	0.0003	0.0007	0.0001	0.0001
	0.5	$\hat{P}_{HT}$	0.0001	0.0003	0.0003	0.0003	0.0001	0.0001	0.0001	0.0001
		$\hat{P}_{GREG}$	0.0001	0.0002	0.0003	0.0002	0.0005	0.0006	0.0001	0.0005
		$\hat{P}_{LGREG}$	0.0001	0.0002	0.0003	0.0002	0.0005	0.0006	0.0002	0.0005

Quadro 6.9. Eficiência relativa do estimador de P usando um plano BE.

n	P	Estimador	Razão de chances (OR)							
			0.1	0.2	0.5	0.65	1.5	2	5	10
50	0.2	$\hat{P}_{GREG}$	0.7851	0.8116	0.8375	0.8404	0.8355	0.8262	0.7931	0.7669
		$\hat{P}_{LGREG}$	0.7854	0.8115	0.8371	0.8399	0.8354	0.8258	0.7924	0.7670
	0.3	$\hat{P}_{GREG}$	0.6797	0.6935	0.7433	0.7213	0.7322	0.7116	0.6891	0.6566
		$\hat{P}_{LGREG}$	0.6776	0.6917	0.7418	0.7209	0.7310	0.7110	0.6884	0.6550
	0.5	$\hat{P}_{GREG}$	0.4614	0.4858	0.5075	0.5143	0.5091	0.5041	0.4823	0.4645
		$\hat{P}_{LGREG}$	0.4593	0.4844	0.5063	0.5132	0.5082	0.5030	0.4808	0.4624
100	0.2	$\hat{P}_{GREG}$	0.7810	0.8030	0.8218	0.8228	0.8236	0.8157	0.7846	0.7710
		$\hat{P}_{LGREG}$	0.7808	0.8032	0.8223	0.8234	0.8240	0.8161	0.7841	0.7714
	0.3	$\hat{P}_{GREG}$	0.6624	0.6967	0.6995	0.7086	0.7059	0.6999	0.6736	0.6504
		$\hat{P}_{LGREG}$	0.6610	0.6967	0.6991	0.7086	0.7057	0.6996	0.6736	0.6509
	0.5	$\hat{P}_{GREG}$	0.4681	0.4898	0.5073	0.5080	0.5029	0.4954	0.4719	0.4525
		$\hat{P}_{LGREG}$	0.4673	0.4892	0.5069	0.5078	0.5027	0.4951	0.4714	0.4515
250	0.2	$\hat{P}_{GREG}$	0.7525	0.7741	0.7997	0.8054	0.8065	0.7975	0.7800	0.7533
		$\hat{P}_{LGREG}$	0.7512	0.7742	0.8000	0.8055	0.8064	0.7975	0.7796	0.7521
	0.3	$\hat{P}_{GREG}$	0.6381	0.6851	0.7099	0.7073	0.6933	0.7018	0.6643	0.6386
		$\hat{P}_{LGREG}$	0.6365	0.6849	0.7100	0.7071	0.6932	0.7020	0.6631	0.6374
	0.5	$\hat{P}_{GREG}$	0.4549	0.4756	0.4937	0.4994	0.4993	0.4931	0.4703	0.4460
		$\hat{P}_{LGREG}$	0.4548	0.4755	0.4936	0.4994	0.4992	0.4931	0.4701	0.4454
500	0.2	$\hat{P}_{GREG}$	0.7643	0.7950	0.8122	0.8164	0.8063	0.8037	0.7781	0.7571
		$\hat{P}_{LGREG}$	0.7634	0.7949	0.8123	0.8165	0.8064	0.8039	0.7785	0.7566
	0.3	$\hat{P}_{GREG}$	0.6495	0.6929	0.7063	0.6985	0.6996	0.6861	0.6734	0.6450
		$\hat{P}_{LGREG}$	0.6480	0.6931	0.7061	0.6986	0.6998	0.6863	0.6731	0.6442
	0.5	$\hat{P}_{GREG}$	0.4584	0.4856	0.5101	0.5154	0.5166	0.5121	0.4902	0.4651
		$\hat{P}_{LGREG}$	0.4582	0.4855	0.5101	0.5154	0.5166	0.5120	0.4901	0.4648
1000	0.2	$\hat{P}_{GREG}$	0.7563	0.7811	0.8016	0.8056	0.8078	0.7989	0.7581	0.7554
		$\hat{P}_{LGREG}$	0.7529	0.7796	0.8014	0.8055	0.8079	0.7987	0.7573	0.7527
	0.3	$\hat{P}_{GREG}$	0.6463	0.6692	0.7132	0.6943	0.6917	0.7032	0.6634	0.6387
		$\hat{P}_{LGREG}$	0.6454	0.6682	0.7131	0.6942	0.6916	0.7034	0.6620	0.6362
	0.5	$\hat{P}_{GREG}$	0.4524	0.4765	0.4914	0.4977	0.4932	0.4885	0.4728	0.4498
		$\hat{P}_{LGREG}$	0.4521	0.4764	0.4914	0.4977	0.4932	0.4885	0.4728	0.4496

Quadro 6.10. Eficiência do ponto de vista do EQM do estimador de P usando um plano BE.

n	P	Estimador	Razão de chances (OR)							
			0.1	0.2	0.5	0.65	1.5	2	5	10
50	0.2	$\hat{P}_{GREG}$	0.7854	0.8117	0.8375	0.8404	0.8356	0.8262	0.7931	0.7670
		$\hat{P}_{LGREG}$	0.7854	0.8115	0.8369	0.8399	0.8354	0.8258	0.7924	0.7670
	0.3	$\hat{P}_{GREG}$	0.6797	0.6935	0.7431	0.7214	0.7324	0.7116	0.6891	0.6567
		$\hat{P}_{LGREG}$	0.6776	0.6917	0.7417	0.7210	0.7312	0.7110	0.6884	0.6550
	0.5	$\hat{P}_{GREG}$	0.4614	0.4858	0.5075	0.5143	0.5091	0.5041	0.4823	0.4645
		$\hat{P}_{LGREG}$	0.4593	0.4845	0.5063	0.5132	0.5082	0.5031	0.4808	0.4624
100	0.2	$\hat{P}_{GREG}$	0.7810	0.8030	0.8218	0.8229	0.8237	0.8157	0.7846	0.7710
		$\hat{P}_{LGREG}$	0.7808	0.8032	0.8223	0.8234	0.8241	0.8162	0.7842	0.7714
	0.3	$\hat{P}_{GREG}$	0.6625	0.6967	0.6995	0.7085	0.7058	0.6999	0.6736	0.6504
		$\hat{P}_{LGREG}$	0.6611	0.6968	0.6991	0.7084	0.7055	0.6996	0.6736	0.6509
	0.5	$\hat{P}_{GREG}$	0.4682	0.4899	0.5074	0.5081	0.5030	0.4955	0.4720	0.4526
		$\hat{P}_{LGREG}$	0.4673	0.4892	0.5070	0.5079	0.5028	0.4952	0.4714	0.4515
250	0.2	$\hat{P}_{GREG}$	0.7525	0.7741	0.7997	0.8054	0.8064	0.7974	0.7800	0.7531
		$\hat{P}_{LGREG}$	0.7514	0.7742	0.7999	0.8055	0.8065	0.7974	0.7796	0.7519
	0.3	$\hat{P}_{GREG}$	0.6382	0.6851	0.7105	0.7073	0.6933	0.7018	0.6642	0.6385
		$\hat{P}_{LGREG}$	0.6365	0.6849	0.7100	0.7072	0.6933	0.7020	0.6630	0.6373
	0.5	$\hat{P}_{GREG}$	0.4549	0.4756	0.4937	0.4994	0.4992	0.4931	0.4702	0.4459
		$\hat{P}_{LGREG}$	0.4548	0.4755	0.4936	0.4993	0.4992	0.4930	0.4700	0.4453
500	0.2	$\hat{P}_{GREG}$	0.7643	0.7950	0.8122	0.8164	0.8063	0.8037	0.7781	0.7570
		$\hat{P}_{LGREG}$	0.7634	0.7949	0.8123	0.8164	0.8065	0.8039	0.7785	0.7565
	0.3	$\hat{P}_{GREG}$	0.6494	0.6930	0.7063	0.6985	0.6997	0.6861	0.6734	0.6450
		$\hat{P}_{LGREG}$	0.6479	0.6931	0.7061	0.6986	0.6998	0.6863	0.6730	0.6442
	0.5	$\hat{P}_{GREG}$	0.4585	0.4856	0.5101	0.5154	0.5166	0.5120	0.4902	0.4650
		$\hat{P}_{LGREG}$	0.4582	0.4855	0.5101	0.5154	0.5166	0.5120	0.4901	0.4648
1000	0.2	$\hat{P}_{GREG}$	0.7563	0.7811	0.8016	0.8056	0.8071	0.7989	0.7583	0.7553
		$\hat{P}_{LGREG}$	0.7529	0.7796	0.8014	0.8055	0.8079	0.7987	0.7575	0.7527
	0.3	$\hat{P}_{GREG}$	0.6463	0.6692	0.7132	0.6944	0.6918	0.7030	0.6634	0.6387
		$\hat{P}_{LGREG}$	0.6454	0.6682	0.7132	0.6945	0.6917	0.7032	0.6620	0.6362
	0.5	$\hat{P}_{GREG}$	0.4524	0.4765	0.4915	0.4977	0.4932	0.4885	0.4728	0.4498
		$\hat{P}_{LGREG}$	0.4521	0.4764	0.4914	0.4977	0.4932	0.4885	0.4728	0.4496

Quadro 6.11. Viés relativo do estimador da variância do estimador de P usando um plano BE.

n	P	Estimador	Razão de chances (OR)							
			0.1	0.2	0.5	0.65	1.5	2	5	10
50	0.2	$\hat{P}_{HT}$	0.0256	0.0166	0.0078	0.0057	0.0071	0.0069	0.0132	0.0005
		$\hat{P}_{GREG}$	0.0096	0.0204	0.0341	0.0368	0.0336	0.0298	0.0455	0.0207
		$\hat{P}_{LGREG1}$	0.0245	0.0349	0.0526	0.0215	0.0254	0.0326	0.0567	0.0325
		$\hat{P}_{LGREG2}$	0.0534	0.0610	0.0726	0.0745	0.0716	0.0680	0.0809	0.0652
	0.3	$\hat{P}_{HT}$	0.0118	0.0065	0.0189	0.0054	0.0130	0.0165	0.0045	0.0055
		$\hat{P}_{GREG}$	0.0575	0.0217	0.0500	0.0268	0.0411	0.0449	0.0430	0.0342
		$\hat{P}_{LGREG1}$	0.0694	0.0355	0.0653	0.0433	0.0556	0.0597	0.0587	0.0472
		$\hat{P}_{LGREG2}$	0.0969	0.0588	0.0794	0.0603	0.0797	0.0835	0.0822	0.0750
	0.5	$\hat{P}_{HT}$	0.0034	0.0004	0.0073	0.0030	0.0038	0.0044	0.0053	0.0009
		$\hat{P}_{GREG}$	0.0274	0.0262	0.0371	0.0388	0.0312	0.0302	0.0290	0.0320
		$\hat{P}_{LGREG1}$	0.0393	0.0414	0.0545	0.0563	0.0490	0.0473	0.0433	0.0429
		$\hat{P}_{LGREG2}$	0.0641	0.0631	0.0733	0.0751	0.0682	0.0672	0.0660	0.0686
100	0.2	$\hat{P}_{HT}$	0.0032	0.0057	0.0019	0.0009	0.0010	0.0022	0.0062	0.0112
		$\hat{P}_{GREG}$	0.0276	0.0224	0.0229	0.0240	0.0281	0.0228	0.0120	0.0180
		$\hat{P}_{LGREG1}$	0.0361	0.0308	0.0318	0.0331	0.0370	0.0319	0.0189	0.0272
		$\hat{P}_{LGREG2}$	0.0504	0.0436	0.0432	0.0442	0.0481	0.0432	0.0333	0.0412
	0.3	$\hat{P}_{HT}$	0.0078	0.0133	0.0208	0.0102	0.0227	0.0154	0.0109	0.0129
		$\hat{P}_{GREG}$	0.0310	0.0204	0.0234	0.0015	0.0062	0.0299	0.0289	0.0352
		$\hat{P}_{LGREG1}$	0.0393	0.0287	0.0318	0.0076	0.0025	0.0375	0.0377	0.0447
		$\hat{P}_{LGREG2}$	0.0501	0.0410	0.0410	0.0206	0.0165	0.0484	0.0496	0.0581
	0.5	$\hat{P}_{HT}$	0.0027	0.0018	0.0074	0.0049	0.0075	0.0001	0.0103	0.0149
		$\hat{P}_{GREG}$	0.0424	0.0393	0.0246	0.0224	0.0116	0.0119	0.0159	0.0252
		$\hat{P}_{LGREG1}$	0.0482	0.0463	0.0334	0.0314	0.0208	0.0209	0.0235	0.0311
		$\hat{P}_{LGREG2}$	0.0608	0.0574	0.0431	0.0411	0.0305	0.0308	0.0351	0.0440
250	0.2	$\hat{P}_{HT}$	0.0134	0.0165	0.0138	0.0069	0.0079	0.0134	0.0155	0.0148
		$\hat{P}_{GREG}$	0.0183	0.0233	0.0143	0.0036	0.0157	0.0179	0.0296	0.0228
		$\hat{P}_{LGREG1}$	0.0150	0.0193	0.0104	0.0193	0.0192	0.0215	0.0323	0.0260
		$\hat{P}_{LGREG2}$	0.0094	0.0149	0.0070	0.0032	0.0220	0.0241	0.0370	0.0291
	0.3	$\hat{P}_{HT}$	0.0058	0.0066	0.0164	0.0029	0.0073	0.0156	0.0016	0.0116
		$\hat{P}_{GREG}$	0.0095	0.0120	0.0007	0.0129	0.0045	0.0341	0.0033	0.0169
		$\hat{P}_{LGREG1}$	0.0073	0.0150	0.0030	0.0162	0.0009	0.0376	0.0060	0.0202
		$\hat{P}_{LGREG2}$	0.0008	0.0194	0.0079	0.0211	0.0025	0.0408	0.0088	0.0234
	0.5	$\hat{P}_{HT}$	0.0074	0.0060	0.0077	0.0105	0.0011	0.0005	0.0045	0.0010
		$\hat{P}_{GREG}$	0.0208	0.0154	0.0028	0.0002	0.0104	0.0077	0.0079	0.0004
		$\hat{P}_{LGREG1}$	0.0229	0.0180	0.0007	0.0034	0.0143	0.0115	0.0114	0.0022
		$\hat{P}_{LGREG2}$	0.0276	0.0219	0.0037	0.0063	0.0169	0.0141	0.0142	0.0054

Continuação da Quadro 6.11

n	P	Estimador	Razão de chances (OR)							
			0.1	0.2	0.5	0.65	1.5	2	5	10
500	0.2	$\hat{P}_{HT}$	0.0193	0.0305	0.0186	0.0240	0.0128	0.0199	0.0190	0.0169
		$\hat{P}_{GREG}$	0.0090	0.0112	0.0049	0.0085	0.0073	0.0102	0.0078	0.0077
		$\hat{P}_{LGREG1}$	0.0063	0.0089	0.0029	0.0067	0.0055	0.0082	0.0059	0.0052
		$\hat{P}_{LGREG2}$	0.0040	0.0068	0.0008	0.0044	0.0030	0.0056	0.0036	0.0012
	0.3	$\hat{P}_{HT}$	0.0012	0.0097	0.0028	0.0210	0.0151	0.0051	0.0180	0.0234
		$\hat{P}_{GREG}$	0.0036	0.0188	0.0131	0.0238	0.0264	0.0100	0.0018	0.0109
		$\hat{P}_{LGREG1}$	0.0054	0.0213	0.0148	0.0219	0.0281	0.0082	0.0001	0.0090
		$\hat{P}_{LGREG2}$	0.0065	0.0233	0.0177	0.0191	0.0293	0.0060	0.0031	0.0051
	0.5	$\hat{P}_{HT}$	0.0068	0.0144	0.0077	0.0045	0.0185	0.0261	0.0299	0.0355
		$\hat{P}_{GREG}$	0.0133	0.0146	0.0277	0.0348	0.0245	0.0166	0.0124	0.0018
		$\hat{P}_{LGREG1}$	0.0149	0.0164	0.0296	0.0367	0.0263	0.0182	0.0136	0.0026
		$\hat{P}_{LGREG2}$	0.0171	0.0185	0.0316	0.0387	0.0285	0.0206	0.0165	0.0059
1000	0.2	$\hat{P}_{HT}$	0.0236	0.0158	0.0232	0.0169	0.0015	0.0077	0.0149	0.0052
		$\hat{P}_{GREG}$	0.0244	0.0146	0.0229	0.0151	0.0062	0.0037	0.0003	0.0022
		$\hat{P}_{LGREG1}$	0.0258	0.0151	0.0223	0.0143	0.0069	0.0031	0.0007	0.0015
		$\hat{P}_{LGREG2}$	0.0246	0.0139	0.0210	0.0131	0.0082	0.0018	0.0013	0.0032
	0.3	$\hat{P}_{HT}$	0.0002	0.0074	0.0253	0.0105	0.0022	0.0038	0.0140	0.0175
		$\hat{P}_{GREG}$	0.0007	0.0136	0.0040	0.0015	0.0024	0.0222	0.0121	0.0138
		$\hat{P}_{LGREG1}$	0.0008	0.0139	0.0032	0.0025	0.0016	0.0230	0.0126	0.0149
		$\hat{P}_{LGREG2}$	0.0017	0.0126	0.0025	0.0045	0.0005	0.0248	0.0112	0.0133
	0.5	$\hat{P}_{HT}$	0.0012	0.0084	0.0082	0.0020	0.0001	0.0029	0.0036	0.0006
		$\hat{P}_{GREG}$	0.0062	0.0018	0.0096	0.0028	0.0030	0.0072	0.0028	0.0053
		$\hat{P}_{LGREG1}$	0.0065	0.0025	0.0087	0.0037	0.0021	0.0063	0.0035	0.0059
		$\hat{P}_{LGREG2}$	0.0078	0.0036	0.0076	0.0048	0.0010	0.0052	0.0048	0.0074

Quadro 6.12. Coeficiente de variação do estimador de P usando um plano BE.

n	P	Estimador	Razão de chances (OR)							
			0.1	0.2	0.5	0.65	1.5	2	5	10
50	0.2	$\hat{P}_{HT}$	31.595	31.769	31.949	31.949	31.810	31.604	31.642	31.649
		$\hat{P}_{GREG}$	27.958	28.295	28.553	28.555	28.289	27.995	27.507	27.541
		$\hat{P}_{LGREG1}$	29.415	29.049	27.564	25.546	26.542	26.458	28.372	26.789
		$\hat{P}_{LGREG2}$	26.822	26.839	26.804	26.729	26.573	26.366	26.297	26.604
	0.3	$\hat{P}_{HT}$	26.111	25.845	25.615	25.822	25.314	25.829	25.813	25.867
		$\hat{P}_{GREG}$	21.944	21.257	20.873	20.749	20.312	20.487	21.386	21.776
		$\hat{P}_{LGREG1}$	23.482	22.051	20.982	20.929	20.404	20.620	22.396	23.526
		$\hat{P}_{LGREG2}$	21.072	20.542	19.549	19.664	19.178	19.764	20.622	21.355
	0.5	$\hat{P}_{HT}$	20.011	19.929	19.967	19.932	19.939	19.936	19.989	19.996
		$\hat{P}_{GREG}$	18.541	17.147	15.888	15.754	15.854	16.110	17.471	18.814
		$\hat{P}_{LGREG1}$	20.352	18.102	16.047	15.762	15.904	16.285	18.335	20.451
		$\hat{P}_{LGREG2}$	17.992	16.734	15.660	15.517	15.566	15.740	16.893	18.154
100	0.2	$\hat{P}_{HT}$	22.470	22.465	22.548	22.552	22.467	22.274	22.079	22.124
		$\hat{P}_{GREG}$	18.964	19.059	19.127	19.087	19.077	18.825	18.586	18.574
		$\hat{P}_{LGREG1}$	19.808	19.489	19.244	19.166	19.179	18.988	19.053	19.530
		$\hat{P}_{LGREG2}$	18.618	18.540	18.551	18.533	18.465	18.274	17.991	18.227
	0.3	$\hat{P}_{HT}$	18.378	18.160	18.472	18.183	17.797	18.198	18.233	18.282
		$\hat{P}_{GREG}$	14.715	14.424	13.869	13.873	13.688	13.771	14.302	14.728
		$\hat{P}_{LGREG1}$	15.908	15.048	14.004	13.956	13.750	13.899	15.165	16.099
		$\hat{P}_{LGREG2}$	14.681	14.025	13.857	13.630	13.267	13.539	14.269	14.718
	0.5	$\hat{P}_{HT}$	14.100	14.053	13.959	13.975	13.959	14.005	14.113	14.195
		$\hat{P}_{GREG}$	12.442	11.436	10.515	10.399	10.477	10.659	11.699	12.680
		$\hat{P}_{LGREG1}$	13.740	12.142	10.635	10.425	10.568	10.857	12.520	14.138
		$\hat{P}_{LGREG2}$	12.276	11.382	10.613	10.477	10.487	10.599	11.357	12.216
250	0.2	$\hat{P}_{HT}$	14.015	14.011	14.051	14.088	14.145	14.085	14.012	14.059
		$\hat{P}_{GREG}$	11.382	11.452	11.574	11.623	11.639	11.504	11.406	11.409
		$\hat{P}_{LGREG1}$	11.865	11.689	11.636	11.662	11.697	11.592	11.696	11.985
		$\hat{P}_{LGREG2}$	11.462	11.445	11.461	11.488	11.504	11.414	11.265	11.441
	0.3	$\hat{P}_{HT}$	11.518	11.427	11.378	11.467	11.239	11.417	11.372	11.467
		$\hat{P}_{GREG}$	8.891	8.808	8.537	8.457	8.233	8.404	8.714	8.952
		$\hat{P}_{LGREG1}$	9.609	9.186	8.628	8.491	8.268	8.496	9.241	9.842
		$\hat{P}_{LGREG2}$	9.034	8.677	8.421	8.444	8.217	8.339	8.698	9.032
	0.5	$\hat{P}_{HT}$	8.871	8.841	8.763	8.751	8.792	8.796	8.831	8.842
		$\hat{P}_{GREG}$	7.642	7.039	6.474	6.383	6.394	6.484	7.023	7.601
		$\hat{P}_{LGREG1}$	8.556	7.538	6.556	6.398	6.463	6.645	7.661	8.675
		$\hat{P}_{LGREG2}$	7.553	6.972	6.442	6.376	6.397	6.481	7.017	7.596

Continuação da Quadro 6.12

n	P	Estimador	Razão de chances (OR)							
			0.1	0.2	0.5	0.65	1.5	2	5	10
500	0.2	$\hat{P}_{HT}$	9.766	9.722	9.796	9.761	9.783	9.683	9.622	9.675
		$\hat{P}_{GREG}$	7.951	8.027	8.079	8.074	8.004	7.926	7.797	7.835
		$\hat{P}_{LGREG1}$	8.308	8.194	8.109	8.086	8.026	7.970	7.994	8.223
		$\hat{P}_{LGREG2}$	7.947	7.899	7.966	7.936	7.927	7.810	7.774	7.766
	0.3	$\hat{P}_{HT}$	8.008	7.972	8.020	7.912	7.938	7.890	7.882	7.876
		$\hat{P}_{GREG}$	6.181	6.086	5.882	5.833	5.737	5.826	6.069	6.210
		$\hat{P}_{LGREG1}$	6.760	6.375	5.944	5.861	5.767	5.893	6.432	6.828
		$\hat{P}_{LGREG2}$	6.224	6.030	5.876	5.823	5.745	5.795	5.967	6.123
	0.5	$\hat{P}_{HT}$	6.149	6.110	6.120	6.129	6.090	6.065	6.0667	6.070
		$\hat{P}_{GREG}$	5.207	4.814	4.452	4.401	4.430	4.501	4.9079	5.318
		$\hat{P}_{LGREG1}$	5.916	5.210	4.525	4.417	4.462	4.595	5.3240	6.045
		$\hat{P}_{LGREG2}$	5.280	4.860	4.458	4.404	4.411	4.469	4.8384	5.247
1000	0.2	$\hat{P}_{HT}$	6.708	6.740	6.725	6.741	6.769	6.703	6.730	6.697
		$\hat{P}_{GREG}$	5.442	5.492	5.501	5.511	5.526	5.447	5.359	5.404
		$\hat{P}_{LGREG1}$	5.686	5.603	5.520	5.518	5.539	5.472	5.490	5.654
		$\hat{P}_{LGREG2}$	5.435	5.471	5.466	5.476	5.483	5.423	5.424	5.430
	0.3	$\hat{P}_{HT}$	5.518	5.492	5.443	5.532	5.413	5.457	5.433	5.433
		$\hat{P}_{GREG}$	4.234	4.144	4.102	4.036	3.937	4.014	4.130	4.234
		$\hat{P}_{LGREG1}$	4.634	4.330	4.136	4.062	3.952	4.055	4.376	4.660
		$\hat{P}_{LGREG2}$	4.288	4.133	4.002	4.056	3.929	3.978	4.148	4.292
	0.5	$\hat{P}_{HT}$	4.243	4.217	4.211	4.224	4.229	4.221	4.2293	4.251
		$\hat{P}_{GREG}$	3.658	3.363	3.081	3.038	3.036	3.076	3.3317	3.600
		$\hat{P}_{LGREG1}$	4.122	3.624	3.129	3.052	3.060	3.143	3.6252	4.096
		$\hat{P}_{LGREG2}$	3.590	3.314	3.063	3.028	3.051	3.098	3.3794	3.655

Quadro 6.13. Taxas de cobertura para um intervalo de confiança de 95% do estimador de P usando um plano BE.

n	P	Estimador	Razão de chances (OR)							
			0.1	0.2	0.5	0.65	1.5	2	5	10
50	0.2	$\hat{P}_{HT}$	0.9179	0.9166	0.9132	0.9170	0.9188	0.9211	0.9282	0.9258
		$\hat{P}_{GREG}$	0.9299	0.9296	0.9262	0.9238	0.9263	0.9253	0.9280	0.9281
		$\hat{P}_{LGREG1}$	0.9277	0.9269	0.9255	0.9238	0.9245	0.9217	0.9271	0.9262
		$\hat{P}_{LGREG2}$	0.9187	0.9161	0.9151	0.9163	0.9150	0.9130	0.9179	0.9212
	0.3	$\hat{P}_{HT}$	0.9470	0.9502	0.9556	0.9510	0.9358	0.9199	0.9206	0.9165
		$\hat{P}_{GREG}$	0.9340	0.9346	0.9334	0.9337	0.9349	0.9313	0.9332	0.9377
		$\hat{P}_{LGREG1}$	0.9310	0.9339	0.9322	0.9319	0.9324	0.9297	0.9304	0.9335
		$\hat{P}_{LGREG2}$	0.9247	0.9227	0.9265	0.9249	0.9258	0.9201	0.9217	0.9255
	0.5	$\hat{P}_{HT}$	0.9480	0.9259	0.9319	0.9314	0.9320	0.9321	0.9253	0.9466
		$\hat{P}_{GREG}$	0.9408	0.9421	0.9374	0.9393	0.9397	0.9396	0.9389	0.9388
		$\hat{P}_{LGREG1}$	0.9385	0.9384	0.9345	0.9354	0.9371	0.9376	0.9363	0.9363
		$\hat{P}_{LGREG2}$	0.9288	0.9306	0.9260	0.9275	0.9294	0.9287	0.9288	0.9266
100	0.2	$\hat{P}_{HT}$	0.9433	0.9426	0.9418	0.9410	0.9451	0.9491	0.9314	0.9504
		$\hat{P}_{GREG}$	0.9379	0.9410	0.9371	0.9390	0.9379	0.9409	0.9405	0.9408
		$\hat{P}_{LGREG1}$	0.9351	0.9402	0.9359	0.9372	0.9374	0.9388	0.9395	0.9398
		$\hat{P}_{LGREG2}$	0.9309	0.9333	0.9325	0.9337	0.9329	0.9345	0.9335	0.9326
	0.3	$\hat{P}_{HT}$	0.9452	0.9517	0.9465	0.9486	0.9477	0.9372	0.9320	0.9302
		$\hat{P}_{GREG}$	0.9405	0.9420	0.9398	0.9418	0.9467	0.9407	0.9425	0.9394
		$\hat{P}_{LGREG1}$	0.9386	0.9413	0.9394	0.9414	0.9455	0.9397	0.9404	0.9366
		$\hat{P}_{LGREG2}$	0.9374	0.9389	0.9367	0.9390	0.9409	0.9384	0.9377	0.9334
	0.5	$\hat{P}_{HT}$	0.9500	0.9435	0.9452	0.9447	0.9453	0.9453	0.9359	0.9454
		$\hat{P}_{GREG}$	0.9438	0.9408	0.9427	0.9464	0.9494	0.9459	0.9425	0.9429
		$\hat{P}_{LGREG1}$	0.9425	0.9411	0.9416	0.9456	0.9481	0.9450	0.9424	0.9409
		$\hat{P}_{LGREG2}$	0.9377	0.9368	0.9376	0.9395	0.9419	0.9407	0.9376	0.9376
250	0.2	$\hat{P}_{HT}$	0.9426	0.9556	0.9529	0.9530	0.9419	0.9490	0.9435	0.9518
		$\hat{P}_{GREG}$	0.9514	0.9503	0.9499	0.9476	0.9461	0.9445	0.9403	0.9449
		$\hat{P}_{LGREG1}$	0.9513	0.9492	0.9492	0.9470	0.9454	0.9440	0.9407	0.9438
		$\hat{P}_{LGREG2}$	0.9478	0.9485	0.9468	0.9450	0.9435	0.9432	0.9408	0.9424
	0.3	$\hat{P}_{HT}$	0.9408	0.9434	0.9455	0.9443	0.9483	0.9514	0.9427	0.9504
		$\hat{P}_{GREG}$	0.9490	0.9494	0.9470	0.9476	0.9479	0.9486	0.9483	0.9456
		$\hat{P}_{LGREG1}$	0.9502	0.9485	0.9460	0.9478	0.9473	0.9478	0.9478	0.9457
		$\hat{P}_{LGREG2}$	0.9476	0.9469	0.9447	0.9433	0.9463	0.9462	0.9466	0.9434
	0.5	$\hat{P}_{HT}$	0.9468	0.9461	0.9519	0.9537	0.9519	0.9519	0.9437	0.9511
		$\hat{P}_{GREG}$	0.9450	0.9467	0.9499	0.9511	0.9517	0.9512	0.9484	0.9515
		$\hat{P}_{LGREG1}$	0.9448	0.9459	0.9490	0.9506	0.9510	0.9505	0.9466	0.9513
		$\hat{P}_{LGREG2}$	0.9436	0.9446	0.9470	0.9484	0.9500	0.9505	0.9461	0.9496

Continuação da Quadro 6.13

n	P	Estimador	Razão de chances (OR)							
			0.1	0.2	0.5	0.65	1.5	2	5	10
500	0.2	$\hat{P}_{HT}$	0.9489	0.9487	0.9410	0.9497	0.9532	0.9487	0.9469	0.9473
		$\hat{P}_{GREG}$	0.9510	0.9491	0.9482	0.9495	0.9503	0.9497	0.9485	0.9495
		$\hat{P}_{LGREG1}$	0.9495	0.9488	0.9485	0.9490	0.9503	0.9492	0.9483	0.9491
		$\hat{P}_{LGREG2}$	0.9482	0.9477	0.9465	0.9466	0.9482	0.9493	0.9473	0.9473
	0.3	$\hat{P}_{HT}$	0.9469	0.9501	0.9480	0.9472	0.9441	0.9480	0.9516	0.9498
		$\hat{P}_{GREG}$	0.9475	0.9464	0.9504	0.9522	0.9458	0.9495	0.9483	0.9493
		$\hat{P}_{LGREG1}$	0.9462	0.9458	0.9505	0.9519	0.9456	0.9492	0.9478	0.9482
		$\hat{P}_{LGREG2}$	0.9462	0.9449	0.9493	0.9517	0.9442	0.9485	0.9485	0.9479
	0.5	$\hat{P}_{HT}$	0.9470	0.9502	0.9438	0.9445	0.9481	0.9500	0.9502	0.9539
		$\hat{P}_{GREG}$	0.9441	0.9445	0.9453	0.9449	0.9427	0.9428	0.9450	0.9482
		$\hat{P}_{LGREG1}$	0.9440	0.9445	0.9446	0.9445	0.9425	0.9425	0.9448	0.9476
		$\hat{P}_{LGREG2}$	0.9443	0.9455	0.9464	0.9439	0.9428	0.9424	0.9436	0.9459
1000	0.2	$\hat{P}_{HT}$	0.9570	0.9514	0.9504	0.9499	0.9491	0.9486	0.9471	0.9477
		$\hat{P}_{GREG}$	0.9527	0.9490	0.9527	0.9511	0.9461	0.9490	0.9500	0.9493
		$\hat{P}_{LGREG1}$	0.9523	0.9496	0.9526	0.9510	0.9463	0.9493	0.9502	0.9493
		$\hat{P}_{LGREG2}$	0.9520	0.9492	0.9514	0.9505	0.9456	0.9486	0.9495	0.9487
	0.3	$\hat{P}_{HT}$	0.9489	0.9491	0.9553	0.9493	0.9500	0.9506	0.9514	0.9503
		$\hat{P}_{GREG}$	0.9486	0.9510	0.9493	0.9489	0.9491	0.9495	0.9512	0.9521
		$\hat{P}_{LGREG1}$	0.9486	0.9513	0.9493	0.9487	0.9493	0.9496	0.9530	0.9534
		$\hat{P}_{LGREG2}$	0.9475	0.9508	0.9491	0.9475	0.9496	0.9495	0.9522	0.9516
	0.5	$\hat{P}_{HT}$	0.9463	0.9488	0.9491	0.9465	0.9492	0.9488	0.9468	0.9480
		$\hat{P}_{GREG}$	0.9478	0.9480	0.9497	0.9475	0.9518	0.9512	0.9501	0.9474
		$\hat{P}_{LGREG1}$	0.9479	0.9481	0.9496	0.9478	0.9513	0.9511	0.9501	0.9476
		$\hat{P}_{LGREG2}$	0.9482	0.9481	0.9493	0.9480	0.9505	0.9501	0.9499	0.9475

6.2.3 Resultados para Amostragem Estratificada

Nos Quadros 6.14 a 6.19 encontra-sem as medidas que permitem comparar e avaliar os estimadores $\hat{P}_{HT}$, $\hat{P}_{GREG}$ e $\hat{P}_{LGREG}$ quando o plano usado para a seleção da amostra é um AAE, considerando os cenários separados e combinado para os estimadores $\hat{P}_{GREG}$ e $\hat{P}_{LGREG}$.
No Quadro 6.14 é apresentado o viés relativo dos estimadores. Os estimadores $\hat{P}_{GREG}$ e $\hat{P}_{LGREG}$ podem ser considerados como não-viesados em todos os cenários de simulação. Pois, este resultado não está sendo influenciado pelo tamanho da amostra, pelo grau de associação entre as variáveis nos estratos, nem pela proporção de indivíduos com a característica de interesse na população.
No Quadro 6.15 encontra-se a eficiência dos estimadores. Os resultados obtidos para esta medida podem ser explicados da seguinte forma:

- Cenário 1: Quando são considerados diferentes OR em cada estrato, são mais eficientes os estimadores de tipo separados. Esta eficiência vai aumentando na medida que P aumenta, e não está influenciada pelo tamanho da amostra. Fazendo uma comparação entre os cenários 1_a e 1_b, pode ser notado que em 1_a, os estimadores combinados são pelo menos tão eficientes quanto o estimador de Horvitz-Thompson.

- Cenário 2: neste cenário são considerados OR iguais para cada estrato. Os estimadores separados e os combinados apresentam eficiência semelhante, embora que este cenário apresenta estimadores mais eficientes no caso dos estimadores combinados. Comparando os cenários 2_a e 2_b, a eficiência dos estimadores no cenário 2_b é maior. Isto é devido a que o grau de associação considerada entre as variáveis de interesse e auxiliar é mais forte.

Nos cenários de simulação 1 e 2, os estimadores do tipo $\hat{P}_{LGREG}$ apresentam-se como os mais eficientes, considerando o caso separado e combinado, embora que, os resultados de $\hat{P}_{GREG}$ e $\hat{P}_{LGREG}$ estejam muito próximos.
No Quadro 6.16 é apresentada a eficiência dos estimadores do ponto de vista do erro quadrático médio. Este resultado é muito similar aos descritos

no Quadro 6.15, pois como já foi mencionado antes, estas duas medidas são equivalentes quando o estimador é não-viesado.

No Quadro 6.17, encontra-se o viés relativo do estimador da variância do estimador. Os estimadores da variância apresentam-se como não-viesados, pois em todos os cenários o viés é menor do $5\%(0.05)$, ou seja, este resultado não está sendo influenciado pelo grau de associação entre as variáveis (OR), nem pela proporção de indivíduos na população com a caraterística de interesse (P), nem pelo tamanho da amostra (n).

No Quadro 6.18 podem ser observados os coeficientes de variação dos estimadores. Esta medida mostra que o cenário ideal para ter um coeficiente de variação bem pequeno, está dado quando o tamanho da amostra cresce, e P aumenta. Entretanto, a mudança do grau de associação em cada um dos estratos não afeta este resultado. No Quadro 6.19, apresenta-se a taxa de cobertura dos estimadores, que em todos os casos é maior que 0.93.

Concluindo, a estimação intervalar dos estimadores $\hat{P}_{HT}$, $\hat{P}_{GREG}$ e $\hat{P}_{LGREG}$ não está sendo afetada quando é usado um plano estratificado, não importando que os estratos sejam ou não homogêneos. Em geral, o ganho de usar a estratificação está no sentido de obter maior eficiência dos estimadores quando são usados os estimadores de regressão separados em estratos heterogêneos.

Com base nos resultados das simulações é recomendável usar o estimador $\hat{P}_{LGREGC}$ nos casos em que a relação entre a variável de interesse e as variáveis auxiliares é assumida igual para toda a população, isto é, sem fazer diferença da relação de um estrato para outro. Analogamente, é recomendado usar o estimador $\hat{P}_{LGREGS}$ quando é assumida uma relação diferente entre a variável de interesse e a variável auxiliar em cada estrato, pois desta maneira, são obtidos estimadores mais eficientes. Com respeito às outras medidas observadas, estes estimadores têm comportamentos similares.

Quadro 6.14. Viés relativo do estimador de P usando um plano AAE.

n	P	Cenário	Estimador				
			$\hat{P}_{HT}$	Separado		Combinado	
				$\hat{P}_{GREG}$	$\hat{P}_{LGREG}$	$\hat{P}_{GREG}$	$\hat{P}_{LGREG}$
100	0.2	1_a	0.0026	0.0083	0.0039	0.0010	0.0008
		1_b	0.0014	0.0081	0.0006	0.0013	0.0010
		2_a	0.0010	0.0047	0.0014	0.0028	0.0020
		2_b	0.0080	0.0154	0.0001	0.0091	0.0049
	0.3	1_a	0.0016	0.0046	0.0008	0.0003	0.0002
		1_b	0.0012	0.0043	0.0007	0.0004	0.0001
		2_a	0.0002	0.0022	0.0005	0.0012	0.0007
		2_b	0.0006	0.0034	0.0015	0.0005	0.0014
	0.5	1_a	0.0002	0.0007	0.0006	0.0003	0.0003
		1_b	0.0003	0.0002	0.0034	0.0008	0.0007
		2_a	0.0004	0.0001	0.0001	0.0005	0.0005
		2_b	0.0014	0.0008	0.0006	0.0007	0.0007
500	0.2	1_a	0.0012	0.0065	0.0022	0.0027	0.0027
		1_b	0.0020	0.0063	0.0025	0.0030	0.0028
		2_a	0.0026	0.0039	0.0025	0.0031	0.0026
		2_b	0.0008	0.0048	0.0013	0.0024	0.0011
	0.3	1_a	0.0007	0.0040	0.0016	0.0015	0.0015
		1_b	0.0011	0.0038	0.0019	0.0017	0.0016
		2_a	0.0010	0.0018	0.0010	0.0013	0.0010
		2_b	0.0018	0.0035	0.0014	0.0023	0.0013
	0.5	1_a	0.0003	0.0003	0.0004	0.0003	0.0003
		1_b	0.0006	0.0005	0.0001	0.0005	0.0005
		2_a	0.0006	0.0007	0.0007	0.0007	0.0007
		2_b	0.0005	0.0003	0.0003	0.0003	0.0003
1000	0.2	1_a	0.0001	0.0022	0.0003	0.0005	0.0004
		1_b	0.0002	0.0022	0.0006	0.0007	0.0006
		2_a	0.0003	0.0011	0.0005	0.0007	0.0005
		2_b	0.0010	0.0005	0.0013	0.0006	0.0013
	0.3	1_a	0.0001	0.0013	0.0003	0.0002	0.0002
		1_b	0.0001	0.0014	0.0005	0.0005	0.0004
		2_a	0.0001	0.0003	0.0009	0.0001	0.0004
		2_b	0.0005	0.0011	0.0002	0.0005	0.0002
	0.5	1_a	0.0002	0.0005	0.0001	0.0003	0.0003
		1_b	0.0002	0.0002	0.0003	0.0005	0.0005
		2_a	0.0001	0.0003	0.0003	0.0002	0.0002
		2_b	0.0004	0.0002	0.0002	0.0004	0.0002

Quadro 6.15. Eficiência do estimador de P usando um plano AAE.

n	P	Cenário	Estimador			
			Separado		Combinado	
			$\hat{P}_{GREG}$	$\hat{P}_{LGREG}$	$\hat{P}_{GREG}$	$\hat{P}_{LGREG}$
100	0.2	1_a	0.8252	0.7302	1.0099	1.0052
		1_b	0.8443	0.7771	0.9753	0.9667
		2_a	0.9365	0.9320	0.9271	0.9257
		2_b	0.8292	0.7796	0.8271	0.7782
	0.3	1_a	0.6776	0.6237	1.0033	1.0016
		1_b	0.7741	0.7362	0.9882	0.9796
		2_a	0.9500	0.9498	0.9274	0.9294
		2_b	0.7262	0.6959	0.7166	0.6891
	0.5	1_a	0.6107	0.5789	1.0032	1.0032
		1_b	0.7399	0.7240	0.9877	0.9864
		2_a	0.9411	0.9369	0.9285	0.9262
		2_b	0.7413	0.7171	0.7310	0.7078
500	0.2	1_a	0.7467	0.6646	1.0021	1.0001
		1_b	0.8285	0.7721	0.9896	0.9830
		2_a	0.9564	0.9491	0.9538	0.9477
		2_b	0.7856	0.7370	0.7817	0.7301
	0.3	1_a	0.6811	0.6242	1.0005	0.9997
		1_b	0.7781	0.7440	0.9899	0.9858
		2_a	0.9459	0.9392	0.9427	0.9373
		2_b	0.7169	0.6776	0.7117	0.6724
	0.5	1_a	0.6347	0.6043	1.0013	1.0012
		1_b	0.7384	0.7107	0.9891	0.9880
		2_a	0.9126	0.9104	0.9086	0.9074
		2_b	0.6774	0.6618	0.6736	0.6598
1000	0.2	1_a	0.7295	0.6502	1.0007	0.9999
		1_b	0.8028	0.7547	0.9802	0.9745
		2_a	0.9360	0.9304	0.9346	0.9288
		2_b	0.8058	0.7502	0.8039	0.7480
	0.3	1_a	0.6875	0.6354	0.9989	0.9985
		1_b	0.7750	0.7428	0.9811	0.9780
		2_a	0.9204	0.9177	0.9177	0.9150
		2_b	0.7227	0.6932	0.7202	0.6903
	0.5	1_a	0.6396	0.6161	0.9995	0.9995
		1_b	0.7311	0.7103	0.9710	0.9701
		2_a	0.8938	0.8925	0.8929	0.8923
		2_b	0.6699	0.6553	0.6681	0.6539

Quadro 6.16. Eficiência do ponto de vista do EQM do estimador de P usando um plano AAE.

n	P	Cenário	Estimador			
			Separado		Combinado	
			$\hat{P}_{GREG}$	$\hat{P}_{LGREG}$	$\hat{P}_{GREG}$	$\hat{P}_{LGREG}$
100	0.2	1_a	0.8351	0.7304	1.0121	1.0071
		1_b	0.8460	0.7770	0.9751	0.9665
		2_a	0.9413	0.9390	0.9304	0.9278
		2_b	0.8337	0.7795	0.8283	0.7783
	0.3	1_a	0.6905	0.6407	1.0047	1.0031
		1_b	0.7757	0.7358	0.9891	0.9800
		2_a	0.9248	0.9209	0.9163	0.9147
		2_b	0.7463	0.7172	0.7329	0.7062
	0.5	1_a	0.6516	0.6161	1.0025	1.0023
		1_b	0.7387	0.7230	0.9881	0.9868
		2_a	0.9047	0.8986	0.8950	0.8934
		2_b	0.7142	0.6955	0.6996	0.6901
500	0.2	1_a	0.7516	0.6651	1.0028	1.0008
		1_b	0.8331	0.7725	0.9902	0.9835
		2_a	0.9575	0.9490	0.9542	0.9478
		2_b	0.7882	0.7371	0.7823	0.7302
	0.3	1_a	0.6843	0.6246	1.0009	1.0001
		1_b	0.7812	0.7445	0.9903	0.9860
		2_a	0.9464	0.9393	0.9429	0.9373
		2_b	0.7191	0.6776	0.7123	0.6724
	0.5	1_a	0.6348	0.6043	1.0013	1.0012
		1_b	0.7383	0.7106	0.9891	0.9879
		2_a	0.9127	0.9105	0.9087	0.9075
		2_b	0.6773	0.6618	0.6736	0.6598
1000	0.2	1_a	0.7308	0.6502	1.0007	1.0001
		1_b	0.8042	0.7548	0.9803	0.9747
		2_a	0.9363	0.9305	0.9347	0.9288
		2_b	0.8056	0.7505	0.8038	0.7482
	0.3	1_a	0.6883	0.6354	0.9989	0.9985
		1_b	0.7761	0.7430	0.9812	0.9780
		2_a	0.9205	0.9177	0.9177	0.9150
		2_b	0.7232	0.6931	0.7202	0.6902
	0.5	1_a	0.6396	0.6160	0.9996	0.9995
		1_b	0.7312	0.7104	0.9710	0.9701
		2_a	0.8939	0.8926	0.8930	0.8923
		2_b	0.6698	0.6552	0.6680	0.6538

Quadro 6.17. Viés relativo do estimador da variância do estimador de P usando um plano AAE.

n	P	Cenário	Estimador						
			$\hat{P}_{HT}$	Separado			Combinado		
				$\hat{P}_{GREG}$	$\hat{P}_{LGREG1}$	$\hat{P}_{LGREG2}$	$\hat{P}_{GREG}$	$\hat{P}_{LGREG1}$	$\hat{P}_{LGREG2}$
100	0.2	1_a	0.0074	0.0636	0.0334	0.0709	0.0047	0.0153	0.0083
		1_b	0.0313	0.0112	0.0301	0.0099	0.0419	0.0368	0.0401
		2_a	0.0176	0.0235	0.0256	0.0443	0.0115	0.0190	0.0193
		2_b	0.0502	0.0087	0.0135	0.0077	0.0112	0.0118	0.0056
	0.3	1_a	0.0094	0.0068	0.0089	0.0359	0.0059	0.0041	0.0004
		1_b	0.0353	0.0524	0.0519	0.0243	0.0212	0.0181	0.0187
		2_a	0.0196	0.0234	0.0347	0.0395	0.0155	0.0217	0.0221
		2_b	0.0175	0.0030	0.0094	0.0355	0.0139	0.0033	0.0069
	0.5	1_a	0.0172	0.0205	0.0226	0.0383	0.0184	0.0246	0.0246
		1_b	0.0326	0.0178	0.0365	0.0457	0.0413	0.0456	0.0462
		2_a	0.0226	0.0236	0.0320	0.0355	0.0165	0.0204	0.0210
		2_b	0.0177	0.0064	0.0133	0.0293	0.0050	0.0054	0.0117
500	0.2	1_a	0.0158	0.0185	0.0095	0.0233	0.0183	0.0228	0.0201
		1_b	0.0210	0.0299	0.0205	0.0308	0.0275	0.0305	0.0291
		2_a	0.0015	0.0189	0.0209	0.0237	0.0157	0.0163	0.0171
		2_b	0.0124	0.0241	0.0241	0.0114	0.0286	0.0323	0.0283
	0.3	1_a	0.0226	0.0184	0.0150	0.0246	0.0230	0.0270	0.0254
		1_b	0.0127	0.0149	0.0106	0.0037	0.0019	0.0002	0.0004
		2_a	0.0484	0.0208	0.0201	0.0178	0.0239	0.0250	0.0244
		2_b	0.0122	0.0069	0.0156	0.0049	0.0121	0.0224	0.0181
	0.5	1_a	0.0036	0.0184	0.0174	0.0124	0.0015	0.0010	0.0009
		1_b	0.0115	0.0097	0.0128	0.0092	0.0006	0.0023	0.0023
		2_a	0.0095	0.0037	0.0071	0.0086	0.0008	0.0015	0.0019
		2_b	0.0107	0.0116	0.0093	0.0030	0.0137	0.0123	0.0099
1000	0.2	1_a	0.0405	0.0614	0.0678	0.0610	0.0397	0.0372	0.0385
		1_b	0.0402	0.0635	0.0638	0.0587	0.0435	0.0415	0.0422
		2_a	0.0534	0.0545	0.0524	0.0515	0.0566	0.0563	0.0582
		2_b	0.0300	0.0148	0.0222	0.0154	0.0168	0.0248	0.0221
	0.3	1_a	0.0617	0.0537	0.0491	0.0445	0.0626	0.0604	0.0612
		1_b	0.0395	0.0439	0.0387	0.0354	0.0374	0.0357	0.0361
		2_a	0.0421	0.0419	0.0391	0.0384	0.0451	0.0439	0.0439
		2_b	0.0204	0.0295	0.0236	0.0182	0.0322	0.0276	0.0252
	0.5	1_a	0.0162	0.0206	0.0113	0.0088	0.0151	0.0139	0.0139
		1_b	0.0362	0.0419	0.0385	0.0366	0.0427	0.0418	0.0419
		2_a	0.0228	0.0298	0.0281	0.0276	0.0301	0.0295	0.0295
		2_b	0.0292	0.0611	0.0604	0.0507	0.0621	0.0628	0.0612

Quadro 6.18. Coeficiente de variação do estimador de P usando AAE.

n	P	Cenário	Estimador						
			$\hat{P}_{HT}$	Separado			Combinado		
				$\hat{P}_{GREG}$	$\hat{P}_{LGREG1}$	$\hat{P}_{LGREG2}$	$\hat{P}_{GREG}$	$\hat{P}_{LGREG1}$	$\hat{P}_{LGREG2}$
100	0.2	1_a	11.753	11.629	18.828	14.269	12.068	12.476	12.008
		1_b	8.831	8.642	12.152	10.060	8.817	9.110	8.940
		2_a	8.551	8.785	9.853	8.926	8.649	9.673	8.819
		2_b	9.202	8.864	13.225	10.533	8.782	13.240	10.319
	0.3	1_a	8.023	9.798	17.710	13.615	8.337	8.645	8.278
		1_b	6.678	7.851	11.842	9.566	7.069	7.764	7.180
		2_a	1.700	6.007	7.644	5.593	5.746	7.581	5.514
		2_b	7.094	8.656	14.025	10.679	8.434	13.841	10.376
	0.5	1_a	1.217	7.986	13.310	9.734	1.513	1.586	1.348
		1_b	1.305	6.176	9.795	7.031	2.543	3.320	2.464
		2_a	1.249	4.866	6.492	4.684	4.622	6.403	4.593
		2_b	1.179	7.550	12.006	8.686	7.282	11.89	8.520
500	0.2	1_a	7.136	6.456	11.318	8.700	7.188	7.310	7.200
		1_b	6.819	6.555	9.223	7.704	6.834	7.100	6.874
		2_a	6.632	6.748	7.407	6.782	6.697	7.341	6.733
		2_b	7.385	6.837	10.468	8.448	6.784	10.43	8.347
	0.3	1_a	4.271	5.246	9.685	7.049	4.278	4.356	4.287
		1_b	3.937	4.655	7.315	5.627	4.049	4.403	4.092
		2_a	3.729	4.326	5.222	4.364	4.281	5.184	4.327
		2_b	4.211	5.213	8.757	6.485	5.140	8.753	6.438
	0.5	1_a	0.464	4.640	7.952	5.772	0.619	0.782	0.594
		1_b	0.483	3.677	6.056	4.347	1.313	1.876	1.307
		2_a	0.475	2.887	3.967	2.835	2.803	3.959	2.810
		2_b	0.498	4.580	7.281	5.286	4.486	7.245	5.248
1000	0.2	1_a	4.776	4.395	7.670	5.847	4.780	4.817	4.784
		1_b	4.534	4.365	6.177	5.152	4.522	4.685	4.590
		2_a	4.428	4.433	4.888	4.533	4.409	4.860	4.519
		2_b	5.024	4.727	7.257	5.747	4.703	7.256	5.711
	0.3	1_a	2.801	3.566	6.635	4.767	2.783	2.810	2.795
		1_b	2.656	3.179	4.984	3.810	2.718	2.952	2.766
		2_a	2.544	2.928	3.535	3.008	2.897	3.515	2.997
		2_b	2.847	3.531	6.062	4.412	3.505	6.057	4.392
	0.5	1_a	0.225	3.128	5.464	3.907	0.337	0.452	0.326
		1_b	0.237	2.528	4.158	2.995	0.855	1.256	0.866
		2_a	0.238	1.907	2.672	1.897	1.880	2.671	1.891
		2_b	0.233	3.034	5.046	3.551	3.011	5.036	3.537

Quadro 6.19. Taxas de cobertura para um intervalo de confiança de 95% do estimador de P usando AAE.

n	P	Cenário	Estimador						
			$\hat{P}_{HT}$	Separado			Combinado		
				$\hat{P}_{GREG}$	$\hat{P}_{LGREG1}$	$\hat{P}_{LGREG2}$	$\hat{P}_{GREG}$	$\hat{P}_{LGREG1}$	$\hat{P}_{LGREG2}$
100	0.2	1_a	0.9500	0.9458	0.9420	0.9460	0.9478	0.9458	0.9474
		1_b	0.9394	0.9416	0.9490	0.9491	0.9458	0.9462	0.9454
		2_a	0.9460	0.9402	0.9396	0.9392	0.9431	0.9411	0.9416
		2_b	0.9416	0.9408	0.9411	0.9411	0.9426	0.9436	0.9428
	0.3	1_a	0.9542	0.9480	0.9454	0.9451	0.9526	0.9516	0.9518
		1_b	0.9452	0.9456	0.9511	0.9471	0.9478	0.9472	0.9468
		2_a	0.9474	0.9432	0.9418	0.9432	0.9436	0.9428	0.9438
		2_b	0.9538	0.9481	0.9448	0.9424	0.9518	0.9446	0.9466
	0.5	1_a	0.9504	0.9446	0.9418	0.9388	0.9462	0.9456	0.9454
		1_b	0.9508	0.9486	0.9440	0.9431	0.9514	0.9510	0.9506
		2_a	0.9436	0.9448	0.9444	0.9436	0.9462	0.9444	0.9460
		2_b	0.9496	0.9486	0.9478	0.9460	0.9508	0.9470	0.9466
500	0.2	1_a	0.9466	0.9432	0.9430	0.9451	0.9428	0.9420	0.9426
		1_b	0.9460	0.9408	0.9450	0.9428	0.9428	0.9422	0.9434
		2_a	0.9426	0.9412	0.9410	0.9398	0.9418	0.9424	0.9424
		2_b	0.9446	0.9494	0.9476	0.9466	0.9502	0.9492	0.9498
	0.3	1_a	0.9502	0.9431	0.9452	0.9424	0.9482	0.9482	0.9482
		1_b	0.9434	0.9480	0.9510	0.9498	0.9461	0.9461	0.9456
		2_a	0.9508	0.9496	0.9482	0.9481	0.9490	0.9496	0.9490
		2_b	0.9468	0.9482	0.9482	0.9458	0.9496	0.9486	0.9472
	0.5	1_a	0.9502	0.9508	0.9506	0.9502	0.9524	0.9522	0.9522
		1_b	0.9540	0.9534	0.9534	0.9528	0.9518	0.9518	0.9516
		2_a	0.9448	0.9464	0.9450	0.9462	0.9474	0.9456	0.9460
		2_b	0.9456	0.9532	0.9506	0.9494	0.9528	0.9522	0.9516
1000	0.2	1_a	0.9524	0.9576	0.9590	0.9592	0.9541	0.9538	0.9538
		1_b	0.9508	0.9530	0.9544	0.9534	0.9522	0.9521	0.9514
		2_a	0.9528	0.9594	0.9568	0.9568	0.9578	0.9562	0.9574
		2_b	0.9518	0.9531	0.9530	0.9524	0.9532	0.9532	0.9528
	0.3	1_a	0.9548	0.9568	0.9532	0.9516	0.9546	0.9544	0.9544
		1_b	0.9556	0.9522	0.9500	0.9502	0.9556	0.9548	0.9552
		2_a	0.9560	0.9540	0.9546	0.9548	0.9536	0.9536	0.9530
		2_b	0.9508	0.9550	0.9514	0.9511	0.9558	0.9516	0.9510
	0.5	1_a	0.9478	0.9494	0.9498	0.9510	0.9462	0.9461	0.9462
		1_b	0.9536	0.9548	0.9520	0.9516	0.9546	0.9544	0.9544
		2_a	0.9562	0.9536	0.9528	0.9532	0.9556	0.9542	0.9548
		2_b	0.9522	0.9568	0.9552	0.9532	0.9568	0.9568	0.9554

CAPÍTULO 7

Ilustração do Uso dos Estimadores GREG's

7.1 A Pesquisa Mensal de Emprego (PME)

A PME foi iniciada em 1980, e foi submetida a uma revisão completa em 1982 e duas parciais nos anos 1988 e 1993. Nestas revisões foram realizados ajustes restritos somente ao plano amostral. Em 2001, foi feita uma revisão metodológica com o objetivo de atualizar sua cobertura temática e se adequar às recomendações da Organização Internacional do Trabalho (OIT). Seu objetivo é produzir indicadores mensais sobre a força de trabalho que permitam avaliar as flutuações e a tendência, a médio e a longo prazo, do mercado de trabalho. A PME é usada para dar indicativo ágil dos efeitos da conjuntura econômica sobre o mercado de trabalho. Além disso, atende a outras necessidades importantes para o planejamento socioeconômico do País.

Atualmente a PME abrange as Regiões Metropolitanas de Recife, Salvador, Belo Horizonte, Rio de Janeiro, São Paulo e Porto Alegre. Esta pesquisa é domiciliar, onde a seleção dos domícilios é feita usando uma amostra probabilística. Sua periodicidade é mensal e investiga características da população residente de 10 anos ou mais de idade na área urbana das regiões metropolitanas estudadas.

7.1.1 Conceitos Básicos

Na pesquisa mensal de emprego (PME) a unidade de investigação é a pessoa, entretanto, a unidade amostral é definida como o domícilio. Portanto, com a finalidade de fazer diferenciação entre a unidade domiciliar e as pessoas que são objeto da pesquisa, são adotadas as seguintes definições:
Domicílio: É o local de moradia de uma ou mais pessoas, o qual verifica as condições de separação e independência de acesso. Os domicílios são classificados em particulares ou coletivos. Sendo os particulares aqueles onde o relacionamento é ditado por laços de parentesco, dependência doméstica ou normas de convivência. Os coletivos são moradias onde prevalece o cumprimento de normas administrativas.
Unidade domiciliar: É o domicílio particular ou a unidade de habitação em domicílio coletivo.
Morador: É a pessoa que tem a unidade domiciliar como residência na data da entrevista.
Pessoas abrangidas pela pesquisa: A PME pesquisa a população residente, excluindo as pessoas moradoras em embaixadas, consulados ou legações e as pessoas institucionalizadas moradores em domicílios coletivos de estabelecimentos institucionais.

7.1.2 Características Investigadas

Para os domicílios que compõem a amostra são observadas todas as informações de localização e caracterização da espécie do domicílio. Para todos os moradores que residem nos domicílios selecionados na amostra, são medidas características sociodemográficas. Além disso, para os moradores que, na data de referência, tinham 10 anos ou mais de idade, são levantadas as informações de educação com o objetivo de classificar a população em idade ativa com o nível de escolaridade alcançado.

As variáveis correspondentes à força de trabalho são investigadas nos indivíduos de 10 anos ou mais de idade na data de referência, pois estes são os considerados em idade ativa e são divididos em três subgrupos dis-

juntos: ocupados, desocupados e não-economicamente ativos. A população ocupada está composta pelas pessoas que exerceram trabalho, remunerado ou sem remuneração, durante pelo menos uma hora completa na semana de referência, ou que tinham trabalho remunerado do qual estavam temporariamente afastadas nessa semana. O subgrupo que corresponde à população desocupada está formado pelas pessoas que não tem trabalho, mas que estavam disponíveis para assumir um trabalho nessa semana e que tomaram alguma providência efetiva para conseguir trabalho no período de referência de 30 dias, sem terem tido qualquer trabalho, ou após terem saído do último trabalho que tiveram nesse período. A população não-economicamente ativa é constituída pelas pessoas em idade ativa que não foram classificadas como ocupadas nem como desocupadas. A divisão da população procura obter, de cada um dos subgrupos um conjunto de informações detalhadas que expliquem a dinâmica do mercado de trabalho.

Para as pessoas que declaram ter trabalho remunerado e que não foi exercido, durante pelo menos uma hora completa na semana de referência, investigam-se o motivo por não ter exercido o trabalho e o tempo em que esteve afastada do trabalho que tinha. Esta investigação é feita com a finalidade de determinar a classificação da pessoa como ocupada ou não. No caso dos ocupados, identificam-se quantos trabalhos possuem na semana de referência e qual é o principal. É coletada também a informação referente à remuneração mensal habitual, a remuneração efetivamente recebida no mês de referência, as horas habitualmente trabalhadas por semana, as horas efetivamente trabalhadas na semana de referência e a contribuição para instituto de previdência para o trabalho principal e para os outros trabalhos exercidos na semana de referência. Os classificados como ocupados são desagregados em quatro categorias de posição na ocupação: empregados, conta-própria, empregadores e trabalhadores não-remunerados de membros da unidade domiciliar que era conta-própria ou empregador, mostrando, de forma clara, as relações de trabalho.

Para os indivíduos sem trabalho na semana de referência é investigado se já tiveram um trabalho antes dessa semana. Para as que tiveram, investiga-se o tempo decorrido desde a saída do último trabalho, além de uma série de

perguntas que procuram fornecer uma descrição geral deste grupo.

7.1.3 Plano Amostral

O cadastro usado para a seleção da amostra está baseado na divisão política administrativa do 2000, que compõe as regiões metropolitanas de abrangência da pesquisa. Os setores da amostra são atualizados anualmente com a operação de atualização da listagem, a qual possibilita a localização e identificação das unidades domiciliares que pertencem a cada setor.
O plano amostral empregado na PME seleciona domicílios (conglomerados) dentro dos quais são levantadas as informações de todas as pessoas que pertencem a ele. Assim, adouto-se uma amostra estratificada e conglomerada em dois estágios, para cada região metropolitana considerada na pesquisa. A população de interesse nesta pesquisa corresponde aos municípios e pseudomunicípios (conjuntos de municípios de menor porte em quantidade de domicílios segundo o Censo Demográfico 2000) das regiões metropolitanas, sendo cada um, um estrato independente de seleção, o que garante o espalhamento da amostra pela região metropolitana.
Em cada município ou pseudo-município são selecionadas as unidades primárias de amostragem (UPA's) e posteriormente as unidades secundárias de amostragem (USA's). Sendo as unidades primárias de amostragem da pesquisa os setores censitários, enquanto as unidades secundárias de amostragem são as unidades domiciliares.
A seleção dos setores (UPA's) é feita usando amostragem sistemática com probabilidade proporcional ao total de domicílios particulares ocupados, obtido pelo Censo Demográfico 2000. Após da seleção dos setores, e baseado na listagem atualizada de unidades domiciliares nestes setores, é feita a seleção de domicílios por meio de amostragem sistemática simples. Assim, a seleção das unidades domiciliares que pertencem à amostra é feita partindo de intervalos de seleção fixos por setor e estabelecidos considerando-se 16 unidades domiciliares por setor.

Uma vez que a amostra é selecionada procede-se à coleta dos dados. As entrevistas são feitas pessoalmente com auxílio de um Pocket PC. O entre-

vistador primeiro obtém a informação sociodemográfica de cada membro do domicílio e, em seguida, as informações sobre educação e trabalho para todas as pessoas de 10 anos ou mais de idade. O ideal seria que cada morador respondesse sobre si mesmo, entretanto, o número de vezes que o entrevistador teria que retornar ao domicílio reverteria em aumento do tempo para obter a entrevista completa e em alto custo. A alternativa utilizada é obter as informações através de um dos moradores apto a prestá-las

É usado um esquema de rotação da amostra, onde a amostra de unidades domiciliares da pesquisa é distribuída pelas quatro semanas de referência do mês. Assim, o resultado do mês é obtido pela média dessas quatro semanas de referência. Na PME cada unidade domiciliar selecionada fica quatro meses consecutivos sendo pesquisada, oito meses sem ser pesquisada e, após este período, é pesquisada novamente por mais quatro meses, e finalmente excluída da amostra. Cabe ressaltar que, se durante o período (12 meses) em que a unidade domiciliar permanece na amostra, a família mudar de endereço e outra família passar a ocupar a unidade domiciliar, a informação será obtida com a nova família pelo período restante.

O esquema de rotação da amostra representa a tentativa de obter ganhos em variâncias de estimativas, de mudanças tanto mês a mês como ano a ano, além de proporcionar um maior grau de precisão na comparabilidade de seus resultados.

7.2 Ilustração do Uso dos Estimadores de Regressão Generalizados

Nesta seção os dados da PME, são usados com o objetivo de mostrar como podem ser aplicados os estimadores de regressão generalizados na estimação de parâmetros de interesse. Para obter a estimativa de um parâmetro de interesse usando o estimador de regressão generalizado logístico, é necessário conhecer os valores das variáveis auxiliares $\mathbf{x}_k$ para todos os elementos que compõem a população em estudo. Por esta razão, é considerada uma amostra do banco de dados da PME, ou seja, este exercício vai ser desenvolvido usando uma subamostra da amostra obtida pelo IBGE, para usar como

variáveis auxiliares da população aquelas que pertencem à amostra, sendo a variável de interesse considerada apenas na subamostra. No banco de dados disponível, tem-se os dados correspondentes à PME do mês de outubro de 2005 das 6 regiões metropolitanas que abrange a pesquisa. Neste exemplo será considerada só a Região Metropolitana de Recife. O objetivo é estimar a taxa de desemprego do mês de outubro do ano 2005, usando o estimador Horvitz-Thompson, o estimador de regressão generalizado (GREG) e o estimador de regressão generalizado logístico (LGREG). Na Quadro 7.1 são apresentadas as variáveis usadas neste exercício.

Quadro 7.1. Variáveis usadas na estimação da taxa de desemprego.

Variável	Código	Descrição
VD2	1	Pessoas desocupadas na semana de referência com procura de trabalho no período de referência de 30 dias que já trabalharam anteriormente.
	2	Pessoas desocupadas na semana de referência com procura de trabalho no período de referência de 30 dias que nunca trabalharam.
VD3	1	Pessoas Economicamente Ativas na semana de referência
SEXO	1	Homem
	2	Mulher
IDADE	Numérico	Idade calculada

Com estas variáveis pode ser definida a proporção a ser estimada e as variáveis auxiliares que serão usadas para o modelo que vai a assistir à estimação. A variável de interesse y é dada por

$$y_k = \begin{cases} 1, & \text{se VD2=1,2;} \\ 0, & \text{se VD3=1,} \end{cases}$$

e as variáveis auxiliares são o sexo e a idade. O tamanho da população (amostra) é $N = 6067$, para selecionar a subamostra serão considerados dois planos amostrais AAS e AE, onde será avaliado o desempenho dos estimadores $\hat{P}_{_{HT}}$, $\hat{P}_{_{GREG}}$ e $\hat{P}_{_{LGREG}}$. Os tamanhos de amostra considerados estão

dados pelas frações amostrais $f_1 = 5\%$, $f_2 = 10\%$ e $f_3 = 20\%$, obtendo assim, $n = 303, 606, 1213$. O valor da taxa de desemprego na população de interesse, ou seja, o valor do parâmetro de interesse foi $P = 0.14735$

7.2.1 Amostragem Aleatória Simples

Para as amostras selecionadas sob este plano foram calculadas as estimativas da taxa de desemprego P, as variâncias dos estimadores e os estimadores das variâncias dos estimadores, usando as expressões apresentadas a seguir:

- Estimador de Horvitz-Thompson $\hat{P}_{HT}$. Foi usada a expressão dada em (5.6). Para calcular a variância do estimador de Horvitz-Thompson foi usada e seguinte expressão

$$V(\hat{P}_{HT}) = \frac{(N-n)P(1-P)}{n(N-1)}.$$

  O estimador de $V(\hat{P}_{HT})$ utilizado foi o da expressão dada em (6.8).

- Estimador GREG $\hat{P}_{GREG}$. Este estimador foi calculado usando a seguinte expressão

$$\hat{P}_{GREG} = \frac{\hat{t}_{GREG}}{N},$$

  em que $\hat{t}_{GREG}$ é dado na expressão (3.9), e o modelo formulado entre a variável de interesse e as variáveis auxiliares pode ser escrito como

$$\begin{cases} E(Y_k) = \alpha + \beta_1 x_{1k} + \beta_2 x_{2k}; \\ V(Y_k) = \sigma^2. \end{cases}$$

  A variância do estimador foi calculada usando (5.7) e o estimador da variância de $\hat{P}_{GREG}$ é expresso como em (5.9) onde $\hat{\boldsymbol{\beta}}_S^{\pi}$ é estimado usando a expressão (2.9).

- Estimador LGREG $\hat{P}_{LGREG}$. A expressão para calcular a estimativa deste estimador foi apresentada em (5.11), onde o modelo formulado entre a variável de interesse e as variáveis auxiliares foi dado em (5.4) considerando intercepto. A variância do estimador é expressa como em (5.7) e o estimador da variância do estimador foi calculado usando a expressão apresentada em (5.14).

Como foi mencionado anteriormente a amostra da PME obtida pelo IBGE foi tratada como a população. Neste contexto é possível obter o valor do parâmetro de interesse e as variâncias dos estimadores. Os resultados obtidos são apresentados nos Quadros 7.2 e 7.3.

No Quadro 7.2 podem-se observar as estimativas de P, obtidas usando os estimadores $\hat{P}_{HT}$, $\hat{P}_{GREG}$ e $\hat{P}_{LGREG}$, as suas variâncias e os valores correspondentes ao estimador da variância do estimador $\hat{P}$, para os diferentes tamanhos de amostra selecionados. Com estas quantidades, foram obtidos intervalos de confiança de 95% para cada estimador. Nos intervalos de confiança pode ser observado que em todos os casos este contém o verdadeiro valor do parâmetro P.
No Quadro 7.3 apresenta-se a eficiência dos estimadores $\hat{P}_{GREG}$ e $\hat{P}_{LGREG}$ com respeito ao estimador $\hat{P}_{HT}$, esta medida é calculada usando a expressão (2.5). Os estimadores $\hat{P}_{GREG}$ e $\hat{P}_{LGREG}$ são em todos os casos mais eficientes do que o estimador de Horvitz-Thompson, ou seja, $eff(\hat{P}) < 1$. Embora, estes valores sejam próximos, o estimador $\hat{P}_{LGREG}$ é o mais eficiente em todos os casos.

Quadro 7.2. Estimativas de P, do estimador da variância e IC 95% usando AAS. ($P = 0.14735$)

n	Estimador	$\hat{P}$	$V(\hat{P})$	$\hat{V}(\hat{P})$	IC 95%	
					LI	LS
303	$\hat{P}_{HT}$	0.14802	0.000056000	0.000056000	0.133352	0.162687
	$\hat{P}_{GREG}$	0.14487	0.000053962	0.000055082	0.130323	0.159416
	$\hat{P}_{LGREG}$	0.14427	0.000053682	0.000055021	0.129731	0.158808
606	$\hat{P}_{HT}$	0.14662	0.000053000	0.000051000	0.132622	0.160617
	$\hat{P}_{GREG}$	0.14318	0.000051125	0.000050279	0.129282	0.157077
	$\hat{P}_{LGREG}$	0.13301	0.000050860	0.000050233	0.119118	0.146901
1213	$\hat{P}_{HT}$	0.14991	0.000047000	0.000045000	0.136761	0.163058
	$\hat{P}_{GREG}$	0.15158	0.000045433	0.000043843	0.138602	0.164557
	$\hat{P}_{LGREG}$	0.14089	0.000045197	0.000043652	0.127940	0.153839

Quadro 7.3. Eficiência do estimador P usando AAS.

n	Estimador	
	$\hat{P}_{GREG}$	$\hat{P}_{LGREG}$
303	0.963600	0.958607
606	0.964622	0.959622
1213	0.966653	0.961638

7.2.2 Amostragem Estratificada

Para ilustrar o uso deste plano foram definidos 3 estratos, dividindo a população de interesse como é mostrado no Quadro 7.4, em que Outros corresponde a os outros municípios que compõem a Região Metropolitana de Recife.

Quadro 7.4. Estratos usados no plano AE.

Estrato	Cod. Município	Tamanho
1	260011	2799
2	260006-260008	1658
3	Outros	1610

Os tamanhos das amostras em cada estrato são proporcionais ao tamanho do estrato e a sua seleção foi feita seguindo um AAS. Neste contexto, é possível calcular as versões separadas e combinadas dos estimadores $\hat{P}_{GREG}$ e $\hat{P}_{LGREG}$. As expressões que adotam os estimadores $\hat{P}_{HT}$, $\hat{P}_{GREG}$ e $\hat{P}_{LGREG}$, suas variâncias e os estimadores das variâncias sob um plano AE foram apresentadas na seção 6.1.3. Usando essas expressões, foram obtidas as estimativas do parâmetro de interesse, as variâncias e os estimadores das variâncias dos estimadores. Os resultados são apresentados nos Quadros 7.5 e 7.6

No Quadro 7.5 encontram-se as estimativas da taxa de desemprego P, obtidas usando o estimador $\hat{P}_{HT}$ e os estimadores $\hat{P}_{GREG}$ e $\hat{P}_{LGREG}$ em suas versões combinada e separada. Além disso, são apresentadas as variâncias e os estimadores da variância, para os tamanhos de amostra considerados e intervalos de confiança de 95% para P. Pode ser observado que a amplitude dos

intervalos de confiança vai diminuindo na medida que aumenta o tamanho da amostra.

No Quadro 7.6 pode ser observada a eficiência dos estimadores $\hat{P}_{_{GREG}}$ e $\hat{P}_{_{LGREG}}$ combinados e separados, com respeito ao estimador $\hat{P}_{_{HT}}$, esta medida é calculada usando a expressão (2.5). A eficiência dos estimadores aumenta na medida que aumenta o tamanho da amostra, ainda que, os valores para os estimadores combinados e separados sejam muito parecidos, observa-se uma diferença muito pequena que favorece os estimadores de tipo separados. Esta semelhança pode ser explicada pelo fato de, o modelo que assiste à estimação considera a mesma relação entre a variável de interesse e as variáveis auxiliares em cada estrato. Neste cenário, é recomendável usar os estimadores combinados, pois a obtenção dos estimadores de tipo separado requer maior numero de calculo e o ganho na eficiência ao usá-lo é quase desprezível.

Em geral, os estimadores $\hat{P}_{_{GREG}}$ e $\hat{P}_{_{LGREG}}$ apresentam-se mais eficientes usando um plano estratificado do que usando um plano AAS. Sendo os mais eficientes os de tipo separado. Em todos os casos o estimador $\hat{P}_{_{LGREG}}$ é o mais eficiente.

Quadro 7.5. Estimativas de P, do estimador da variância e IC 95% usando AE ($P = 0.14735$).

n	Estimador	$\hat{P}$	$V(\hat{P})$	$\hat{V}(\hat{P})$	IC 95%	
					LI	LS
303	$\hat{P}_{HT}$	0.14175	0.000393	0.00038	0.10354	0.17995
	$\hat{P}_{GREGC}$	0.13945	0.000371	0.00037	0.10174	0.17715
	$\hat{P}_{LGREGC}$	0.13989	0.000366	0.00037	0.10218	0.17759
	$\hat{P}_{GREGS}$	0.14105	0.000368	0.00036	0.10386	0.17823
	$\hat{P}_{LGREGS}$	0.14057	0.000362	0.00035	0.10390	0.17723
606	$\hat{P}_{HT}$	0.13845	0.000186	0.00019	0.11143	0.16546
	$\hat{P}_{GREGC}$	0.14182	0.000176	0.00018	0.11552	0.16811
	$\hat{P}_{LGREGC}$	0.14132	0.000174	0.00018	0.11502	0.16761
	$\hat{P}_{GREGS}$	0.14136	0.000172	0.00018	0.11506	0.16765
	$\hat{P}_{LGREGS}$	0.14161	0.000170	0.00018	0.11531	0.16790
1213	$\hat{P}_{HT}$	0.15905	0.000083	0.00009	0.14045	0.17764
	$\hat{P}_{GREGC}$	0.15757	0.000078	0.00008	0.14003	0.17510
	$\hat{P}_{LGREGC}$	0.15633	0.000077	0.00008	0.13879	0.17386
	$\hat{P}_{GREGS}$	0.15674	0.000075	0.00008	0.13920	0.17427
	$\hat{P}_{LGREGS}$	0.15622	0.000074	0.00008	0.13868	0.17375

Quadro 7.6. Eficiência do estimador de P usando AE.

n	Estimador			
	Combinado		Separado	
	$\hat{P}_{GREG}$	$\hat{P}_{LGREG}$	$\hat{P}_{GREG}$	$\hat{P}_{LGREG}$
303	0.946020	0.935297	0.936386	0.921119
606	0.944236	0.931483	0.924731	0.913978
1213	0.939759	0.927710	0.903614	0.891566

CAPÍTULO 8

Considerações Finais

Inicialmente, neste trabalho foram apresentados os modelos da família exponencial, como uma opção para assistir ao processo de estimação dos estimadores de tipo regressão, em populações finitas. Estes permitem analisar conjuntos de dados com resposta discreta e contínua, restrita ao intervalo $(0, \infty)$, proporcionando grande flexibilidade para a especificação da relação entre a variável resposta e as variáveis explicativas. A inferência nos modelos da família exponencial está baseada na metodologia de máxima verossimilhança e de míninos quadrados ordinários. Neste trabalho foi considerada uma adaptação quando só está disponível uma amostra da população. Nesse caso, os parâmetros de interesse são estimados usando o método de pseudo máxima-verossimilhança. Baseado nestes modelos, foi proposta uma classe de estimadores de regressão, chamados estimadores de regressão generalizados (GREG). Foi proposto o Lema 1 que sob certas condições permite que o termo de ajuste do estimador GREG desapareça.

Na definição dos modelos da família exponencial para assistir à estimação de parâmetros em populações finitas foi definido ϕ_k, como o parâmetro de dispersão, o que permite especificar modelos heterocedasticos, isto é, quando a variância do modelo que assiste à estimação é não constante, como no caso do *estimador de razão*, e modelos homocedasticos como no caso do estimador de regressão.

Por outro lado, foi discutida a aplicação dos estimadores GREG no contexto de estratificação, propondo formas para o estimador de regressão genera-

lizado combinado (GREGC) e o estimador de regressão generalizado separado (GREGS). Além disso, foi apresentada a abordagem de probabilidades condicionais de inclusão como forma alternativa de derivação do estimador de GREG, quando o modelo que assiste à estimação é linear.

Na segunda parte desta dissertação foi apresentado o estimador de regressão generalizado logístico (LGREG) como uma alternativa para a estimação de proporções na presença de informação auxiliar, onde o processo de estimação é assistido por um modelo de regressão logística entre a variável de interesse e as variáveis auxiliares. Foram apresentados ainda estudos de simulação de Monte Carlo, com o objetivo de avaliar e comparar as propriedades dos estimadores de Horvitz-Thompson, GREG e LGREG quando o parâmetro de interesse é uma proporção, no caso de amostragem aleatória simples, Bernoulli e estratificada. Neste último, foi avaliado o desempenho dos estimadores no contexto da estratifição, considerando os casos separado e combinado.

Os estudos de simulação mostraram em geral que, os estimadores $\hat{P}_{_{GREG}}$ e $\hat{P}_{_{LGREG}}$ apresentam características semelhantes ao estimador $\hat{P}_{_{HT}}$, com respeito ao viés, o viés do estimador da variância e taxas de cobertura. A diferença está na eficiência deles, pois $\hat{P}_{_{GREG}}$ e $\hat{P}_{_{LGREG}}$ apresentam-se mais eficientes do que o estimador $\hat{P}_{_{HT}}$. Embora que, os resultados do estimador $\hat{P}_{_{GREG}}$ e $\hat{P}_{_{GREG}}$ estejam muito próximos, o estimador $\hat{P}_{_{LGREG}}$ é o mais eficiente.
O desempenho dos estimadores de tipo combinado e separado, estão deacordo com o objetivo de sua aplicação, ou seja, quando é suposta a mesma relação entre a variável de interesse e as variáveis auxiliares para toda a população sem fazer diferença entre estratos, é apropriado aplicar os estimadores combinados e quando tem-se que supor relações diferentes em cada estrato entre estas variáveis, os separados são os mais adequados.
É recomendável usar como estimador da variância de $\hat{P}_{_{LGREG}}$, $\hat{V}_2(\hat{P}_{_{LGREG}})$, pois este estimador possue coeficientes de variação mais pequenos do que $\hat{V}_1(\hat{P}_{_{LGREG}})$ e seu calculo é mais fácil de obter.

Uma das principais diferenças entre o estimador de regressão generalizado e

o estimador de regressão generalizado logístico está na sua aplicação, pois, para o estimador GREG, só é necessário conhecer os valores das variáveis auxiliares $\mathbf{x}_k$ para os elementos da amostra, e o total populacional delas. Entretanto, para a aplicação do LGREG é necessário conhecer $\mathbf{x}_k$ para todos os elementos na população.

Este trabalho contribui em alguma medida na tarefa de difundir os estimadores de regressão generalizados, mais especificamente, os estimadores de regressão generalizados assistidos por modelos pertencentes à família exponencial. Como trabalhos futuros poderiam-se estudar as características e propriedades dos estimadores GREG, onde o modelo que assiste a estimação use como função de ligação a *probit, log-log, complemento log-log*, entre outras.

APÊNDICE A

Prova do Lema 1

Lema 1. *Se o estimador de regressão generalizado (GREG) descrito na expressão (3.1) considera um modelo de regressão da família exponencial, onde tem-se:*

S1. Homogeneidade no parâmetro de dispersão, ou seja, $\phi_k = \phi$*, para todo* $k \in U$*;*

S2. Componente sistemática com intercepto, ou seja, existe β_j *em* $\boldsymbol{\beta}$ *tal que* $\dfrac{\partial \eta_k}{\partial \beta_j} = C$*,* $\forall k \in U$*, com* C *uma constante;*

S3. Componente sistemática com ligação canônica, ou seja, $\theta_k = \eta_k$*, para todo* $k \in U$*;*

Então, o estimador de regressão generalizado pode ser escrito como:

$$\hat{t}_{_{GREG}} = \sum_{k \in U} \hat{\mu}_k^S = \sum_{k \in U} g^{-1}(h(\hat{\boldsymbol{\beta}}_S^{\pi}; \mathbf{x}_k)). \tag{A.1}$$

Além disso, o total de y *pode ser expresso da seguinte forma*

$$t_y = \sum_{k \in U} \hat{\mu}_k^U = \sum_{k \in U} g^{-1}(h(\hat{\boldsymbol{\beta}}_U; \mathbf{x}_k)). \tag{A.2}$$

Prova: A prova deste Lema está divida em duas partes: a primeira para provar a expressão dada em (A.1) e a segunda, para provar a expressão dada (A.2).

Primera parte: O estimador de pseudo máxima-verossimilhança de $\boldsymbol{\beta}$ pode ser escrito como:

$$\hat{\boldsymbol{\beta}}_S^{\pi} \;=\; \arg\max_{\boldsymbol{\beta}} L_S(\boldsymbol{\beta}),$$

com

$$L_S(\boldsymbol{\beta}) = \sum_{k\in S} \frac{\phi\{y_k\theta_k - b(\theta_k)\}}{\pi_k}, \qquad \text{pela suposição S1.}$$

O estimador $\hat{\beta}_S^{\pi}$ verifica a restrição $\mathbf{U}_S(\boldsymbol{\beta}) = \mathbf{0}$, em que a função $\mathbf{U}_S(\boldsymbol{\beta}) = \dfrac{\partial L_S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ e chamada de função de escore.

Dado que a componente sistemática do modelo formulado tem intercepto, pode-se escrever o j-ésimo componente de $\mathbf{U}_S(\boldsymbol{\beta})$ na seguinte forma:

$$\begin{aligned}
\mathbf{U}_S^j(\boldsymbol{\beta}) &= \frac{\partial L_S(\boldsymbol{\beta})}{\partial \beta_j} \\
&= \sum_{k\in S} \frac{\phi}{\pi_k}\left\{ y_k \frac{\partial\theta_k}{\partial\eta_k}\frac{\partial\eta_k}{\partial\beta_j} - b^{'}(\theta_k)\frac{\partial\theta_k}{\partial\eta_k}\frac{\partial\eta_k}{\partial\beta_j}\right\} \qquad \text{(S3.)} \\
&= \phi\sum_{k\in S} \frac{1}{\pi_k}\left\{ y_k \frac{\partial\eta_k}{\partial\beta_j} - b^{'}(\theta_k)\frac{\partial\eta_k}{\partial\beta_j}\right\} \qquad \text{(S2.)} \\
&= \phi\, C \sum_{k\in S} \frac{1}{\pi_k}\{y_k - \mu_k\}, \qquad \text{com } g(\mu_k) = h(\boldsymbol{\beta}; \mathbf{x}_k).
\end{aligned}$$

Então, dado que $\mathbf{U}_S^j(\hat{\boldsymbol{\beta}}_S^{\pi}) = 0$, pode-se escrever:

$$\mathbf{U}_S^j(\hat{\boldsymbol{\beta}}_S^{\pi}) = \phi\, C \sum_{k\in S} \frac{(y_k - \hat{\mu}_k^S)}{\pi_k} = 0,$$

o que implica que

$$\sum_{k\in S} \frac{(y_k - \hat{\mu}_k^S)}{\pi_k} = 0,$$

provando assim, o resultado em (A.1).

Segunda Parte: O estimador de máxima-verossimilhança de $\boldsymbol{\beta}$ pode ser escrito como:

$$\hat{\boldsymbol{\beta}}_U \;=\; \arg\max_{\boldsymbol{\beta}} L_U(\boldsymbol{\beta}),$$

com

$$L_U(\boldsymbol{\beta}) = \sum_{k\in U} \phi\{y_k\theta_k - b(\theta_k)\}, \qquad \text{pela suposição S1.}$$

O estimador $\hat{\beta}_U$ verifica a restrição $\mathbf{U}_U(\boldsymbol{\beta}) = \mathbf{0}$, em que a função $\mathbf{U}_U(\boldsymbol{\beta}) = \dfrac{\partial L_U(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ é chamada de função de escore.

Dado que a componente sistemática do modelo formulado tem intercepto, pode-se escrever o j-ésimo componente de $\mathbf{U}_U(\boldsymbol{\beta})$ na seguinte forma:

$$\begin{aligned}
\mathbf{U}_U^j(\boldsymbol{\beta}) &= \frac{\partial L_U(\boldsymbol{\beta})}{\partial \beta_j} \\
&= \sum_{k\in U} \phi \left\{ y_k \frac{\partial \theta_k}{\partial \eta_k}\frac{\partial \eta_k}{\partial \beta_j} - b'(\theta_k)\frac{\partial \theta_k}{\partial \eta_k}\frac{\partial \eta_k}{\partial \beta_j} \right\} \qquad \text{(S3.)} \\
&= \phi \sum_{k\in U} \left\{ y_k \frac{\partial \eta_k}{\partial \beta_j} - b'(\theta_k)\frac{\partial \eta_k}{\partial \beta_j} \right\} \qquad \text{(S2.)} \\
&= \phi\, C \sum_{k\in U} \{y_k - \mu_k\}, \qquad \text{com } g(\mu_k) = h(\boldsymbol{\beta}; \mathbf{x}_k).
\end{aligned}$$

Então, dado que $\mathbf{U}_U^j(\hat{\boldsymbol{\beta}}_U) = 0$ pode-se escrever:

$$\mathbf{U}_U^j(\hat{\boldsymbol{\beta}}_U) = \phi\, C \sum_{k\in U} (y_k - \hat{\mu}_k^U) = 0,$$

o que implica que

$$\sum_{k\in U} (y_k - \hat{\mu}_k^U) = 0,$$

provando assim, o resultado em (A.2). □

APÊNDICE B

Prova do Resultado 1

Os seguintes lemas são usados na prova do resultado 1:

Lema 2. *Se $\gamma_k = V_x^{-1/2}(\bar{x}_{|k} - \bar{x})$, então $\gamma_k = O(n^{-1/2})$, onde $n^{\alpha}O(n^{-\alpha})$, $\alpha < 0$, é uma quantidade que permanece limitada quando $n \to \infty$.*

Prova: Das expressões (4.5) e (4.6) tem-se que

$$\begin{aligned}\gamma_k &= \left(\frac{N-n}{N-1}\frac{\sigma_x^2}{n}\right)^{-1/2}\frac{N-n}{N-1}\frac{x_k-\bar{x}}{n} \\ &= \left(\frac{N-n}{n(N-1)}\right)^{1/2}\frac{x_k-\bar{x}}{\sigma_x} = O(n^{-1/2}).\end{aligned}$$

□

Lema 3.

$$\frac{V_{x|k}}{V_x} = 1 + O\left(\frac{1}{n}\right)$$

Prova: Das equações (4.6) e (4.7) pode ser mostrado que

$$\begin{aligned}\frac{V_{x|k}}{V_x} &= \frac{N(n-1)}{(N-2)n}\left\{1 - \frac{(x_k-\bar{x})^2}{(N-1)\sigma_x^2}\right\} \\ &= 1 - \frac{1}{n}\left\{\frac{N-2n}{(N-2)} + \frac{N(n-1)}{(N-2)}\frac{(x_k-\bar{x})^2}{(N-1)\sigma_x^2}\right\} \\ &= 1 + O\left(\frac{1}{n}\right)\end{aligned}$$

□

Lema 4. *Se* $V_{x|k}V_x^{-1} - 1 = O(n^{-1})$*, então* $V_{x|k}V_x^{-1} = 1 + O(n^{-1})$.

Agora, é definido $\hat{\tilde{x}}_c = V_x^{-1/2}(\hat{\tilde{x}} - \bar{x})$, $\gamma_x = V_x^{-1/2}(x_{|k} - \bar{x})$, e

$$c_k = \left(\frac{V_{x|k}}{V_x}\right)^{1/2} \exp\left[\frac{\hat{\tilde{x}}_c^2}{2}\left(\frac{V_x}{V_{x|k}} - 1\right)\right], \tag{B.1}$$

então, a expressão (4.8) pode ser adotar a seguinte forma

$$a_k(\hat{\tilde{x}}) = \left(\frac{V_{x|k}}{V_x}\right)^{1/2} \exp\left[-\frac{\hat{\tilde{x}}_c^2}{2} + \frac{(\hat{\tilde{x}}_c - \gamma_k)^2 V_x}{2V_{x|k}}\right] = c_k \exp\left[\gamma_k \frac{V_x}{2V_{x|k}}(\gamma_k - 2\hat{\tilde{x}}_c)\right]$$

$$a_k(\hat{\tilde{x}}) = c_k\left(1 - \gamma_k \frac{V_x}{V_{x|k}}\hat{\tilde{x}}_c\right) + \mathcal{R}(\gamma_k^{(0)}), \tag{B.2}$$

em que

$$\mathcal{R}(\gamma_k^{(0)}) = c_k \exp\left\{\gamma_k^{(0)}\frac{V_x}{2V_{x|k}}(\gamma_k^{(0)} - 2\hat{\tilde{x}}_c)\right\}\gamma_k^2\frac{V_x}{V_{x|k}}\left\{\frac{V_x}{V_{x|k}}(\gamma_k^{(0)} - \hat{\tilde{x}}_c)^2 - 1\right\},$$

onde $\gamma_k^{(0)}$ é um vetor cujos elementos são incluidos entre os elementos correspondentes a γ_k e 0. Usando o Lema 2 tem-se que $\gamma_k = O(n^{-1})$ e $\mathcal{R}(\gamma_k^{(0)}) = O_p(n^{-1})$. Além disso, usando os resultados dos Lemas 3 e 4 na expressão (B.1) tem-se que

$$c_k = [1 + O(n^{-1})]^{-1/2} \exp\left\{\frac{\hat{\tilde{x}}_c^2}{2}O(n^{-1})\right\} = 1 + O_p(n^{-1}). \tag{B.3}$$

Substituindo a expressão (B.3) na (B.2)

$$\begin{aligned} a_k(\hat{\tilde{x}}) &= (1 + O_p^{-1})\left\{1 - \gamma_k(1 + O(n^{-1}))\hat{\tilde{x}}_c + O_p(n^{-1})\right\} \\ &= \left\{1 - \gamma_k\hat{\tilde{x}}_c + O_p(n^{-1})\right\} \\ &= \left\{1 + (\bar{x} - \hat{\tilde{x}})V_x^{-1}(\bar{x}_{|k} - \bar{x} + O_p(n^{-1}))\right\}. \end{aligned}$$

Finalmente

$$\begin{aligned} \hat{\tilde{y}}_{|\hat{\tilde{x}}} &= \frac{1}{n}\sum_{k\in S} a_k(\hat{\tilde{x}})y_k \\ &= \hat{\tilde{y}}_\pi + (\bar{x} - \hat{\tilde{x}})V_x^{-1}\frac{1}{n}\sum_{k\in S}(\bar{x}_{|k} - \bar{x})y_k + O_p(n^{-1}) \\ &= \hat{\tilde{y}}_\pi + (\bar{x} - \hat{\tilde{x}})\frac{1}{n\sigma_x^2}\sum_{k\in S}(x_k - \bar{x})y_k + O_p(n^{-1}). \end{aligned}$$

Provando assim o resultado em (4.9).

APÊNDICE C

Obtenção de β_0

Neste apêndice é apresentada a obtenção da expressão dada na equação (6.2), através da qual é possível simular uma população cuja probabilidade de sucesso seja P.

Prova: Para gerar uma população tem-se

$$\log\left[\frac{\pi(x_k)}{1-\pi(x_k)}\right] = \beta_0 + \beta_1 x_k, \qquad k \in U,$$

em que X_k e Y_k são variáveis aleatórias, y_k uma realização de Y_k e

$$Y_k = \begin{cases} 1, & \text{com probabilidade } \pi(x_k); \\ 0, & \text{com probabilidade } 1-\pi(x_k). \end{cases}$$

Uma população com a proporção de sucessos P deve satisfazer:

$$\frac{1}{N}\sum_{k\in U} y_k = P,$$

assim, a restrição é dada por

$$E\left(\frac{1}{N}\sum_{k\in U} Y_k\right) = \frac{1}{N}E\left(\sum_{k\in U} Y_k\right) = \frac{1}{N}E_x\left(E\left(Y_k|X_k\right)\right) = P,$$

onde $E_x(\cdot)$ é a esperança sob X e $E(Y_k|X_k = x_k)$ é a esperança de Y_k dado $X_k = x_k$. Tem-se que

$$Y_k|X_k = x_k \ \sim \ \text{Bernoulli}(\pi(\text{x}_\text{k})).$$

Daí, tem-se que

$$\frac{1}{N}E\left(\sum_{k\in U}Y_k\right)=\frac{1}{N}E_x\left(\pi(X_k)\right)=P. \tag{C.1}$$

Neste caso, $\pi(x)$ pode ser aproximada usando a expansão de Taylor de segunda ordem dada por

$$\pi(x)=\pi(x_0)+\pi'(x_0)(x-x_0)+\frac{1}{2}\pi''(x_0)(x-x_0)^2+e.$$

Usando que $x_0=\mu_x=E(X)$ e supondo que $\mu_x=0$ tem-se que

$$\pi(x)\approx\pi(0)+\pi'(0)(x)+\frac{1}{2}\pi''(0)(x)^2+e. \tag{C.2}$$

Entretanto, por simplicidade, podem ser considerados somente os dois primeiros termos de (C.2)

$$\pi(x)\approx\pi(0)+\pi'(0)(x). \tag{C.3}$$

Aplicando a esperança na expressão em (C.3), tem-se que

$$\begin{aligned}E(\pi(x))\approx\pi(0)+\pi'(0)E(x)&=\pi(0)\\&=\frac{\exp(\beta_0+\beta_1(0))}{1+\exp(\beta_0+\beta_1(0))}\\&=\frac{\exp(\beta_0)}{1+\exp(\beta_0)}.\end{aligned}$$

Então da equação (C.1)

$$\frac{1}{N}E_x\left(\pi(X_k)\right)=\frac{1}{N}\sum_{k\in U}\left[\frac{\exp(\beta_0)}{1+\exp(\beta_0)}\right]=P,$$

daí

$$\left[\frac{\exp(\beta_0)}{1+\exp(\beta_0)}\right]=P,$$

o que implica que

$$\beta_0=\log\left(\frac{P}{1-P}\right).$$

□

APÊNDICE D

Uso do computador

Os métodos usados na coleta dos dados de pesquisas que usam a teoria de amostragem introduzem uma complexidade na análise, que deve ser considerada na estimação dos parâmetros de interesse para obter uma maior precisão. Porém, esta complexidade nos dados não é considerada pelas técnicas de análise estatística disponíveis nos pacotes computacionais estatísticos padrões de uso generalizado, o que leva ao investigador a conclusões incorretas do problema de estudo. Por esta razão, as vezes é necessário usar pacotes estatísticos especializados para a análise de dados de pesquisas amostrais complexas.

Neste capítulo pretende-se mostrar como utilizar o pacote computacional SAS que traz consigo rotinas especiais para a incorporação do esquema amostral na análise de dados. Em particular, foi estudado o procedimento *Survey-logistic* do pacote SAS versão 9.1, no módulo SAS/STAT. Para esta finalidade foi usado como guia o artigo de An (2002) e o manual do SAS disponível na página http://support.sas.com/onlinedoc/913/docMainpage.jsp.

D.1 SAS

O SAS é o pacote estatístico mais utilizado pelas grandes corporações, em mais de 100 países diferentes. O seu nome nasceu como um acrônimo: Statistical Analysis System (SAS), mas a quantidade de serviços e produtos

oferecidos pela SAS (a companhia que produz o SAS) foi se tornando tão diversa que hoje em dia o nome é simplesmente SAS. A SAS é a maior empresa de software do mundo de capital privado. O SAS é um software integrado para análise de dados que consiste de vários produtos, que proporcionam: recuperação de dados, gerenciamento de arquivos, análise estatística, acesso a banco de dados (ORACLE, DB2, etc), geração de gráficos, geração de relatórios, geração de aplicativos e soluções de negócios. Além disso, o SAS é um software de grande portabilidade, podendo operar em diversos ambientes computacionais.

As origens do software datam da década de 70, quando os computadores ainda eram operados por cartões perfurados (o comando CARDS, dentro do passo DATA, vem justamente daí) e o poder de processamento era muito baixo. O software esta composto por diversos módulos, que provém soluções para problemas específicos, estes módulos são:

- SAS/Base - Sistema básico do SAS, necessário para rodar qualquer outro produto SAS. Ele contém o passo DATA, para manipulação de dados e alguns procedimentos estatísticos simples.
- SAS/STAT - Módulo que fornece uma grande quantidade de métodos estatísticos, como regressão, ANOVA, análise multivariada, análise de sobrevivência entre outros.
- SAS/GRAPH - Módulo para fazer gráficos em alta resolução.
- SAS/ETS - Módulo de econometria, (análise de séries temporais).
- SAS/ASSIST - Módulo de programação interativa e amigável.
- SAS/ACCESS - Módulo para acesso aos diversos tipos de banco de dados.
- SAS/AF - Módulo para desenvolvimento de aplicações.
- SAS/CONNECT - Módulo para conexão entre ambientes operacionais heterogêneos.

- SAS/FSP - Módulo para agilizar o acesso a arquivos.
- SAS/IML - Módulo para análise e operação de matrizes.
- SAS/OR - Módulo de análise e pesquisa operacional (Programação linear, Análise de caminho crítico).
- SAS/QC - Módulo para análise de controle de qualidade.
- SAS/EG (ou Enterprise Guide) - Interface gráfica para o SAS, permitindo fazer algumas análises estatísticas apontando e clicando.

A funcionalidade do Sistema SAS foi construída em torno de quatro idéias básicas no tratamento de dados: acessar, administrar, analisar e apresentar dados.

D.1.1 PROC SURVEYLOGISTIC

No SAS são vários os procedimentos para a análise de dados amostrais. O *Surveyselect* que é usado para selecionar uma amostra probabilística pelo meio de diferentes planos amostrais, incluindo os planos estratificados e proporcional ao tamanho. O *Surveymeans* calcula as estatísticas descritivas para dados de levantamentos amostrais, tais como totais, médias, razões, e dominios. O *Surveyreg* é usado para ajustar modelos de regressão lineares, obter testes de hipóteses e estimativas dos parâmetros de interesse para dados amostrais.
O *Surveylogistic* está baseado no procedimento *Logistic* que calcula as estatísticas sob o suposto que os dados a serem analisados (amostra) foram coletados de uma população infinita usando amostragem aleatória simples. Entretanto, muitos dos estudos são coletados de populações finitas baseados em planos amostrais complexos. Com o objetivo de fazer inferências válidas para este tipo de estudos, o plano amostral deve ser incorporado nas análises, então é implementado o procedimento *Surveylogistic* baseado no *Logistic* como resposta a esta necessidade.

No *Surveylogistic* para o cálculo das estimativas de máxima verossimilhança dos coeficientes de regressão para o modelo de regressão logística são usa-

dos métodos iterativos como, escoring de Fisher e Newton Raphson. Para o cálculo da variância é usada a técnica de linearização de taylor incorporando a informação do plano amostral. O *Surveylogistic* fornece as variâncias para cada estrato e depois faz a soma delas. Um ajuste proposto por Morel (1989) é usado na estimação da variância com a finalidade de reduzir o viés quando o tamanho da amostra é pequeno. Além disso, o fator de correção da população finita é incluido na estimação da variância se a amostra foi selecionada sem reemplazamento.

Estrutura do *Surveylogistic*

PROC SURVEYLOGISTIC $<$ opções $>$;
BY variáveis;
CLASS variáveis $<$(v-opções)$>$
$<$variáveis $<$(v-opções)$> \ldots ><$ / v-opções$>$;
CLUSTER variáveis;
CONTRAST 'rótulo' efeito valor $<, \ldots$ efeito$><$ /opções $>$;
FREQ variável;
MODEL events/trials $=<$ efeitos $><$ / opções $>$;
MODEL variável $<$ (variável_ opções) $>=<$ efeitos $><$ / opções $>$;
STRATA variáveis $<$ / opções $>$;
$<$ rótulo: $>$ **TEST** equação1 $<, \ldots, <$ equaçãok $>><$ /opção$>$;
UNITS independente1=lista1$< \ldots$independente$k$=lista$k$ $><$ /opção$>$;
WEIGHT variável $<$ / opção $>$;

O PROC SURVEYLOGISTIC e as opções da declaração do MODELO são necessárias. As declarações CLASS, CLUSTER, STRATA e CONTRAST podem aparecer várias vezes, em um mesmo chamado do procedimento. As declarações MODELO e WEIGHT podem ser usadas só uma vez. A declaração CLASS (se ela for usada) deve preceder a do MODELO, e a CONTRAST (se for usada) deve estar depois da declaração do MODELO. A seguir será feita uma breve descrição das opções mais importantes disponíveis para cada uma das declarações listadas acima.

PROC SURVEYLOGISTIC faz o chamado do procedimento *Surveylogistic*, adicionalmente identifica o conjunto de dados de entrada e a classificação dos níveis da resposta.

< opções > DATA= Nome do banco de dados para ser processado.

INEST= Nome do banco de dados que contém as estimativas inicias para os parâmetros do modelo.

R= Fator de correção da população finita.

TOTAL= Especifica o número total de UPA's na população.

BY especifica uma o mais variáveis que possibilitam o agrupamento dos dados. Esta declaração não é obrigatoria, mas quando aparece o procedimento ordena as variáveis do banco de dados segundo esteja especificado nesta declaração.

CLASS nomeia as variáveis de classificação que vão a ser usadas na análise. Esta declaração deve preceder a declaração do MODELO. Para cada uma das variáveis de classificação podem ser especificadas várias opções, incluindo um parênteses depois do nome da variável, se estas opções são globais podem ser especificadas usando /. Entretanto, as opções individuas anulam as globais.

< opções > DESC ou DESCENDING inverte a ordem da variável de classificação.

LPREFIX= n, especifica que, no máximo, os primeiros n caráteres do rótulo da variável sejam usados para criar os rótulos de variáveis dummy.

ORDER= <opção>:<DATA||FORMATTED||FREQ||INTERNAL>, especifica a ordem da escolha para os níveis das variáveis de classificação.

PARAM= <opção>, especifica o método de parametrização usado para a variável de classificação.

REF= <opção>: <nível—palavra chave>, especifica o nível de referência usado quando PARAM=EFFECT ou PARAM=REF.

CLUSTER nomeia as variáveis que identificam os conglomerados, quando é usada amostragem de conglomerados. As combinações das categorias das variáveis em CLUSTER podem definir os conglomerados na amostra. Se houver uma declaração STRATA, os conglomerados são embutidos dentro dos estratos. Se o plano que esta sendo usado é conglomerado em múltiplos estágios, podem então ser identificadas só as unidades primárias de amostragem (UPA's) em CLUSTER. As variáveis especificadas em CLUSTER podem ser de tipo caráter ou numérico.

CONTRAST provê uma estrutura que permite definir contrastes para testes de hipóteses. Com este objetivo pode ser definida pelo usuário uma matriz $\mathbf{L}$, para testar o sistema de hipóteses definido por $\mathbf{L}\boldsymbol{\theta} = 0$, onde $\boldsymbol{\theta}$ é o vetor de parâmetros. Adicionalmente, o CONTRAST permite estimar cada linha, $\mathbf{l}_i^T\boldsymbol{\theta}$, de $\mathbf{L}\boldsymbol{\theta}$ e testar a hipótese $\mathbf{l}_i^T\boldsymbol{\theta} = 0$. Não há um limite para o número de testes que o usuário deseje definir, mas estes têm que preceder a declaração do MODELO. A seguir são aprensentados os parâmetros que podem ser especificados na declaração CONTRAST.

rótulo	permite identificar a hipótese que esta sendo testada.
efeito	identifica um efeito que pertence ao modelo.
valor	os elementos da matriz $\mathbf{L}$ associados com o efeito.

Agora serão listadas as opções disponíveis para a declaração CONTRAST.

$<$ opções $>$	ALPHA= α, nível de confiança.
	E, permite que a matriz $\mathbf{L}$ seja exibida.
	ESTIMATE= $<$opção$>$, permite que cada as hipóteses sejam testadas, neste caso são apresentadas as estimativas pontuais, seu erro padrão, o intervalo de confiança e as estatística de Wald para cada hipótese.

FREQ identifica a variável que contém a freqüência de ocorrência de cada observação. O PROC SURVEYLOGISTIC trata cada observação como se

ela aparecera n vezes, onde n é o valor da variável em FREQ. Quando a declaração FREQ não é especificada, para cada observação é asignada a freqüência de 1.

MODEL nesta declaração deve ser nomeada a variável resposta e as variáveis explicativas, incluindo as covariáveis, efeitos principais, interações e efeitos embutidos. Se fossem omitidas as variáveis explicativas, o procedimento ajusta um modelo só com o intercepto.

Podem ser especificadas duas formas na declaração do MODELO. A primeira, é chamada por ensaio, é aplicável a dados de resposta binária, ordinal e nominal, e é usada quando o banco de dados contém a informação de um só ensaio para cada um dos indivíduos. A segunda forma, chamada por freqüência, restrita ao caso de dados binários.

A seguir serão listadas as opções disponíveis para a variável resposta no modelo, estas opções devem ir especificadas dentro de um parênteses logo da declaração da variável resposta .

< opções > DESDENDING ou DESC, inverte a ordem das categorias da variável resposta.

EVENT= <opção>, especifica a categoria da variável resposta que vai ser usada como sucesso, esta opção não tem nenhum efeito se a variável resposta tiver mais de duas categorias.

ORDER= <opção>:
<DATA‖FORMATTED‖FREQ‖INTERNAL>,
especifica a ordem da escolha para os níveis da variável resposta, quando esta opção não é usada os níveis são ordenados pelos valores da variável.

REF= <opção>, especifica a categoria que vai ser usada como referência para a variável resposta.

Agora, serão apresentadas algumas das opções que podem ser usadas no modelo, estas devem ir especificadas logo da declaração do modelo, usando /.

$<$ opções $>$ LINK= $<$opção$>$, especifica função de ligação, podem ser usadas as seguintes opções, CLOGLOG (complemento log-log), GLOGIT (logit generalizada), LOGIT e PROBIT. A opção por defeito é LINK=LOGIT.

NOINT, ajusta um modelo sem intercepto.

VADJUST= $<$opção$>$, permite escolher o método para ajustar a variância. As disponíveis são DF, MOREL, NONE.

ALPHA= α, nível de confiança para o intervalo.

CLPARM, calcula os intervalos de confiança para os parâmetros do modelo.

CLODDS, calcula os intervalos de confiança para as razões de chances (odds ratios).

CORRB, apresenta a matriz de correlação.

COVB, mostra a matriz de covariâncias.

ITPRINT, apresenta o historial das iterações.

PARMLABEL, mostra os rótulos dos parâmetros.

RSQUARE, apresenta o R^2.

STRATA especifica as variáveis que formam os estratos em uma amostragem estratificada. As combinações dos níveis das variáveis em STRATA definem os estratos na amostra. Se o plano que esta sendo usado é estratificado em múltiplos estágios, podem então ser identificados os estratos no primeiro estágio em STRATA. As variáveis especificadas em STRATA podem ser de tipo caráter ou numérico.

$<$ opções $>$ LIST, apresenta uma tabela com a informação dos estratos, que inclui os valores das variávies em STRATA e a taxa amostral para cada estrato. Além disso, esta tabela apresenta o número de observações e número de conglomerados para cada estrato e a variável de análise.

TEST testa hipóteses lineares sobre os coeficientes de regressão. O teste de Wald é usado para testar as hipóteses nulas conjuntamente ($H_0 : \mathbf{L}\boldsymbol{\theta} =$

c). Quando $c = 0$ é mais apropriado usar a declaração CONTRAST. Cada equação especifica uma hipótese linear (a linha da matriz de **L** e o seu correspondente elemento no vetor de c), se é preciso usar equações múltiplas estas devem ir separadas através de vírgulas.

Para um modelo de resposta binária, o parâmetro correspondente ao intercepto é nomeado INTERCEPT, em um modelo de resposta ordinal, o intercepto do modelo é rotulado INTERCEPT, INTERCEPT2, INTERCEPT3, e assim por diante. Quando não aparece o sinal $=$, a expressão a testar é igualada a 0. O código seguinte ilustra possíveis usos da declaração de TEST:

```
proc surveylogistic;
      model y= a1 a2 a3 a4;
      test1: test intercept + .5 * a2 = 0;
      test2: test intercept + .5 * a2;
      test3: test a1=a2=a3;
      test4: test a1=a2, a2=a3;
run;
```

onde test1 é equivalente a test2, assim como test3 e test4.

UNITS permite especificar unidades de câmbio para as variáveis explicativas contínuas de forma que as razões de chance possam ser estimadas. Uma estimativa da razão de chance correspondente é calculada para cada unidade de câmbio especificada para uma variável explicativa. A declaração UNITS é ignorada para as variáveis que foram especificadas na declaração CLASS. A sintaxe desta declaração é a seguinte **UNITS** independente1 = lista1 $< \ldots$ independente$k$ = lista$k$ $><$ / opção $>$, onde o termo independente faz referência ao nome da variável explicativa e lista representa a lista das unidades de câmbio que são de interesse para a variável, estas unidades são separadas por espaços. Cada unidade de câmbio na lista tem a seguinte forma,

- *número*
- SD ou -SD
- *número**SD

em que *número* é algum número diferente de zero, e SD é o erro padrão da correspondente variável independente.

< opções > DEFAULT= lista, apresenta uma lista das unidades de câmbio para todas as variáveis explicativas que não foram especificadas no UNITS. Se a opção DEFAUL não é usada, o PROC SURVEYLOGISTIC não calcula a razão de chance para as variáveis que não são especificadas no UNITS.

WEIGHT nomeia a variável que contém os pesos amostrais. Esta variável deve ser de tipo numérico. Se a declaração WEIGHT não for especificada, o PROC SURVEYLOGISTIC asigna para todas as observações um peso de 1. Estes pesos devem ser números positivos. Se uma observação tem um peso que é não positivo ou faltante, o procedimento omite aquela observação na análise. Se o usuário especifica mais de uma variável em WEIGHT, o procedimento usa só a primeira e ignora o as outras.

APÊNDICE E

Programas de Simulação

Neste apêndice são apresentados os programas de simulação utilizados neste trabalho, resumidos ao caso em que é selecionada só uma amostra*. Estes programas permitem calcular as medidas usadas na avaliação dos estimadores $\hat{P}_{HT}$, $\hat{P}_{GREG}$ e $\hat{P}_{LGREG}$ apresentada no capítulo 6, sobre os diferentes cenários considerados.

Estes programas foram desenvolvidos no pacote estatístico R em sua versão 2.1.3. Para maoires informações ver http://www.r-project.org.

E.1 Amostragem Aleatória Simples

```
# Programa para a obtenção das medidas para a avaliação dos estimadores sob AAS #
rm(list = ls())
OR  <- 0.1    #Sentido e grau da associaçao entre a variável auxiliar e a variável de interesse
Pr  <- 0.2    #Proporçao populacional com a caracteristica de interesse
N   <- 10000  #Tamanho da populaçao
n1  <- 50     #Tamanho da amostra 1
r   <- 10000  #Numero de replicas

# matrizes que para armazenar os dados #
PHT   <- matrix(0,r,5)
PGREG <- matrix(0,r,5)
LGREG <- matrix(0,r,5)
VPHT  <- matrix(0,r,5)
VPGREG <- matrix(0,r,5)
VLGREG <- matrix(0,r,5)
VLGREG2 <- matrix(0,r,5)
TCPHT  <- matrix(0,r,5)
TCPGREG <- matrix(0,r,5)
TCLGREG <- matrix(0,r,5)
TCLGREG2 <- matrix(0,r,5)
IC  <- matrix(0,4,10)
sec  <- seq(1,N,by=1)
```

*O codigo completo dos programas pode ser obtido entrando en contato com o autor.

```
x  <- rnorm(N)   #Variável auxiliar
b1 <- log(OR)
bo <- log(Pr/(1-Pr))
p  <- exp(bo + b1*x)/(exp(bo + b1*x)+1) #Probabilidade de sucesso
u  <- runif(N)
y  <- ifelse(u<=p,1,0)  # Geração da variável de interesse

#### laço de monte carlo
for(i in 1:r){
#seleção das amostras
s1 <- sample(sec,n1)

#Valores de las variáveis de interesse e auxiliar na amostra
y_s1  <- y[s1]
x_s1  <- x[s1]

#Ajustando a regressão logística
templ1 <- glm(y_s1~x_s1, family=binomial)

#estimativas dos parâmetros para o modelo de regressão logística
beta_hat1  <- templ1$coeff
p_u1  <- exp(beta_hat1[1] + beta_hat1[2]*x)/(exp(beta_hat1[1] + beta_hat1[2]*x) + 1)

#ajustando um modelo de regressão linear
tempr1 <- lm(y_s1~x_s1)

#estimativas dos estimadores
#estimador de Horvitz-Thompson
PHT[i,1]  <- mean(y_s1)
#Estimador de regressão generalizado GREG
PGREG[i,1]  <- PHT[i,1] + tempr1$coeff[2]*(mean(x)-mean(x_s1))
#Estimador de regressão generalizado logístico LGREG
LGREG[i,1]  <- sum(p_u1)/N

#estimadores das variâncias
#estimador da variância do estimador de Horvitz-Thompson
VPHT[i,1]  <- (N-n1)*PHT[i,1]*(1-PHT[i,1])/(N*(n1-1))
#estimador da variância do estimador do GREG
er1  <- tempr1$resid
g1 <- 1 + n1*(x_s1-mean(x_s1))*(mean(x)-mean(x_s1))/sum((x_s1-mean(x_s1))^2)
VPGREG[i,1]  <- (N-n1)*sum((er1*g1)^2)/(n1*N*(N-1))
#estimadores da variância do estimador do LGREG
el1  <- y_s1-exp(beta_hat1[1] + beta_hat1[2]*x_s1)/(exp(beta_hat1[1] + beta_hat1[2]*x_s1) + 1)
VLGREG[i,1]  <- (1 + (LGREG[i,1]-mean(y_s1))/PHT[i,1])^2*(N-n1)*sum(el1^2)/(n1*N*(n1-1))
VLGREG2[i,1] <- (N-n1)*sum(el1^2)/(n1*N*(n1-1))

#intervalos de confiança
var_est1 <- cbind(VPHT[i,1],VPGREG[i,1],VLGREG[i,1],VLGREG2[i,1])
p_est1  <- cbind(PHT[i,1],PGREG[i,1],LGREG[i,1],LGREG[i,1])

P  <- mean(y)
IC[,1] <- t(p_est1 - 1.96*sqrt(var_est1))
IC[,2] <- t(p_est1 + 1.96*sqrt(var_est1))
#taxas de cobertura
TCPHT[i,1] <- (P>=IC[1,1])*(P<=IC[1,2])
TCPGREG[i,1] <- (P>=IC[2,1])*(P<=IC[2,2])
TCLGREG[i,1]  <- (P>=IC[3,1])*(P<=IC[3,2])
TCLGREG2[i,1] <- (P>=IC[4,1])*(P<=IC[4,2])
} # fim do laço de monte carlo

#variâncias
VAPHT1 <- var(PHT[,1])
VAPGREG1 <- var(PGREG[,1])
VALGREG1  <- var(LGREG[,1])
```

```
#erro quadrático médio
EQMHT1 <- VAPHT1 + (mean(PHT[,1])-P)^2
EQMGREG1 <- VAPGREG1 + (mean(PGREG[,1])-P)^2
EQMLGREG1  <- VALGREG1 + (mean(LGREG[,1])-P)^2

#medidas para a avaliação dos estimadores
sink("C:\\Luz_Marina\\AAS.txt")
try ("OR: Associaçao entre a variável auxiliar e a variável de interesse")
OR
try("Proporçao populacional com a caracteristica de interesse")
Pr

#viés relativo dos estimadores
try("VIÉS RELATIVO DOS ESTIMADORES")
names1  <- cbind('PHT','PGREG','LGREG')
medias1 <- cbind(mean(PHT[,1]),mean(PGREG[,1]),mean(LGREG[,1]))
vies1 <- abs(medias1-P)/P
cbind(t(names1),t(vies1))

#eficiência dos estimadores
try("EFICIENCIA DOS ESTIMADORES")
deff1 <- cbind(1,VAPGREG1/VAPHT1,VALGREG1/VAPHT1)
cbind(t(names1),t(deff1))

#eficiência usando o erro quadrático médio
try("EFICIENCIA ERRO QUADRATICO MEDIO")
EQM1 <- cbind(1,EQMGREG1/EQMHT1,EQMLGREG1/EQMHT1)
cbind(t(names1),t(EQM1))

#viés relativo do estimador da variância
try("VIES RELATIVO DO ESTIMADOR DA VARIANCIA")
names2  <- cbind('PHT','PGREG','LGREG1','LGREG2')
mediasv1 <- cbind(mean(VPHT[,1]),mean(VPGREG[,1]),mean(VLGREG[,1]),mean(VLGREG2[,1]))
v1 <- cbind(VAPHT1,VAPGREG1,VALGREG1,VALGREG1)
vies21 <- abs(mediasv1-v1)/v1
cbind(t(names2),t(vies21))

#coeficientes de variação
try("COEFICIENTES DE VARIÇÃO")
vies31 <- cbind(100*sqrt(var(VPHT[,1]))/mean(VPHT[,1]),100*sqrt(var(VPGREG[,1]))/mean(VPGREG[,1]),
100*sqrt(var(VLGREG[,1]))/mean(VLGREG[,1]),100*sqrt(var(VLGREG2[,1]))/mean(VLGREG2[,1]))
cbind(t(names2),t(vies31))

#taxas de cobertura
try("TAXAS DE COBERTURA ")
vies41 <- cbind(mean(TCPHT[,1]),mean(TCPGREG[,1]),mean(TCLGREG[,1]),mean(TCLGREG2[,1]))
cbind(t(names2),t(vies41))
sink()
```

E.2 Amostragem de Bernoulli

```
# Programa para a obtenção das medidas para a avaliação dos estimadores sob BE#
rm(list = ls())
OR  <- 1.5    #Sentido e grau da associaçao entre a variável auxiliar e a variável de interesse
Pr  <- 0.5    #Proporçao populacional com a caracteristica de interesse
N   <- 10000  #Tamanho da populaçao
n1  <- 50     #Tamanho da amostra 1
r   <- 10000  #Numero de replicas

# matrizes que para armazenar os dados #
PHT   <- matrix(0,r,5)
```

```
PGREG <- matrix(0,r,5)
LGREG <- matrix(0,r,5)
VPHT  <- matrix(0,r,5)
VPGREG <- matrix(0,r,5)
VLGREG <- matrix(0,r,5)
VLGREG2 <- matrix(0,r,5)
TCPHT  <- matrix(0,r,5)
TCPGREG <- matrix(0,r,5)
TCLGREG <- matrix(0,r,5)
TCLGREG2 <- matrix(0,r,5)
IC   <- matrix(0,4,10)
sec  <- seq(1,N,by=1)

x  <- rnorm(N)   #Variável auxiliar
b1 <- log(OR)
bo <- log(Pr/(1-Pr))
p  <- exp(bo + b1*x)/(exp(bo + b1*x)+1) #Probabilidade de sucesso
u  <- runif(N)
y  <- ifelse(u<=p,1,0)  # Geração da variável de interesse

#### laço de monte carlo
for(i in 1:r){
#seleçao da amostra de tamanho esperado n1=50
pik1 <- n1/N
U1 <- runif(N)
sel1 <- ifelse( U1 <= pik1, 1, 0)
ns1 <- sum(sel1)
s1 <- (1:N)[sel1 ==1]

#Valores de las variáveis resposta e auxiliar na amostra
y_s1  <- y[s1]
x_s1  <- x[s1]

#Ajustando a regressão logistica
templ1 <- glm(y_s1~x_s1, family=binomial)

#estimativas dos parâmetros para o modelo de regressão logistica
beta_hat1  <- templ1$coeff
p_u1  <- exp(beta_hat1[1] + beta_hat1[2]*x)/(exp(beta_hat1[1] + beta_hat1[2]*x) + 1)

#ajustando um modelo de regressão linear
tempr1 <- lm(y_s1~x_s1)

#estimadores
#estimador de Horvitz-Thompson
ajust1 <- length(y_s1)/n1
PHT[i,1]  <- ajust1*mean(y_s1)
#Estimador de regressão generalizado GREG
PGREG[i,1]  <- PHT[i,1]/ajust1 + tempr1$coeff[2]*(mean(x)-mean(x_s1))
#Estimador de regressão generalizado logístico LGREG
LGREG[i,1]  <- sum(p_u1)/N

#estimadores das variâncias
#estimador da variância do estimador de Horvitz-Thompson
VPHT[i,1]  <- (N-n1)*PHT[i,1]/(N*n1)
#estimador da variância do estimador do GREG
er1  <- tempr1$resid
g1 <- (1 + (length(y_s1)*(x_s1-mean(x_s1))*(mean(x)-mean(x_s1))/sum((x_s1-mean(x_s1))^2)))/ajust1
VPGREG[i,1]  <- (N-n1)*sum((er1*g1)^2)/(n1*n1*N)
#estimadores da variância do estimador do LGREG
el1  <- y_s1-exp(beta_hat1[1] + beta_hat1[2]*x_s1)/(exp(beta_hat1[1] + beta_hat1[2]*x_s1) + 1)
VLGREG[i,1]  <- (LGREG[i,1]/PHT[i,1])^2*(N-n1)*sum(el1^2)/(n1*n1*N)
VLGREG2[i,1] <- (N-n1)*sum(el1^2)/(n1*n1*N)

#intervalos de confiança
```

```
var_est1 <- cbind(VPHT[i,1],VPGREG[i,1],VLGREG[i,1],VLGREG2[i,1])
p_est1  <- cbind(PHT[i,1],PGREG[i,1],LGREG[i,1],LGREG[i,1])
P  <- mean(y)
IC[,1] <- t(p_est1 - 1.96*sqrt(var_est1))
IC[,2] <- t(p_est1 + 1.96*sqrt(var_est1))

# taxas de cobertura
TCPHT[i,1] <- (P>=IC[1,1])*(P<=IC[1,2])
TCPGREG[i,1] <- (P>=IC[2,1])*(P<=IC[2,2])
TCLGREG[i,1]  <- (P>=IC[3,1])*(P<=IC[3,2])
TCLGREG2[i,1] <- (P>=IC[4,1])*(P<=IC[4,2])
}
#variâncias
VAPHT1 <- var(PHT[,1])
VAPGREG1 <- var(PGREG[,1])
VALGREG1  <- var(LGREG[,1])
#erro quadrático médio
EQMHT1 <- VAPHT1 + (mean(PHT[,1])-P)^2
EQMGREG1 <- VAPGREG1 + (mean(PGREG[,1])-P)^2
EQMLGREG1  <- VALGREG1 + (mean(LGREG[,1])-P)^2

#medidas para a avaliação dos estimadores
sink("C:\\Luz_Marina\\BE.txt")
try ("OR: Associaçao entre a variável auxiliar e a variável de interesse")
OR
try("Proporçao populacional com a caracteristica de interesse")
Pr
#VIÉS RELATIVO DOS ESTIMADORES
names1  <- cbind('PHT','PGREG','LGREG')
medias1 <- cbind(mean(PHT[,1]),mean(PGREG[,1]),mean(LGREG[,1]))
vies1 <- abs(medias1-P)/P
try("VIÉS RELATIVO DOS ESTIMADORES")
cbind(t(names1),t(vies1))

#EFICIENCIA DO ESTIMADOR
deff1 <- cbind(1,VAPGREG1/VAPHT1,VALGREG1/VAPHT1)
try("EFICIENCIA DOS ESTIMADORES")
cbind(t(names1),t(deff1))

#ERRO QUADRATICO MEDIO
EQM1 <- cbind(1,EQMGREG1/EQMHT1,EQMLGREG1/EQMHT1)
try("ERRO QUADRATICO MEDIO")
cbind(t(names1),t(EQM1))

#VIES RELATIVO DO ESTIMADOR DA VARIANCIA
names2  <- cbind('PHT','PGREG','LGREG1','LGREG2')
mediasv1 <- cbind(mean(VPHT[,1]),mean(VPGREG[,1]),mean(VLGREG[,1]),mean(VLGREG2[,1]))
v1 <- cbind(VAPHT1,VAPGREG1,VALGREG1,VALGREG1)
vies21 <- abs(mediasv1-v1)/v1
try("VIES RELATIVO DO ESTIMADOR DA VARIANCIA")
cbind(t(names2),t(vies21))

#COEFICIENTES DE VARIÇÃO
vies31 <- cbind(100*sqrt(var(VPHT[,1]))/mean(VPHT[,1]),100*sqrt(var(VPGREG[,1]))/mean(VPGREG[,1]),
100*sqrt(var(VLGREG[,1]))/mean(VLGREG[,1]),100*sqrt(var(VLGREG2[,1]))/mean(VLGREG2[,1]))
try("COEFICIENTES DE VARIÇÃO")
cbind(t(names2),t(vies31))

#TAXAS DE COBERTURA
vies41 <- cbind(mean(TCPHT[,1]),mean(TCPGREG[,1]),mean(TCLGREG[,1]),mean(TCLGREG2[,1]))
try("TAXAS DE COBERTURA ")
cbind(t(names2),t(vies41))
sink()
```

E.3 Amostragem Estratificada

```
# Programa para a obtenção das medidas para a avaliação dos estimadores sob AAE#

rm(list = ls())
#funções para o calculo dos estimadores
Var_AAS <- function(N,n,y){
N*N*(1-n/N)*sum((y-mean(y))^2)/(n*(n-1))
}

Lgreg <- function(y,x,N,n,s){
y_s <- y[s]
x_s <- x[s]
templ <- glm(y_s~x_s, family=binomial)
beta_hat  <- templ$coeff
p_u  <- exp(beta_hat[1] + beta_hat[2]*x)/(exp(beta_hat[1] + beta_hat[2]*x) + 1)
tempr <- lm(y_s~x_s)
PHT  <- mean(y_s)
PGREG  <- PHT + tempr$coeff[2]*(mean(x)-mean(x_s))
LGREG  <- sum(p_u)/N
VPHT   <- Var_AAS(N,n,y_s)
er  <- tempr$resid
g <- 1 + n*(x_s-mean(x_s))*(mean(x)-mean(x_s))/sum((x_s-mean(x_s))^2)
eg <- er*g
VPGREG  <- Var_AAS(N,n,eg)
el  <- y_s-exp(beta_hat[1] + beta_hat[2]*x_s)/(exp(beta_hat[1] + beta_hat[2]*x_s) + 1)
VLGREG  <- (1 + (LGREG-mean(y_s))/PHT)^2*Var_AAS(N,n,el)
VLGREG2 <- Var_AAS(N,n,el)
estp <- N*cbind(PHT,PGREG,LGREG)
estv <- cbind(VPHT,VPGREG,VLGREG,VLGREG2)
cbind(estp,estv)
}

Lgregcomb <- function(popu,muestra,tamano){
N <- length(popu[,1])
n <- length(muestra[,1])
pesos <- muestra[,3]
y_s <- muestra[,1]
x_s <- muestra[,2]
x <- popu[,1]        #no seria popu[,1]
templ <- glm(y_s~x_s,family=binomial)
beta_hat  <- templ$coeff
p_u  <- exp(beta_hat[1] + beta_hat[2]*x)/(exp(beta_hat[1] + beta_hat[2]*x) + 1)
tempr <- lm(y_s~x_s)
PHT  <- sum(pesos*y_s)/N
PGREG  <-  sum(pesos*y_s)/sum(pesos) + tempr$coeff[2]*(mean(x)-sum(pesos*x_s)/sum(pesos))
LGREG  <-  sum(p_u)/N
s1 <- (1:n)[muestra[,4] ==1]
s2 <- (1:n)[muestra[,4] ==2]
s3 <- (1:n)[muestra[,4] ==3]
VPHT <- Var_AAS(tamano[1,1],tamano[1,2],y_s[s1])+Var_AAS(tamano[2,1],tamano[2,2],y_s[s2])
VPHT <- VPHT + Var_AAS(tamano[3,1],tamano[3,2],y_s[s3])
er  <- tempr$resid
x_m <- sum(pesos*x_s)/sum(pesos)
g <- (N/sum(pesos))*(1 + (x_s-x_m)*(mean(x)-x_m)*sum(pesos)/(sum(pesos*(x_s-x_m)^2)))
eg <- er*g
VPGREG <- Var_AAS(tamano[1,1],tamano[1,2],eg[s1])+Var_AAS(tamano[2,1],tamano[2,2],eg[s2])
VPGREG <- VPGREG + Var_AAS(tamano[3,1],tamano[3,2],eg[s3])
el  <- y_s-exp(beta_hat[1] + beta_hat[2]*x_s)/(exp(beta_hat[1] + beta_hat[2]*x_s) + 1)
VLGREG2 <- Var_AAS(tamano[1,1],tamano[1,2],el[s1])+Var_AAS(tamano[2,1],tamano[2,2],el[s2])
VLGREG2 <- VLGREG2 + Var_AAS(tamano[3,1],tamano[3,2],el[s3])
VLGREG <- VLGREG2*(LGREG/PHT)^2
estp <- cbind(PHT,PGREG,LGREG)
estv <- cbind(VPHT,VPGREG,VLGREG,VLGREG2)/(N*N)
```

```
cbind(estp,estv)
}

Pr  <-  0.5  #proporção de indivíduos na população com a característica de interesse
OR1  <- 2    #OR para o estrato 1
OR2  <- 0.1  #OR para o estrato 2
OR3  <- 5    #OR para o estrato 3
N1  <- 5000  #tamanho do estrato 1
N2  <- 3000  #tamanho do estrato 2
N3  <- 2000  #tamanho do estrato 3
N<- N1+N2+N3 #tamanho da população
n1  <- 200   #tamanho da amostra total
n11 <- ceiling(n1*N1/N) #tamanho da amostra no estrato 1
n12 <- ceiling(n1*N2/N) #tamanho da amostra no estrato 2
n13 <- ceiling(n1*N3/N) #tamanho da amostra no estrato 3
r   <- 10000  #número de réplicas

############## GERAÇAO DA VARIAVEL AUXILIAR ###
x1  <- rnorm(N1)
x2  <- rnorm(N2)
x3  <- rnorm(N3)
b11 <- log(OR1)
b12 <- log(OR2)
b13 <- log(OR3)
bo1 <- log(Pr/(1-Pr))
bo2 <- log(Pr/(1-Pr))
bo3 <- log(Pr/(1-Pr))

################  PROBABILIDADES DE SUCESSO  ###
p1  <- exp(bo1 + b11*x1)/(exp(bo1 + b11*x1)+1)
p2  <- exp(bo2 + b12*x2)/(exp(bo2 + b12*x2)+1)
p3  <- exp(bo3 + b12*x3)/(exp(bo3 + b12*x3)+1)

#############  GERAÇAO DA VARIAVEL RESPOSTA ####
u1  <- runif(N1)
y1  <- ifelse(u1<=p1,1,0)
u2  <- runif(N2)
y2  <- ifelse(u2<=p2,1,0)
u3  <- runif(N3)
y3  <- ifelse(u3<=p3,1,0)
y <-t(cbind(t(y1),t(y2),t(y3)))
x <-t(cbind(t(x1),t(x2),t(x3)))
popu <- cbind(x,y)

############ Laço de Monte Carlo
for(i in 1:r){
sec1  <- seq(1,N1,by=1)
sec2  <- seq(1,N2,by=1)
sec3  <- seq(1,N3,by=1)
#seleção das amostras em cada estrato
s11 <- sample(sec1,n11)
s12 <- sample(sec2,n12)
s13 <- sample(sec3,n13)
muestra11 <- cbind(y1[s11],x1[s11],N1/n11,1)
muestra12 <- cbind(y2[s12],x2[s12],N2/n12,2)
muestra13 <- cbind(y3[s13],x3[s13],N3/n13,3)
muestra1 <- rbind(muestra11,muestra12,muestra13)
tamano1 <- cbind(rbind(N1,N2,N3),rbind(n11,n12,n13))

#estimadores combinados
m1 <- Lgregcomb(popu,muestra1,tamano1)
estp <- cbind(PHT,PGREG,LGREG)
estv <- cbind(VPHT,VPGREG,VLGREG,VLGREG2)/(N*N)
PHTC[i,1] <- m1[1]
LGREGC[i,1] <- m1[3]
```

```
VPHTC[i,1] <- m1[4]
VPGREGC[i,1] <- m1[5]
VLGREGC[i,1] <- m1[6]
VLGREGC2[i,1] <- m1[7]

#estimadores separados
t1 <- Lgreg(y1,x1,N1,n11,s11) + Lgreg(y2,x2,N2,n12,s12) + Lgreg(y3,x3,N3,n13,s13)
PHT[i,1] <- t1[1]/N
PGREG[i,1]   <- t1[2]/N
LGREG[i,1]   <- t1[3]/N
VPHT[i,1] <- t1[4]/(N*N)
VPGREG[i,1]   <- t1[5]/(N*N)
VLGREG[i,1]   <- t1[6]/(N*N)
VLGREG2[i,1]   <- t1[7]/(N*N)

#intervalos de confiança estimadores separados
var_est1 <- cbind(VPHT[i,1],VPGREG[i,1],VLGREG[i,1],VLGREG2[i,1])
p_est1  <- cbind(PHT[i,1],PGREG[i,1],LGREG[i,1],LGREG[i,1])
P  <- (sum(y1)+sum(y2)+sum(y3))/N
IC[,1] <- t(p_est1 - 1.96*sqrt(var_est1))
IC[,2] <- t(p_est1 + 1.96*sqrt(var_est1))
TCPHT[i,1] <- (P>=IC[1,1])*(P<=IC[1,2])
TCPGREG[i,1] <- (P>=IC[2,1])*(P<=IC[2,2])
TCLGREG[i,1]  <- (P>=IC[3,1])*(P<=IC[3,2])
TCLGREG2[i,1] <- (P>=IC[4,1])*(P<=IC[4,2])

#intervalos de confiança estimadores combinados
var_est1C <- cbind(VPHTC[i,1],VPGREGC[i,1],VLGREGC[i,1],VLGREGC2[i,1])
p_est1C  <- cbind(PHTC[i,1],PGREGC[i,1],LGREGC[i,1],LGREGC[i,1])
P  <- (sum(y1)+sum(y2)+sum(y3))/N
ICC[,1] <- t(p_est1C - 1.96*sqrt(var_est1C))
ICC[,2] <- t(p_est1C + 1.96*sqrt(var_est1C))

###taxas de cobertura
TCPHTC[i,1] <- (P>=ICC[1,1])*(P<=ICC[1,2])
TCPGREGC[i,1] <- (P>=ICC[2,1])*(P<=ICC[2,2])
TCLGREGC[i,1]  <- (P>=ICC[3,1])*(P<=ICC[3,2])
TCLGREGC2[i,1] <- (P>=ICC[4,1])*(P<=ICC[4,2])
}

#VARIANCIAS
VAPHT1 <- var(PHT[,1])
VAPGREG1 <- var(PGREG[,1])
VALGREG1  <- var(LGREG[,1])
VAPHTC1 <- var(PHTC[,1])
VAPGREGC1 <- var(PGREGC[,1])
VALGREGC1  <- var(LGREGC[,1])
#ERRO QUADRATICO MEDIO
EQMHT1 <- VAPHT1 + (mean(PHT[,1])-P)^2
EQMGREG1 <- VAPGREG1 + (mean(PGREG[,1])-P)^2
EQMLGREG1  <- VALGREG1 + (mean(LGREG[,1])-P)^2
EQMHTC1 <- VAPHTC1 + (mean(PHTC[,1])-P)^2
EQMGREGC1 <- VAPGREGC1 + (mean(PGREGC[,1])-P)^2
EQMLGREGC1  <- VALGREGC1 + (mean(LGREGC[,1])-P)^2
#médias dos estimadores similar aos outro programas
```

APÊNDICE F

Programa em SAS

Neste apêndice é apresentado o programa no pacote estatístico SAS em sua versão 9.1.3, para a obtenção dos resultados apresentados no capítulo 7.

F.1 Amostragem Aleatória Simples

```
/* Programa para a obtenção das estimativas da taxa de desemprego no mês de outubro de 2005,
na Região Metropolinata de Recife, usando uma plano AAS*/
/* Leitura dos dados da PME*/
Data PMENOVA;
INFILE 'C:\LUZ_MARINA\PMEnova.102005.TXT' LRECL=490 MISSOVER;
INPUT
@00001  V035    2. /* RM   */
... /* variáveis da pesquisa*/
@00478  VD28    3. /* Número de horas efetivamente trabalhadas na semana de referência */
;
/* selecionando o banco de dados só para Recife */
data pme;set pmenova;
if v035=26;
run;

/*macro para obter as estimativas do parâmetro de interesse */
/*N=Tamanho da populaçao 6067, m=Tamanho da amostra, p=fraccion muestral*/
%macro pme(N,m,p);
/*Retirando as variáveis do banco de dados de Recife*/
data pme;set pme.pme;if vd3=1;y=0;
if vd2 in (1,2) then y=1;v2032=0; if v203=1 then v2032=1;keep y v203 v2032 v234;run;

ods listing close;
proc surveyselect data=pme method=srs rate=&p out=muestra1;run;
ods output ParameterEstimates=linear_par;
proc genmod data=muestra1;
class v203;model y=v203 v234/dist=normal link=identity;run;
ods output ParameterEstimates=logistic_par;
proc surveylogistic data=muestra1 descending;
class v203;model y=v203 v234;run;
ods listing;

data linear_par;set linear_par;
if Parameter='V203' and Level1='2' then delete;b0=0;b1=0;b2=0;
```

```
if Parameter='Intercept' then b0=Estimate;
if Parameter='V203' then b1=Estimate;
if Parameter='V234' then b2=Estimate;model='linear';
keep model b0 b1 b2;run;

data logistic_par;set logistic_par;b0=0;b1=0;b2=0;
if Variable='Intercept' then b0=Estimate;
if Variable='V203' then b1=Estimate;
if Variable='V234' then b2=Estimate;model='logistic';
keep model b0 b1 b2;run;
data par;set linear_par logistic_par;run;

proc sql;create table t1 as
select model, sum(b0) as b0, sum(b1) as b1, sum(b2) as b2
from par group by model;
create table t2 as select pme.*,t1.* from pme, t1;
create table r1 as select muestra1.*,t1.* from muestra1, t1;quit;

data t2;set t2;preditor=b0 + b1*v2032 + b2*v234;mu=preditor;
if model='logistic' then mu=exp(preditor)/(1+exp(preditor));run;
data r1;set r1;preditor=b0 + b1*v2032 + b2*v234;mu=preditor;
if model='logistic' then mu=exp(preditor)/(1+exp(preditor));run;

proc sql;create table result as
select model, sum(mu)/count(*) as P_estimado from t2 group by model;
create table result2 as
select model, (&N-&m)*std(y-mu)/(&N*(&N-1)) as Var_estimada from r1 group by model;quit;

ods listing close;
ods output ParameterEstimates=linear_par_u;
proc genmod data=pme;
class v203;model y=v203 v234/dist=normal link=identity;run;
ods output ParameterEstimates=logistic_par_u;
proc surveylogistic data=pme descending;
class v203;model y=v203 v234;run;
ods listing;

data linear_par_u;set linear_par_u;
if Parameter='V203' and Level1='2' then delete;b0=0;b1=0;b2=0;
if Parameter='Intercept' then b0=Estimate;
if Parameter='V203' then b1=Estimate;
if Parameter='V234' then b2=Estimate;model='linear';
keep model b0 b1 b2;run;

data logistic_par_u;set logistic_par_u;b0=0;b1=0;b2=0;
if Variable='Intercept' then b0=Estimate;
if Variable='V203' then b1=Estimate;
if Variable='V234' then b2=Estimate;model='logistic';
keep model b0 b1 b2;run;
data par_u;set linear_par_u logistic_par_u;run;

proc sql; create table tu1 as select model, sum(b0) as b0, sum(b1) as b1, sum(b2) as b2
from par_u group by model;
create table tu2 as select pme.*,tu1.* from pme, tu1;quit;

data tu2;set tu2;preditor=b0 + b1*v2032 + b2*v234;mu=preditor;
if model='logistic' then mu=exp(preditor)/(1+exp(preditor));run;
proc sql;create table result3 as
select model, (&N-&m)*std(y-mu)/(&N*(&N-1)) as Var_popu
from tu2 group by model;quit;

proc sql;select mean(y) as P from pme;quit;
proc print data=result noobs;run;
proc print data=result2 noobs;run;
proc print data=result3 noobs;run;
```

```
proc sql;select 'PHT' as estimador, (mean(y)) as P_estimado,
(&N-&m)*std(y)/(&N*(&N-1)) as Var_est from muestra1;
select 'PHT' as estimador, (&N-&m)*std(y)/(&N*(&N-1)) as Var_popu from pme;quit;
%mend pme;
%pme(6067,1214,0.2);
```

F.2 Amostragem Estratificada

```
/* Programa para a obtenção das estimativas da taxa de desemprego no mês de outubro de 2005,
na Região Metropolinata de Recife, usando uma plano AE*/

%macro estratificado(N,p);
options nodate nonumber;
/*selecionando os estratos*/
data pme;set pme.pme;if vd3=1;y=0;
if vd2 in (1,2) then y=1;estrato=3;
if v112 in (260010) then estrato=1;
if v112 in (260006,260008) then estrato=2;obs=_n_;
keep y v203 v234 estrato obs;run;

proc sort data=pme;by estrato;run;
ods listing close;
proc surveyselect data=pme method=srs rate=&p out=muestra1;
strata estrato;run;
data muestra1;set muestra1;obs=_n_;run;
ods output ObStats=salida;
proc genmod data=muestra1;
class v203;model y=v203 v234/dist=normal residual predicted link=identity;
by estrato;run;
data salida;set salida;keep estrato pred resraw y;run;
ods output ObStats=salida2;
proc surveylogistic data=muestra1 descending;
class v203;model y=v203 v234/dist=binomial residual predicted link=logit;
by estrato;run;
data salida2;set salida2;keep estrato pred resraw y;run;
ods output ObStats=salida3;
proc genmod data=muestra1;
class v203;model y=v203 v234/dist=normal residual predicted link=identity;run;
data salida3;set salida3;keep observation pred resraw y;run;
ods output ObStats=salida4;
proc surveylogistic data=muestra1 descending;
class v203;model y=v203 v234/dist=binomial residual predicted link=logit;run;
data salida4;set salida4;keep observation pred resraw y;run;

ods output ObStats=universo;
proc genmod data=pme;
class v203;model y=v203 v234/dist=normal residual predicted link=identity;
by estrato;run;
data universo;set universo;keep estrato pred resraw y;run;
ods output ObStats=universo2;
proc surveylogistic data=pme descending;
class v203;model y=v203 v234/dist=binomial residual predicted link=logit;
by estrato;run;
data universo2;set universo2;keep estrato pred resraw y;run;
ods output ObStats=universo3;
proc genmod data=muestra1;
class v203;model y=v203 v234/dist=normal residual predicted link=identity;run;
data universo3;set universo3;keep observation pred resraw y;run;
ods output ObStats=universo4;
proc surveylogistic data=pme descending;
class v203;model y=v203 v234/dist=binomial residual predicted link=logit;run;
data universo4;set universo4;keep observation pred resraw y;run;
```

```
ods listing;

proc sql;create table salida31 as
select salida3.*, estrato
from salida3 inner join muestra1 on (obs=observation);
create table salida41 as
select salida4.*, estrato
from salida4 inner join muestra1 on (obs=observation);
create table universo31 as
select universo3.*, estrato
from universo3 inner join muestra1 on (obs=observation);
create table universo41 as
select universo4.*, estrato
from universo4 inner join muestra1 on (obs=observation);

title 'ESTIMATIVA PHT DE P
       Aporte de cada estrato a la estimativa de P';
select estrato, mean(y)*count(*)/(&p*&N) as contribucion
from salida group by estrato;
title 'Estimativa de P';
select mean(y) as P_estimado from salida;
title 'Aporte de cada estrato a la estimativa de Var(P)';
create table temp as
select estrato, ((1-&p)/count(*))*(std(y)*count(*)/(&p*&N))*((1-&p)/count(*))*(std(y)*count(*)/(&p*&N))
as contribucion
from salida group by estrato;
select temp.* from temp;
title 'Estimativa de Var(P)';
select sum(contribucion) as Var_estimada from temp;
title 'Aporte de cada estrato a Var(P)';
create table temp as
select estrato, ((1-&p)/(count(*)*&p))*(std(y)*count(*)/&N)*(std(y)*count(*)/&N)
as contribucion
from universo group by estrato;
select temp.* from temp;
title 'Var(P)';
select sum(contribucion) as Var_estimada from temp;

title 'ESTIMATIVA SEPARADA PGREG DE P
       Aporte de cada estrato a la estimativa de P';
select estrato, mean(pred)*count(*)/&N as contribucion
from universo group by estrato;
title 'Estimativa de P';
select sum(pred)/&N as P_estimado
from universo;
title 'Aporte de cada estrato a la estimativa de Var(P)';
create table temp as
select estrato, ((1-&p)/count(*))*(std(resraw)*count(*)/(&p*&N))*(std(resraw)*count(*)/(&p*&N))
as contribucion
from salida group by estrato;
select temp.* from temp;
title 'Estimativa de Var(P)';
select sum(contribucion) as Var_estimada from temp;
title 'Aporte de cada estrato a Var(P)';
create table temp as
select estrato, ((1-&p)/(count(*)*&p))*(std(resraw)*count(*)/&N)*(std(resraw)*count(*)/&N)
as contribucion
from universo group by estrato;
select temp.* from temp;
title 'Var(P)';
select sum(contribucion) as Var_estimada from temp;

title 'ESTIMATIVA SEPARADA LGREG DE P
       Aporte de cada estrato a la estimativa de P';
select estrato, mean(pred)*count(*)/&N as contribucion
```

```
from universo2 group by estrato;
title 'Estimativa de P';
select sum(pred)/&N as P_estimado
from universo2;
title 'Aporte de cada estrato a la estimativa de Var(P)';
create table temp as
select estrato, ((1-&p)/count(*))*(std(resraw)*count(*)/(&p*&N))*(std(resraw)*count(*)/(&p*&N))
as contribucion
from salida2 group by estrato;
select temp.* from temp;
title 'Estimativa de Var(P)';
select sum(contribucion) as Var_estimada from temp;
title 'Aporte de cada estrato a Var(P)';
create table temp as
select estrato, ((1-&p)/(count(*)*&p))*(std(resraw)*count(*)/&N)*(std(resraw)*count(*)/&N)
as contribucion
from universo2 group by estrato;
select temp.* from temp;
title 'Var(P)';
select sum(contribucion) as Var_estimada from temp;

title 'ESTIMATIVA COMBINADA PGREG DE P
       Estimativa de P';
select sum(pred)/&N as P_estimado
from universo31;
title 'Aporte de cada estrato a la estimativa de Var(P)';
create table temp as
select estrato, ((1-&p)/count(*))*(std(resraw)*count(*)/(&p*&N))*(std(resraw)*count(*)/(&p*&N))
as contribucion
from salida31 group by estrato;
select temp.* from temp;
title 'Estimativa de Var(P)';
select sum(contribucion) as Var_estimada from temp;
title 'Aporte de cada estrato a Var(P)';
create table temp as
select estrato, ((1-&p)/(count(*)*&p))*(std(resraw)*count(*)/&N)*(std(resraw)*count(*)/&N)
as contribucion
from universo31 group by estrato;
select temp.* from temp;
title 'Var(P)';
select sum(contribucion) as Var_estimada from temp;

title 'ESTIMATIVA COMBINADA LGREG DE P
       Estimativa de P';
select sum(pred)/&N as P_estimado
from universo41;
title 'Aporte de cada estrato a la estimativa de Var(P)';
create table temp as
select estrato, ((1-&p)/count(*))*(std(resraw)*count(*)/(&p*&N))*(std(resraw)*count(*)/(&p*&N))
as contribucion format 2.10
from salida41 group by estrato;
select temp.* from temp;
title 'Estimativa de Var(P)';
select sum(contribucion) as Var_estimada from temp;
title 'Aporte de cada estrato a Var(P)';
create table temp as
select estrato, ((1-&p)/(count(*)*&p))*(std(resraw)*count(*)/&N)*(std(resraw)*count(*)/&N)
as contribucion
from universo41 group by estrato;
select temp.* from temp;
title 'Var(P)';
select sum(contribucion) as Var_estimada from temp;quit;
%mend estratificado;
%estratificado(6067,0.2);
```

Referências

[1] Agresti, A. (1990). *Categorical Data Analysis.* New York: Willey.

[2] An, B.A. (2002). Performing logistic regression on survey data with the new SURVEYLOGISTIC procedure. *Proceedings of the Twenty-Seventh Annual SAS Users Group International Conference (SUGI27),* Contributed paper, 258-27.

[3] Casady, R.J. e Valliant, R. (1993). Conditional properties of post-stratified estimators under normal theory. *Survey Methodology*. **19**,183-192.

[4] Cochran, W.G. (1977). *Sampling Techniques, 3rd Edition*. New York: Wiley.

[5] Deville, J.C. (1992). Constrained samples, conditional inference, weighting: three aspects of the utilisation of auxiliary information. *Proceeding of the Workshop: Auxiliary Information in Surveys,* Örebro.

[6] Deville, J.C. e Särndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association.* **87**, 376-382.

[7] Duchesne, P. (2003). Estimation of a proportion with survey data. *Journal of Statistics Education*. **11**, 1-24.

[8] Ferraz, C. (2005). *Notas de aula de MES921-Amostragem, não publicadas.*

[9] Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology* **28**, 5-23.

[10] Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Pub. Math. Inst. Hungrian Acad. Sci.* **5**, 361-374.

[11] Holt, D., Smith, T.M.F. e Winter, P.D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society, Series A.* **143**, 474-487.

[12] Horvitz, D.G. e Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association.* **47**, 663-685.

[13] Hosmer, D.W. e Lemeshow, S. (1989). *Applied Logistic Regression.* New York: Wiley.

[14] IBGE (2005). *Pesquisa Mensal de Emprego*. Rio de Janeiro, Fundação Instituto Brasileiro de Geografia e Estatística.

[15] Isaki, C.T. e Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association.* **77**, 89-96.

[16] Jørgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society B.* **49**, 127-144.

[17] Lehtonen, R. e Veijanen, A. (1998a). Logistic generalized regression estimators. *Survey Methodology.* **24**, 51-55.

[18] Lehtonen, R. e Veijanen, A. (1998b). On multinomial logistic generalized regression estimators. *Preprint from Department of Statistics, University of Jyväskylä.* **22**.

[19] Lehtonen, R. e Pahkinen E.J. (2004). *Practical Methods for Design and Analysis of Complex Surveys, 2nd Edition*. Chichester: Wiley.

[20] Lohr, S. (1999). *Sampling:Design and Analysis.* Belmont, CA: Duxbury Press.

[21] Madow, W.G. (1948). On the limiting distributions of estimates based on samples from finites universes. *Annals of Mathematical Statistics.* **19**. 535-545.

[22] McCullagh, P. e Nelder, J.A. (1989). *Generalized Linear Models, 2nd. Edition*. London: Chapman and Hall.

[23] Montanari, G.E. (1997). On conditional properties of finite population mean estimators . *Proceeding of the 51st Session of the International Statistical Institute,* Contributed paper, 351-352.

[24] Montanari, G.E. (1998). On regression estimation of finite population means. *Survey Methodology*. **24**, 69-77.

[25] Morel, G. (1989). Logistic regression under complex survey design. *Survey Methodology* **15**, 203-223.

[26] Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics.* **10**, 153-165.

[27] Rao, J.N.K. (1997). Development in sample survey theory: an appraisal. *Canadian Journal of Statistics.* **25**, 1-21.

[28] Särndal, C.E., Swensson, B. e Wretman, J. (1992). *Model Assisted Survey Sampling.* New York: Springer Verlag.

[29] Särndal, C.E. (2001). Design-Based Methodologies for Domain Estimation. *Lectures Notes on Estimation for Population Domains and Small Areas.* **2001/5**, 5-49.

[30] SAS Institute Inc. (2006). *SAS OnlineDoc 9.1.3*. Cary, NC: SAS Institute Inc.

[31] Tillé, Y. (1998). Estimation in surveys using conditional inclusion probabilities: simple random sampling. *International Statistical Review.* **66** 303-322.

[32] Wei, B.C. (1998). Exponential Family Nonlinear Models. *Lecture Notes in Statistics.* **130**, New York:Springer.

[33] Wright, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association.* **78**, 879-884.