Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

IDENTIFICAÇÃO DOS FATORES QUE POSSIBILITAM A INCLUSÃO NO PROGRAMA BOLSA FAMÍLIA

Camila Marques Mendes Tavares **07/30831**
Iracema Veiga Madeira Mauriz **05/84312**

Brasília

2011

Camila Marques Mendes Tavares **07/30831**

Iracema Veiga Madeira Mauriz **05/84312**

IDENTIFICAÇÃO DOS FATORES QUE POSSIBILITAM A INCLUSÃO NO PROGRAMA BOLSA FAMÍLIA

Relatório apresentado à disciplina Estágio Supervisionado II do curso de graduação em Estatística, Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Orientador: Alan Ricardo da Silva

Brasília

2011

"O conhecimento de estatística é como o de uma língua estrangeira ou álgebra; ele poderá ser útil a qualquer tempo ou circunstância.".(A. L. Bowley)

"À Deus, por ser minha luz e guia. Às nossas famílias, pelo apoio e torcida incondicionais. Aos nossos colegas, amigos, professores, a todas as pessoas que estiveram presentes em nossas vidas, desde as que nos ajudaram a superar os maiores desafios ou simplesmente aos que nos deram um abraço de bom dia. "

Agradecimentos

À Deus, que criou e tornou possível a realização de mais esta etapa. Ao nosso orientador e amigo Alan, por nos acompanhar em cada passo desse trabalho. Ao MDS pelo apoio e pelas informações cedidas. Às nossas famílias por sempre estarem ao nosso lado. A todos que contribuíram de alguma forma para conseguirmos atingir os objetivos desse estudo.

Resumo

O presente trabalho tratará sobre a identificação dos fatores que possibilitam a inclusão de pessoas no Program Bolsa Família. Esse programa, criado em 2003 no Governo Lula, representa o maior programa de transferência condicionada de renda do mundo, consistindo na unificação, integração e ampliação de programas sociais já existentes, num único programa social que hoje atende mais de 12,5 milhões de famílias. O Bolsa Família seleciona as famílias com base nas informações inseridas pelo município no Cadastro Único para Programas Sociais, e a determinados critérios pré-estabelecidos. Tal banco foi gentilmente fornecido pelo Ministério do Desenvolvimento Social e Combate a Fome (MDS). Portanto, o trabalho tem por objetivo identificar os fatores que possibilitam a inclusão de famílias no Programa Bolsa Família, por meio de uma análise exploratória dos dados e de um modelo logístico, além de analisar as diferenças regionais desses fatores de inclusão.

Sumário

Capítulo 1

INTRODUÇÃO

Programas de transferências de renda contra a pobreza são políticas sociais correntemente empregadas em várias partes do mundo para combater e reduzir a pobreza. Segundo Duarte et al. (2007), no curto prazo, esses programas objetivam aliviar os problemas decorrentes da situação de pobreza e, no longo prazo, investir no capital humano, interrompendo o ciclo da pobreza de geração a geração. A idéia dos programas de transferências condicionadas começou a ganhar força em 1997, quando só havia três países no mundo com essa experiência: Bangladesh, México e Brasil (Lindert, 2005). Alguns anos depois, outros países implantaram programas similares, alguns deles até inspirados no Bolsa Família. Este programa tem sido recomendado pela Organização das Nações Unidas para adoção em outros países em desenvolvimento.

O Bolsa Família é hoje o maior programa de transferência condicionada de renda do mundo. O programa, criado em 2003, no governo Lula, consistiu na unificação, integração e ampliação de programas sociais já existentes, como o Fome Zero, Bolsa Escola, Auxilio Gás e o Cartão Alimentação, num único programa social que hoje atende mais de 12,5 milhões de famílias. O Bolsa Família é citado por alguns ana-

listas como sendo um dos responsáveis pela redução do índice de miséria no Brasil, que caiu 27,7% entre 2002 e 2006 (SENARC, 2009).

O Bolsa Família seleciona as famílias com base nas informações inseridas pelo município no Cadastro Único para Programas Sociais. O Cadastro é um instrumento de coleta de dados que tem como objetivo identificar todas as famílias de baixa renda existentes no País (MDS, 2010).

Dentre os diversos fatores que possibilitam ou não a inclusão das famílias no programa Bolsa Família, se destacam a renda inferior a R$ 140,00 (percapita), frequência escolar regular e vacinação em dia. Portanto, o trabalho visa identificar e quantificar os fatores que possibilitam a inclusão no Programa Bolsa Família, através da análise do banco de dados do Cadastro Único, gentilmente fornecido pelo Ministério do Desenvolvimento Social e Combate a Fome (MDS).

Em 2011, com a mudança no governo ocorreram algumas alterações em relação ao programa. O valor dos benefícios aumentou para atualizar os valores à nova realidade econômica do país e manter o poder de compra das famílias beneficiárias. Sabe-se, ainda, que os recursos do Programa Bolsa Família são limitados, apesar de serem considerávelmente grandes. Outro fato de importância foi o anúncio, pelo atual ministro da Fazenda, de um corte de cinquenta bilhões de reais no Orçamento para o primeiro ano de mandato da presidenta Dilma. Porém, alguns programas como o PAC e o PBF tiveram seus recursos aumentados. Logo, em um contexto de uma possivel redução dos recursos, o modelo que será proposto indicará a probabilidade de uma familia pertencer ao programa. Por exemplo, objetivaremos responder

perguntas como: em uma situação onde só há a possibilidade de inclusão de mais uma família para receber o beneficio, a família escolhida seria a que tem em sua composição 3 filhos em idade escolar e que frequentam a escola e um membro aposentado que sustenta a família com a renda de sua aposentadoria ou a família que tem em sua composição quatro filhos, dois em idade escolar e dois jovens e ainda dois membros que trabalham?

1.1 Objetivos

O objetivo geral do trabalho é identificar os fatores que possibilitam a inclusão no Programa Bolsa Família, por meio de um modelo logístico e analisar as diferenças regionais desses fatores de inclusão.

Os objetivos específicos são:

- Desenvolver a análise exploratória do banco de dados utilizando o software SAS para identificar as regiões de maior concentração de famílias em situação de extrema pobreza;
- Identificar as variáveis do modelo logístico;
- Constatar a eficácia do programa Bolsa Família e o impacto na renda dos beneficiários;

Capítulo 2

PROGRAMA BOLSA FAMÍLIA

2.1 INTRODUÇÃO

O Bolsa Família é um programa do Governo Federal destinado às ações de transferência de renda com condicionalidades, instituído pelo Governo Federal em outubro de 2003, por meio da Medida Provisória n.º 132. Posteriormente criado pela lei n.º 10.836 de 9 de janeiro de 2004 e regulamentado pelo decreto n.º5.209 de 17 de setembro de 2004. Esse programa é gerido pelo Ministério do Desenvolvimento Social e Combate a Fome (MDS).

O programa unificou os procedimentos de gestão e execução das ações de transferência de renda, como o Fome Zero, Bolsa Escola, Auxílio Gás, Cartão Alimentação, Bolsa Alimentação, entre outros. O Bolsa Família tem por objetivos (CAIXA, 2007):

1. Combater a fome e promover a segurança alimentar e nutricional;

2. Combater a pobreza e outras formas de privação das famílias;

3. Promover o acesso à rede de serviços públicos, em especial, saúde, educação, segurança alimentar e assistência social;

4. Criar possibilidades de emancipação sustentada dos grupos familiares e desenvolvimento local dos territórios.

O público alvo do programa são as famílias com renda per capita mensal de até R$70,00 com ou sem filhos, que são consideradas como em situação de extrema pobreza. E também as famílias com renda per capita mensal entre R$70,00 e R$140,00, desde que tenham crianças ou adolescentes com idade entre 0 e 17 anos ou gestantes em sua composição, estas famílias são consideradas em situação de pobreza.

2.2 Benefícios

O valor total do benefício recebido pode variar de R$22,00 a R$200,00 por família. Existem três tipos de benefícios: o básico, o variável e o variável jovem. O benefício básico (BB), no valor de R$68,00 por família, é destinado às famílias com renda de até R$70,00, independente da composição e do número do grupo familiar. O benefício variável (BV), no valor de R$22,00 por criança/adolescente, é destinado às famílias que tenham em sua composição gestantes ou crianças e adolescentes com até 15 anos de idade e cada família pode acumular até três benefícios (um por filho e até três filhos). Já o Benefício Variável vinculado ao Jovem (BVJ), no valor de R$ 33,00 por jovem, é destinado a famílias que tenham em sua composição adolescentes entre 16 e 17 anos e cada família pode acumular até dois benefícios (CAIXA, 2010).

O Programa Bolsa Família permite às famílias em situação de extrema pobreza acumular o benefício básico, o variável, até o máximo de três benefícios por família e o variável para jovem, até o máximo de dois benefícios por família, totalizando R$200,00 por mês. Já para as famílias em situação de pobreza, é permitido acumular

o benefício variável, até o máximo de três benefícios por família e o variável para jovem, até o limite de dois benefícios por família, totalizando no máximo R$132,00 por mês. A Tabela 2.2 apresenta os valores por tipo de benefício.

Tabela 2.1: Valor e Tipo de Benefício.

Perfil/Tipo da Família*	**BB**	**BV**	**BVJ**
Familia com renda por pessoa de até R$70,00 por mês	R$68,00	R$22,00 a R$66,00 (máximo de 3 benefícios variáveis por família)	R$33,00 a R$66,00 (máximo de 2 2 BVJ por família)
Família com renda por pessoa de R$70,01 até R$140,00 por mês	-	R$22,00 a R$66,00 (máximo de 3 benefícios variáveis por família)	R$33,00 a R$66,00 (máximo de 2 BVJ por família)

Fonte:Datasus (2009).

2.3 Condicionalidades

As famílias beneficiárias só recebem o valor estabelecido se cumprirem as seguintes exigências (CAIXA, 2007):

- No caso de existência de gestantes, o comparecimento às consultas de pré-natal, conforme calendário preconizado pelo Ministério da Saúde (MS);
- Participação em atividades educativas ofertadas pelo MS sobre aleitamento materno e alimentação saudável, no caso de inclusão de nutrizes;
- Manter em dia o cartão de vacinação das crianças de 0 a 6 anos;
- Garantir freqüência mínima de 85% na escola, para crianças e adolescentes de 6 a 15 anos;
- Garantir frequência mínima de 75% na escola, para adolescentes de 16 e 17 anos;

- Participar, quando for o caso, de programas de alfabetização de adultos.

Cabe à prefeitura municipal realizar o cadastramento dessas famílias, por meio do Cadastro Único dos Programas Sociais do Governo Federal (CadÚnico), desde que a família procure o setor responsável pelo Bolsa Família no seu município. Depois de cadastrada, é feita a seleção das famílias aptas a receber o benefício pelo Ministério do Desenvolvimento Social e Combate à Fome (MDS), com base nos dados inseridos pelas prefeituras no CadÚnico. A seleção, realizada mensalmente, tem como critério principal a renda per capita da família e é regulamentada pela Portaria GM/MDS nº. 341 de 7 de outubro de 2008.

É necessário escolher um representante legal para a família que será quem poderá sacar o beneficio e quem deve apresentar um documento de identificação na hora de se cadastrar. É preferível que o representante legal seja mulher, o que acontece com cerca de 96% das famílias, pois a família pode escolher como quer gastar o benefício e as mulheres são geralmente mais consciente, gastando-o com os filhos. Essa é também uma forma de emancipação dessas mulheres e acaba por elevar a auto-estima delas (Datasus, 2009).

2.4 Estatísticas Básicas

O Bolsa Família é citado por alguns analistas como sendo um dos responsáveis pela redução do índice de miséria no Brasil, que caiu 27,7% entre 2002 e 2006 (SENARC, 2009). O programa, que hoje atende mais de 12,5 milhões de famílias, tem sido recomendado pela Organização das Nações Unidas para adoção em outros países em desenvolvimento.

Dessas 12,5 milhões de famílias atendidas pelo programa, 70% delas residem em área urbana, 61,6% em domicilios próprios e 92,6% em casas. Pouco mais de 40% dos domicílios atendidos dispõem simultaneamente de abastacimento de água por rede pública, acesso a coleta de lixo e escoamento sanitário adequado, ou seja, por fossa séptica ou por rede pública. Precariamente, 21,6% dos domicílios não dispõem de água tratada, mas esse índice tem apresentado melhoras ao longo dos anos desde a criação do PBF. O perfil das pessoas que compõem as famílias beneficiárias aponta para o predomínio de mulheres (54% do total), de cor/raça parda(64,1%) e com idade inferior a 20 anos (54,5%). A escolaridade entre os adultos beneficiários é baixa, entre os que tem 25 anos ou mais, 16,7% são analfabetos e 65,4% tem ensino fundamental incompleto SENARC (2009).

O impacto do PBF no alívio imediato da pobreza pode ser avaliado pelo seu efeito positivo na renda das famílias pobres. Os benefícios monetários do programa elevaram a média da renda familiar mensal per capita de R$48,69, antes do benefício, para R$72,42, resultando em um aumento de 48,7%. O impacto é diferenciado por região, sendo ainda mais significativo no Norte e no Nordeste, resultando em um aumento de 58,96% e 62,9% nessas regiões, respectivamente SENARC (2009).

Para identificar quais fatores permitem a inclusão de famílias no progrma Bolsa Família, além de quantificar a probablidade dessa inclusão, faz-se uso do modelo de regressão logística, que será explicado a seguir.

Capítulo 3

REGRESSÃO LOGÍSTICA

3.1 INTRODUÇÃO

Uma variável categorizada é aquela na qual a escala de medida consiste em um conjunto de categorias. Muitas variáveis respostas são categorizadas, por exemplo, a variável "gênero" que tem as categorias: masculino e feminino ou então a variável "possui filhos" que tem apenas as categorias sim ou não. Denota-se uma variável resposta binária por Y e suas possíveis categorias por 1(sucesso/sim) e 0 (falha/não). Esses dados binários são as formas mais comuns de representar os dados categorizados.

A distribuição de Y é especificada pela probabilidade $P(Y = 1) = \pi$ de sucessos e $P(Y = 0) = (1 - \pi)$ de falha. Logo, sua média é $E(Y) = \pi$. Para n observações independentes, o número de sucessos tem uma distribuição binomial com parâmetros (n, π), ou seja, cada observação binária é uma variável com distribuição de bernoulli (Agresti, 2002).

Para variáveis respostas binárias tem-se o modelo de probabilidade linear dado por:

$$\pi(x) = \alpha + \beta x \tag{3.1}$$

onde o parâmetro β representa a mudança na probabilidade por unidade de mudança em x.

A função logística é dada por:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3.2}$$

Tem-se, também, o modelo de regressão logística, onde a relação entre $\pi(x)$ e x normalmente é não-linear. Verifica-se que uma mudança fixa em x pode ter menos impacto quando π está perto de 0 ou 1 do que quando π está perto do meio da amplitude. Na prática, $\pi(x)$ ou cresce continuamente ou decresce continuamente quando x aumenta. A função matemática mais importante que descreve essa relação é a probabilidade de $\pi(x)$, que é obtida da fórmula de regressão logística dada em (3.2). Assim,

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}} \tag{3.3}$$

Supondo uma única variável explicativa x, para uma variável resposta binária y, e tendo que $\pi(x)$ denota a probabilidade de sucesso da variável x, o modelo de regressão logística tem uma forma linear para o *logit* de sua probabilidade. Então (3.3) também pode ser escrito como um modelo de regressão logística dado por Agresti (2002):

$$logit[\pi(x)] = log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \alpha + \beta x \tag{3.4}$$

3.2 Interpretação do Modelo de Regressão Logística

O modelo de regressão linear descrito em (3.4) indica que o *logit* aumenta de acordo com o β para cada aumento de uma unidade de x. Já o parâmetro β nas

duas equações determina a razão de aumento ou diminuição da curva de $\pi(x)$. O sinal de β indica se a curva é crescente ($\beta > 0$) ou decrescente ($\beta < 0$), e que a razão de chance aumenta assim que $|\beta|$ aumenta. Quando $\beta = 0$, então o lado direito da equação (3.3) simplifica para uma constante e, consequentemente, $\pi(x)$ é idêntico para toda x, e a curva se torna uma linha horizontal reta. Outro fato importante é que a variável resposta binária y fica independente de x (Agresti, 2002).

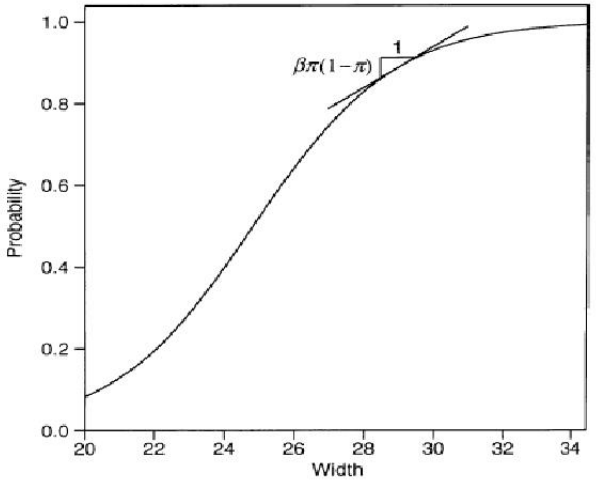

Figure 4.1. Linear approximation to logistic regression curve.

Figura 3.1: Aproximação Linear para a Curva de Regressão Logística

Fonte:Hosmer and Lemeshow (2000).

A análise dessa curva é dada por meio do parâmetro β da regressão logística, que pode ser valorado traçando uma reta tangente à curva, que descreverá a razão de chance nesse ponto. Essa linha é medida pelo seu ângulo com a superfície por meio da seguinte equação: $\beta\pi(x)[1 - \pi(x)]$. Quando $\pi(x) = 0.50$, ocorre o chamado *median effective level*, que representa o nível em que cada resultado tem 50% de chance. Esse evento é muitas vezes denotado por EL_{50} (Agresti, 2002).

Uma importante interpretação do modelo de regressão logística utiliza a razão entre a probabilidade de que um evento vai ocorrer e do que não vai ocorrer, mais conhecida como *odds*; e a razão de chance, *odds ratio* é a razão entre as *odds*. Para o modelo (3.4), a *odds* da reposta 1, de obter sucesso, é dada por:

$$\frac{\pi(x)}{1-\pi(x)} = exp(\alpha + \beta x) = exp^{\alpha}(e^{\beta})^{x} \tag{3.5}$$

Essa relação exponencial possibilita uma interpretação para β, isto é, a *odds* é multiplicada por e^{β} para cada uma unidade de aumento em x, significando que a *odds* no nível $x+1$ é igual a *odds* no nível x multiplicada por e^{β}. Se $\beta = 0$, então $e^{\beta} = 1$ e a *odds* não muda conforme o valor de x (Agresti, 2002).

Uma outra importante interpretação do modelo em estudo é quando há estudos retrospectivos, ou seja, quando a variável explicativa x em vez da variável resposta y é aleatória. Esse estudo é indicado para experimentos em que uma das categorias das respostas raramente ocorrem, e um estudo prospectivo talvez tenha tão poucos casos que impossibilite um bom estudo das estimativas dos efeitos dos preditores. Esses experimentos retrospectivos são normalmente utilizados em estudos de casos controle na área biomédica.

A grande questão desse estudo é que não é possível estimar efeitos em modelos binários com funções diversas, apenas com a função *logit*; outro aspecto é que diferentemente da razão de chance, o efeito para a distribuição condicional de x dado y não é igual para o de y dado x, o que fornece uma significante vantagem do uso da função *logit* sobre as outras (*probit*).

Um caso específico em que o modelo de regressão para Y descreve bem a relação,

independentemente do tipo de amostragem utilizado é quando tem-se uma distribuição X para o objeto em análise, tendo $Y = 1$ com uma distribuição normal, $N(\mu_1, \sigma)$ ou, ainda, uma distribuição de X para análise de $Y = 0$ tem uma distribuição normal,$N(\mu_0, \sigma)$, isto é, uma distribuição com médias diferentes, mas com o mesmo desvio padrão. Nesse caso, o teorema de Bayes nos fornece que de uma distribuição de X dado $Y = y$ para uma distribuição de Y dado $X = x$, tem-se que $P(Y = 1|x)$ satisfaz a curva da regressão logística. Para essa curva, o efeito de x é $\beta = \frac{\mu_1 - \mu_0}{\sigma^2}$. Em particular, β tem o mesmo sinal de $\mu_1 - \mu_0$ (Agresti, 2002).

Na próxima seção apresenta-se algumas inferências para os parâmetros do modelo, o que ajudará no julgamento da significância e do tamanho de seus efeitos.

3.3 Inferências para o Modelo de Regressão Logística

A primeira observação a ser feita é quanto ao caso dos dados binários agrupados e desagrupados. No primeiro caso, um conjunto de observações tem o mesmo valor do preditor das variáveis, como por exemplo quando as variáveis explicativas são discretas, e o modelo ajustado pode tratar as observações como um conjunto de sucessos de uma binomial retiradas de um certo tamanho de amostra, com a variada combinação dos valores dos preditos. No segundo caso, cada observação é uma única variável binária.

Um intervalo de confiança usado em modelos de regressão logística é o $Wald$, sendo geralmnte para grandes amostras. Para o parâmetro β da fórmula (3.4) é

dado por (Agresti, 2002):

$$\hat{\beta} \pm z_{\frac{\alpha}{2}}(EP) \tag{3.6}$$

onde (EP) é o erro padão.

Para obter um intervalo para e^{β}, basta calcular a exponencial dos pontos finais, ou seja, o efeito multiplicativo da *odds* de uma unidade aumenta em x.

Quando n é pequeno ou as probabilidades ajustadas são próximas de 0 ou 1, é preferível construir um intervalo de confiança baseado no teste de razão de verossimilhança. Esse intervalo contem todos os valores de β_0 para o teste de hipóteses, onde $H_0 : \beta = \beta_0$ com $p - valor > \alpha$.

Para obter um teste de significância para o modelo de regressão logística, onde $H_0 : \beta = 0$ significa que a probabilidade de sucesso é independente de X, tem-se o teste Wald para grandes amostras em que z tem um desvio padrão normal quando não se rejeita H_0. Logo,

$$z = \frac{\widehat{\beta}}{EP} \tag{3.7}$$

Equivalentemente, para $H_a : \beta \neq 0$, tem-se que z^2 segue uma distribuição qui-quadrado com 1 grau de liberdade (gl), onde

$$z^2 = \left(\frac{\widehat{\beta}}{EP}\right)^2 \tag{3.8}$$

Todavia, o teste de Wald é mais indicado para grandes amostras, enquanto que o teste de verossimilhança é mais poderoso e confiável para pequenas amostras, que são mais usadas na prática do dia-a-dia. Este teste compara o máximo de L_0 do log da função de verossimilhança quando $\beta = 0$ com o máximo de L_1 do log da função de verossimilhança para β irrestrito. Então, a estatística desse teste também tem

tem distribuição qui-quadrado com $gl = 1$ quando a amostra é grande e é calculado por:

$$D = -2(L_0 - L_1) \tag{3.9}$$

Para o modelo de regressão logística de $P(Y = 1)$ com um x fixo, tem-se que

$$\widehat{\pi(x)} = \frac{\hat{\alpha} + \hat{\beta}x}{1 + exp(\hat{\alpha} + \hat{\beta}x)} \tag{3.10}$$

onde, essa estimativa de $\pi(x)$ é considerada um intervalo de confiança para a probabilidade de $\pi(x)$, quando essa é verdadeira, dado um nível de significância.

Pode-se, entretanto, ao invés de usar $\widehat{\pi(x)}$ para ajustar o modelo, poderia simplesmente usar uma amostra proporcional para estimar a probabilidade, calculando: a proporção, o $EP = \sqrt{\frac{p(1-p)}{n}}$ e o intervalo de confiança. Logo, verifica-se que quando o modelo de regressão logística representa muito bem a relação entre $\pi(x)$ e x, então o estimador de $\widehat{\pi}(x)$ é bem melhor do que o estimador da proporção.

Como é difícil que apenas uma variável explique o fenôeno estudado, necessita-se fazer o estudo das possíveis variáveis que influenciem a variável em estudo. Desse fato, aborda-se na próxima seção o modelo com múltiplas variáveis.

3.4 Regressão Logística Múltipla

Nesta seção, objetiva-se generalizar o modelo logístico para o caso em que se tem mais de uma variável independente, mais conhecido como o caso multivariado, ou ainda, regressão logística múltipla. A seguir faz-se considerações sobre a estimação dos coeficientes no modelo, bem como de sua significância.

Considere uma coleção de p variáveis independentes denotada pelo vetor $X' =$

$(x_1, x_2, x_3, ..., x_p)$. A princípio assume-se que as variáveis analisadas são escalares intervalar. Supondo que a probabilidade condicional que a resposta desejada está presente é expressa por $P(Y = 1|\mathbf{x}) = \pi(\mathbf{x})$, o logit para o modelo de regressão logística múltipla é dado por (Hosmer and Lemeshow, 2000):

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \tag{3.11}$$

onde, o modelo de regressão logística é dado por:

$$\pi(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}} \tag{3.12}$$

Há casos em que as variáveis independentes são discretas, nominais escalares (como raça, gênero) em que é inapropriado incluí-las no modelo como se fossem variáveis escalares intervalar, pois os números usados para representar os diversos níveis dessa variável nominal são meramente identificadores, não tendo, então, nenhum significado numérico. Nessa situação, o método alternativo é usar variáveis *dummy*, ou seja, usa-se uma codificação para essas variáveis. Suponha,por exemplo, a variável gênero: quando a resposta é feminino, F é 0 e, masculino, M é 1.

Em geral, se a variável nominal escalar tem k valores possíveis, então será necessário $k - 1$ variáveis *dummy*, salvo quando todos os modelos tem um termo constante. Denota-se o modelo logit para p tipos dessa variável *dummy* e a j-ésima variável sendo discreta por(Hosmer and Lemeshow, 2000):

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + ... + \sum_{l=1}^{k_j - 1} \beta_{jl} D_{jl} + \beta_p x_p \tag{3.13}$$

onde, a j-ésima variável independente x_j tem k_j níveis e as $k_j - 1$ variáveis *dummy* são representadas por D_{jl} e seus coeficientes por β_{jl}, para todo $l = 1, 2, ..., k_j - 1$.

O modelo para ser ajustado requer a obtenção dos estimadores do vetor β'=$(\beta_0, \beta_1, ..., \beta_p)$, dado uma amostra de n observações independentes $(\mathbf{x}_i, \mathbf{y}_i)$, para todo $i = 1, 2, ..., n$. O método usado será o de Máxima Verossimilhança, como explicado na situação univariada. A única diferença será na fórmula do $\pi(\mathbf{x})$. Logo, existirão $p + 1$ equações de verossimilhança que são obtidas pela diferenciação do logaritmo da função de máxima verossimilhança com respeito aos $p+1$ coeficientes. As equações de verossimilhança obtidas por esse processo são:

$$\sum_{i=1}^{n}[y_i - \pi(\mathbf{x}_i)] = 0 \quad e \quad \sum_{i=1}^{n} x_{ij}[y_i - \pi(\mathbf{x}_i)] = 0 \tag{3.14}$$

para $j = 1, 2, ..., p$

A solução dessas equações será representada por $\widehat{\beta}$. O valor ajustado será denotado por $\widehat{\pi(x)}$. Além disso, tem-se que o erro padrão estimado dos coeficientes estimados é dado por:

$$\widehat{EP(\widehat{\beta}_j)} = [\widehat{Var}(\widehat{\beta}_j)]^{1/2} \tag{3.15}$$

para $j = 1, 2, ..., p$

A notação matricial será útil quando discutido o ajuste do modelo, que é $\widehat{\mathbf{I}}(\widehat{\beta})$=$\mathbf{X'VX}$, onde $\mathbf{X}$ é uma matriz $n\ x\ p + 1$; $\mathbf{V}$ é uma matriz diagonal $n\ x\ n$ com elementos $\widehat{\pi}(1 - \widehat{\pi})$. A matriz $\mathbf{X}$ é dada por

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

e a matriz $\mathbf{V}$ é

$$\mathbf{V} = \begin{bmatrix} \widehat{\pi}_1(1-\widehat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \widehat{\pi}_2(1-\widehat{\pi}_2) & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & 0 & \widehat{\pi}_n(1-\widehat{\pi}_n) \end{bmatrix}$$

Para testar a significância do modelo, podemos utilizar o teste da razão de verossimilhança nos coeficientes das variáveis independentes. O teste é realizado na mesma maneira que no caso univariado onde a estatística do teste G é dada pela equação

$$G = -2ln\left[\frac{Verossimilhanca\ sem\ a\ variavel}{Verossimilhanca\ com\ variavel}\right] \tag{3.16}$$

Os valores ajustados, $\widehat{\pi}$, do modelo estão em um vetor de $p+1$ parâmetros, $\widehat{\beta}$. Sob a hipótese nula G tem uma distribuição qui-quadrado com p graus de liberdade. Para se certificar que algum ou todos coeficientes sejam diferentes de zero, pode-se utilizar o teste de Wald citado na seção anterior.

O Teste de Wald multivariado é obtido por

$$W = \widehat{\beta}'\left[\widehat{Var}(\widehat{\beta})\right]^{-1}\widehat{\beta} = \widehat{\beta}'\left(\mathbf{X}'\mathbf{V}\mathbf{X}\right)\widehat{\beta}, \tag{3.17}$$

onde W tem distribuição qui-quadrado com $p+1$ graus de liberdade e sob a hipótese nula de cada um dos $p+1$ coeficientes são iguais a zero.

Os métodos de estimação de intervalo de confiança para o caso multivariado são essencialmente os mesmos do caso univariado, somente ampliados pelo fato de haver mais do que uma variável independente no modelo. A estimação do intervalo de confiança do logito não é tão trivial quanto a estimação no caso univariado,

portanto o uso de matrizes facilita a obtenção do intervalo. A expressão do logito

$$\widehat{g}(\mathbf{x}) = \mathbf{x}'\widehat{\beta}$$

onde $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, ..., \widehat{\beta}_p)$ denota o estimador de $p+1$ coeficientes e o vetor $\mathbf{x}' = (x_0, x_1, x_2, ..., x_3)$ representa a constante e o conjunto de valores das p-covariáveis do modelo onde $x_0 = 1$.

Pode-se expressar a estimativa da variância do vetor β assim

$$\widehat{Var}(\widehat{\beta}) = (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1} \tag{3.18}$$

logo pode-se expressar a estimativa da variância do logito como

$$\widehat{Var}(\widehat{g}(\mathbf{x})) = \mathbf{x}'\widehat{Var}(\widehat{\beta})\mathbf{x} = \mathbf{x}'(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{x}. \tag{3.19}$$

Capítulo 4

MATERIAL E MÉTODOS

4.1 Introdução

O presente trabalho objetiva identificar quais são as características que uma família deve ter para ser beneficiária do Programa Bolsa Família, dado que ela está inscrita no CadÚnico. Isso significa que, dado que a família está em uma condição de vulnerabilidade social, quais fatores são importantes e fundamentais na seleção para receber algum benefício do PBF.

A identificação dessas características será feita utilizando diversas variáveis da base de dados, como a renda, o número de filhos, escolaridade, entre outros. Portanto, como várias variáveis interferem no modelo e muitas delas são independentes, temos um caso multivariado.

A regressão logística múltipla, permite criar um modelo que estime a probabilidade de uma família cadastrada no CadÚnico, ser beneficiária do Programa Bolsa Família, dadas certas características da família. Assim, a partir de uma análise exploratória dos dados serão selecionadas as variáveis mais importantes para o modelo em questão. Em seguida será construído o modelo de regressão logística múltipla

baseado nas variáveis escolhidas e que deve ter a forma da equação (3.12).

4.2 Base de Dados - CadÚnico

O Cadastro Único para Programas sociais, regulamentado pelo Decreto n.º 6.135/07 e coordenado pelo Ministério do Desenvolvimento Social e Combate à Fome (MDS), é uma base de dados que deve ser obrigatoriamente utilizada para seleção de beneficiários e integração de programas sociais do Governo Federal, como o Bolsa Família. Suas informações também podem ser utilizadas pelos governos estaduais e municipais para obter o diagnóstico socioeconômico das famílias cadastradas, possibilitando a análise das suas principais necessidades. Atualmente o Cadastro Único conta com mais de 19 milhões de famílias inscritas (MDS, 2010).

Esta base de dados, conhecida também como CadÚnico, é um instrumento que identifica e caracteriza as famílias com renda mensal de até meio salário mínimo per capita ou de três salários mínimos no total (MDS, 2010).

A base contém informações:

- Do núcleo familiar, como por exemplo o número de pessoas na família;
- Das características do domicílio, como por exemplo o tipo de material usado na sua construção;
- Das formas de acesso a serviços públicos essenciais, como por exemplo a presença ou não de coleta de lixo e abastecimento de água por rede pública;

- Das informações de cada pessoa da família, como por exemplo a idade, escolaridade, raça/cor.

O cadastramento das famílias possibilita ao poder público formular e implementar políticas específicas, que possam contribuir para a redução das vulnerabilidades sociais a que essas famílias estão expostas (MDS, 2010).

A base de dados utilizada será a base de dados do Cadastro Único referente a julho de 2010, cedida pelo Ministério do Desenvolvimento Social e Combate à Fome. Basicamente está dividida em quatro tabelas principais. São elas: Tabela de Domicílios, Tabela de Pessoas, Tabela de Agricultor Rural e Folha de Pagamentos. Neste trabalho será usado algumas variáveis da Tabela de Domicílios, da Tabela de Pessoas e da Folha de Pagamentos.

4.3 Calibração do Modelo

Primeiramente será feita a análise exploratória dos dados, buscando identificar a distribuição espacial das famílias cadastradas e beneficiárias do Program Bolsa Família. Alguns mapas serão utilizados na ilustração das diferenças regionais dessas características. As variáveis análisadas foram escolhidas por meio de um conhecimento prévio sobre a base de dados e sobre o programa.

Em seguida, será feita a seleção das variáveis do modelo por meio do método *Stepwise*, que tem o objetivo de identificar as variáveis de maior influência. A partir das variaveis selecionadas estarão definidas quais caracteristicas são importantes para calcular a probabilidade de uma família ser atendida pelo Program Bolsa Família.

Posteriormente, será feita a análise do modelo utilizando a *odds ratio*. Serão verificadas a validade e a qualidade do modelo logístico. Ressalta-se que a análise do modelo logístico multivariado terá o vetor $\boldsymbol{\beta}$ como um parâmetro e não como uma estatística, pois a base dados utilizada é censitária.

As variáveis em análise estão citadas em grupos: domicílio, raça, trabalha/aposentado, construção, localidade, família, região, outros/serviços básicos, para facilitar o entendimento e são:

- DOMICILIO_ALUGADO: indica que o domicílio onde a família reside é alugado;
- DOMICILIO_ARRENDADO: indica que o domicílio onde a família reside é arrendado;
- DOMICILIO_CEDIDO: indica que o domicílio onde a família reside é cedido;
- DOMICILIO_FINANCIADO: indica que o domicílio onde a família reside é financiado;
- DOMICILIO_INVASAO: indica que o domicílio onde a família reside é invasão;
- DOMICILIO_PROPRIO: indica que o domicílio onde a família reside é próprio;
- AMARELO: quantidade de pessoas que se declararam amarelas na família;
- BRANCO: quantidade de pessoas que se declararam brancas na família;
- INDIO: quantidade de pessoas que se declararam indígenas na família;
- NEGRO: quantidade de pessoas que se declararam negras na família;

- PARDO: quantidade de pessoas que se declararam pardas na família;
- VL_RENDA_PERCAPITA: indica a renda percapita da família sem contar o valor do benfício se a família for beneficiária;
- APOSENTADO: indica quantas pessoas na família declararam que são aposentadas;
- N_TRABALHA: indica quantas pessoas na família declararam que não trabalham;
- OUTRO: indica quantas pessoas na família declararam alguma outra situação no mercado de trabalho além das citadas acima;
- CD_TIPO_LOCALIDADE: indica se a família mora em região urbana ou rural;
- CONTRUCAO_NAO_INFOR: indica que a família não declarou o tipo de material utilizado na construção do domicílio onde reside;
- NU_COMODOS: indica a quantidade de cômodos do domicilio da família;
- MORADOR_RUA: indica se a família é moradora de rua;
- BEBE_NA_FAMILIA: indica a presença de bebê na família;
- DEFICIENTE_NA_FAMILIA: indica a presença de pessoa com deficiência na família;
- N_CRIANCAS: quantidade de crianças na família, de 0 a 17 anos;
- N_CRIANCAS_IDADE_ESC: quantidade de crianças de 7 a 17 anos;

- N_CRIANCA_ESCOLA: quantidade de crianças que frequentam a escola;
- QT_PESSOAS_CALCULADA: indica o número de pessoas na família;
- REGIAO: indica a região em que a família está cadastrada;
- CD_IBGE: código do Município onde a família está cadastrada;
- SERVICOS_BASICOS_ADEQUADOS: indica que o domicílio onde a família reside possui abastecimento de água por rede pública, água tratada por filtração, iluminação por relógio próprio, coleta de lixo e escoamento sanitário por rede pública ou fossa séptica;

Capítulo 5

ANÁLISE DOS RESULTADOS

Este Capítulo analisára a base de dados e os resultados do modelo logístico, de acordo com o método proposto anteriormente.

5.1 Análise Exploratória dos Dados

Para ter uma idéia do tamanho da base foi calculado por região o número total de pessoas cadastradas, o número total de famílias cadastradas e o número total de famílias beneficiárias. Em julho de 2010, o programa atendia mais de 12 milhões de famílias, ou seja 61,54% das famílias cadastradas estavam sendo assistidas pelo PBF. Esses totais estão na Tabela 5.1.

Tabela 5.1: Tabela de Frequências de pessoas e famílias beneficiárias e cadastradas por Região

Regiões	Pessoas Cadastradas	Famílias Cadastradas	Famílias Beneficiárias
Norte	7.606.326	1.959.389	1.337.881
Nordeste	32.437.333	9.256.792	6.359.947
Sudeste	7.586.966	2.137.828	1.072.185
Sul	4.471.297	1.251.367	643.081
Centro Oeste	20.797.337	5.711.697	3.089.912
Brasil	72.899.259	20.317.073	12.503.006

Complementando a análise da distribuição territorial das famílias cadastradas, verificou-se que aproximadamente 70% delas vivem em área urbana enquanto que apenas 30% delas vive em área rural. Um fato interessante é que essa proporção é

mantida se for analisado somente as famílias beneficiárias.

Outra análise importante é a da renda média per capita das famílias cadastradas comparada com a renda das famílias beneficiárias, dada pela Tabela 5.1. Como esperado, em todos os estados a média da renda percapita das famílias cadastradas é maior do que quando se considera somente as famílias beneficiárias.

Tabela 5.2: Renda média per capita das famílias cadastradas e das famílias beneficiárias

Estado	**Famílias Cadastradas**	**Famílias Beneficiárias**
Rondônia	95,73	55,86
Acre	106,32	72,08
Amazonas	71,80	42,28
Roraima	93,54	43,96
Pará	76,62	46,80
Amapá	100,60	47,12
Tocantins	152,28	59,39
Maranhão	64,75	38,97
Piauí	103,38	39,07
Ceará	82,05	48,45
Rio Grande do Norte	143,53	50,21
Paraíba	121,05	43,56
Pernambuco	83,58	45,37
Alagoas	69,29	48,19
Sergipe	92,57	48,94
Bahia	96,75	48,67
Minas Gerais	153,91	67,50
Espirito Santo	117,49	69,05
Rio de Janeiro	160,85	56,87
São Paulo	182,08	72,76
Paraná	188,29	77,46
Santa Catarina	218,20	79,05
Rio Grande do Sul	196,27	65,33
Mato Grosso do Sul	118,00	71,29
Mato Grosso	139,77	66,97
Goiás	165,62	84,53
Distrito Federal	195,08	67,92

A Figura 5.1 foi feita com o objetivo de melhor representar a distribuição da renda entre as famílias cadastradas, onde a cor mais escura representa quais regiões possuem maior renda média per capita das famílias do Cadastro Único e, a cor mais clara, as regiões de menor renda.

A análise da Figura 5.2 foi feita em relação ao número de cômodos das famílias do cadastro. Verifica-se uma variação de aproximadamente 1 a 7 cômodos. Novamente,

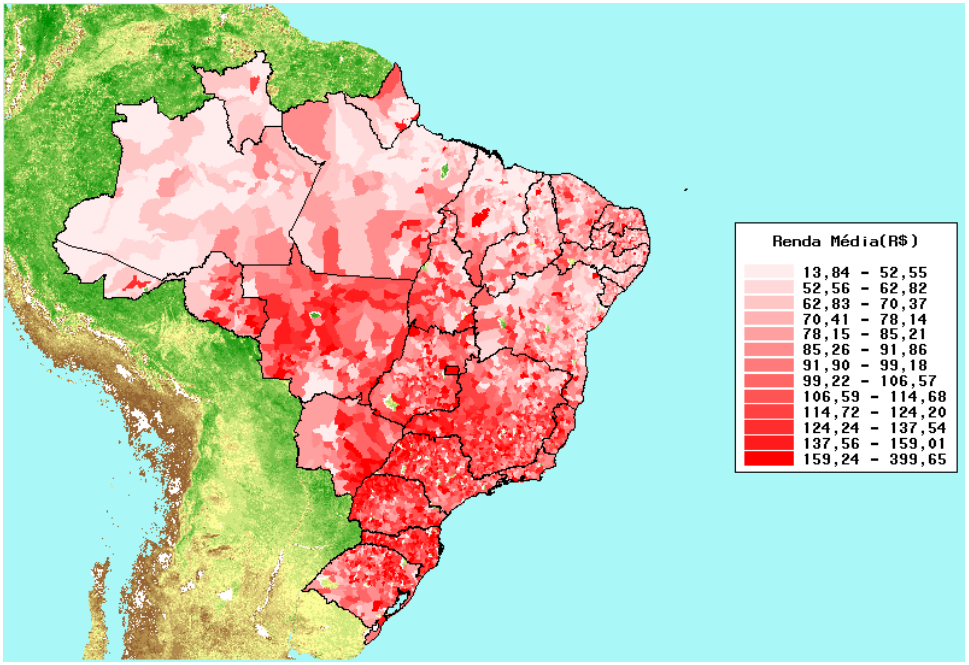

Figura 5.1: Distribuição da Renda Média Per Capita

a cor mais forte indica uma região com maior quantidade da variável analisada e, a mais fraca, uma menor quantidade.

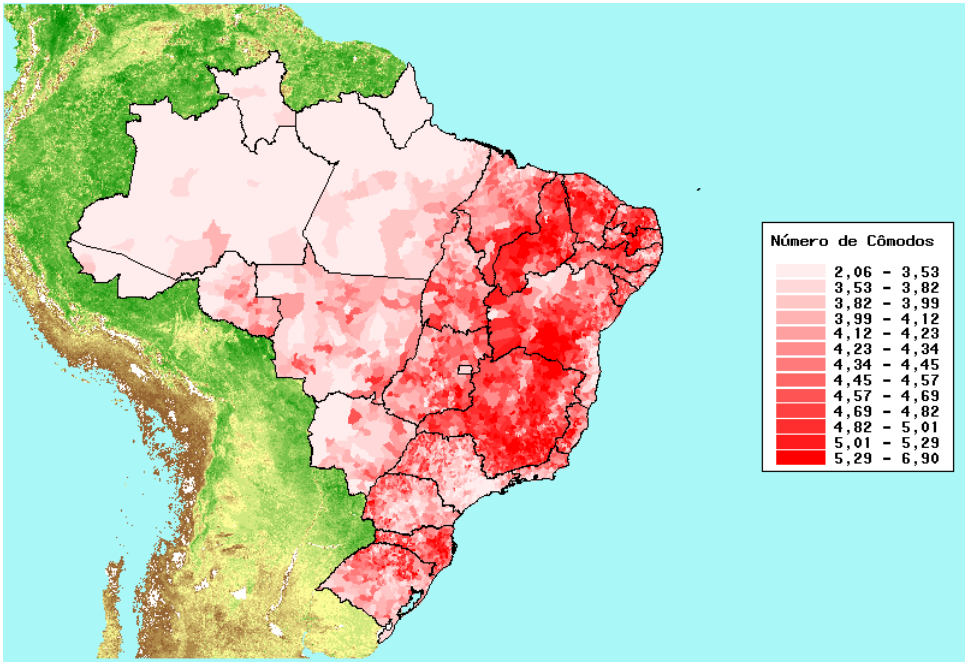

Figura 5.2: Número de Cômodos por Domicílio

A Figura 5.5 representa a distribuição do número médio de pessoas nas famílias cadastradas, sendo que a variação desse número é de aproximadamente 2 a 7 pessoas. Tanto para as famílias cadastradas, como quando se considera somente as famílias beneficiárias, a composição familiar típica é de 3 ou 4 pessoas, ou seja, aproximadamente 50% das famílias tem 3 ou 4 pessoas em sua composição.

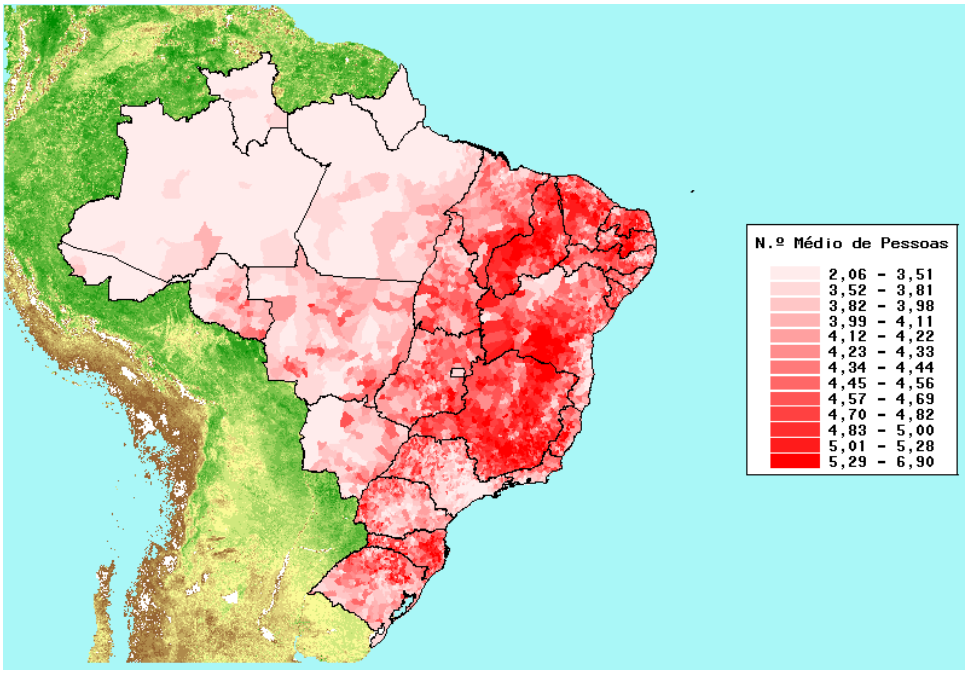

Figura 5.3: Número Médio de Pessoas por Domicílio

A Figura 5.4 mostra que as regiões Nordeste e Norte apresentam maior número de domicílios próprios, o que pode ser explicado pelo tipo de casa construída, como por exemplo o tapume.

Por meio da análise do banco de dados consegui-se extrair a Tabela 5.1, indicando o número de famílias que poderiam ser atendidas pelo Programa Bolsa Família e estão fora deste.

Nessa tabela, os seis estados que apresentam o maior número de famílias

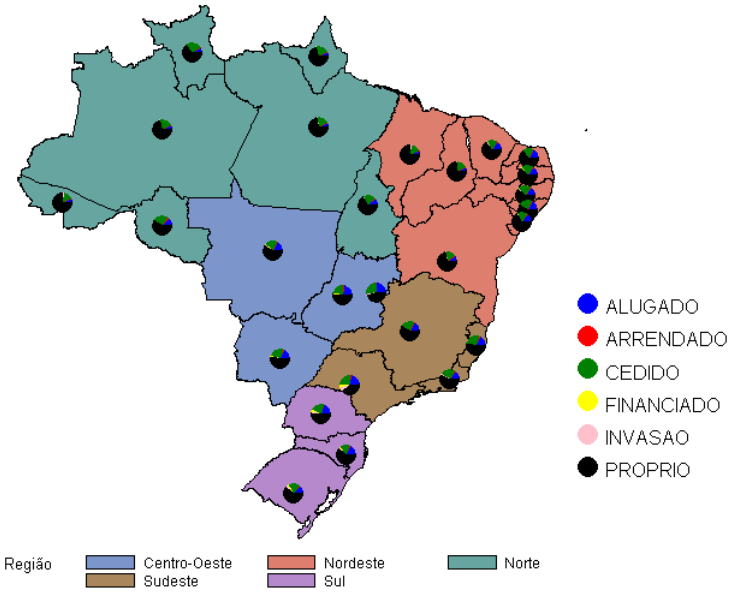

Figura 5.4: Tipo de Domicílio por Região

Tabela 5.3: Número de famílias que não receberam o PBF mas tem as características exigidas

UF	Perfil Benefício Variável	Perfil Benefício Básico	Perfil PBF
RO	3.644	10.015	13.659
AC	9.297	20.156	29.453
AM	20.311	53.370	73.681
RR	15.696	24.904	40.600
PA	48.189	463.366	211.555
AP	5.159	21.431	26.590
MA	42.038	171.258	213.296
PI	33.654	83.279	116.933
CE	93.015	150.321	243.336
RN	38.977	63.749	102.726
PB	40.285	103.036	143.321
PE	84.443	209.221	293.664
AL	29.184	54.086	83.270
SE	18.281	46.023	64.304
BA	138.317	340.411	478.728
MG	276.788	244.504	521.292
ES	41.374	47.686	89.060
RJ	93.366	137.831	231.197
SP	295.709	311.267	606.976
PR	126.314	90.889	217.203
SC	54.296	39.744	94.040
RS	91.987	111.831	203.818
MS	24.135	23.719	47.854
MT	36.254	32.099	68.353
GO	67.243	66.268	133.511
DF	46.686	61.551	108.237
TOTAL	1.793.680	2.704.263	4.497.943

passíveis de serem assistidas pelo programa são: SP(606.976), MG(521.292), BA(478.728), PE(293.664), CE(243.336), RJ(231.197). Dentre esses, os que apresentam maior número de benefício variável são SP(295.709), MG(276.788), BA(138.317), ressaltando-se ainda PR(126.314); o de benefício básico são PA(463.66), BA(340.411), SP(311.267), MG(244.504), PE(209.221), MA(171.258), sendo expressivo também os estados do CE(150.321), RJ(137.831) e RS(111.831). Essa realidade é melhor representada na Figura 5.5.

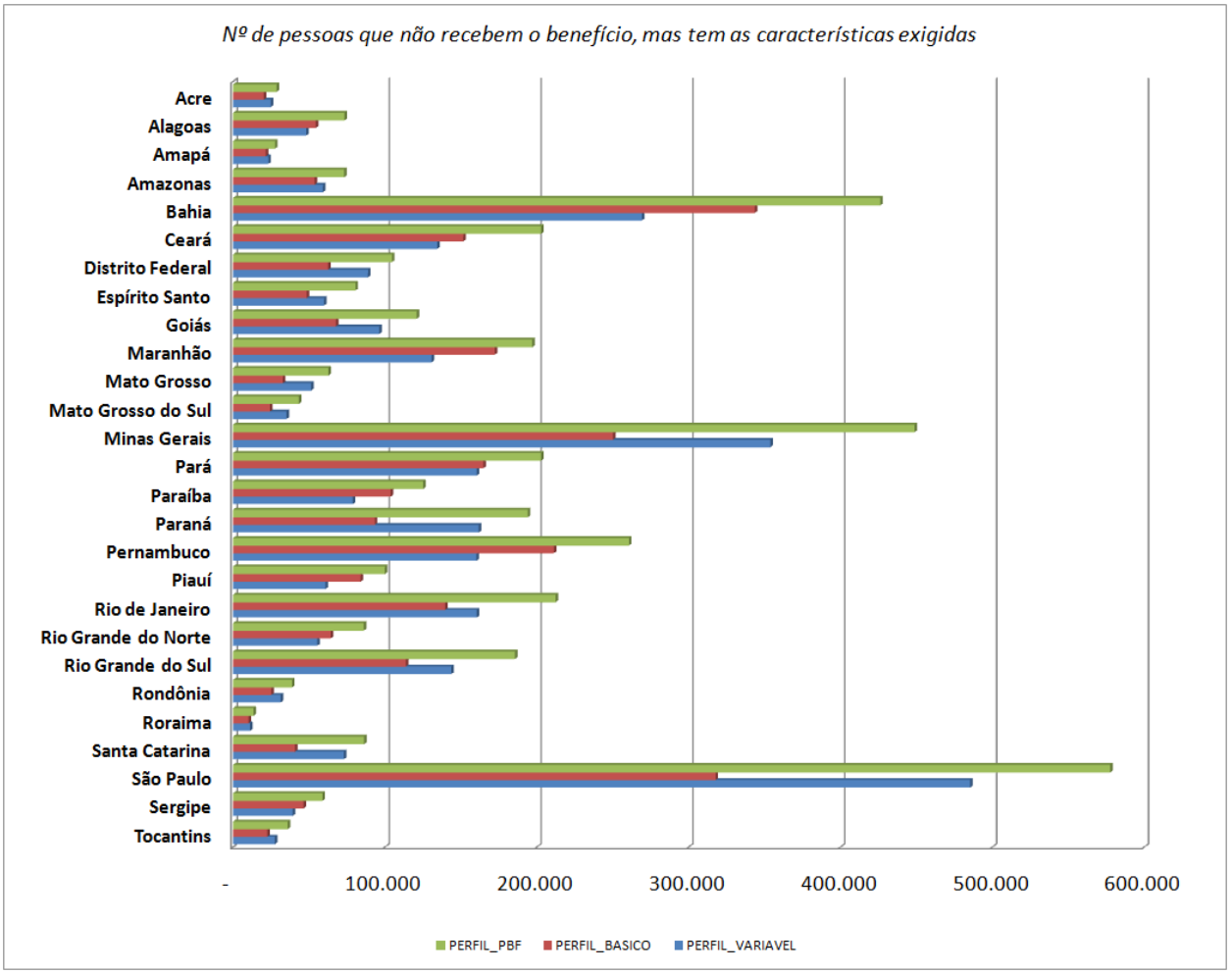

Figura 5.5: Gráfico de Barras das possíveis famílias a serem beneficiadas

A Figura 5.6 indica a proporção entre as famílias com e sem o benefício do Bolsa

Família nas regiões do Brasil. Dentre essas, o Nordeste e o Norte apresentam o maior número com o Bolsa Família, enquanto que o DF é a Unidade da Federação que tem maior proporção de famílias sem o Bolsa Família.

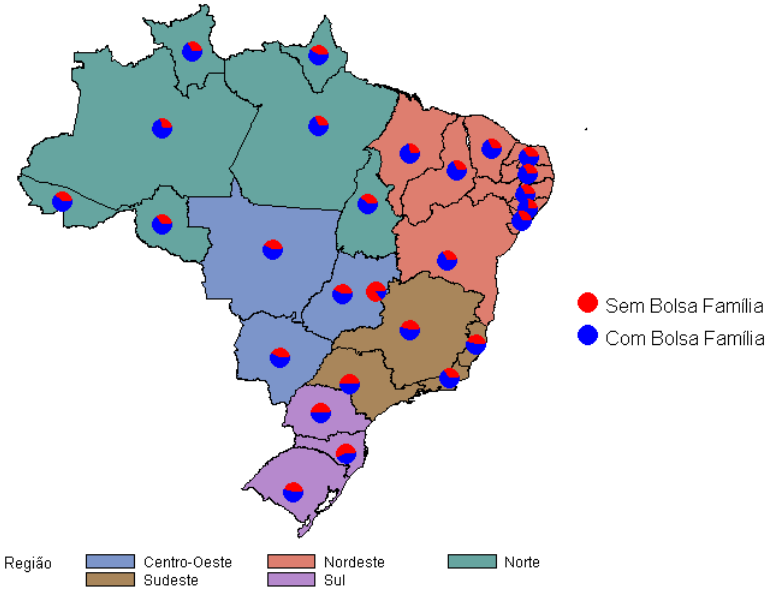

Figura 5.6: Distribuição dos Beneficários do Programa Bolsa Família, por Estado

Para atender ao objetivo de constatar a eficácia do programa Bolsa Família e o impacto na renda dos beneficiários, a Tabela 5.4 apresenta o aumento médio e o percentual que o programa tem na renda percapita famíliar dos beneficiários.

Tabela 5.4: Aumento da renda dos beneficiários do PBF

Estado	Aumento médio	Percentual
Rondônia	97,14	173,87%
Acre	108,92	151,09%
Amazonas	109,49	258,96%
Roraima	109,67	249,45%
Pará	107,39	229,45%
Amapá	111,63	236,87%
Tocantins	94,29	158,74%
Maranhão	104,54	268,24%
Piauí	98,89	253,06%
Ceará	97,40	201,00%
Rio Grande do Norte	95,24	189,66%
Paraíba	97,71	224,29%
Pernambuco	97,38	214,61%
Alagoas	98,71	204,83%
Sergipe	99,73	203,76%
Bahia	98,25	201,84%
Minas Gerais	87,77	130,02%
Espirito Santo	88,12	127,61%
Rio de Janeiro	91,80	161,42%
São Paulo	85,92	118,08%
Paraná	82,34	106,29%
Santa Catarina	82,96	104,94%
Rio Grande do Sul	89,46	136,92%
Mato Grosso do Sul	89,50	133,63%
Mato Grosso	88,09	123,56%
Goiás	87,97	104,07%
Distrito Federal	89,54	131,83%

5.2 Procedimento *Stepwise* - Seleção das variáveis do modelo

Depois de uma pré-seleção das variáveis via análise exploratória, onde se percebeu as características mais importantes das famílias cadastradas, utilizou-se o método de seleção *Stepwise* para determinar quais variáveis iniciais são significativas para o modelo logístico em estudo.

O procedimento de seleção é baseado em um algoritmo que verifica a importância das variáveis, incluindo ou excluindo-as por meio da medida de significância do coeficiente associado à variável para o modelo. Essa estatística depende das suposições do modelo. Na regressão logística os erros seguem distribuição binomial e a significância é assegurada via Teste da Razão de Verossimilhança. Assim, em cada passo do procedimento a variável mais importante é aquela que produz a maior

mudança no logaritmo da verossimilhança em relação ao modelo que não contém a variável.

Utilizando o *software* SAS seleciou-se as variáveis estabelecendo um critério onde o $p-valor$ associado ao teste da razão de verossimilhança é de 0,05; ou seja, a cada inclusão de uma variável foi analisado se o aumento no log da verossimilhança do novo modelo em relação ao anterior é significativo ao nível de 5%.

As variáveis selecionadas estão descritas na Tabela 5.2.

Tabela 5.5: Tabela de variáveis selecionadas no *Stepwise*

Passo	**Entrada**	**GL**	**Score Qui-Quadrado**	**P-valor**
1	N_CRIANCAS	1	3057020.94	< .0001
2	APOSENTADO	1	786.794.279	< .0001
3	CD_TIPO_LOCALIDADE	1	433.549.355	< .0001
4	TRABALHA	1	289.390.913	< .0001
5	PARDO	1	205.370.151	< .0001
6	CONTRUCAO_NAO_INFOR	1	132.130.149	< .0001
7	N_CRIANCA_ESCOLA	1	746.512.547	< .0001
8	N_CRIANCAS_IDADE_ESC	1	237.841.146	< .0001
9	DOMICILIO_ALUGADO	1	383.914.418	< .0001
10	NEGRO	1	259.702.454	< .0001
11	DOMICILIO_FINANCIADO	1	215.579.710	< .0001
12	QT_PESSOAS_CALCULADA	1	106.357.732	< .0001
13	INDIO	1	70.993.608	< .0001
14	BRANCO	1	126.127.335	< .0001
15	SERVICOS_BASICOS_ADE	1	50.972.522	< .0001
16	AMARELO	1	42.003.434	< .0001
17	DOMICILIO_INVASAO	1	35.719.471	< .0001
18	MORADOR_RUA	1	30.153.250	< .0001
19	N_TRABALHA	1	28.797.500	< .0001
20	NU_COMODOS	1	21.182.878	< .0001
21	DOMICILIO_CEDIDO	1	13.084.668	< .0001
22	DEFICIENTE_NA_FAMILI	1	6.674.583	< .0001
23	VL_RENDA_PERCAPITA	1	6.308.693	< .0001
24	DOMICILIO_ARRENDADO	1	3.331.159	< .0001
25	BEBE_NA_FAMILIA	1	422.463	< .0001

5.3 Análise do Modelo

Nesse capítulo apresenta-se a análise do modelo logístico do Programa Bolsa Família. Verificou-se que a frequência total de não beneficiários e beneficiários é de, respectivamente, 8.069.163 e de 12.658.567; ou seja, da base de dados, 38.93% não recebem o benefício enquanto 61.07% recebem.

Tabela 5.6: Hipótese nula: $\boldsymbol{BETA}$=0

Teste	Qui-quadrado	gl	P-valor
Razão de verossimilhança	7958695.43	25	<.0001
Score	5038763.86	25	<.0001
Wald	4064091.02	25	<.0001

A Tabela 5.6 mostra que o modelo testado existe, utilizando o teste da razão de verossimilhança. Os testes Score e Wald convergem assintoticamente para o teste da razão de verossimilhança.

Tabela 5.7: Análise das estimativas de máxima verossimilhança

Paramêtro	gl	Estimativa	Erro padrão	Qui-quadrado Wald	P-valor
Intercepto	1	0.4487	0.00270	27525.8879	<.0001
DOMICILIO_ALUGADO	1	−0.1342	0.00167	6493.2209	<.0001
DOMICILIO_ARRENDADO	1	−0.1702	0.00933	332.4938	<.0001
DOMICILIO_CEDIDO	1	−0.0582	0.00141	1693.3139	<.0001
DOMICILIO_FINANCIADO	1	−0.4303	0.00523	6773.3956	<.0001
DOMICILIO_INVASAO	1	−0.1528	0.00459	1106.8838	<.0001
AMARELO	1	0.2937	0.00452	4226.9221	<.0001
BRANCO	1	0.2432	0.00130	35096.5426	<.0001
INDIO	1	0.4158	0.00327	16166.0337	<.0001
NEGRO	1	0.3161	0.00144	48067.2454	<.0001
PARDO	1	0.3291	0.00127	66867.0633	<.0001
VL_RENDA_PERCAPITA	1	−0.0137	0.000012	1361016.10	<.0001
APOSENTADO	1	−0.4455	0.00195	52421.5617	<.0001
N_TRABALHA	1	−0.0290	0.00110	694.7969	<.0001
OUTRA	1	0.2266	0.00123	33901.1187	<.0001
CD_TIPO_LOCALIDADE	1	0.3760	0.00139	73637.2274	<.0001
CONTRUCAO_NAO_INFOR	1	−1.5914	0.00409	151263.429	<.0001
NU_COMODOS	1	0.0113	0.000321	1249.5675	<.0001
MORADOR_RUA	1	0.3726	0.0114	1075.3800	<.0001
BEBE_NA_FAMILIA	1	−0.0144	0.00222	42.2450	<.0001
DEFICIENTE_NA_FAMILI	1	0.0402	0.00262	235.7853	<.0001
N_CRIANCAS	1	0.6411	0.00120	285477.084	<.0001
N_CRIANCAS_IDADE_ESC	1	−0.6271	0.00140	200468.134	<.0001
N_CRIANCA_ESCOLA	1	0.7517	0.00130	336450.820	<.0001
QT_PESSOAS_CALCULADA	1	−0.3511	0.00158	49179.1433	<.0001
SERVICOS_BASICOS_ADE	1	−0.0389	0.00144	729.3353	<.0001

A Tabela 5.7: mostra a ordem das variáveis que mais contribuem para o recebimento do Bolsa Família e aponta a significância dos coeficientes das variáveis explicativas do modelo. O modelo de regressão logística múltipla 3.11 foi obtido em 25 passos, e pode ser escrito como:

$$logit[\pi(x)] = g(\mathbf{x}) = 0.4487 + 0.2937 \times Amarelo - 0.0137 \times Renda - 0.4455 \times Aposentado$$
$$-0.0144 \times Bebe + 0.2432 \times Branco + 0.3760 \times Tipolocalidade - 1.5914 \times construcaonaoinform$$
$$+0.0402 \times Deficientenafamilia - 0.1342 \times Domicilioalugado - 0.1702 \times Domicilioarrendado$$
$$-0.0582 \times Domiciliocedido + 0.3726 \times Moradorderua - 0.4303 \times Domiciliofinanciado$$
$$-0.1528 \times Domicilioinvasao + 0.4158 \times Indio + 0.3161 \times Negro + 0.0113 \times Nucomodos$$
$$+0.6411 \times Ncriancas - 0.6271 \times Ncriancasidadeescolar + 0.7517 \times Ncriancasescola$$
$$-0.0290 \times Ntrabalha + 0.3291 \times Pardo - 0.3511 \times Qtpessoascalculada + 0.2266 \times Outraocupacao$$
$$- 0.0389 \times Servicosbasicosadequados$$

Então, o modelo de regressão logístico 3.12 é dado por:

$$\pi(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}}$$

O cálculo do $\pi(\mathbf{x})$ é o que mostra a probabilidade da família pertencer ao programa. Por exemplo, uma família de 7 pessoas pardas, pai, mãe e 5 filhos,sendo 1 deles um bebe e 3 deles em idade escolar e na escola, com o pai aposentado, renda de R$100,00 e que mora de alugel em uma casa com 5 comodos, teria a probabilidade de pertencer ao programa dada por:

$$g(\mathbf{x}) = 0.4487 - (0.0137 \times 100) - (0.6721 \times 1) - (0.0144 \times 1) - (0.1342 \times 1) + (0.0113 \times 5)$$

$$+(0.6411 \times 5) - (0.6271 \times 3) + (0.7517 \times 3) + (0.3291 \times 7) - (0.3511 \times 7) = 1,7398$$

$$\pi(\mathbf{x}) = \frac{e^{1,7398}}{1 + e^{1,7398}} = 0,85066$$

Ou seja, a família com as características descritas tem uma probabilidade de 85,066% de ser beneficiária do Programa Bolsa Família.

Tabela 5.8: Estimativas de razão de chances

Efeito	Estimativa	Intervalo de confiança de 95 de Wald	
DOMICILIO_ALUGADO	0.874	0.872	0.877
DOMICILIO_ARRENDADO	0.843	0.828	0.859
DOMICILIO_CEDIDO	0.944	0.941	0.946
DOMICILIO_FINANCIADO	0.650	0.644	0.657
DOMICILIO_INVASAO	0.858	0.851	0.866
AMARELO	1.341	1.330	1.353
BRANCO	1.275	1.272	1.279
INDIO	1.516	1.506	1.525
NEGRO	1.372	1.368	1.376
PARDO	1.390	1.386	1.393
VL_RENDA_PERCAPITA	0.986	0.986	0.986
APOSENTADO	0.641	0.638	0.643
N_TRABALHA	0.971	0.969	0.973
OUTRO	1.254	1.251	1.257
CD_TIPO_LOCALIDADE	1.456	1.452	1.460
CONTRUCAO_NAO_INFOR	0.204	0.202	0.205
NU_COMODOS	1.011	1.011	1.012
MORADOR_RUA	1.452	1.420	1.484
BEBE_NA_FAMILIA	0.986	0.981	0.990
DEFICIENTE_NA_FAMILI	1.041	1.036	1.046
N_CRIANCAS	1.899	1.894	1.903
N_CRIANCAS_IDADE_ESC	0.534	0.533	0.536
N_CRIANCA_ESCOLA	2.121	2.115	2.126
QT_PESSOAS_CALCULADA	0.704	0.702	0.706
SERVICOS_BASICOS_ADE	0.962	0.959	0.965

A Tabela 5.8 apresenta as estimativas da razão de chance. O valor 1.456 para a variável CD_TIPO_LOCALIDADE, por exemplo, significa que uma família que more na ára rural tem $45,6\%$ mais chance de ser beneficiária do que uma família que more na área urbana. Dentre as raças, as pessoas que possuem maiores chances de serem contempladas com o programa bolsa família são os índios$(51,\%6)$ e pardos

$(39,0\%)$, seguido pelos negros$(37,2\%)$, amarelos$(34,1\%)$ e brancos$(27,5\%)$. Essa mesma relação pode ser feita em relação aos domicílios, sendo que o domicílio financiado tem 35% mais chance de participar do programa em relação aos outros domicílios.

O ajuste do modelo pode ser avaliado pela média dos percentuais de acerto dados na tabela 5.9. Utilizando os dados da tabela o percentual médio de acerto do modelo é 76,7%, resultado da média de 63,07 e 90,33. Isso significa que as variaveis estão explicando 76,7% da variável resposta, ou seja, as características da família explicam aproximadamente 77% da probabilidade dela pertencer ao programa.

Tabela 5.9: Frequencias Recebem e Não Recebem PBF

Frequencia / Percentuais	**Não Recebe**	**Recebe**	**Total**
Não Recebe	5089541	2979622	8069163
	24,55	14,38	38,93
	63,07	36,93	
	80,62	20,67	
Recebe	1223784	1,143E7	1,266E7
	5,90	55,17	61,07
	9,67	90,33	
	19,38	79,33	
Total	6313325	1,441E7	2,073E7
	30,46	69,54	100,00

Tabela 5.10: Associação das respostas observadas e probabilidades ajustadas

Percentual Concordante	84.4	Somers' D	0.690
Percentual Discordante	15.4	Gamma	0.691
Percentual Empatado	0.1	Tau-a	0.328
Pares	1.0213754E14	c	0.845

A Tabela 5.10 apresenta os testes (variáveis ordinárias) considerando as respostas observadas com as respostas estimadas. O percentual concordante indica o número de pares de observações que foram concordantes; o percentual discordante indica o número de pares de observações que foram discordantes; o percentual empatado indica o número de pares de observações que não foram nem discordantes e nem

concordantes; e pares indica o Total de pares únicos analisados. O Somer's D é o teste que usado para determinar a força e a direção da relação entre os pares, varia entre -1 e 1. O valor Gamma avalia o mesma que o Somer's D, mas não considera os empates tendo um valor maior que o D, varia entre -1 e 1. A Tau-a funciona do mesmo jeito que o Somer's D, mas considera outros fatores e por isso tem um valor menor, varia entre -1 e 1. Por fim, c é outro coeficiente de correlação entre variáveis ordinárias. Verifica-se uma correlação positiva entre as variáveis.

Como exemplo, podemos utilizar o modelo para analisar a situação de quatro famílias que declararem as seguintes características:

- Uma família de 4 pessoas negras sendo uma criança, tem o modelo, $g(\mathbf{x}) = 0.4487 + (0.3161 \times 4) + (0.6411 \times 1) = 2.3542$, tendo a probabilidade de 91,33% de ser contemplado com o programa;
- Uma família de 4 pessoas brancas sendo uma criança, sua probabilidade de ser contemplado com o programa é de 0,7913, com o modelo $g(\mathbf{x}) = 0.4487 + (0.2432 \times 4) + (0.6411 \times 1) = 2,0626$;
- Ter duas crianças, ter algum aposentado na famíla e ter renda familiar per-capita de R$140,00, com modelo $g(\mathbf{x}) = 0.4487 - (0.4455 \times 1) + (0.6411 \times 2) - (0.0137 \times 140) = -0,6326$ sua probabilidade de ser contemplado com o programa é de 34,69%;
- Ter duas criança, ter algum aposentado na família e renda familiar percapita de R$59,00, com modelo $g(\mathbf{x}) = 0.4487 - (0.4455 \times 1) + (0.6411 \times 2) - (0.0137 \times 59) = 0,4771$, sua probabilidade de ser contemplado com o programa é de 61,7%.

Capítulo 6

CONCLUSÕES

A partir da análise do banco de dados do CadUnico, fornecido pelo MDS, o trabalho identificou os fatores que possibilitam a inclusão no Programa Bolsa Família por meio da análise exploratória desse banco e de um modelo de regressão logística múltipla estabelecido. É importante considerar que os resultados podem ser alterados se variáveis forem retiradas ou se mais variáveis forem adicionadas. Para a manipulação do Cadastro e para gerar o modelo foram necessários alguns dias de processamento, pois a base tinha mais de 20 milhões de famílias. Esse fato limitou o estudo porque não foi possível gerar todos os resultados desejados, como por exemplo, mais mapas e tabelas para a análise exploratória e um modelo politômico para estudar a distribuição dos beneficiários em cada tipo de benefício pago pelo programa.

Verificou-se que a frequência total de não beneficiários é de 38.93% e a de beneficiários é de 61.07%. Além disso, as regiões Norte e Nordeste têm uma maior quantidade de pessoas atendidas pelo Programa. Em termos percentuais, todos os estados apresentaram um aumento maior que 100% na renda das famílias beneficiárias. Por fim, o modelo de regressão logística múltipla possibilita a determinação da probabil-

idade de inclusão das pessoas cadastradas no Programa Bolsa Família conforme as suas características familiares, ou seja, havendo uma eventual redução de recursos, pode-se determinar as pessoas que tem maior propensão para receber o benefício.

O estudo estatístico realizado neste trabalho foi de suma importância para a percepção de que os critérios preestabelecidos nem sempre são suficientes e necessários para a determinação de entrada no Programa Bolsa Família, dado que esse tem recursos limitados. Os critérios preestabelecidos não são a forma mais completa para determinar se a família realmente deve receber o benefício em detrimento de outra família em situações similares.

O novo plano do governo, o Plano Brasil Sem Miséria, tem como meta tirar 16,2 milhões de pessoas da extrema pobreza. Serão beneficiados cidadãos que ganham até R$ 70 por mês. Entre os objetivos está acrescentar ao Bolsa Família mais 1,3 milhão de famílias, portanto o modelo apresentado nesse trabalho pode ser um instrumento de seleção das novas famílias que irão integrar o programa.

Referências Bibliográficas

AGRESTI, A. *Categorical Data Analysis.* 2. ed. Wiley, 2002.

CAIXA, 2007, *BOLSA FAMÍLIA.* URL <http://www1.caixa.gov.br/gov/gov_social/estadual/distribuicao_servicos_cidadao/bolsa_familia/saiba_mais.asp>. Acesso em 31 out.2010.

CAIXA, 2010, *BOLSA FAMÍLIA-CAIXA*.URL <http://www.caixa.gov.br/voce/social/transferencia/bolsa_familia/index.asp> Acesso em 9 out. 2010.

Datasus, 2009, Manual pbf saúde 2009. URL <http://bolsafamilia.datasus.gov.br/documentos_bfa/MANUAL_PBF_BOLSAFAMILIA_SAUDE2009.PDF>. Acesso em 9 out. 2010.

Duarte, G. B.; Sampaio, B.; Sampaio, Y, 2007, Impactos do programa bolsa família sobre os gastos com alimentos de famílias rurais. Banco do Nordeste (BNB).URL <www.bnb.gov.br/content/aplicacao/Eventos/ForumBNB2007/docs/impactos-do-program.pdf>. Acesso em 17 out.2010.

HOSMER, D. W. ; LEMESHOW, S. *Applied Logistic Regression.* 2 ed. Wiley, 2000.

Langellier, J. P, 2008, No brasil, governo paga bolsa para quem vai à escola.URL < http://www.vermelho.org.br/ma/noticia.php?id_noticia=40871&id_secao=2 >.Acesso em 17 out. 2010.

Lindert, K. Banco Mundial, 2005, Brazil:Bolsa família program scaling-up cash transferes for the poor.URL <http:// www.mfdr. Org/sourcebook/6-1Brazil-BolsaFamilia.pdf>. Acesso em 17 out. 2010.

MDS, 2010, Cadastro Único.URL <www.mds.gov.br/bolsafamilia/cadastrounico>. Acesso em 9 out. 2010.

MORRETIN, P. A.; BUSSAB, W. O. *Estatística Básica.* 3 ed. Editora Saraiva, 2002.

SENARC. *Perfil das famílias beneficiadas pelo programa bolsa família 2009.* Secretaria Nacional de Renda e Cidadania, Ministério do Desenvolvimento Social e Combate a Fome (MDS), 2009.

SAS Institute Inc., Version 9.2, Cary, NC: SAS Institute Inc., 2008.