UNIVERSIDADE FEDERAL DE SERGIPE

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA E CIÊNCIAS ATUARIAIS

Evany Dandara Souza dos Santos

IDENTIFICAÇÃO DE SUBDECLARAÇÕES DE RENDA NO CADASTRO ÚNICO PARA PROGRAMAS SOCIAIS UTILIZANDO REGRESSÃO LOGÍSTICA

São Cristóvão – SE

2025

Evany Dandara Souza dos Santos

Identificação de subdeclarações de renda no cadastro único para programas sociais utilizando regressão logística

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística e Ciências Atuariais da Universidade Federal de Sergipe, como parte dos requisitos para obtenção do grau de Bacharel em Estatística.

Orientador (a): Luiz Henrique Gama Dore de Araujo

São Cristóvão – SE

2025

Evany Dandara Souza dos Santos

Identificação de subdeclarações de renda no cadastro único para programas sociais utilizando regressão logística

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística e Ciências Atuariais da Universidade Federal de Sergipe, como um dos pré-requisitos para obtenção do grau de Bacharel em Estatística.

Aprovado em ____/____/____, Nota Final_____.

Banca Examinadora:

_____

Prof. Dr. Luiz Henrique Gama Dore de Araujo

Orientador

_____

Prof. Dr. Allan Robert da Silva

1° Examinador

_____

Prof. Dr. Cleber Martins Xavier

2° Examinador

"A análise estatística é como uma janela através da qual podemos ver o mundo, mas devemos ter cuidado com os vidros que escolhemos."

— *John Tukey*

À minha mãe Ana Maria Felix de Souza, principal figura que me motivou a ser tudo o que eu sempre quis. Sou e sempre serei sua maior fã! Seu amor sempre me orientou.

AGRADECIMENTOS

Agradeço ao meu pai Evaldo Evangelista dos Santos por todas as orações, toda energia emanada e toda motivação depositada desde a infância, juntamente com minha mãe a quem devo tudo o que sou hoje, sempre me direcionando, me motivando e sendo minha maior inspiração.

Agradeço à minha namorada por todo apoio, cumplicidade, paciência e compreensão, além do incentivo que me foi dado.

Agradeço ao meu orientador Professor Doutor Luiz Henrique Gama Dore de Araujo, pela orientação, ajuda e repartição de conhecimentos, sendo peça fundamental também para concretização deste trabalho.

Aos meus amigos, e a todos aqueles que fizeram parte dessa jornada, meu muito obrigada pelo apoio e torcida, vocês foram peças fundamentais para a realização desse sonho.

RESUMO

Os programas de transferência de renda são muito importantes para garantir o direito básico do ser humano: moradia e alimentação. São iniciativas do Governo por meio do Ministério do Desenvolvimento e Assistência Social, Família e Combate à Fome, sendo o Programa Bolsa Família o mais importante e que abrange o maior número de famílias. Um dos requisitos que uma família deve cumprir para se tornar beneficiária do programa é não possuir uma renda familiar per capta superior ao patamar estabelecido pelo Governo Federal. Para verificar a elegibilidade das famílias quanto a esse critério de renda, o Governo Federal utiliza a renda familiar per capta, calculada a partir das rendas individuais de cada membro da família, informadas no Cadastro Único. Um problema com essa abordagem é que, com o intuito de ser beneficiada pelo programa, uma pessoa pode declarar ter renda individual inferior àquela que de fato possui ou até mesmo omitir a sua renda individual, reduzindo sua renda familiar per capta. Um dos mecanismos adotados pelo Governo Federal para tentar identificar esse tipo de fraude é o cruzamento dos dados do CadÚnico com outras bases de dados, como o RAIS e o CAGED. Além desse mecanismo, o uso de modelos matemáticos/estatísticos para identificar famílias com renda subdeclarada no CadÚnico vêm sendo proposto. O presente trabalho propõe um procedimento, baseado em regressão logística, para identificar famílias suspeitas de terem declarado, de maneira fraudulenta, renda inferior ao patamar estabelecido pelo Governo Federal. As famílias são consideradas suspeitas quando declaram ter renda inferior ao patamar estabelecido pelo Governo Federal, porém, segundo o modelo de regressão logística ajustado, possuem uma probabilidade “pequena” de preencherem este requisito. O procedimento foi aplicado aos dados do CadÚnico, referentes ao ano de 2018, e identificou 33 famílias suspeitas. Destas, 14 eram beneficiárias do programa.

Palavras-chave: CadÚnico; Bolsa Família; Renda; Subdeclaração; Fraude; Regressão logística

ABSTRACT

Income transfer programs are very important to guarantee the basic human rights of housing and food. They are government initiatives through the Ministry of Development and Social Assistance, Family and Fight against Hunger, with the Bolsa Família Program being the most important and covering the largest number of families. One of the requirements that a family must meet to become a beneficiary of the program is not to have a per capita family income higher than the level established by the Federal Government. To verify the eligibility of families regarding this income criterion, the Federal Government uses the per capita family income, calculated based on the individual income of each family member, reported in the Single Registry. One problem with this approach is that, in order to benefit from the program, a person may declare having a lower individual income than they actually have or even omit their individual income, reducing their per capita family income. One of the mechanisms adopted by the Federal Government to try to identify this type of fraud is to cross-reference CadÚnico data with other databases, such as RAIS and CAGED. In addition to this mechanism, the use of mathematical/statistical models to identify families with under-declared income in CadÚnico has been proposed. This paper proposes a procedure, based on logistic regression, to identify families suspected of having fraudulently declared income below the level established by the Federal Government. Families are considered suspicious when they declare having an income below the level established by the Federal Government, but according to the adjusted logistic regression model, they have a "small" probability of meeting this requirement. The procedure was applied to CadÚnico data for the year 2018 and identified 33 suspicious families. Of these, 14 were beneficiaries of the program.

Keywords: CadÚnico; Bolsa Família; Income; Under-declaration; Fraud; Logistic regression

LISTA DE ILUSTRAÇÃO

SUMÁRIO

1 INTRODUÇÃO

Em tempos de crise econômica, combate à fome, empenho no desenvolvimento social e o desequilíbrio financeiro, é preciso uma interferência do Governo, ao qual detém recursos financeiros do país e os transfere diretamente da União (Estado ou município) aos cidadãos que participam de programas sociais para combater tais condições, os chamados: Programas de Transferência de Renda. “[...] O Brasil assume uma posição de destaque com o Programa Bolsa Família” (SILVA, Tiago Falcão (org.) ENAP, 2018).

De acordo com o Ministério do Desenvolvimento e Assistência Social, Família e Combate à Fome, PBF surgiu com a proposta de unificar iniciativas já existentes, como o Auxílio Gás e Fome Zero. Esse programa vem apresentando bons resultados como, por exemplo, em 2014, onde mais de 14 milhões de famílias brasileiras já eram assistidas, e no mesmo ano a Organização das Nações Unidas para Alimentação e Agricultura (FAO) anuncia a remoção do Brasil do Mapa da Fome como afirma o Memorial da Democracia "Combate à fome" em seu site.

Para participarem do PBF, as famílias têm que cumprir uma série de requisitos, dentre eles, não possuir renda familiar per capta superior a um patamar estabelecido pelo Governo Federal que, atualmente, é de até R$ 218,00, além de precisarem estar inscritas no Cadastro único – CadÚnico. (Ministério do Desenvolvimento e Assistência Social, Família e Combate à Fome)

O CadÚnico é um registro que reúne informações associadas aos perfis sociodemográficos e de moradia de pessoas e famílias brasileiras (Ministério do Desenvolvimento e Assistência Social, Família e Combate à Fome). Um dos seus principais objetivos é identificar as famílias brasileiras de baixa renda e saber suas condições de vida. Para verificar a elegibilidade das famílias quanto ao critério da renda familiar per capta, o Governo Federal utiliza a renda familiar per capta, calculada a partir das rendas individuais de cada membro da família, informadas no CadÚnico. Um problema com essa abordagem é que, com o intuito de ser beneficiada pelo PBF, uma

pessoa pode declarar ter renda individual inferior àquela que de fato possui ou até mesmo omitir a sua renda individual, reduzindo sua renda familiar per capta. Um dos mecanismos adotados pelo Governo Federal para tentar identificar esse tipo de fraude é o cruzamento dos dados do CadÚnico com outras bases de dados, como o RAIS e o CAGED (De acordo pelo Boletim divulgado em 2017 pela Secretária Nacional de Renda de Cidadania).

Para além do cruzamento dos dados do CadÚnico com outras bases de dados, alguns pesquisadores vêm propondo o uso de modelos matemáticos/estatísticos para identificar famílias com renda subdeclarada no CadÚnico (MENDES & SAMPAIO, 2008; MOSTAFA & SANTOS, 2016).

O presente trabalho propõe um procedimento, baseado em regressão logística, para identificar famílias suspeitas de terem declarado, de maneira fraudulenta, renda inferior ao patamar estabelecido pelo Governo Federal. Considera-se como variável resposta a variável indicadora que indica se a família atende (ou não) ao critério da renda. Como variáveis preditoras, são consideradas várias características socioeconômicas e de moradia das famílias. As famílias são consideradas suspeitas quando declaram terem renda inferior ao patamar estabelecido pelo Governo Federal, porém, segundo o modelo de regressão logística ajustado, possuem uma probabilidade “pequena” de preencherem esse requisito. Foram utilizados os dados do CadÚnico referentes ao ano de 2018.

Esse trabalho encontra-se organizado da seguinte maneira: capítulo 1, a introdução deste trabalho e tema; Capítulo 2, temos os objetivos com a realização deste estudo; Capítulo 3, a justificativa para a realização deste trabalho; Capitulo 4, a revisão literária; Capítulo 5, a Metodologia; E capítulo 6 e 7, os resultados e conclusões, respectivamente.

2 OBJETIVOS

2.1 Geral

Aplicar a técnica de regressão logística para identificar famílias com indícios de subdeclaração de renda, com base nos critérios de elegibilidade do Programa Bolsa Família.

2.2 Específicos

- Buscar relações relevantes entre as variáveis socioeconômicas.
- Elaborar um modelo baseado em regressão logística que permita identificar inconsistências nas autodeclarações de renda.
- Propor esse método como forma de ferramenta para a identificação de famílias potencialmente irregulares no cadastramento do programa.

3 JUSTIFICATIVA

A presente pesquisa justifica-se pela relevância do seu tema acerca de um importante programa que é o Bolsa Família para a distribuição de renda no Brasil, garantindo que esses recursos sejam transferidos de forma mais eficiente. Assim, minimizando um dos seus grandes desafios que é o controle e a verificação das rendas declaradas, visto que para se inscrever no programa não há uma comprovação de renda para os aderentes ao CadÚnico. Este estudo busca contribuir para que se tenha um controle mais criterioso, possibilitando formas e mecanismos mais assertivos para um melhor monitoramento, sendo o processo e o resultado muito mais transparente e justo para a população.

4. REVISÃO LITERÁRIA

De acordo com o relatório do Ministério do Desenvolvimento Social (2017), foram identificadas inconsistências em aproximadamente 1,1 milhão de benefícios, resultando em bloqueios e cancelamentos em novembro do mesmo ano. E para compreender maneiras de amenizar essas inconsistências e minimizar tais efeitos é imprescindível a análise do contexto histórico ao longo dos tempos (Cadernos SUAS, Volume III, 2017).

4.1 Origem

O Brasil tem uma luta constante contra a fome, seca, e pobreza, retratada desde a realidade de quem vive e viveu, presente na literatura e cinema brasileiros. Podemos ver em livros como “Vidas secas” de Graciliano Ramos, “Os Sertões” de Euclides da Cunha e “Menino de Engenho” de José Lins do Rego, e em filmes como “Fome” de Cristiano Burlan, “Histórias da Fome no Brasil” de Camilo Tavares e “Garapa” de José Padilha. As obras retratam parte do que se foi vivido por quem mais queria viver sua vida dignamente, ou até que eram gratos pelo pouco que se tinha.

Há então a grande luta para romper essa realidade estendida desde os primórdios, e que é luta constante para o povo brasileiro, principalmente para o norte e nordeste, em especial o povo do Sertão brasileiro.

O pilar que motivou a criação do Programa Bolsa Família criado pelo Governo de Fernando Henrique Cardoso e impulsionado e oficializado pelo Presidente Luiz Inácio Lula da Silva, foi o Combate à Fome e a Pobreza.

Em 2003, cerca de 44 milhões de brasileiros sofriam gravemente da extrema pobreza, e partindo disso foram instauradas políticas públicas eficientes direcionadas a beneficiar quem mais necessitava: os mais pobres. Tornando-se referência mundial no combate à fome. Com o principal objetivo de diminuir índices alarmantes de fome no país, com o auxílio dos programas de transferência de renda, sendo o Bolsa Família o maior agora dentre todos, o Brasil em 2004, sai do mapa da fome.

4.2 Programa Vs Fraude: O controle de qualidade

A efetividade do programa é evidenciada pela saída do Brasil do Mapa da Fome. No entanto, para garantir que os benefícios alcancem exclusivamente os indivíduos que realmente atendem aos critérios do programa, é essencial a implementação de um controle de qualidade rigoroso e eficaz.

Uma das principais problemáticas enfrentadas é a ausência de comprovação de renda no CadÚnico, o que abre brechas para autodeclarações imprecisas. Nesse contexto, Mesquita (2007, p. 16) destaca:

> […] um outro fator que compromete a veracidade das informações é o fato dessa coleta de dados estar vinculada à concessão de um benefício que utiliza um determinado corte de renda como critério de elegibilidade. Isso faz com que, por estratégia de sobrevivência, muitas famílias omitam ou subdeclarem alguma fonte de rendimento. Importante ainda ressaltar que as informações são autodeclaradas, não se fazendo necessário nenhum tipo de comprovação. […] (Mesquita (2007, p. 16))

Neste âmbito, Mesquita (2007) considera também a qualidade dessas informações inseridas no CadÚnico visto que são preenchidas por agentes municipais com pouco ou nenhum treinamento. Similarmente apontado por Magalhães (2007 apud Melgarejo, 2011) na implementação do PBF em Duque de Caxias (RJ) ao qual cita que alguns profissionais ligados diretamente com o programa não conhecem de fato suas condições ou não concorda com elas.

Em Relatórios de Gestão da CGU do ano de 2006 foram encontradas falhas e irregularidades denunciadas ou detectadas, que foram agrupadas por Filgueiras (2008), mensurando como: cadastramento e visitas às famílias, agente operador CAIXA, atuação dos gestores municipais, acompanhamento de condicionalidades e controle social. Sendo visto na aba de cadastramento e visitas às famílias pessoas e famílias incluídas indevidamente ou que recebem o benefício com duplicidade. E como sinaliza Lindert et al (2006 apud Filgueiras, 2008), a eficiência de tais programas, dependente da

confiabilidade do processo do cadastramento e da manutenção do banco de dados, ressaltando que com inconsistências o cadastro perde sua utilidade.

Entre 2007 e 2014, com a expansão do programa, houve uma redução nos erros de exclusão, acompanhada por um aumento nos erros de inclusão. A partir de 2014, essa tendência se inverteu, com a diminuição dos erros de inclusão e o aumento dos erros de exclusão. (Souza e Bruce, 2022)

Segundo Cutrim (2019) os municípios têm papel fundamental no sucesso do programa e mesmo com o aprimoramento de capacidades técnicas da gestão municipal, ao qual tem conseguido identificar irregularidades, ainda se espera que o cumprimento dos requisitos seja realizado à risca e seja nulo o grupo de pessoas fora dos critérios. E Cutrim (2019, p. 03) complementa:

> […] Nessa abordagem, os casos de descumprimento de regras do Programa se tratam, na maioria das vezes, de situações de "fraude" ou "corrupção", que devem ser minimizados com o aprimoramento das ações de controle antes e depois da entrada das famílias no cadastro, o que seria garantido pelo aperfeiçoamento da gestão municipal do Programa e do Cadastro Único e de seus mecanismos de liderança, estratégia, monitoramento e controle postos em prática visando a condução das políticas públicas. […] (Cutrim (2019, p. 03))

Como forma de controle e monitoramento imediato para identificação de indivíduos com subdeclaração ou omissão de renda, como menciona Mostafa & Santos (2016) é realizado cruzamentos de dados de maneira sistêmica com registros administrativos com informação de rendimento informal, de óbito, beneficiários do INSS (Instituto Nacional de Seguro Social).

Na ação realizada pelo Ministério do Desenvolvimento Social (2017) cruzou-se 06 bases de dados governamentais:

- Relação Anual de Informações Sociais (RAIS);
- Cadastro Geral de Empregados e Desempregados (CAGED), do Ministério do Trabalho;

- Sistema de benefícios permanentes e auxílios pagos pelo INSS;
- Sistema de Controle de Óbitos (SISOBI);
- Sistema Integrado de Administração de Recursos Humanos (SIAPE), de servidores públicos do governo federal;
- Cadastro Nacional de Pessoas Jurídicas (CNPJ).

Do cruzamento dessas bases com a Folha de pagamento do PBF foram identificadas inconsistências cadastrais, em torno de 1,1 milhão dentre os 13,9 milhões de beneficiários no ano de 2016.

A verificação da renda após o cadastro no programa é realizada em todo o país com base na Relação Anual de Informações Sociais (RAIS), do Ministério do Trabalho e Emprego (MTE), ou nos registros de benefícios do INSS. No entanto, devido ao atraso na captura e sistematização dos dados, a erros nos registros de trabalho e à alta rotatividade do mercado formal, essa verificação identifica apenas indícios de inconsistências. A partir dessas informações, os dados são enviados ao gestor de cada unidade municipal para atualização cadastral. Caso a atualização não seja realizada ou os critérios do programa não sejam cumpridos, o benefício é cancelado (Mostafa & Santos, 2016).

Mendes e Sampaio (2008) refletem sobre os custos gerados para os cofres públicos devido a fraudes decorrentes da seleção adversa. Segundo os autores, essa seleção ocorre devido à assimetria de informação entre o governo e o potencial beneficiário, impossibilitando a correta avaliação da real necessidade do candidato e a verificação do cumprimento das exigências para a inclusão no programa. Diante disso, destaca-se a importância de estudos voltados para o tema.

4.3 A Estatística e a Matemática como ferramentas para detecção de fraudes

A estatística e a matemática se constituem como poderosa ferramenta para análise e interpretação de dados, desde a organização e tabulação, até a modelagem e inferência. Nesta seção, serão apresentados alguns métodos analíticos utilizados por diferentes autores como base metodológica em seus estudos.

Cutrim (2019) utilizou um modelo de regressão logística múltipla onde investigou a relação entre as variáveis socioeconômicas e o nível agregado de cancelamentos (Apenas por descumprimentos de regras do programa). Com esta técnica é possível identificar como cada variável independente se comporta e contribui na mudança da variável dependente (a taxa de cancelamento). E com isso traçar um modelo que contribua para a compreensão das características sociais, e sua relevância no contexto do tema.

Mostafa e Santos (2016) destacam o pioneirismo do Programa Bolsa Família (PBF) na aplicação de testes de predição estatística em escala nacional, com ênfase no uso do *Proxy Means Test* (PMT). As autoras abordam os desafios relacionados à pobreza e discutem as limitações desse modelo estatístico na revisão dos benefícios. O estudo teve como objetivo avaliar a aplicação de um preditor de renda como instrumento complementar à renda declarada pelas famílias no CadÚnico, buscando identificar a viabilidade desse método para detectar omissão ou subdeclaração de renda.

No trabalho de Mendes & Sampaio (2008) buscou-se evidenciar as ineficiências existentes no PBF, devido assimetria em um jogo de informação, apresentando algumas alternativas a serem tomadas como forma de amenizá-las. O equilíbrio do jogo se dá pela crença do governo na honestidade dos candidatos ao programa, influenciado pela regra de Bayes. Mendes & Sampaio (2008) concluem em seu estudo através da modelagem, que sem punições efetivas, há um forte incentivo para tentativas de fraudes, sendo fundamental uma fiscalização apurada.

5 METODOLOGIA

Este capítulo apresenta a abordagem metodológica e estatística que foi utilizada durante o presente estudo, explicando cada etapa da execução, bem como o desenvolvimento do modelo de regressão logística realizado. Toda análise foi realizada em linguagem R, utilizando o software estatístico Rstudio e seus principais pacotes como “hnp”, “tidyverse” e “statmod”, ao qual foi realizado todos os cálculos, a estimação do modelo pela máxima verossimilhança e o ajuste dos parâmetros da regressão logística.

5.1 Análise Exploratória de Dados

O estudo adotou uma abordagem quantitativa, com análise estatística dos dados extraídos do CadÚnico (Cadastro Único para Programas Sociais – 2018). Para delimitar a análise, a amostra foi limitada à cidade de Aracaju/SE, somando 10.555 observações. Essa escolha teve por iniciativa a proximidade geográfica com o local de realização da pesquisa.

O tratamento dos dados incluiu:

- Exploração inicial: verificação da estrutura da base.
- Criação de um banco secundário: evitou-se alterações diretas na base original.
- Remoção de variáveis irrelevantes: as variáveis identificação das famílias (ID) e peso da família foram excluídas.

Seleção de variáveis preditoras: foram escolhidas as mesmas variáveis do estudo base (MOSTAFA & SANTOS, 2016), incluindo:

- Características do domicilio (número de cômodos, abastecimento de água, escoamento sanitário, iluminação, destino do lixo).
- Demografia familiar (número de homens e mulheres por faixa etária, presença de cônjuge, migração).
- Nível educacional (grau de instrução, alfabetização, tipo de escola frequentada por menores de 17 anos).
- Ocupação (número de pessoas empregadas por gênero e idade).

Os dados foram analisados inicialmente por meio de resumos estatísticos usualmente usados em pesquisas para entendimento do comportamento dos dados, buscando uma visão geral, e identificação de padrões. Ao todo temos 5877 beneficiários do programa e 4678 que não fazem parte do PBF.

5.2 Modelagem Estatística: Regressão Logística

Regressão logística é um caso particular de MLG – Modelos Lineares Generalizados, onde a ideia central é ter opções de distribuições de probabilidades dos erros associados ao modelo de regressão que surgem quando temos variáveis respostas binárias ou de contagem, modeladas pelas distribuições. (PAULA, 2013 apud SILVA, 2018)

Seja, X = (1, X1, X2, ..., Xn) um vetor onde o primeiro elemento é igual a 1 (constante) e os demais representam as n variáveis preditoras do modelo. O modelo de Regressão Logística é um caso particular dos Modelos Lineares Generalizados (Neter et al., 1996, apud Ferreira 2021), tal que:

$$\ln\left(\frac{p(X)}{1-p(X)}\right) = \beta' X$$

onde β' = $(\beta_1, \beta_2, \beta_3 \ldots \beta_n)$ é o vetor parâmetro associado às variáveis e p(X) = E(Y=1|X) sendo a probabilidade de a família ser classificada como dentro ou fora dos critérios, e portanto, possível fraudulenta ou não (Neter et al., 1996, apud Ferreira 2021).

O método de Regressão Logística foi escolhido por ser um modelo adequado para prever as probabilidades de um evento binário acontecer, neste caso, as subdeclarações de renda. Dito isto, o modelo que segue a distribuição binomial, sendo o resultado de interesse binário, foi aplicado (McNulty, 2021).

5.2.1 Parâmetros de análise da Regressão Logística

Um ponto muito importante é a relação dos resíduos, como verificar o afastamento dos pontos, com os pontos do modelo estimado, sempre de acordo com a parte aleatória e sistêmica. Um dos resíduos mais utilizados é o desvio residual. E uma das formas gráficas de verificação mais comum é a de envelope, utilizada nesse trabalho.

A estimação do coeficiente é efetuada pela máxima verossimilhança, esta busca localizar as estimativas mais prováveis dos coeficientes e maximizar a probabilidade de que tal evento ocorre, sempre observando o R^2, a precisão preditiva e o p-valor.

Como os valores são apresentados de forma logarítmica, é realizada a transformação exponencial da variável da regressão, sendo assim a chamada Odd, ou "chances" para as variáveis independentes. A determinação do intervalo de confiança também é de extrema importância para análise da regressão logística, estima-se um intervalo sempre em média de 90-95% de confiança. E como lida-se com probabilidades sempre temos a predição dela como método de análise, o que permite a comparação entre as variáveis e seus resultados.

De maneira prática temos a qualificação o ajuste do modelo na tabela, que chamamos de matriz de confusão, ao qual cruzasse o resultado da classificação cruzada da variável resposta de acordo com a variável dicotômica a definir pelos valores das probabilidades estimadas, sempre determinando um ponto de corte, ao qual foi utilizado neste estudo o de 0.5 o usual em análises.

5.3 Análise diagnóstico

Nesta seção serão apresentadas técnicas de validação e avaliação do ajuste do modelo de regressão logística descrito acima.

5.3.1 Validação do Modelo: Aplicando o teste de Hosmer-Lemeshow

O teste de Hosmer Lemeshow busca identificar falhas na estimação. Com base nas previsões do modelo, é feita a divisão dos dados em grupos onde é calculado através da soma dos quadrados das diferenças entre os valores observados e esperados se caracterizando como um teste de ajuste Qui quadrado (X^2). As hipóteses do teste são:

H0: o modelo se ajusta bem aos dados
H1: o modelo não está bem ajustado aos dados

O teste avalia o modelo ajustado através das distâncias entre as probabilidades ajustadas e as probabilidades observadas. (Hosmer e Lemeshow, 1980 apud Cordeiro, 2016, p.14). Interpreta-se o modelo como adequado de acordo com seu p-valor (uma condição estatística que demonstra a compatibilidade dos dados esperados e observados, tornando a hipótese nula verdadeira, para isso o p-valor deve ser > 0.05); com aplicação através da função “logit” foi verificada a adequabilidade do modelo de regressão.

5.3.2 Gráfico de normalidade de probabilidades com envelope

Este tipo de gráfico é utilizado para avaliar visualmente o ajuste do modelo, principalmente partindo da relação de seus resíduos (diferença entre os valores observados e previstos, sendo normalmente utilizado para análise de resíduos, gerando portanto o gráfico, permitindo a verificação em modelos de regressão normal, como menciona Fernandes (2019), baseado em constatações de Flack & Flores (1989).

É possível visualizar se o comportamento dos resíduos segue a área de confiança (como propôs Atkison, 1981, apud Colosimo, 2025, em simulações de Monte Carlo) do envelope, se assim for, podemos afirmar que os resíduos do modelo seguem a distribuição normal. De acordo com Colosimo (2025), o envelope é formado pela mediana e o percentis dos resíduos do modelo.

5.3.3 Gráfico do resíduo componente do desvio padronizado Vs índice

Importante recurso para avaliar a distribuição e comportamento dos resíduos em relação ao índice das observações. Ajudando a identificar outliers e erros de ajustes, além de verificar a homocedasticidade, ou seja, a variabilidade dos erros (resíduos) deve ser constante ao longo de todas as observações. Caso não seja identificado esse padrão, pode indicar heterocedasticidade, e os resíduos não têm essa variação constante (Gori, 2005).

5.3.4 Gráfico da distância de Cook vs o índice

Cook (1986), propôs a avaliação da influência conjunta das observações sob pequenas

mudanças, a chamada: Influência local.

Gráfico muito utilizado também para verificação de outliers, ao qual podemos também identificar as observações mais influentes de forma bem visual, ou seja, a distância de Cook nada mais é que a distância entre uma observação e a média da amostra, o que nos dá a equiparação de sua influência.

5.3.5 Identificação de famílias suspeitas

O modelo de regressão logística foi utilizado para estimar a probabilidade de uma família ter renda superior ao limite permitido. Inicialmente foram classificadas através do modelo quais famílias declara renda inferior, mas que o modelo afirma que tem renda superior; adotou-se um limiar de 50% de modo padrão para classificar, portanto, onde a probabilidade apontada for maior que 50%, o modelo classifica aquela família como superior, e menor que 50% o modelo classifica como inferior à renda permitida.

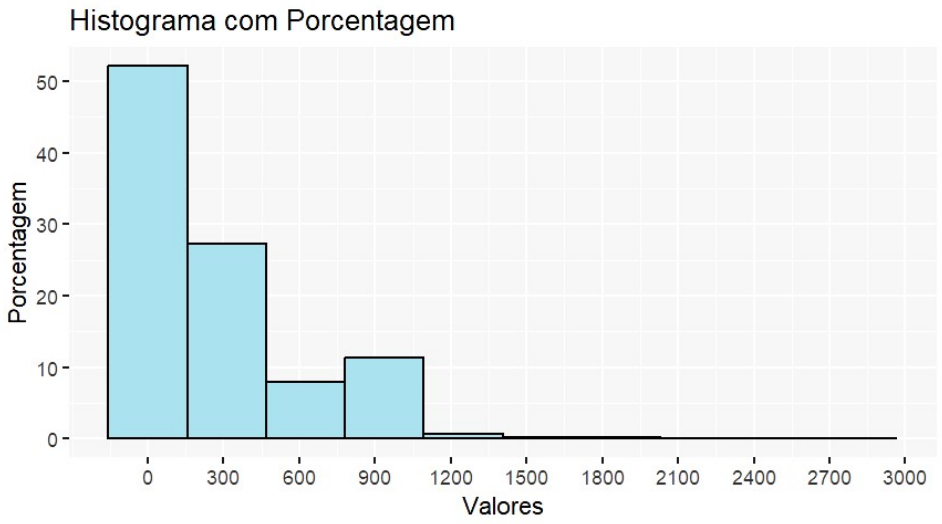
Dentre essas famílias que o modelo afirma que tem a renda superior, foram filtradas aquelas com o resíduo do componente do desvio padronizado maior que o critério estabelecido; foi estabelecido o critério de ±2, equivalendo a 90% das observações. O que nos permite identificar com maior discrepância entra a renda declarada e a estimada pelo modelo. Essas famílias seriam detectadas como potenciais suspeitas para subdeclarações de renda ou inconsistências nas informações.

6 RESULTADOS

6.1 Análise exploratória de dados

Foi iniciada as análises, e o ponto de partida foi o entendimento da renda declarada das famílias inscritas no CadÚnico, observamos que a média da renda foi de R$287,31, com valores extremos de R$2.811,00 como o máximo e R$0,00 como o mínimo declarado.

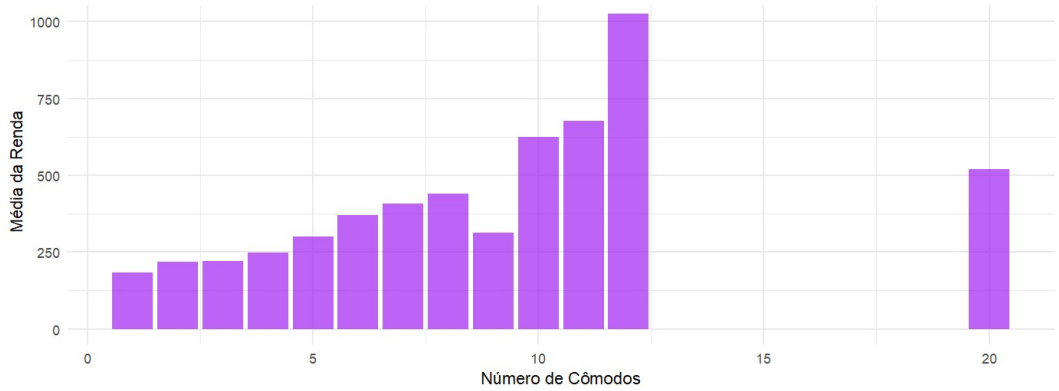
Figura 1: Distribuição de Renda familiar. As porcentagens baseadas nas rendas declaradas.



Fonte: Elaborado pela autora baseado nos dados extraídos do CadÚnico

Podemos observar que a frequência de rendas consideradas baixas é bem maior que a renda considerada alta, sendo mais da metade de observações dentro do intervalo de R$0,00 – R$300,00.

Figura 2: Média da renda X número de cômodos

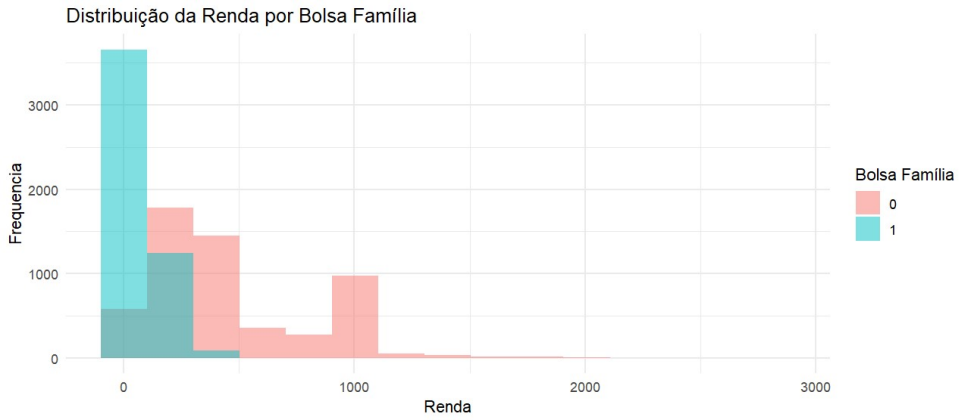


Fonte: Elaborado pela autora baseado nos dados extraídos do CadÚnico

O número de cômodos acompanha o valor da renda, quanto maior o número de cômodos maior a renda. Com uma observação no indivíduo que está em último no gráfico com alto

número de cômodos e um valor razoável de renda declarada.

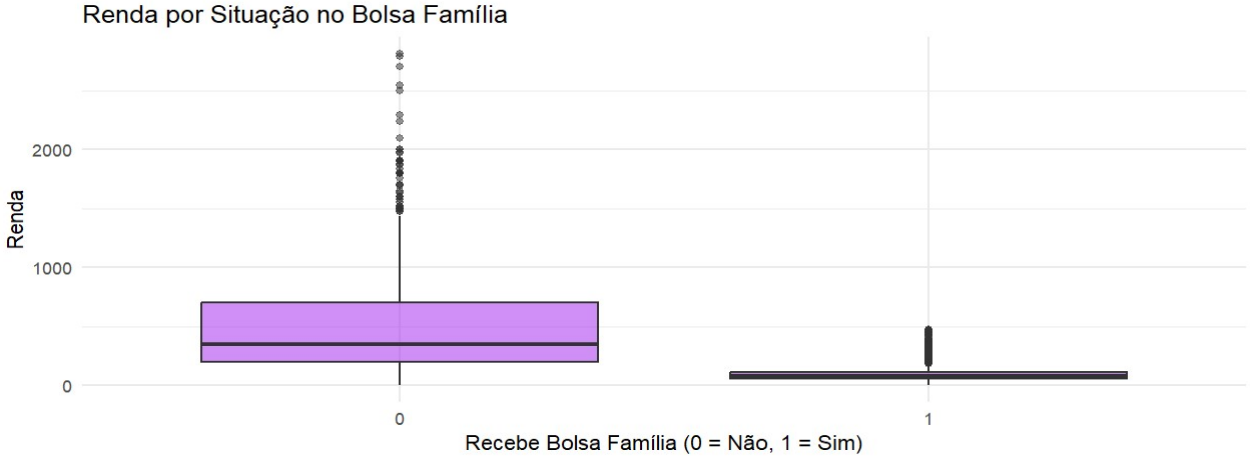
Gráfico 3 Relação entre a renda e os Beneficiários do PBF



Fonte: Elaborado pela autora baseado nos dados extraídos do CadÚnico

No gráfico 3 e 4, temos uma comparação entre a frequência e a renda de cada família beneficiária do PBF. Foi designado de modo binário 1 como beneficiário e 0 como não beneficiário. Podemos observar que os beneficiários estão no eixo de renda baixa, como já era de se esperar, bem como é mostrado no gráfico 4 no boxplot da direito onde representa todos os beneficiários e sua concentração. Na esquerda podemos ver os não beneficiários e como suas rendas variam.

Gráfico 4 Relação entre renda e os Beneficiários do PBF

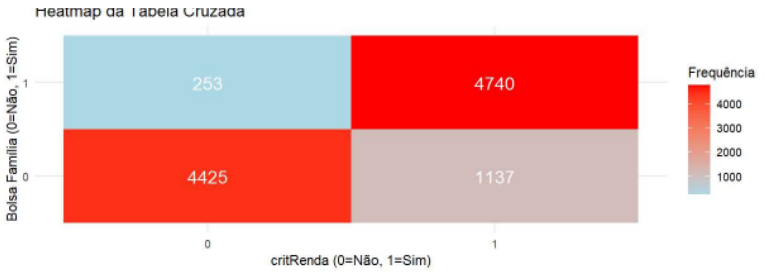


Fonte: Elaborado pela autora baseado nos dados extraídos do CadÚnico

Com isso e sabendo que o critério de renda para participação permitido era de até R$178,00 (SILVA, Tiago Falcão (org.) ENAP, 2018), cruzamos os valores por meio da

tabela de contingência e foi identificado que 253 famílias, o equivalente a 2,4% que são beneficiárias, porém tem renda maior que R$178,00, o que nos leva a indagações sobre essas famílias.

Figura 1 Heatmap da tabela cruzada entre beneficiários e o critério de renda

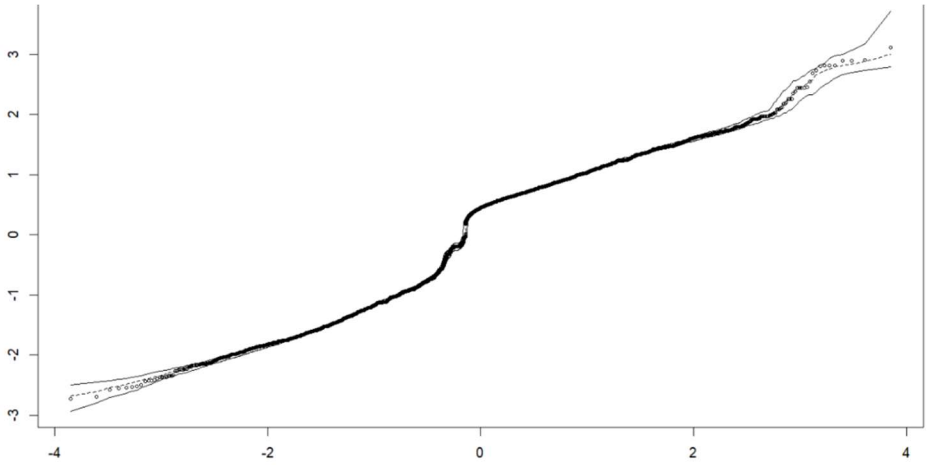


Fonte: Elaborado pela autora baseado nos dados extraídos do CadÚnico

6.2 Detecção das famílias suspeitas

Como o método escolhido foi de Regressão logística por prever a probabilidade de a família ser classificada adepta ou não ao Programa Social, no caso, mais especificamente, adepta ou não às subdeclarações de renda, tivemos um modelo com 0.05 de p-valor e portanto, com bom desempenho validado pelo teste de Hosmer-Lemershow indicando que não há evidências suficientes para rejeitar H0, dito isso, o modelo se ajusta bem aos dados.

Gráfico 5 Probabilidade Normal de Envelope (resíduos)



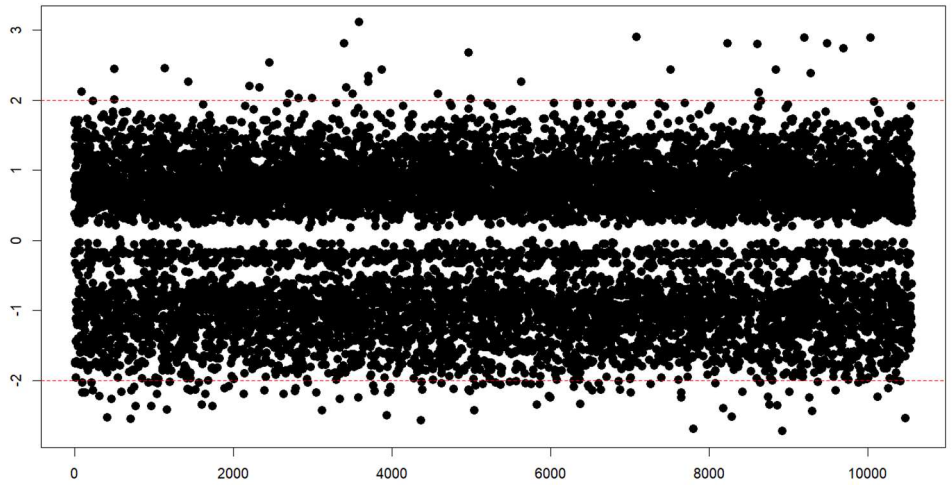
Fonte: Elaborado pela autora baseado nos dados extraídos do CadÚnico

No gráfico 5, é possível visualizar que o comportamento dos resíduos segue a área de confiança (como propôs Atkison, 1981, apud Colosimo, 2025, em simulações de Monte Carlo) do envelope, e também podemos visualizar que todos os valores estão dentro do

intervalo de (-4,4) então, podemos afirmar que os resíduos do modelo seguem a distribuição normal.

Ao analisar a homoscedasticidade, temos o comportamento dos dados com uma variabilidade constante.

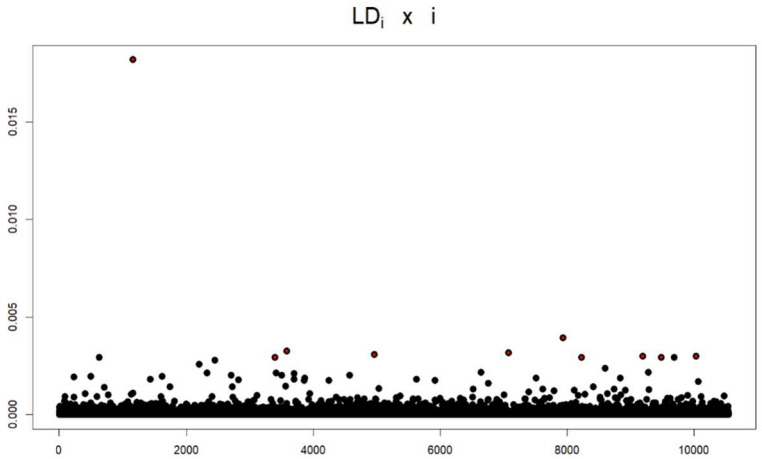
Gráfico 6 Resíduo componente do desvio padronizado vs índice



Fonte: Elaborado pela autora baseado nos dados extraídos do CadÚnico

No gráfico 6 é possível ver um comportamento similar ao longo dos índices (considerou-se como limites de referência os pontos (-2,2)), portanto prevalece o comportamento homocedástico. É possível observar alguns pontos muito afastados da linha de zero (centralizada) e também dos eixos traçados, isso pode indicar presença de outlier. Os resíduos também se distribuem de modo constante em torno de zero, logo, podemos afirmar novamente que é um bom modelo, e um bom ajuste.

Gráfico 7 Distância de Cook vs Índice



Fonte: Elaborado pela autora baseado nos dados extraídos do CadÚnico

Observa-se um outlier que se distingue claramente dos demais dados, apresentando um comportamento atípico. Esse ponto foi identificado na amostra e refere-se a um indivíduo (575 na amostra) que, embora possua saneamento regular, utiliza iluminação à vela; o dado permaneceu no conjunto de dados.
Essa observação, por ser atípica, pode influenciar indevidamente os resultados do modelo.

Em seguida, definiu-se primeiramente o ponto de corte das predições como 0,5, onde se a probabilidade da classe positiva for maior que 0,5, teremos a previsão igual a 1, em contrapartida, será 0. As probabilidades foram convertidas em binárias com base no ponto de corte. Também foi calculado a proporção de erro/observações mal classificadas, além da matriz de confusão, acurácia, precisão, recall, e F1-Score.

Tabela 1: Matriz de confusão

	Predito:0	Predito:1
Real: 0	3020	1658
Real: 1	1158	4719

Fonte: Elaborado pela autora baseado nos dados extraídos do CadÚnico

Na matriz de confusão da tabela 1, podemos verificar:

3020 - Casos corretamente classificados como 0. (Verdadeiros Negativos)
1658 - Casos reais 0, mas classificados como 1. (Falsos Positivos)
1158 - Casos reais 1, mas classificados como 0. (Falsos Negativos)
4719 - Casos corretamente classificados como 1. (Verdadeiros Positivos)

Com uma proporção de erro apresentado de 0,2668, ou 26,68%, acurácia de 0,7332. 73,32%; proporção de previsões corretas sobre o total de dados; Precisão de 0,74. 74%, logo, 26% são falsos positivos. Um recall de 80.3% garantindo que o modelo capture a maior parte das observações. E F1-Score: 0,7702, 77,02% equiparando o recall com a precisão.

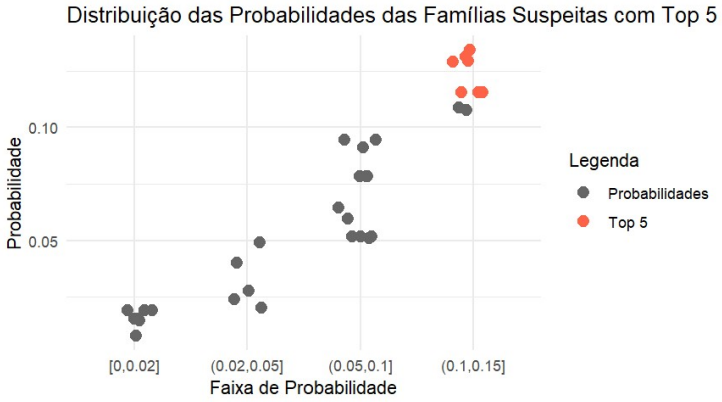
O uso apenas do classificador aponta 1546 famílias suspeitas. Observa-se que, entre as famílias classificadas como suspeitas, aproximadamente 50% delas apresentam uma

probabilidade superior a 40%, enquanto o ponto de corte utilizado é de 50%. Isso indica que, embora essas famílias sejam classificadas como não pertencentes ao grupo de baixa renda, as probabilidades associadas a elas ainda são relativamente altas. Assim, o próximo passo será identificar, entre essas famílias suspeitas, aquelas cujas probabilidades são substancialmente mais baixas do que o valor observado para a resposta (y == 1), que serão consideradas como pontos aberrantes.

Para isso, utilizaremos o resíduo componente do desvio padronizado, que identifica o quanto cada caso se afasta do que era esperado pelo modelo; adota-se como critério a seleção das famílias cujos resíduos sejam menores que 2 ou maiores que 2, características indicativas de observações atípicas que merecem atenção; Vale ressaltar que a escolha do intervalo é conservadora pois destaca principalmente os extremos casos que estão bem fora do esperado.

Foi calculado a quantidade de famílias suspeitas que atendem a esses critérios (falsos negativos com resíduos atípicos), e o modelo encontrou 33 famílias com comportamentos suspeitos que merecem uma investigação acerca do tema.

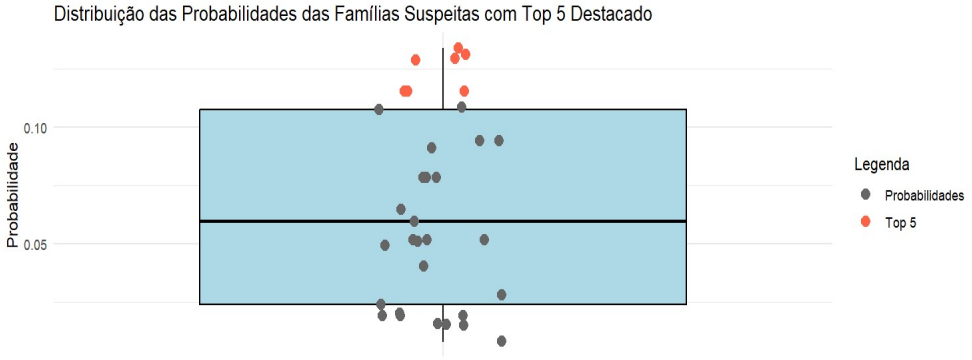
Gráfico 8 Gráfico da distribuição das probabilidades das 33 famílias



Fonte: Elaborado pela autora baseado nos dados extraídos do CadÚnico

Temos no gráfico 8 e 9 o top 5 das probabilidades apresentadas na faixa entre 0.1 – 0.15: 0.13408, 0.13125, 0.12941, 0.12895, 0.11537. Ou seja, a família com maior probabilidade tem aproximadamente 13% de estar dentro dos critérios.

Gráfico 9 Boxplot com a distribuição das probabilidades com as 5 maiores destacadas



Fonte: Elaborado pela autora baseado nos dados extraídos do CadÚnico

Em resumo as probabilidades têm média de 0.0667, mínimo de 0.0078, mediana de 0.0596 e o máximo atingindo 0.1340, evidenciando as probabilidades muito baixas dessas famílias se enquadrar nos requisitos de renda do programa. Dentre estas 14 são beneficiárias do PBF. No top 5 temos famílias com quantidade de cômodos de 2 – 7 cômodos, renda entre R$66,00 – R$954,00 e com quartos entre 1 – 3, sendo a renda o maior influenciador neste trabalho. A identificação desses casos é fundamental, pois permite isolar observações que se desviam do comportamento esperado, permitindo análises mais robustas nas próximas etapas.

7 CONCLUSÃO

Observou-se como os dados foram se comportando ao longo do ajuste do modelo, permeando sobre as relevâncias das variáveis como por exemplo a renda e a quantidade de cômodos. Bem como, é comprovado através do próprio modelo, resíduos e normalidade, a aplicabilidade da regressão logística, cabendo de maneira crucial e peculiar neste tipo de pesquisa.

Foi tido como resultado probabilidades a fim de identificar famílias que não atendem aos critérios do Programa Bolsa Família, e assim foi realizado, identificado famílias suspeitas de não se adequar ao critério de renda estipulado, e que consequentemente, merecem uma investigação aprofundada, ao qual a maior das probabilidades de o indivíduo (família) estiver correto em suas afirmações é de 13%, aproximadamente.

Este presente trabalho focou em identificar as subdeclarações, principalmente partindo das rendas declaradas, comparando com o critério de corte e observando as probabilidades das afirmações serem verídicas, ou falsos positivos, e falsos negativos, do ponto de influência utilizado comumente, e tendo a regressão logística como protagonista, que nos retornou 33 famílias suspeitas, e dentre elas 14 beneficiárias do PBF, norte esse que nos dá embasamento para a investigação. Para futuros trabalhos recomenda-se trabalhar outros limiares e pontos de influências, além da sugestão de trabalho com a curva ROC, e utilizar outras variáveis conforme achar necessário.

Conclui-se que é viável e confiável estimar um modelo de regressão logística para identificar possíveis subdeclarações de renda e fraudes, contribuindo não só para uma fiscalização, mas também para a construção de um programa mais transparente, eficiente e, acima de tudo, justo para todos os cidadãos.

REFERÊNCIAS

AGENCIA BRASIL. Bolsa Família: a trajetória do programa que tirou o Brasil do mapa da fome. 2023. Disponível em: https://agenciagov.ebc.com.br/noticias/202310/bolsa-familia-a-trajetoria-do-programa-que-tirou-o-brasil-do-mapa-da-fome. Acesso em: 15 out. 2025.

AKAIKE, H. A new look at the statistical model identification. IEEE Transactions on Automatic Control, v. 19, n. 6, p. 716–726, 1974. Acesso em: 15 mar. 2025.

BRASIL. Ministério do Desenvolvimento Social. Cadernos SUAS: Evolução e Recursos – Volume III. 2017. Disponível em: https://www.mds.gov.br/webarquivos/publicacao/assistencia_social/Cadernos/Suas_Evolucao_Recursos_III.pdf. Acesso em: 10 mar. 2025.

BRASIL. Ministério do Desenvolvimento Social. Reconstrução e dignidade ao povo brasileiro: Governo Federal celebra os 20 anos do Programa Bolsa Família. 2023. Disponível em: https://www.gov.br/mds/pt-br/noticias-e-conteudos/desenvolvimento-social/noticias-desenvolvimento-social/reconstrucao-e-dignidade-ao-povo-brasileiro-governo-federal-celebra-os-20-anos-do-programa-bolsa-familia. Acesso em: 11 mar. 2025.

BRASIL. Ministério do Desenvolvimento Social. Relatório Bolsa Família - Janeiro de 2017. 2017. Disponível em: https://mds.gov.br/webarquivos/sala_de_imprensa/boletins/boletim_bolsa_familia/relatorios/relatorio_11012017/1478611065.html. Acesso em: 10 fev. 2025.

COCHRANE. Entendendo o valor de p. Disponível em: https://eme.cochrane.org/entendendo-o-valor-de-p/. Acesso em: 15 mar. 2025.

COLOSIMO, Enrico A. Análise de dados categóricos: modelo de regressão de Poisson. Departamento de Estatística, Universidade Federal de Minas Gerais. Disponível em: https://www.est.ufmg.br/~enricoc/pdf/categoricos/Modelo_Poisson.pdf. Acesso em: 29 mar. 2025.

CNN BRASIL. Identificamos 3,7 milhões de fraudes em benefícios, diz ministro à CNN. 2023. Disponível em: https://www.cnnbrasil.com.br/economia/macroeconomia/identificamos-37-milhoes-de-fraudes-em-beneficios-diz-ministro-a-cnn/. Acesso em: 15 mar. 2025.

FILGUEIRAS, Cristina Almeida C. Controle e transparência na gestão do Programa Bolsa Família. Primeira Mostra Nacional de Estudos sobre o Programa Bolsa Família, 24 e 25 de novembro de 2008. Disponível em: https://ipcid.org/publication/mds/28M.pdf. Acesso em: 25 mar. 2025.

FERREIRA, Raquel Rossi. Regressão Logística Geograficamente Ponderada na Análise de Risco de Crédito. 2021. Universidade Federal do Rio Grande do Sul, Porto Alegre. Disponível em: https://lume.ufrgs.br/handle/10183/235629. Acesso em: 10 fev. 2025.

FERNANDES, Victor Vinícius. Contribuições sobre o envelope simulado na análise diagnóstico em modelos de regressão. 2019. Universidade de São Paulo. Disponível em: https://www.teses.usp.br/teses/disponiveis/104/104131/tde-07082019-113800/publico/VictorViniciusFernandes_revisada.pdf. Acesso em: 30 mar. 2025.

FLACK, V.; FLORES, R. Using simulated envelopes in the evaluation of normal probability plots of regression residuals. 1989. Acesso em: 15 mar. 2025.

GORI, Ricardo. Econometria I: Capítulo 12 - Heterocedasticidade. 2005. Disponível em: https://www4.eco.unicamp.br/docentes/gori/images/arquivos/EconometriaI/Econometria_Cap12_Heterocedasticidade.pdf. Acesso em: 29 mar. 2025.

LAMFO - Laboratório de Aprendizado de Máquina em Finanças e Organizações. Diagnóstico em regressão. 2019. Disponível em: https://lamfo-unb.github.io/2019/04/13/Diagnostico-em-Regressao/. Acesso em: 29 mar. 2025.

MARTINS, A. A. Aplicação de Análise de Risco de Crédito com o uso das Técnicas de Regressão Logística e Árvores de Decisão. 2019. Monografia de especialização, Universidade Federal de Minas Gerais. Disponível em: https://repositorio.ufmg.br/bitstream/1843/35524/1/TCC%20UFMG%20CURSO%20ESTAT%C3%8DSTICA%20ANDRIGO%20ANDRADE%20MARTINS.pdf. Acesso em: 15 mar. 2025.

McNULTY, K. Handbook of Regression Modeling in People Analytics: With Examples in R, Python and Julia. Chapman & Hall/CRC, 2021. Disponível em: https://peopleanalytics-regression-book.org/bin-log-reg.html. Acesso em: 29 mar. 2025.

MEMORIAL DA DEMOCRACIA. Combate à fome. Disponível em: https://memorialdademocracia.com.br/card/combate-a-fome/7. Acesso em: 15 mar. 2025.

MELGAREJO, Ana Paula Bento. Eficiência do Controle do Programa Bolsa Família na Perspectiva da Gestão por Resultados. 2011. Dissertação (Mestrado Profissional em Gestão Empresarial) – FGV, Brasília. Disponível em: https://repositorio.fgv.br/server/api/core/bitstreams/64a82b2e-8969-4717-a171-dc54fae4c1aa/content. Acesso em: 25 mar. 2025.

MESQUITA, Camile Sahb. O Programa Bolsa Família: uma análise de seu impacto e alcance social. 2007. Dissertação (Mestrado em Política Social) – Universidade de Brasília, Brasília. Disponível em: https://repositorio.unb.br/bitstream/10482/3144/1/2007_CamileSahbMesquita.pdf. Acesso em: 25 mar. 2025.

MOSTAFA, Joana; SANTOS, Thuany dos. Limitações de um teste de meios via predição de renda: evidências de uma aplicação no Programa Bolsa Família. Instituto de Pesquisa Econômica Aplicada (IPEA), 2017. Disponível em: https://repositorio.ipea.gov.br/handle/11058/7234. Acesso em: 10 fev. 2025.

MENDES, Cassandro; SAMPAIO, Luciano. Programa Bolsa Família e a importância da credibilidade do Governo: Uma digressão através da teoria dos jogos. XXXVI Encontro Nacional de Economia, 2008. Acesso em: 25 mar. 2025.

SILVA, Tiago Falcão (org.). Bolsa Família 15 Anos (2003-2018). Brasília: ENAP, 2018. 530 p. ISBN 978-85-256-0100-1. Disponível em: https://repositorio.enap.gov.br/jspui/bitstream/1/3647/4/15%20Anos%20Bolsa%20Fam%C3%ADlia.pdf. Acesso em: 25 mar. 2025.

SOUZA, Pedro H. G. FERREIRA de; BRUCE, Raphael. Uma avaliação final da focalização e da efetividade contra a pobreza do Programa Bolsa Família, em perspectiva comparada. Brasília: IPEA, 2022. Disponível em: https://repositorio.ipea.gov.br/handle/11058/11560. Acesso em: 25 mar. 2025.