Universidade de Brasília

Departamento de Estatística

Violência Doméstica e Familiar contra a Mulher: Uma aplicação da Regressão Logística

Álefe Lacerda Gomes Santos

Brasília

2021

Álefe Lacerda Gomes Santos

Violência Doméstica e Familiar contra a Mulher: Uma aplicação da Regressão Logística

Orientadora: Juliana Betini Fachini Gomes

Coorientadora: Maria Teresa Leão Costa

Relatório Final apresentado para o Departamento de Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Brasília

2021

Agradecimentos

À Deus, pela minha vida e por todos os obstáculos superados. À Aloísia Gomes, minha avó paterna, por me proporcionar oportunidades de educação e crescimento profissional. Ao Áthila Côrtes, meu noivo, por todo os momentos difíceis em que esteve ao meu lado, sempre me dando apoio, carinho e amor. À minha família e amigos queridos, que me incentivaram nos momentos mais difíceis e compreenderam a minha ausência enquanto eu me dedicava à realização deste trabalho. Aos professores, pelos ensinamentos que agregaram meu crescimento pessoal e profissional. Por fim, e não menos importante, as minhas orientadoras, exemplos de mulher, que me tranquilizaram em momentos de aflição e me motivaram a dar o melhor de mim.

Sumário

Resumo

Este estudo buscou verificar quais fatores socioeconômicos da população feminina brasileira contribuem para explicar a percepção da ocorrência de Violência Doméstica e Familiar contra a Mulher. As análises foram fundamentadas na Pesquisa de Opinião sobre Violência Doméstica realizada pelo Instituto DataSenado, em que foram entrevistadas 2400 mulheres com idade superior a 15 anos e com acesso a telefonia fixa e/ou móvel. Como resultado, utilizando técnicas de regressão logística para dados amostrais complexos, foi constatado que religião, renda, idade e força de trabalho são fatores significativos. Verificou-se também que o ciclo de violência é alimentado pela tolerância e pela falta de conhecimento sobre as especificidades desse crime. Propõem-se outros estudos nesta esfera com finalidade de maior compreensão do fenômeno da violência contra mulheres, em especial estudos que tenham como foco os tipos de violência classificados pela Lei Maria da Penha.

Palavras-chave: violência doméstica; mulheres; regressão logística; violência de gênero; perfil de vítima.

Abstract

This study sought to verify which socioeconomic factors of the Brazilian female population contribute to explain the perception of the occurrence of Domestic and Family Violence against Women. The analyzes were based on the Opinion Poll on Domestic Violence conducted by the DataSenado Institute, in which 2400 women aged over 15 years and with access to landline and/or mobile telephony were interviewed. As a result, using logistic regression techniques for complex sample data, it was found that religion, income, age and workforce are significant factors. It was also found that the cycle of violence is fueled by tolerance and lack of knowledge about the specifics of this crime. Other studies in this sphere are proposed in order to better understand the phenomenon of violence against women, especially studies that focus on the types of violence classified by the Maria da Penha Law.

Keywords: domestic violence; women; logistic regression; gender violence; victim profile.

1 Introdução

A violência contra a mulher, mais que um problema individual, é um fenômeno histórico e social. Na colônia, no Império e até nos primórdios da República, as mulheres eram submetidas a padrões estéticos e comportamentais. Mecanismos sociais eram usados para inibir atitudes divergentes das funções designadas a elas.

> O modelo familiar da época era hierarquizado pelo homem, sendo que desenvolvia um papel paternalista de mando e poder, exigindo uma postura de submissão da mulher e dos filhos. Esse modelo veio à sofrer modificações a partir da Revolução Industrial, quando as mulheres foram chamadas ao mercado de trabalho, descobrindo assim, a partir de então, o direito à liberdade, passando a almejar a igualdade e a questionar a discriminação de que sempre foram alvos. Com essas alterações, a mulher passou a participar, com o fruto de seu trabalho, da mantença da família, o que lhe conferiu certa independência. Começou ela a cobrar uma participação do homem no ambiente doméstico, impondo a necessidade de assumir responsabilidade dentro de casa e partilhar cuidado com os filhos (DIAS, 2004, p. 22-24).

A busca por liberdade, independência financeira e igualdade de gênero foi fortalecida com a inserção da mulher no mercado. A mídia impressa permitia ressaltar a importância das mulheres no país e expor a necessidade de educação em prol delas mesmas e da participação política, reclamando o direito de voto. A reivindicação dos direitos femininos permitiu que as mulheres ocupassem seu lugar na sociedade, ainda que de forma lenta.

O Brasil vem evoluindo com o avanço dos movimentos feministas no país e no mundo, mas o controle do homem sobre a mulher persiste na memória social (PRIORE, 2011, p 300).

> No decorrer do atual século, a sociedade reproduz a subordinação da mulher perante o homem através das tradições e costumes, e desse modo, banaliza e naturaliza uma opressão sofrida por décadas e que até hoje reflete em diversos setores sociais dos quais o sexo feminino esteja presente não somente na esfera familiar, tampouco apenas no âmbito trabalhista, na mídia ou na política (ESSY, 2017, p. 1).

A violência doméstica e familiar contra a mulher é a manifestação extrema de concepções desiguais de gênero, historicamente construídas, que determinam os comportamentos femininos e masculinos tidos como socialmente adequados. Essas concepções vigoram, com pequenas variações, nos campos social, político, cultural e econômico da maioria absoluta das sociedades e culturas e atuam em muitos casos em que agressões acontecem para 'justificar' ou minimizar a responsabilização de quem cometeu o ato violento, atribuindo as ações praticadas por uma pessoa à biologia ou, pior ainda, a quem foi vítima da agressão. (DOSSIÊ - Agência Patrícia Galvão).

Fatores como uso de álcool, drogas, ciúmes, términos ou traições são meros estopins para a ocorrência da violência e não justificam as agressões ocorridas. Matthew Gutmann, antropólogo especialista em masculinidade da Universidade Brown (EUA), afirma que "É muito comum o uso de termos como genes, hormônios ou hereditariedade para explicar ou desculpar o comportamento humano". Sergio Barbosa, filósofo coordenador do trabalho realizado com homens autores de agressão na cidade de São Paulo pela ONG Coletivo Feminista Sexualidade e Saúde, complementa que "[...] a própria construção da masculinidade que desencadeia esse exercício da violência sobre as mulheres."

Os primeiros estudos relacionados a violência contra a mulher evidenciavam que tais agressões ocorriam dissociadas de grupo social, religioso ou cultural. Em contrapartida, estudos recentes apontam pobreza familiar e baixo nível de escolaridade como fatores associados ao risco de violência contra a mulher, acarretando às mulheres pobres e negras uma carga mais pesada e maior exposição às violências. (KRONBAUER,2005)

É importante ressaltar uma variação considerável em estudos devido às diferentes populações estudadas. Muitas pesquisas relacionadas a violência de gênero incluem todas as mulheres de uma determinada faixa etária, etnia, raça, religião ou renda enquanto outras entrevistam apenas um nicho de mulheres com estado civil, poder aquisitivo e perfis predominantes, por exemplo. Nesses casos, tanto a idade quanto estado civil ou determinada característica pode estar associada ao risco de uma mulher ser vítima de abuso por parte do parceiro, pois não foram utilizados mecanismos para controlar o viés das informações.

O processo de amostragem da pesquisa e os critérios de seleção utilizados podem, assim, acometer consideravelmente as estimativas sobre a predominância de um determinado perfil nas análises. Essas estimativas também podem variar segundo a fonte dos dados.

> Diversos estudos nacionais produziram estimativas sobre a predominância de violência de gênero – estimativas essas que geralmente estão abaixo daquelas obtidas em estudos menores e em profundidade acerca das experiências das mulheres com relação à violência. Os estudos menores e em profundidade tendem a se concentrar mais na interação entre os entrevistadores e os entrevistados. Esses estudos também tendem a cobrir o assunto muito mais detalhadamente do que a maioria das pesquisas nacionais. As estimativas de predominância entre os dois tipos de estudos também podem variar devido a alguns fatores anteriormente mencionados, inclusive diferenças nas populações do estudo e nas definições de perfis (ORGANIZAÇÃO MUNDIAL DA SAÚDE, 2002, p. 94).

Por outro lado, quando analisamos aspectos referentes a dinâmica das relações familiares, em especial vínculos conjugais, notamos diversas transformações ao longo das gerações, ora com avanços, ora com retrocessos. Consequentemente, novas formas

e padrões de comportamentos e relacionamentos, além de novos ideais, surgem proporcionalmente às rápidas e amplas mudanças de pensamento na sociedade.

Diante do exposto, o presente trabalho tem por objetivo verificar quais fatores socioeconômicos da população feminina brasileira contribuem para explicar a percepção da ocorrência de Violência Doméstica e Familiar contra a Mulher.

A análise será fundamentada na Pesquisa de Violência Doméstica e Familiar contra a Mulher realizada pelo Instituto DataSenado, utilizando como ferramente de análise o *software* R (versão 4.1). Essa pesquisa aborda temas relacionados à percepção de desrespeito à Lei Maria da Penha, sensação de proteção, experiências de violência e de denúncias, perfil dos agressores, perfil das vítimas entre outros. O levantamento é realizado por entrevistas telefônicas, a partir de uma amostra probabilística representativa de mulheres com acesso ao telefone fixo e/ou móvel.

Os capítulos conseguintes abrangem a metodologia a ser aplicada neste trabalho, bem como as análises dos dados e conclusões obtidas.

2 Regressão Logística

Uma variável qualitativa constitui-se de uma escala de medidas divididas em categorias. Compostas por duas ou mais categorias, as variáveis demográficas em estudo (renda, escolaridade, religião, etc.) são classificadas em nominal, ausência de hierarquia entre as categorias, ou ordinal, as categorias seguem uma ordem previamente estabelecida. Caso possua apenas duas categorias, são denominadas variáveis qualitativas binárias ou dicotômicas. A escala de medidas de uma variável determina quais métodos estatísticos são apropriados.

A análise de regressão é uma metodologia estatística utilizada para estudar e quantificar a relação existente entre duas ou mais variáveis. Essa metodologia permite que a variável de interesse seja prevista a partir de outra, ou outras, sendo amplamente utilizada em negócios, ciências sociais e comportamentais, ciências biológicas e muitas outras disciplinas.

Em um experimento realizado para estudar a letalidade de uma determinada toxina, animais eram expostos a diferentes dosagens dessa substância com o intuito de identificar qual quantidade administrada era letal. Foi nesse Bioensaio que utilizou-se pela primeira vez o modelo de regressão para dados binários, em que a variável de interesse no estudo era a proporção de animais mortos pelas dosagens dessa substância tóxica. (Finney, 1973).

Este capítulo compreende o modelo de regressão em que a variável resposta é qualitativa com dois resultados possíveis (dicotômica) e, portanto, pode ser representada por uma variável indicadora assumindo valores 0 e 1, tais como sim (1) e não (0) para ocorrência de violência doméstica e familiar contra a mulher.

2.1 Formulação do Modelo

A primeira etapa do processo de modelagem estatística é a formulação do modelo, que consite na escolha da distribuição de probabilidade adequada para a variável resposta, das variáveis explicativas a serem estudadas e da função de ligação apropriada para descrever as principais características da variável de interesse (NETER,2005).

Considere um modelo de regressão linear

$$Y_i = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \varepsilon_i = \boldsymbol{X_i'\beta} + \varepsilon_i \qquad Y_i = 0, 1$$

em que a variável resposta é binária, assumindo valores 0 e 1. Uma vez que $E[\varepsilon_i] = 0$, o valor esperado de Y_i é dado por:

$$E[Y_i] = \boldsymbol{X_i'\beta}. \tag{2.1.1}$$

Considere que a variável aleatória Y_i tem distribuição Bernoulli com função de probabilidade $f(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$, em que:

Tabela 1: Função de Probabilidade da Variável Resposta

Y_i	Probabilidade
1	$P(Y_i = 1) = \pi_i$
0	$P(Y_i = 0) = 1 - \pi_i$

Por definição, o valor esperado de uma variável aleatória é

$$E[Y_i] = \sum_i Y_i P(Y_i) = 1(\pi_i) + 0(1-\pi_i) = \pi_i. \tag{2.1.2}$$

A partir das equações 5.1.1 e 5.1.2 conclui-se que:

$$E[Y_i] = \boldsymbol{X_i'\beta} = \pi_i. \tag{2.1.3}$$

Quando comparado à regressão linear usual, o modelo logístico apresenta divergências, pois a variável resposta Y_i é binária. Considera-se três problemas a seguir, usando um modelo de regressão linear como ilustração.

- **Erros não normais:** Para uma variável de resposta binária, 0 e 1, cada termo de erro $\varepsilon_i = Y_i - \boldsymbol{X_i'\beta}$ assume apenas dois valores:

$$\varepsilon_i = 1 - \boldsymbol{X_i'\beta}, \qquad Y_i = 1$$

$$\varepsilon_i = -\boldsymbol{X_i'\beta}. \qquad Y_i = 0$$

- **Heterocedasticidade:** As variações do erro serão desiguais em níveis diferentes do vetor de variáveis aleatórias $\boldsymbol{X_i}$, pois dependem dele.

$$Var[Y_i] = E[Y_i - E(Y_i)] = \pi_i(1-\pi_i) = (\boldsymbol{X_i'\beta})(1 - \boldsymbol{X_i'\beta})$$

- **Limitações do valor esperado:** Visto que o valor esperado representa uma probabilidade, está restrito entre

$$0 \leq E[Y_i] = \pi_i \leq 1.$$

A Figura 1, cuja curva apresenta um formato de sigmoide, ilustra uma função de resposta não linear para o valor esperado da variável indicadora. O parâmetro β indica

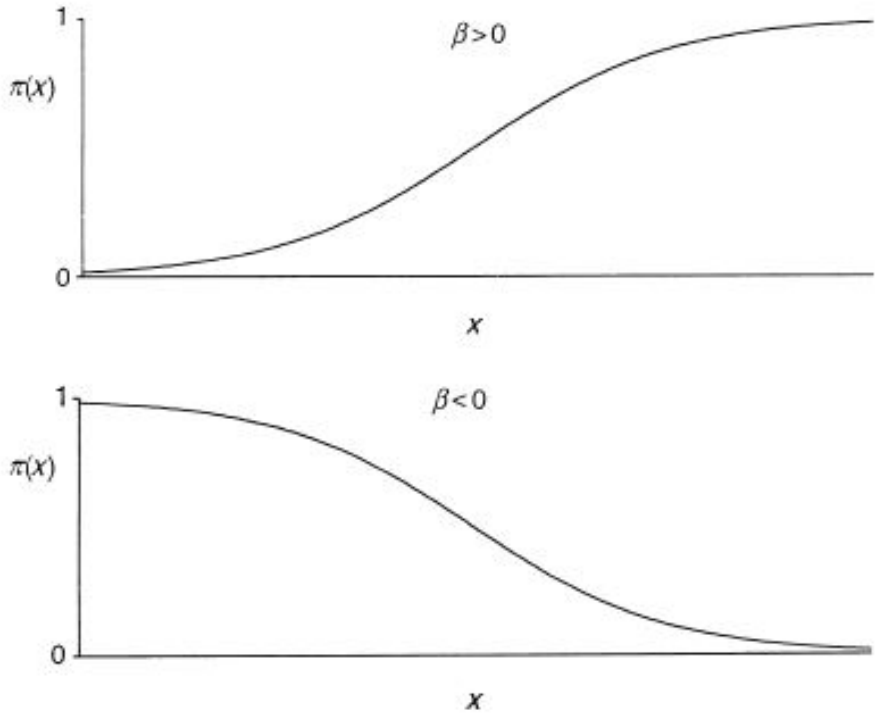

Figura 1: Comportamento da Curva

relação crescente ($\beta > 0$) ou decrescente ($\beta < 0$).

Para o caso em que (β=0), dizemos que a covariável X_i não tem efeito significativo sob a variável de interesse Y_i. Neste estudo, a variável indicadora Y_i refere-se a ocorrência de violência doméstica e familiar contra a mulher, conforme dito anteriormente. Portanto, o valor esperado $E[Y_i] = \pi_i$, limitado no intervalo [0,1], consiste na probabilidade de ocorrência do evento de interesse $P(Y_i = 1)$ quando o nível da variável preditora é $\boldsymbol{X_i}$.

Para resolver o problema de não linearidade, apresentado na Figura 1, utiliza-se uma **função de ligação** entre o valor esperado $E[Y_i]$ e o preditor linear $\boldsymbol{X_i'\beta}$. A função *logit* é amplamente utilizada em modelos cuja variável resposta é dicotômica.

Dessa forma, o modelo de regressão logística é definido por:

$$Y_i = E[Y_i] + \varepsilon_i. \tag{2.1.4}$$

Visto que os erros ε_i dependem da distribuição Bernoulli associada a variável aleatória Yi em que:

$$E[Y_i] = \pi_i = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}. \tag{2.1.5}$$

A função de ligação *logit*, obtida de forma inversa à equação 3.1.5 e escrita como o logaritmo da chance (odds),

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{X_i'\beta}, \tag{2.1.6}$$

é a função de ligação canônica (padrão) da Regressão Logística.

2.2 Inferência

Após a formulação do modelo, seguimos para a etapa de ajuste, que compreende o processo de estimação dos parâmetros pelo método da máxima verossimilhança e das medidas de adequação dos valores estimados (NETER,2005).

Considere as variáveis aleatórias $Y_1, ..., Y_n$ independentes com distribuição $Bernoulli$,

$$f(y_i, \pi) = \pi^{y_i}(1-\pi)^{1-y_i}.$$

O estimador de máxima verossimilhança $\widehat{\boldsymbol{\beta}}$ para $\boldsymbol{\beta}$, vetor de parâmetros do modelo, é o valor que maximiza a função de verossimilhança $L(\boldsymbol{\beta}, y_i) = \prod_{i=1}^{n} f(y_i, \boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$. Uma vez que a função logarítmica é monotônica, o parâmetro $\widehat{\boldsymbol{\beta}}$ equivale também ao ponto de máximo da função de log-verossimilhança $l(\boldsymbol{\beta}, y_i) = log[L(\boldsymbol{\beta}, y_i)]$,

$$l(\boldsymbol{\beta}, y_i) = log_e\left[\prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}\right] = \sum_{i=1}^{n}\left[y_i\ log_e\left(\frac{\pi_i}{1-\pi_i}\right)\right] + \sum_{i=1}^{n} log_e(1-\pi_i). \tag{2.2.1}$$

O estimador de máxima verossimilhança para β_i é obtido derivando-se a função de log-verossimilhança em relação a β_i e em seguida igualando-se a equação obtida a zero,

$$\frac{\partial l(\boldsymbol{\beta}, y_i)}{\partial \beta_i} = U_i = 0. \tag{2.2.2}$$

Encontradas as estimativas de máxima verossimilhança, substituem-se esses valores no modelo de regressão logística para obter a função de resposta ajustada. Denota-se $\hat{\pi}_i$, o valor ajustado ou predito para a i-ésima observação,

$$\hat{\pi}_i = \frac{exp[\boldsymbol{X_i'\hat{\beta}}]}{1 + exp[\boldsymbol{X_i'\hat{\beta}}]}. \tag{2.2.3}$$

Após o modelo logístico ser ajustado, os próximos passos usuais são examinar a adequação da função de resposta ajustada e, se o ajuste for adequado, realizar inferências e previsões.

Os métodos utilizados na regressão logística são similares aos da regressão linear,

em que são testados os coeficientes de regressão, estimativas das respostas médias e previsões de novas observações. Primeiramente, seleciona-se o modelo mais adequado. Após a seleção, analisa-se a adequabilidade do modelo e compara-se os valores observados com os valores preditos da variável resposta para posteriormente interpretá-lo.

Os procedimentos de inferência apresentados nessa seção dependem de grandes tamanhos de amostra. Para grandes amostras, sob condições geralmente aplicáveis, os estimadores de máxima verossimilhança para regressão logística são aproximadamente normais, com pouco ou nenhum viés, e com variâncias e covariâncias aproximadas que são funções das derivadas parciais de segunda ordem do logaritmo da função de verossimilhança.

As inferências sobre os coeficientes de regressão para o modelo de regressão logística são baseadas no seguinte resultado aproximado quando o tamanho da amostra é suficientemente grande:

$$\frac{\hat{\beta}_i - \beta_i}{\widehat{SE}(\widehat{\beta}_i)} \sim N(0,1), \tag{2.2.4}$$

em que $\widehat{SE}$ é a estimativa do erro padrão obtidas atavés da matrix de derivadas parciais.

2.2.1 Teste de Wald

Sob a hipótese de ausência de efeito dos parâmetros,

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

esse teste é obtido comparando-se o estimador de máxima verossimilhança do parâmetro $\hat{\beta}_i$ à estimativa do respectivo erro padrão $\widehat{SE}$. A razão resultante

$$W = \frac{\widehat{\beta}_i}{\widehat{SE}(\widehat{\beta}_i)},$$

segue uma distribuição normal. A regra de decisão é dada por:

$$|W| \leq z_{1-\alpha/2}, \qquad \text{Não se rejeita } H_0,$$

$$|W| > z_{1-\alpha/2}, \qquad \text{Rejeita-se } H_0.$$

Segundo Hosmer e Lemeshow (2000), uma maneira alternativa de testar a significância do modelo é calculando e interpretando o intervalo de confiança para os parâmetros

de interesse,

$$IC[\beta_i, 100(1-\alpha)\%] = [\beta_i - z_{1-\alpha/2}\ \widehat{SE}(\widehat{\beta}_i)\ ;\ \beta_i + z_{1-\alpha/2}\ \widehat{SE}(\widehat{\beta}_i)],$$

em que $z_{1-\alpha/2}$ denota o quantil da distribuição normal e $\widehat{SE}(.)$ é o estimador do erro padrão do respectivo parâmetro estimado. Em particular, o intervalo de confiança para as estimativas do vetor de parâmetros $\boldsymbol{\beta}$ são baseadas no teste de Wald.

2.2.2 Teste da Razão de Verossimilhança

Frequentemente, há interesse em determinar se um subconjunto das variáveis explicativas X_i em um modelo de regressão logística múltipla pode ser descartado, isto é, testar se os coeficientes de regressão associados são significativos para o modelo parcimonioso. O teste da razão de verossimilhança compara modelos completos e reduzidos (NETER,2005).

Começamos com o modelo logístico completo com função de resposta:

$$\pi_i = [1 + exp(-\boldsymbol{X_i'\beta_F})]^{-1} \quad \text{Modelo Completo} \tag{2.2.5}$$

em que $\boldsymbol{X_i'\beta_F} = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$.

Em seguida, calcula-se as estimativas de máxima verossimilhança para o modelo completo, agora denotado por $\hat{\boldsymbol{\beta}}_{\boldsymbol{F}}$, e avalia-se a função de verossimilhança $L(\boldsymbol{\beta}, Y_i)$ quando $\boldsymbol{\beta_F} = \hat{\boldsymbol{\beta}}_{\boldsymbol{F}}$.

Sob a hipótese de ausência de efeito dos parâmetros,

$$H_0 : \beta_1 = ... = \beta_p = 0,$$

$$H_1 : \beta_i \neq 0, i = 1, 2, ..., p$$

em que, por conveniência, organizamos o modelo de forma que os últimos p - k coeficientes sejam aqueles testados. O modelo logístico reduzido, portanto, tem a função de resposta

$$\pi = [1 + exp(-\boldsymbol{X_i'\beta_R})]^{-1} \quad \text{Modelo Reduzido} \tag{2.2.6}$$

em que $\boldsymbol{X_i'\beta_R} = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$.

Em seguida, calcula-se as estimativas de máxima verossimilhança para o modelo reduzido, agora denotado por $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$, e avalia-se a função de verossimilhança $L(\boldsymbol{\beta}, y_i)$ quando $\boldsymbol{\beta_R} = \hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$.

O teste da razão de verossimilhança, que consiste em obter a diferença entre o máximo do log dessa função para o modelo de interesse e para o modelo mais completo, é baseado na estatística

$$\boldsymbol{G^2} = -2log_e\left[\frac{L_R}{L_F}\right] = -2[log_e L_R - log_e L_F], \tag{2.2.7}$$

em que

- L_F: Logaritmo da função de verossimilhança do modelo completo (p parâmetros).
- L_R: Logaritmo do modelo reduzido em investigação (k parâmetros), onde $k < p$.

Observe que, se a razão L_R/L_F for pequena, indicando que H_1 é a conclusão apropriada, então $\boldsymbol{G^2}$ é grande. Assim, valores elevados de $\boldsymbol{G^2}$ rejeitam H_0. A teorema de grandes amostras afirma que quando **n** é grande, a estatística $\boldsymbol{G^2}$ tem distribuição χ^2_p com $p - k$ graus de liberdade, em que p e k são os parâmetros a serem estimados.

Ambos os testes, $\boldsymbol{G}^2$ e $\boldsymbol{W}$, são baseados no estimador de máxima verossimilhança dos parâmetros.

Outros dois critérios utilizados para avaliar a qualidade de ajuste do modelos são o AIC (Akaike's Information Criterion) e o BIC (Critério de Informação Bayesiano). Proposto por Akaike, 1974, a medida AIC utiliza a Informação de Kullback-Leibler para verificar a adequabilidade do modelo estudado (Paula,2019). Em caso de empate nos valores, utiliza-se o *deviance* com a mesma interpretação, quanto menor o valor da medida, melhor é o ajuste do modelo.

As medidas AIC e BIC são calculadas da seguinte forma

$$AIC = -2log(\boldsymbol{L}(\widehat{\boldsymbol{\beta}})) + 2p;$$
$$BIC = -2log(\boldsymbol{L}(\widehat{\boldsymbol{\beta}})) + p\ log(n)$$

em que $\boldsymbol{L}(\widehat{\boldsymbol{\beta}})$ é a verossimilhança do modelo, p é o número de parâmetros e n o tamanho da amostra.

2.3 Seleção de Variáveis

Assim como na Regressão Linear, o processo de seleção de variáveis explicativas torna-se mais desafiador à medida que seu número aumenta, devido ao rápido aumento dos possíveis efeitos e interações, além da interferência nos coeficientes resultantes. O modelo deve ser completo o suficiente para ter um bom ajuste aos dados, porém simples de interpretar. Popularmente conhecido como **modelo parcimonioso** o que explica melhor

o efeito das variáveis explicativas com menor número de parâmetros possível. Inicia-se esse processo pela análise cuidadosa das variáveis explicativas, um passo importante quando há muitas variáveis como potenciais preditoras. Algumas dessas escolhas são puramente a critério do analista (Agresti, 2019).

De acordo com Thomas Lumley (2010), as variáveis preditoras podem ser classificadas em três categorias:

- **Variáveis de Interesse**: Questiona-se subjetivamente a relação entre essa variável e a variável resposta. O pesquisador é livre para escolher como resumir essa relação, seja como uma diferença entre categorias, seja com uma tendência linear, seja com uma curva de exposição detalhada. A escolha dependerá do tamanho da amostra, do nível de conhecimento prévio e do uso para o qual as conclusões serão colocadas.
- **Variáveis Confundidoras**: também conhecida como **fator de confusão**, essa variável afeta não somente a variável resposta, como também variáveis explicativas, causando uma associação espúria. Devem ser modeladas com precisão a fim de evitar efeitos adversos provindos da multicolinearidade e possíveis associações sejam controladas.
- **Variáveis de Precisão**: Estas não são associadas às variáveis de interesse, e portanto não afetam a interpretação dos coeficientes. Porém, elas ajudam a explicar melhor a variação existente na variável resposta e reduzem os erros. Ususalmente, não há interesse em avaliar seus coeficientes. Variáveis de precisão são relevantes se possuem forte associação com a variável resposta. Pouca interferência nos resíduos torna-se irrelevante.

Variáveis que não estão inclusas nessas três categorias são aquelas cuja associação com a variável resposta é irrelevante. De acordo com Hosmer e Lemeshow (2000), o critério para inclusão de uma variável no modelo pode sofrer alterações de acordo com o problema em estudo ou o pesquisador que o está avaliando. O objetivo principal da seleção de variáveis é encontrar o modelo mais parciomonioso, ou seja, menor conjunto de parâmetros possível que melhor se adeque ao modelo e explique os dados. A razão em minimizar o número de variáveis no modelo está em encontrar um modelo numericamente estável e mais facilmente generalizado. Quanto maior o número de variáveis incluídas no modelo, maior serão as estimativas do erro padrão e mais dependente o modelo se torna das observações.

O processo de seleção de variáveis inicia-se analisando cautelosamente cada variável de forma individual. Para variáveis nominais, ordinais e contínuas com poucos valores inteiros, sugere-se a elaboração de uma tabela de contingência entre a variável de saída (y = 0,1) e a variável de entrada em k níveis. O teste da razão de verossimilhança com

k-1 graus de liberdade é equivalente ao teste da razão de verossimilhança para a significância dos coeficientes para as k-1 dummies em um modelo univariado. Uma vez que o teste de Qui-Quadrado de Pearson é assintoticamente equivalente ao teste da razão de verossimilhança, este pode ser usado.

Sugere-se estimar a razão de chances individual entre as variáveis que tenham associação no mínimo moderada. Deve-se também ter cautela com tabelas de contingência que apresentem valores nulos, pois estes podem afetar consideravelmente a análise. Nesses casos, há algumas soluções: reorganizar as categorias de forma a eliminar os valores nulos, eliminar completamente determinada categoria ou, caso a variável seja ordinal, modelar a variável de forma que esta seja considerada contínua. Após a análise univariada, a partir do teste de Qui-Quadrado de Pearson, seleciona-se as variáveis para a análise multivariada. Quaisquer variáveis com p-valor inferior a 0.25 são candidatas juntamente com aquelas consideradas importantes pelo pesquisador. Uma vez indentificadas as variáveis, inicia-se o processo de modelagem. O critério de seleção acima mencionado para a regressão logística foi sugerido por Mickey e Greenland (1989).

Esses autores mostram que níveis de significância tradicionalmente utilizados frequentemente falham em identificar variáveis importantes. Em contrapartida, o uso de níveis maiores podem incluir variáveis não significativas para o modelo. Vale lembrar que, apesar de não ter uma associação direta com a variável resposta, algumas variáveis explicativas, em conjunto com as demais presentes no modelo, podem melhorar o seu ajuste e sua acurácia. Por esses motivos, é de suma importância a revisão criteriosa de todas as variáveis utilizadas antes de obter o modelo final.

Um modelo não deve conter muitas variáveis explicativas quando há poucas respostas para determinada categoria. Peduzzi et al. (1996) sugerem p variáveis explicativas inferiores ao número de observações dividido por 10 na categoria de menor ocorrência. Em pesquisas e estudos, métodos de seleção de variáveis podem ser informativos quando usados com cautela. De acordo com Agresti(2019), são classificados da seguinte forma:

- **Método *Backward***: Também conhecido como eliminação reversa, o processo inicia-se com o modelo complexo (modelo com todas as variáveis explicativa) e prossegue removendo sequencialmente as variáveis sem efeito significativo. Exemplificando, elimina-se a variável com maior p valor, em que não se rejeita a hipótese de parâmetro nulo (igual a zero) no modelo. O processo é interrompido quando qualquer outra exclusão leva a um ajuste significativamente pior. Nesse método, não é recomendado remover variáveis com efeito significativo.

- **Método *Forward***: Esse processo tem início no modelo mais simples, onde não há variáveis explicativas, somente o intercepto. As variáveis explicativas, com efeito significativo, são adicionadas de forma sequencial para melhorar o ajuste do modelo.

O processo é interrompido quando não há mais melhorias no ajuste do modelo. Recomenda-se manter variáveis com efeito não significativo apenas quando estas contribuem para melhoria na adequabilidade do modelo.

- **Método *Stepwise***: É uma junção dos dois métodos anteriores. Inicia-se o processo com o modelo mais simples e adiciona-se ou remove-se sequencialmente variáveis explicativas até obter o melhor ajuste do modelo.

Em seguida, a importância de cada variável selecionada pode ser verificada através da análise da estatística de Wald e a comparação dos coeficientes estimados no modelos com os coeficientes estimados para o modelo univariado. Variáveis que não contribuem estatisticamente para o modelo, podem ser retiradas. Sugere-se que as variáveis inicialmente descartadas do modelo sejam incluídas posteriormente a fim de avaliar sua contribuição em conjunto com as demais.

Em qualquer um dos processos, para variáveis explicativas com mais de duas categorias, o processo deve considerar toda a variável em qualquer estágio, ao invés de apenas o indicador individual das variáveis. Caso contrário, o resultado dependerá da escolha da categoria as variáveis do indicador. Agresti (2019) recomenda adicionar ou eliminar a variável por inteiro e não somente seus indicadores.

2.4 Diagnóstico da Qualidade de Ajuste do Modelo

Após a escolha de modelos preliminares (candidatos), o próximo passo consiste em verificar o ajuste do modelo por meio de testes de adequação e resíduos antes de quaisquer interpretações do modelo. Em particular, precisamos examinar se a função de resposta estimada para os dados é monotônica e sigmoidal (NETER,2005).

Nessa etapa, pressume-se que os modelos preliminares contenham as variáveis principais bem como as interações necessárias. Os testes de qualidade de ajuste fornecem medidas gerais de ajuste do modelo e geralmente não são sensíveis quando o ajuste é ruim em apenas alguns casos.

O modelo está bem ajustado se (1) os valores preditos $\hat{y}_i$ estão próximos aos valores observados y_i e (2) a contribuição de cada par de valores $(\hat{y}_i, y_i)$ para as medidas resumo do modelo são assistemáticas e pequenas em relação à estrutura de erro do modelo. Na regressão logística, há diversas formas de mensurar a diferença entre os valores observados e os valores preditos da variável resposta.

As suposições para $Y_i|(X_1 = x_1, ..., X_n = x_n) \sim Ber(logit(\boldsymbol{X'\beta}))$ que devem ser verificadas na etapa de diagnóstico do modelo são:

- **Linearidade** na esperança transformada $E[Y_i|(X_1 = x_1, ..., X_n = x_n] = g^{-1}(\boldsymbol{X'\beta})$;

- **Distribuição de Probabilidade da Variável Resposta**, $Y \sim Bin(n, \pi_i)$;
- **Independência**, $Y_i's$ são independentes condicionalmente a $X_i's$.

Duas medidas podem ser utilizadas para verificar a adequabilidade e o ajuste do modelo: o resíduo de Pearson para o teste $\boldsymbol{\chi}^2$ de Pearson e o resíduo deviance para o teste de Verossimilhança $\boldsymbol{G^2}$.

2.4.1 Teste de Adequação Qui-quadrado de Pearson

Esse teste permite verificar a suposição de independência das variáveis aleatórias Y_i, porém detecta apenas grandes desvios de uma função de resposta logística por não ser sensível a pequenos desvios. As hipóteses de interesse são:

$$H_0 : E[Y_i] = 1 + exp[-\boldsymbol{X_i'\beta}]^{-1}$$

$$H_1 : E[Y_i] \neq 1 + exp[-\boldsymbol{X_i'\beta}]^{-1}$$

Dada a variável aleatória Y_i com distribuição Bernoulli cujos resultados são 1 e 0, os valores observados e esperados são obtidos conforme a seguir:

Tabela 2: Caption

	$Y_i = 1$	$Y_i = 0$
O_i	$\sum_i y_i$	$n - \sum_i y_i$
E_i	$n\pi_i$	$n(1 - \pi_i)$

Para a i-ésima variável, a estatística do teste de Pearson $\chi^2 = \sum_{i=1}^{n} r(y_i, \widehat{\pi}_i)$ é baseada no resíduo de Pearson que pode ser calculado pela seguinte equação:

$$r(y_i, \widehat{\pi}_i) = \frac{(O_i - E_i)^2}{E_i}.$$

Se a função de resposta logística for apropriada, X^2 segue aproximadamente uma distribuição χ^2 com $n - (p + 1)$ graus de liberdade quando n é grande e $p + 1 < n$. Tal como acontece com outros testes de qualidade de ajuste do qui-quadrado, é aconselhável que a maioria das frequências esperadas E_i sejam superiores a 5 e nenhuma inferior a 1 (NETER,2005).

Valores elevados da estatística de teste X^2 indicam que a função de resposta logística não é apropriada. A regra de decisão para testar as hipóteses anteriormente estabelecidas, ao controlar o nível de significância em α, é, portanto:

$$X^2 \leq \chi^2(1-\alpha, n-(p+1)), \qquad \text{Não se rejeita } H_0,$$
$$X^2 > \chi^2(1-\alpha, n-(p+1)), \qquad \text{Rejeita-se } H_0.$$

2.4.2 Teste de Adequação *Deviance*

O teste *Deviance* para modelos de regressão logística é completamente análogo ao teste F para falta de ajuste nos modelos de regressão linear simples e múltipla. Assim como o teste F para falta de ajuste e o teste de qualidade de ajuste do qui-quadrado de Pearson, assumimos que há p combinações únicas dos preditores denotados $X_1, ..., X_p$, o número de observações binárias repetidas em X_i é n_i, e a i-ésima resposta binária na combinação de preditor X_i é denotada Y_i (NETER,2005).

Esse teste é baseado no teste da razão de verossimilhança do modelo reduzido, em que o resíduo *deviance* mede o desvio, em termos de $-2log_eL(\boldsymbol{\beta}, y_i)$, entre o modelo saturado e o modelo reduzido (ver seção 3.2).

A soma dos quadrados dos resíduos (SQE) da regressão linear é denominado *Deviance* no modelo logístico, dado por $\boldsymbol{D} = \sum_{i=1}^{n} d(y_i, \widehat{\pi}_i)^2$, onde cada $d(y_i, \widehat{\pi}_i)$ é um componente da estatística *Deviance*, definido como:

$$d(y_i, \widehat{\pi}_i) = 2\left[y_i\ ln\left(\frac{y_i}{n\hat{\pi}_i}\right) + (n-y_i)\ ln\left(\frac{n-y_i}{n(1-\hat{\pi}_i)}\right)\right]^{1/2}, \tag{2.4.1}$$

Se modelo logístico ajustado for adequado e os tamanhos de amostra forem grandes, então o desvio seguirá aproximadamente uma distribuição qui-quadrado com $n-(p+1)$ graus de liberdade. Grandes valores do **Deviance** indicam que o modelo logístico ajustado é inadequado. Analogamente ao teste de qui-quadrado de pearson, testa-se as alternativas:

$$H_0 : E[Y_i] = 1 + exp[-\boldsymbol{X_i'\beta}]^{-1}$$

$$H_1 : E[Y_i] \neq 1 + exp[-\boldsymbol{X_i'\beta}]^{-1}.$$

A regra de decisão é dada por:

$$D \leq \chi^2(1-\alpha, n-(p+1)), \qquad \text{Não se rejeita } H_0,$$
$$D > \chi^2(1-\alpha, n-(p+1)), \qquad \text{Rejeita-se } H_0.$$

Ambas as estatísticas dos testes $\boldsymbol{G^2}$ e $\mathbf{D}$, sob a suposição de que o modelo esteja

adequado, tem distribuição χ^2 com $n-(p+1)$ graus de liberdade, em que $p+1$ é o número de parâmetros a serem estimados.

Em pesquisas amostrais complexas, são necessárias correções na estatística do teste de Pearson, tais como os ajustes de Rao-Scott. Alternativamente, a estatística de Wald, devido ao uso de uma estimativa apropriada da variância, que incorpora a complexidade do plano amostral e do efeito da estratificação, fornece uma estatística de teste assistoticamente válida. Sendo assim, não são necessários ajustes como na estatística de Pearson. Esta pode ser considerada uma vantagem quando o objetivo é obter inferências válidas (IBGE,2018).

A estatística de Wald, sob hipótese nula $H_0 : \beta_i = 0$, pode ser obtida calculando-se

$$\chi^2_W = \frac{(\hat{\pi}_i - \pi_i)^2}{\hat{V}(\hat{\pi}_i)} \tag{2.4.2}$$

em que $\hat{V}(\hat{\pi}_i)$ é uma estimativa da variância de aleatorização de $\hat{\pi}_i$, correspondente ao plano amostral efetivamente utilizado.

2.5 Métodos Gráficos de Diagnóstico

Nesta seção, será abordada a análise de resíduos para a regressão logística por métodos gráficos. Em todo o processo, devemos supor que as respostas são binárias, ou seja, nos concentramos no caso desagrupado, que resulta em duas tendências lineares com inclinação -1 (NETER, 2005).

A análise de resíduos na regressão logística é mais difícil que em modelos de regressão linear, pois a variável resposta Y_i assume apenas valores 0 e 1. Consequentemente, o i-ésimo resíduo comum, ε_i, assumirá um de dois valores possíveis:

$$\varepsilon_i = \begin{cases} 1 - \hat{\pi}_i, Y_i = 1 \\ -\hat{\pi}_i, Y_i = 0 \end{cases}$$

Os resíduos comuns não são normalmente distribuídos e, de fato, sua distribuição sob a suposição de que o modelo ajustado está correto é desconhecida. Gráficos de resíduos comuns *vs* valores ajustados ou variáveis preditoras geralmente não são informativos. Por isso, outros resíduos precisam ser definidos e utilizados na análise de resíduo.

Resíduo de Pearson Para uma melhor análise, os resíduos comuns são transformados, divide-se pelo erro padrão estimado de Y_i, para que esse seja melhor comparável. Os resíduos de Pearson resultantes são dados por:

$$r_{P_i} = \frac{Y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}} \tag{2.5.1}$$

Os resíduos de Pearson estão diretamente relacionados à estatística de qualidade do ajuste do qui-quadrado de Pearson, pois a soma dos quadrados dos resíduos de Pearson é numericamente equivalente à estatística do teste qui-quadrado de Pearson. Portanto, o quadrado de cada resíduo de Pearson mede a contribuição de cada resposta binária para a estatística do teste qui-quadrado de Pearson. Observe que a estatística de teste não segue uma distribuição qui-quadrado aproximada para dados binários sem replicações.

Resíduo Estudentizado de Pearson

Os resíduos de Pearson não têm variância unitária, uma vez que nenhuma provisão foi feita para a variação inerente no valor ajustado $\hat{\pi}_i$. Um procedimento melhor é dividir os resíduos comuns por seu respectivo desvio padrão estimado. Este valor é aproximado por $\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)(1-h_{ii})}$, em que h_{ii} é o i-ésimo elemento diagonal da matriz hessiana estimada para a regressão logística, $H = \hat{W}^{1/2}X(X'\hat{W}X)^{-1}X'\hat{W}^{1/2}$, em que $\hat{W}$ é a diagonal da matriz de covariância. Os resíduos estudantizados de Pearson são definidos como:

$$r_{SP_i} = \frac{r_{P_i}}{\sqrt{1-h_{ii}}} \tag{2.5.2}$$

O resíduo *Deviance* foi definido anteriormente (ver seção 3.4.2).

2.5.1 Linearidade

Os gráficos de resíduos $Y_i - \hat{Y}_i$ *vs* valores ajustados (primeira linha) para o conjunto de dados (segunda linha) respeitando a suposição de linearidade na regressão logística são apresentados na Figura 2. O pressuposto de linearidade entre o valor esperado de Y_i e as variáveis aleatórias X_i é essencial na aplicação de GLMs. Para verificar através do gráfico, a linha vermelha precisa estar bem ajustada aos valores (centralizada), sem apresentar aspectos curvilíneos severos ou estimar somente um dos valores (0 ou 1). Se essa suposição falhar, todas as conclusões que poderiam ser extraídas da análise são suspeitas de serem falhas. Quando isso ocorre, pode-se aplicar uma transformação linear nas variáveis preditoras ou adicionar possíveis interações. Alternativamente, considerar uma transformação não linear para a variável resposta Y também pode ser útil. Portanto, a suposição de linearidade é fundamental e pode ser investigada através do gráfico dos resíduos *vs* valores ajustados (García-Portugués,2021).

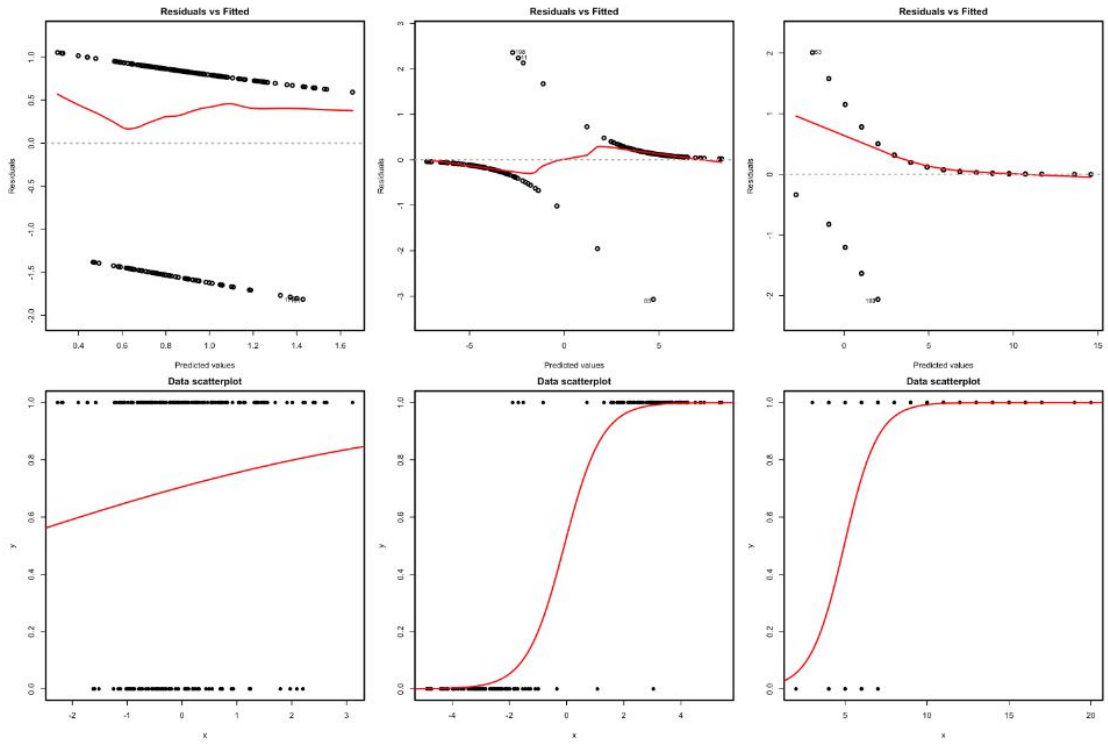

Figura 2: Investigando o pressuposto de linearidade. Fonte: Extraído de García-Portugués

2.5.2 Distribuição da Variável Resposta

QQ-plot dos resíduos *deviance* (primeira linha) para o conjunto de dados (segunda linha) respeitando a suposição de linearidade na regressão logística são apresentados na Figura 3. A normalidade assintótica dos resíduos deviance permite avaliar o quão satisfeita é a suposição da distribuição da variável resposta. A convergência assintótica e a validade efetiva desses resíduos depende muito de vários aspectos: distribuição da variável resposta, tamanho da amostra e distribuição das variáveis preditoras (García-Portugués,2021).

O gráfico QQ-plot permite verificar se os resíduos padronizados seguem uma distribuição $N(0,1)$. Nesse caso, espera-se que os pontos se alinhem a reta diagonal. É comum haver desvios nos extremos diferentes do centro, mesmo em normalidade, embora esses desvios sejam mais evidentes se os dados não forem normais. Infelizmente, também é possível ter desvios severos da normalidade, mesmo se o modelo estiver perfeitamente correto, conforme Figura 3. A razão é que os resíduos deviance não são normais, o que ocorre frequentemente na regressão logística (García-Portugués,2021).

Quando o pressuposto não é atendido e a distribuição da variável resposta não está bem ajustada, corrigir esse problema não é tão trivial e deve-se considerar o uso de modelos mais flexíveis. Uma alternativa é aplicar transformações adequadas na variável resposta Y, como a transformação de Box e Cox (1964) por exemplo, e remodelar a partir da variável transformada.

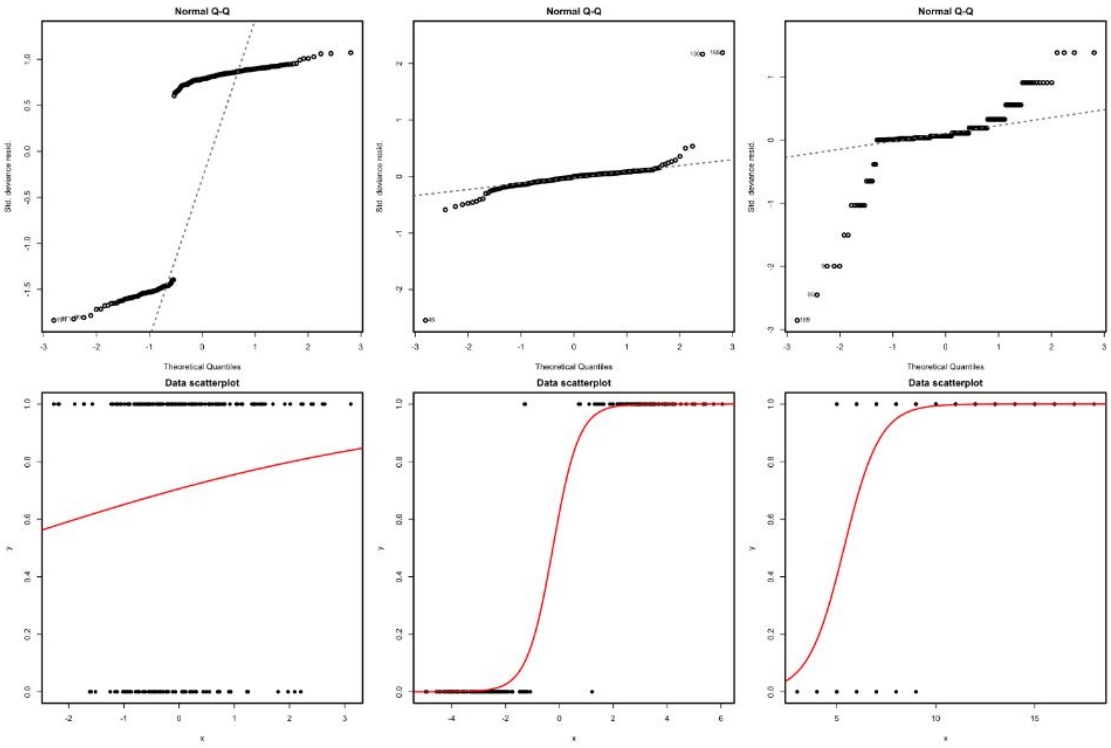

Figura 3: Investigando a Distribuição da Variável Resposta. Fonte: Extraído de García-Portugués

2.5.3 Independência

Os gráfico de dispersão dos resíduos (primeira linha) para o conjunto de dados (segunda linha) respeitando a suposição de independência na regressão logística são apresentados na Figura 4. Também considerado um pressuposto fundamental para a regressão logística, essa suposição é atendida quando, analisando a Figura 4, não são observados resíduos alternados, ou seja, devem ter comportamento aleatório e constante sem divisões ou falhas. Se houver dependência nos dados, pouco pode ser feito, uma vez que os dados já foram coletados. Uma alternativa para corrigir esse problema na regressão linear é aplicar uma diferenciação na variável resposta, que pode ocasionar observações independentes. A presença de autocorrelação nos resíduos pode ser examinada utilizando-se um gráfico de disperção dos resíduos.

2.5.4 Multicolinearidade

De acordo com García Portugués (2021), ainda que os preditores não tenham efeito linear direto sobre a variável resposta, estes são combinados linearmente. Portanto, nos modelos lineares generalizados, a multicolinearidade também pode estar presente. Se dois ou mais preditores estiverem altamente correlacionados entre si, o ajuste do modelo ficará comprometido, uma vez que o reconhecimento do efeito linear individual de cada

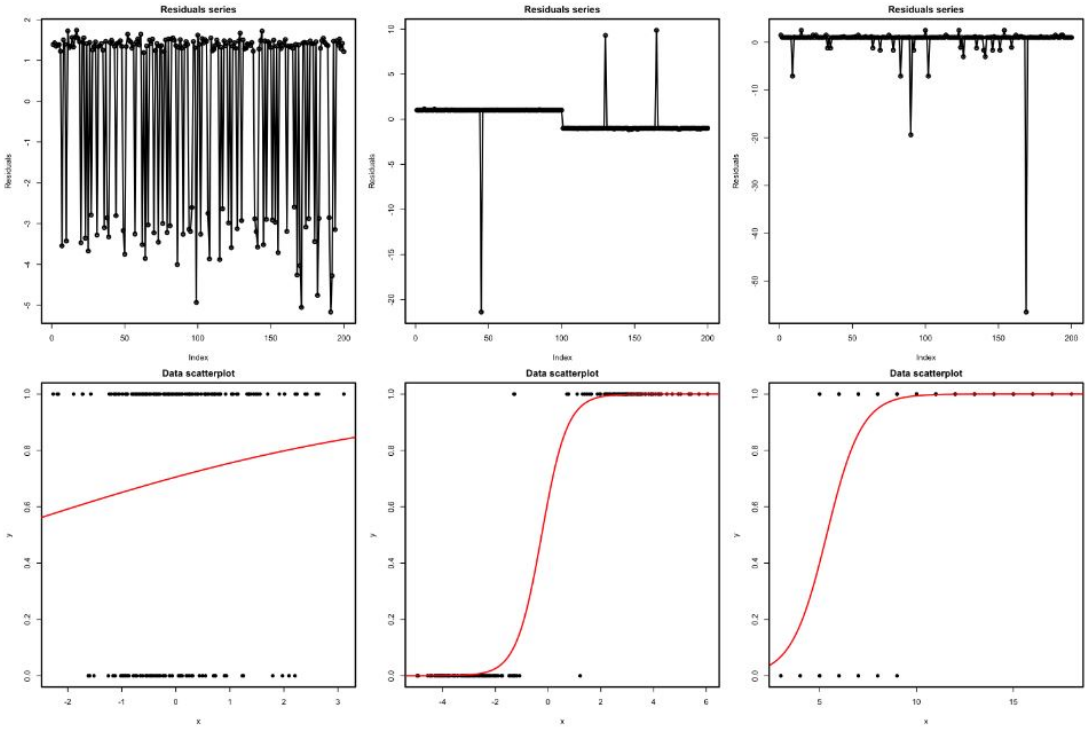

Figura 4: Investigando o pressuposto de independência da variável resposta. Fonte: Extraído de García-Portugués

preditor é afetado.

Uma maneira útil de detectar a multicolinearidade é inspecionar o VIF de cada coeficiente. A situação é exatamente a mesma que na regressão linear, uma vez que VIF olha apenas para as relações lineares entre os preditores. Portanto, a regra prática é a mesma:

- **VIF próximo a 1**: ausência de multicolinearidade.
- **VIF maior que 5 ou 10**: quantidade problemática de multicolinearidade.

Sugere-se remover do modelo o preditor com maior VIF.

2.5.5 Razão de Chances e Interpretação do Modelo

No modelo de regressão usual, existe uma relação linear entre o valor médio π_i e as covariáveis X_i, pois a função de ligação é a identidade, permitindo que valores esperados e preditores lineares assumam qualquer valor real. Portanto, com o aumento de uma unidade para covariável X_k, o valor médio π_i é acrescido de β_k. Contudo, para modelos em que essa relação não é linear, a interpretação não é direta.

Nos modelos logísticos, existe uma relação linear entre o $log(\pi_i)$ e as covariáveis X_i, ou seja, $log(\pi_i)$ é acrescido de β_i a cada aumento em uma unidade de X_i ou mudança

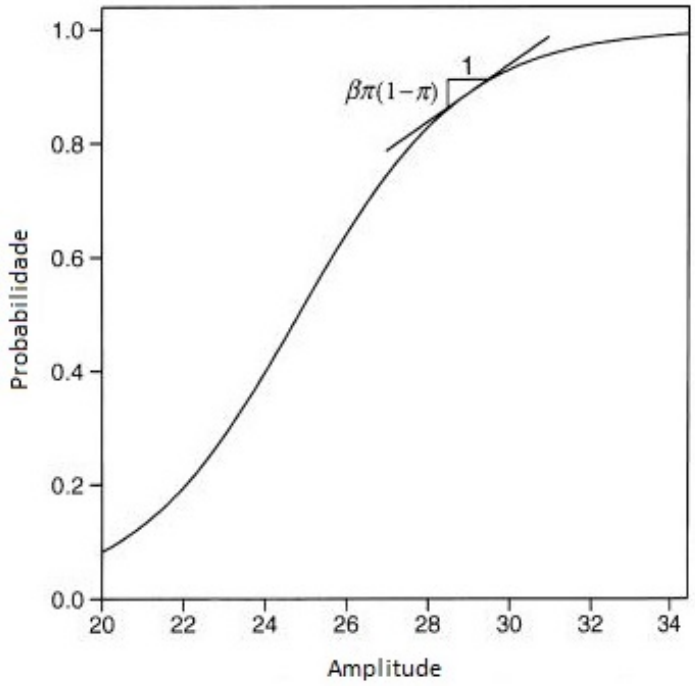

Figura 5: **Aproximação Linear para a Curva da Regressão Logística**

de categoria em caso de variáveis preditoras qualitativas. Todavia, não é trivial pensar em uma escala logarítmica. Para resolver esse problema, há uma outra maneira de interpretar o modelo. Exponenciando ambos os lados da equação de regressão logística (equação 3.1.6) e considerando uma única covariável X_1, obtemos a seguinte relação

$$OR = \frac{\pi_i(X)}{1-\pi_i(X)} = e^{\beta_0+\beta_1 X} = e^{\beta_0}\ e^{(\beta_1)X}$$

que se refere à razão de chances e à probabilidade em si, numa interpretação mais simples. Dessa forma, a probabilidade é multiplicada por e^{β_1} para cada aumento de 1 unidade em X_i ou mudança de categoria em relação à categoria de referência. Para melhor compreensão do que foi abordado, pode-se observar o gráfico apresentado na Figura 5 em que é possível visualizar um formato de S (sigmóide) para a probabilidade π do modelo.

Dado um comportamento curvilíneo no gráfico (Figura 5), ao invés de linear, a taxa de variação em $\pi(x)$ a cada aumento de 1 unidade em X depende do valor da covariável. Uma linha reta desenhada tangente à curva em um determinado valor de X, como mostra a Figura 5, descreve a taxa de mudança nesse ponto. Para o parâmetro de regressão logística β, essa linha tem inclinação igual a $\beta\pi(x)[1-\pi(x)]$. (Agresti,2019)

3 Banco de Dados

O objetivo deste trabalho, conforme introduzido anteriormente, consiste em avaliar e compreender se as características socioeconômicas das mulheres no país têm efeito na ocorrência de violência doméstica e familiar, ou seja, se tornam as mulheres mais suscetíveis às agressões. Para isso, serão analisados os dados da Pesquisa de Violência Doméstica e Familiar, realizada pelo Datasenado em 2019.

A metodologia de coleta dos dados consiste em selecionar números de telefone aleatoriamente, a partir do cadastro telefônico disponibilizado pelo Anatel. Esse cadastro contém todos os números habilitados do país. A seleção foi alocada para cada UF e realizaram as ligações até atingirem 2400 entrevistas, desde que a entrevistada fosse mulher e autorizasse a participação na pesquisa. As respondentes foram selecionadas por meio de técnicas de amostragem aleatória estratificada em dois fatores (*two way sample*), sendo os estratos a Unidade da Federação e o tipo de acesso a telefonia (fixo ou móvel), respectivamente.

O questionário da Pesquisa de Violência Doméstica e Familiar foi elaborado com 33 questões, das quais 7 coletam informações a respeito do perfil das mulheres entrevistadas. A população alvo da pesquisa é reservada para pessoas do sexo feminino, residentes no país, com idade superior a 15 anos e que possuam acesso à telefones móveis ou fixos (Datasenado, 2020).

Baseada na metodologia *rake*[1], um procedimento interativo de ajuste e criação de pesos que visa reduzir o viés da amostra coletada, considerando a distribuição estimada da população feminina do Brasil por Grande Região, idade, escolaridade, raça ou cor e força de trabalho (ocupada, desocupada ou fora da força de trabalho), aplicou-se uma ponderação ao cômputo dos resultados para as pesquisas amostrais. Para esse cálculo, foi utilizada a Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua) do 2º trimestre de 2019.

A base de dados não somente é composta por mulheres que foram vítimas de violência e/ou presenciaram algum ato de violência doméstica e familiar, como também aquelas que nunca sofreram nenhum tipo de agressão física, verbal ou psicológica provocadas por um homem. Logo, a variável de interesse ou variável resposta assume dois valores, $Y_i = 0$ e $Y_i = 1$, denominados "fracasso"e "sucesso", respectivamente. Neste caso, o evento de interesse é a ocorrência da violência doméstica, ou seja, mulheres que se declararam vítimas de agressões.

1

[1] https://www.teses.usp.br/teses/disponiveis/6/6132/tde-02042013-104046/publico/NeuberSegri.pdf

4 Resultados

4.1 Análise Descritiva

Em uma relação íntima, a violência de gênero refere-se a qualquer comportamento que cause dano físico, psicológico, moral, patrimonial ou sexual àqueles que fazem parte dessa relação. Em 2005, ano anterior à promulgação da Lei Maria da Penha, o Instituto DataSenado, em parceria com o Observatório da Mulher contra a Violência, iniciou um levantamento de dados a respeito da Violência Doméstica e Familiar. Com periodicidade de 2 anos, esses dados representam a opinião e vivência da população feminina brasileira com acesso a telefone fixo e celular. Em 2019, foi realizada a 8º edição da Pesquisa, em que 27% das 2400 mulheres entrevistadas já haviam sofrido algum tipo de violência doméstica ou familiar provocada por um homem. Contudo, pôde-se observar na última edição que um percentual das entrevistadas desconhecia algumas das situações de violência elencadas. Isso mostra que, apesar das variadas campanhas de conscientização acerca da violência doméstica, uma parcela da população feminina ainda carece de informação. De acordo com o Relatório de Pesquisa do DataSenado, publicado em dezembro de 2019:

> Os resultados mostraram que, além de 27% das mulheres que reconheceram inicialmente terem sido vítimas de violência em algum momento da vida, outras **9%** relataram já ter vivenciado, no último ano, pelo menos uma das doze situações elencadas provocadas por parceiro ou ex-parceiro. Assim, pode-se afirmar que pelo menos 36% das brasileiras já sofreram violência doméstica. Conclui-se que atos como humilhar a mulher em público, tomar seu salário ou outras situações nem sempre são reconhecidos por elas como violência.

Em pesquisas sobre a violência de gênero, é comumente questionada a ocorrência de abuso baseada em situações específicas de agressão, listadas para facilitar a compreensão das respondentes. Essa metodologia foi inserida ao questionário do DataSenado em 2018, fundamentada em pesquisas anteriores realizadas por outros países. Questões específicas de comportamentos agressivos como ser estapeada, ameaçada, financeiramente lesada, moralmente e sexualmente exposta, quando utilizadas em pesquisas, produzem maiores índices de resposta positiva em relação a perguntas como "Você já sofreu algum tipo de violência doméstica ou familiar provocada por um homem?"(Organização Mundial da Saúde, 2002), conforme pôde ser observado no relatório de pesquisa do DataSenado em 2019. Vale ressaltar que, para a primeira situação, o questionário da pesquisa refere-se aos últimos doze meses, enquanto para o segundo questionamento, a referência é o tempo de vida da mulher entrevistada.

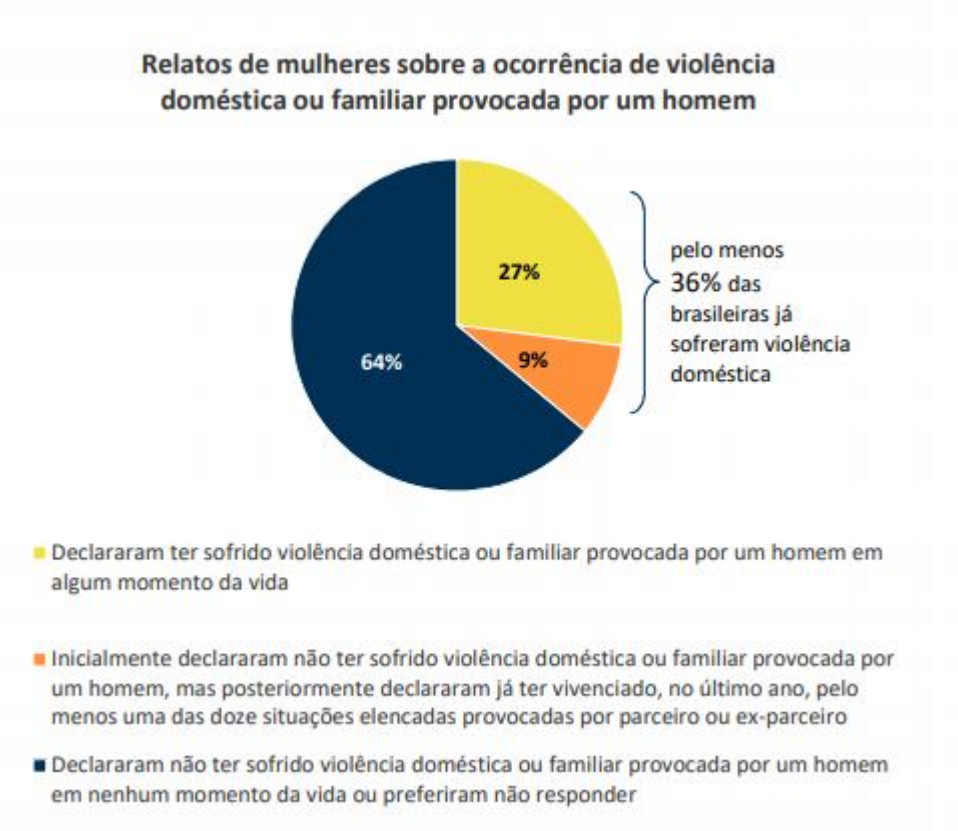

Figura 6: Relatos de Mulheres sobre a ocorrência de violência doméstica ou familiar provocada por um homem. Fonte: DataSenado(2019)

Para mensurar essas informações, uma nova variável foi acrescida ao banco de dados, composta pela junção das informações referentes a ocorrência ou não de violência doméstica, em algum momento da vida, e a ocorrência ou não de doze situações elencadas de violência doméstica nos últimos doze meses previamente classificadas em moral, física, psicológica, patrimonial ou sexual. Somando essas informações, totalizou-se oito mulheres que não souberam ou preferiram não responder às perguntas, representadas nos gráficos como **NR**, uma parcela muito pequena do total de mulheres entrevistadas.

Portanto, a variável resposta do estudo refere-se a ocorrência de violência doméstica e familiar provocada por um homem em algum momento da vida.

Evidenciando as características das mulheres respondentes à pesquisa, na Figura 7, o percentual de mulheres declaradas vítimas ou que reconheceram ter sofrido quaisquer das doze situações de violência doméstica elencadas é superior para as categorias de raça preta e parda. Por outro lado, prevalecem mulheres que declararam não terem sofrido nenhum tipo de agressão para as categorias de raça branca e amarela. O percentual de não resposta é baixo.

Observando a Figura 8, destacam-se mulheres que não sofreram ou não percebem a ocorrência de agressões entre aquelas com 60 anos ou mais. Considerando às mulheres

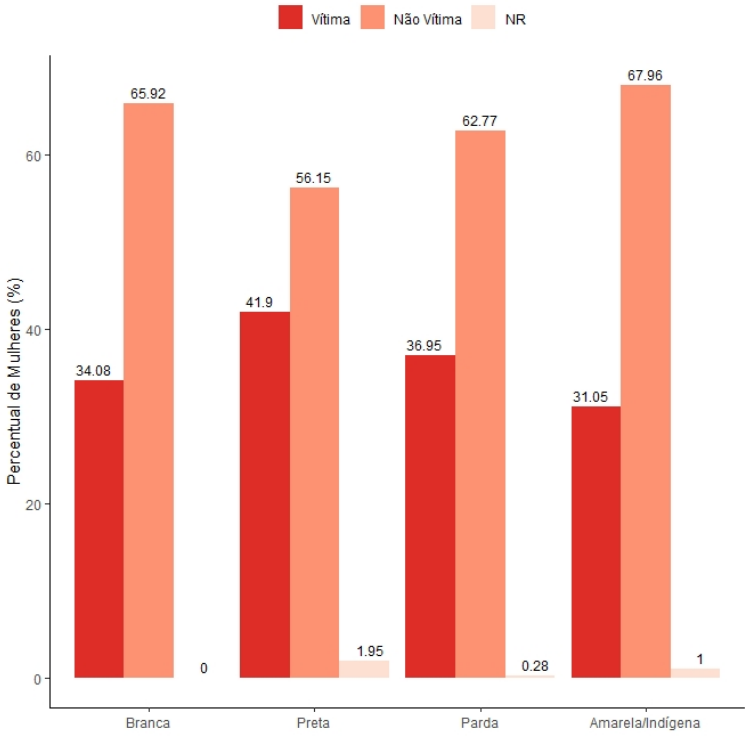

Figura 7: Raça/Cor das Mulheres Entrevistadas

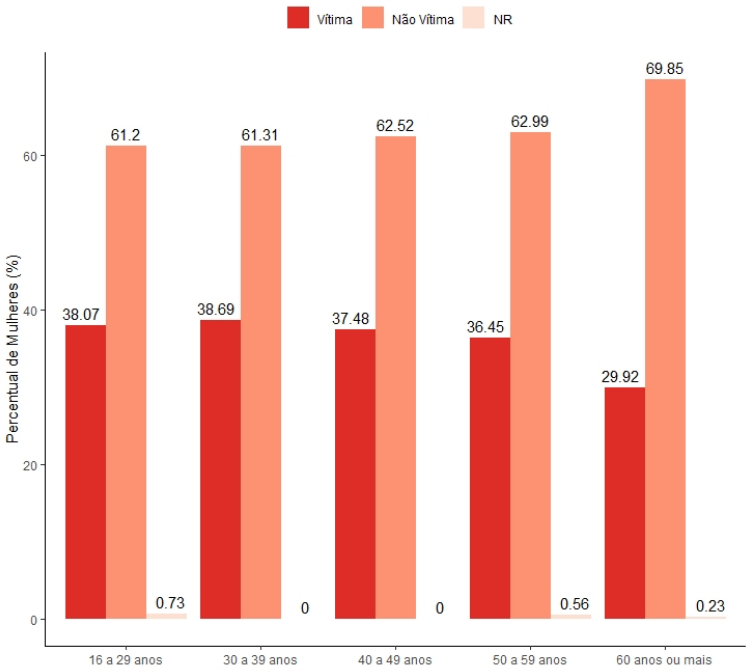

Figura 8: Faixa Etária das Mulheres Entrevistadas

com idade entre 16 e 59 anos, os respectivos percentuais de mulheres declaradas vítimas para cada categoria são semelhantes, oscilando entre 36% e 39%. Novamente, o percentual de não resposta é aproximadamente nulo.

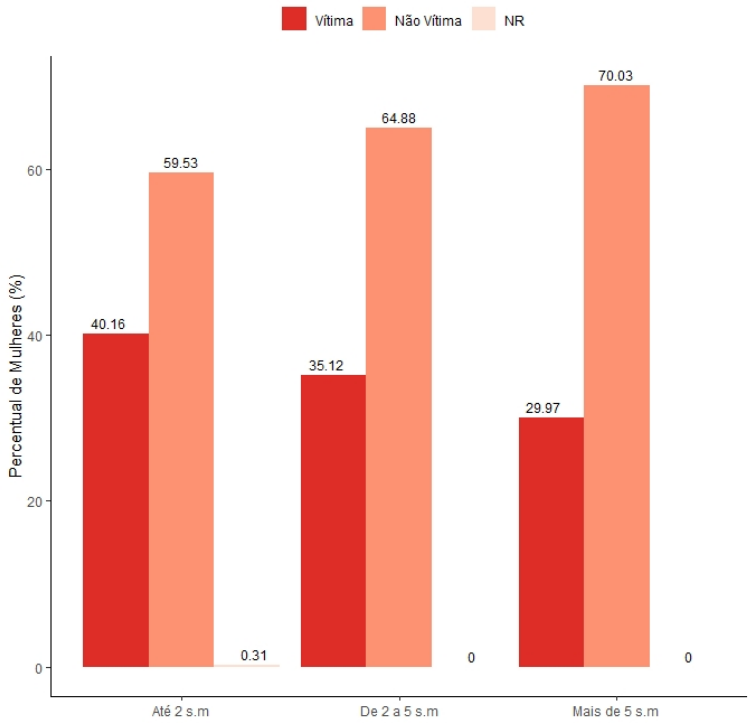

Figura 9: Renda das Mulheres Entrevistadas

Há um comportamento interessante na Figura 9, em que pode-se perceber um aumento do percentual de mulheres que não foram vítimas ou não perceberam a ocorrência da violência doméstica proporcional ao aumento da renda salarial. Em contrapartida, a proporção de vítimas reduz, de forma gradativa em relação ao aumento salarial. Portanto, de acordo com a Figura 9, a medida que a renda aumenta a proporção de mulheres vítimas de violência reduz e a de não vítimas aumenta. A proporção de não resposta é próxima ou igual a zero.

Conforme resultados observados na Figura 10, destaca-se o comportamento da categoria de mulheres fora da força de trabalho, que estão inaptas a exercer uma profissão, com percentual mais elevado dentre as mulheres declaradas não vítimas e inferior para as declaradas vítimas de violência doméstica. A proporção de mulheres vítimas é mais elevado para a categoria desocupada, ou seja, que estão aptas a trabalhar, mas se encontram desempregadas no momento da pesquisa.

Analisando a variável escolaridade na Figura 11, para o nível médio, completo ou incompleto, o percentual de mulheres declaradas vítimas é superior. Quando observados os níveis fundamental incompleto e superior ou mais, a resposta mais frequente é de não ocorrência da violência doméstica.

Observando a região de origem da respondente, não é possível notar grandes diferenças de opinião em relação a ocorrência de violência nas diferentes regiões do Brasil,

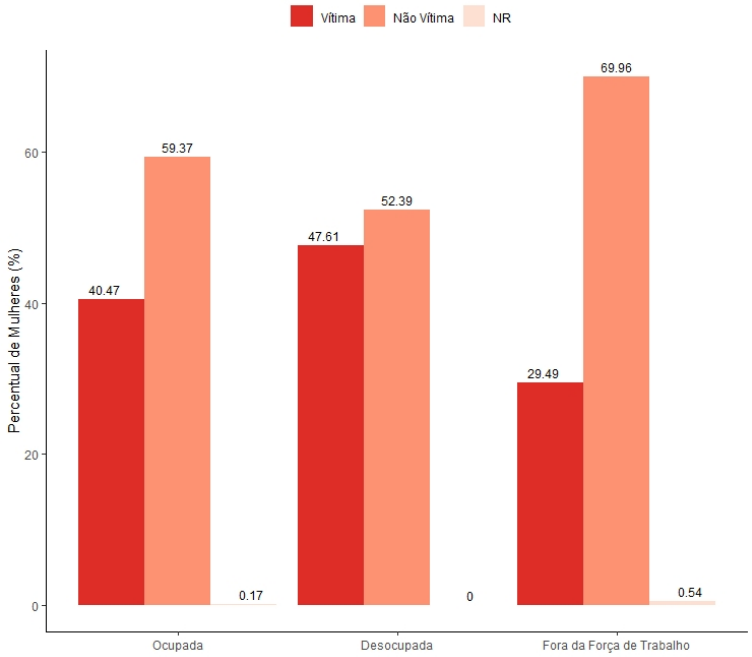

Figura 10: Força de Trabalho das Mulheres Entrevistadas

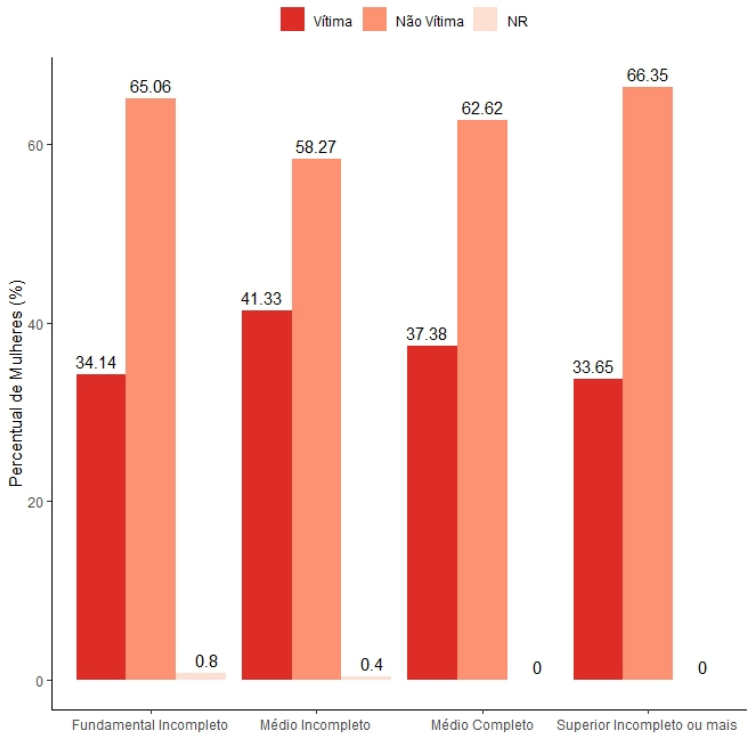

Figura 11: Escolaridade das Mulheres Entrevistadas

conforme apresentado na Figura 12. A amostra da pesquisa é representativa, e um dos estratos é composto pelo estado de origem da respondente. Logo, o efeito da região pode ter sido anulado.

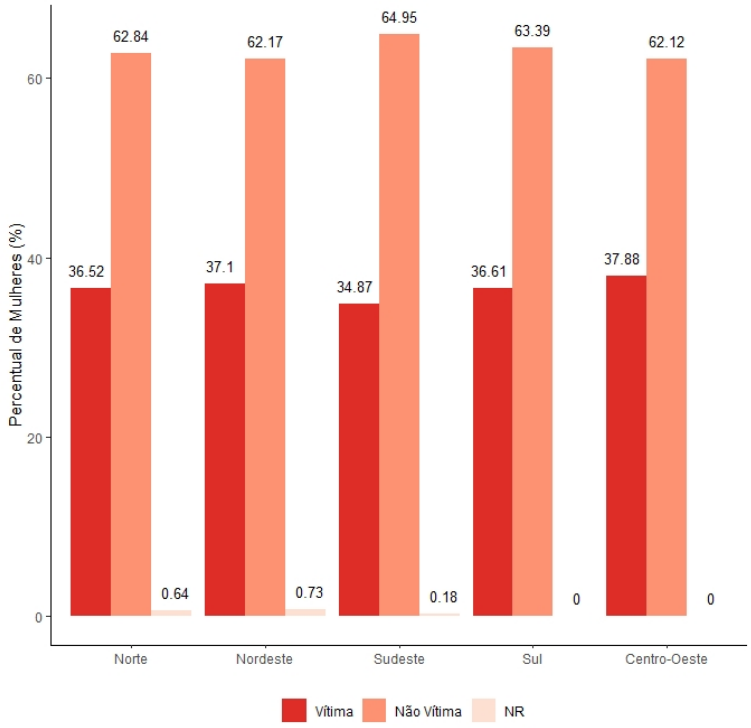

Figura 12: Região das Mulheres Entrevistadas

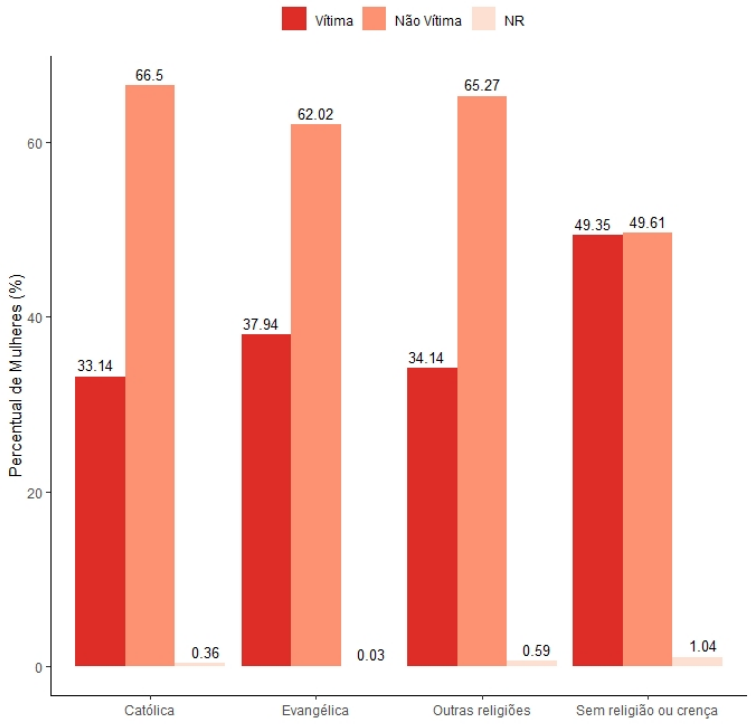

Figura 13: Religião das Mulheres Entrevistadas

Em relação a ocorrência de violência doméstica, a Figura 13 mostra que mulheres sem religião ou crença têm índices mais elevados quando comparadas às mulheres que

possuem religião ou crença. Além disso, a percentual de resposta sim e não para a categoria sem religião é aproximadamente o mesmo. Considerando mulheres que possuem religião ou crença, a proporção de resposta para a não ocorrência de violência é maior, tendo comportamentos semelhantes entres as religiões, seja evangélica, católica ou outras.

4.2 Seleção de Variáveis e Modelos Candidatos

Definindo como evento de interesse a ocorrência de Violência Doméstica e Familiar, temos as respostas "Sim"e "Não", assumindo valores 1 e 0 respectivamente.

Conforme exposto anteriormente, trata-se de uma amostra aleatória estratificada em dois fatores, com ponderação baseada na metodologia *rake*. O peso amostral foi calculado com auxílio da PNAD e considerou-se a distribuição estimada da população feminina do Brasil por grande região, idade, escolaridade, raça ou cor e força de trabalho (ocupada, desocupada ou fora da força de trabalho).

A variável composta pelos respectivos pesos amostrais foi utilizada em todos os cálculos. Em um primeiro momento, deseja-se conhecer quais variáveis de perfil possuem associação com a variável de interesse, ocorrência ou não de violência doméstica. Para isso, aplicou-se o Teste de Qui-Quadrado, com resultados conforme a Tabela 3.

Tabela 3: Teste de Qui-Quadrado entre a variável resposta e as possíveis variáveis explicativas

Variável	**Estatística**	**G.L.**	**P-valor**
Idade	9,28	4	0.0544
Raça/Cor	4,75	3	0.1912
Escolaridade	9,34	3	0.0251
Força de Trabalho	40,16	2	$1.9e^{-9}$
Religião	20,99	3	0.0001
Renda	12,38	2	0.002
Região	1,78	4	0.775

A fim de investigar com maior profundidade e detalhamento a relação entre as variáveis, quando os valores amostrais das variáveis de interesse da pesquisa podem ser considerados resultados de um conjunto de vetores aleatórios, sugere-se a aplicação de técnicas de modelagem estatística. Modelos podem ser especificados, ajustados, testados e reformulados usando procedimentos estatísticos padrões como os apresentados, por exemplo, em (Bickel and Doksum 1977) e (Garthwaite, Jollife, and Jones 1995).

De acordo com a seção 3.5.4, quaisquer variáveis com p-valor inferior a 0.25 são candidatas ao modelo juntamente com aquelas consideradas importantes pelo pesquisador. Pela Tabela 1, inicialmente descarta-se o uso da variável **Região**, pois há evidências de independência entre a região de origem da respondente e sua percepção em relação à

ocorrência de violência doméstica.

Logo, serão mantidas no modelo completo as variáveis: Idade; Raça/Cor; Escolaridade; Força de Trabalho; Religião e Renda. Contudo, variáveis como escolaridade, renda, força de trabalho e idade sugerem uma associação natural, que deve ser investigada. A Tabela 4 apresenta os resultados dos testes de independência entre as variáveis que serão utilizadas no modelo.

Tabela 4: Teste de Qui-Quadrado e Coeficiente de Contingência Modificado das Variáveis Explicativas 2 a 2.

Variáveis	**Estatística**	**P-valor**	**Coef. Conting. Mod.**
Idade - Raça/Cor	29,13	0,0038	0,14
Idade - Escolaridade	286,88	< 0.0001	$\mathbf{0,41}$
Idade - Força de Trabalho	434,46	< 0.0001	$\mathbf{0,52}$
Idade - Religião	67,85	< 0.0001	0,21
Idade - Renda	53,79	< 0.0001	0,20
Idade - Região	22,65	0,1234	0,12
Raça/Cor - Escolaridade	33,85	< 0.0001	0,15
Raça/Cor - Força de Trabalho	9,19	0,1631	0,08
Raça/Cor - Religião	18,23	0,0326	0,11
Raça/Cor - Renda	79,63	< 0.0001	0,24
Raça/Cor - Região	305,34	< 0.0001	0,37
Escolaridade - Força de Trabalho	255,65	< 0.0001	$\mathbf{0,42}$
Escolaridade - Religião	87,00	< 0.0001	0,24
Escolaridade - Renda	594,37	< 0.0001	$\mathbf{0,59}$
Escolaridade - Região	22,36	0,0337	0,11
Força de Trabalho - Religião	29,86	< 0.0001	0,12
Força de Trabalho - Renda	212,27	< 0.0001	0,38
Força de Trabalho - Região	21,97	0,005	0,11
Religião - Renda	31,61	< 0.0001	0,15
Religião - Região	40,26	< 0.0001	0,14
Renda - Região	92,10	< 0.0001	0,21

Analisando a Tabela 4, há evidências de associação entre quase todas as variáveis explicativas, exceto Idade - Região e Raça/Cor - Força de Trabalho, para as quais o teste não rejeitou a hipótese de independência. Contudo, entre as variáveis Idade - Escolaridade, Idade - Força de Trabalho, Escolaridade - Força de Trabalho e Escolaridade - Renda, o coeficiente de contingência modificado indica associação moderada. Optou-se, então, por construir 3 modelos iniciais, separando essas variáveis a fim de evitar possíveis efeitos de multicolinearidade.

Em pesquisas amostrais complexas, há situações em que a variância dos dados é superior a média, pois os pesos atribuídos pelo processo amostral ocasionam um sobredispersão nos dados. Nesse contexto, sugere-se o uso da distribuição quasi-binomial, onde

é acrescido o parâmetro de dispersão. Apesar de a distribuição não ser especificada, a mesma estrutura do modelo para a função de ligação e preditor é mantida (ZUUR, 2009).

Wedderburn (1974) definiu as funções de quasi-verossimilhança para exibir propriedades semelhantes ao log da máxima verossimilhança, mas sem possuir correspondência com nenhuma distribuição de probabilidade – os modelos são caracterizados apenas por média e variância. Essa alternativa permite que o modelo disponha de parâmetros em um estado natural e interpretável e proporciona diagnósticos padrões sem a perda de eficiência no ajuste de algoritmos (HOEF; BOVENG, 2007).

A função de quasi-verossimilhança contém um fator multiplicativo, isto é, o parâmetro de sobredispersão (do inglês overdispersion parameter ou scale parameter), o qual é estimado a partir dos dados. Porém, dificulta a aplicação de métodos usuais de diagnóstico do ajuste de modelos. Alguns gráficos de resíduos e outros procedimentos da inferência clássica (tais como o teste estatístico da Razão de Verossimilhança) não podem ser utilizados. Por esse motivo, durante a seleção de variáveis dos modelos candidatos, utiliza-se o Teste de Wald e AIC como critério de escolha (IBGE, 2018).

A partir do modelo inicial, é realizado o teste de Wald para modelos encaixados, retirando-se uma variável por vez. Essa técnica é semelhante à ANOVA (que normalmente executa testes de razão de verossimilhança para modelos de regressão usuais), mas com algumas diferenças. Se apenas um modelo ajustado for especificado, este é comparado ao modelo simples (com apenas o intercepto). O teste pode ser realizado utilizando-se a estatística F de amostra finita ou a estatística Qui-quadrado assintótica , $F = Chisq/k$ se k for a diferença em graus de liberdade (R Documentation).

Os modelos de regressão logística foram ajustados com auxílio do *software* R (versão 1.4.11) através da função *svyglm()* do pacote *survey*, para pesquisas amostrais complexas . A fórmula do modelo e as características do levantamento amostral são informados através do código *svydesign()*, em que são especificados os pesos, estratos, entre outros elementos importantes para o *design* do modelo. Para trabalhar com os pesos amostrais não inteiros, usa-se *family = quasibinomial(link="logit")*. As Tabelas 5,6 e 7 apresentam o P-valor dos testes de Wald para os modelos encaixados.

Em todos os modelos iniciais, utilizou-se o método de seleção **Backward**, em que as variáveis são retiradas, uma a uma, testando se seus respectivos coeficientes de regressão são significativos, ou seja, $\beta_i \neq 0$. O Modelo 1 é composto pelas variáveis idade, raça/cor, religião e renda. Idade e Raça/Cor não foram consideradas significativas, com p-valor $0,3426$ e $0,6169$, respectivamente. O modelo contendo apenas as informações de renda e religião (modelo 1.21) teve menor AIC. Porém, a diferença entre os modelos 1.2 e 1.21 foi pequena. Desse modo, optou-se pelo modelo com a variável Idade, considerada uma informação importante. Logo, com AIC resultante de 2621.96, o modelo composto

Tabela 5: Passo 1 - Seleção de Variáveis do Modelo 1

Modelo	Variáveis Explicativas	AIC	Chisq	P-Valor
1	Idade + Raça + Religião + Renda	2627.53		
1.1	Raça + Religião + Renda	2623.27	4,5	0,3426
1.2	**Idade + Religião + Renda**	**2621.96**	**1, 79**	**0, 6169**
1.3	Idade + Raça + Renda	2637.17	14,02	0,0029
1.4	Idade + Raça + Religião	2637.717	11,71	0,0028
1.21	**Religião + Renda**	**2618.178**	**4, 78**	**0, 3102**
1.22	Idade + Renda	2632.023	14,41	0,0024
1.23	Idade + Religião	2633.908	12,9	0,0016
1.211	Renda	2632.71	17,96	0,0004

Tabela 6: Passo 2 - Seleção de Variáveis do Modelo 2

Modelo	Variáveis Explicativas	AIC	Chisq	P-Valor
2	Raça + Escolaridade + Religião	2629.92		
2.1	**Escolaridade + Religião**	**2626.64**	**3, 14**	**0, 37**
2.2	Raça + Religião	2630.82	8,25	0,0411
2.3	Raça + Escolaridade	2641.35	15,22	0,0016
2.11	Escolaridade	2639.11	16,23	0,001
2.12	Religião	2627.6	8,7	0,0335

Tabela 7: Passo 3 - Seleção de Variáveis do Modelo 3

Modelo	Variáveis Explicativas	AIC	Chisq	P-Valor
3	Raça + Força de Trabalho + Religião + Renda	2582.62		
3.1	**Força de Trabalho + Religião + Renda**	**2576.86**	**2, 06**	**0, 56**
3.2	Raça + Religião + Renda	2623.27	30,65	$2{,}21.10^{-7}$
3.3	Raça + Força de Trabalho + Renda	2590.58	12,54	0,0057
3.4	Raça + Força de Trabalho + Religião	2600,39	16,98	0,0002
3.11	Religião + Renda	2618.18	31,08	$1{,}77.10^{-7}$
3.12	Força de Trabalho + Renda	2585.17	12,93	0.0048
3.13	Força de Trabalho + Religião	2596.46	18,02	0.0001

pelas variáveis **Idade, Religião e Renda** será o primeiro modelo candidato.

A Tabela 6 apresenta a seleção de variáveis do modelo 2, composto por **Raça/Cor, Escolaridade e Religião**. No primeiro passo, retirando uma variável, a variável Raça/Cor não rejeitou hipótese nula $H_0 : \beta_i = 0$, ou seja, o coeficiente não é significativo e pode ser retirado do modelo. No segundo passo, todas as variáveis foram significativas e devem permanecer, resultando no modelo composto por escolaridade e religião, com AIC resultante de $2626,64$. Porém, a seleção do modelo 2, analisando-a subjetivamente, não teve resultado tão satisfatório e relevante quanto o modelo 1,. Os valores de AIC são mais elevados.

O terceiro modelo inicial é composto pelas variáveis raça/cor, força de trabalho, religião e renda. Aplicando-se novamente o método **Backward**, o coeficiente da variável

raça/cor não foi considerado significativa para o modelo, com p-valor de 0.56, não rejeita a hipótese nula $H_0 : \beta_i = 0$ e pode ser removida do modelo. No passo 3, os modelos subsequentes rejeitam a hipótese nula para o teste de Wald, indicando que os demais coeficientes $\beta_i \neq 0$ são significativos, ou seja, as variáveis devem permanecer no modelo. Portanto, o segundo modelo candidato, com menor AIC de $2576,86$ entre os modelos construídos durante a etapa de seleção, será composto pelas variáveis **Força de Trabalho, Religião e Renda**.

Assim, temos:

Modelo Candidato 1:

Tabela 8: Modelo Candidato 1, com estimativa dos parâmetros, erro padrão, teste de wald e p-valor

Coeficientes	Estimativa	Erro Padrão	t	P-valor
Intercepto	-0.36131	0.15075	-2.397	0,016636
30 a 39 anos	-0.07097	0.15886	-0.447	0,655136
40 a 49 anos	-0.05975	0.16783	-0.356	0,721870
50 a 59 anos	-0.26509	0.18944	-1.399	0,161865
60 anos ou mais	-0.36205	0.19622	-1.845	0,065175
Evangélica	0.19096	0.13020	1.467	0,142618
Outras religiões	0.12495	0.19636	0.636	0,524632
Sem religião ou crença	0.76274	0.20237	3.769	0,000169
Mais de 2 a 5 salários mínimos	-0.28619	0.12736	-2.247	0,024741
Mais de 5 salários mínimos	-0.52035	0.15149	-3.435	0,000605

Os coeficientes significativos, de acordo com a Tabela 8 e considerando $\alpha = 0.1$, são da categoria **60 anos ou mais** para a Faixa etária das entrevistadas, **Sem religião ou crença**, para a variável religião e a variável Renda. Nesse caso, as características de referência são mulher brasileira de 16 a 29 anos, católica e de baixa renda (até 2 salários mínimos).

Modelo Candidato 2:

Tabela 9: Modelo Candidato 2, com estimativa dos parâmetros, erro padrão, teste de wald e p-valor

Coeficientes	Estimativa	Erro Padrão	t	P-valor
Intercepto	-0.19488	0.12195	-1.598	0,110189
Desocupada	0.15894	0.14456	1.099	0,271697
Fora da Força de trabalho	-0.64515	0.13609	-4.741	$2{,}28.10^{-6}$
Evangélica	0.19427	0.12947	1.500	0,133653
Outras religiões	0.09947	0.20090	0.495	0,620574
Sem religião ou crença	0.71997	0.20329	3.542	0,000407
Mais de 2 a 5 salários mínimos	-0.40615	0.13189	-3.079	0,002102
Mais de 5 salários mínimos	-0.62564	0.15935	-3.926	$8{,}93.10^{-5}$

O modelo 2 (Tabela 9) tem como referência mulheres ocupadas (exercendo ati-

vidade remunerada), católicas e de baixa renda. Considerando esse perfil, os coeficientes significativos são da categoria **Fora da Força de Trabalho** para a variável Força de Trabalho, **Sem religião ou crença** para a variável Religião e, também, a variável Renda. Os coeficientes das demais categorias não foram significativos em relação ao perfil de referência, considerando um nível de significância de 10%.

Após o ajuste dos modelos, a ênfase muda do cálculo e avaliação da significância dos coeficientes estimados à interpretação dos seus valores. Entretanto, uma avaliação da adequação dos modelos ajustados deve preceder qualquer tentativa de interpretá-lo (HOSMER, 2013)

4.3 Diagnóstico da Qualidade de Ajuste dos Modelos

Em planos amostrais complexos, conforme exposto anteriormente, os procedimentos de diagnósticos usuais sofrem ajustes ou não são indicados, em razão do usa da pseudo-verossimilhança (quasi-verossimilhança), para mais informações veja García-Portugués, E. (2021). Isso ocorre por efeito do uso de conglomeração, estratificação e/ou pesos desiguais e não inteiros no processo amostral. Portanto, optou-se por utilizar métodos gráficos de diagnóstico e o Teste de Wald para o Ajuste Global.

A seguir estão os testes de qualidade de ajuste global dos modelos candidatos:

Tabela 10: Teste de Wald para Ajuste Global do Modelo

Modelo	Variáveis Explicativas	AIC	Estatística	P-valor
1	Idade + Religião + Renda	2621.9	32.8	0.0001
2	Força de Trabalho + Religião + Renda	¨2576.8	62.1	$5{,}81.10^{-11}$

Analogamente à seção 3.2.1, as hipóteses do teste de Wald são $H_0 : \beta_1 = ... = \beta_p = 0$ e $H_1 : \beta_i \neq 0$. Os P-valores dos testes foram inferiores ao nível de significância $\alpha = 0.05$ (¡0.001). Portanto, ambos os testes possuem coeficientes com efeito significativo, ou seja, há algum $\beta_i \neq 0$. Posteriormente, serão avaliados os efeitos individuais das covariáveis. Para investigar a presença de Multicolinaeridade nos modelos, calculou-se o VIF de cada variável preditiva conforme a seguir:

Tabela 11: VIF: Variance Inflation Factor

Modelo	Variável	GVIF	Df	GVIF$^{(}1/(2 * Df))$
1	Idade	1.07	4	1.01
	Religião	1.06	3	1.01
	Renda	1.04	2	1.01
2	Força de Trabalho	1.17	2	1.04
	Religião	1.03	3	1.00
	Renda	1.19	2	1.04

A decisão tomada anteriormente, de separar as covariáveis com associações moderadas, contribuiu para amenizar ou até mesmo anular um possível efeito de multicolinearidade. Como pode ser observado na Tabela 11, os valores de VIF são aproximadamente **1**, indicando ausência de multicolinearidade nos modelos candidatos.

4.3.1 Modelo 1

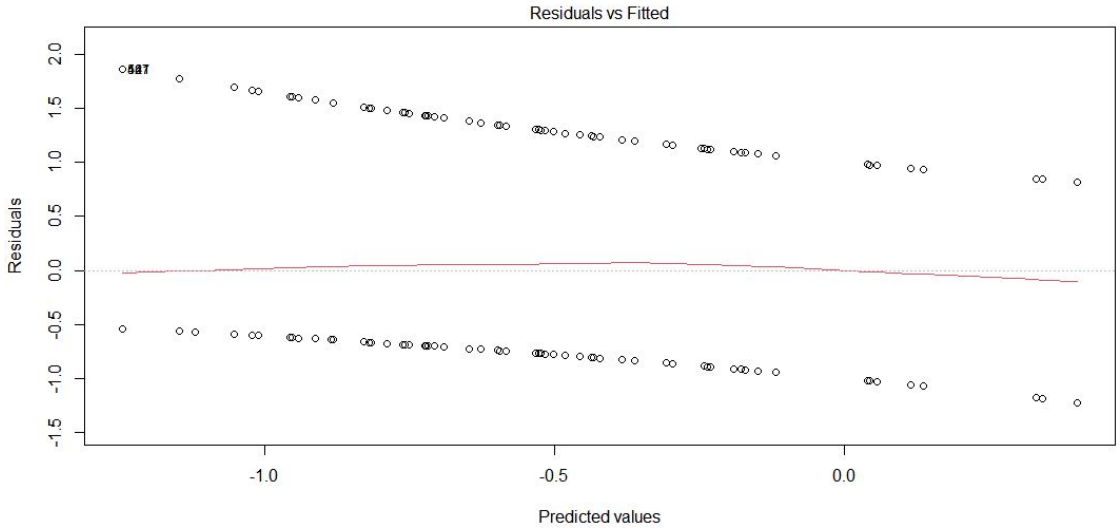

Figura 14: Gráfico de Resíduos *vs* Valores Ajustados

A seção 3.5 apresenta os padrões de cada método gráfico utilizado para investigar as suposições do modelo. Na Figura 14, gráfico de resíduo *vs* valores ajustados, a linha vermelha central indica um comportamento linear, respeitando o pressuposto de linearidade estabelecido.

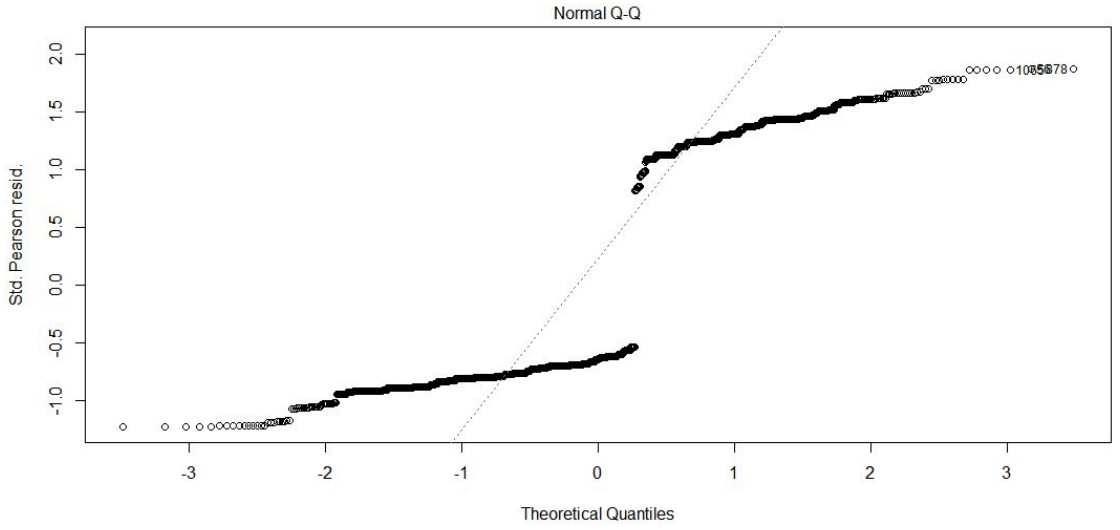

Figura 15: QQ-plot dos Resíduos

Em modelos logísticos é comum que os resíduos *deviance* não sejam significativamente normais, ainda que a distribuição escolhida para a variável resposta esteja adequada e o modelo bem ajustado. Isso afeta a análise visual, uma vez que os pontos nem sempre se alinham à reta diagonal do gráfico QQ-plot. A Figura 15 está de acordo com o comportamente previamente estabelecido na seção 3.5.6, em que se observa uma curvatura dos pontos no extremo centro do gráfico, indicando que o pressuposto da distribuição escolhida ser adequada foi atendido.

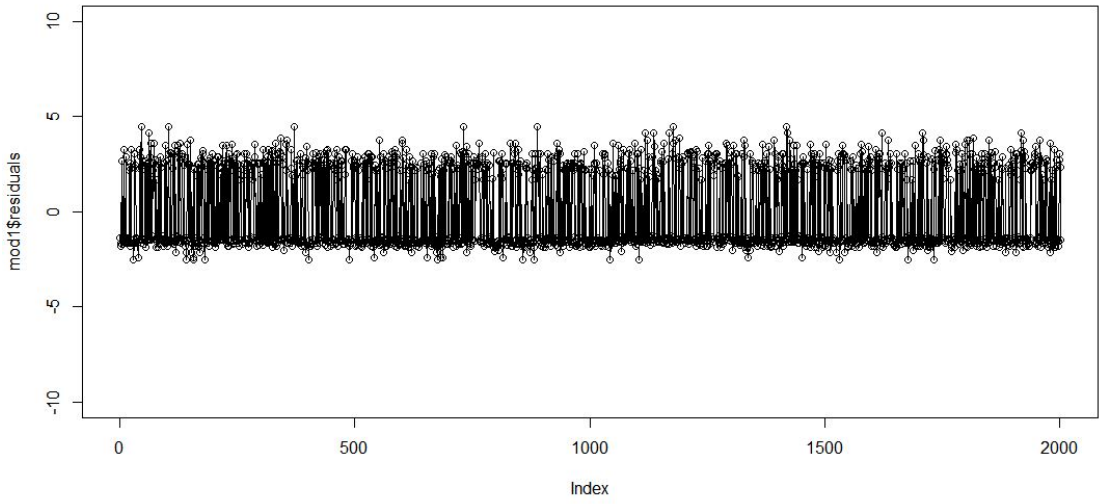

Figura 16: Gráfico de Dispersão dos Resíduos

Investiga-se o pressuposto de independência através da Figura 16. Os resíduos estão dispersos de forma homogênea e constante em torno do zero. Não são observados pontos muito discrepantes nem comportamentos alternados crescentes e decrescentes. Também não há aumento ou diminuição na amplitude. Logo, o pressuposto de independência também foi atendido.

4.3.2 Modelo 2

Os pressupostos de linearidade e distribuição adequada da variável resposta para o modelo 2, Figuras 18 e 17, foram atendidos. A Figura 19 também apresenta um comportamento constante e aleatório dos resíduos, apesar de os pontos superiores estarem um pouco mais distantes quando comparado ao modelo 1.

Portanto, ambos os modelos estão adequados e bem ajustados. A interpretação dos efeitos das covariáveis será descrita a seguir.

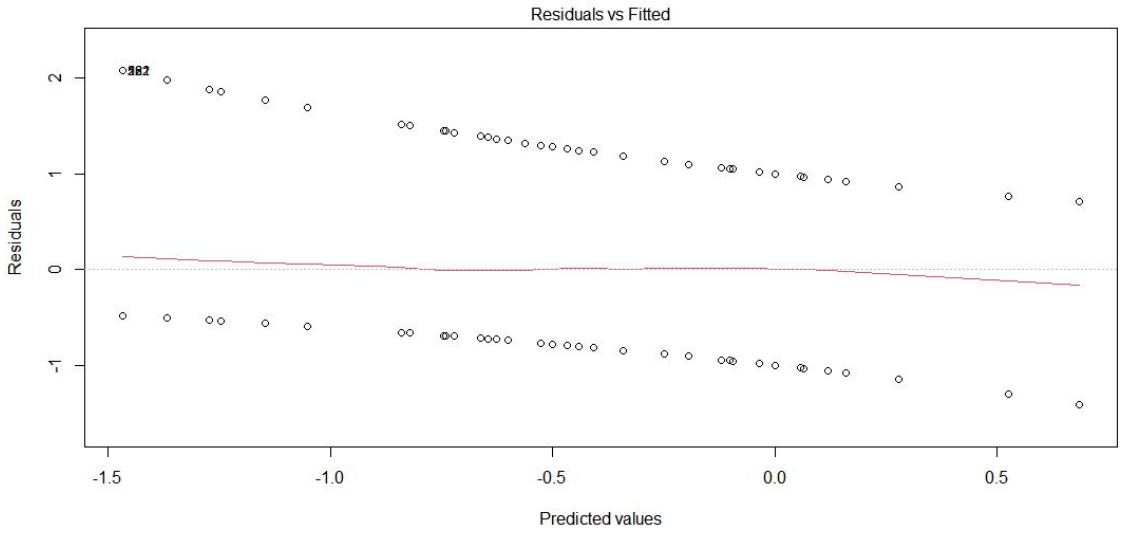

Figura 17: Gráfico de Resíduos *vs* Valores Ajustados

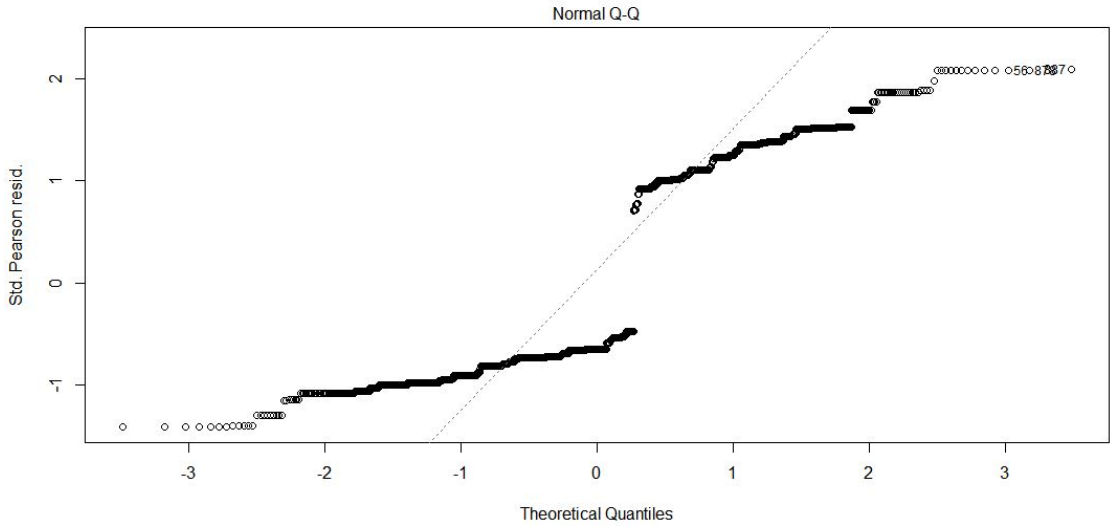

Figura 18: QQ-plot dos Resíduos

4.3.3 Razão de Chances e Interpretação do Modelo

Neste estudo, considerando que as variáveis independentes são nominais dicotômicas (medidas em dois níveis sendo 1, possuir determinada característica e 0, caso contrário), a medida de intepretação mais adequada é a **razão de chances**, definida como a razão entre as chances de $x = 1$ e as chances de $x = 0$. Para um modelo de regressão logística, o coeficiente de inclinação β_i é igual à diferença entre o valor da variável dependente em $x + 1$ ou categoria de referência e o valor da variável dependente em x ou mudança de categoria analisada para qualquer valor de x, ou seja, a razão de chances aumenta em β_i a cada unidade acrescida em x ou mudança de categoria em caso de variáveis preditoras

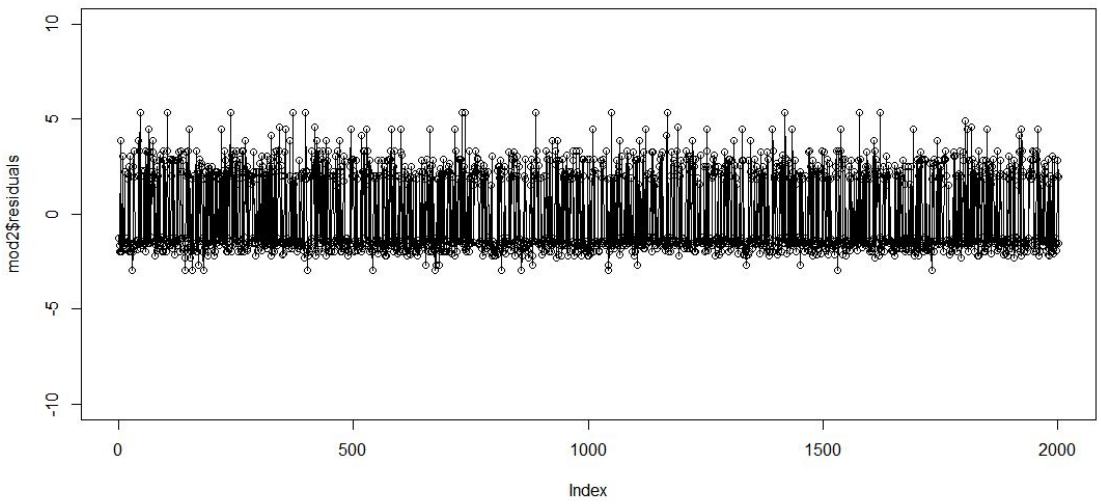

Figura 19: Gráfico de Dispersão dos Resíduos

qualitativas (HOSMER, 2013).

A seguir estão descritas as razões de chances e seus respectivos intervalos, referentes ao modelo 1 e 2.

Tabela 12: Razão de Chances do Modelo 1

Categoria	β_i	Razão de Chances	L.I 95%	L.S 95%
Intercepto	-0.36	0.6968	0.5185	0.9363
30 a 39 anos	-0.07	0.9315	0.6823	1.2718
40 a 49 anos	-0.06	0.942	0.6779	1.3089
50 a 59 anos	-0.265	0.7671	0.5292	1.112
60 anos ou mais	-0.362	0.6962	0.4739	1.0228
Evangélica	0.19	1.2104	0.9378	1.5623
Outras religiões	0.125	1.1331	0.7711	1.665
Sem religião ou crença	0.76	2.1441	1.442	3.188
Mais de 2 a 5 salários mínimos	-0.29	0.7511	0.5852	0.9641
Mais de 5 salários mínimos	-0.52	0.5943	0.4416	0.7997

As perguntas de interesse da pesquisa possuem respostas únicas, ou seja, não é possível responder a duas ou mais característica simultaneamente, como por exemplo ser Evangélica e Católica. Somente uma alternativa pode ser escolhida. A Tabela 8 tem por referência mulheres entrevistadas com faixa etária de 16 a 29 anos, católicas e de baixa renda (até 2 salários mínimos). Para esse perfil, a probabilidade de responder positivamente à pergunta: "Você já sofreu algum tipo de violência doméstica ou familiar provocada por um homem?"ou citar quaisquer das 12 situações de violência elencadas na pesquisa tendo como referência os últimos doze meses, observando os coeficientes relacionados na Tabela 12, é de:

$$E[Y] = \pi = \frac{e^{-0.36}}{1 + e^{-0.36}} = 0,41.$$

A partir da Tabela 12, mulheres sem religião ou crença, com renda superior a 5 salários mínimos e com faixa etária de 16 a 29 anos, têm probabilidade de perceber a ocorrência de violência doméstica e familiar e, possivelmente, quebrar o ciclo de violência de

$$E[Y] = \pi = \frac{e^{-0.36+0.76-0.52}}{1 + e^{-0.36+0.76-0.52}} = \frac{e^{-0.12}}{1 + e^{-0.12}} = 0,47.$$

Tabela 13: Razão de Chances do Modelo 2

Categoria	β_i	Razão de Chances	L.I 95%	L.S 95%
Intercepto	-0.19	0.8229	0.648	1.0451
Desocupada	0.16	1.1723	0.883	1.5562
Fora da Força de Trabalho	-0.64	0.5246	0.4018	0.6849
Evangélica	0.19	1.2144	0.9422	1.5652
Outras religiões	0.1	1.1046	0.7450	1.6376
Sem religião ou crença	0.72	2.0544	1.3792	3.06
Mais de 2 a 5 salários mínimos	-0.41	0.6662	0.5144	0.8627
Mais de 5 salários mínimos	-0.626	0.5349	0.3914	0.7310

O segundo modelo, Tabela 13, tem como referência mulheres ocupadas (exercem atividade remunerada), católicas e de baixa renda. Para esse perfil, a probabilidade de responder positivamente à pergunta P16 ou P26, de acordo com os coeficientes estimados, é:

$$E[Y] = \pi = \frac{e^{-0.195}}{1 + e^{-0.195}} = 0,45.$$

Considerando mulheres fora da força de trabalho, sem religião ou crença e de baixa renda, essa probabilidade é de:

$$E[Y] = \pi = \frac{e^{-0.195-0.64+0.72}}{1 + e^{-0.195-0.64+0.72}} = \frac{e^{-0.115}}{1 + e^{-0.115}} = 0,47.$$

Considerando a variável **Força de trabalho**, mulheres fora da força de trabalho tiveram coeficiente significativo, ao nível de significânica de 10%. Quando ocupada (exercendo atividade remunerada), a chance de responder positivamente às perguntas de interesse da pesquisa é aproximadamente duas vezes maior comparada às mulheres fora da força de trabalho.

Em ambos os modelos, a razão de chances de mulheres sem religião ou crença indica que essa categoria tem aproximadamente 100% mais chance (duas vezes maior) de

perceber a ocorrência de violência quando comparadas às mulheres católicas. O efeito dos coeficientes β_i, de acordo com a tabela 8, não foi significativo ao nível de 10% ($\alpha = 0.1$) para as demais religiões: evangélica, espírita, umbanda, candomblé, islâmica, judaica. Sob mesmo nível de significância, a razão de chances da faixa etária de 60 anos ou mais indica que essa categoria tem 30% menos chance de percepção da violência em relação às mulheres de 16 a 29 anos.

Em geral, a renda é um fator significativo em se tratando da ocorrência de violência. Isso gera reflexão a respeito de fatores como disponibilidade de informação e conhecimento a respeito do que é a violência doméstica para a mulher entrevistada. A razão de chances para mulheres com renda de 2 a 5 salários mínimos indica que essa categoria tem 25% menos chance de responder positivamente sobre a ocorrência de violência doméstica e familiar em relação a mulheres com renda inferoir a 2 salários mínimos mantidas constantes as demais varáveis.. Além disso, a razão de chances para mulheres com renda superior a 5 salários mínimos indica que essas têm aproximadamente 40% menos chance de responder positivamente à pergunta sobre ocorrência de violência, ou seja, quanto maior aa renda, menor a chance de resposta positiva.

5 Conclusão

Ambos os modelos de regressão logística estudados foram adequados para prever a probabilidade de ocorrência de violência doméstica e familiar contra a mulher, provocada por um homem em algum momento da vida, ou seja, identificam quais fatores têm influência sobre a percepção da mulher em relação as agressões.

A diferença entre os modelos resultantes foram as variáveis idade e força de trabalho. Em algumas pesquisas, é mais comum haver informações relacionadas a faixa etária das respondentes. As variáveis religião e renda das mulheres entrevistadas tiveram efeito significativo sobre a probabilidade de ocorrência da violência doméstica e familiar contra a mulher para os dois modelos. Essas características permitem refletir a respeito dos ambientes nos quais essa mulher se encontra, os convívios que ela possue e o compartilhamento de informação que possa existir.

Constatou-se que o fato de uma mulher não possuir religião ou crença torna duas vezes maior a chance de perceber a violência em relação às mulheres católicas, mantidas constantes as demais varáveis. Ainda considerando como referência mulheres católicas, o modelo não apresenta diferenças significativas para as demais religiões ou crenças.

Do acolhimento ao silenciamento, são muitos os papéis possíveis de intituições e líderes religiosos, exercidos diante das situações de violência doméstica e familiar. Esses grupos têm grande influência na vida das mulheres vítimas de agressões, ressalta Camila da Silva (2019), em seu artigo sobre "Qual o papel da religião na violência doméstica?".

De acordo com a teóloga Valéria Vilhena, fundadora do grupo Evangélicas pela Igualdade de Gênero: "A igreja é uma instituição social, como outras que formam a nossa sociedade. E portanto, ela faz parte do problema. Muitas vezes ela reproduz essa violência, ela perpetua essa violência". É importante também que haja um estudo em relação ao agressor para entender o que motiva as agressões, qual educação ele teve ou está tendo, e como evitá-las.

"Infelizmente, falsos entendimentos da Bíblia são usados para acobertar a violência contra as mulheres. A proteção, a defesa e a promoção da mulher também precisam passar por um aprofundamento da religiosidade, vencendo toda forma de mascaramento da violência", escreveu o padre Cleiton Viana da Silva, em um artigo sobre o papel da Igreja Católica no enfrentamento da violência doméstica.

Além disso, a variável renda, ainda que indiretamente, reflete o conhecimento e grau de escolaridade que essa mulher possui, o que transparece o discernimento das situações classificadas como violência doméstica e familiar.

A Delegada de Polícia Civil Desirée Cristina, titular da Delegacia da Mulher de

João Pessoa no ano da publicação de seu artigo, apontava que "A realidade das delegacias especializadas da mulher, multiplicadas em todo o país dão notícia de que pobreza, baixa escolaridade e violência andam juntos [...] "(VASCONCELOS, 2018). Desirée integra que uma mulher se torna duplamente vulnerável ao carecer de educação: "Na prática, sofrerá violência social proveniente da desigualdade social, como também as violências de gênero, domésticas e/ou familiares."(VASCONCELOS, 2018).

> Sabe-se que a violência contra a mulher é resultado de características individuais, contextuais e ambientais, que se presentes, aumentam a sua possibilidade. Várias características podem moldar os padrões e variações nas taxas de violência intrafamiliar, ainda que não necessariamente definam quem se tornara perpetrador ou vítima. São citadas as idades, o estado civil ou a personalidade de discriminar, hábitos de vida como o abuso de álcool e drogas ilícitas, a inserção social da família envolvendo baixa renda, pouca escolaridade e desemprego, ou ainda o papel de gênero nas relações familiares, quer presentes, quer históricas nas famílias de origem (ALBUQUERQUE, 2019).

A aplicação da regressão logística, além das análises descritivas realizadas, permitem compreender que trata-se não somente do risco de ocorrência da violência doméstica, mas da percepção e conhecimento que essas mulheres possuem em relação à ela, principalmente por se tratar da análise de uma pesquisa de opinião.

Neste cenário complexo, a falta de conhecimento sobre as especificidades desse crime dificultam o rompimento do silêncio e interrupção do ciclo de violência. Denunciar, porém, "não é uma tarefa fácil quando as agressões partem de uma pessoa com quem a vítima mantém relações íntimas de afeto, cujo rompimento coloca questões emocionais e objetivas, que envolvem a desestruturação do cotidiano e até mesmo o risco de morte para a mulher"(Dossiê Violência Contra a Mulher).

A dinâmica da violência doméstica costuma se repetir e se tornar cada vez mais grave e frequente. Portanto, reafirma-se a necessidade de constantes campanhas de conscientização sobre os direitos da mulher e as redes de atendimento. Além disso, proporcionar meios que minimizem a dependência econômica do agressor para criação dos filhos e melhorarem o acesso e a confiabilidade dos serviços de atendimento das mulheres em situação de violência.

6 Referências

AGRESTI, Alan. **An introduction to categorical data analysis**. John Wiley Sons, 2018.

ALBUQUERQUE, Neimar de Figueiredo. Violência doméstica e familiar: O impacto na relação com a Lei Maria da Penha. **DireitoNet**, 2019. Disponível em: https://www.direitonet.com.br/artigos/exibir/11306/Violencia-domestica-e-familiar-o-impacto-na-relacao-com-a-Lei-Maria-da-Penha. Acessado em: 05 de abril de 2021.

AMARAL, Luana Bandeira de Mello et al. Violência doméstica e a Lei Maria da Penha: perfil das agressões sofridas por mulheres abrigadas em unidade social de proteção. **Revista Estudos Feministas**, v. 24, n. 2, p. 521-540, 2016.

ARJONA, Reciane Cristina. Violência Doméstica contra Mulher. **JUS**, 2019. Disponível em: https://jus.com.br/artigos/74965/violencia-domestica-contra-mulher. Acesso em: 28 de set de 2020.

BINDER, D. et al. Analytic uses of survey data: a review. **Advances in the Statistical Sciences: Applied Probability, Stochastic Processes, and Sampling Theory**, p. 243-264, 1987.

BITTENCOURT, Hélio Radke. Regressão logística politômica: revisão teórica e aplicações. **Acta Scientiae**, v. 5, n. 1, p. 77-86, 2003.

CARVALHO, Fábio Janoni et al. Modelos lineares generalizados na agronomia: análise de dados binomiais e de contagem, zeros inflacionados e enfoque bayesiano. 2019.

CORDEIRO, Gauss Moutinho; DEMÉTRIO, Clarice GB. Modelos lineares generalizados e extensões. **Piracicaba: USP**, 2008.

COX, David Roxbee; HINKLEY, David Victor. **Theoretical statistics**. CRC Press, 1979.

Violência Doméstica e Familiar contra a Mulher. **DATASENADO**, 2019. Disponível em: https://www12.senado.leg.br/institucional/datasenado/arquivos/violencia-contra-a-mulher-agressoes-cometidas-por-2018ex2019-aumentam-quase-3-vezes-em-8-anos-1. Acesso em: 01 de out de 2020.

Violência Doméstica e Familiar contra a Mulher. **DATASENADO**, 2019. Disponível em: https://www12.senado.leg.br/institucional/datasenado/arquivos/violencia-contra-a-mulher-agressoes-cometidas-por-2018ex2019-aumentam-quase-3-vezes-em-8-anos-1. Acesso em: 18 de fev de 2021.

DIAS, Maria Berenice. Lei Maria da penha. **São Paulo: Ed. Revistas dos Tribunais**, 2015.

DO NASCIMENTO SILVA, Pedro Luis; DUARTE, Renata Pacheco Nogueira. Análise Estatística de Dados de Pesquisas por Amostragem: Problemas no Uso de Pacotes-Padrões. **INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA-IBGE**, p. 53.

DOBSON, Annette J.; BARNETT, Adrian G. **An introduction to generalized linear models**. CRC press, 2018.

DOMINONI, Thiago Dantas Bhering. Análise das propriedades do estimador Horvitz-Thompson. 2012.

DORNELLES FILHO, Adalberto Ayjara; MINCATO, Ramone; GRAZZI, Paula Cervelin. Perfil da mulher vítima de violência doméstica no Brasil, Rio Grande Sul e Caxias do Sul. **Anais do XIII Encontro Aspectos Econômicos e Sociais da Região Nordeste do RS**, 2014.

DUNN, Peter K.; SMYTH, Gordon K. **Generalized linear models with examples in R**. New York, NY: Springer, 2018.

ESSY, Daniela Benevides. A evolução histórica da violência contra a mulher no cenário brasileiro: do patriarcado à busca pela efetivação dos direitos humanos femininos. **Conteudo Juridico, Brasilia-DF**, v. 26, 2017.

FULLER, Wayne A. Least squares and related analyses for complex survey designs. **Survey Methodology**, v. 10, n. 1, p. 97-118, 1984.

FULLER, Wayne A. Regression analysis for sample survey. **Sankhya**, v. 37, n. 3, p. 117-132, 1975.

García-Portugués, E. *Notes for Predictive Modeling*. Version 5.8.6. ISBN 978-84-09-29679-8, 2019. Disponível em: https://bookdown.org/egarpor/PM-UC3M/. Acesso em: 18 de fev de 2021.

HEISE, Lori. Violência e gênero: uma epidemia global. **Cad Saúde Pública**, v. 10, n. Supl 1, 1994.

HOEF, Jay M.; BOVENG, Peter L. Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data?. **Ecology**, v. 88, n. 11, p. 2766-2772, 2007.

HOLT, D.; SMITH, T. M. F.; WINTER, P. D. Regression analysis of data from complex surveys. **Journal of the Royal Statistical Society: Series A (General)**, v. 143, n. 4, p. 474-487, 1980.

HORVATH, Ben. Generalized Linear Models: Residuals and Diagnostics. 2019.

HOSMER, David W.; LEMESHOW, Stanley; STURDIVANT, Rodney X. **Applied logistic regression**. John Wiley Sons, 2013.

HOW CAN I DO REGRESSION ESTIMATION WITH SURVEY DATA? — R FAQ. Disponível em: https://stats.idre.ucla.edu/r/faq/how-can-i-do-regression-estimation-with-survey-data/. Acesso em: 25 de fev de 2021.

IBGE, Djalma Galvão Carneiro Pessoa Consultor M. Análise de Dados Amostrais Complexos, 2018. Disponível em: https://djalmapessoa.github.io/adac/index.html. Acesso em: 20 de fev de 2021.

KRUG, Etienne G. et al. Informe mundial sobre la violencia y la salud. 2003.

NATHAN, Gad; HOLT, D. The effect of survey design on regression analysis. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 42, n. 3, p. 377-386, 1980.

NETER, John et al. Applied linear statistical models. 1996.

ORGANIZAÇÃO MUNDIAL DA SAÚDE; KRUG, Etienne G. **Relatório mundial sobre violência e saúde**. Genebra: Organização Mundial da Saúde, 2002.

PAULA, Gilberto Alvarenga. **Modelos de regressão: com apoio computacional**. São Paulo: IME-USP, 2004.

PEDUZZI, Peter et al. A simulation study of the number of events per variable in logistic regression analysis. **Journal of clinical epidemiology**, v. 49, n. 12, p. 1373-1379, 1996.

PESSOA, Djalma GC; SILVA, PL Nascimento; DUARTE, Renata PN. Análise estatística de dados de pesquisas por amostragem: problemas no uso de pacotes-padrão. **Revista Brasileira de Estatística**, v. 58, n. 210, p. 53-75, 1997.

PRIORE, Mary del. Histórias íntimas: sexualidade e erotismo na história do Brasil. **São Paulo: Editora Planeta do Brasil, 2011. Entrevista de Daniela Galdino, concedida à SELFIEPOESIA**. Disponível em: /https://www. youtube. com/watch, 2005. Acesso em: 5 de fev de 2021.

ROSEMBERG, Vítor Teófilo. Estudo sobre a incidência de acidentes fatais no trânsito do Distrito Federal. 2018. 60 f., il. Trabalho de Conclusão de Curso (Bacharelado em Estatística)—Universidade de Brasília, Brasília, 2018.

SALLES, Taian Cristal Ferreira. Aplicação de modelos de regressão logística em um estudo de neurocirurgia. 2018. 31 f., il. Trabalho de Conclusão de Curso (Bacharelado em Estatística)—Universidade de Brasília, Brasília, 2018.

SAFFIOTI, Heleieth IB. O poder do macho. São Paulo. **Moderna**, 1987.

SAGIM, Mirian Botelho et al. Violência doméstica: a percepção que as vítimas têm de seu parceiro, do relacionamento mantido e das causas da violência. **Cogitare Enfermagem**, v. 12, n. 1, p. 30-36, 2007.

SILVA, Camila. Qual o papel da religião na violência doméstica?. **AZMina**, 2019. Disponível em: https://azmina.com.br/reportagens/qual-o-papel-da-religiao-na-violencia-domestica/. Acessado em: 26 de abril de 2021.

SOUZA, Francisco das Chagas Silva. Histórias íntimas: sexualidade e erotismo na história do Brasil. **Cadernos de Pesquisa**, v. 42, n. 146, p. 672-678, 2012.

TEAM, R. Core et al. R: A language and environment for statistical computing. 2013.

VASCONCELOS, Desirée Cristina Rodrigues Vasconcelos. EDUCAÇÃO, DIREITOS HUMANOS E VIOLÊNCIA CONTRA MULHER: ASPECTO PEDAGÓGICO DA LEI MARIA DA PENHA. **Revista Jurídica Faculdade Paraibana-FAP**, p. 73, 2018.

ZART, Louise; SCORTEGAGNA, Silvana Alba; PIBIC, P. Perfil sociodemográfico de mulheres vítimas de violência doméstica e circunstâncias do crime. **Erechim: Perspectiva**, v. 39, n. 148, p. 85-93, 2015.

ZUUR, Alain et al. **Mixed effects models and extensions in ecology with R**. Springer Science & Business Media, 2009.