Fábio Mello Valladão

Avaliação do impacto da função de ligação na qualidade do ajuste de modelo linear generalizado para um desfecho binário de consulta ao médico

Niterói - RJ, Brasil

07 de dezembro de 2020

Universidade Federal Fluminense

Fábio Mello Valladão

Avaliação do impacto da função de ligação na qualidade do ajuste de modelo linear generalizado para um desfecho binário de consulta ao médico

Trabalho de Conclusão de Curso

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientador: Prof. Dr. José Rodrigo de Moraes

Niterói - RJ, Brasil

07 de dezembro de 2020

Universidade Federal Fluminense

Fábio Mello Valladão

Avaliação do impacto da função de ligação na qualidade do ajuste de modelo linear generalizado para um desfecho binário de consulta ao médico

Monografia de Projeto Final de Graduação sob o título *"Avaliação do impacto da função de ligação na qualidade do ajuste de modelo linear generalizado para um desfecho binário de consulta ao médico"*, defendida por Fábio Mello Valladão em 07 de dezembro de 2020, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

_____

Prof. Dr. José Rodrigo de Moraes

Orientador

Departamento de Estatística – UFF

_____

Prof. Dr. Licinio Esmeraldo da Silva

Departamento de Estatística – UFF

_____

Profa. Dra. Luciane Ferreira Alcoforado

Departamento de Estatística – UFF

Niterói, 07 de dezembro de 2020

Ficha catalográfica automática - SDC/BIME
Gerada com informações fornecidas pelo autor

V176a Valladão, Fábio Mello
Avaliação do impacto da função de ligação na qualidade do ajuste de modelo linear generalizado para um desfecho binário de consulta ao médico / Fábio Mello Valladão ; José Rodrigo de Moraes, orientador. Niterói, 2020.
57 f. : il.

Trabalho de Conclusão de Curso (Graduação em Estatística)-Universidade Federal Fluminense, Instituto de Matemática e Estatística, Niterói, 2020.

1. Modelo probit. 2. Modelo logit. 3. Modelo complemento log-log. 4. Plano amostral complexo. 5. Produção intelectual. I. Moraes, José Rodrigo de, orientador. II. Universidade Federal Fluminense. Instituto de Matemática e Estatística. III. Título.

CDD -

Bibliotecário responsável: Sandra Lopes Coelho - CRB7/3389

Resumo

A função de ligação é uma das três componentes de um modelo linear generalizado, e a escolha de uma função de ligação inapropriada pode influenciar a significância dos parâmetros e a qualidade do ajuste do modelo. Este trabalho teve como objetivo avaliar o impacto das funções de ligação logit, probit e complemento log-log na qualidade do ajuste, bem como no sentido, magnitude e significância das associações entre um conjunto de características sociodemográficas e de saúde de idosos e um desfecho binário referente ao tempo da última consulta médica realizada. Os modelos foram ajustados a partir dos dados da Pesquisa Nacional de Saúde 2013, usando o método de Máxima Pseudo-Verossimilhança (MPV). Com relação aos resultados do trabalho, nos modelos probit, logit e complemento log-log, as mesmas variáveis sociodemográficas e de saúde dos idosos apresentaram associação significativa com o desfecho de estudo. Além disso, os sinais das estimativas pontuais dos parâmetros nos três modelos foram iguais. Entretanto, as estimativas pontuais (padronizadas) do modelo complemento log-log foram maiores (variação relativa superior a 10%) que as respectivas estimativas obtidas nos modelos probit e logit, sendo que entre estes dois últimos modelos as estimativas pontuais foram similares (variação relativa inferior a 5%). As medidas de Pseudo-R^2 de Cox-Snell ($R^2_{CS} \cong 0{,}11$) e Nagelkerke ($R^2_N \cong 0{,}18$), as medidas de sensibilidade e especificidade, assim como as áreas sob a curva ROC ($A \cong 0{,}75$), também foram bem similares nos três modelos. Conclui-se, portanto, que para os dados utilizados no presente estudo, a escolha da função de ligação não influenciou a significância e o sentido das associações entre as características dos idosos e o desfecho do estudo, e nem na qualidade do ajuste e na capacidade preditiva dos modelos. Desse modo, qualquer uma das três funções de ligação poderia ser escolhida, entretanto, em investigações na área de saúde, o modelo logit é o mais frequentemente utilizado pela facilidade de interpretação das estimativas e por possibilitar o cálculo de razão de chances (*odds ratio*), que é uma medida de associação utilizada em estudos epidemiológicos.

Palavras-chaves: Modelo probit. Modelo logit. Modelo complemento log-log. Função de ligação. Plano amostral complexo.

Agradecimentos

A Deus, que me abençoa com infinitamente mais do que mereço.

À minha família e amigos, pessoas que amo e sem as quais eu não teria chegado até aqui.

Ao professor José Rodrigo de Moraes, cuja dedicada orientação em cada etapa dessa tarefa foi indispensável para que ela fosse completada.

Aos professores Licinio da Silva e Luciane Alcoforado, pelas contribuições a este trabalho ao participarem da banca avaliadora, e a Germán Rodríguez, professor de Princeton, que solicitamente nos esclareceu certos pontos da seção 2.2.3., sobre os quais havíamos encontrado pouca literatura.

A todos os professores e mestres que contribuíram para minha formação ao longo dos anos.

A vocês, minha mais profunda gratidão.

Sumário

Lista de Figuras

Lista de Quadros

Lista de Tabelas

1. Introdução

A inferência clássica assume que os dados da amostra são provenientes de um plano aleatório simples com reposição, sendo considerados realizações de variáveis aleatórias independentes e identicamente distribuídas. Neste caso, os parâmetros de Modelos Lineares Generalizados (MLGs) são geralmente estimados usando o método de Máxima Verossimilhança (CORDEIRO; DEMÉTRIO, 2013). Entretanto, este método de estimação é inadequado quando as amostras são selecionadas utilizando planos amostrais complexos, e isso se dá porque as propriedades de independência e identidade de distribuição (IID) não são válidas nesses casos. A alternativa, portanto, é realizar a estimação dos parâmetros dos MLGs usando o método de Máxima Pseudo-Verossimilhança (MPV) (PESSOA; SILVA, 1998).

Modelos Lineares Generalizados (MLGs) são utilizados em diferentes áreas do conhecimento para avaliar a associação entre um conjunto de variáveis explicativas e uma determinada variável resposta (desfecho), numérica ou categórica, bem como para identificar um subconjunto de variáveis explicativas significativamente relacionadas com a variável resposta e que produzam um modelo mais parcimonioso (DOBSON; BARNETT, 2018).

MLGs são definidos por três componentes: uma *componente aleatória* composta de uma variável resposta com n observações com distribuição de probabilidade pertencente à família exponencial; uma *componente determinística* caracterizada pelo preditor linear, formado pelas variáveis explicativas e pelos parâmetros do modelo; e a *função de ligação*, que relaciona a média da variável resposta ao preditor linear (KUTNER *et al*., 2005, CORDEIRO; NETO, 2004).

A função de ligação é uma componente fundamental de um MLG, e a escolha da função de ligação é normalmente definida pelas próprias características da variável resposta, como, por exemplo, positividade ou ainda devido a facilidade de interpretação (FARAWAY, 2006). As principais funções de ligação adotadas em MLGs para respostas binárias são as funções logit, probit e complemento log-log, e sua escolha deve-se ao fato delas garantirem que as probabilidades estimadas do evento de interesse estejam compreendidas no intervalo $[0,1]$. Embora as demais funções também sejam utilizadas na prática, a logit possui características que a tornam a função de ligação preferida na análise de dados binários, como propriedades teóricas mais simples e a interpretação dos

parâmetros como “efeitos” das variáveis explicativas no logaritmo da chance (*odds*) em favor da ocorrência do evento de interesse, o que faz com que seja bastante utilizada em estudos toxicológicos e epidemiológicos (CORDEIRO; NETO, 2004). A função de ligação logit é frequentemente utilizada em vários estudos, sobretudo na área da saúde (NORONHA; ANDRADE, 2005, OSORIO; SERVO; PIOLA, 2011, FONTES; CONCEIÇÃO; MACHADO, 2017; REIS *et al*., 2000; KELLES *et al*., 2015), e sua utilização possibilita a estimação de razão de chances (*odds ratio*), medida de associação utilizada em diferentes estudos epidemiológicos.

Todavia, cabe destacar que a escolha de uma função de ligação inapropriada pode influenciar a significância do modelo e a sua capacidade preditiva (MARÔCO, 2010).

Usando os dados da Pesquisa Nacional de Saúde (PNS) 2013, este trabalho teve como objetivo analisar o impacto da escolha da função de ligação, seja ela logit, probit ou complemento log-log, na qualidade do ajuste de MLGs adotados para avaliar a associação entre variáveis sociodemográficas e de saúde e um desfecho binário de consulta médica por idosos (60 anos ou mais). No presente estudo, optou-se por utilizar esse grupo etário em função do rápido envelhecimento populacional ocorrido no Brasil, a partir da segunda metade do século XX, o que demanda a redefinição de políticas e ações preventivas específicas para os idosos, a fim de promover melhorias na sua qualidade de vida e seu estado de saúde (BRASIL, 2014, MARTINE; MCGRANAHAN, 2010, WONG; CARVALHO, 2006).

A PNS é uma pesquisa que utilizou um plano amostral conglomerado em três estágios de seleção, com estratificação das unidades primárias de amostragem (UPAs), onde as UPAs são os setores censitários, ou conjuntos de setores, os domicílios são as unidades secundárias de amostragem (USAs), e os moradores com 18 anos ou mais são as unidades terciárias de amostragem (UTAs) (IBGE, 2014). Esta estrutura denota um plano amostral complexo, o que sinaliza a necessidade de se considerar na modelagem estatística os pesos amostrais distintos referentes às unidades da amostra, a conglomeração e a estratificação, de forma a não produzir estimativas fundamentalmente incorretas para parâmetros do modelo e/ou para variâncias dos estimadores dos referidos parâmetros (PESSOA; SILVA, 1998, MORAES; MOREIRA; LUIZ, 2012).

1.1. Objetivos Geral e Específicos

O objetivo geral é avaliar o impacto das funções de ligação logit, probit e complemento log-log na qualidade do ajuste de modelo linear generalizado usando um desfecho binário de consulta ao médico.

Entre os objetivos específicos, pode-se citar:

- Analisar a distribuição do desfecho de consulta médica, segundo as características sociodemográficas e de saúde dos idosos.
- Avaliar o sentido e a significância das associações entre características sociodemográficas e de saúde de idosos e o desfecho de consulta ao médico, usando os modelos logit, probit e complemento log-log.
- Analisar e comparar as medidas de qualidade do ajuste e de capacidade discriminatória dos modelos logit, probit e complemento log-log.
- Estimar a probabilidade do idoso nunca ter ido ao médico ou ter ido há 1 ano ou mais, segundo os níveis de escolaridade e os níveis de autoavaliação de saúde, considerando um dado perfil a partir das demais características do idoso.

2. Material e Métodos

2.1. PNS 2013

A Pesquisa Nacional de Saúde (PNS), realizada pela primeira vez em 2013, faz parte do Sistema Integrado de Pesquisas Domiciliares (SIPD), um projeto que teve como objetivo reformular as pesquisas domiciliares amostrais já existentes no Brasil e padronizar futuros levantamentos, de modo a assegurar a mesma infraestrutura de amostragem para todas as pesquisas pertencentes ao SIPD (IBGE, 2007). Isso foi metodologicamente possível a partir da adoção da Amostra Mestra, definida como "*um conjunto de unidades de áreas que são selecionadas para atender a diversas pesquisas*" (SOUZA-JÚNIOR *et al*., 2015, p. 208).

A PNS teve por objetivo coletar informações sobre o desempenho do sistema de saúde do país, sobre as condições de saúde da população brasileira, e aumentar a vigilância a respeito de doenças crônicas não transmissíveis e de seus fatores associados. Procurou-se também estimar indicadores com a precisão desejada e monitorar outros indicadores previamente levantados no Suplemento Saúde da Pesquisa Nacional por Amostras de Domicílios (PNAD) (IBGE, 2014).

A PNS levantou informações sobre os moradores de domicílios particulares permanentes (DPPs) no Brasil, excetuando-se aqueles moradores cujos domicílios estão localizados em setores censitários especiais, como bases militares, alojamentos, penitenciárias, asilos, orfanatos, hospitais, entre outros (IBGE, 2014).

2.1.1. Plano de amostragem da PNS

O plano amostral utilizado na PNS foi o de amostragem por conglomerado em três estágios de seleção, com estratificação das unidades primárias de amostragem (UPAs). As UPAs são setores censitários ou conjuntos de setores censitários, os domicílios são as unidades secundárias de amostragem (USAs), e os moradores com 18 anos ou mais são as unidades terciárias de amostragem (UTAs) (IBGE, 2014).

As unidades que formam as UPAs na PNS são UPAs em todas as pesquisas cujas amostras são subamostras da *Amostra Mestra* do SIPD, e são estratificadas da mesma maneira, de acordo com quatro critérios: administrativo, geográfico, de situação

censitária e estatístico. Sob o *critério administrativo*, essas UPAs são estratificadas primeiramente em função da Unidade da Federação (UF) e depois são subdivididas em áreas de capital, resto da Região Metropolitana (RM), Região Integrada de Desenvolvimento Econômico (RIDE), ou resto da UF. Já o sob o *critério geográfico*, as capitais e demais municípios de grande porte são novamente estratificados, em distritos, subdistritos e bairros. O *critério de situação censitária* subdivide os estratos anteriores em urbano e rural, e, por fim, segundo o *critério estatístico*, os estratos resultantes são subdivididos de forma homogênea de acordo com o rendimento total dos domicílios e total de DPP, objetivando melhorar a precisão das estimativas (SOUZA-JÚNIOR *et al*., 2015).

Sobre os critérios de seleção das unidades da *Amostra Mestra* que comporiam a amostra da PNS, tem-se que a escolha dos setores censitários (UPAs) foi por amostragem aleatória simples (AAS). As unidades domiciliares (USAs) também foram escolhidas por AAS, selecionando-se uma quantidade fixa de DPPs em cada setor da amostra (UPA) a partir do Cadastro Nacional de Endereços para Fins Estatísticos (CNEFE). E, finalmente, dentro de cada domicílio da amostra (USA), selecionou-se por AAS um dentre os moradores adultos com 18 anos ou mais de idade, formando assim as UTAs (IBGE, 2014).

2.1.2. Questionário

O questionário da PNS foi separado em três partes de acordo com os temas abordados. Na primeira parte, a domiciliar, buscaram-se informações sobre o domicílio, como o material que predomina na construção de diferentes partes do domicílio, a forma de abastecimento de água, a quantidade de cômodos que servem como dormitório, entre outras. Além disso, foram levantadas informações sobre visitas domiciliares, seja por agentes de combate a endemias ou pela equipe de Saúde da Família. Todas essas questões deveriam ser respondidas pelo responsável pelo domicílio ou por outro morador que possuísse as informações no momento da entrevista.

Na segunda parte, foram coletadas informações sobre os moradores daquele domicílio, com o objetivo de conhecer as características gerais dos moradores, como escolaridade, renda e saúde, de forma que cada morador deveria responder ao questionário.

Já na terceira parte do questionário, a individual, apenas um morador adulto, escolhido aleatoriamente, deveria ser entrevistado por domicílio, respondendo os quesitos sobre trabalho, apoio social, e questões relacionadas à sua saúde (SOUZA-JÚNIOR *et al*., 2015).

2.2. Modelos Lineares Generalizados

2.2.1. Componentes de um MLG

Um modelo linear generalizado (MLG) é definido em termos de um conjunto de n variáveis aleatórias, $Y_1, Y_2, \ldots, Y_n$, que formam a componente *aleatória* do modelo, e são independentes, com distribuição pertencente à família exponencial[1], e com as seguintes propriedades (DOBSON; BARNETT, 2018):

1) A distribuição de probabilidade de cada Y_i tem a forma canônica (ou seja, tal que $a(y_i) = y_i$), e depende de apenas um parâmetro p_i. Portanto, pode ser escrita do seguinte modo:

$$P(Y_i = y_i, p_i) = e^{a(y_i)\,\mathrm{b}(p_i)+\mathrm{c}(p_i)+\mathrm{d}(y_i)}$$

onde a e d são funções que dependem apenas da observação y_i, e b e c são funções que dependem apenas do parâmetro p_i.

2) A distribuição de todas as variáveis aleatórias $Y_1, Y_2, \ldots, Y_n$ tem a mesma forma, e por isso a distribuição de probabilidade conjunta dessas variáveis aleatórias pode ser escrita da seguinte maneira:

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n, p_1, p_2, \ldots, p_n) = e^{\sum_{i=1}^{n}[a(y_i)\,\mathrm{b}(p_i)+\mathrm{c}(p_i)+\mathrm{d}(y_i)]} \quad (1.1)$$

---

[1] A distribuição de probabilidade de uma variável aleatória Y, que depende apenas de um parâmetro p, será pertencente à família exponencial se puder ser escrita como: $P(Y = y, p) = e^{a(y)\,\mathrm{b}(p)+\mathrm{c}(p)+\mathrm{d}(y)}$, onde a, b, c e d são funções conhecidas. Além disso, se $a(y) = y$, diz-se que a distribuição tem forma canônica, e que $\mathrm{b}(p)$ é o parâmetro natural dessa distribuição (DOBSON; BARNETT, 2018).

Seja $E[Y_i] = \mu_i$, onde μ_i é uma função do parâmetro p_i. Também fazem parte do modelo:

a) o preditor linear (η_i), que forma a componente *determinística* do modelo, e é função de $\boldsymbol{x}_i = [x_{i1} \quad x_{i2} \quad \cdots \quad x_{ip}]$, o i-ésimo vetor de variáveis explicativas, com dimensão $p \times 1$, e do vetor $\boldsymbol{\beta}$ de parâmetros a serem estimados, também com dimensão $p \times 1$:

$$\eta_i = \boldsymbol{x}_i^t \boldsymbol{\beta} = [x_{i1} \quad x_{i2} \quad \cdots \quad x_{ip}] \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

b) g, a função de ligação, que é monótona e diferenciável, e que relaciona a média da variável resposta ao preditor linear:

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i^t \boldsymbol{\beta}$$

A distribuição binomial pertence à família exponencial e pode ser escrita na forma canônica (DOBSON; BARNETT, 2018), portanto, o mesmo pode ser afirmado para a distribuição de Bernoulli, uma vez que esta distribuição é um caso particular da binomial. Tem-se então que as n variáveis com distribuição de Bernoulli possuem a propriedade 1.

Seja $(Y_1, Y_2 \ldots, Y_n)$ uma amostra aleatória de tamanho n, onde Y_i tem distribuição de probabilidade de Bernoulli com parâmetro $p_i, \forall\, i = 1,2, \ldots, n$, onde p_i é a probabilidade de sucesso. Então, a função de probabilidade de Y_i é dada por:

$$P(Y_i = y_i, p_i) = {p_i}^{y_i} \, (1 - p_i)^{1 - y_i};\; y_i = 0, 1 \tag{1.2}$$

Como as variáveis aleatórias componentes da amostra são independentes, a função de probabilidade conjunta de $Y_1, Y_2, \ldots, Y_n$ é dada por:

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n, p_1, \ldots, p_n) = \prod_{i=1}^{n} P(Y_i = y_i, p_i)$$

$$= \prod_{i=1}^{n} e^{\ln P(Y_i = y_i, p_i)} \qquad (1.3)$$

Substituindo (1.2) em (1.3), obtém-se a equação a seguir:

$$\begin{aligned} P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n, p_1, \dots, p_n) &= \prod_{i=1}^{n} e^{\ln\left[p_i{}^{y_i} \cdot (1-p_i)^{1-y_i}\right]} \\ &= \prod_{i=1}^{n} e^{y_i \ln p_i + (1-y_i)\ln(1-p_i)} \\ &= \prod_{i=1}^{n} e^{y_i \ln p_i + \ln(1-p_i) - y_i \ln(1-p_i)} \\ &= \prod_{i=1}^{n} e^{y_i \ln\left(\frac{p_i}{1-p_i}\right) + \ln(1-p_i)} \qquad (1.4) \end{aligned}$$

Finalmente, calculando o produtório, tem-se que:

$$P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n, p_1, \dots, p_n) = e^{\sum_{i=1}^{n}\left[y_i \ln\left(\frac{p_i}{1-p_i}\right) + \ln(1-p_i)\right]} \qquad (1.5)$$

A equação (1.5) mostra que a distribuição de probabilidade conjunta das n variáveis com distribuição de Bernoulli pode ser escrita como em (1.1), onde:

$$a(y_i) = y_i;\ \ \mathrm{b}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right);\ \ c(p_i) = \ln(1-p_i);\ \ d(y_i) = 0$$

O que demonstra que as variáveis aleatórias com distribuição de Bernoulli também possuem a propriedade 2, e que, portanto, podem formar a componente *aleatória* de um MLG, que terá desfecho binário.

2.2.2. Modelos para desfecho binário

A seguir são apresentados três tipos de MLGs para desfecho binário, ou seja, desfecho cujos valores possíveis são 0 (fracasso) ou 1 (sucesso). Cada modelo é obtido

ao igualar a função de ligação escolhida ao preditor linear, como visto no item b) da seção 2.2.1. No caso de p variáveis explicativas, o modelo assume a seguinte forma:

$$\begin{aligned} g(p_i) &= \boldsymbol{x}_i^t \boldsymbol{\beta} \\ &= \beta_1 x_{i1} + \beta_2 x_{i2} \ldots + \beta_p x_{ip} \end{aligned}$$

2.2.2.1. Modelo logit binário

Modelo no qual a função de ligação $g(p_i)$ é a logit, ou *logística*, dado por:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_1 x_{i1} + \beta_2 x_{i2} \ldots + \beta_p x_{ip}$$

onde: $g(p_i) = \ln\left(\frac{p_i}{1-p_i}\right)$ é o logaritmo da chance em favor da ocorrência do evento de interesse (sucesso).

A partir da equação anterior, pode-se determinar a expressão da probabilidade de sucesso para o *i*-ésimo elemento da amostra:

$$p_i = \frac{e^{\beta_1 x_{i1} + \beta_2 x_{i2} \ldots + \beta_p x_{ip}}}{1 + e^{\beta_1 x_{i1} + \beta_2 x_{i2} \ldots + \beta_p x_{ip}}}$$

2.2.2.2. Modelo probit binário

Modelo no qual a função de ligação $g(p_i)$ é a probit, ou *probito*, dado por:

$$F_Z^{-1}(p_i) = \beta_1 x_{i1} + \beta_2 x_{i2} \ldots + \beta_p x_{ip}$$

onde $g(p_i) = F_Z^{-1}(p_i)$ é a inversa da função distribuição acumulada da Normal padrão no ponto p_i.

A probabilidade de sucesso é dada pela função de distribuição acumulada da Normal padrão, avaliada no preditor linear, ou seja:

$$p_i = F_Z(\boldsymbol{x}_i^t\boldsymbol{\beta}) = P(Z \leq \boldsymbol{x}_i^t\boldsymbol{\beta}) = \int_{-\infty}^{\boldsymbol{x}_i^t\boldsymbol{\beta}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

2.2.2.3. Modelo complemento log-log

Modelo no qual a função de ligação $g(p_i) = \ln[-\ln(1-p_i)]$ é o complemento log-log, representado pela equação a seguir:

$$\ln[-\ln(1-p_i)] = \beta_1 x_{i1} + \beta_2 x_{i2} \ldots + \beta_p x_{ip}$$

A partir da equação anterior, pode-se determinar a expressão da probabilidade de sucesso para o *i*-ésimo elemento da amostra:

$$p_i = 1 - e^{-e^{x_i^t\boldsymbol{\beta}}}$$

Os três modelos apresentados são transformações de p que garantem que a probabilidade estimada $\hat{p}$ pertença ao intervalo $[0,1]$, quaisquer que sejam os valores de $\boldsymbol{x}$ e de $\boldsymbol{\beta}$ (POWERS; XIE, 1999).

Na Figura 1, apresenta-se o comportamento da inversa da função de distribuição acumulada (F^{-1}) de cada um desses modelos, para os possíveis valores de p. Observa-se que as três funções de ligação têm comportamentos bem semelhantes e praticamente linear para valores de p compreendidos no intervalo $[0{,}1; 0{,}9]$. Para valores pequenos de p, as funções logit e complemento log-log são bem próximas, decaindo de forma mais rápida que a função probit. Já para maiores valores de p, quando se aproximam de 1, a função complemento log-log cresce de forma mais lenta que as funções logit e probit.

Além disso, a partir da Figura 1 é possível notar o comportamento simétrico dos modelos logit e probit ao redor de $p = 0{,}5$, o que não ocorre com o modelo complemento log-log.

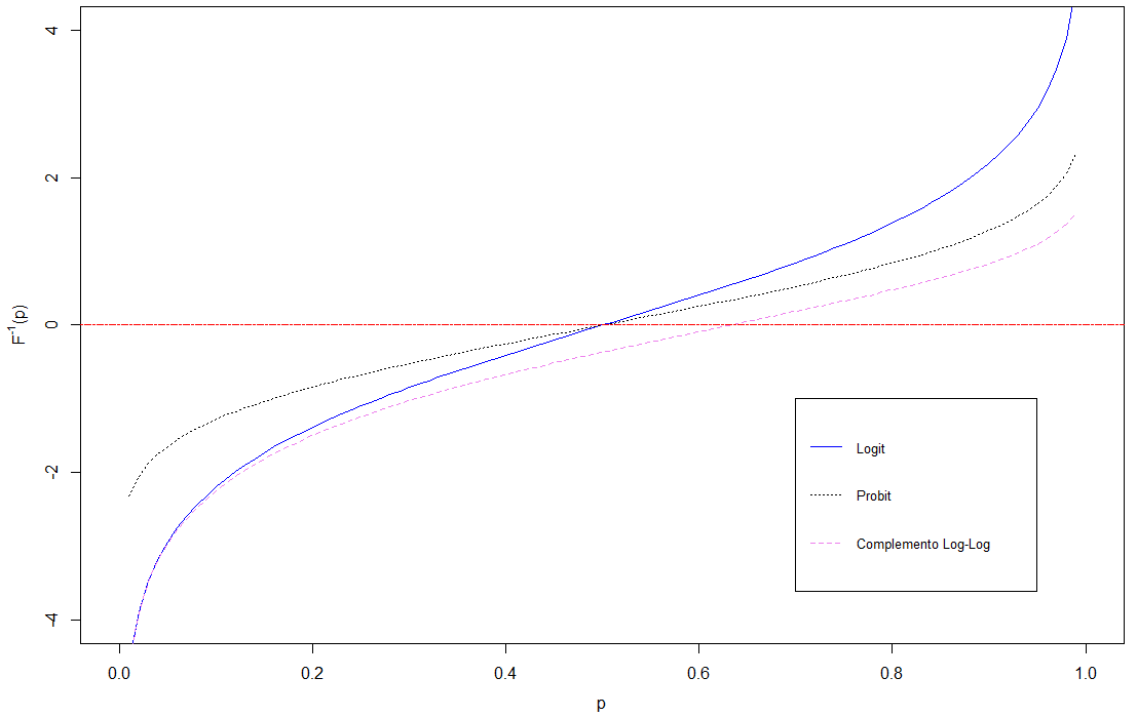


Figura 1: Inversas das funções de distribuição acumulada dos modelos logit, probit e complemento log-log avaliadas no intervalo de valores possíveis de p.

2.2.3. Variáveis latentes

Outra maneira de definir os modelos logit, probit e complemento log-log é por meio de equações lineares usando variáveis respostas latentes, que são variáveis aleatórias contínuas, não observáveis, que variam no intervalo$]-\infty, +\infty[$, e que geram os valores da variável resposta observada. Segundo Powers e Xie (1999), uma variável latente Y^* pode ser vista como índice de propensão para a variável resposta observada Y assumir o valor 1 ($Y = 1$).

No caso de um MLG com resposta binária, pode-se representar por Y_i^* uma variável aleatória contínua latente (ou não-observável), definida na reta real, que dá origem a Y_i, variável aleatória binária observável que assume os valores 0 (fracasso) ou 1 (sucesso), através da seguinte relação:

$$Y_i = \begin{cases} 1 & se \quad Y_i^* > \alpha \\ 0 & se \quad Y_i^* \leq \alpha \end{cases}$$

A variável binária assume o valor 1 quando a variável latente Y_i^* é maior que o ponto de corte α, e assume o valor 0 quando a variável latente é menor ou igual a este ponto de corte. Em termos probabilísticos, pode-se dizer que a probabilidade de sucesso (p_i) da variável observável referente a i-ésima unidade depende da variável latente da seguinte forma:

$$p_i = P(Y_i = 1) = P(Y_i^* > \alpha)$$

Supondo que Y_i^* dependa de um vetor de variáveis explicativas $\boldsymbol{x_i}$, pode-se representar a variável latente através do seguinte modelo linear:

$$Y_i^* = \boldsymbol{x}_i^t \boldsymbol{\beta} + \varepsilon_i$$

onde ε_i é o erro aleatório do modelo, com determinada distribuição de probabilidade e função de distribuição acumulada (fda) denotada por $F(\varepsilon)$. Utilizando o ponto de corte $\alpha = 0$ e especificando a distribuição de probabilidade para o erro aleatório ε_i, e consequentemente para a varíavel latente Y_i^* do modelo, pode-se determinar a probabilidade de sucesso para cada unidade da amostra, do seguinte modo:

$$p_i = P(Y_i^* > 0) = P(\varepsilon_i > -\boldsymbol{x}_i^t \boldsymbol{\beta}) = 1 - F(-\boldsymbol{x}_i^t \boldsymbol{\beta}) = F(\boldsymbol{x}_i^t \boldsymbol{\beta}) \quad (1.6)$$

Observa-se então que a probabilidade de sucesso é encontrada a partir da fda avaliada no ponto do preditor linear $\boldsymbol{x}_i^t \boldsymbol{\beta}$, como indicado em (1.6). Daí resulta o modelo estatístico geral a seguir, obtido ao se aplicar a função inversa da fda, explicitando o preditor linear do modelo:

$$F^{-1}(p_i) = \boldsymbol{x}_i^t \boldsymbol{\beta} \quad (1.7)$$

2.2.3.1. Modelo logit binário

No modelo logit binário, o erro aleatório ε_i tem distribuição logística padrão com média 0 e variância $\frac{\pi^2}{3}$. Portanto, a fda do erro aleatório ε_i no ponto η_i, tal que $\eta_i \in \mathbb{R}$, é dada por:

$$F(\eta_i) = P(\varepsilon_i \leq \eta_i) = \frac{1}{1+e^{-\eta_i}} = \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

Como definido genericamente na expressão (1.6), a probabilidade de sucesso para *i*-ésima unidade da amostra pode ser obtida diretamente da fda da distribuição logística padrão, no ponto $\eta_i = \boldsymbol{x}_i^t\boldsymbol{\beta}$, como representado a seguir:

$$p_i = F(\boldsymbol{x}_i^t\boldsymbol{\beta}) = \frac{1}{1+e^{-\boldsymbol{x}_i^t\boldsymbol{\beta}}} = \frac{e^{\boldsymbol{x}_i^t\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i^t\boldsymbol{\beta}}}$$

Aplicando a função inversa, como indicado em (1.7), a seguinte expressão é encontrada algebricamente: $\boldsymbol{x}_i^t\boldsymbol{\beta} = F^{-1}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right)$, que é o modelo apresentado em 2.2.2.1. (modelo logit), onde a função de ligação é a logit.

2.2.3.2. Modelo probit binário

No modelo probit binário, a distribuição do erro aleatório é a distribuição normal padrão, isto é, $\varepsilon_i \sim N(0,1)$. Neste caso, a fda no ponto η_i é dada por:

$$F_Z(\eta_i) = P(\varepsilon_i \leq \eta_i) = \int_{-\infty}^{\eta_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

Então, a probabilidade de sucesso é obtida calculando o valor da fda da distribuição normal padrão no ponto $\eta_i = \boldsymbol{x}_i^t\boldsymbol{\beta}$, como mostrado a seguir:

$$p_i = F_Z(\boldsymbol{x}_i^t\boldsymbol{\beta}) = \int_{-\infty}^{\boldsymbol{x}_i^t\boldsymbol{\beta}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

Por não possuir forma fechada, a aplicação da função inversa é feita apenas por métodos iterativos, e a expressão $\boldsymbol{x}_i^t\boldsymbol{\beta} = F_Z^{-1}(p_i)$ representa o modelo mostrado em

2.2.2.2. (modelo probit), aquele em que a função de ligação é a inversa da fda da distribuição normal padrão, também chamada função probit.

Cabe mencionar que tanto a distribuição normal padrão quanto a distribuição logística padrão tem média igual a zero, mas as variâncias são diferentes, isto é, a variância do erro é igual a 1 no modelo probit e $\pi^2/3$ no modelo logit. Assim, para comparar as estimativas pontuais dos parâmetros do modelo logit com as respectivas estimativas do modelo probit, deve-se padronizar as estimativas pontuais do modelo logit dividindo-as por $\pi/\sqrt{3} \cong 1{,}8138$ (GUJARATI; PORTER, 2011, RODRÍGUEZ, 2007).

2.2.3.3. Modelo complemento log-log

O modelo complemento log-log, por sua vez, é baseada na suposição de que o erro (ou a variável latente) tem distribuição de valor extremo (POWERS; XIE, 1999), também chamada de distribuição de Gompertz, que possui média γ (chamada constante de Euler) e variância $\frac{\pi^2}{6}$. A fda da distribuição de Gompertz, no ponto η_i, é dada por:

$$F(\eta_i) = P(\varepsilon_i \leq \eta_i) = \ 1 - \ e^{-e^{\eta_i}}$$

Analogamente, a probabilidade de sucesso da *i*-ésima unidade é determinada calculando o valor da fda da distribuição de Gompertz no ponto $\eta_i = \boldsymbol{x}_i^t\boldsymbol{\beta}$, como indicado abaixo:

$$p_i = \ F(\boldsymbol{x}_i^t\boldsymbol{\beta}) = 1 - \ e^{-e^{x_i^t\boldsymbol{\beta}}}$$

Aplicando a função inversa, pode-se explicitar o preditor linear $\boldsymbol{x}_i^t\boldsymbol{\beta}$ do modelo, obtendo a seguinte expressão algébrica:

$$\boldsymbol{x}_i^t\boldsymbol{\beta} = F^{-1}(p_i) = \ \ln\left[-\ln\left(1 - p_i\right)\right]$$

Desse modo, apreende-se que o modelo é linear na inversa da função de distribuição acumulada, que é dada pelo logaritmo do negativo do logaritmo do complemento da probabilidade de sucesso p_i, ou seja, $\ln\left[-\ln\left(1 - p_i\right)\right] = \boldsymbol{x}_i^t\boldsymbol{\beta}$. Essa

equação representa o modelo apresentado em 2.2.2.3. (modelo complemento log-log**)**, onde a função de ligação é o complemento log-log.

Cabe mencionar ainda que as estimativas pontuais dos parâmetros do modelo complemento log-log somente podem ser comparadas com as respectivas estimativas dos modelos probit e logit após a aplicação da padronização de suas estimativas pontuais. Então, para comparar com as estimativas pontuais dos parâmetros do modelo probit, deve-se padronizar as estimativas pontuais do modelo complemento log-log dividindo-as por $\pi/\sqrt{6} \cong 1{,}2825$. Após a padronização das estimativas do modelo complemento log-log e do modelo logit para comparar com as estimativas do modelo probit, elas também podem ser comparadas entre si (complemento log-log *versus* logit) (RODRÍGUEZ, 2007).

2.3. Estimação dos parâmetros do modelo

Dentre os diversos métodos de estimação para os parâmetros de um MLG, o de Máxima Verossimilhança (MV) é o mais utilizado (CORDEIRO; DEMÉTRIO, 2013). No software R, o método de MV está implementado na função *glm (generalised linear model)*. Entretanto, a utilização desse método de estimação tem como hipótese que as observações amostrais $y_1, y_2, \ldots, y_n$ sejam independentes e identicamente distribuídas, o que não é válido para uma pesquisa com plano amostral complexo, como o da PNS. Em casos como esse, uma alternativa é realizar a estimação dos parâmetros de um MLG pelo método de Máxima Pseudo-Verossimilhança (MPV), que permite incorporar as informações do plano amostral, como pesos das unidades da amostra, estratos e conglomerados. No software R isso se dá pela função *svyglm*, do pacote *survey*.

2.3.1. Método de Máxima Pseudo-Verossimilhança para modelo binário

Seja uma população finita U de tamanho N, tal que para cada elemento i dessa população esteja associada uma variável aleatória Y_i com distribuição de Bernoulli com parâmetro p_i. Suponha ainda que cada Y_i seja independente das demais, e que gere a observação y_i. Então, o logaritmo da função de verossimilhança populacional é:

$$\ln L_U(\boldsymbol{\beta}) = \ln \prod_{i=1}^{N} P(Y_i = y_i, p_i) = \ln \prod_{i=1}^{N} {p_i}^{y_i} (1 - p_i)^{1-y_i}$$

$$= \sum_{i=1}^{N} \ln \left[{p_i}^{y_i} (1 - p_i)^{1-y_i} \right] \qquad (1.8)$$

Derivando (1.8) em função de $\boldsymbol{\beta}$, obtém-se $\boldsymbol{u}(\boldsymbol{\beta})$, o vetor de escores na população, de dimensão $p \times 1$:

$$\boldsymbol{u}(\boldsymbol{\beta}) = \frac{\partial \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\beta}} \ln \left[{p_i}^{y_i} (1 - p_i)^{1-y_i} \right] = \sum_{i=1}^{N} \boldsymbol{u}_i(\boldsymbol{\beta}) \qquad (1.9)$$

onde $\boldsymbol{u}_i(\boldsymbol{\beta})$, de dimensão $p \times 1$, é o escore do i-ésimo elemento populacional, $i = 1, 2, \ldots, N$. A expressão deste escore, que é descrita na seção 2.3.2, depende da função de ligação utilizada.

Caso todos os elementos da população U fossem conhecidos, seria possível obter $\boldsymbol{\beta_{MV}}$, o estimador de Máxima Verossimilhança de $\boldsymbol{\beta}$, ao resolver o sistema de equações $\boldsymbol{u}(\boldsymbol{\beta}) = \mathbf{0}$ através de métodos iterativos, como Newton-Raphson ou Escore de Fisher (HEERINGA; WEST; BERGLUND, 2010). Pode-se considerar $\boldsymbol{\beta_{MV}}$, que é um pseudo-parâmetro, como uma Quantidade Descritiva Populacional Correspondente (QDPC) a $\boldsymbol{\beta}$. Embora não seja calculável sem a realização de um censo, essa QDPC pode ser estimada, e será o alvo da inferência. Isso porque, sob certas condições de regularidade, um estimador adequado para $\boldsymbol{\beta_{MV}}$ também deverá estimar $\boldsymbol{\beta}$ adequadamente, o que geralmente é satisfeito para amostras grandes, como é o caso da PNS (PESSOA; SILVA, 1998).

Em (1.9) foi visto que $\boldsymbol{u}(\boldsymbol{\beta})$ é a soma dos escores da população, o que o faz também um vetor de totais populacionais, que pode ser estimado através do estimador linear ponderado $\hat{\boldsymbol{u}}(\boldsymbol{\beta})$:

$$\hat{\boldsymbol{u}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i \boldsymbol{u}_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i \frac{\partial}{\partial \boldsymbol{\beta}} \ln P(Y_i = y_i, p_i) = \frac{\partial}{\partial \boldsymbol{\beta}} \ln L_S(\boldsymbol{\beta}) \qquad (1.10)$$

onde n é o tamanho da amostra, w_i é o peso amostral de cada unidade da amostra e $\ln L_S(\boldsymbol{\beta})$ é o logaritmo da função de pseudo-verossimilhança (ponderada), que é definida como a soma dos produtos dos pesos amostrais pelos logaritmos da função de probabilidade referentes às unidades da amostra (LUMLEY, 2017):

$$\ln L_S(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i \ln P(Y_i = y_i, p_i)$$

Segundo Lumley (2017), o logaritmo da função de pseudo-verossimilhança (ponderada), denotada no presente trabalho por $\ln L_S(\boldsymbol{\beta})$ é um estimador não viciado para o logaritmo da função de verossimilhança populacional $\ln L_U(\boldsymbol{\beta})$, isto é, $E[\ln L_S(\boldsymbol{\beta})] = \ln L_U(\boldsymbol{\beta})$.

Assim, o estimador de Máxima Pseudo-Verossimilhança para $\boldsymbol{\beta}$ será obtido ao resolver o sistema de equações de pseudo-verossimilhança:

$$\widehat{\boldsymbol{u}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i \boldsymbol{u}_i(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \ln L_S(\boldsymbol{\beta}) = \mathbf{0} \qquad (1.11)$$

A solução desse sistema é o estimador de MPV de $\boldsymbol{\beta}$, denotado por $\widehat{\boldsymbol{\beta}}_{\boldsymbol{MPV}}$.

Utilizando a técnica de linearização de Taylor, também é possível obter um estimador consistente para a variância assintótica de $\widehat{\boldsymbol{\beta}}_{\boldsymbol{MPV}}$. Assim, a matriz estimada de variância-covariância dos estimadores de MPV dos parâmetros do modelo é apresentada a seguir (PESSOA et al., 1998):

$$\widehat{\boldsymbol{VAR}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{MPV}}) \cong \left[\widehat{\mathfrak{J}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{MPV}})\right]^{-1} \boldsymbol{VAR}\left[\sum_{i=1}^{n} w_i \boldsymbol{u}_i(\widehat{\boldsymbol{\beta}}_{\boldsymbol{MPV}})\right] \left[\widehat{\mathfrak{J}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{MPV}})\right]^{-1} \qquad (1.12)$$

onde, $\widehat{\mathfrak{J}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{MPV}})$ é uma matriz simétrica de derivadas segundas, de dimensão $p \times p$, dada por:

$$\widehat{\mathfrak{I}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{MPV}}) = \left.\frac{\partial \widehat{\boldsymbol{u}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{\boldsymbol{MPV}}} = \sum_{i=1}^{n} w_i \left.\frac{\partial \boldsymbol{u}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{\boldsymbol{MPV}}} \quad (1.13)$$

e $\boldsymbol{VAR}\left[\sum_{i=1}^{n} w_i \boldsymbol{u}_i(\widehat{\boldsymbol{\beta}}_{\boldsymbol{MPV}})\right]$ é a matriz estimada de variância-covariância dos totais amostrais dos escores ponderados.

Sabe-se também que a distribuição assintótica do estimador $\widehat{\boldsymbol{\beta}}_{\boldsymbol{MPV}}$ é Normal Multivariada (BINDER, 1983):

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{MPV}} \sim NM\left[\boldsymbol{\beta}, \widehat{\boldsymbol{VAR}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{MPV}})\right] \quad (1.14)$$

2.3.2 Expressões dos escores para diferentes funções de ligação

Na seção anterior, em (1.9), foi apresentado $\boldsymbol{u}_i(\boldsymbol{\beta})$, o escore do *i*-ésimo elemento da população. Entretanto, considerou-se a probabilidade de sucesso (p_i) de maneira genérica, e na presente seção será agregada a informação referente à probabilidade de sucesso específica a cada uma das funções de ligação apresentadas em 2.2.2.

2.3.2.1. Função logit

Na seção 2.2.2.1, mostrou-se que um MLG com desfecho binário e função de ligação logit possui $p_i = \frac{e^{x_i^t \boldsymbol{\beta}}}{1+ e^{x_i^t \boldsymbol{\beta}}}$. Portanto, o escore do *i*-ésimo elemento populacional é dado por:

$$\begin{aligned}
\boldsymbol{u}_i(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}} \ln[{p_i}^{y_i}\,(1-p_i)^{1-y_i}] \\
&= \frac{\partial}{\partial \boldsymbol{\beta}}\,[y_i \ln p_i + \ln(1-p_i) - \, y_i \ln(1-p_i)] \\
&= \frac{\partial}{\partial \boldsymbol{\beta}}\left[y_i \ln\left(\frac{p_i}{1-p_i}\right) + \ln(1-p_i)\right] \\
&= \frac{\partial}{\partial \boldsymbol{\beta}}\left[y_i \boldsymbol{x}_i^t\,\boldsymbol{\beta} - \ln\left(1 + \, e^{\boldsymbol{x}_i^t \boldsymbol{\beta}}\right)\right] \\
&= y_i \boldsymbol{x}_i^t - \frac{x_i^t e^{\boldsymbol{x}_i^t \boldsymbol{\beta}}}{1 + \, e^{\boldsymbol{x}_i^t \boldsymbol{\beta}}}
\end{aligned}$$

$$= \boldsymbol{x}_i^t[y_i - p_i]$$

2.3.2.2 Função probit

Como visto na seção 2.2.2.2, um MLG com desfecho binário e função de ligação probit possui $p_i = F_Z(\boldsymbol{x}_i^t\,\boldsymbol{\beta})$. Portanto, pode-se demonstrar que:

$$\begin{aligned}
\boldsymbol{u}_i(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}} \ln[{p_i}^{y_i}\,(1-p_i)^{1-y_i}] \\
&= \frac{\partial}{\partial \boldsymbol{\beta}} \left[y_i \ln F_Z(\boldsymbol{x}_i^t\,\boldsymbol{\beta}) + (1-y_i)\ln\left(1 - F_Z(\boldsymbol{x}_i^t\,\boldsymbol{\beta})\right)\right] \\
&= y_i\,\frac{1}{F_Z(\boldsymbol{x}_i^t\,\boldsymbol{\beta})}\,f_Z(\boldsymbol{x}_i^t\,\boldsymbol{\beta}) - (1-y_i)\,\frac{1}{F_Z(\boldsymbol{x}_i^t\,\boldsymbol{\beta})}\,f_Z(\boldsymbol{x}_i^t\,\boldsymbol{\beta})\,\boldsymbol{x}_i^t \\
&= \left[\frac{y_i\,f_Z(\boldsymbol{x}_i^t\,\boldsymbol{\beta})\left(1 - F_Z(\boldsymbol{x}_i^t\,\boldsymbol{\beta})\right) - (1-y_i)\,f_Z(\boldsymbol{x}_i^t\,\boldsymbol{\beta})\,F_Z(\boldsymbol{x}_i^t\,\boldsymbol{\beta})}{F_Z(\boldsymbol{x}_i^t\,\boldsymbol{\beta})\left(1 - F_Z(\boldsymbol{x}_i^t\,\boldsymbol{\beta})\right)}\right]\boldsymbol{x}_i^t \\
&= \left[\frac{y_i\,f_Z(\boldsymbol{x}_i^t\,\boldsymbol{\beta}) - f_Z(\boldsymbol{x}_i^t\,\boldsymbol{\beta})\,F_Z(\boldsymbol{x}_i^t\,\boldsymbol{\beta})}{F_Z(\boldsymbol{x}_i^t\,\boldsymbol{\beta})\left(1 - F_Z(\boldsymbol{x}_i^t\,\boldsymbol{\beta})\right)}\right]\boldsymbol{x}_i^t = \\
&= \left[\frac{(y_i - p_i)\,f_Z(\boldsymbol{x}_i^t\,\boldsymbol{\beta})}{p_i(1-p_i)}\right]\boldsymbol{x}_i^t
\end{aligned}$$

2.3.2.3. Função complemento log-log

Como indicado na seção 2.2.2.3, num MLG com desfecho binário e função de ligação complemento log-log, a probabilidade de sucesso é dada pela seguinte expressão: $p_i = 1 - e^{-e^{x_i^t\,\beta}}$. Dessa forma, o escore do *i*-ésimo elemento populacional é dado por:

$$\begin{aligned}
\boldsymbol{u}_i(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}} \ln[{p_i}^{y_i}\,(1-p_i)^{1-y_i}] \\
&= \frac{\partial}{\partial \boldsymbol{\beta}} \ln\left\{\left[1 - e^{-e^{x_i^t\,\beta}}\right]^{y_i}\left(1 - \left[1 - e^{-e^{x_i^t\,\beta}}\right]\right)^{1-y_i}\right\} \\
&= \frac{y_i\boldsymbol{x}_i^t e^{x_i^t\,\boldsymbol{\beta} - e^{x_i^t\,\boldsymbol{\beta}}}}{1 - e^{-e^{x_i^t\,\boldsymbol{\beta}}}} + (1-y_i)\left(-e^{x_i^t\,\boldsymbol{\beta}}\right)
\end{aligned}$$

$$= \frac{\boldsymbol{x}_i^t e^{\boldsymbol{x}_i^t \boldsymbol{\beta}} \left(y_i - 1 + e^{-e^{\boldsymbol{x}_i^t \boldsymbol{\beta}}} \right)}{1 - e^{-e^{\boldsymbol{x}_i^t \boldsymbol{\beta}}}}$$

$$= \frac{\boldsymbol{x}_i^t e^{\boldsymbol{x}_i^t \boldsymbol{\beta}} (y_i - p_i)}{p_i}$$

2.3.3. Inferência sobre os parâmetros do modelo

Uma vez obtidas as estimativas pontuais dos parâmetros do modelo e a matriz de variância-covariância estimada, o teste de Wald de significância individual pode ser utilizado para verificar se a relação entre cada variável explicativa e o desfecho do modelo é estatisticamente significante. Esse teste se baseia no seguinte resultado, válido para grandes amostras (KUTNER et al., 2005):

$$Z = \frac{\widehat{\beta_j} - \beta_j}{\sqrt{V\hat{A}R\left(\widehat{\beta_j}\right)}} \sim N(0,1), \quad \forall\, j = 1, 2, \dots, p \tag{1.15}$$

onde β_j é o efeito da *j*-ésima variável explicativa (covariável ou variável *dummy*), $\widehat{\beta_j}$ é a estimativa por MPV de β_j, $\sqrt{V\hat{A}R\left(\widehat{\beta_j}\right)}$ é o desvio padrão (ou erro padrão) estimado de $\widehat{\beta_j}$, e Z é uma estatística com distribuição normal padrão.

Alguns pacotes estatísticos de análise de dados amostrais complexos, diferentemente do programa R, utilizam como alternativa à estatística Z a estatística W, apresentada abaixo:

$$W = \frac{\left(\widehat{\beta_j} - \beta_j\right)^2}{V\hat{A}R\left(\widehat{\beta_j}\right)} \sim \chi_1^2, \quad \forall\, j = 1, 2, \dots, p \tag{1.16}$$

O teste de Wald de significância individual possui as seguintes hipóteses, nula e alternativa, respectivamente:

$$\begin{cases} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0 \end{cases}$$

A estatística de teste é a mesma apresentada em (1.15), e seu valor observado, sob H_0, é denotado por:

$$z_{obs} = \frac{\widehat{\beta}_j}{\sqrt{V\hat{A}R(\widehat{\beta}_j)}}$$

A região crítica é $RC = \left\{z \in \mathbb{R} \,/\, |z| > z_{\left(1-\frac{\alpha}{2}\right)}\right\}$, onde α é o nível de significância adotado para o teste, e $z_{\left(1-\frac{\alpha}{2}\right)}$ é o valor crítico da distribuição normal padrão (ou reduzida) no percentil $\left(1-\frac{\alpha}{2}\right)$.

O critério de decisão é rejeitar H_0 ao nível de significância α, se $z_{obs} \in RC$. Quando isso acontece, diz-se que existe relação estatisticamente significante entre a *j*-ésima variável explicativa e o desfecho do modelo. Já se $z_{obs} \notin RC$, não há evidências para se rejeitar H_0, e portanto o entendimento é contrário ao anteriormente exposto.

Outro critério de decisão, que leva à mesma conclusão obtida a partir do critério descrito anteriormente, é verificar se o p-valor do teste é menor que o α adotado. Em caso afirmativo, rejeita-se H_0 ao nível de significância α. Caso contrário, não há evidências para se rejeitar H_0 ao nível de significância adotado.

Além do teste de Wald de significância individual, pode ser empregado o teste de Wald de significância geral para avaliar se múltiplos parâmetros do modelo são iguais a zero. Como apontado por Heeringa, West e Berglund (2010), através do teste de Wald geral pode-se avaliar a significância do efeito de variáveis qualitativas com várias categorias em modelos lineares generalizados. Para ilustrar o teste de Wald geral (LUMLEY; SCOTT, 2017, CHAMBERS; SKINNER, 2003), suponha que $\boldsymbol{\beta}$, de dimensão $p \times 1$, é particionado em $(\boldsymbol{\beta_1}, \boldsymbol{\beta_2})$, onde $\boldsymbol{\beta_1}$ tem dimensão $p_1 \times 1$ e $\boldsymbol{\beta_2}$ tem dimensão $p_2 \times 1$, tal que $p_1 + p_2 = p$. Se o objetivo é testar as seguintes hipóteses sobre $\boldsymbol{\beta_1}$:

$$\begin{cases} H_0: \boldsymbol{\beta_1} = \mathbf{0} \\ H_1: \boldsymbol{\beta_1} \neq \mathbf{0} \end{cases}$$

A estatística de teste, denotada por W, é dada por:

$$W = (\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}})^{t}\left[\widehat{\boldsymbol{VAR}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}})\right]^{-\mathbf{1}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}})$$

Onde $W \sim \chi^2_{p_1}$ assintoticamente sob a hipótese nula H_0, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}$, de dimensão $p_1 \times 1$, é o estimador de MPV de $\boldsymbol{\beta}_{\boldsymbol{1}}$, e $\widehat{\boldsymbol{VAR}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}})$ é o estimador da matriz de variância-covariância assintótica de $\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}$, de dimensão $p_1 \times p_1$. Ao invés de W, pode-se também utilizar a estatística aproximada a seguir (LUMLEY; SCOTT, 2017):

$$F_W = \frac{W}{p_1}$$

em que $F_W \sim F_{p_{1,m}}$, onde $F_{p_{1,m}}$ representa a distribuição F de Snedecor com p_1 graus de liberdade no numerador e m graus de liberdade no denominador, sendo m chamado "graus de liberdade do plano amostral", que é o número de UPAs menos o número de estratos.

A região crítica do teste baseado na estatística F_W é: $RC = \left\{f \in \mathbb{R} \,/\, f < f_{\alpha/2,p_1,m} \;\; ou \;\; f > f_{1-\alpha/2,p_1,m}\right\}$, onde α é o nível de significância adotado para o teste, $f_{\alpha/2,p_1,m}$ é o valor crítico da distribuição $F_{p_{1,m}}$ no percentil $\frac{\alpha}{2}$, e $f_{1-\alpha/2,p_1,m}$ é o valor crítico da distribuição $F_{p_{1,m}}$ no percentil $1 - \frac{\alpha}{2}$.

O critério de decisão é rejeitar H_0 ao nível de significância α, se $f_{obs} \in RC$. Quando isso acontece, diz-se que existe relação estatisticamente significante entre a variável qualitativa com múltiplas categorias (cujos parâmetros compõem o vetor $\boldsymbol{\beta}_{\boldsymbol{1}}$) e o desfecho do modelo. Já se $f_{obs} \notin RC$, diz-se não haver evidências para se rejeitar H_0, isto é, não existe associação significativa entre a variável qualitativa (com múltiplas categorias) e o desfecho.

2.3.4. Medidas da capacidade discriminatória do modelo

Uma vez ajustado o modelo binário, pode-se obter a probabilidade estimada de sucesso do i-ésimo elemento da amostra, $\forall\, i = 1, 2, \ldots, n$, o que permite avaliar a capacidade discriminatória do modelo. Para isso, estipula-se um ponto de corte γ — geralmente $\gamma = 0{,}5$, embora esse valor seja arbitrário — de maneira que se $\widehat{p}_i$ for maior ou igual que o ponto de corte, então $\widehat{Y}_i = 1$, e, caso contrário, $\widehat{Y}_i = 0$. Outra possibilidade

é utilizar o ponto de corte ótimo, aquele para o qual a proporção de predições incorretas realizadas pelo modelo é mínima (KUTNER *et al*., 2005).

Uma vez obtidas as predições de sucesso e fracasso para cada elemento da amostra, pode-se classificá-las em razão de seu acerto ou erro em relação ao desfecho observado. Dessa forma, é possível construir um quadro que apresenta em cada célula a frequência de elementos na amostra, de acordo com as categorias observadas e preditas do desfecho.

Quadro 1: Frequência dos elementos amostrais de acordo com as categorias observadas e preditas por MLG com desfecho binário.

Categorias Preditas	Categorias observadas		Total
	Sucesso ($Y = 1$)	Fracasso ($Y = 0$)	
Sucesso ($\hat{Y} = 1$)	f_{11} (Verdadeiros Positivos)	f_{12} (Falsos Positivos)	$f_{11} + f_{12}$
Fracasso ($\hat{Y} = 0$)	f_{21} (Falsos Negativos)	f_{22} (Verdadeiros Negativos)	$f_{21} + f_{22}$
Total	$f_{11} + f_{21}$	$f_{12} + f_{22}$	$\sum_{i=1}^{2}\sum_{j=1}^{2} f_{ij}$

O Quadro 1 auxilia na obtenção de medidas quantitativas da capacidade discriminatória do modelo, como a taxa global de acertos, a sensibilidade e a especificidade.

A taxa global de acertos (TG) é a razão entre o total de elementos em que o resultado predito é o mesmo que observado — ou seja, a quantidade de acertos — e o total de elementos. Seu resultado é multiplicado por 100, para que se forneça o percentual de acertos do modelo:

$$TG = \left(\frac{f_{11} + f_{22}}{\sum_{i=1}^{2}\sum_{j=1}^{2} f_{ij}}\right) \cdot 100$$

A sensibilidade (S) é a probabilidade do modelo ajustado predizer o evento "sucesso" para um elemento, dado que seu desfecho observado é de fato "sucesso". Portanto,

$$S = P\left[\hat{Y} = 1 \middle| Y = 1\right] = \left(\frac{f_{11}}{f_{11} + f_{21}}\right) \cdot 100$$

A especificidade (E) é a probabilidade do modelo predizer o evento "fracasso" para um elemento, dado que seu desfecho observado é de fato "fracasso". Portanto,

$$E = P\left[\hat{Y} = 0 \middle| Y = 0\right] = \left(\frac{f_{22}}{f_{12} + f_{22}}\right) \cdot 100$$

Um modelo é classificado como *bom* segundo sua capacidade preditiva, se tanto sua *sensibilidade* quanto sua *especificidade* forem maiores que 80%. Analogamente, é dito *razoável* se S e E estiverem entre 50% e 80%, e, abaixo disso, a capacidade preditiva do modelo é dita *medíocre* (MARÔCO, 2010).

Outra medida da qualidade da predição de um modelo é a área sob a curva ROC (*Receiver Operating Characteristic*). Essa área, também chamada AUC — do inglês, "*Area Under the (ROC) Curve*" — varia entre 0 e 1, e quanto mais próxima de 1, maior é a capacidade do modelo em distinguir corretamente os sucessos e os fracassos. A curva ROC é obtida ao plotar no gráfico pares de sensibilidade e complemento da especificidade ($\bar{E}$), onde $\bar{E} = 1 - E$, e cada par é resultado da escolha de um ponto de corte diferente. Segundo Marôco (2010), a área sob a curva ROC, denotada neste trabalho por A, indica a capacidade discriminatória do modelo da seguinte maneira (Quadro 2):

Quadro 2: Classificação da capacidade preditiva de um MLG para desfecho binário, segundo a área sob a curva ROC.

Área sob a curva ROC (A)	**Capacidade discriminatória do modelo**
$A = 0.5$	Sem poder discriminatório
$0{,}5 < A < 0{,}7$	Discriminação ruim
$0{,}7 \leq A < 0{,}8$	Discriminação aceitável
$0{,}8 \leq A < 0{,}9$	Discriminação boa
$A \geq 0{,}9$	Discriminação excepcional

Fonte: Adaptado de Marôco (2010).

2.3.5. Medidas de Pseudo-R^2

Em modelos clássicos de regressão linear, o coeficiente de determinação (R^2) é uma medida, que varia entre 0 e 1, utilizada para avaliar a qualidade do ajuste de um modelo estimado em geral através do método de mínimos quadrados, e é interpretada como a proporção da variação total dos dados que é explicada pelo modelo ajustado (DOBSON; BARNETT, 2018), ou melhor, como o percentual de variação da variável dependente (desfecho) que é explicada pelas variáveis independentes.

No caso de modelos lineares generalizados para desfecho binário, não é possível calcular o R^2, uma vez que "*a variância da variável dependente, depende da probabilidade em que ocorrem os seus valores*" (MARÔCO, 2010, p. 803)". Em contrapartida, a qualidade do ajuste pode ser avaliada através da utilização das medidas de pseudo-R^2, em que maiores valores de pseudo-R^2 indicam melhore ajustes. Entretanto, essas medidas não são interpretáveis separadamente, e a comparação entre modelos só é significativa se realizada utilizando a mesma medida de pseudo-R^2, o mesmo conjunto de dados e as mesmas variáveis explicativas (GUO; FRASER, 2014).

O pseudo-R^2 de Cox-Snell é uma das medidas de qualidade do ajuste, que indica o ganho de informação do modelo sob consideração (aquele com variáveis explicativas) em comparação ao modelo nulo (modelo considerando apenas a constante). Algumas adaptações à estatística original foram propostas por Lumley (2017) para se obter o R^2 de Cox-Snell para utilização em MLGs ajustados a partir de amostras provenientes de planos amostrais complexos, e seu cálculo é dado por:

$$R^2_{CS} = 1 - e^{2\left[\frac{\ln L_S(\widehat{\boldsymbol{\beta}}^{(\mathbf{0})}) - \ln L_S(\widehat{\boldsymbol{\beta}})}{\sum_{i=1}^{n} w_i}\right]}$$

onde $\ln L_S\left(\widehat{\boldsymbol{\beta}}^{(\mathbf{0})}\right)$ é o logaritmo da função de pseudo-verossimilhança do modelo nulo, $\ln L_S\left(\widehat{\boldsymbol{\beta}}\right)$ é o logaritmo da função de pseudo-verossimilhança do modelo sob consideração e w_i é o peso amostral do *i*-ésimo indivíduo.

Da expressão anterior é possível perceber que R^2_{CS} nunca assume valor 1, que indicaria o caso em que o modelo sob consideração fornece um ajuste perfeito. O mesmo acontece com o R^2 de Cox-Snell original, e, portanto, visando obter uma medida de qualidade do ajuste que pudesse assumir o valor 1, Nagelkerke propôs sua versão dessa

estatística (MARÔCO, 2010), cuja adaptação para que o plano amostral complexo seja considerado também foi proposta por Lumley (2017):

$$R_N^2 = \frac{R_{CS}^2}{1 - e^{\frac{2 \ln L_S(\hat{\boldsymbol{\beta}}^{(\mathbf{0})})}{\sum_{i=1}^{n} w_i}}}$$

2.4. Variáveis utilizadas no estudo

2.4.1. Desfecho de realização de consulta médica

A variável escolhida como desfecho dos modelos no presente estudo foi originada a partir do questionamento "*Quando consultou um médico pela última vez*?", pertencente ao módulo J ("Utilização de Serviços de Saúde") do questionário da PNS 2013. A pergunta possui cinco alternativas de respostas possíveis, a saber: 1) "Nos doze últimos meses", 2) "De 1 ano a menos de 2 anos", 3) "De 2 anos a menos de 3 anos", 4) "3 anos ou mais" e 5) "nunca foi ao médico".

No presente trabalho, utilizou-se um desfecho binário obtido a partir da agregação dessas alternativas de resposta. A categoria 1 foi renomeada para "há menos de 1 ano", enquanto as categorias de 2 a 5 foram reclassificadas como "nunca foi ao médico ou foi há 1 ano ou mais".

2.4.2. Variáveis sociodemográficas e de saúde

Para compor a listagem de variáveis explicativas, procurou-se verificar quais variáveis têm sido utilizadas para explicar o desfecho "consulta ao médico" em alguns estudos publicados nesta temática (MOREIRA; MORAES; LUIZ, 2011, CAPILHEIRA; SANTOS, 2006). A seguir são apresentadas as variáveis presentes no questionário da PNS que serão utilizadas como variáveis explicativas dos modelos ajustados.

Quadro 3: Listagem das variáveis explicativas dos modelos binários, e suas categorias.

Variáveis explicativas	Categorias
Sexo	Masculino, Feminino
Faixa etária	60 a 69 anos, 70 anos ou mais
Cor / raça	Branca, Não branca
Situação conjugal	Possui cônjuge, Não possui cônjuge
Escolaridade	Sem instrução ou com no máximo o ensino fundamental incompleto, Com ensino fundamental ou médio completos, Com ensino superior completo
Posse de plano de saúde (médico ou odontológico)	Sim, Não
Região de residência	Norte, Nordeste, Centro-Oeste, Sul, Sudeste
Autoavaliação de saúde	Boa, Regular, Ruim
Área de localização do domicílio	Urbana, Rural
Diagnóstico de alguma doença crônica, física ou mental, ou doença de longa duração (de mais de 6 meses de duração)	Sim, Não
Grau de dificuldade em ir ao médico sozinho(a)?	Não consegue ou tem grande dificuldade, Tem pequena dificuldade ou não tem dificuldade

2.5. População de estudo

A população de estudo é composta por idosos de 60 anos ou mais de idade que residem em domicílios particulares permanentes. Foram excluídos os idosos que não declararam informação sobre a sua última consulta médica e/ou que não informaram sobre quaisquer das características sociodemográficas ou de saúde consideradas na análise.

3. Resultados

No presente trabalho, foi observado que 16,5% dos idosos de 60 anos ou mais de idade nunca consultaram médico ou consultaram o médico pela última vez há 1 ano ou mais. Os demais idosos (83,5%) informaram ter consultado o médico pela última vez há menos de 1 ano.

A Tabela 1 apresenta a distribuição percentual (%) dos idosos segundo as suas características sociodemográficas e de saúde, bem como a distribuição conjunta dos idosos segundo o desfecho binário referente ao tempo da última consulta médica realizada pelos idosos, para cada uma das características mencionadas.

Tabela 1: Distribuição percentual de idosos com 60 anos ou mais, por suas características sociodemográficas e de saúde, segundo o tempo da última consulta médica realizada pelos idosos.

Variáveis Explicativas	Percentual de idosos (%)	Tempo da última consulta médica	
		Menos de 1 ano	Há 1 ano ou mais ou nunca se consultou
Sexo			
Masculino	43,6	77,8	22,2
Feminino	56,4	87,9	12,1
Faixa etária			
60 a 69 anos	43,9	81,9	18,1
70 anos ou mais	56,1	85,5	14,5
Escolaridade			
Sem inst./Fund. incomp.	71,1	82,4	17,6
Fund. ou médio comp.	20,3	84,5	15,5
Superior completo	8,7	89,8	10,2
Plano de saúde			
Sim	30,8	91,4	8,6
Não	69,2	80,0	20,0
Região			
Sul	15,1	87,6	12,4
Sudeste	47,9	86,3	13,7
Centro-Oeste	6,4	83,6	16,4
Nordeste	25,2	77,4	22,6
Norte	5,4	75,7	24,3
Autoaval. de saúde			
Bom	45,5	77,8	22,2
Regular	42,1	87,2	12,8
Ruim	12,4	92,0	8,0
Área			
Urbana	85,3	85,2	14,8
Rural	14,7	73,6	26,4
Diag. de doenças*			
Sim	39,5	93,0	7,0
Não	60,5	77,3	22,7
Dificuldade**			
Não cons./Grande dif.	13,3	90,3	9,7
Peq. dif./ Não tem dif.	86,7	82,4	17,6
Cor / raça			
Branca	53,6	86,0	14,0
Não branca	46,4	80,6	19,4
Situação conjugal			
Possui cônjuge	57	82,4	17,6
Não possui cônjuge	43	84,9	15,1

*Diagnóstico de doença crônica, física, mental ou doença de longa duração.
**Dificuldade em ir ao médico sozinho(a).
Amostra não expandida: 23811. Amostra expandida: 26404585.

Na Tabela 1, observa-se que o sexo feminino compõe 56,4% da população idosa brasileira, e que as mulheres tendem a evitar uma demora maior entre consultas, visto que apenas 12,1% delas se consultaram pela última vez há 1 ano ou mais, contra 22,2% entre os homens.

A divisão por faixa etária mostra que 56,1% dos idosos possuem 70 anos ou mais, e que 14,5% desse grupo consultou-se há 1 ano ou mais, contra 18,1% do outro grupo, o de idosos entre 60 e 69 anos.

Com relação aos níveis de escolaridade dos idosos, observa-se que a maioria (71,1%) dos idosos não possuem instrução ou possuem no máximo o ensino fundamental incompleto. Além disso, observa-se um gradiente decrescente entre o nível de escolaridade e o percentual de idosos que nunca realizaram consulta médica ou consultaram o médico há 1 ou mais, isto é, há uma tendência a esperar menos tempo para se consultar conforme a escolaridade aumenta. Apenas 10,2% dos idosos mais escolarizados nunca consultaram o médico ou realizaram consulta médica há 1 ano ou mais, enquanto entre os idosos menos escolarizados o percentual dos que nunca consultaram o médico, ou que o procuraram há 1 ano ou mais, foi de 17,6%.

Quanto ao plano de saúde, vê-se que apenas 30,8% dos idosos podem utilizar-se desse benefício, e que a posse do plano parece reduzir o tempo de realização de consultas médicas, visto que o percentual de idosos que nunca consultaram o médico ou que se consultaram há 1 ano ou mais foi de apenas 8,6% entre os que possuem plano de saúde, enquanto esse percentual é de 20,0% entre os idosos que não possuem plano de saúde.

A distribuição entre as grandes regiões brasileiras mostra que quase metade dos idosos brasileiros se encontra no Sudeste (47,9%), 25,2% no Nordeste, 15,1% no Sul, 6,4% no Centro-Oeste e 5,4% no Norte. Quanto ao tempo desde a última consulta, as regiões Sul e Sudeste apresentaram as menores porcentagens de idosos que nunca consultaram médico ou que se consultaram há 1 ano ou mais (12,4% e 13,7%, respectivamente), seguidas da região Centro-Oeste, com 16,4%, e das regiões Norte e Nordeste com as maiores porcentagens (22,6 e 24,3%, respectivamente).

Com relação a autoavaliação de saúde 45,5% dos idosos classificaram sua saúde como boa, 42,1% como regular e apenas 12,4% como ruim. Observou-se um gradiente crescente entre o nível de autoavaliação de saúde e o percentual de idosos que nunca realizaram consulta médica ou que se consultaram há 1 ano ou mais, indicando maiores demoras na realização de consultas médicas associadas a melhores níveis de autoavaliação de saúde, visto que 22,2% dos idosos que reportaram saúde boa nunca se consultaram ou se consultaram pela última vez há 1 ano ou mais, contra 12,8% dos que consideram sua saúde regular, e apenas 8% dentre os que avaliaram a saúde como ruim.

Quanto à área de residência, tem-se que 85,3% dos idosos vivem em área urbana, e entre eles 14,8% nunca realizaram consulta médica ou realizaram sua última consulta médica há 1 ano ou mais. Já na área rural esse percentual é de 26,4%.

Os idosos que não possuem diagnóstico de doença crônica, física, mental ou doença de longa duração compõem 60,5% do total, e 22,7% deles nunca consultaram o médico ou se consultaram com um médico pela última vez há 1 ano ou mais. O percentual de idosos que nunca realizaram consulta médica ou consultaram o médico há 1 ou mais é menor (7%) entre aqueles que possuem algum diagnóstico de doença.

Pequena dificuldade, ou ausência dela, em ir ao médico sozinho foi relatada por 86,7% dos idosos. Destes, 17,6% nunca realizaram consulta médica ou realizaram há 1 ano ou mais. Entre os idosos que não conseguem ou tem grande dificuldade em ir ao médico sozinhos, 9,7% nunca se consultaram ou se consultaram há 1 ano ou mais.

A cor/raça de 53,6% dos idosos é branca, e 14% desses idosos nunca se consultaram ou se consultaram com o médico pela última vez há 1 ano ou mais. Já para ~~a~~ os idosos de cor/raça não branca, esse percentual foi de 19,4%.

Por fim, tem-se que, a propósito da situação conjugal, 57% dos idosos possuem cônjuge, e que 17,6% deles consultaram-se pela última vez há 1 ano ou mais ou nunca consultaram o médico. Dentre os que não possuem cônjuge, esse percentual de não consulta ou de consulta há 1 ano ou mais foi de 15,1%.

A Tabela 2 apresenta os resultados do ajuste dos modelos logit, probit e complemento log-log para o desfecho de estudo, considerando todas as variáveis explicativas. Entre os resultados, estão as estimativas pontuais dos parâmetros de cada modelo binário e os seus respectivos intervalos com 95% de confiança, bem como os p-valores do teste de Wald de significância individual e Geral.

Tabela 2: Resultados do ajuste dos modelos logit, probit e complemento log-log para o desfecho binário do tempo da última consulta médica realizada pelos idosos brasileiros, considerando todas as variáveis explicativas.

Variáveis Explicativas	Modelo logit			Modelo probit			Modelo complemento log-log		
	Est.	IC 95%	p-valor	Est.	IC 95%	p-valor	Est.	IC 95%	p-valor
Sexo									
Masculino	0,640	[0,53; 0,75]	<0,001	0,359	[0,30; 0,42]	<0,001	0,554	[0,46; 0,65]	<0,001
Feminino	1	-	-	1	-	-	1	-	-
Faixa etária									
60 a 69 anos	0,156	[0,04; 0,28]	0,011	0,086	[0,02; 0,15]	0,013	0,128	[0,02; 0,23]	0,016
70 anos ou mais	1	-	-	1	-	-	1	-	-
Escolaridade			0,029			0,032			0,034
Sem inst./Fund. incomp.	0,355	[0,09; 0,62]	0,008	0,190	[0,05; 0,33]	0,009	0,311	[0,08; 0,55]	0,010
Fund. ou médio comp.	0,321	[0,05; 0,59]	0,021	0,169	[0,02; 0,32]	0,024	0,291	[0,05; 0,54]	0,020
Superior completo	1	-	-	1	-	-	1	-	-
Plano de saúde									
Sim	1	-	-	1	-	-	1	-	-
Não	0,852	[0,67; 1,03]	<0,001	0,464	[0,37; 0,56]	<0,001	0,766	[0,60; 0,93]	<0,001
Região			<0,001			<0,001			<0,001
Sul	1	-	-	1	-	-	1	-	-
Sudeste	0,144	[-0,05; 0,34]	0,147	0,080	[-0,03; 0,19]	0,146	0,127	[-0,05; 0,30]	0,151
Centro-Oeste	0,354	[0,13; 0,58]	0,002	0,206	[0,08; 0,33]	0,001	0,302	[0,11; 0,50]	0,002
Nordeste	0,639	[0,45; 0,83]	<0,001	0,360	[0,25; 0,47]	<0,001	0,546	[0,38; 0,72]	<0,001
Norte	0,563	[0,33; 0,79]	<0,001	0,325	[0,20; 0,45]	<0,001	0,474	[0,28; 0,67]	<0,001
Autoaval. de saúde			<0,001			<0,001			<0,001
Bom	1,234	[1,03; 1,44]	<0,001	0,674	[0,57; 0,78]	<0,001	1,089	[0,90; 1,28]	<0,001
Regular	0,493	[0,29; 0,69]	<0,001	0,260	[0,16; 0,36]	<0,001	0,450	[0,27; 0,63]	<0,001
Ruim	1	-	-	1	-	-	1	-	-
Área									
Urbana	1	-	-	1	-	-	1	-	-
Rural	0,436	[0,30; 0,57]	<0,001	0,248	[0,17; 0,33]	<0,001	0,365	[0,25; 0,48]	<0,001
Diag. de doenças									
Sim	1	-	-	1	-	-	1	-	-
Não	1,059	[0,92; 1,20]	<0,001	0,563	[0,49; 0,64]	<0,001	0,968	[0,83; 1,10]	<0,001
Dificuldade									
Não cons./Gran. dif.	1	-	-	1	-	-	1	-	-
Peq. dif./ Não tem dif.	0,278	[0,09; 0,47]	0,004	0,147	[0,05; 0,25]	0,004	0,248	[0,08; 0,42]	0,005
Cor / raça									
Branca	1	-	-	1	-	-	1	-	-
Não branca	0,105	[-0,02; 0,23]	0,099	0,059	[-0,01; 0,13]	0,097	0,095	[-0,01; 0,20]	0,081
Situação conjugal									
Possui cônjuge	0,040	[-0,02; 0,10]	0,192	0,023	[-0,01; 0,06]	0,183	0,033	[-0,02; 0,09]	0,213
Não possui cônjuge	1	-	-	1	-	-	1	-	-

Est. = Estimativa pontual do parâmetro do modelo.

Com relação à significância estatística dos parâmetros, observa-se que em todos os modelos binários (logit, probit e complemento log-log) as variáveis “cor/raça” e “situação conjugal” não apresentaram relação estatisticamente significante com a probabilidade do idoso nunca ter ido ao médico ou ter ido há 1 ano ou mais (p-valor>0,05). Todas as demais variáveis apresentaram associação significativa e de mesmo sentido com a probabilidade do evento de interesse, considerando o nível de significância de 5% (Tabela 2).

A Tabela 3 apresenta os resultados do ajuste dos modelos logit, probit e complemento log-log para o desfecho referente ao tempo da última consulta médica realizada pelos idosos, considerando as nove variáveis restantes.

Tabela 3: Resultados do ajuste dos modelos logit, probit e complemento log-log para o desfecho binário do tempo da última consulta médica realizada pelos idosos brasileiros, considerando somente as variáveis explicativas selecionadas.

Variáveis Explicativas	Modelo logit			Modelo probit			Modelo complemento log-log		
	Est.	IC 95%	p-valor	Est.	IC 95%	p-valor	Est.	IC 95%	p-valor
Sexo									
Masculino	0,612	[0,52; 0,71]	<0,001	0,344	[0,29; 0,40]	<0,001	0,532	[0,45; 0,62]	<0,001
Feminino	1	-	-	1	-	-	1	-	-
Faixa etária									
60 a 69 anos	0,151	[0,03; 0,27]	0,013	0,083	[0,02; 0,15]	0,014	0,125	[0,02; 0,23]	0,018
70 anos ou mais	1	-	-	1	-	-	1	-	-
Escolaridade			0,017			0,019			0,020
Sem inst./Fund. incomp.	0,377	[0,12; 0,64]	0,005	0,202	[0,06; 0,34]	0,005	0,331	[0,10; 0,57]	0,005
Fund./Médio comp.	0,335	[0,06; 0,61]	0,016	0,177	[0,03; 0,32]	0,018	0,303	[0,06; 0,55]	0,015
Superior completo	1	-	-	1	-	-	1	-	-
Plano de saúde									
Sim	1	-	-	1	-	-	1	-	-
Não	0,871	[0,69; 1,05]	<0,001	0,474	[0,38; 0,57]	<0,001	0,782	[0,62; 0,94]	<0,001
Região			<0,001			<0,001			<0,001
Sul	1	-	-	1	-	-	1	-	-
Sudeste	0,170	[-0,02; 0,36]	0,082	0,094	[-0,01; 0,20]	0,083	0,150	[-0,02; 0,32]	0,083
Centro-Oeste	0,395	[0,18; 0,61]	<0,001	0,228	[0,11; 0,35]	<0,001	0,338	[0,15; 0,53]	<0,001
Nordeste	0,693	[0,51; 0,88]	<0,001	0,391	[0,29; 0,49]	<0,001	0,596	[0,44; 0,76]	<0,001
Norte	0,625	[0,41; 0,84]	<0,001	0,359	[0,24; 0,48]	<0,001	0,530	[0,34; 0,72]	<0,001
Autoaval. de saúde			<0,001			<0,001			<0,001
Bom	1,231	[1,02; 1,44]	<0,001	0,674	[0,57; 0,78]	<0,001	1,086	[0,90; 1,27]	<0,001
Regular	0,488	[0,29; 0,69]	<0,001	0,259	[0,16; 0,36]	<0,001	0,446	[0,26; 0,63]	<0,001
Ruim	1	-	-	1	-	-	1	-	-
Área									
Urbana	1	-	-	1	-	-	1	-	-
Rural	0,431	[0,30; 0,56]	<0,001	0,245	[0,17; 0,32]	<0,001	0,361	[0,25; 0,47]	<0,001
Diag. de doenças									
Sim	1	-	-	1	-	-	1	-	-
Não	1,059	[0,92; 1,20]	<0,001	0,563	[0,49; 0,64]	<0,001	0,968	[0,83; 1,10]	<0,001
Dificuldade									
Não cons./ Grande dif.	1	-	-	1	-	-	1	-	-
Peq. dif./Não tem dif.	0,270	[0,08; 0,46]	0,005	0,142	[0,04; 0,24]	0,006	0,242	[0,07; 0,41]	0,006

Após a exclusão das variáveis “cor/raça” e “situação conjugal”, observou-se que nos três modelos as demais variáveis permaneceram associadas com a probabilidade do idoso nunca ter ido ao médico ou ter ido há 1 ano ou mais. Além disso, independentemente da escolha da função de ligação, é possível observar que os sentidos

das associações se mantiveram, isto é, os sinais das estimativas pontuais dos parâmetros nos três modelos binários foram iguais (Tabela 3).

Desse modo, pode-se observar que a probabilidade de consulta médica nunca ter sido realizada, ou ter ocorrido há 1 ano ou mais, foi maior entre os idosos do sexo masculino, mais jovens (60 a 69 anos), residentes da área rural, e das regiões Norte e Nordeste.

Com relação às características de saúde, observam-se maiores probabilidades de consulta médica nunca ter sido realizada, ou ter ocorrido há 1 ano ou mais, entre idosos que não têm plano de saúde, que não têm diagnóstico de doenças crônica, física ou mental e entre idosos que reportaram no máximo pequena dificuldade para ir ao médico sozinhos.

Nos três modelos foi observado um gradiente na probabilidade do idoso nunca ter ido ao médico ou ter ido há 1 ano ou mais, segundo a escolaridade e a autoavaliação de saúde geral. Idosos menos escolarizados e aqueles que reportaram melhores níveis de saúde apresentaram maiores probabilidades da última consulta médica ter ocorrido há 1 ano ou mais ou nunca ter sido realizada.

A Tabela 4 apresenta as estimativas pontuais padronizadas dos modelos logit e complemento log-log, a fim de torná-las comparáveis entre si e com as estimativas dos parâmetros do modelo probit.

Tabela 4: Padronização das estimativas pontuais dos parâmetros dos modelos logit e complemento log-log para o desfecho binário referente ao tempo da última consulta médica realizada pelos idosos brasileiros, considerando somente as variáveis explicativas selecionadas.

Variáveis Explicativas	Modelo Logit	Modelo Probit	Modelo Clog-log	Est.P (Logit)/ Est. (Probit)	Est.P (Cloglog)/ Est. (Probit)	Est.P (Cloglog)/ Est.P (Logit)
	Est.P	Est.	Est.P			
Sexo						
Masculino	0,338	0,344	0,415	0,983	1,208	1,229
Feminino	1	1	1	1	1	1
Faixa etária						
60 a 69 anos	0,083	0,083	0,097	1,004	1,170	1,165
70 anos ou mais	1	1	1	1	1	1
Escolaridade						
Sem inst./Fund. incomp.	0,208	0,202	0,258	1,027	1,276	1,243
Fund./Médio comp.	0,185	0,177	0,236	1,045	1,337	1,280
Superior completo	1	1	1	1	1	1
Plano de saúde						
Sim	1	1	1	1	1	1
Não	0,480	0,474	0,610	1,013	1,286	1,270
Região						
Sul	1	1	1	1	1	1
Sudeste	0,094	0,094	0,117	0,999	1,249	1,250
Centro-Oeste	0,218	0,228	0,263	0,953	1,154	1,210
Nordeste	0,382	0,391	0,465	0,979	1,191	1,216
Norte	0,344	0,359	0,413	0,958	1,150	1,200
Autoaval. de saúde						
Bom	0,679	0,674	0,846	1,007	1,256	1,247
Regular	0,269	0,259	0,347	1,040	1,342	1,290
Ruim	1	1	1	1	1	1
Área						
Urbana	1	1	1	1	1	1
Rural	0,238	0,245	0,282	0,969	1,148	1,184
Diag. de doenças						
Sim	1	1	1	1	1	1
Não	0,584	0,563	0,755	1,038	1,342	1,293
Dificuldade						
Não cons./ Grande dif.	1	1	1	1	1	1
Peq. dif./Não tem dif.	0,149	0,142	0,188	1,047	1,324	1,264

Est. = Estimativas pontuais dos parâmetros do modelo Probit.
Est.P = Estimativas pontuais padronizadas dos parâmetros dos modelos Logit e Complemento Log-Log.

Cabe mencionar que as estimativas pontuais dos parâmetros do modelo probit não precisam ser padronizadas, pois a função probit já é baseada na distribuição normal padrão, que possui variância 1, e portanto padronizar essas estimativas seria dividi-las por $\sqrt{1}$. As estimativas pontuais dos parâmetros do modelo probit apresentadas na Tabela 3

foram repetidas na Tabela 4, e são utilizadas como base de comparação para os modelos logit e complemento log-log.

Com relação à magnitude das estimativas pontuais, observa-se que todas as estimativas do modelo logit são próximas daquelas obtidas pelo modelo probit, ou seja, não divergem em mais do que 5%. A maior diferença positiva (variação de +4,7%) ocorreu na estimativa do parâmetro referente a categoria "pequena dificuldade ou nenhuma dificuldade para ir ao médico sozinho" e a maior diferença negativa (variação de -4,7%) ocorreu para a categoria "região Centro-Oeste".

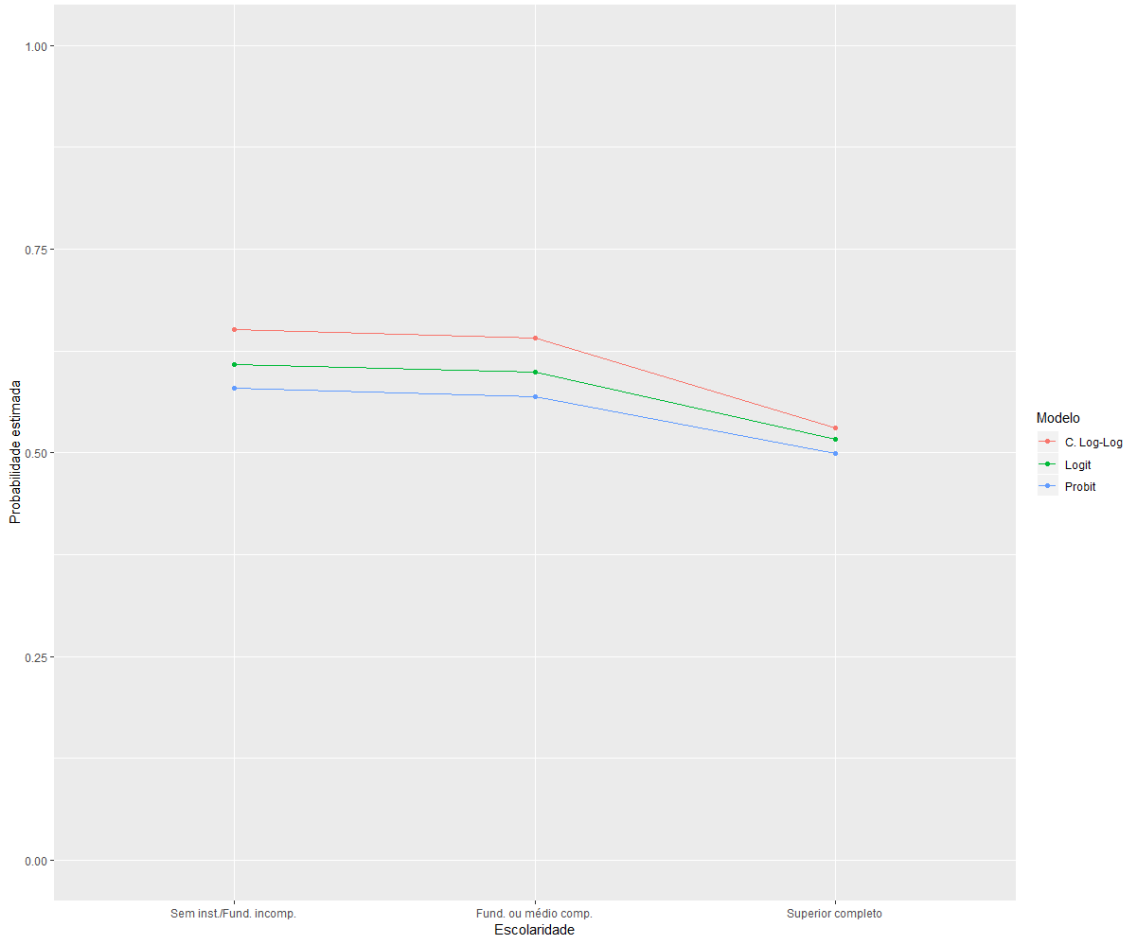
Já ao comparar tanto as estimativas pontuais do modelo logit quanto as do modelo probit com aquelas provenientes do modelo complemento log-log, é possível observar no modelo complemento log-log estimativas substancialmente maiores que aquelas obtidas nos outros dois modelos. As maiores diferenças foram observadas para a estimativa pontual do parâmetro associado à autoavaliação de saúde "regular", que foi 34,2% maior no modelo complemento log-log em comparação com o modelo probit, e 29% maior no modelo complemento log-log comparativamente ao modelo logit. Além disso, a estimativa pontual referente à não existência de diagnóstico de doença no modelo complemento log-log foi 34,2% maior que a do modelo probit, e 29,3% maior que a do logit. Comparando os modelos complemento log-log *versus* probit, a menor diferença foi observada para a estimativa do parâmetro referente à área "rural", cuja estimativa no modelo complemento log-log foi 14,8% maior que a do modelo probit. Já ao comparar os modelos complemento log-log *versus* logit, a menor diferença foi verificada para a estimativa pontual referente a faixa etária de "60 a 69 anos", a qual foi 16,5% maior que a do modelo logit.

A Figura 2 apresenta as probabilidades estimadas dos idosos nunca terem ido ao médico ou terem ido há 1 ano ou mais segundo os níveis de escolaridade, fixando as demais características dos idosos. O perfil fixado foi formado por idosos da região Nordeste, do sexo masculino, com idade entre 60 e 69 anos, sem posse de plano de saúde, com boa autoavaliação de saúde, sem diagnóstico de doença crônica, física, mental ou doença de longa duração, com pequena ou nenhuma dificuldade de ir ao médico sozinho e residente em área rural.

A Figura 3 apresenta o mesmo tipo de comparação, porém apresenta as probabilidades estimadas do desfecho de interesse segundo os níveis de autoavaliação de saúde de idoso. Nesta figura, fixou-se o seguinte perfil: idosos residentes na região Nordeste, do sexo masculino, com idade entre 60 e 69 anos, sem plano de saúde, sem

instrução ou apenas com fundamental incompleto, sem diagnóstico de doença crônica, física, mental ou doença de longa duração, com pequena ou nenhuma dificuldade de ir ao médico sozinho e residente em área rural.

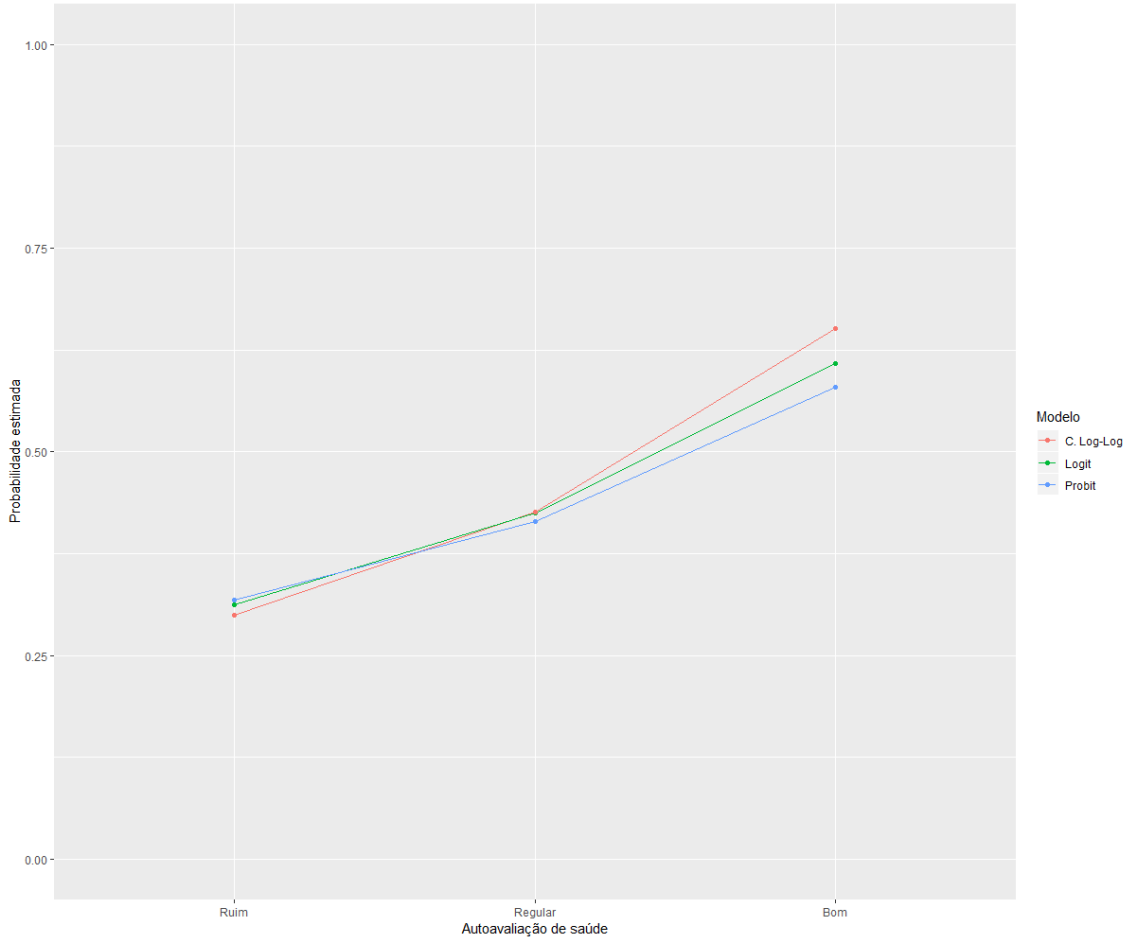
Figura 2: Análise comparativa das probabilidades estimadas dos idosos nunca terem ido ao médico ou terem ido há 1 ano ou mais por cada modelo selecionado, segundo os níveis de escolaridade, e para um dado perfil escolhido.



Na Figura 2, observa-se para os três modelos binários que ao aumentar o nível de escolaridade do idoso, ocorre uma diminuição em sua probabilidade de nunca ter ido ao médico ou ter ido há um ano ou mais. Para os três níveis de escolaridade considerados na análise, as probabilidades estimadas pelo modelo complemento log-log foram maiores que as probabilidades estimadas pelos modelos logit (aproximadamente 7% maiores para os dois níveis de escolaridades mais baixos) e probit (cerca de 12% maiores para os dois níveis mais baixos). Quanto ao nível superior, a probabilidade estimada pelo modelo complemento log-log foi 6,4% maior do que a estimada pelo modelo probit, e apenas

2,7% maior que a do modelo logit. As diferenças entre as estimativas das probabilidades obtidas pelos modelos logit e probit foram de no máximo 5%, aproximadamente.

Figura 3: Análise comparativa das probabilidades estimadas dos idosos nunca terem ido ao médico ou terem ido há 1 ano ou mais por cada modelo selecionado, segundo os níveis de autoavaliação de saúde, e para um dado perfil escolhido.



Na Figura 3, observa-se para os três modelos binários a mesma tendência de aumento da probabilidade do idoso nunca ter ido ao médico ou ter ido há um ano ou mais, à medida que melhora o nível de saúde reportado por ele. As probabilidades estimadas pelos modelos logit e probit são próximas, alcançando apenas uma variação relativa de aproximadamente 5% para o nível "bom" de autoavaliação de saúde. A maior diferença nas probabilidades estimadas para o nível "bom" de autoavaliação de saúde foi observada entre o modelo complemento log-log e os modelos logit e probit, isto é, a probabilidade do idoso nunca ter ido ao médico ou ter ido há 1 ano ou mais obtida pelo modelo

complemento log-log foi 7% maior que a probabilidade estimada através do modelo logit, e 12,5% maior que a estimada pelo modelo probit.

A Tabela 5 apresenta algumas medidas de qualidade do ajuste para os modelos logit, probit e complemento log-log selecionados. A variação entre as medidas de pseudo-R^2 de Cox-Snell e de Nagelkerke observadas entre os modelos é pequena, o que sugere não haver real diferença entre os três modelos binários no quesito qualidade do ajuste baseando-se nestas medidas. Quanto ao teste de Wald de significância geral, os p-valores são menores que 5%, apontando que as variáveis sociodemográficas e de saúde utilizadas possuem relação estatisticamente significante com o desfecho, e que, portanto, os três modelos que utilizam essas variáveis explicativas se ajustam melhor aos dados observados do que os modelos nulos (somente com a constante).

Tabela 5: Teste de Wald Geral e Medidas de qualidade do ajuste para cada um dos três modelos binários selecionados.

Medidas de qualidade do ajuste	**Modelo logit**	**Modelo probit**	**Modelo complemento log-log**
R^2 de Cox-Snell	0,1078	0,1074	0,1077
R^2 de Nagelkerke	0,1821	0,1814	0,1819
p-valor do teste de Wald geral	<0,001	<0,001	<0,001

A Tabela 6 mostra que, para o presente caso, o modelo logit possui uma medida de sensibilidade ligeiramente maior dentre os modelos analisados, ou seja, usando o modelo logit, 69.59% dos idosos que nunca foram ao médico ou que foram há 1 ano ou mais foram classificados corretamente. Já a medida de especificidade foi ligeiramente maior para o modelo complemento log-log, onde 69,40% dos idosos que se consultaram com o médico há menos de 1 ano foram corretamente classificados pelo modelo neste grupo. Entretanto, cabe mencionar que os valores de sensibilidade e especificidade estão no intervalo entre 50% e 80%, indicando uma capacidade preditiva razoável para os três modelos binários. As taxas globais de classificações corretas também foram similares nos três modelos

Tabela 6: Medidas de avaliação da capacidade preditiva dos modelos logit, probit e complemento log-log para o desfecho binário referente ao tempo da última consulta médica realizada pelos idosos.

Medidas de capacidade preditiva	**Modelo Logit**	**Modelo Probit**	**Modelo Complemento Log-Log**
Sensibilidade (S)*	69,59%	68,87%	67,26%
Especificidade (E)*	67,18%	67,87%	69,40%
Taxa global de classificações corretas (TG)*	67,58%	68,03%	69,05%
Área sob a Curva ROC (A)	0,7500	0,7500	0,7498

*Pontos de corte (γ) ótimos: Logit = 0,1726; Probit = 0,1838; Comp. Log-Log = 0,1792

A Figura 4 apresenta a curva ROC para os modelos probit, logit e complemento log-log, e sugere grande semelhança entre elas. Além disso, tanto o modelo logit quanto o probit apresentaram a mesma área sob a curva ROC, área essa que também é muito próxima daquela observada sob a curva ROC do modelo complemento log-log. A diferença relativa entre essas áreas é de apenas 0,027%, o permite concluir que não há diferença entre as capacidades discriminatórias dos três modelos, e as suas capacidades preditivas são consideradas "aceitáveis" ou "razoáveis" (Tabela 6).

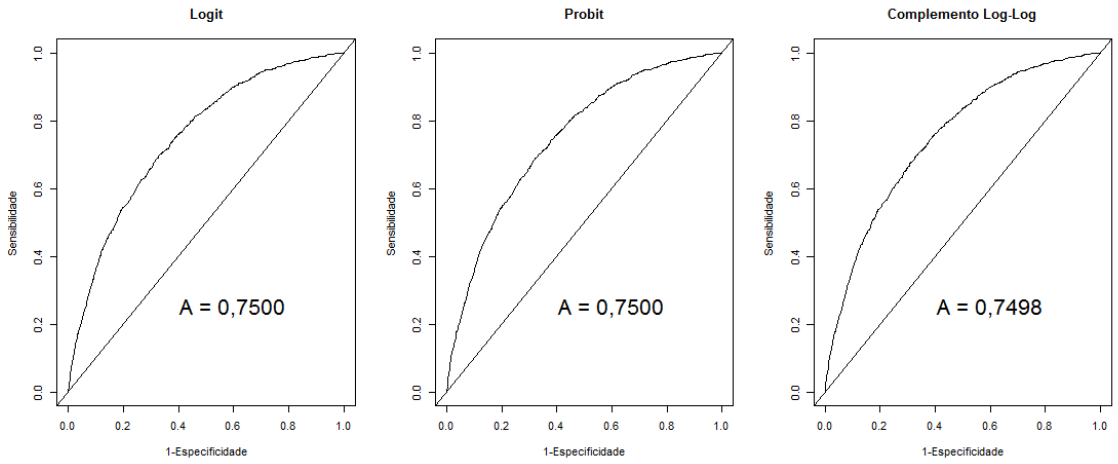


Figura 4: Curvas ROC referentes aos modelos logit, probit e complemento log-log, com as respectivas áreas sob as curvas.

4. Discussão, conclusões e considerações finais

No presente trabalho buscou-se avaliar, sobretudo, o impacto que a escolha das funções de ligação logit, probit e complemento log-log tem na qualidade do ajuste e na capacidade discriminatória de um modelo linear generalizado, considerando um desfecho binário referente ao tempo da última consulta ao médico realizada por idosos brasileiros e as características sociodemográficas e de saúde desses indivíduos como variáveis explicativas. Buscou-se também avaliar se a função de ligação adotada influencia na significância dos parâmetros, na magnitude e no sentido das associações entre as características dos idosos e o tempo desde a última consulta médica.

Verificou-se que a escolha da função de ligação não produziu mudanças na significância das associações entre as características dos idosos e o desfecho de estudo, ou seja, independentemente da função de ligação utilizada nos modelos binários, das onze características sociodemográficas e de saúde inicialmente incluídas na modelagem, nove destas características (as mesmas) apresentaram relação estatisticamente significante com o tempo da última consulta ao médico realizada pelos idosos. Além disso, ao comparar os modelos logit, probit e complemento log-log, notou-se também que os sinais das estimativas pontuais dos parâmetros foram os mesmos, ou seja, essas características dos idosos influenciaram no mesmo sentido a probabilidade do idoso nunca ter ido ao médico ou ter ido há um ano ou mais.

Cabe mencionar que, dentre as características inicialmente consideradas na modelagem estatística, a situação conjugal e a cor/raça do idoso não apresentaram associação estatisticamente significante com o desfecho de estudo em quaisquer dos modelos avaliados. Quanto à cor/raça, a falta de associação significativa com a consulta ao médico já havia sido observada por Moreira, Moraes e Luiz (2011) e Capilheira e Santos (2006), embora em ambos os artigos a população de estudo tenha sido composta por adultos (20 anos ou mais), e não especificamente por idosos. Já em relação à situação conjugal, Moreira, Moraes e Luiz (2011) encontraram associação significativa com o desfecho de consulta ao médico por adultos, associação que não foi observada em relação a idosos no presente trabalho.

Segundo as medidas utilizadas no presente estudo, não foram encontradas diferenças significativas na qualidade do ajuste e na capacidade preditiva dos MLGs, com indicações pouco contundentes a favor dos modelos logit (sensibilidade levemente maior)

e complemento log-log (valores ligeiramente maiores de especificidade e taxa global de acertos), sem que haja argumento conclusivo para apontar que o uso de algum dos modelos seja melhor que o dos demais para representar os dados observados.

Cabe ressaltar, entretanto, que apesar de não haver diferenças em termos de seleção de variáveis explicativas, significância dos parâmetros e sentido das associações nos três modelos, o modelo complemento log-log gerou estimativas pontuais padronizadas consideravelmente maiores que as geradas pelos modelos logit e probit (variações relativas superiores a 10%). Campos (2016) comparou MLGs com desfecho binário e as mesmas funções de ligação utilizadas no presente estudo, e encontrou resultados bastante semelhantes. Em seu trabalho, a qualidade do ajuste dos modelos foi verificada pelo critério de informação de Akaike (AIC), além da taxa global de acertos como medida de avaliação da capacidade preditiva, através do qual perceberam-se diferenças apenas marginais entre os MLGs, com pequena vantagem obtida pelo modelo logit. O autor optou pelo uso do modelo logit, citando algumas vantagens deste modelo em relação aos demais, sobretudo em termos de facilidade de interpretação. Diferentemente do presente estudo, a amostra utilizada por Campos (2016) não provém de plano amostral complexo.

Entre as potencialidades deste trabalho, pode-se apontar a utilização do método de MPV no ajuste dos modelos logit, probit e complemento log-log, cujo método de estimação permite agregar as informações do plano amostral complexo (estratificação, conglomeração e pesos amostrais) da PNS 2013, de forma a obter estimativas pontuais e medidas de precisão de forma adequada. Ignorar as informações do plano amostral resultaria em incorreções, seja por subestimação ou superestimação, nas estimativas pontuais dos parâmetros dos modelos e nas variâncias estimadas dessas estimativas (PESSOA; SILVA, 1998), o que afetaria os resultados dos testes de Wald de significância individual e geral e as probabilidades estimadas de sucesso de cada indivíduo, podendo afetar inclusive a seleção do modelo. Outra possível potencialidade, do ponto de vista social, é a análise da consulta ao médico dirigida ao grupo de idosos, diferindo do uso do grupo mais geral formado por adultos. O foco no grupo de indivíduos com 60 anos ou mais visa contribuir para a identificação dos fatores sociodemográficos e de saúde que precisam ser priorizados para que esse grupo tenha suas necessidades de saúde atendidas. Devido às limitações do sistema de saúde pública brasileiro, o rápido processo de envelhecimento populacional aponta para a necessidade de reformulação das políticas

públicas a fim de prevenir uma saúde precária e a incapacidade funcional de idosos (WONG; CARVALHO, 2006, ALVES; RODRIGUES, 2005).

Quanto às limitações deste trabalho, tem-se a realização da modelagem utilizando apenas os efeitos principais das variáveis, sem a verificação de possíveis efeitos de interação entre elas, o que tornaria o processo de modelagem e a interpretação das estimativas mais complexos. Entende-se, entretanto, que potencialmente seria possível obter modelos mais bem ajustados, para os quais talvez houvesse maiores diferenças entre as medidas de qualidade do ajuste e de capacidade preditiva. Outra limitação que pode ser apontada deve-se à própria natureza da pesquisa utilizada neste trabalho. Como a PNS é um estudo seccional, os resultados obtidos não podem ser interpretados como relações de causa e efeito, pois neste tipo de estudo a amostra é selecionada num único momento (período) do tempo, não havendo garantia de temporalidade da associação entre as variáveis explicativas e a ocorrência do desfecho. Segundo Medronho *et al*. (2009), estudos seccionais não são considerados estratégias válidas para testar vínculos causais entre eventos, e sim para testar se há associações estatísticas entre eles.

A partir do presente estudo, conclui-se que, para os dados amostrais utilizados, qualquer uma das três funções de ligação poderia ser escolhida, a fim de ajustar um MLG para avaliar a associação entre as características sociodemográficas e de saúde dos idosos e o tempo da última consulta médica realizada por eles. Entretanto, em investigações na área de saúde, o modelo logit é o mais frequentemente utilizado pelos pesquisadores, por possibilitar interpretações mais fáceis das associações entre as variáveis explicativas e um dado desfecho binário. Entre as vantagens do modelo logit, em comparação aos modelos probit e complemento log-log, Faraway (2006) aponta a interpretação mais fácil das estimativas dos parâmetros do modelo proporcionada pelo uso de chances (no inglês, *odds*) e a possibilidade de manipulação algébrica para a função logit. Faraway (2006), menciona ainda que como o comportamento das três funções é bastante similar quando as probabilidades não estão próximas de 0 ou 1, a escolha da função de ligação a ser usada no modelo binário deve ser baseada não somente a partir dos dados, mas também em critérios subjetivos.

Referências

ALVES, L. C.; RODRIGUES, R. N. Determinantes da autopercepção de saúde entre idosos do Município de São Paulo, Brasil. *Revista Panamericana de Salud Pública*. 17(5-6):333-41, 2005.

BINDER, D. A. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, p.279-292, 1983.

BRASIL. *Manual de enfrentamento à violência contra a pessoa idosa: É possível prevenir. É necessário superar*. Brasília: Secretaria de Direitos Humanos do Brasil, 2014.

CAMPOS, M. H. C. *Aplicação do Modelo Binomial aos Dados do Sítio Consumidor.gov.br*. Trabalho de Conclusão de Curso (Graduação em Estatística) — Universidade Federal de Juiz de Fora, 2016.

CAPILHEIRA, M. F.; SANTOS, I. S. Fatores individuais associados à utilização de consultas médicas por adultos. *Rev Saúde Pública*, 40(3):436-43, 2006.

CORDEIRO, G. M.; NETO, E. A. L. *Modelos paramétricos*. São Paulo: ABE, 2004.

CORDEIRO, G. M.; DEMÉTRIO, C. G. B. *Modelos Lineares Generalizados e Extensões*. Piracicaba, 2013.

DOBSON, A. J.; BARNETT, A. G. *An introduction to generalized linear models*. 4. ed. Boca Raton: CRC Press, 2018.

FARAWAY, J. J. *Extending the Linear Model with R: Generalized Linear, Mixed Effects, and Nonparametric Regression Models*. 2. ed. Boca Raton: CRC Press, 2006.

FONTES, L. F. C.; CONCEIÇÃO, O. C.; MACHADO, S. Violência sexual na adolescência, perfil da vítima e impactos sobre a saúde mental. *Ciência & Saúde Coletiva*, 22(9): 2919-2928, 2017.

GUJARATI, D. N.; PORTER, D. C. *Econometria Básica*. 5. ed. São Paulo: AMGH Editora, 2011.

GUO, S.; FRASER, M. W. *Propensity Score Analysis: Statistical Methods and Applications*. 2. ed. California: SAGE Publications, 2015.

HEERINGA, S. G.; WEST, B. T.; BERGLUND, P. A. *Applied Survey Data Analysis*. Boca Raton: CRC Press, 2010.

INTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). *Sistema Integrado de Pesquisas Domiciliares — SIPD*. Rio de Janeiro: IBGE, 2007.

INTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). *Pesquisa Nacional de Saúde 2013: Percepção do estado de saúde, estilos de vida e doenças crônicas – Brasil, Grandes Regiões e Unidades da Federação*. Rio de Janeiro: IBGE, 2014.

KELLES, S. M. B.; DINIZ, M. D. F. H. S.; MACHADO, C. J.; BARRETO, S. M. Perfil de pacientes submetidos à cirurgia bariátrica, assistidos pelo Sistema Único de Saúde do Brasil: revisão sistemática. *Cadernos de Saúde Pública*, 31(8): 1587-1601, 2015.

KUTNER, M. H.; NACHTSHEIM, C. J.; NETER, J.; LI, W. *Applied Linear Regression Models*. 5. ed. New York: McGraw-Hill, 2004.

MARÔCO, J. *Análise Estatística com o PASW Statistics*. Pêro Pinheiro: ReportNumber, 2010.

MARTINE, G.; MCGRANAHAN, G. A transição urbana brasileira: trajetória, dificuldades e lições aprendidas. In: BAENINGER, R. (Org.). População e cidades: subsídios para o planejamento e para as políticas sociais. Campinas: Núcleo de Estudos de População NEPO/Unicamp; Brasília: UNFPA, 2010.

MEDRONHO, R. A.; BLOCH, K. V.; LUIZ, R. R.; WERNECK, G. L. *Epidemiologia*. 2.ed. Rio de Janeiro: Atheneu; 2009.

MORAES J. R.; MOREIRA J. P. L.; LUIZ R. R. Efeito do plano amostral em modelo logístico ordinal: uma análise do estado de saúde autorreferido de adultos no Brasil usando a Pesquisa Nacional por Amostra de Domicílios de 2008. *Cadernos de Saúde Pública*, 28(5): 913-924, 2012.

MOREIRA, J. P. L.; MORAES, J. R.; LUIZ, R. R. Utilização de consulta médica e hipertensão arterial sistêmica nas áreas urbanas e rurais do Brasil, segundo dados da PNAD 2008. *Ciência & Saúde Coletiva*, 16(9): 3781-3793, 2011.

NORONHA, K. V. M. D. S.; ANDRADE, M. V. Desigualdades sociais em saúde e na utilização dos serviços de saúde entre os idosos na América Latina. *Revista Panamericana de Salud Pública*, 17(5-6): 410-418, 2005.

OSORIO, R. G.; SERVO, L. M. S.; PIOLA, S. F. Necessidade de saúde insatisfeita no Brasil: uma investigação sobre a não procura de atendimento. *Ciência & Saúde Coletiva*, 16(9): 3741-3754, 2011.

PESSOA, D. G. C.; SILVA, P. L. N. *Análise de Dados Amostrais Complexos*. São Paulo: ABE, 1998.

POWERS, D. A.; XIE, Y. *Statistical Methods for Categorical Data Analysis*. Academic Press, 1999.

REIS, R. J.; PINHEIRO, T. M.; NAVARRO, A.; MARTIN, M. Perfil da demanda atendida em ambulatório de doenças profissionais e a presença de lesões por esforços repetitivos. *Revista de Saúde Pública*, 34(3): 292-298, 2000.

RODRÍGUEZ, G. *Lecture Notes on Generalized Linear Models*. Princeton. 2007. Disponível em: <https://data.princeton.edu/wws509/notes/>. Acesso em: 18/11/2020.

SOUZA-JÚNIOR, P. R. B.; FREITAS, M. P. S.; ANTONACI, G. A.; SZWARCWALD, C. L. Desenho da amostra da Pesquisa Nacional de Saúde 2013. *Epidemiol. Serv. Saúde*, Brasília, 24(2): 207-216, 2015.

WONG, L. L.; CARVALHO, J. A. O rápido processo de envelhecimento populacional do Brasil: sérios desafios para as políticas públicas. *Revista Brasileira de Estudos de População*, 23(1): 5-26, 2006.