Universidade de Brasília
Departamento de Estatística

Fatores Associados à Autopercepção de Saúde: Uma Aplicação de Regressão Logística

Alisson Moreira Ferreira

Projeto apresentado para obtenção do título de Bacharel em Estatística.

Brasília

2017

Alisson Moreira Ferreira

Fatores Associados à Autopercepção de Saúde: Uma Aplicação de Regressão Logística

Orientadora:
Prof(a). **Maria Teresa Leão Costa**

Projeto apresentado para obtenção do título de Bacharel em Estatística.

Brasília

2017

Agradecimentos

À professora Maria Teresa Leão Costa pela paciência e dedicação.

Aos professores do departamento de estatística da UnB que me auxiliaram com seus conhecimentos.

Aos meus familiares pelo carinho, amor e proteção que sempre me deram.

Aos meus amigos Alan, Davi, Talita, Bruna e Luciano pelo companheirismo durante o curso. Vocês fizeram a diferença na minha vida.

A todos que direta ou indiretamente contribuíram para conclusão deste trabalho.

Sumário

Resumo

Regressão Logística

O estudo de regressão logística está cada vez mais presente na estatística, principalmente em artigos científicos da área médica. Por se tratar de uma técnica que pode ser utilizada em diversas áreas e suas conclusões são de fácil entendimento para os profissionais, faz-se necessário que o modelo esteja bem ajustado. A aplicação da análise de regressão logística tem por objetivo principal encontrar fatores de risco para a situação problema em questão.

Para verificar a adequabilidade do modelo e a presença de observações discrepantes que causam modificações nas estimativas do modelo, é fundamental realizar a análise de diagnóstico. As técnicas empregadas no diagnóstico da função logito são bastante semelhantes às técnicas de diagnóstico da regressão linear.

No presente trabalho será apresentado um estudo de regressão logística aplicado nos dados do Programa Nacional de Saúde (PNS - 2013). O estudo será focado na auto percepção investigando a influência de determinantes geográficas, dificuldades locomotoras, auditivas e visuais e variáveis psicológicas. Serão apresentadas algumas das técnicas mais importantes para o diagnóstico de regressão logística. Algumas dessas técnicas usam resíduos calculados a partir do modelo ajustado e identificam através de gráficos os pontos atípicos.

Capítulo 1

1 Introdução e Justificativa

Variáveis dicotômicas são comuns em estudos estatísticos experimentais e observacionais. Essas variáveis estão presentes em diversas áreas do conhecimento. Pode-se citar como exemplo: pesquisas médicas com o objetivo de estudar o rompimento ou não da artéria do cérebro, causando aneurisma cerebral, estudos educacionais para avaliar a aprovação/reprovação dos estudantes, estudos na área bancária para avaliar a inadimplência ou não do consumidor, entre outros. Mesmo diante de variáveis quantitativas é possível fazer a classificação em dois grupos, desde que o ponto de corte escolhido seja coerente com os estudos e a análise dos dados.

David Collet , em seu livro Modelling Binary Date, destaca que as principais vantagens de utilizar o modelo de regressão logística são:

- A conveniência de seu uso do ponto de vista computacional;
- A sua interpretação direta em termos do logaritmo de chance de sucesso, sendo útil na análise de dados de estudos epidemiológicos;
- o fato de modelos com transformação logística, por suas propriedades, serem mais adequados para a análise de dados que foram coletados retrospectivamente, tal como em estudos de caso-controle.

Modelos de regressão tem sido amplamente utilizados quando o objetivo do estudo é explicar o comportamento de uma característica do fenômeno estudado (variável resposta) em função de outras características associadas (variáveis explicativas). Nos modelos de regressão linear, a variável resposta é quantitativa, já no caso da variável resposta ser binária utiliza-se a regressão logística. O modelo logístico foi apresentado primeiramente por Fisher e Yates (1938). Duas décadas depois, Cox (1958) escreveu um importante trabalho sobre o assunto, o que contribuiu para que o tema ganhasse mais notoriedade entre os pesquisadores. Nos últimos anos, a utilização do modelo logístico tem crescido ainda mais

nas diversas áreas do conhecimento, em especial em estudos na área de saúde. A principal razão da aplicabilidade da técnica de regressão logística é a possibilidade de interpretar os resultados do modelo como razões de chance. Os métodos aplicados na regressão logística são semelhantes aos aplicados na regressão linear, porém com algumas restrições e suposições que serão explanadas ao longo do trabalho.

Para que um modelo seja considerado adequado, é necessário que:

- As principais variáveis explicativas que contribuem para o melhor ajuste dos dados estejam devidamente especificadas;
- O modelo não seja influenciado por determinadas características dos dados, como observações que causem alguma mudança nas estimativas dos coeficientes, superestimando-as ou subestimando-as.

Para identificar a ocorrência de observações atípicas, são utilizadas as técnicas de diagnóstico do modelo. Essas técnicas também verificam se as suposições do modelo estão satisfeitas, se há presença de outliers e se o modelo está bem ajustado para o conjunto completo de covariáveis. Uma das técnicas de diagnóstico mais usada para modelos de regressão é a análise de resíduos. O principal objetivo da análise de resíduos na regressão logística é identificar casos para os quais as estimativas do modelo se distanciam muito dos valores observados, ou casos em que exerçam uma influência maior do que deveriam nas estimativas dos parâmetros do modelo.

O objetivo principal do presente estudo consiste em identificar os fatores associados à autopercepção de saúde dos brasileiros.

Capítulo 2

2 Revisão da Literatura

2.1 Regressão Logística

O modelo de regressão logística binária é um caso particular dos modelos lineares generalizados, onde a variável resposta é qualitativa e binária.

Uma variável binária assume dois valores, por exemplo Y_i=0 ou Y_i=1, denominados "sucesso"e "não-sucesso", respectivamente.

Esse modelo de regressão pode ser utilizado em estudos observacionais e experimentais. No modelo linear temos:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

e a E($\beta_0 + \beta_1 x_i + \epsilon_i$) = β_0+$\beta_1 x_i$, assumindo que E(ϵ_i)=0 e V(Y$_i$=σ^2) para qualquer i. Além disso, assume-se a normalidade do erro.

A variável Y tem distribuição de Bernoulli (1,π) com probabilidade de sucesso P(Y_i=1)=π_i e de não-sucesso P(Y_i=0)=1-π_i.

Desse modo, alguns problemas surgem quando a variável resposta é binária.

1. Não normalidade do erro

   Como Y_i=$\beta_0 + \beta_1 x_i + \epsilon_i \rightarrow \epsilon_i$=0-$\beta_0$-$\beta_1 x_i$ ou ϵ_i=1-β_0-$\beta_1 x_i$

   Ou seja, não faz sentido assumir normalidade no erro.

2. Variância do erro não constante

   A variância da função de Bernoulli é dada por $\pi_i(1 - \pi_i)$. Logo, temos que:

   Var(ϵ_i)=$\pi_i(1 - \pi_i)$=β_0+$\beta_1 x_i$(1-β_0-$\beta_1 x_i$).

   Como a variância do erro depende de x_i, conclui-se que ela não é constante, ou seja, a hipótese de homocedasticidade foi violada.

3. Limitação para a resposta média

   Como a resposta média é uma probabilidade, temos que $0 \leq E(Y_i) \leq 1$, ou seja, $0 \leq \beta_0 + \beta_1 x_i \leq 1$. Como a função $\beta_0 + \beta_1 x_i$ representa a função de uma reta, seus limites são $(-\infty, +\infty)$.

2.2 Regressão Logística Simples

Muitas funções foram propostas para análise de variáveis com respostas dicotômicas. Dentre elas, a mais simples é a que dá origem ao modelo logístico. Em relação a estatística, esse modelo é o mais simples, bastante flexível e com interpretação de fácil entendimento.

Assim, π_i deve variar entre 0 e 1. Uma representação linear simples pra π_i sobre todos os valores possíveis de x não é adequada, desse modo considera-se a transformação logística π_i sob a forma linear:

$$ln\left(\frac{\pi_i}{1-\pi_i}\right) = g(x)$$

Onde g(x)=$\beta_0 + \beta_1 x_i$. Ou equivalentemente:

$$\pi_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \tag{1}$$

Para provar a equivalência das equações acima é necessário usar a definição de chance que é definida como a razão entre a probabilidade de um evento ocorrer sobre a probabilidade do evento não ocorrer. No estudo da regressão logística, essa chance é definida como:

$$Chance = \left(\frac{\pi_i}{1-\pi_i}\right)$$

A equação acima pode ser reescrita da seguinte forma:

$$\left(\frac{\pi_i}{1-\pi_i}\right) = \left(\frac{\frac{e^{\beta_0 + \beta_1 x_i}}{1+\beta_0 + \beta_1 x_i}}{1 - \frac{e^{\beta_0 + \beta_1 x_i}}{1+e^{\beta_0 + \beta_1 x_i}}}\right)$$

Com alguns cálculos e manipulações algébricas, chegamos a seguinte expressão:

$$ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i \tag{2}$$

De acordo com a equação (2), podemos tirar algumas conclusões a respeito dos estimadores:

- Quando $\beta_1 < 0$, π_i é função crescente, conforme pode ser visto no gráfico abaixo.

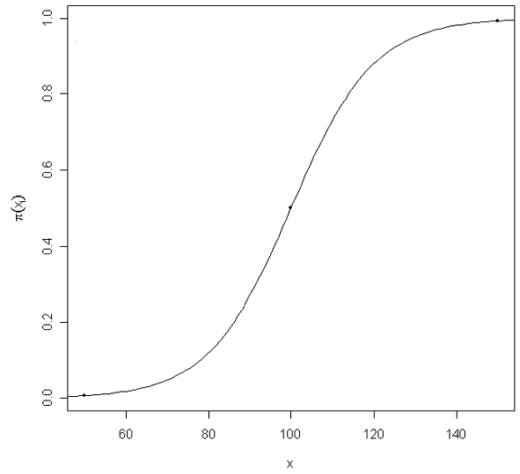

Figura 1: Modelo Logístico com β_1 positivo

- Quando $\beta_1 > 0$, π_i é função decrescente, conforme pode ser visto no gráfico abaixo.

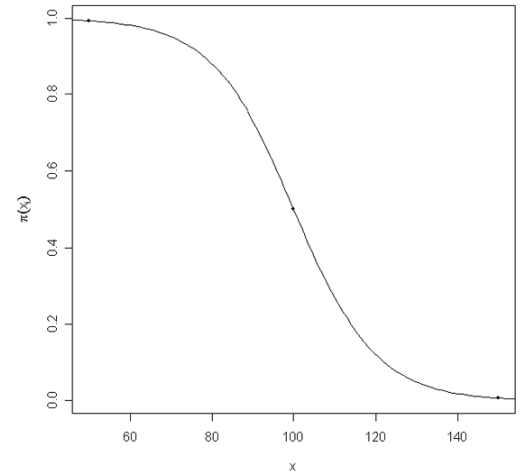

Figura 2: Modelo Logístico com β_1 negativo

- Quando $\beta_1 = 0$, a variável resposta Y é independente da variável x.

2.2.1 Estimação dos Parâmetros do Modelo

Em regressões com uma única variável explicativa, usa-se o modelo regressivo simples. Suponha uma sequência de ensaios de Bernoulli que satisfaz as seguintes condições:

- Em cada ensaio considera-se somente a ocorrência ou a não-ocorrência de um certo evento que será denominado sucesso (S) e cuja não-ocorrência será denominada falha (F)
- Os ensaios são independentes
- A probabilidade de sucesso, que denotaremos por π_i é a mesma em cada ensaio. A probabilidade de falha será denotada por 1-π_i

Suponha uma amostra independente de n observações que possui a seguinte função densidade de probabilidade:

$$P(Y_i = y_i) = \binom{m_i}{y_i} \pi^{y_i} (1 - \pi_i)^{m_i - y_i}$$

Onde:

- x_i é o valor da variável explicativa
- m_i é o número de ensaios na amostra
- y_i é o número de ocorrências do evento em m_i ensaios
- n é o tamanho da amostra

Para estimar os parâmetros (β_0, β_1) do modelo de regressão logística simples, utiliza-se o método de Máxima Verossimilhança. Esse método consiste em estimar os parâmetros de um modelo utilizando as estimativas que tornam máximo o valor da função verossimilhança. Isso é equivalente a achar o máximo do logaritmo da função de verossimilhança.

Assumindo que os valores da amostra aleatória são independentes e identicamente distribuídas (iid), a função de verossimilhança é da seguinte forma:

$$L(\beta_0, \beta_1/Y = y_1, ..., y_n) = P[Y = y_1, ..., y_n/\beta_0, \beta_1] = \prod_{i=1}^{n} \binom{m_i}{y_i} \pi^{y_i} (1 - \pi_i)^{m_i - y_i}$$

$$P[Y = y_1, ..., y_n/\beta_0, \beta_1] = \prod_{i=1}^{n} \binom{m_i}{y_i} \left(\frac{\pi_i}{1 - \pi_i}\right)^{y_i} (1 - \pi_i)^{m_i}$$

O termo $\binom{m_i}{y_i}$ é uma constante que não depende de x_i. Tomando logaritmo em todos os lados da equação anterior, temos:

$$l(\beta_0, \beta_1/Y = y_1, ..., y_n) = \sum_{i=1}^{n} ln \left(\left(\frac{\pi_i}{1 - \pi_i}\right)^{y_i} (1 - \pi_i)^{m_i} \right)$$

$$l(\beta_0, \beta_1/Y = y_1, ..., y_n) = \sum_{i=1}^{n} \left(ln \left(\frac{\pi_i}{1 - \pi_i}\right)^{y_i} + ln(1 - \pi_i)^{m_i} \right)$$

$$l(\beta_0, \beta_1/Y = y_1, ..., y_n) = \sum_{i=1}^{n} \left(y_i ln \left(\frac{\pi_i}{1 - \pi_i}\right) + m_i ln(1 - \pi_i) \right) \tag{3}$$

Considerando a equação (1) e (2) e sabendo que:

1 - $\pi_i = 1 - \left(\frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}\right) = \frac{1}{1 + e^{\beta_0 + \beta_1 x_i}} = \left(1 + e^{\beta_0 + \beta_1 x_i}\right)^{-1}$,

A expressão (3) pode ser escrita como:

$$L(\beta_0, \beta_1/Y = y_1, ..., y_n) = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} m_i ln(1 + e^{\beta_0 + \beta_1 x_i})$$

Os estimadores de máxima verossimilhança para os parâmetros β_0 e β_1 são os valores de $\hat{\beta}_0$ e $\hat{\beta}_1$ que maximizam o logaritmo da função de verossimilhança.

Para maximizar a função de verossimilhança é necessário derivar em relação aos parâmetros e igualar a zero.

$$\frac{\partial \ln(L(\beta_0;\beta_1))}{\partial \beta} = 0 \tag{4}$$

Portanto:

$$\frac{\partial \ln(L(\beta_0;\beta_1))}{\partial \beta_0} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} m_i \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} = 0$$

$$\frac{\partial \ln(L(\beta_0;\beta_1))}{\partial \beta_1} = \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} m_i x_i \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} = 0$$

Porém, essas equações são não-lineares no parâmetro e para resolvê-las é necessário recorrer a métodos numéricos interativos, como Newton-Rapson.

2.2.2 Interpretação dos Parâmetros do Modelos

A interpretação dos parâmetros de um modelo de regressão logística é obtida comparando a probabilidade de sucesso com a probabilidade de não-sucesso, usando a função odds ratio, conforme exemplificado na equação (2).

Assim, ao tomarmos a função logarítmica dois valores distintos da variável explicativa x_j e x_{j+1}, obtemos:

$ln(OR) = ln\left[\frac{g(x_{j+1})}{g(x_j)}\right]$ =ln[g(x_{j+1})] $- ln[g(x_j)] = \beta_0 + \beta_1 x_{j+1} - \beta_0 - \beta_1 x_j = \beta_1(x_{j+1} - x_j)$

Fazendo $x_{j+1} - x_j = 1$, temos:

$$ln(OR) = ln(e^{\beta_1}) = \beta_1$$

Assim, temos o quão provável o resultado ocorrerá entre os indivíduos x_{j+1} tem em relação aos indivíduos x_j.

Para variáveis quantitativas, o parâmetro β_1 representa o efeito médio que uma unidade a mais na variável representada pelo parâmetro interfere na variável resposta.

2.2.3 Variáveis Qualitativas Explicativas com Mais de Duas Categorias

Em um modelo de regressão, quando a variável explicativa é qualitativa com k categorias, é necessário codificar a variável explicativa. Quando temos mais de duas categorias, as variáveis codificadas são chamadas de variáveis dummies, além disso deve-se

especificar a variável de referência em que a odds ratio das demais categorias são comparadas com ela. Desse modo, caso exista k categorias da variável explicativa, temos então k-1 variáveis dummies.

2.2.4 Intervalo de Confiança Para os Parâmetros

A base de construção das estimativas do intervalo de confiança para os parâmetros é a mesma teoria estatística que usamos para os testes de significância do modelo. O intervalo de confiança de 100(1-α)% para o parâmetro β_1 é:

$$\left[\hat{\beta}_1 - z_{1-\alpha/2}DP(\hat{\beta}_1), \hat{\beta}_1 + z_{1-\alpha/2}DP(\hat{\beta}_1)\right]$$

De maneira análoga, o intervalo de confiança do intercepto é dado por:

$$\left[\hat{\beta}_0 - z_{1-\alpha/2}DP(\hat{\beta}_0), \hat{\beta}_0 + z_{1-\alpha/2}DP(\hat{\beta}_0)\right]$$

Onde $z_{1-\alpha/2}$ é o ponto da curva normal padrão que corresponde a 100(1-α)%.

2.2.5 Intervalo de Confiança Para Odds Ratio

Sejam os limites de confiança para o parâmetro β_1:

$\beta_I = \hat{\beta}_1 - z_{1-\alpha/2}DP(\hat{\beta}_1)$ e $\beta_S = \hat{\beta}_1 + z_{1-\alpha/2}DP(\hat{\beta}_1)$

O intervalo de confiança para odds ratio é:

IC(Odds Ratio, 1-α) = $[e^{\beta_I}, e^{\beta_S}]$

2.3 Regressão Logística Múltipla

Assim como no modelo linear, podemos ajustar um modelo para a variável resposta levando em conta mais de uma variável explicativa. Nesse caso, fazemos o uso de notações matriciais para representar o modelo.

Considere um conjunto de p variáveis explicativas denotadas por

X=$(X_1, X_2, ..., X_p)$.

Nesse caso, a função de ligação é denotada por:

$$\text{g(x)=ln } \left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

$$\text{E[Y]=}\pi(X) = \frac{e^{g(x)}}{1+e^{g(x)}}.$$

sendo:

$$\hat{\beta_{px1}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta_{p-1}} \end{pmatrix} \quad X_{px1} = \begin{pmatrix} 1 \\ X_1 \\ \vdots \\ X_{p-1} \end{pmatrix} \quad X_{pxi} = \begin{pmatrix} 1 \\ X_{i1} \\ \vdots \\ X_{i,p-1} \end{pmatrix}$$

As estimativas para os parâmetros segue processo análogo às estimativas dos parâmetros de regressão logística simples.

2.3.1 Inferência no modelo de regressão logística múltipla

Similar ao modelo de regressão logística simples, é necessário testar o nível de significância dos parâmetros. Para isso, podemos utilizar o Teste da Razão de Verossimilhança ou Teste de Wald.

2.3.2 Teste de Razão de Verossimilhança

O teste da razão de verossimilhança avalia a significância dos p coeficientes das variáveis independentes do modelo. As hipóteses de interesse são:

$$\begin{cases} H_0 : \beta_1 = ... = \beta_p \\ H_1 : \beta_j \text{ é diferente de zero para algum j} \end{cases} \tag{5}$$

A estatística G do teste é dada por:

$$G = -2ln \left[\frac{(verossimilhança\ sem\ a\ variável)}{(verossimilhança\ com\ a\ variável)}\right]$$

No caso da regressão logística múltipla, temos o interesse em saber se pelo menos uma variável é significativa para o modelo. Sob a hipótese nula, os p coeficientes são iguais a zero, assim a estatística G tem distribuição Qui-Quadrado com p graus de liberdade.

2.3.3 Teste de Wald

Nesse teste, as hipóteses de interesse são:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases} \tag{6}$$

A estatística-teste é obtida da seguinte forma:

$$W_j = \frac{\hat{\beta}_j}{DP(\hat{\beta}_j)}.$$

Se não rejeitarmos H_0, temos que a variável X_j não explica a variável resposta. De forma equivalente, o teste de Wald pode ser obtido pela multiplicação dos seguintes vetores:

$$W = \hat{\beta}'[I\hat{(}\beta)]^{-1}\hat{\beta} = \hat{\beta}'(X'VX)\hat{\beta}$$

Onde $I\hat{(}\beta)$ é a matriz de informação de Fisher estimada.

2.4 Seleção de Variáveis

A abordagem tradicional na construção de modelos estatísticos é encontrar o modelo mais adequado para explicar os dados. Qualquer procedimento para seleção ou exclusão de variáveis de um modelo é baseado em um algoritmo que verifica a importância das variáveis, incluindo ou excluindo-as do modelo se baseando em uma regra de decisão. Existem algumas técnicas para auxiliar na seleção de variáveis para um modelo de Regressão Logística.

2.4.1 Seleção Forward

Esse procedimento parte da suposição de que não há variável no modelo, apenas o intercepto. A ideia do método é adicionar uma variável de cada vez. A primeira variável selecionada é aquela com maior correlação com a resposta.

2.4.2 Seleção Backward

Esse procedimento faz o processo inverso do método de Forward. Incorpora inicialmente todas as variáveis e depois, por etapas, cada uma pode ser ou não eliminada.

2.4.3 Seleção Stepwise

Stepwise é uma modificação da seleção Forward em que cada passo todas as variáveis do modelo são previamente verificadas com o objetivo de identificar que variável pode ser removida do modelo e que variável pode ser excluída do modelo.

2.5 Técnicas de Diagnóstico em Análise de Regressão Logística

Estatística de ajuste do modelo como χ^2 de Pearson, Deviance e o Teste de Hosmer e Lemeshow são indicadores da qualidade do modelo como um todo. Entretanto, é necessário verificar também o diagnóstico do modelo para cada ponto, isto é, verificar a dependência do modelo estatístico em relação às várias observações que foram coletadas. Assim, será realizado técnicas de diagnóstico para verificar se as suposições iniciais do modelo de regressão logística estão sendo satisfeitas e identificar características que podem estar atrapalhando na aplicação da técnica, tais como observações influentes, que causem alguma mudança na estimativa dos coeficientes, levando a problemas nas conclusões geradas pelo modelo.

Para detectar esses problemas, existem várias técnicas da regressão linear normal que podem ser utilizadas na Regressão Logística. A seguir, serão apresentadas algumas dessas técnicas, considerando um modelo com p variáveis explicativas que formam J padrões de covariáveis, m_j é o número de indivíduos que apresentam os mesmos valores para cada uma das p variáveis independentes e segue que $\sum m_j = 0$.

2.5.1 Teste de Hosmer & Lemeshow

O Teste de Hosmer & Lemeshow é muito utilizado em regressão logística com a finalidade de testar a bondade do ajuste, ou seja, o teste comprova se o modelo proposto pode explicar bem o que se observa. Hosmer & Lemeshow (1980, 1982) propuseram um procedimento que utiliza os valores de probabilidade preditos para criar grupos. As hipóteses do teste são:

$$\begin{cases} H_0 : E(Y) = \frac{e^{X'\beta}}{1+e^{X'\beta}}, \text{ou seja, o modelo está bem ajustado.} \\ H_1 : E(Y) \neq \frac{e^{X'\beta}}{1+e^{X'\beta}}, \text{ou seja, o modelo não está bem ajustado.} \end{cases} \tag{7}$$

A bondade do teste é baseada na divisão da amostra segundo suas probabilidades ajustadas com base nos valores dos parâmetros estimados pela regressão logística. Os valores ajustados são dispostos do menor para o maior e em seguida separados em g grupos de tamanho aproximadamente igual. Os autores recomendam criar g=10 grupos de aproximada-

mente mesmo tamanho. Os grupos são criados de modo que o primeiro tenha probabilidade predita entre 0,0 e 0,1, o segundo, entre 0,1 e 0,2 e assim sucessivamente até que o décimo grupos tenha valores de probabilidade predita entre 0,9 e 1,0. Se as frequências esperadas em alguns dos grupos forem muito pequenas, a estatística do teste de Hosmer & Lemeshow é calculada, entretanto, pode não ser confiável. Neste caso, devemos especificar um número menor de grupos, contudo não se pode utilizar menos de 3 grupos, pois com g ≤ 3 a estatística do teste é impossibilitada de ser calculada.

Após a divisão dos valores preditos em grupos, são calculados os valores esperados para cada grupos e comparados com os valores observados usando a estatística Qui-Quadrado de Pearson. Hosmer & Lemeshow (1980) mostrou por simulação que a estatística do teste segue, aproximadamente, uma distribuição Qui-Quadrado com g-2 graus de liberdade quando o modelo está especificado corretamente.

2.5.2 *Leverage*

Assim como no modelo linear, uma métrica para diagnosticar outliers é a leverage (diagonal da matriz chapéu). No modelo linear, a matriz chapéu é definida por:

$$H = X(X'X)^{-1}X'$$

Os resíduos, $\hat{\epsilon}$= Y-$\hat{Y}$, podem ser expressos em função da matriz H, como $\hat{\epsilon}$=(I-H)y, onde I é a matriz identidade JxJ. Os elementos da diagonal principal da matriz H são denominados h_{ii}.

Usando a regressão de Mínimos Quadrados ponderados como modelo, Pregibon (1981) realizou uma aproximação linear para os valores ajustados definindo a matriz H para regressão logística por:

$$H = V^{1/2}X(X'VX)^{-1}X'V^{1/2}$$

Sendo:

- V é a matriz diagonal JxJ com elemento $v_j = m_j\hat{\pi}(x_j)(1 - \hat{\pi}(x_j))$
- $b_j = x_j^{'}(X'VX)^{-1}x_j^{'}$

Desse modo, o modelo de regressão logística, h_j para j-ésima diagonal de H são dados por:

$$h_j = m_j\hat{\pi}_j(1-\hat{\pi}_j)x_j^{'}(X'VX)^{-1}x_j^{'} = v_j b_j \tag{8}$$

A matriz H é simétrica e idempotente. Tem-se que:

- $0 \leq h_j \leq 1$
- tr(H)=$\sum_{j=1}^{n} h_j = p+1$

Como se pode ver por (7), o elemento h_j só depende dos valores das variáveis explicativas, ou seja, da matriz X, e não envolve as observações de y. Se h_j é grande, os valores das variáveis explicativas associados a j-ésima observação são atípicos, ou seja, estão distantes do vetor de valores médios das variáveis explicativas.

Segundo Cordeiro e Lima Neto (2004), esses pontos exercem um papel importante no ajuste final dos parâmetros de um modelo estatístico, ou seja, sua exclusão pode implicar mudanças dentro de uma análise estatística. Esses pontos discrepantes podem ser também informativos com relação à estimativa de β.

No caso da regressão linear, se $h_y \geq 2(p+1)/n$, onde p é o número de parâmetros do modelo, então a observação j é considerada um outlier. Entretanto, Hosmer & Lemeshow advertem que quando o número de padrões das covariáveis (J) é muito menor do que n, existe ponto de falha para identificar os pontos de influência. Desse modo, outras medidas devem ser utilizadas para confirmar esse primeiro diagnóstico.

2.5.3 Resíduos

Os resíduos são utilizados para avaliar a adequabilidade de um modelo, sendo utilizado para avaliar sua capacidade preditiva e definido a partir dos próprios dados utilizados na determinação do modelo. Modelos de regressão com bom desempenho estatístico apresentam pequena discrepância residual. Em muitas aplicações, o coeficiente de determinação (R^2) é utilizado como indicador numérico que permite comparar o desempenho de diferentes modelos. A seguir são definidos os resíduos de Pearson, resíduos Deviance e resíduos Stu-

dentizados, que são úteis para identificar observações que não estão se adequando bem ao modelo.

1. Resíduos Padronizados de Pearson

   Os resíduos de Pearson são definidos como uma comparação entre a diferença de um valor observado e o seu respectivo valor estimado com o desvio padrão estimado para essa observação, conforme a equação a seguir:

   $$r(y_j, \hat{\pi}_j) = \frac{y_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}}.$$

   O resíduo padronizado de Pearson para um padrão de covariável x_j é definido por:

   $$r_{sj} = \frac{r(y_j, \hat{\pi}_j)}{\sqrt{1-h_j}} \tag{9}$$

2. Resíduo Deviance

   O resíduo Deviance mede o grau de discordância entre o máximo da função de verossimilhança observada e da estimada. Como a regressão logística usa o princípio da máxima verossimilhança, o objetivo é minimizar a soma dos resíduos Deviance.

   Os resíduos Deviance são definidos por:

   $$r(y_j, \hat{\pi}_j) = \pm \left\{ 2 \left[y_j ln\left(\frac{y_j}{m_j\hat{\pi}_j}\right) + (m_j - \hat{y}_j) ln\left(\frac{(m_j - y_j)}{m_j(1-\hat{\pi}_j)}\right) \right] \right\}^{1/2} \tag{10}$$

3. Resíduos Studentizados

   O resíduo studentizado é o quociente entre o resíduo e a estimativa de seu desvio-padrão. O resíduo studentizado é definido de maneira a garantir a independência entre numerador e denominador na padronização dos resíduos. Assim, define-se o resíduo studentizado por:

   $$r_{ij} = \frac{y_j - \hat{\pi}_j}{\sqrt{\hat{\pi}_j(1-\hat{\pi}_j)}\sqrt{(1-h_j)}} \tag{11}$$

Medidas como h_i ou os resíduos são úteis para detectar valores extremos mas não mostram qual impacto que esses valores têm nos vários aspectos do ajuste do modelo, como por exemplo, nos parâmetros estimados e nas estatísticas de qualidade de ajuste do modelo. A seguir serão apresentadas três técnicas para detectar medidas de influência e que tem como objetivo quantificar o efeito de cada observação no ajuste do modelo.

2.5.4 Medidas de Influência

As medidas de influência verificam quanto uma observação contribui em modificações nas estimativas dos parâmetros do modelo de regressão. Geralmente, o efeito que uma observação tem nos parâmetros do modelo é verificado através da exclusão dessa observação. As técnicas do diagnóstico do modelo mais usadas para detectar esse tipo de influência são: distância de Cook, DFFITS e DFBETAS.

- Distância de Cook

  A distância de Cook (1977) tem como objetivo medir a influência de cada observação nos parâmetros do modelo. Essa medida é calculada através da diferença entre $\hat{\beta}$ e $\hat{\beta}_{-j}$ que representam, respectivamente, as estimativas de máxima verossimilhança calculadas usando todos os j das variáveis e excluindo as observações com padrão j e ainda padronizada pela matriz de covariância de $\hat{\beta}$. Pregibon (1981) mostrou através de uma aproximação linear que a distância de Cook para a regressão logística é dada por:

$$\triangle\hat{\beta}_j = (\hat{\beta} - \hat{\beta}_{-j})'(X'VX)(\hat{\beta} - \hat{\beta}_{-j}) = \frac{r_j^2 h_j}{(1-h_j)^2} = \frac{r_{sj}^2 h_j}{(1-h_j)} \tag{12}$$

  Alguns autores,como Neter et. al. (1996) e Cordeiro e Lima Neto (2004) sugerem que observações que apresentam $\triangle\hat{\beta}_j \leq$F(p+1,n-p-1) podem ser consideradas influentes. Nesse caso, recomendas-se observar o que acontece com o ajuste do modelo quando essas observações são excluídas.

- DFFITS

  Proposta por Besley et al (1980), DFFIT é uma medida alternativa à distancia de Cook e tem como objetivo medir a influência da j-ésima observação nos parâmetros de locação

e escala do modelo. O DFFIT está em função do resíduo studentizado r_{ij} e da medida de Levarage h_j, dado por:

$$DFFITS_j = r_{ij} \left\{ \frac{h_j}{1 - h_j} \right\} \tag{13}$$

Neter et al (1996) indicam verificar os casos com valor absoluto de DFFITS superiores a 1 para amostras pequenas e observações com valor absoluto de DFFITS maior ou igual a $2\sqrt{(p+1)/n}$ para amostras grandes.

- DFBETAS

  $DFBETA_j$ mede o quanto cada coeficiente de regressão relacionada a uma variável independente X_j se modifica quando a observação j é excluída. A medida de $DFBETA_j$ é definida por:

$$\hat{\beta} - \hat{\beta}_j = (X'X)^{-1} x^{'} (1 - h_j)^{-1} \hat{\epsilon}_j \tag{14}$$

  Segundo Neter et al (1996) , deve-se dar maior atenção a observações que apresentarem valores absolutos de DFBBETAS superiores a 1, para amostras pequenas, e valores absolutos de DFBETAS maiores do que $1/\sqrt{n}$, para amostras grandes.

2.5.5 Análise gráfica

Geralmente, examinam-se as medidas definidas anteriormente através de gráficos, plotando os valores de h_j versus os valores preditos ou mesmo versus o número de cada observação. Esses dois tipos de gráficos apresentam resultados similares. Através desses gráficos é possível localizar as observações que estão muito afastadas do restante do conjunto de dados. Deve-se dar mais atenção aos pontos que apresentam valores muito altos para as medidas de diagnóstico.

Para ilustrar o uso da regressão logística e os gráficos para identificar as medidas de influência que auxiliam na análise de diagnóstico, serão usados os dados do exemplo 14.1 do livro Applied Linear Statistical Models, Neter et. al. (1996). O problema consiste em um estudo realizado com 25 estudantes para analisar a capacidade concluir com êxito um

complexo teste de programação. Os estudantes selecionados tinham quantidades diferentes de tempo de estudo na área de computação, dada em meses, conforme mostra a tabela 1 coluna 1. A todos os estudantes foi dado o mesmo teste de programação e o resultado está na coluna 2, sendo essa uma variável dicotômica, com Y=1 se o estudante concluiu o teste corretamente e Y=0 se o estudante não conseguiu concluir o teste no tempo determinado. O modelo de regressão é dado por:

$$\hat{g}(x) = \beta_0 + \beta_1 \ln(\text{meses de experiência})$$

Tabela 1: Dados

Estudante (i)	**Meses de Experiência**	**Resultado no Teste**	**Valor Ajustado**
1	14	0	0.310262
2	29	0	0.835263
3	06	0	0.109996
4	25	1	0.726602
5	18	1	0.461837
6	04	0	0.082130
7	18	0	0.461837
8	12	0	0.245666
9	22	1	0.620812
10	6	0	0.109996
11	30	1	0.856299
12	11	0	0.216980
13	30	1	0.856299
14	5	0	0.095154
15	20	1	0.542404
16	13	0	0.276802
17	9	0	0.167100
18	32	1	0.891664
19	24	0	0.693379
20	13	1	0.276802
21	19	0	0.502134
22	4	0	0.082130
23	28	1	0.811825
24	22	1	0.620812
25	8	1	0.145815

As estimativas obtidas para os parâmetros são $\hat{\beta}_0$= -3,0597 e $\hat{\beta}_1$= 0,1615. Como o teste de Hosmer & Lemeshow apresenta χ^2 = 5,1453 (p= 0,5253), pode-se concluir que o modelo está bem ajustado pela regressão logística.

Na figura 3, os valores de h_j aparecem plotados versus o número das observações. Como sugerido por Hosmer e Lemeshow (1989), ao invés de usar o resíduo Deviance, será usado o Resíduo Deviance elevado ao quadrado. Analisando esse gráfico pode-se perceber que os pontos 2 e 25 distanciam-se dos demais pontos, destacando-se.

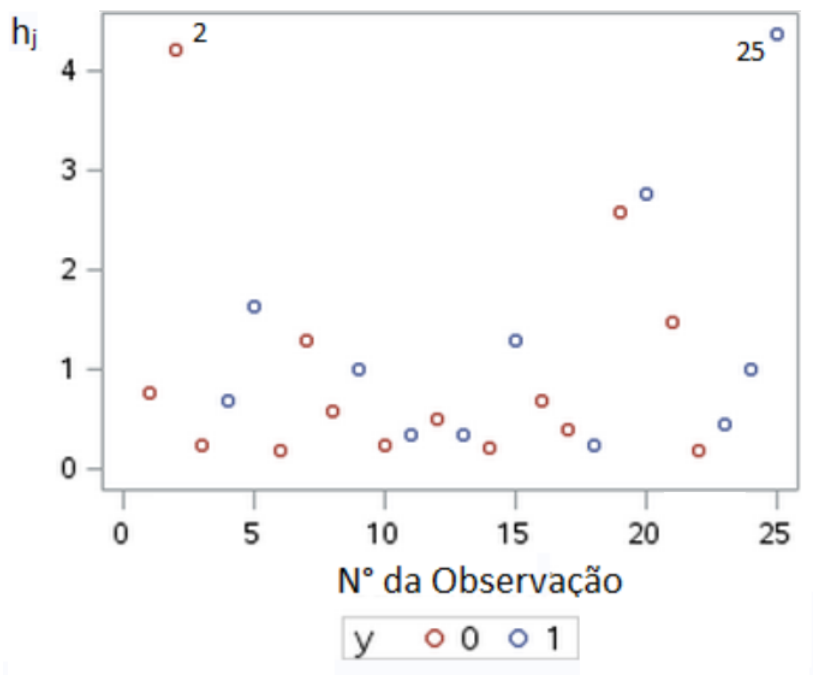

Figura 3: Valores de h_j versus o número da observação

As figuras 4, 5 e 6 apresentam os gráficos dos Resíduos versus o número das observações. É importante observar que em todos os gráficos residuais os pontos 2 e 25 são observações discrepantes.

Quando se avalia o gráfico de DfBetas versus número das observações, pode-se perceber que as observações 2 e 25 estão afastadas do restante do conjunto de dados.

Ao eliminar as observações 2 e 25, obtêm-se as novas estimativas para os parâmetros do modelo: $\hat{\beta}_0$= -5,6016 e $\hat{\beta}_1$= 0,3022. As novas estimativas representam modificação de aproximadamente 83% e 87%, respectivamente.

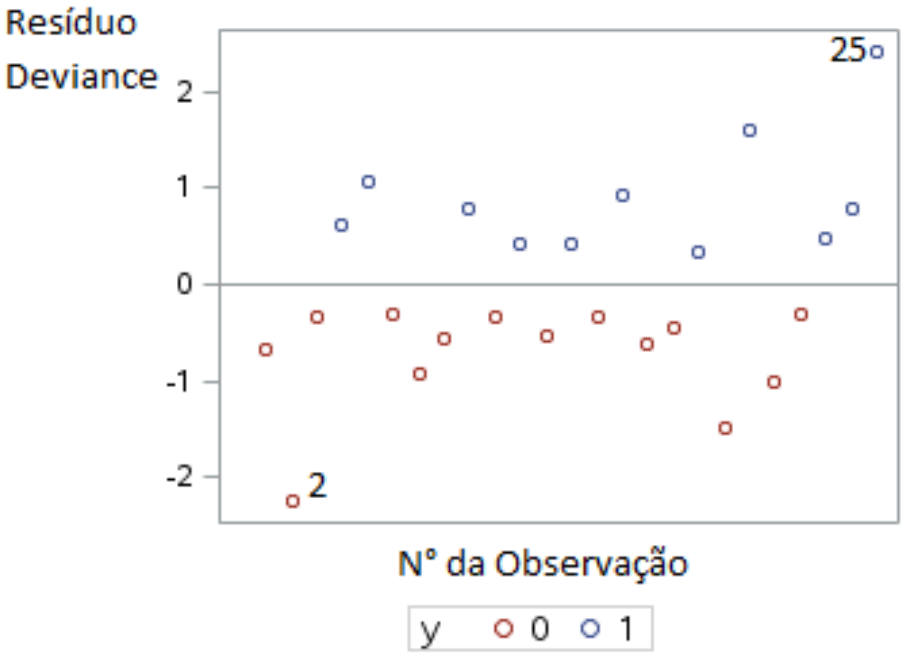

Figura 4: Resíduo Deviance versus o número da observação

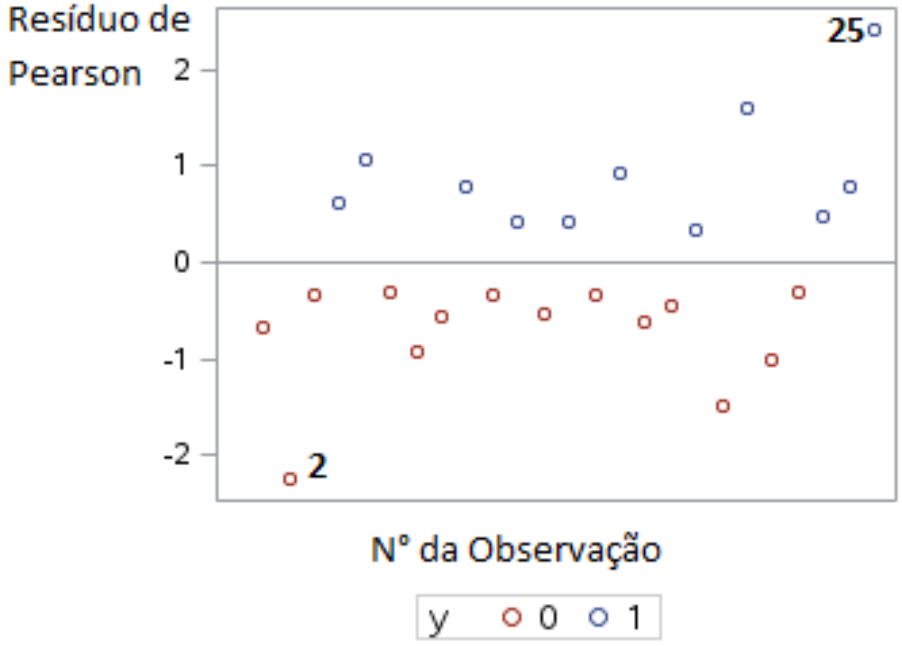

Figura 5: Resíduo de Pearson versus o número da observação

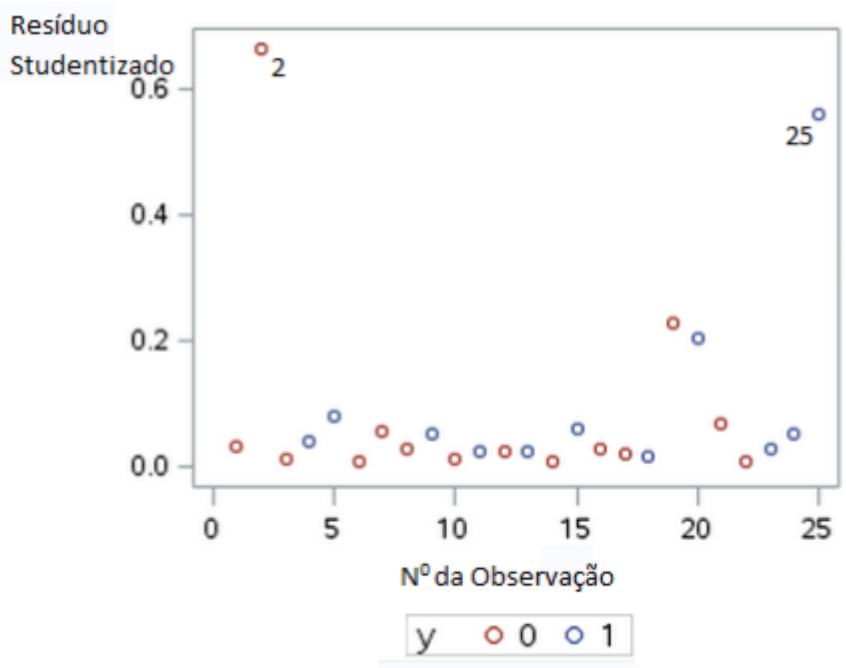

Figura 6: Resíduo Studentizado versus o número da observação

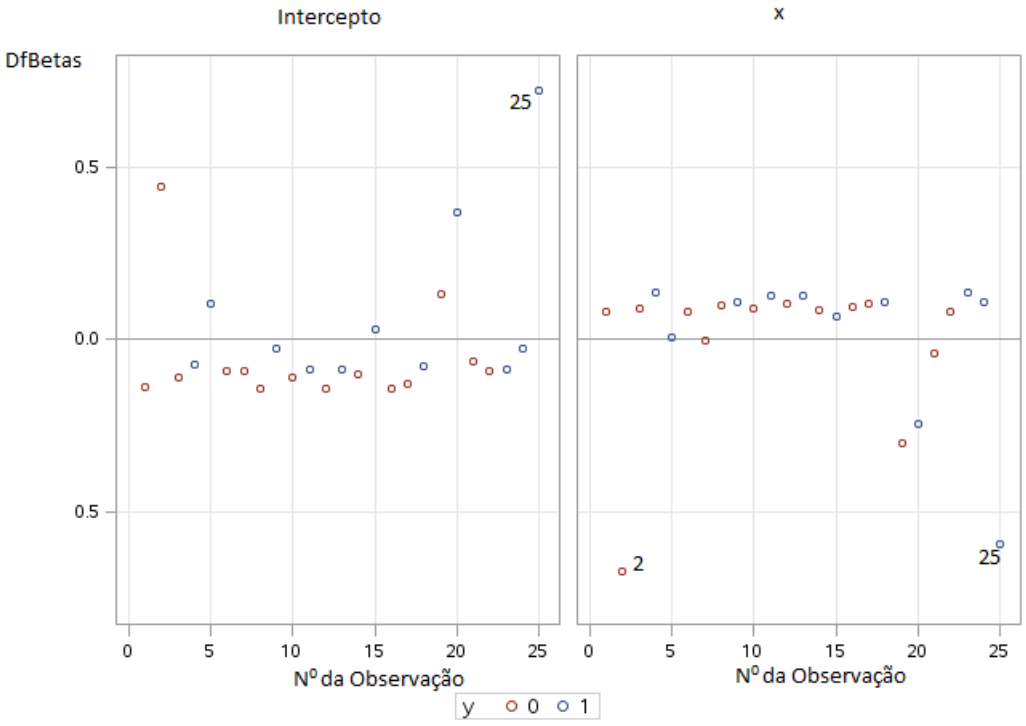

Figura 7: DfBetas versus o número da observação

Capítulo 3

3 Aplicação

3.1 Introdução

A Pesquisa Nacional de Saúde (PNS) é uma pesquisa de base domiciliar, de âmbito nacional, realizada em parceria com o Instituto Brasileiro de Geografia e Estatística (IBGE) com periodicidade de 5 anos. O inquérito é composto por três questionários: o domiciliar, referente às características do domicílio, nos moldes do censo demográfico e da PNAD; o relativo a todos os moradores do domicílio, que dará continuidade ao Suplemento Saúde da PNAD; e o individual, a ser respondido por um morador de 18 anos e mais do domicílio, selecionado com equiprobabilidade entre todos os residentes elegíveis, com enfoque nas principais doenças crônicas não transmissíveis, aos estilos de vida, e ao acesso ao atendimento médico. Neste capítulo, serão analisados os dados da Pesquisa Nacional de Saúde – PNS 2013, cujo objetivo foi avaliar a percepção do estado de saúde, estilos de vida e doenças crônicas.

Para descrever aspectos relacionados as condições de saúde da população brasileira, a PNS abordou a percepção individual de saúde, que consiste na percepção que os indivíduos possuem de sua própria saúde, em várias dimensões. Investigou-se, particularmente, a auto avaliação de saúde, indicador que tem sido utilizado, nacional e internacionalmente, para estabelecer diferenças de morbidade em subgrupos populacionais, comparar necessidades de serviços e recursos de saúde por área geográfica, bem como para calcular outros indicadores de morbi-mortalidade, tais como a esperança de vida saudável. Esse indicador engloba tanto componentes físicos quanto emocionais dos indivíduos, além de aspectos do bem-estar e da satisfação com a própria vida. Os dados foram obtidos através de uma questão única em que o próprio morador classifica sua saúde em uma escala de 5 graus: Muito boa, boa, regular, ruim ou muito ruim. A percepção do indivíduo sobre a saúde vai além das sensações físicas de dor e desconforto, mas, sobretudo, das consequências sociais e psicológicas da presença da enfermidade.

O presente trabalho tem como objetivo identificar os fatores associados a auto-

percepção de saúde utilizando regressão logística.

3.2 Dados

A elaboração da PNS foi fundamentada em três eixos principais: o desempenho do sistema nacional de saúde; as condições de saúde da população brasileira; a vigilância das doenças crônicas não transmissíveis e fatores de risco associado. Com desenho próprio, elaborado especificamente, para coletar informações de saúde, a PNS foi planejada para a estimação de vários indicadores com a precisão desejada e para assegurar a continuidade no monitoramento da grande maioria dos indicadores do Suplemento Saúde da PNAD. O planejamento amostral foi feito com base na Amostra Mestra. Esse tipo de amostragem caracteriza-se em um conjunto de unidades de áreas que são consideradas Unidades Primárias de Amostragem (UPA's). Diante disso, cada UPA é estratificada e ocorre uma seleção, com probabilidade proporcional ao tamanho, dos domicílios particulares contidos na UPA. A amostra da PNS é por conglomerados em três estágios de seleção:

- 1° estágio: Seleção da subamostra de UPA's em cada estrato da amostra mestra;
- 2° estágio: seleção por amostragem aleatória simples de domicílios em cada UPA selecionada no primeiro estágio;
- 3° estágio: seleção por amostragem aleatória simples do adulto (pessoa com 18 ou mais anos de idade) entre todos os moradores adultos do domicílio.

Para analisar os fatores relacionados à doenças crônicas e problemas psicológicos, foram utilizadas 21 variáveis para explicar a percepção individual de saúde. Essas variáveis explicativas podem ser agrupadas em 5 grupos: problemas de locomoção, dor ou desconforto no peito, incômodos no dia a dia, problemas de audição e problemas de visão.

As variáveis demográficas inseridas no modelo podem ajudar a identificar um comportamento diferenciado e serão úteis para ajudar no ajuste e nas interpretações da regressão logística.

Tabela 2: Variáveis explicativas usadas no modelo de regressão logística

Variável	**Nomenclatura**
Locomoção	
NOO2	O Sr(a) utiliza algum recurso para auxiliar na locomoção ?
NOO3	Que grau de dificuldade o Sr(a) tem para se locomover ?
Dores no peito	
NOO4	O Sr(a) sente algum desconforto no peito ao exercitar-se?
NOO5	O Sr(a) sente algum desconforto no peito ao caminhar ?
NOO6	O quer o Sr(a) faz quando sente dor no peito ?
NOO7	Se o/a Sr(a) parar, o que acontece com a dor no peito ?
NOO8	O/A Sr(a) pode me mostrar onde sente essa dor no peito ?
Incômodos no dia a dia	
NO10	Com que frequência o Sr(a) tem problemas com sono ?
NO11	Com que frequência o Sr(a) sente-se indisposto durante o dia ?
NO12	Com que frequência o Sr(a) não sentiu prazer no dia a dia ?
NO13	Com que frequência o Sr(a) teve problemas de concentração ?
NO14	Com que frequência o Sr(a) teve problemas na alimentação ?
NO15	Com que frequência o Sr(a) teve lentidão para movimentar-se ?
NO16	Com que frequência o Sr(a) se sentiu deprimido ?
NO17	Com que frequência o Sr(a) se sentiu mal consigo mesmo ?
NO18	Com que frequência o Sr(a) pensou em se ferir ou suicidar-se ?
Problemas auditivos	
NO19	O Sr(a) faz uso de aparelho auditivo ?
NO20	Em geral, que grau de dificuldade o Sr(a) tem para ouvir ?
Problemas de visão	
NO21	O Sr(a) usa algum tipo de recurso para auxiliar a enxergar ?
NO22	Em geral, qual grau de dificuldade para ver de longe ?
NO23	Em geral, qual grau de dificuldade para ver de perto ?

Tabela 3: Variáveis demográficas usadas no modelo de regressão logística

Variável demográfica	**Nomenclatura**
COO6	Sexo
COO8	Idade
COO9	Cor ou raça
CO10	Vive com cônjuge ou companheiro
CO11	Estado civil
CO12	Informante
DOO3	Escolaridade
EOO1	Ocupação

No presente trabalho foi utilizado o software estatístico SAS para análise dos resultados. No SAS, a análise dessa técnica é feita através do PROC LOGISTIC sendo obrigatório colocar a classe de referência para as variáveis qualitativas e avaliar o grau de ajustamento do modelo, de acordo com o Teste de Hosmer & Lemeshow.

3.3 Análise univariada

Neste estudo serão consideradas apenas as duas categorias de percepção extremas tendo em vista que em perguntas de opinião as pessoas tendem a responder os valores intermediários. Optou-se então por utilizar aqueles que tinham uma percepção mais definida sobre sua saúde.

Tabela 4: Distribuição empírica dos indivíduos segundo auto avaliação da saúde

Classificação	**n**	**%**
Muito ruim	765	9,33
Muito boa	7433	90,67
TOTAL	8198	100

Pode-se observar que a maioria auto avaliou sua saúde como muito boa. Apenas 9,33% se auto avaliaram como muito ruim.

3.3.1 Análise descritiva

Segundo a PNS, em 2013, no Brasil, havia 146,3 milhões de pessoas com 18 anos ou mais de idade, destas, 66,1% autoavaliaram sua saúde como boa ou muito boa. As estimativas variaram de 56,7%, no Nordeste, a 71,5%, no Sudeste. O gráfico da figura 8, extraída do sítio da biblioteca virtual do Instituto Brasileiro de Geografia e Estatística, torna evidente esse fato.

Em relação ao sexo, 70,3% dos homens consideraram sua saúde como boa ou muito boa, contra 62,4% das mulheres. Em relação aos grupos de idade, quanto maior a faixa etária menor o percentual, que variou de 81,6%, para aqueles de 18 a 29 anos de idade, a 39,7%, para as pessoas de 75 anos ou mais de idade. Em relação à escolaridade, observou-se que, conforme maior o grau de instrução, maior o percentual daqueles que consideraram sua saúde boa ou muito boa. Entre as pessoas sem instrução ou com o fundamental incompleto, o percentual foi de 49,2%, enquanto para aquelas com superior completo foi de 84,1%. O gráfico da figura 9, extraída do sítio da biblioteca virtual do Instituto Brasileiro de Geografia e Estatística, mostra a percepção do estado de saúde segundo sexo, idade, cor ou raça e escolaridade.

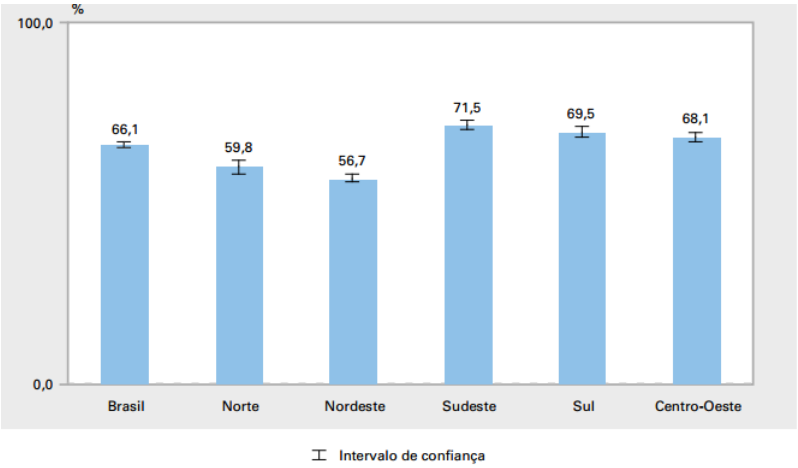

Figura 8: Proporção de pessoas de 18 anos ou mais de idade com autoavaliação de saúde boa ou muito boa, com indicação do intervalo de confiança de 95%, segundo as Grandes Regiões

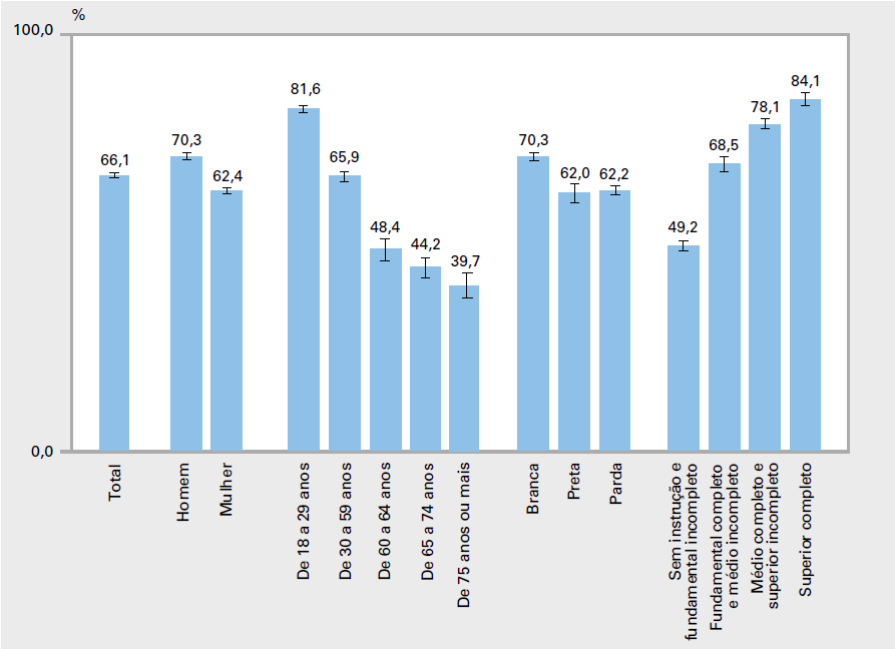

Figura 9: Proporção de pessoas de 18 anos ou mais de idade com autoavaliação de saúde boa ou muito boa, com indicação do intervalo de confiança de 95%, segundo sexo, grupos de idade, cor ou raça e escolaridade

Para o banco de dados que foi aplicado a regressão logística, as variáveis apresentam valores similares.

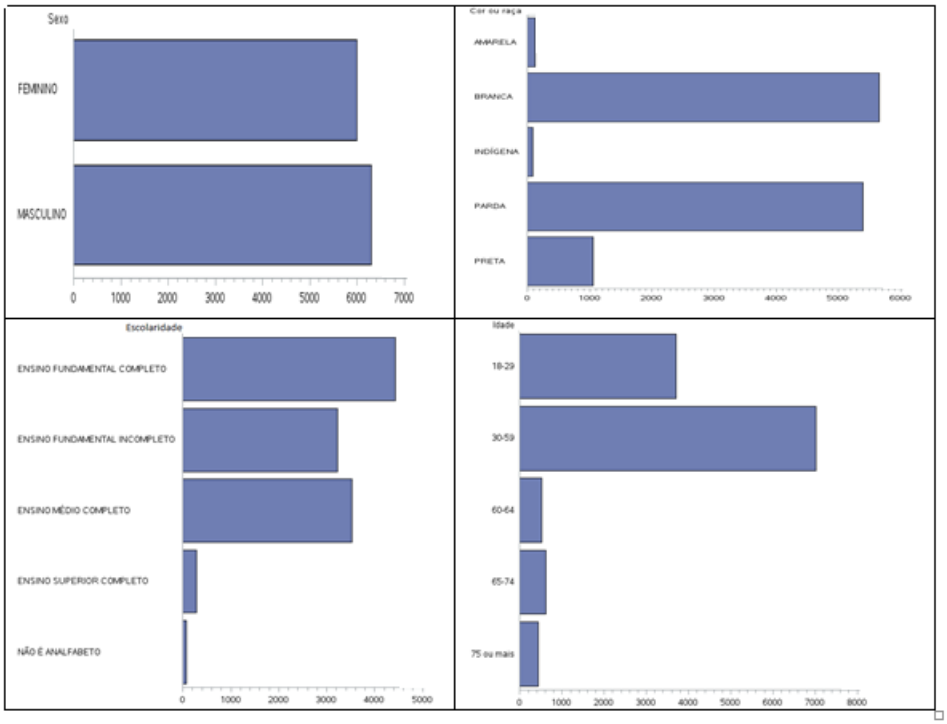

Figura 10: Proporção de pessoas da amostra de regressão logística segundo sexo, grupos de idade, cor ou raça e escolaridade

3.3.2 Teste χ^2 de independência

Para verificar se há relação entre as variáveis explicativas e a variável independente, é necessário fazer o teste de χ^2. O teste do χ^2 verifica se as proporções de sujeitos ou dados em cada categoria são iguais ou diferentes das que seriam esperadas pelo acaso ou pelo conhecimento prévio a respeito do problema (ou seja, se há alguma relação entre as duas variáveis capaz de desviar essas proporções de maneira significativa dos valores que elas deveriam ter caso não houvesse qualquer relação).

Para o banco de dados usado para regressão logística, os valores são bastante similares.

Tabela 5: χ^2 das variáveis explicativas em relação à variável independente (continua)

Variável explicativa		**Sucesso**	**Insucesso**	χ^2	p-valor
COO6	Masculino	264	3617	55,7208	<.0001
	Feminino	501	3816		
	Branca	291	3846	53,8323	<.0001
	Preta	81	589		
COO9	Amarela	6	79		
	Parda	379	2884		
	Indígena	8	35		
CO10	Sim	368	3947	6,9456	0,0084
	Não	397	3486		
	Casado(a)	275	2923	263,6527	<.0001
	Seperado(a)	34	205		
CO11	Divorciado(a)	62	476		
	Viúvo(a)	136	322		
	Solteiro(a)	258	3507		
	A própria	571	5341	135,0444	<.0001
CO12	Outro morador	170	2004		
	Não morador	24	88		
	Não é analfabeto	16	20	784,4545	<.0001
	Ens. Fund. Inc.	395	1515		
DOO3	Ens. Fun. Comp.	111	2832		
	Ens. Méd. Comp.	40	2650		
	Ens. Sup. Comp.	3	219		
EOO1	Ocupado	243	5465	571,9336	<.0001
	Não ocupado	522	1968		
NOO2	Sim	139	52	930,3226	<.0001
	Não	626	7381		
	Nenhum	342	7292	3241,4338	<.0001
	Leve	113	66		
NOO3	Médio	109	20		
	Intenso	143	9		
	Não Consegue	58	46		
	Sim	266	243	1619,7099	<.0001
NOO4	Não	402	7133		
	Não se aplica	39	11		
NOO5	Sim	154	69	1111,7592	<.0001
	Não	514	7307		

Tabela 6: χ^2 das variáveis explicativas em relação à variável independente (continua)

	Variável explicativa	**Sucesso**	**Insucesso**	χ^2	p-valor
	Para ou diminui a velocidade	234	202	19,9463	<.0001
NOO6	Continua após tomar remédio	15	2		
	Continua caminhando	22	46		
	É aliviada em menos de 10 min.	158	207	37,3701	<.0001
NOO7	É aliviada em mais de 10 min.	88	32		
	Não é aliviada	25	11		
	Acima ou no meio do peito	210	181	1,8896	.5956
NOO8	Abaixo do peito	45	52		
	Braço esquerdo	10	10		
	Outros	6	7		
	Nenhum dia	228	6044	1298,9596	<.0001
NO10	Menos da metade dos dias	142	745		
	Mais da metade dos dias	123	213		
	Quase todos os dias	242	431		
	Nenhum dia	224	5848	1471,5330	<.0001
NO11	Menos da metade dos dias	158	1076		
	Mais da metade dos dias	124	223		
	Quase todos os dias	259	286		
	Nenhum dia	297	6392	1724,2650	<.0001
NO12	Menos da metade dos dias	160	799		
	Mais da metade dos dias	87	113		
	Quase todos os dias	221	129		
	Nenhum dia	357	6646	1579,4755	<.0001
NO13	Menos da metade dos dias	137	590		
	Mais da metade dos dias	88	101		
	Quase todos os dias	183	96		
	Nenhum dia	401	6619	898,4969	<.0001
NO14	Menos da metade dos dias	137	483		
	Mais da metade dos dias	80	156		
	Quase todos os dias	147	175		
	Nenhum dia	385	6889	1482,9446	<.0001
NO15	Menos da metade dos dias	141	353		
	Mais da metade dos dias	71	90		
	Quase todos os dias	168	101		
	Nenhum dia	284	6540	1901,6780	<.0001
NO16	Menos da metade dos dias	169	671		
	Mais da metade dos dias	106	115		
	Quase todos os dias	206	107		

Tabela 7: χ^2 das variáveis explicativas em relação à variável independente (conclusão)

Variável explicativa		**Sucesso**	**Insucesso**	χ^2	p-valor
NO17	Nenhum dia	448	6935	1212,4082	<.0001
	Menos da metade dos dias	125	365		
	Mais da metade dos dias	69	66		
	Quase todos os dias	123	67		
NO18	Nenhum dia	601	7338	942,6666	<.0001
	Menos da metade dos dias	84	64		
	Mais da metade dos dias	26	14		
	Quase todos os dias	54	17		
NO19	Sim	4	11	5,3374	0,0209
	Não	761	7422		
NO20	Nenhum	558	7116	682,5033	<.0001
	Leve	116	235		
	Médio	54	62		
	Intenso	34	14		
NO21	Não Consegue	3	6		
	Sim	375	2690	48,7709	0,0209
	Não	390	4743		
NO22	Nenhum	354	5912	836,7372	<.0001
	Leve	133	853		
	Médio	109	494		
	Intenso	137	156		
	Não Consegue	32	18		
NO23	Nenhum	313	5854	961,7372	<.0001
	Leve	149	1000		
	Médio	148	433		
	Intenso	136	130		
	Não Consegue	19	16		

Apenas a variável NOO8 possui alto p-valor, indicando que a mesma não possui associação com o logito.

3.4 Análise multivariada: seleção do modelo

Para a construção do modelo multivariado, serão considerados os métodos de seleção de variáveis apresentados na seção 2.4.

O modelo foi construído considerando um esquema de amostragem aleatória simples, uma vez que o tamanho da amostra é grande e portanto as estimativas das variâncias dos estimadores não se alterariam muito se fosse considerado o plano amostral na estimação dos parâmetros.

Como modelo com todas as variáveis em estudo necessita de um considerável esforço computacional, utilizou-se o método de seleção automática stepwise para selecionar o

melhor subconjunto de variáveis. O critério de adição ou remoção de covariáveis é baseado na estatística G^2, comparando modelos com e sem as variáveis em questão. Ao aplicar esse método, o modelo reduziu-se a dez variáveis: COO6, COO8, COO9, CO11, DOO3, EOO1, NOO2, NOO5, NO14, NO20.

O coeficiente estimado, o erro padrão e o valor da estatística do Teste de Wald relativo às variáveis explicativas com o logito são apresentados na tabela 8.

Tabela 8: Coeficiente estimado ($\hat{\beta}$), erro padrão (EP($\hat{\beta}$)) e χ^2 - Wald referentes à regressão logística multivariada das variáveis explicativas em relação à variável resposta.

Variável explicativa		n	$\hat{\beta}$	EP($\hat{\beta}$)	χ^2 - Wald
COO6	Feminino	6004	0,57	0,13	18,94
	Masculino	6300			
COO8	Idade	12304	0,04	0,004	109,22
COO9	Indígena	80	0,59	0,65	0,84
	Preta	1053	0,34	0,21	2,62
	Amarela	114	0,08	0,69	0,01
	Parda	5395	0,56	0,13	18,92
	Branca	5662			
CO11	Solteiro(a)	5995	0,15	0,14	1,13
	Separado(a)	318	0,40	0,29	1,91
	Divorciado(a)	697	0,35	0,22	2,59
	Viúvo(a)	664	-1,09	0,24	21,53
	Casado(a)	4630			
DOO3	Não é analfabeto	65	2,75	0,46	35,00
	Ens. Fund. Inc.	3220	2,41	0,20	142,63
	Ens. Fun. Comp.	4433	0,85	0,22	15,13
	Ens. Méd. Comp.	3529			
	Ens. Sup. Comp.	282	-1,23	1,09	1,28
EOO1	Ocupado	8548	-0,67	0,12	28,10
	Não ocupado	3756			
NOO2	Sim	191	2,84	0,39	53,26
	Não	8007			
NOO5	Sim	223	2,56	0,22	137,46
	Não	7821			
NO14	Quase todos os dias	322	2,24	0,19	130,95
	Mais da metade dos dias	236	2,05	0,22	80,59
	Menos da metade dos dias	620	1,56	0,16	88,42
	Nenhum dia	7020			
NO20	Intenso	48	0,32	0,61	0,27
	Médio	116	-0,22	0,37	0,34
	Leve	351	0,98	0,20	24,69
	Nenhum	7674			

3.5 Qualidade do modelo ajustado

Para verificar a qualidade do ajuste do modelo, será usado o Teste de Hosmer & Lemeshow. Esse teste, com 8 graus de liberdade apresenta o valor de $\chi^2 = 5{,}3262$ e p-valor = 0,72. Assim, não é possível rejeitar H_0 com nível de significância de 5% e conclui-se que o modelo está bem ajustado.

3.6 Diagnóstico do modelo

As análises de diagnóstico para o modelo de regressão logística multivariado serão realizadas através de análise gráfica.

3.6.1 *Leverage*

A figura 11 mostra o gráfico do número da observação versus os valores h_j. Pode-se observar que a maior partes dos pontos se concentram abaixo de 0,1

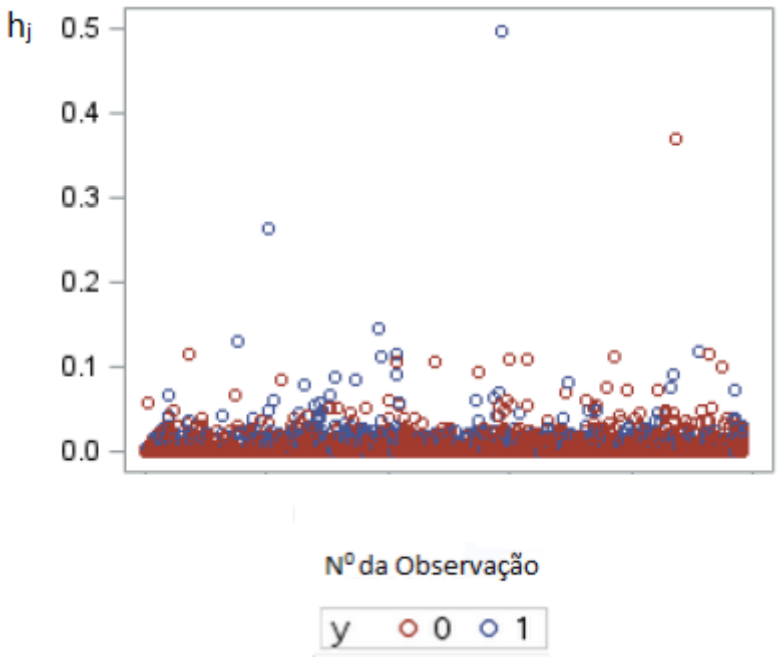

Figura 11: Valores de h_j versus o número da observação

As observações que possuem $h_j \geq 0{,}1$ são aquelas relativas aos indivíduos que classificam-se com outlier, o que concorda com o conceito de pontos de alavanca, que são

aqueles mais afastados do restante dos pontos de uma variável independente. Se usarmos como base para obter ponto de alavanca os pontos $h_j \geq 2p/n$=0,001, como foi sugerido por Belsey et al (1980), aumentaria o número de possíveis pontos de alavanca.

3.6.2 Resíduos

Nas figuras 12 e 13, são apresentados gráficos para os resíduos de Pearson e Deviance versus o número da observação através do modelo de regressão logística. Observa-se um comportamento similar nos pontos de ambos os gráficos. Os casos nos quais as pessoas auto avaliaram seu estado de saúde como muito ruim estão acima do eixo zero e os casos nos quais as pessoas auto avaliaram sua saúde como muito boa estão abaixo de eixo zero. Observa-se, também, que os casos de sucesso estão mais dispersos, representando maior variabilidade.

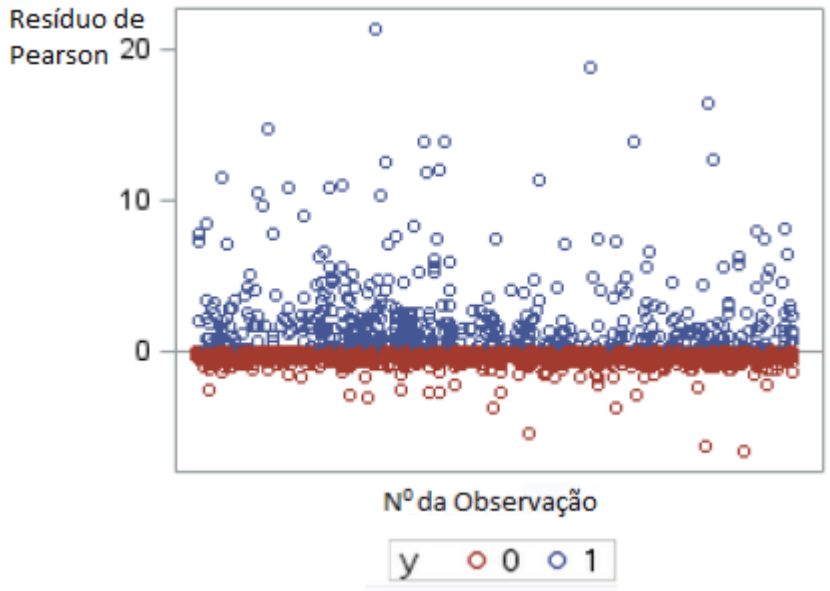

Figura 12: Resíduo de Pearson versus o número da observação

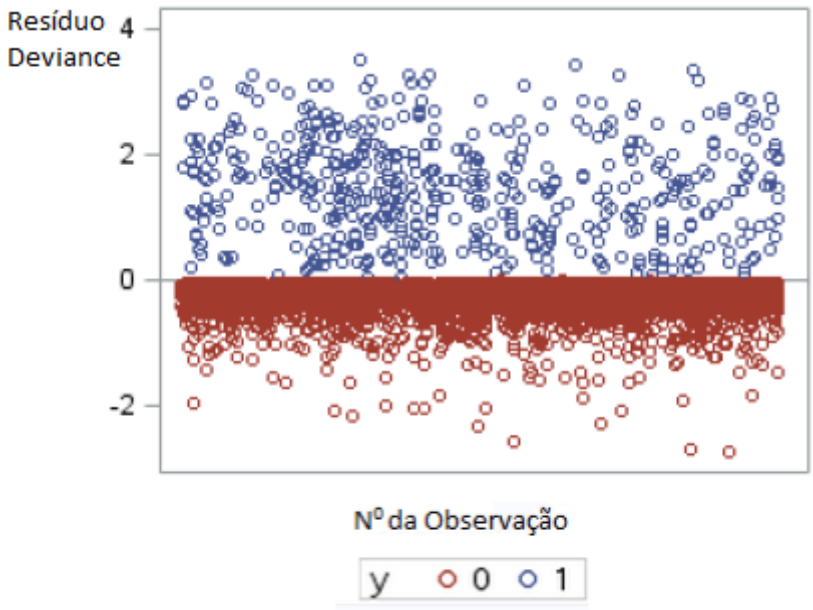

Figura 13: Resíduo Deviance versus o número da observação

Como o modelo possui muitos parâmetros e diversas categorias, representar o gráfico para os DFBETAS de cada categoria por parâmetros do modelo tornaria o relatório longo e cansativo. Dessa forma, selecionou-se dois gráficos para os DFBETAS com as características mais divergentes.

O gráfico abaixo mostra o comportamento da classe amarela da variável COO9 (Raça/cor). Infere-se que apenas 12 pontos estão fora da do eixo zero. Levando em conta que a amostra possui 8.198 observações, esse número torna-se desprezível.

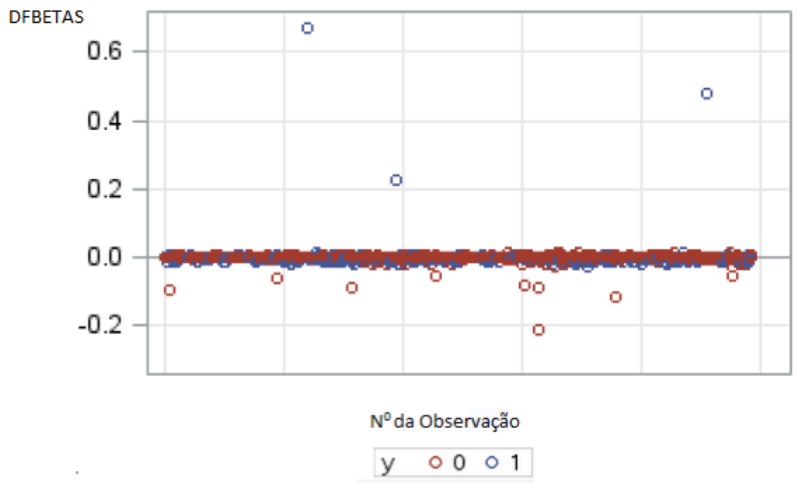

Figura 14: DFBETAS versus o número da observação

Já o gráfico a seguir representa o comportamento da variável EOO1 (ocupação). Infere-se que a variabilidade dos dados é intensa, mostrando que cada observação i possui relevante influência sobre o coeficiente de X_j.

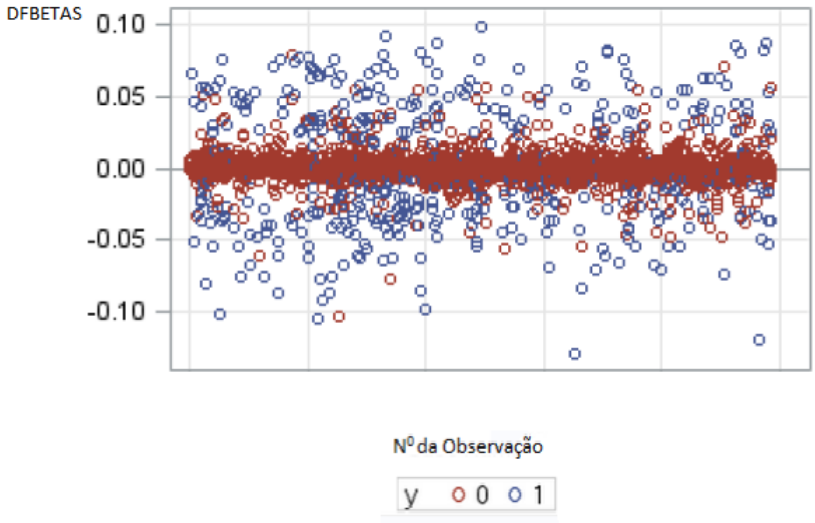

Figura 15: DFBETAS versus o número da observação

Pode-se perceber que tanto na figura 13 quanto na 14 existem algumas observações que se distanciam das demais. Como sugerido por Neter et al (1996), usando como base para detectar valores absolutos de DFBETAS superiores a $1/\sqrt{n}$=$1/\sqrt{8198}$=0,011, aumentaria o número de observações discrepantes.

3.7 Interpretação do modelo

A interpretação de cada variável explicativa em relação ao logito será feira através da medida razão de chances. A razão de chances e o intervalo de confiança de cada variável explicativa selecionada para o modelo são apresentados na tabela 7.

Tabela 9: Razão de chances (RC) e intervalo de confiança (IC) referentes à regressão logística multivariada das variáveis explicativas em relação à variável resposta.

Variável explicativa		RC	IC 95% para RC
COO6	Feminino	1,76	(1,37-2,28)
	Masculino	1	
COO8	Idade	1,047	(1,03-1,05)
COO9	Indígena	1,81	(0,51-6,48)
	Preta	1,41	(0,96-2,13)
	Amarela	1,09	(0,283-4,2)
	Parda	1,75	(1,36-2,25)
	Branca	1	
CO11	Solteiro(a)	1,16	(0,88-1,53)
	Separado(a)	1,50	(0,84-2,65)
	Divorciado(a)	1,42	(0,92-2,18)
	Viúvo(a)	0,33	(0,21-0,53)
	Casado(a)	1	
DOO3	Não é analfabeto	15,76	(6,32-39,30)
	Ens. Fund. Inc.	11,16	(7,51-16,59)
	Ens. Fun. Comp.	2,33	(1,52-3,57)
	Ens. Méd. Comp.	1	
	Ens. Sup. Comp.	0,29	(0,03-2,46)
EOO1	Ocupado	0,51	(0,40-0,65)
	Não ocupado	1	
NOO2	Sim	17,08	(7,97-36,61)
	Não	1	
NOO5	Sim	12,91	(8,42-19,80)
	Não	1	
NO14	Quase todos os dias	9,41	(6,41-13,82)
	Mais da metade dos dias	7,80	(4,50-12,22)
	Menos da metade dos dias	4,77	(3,44-6,61)
	Nenhum dia	1	
NO20	Intenso	3,70	(1,15-11,81)
	Médio	2,15	(1,11-4,15)
	Leve	2,68	(1,82-3,95)
	Nenhum	1	

Considerando a variável COO6 (sexo), vemos que, a chance de uma pessoa do sexo feminino auto avaliar sua saúde como muito ruim é 76% maior se comparada a uma pessoa do sexo masculino, evidenciando que as mulheres estão menos saudáveis física e psicologicamente.

A chance da pessoa classificar sua saúde como muito ruim aumenta 0,047% para cada aumento de 1 ano na idade.

Em relação à variável COO9 (cor ou raça) percebe-se que, tendo uma pessoa de cor branca como referência, todas as demais cor/raça envolvidos no estudos possuem maiores chances de auto avaliar seu estado de saúde muito ruim , sendo máximo de 81% para a raça indígena e o mínimo de 9% para a cor amarela.

As chances das pessoas solteiras, separadas ou divorciadas auto avaliarem sua saúde como muito ruim é superior aos casados. Porém, em relação aos viúvos, essa interpretação muda, mostrando que a chance de um(a) viúvo(a) auto avaliar seu estado de saúde como muito ruim é 67% inferior se comparada a uma pessoa casada. Esse fato é bastante interessante tendo em vista que na amostra selecionada a maioria das pessoas (54,01%) ainda vivem com cônjuge. E os viúvos correspondem a 5,40% da amostra.

O estudo mostrou que para a variável DOO3 (escolaridade), pessoas que possuem alfabetização mínima e ensino fundamental incompleto possuem razão de chances alta em relação aos que possuem ensino médio completo. Já em relação aos que possuem ensino superior completo, as interpretações mudam, pois a chance de uma pessoa com ensino superior completo auto avaliar sua saúde como muito ruim é 71% menor se comparada as pessoas que possuem ensino médio completo. Isso evidencia que a importância do nível superior vai muito além do profissionalismo, interferindo também na saúde física e psicológica.

Já em relação a variável EOO1 (ocupação), nota-se que a razão de chances é menor que 1, mostrando que a chace de uma pessoa ocupada classificar sua saúde como muito ruim é 49% menor se comparada à pessoa que não possui ocupação. Esse fato mostra como a ocupação, ou seja, a pessoa possuir emprego, interfere no psicológico das pessoas.

Para as variáveis NOO2 (utilização de recurso para locomoção) e NOO5 (problemas cardíacos) as razões de chances são extremamente elevadas para pessoas que utilizam recursos para locomoção e problemas cardíacos, respectivamente.

As pessoas que possuem problemas relacionados à alimentação, evidenciados na variável NO14, possuem chances de auto avaliar sua saúde como muito ruim proporcional à frequência na qual possuem problemas alimentares. Pois a chance de uma pessoa que quase todos os dias do mês possui problemas alimentares se auto avaliar como muito ruim a saúde é 9,41 vezes maior em relação a quem não possui problemas com alimentação. Para os que possuem mais da metade dos dias problemas alimentares, esse número cai para 7,8 e menos da metade dos dias, 4,77.

A última variável explicativa do modelo foi a audição, mostrando que a razão de chances é proporcional ao grau de dificuldade para ouvir, ou seja, quanto maior a deficiência auditiva, maior a chance da pessoa auto avaliar como muito ruim seu estado de saúde.

É interessante observar que das variáveis explicativas relacionadas a aspectos físicos e psicológicos (mostradas na tabela 2), apenas variáveis relacionadas à visão (NO21, NO22, NO23) não foram incluídas no modelo. Isso sugere que a deficiência visual é tão comum, que deixa de ser um aspecto relevante na auto avaliação da saúde. De acordo com os dados, aproximadamente 62% não utiliza recursos para auxiliar a enxergar e a maioria (em torno de 75%) não possui nenhuma dificuldade para enxergar de longe e de perto.

Capítulo 4

4 Conclusão

A análise dos resultados obtidos através da modelagem indicaram que as variáveis demográficas sexo, idade, raça/cor, estado civil , escolaridade e ocupação são significativas na predição do modelo. Em relação às variáveis físicas e psicológicas, problemas de locomoção, cardíacos, alimentares e auditivos foram selecionadas na predição do modelo de regressão logística. Desse modo a autopercepção da saúde mostra-se influenciada pela cultura, paradigmas, estado psicológico ou conceitos individuais do indivíduo. As maiores diferenças foram observadas nas variáveis escolaridade, problemas cardíacos e de locomoção, uma vez que as categorias dessas variáveis foram as que apresentaram maiores medidas de razão de chances. Em relação às variáveis explicativas DOO3 (escolaridade), NO14 (problemas alimentares) e NO20 (problemas auditivos), infere-se que a auto avaliação da saúde como muito ruim é proporcional à classificação das classes da variável.

É importante fazer o diagnóstico do modelo e verificar quais observações estão interferindo nas estimativas do modelo ajustado através do modelo de regressão logística. Porém, deve-se ter bastante cuidado quanto à decisão sobre o que fazer com essas observações, que podem ser tanto pontos influentes quanto valores extremos. É interessante observar que diferentes autores apontam para diferentes pontos de corte que definem quais observações estão interferindo no ajustamento do modelo. Diante desse tipo de situação, torna-se essencial consultar algum profissional especializado da área ou com o pesquisador que sugeriu a pesquisa a fim de verificar se o ponto de corte condiz com o estudo em questão. Deve-se estar "familiarizado" com a natureza dos dados de maneira a conhecer qual o comportamento dos diversos elementos na população e, quando isso não ocorrer, estar preparado para fazer uma análise mais detalhada da situação para decidir o quão importante é manter ou eliminar uma observação ou modificar o modelo de regressão utilizado, de maneira que as inferências para esse novo banco de dados sejam mais "eficientes".

Essa monografia apresentou modelos de regressão logística para avaliar as variáveis demográficas e os fatores físicos e psicológicos que interferem na auto avaliação

da percepção de saúde do indivíduo. O estudo desenvolvido poderá fomentar aplicações mais sofisticadas para o tema abordado e podem servir de base na elaboração de políticas públicas.

5 Referências

AGRESTI, D. **Categorical Data Analisys**. second edition, New York: John Wiley & sons. 1990.

COLLET, D. **Modelling Binary Data**. second edition, London: Chapman & Hall/CRC. 2002.

CORDEIRO, G. M. , LIMA NETO, E. A. **Modelos Paramétricos. Livros Texto de Minicurso, XVI Simpósio Nacional de Probabilidade e Estatística**. 2004.

HOSMER, D., LEMESHOW, S. **Applied Logistic Regression** . second edition, New York: Jonh Wiley Sons. 1989.

NETER, J.,KUTNER.M., NACHTSHEIM, C. J. e WASSERMAN, W. **Applied Linear Statistical Models**. fifth edition, Illinois: Irwin, 2005.

LAND, J.M.,PREGIBON, D. & SHOEMAKER, A.C. **Graphical methods for assessing logistic regression models (with discussion)**. Journal of the American Statistical Association, 61-71.

WEISBERG, S. **Applied Linear Regression** . third edition, New York:John Wiley Sons, Inc, 2005.