Diagnóstico no modelo de regressão logística ordinal

Marina Calais de Freitas Moura

DISSERTAÇÃO DE MESTRADO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO

Programa: Estatística
Orientadora: Profa. Dra. Mônica Carneiro Sandoval

São Paulo, junho de 2019

Diagnóstico no modelo de regressão logística ordinal

Esta versão da dissertação contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 11/06/2019. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Profª. Drª. Mônica Carneiro Sandoval - IME-USP
- Prof. Dr. Gustavo Henrique de Araújo Pereira - UFSCar
- Prof. Dr. João Ricardo Saito - UFABC

Agradecimentos

A Deus por permitir que continuasse os estudos.

Aos meus pais Eivanyr de Moura e Maria José de Freitas Moura e minhas irmãs Mônica Moura da Silveira Lima, Bárbara de Freitas Moura e Ana Maria Moura por todo amor e apoio.

Às professoras Mônica Carneiro Sandoval e Denise Aparecida Botter por todos ensinamentos.

A todos professores e amigos que me incentivaram.

Resumo

MOURA, M. C. F. **Diagnóstico no modelo de regressão logística ordinal**. 2019. 66 f. Dissetação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2019.

Os modelos de regressão logística ordinais são usados para descrever a relação entre uma variável resposta categórica ordinal e uma ou mais variáveis explanatórias. Uma vez ajustado o modelo de regressão, se faz necessário verificar a qualidade do ajuste do modelo. As estatísticas qui-quadrado de Pearson e da razão de verossimilhanças não são adequadas para acessar a qualidade do ajuste do modelo de regressão logística ordinal quando variáveis contínuas estão presentes no modelo. Para este caso, foram propostos os testes de Lipsitz, a versão ordinal do teste de Hosmer-Lemeshow e os testes qui-quadrado e razão de verossimilhanças de Pulkistenis-Robinson. Nesta dissertação é feita uma revisão das técnicas de diagnóstico disponíveis para os Modelos logito cumulativo, Modelos logito categorias adjacentes e Modelos logito razão contínua, bem como uma aplicação a fim de investigar a relação entre a perda auditiva, o equilíbrio e aspectos emocionais nos idosos.

Palavras-chave: teste de Lipsitz, testes qui-quadradado e da razão de verossimilhanças de Pulkstenis-Robinson, versão ordinal do teste Hosmer-Lemeshow.

Abstract

MOURA, M. C. F. **Diagnostic of ordinal logistic regression model**. 2019. 66 f. Dissertation (Master´s degree) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2019.

Ordinal regression models are used to describe the relationship between an ordered categorical response variable and one or more explanatory variables which could be discrete or continuous. Once the regression model has been fitted, it is necessary to check the goodness-of-fit of the model. The Pearson and likelihood-ratio statistics are not adequate for assessing goodness-of-fit in ordinal logistic regression model with continuous explanatory variables. For this case, the Lipsitz test, the ordinal version of the Hosmer-Lemeshow test and Pulkstenis-Robinson chi-square and likelihood ratio tests were proposed. This dissertation aims to review the diagnostic techniques available for the cumulative logit models, categories adjacent logit models and continuous ratio logistic models. In addition, an application was developed in order to investigate the relationship between hearing loss, balance and emotional aspects in the elderly.
Keywords: Lipsitz test, Pulkstenis-Robinson chi-squared and likelihood-ratio test, ordinal version of the Hosmer-Lemeshow test.

Sumário

Lista de Abreviaturas

AIC	Critério de Informação Akaike (*Akaike Information Criterion)*
IC	Intervalo de Confiança (*Confidence Interval*)
MLG	Modelo Linear Generalizado (*Generalized Linear Model*)

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Ao longo dos anos, métodos para análise de dados categorizados têm recebido uma considerável atenção, devido ao crescente uso desse tipo de dados em diversas aplicações (Paulino e Singer, 2006).

No âmbito da Psicologia ou da Ciências Sociais, o uso de dados categorizados é muito comum para medir atitudes, opiniões ou preferências. Dados categorizados também podem ser encontrados nas áreas ligadas à Saúde, por exemplo, quando o interesse é avaliar a melhora ou não de um paciente.

Dados categóricos ou qualitativos discriminam-se em três tipos: nominal, ordinal e intervalar. O primeiro tipo ocorre quando a escala categórica não é ordenada, ou seja, a permutação das categorias não afeta a análise estatística, por exemplo: estado conjugal (solteira, casada, divorciada, viúva). Já, o segundo tipo ocorre quando suas categorias seguem uma ordenação clara e a permutação delas influencia na análise estatística. Como ilustração tem-se a variável nível de escolaridade (ensino fundamental, ensino médio, graduação, pós-graduação). O tipo intervalar acontece quando uma variável contínua é resumida, agrupando-se os valores em categorias, por exemplo, idade (0-20, 20-40, 40-60, 60-80, acima de 80).

Métodos desenvolvidos para análise de variáveis nominais podem ser utilizados para variáveis nominais e ordinais, uma vez que estes requerem apenas escala categórica, não importando a ordenação das categorias. Contudo, resultados diferentes podem ser obtidos quando são utilizados métodos desenvolvidos para análise de variáveis ordinais. Assim, estes métodos, só podem ser utilizados para variáveis ordinais, pois levam em conta a ordenação das categorias.

Quando a classificação dos dados ordinais não é totalmente explorada, ou seja, quando se trata variáveis ordinais ou intervalares como nominais, as permutações das categorias são irrelevantes e, consequentemente, muita informação é perdida. O que distingue o uso dos modelos ordinais dos outros é que estes produzem inferências mais poderosas sobre as características da população.

O Modelo logito cumulativo para análise de variáveis resposta ordinais foi originalmente proposto

por Walker e Duncan (1967) e posteriormente chamado de Modelo de chances proporcionais por Cullagh (1980). Este modelo é uma extensão do Modelo de regressão logística que por sua vez é um caso especial do Modelo Linear Generalizado (MLG), o qual foi proposto por Nelder e Wedderburn (1972), que permite o uso de distribuições pertencentes à família exponencial para a variável resposta, além da distribuição normal.

Além do Modelo logito cumulativo para análise de dados com variáveis respostas ordinais, Agresti (1984) sugeriu o Modelo logito categorias adjacentes e Feinberg (1980) propôs o Modelo logito razão contínua. Segundo Agresti (2010), a escolha do modelo está associada ao tipo de comparação que faz mais sentido para o estudo.

Os modelos supracitados são úteis para descrever a relação entre a variável resposta ordinal e uma ou mais variáveis explanatórias. Estes modelos diferem em como as categorias da variável resposta são comparadas.

Depois de ajustado o modelo, se faz necessário verificar a qualidade desse ajuste, ou seja, verificar o quão próximo os valores preditos por este modelo se encontram de seus correspondentes valores observados. Para os modelos citados anteriormente, esta qualidade pode ser averiguada pela estatística de Pearson, X^2; e pela estatística da razão de verossimilhanças, G^2, desde que as variáveis explanatórias sejam categóricas.

Para o caso em que o modelo contém pelo menos uma variável explanatória contínua, Lipsitz *et al.* (1996) propôs um teste de qualidade de ajuste, baseado parcialmente no teste de Hosmer-Lemeshow para modelos de regressão logística.

Pulkstenis e Robinson (2004) modificaram a estatística de Pearson e da razão de verossimilhanças para quando o modelo contém variáveis explanatórias contínuas e categóricas. Fagerland e Hosmer (2013) desenvolveram a versão ordinal do teste de Hosmer-Lemeshow.

Os testes mencionados foram desenvolvidos com base no Modelo logito cumulativo com chances proporcionais, que é o modelo mais utilizado na análise de dados categóricos ordinais. Contudo, Fagerland e Hosmer (2016) estenderam os testes para os Modelos logito categorias adjacentes com chances proporcionais e Modelo logito razão contínua com chances proporcionais.

1.1 Objetivos

Um dos objetivos deste trabalho é revisar os métodos disponíveis para averiguar a qualidade do ajuste dos Modelos de regressão logística ordinais mais utilizados, dando ênfase aos modelos em que variáveis explanatórias contínuas estão presentes ou quando há dados esparsos (quando poucas observações forem registradas em algumas das categorias da variável resposta). Outro objetivo é

ilustrar os métodos estudados a conjuntos de dados reais.

1.2 Organização do texto

No Capítulo 2, são abordados os conceitos básicos relacionados aos principais modelos utilizados na análise de dados com variável resposta ordinal: Modelos logito cumulativo, Modelo logito categorias adjacentes e Modelo logito razão contínua.

Os métodos utilizados para verificar a qualidade do ajuste dos modelos estudados no Capítulo 2 são discutidos no Capítulo 3.

O Capítulo 4 apresenta uma aplicação referente ao estudo da perda auditiva em idosos.

Por fim, no Capítulo 5 são estabelecidas conclusões sobre o estudo realizado nessa dissertação.

Capítulo 2

Modelos de Regressão Logística Ordinal

Neste capítulo será apresentada uma revisão dos Modelos lineares generalizados e dos Modelos de regressão logística. Serão abordados, ainda, os modelos mais utilizados na análise de variáveis respostas ordinais, que são os Modelos logito cumulativo, Modelos logito categorias adjacentes e Modelos logito razão contínua.

2.1 Modelos lineares generalizados

Ao longo de muitos anos, utilizou-se a estrutura dos modelos de regressão normais com o intuito de descrever a maioria dos fenômenos aleatórios, ainda que, estes fenômenos não apresentassem os critérios que justificassem tal suposição. Nestes casos era sugerida uma transformação a fim de atender a normalidade, porém, além de dificultar a interpretação das variáveis transformadas, o ajuste do modelo não era muitas vezes satisfatório.

Nelder e Wedderburn (1972) propuseram os Modelos lineares generalizados (MLG) que possibilitam a utilização de outras distribuições, desde que pertençam à família exponencial de distribuições. Com a teoria e metodologia dos MLG, a transformação dos dados não é necessária e a escolha da distribuição não se restringe à normalidade. O MLG foi introduzido com a finalidade de unificar uma ampla variedade de modelos estatísticos.

Os Modelos lineares generalizados têm três componentes:

- Componente Aleatório - seleciona a distribuição de probabilidade da variável resposta Y;

  Este componente é composto por variáveis aleatórias $Y_1, \ldots, Y_n$ independentes, pertencentes à família exponencial, cada uma com função densidade ou função de probabilidade na forma:

$$f(y_i; \theta_i, \phi) = exp\{\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i; \phi)\},$$

  em que b(.) e c(.) são funções reais conhecidas, θ_i é o parâmetro canônico, ϕ é o parâmetro de dispersão e o domínio de $y_i \subset \mathbb{R}$ e não depende de θ_i.

Muitas distribuições são casos especiais da família exponencial. Para os casos em que cada observação Y_i é binária, como em ensaios do tipo sucesso ou fracasso, assume-se a distribuição de Bernoulli para cada elemento do componente aleatório. Em outras aplicações, em que cada resposta é uma contagem não negativa, assume-se a distribuição de Poisson para o componente aleatório. Para os casos em que as observações são contínuas, assume-se, por exemplo, que o componente aleatório tem distribuição Normal ou distribuição Gama.

- Componente sistemático – constituído pelas variáveis explanatórias usadas como preditores no modelo.

  Este componente, denominado preditor linear, é a combinação linear das p variáveis explanatórias do modelo, o qual tem a seguinte forma:

$$\eta_i = \sum_j \beta_j x_{ij},\ i = 1, ..., n\ e\ j = 1, ..., p,$$

  em que β_j é um parâmetro desconhecido que descreve o efeito da j-ésima variável explanatória e x_{ij} é o valor da j-ésima variável explanatória, associado à i-esima unidade experimental.

- Função de ligação – descreve a relação entre o componente sistemático e o valor esperado do componente aleatório.

  A função de ligação é uma função monótona e diferenciável, representada por $g(\mu_i)$, a qual relaciona o componente sistemático (η_i) e a média do componente aleatório (μ_i) da seguinte forma:

$$\eta_i = g(\mu_i).$$

  Quando o parâmetro canônico (θ) coincide com o preditor linear $(\theta = \eta)$, as funções de ligação são chamadas de *funções de ligação canônicas*. Desta forma, cada distribuição possui a sua *função de ligação canônica* correspondente.

2.2 Modelo de regressão logística

O Modelo de regressão logística é um caso especial do MLG, em que a variável resposta é binária, ou seja, admite apenas dois resultados, como por exemplo, o resultado do diagnóstico de um exame

de laboratório, positivo ou negativo (sucesso ou fracasso). Frequentemente chamamos de sucesso o resultado de interesse para o estudo.

Denote Y_i como a variável resposta binária. Desta forma, Y_i segue uma distribuição de *Bernoulli* em que $P(Y_i = 1) = \pi_i$ é a probabilidade de sucesso e $P(Y_i = 0) = 1 - \pi_i$ é a probabilidade de fracasso. Esta distribuição tem média $E(Y_i) = \pi_i$ e variância $Var(Y_i) = \pi_i(1 - \pi_i)$.

Considere $P(Y_i = 1|x_i) = \pi_i = \pi(x_i)$ a probabilidade de sucesso dado o valor x_i de uma variável explanatória (x). O modelo de regressão logística, também chamado de *modelo logito*, tem a seguinte forma:

$$logito[\pi(x_i)] = log\frac{\pi(x_i)}{1 - \pi(x_i)} = \alpha + \beta x_i, \tag{2.1}$$

em que α e β são os parâmetros desconhecidos e $\eta_i = \alpha + \beta x$ é o componente sistemático do MLG.

Exponencializando (2.1) dos dois lados, é possível obter as interpretações dos parâmetros do modelo. Suponha que haja o interesse em analisar a associação entre a artrite reumatóide (presença ou ausência) e o sexo (masculino, feminino). Desta maneira, seria coletada uma amostra de n_1 indivíduos do sexo masculino $(x_i = 1)$ e n_2 indivíduos do sexo feminino $(x_i = 0)$. Assim, a chance de desenvolvimento da doença para um indivíduo do sexo masculino é:

$$\frac{\pi(1)}{1 - \pi(1)} = e^{\alpha+\beta},$$

enquanto que a chance de desenvolvimento da doença para um indivíduo do sexo feminino é:

$$\frac{\pi(0)}{1 - \pi(0)} = e^{\alpha}.$$

.

Logo a razão de chances fica dada por:

$$\frac{\pi(1)(1 - \pi(0))}{\pi(0)(1 - \pi(1))} = e^{\beta}.$$

Este modelo provê a seguinte interpretação: a chance de um indivíduo do sexo masculino ter artrite reumatóide é e^{β} vezes a chance de um indivíduo do sexo feminino ter artrite reumatóide.

Suponha agora que haja interesse em analisar a relação entre infarto do miocárdio (sim ou não)

e o número de cigarros fumados por dia. Quanto à interpretação, tem-se que para cada unidade de aumento em x, ou seja, para cada unidade de aumento no número de cigarros, a chance de ter infarto do miocárdio é multiplicada por e^{β}.

Além do *logito*, existem outras funções de ligação que podem ser utilizadas para a variável resposta binária, tais como, *probito* e *complemento log-log*. Os modelos que utilizam essas ligações podem ser verificados em Agresti (2003).

O modelo descrito neste seção é indicado quando a variável resposta é binária ou dicotômica. Contudo, quando a variável resposta tem mais de duas categorias, outros modelos são apropriados. Para análise de dados que apresentam variável resposta nominal com mais de duas categorias, um modelo proposto é o modelo logito categoria de referência (Agresti, 1996). Este modelo não leva em conta a ordenação das categorias.

2.3 Principais modelos ordinais

O Modelo de regressão logística ordinal é aplicado quando o número de categorias da variável resposta excede dois e quando estas são ordenadas. Existe uma variedade de modelos para variáveis ordinais que respeitam a natureza ordinal dos dados, os quais são aplicados quando há interesse em verificar a relação entre a variável resposta ordinal e as variáveis explanatórias de relevância para o estudo, sendo que as variáveis explanatórias podem ser contínuas ou categóricas.

Quando se utiliza modelos nominais para variáveis ordinais, as permutações das categorias são irrelevantes e, consequentemente, muita informação é perdida. Métodos ordinais possibilitam descrição simples dos dados e permitem inferências mais poderosas sobre as características da população do que os modelos para variáveis nominais que ignoram a informação ordinal.

A seguir serão apresentados os modelos utilizados quando a variável resposta é ordinal, que utilizam a ligação *logito*. Assim como no Modelo de regressão logística binária, outros tipos de ligações podem ser usados, como *probito* e *complemento log-log*. Os modelos que utilizam essas ligações podem ser verificados em Agresti (2010).

2.3.1 Modelo logito cumulativo

Considerando uma variável resposta Y_i com c categorias ordinais, o *logito cumulativo* é definido por:

$$logito[P(Y_i \leq j|\mathbf{x_i})] = log\frac{P(Y_i \leq j|\mathbf{x_i})}{1 - P(Y_i \leq j|\mathbf{x_i})} = log\frac{\pi_1(\mathbf{x_i}) + ... + \pi_j(\mathbf{x_i})}{\pi_{j+1}(\mathbf{x_i}) + ... + \pi_c(\mathbf{x_i})},\ j = 1, ..., c-1,$$

em que j representa cada categoria ordenada da variável resposta e $P(Y_i = j|\mathbf{x_i}) = \pi_j(\mathbf{x_i})$ representa a probabilidade de ocorrência da j-ésima categoria de resposta para um dado vetor $\mathbf{x_i}$ de p covariáveis, com $\sum_{j=1}^{c} \pi_j(\mathbf{x_i}) = 1$.

2.3.1.1 Modelo logito cumulativo com chances proporcionais

O Modelo logito cumulativo com chances proporcionais foi popularizado por Cullagh (1980) com o nome de Modelo de chances proporcionais (*proportional odds model*). Entretanto, Agresti (2010) referiu-se a esse modelo por Modelo logito cumulativo com chances proporcionais (*proportional odds version of the cumulative logit model*) devido à estrutura de chances proporcionais também poder ser utilizada para outros tipos de modelos, como o Modelo logito categorias adjacentes ou Modelo logito razão contínua.

O Modelo logito cumulativo com chances proporcionais tem a seguinte forma:

$$logito[P(Y_i \leq j|\mathbf{x_i})] = \alpha_j + \boldsymbol{\beta}' \boldsymbol{x_i} = \alpha_j + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip},\ j = 1, ..., c-1, \tag{2.2}$$

em que $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)'$ é um vetor de p parâmetros desconhecidos, $\mathbf{x_i}$ denota o vetor dos valores das variáveis explanatórias para a observação i e j representa cada categoria ordenada da variável resposta. Os interceptos α_j são tais que $\alpha_1 < \alpha_2 < ... < \alpha_j$, pois as probabilidades cumulativas $P(Y_i \leq j)$ aumentam em j para cada valor fixo de $\mathbf{x_i}$.

O Modelo logito cumulativo com chances proporcionais tem para cada logito cumulativo um intercepto α_j e o mesmo coeficiente angular $\boldsymbol{\beta}$. Este efeito comum implica que as probabilidades cumulativas têm a mesma curvatura (mesma forma). Esse modelo liga os $c-1$ logitos a um único modelo, ou seja, o modelo é mais simples de se interpretar do que se fossem ajustados $c-1$ modelos separados.

Para exemplificar, considere esse modelo aplicado a uma variável resposta ordinal com $c = 4$ categorias ordinais e uma variável explanatória contínua. O parâmetro $\boldsymbol{\beta}$ comum aos três logitos implica que a curva para as três probabilidades cumulativas tem o mesmo formato, conforme mostra a Figura 2.1.

Por vezes, o Modelo logito cumulativo com chances proporcionais pode ser expresso por uma parametrização alternativa que utiliza um sinal negativo precedendo $\boldsymbol{\beta}' \boldsymbol{x_i}$, ou seja,

$$logito[P(Y_i \leq j|\mathbf{x_i})] = \alpha_j - \boldsymbol{\beta}' \boldsymbol{x_i},\ j = 1, ..., c-1.$$

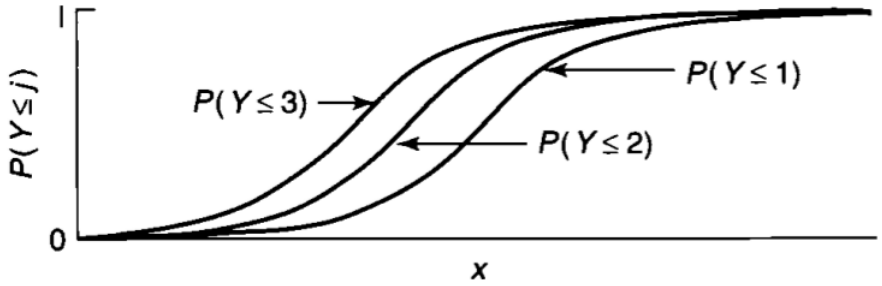

Figura 2.1: *Probabilidades cumulativas no Modelo logito cumulativo com chances proporcionais*
Fonte: Agresti (2010, p.47)

Existem alguns softwares que utilizam esta parametrização, como o SPSS (Corporation, 2017) e outros que utilizam a parametrização positiva, como o SAS (SAS, 2013) e o R (R Core Team, 2017). Para o caso de parametrização negativa, outra forma de obter o sinal usual para a interpretação do modelo é expressar o modelo como:

$$logito[P(Y_i > j|\mathbf{x_i})] = log\frac{P(Y_i > j|\mathbf{x_i})}{1 - P(Y_i > j|\mathbf{x_i})} = \alpha_j + \boldsymbol{\beta}'\boldsymbol{x_i},\ j = 1, ..., c-1.$$

A estimação dos parâmetros do modelo pode ser realizada pelo método de máxima verossimilhança. A função de verossimilhança L é dada por:

$$L = \prod_{i=1}^{n}\left[\prod_{j=1}^{c}\pi_j(\mathbf{x_i})^{y_{ij}}\right] = \prod_{i=1}^{n}\left\{\prod_{j=1}^{c}[P(Y_i \leq j|\mathbf{x_i}) - P(Y \leq j-1|\mathbf{x_i})]^{y_{ij}}\right\} =$$
$$\prod_{i=1}^{n}\left\{\prod_{j=1}^{c}\left[\frac{exp(\alpha_j + \boldsymbol{\beta}'\boldsymbol{x_i})}{1 + exp(\alpha_j + \boldsymbol{\beta}'\boldsymbol{x_i})} - \frac{exp(\alpha_{j-1} + \boldsymbol{\beta}'\boldsymbol{x_i})}{1 + exp(\alpha_{j-1} + \boldsymbol{\beta}'\boldsymbol{x_i})}\right]^{y_{ij}}\right\},$$

em que $y_{ij} = 1$, se a resposta do indivíduo i, $i = 1, 2, ...n$, está na categoria j, $j = 1, 2, ..., c$ e $y_{ij} = 0$, caso contrário, com $\sum_{j=1}^{p} y_{ij} = 1$.

Métodos iterativos são utilizados para obter os estimadores de máxima verossimilhança. Cullagh (1980) e Walker e Duncan (1967) usaram o método escore de Fisher. A interpretação do parâmetro β é análoga à do modelo de regressão logística.

Nos casos em que a suposição de chances proporcionais não é válida, podem ser utilizados os modelos descritos a seguir.

2.3.1.2 Modelo logito cumulativo com chances proporcionais parciais

Peterson e Harrell Jr (1990) propuseram o Modelo logito cumulativo com chances proporcionais

parcias em que é suposto que parte das variáveis explanatórias apresentam a propriedade de chances proporcionais e parte não. Esse modelo fica expresso por:

$$logito[P(Y_i \leq j|\mathbf{x_i})] = \alpha_j + \boldsymbol{\beta}^{'}\boldsymbol{x_i} + \boldsymbol{\gamma_j}^{'}\boldsymbol{z_i},\ j = 1, ..., c-1,$$

em que $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^{'}$ é um vetor $p \times 1$ de parâmetros desconhecidos associados às variáveis em $\mathbf{x_i}$, $\mathbf{x_i}$ denota o vetor $p \times 1$ de variáveis explanatórias para a i-ésima observação, $\mathbf{z_i}$ é um subgrupo das p variáveis explanatórias (vetor $q \times 1, q < p$) para i-ésima observação para o qual a suposição de chances proporcionais não é válida e $\boldsymbol{\gamma_j}$ é um vetor $q \times 1$ dos parâmetros desconhecidos associados às q variáveis em $\mathbf{z_i}$. Então, $\boldsymbol{\gamma_j}$ é definido como um incremente associado somente com o j-ésimo logito cumulativo.

Caso $\boldsymbol{\gamma_j} = \mathbf{0}$ para todo j, então esse modelo se reduz ao Modelo logito cumulativo com chances proporcionais. Quanto à estimação dos parâmetros desse modelo, utiliza-se o método de máxima verossimilhança.

2.3.1.3 Modelo logito cumulativo sem chances proporcionais

O Modelo logito cumulativo sem chances proporcionais permite diferentes efeitos para as variáveis explanatórias:

$$logito[P(Y_i \leq j|\mathbf{x_i})] = \alpha_j + \boldsymbol{\beta_j}^{'}\boldsymbol{x_i},\ j = 1, ..., c-1, \tag{2.3}$$

em que α_j é um parâmetro desconhecido, $\boldsymbol{\beta_j} = (\beta_{1j}, \beta_{2j}, ..., \beta_{pj})^{'}$ denota o vetor $p \times 1$ de parâmetros desconhecidos e $\mathbf{x_i}$ é o vetor dos valores das variáveis explanatórias para a i-ésima observação.

Segundo Agresti (2010), esse modelo é mais indicado quando não houver variáveis explanatórias contínuas no modelo ou quando poucas observações forem registradas em uma categoria.

2.3.2 Modelo logito categorias adjacentes

O logito categorias adjacentes é definido pelo logito da probabilidade condicional da resposta na categoria j, dada a resposta na categoria j ou $j+1$:

$$logito\left[\frac{P(Y_i = j|\mathbf{x_i})}{P(Y_i = j|\mathbf{x_i}\ ou\ Y_i = j+1|\mathbf{x_i})}\right] = logito\frac{\pi_j(\mathbf{x_i})}{\pi_j(\mathbf{x_i}) + \pi_{j+1}(\mathbf{x_i})} = log\frac{\pi_j(\mathbf{x_i})}{\pi_{j+1}(\mathbf{x_i})},\ j = 1, ..., c-1.$$

Esse logito compara a categoria j com a categoria de referência $j+1$. Entretanto, eventualmente

é possível comparar a categoria j com outra categoria. Normalmente, utiliza-se a última categoria como referência ou a mais comum. Note que, quando se utiliza a última categoria como referência, tem-se a seguinte equação:

$$log\frac{\pi_j(\mathbf{x_i})}{\pi_c(\mathbf{x_i})} = log\frac{\pi_j(\mathbf{x_i})}{\pi_{j+1}(\mathbf{x_i})} + log\frac{\pi_{j+1}(\mathbf{x_i})}{\pi_{j+2}(\mathbf{x_i})} + \cdots + log\frac{\pi_{c-1}(\mathbf{x_i})}{\pi_c(\mathbf{x_i})}. \tag{2.4}$$

2.3.2.1 Modelo logito categorias adjacentes com chances proporcionais

O Modelo logito categorias adjacentes com chances proporcionais é definido por:

$$log\frac{\pi_j(\mathbf{x_i})}{\pi_{j+1}(\mathbf{x_i})} = \alpha_j + \boldsymbol{\beta}'\boldsymbol{x_i},\ j = 1, ..., c-1, \tag{2.5}$$

em que α_j é um parâmetro desconhecido, $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)'$ é um vetor $p \times 1$ de parâmetros desconhecidos, $\mathbf{x_i}$ denota o vetor dos valores das variáveis explanatórias para a i-ésima observação e j representa cada categoria ordenada da variável resposta.

Utilizando a categoria c como referência conforme (2.4), o modelo tem a seguinte forma:

$$log\frac{\pi_j(\mathbf{x_i})}{\pi_c(\mathbf{x_i})} = \sum_{k=j}^{c-1}\alpha_k + (c-j)\boldsymbol{\beta}'\boldsymbol{x_i},\ j = 1, ..., c-1.$$

A versão de chances proporcionais para o modelo logito de categorias adjacentes reconhece a ordem das categorias da variável resposta e assume que as variáveis explicativas têm o mesmo efeito para as c categorias. Similar ao que foi apresentado para o Modelo logito cumulativo com chances proporcionais, a estimação dos parâmetros é feita por máxima verossimilhança.

2.3.2.2 Modelo logito categorias adjacentes com chances proporcionais parciais

O Modelo logito categorias adjacentes com chances proporcionais parciais, da mesma maneira que o Modelo logito cumulativo na Subseção 2.3.1.2, permite uma estrutura simples para algumas variáveis explanatórias, em que é suposto que parte das variáveis explanatórias apresentam a propriedade de chances proporcionais e parte não. Esse modelo é expresso por:

$$log\frac{\pi_j(\mathbf{x_i})}{\pi_{j+1}(\mathbf{x_i})} = \alpha_j + \boldsymbol{\beta}'\boldsymbol{x_i} + \gamma_{\mathbf{j}}'\mathbf{z_i},\ j = 1, ..., c-1,$$

em que $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)'$ é um vetor $p \times 1$ de parâmetros desconhecidos associados às variáveis em $\mathbf{x_i}$, $\mathbf{x_i}$ denota o vetor $p \times 1$ de variáveis explanatórias para a i-ésima observação, $\mathbf{z_i}$ é um subgrupo das p variáveis explanatórias (vetor $q \times 1, q < p$) para i-ésima observação para o qual a suposição de chances proporcionais não é válida e $\boldsymbol{\gamma_j}$ é um vetor $q \times 1$ dos parâmetros desconhecidos associados às q variáveis em $\mathbf{z_i}$.

Com relação à estimação dos parâmetros para este modelo, também é utilizado o método de máxima verossimilhança.

2.3.2.3 Modelo logito categorias adjacentes sem chances proporcionais

O Modelo logito categorias adjacentes sem chances proporcionais permite diferentes efeitos para as variáveis explanatórias, o qual tem a seguinte forma:

$$log\frac{\pi_j(\mathbf{x_i})}{\pi_{j+1}(\mathbf{x_i})} = \alpha_j + \boldsymbol{\beta_j'}\boldsymbol{x_i},\ j = 1, ..., c-1, \tag{2.6}$$

em que α_j é um parâmetro desconhecido, $\boldsymbol{\beta_j} = (\beta_{1j}, \beta_{2j}, ..., \beta_{pj})'$ é o vetor $p \times 1$ de parâmetros desconhecidos e x_i é o vetor dos valores das variáveis explanatórias para a i-ésima observação.

Quanto à estimação dos parâmetros para este modelo, utiliza-se o método de máxima verossimilhança.

Segundo Agresti (2010), a escolha do modelo deve ser baseada em qual comparação entre as categorias da variável resposta é mais informativa para o problema.

Como ilustração, suponha que haja interesse em analisar a relação entre os graus da doença renal crônica (normal, leve, moderada e severa) e ter diabetes (sim e não). O Modelo logito cumulativo deve ser escolhido quando o interesse for avaliar a chance de um indivíduo diabético não ter a doença em relação a apresentar insuficiência renal (leve, moderada ou severa); ou avaliar a chance de um indivíduo diabético não ter a doença ou apresentar insuficiência renal leve em relação à insuficiência renal moderada ou severa; ou ainda, avaliar a chance de um indivíduo diabético não ter a doença ou apresentar insuficiência renal leve ou moderada em relação à insuficiência renal severa. Já, o Modelo logito categorias adjacentes deve ser escolhido quando o interesse for avaliar a chance de um indivíduo diabético não ter a doença em relação à ter insuficiência renal leve; ou a chance de um indivíduo diabético ter insuficiência renal leve em relação à moderada; ou a chance de um indivíduo diabético ter insuficiência renal moderada em relação à severa.

2.3.3 Modelo logito razão contínua

O logito razão contínua é definido por:

$$logito[w_j(\mathbf{x_i})] = log\frac{w_j(\mathbf{x_i})}{(1-w_j(\mathbf{x_i}))} = log\frac{\pi_j(\mathbf{x_i})}{\pi_{j+1}(\mathbf{x_i})+...+\pi_c(\mathbf{x_i})},\ j=1,...,c-1,$$

em que $w_j(\mathbf{x_i}) = \dfrac{P(Y=j|\mathbf{x_i})}{P(Y\geq j|\mathbf{x_i})} = \dfrac{\pi_j(\mathbf{x_i})}{\pi_j(\mathbf{x_i})+...+\pi_c(\mathbf{x_i})}$,
ou

$$logito[w^*_{j+1}(\mathbf{x_i})] = log\frac{w^*_{j+1}(\mathbf{x_i})}{(1-w^*_{j+1}(\mathbf{x_i}))} = log\frac{\pi_{j+1}(\mathbf{x_i})}{\pi_1(\mathbf{x_i})+...+\pi_j(\mathbf{x_i})},\ j=1,...,c-1,$$

em que $w^*_{j+1}(\mathbf{x_i}) = \dfrac{P(Y=j+1|\mathbf{x_i})}{P(Y\leq j+1|\mathbf{x_i})} = \dfrac{\pi_{j+1}(\mathbf{x_i})}{\pi_1(\mathbf{x_i})+...+\pi_{j+1}(\mathbf{x_i})}$.

De acordo com Agresti (2010), os logitos razão contínua são úteis quando a variável Y for caracterizada por um processo sequencial em que cada observação deve passar pela categoria resposta j antes de passar para uma categoria superior. Por exemplo, Tutz (1991) utilizou o Modelo logito razão contínua para analisar a relação entre o aumento de amídala (não ampliada, ampliada e grandemente ampliada) e ser portador ou não de bactérias. Tutz (1991) argumentou que estes dados formam um grupo sequencial pois a amídala começa no estágio de não ampliada e depois torna-se ampliada.

2.3.3.1 Modelo logito razão contínua com chances proporcionais

Sendo a resposta Y caracterizada por um processo sequencial, o Modelo logito razão contínua com chances proporcionais pode ser expresso por:

$$log\frac{\pi_j(\mathbf{x_i})}{\pi_{j+1}(\mathbf{x_i})+...+\pi_c(\mathbf{x_i})} = \alpha_j + \boldsymbol{\beta}'\boldsymbol{x_i},\ j=1,...,c-1, \tag{2.7}$$

em que $\boldsymbol{\beta} = (\beta_1,\beta_2,...,\beta_p)'$ é um vetor $p\times 1$ de parâmetros desconhecidos, $\mathbf{x_i}$ denota o vetor dos valores das variáveis explanatórias para a observação i e j representa cada categoria ordenada da variável resposta.

Procedimentos similares aos apresentados para os Modelos logitos cumulativos são realizados para os Modelos logitos razão contínua, quanto à estimação dos parâmetros.

2.3.3.2 Modelo logito razão contínua com chances proporcionais parciais

Analogamente ao Modelo logito cumulativo da Subseção 2.3.1.2 e ao Modelo logito categorias adjacentes da Subseção 2.3.2.2, o Modelo logito razão contínua com chances proporcionais parciais supõe que parte das variáveis explanatórias apresentam a propriedade de chances proporcionais e parte não. Esse modelo pode ser expresso por:

$$log\frac{P(Y_i = j|\mathbf{x_i})}{P(Y_i > j|\boldsymbol{x_i})} = \alpha_j + \beta^{'}\mathbf{x_i} + \gamma_{\mathbf{j}}^{'}\mathbf{z_i},\ j = 1, ..., c-1,$$

em que $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^{'}$ é um vetor $p \times 1$ de parâmetros desconhecidos associados às variáveis em $\mathbf{x_i}$, $\mathbf{x_i}$ denota o vetor $p \times 1$ das variáveis explanatórias para a i-ésima observação, $\mathbf{z_i}$ é um subgrupo das p variáveis explanatórias (vetor $q \times 1$, $q < p$) para i-ésima observação para o qual a suposição de chances proporcionais não é válida e $\boldsymbol{\gamma_j}$ é um vetor $q \times 1$ dos parâmetros desconhecidos associados às q variáveis em $\mathbf{z_i}$.

Quanto à estimação dos parâmetros para este modelo, utiliza-se o método de máxima verossimilhança.

2.3.3.3 Modelo logito razão contínua sem chances proporcionais

O Modelo logito razão contínua sem chances proporcionais é expresso por:

$$log\frac{P(Y_i = j|\mathbf{x_i})}{P(Y_i > j|\mathbf{x_i})} = \alpha_j + \boldsymbol{\beta_j^{'}}\boldsymbol{x_i},\ j = 1, ..., c-1, \tag{2.8}$$

em que α_j é um parâmetro desconhecido, $\boldsymbol{\beta_j} = (\beta_{1j}, \beta_{2j}, ..., \beta_{pj})^{'}$ é o vetor de $p \times 1$ parâmetros desconhecidos e $\mathbf{x_i}$ é o vetor $p \times 1$ dos valores das variáveis explanatórias para a i-ésima observação.

Este modelo possui diferentes efeitos para as diferentes categorias. Quanto à estimação dos parâmetros para este modelo, utiliza-se o método de máxima verossimilhança.

Neste capítulo foram apresentados os modelos de regressão logística ordinal que fazem uso do logito cumulativo, logito categorias adjacentes e logito razão contínua. Dentre os modelos expostos, o Modelo de regressão logística ordinal mais utilizado na literatura é o Modelo logito cumulativo.

Capítulo 3

Técnicas de Diagnóstico

Neste capítulo serão apresentadas técnicas de diagnóstico para verificar a adequabilidade dos modelos de regressão logística ordinal citados no Capítulo 2, bem como, métodos para comparação de modelos e análise de resíduos.

3.1 Conceitos básicos

Os modelos descritos no Capítulo 2 permitem o uso de variáveis explanatórias categóricas e contínuas. Quando se tem apenas variáveis explanatórias categóricas no modelo, é possível formar uma tabela de contingência por meio das categorias da variável resposta e das combinações das categorias das variáveis explanatórias.

Suponha que em um modelo, a variável resposta tenha c categorias ordinais, as variáveis explanatórias sejam todas categóricas e o número de combinações dos níveis dessas variáveis explanatórias seja k. Seja $\boldsymbol{z_l}$ o vetor de valores das variáveis explanatórias que representa a l-ésima combinação, $l = 1, ..., k$. A Tabela 3.1 exemplifica uma tabela de contingência em que as k linhas são as combinações das categorias das variáveis explanatórias e as c colunas são as categorias da variável resposta. O número de indivíduos em cada célula é denotado por n_{lj}, com $n_l = \sum_j n_{lj}$ denotando a frequência marginal da linha l, $l = 1, ..., k$ e com $n = \sum_{l,j} n_{lj}$ denotando o total da amostra. Os valores das células são a contagem das observações que pertencem à categoria de resposta j e à combinação l das categorias das variáveis explanatórias.

Combinação das categorias das variáveis explanatórias	Categorias de resposta				
	1	2	...	c	**Total**
1	n_{11}	n_{12}	...	n_{1c}	$\mathbf{n_1}$
2	n_{21}	n_{22}	...	n_{2c}	$\mathbf{n_2}$
.	.	.	...	.	.
.	.	.	...	.	.
.	.	.	...	.	.
k	n_{k1}	n_{k2}	...	n_{kc}	$\mathbf{n_k}$

Tabela 3.1: *Tabela de contingência*

A partir de cada modelo, apresentado no Capítulo 2, é possível estimar a probabilidade da

variável resposta assumir a categoria j, para um dado vetor $\boldsymbol{z_l}$, $\pi_{lj} = P(Y = j|\boldsymbol{z_l})$. Por exemplo, no caso do Modelo logito cumulativo, as probabilidades cumulativas são expressas por:

$$P(Y \leq j|\boldsymbol{z_l}) = \frac{e^{\alpha_j + \boldsymbol{\beta}'\boldsymbol{z_l}}}{1 + e^{\alpha_j + \boldsymbol{\beta}'\boldsymbol{z_l}}}, j = 1, 2, ..., c-1.$$

Logo,

$$\pi_{lj} = P(Y = j|\boldsymbol{z_l}) = P(Y \leq j|\boldsymbol{z_l}) - P(Y \leq j-1|\boldsymbol{z_l}),$$

em que $\sum_j P(Y = j|\boldsymbol{z_l}) = 1$.

Com base nesta probabilidade, o número esperado E_{lj} de observações da variável resposta pertencentes à categoria j e à combinação l das variáveis explanatórias é dado por $E_{lj} = n_l\hat{\pi}_{lj}$, sendo $\hat{\pi}_{lj}$ a estimativa de π_{lj} obtida pelo modelo adotado.

3.2 Testes da qualidade do ajuste do modelo

Nesta seção serão apresentados testes para checar a adequabilidade dos modelos ajustados.

3.2.1 Teste de Pearson e teste da razão de verossimilhanças

Nos casos em que a tabela de contingência não possui valores esparsos, duas estatísticas são usualmente utilizadas para testar a qualidade do ajuste de um modelo: a estatística X^2 de Pearson e a estatística G^2 da razão de verossimilhanças.

Os testes comparam os valores observados na tabela de contingência com as frequências esperadas estimadas com base no modelo ajustado. A hipótese nula equivale ao modelo proposto estar bem ajustado.

A estatística X^2 de Pearson para testar a qualidade do ajuste do modelo é expressa por:

$$X^2 = \sum_l \sum_j \frac{(n_{lj} - E_{lj})^2}{E_{lj}}.$$

Já, a estatística G^2 da razão de verossimilhanças tem a seguinte forma:

$$G^2 = 2\sum_{l}\sum_{j} n_{lj} log\frac{n_{lj}}{E_{lj}}.$$

Sob a hipótese nula, X^2 e G^2 têm distribuição assintótica qui-quadrado (χ^2). Segundo Agresti (2010), o número de graus de liberdade é igual ao número de logitos menos o número de parâmetros do modelo ajustado.

Suponha, por exemplo, que o modelo proposto seja o Modelo logito cumulativo com chances proporcionais com apenas uma variável explanatória, com m ($m \geq 2$) categorias. O número de parâmetros do modelo ajustado é igual ao número de interceptos ($\{\alpha_j\}$), $j = 1, ..., c-1$, mais os $m-1$ parâmetros associados a variável explanatória, ou seja, $(c-1)+(m-1)$ parâmetros. Então o número de graus de liberdade para os testes é $m(c-1)-c-m+2$.

Giolo (2017) adotou a seguinte expressão para os graus de liberdade, $(k-1)(c-1)-q$, em que k é a combinação dos níveis das variáveis explanatórias, c é o número de categorias da variável resposta e q é o número de parâmetros do modelo associados às variáveis explanatórias (não leva em conta os parâmetros do intercepto, α_j). Esta expressão é equivalente à expressão utilizada por Agresti (2010).

Para que a distribuição das estatísticas X^2 e G^2 tenha uma boa aproximação pela distribuição χ^2, pelo menos 80% dos valores esperados devem ser maiores que 5 (Freeman Jr, 1987).

De acordo com Agresti (1996), quando o modelo contém um preditor contínuo, os testes X^2 e G^2 não são válidos, pois cada valor deste preditor contínuo representaria uma categoria e, o número de combinações k das categorias da variáveis explanatórias seria muito grande, com a provável ocorrência de valores esparsos.

Nas próximas subseções serão apresentados quatro testes para checar a qualidade do ajuste de modelos de regressão logística ordinais com chances proporcionais, Modelo logito cumulativo, Modelo logito categorias adjacentes e Modelo logito razão contínua, que devem ser utilizados quando há valores esparsos ou quando há preditores contínuos no modelo. Uma opção para checar a qualidade do ajuste de modelos de regressão logística ordinais sem chances proporcionais e que tenham variáveis explicativas contínuas no modelo, seria categorizar as variáveis explicativas contínuas e aplicar os métodos utilizados quando se tem apenas variáveis explanatórias categóricas no modelo.

Estes testes encontram-se implementados em vários sofwares, tais como R e SAS.

3.2.2 Teste de Lipsitz

O teste de Lipsitz *et al.* (1996) é baseado parcialmente no teste de Hosmer e Lemeshow (1980), utilizado para verificar a qualidade do ajuste de modelos de regressão logística com resposta binária.

Suponha uma amostra de n observações independentes, $(\boldsymbol{x_i}, y_i)$, $i = 1, ..., n$. Denote $\pi_{ij} = P(Y_i = j|\boldsymbol{x_i})$. Assim, $\hat{\pi}_{ij}$ é a probabilidade estimada de cada categoria da variável resposta para cada $\boldsymbol{x_i}$. Primeiramente, calcula-se a probabilidade estimada pelo modelo de regressão ordinal ajustado ($\hat{\pi}_{ij}$). Depois, atribui-se um escore (s_i) para cada observação, usando pesos igualmente espaçados:

$$s_i = \hat{\pi}_{i1} + 2\hat{\pi}_{i2} + ... + c\hat{\pi}_{ic},\ i = 1, ..., n. \tag{3.1}$$

Depois de atribuídos os escores, ordenam-se as observações com base nos escores obtidos e formam-se g grupos, de modo que, o 1° grupo contenha as n/g observações com os menores escores e o grupo g tenha as n/g observações com os maiores escores. Após a criação dos grupos, criam-se $g - 1$ variáveis indicadoras binárias, da seguinte forma:

$$\boldsymbol{I_{iv}} = \begin{cases} 1, & \text{se a observação } i \text{ está no grupo } v. \\ 0, & \text{caso contrário.} \end{cases}$$

para $i = 1, ..., n$ e $v = 1, ..., g - 1$.

Então, ajusta-se um novo modelo de regressão ordinal que inclua as variáveis indicadoras,

$$h_{ij} = \alpha_j + \boldsymbol{\beta}^{'}\boldsymbol{x} + \sum_{v=1}^{g-1} \gamma_v I_v,\ j = 1, ..., c - 1,$$

em que h_{ij} representa a função de ligação relacionada ao modelo proposto.

De acordo com o Capítulo 2, para o Modelo logito cumulativo com chances proporcionais,

$$h_{ij} = log\frac{\pi_1(\boldsymbol{x_i}) + ... + \pi_j(\boldsymbol{x_i})}{\pi_{j+1}(\boldsymbol{x_i}) + ... + \pi_c(\boldsymbol{x_i})}.$$

Já para o Modelo categorias adjacentes com chances proporcionais,

$$h_{ij} = log\frac{\pi_j(\boldsymbol{x_i})}{\pi_{j+1}(\boldsymbol{x_i})},$$

e, para o Modelo logito razão contínua com chances proporcionais,

$$h_{ij} = log\frac{\pi_j(\boldsymbol{x_i})}{\pi_{j+1}(\boldsymbol{x_i}) + ... + \pi_c(\boldsymbol{x_i})}.$$

A princípio, outras funções de ligação, além da função de ligação *logito*, podem ser usadas, tais como, a ligação *probito* ou a ligação *complemento log-log*.

Pode-se utilizar a estatística da razão de verossimilhanças, a estatística de Wald ou a estatística escore para testar:

$$H_0 : \gamma_1 = ... = \gamma_{g-1} = 0.$$

Se H_0 não for rejeitada, há a indicação de que o modelo adotado está corretamente especificado.

Denotemos L_0 e L_1 como o logaritmo da função de verossimilhança do modelo ajustado sem e com as variáveis indicadoras, respectivamente. O valor-p do teste associado à estatística da razão de verossimilhanças é obtido aproximando-se a estatística $-2(L_0 - L_1)$ pela distribuição χ^2 com $g - 1$ graus de liberdade.

De acordo com Lipsitz *et al.* (1996), a estatística $-2(L_0 - L_1)$ pode ser escrita como uma combinação linear da diferença $(O_{vj} - E_{vj})$, em que $O_{vj} = \sum_{i=1}^{n} I_{iv}Y_{ij}$ é o número de observações da variável resposta pertencentes à categoria j e ao grupo v e $E_{vj} = \sum_{i=1}^{n} I_{iv}\hat{\pi}_{ij}$ é o número esperado de observações da variável resposta pertencentes à categoria j e ao grupo v.

Para se garantir que a estatística do teste tenha distribuição aproximadamente χ^2 com $g - 1$ graus de liberdade, a amostra deve ser grande. Espera-se que o número esperado de indivíduos na categoria de resposta j seja maior que 1 em cada um dos g grupos, e que em pelo menos 80% dos g grupos, o número esperado seja maior que 5 (Freeman Jr, 1987). Para situações em que isso não acontece, a aproximação χ^2 pode ser pobre.

Uma solução possível para a situação de amostra pequena é usar um número pequeno de grupos. Na aproximação de Hosmer e Lemeshow (1980) para dados binários, é recomendado formar 10 grupos ($g = 10$) de tamanhos aproximadamente iguais. Lipsitz *et al.* (1996) sugerem que o número

de grupos seja escolhido no intervalo $6 \leq g < \frac{n}{5c}$.

Este teste encontra-se implementado no R por meio do pacote "generalhoslem" para o Modelo logito cumulativo com chances proporcionais. Atualmente este teste também foi implementado no *software* Stata (Stata *et al.*, 2017) para o Modelo logito categorias adjacentes com chances proporcionais e para o Modelo logito razão contínua com chances proporcionais.

3.2.3 Teste de Pulkstenis-Robinson

Pulkstenis e Robinson (2004) apresentaram uma modificação das estatísticas X^2 de Pearson e G^2 da razão de verossimilhanças para ser utilizada quando preditores contínuos e categóricos estão presentes simultaneamente no modelo.

Primeiramente, são determinadas as combinações das variáveis que serão utilizadas usando somente as variáveis categóricas do modelo, em que as categorias não observadas são desconsideradas. Em seguida, calculam-se os escores da mesma maneira que em (3.1) e então cada combinação l das variáveis explanatórias é dividida em duas com base na mediana dos escores pertencentes a cada combinação. É construída uma tabela com as frequências observadas e estimadas e, a partir desta tabela é obtida a estatística modificada de Pearson e da razão de verossimilhanças, por meio das fórmulas descritas abaixo, respectivamente:

$$X^2_{PR} = \sum_{l=1}^{k}\sum_{t=1}^{2}\sum_{j=1}^{c}\frac{(n_{ltj} - E_{ltj})^2}{E_{ltj}}.$$

$$G^2_{PR} = \sum_{l=1}^{k}\sum_{t=1}^{2}\sum_{j=1}^{c} n_{ltj} log\frac{n_{ltj}}{E_{ltj}}.$$

em que l representa as combinações das variáveis explanatórias, t representa os dois subgrupos baseados nos escores ordinais, j representa as categorias da variável resposta, n_{ltj} representa o número de indivíduos pertencentes à categoria j, à l-ésima combinação das variáveis explanatórias e ao t-ésimo subgrupo e E_{ltj} representa o número esperado de observações da variável resposta pertencentes à categoria j, à l-ésima combinação das variáveis explanatórias e ao t-ésimo subgrupo, ou seja, é a soma das probabilidades estimadas pelo modelo ajustado para as $n_{lt} = \sum_{j=1}^{c} n_{ltj}$ observações.

As duas estatísticas seguem uma distribuição χ^2 com $(2k-1)(c-1)-m-1$ graus de liberdade,em que $2k$ é o número de linhas da tabela de contingência, c é o número de categorias da variável resposta, m é o número de termos categóricos do modelo e foi subtraído um grau de liberdade

devido a contribuição das variáveis contínuas. Por exemplo, caso fosse ajustado um modelo com uma variável explanatória dicotômica e outra variável explanatória categórica com quatro níveis, então $m = 4$. A fim de garantir que a estatística do teste tenha a distribuição aproximada χ^2, foi sugerido que a maioria (80%) dos valores esperados seja maior que 5.

Este teste encontra-se implementado no R por meio do pacote "generalhoslem" para o Modelo logito cumulativo com chances proporcionais. Atualmente este teste também foi implementado no *software* Stata para o Modelo logito categorias adjacentes com chances proporcionais e para o Modelo logito razão contínua com chances proporcionais.

3.2.4 Versão ordinal do teste de Hosmer-Lemeshow

Para desenvolver este teste, Fagerland e Hosmer (2013) basearam-se no teste de Hosmer - Lemeshow para regressão logística binária. Assim como nos outros testes, calcula-se a probabilidade estimada do modelo de regressão ordinal ajustado ($\hat{\pi}_{ij}$) e atribui-se um escore (s_i) para cada observação, como em (3.1).

A seguir, ordenam-se as observações com base nos escores obtidos e formam-se g grupos, de modo que, o 1° grupo contenha as observações com os n/g menores escores e o grupo g tenha as n/g observações com os maiores escores. Foi sugerido formar 10 grupos ($g = 10$) de tamanhos aproximadamente iguais, assim como na aproximação de Hosmer-Lemeshow para dados binários.

Constrói-se, então, uma tabela de contingência com g linhas e c colunas, conforme mostra a Tabela 3.2, em que n_{vj} representa o número de observações pertencentes ao v-ésimo grupo e à j-ésima categoria da resposta e E_{vj} representa o número esperado de observações pertencentes ao v-ésimo grupo e à j-ésima categoria da resposta, conforme definido na Seção 3.2.2. A versão ordinal da estatística do teste de Hosmer-Lemeshow é dada por:

$$C_g = \sum_{v=1}^{g} \sum_{j=1}^{c} \frac{(n_{vj} - E_{vj})^2}{E_{vj}}.$$

A distribuição de C_g é aproximadamente χ^2 com $(g-2)(c-1) + (c-2)$ graus de liberdade, em que $(g-2)(c-1)$ é referente ao teste de Hosmer-Lemeshow binário e $(c-2)$ é uma correção feita para refletir o fato de que a soma das frequências estimadas não é igual à das observadas em cada um dos c níveis da resposta. Um grau de liberdade é perdido por conta da ordenação dos interceptos.

Este teste encontra-se implementado no R por meio do pacote "generalhoslem" para o Modelo

	Categorias de resposta				
Grupo	1	2	...	c	**Total**
1	n_{11}	n_{12}	...	n_{1c}	n_1
2	n_{21}	n_{22}	...	n_{2c}	n_2
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
g	n_{g1}	n_{g2}	...	n_{gc}	n_g

Tabela 3.2: *Frequências observadas*

logito cumulativo com chances proporcionais. Atualmente este teste também foi implementado no *software* Stata para o Modelo logito categorias adjacentes com chances proporcionais e para o Modelo logito razão contínua com chances proporcionais.

3.2.5 Comparação entre os testes

Fagerland e Hosmer (2013) examinaram o poder do teste de Lipsitz, da versão ordinal do teste Hosmer-Lemeshow e dos testes qui-quadrado e razão de verossimilhanças de Pulkstenis-Robinson para o Modelo logito cumulativo com chances proporcionais para seis diferentes tipos de situações. Já, Fagerland e Hosmer (2016) avaliaram o poder do teste de Lipsitz, da versão ordinal do teste Hosmer-Lemeshow e dos testes qui-quadrado e da razão de verossimilhanças de Pulkstenis-Robinson para o Modelo logito categorias adjacentes e para o Modelo logito razão contínua, com chances proporcionais, em oito diferentes tipos de situações. A seguir serão apresentados os resultados obtidos nos artigos supramencionados.

No Modelo logito cumulativo com chances proporcionais, os testes não apresentaram bom poder para detectar omissão do termo quadrático. O teste Lipsitz foi o que apresentou melhor performance, entretanto para obter um poder maior que 50% é preciso ter uma amostra grande e um termo quadrático influente. Já nos Modelos logito categorias adjacentes e razão contínua, o teste de Lipsitz obteve poder maior do que a versão ordinal do teste de Hosmer-Lemeshow para detectar a omissão de termo quadrático da variável explanatória contínua, enquanto os testes de Pulkstenis-Robinson obtiveram poder baixo.

Os quatro testes obtiveram poder maior que 90% para detectar a falta do termo de interação, tanto entre uma variável explanatória contínua e uma variável explanatória categórica dicotômica quanto entre duas variáveis explanatórias categóricas dicotômicas no Modelo logito cumulativo com chances proporcionais. Os testes Lipsitz e Pulkstenis-Robinson apresentaram maior poder do que a versão ordinal do teste Hosmer-Lemeshow.

Todos os testes obtiveram excelente poder para detectar a falta de interação entre uma variável explanatória contínua e uma variável explanatória categórica dicotômica para os modelos logito

categorias adjacentes e logito razão contínua.

Para o Modelo logito categorias adjacentes, todos os testes apresentaram bom poder para detectar a falta de interação entre duas variáveis explanatórias dicotômicas. Já para o Modelo logito razão contínua, os testes de Pulkstenis-Robinson apresentaram excelente poder, enquanto os testes Lipsitz e Hosmer-Lemeshow obtiveram poder menor.

Nenhum dos quatro testes detectou erro na função da variável explanatória contínua, quando o Modelo logito cumulativo com chances proporcionais foi ajustado com o termo x ao invés de $log(x)$.

Para o Modelo logito categorias adjacentes, a versão ordinal do teste Hosmer-Lemeshow apresentou poder baixo a moderado para detectar erro na função da variável explanatória contínua, enquando o teste Lipsitz apresentou poder moderado a bom e os testes de Pulkstenis-Robinson apresentaram fraco poder. No Modelo logito razão contínua, a versão ordinal do teste Hosmer-Lemeshow apresentou poder baixo, o teste Lipsitz apresentou poder baixo a moderado e os testes de Pulkstenis-Robinson apresentaram poder baixo.

Para o Modelo logito cumulativo com chances proporcionais, os testes de Pulkstenis-Robinson e a versão ordinal do teste de Hosmer-Lemeshow apresentaram boa performance para detectar a falta de proporcionalidade do modelo, enquanto o teste Lipsitz não apresentou.

O poder de todos os testes foi baixo para verificar a falta de uma variável explanatória independente para os Modelos logito categorias adjacentes e razão contínua.

Fagerland e Hosmer (2013) e Fagerland e Hosmer (2016) recomendaram usar os quatro tipos de testes, visto que os quatro testes podem exibir poderes diferentes dependendo do tipo de falta de ajuste. Por conta do poder ser geralmente baixo para amostras de tamanho pequeno, foi recomendado usar um nível de significância de 10%, caso a amostra seja menor que 400.

3.3 Teste de proporcionalidade

Uma característica dos modelos com estrutura de chances proporcionais é que o efeito das variáveis explanatórias é o mesmo para cada um dos c-1 logitos. Assim, é de interesse verificar a propriedade de chances proporcionais. O teste de proporcionalidade está associado às hipóteses:

$$H_0 : \ o\ modelo\ é\ dado\ por\ 2.2,\ 2.5\ ou\ 2.7\ (modelos\ com\ chances\ proporcionais) \tag{3.2}$$

versus,

$$H_1 : \ o\ modelo\ é\ dado\ por\ 2.3,\ 2.6\ ou\ 2.8\ (modelos\ sem\ chances\ proporcionais) \tag{3.3}$$

Para testar H_0, pode-se utilizar a estatística da razão de verossimilhanças, a estatística de Wald, proposta por Brant (1990) ou a estatística escore, proposta por Peterson e Harrell Jr (1990).

A estatística do teste da razão de verossimilhanças é dada por $-2(L_0 - L_1)$, em que L_0 e L_1 denotam o logaritmo da função de verossimilhança do modelo em (3.2) e (3.3) respectivamente.

Este teste segue aproximadamente a distribuição qui-quadrado (χ^2). O número de graus de liberdade equivale à diferença entre o número de parâmetros do modelo sob a hipótese alternativa e o número de parâmetro do modelo proposto, ou seja, o modelo em (3.3) tem $p(c-1)$ parâmetros associados às p variáveis explanatórias, enquanto o modelo em (3.2) tem p parâmetros associados às p variáveis explanatórias. Logo, o número de graus de liberdade é $p(c-2)$.

Segundo Agresti (2010), quando não for possível ajustar o modelo mais complexo, o que acontece por vezes no Modelo logito cumulativo sem chances proporcionais, é possível utilizar o teste escore, visto que este teste compara os modelos usando o log da função de verossimilhança somente sob a hipótese nula. Já, quando for possível ajustar o modelo mais complexo, além do teste da razão de verossimilhanças, é também possível utilizar o teste de Wald.

Kim (2003) sugeriu plotar as probabilidades estimadas obtidas sob a estrutura do modelo proporcional contra as probabilidades estimadas no modelo que permite diferentes efeitos. Em termos práticos, a falta de ajuste não é severa se os pares de de probabilidades estimadas caem próximos da linha de intercepto 0 e declive 1.

3.4 Análise de resíduos

Uma maneira alternativa de checar a falta de ajuste quando se tem apenas variáveis explanatórias categóricas no modelo é por meio da análise de resíduos. Para uma tabela de contingência com o valor da célula n_{lj} e número esperado $E_{lj} = n_l\hat{\pi}_{lj}$ para a l-ésima combinação das variáveis explanatórias e categoria resposta j, o resíduo padronizado é:

$$r_{lj} = \frac{n_{lj} - n_l\hat{\pi}_{lj}}{\sqrt{n_l\hat{\pi}_{lj}[1-\hat{\pi}_{lj}]}}.$$

Os resíduos padronizados têm distribuição aproximadamente normal padrão. Valores grandes,

como excedendo 3 em valor absoluto, indicam falta de ajuste na célula.

3.5 Critério de informação Akaike

Outra maneira que pode ajudar a selecionar um bom modelo é utilizar o método proposto por Akaike (1974), o qual é um processo de minimização e não envolve testes estatísticos.

O critério de informação de Akaike *(Akaike information criterion - AIC)* consiste em encontrar um modelo tal que a quantidade abaixo seja minimizada:

$$AIC = -2(logL(\hat{\theta}) - p),$$

em que θ é o vetor de parâmetros do modelo proposto, $L(\hat{\theta})$ é o logarítimo da função de verossimilhança avaliado na estimativa de máxima verossimilhança de θ e p o número de parâmetros do modelo.

Agresti (2010) mencionou que as variáveis explanatórias ordinais podem ser tratadas como quantitativas, uma vez que escores numéricos sejam atribuídos às categorias. Como o método AIC penaliza o modelo que tem mais parâmetros, ele deve ser usado com cautela na escolha de um modelo que utiliza variáveis explanatórias categóricas como quantitativas.

3.6 Método de seleção

Para selecionar variáveis em um modelo podemos utilizar um dos métodos de seleção como *stepwise*, *forward* ou *backward*, ou utilizar o método AIC.

Capítulo 4

Aplicações

4.1 Dados utilizados

Os dados que ilustram este trabalho foram coletados em uma Unidade de Referência Especializada (URE) do serviço público, na cidade de Belém-PA, no período de junho de 2016 a fevereiro de 2017 (Magrini, 2015). Este estudo tem como objetivo investigar a relação entre perda auditiva, equilíbrio e aspectos emocionais no idoso.

Os dados referem-se a 138 indíviduos com idade superior a 60 anos e com perda auditiva. As variáveis demográficas incluem informações quanto ao sexo (1=masculino, 2=feminino), escolaridade (0=analfabeto, 1=ensino fundamental, 2=ensino médio, 3=graduação), renda mensal (1=sem renda, 2=1 salário mínimo, 3=maior que 1 a 2 salários mínimos, 4=maior que 2 salários mínimos) e idade (em anos).

Para avaliar a depressão foi utilizada a Escala de Depressão Geriátrica formada com 15 perguntas afirmativas e negativas em que o resultado de cinco ou mais pontos caracteriza o diagnóstico de depressão (0=não, 1=sim).

Foram feitas perguntas de auto referência sobre audição, como, quando a família começou a perceber o problema de audição, contendo as respostas 6 meses (código 1), 1 ano (código 2), 2 anos (código 3) e mais de dois anos (código 4).

Foram realizados testes de equilíbrio estático, dinâmico e de mobilidade, entre eles, a prova de Unterberg com olhos abertos. Esta prova consiste em o paciente executar movimento de marcha sem sair do lugar e braços estendidos à sua frente. Considerou-se positivo (código 1), caso o idoso conseguisse realizar a prova e negativo (código 0) se ele não conseguisse realizar.

Outro teste de equilíbrio estático, dinâmico e de mobilidade, realizado pelos idosos, foi a prova Time Up and Go (TUG) que consiste em cronometrar o tempo que o paciente leva para se levantar de uma cadeira, fazer um percurso de três metros e voltar a sentar na cadeira. O resultado com tempo menor que 11 segundos (código 1) é considerado normal, são idosos independentes e sem risco de quedas; entre 11 e 20 segundos (código 2), são idosos que apresentam independência

parcial e com baixo risco de quedas e o tempo acima de 20 segundos (código 3), são idosos que apresentam déficits importantes de mobilidade física e risco de quedas.

Por último, foi realizada uma auto avaliação sobre audição, por meio da pergunta em geral você diria que sua audição é ruim (código 1), regular (código 2), boa (código 3), muito boa (código 4) e excelente (código 5). As categorias muito boa e excelente foram desconsideradas em virtude de não haver esses resultados na amostra.

Para este estudo, foram consideradas como variáveis explanatórias as variáveis sexo, escolaridade, renda mensal, idade, depressão, quando a família começou a perceber e prova de Unterberg com olhos abertos. As variáveis respostas consideradas foram a prova Time Up and Go e em geral você diria que sua audição é. Para o modelo que considerou como variável resposta a prova Time Up and Go, utilizaram-se as variáveis explanatórias supracitadas com exceção da variável prova de Unterberg com olhos abertos, que não era de interesse no estudo.

No Apêndice A encontra-se a distribuição conjunta entre as variáveis explanatórias e cada uma das variáveis respostas.

4.2 Análise inferencial

Para ilustrar algumas aplicações com este banco de dados, optou-se por ajustar o Modelo logito cumulativo com chances proporcionais para a variável prova Time Up and Go e o Modelo logito categorias adjacentes com chances proporcionais para a variável em geral você diria que sua audição é. Utilizou-se o *software* R para fazer as análises. O teste de Lipsitz, a versão ordinal do teste de Hosmer-Lemeshow e os testes qui-quadrado e da razão de verossimilhanças de Pulkstenis-Robinson estão implementados neste *software* por meio do pacote "generalhoslem" para o Modelo logito cumulativo com chances proporcionais. Como estes testes ainda não encontram-se implementados para o Modelo logito categorias adjacentes com chances proporcionais, estes foram desenvolvidos nesta dissertação. Nas análises, adotou-se o número de grupos igual a seis ($g = 6$) para o teste de Lipsitz e para a versão ordinal do teste de Hosmer-Lemeshow. Utilizou-se o nível de significância de 10% para a seleção das variáveis explanatórias.

4.2.1 Prova Time Up and Go

Para a variável prova Time Up and Go, ajustou-se o Modelo logito cumulativo com chances proporcionais considerando-se duas situações: na primeira, as variáveis explanatórias com mais de duas categorias foram tratadas como quantitativas e, na segunda, como qualitativas.

Para a primeira situação, a Tabela 4.1 apresenta os valores das estatísticas, os graus de liberdade e os valores-p do teste de Lipsitz, da versão ordinal do teste Hosmer-Lemeshow e dos testes qui-

quadrado e da razão de verossimilhanças de Pulkstenis-Robinson para verificar a qualidade do ajuste do modelo completo contendo todas as variáveis explanatórias. Os resultados dessa tabela evidenciam o bom ajuste do modelo completo.

Teste	Valor da estatística	Graus de liberdade	valor-p
Lipsitz	$6,96$	5	$0,224$
Hosmer-Lemeshow	$10,94$	9	$0,280$
X^2_{PR}	$10,79$	11	$0,461$
G^2_{PR}	$13,89$	11	$0,239$

Tabela 4.1: *Testes de qualidade do ajuste - Prova Time Up and Go - situação 1*

Para a seleção das variáveis explanatórias utilizou-se o método *backward*. As variáveis que obtiveram efeito significativo foram idade (Z_1) e renda mensal (Z_2).

Para avaliar a validade da suposição de chances proporcionais para as duas variáveis explanatórias que obtiveram efeito significativo, aplicou-se o teste de proporcionalidade. O teste mostrou que o Modelo logito cumulativo com chances proporcionais parece adequado (valor-p $= 0,328$).

A Tabela 4.2 mostra as estimativas dos parâmetros, respectivos erros padrões e valores-p para o modelo final ajustado.

Variável	Parâmetros	Estimativa	Erro Padrão	valor p
Intercepto 1	α_1	$3,8843$	$2,0056$	$0,053$
Intercepto 2	α_2	$8,2238$	$2,1551$	$< 0,001$
Idade	γ_1	$-0,1034$	$0,0276$	$< 0,001$
Renda mensal	γ_2	$0,9283$	$0,2697$	$0,001$

Tabela 4.2: *Estimativas dos parâmetros do modelo final - Prova Time Up and Go - situação 1*

O modelo final ajustado pode ser expresso em termos dos logitos, por:

$$log\left[\frac{\hat{P}(Y_i \leq j|\boldsymbol{z_i})}{1-\hat{P}(Y_i \leq j|\boldsymbol{z_i})}\right] = \hat{\alpha_j} - 0,1034Z_1 + 0,9283Z_2,\ j = 1,2,$$

em que Z_1 representa a idade em anos; $Z_2 = 1$, se o idoso não tem renda; $Z_2 = 2$, se o idoso tem renda mensal de 1 salário mínimo; $Z_2 = 3$, se o idoso tem renda maior que 1 salário a 2 salários mínimos; e $Z_2 = 4$, se o idoso tem salário maior que 2 salários mínimos.

Quanto à interpretação, tem-se que:

- fixada a renda mensal, a chance de um idoso com z_1 anos demorar menos que 11 segundos para realizar a prova Time Up and Go é $\exp(\hat{\gamma_1}) = 0,90$ vezes a chance de um idoso com $(z_1 - 1)$ anos demorar menos que 11 segundos para realizar a prova;

- fixada a idade, a chance de um idoso com renda mensal igual a z_2 salários mínimos demorar

menos que 11 segundos para realizar a prova Time Up and Go é $\exp(\hat{\gamma}_2) = 2,53$ vezes a chance de um idoso com renda mensal igual a $(z_2 - 1)$ salários mínimos demorar menos que 11 segundos para realizar a prova.

Devido à suposição de chances proporcionais assumida para o modelo ajustado aos dados, as mesmas conclusões são obtidas quanto à chance de um idoso demorar até 20 segundos para realizar a prova Time Up and Go.

Na segunda situação, as variáveis explanatórias com mais de duas categorias foram tratadas como qualitativas. A Tabela 4.3 apresenta os valores das estatísticas, os graus de liberdade e os valores-p do teste de Lipsitz, da versão ordinal do teste de Hosmer-Lemeshow e dos testes qui-quadrado e razão de verossimilhanças de Pulkstenis-Robinson para verificar a qualidade do ajuste do modelo completo contendo todas as variáveis explanatórias. Os resultados dessa tabela evidenciam o bom ajuste do modelo completo.

Teste	Valor da estatística	Graus de liberdade	valor-p
Lipsitz	$3,45$	5	$0,630$
Hosmer-Lemeshow	$4,79$	9	$0,852$

Tabela 4.3: *Testes de qualidade do ajuste - Prova Time Up and Go - situação 2*

Para a seleção das variáveis explanatórias utilizou-se o método *backward*. As variáveis que obtiveram efeito significativo foram idade (Z_1) e renda mensal (Z_2).

Para avaliar a validade da suposição de chances proporcionais para as duas variáveis explanatórias que obtiveram efeito significativo, aplicou-se o teste de proporcionalidade. O teste mostrou que o Modelo logito cumulativo com chances proporcionais parece adequado (valor-p $= 0,147$).

A Tabela 4.4 mostra as estimativas dos parâmetros, respectivos erros padrões e valores-p para o modelo final ajustado.

Variável	Parâmetros	Estimativa	Erro Padrão	valor-p
Intercepto 1	α_1	$4,8272$	$1,9805$	$0,015$
Intercepto 2	α_2	$9,1651$	$2,1487$	$<0,001$
Idade	γ_1	$-0,1053$	$0,0287$	$< 0,001$
Renda mensal de 1 salário mínimo	γ_2	$1,1229$	$0,8359$	$0,179$
Renda mensal $>$ 1 a 2 salários mínimos	γ_3	$1,6628$	$0,9195$	$0,075$
Renda mensal $>$ 2 salários mínimos	γ_4	$3,1578$	$1,0254$	$0,002$

Tabela 4.4: *Estimativas dos parâmetros do modelo final - Prova Time Up and Go - situação 2*

O modelo final ajustado pode ser expresso em termos dos logitos, por:

$$log\left[\frac{\hat{P}(Y_i \leq j|\boldsymbol{z_i})}{1-\hat{P}(Y_i \leq j|\boldsymbol{z_i})}\right] = \hat{\alpha_j} - 0,1053Z_1 + 1,1229Z_2 + 1,6628Z_3 + 3,1578Z_4,\ j = 1,2,$$

em que Z_1 representa a idade em anos; $Z_2 = 1$, se o idoso tem renda mensal de 1 salário mínimo; $Z_2 = 0$, caso contrário; $Z_3 = 1$, se o idoso tem renda mensal maior que 1 a 2 salários mínimos; $Z_3 = 0$, caso contrário; e $Z_4 = 1$, se o idoso tem salário maior que 2 salários mínimos; $Z_4 = 0$, caso contrário.

Quanto à interpretação, tem-se que:

- fixada a renda mensal, a chance de um idoso com z_1 anos demorar menos que 11 segundos para realizar a prova Time Up and Go é $\exp(\hat{\gamma_1}) = 0,90$ vezes a chance de um idoso com $(z_1 - 1)$ anos demorar menos que 11 segundos para realizar a prova;

- fixada a idade, a chance de um idoso com renda mensal de 1 salário mínimo demorar menos do que 11 segundos para realizar a prova Time Up and Go é $\exp(\hat{\gamma_2}) = 3,07$ vezes a chance de um idoso que não tenha renda demorar menos de 11 segundos para realizar a prova Time Up and Go;

- fixada a idade, a chance de um idoso com renda mensal maior que 1 salário a 2 salários mínimos demorar menos que 11 segundos para realizar a prova Time Up and Go é $\exp(\hat{\gamma_3} - \hat{\gamma_2}) = 1,72$ vezes a chance de um idoso que tenha renda mensal de 1 salário mínimo demorar menos que 11 segundos para realizar a prova Time Up and Go;

- fixada a idade, a chance de um idoso com renda mensal maior que 2 salários mínimos demorar menos que 11 segundos para realizar a prova Time Up and Go é $\exp(\hat{\gamma_4} - \hat{\gamma_3}) = 4,46$ vezes a chance de um idoso com renda mensal maior que 1 salário a 2 salários mínimos demorar menos que 11 segundos para realizar a prova Time Up and Go.

Devido à suposição de chances proporcionais assumida para o modelo ajustado aos dados, as mesmas conclusões são obtidas quanto à chance de um idoso demorar até 20 segundos para realizar a prova Time Up and Go.

Finalmente, utilizou-se o Critério de Informação Akaike (AIC) para comparar o modelo considerando as variáveis explanatórias com mais de duas categorias como quantitativas com o modelo que considera as variáveis explanatórias como qualitativas. A partir da Tabela 4.5, nota-se que o menor AIC foi para o modelo mais simples, o qual considera as variáveis explanatórias com mais de duas categorias como quantitativas. Entretanto, este método deve ser utilizado com cautela, uma vez que, este critério favorece o modelo mais simples pois penaliza o modelo com mais parâmetros.

Por fim, ajustou-se o Modelo logito categorias adjacentes com chances proporcionais para as mesmas variáveis explanatórias e resposta e obteve-se as mesmas variáveis explanatórias significa-

Modelo	AIC
Modelo (qualitativa)	204,03
Modelo (quantitativa)	199,44

Tabela 4.5: *AIC dos modelos ajustados - Prova Time Up and Go*

tivas, tanto para o modelo que considera as variáveis explanatórias com mais de duas categorias como quantitativas quanto para o modelo que considera como qualitativas.

4.2.2 Em geral você diria que sua audição é

Para a variável em geral você diria que a sua audição é, ajustou-se o Modelo logito categorias adjacentes com chances proporcionais, considerou-se duas situações: na primeira, as variáveis explanatórias com mais de duas categorias foram tratadas como quantitativas e, na segunda, como qualitativas.

Para a primeira situação, a Tabela 4.6 apresenta os valores das estatísticas, os graus de liberdade e os valores-p do teste de Lipsitz, da versão ordinal do teste Hosmer-Lemeshow e dos testes qui-quadrado e razão de verossimilhanças de Pulkstenis-Robinson para verificar a qualidade do ajuste do modelo completo contendo todas as variáveis explanatórias. Com exceção do teste X^2_{PR}, os resultados dessa tabela evidenciam o bom ajuste do modelo completo.

Teste	Valor da estatística	Graus de liberdade	valor-p
Lipsitz	$4,46$	5	$0,486$
Hosmer-Lemeshow	$9,05$	9	$0,432$
X^2_{PR}	$33,27$	22	$0,058$
G^2_{PR}	$12,55$	22	$0,293$

Tabela 4.6: *Testes de qualidade do ajuste - Em geral você diria que sua audição é - situação 1*

Para avaliar a validade da suposição de chances proporcionais no modelo completo, aplicou-se o teste de proporcionalidade. O teste mostrou que o Modelo logito categorias adjacentes com chances proporcionais parece adequado (valor p $= 0,179$).

Excluindo a variável prova de Unterberg com olhos abertos que, como mostra a Tabela A.11, apresenta $93,5\%$ de respostas negativas, todos os testes de qualidade do ajuste apresentaram valores-p superiores a 10%.

Para a seleção das variáveis utilizou-se o método *backward*. A variável explanatória que obteve efeito significativo foi a variável depressão (X_1).

A Tabela 4.7 mostra as estimativas dos parâmetros, respectivos erros padrões e valores-p para o modelo final ajustado.

O modelo final ajustado pode ser expresso em termos dos logitos, por:

Variável	Parâmetros	Estimativa	Erro Padrão	valor p
Intercepto 1	α_1	$-0,6427$	$0,2785$	$0,021$
Intercepto 2	α_2	$1,8990$	$0,3942$	$< 0,001$
Depressão	β_1	$0,5927$	$0,3096$	$0,056$

Tabela 4.7: *Estimativas dos parâmetros do modelo final adotado*

$$log\left[\frac{\hat{P}(Y_i = j|\boldsymbol{x_i})}{\hat{P}(Y_i = j+1|\boldsymbol{x_i})}\right] = \hat{\alpha_j} + 0,5927X_1,\ j = 1, 2,$$

em que $X_1 = 0$, se o idoso não tem depressão; e $X_1 = 1$, se o idoso tem depressão.

Em virtude do modelo final não ter variáveis contínuas, utilizou-se a estatística de Pearson (X^2) e a razão de verossimilhanças (G^2) para verificar a qualidade do ajuste desse modelo. Os resultados dos testes de qualidade do ajuste podem ser vistos na Tabela 4.8, indicando evidências a favor do modelo.

Teste	Valor da estatística	Graus de liberdade	valor p
Pearson (X^2)	$1,56$	1	$0,212$
Razão de Verossimilhanças (G^2)	$1,56$	1	$0,212$

Tabela 4.8: *Testes de qualidade do ajuste - Em geral você diria que sua audição é - situação 1 - modelo final*

A Figura 4.1 mostra o gráfico dos resíduos de Pearson em função das combinações das categorias da variável depressão com as categorias da variável resposta. Como os valores situam-se em torno de zero, há evidências a favor do modelo.

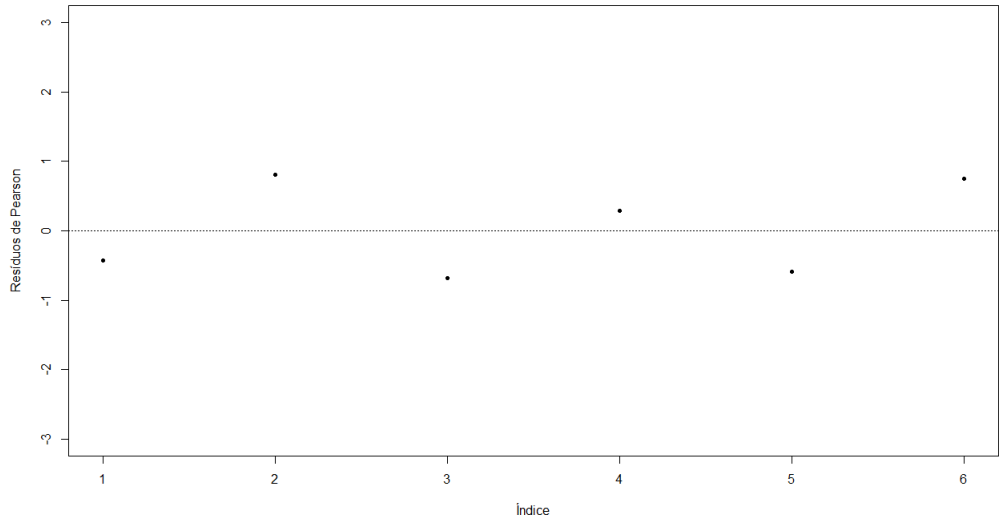

Figura 4.1: *Resíduos do Modelo logito categorias adjacentes - Em geral você diria que sua audição é - situação 1*

Quanto à interpretação, tem-se que:

- a chance de um idoso com depressão dizer que a sua audição é ruim em relação à regular é $\exp(\hat{\beta}_1) = 1,8$ vezes a chance de um idoso que não tem depressão dizer que a sua audição é ruim em relação à regular.

Devido à suposição de chances proporcionais assumida para o modelo ajustado aos dados, a mesma conclusão é obtida quanto à chance de um idoso dizer que a sua audição é regular em relação à boa.

Também é possível verificar a chance do idoso dizer que a audição é ruim em relação à boa. Neste caso o modelo pode ser expresso por:

$$log\left[\frac{\hat{P}(Y_i = j|\boldsymbol{x_i})}{\hat{P}(Y_i = 3|\boldsymbol{x_i})}\right] = \sum_{j=1}^{2} \hat{\alpha_j} + 0,5927(3-j)X_1,\ j = 1, 2,$$

em que $X_1 = 0$, se o indivíduo não tem depressão; e $X_1 = 1$, se o indivíduo tem depressão.

Quanto à interpretação, tem-se que:

- a chance de um idoso com depressão dizer que a sua audição é ruim em relação à boa é $\exp(2.\hat{\beta_1}) = 3,3$ vezes a chance de um idoso que não tem depressão dizer que a sua audição é ruim em relação à boa.

Na segunda situação, as variáveis explanatórias com mais de duas categorias foram tratadas como qualitativas. A Tabela 4.9 apresenta os valores das estatísticas, os graus de liberdade e os valores-p do teste de Lipsitz, da versão ordinal do teste de Hosmer-Lemeshow e dos testes qui-quadrado e razão de verossimilhanças de Pulkstenis-Robinson para verificar a qualidade do ajuste para o modelo completo contendo todas as variáveis explanatórias. Os resultados dessa tabela evidenciam o bom ajuste do modelo completo, exceto para o teste qui-quadrado de Pulkstenis-Robinson.

Teste	**Valor da estatística**	**Graus de liberdade**	**valor-p**
Lipsitz	$4,96$	5	$0,421$
Hosmer-Lemeshow	$4,98$	9	$0,836$

Tabela 4.9: *Testes de qualidade do ajuste - Em geral você diria que sua audição é - situação 2*

Para avaliar a validade da suposição de chances proporcionais no modelo completo, aplicou-se o teste de proporcionalidade. O teste mostrou que o Modelo logito categorias adjacentes com chances proporcionais parece adequado (valor-p $= 0,197$).

Novamente, ao excluir-se a variável prova de Unterberg com olhos abertos, o modelo logito categorias adjacentes com chances proporcionais mostrou-se bem ajustado nos quatro testes.

Para a seleção das variáveis explanatórias utilizou-se o método *backward*. As variáveis que obtiveram efeito significativo foram depressão (X_1) e quando a família começou a perceber a falta de audição (X_2).

A Tabela 4.10 mostra as estimativas dos parâmetros, respectivos erros padrões e valores-p para o modelo final ajustado.

Variável	Parâmetros	Estimativa	Erro Padrão	valor-p
Intercepto 1	α_1	$-2,3920$	$0,9566$	$0,012$
Intercepto 2	α_2	$0,3290$	$0,8874$	$0,711$
Depressão	β_1	$0,5796$	$0,3179$	$0,068$
Família (1 ano)	β_2	$2,2474$	$1,0166$	$0,027$
Família (2 anos)	β_3	$1,3361$	$0,9515$	$0,160$
Família (> 2 anos)	β_4	$1,8146$	$0,9201$	$0,049$

Tabela 4.10: *Estimativas dos parâmetros do modelo selecionado*

O modelo final ajustado pode ser expresso em termos dos logitos, por:

$$log\left[\frac{\hat{\pi}_j(\boldsymbol{x})}{\hat{\pi}_{j+1}(\boldsymbol{x})}\right] = \hat{\alpha_j} + 0,5796X_1 + 2,2472X_2 + 1,3361X_3 + 1,8146X_4,\ j = 1,2,$$

em que $X_1 = 0$, se o indivíduo não tem depressão e $X_1 = 1$, se o indivíduo tem depressão; $X_2 = 1$, se a família começou a perceber o problema de audição em 1 ano e $X_2 = 0$, caso contrário; $X_3 = 1$, se a família começou a perceber o problema de audição em 2 anos e $X_3 = 0$, caso contrário; e $X_4 = 1$, se a família começou perceber o problema de audição depois de 2 anos e $X_4 = 0$, caso contrário.

Assim como no modelo anterior, utilizou-se a estatística de Pearson (X^2) e razão de verossimilhanças (G^2) para verificar a qualidade do ajuste do modelo final. O resultado do ajuste pode ser visto Tabela 4.11, indicando evidências a favor do modelo.

Teste	Valor da estatística	Graus de liberdade	valor p
Pearson (X^2)	$9,10$	10	$0,523$
Razão de verossimilhanças (G^2)	$8,59$	10	$0,571$

Tabela 4.11: *Testes de qualidade do ajuste - Em geral você diria que sua audição é - situação 2 - modelo final*

A Figura 4.2 mostra o gráfico dos resíduos de Pearson em função das combinações das categorias da variável depressão e da variável quando a família começou a perceber a falta de audição com as categorias da variável resposta. Como os valores situam-se em torno de zero, há evidências a favor do modelo.

Quanto à interpretação, tem-se que:

- fixada a variável quando a família começou a perceber o problema de audição, a chance de um idoso com depressão dizer que a sua audição é ruim em relação à regular é $\exp(\hat{\beta}_1) = 1,79$ vezes a chance do idoso sem depressão dizer que a sua audição é ruim em relação à regular;

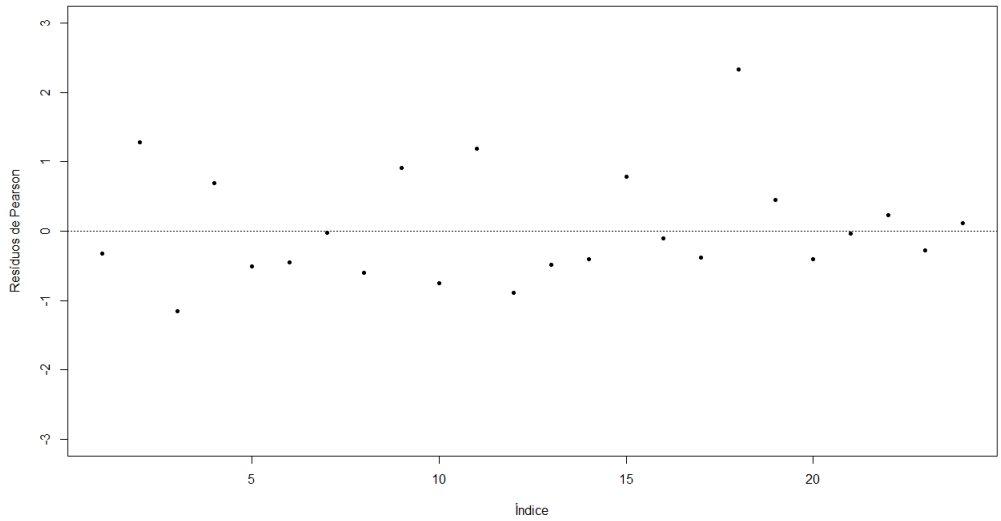

Figura 4.2: *Resíduos do Modelo logito categorias adjacentes - Em geral você diria que sua audição é - situação 2*

- fixada a variável depressão, a chance de um idoso cuja família começou a perceber o problema de audição em 1 ano dizer que a audição é ruim em relação à regular é $\exp(\hat{\beta}_2) = 9,46$ vezes a chance de um idoso cuja família começou a perceber a falta de audição em 6 meses dizer que a audição é ruim em relação à regular;

- fixada a variável depressão, não há evidências de que a chance de um idoso dizer que a audição é ruim em relação à regular seja diferente de quando a família começou a perceber o problema de audição em 6 meses ou em 2 anos;

- fixada a variável depressão, a chance de um idoso cuja família demorou mais de 2 anos para perceber a falta de audição dizer que a audição é ruim em relação à regular é $\exp(\hat{\beta}_4) = 6,14$ vezes a chance de um idoso cuja família começou a perceber a falta de audição em 6 meses dizer que a audição é ruim em relação à regular.

Devido à suposição de chances proporcionais assumida para o modelo ajustado aos dados, as mesmas conclusões são obtidas quanto da chance de um idoso dizer que a sua audição é regular em relação à boa.

Também é possível verificar a chance do idoso dizer que a audição é ruim em relação à boa. Neste caso o modelo fica expresso por:

$$\begin{aligned} log\left[\frac{\hat{P}(Y_i = j|\boldsymbol{x_i})}{\hat{P}(Y_i = 3|\boldsymbol{x_i})}\right] &= \sum_{j=1}^{2} \hat{\alpha_j} + 0,5796(3-j)X_1 + 2,2472(3-j)X_2 + 1,3361(3-j)X_3 \\ &+ 1,8146(3-j)X_4,\ j = 1, 2, \end{aligned}$$

em que $X_1 = 0$, se o indivíduo não tem depressão e $X_1 = 1$, se o indivíduo tem depressão; $X_2 = 1$,

se a família começou a perceber o problema de audição em 1 ano e $X_2 = 0$, caso contrário; $X_3 = 1$, se a família começou a perceber o problema de audição em 2 anos e $X_3 = 0$, caso contrário; e $X_4 = 1$, se a família começou perceber o problema de audição depois de 2 anos; $X_4 = 0$, caso contrário.

Quanto à interpretação, tem-se que:.

- fixada a variável quando a família começou a perceber o problema de audição, a chance de um idoso com depressão dizer que a sua audição é ruim em relação à boa é $\exp(2.\hat{\beta}_1) = 3,19$ vezes a chance de um idoso que não tem depressão dizer que a sua audição é ruim em relação à boa;

- fixada a variável depressão, a chance de um idoso cuja família começou a perceber o problema de audição em 1 ano dizer que a audição é ruim em relação à boa é $\exp(2.\hat{\beta}_2) = 89,55$ vezes a chance de um idoso cuja família começou a perceber a falta de audição em 6 meses dizer que a audição é ruim em relação à boa;

- fixada a variável depressão, não há evidências de que a chance de um idoso dizer que a audição é ruim em relação à boa seja diferente de quando a família começou a perceber o problema de audição em 6 meses ou em 2 anos;

- fixada a variável depressão, a chance de um idoso cuja família demorou mais de 2 anos para perceber a falta de audição dizer que a audição é ruim em relação à boa é $\exp(2.\hat{\beta}_4) = 37,68$ vezes a chance de um idoso cuja família começou a perceber a falta de audição em 6 meses dizer que a audição é ruim em relação à boa.

Utilizou-se o Critério de Informação Akaike (AIC) para comparar o modelo considerando as variáveis explanatórias com mais de duas categorias como quantitativas com o modelo que considera as variáveis explanatórias como qualitativas. A partir da Tabela 4.12 e da mesma forma que nos modelos anteriores, nota-se que o menor AIC foi para o modelo mais simples, o qual considera as variáveis explanatórias com mais de duas categorias como quantitativas. Entretanto, o modelo o qual considera as variáveis com mais de duas categorias como quantitativas apresentou apenas uma variável significativa. Já, o modelo que considera as variáveis com mais de duas categorias como qualitativas apresentou duas variáveis explanatórias significativas.

Por último, ajustou-se o Modelo logito cumulativo com chances proporcionais para as mesmas variáveis explanatórias e resposta e obteve-se as mesmas variáveis explanatórias significativas, tanto

Modelo	AIC
Modelo (qualitativa)	253,04
Modelo (quantitativa)	250,75

Tabela 4.12: *AIC dos modelos ajustados - Em geral você diria que a sua audição é*

para o modelo que considera as variáveis explanatórias com mais de duas categorias como quantitativa quanto para o modelo que considera como qualitativa. Além disso, os efeitos concordaram em ambos modelos.

Capítulo 5

Conclusões e sugestões para pesquisas futuras

O objetivo central deste trabalho foi apresentar uma revisão das técnicas de diagnóstico dos principais modelos ordinais, com ênfase nos Modelos logito cumulativo, Modelos logito categorias adjacentes e Modelos logito razão contínua, úteis para descrever a relação entre a variável resposta ordinal e uma ou mais variáveis explanatórias.

Existem algumas opções para verificar a qualidade do ajuste para estes modelos quando as variáveis explanatórias são categóricas e não há valores esparsos, tais como, teste de Pearson, teste da razão de verossimilhanças e os resíduos de Pearson. Quando variáveis preditoras contínuas estão presentes no modelo ou quando os valores são esparsos, a estatística X^2 de Pearson e a estatística G^2 da razão de verossimilhanças não são adequadas para verificar a qualidade do ajuste. Para esses casos, utiliza-se o teste de Lipsitz, os testes qui-quadrado e razão de verossimilhanças de Pulkstenis-Robinson e a versão ordinal do teste de Hosmer-Lemeshow.

Atualmentes, os Modelos logito cumulativo, Modelos logito categorias adjacentes e Modelos logito razão contínua podem ser ajustados, visto que estes se encontram implementados nos principais pacotes computacionais. Entretanto, o teste de Lipsitz, os testes qui-quadrado e razão de verossimilhanças de Pulkstenis-Robinson e a versão ordinal do teste Hosmer-Lemeshow no *software* R estão disponíveis apenas para o Modelo logito cumulativo com chances proporcionais (pacote "generalhoslem"). Em virtude destes testes não estarem implementados para o Modelo logito categorias adjacentes com chances proporcionais, estes foram desenvolvidos nessa dissertação. Recentemente, Fagerland e Hosmer (2017) apresentaram o comando *ologitgof* implementado no Stata o qual calcula os testes mencionados para avaliar a adequação dos modelos supracitados.

Para a aplicação apresentada neste trabalho selecionaram-se o Modelo logito cumulativo com chances proporcionais e o Modelo logito categorias adjacentes com chances proporcionais para investigar a relação entre a perda auditiva, o equilíbrio e os aspectos emocionais no idoso. Os modelos com chances proporcionais apresentaram melhor ajuste do que os modelos sem chances proporcionais.

Quando ajustados os Modelos logito cumulativo e categorias adjacentes, ambos com chances proporcionais, e com as mesmas variáveis explanatórias e resposta, foram selecionadas as mesmas variáveis explanatórias nos dois modelos.

Ainda há muito o que ser desenvolvido em pesquisas futuras para as técnicas de diagnósticos dos modelos ordinais, tanto na parte teórica quanto em desenvolvimento de *software*. Por exemplo, ainda não há métodos disponíveis para acessar a qualidade do ajuste para os modelos sem chances proporcionais ou com chances proporcionais parciais e que consideram variáveis explanatórias contínuas. Uma opção seria categorizar a variável explanatória contínua ou utilizar o modelo de regressão logística multinomial, os quais dispõem de técnicas de diagnósticos. Entretanto muita informação pode ser perdida com a categorização da variável ou com a utilização do modelo de regressão logística multinomial, visto que este modelo ignora a ordem das categorias da variável resposta.

Apêndice A

Análise descritiva

	Prova Time Up and Go			
Sexo	até 10 segundos	entre 11 e 20 segundos	> 20 segundos	**Total**
Masculino	23 (26%)	61 (68%)	5 (6%)	89 (100%)
Feminino	7 (14%)	36 (74%)	6 (12%)	49 (100%)
Total	30 (22%)	97 (70%)	11 (8%)	138 (100%)

Tabela A.1: *Distribuição conjunta das variáveis Prova Time Up and Go e Sexo*

	Prova Time Up and Go			
Escolaridade	até 10 segundos	entre 11 e 20 segundos	> 20 segundos	**Total**
Analfabeto	3 (25%)	7 (58%)	2 (17%)	12 (100%)
Ensino Fundamental	14 (15%)	72 (76%)	9 (9%)	95 (100%)
Ensino Médio	9 (33%)	18 (67%)	0 (0%)	27 (100%)
Graduação	4 (100%)	0 (0%)	0 (0%)	4 (100%)
Total	30 (22%)	97 (70%)	11 (8%)	138 (100%)

Tabela A.2: *Distribuição conjunta das variáveis Prova Time Up and Go e Escolaridade*

	Prova Time Up and Go			
Renda Mensal	até 10 segundos	entre 11 e 20 segundos	> 20 segundos	**Total**
< 1 SM	0 (0%)	10 (100%)	0 (0%)	10 (100%)
1 Salário Mínimo	16 (18%)	65 (71%)	10 (11%)	91 (100%)
> 1 SM a 2 SM	6 (25%)	17 (71%)	1 (4%)	24 (100%)
> 2 SM	8 (62%)	5 (38%)	0 (0%)	13 (100%)
Total	30 (22%)	97 (70%)	11 (8%)	138 (100%)

Tabela A.3: *Distribuição conjunta das variáveis Prova Time Up and Go e Renda mensal*

	Prova Time Up and Go			
Depressão	até 10 segundos	entre 11 e 20 segundos	> 20 segundos	**Total**
Não	11 (22%)	38 (78%)	0 (0%)	49 (100%)
Sim	19 (21%)	59 (66%)	11 (13%)	89 (100%)
Total	30 (22%)	97 (70%)	11 (8%)	138 (100%)

Tabela A.4: *Distribuição conjunta das variáveis Prova Time Up and Go e Depressão*

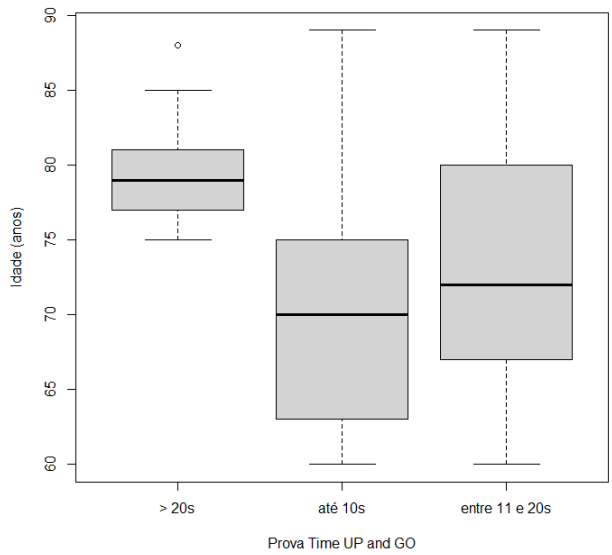

Figura A.1: *Box plot da variável Idade por categoria da variável Prova Time Up and Go*

	Prova Time Up and Go			
Quando a família começou a perceber	até 10s	entre 11 e 20s	> 20s	**Total**
6 meses	2 (50%)	2 (50%)	0 (0%)	4 (100%)
1 ano	4 (23%)	11 (65%)	2 (12%)	17 (100%)
2 anos	2 (8%)	23 (88%)	1 (4%)	26 (100%)
mais de 2 anos	22 (24%)	61 (67%)	8 (9%)	91 (100%)
Total	30 (22%)	97 (70%)	11 (8%)	138 (100%)

Tabela A.5: *Distribuição conjunta das variáveis Prova Time Up and Go e Quando a família começou a perceber*

	Em geral você diria que sua audição é			
Sexo	Ruim	Regular	Boa	**Total**
Masculino	37 (41%)	46 (52%)	6 (7%)	89 (100%)
Feminino	20 (41%)	27 (55%)	2 (4%)	49 (100%)
Total	57 (41%)	73 (53%)	8 (6%)	138 (100%)

Tabela A.6: *Distribuição conjunta das variáveis Em geral você diria que a sua audição é e Sexo*

	Em geral você diria que sua audição é			
Escolaridade	Ruim	Regular	Boa	**Total**
Analfabeto	6 (50%)	6 (50%)	0 (0%)	12 (100%)
Ensino Fundamental	40 (42%)	49 (52%)	6 (6%)	95 (100%)
Ensino Médio	9 (33%)	16 (59%)	2 (8%)	27 (100%)
Graduação	2 (50%)	2 (50%)	0 (0%)	4 (100%)
Total	57 (41%)	73 (53%)	8 (6%)	138 (100%)

Tabela A.7: *Distribuição conjunta das variáveis Em geral você diria que a sua audição é e Escolaridade*

	Em geral você diria que sua audição é			
Renda Mensal	Ruim	Regular	Boa	**Total**
Sem Renda	6(60%)	4 (40%)	0 (0%)	10 (100%)
1 Salário Mínimo	35 (39%)	53 (58%)	3 (3%)	91 (100%)
> 1 SM a 2 SM	10 (42%)	11 (46%)	3 (12%)	24 (100%)
> 2 SM	6 (46%)	5 (39%)	2 (15%)	13 (100%)
Total	57 (41%)	73 (53%)	8 (6%)	138 (100%)

Tabela A.8: *Distribuição conjunta das variáveis Em geral você diria que a sua audição é e Renda mensal*

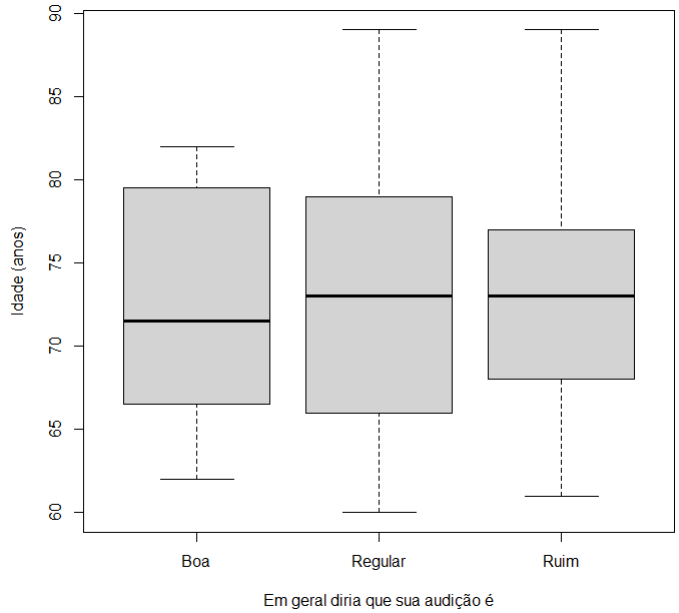

Figura A.2: *Box plot da variável Idade por categoria da variável Em geral você diria que sua audição é*

	Em geral você diria que sua audição é			
Depressão	Ruim	Regular	Boa	**Total**
Não	14 (29%)	32 (65%)	3 (6%)	49 (100%)
Sim	43 (48%)	41 (46%)	5 (6%)	89 (100%)
Total	57 (41%)	73 (53%)	8 (6%)	138 (100%)

Tabela A.9: *Distribuição conjunta das variáveis Em geral você diria que a sua audição é e Depressão*

	Em geral você diria que sua audição é			
Quando a família começou a perceber	Ruim	Regular	Boa	**Total**
6 meses	0 (0%)	3 (75%)	1 (25%)	4 (100%)
1 ano	10 (59%)	6 (35%)	1 (6%)	17 (100%)
2 anos	9 (35%)	14 (54%)	3 (11%)	26 (100%)
mais de 2 anos	38 (42%)	50 (55%)	3 (3%)	91 (100%)
Total	57 (41%)	73 (53%)	8 (6%)	138 (100%)

Tabela A.10: *Distribuição conjunta das variáveis Em geral você diria que a sua audição é e Quando a família começou a perceber*

Prova de Unterbeg com olhos abertos	Em geral você diria que sua audição é			**Total**
	Ruim	Regular	Boa	
Negativo	51 (40%)	71 (55%)	7 (5%)	129 (100%)
Positivo	6 (67%)	2 (22%)	1 (11%)	9 (100%)
Total	57 (41%)	73 (53%)	8 (6%)	138 (100%)

Tabela A.11: *Distribuição conjunta das variáveis Em geral você diria que a sua audição é e Prova de Unterberg com olhos abertos*

Apêndice B

Códigos usados no *software* R

Neste apêndice são apresentados os códigos em R, utilizados na aplicação.

1. Código usado para o ajuste do Modelo logito cumulativo com chances proporcionais.

```
require(MASS)
mlc <- polr(PROVA~DEPRESSAO+FAMILIA+IDADE+SEXO+ESCOLARIDADE
+RENDA.MENSAL, dados2)
```

2. Códigos usados para verificar a qualidade do ajuste do Modelo logito cumulativo com chances proporcionais.

```
require(generalhoslem)
lipsitz.test(mlc, g=6)

logitgof(dados2$PROVA, fitted(mlc), g=6, ord = TRUE)

pulkrob.chisq(mlc, c("DEPRESSAO","FAMILIA","SEXO",
"ESCOLARIDADE","RENDA.MENSAL"))

pulkrob.deviance(mlc, c("DEPRESSAO","FAMILIA","SEXO",
"ESCOLARIDADE","RENDA.MENSAL"))
```

3. Código usado para testar a suposição de chances proporcionais.

```
require(VGAM)
mlc2<-vglm(PROVA~IDADE+RENDA.MENSAL,
                 cumulative(parallel=TRUE,reverse=FALSE), dados2)
mlc3<-vglm(PROVA~IDADE+RENDA.MENSAL,
           cumulative(parallel=FALSE,reverse=FALSE), dados2)
lrtest(mlc2,mlc3)
```

4. Código usado para o ajuste do Modelo logito categorias adjacentes com chances proporcionais.

```
require(VGAM)
catadj<-vglm(OPINIAO~DEPRESSAO+UNTERBERG+FAMILIA+IDADE+SEXO+
ESCOLARIDADE+RENDA.MENSAL,acat(parallel=TRUE,reverse=TRUE), dados2)
```

5. Código usado para o teste de Lipsitz do Modelo logito categorias adjacentes com chances proporcionais.

```
fitted(catadj)
est <- fitted(catadj)%*%c(1,2,3)
matrix <- cbind(est,dados2$ID)
Grupo1 <-c(rep(1,23),rep(0,115))
Grupo2 <-c(rep(0,23),rep(1,23),rep(0,92))
Grupo3 <-c(rep(0,46),rep(1,23),rep(0,69))
Grupo4 <-c(rep(0,69),rep(1,23),rep(0,46))
Grupo5 <-c(rep(0,92),rep(1,23),rep(0,23))
Grupo6 <-c(rep(0,115),rep(1,23))
teste <-data.frame(est, dados2$ID)
teste1 <-teste[order(teste$est),]
teste2 <-data.frame(teste1,Grupo1,Grupo2,Grupo3,Grupo4,Grupo5,Grupo6)
teste3 <-teste2[order(teste2$dados2.ID),]
EST<-fitted(catadj)
colnames(EST)<-c("EST1","EST2","EST3")
teste4 <- data.frame(teste3, OPINIAO=dados2$OPINIAO,EST)

catadj<-vglm(OPINIAO~DEPRESSAO+UNTERBERG+FAMILIA+IDADE+SEXO+
ESCOLARIDADE+RENDA.MENSAL,acat(parallel=TRUE,reverse=TRUE), dados2)

catadjLP<-vglm(OPINIAO~DEPRESSAO+UNTERBERG+FAMILIA+IDADE+SEXO+
ESCOLARIDADE+RENDA.MENSAL+teste3$Grupo1+teste3$Grupo2+teste3$Grupo3
+teste3$Grupo4+teste3$Grupo5,acat(parallel=TRUE,reverse=TRUE), dados2)

lrtest(catadjLP,catadj)
```

6. Código usado para a versão ordinal do teste de Hosmer-Lemeshow do Modelo logito categorias adjacentes com chances proporcionais.

```
N1<-subset(teste4,OPINIAO=="1")
NN1<-c(dim(subset(N1,Grupo1=="1"))[1],
       dim(subset(N1,Grupo2=="1"))[1],
       dim(subset(N1,Grupo3=="1"))[1],
       dim(subset(N1,Grupo4=="1"))[1],
       dim(subset(N1,Grupo5=="1"))[1],
       dim(subset(N1,Grupo6=="1"))[1])
N2<-subset(teste4,OPINIAO=="2")
NN2<-c(dim(subset(N2,Grupo1=="1"))[1],
       dim(subset(N2,Grupo2=="1"))[1],
       dim(subset(N2,Grupo3=="1"))[1],
       dim(subset(N2,Grupo4=="1"))[1],
       dim(subset(N2,Grupo5=="1"))[1],
```

```
       dim(subset(N2,Grupo6=="1"))[1])
N3<-subset(teste4,OPINIAO=="3")
NN3<-c(dim(subset(N3,Grupo1=="1"))[1],
       dim(subset(N3,Grupo2=="1"))[1],
       dim(subset(N3,Grupo3=="1"))[1],
       dim(subset(N3,Grupo4=="1"))[1],
       dim(subset(N3,Grupo5=="1"))[1],
       dim(subset(N3,Grupo6=="1"))[1])
OO<-cbind(NN1,NN2,NN3)
EE1<-c(sum(subset(teste4,Grupo1=="1")$EST1),
       sum(subset(teste4,Grupo2=="1")$EST1),
       sum(subset(teste4,Grupo3=="1")$EST1),
       sum(subset(teste4,Grupo4=="1")$EST1),
       sum(subset(teste4,Grupo5=="1")$EST1),
       sum(subset(teste4,Grupo6=="1")$EST1))
EE2<-c(sum(subset(teste4,Grupo1=="1")$EST2),
       sum(subset(teste4,Grupo2=="1")$EST2),
       sum(subset(teste4,Grupo3=="1")$EST2),
       sum(subset(teste4,Grupo4=="1")$EST2),
       sum(subset(teste4,Grupo5=="1")$EST2),
       sum(subset(teste4,Grupo6=="1")$EST2))
EE3<-c(sum(subset(teste4,Grupo1=="1")$EST3),
       sum(subset(teste4,Grupo2=="1")$EST3),
       sum(subset(teste4,Grupo3=="1")$EST3),
       sum(subset(teste4,Grupo4=="1")$EST3),
       sum(subset(teste4,Grupo5=="1")$EST3),
       sum(subset(teste4,Grupo6=="1")$EST3))
EE<-cbind(EE1,EE2,EE3)
sum(((OO-EE)^2)/EE)
1-pchisq(sum(((OO-EE)^2)/EE),9)
```

Referências Bibliográficas

Agresti(2003) Alan Agresti. *Categorical data analysis*. John Wiley & Sons. Citado na pág. 8

Agresti(2010) Alan Agresti. *Analysis of ordinal categorical data*. John Wiley & Sons. Citado na pág. 2, 8, 9, 11, 13, 14, 19, 26, 27

Agresti(1984) Alan Agresti. *Analysis of Categorical Data*. Wiley New York. Citado na pág. 2

Agresti(1996) Alan Agresti. *An introduction to categorical data analysis*. Wiley New York. Citado na pág. 8, 19

Akaike(1974) Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19:716–723. Citado na pág. 27

Brant(1990) Rollin Brant. Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, 46:1171–1178. Citado na pág. 26

Corporation(2017) IBM Corporation. *IBM SPSS Statistic 25 Core System User´s Guide*, 2017. URL https://www-01.ibm.com/support/docview.wss?uid=swg24043678. Citado na pág. 10

Cullagh(1980) MC Cullagh. Regression models for ordinal data (with discussion). *J. Roy. Statist. Soc., B*, 42:109–142. Citado na pág. 2, 9, 10

Fagerland e Hosmer(2016) Morten W Fagerland e David W Hosmer. Tests for goodness of fit in ordinal logistic regression models. *Journal of Statistical Computation and Simulation*, 86: 3398–3418. Citado na pág. 2, 24, 25

Fagerland e Hosmer(2017) Morten W. Fagerland e David W. Hosmer. How to test for goodness of fit in ordinal logistic regression models. *The Stata Journal*, 17:668–686. Citado na pág. 41

Fagerland e Hosmer(2013) Morten W Fagerland e David W Hosmer. A goodness-of-fit test for the proportional odds regression model. *Statistics in medicine*, 32:2235–2249. Citado na pág. 2, 23, 24, 25

Feinberg(1980) Stephen E Feinberg. The analysis of cross-classified categorical data. *Massachusetts Institute of Technology Press, Cambridge and London*. Citado na pág. 2

Freeman Jr(1987) Daniel H Freeman Jr. *Applied categorical data analysis*. Marcel Dekker, Inc. Citado na pág. 19, 21

Giolo(2017) Suely Ruiz Giolo. *Introdução à análise de dados categóricos com aplicações*. Edgard Blucher. Citado na pág. 19

Hosmer e Lemeshow(1980) David W Hosmer e Stanley Lemeshow. Goodness of fit tests for the multiple logistic regression model. *Communications in statistics-Theory and Methods*, 9:1043–1069. Citado na pág. 20, 21

Kim(2003) Ji-Hyun Kim. Assessing practical significance of the proportional odds assumption. *Statistics & probability letters*, 65:233–239. Citado na pág. 26

Lipsitz *et al.*(1996) Stuart R Lipsitz, Garrett M Fitzmaurice e Geert Molenberghs. Goodness-of-fit tests for ordinal response regression models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 45:175–190. Citado na pág. 2, 20, 21

Magrini(2015) A M Magrini. *Investigar a relação entre a perda auditiva, os aspectos emocionais e o equilíbrio no idoso*. Tese de doutorado, Pontifíca Universidade Católica de São Paulo. Citado na pág. 29

Nelder e Wedderburn(1972) J A Nelder e W Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society Series A*, 135:370–384. Citado na pág. 2, 5

Paulino e Singer(2006) Carlos Daniel Mimoso Paulino e Julio da Motta Singer. *Análise de dados categorizados*. Edgard Blücher. Citado na pág. 1

Peterson e Harrell Jr(1990) Bercedis Peterson e Frank E Harrell Jr. Partial proportional odds models for ordinal response variables. *Applied statistics*, 39:205–217. Citado na pág. 10, 26

Pulkstenis e Robinson(2004) Erik Pulkstenis e Timothy J Robinson. Goodness-of-fit tests for ordinal response regression models. *Statistics in medicine*, 23:999–1014. Citado na pág. 2, 22

R Core Team(2017) R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL http://www.R-project.org. Citado na pág. 10

SAS(2013) SAS. *Base SAS 9.4 procedures guide: Statistical procedures*, 2013. URL https://support.sas.com/documentation/cdl/en/procstat/66703/HTML/default/viewer.htm#titlepage.htm. Citado na pág. 10

Stata *et al.*(2017) A Stata *et al.* *Stata base reference manual release 15*. Stata Press, Texas, 2017. URL https://www.stata.com/bookstore/base-reference-manual/. Citado na pág. 22

Tutz(1991) Gerhard Tutz. Sequential models in categorical regression. *Computational Statistics & Data Analysis*, 11:275–295. Citado na pág. 14

Walker e Duncan(1967) Strother H Walker e David B Duncan. Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54:167–179. Citado na pág. 2, 10