Diagnóstico no modelo de regressão logística ordinal

Marina Calais de Freitas Moura [1], **Mônica Carneiro Sandoval** [2], **Denise Aparecida Botter** [3]

1 Introdução

Os modelos de regressão logística ordinais são usados para descrever a relação entre uma variável resposta categórica ordinal e uma ou mais variáveis explanatórias. Existe uma variedade de modelos para ajustar variáveis ordinais. Dentre eles, os mais utilizados são os Modelos logito cumulativo, Modelos logito categorias adjacentes e Modelos logito razão contínua, com chances proporcionais (Agresti, 2010).

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 explanatórias 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 da razão de verossimilhanças de Pulkistenis-Robinson.

Neste estudo é feita uma revisão das técnicas de diagnóstico disponíveis para os Modelos logito cumulativo, Modelos logito categorias adjacentes e Modelos logito razão contínua, com chances proporcionais, bem como uma aplicação a fim de investigar a relação entre a perda auditiva, o equilíbrio e aspectos emocionais em idosos.

2 Modelo de regressão logística ordinal

O Modelo de regressão logística ordinal é aplicado quando o número de categorias da variável resposta excede dois e quando estas são ordenadas. Nesta seção serão apresentados os modelos de regressão logística ordinais mais utilizados.

2.1 Modelo logito cumulativo com chances proporcionais

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

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

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 as categorias ordenadas 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 $\mathbf{x_i}$.

2.2 Modelo logito categorias adjacentes com chances proporcionais

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

[1]Instituto de Matemática e Estatística, Universidade de São Paulo. e-mail: *marinakalais@hotmail.com*
[2]Instituto de Matemática e Estatística, Universidade de São Paulo. e-mail: *sandoval@ime.usp.br*
[3]Instituto de Matemática e Estatística, Universidade de São Paulo. e-mail: *botter@ime.usp.br*

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

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

2.3 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:

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

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

3 Técnicas de diagnóstico

3.1 Testes de qualidade do ajuste

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

3.1.1 Teste de Pearson e teste da razão de verossimilhanças

Quando se tem apenas variáveis explanatórias categóricas no modelo, é possível formar uma tabela de contingência por meio das categorias da variável resposta ($j = 1, ..., c$) e das combinações das categorias das variáveis explanatórias ($l = 1, ..., k$).

A estatística X^2 de Pearson para testar a qualidade do ajuste do modelo e a estatística G^2 da razão de verossimilhanças são expressas por:

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

em que $E_{lj} = \sum_l \hat{\pi}_{lj} = n_l\hat{\pi}_{lj}$, sendo $\hat{\pi}_{lj}$ a estimativa da probabilidade da variável resposta assumir a categoria j, para um dado vetor $\boldsymbol{z_l}$ de valores das variáveis explanatórias que representa a l-ésima combinação, $l = 1, ..., k$, e n_{lj} é o número de indivíduos em cada célula da tabela de contingência.

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

Nas próximas subseções serão apresentados três 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.

3.1.2 Teste de Lipsitz

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, atribui-se um escore (s_i) para cada observação, usando pesos igualmente espaçados, $s_i = \hat{\pi}_{i1} + 2\hat{\pi}_{i2} + ... + c\hat{\pi}_{ic},\ i = 1, ..., n$.

Depois, 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 logística 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.

O valor-p do teste associado à estatística da razão de verossimilhanças para testar $H_0 : \gamma_1 = ... = \gamma_{g-1} = 0$ é obtido aproximando-se a estatística $-2(L_0 - L_1)$ pela distribuição χ^2 com $g - 1$ graus de liberdade.

3.1.3 Teste de Pulkistenis e 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-se somente as variáveis categóricas do modelo, em que as categorias não observadas são desconsideradas. Em seguida, calculam-se os escores da mesma maneira que no teste de Lipsitz (s_i) e então cada combinação 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:

$$\chi^2 = \sum_{l=1}^{k}\sum_{t=1}^{2}\sum_{j=1}^{c} \frac{(n_{ltj} - E_{ltj})^2}{E_{ltj}}, \qquad G^2 = \sum_{l=1}^{k}\sum_{t=1}^{2}\sum_{j=1}^{c} n_{ltj} log\frac{n_{ltj}}{E_{ltj}}.$$

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.

As duas estatísticas seguem uma distribuiçõ 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.

3.1.4 Versão ordinal do teste de Hosmer e 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.

A seguir, ordenam-se as observações com base nos escores obtidos e formam-se g grupos, da mesma maneira que no teste de Lipsitz (Lipsitz et al, 1996). A versão ordinal da estatística do teste de Hosmer-Lemeshow é dada por:

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

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.

A distribuição de C_g é aproximadamente χ^2 com $(g-2)(c-1)+(c-2)$ graus de liberdade.

3.2 Análise dos 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. Valores grandes, como excedendo 3 em valor absoluto, indicam falta de ajuste na célula.

4 Aplicação

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, 2017). Os dados referem-se a 138 indíviduos com idade superior a 60 anos e com perda auditiva.

Para este estudo, foram consideradas como variáveis explanatórias as variáveis 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), idade (em anos), depressão (0=não, 1=sim) e quando a família começou a perceber a falta de audição (6 meses (código 1), 1 ano (código 2), 2 anos (código 3) e mais de dois anos (código 4)). A variável resposta

considerada foi a prova Time UP and GO, 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 (11 segundos (código 1), entre 11 e 20 segundos (código 2), 20 segundos (código 3)).

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.

A Tabela 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 considerando as variáveis explanatórias com mais de duas categorias como quantitativas (I) e qualitativas (II), respectivamente. Os resultados da Tabela 1 evidenciam o bom ajuste do modelo completo.

Tabela 1: Testes de qualidade do ajuste

Var explanatória	**Teste**	**Estatística**	**Graus de liberdade**	**valor-p**
(I)	Lipsitz	$6,96$	5	$0,224$
	C_6	$10,94$	9	$0,280$
	X^2	$10,79$	11	$0,461$
	G^2	$13,89$	11	$0,239$
(II)	Lipsitz	$3,45$	5	$0,630$
	C_6	$4,79$	9	$0,852$
	X^2	$124,79$	152	$0,948$
	G^2	$112,43$	152	$0,993$

Fonte: Elaborada pelas autoras

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 nas duas situações. A análise de resíduos não foi feita, em virtude de haver variável explanatória contínua no modelo final.

Finalmente, utilizou-se o Critério de Informação Akaike (AIC) para comparar o modelo considerando as variáveis explanatórias com mais de duas categorias como quantitativas ($AIC = 199,44$) com o modelo que considera as variáveis explanatórias como qualitativas ($AIC = 204,03$). Notou-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 penalizando 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 significativas, tanto para o modelo que considera as variáveis explanatórias com mais de duas categorias como quantitativas quanto para o modelo que considera como qualitativas.

5 Conclusão

Os Modelos logito cumulativo, Modelos logito categorias adjacentes e Modelos logito razão contínua podem ser facilmente ajustados, visto que estes se encontram implementados nos principais pacotes computacionais.

O teste de Lipsitz, os testes qui-quadrado e razão de verossimilhanças de Pulkstenis-Robinson e a versão ordinal do teste Hosmer-Lemeshow estão disponíveis no software R apenas para o Modelo logito cumulativo com chances proporcionais (pacote "generalhoslem"). Entretanto, Fagerland e Hosmer (2017) apresentaram o comando *ologitgof* implementado no Stata o qual calcula os testes mencionados para avaliar a adequação dos modelos supracitados.

Os modelos com chances proporcionais apresentaram melhor ajuste do que os modelos sem chances proporcionais. Na aplicação da Seção 4, 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 modelos.

Referências Bibliográficas

AGRESTI, A. Analysis of ordinal categorical data *John Wiley & Sons*. v.656, 2010.
FAGERLAND, M. W.; HOSMER, D. W. A goodness-of-fit test for the proportional odds regression model. *Statistics in medicine*, Washington, v.32, n.13, p.2235-2249, 2013.

FAGERLAND, M. W.; HOSMER, D. W. Tests for goodness of fit in ordinal logistic regression models. *Journal of Statistical Computation and Simulation*, v.86, n.17, p.3398-3418, 2016.

FAGERLAND, M. W.; HOSMER, D. W. How to Test for Goodness of Fit in Ordinal Logistic Regression Models. *The Stata Journal*, v.17, n.3, p.668-686, 2017.

LIPSITZ, S. R.; FITZMAURICE, G. M.. MOLENBERGHS, G. Goodness-of-fit tests for ordinal response regression models. *Applied Statistics*, v.45, n.2, p.175-190, 1996.

MAGRINI, A. M.A. Investigar a relação entre a perda auditiva, os aspectos emocionais e o equilíbrio no idoso. *Pontifíca Universidade Católica de São Paulo*. São Paulo. Tese de doutorado. 2015.

PULKSTENIS, E.; ROBINSON, T. J. Goodness-of-fit tests for ordinal response regression models. *Statistics in medicine*, v.23, n.6, p.999-1014, 2004.