O MODELO DE RESPOSTA GRADUAL NA ESTIMAÇÃO DE PROBABILIDADES DE RESULTADOS NO FUTEBOL

Héliton Tavares[1], Hugo Harada [2], Valcir Farias[3], Franciney Palheta[4]

[1] UFPA, heliton@ufpa.br

[2] Adaptativa Educacional, hugo.harada@adaptativa.com

[3] UFPA, valcir@ufpa.br

[4] UFPA, franciney@ufpa.br

Resumo

O Modelo de Resposta Gradual (MRG) é um dos submodelos adotados na área de Teoria da Resposta ao Item (TRI) que tem sido largamente empregado para modelar itens de testes ou questionários cuja respostas tenham várias categorias de resposta ordenadas. Neste trabalho propomos a aplicação do MRG para a estimação das probabilidades de Vitória, Empate e Derrota pelo MRG, com dados do Campeonato Brasileiro Série A de 2006 a 2024. Os parâmetros do modelo foram estimados pelo Método dos Mínimos Quadrados. Foi feita uma proposta do Efeito-Casa, facilmente estimado com essa modelagem.

Palavras-chave: Scout, Função de resposta ao item.

1. Introdução

Modelos de Resposta Gradual (MRG, ou GRM em inglês, ver Ref) compõe uma família de modelos de traços latentes (também denominados de construtos), ou seja, variáveis que não são medidas diretamente, mas sim através de outras variáveis. Por exemplo, o conhecimento cognitivo em determinada área do conhecimento, tal como matemática ou linguagem, bem como aspectos socioemocionais, como o nível de neurose, conscienciosidade ou abertura de uma pessoa, dentre muitas outras possíveis aplicações. O MRG foi desenvolvido por Fumiko Samejima em 1969 como uma extensão dos modelos ***dicotômicos*** (apenas 2 categorias de resposta, muitas vezes denotadas por 0 e 1, naturalmente ordenadas) e tem sido amplamente adotado desde então para classificar respostas de indivíduos que tenham categorias gradualmente ordenadas, compondo um dos modelos politômicos, com várias categorias de resposta, neste caso, ordenadas. O MRG também é conhecido como Modelo de Respostas

Categóricas Ordenadas, pois lida com categorias politômicas ordenadas que podem se relacionar a itens de resposta construída ou resposta selecionada, onde os examinandos devem obter vários níveis de pontuação, como 0-4 pontos. Neste caso, as categorias são as seguintes: 0, 1, 2, 3 e 4; e são ordenadas. Em itens abertos pode-se estabelecer uma grade de correção em que a categoria "A" representa "Completamente certo" e "E" significa "Completamente errado". Em questionário podemos ter "A": "Concordo completamente a "E": "Discordo completamente".

No ramo do esporte há um bom campo para aplicação de métodos estatísticos, principalmente no futebol, cuja área estatística é denominada Scout. De forma geral, temos 3 categorias bem definidas para um time que joga em sua sede: A: "Vitória", "B": Empate, "C": Derrota. O fato de jogar ou não na própria sede (casa), contando com o apoio de sua torcida e conhecimento do estádio, pode agregar uma nova componente no modelo, dentre outras informações, alterando as probabilidades associadas ao MRG original.

Este artigo propõe o uso do MRG para modelar as probabilidades de Derrota, Empate e Vitória de forma bastante intuitiva e simplificada. Na Seção 2 será feita a apresentação formal do modelo, com parametrização e propriedades, seguida pelo processo de estimação dos parâmetros via Mínimos Quadrados. Na Seção 3 será feita uma aplicação a dados do Campeonato Brasileiro Série A, composto por 20 times, que realizam 380 partidas a cada ano, bem como apresentadas as estimativas dos parâmetros. Na Seção 4 são apresentadas as conclusões e feitas as considerações finais.

2. Metodologia

O modelo de resposta gradual de Samejima (1969) assume que as categorias de resposta de um determinado item podem ser ordenadas entre si. Este modelo visa obter mais informação das respostas politômicas dos indivíduos e itens do que simplesmente se foram dadas respostas totalmente corretas ou incorretas, explorando várias gradações.

Embora um instrumento (prova ou questionário) seja composto por vários itens, de forma que há necessidade de inclusão de um índice (normalmente i) para representá-lo, vamos simplificar a notação omitindo esse índice. Suponha então que as categorias de um item são arranjadas em ordem do menor para o maior e denotados por $k = 0,1,\cdots,m,$ de forma $m+1$ será o número de categorias do item. A probabilidade de um indivíduo j escolher (ou ser categorizado) em uma particular categoria ou outra mais alta do item (função cumulativa, por isso o sinal +) pode ser dada por uma extensão do modelo logístico de 2 parâmetros:

$$P(U_j \geq k|\theta_j) = P_k^+(\theta_j) = \frac{1}{1 + e^{-a(\theta_j - b_k)}} \quad (1)$$

Na expressão (1), U_j representa a resposta categórica do indivíduo j, θ_j é o construto sendo medido correspondente ao indivíduo j, e b_k é o parâmetro da categoria k do item. Naturalmente temos que $P_0^+(\theta_j) = 1$ e podemos incluir artificialmente uma categoria $m + 1$ de forma que $P_{m+1}^+(\theta_j) = 0$. Com essa construção cumulativa, temos que $P_k^+(\theta_j) \geq P_{k+1}^+(\theta_j)$, $k = 0,1, \dots, m$, de forma que surge uma restrição natural:

$$b_1 \leq b_2 \leq \cdots \leq b_m.$$

Ou seja, devemos ter necessariamente uma ordenação entre o nível de dificuldade das categorias de um dado item, de acordo com a classificação de seus escores. Ainda, se temos $m + 1$ categorias, teremos apenas m parâmetros b_k, posto que temos uma forte restrição: a soma das probabilidades associadas às categorias deve ser 1. Para o caso dicotômico, com 2 categorias de resposta, basta modelar a probabilidade de acerto ao item, pois a probabilidade de erro sai por diferença.

A probabilidade de um indivíduo j receber um escore k no item é dada então pela expressão:

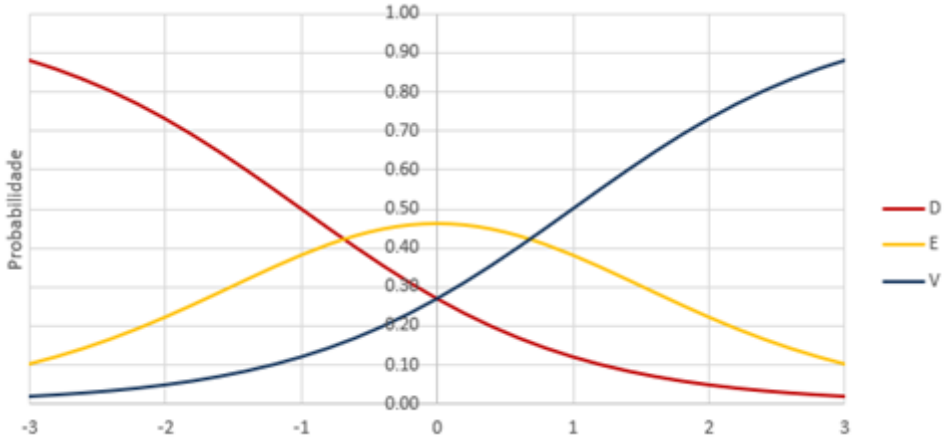
$$P(U_j = k|\theta_j) = P(U_j \geq k|\theta_j) - P(U_j \geq k + 1|\theta_j) = P_k^+(\theta_j) - P_{k+1}^+(\theta_j) \quad (2)$$

Para um item com 3 categorias, adotando a notação simplificada $P_k(\theta_j) = P(U_j = k|\theta_j)$, $k = 0,1,2$, teremos:

$$\begin{aligned} P_0(\theta_j) &= 1 - P_1^+(\theta_j) \\ P_1(\theta_j) &= P_1^+(\theta_j) - P_2^+(\theta_j) \\ P_2(\theta_j) &= P_2^+(\theta_j), \end{aligned} \quad (3)$$

de onde vemos que $P_0(\theta_j) + P_1(\theta_j) + P_1(\theta_j) = 1$, em que P_0 terá parâmetros (a, b_1), P_1 terá (a, b_1, b_2) e P_2 terá (a, b_2). A Figura 1 ilustra um item com 3 categorias.

Figura 1. Item gradual com 3 categorias: 0, 1, 2 ($a = 1, b_1 = -1, b_2 = 1$)



Fonte: Elaborado pelos autores

Neste exemplo, a escala do escore está centrada em zero, no intervalo [-3;3], compatível com uma distribuição Normal com média zero e desvio-padrão 1. No entanto, em aplicações gerais é bastante frequente fazer transformações de escala visando uma melhor interpretação e entendimento por pessoas de áreas diversas.

3. Associação com ranqueamento no futebol

No campeonato brasileiro de futebol da série A, composta por $N = 20$ times, em cada rodada cada time estará em uma colocação (ranking) exclusivo no intervalo discreto de 1 a N, em que o ranking 1 indica a melhor colocação. Definimos:

$$R_i: Ranking\ do\ time\ i\ jogando\ em\ casa$$
$$R_j: Ranking\ do\ Time\ Visitante$$

3.1. Base de dados disponível e resultados preliminares

A Diferença dos Rankings será dada por

$$Dif_{i,j} = R_j - R_i,$$

variando de –(N-1) a (N-1), excluindo-se o valor nulo, pois não há coincidência de colocações em uma mesma rodada. Por exemplo, se o último colocado (R_i =20) for *Mandante* e receber o primeiro colocado ($R_j = 1$) como Visitante, teremos $Dif_{i,j} = -19$.

A Diferença de Ranking Padronizada (que denominaremos de $DifP_{i,j}$) poderá ser construída a partir de todas as $20(20-1) = 380$ possíveis diferenças $Dif_{i,j}$ e obtendo-se o desvio-padrão (S), para ficar compatível com a distribuição normal padrão. Essa construção irá

gerar um desvio-padrão dos $D_{i,j}$ de aproximadamente 8. Com isso, podemos definir a diferença de ranking padronizada por:

$$DifP_{i,j} = Dif_{i,j}/S$$

Na Figura 1, elaborada com os $DifP_{i,j}$, um time com $DifP_{i,j}$ inferior a -0,6 terá maior probabilidade de derrota (D); times com $DifP_{i,j}$ entre -0,6 e 0,6 terão maior probabilidade de empate (E), enquanto times com $DifP_{i,j}$ maior que *0,6* terão maior probabilidade de vitória. Ainda, para um jogo entre times com rankings muito próximos, teremos $DifP_{i,j}$ próximo de zero, e usando o MLG no ponto zero teremos as probabilidades de Vitória e de Derrota próximos de 0,28, e de Empate em 0,44, quando não houver Efeito-Casa. O conjunto de respostas (D, E, V) a cada rodada servirá para estimar os parâmetros (a, b_1, b_2). As estimativas poderão ser diferentes das exemplificadas, indicando, por exemplo, o efeito de jogar em casa.

Embora a padronização seja conveniente para o caso geral, essa operação é opcional, posto que o cálculo de probabilidades ficará mais simples com a diferença direta de rankings. Neste caso, as estimativas dos parâmetros deverão ser devidamente transformadas. Por exemplo, para os valores exemplificados na Figura 1, $(a = 1, b_1 = -1, b_2 = 1)$, teremos

$$a^* = a/S;\ b_1^* = b_1 S;\ b_2^* = b_2 S,$$

de forma que as probabilidades associadas às categorias permanecem as mesmas:

$$a(\theta - b_k) = \frac{a}{S}(\theta S - b_k S) = a^*(\theta^* - b_k^*).$$

Alternativamente, os parâmetros podem ser estimados diretamente com $Dif_{i,j}$, sem necessidade de transformação da escala de [-3;3] para a escala original, de $-(N-1)$ a $(N-1)$.

3.2. Base de dados disponível e resultados preliminares

A base de dados adotada nesta pesquisa é composta pelos resultados do Campeonato Brasileiro da Série A dos anos 2006 a 2024 (este último ainda incompleto, visto que o campeonato ainda não terminou), obtida no site *https://basedosdados.org*. Nos anos anteriores a 2006 o campeonato tinha um outro formato com número diferente de times. Atualmente, com $N = 20$ times disputando o campeonato em partidas de ida e volta, há $N(N-1) = 20 * 19 = 380$ jogos.

Tabela 1. Quantitativos de partidas

Sequência	Campeonato	# Jogos
1	2006	380
2	2007	380
3	2008	380
4	2009	380
5	2010	380
6	2011	380
7	2012	380
8	2013	380
9	2014	380
10	2015	380
11	2016	380
12	2017	380
13	2018	380
14	2019	380
15	2020	380
16	2021	380
17	2022	380
18	2023	380
19	2024	169

Fonte: Elaborada pelos autores

A base de dados tem todas as informações necessárias para estimar as proporções necessárias, obtendo as $Dif_{i,j}$ para cada jogo em cada rodada, bem como as proporções de Derrota (PD), Empate (PE) e Vitória (PV), apresentadas na Tabela 2.

Tabela 2. Proporções de (D,E,V) por Dif

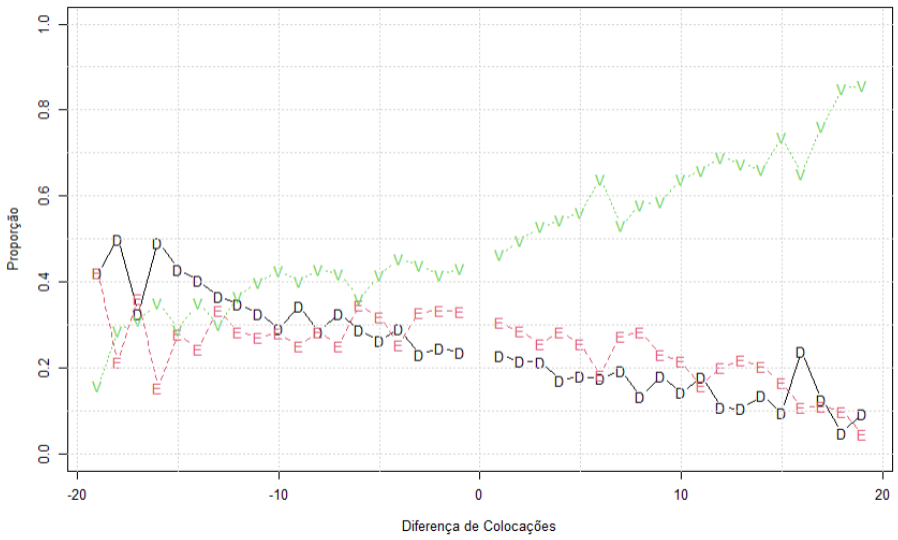
Diff	Freq	PD	PE	PV
-19	19	0.42	0.42	0.16
-18	28	0.50	0.21	0.29
-17	61	0.33	0.36	0.31
-16	71	0.49	0.15	0.35
-15	93	0.43	0.28	0.29
-14	94	0.40	0.24	0.35
-13	123	0.37	0.33	0.30
-12	158	0.35	0.28	0.37
-11	180	0.33	0.27	0.40
-10	178	0.29	0.28	0.43
-9	206	0.34	0.25	0.40
-8	200	0.29	0.29	0.43
-7	238	0.33	0.25	0.42
-6	259	0.29	0.35	0.36
-5	269	0.26	0.32	0.42
-4	275	0.29	0.25	0.45
-3	294	0.23	0.33	0.44
-2	299	0.25	0.33	0.42
-1	366	0.24	0.33	0.43

1	317	0.23	0.31	0.46
2	342	0.22	0.29	0.50
3	280	0.21	0.26	0.53
4	261	0.17	0.28	0.54
5	276	0.18	0.26	0.56
6	245	0.18	0.18	0.64
7	207	0.19	0.28	0.53
8	193	0.13	0.28	0.58
9	177	0.18	0.23	0.59
10	147	0.14	0.22	0.64
11	150	0.18	0.16	0.66
12	139	0.11	0.20	0.69
13	123	0.11	0.22	0.67
14	118	0.14	0.20	0.66
15	72	0.10	0.17	0.74
16	46	0.24	0.11	0.65
17	63	0.13	0.11	0.76
18	40	0.05	0.10	0.85
19	21	0.10	0.05	0.86

Fonte: Elaborada pelos autores

Esses elementos individuais serão representados por $PD[l], PE[l]$ e $PV[l]$, para $l = -(N-1), \ldots, (N-1)$. Graficamente teremos o seguinte comportamento:

Figura 2. Proporções (PD,PE,PV) de acordo com a Dif



Fonte: Elaborado pelos autores

3.3. Processos de estimação dos parâmetros

O processo de estimação dos parâmetros (a, b_1, b_2) pode ser feito pelo Método dos Mínimos Quadrados (Bates & Chambers, 1992), minimizando a função:

$$SQ(a, b_1, b_2) = \sum_{l=-(N-1)}^{N-1} \left\{ \left(PD(l) - P_1(l)\right)^2 + \left(PE(l) - P_2(l)\right)^2 + \left(PV(l) - P_3(l)\right)^2 \right\},$$

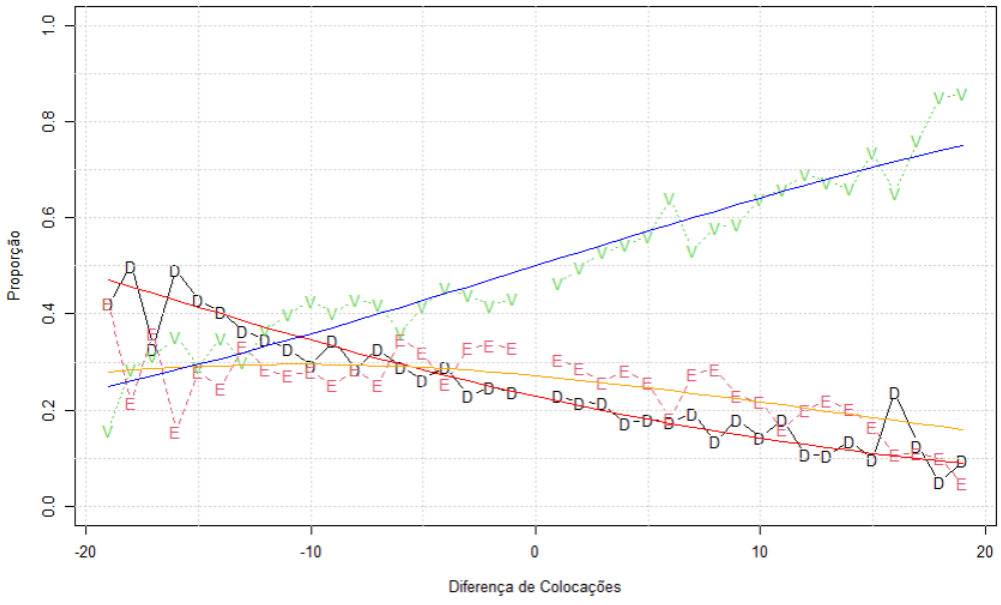
com $P_1(l)$, $P_2(l)$ e $P_3(l)$ dados pela expressão (1) como função de a, b_1 e b_2, e $PD(l), PE(l), PV(l)$ apresentados na Tabela 2. O processo de estimação visa obter os valores de (a, b_1, b_2) que minimizam $SQ(a, b_1, b_2)$, e foi implementado no software R (Ref), restringindo-se a valores inteiros para as estimativas de b_1 e b_2 no intervalo de $-(N-1)$ a $(N-1)$. O processo apresentou as seguintes estimativas:

$$\hat{a} = 0{,}057, \quad \hat{b}_1 = -21.6 \text{ e } \hat{b}_2 = -0{,}3. \tag{4}$$

3.4. Obtenção das probabilidades de Derrota, Empate e Vitória:

Com as estimativas dos parâmetros do MRG obtidas em (4), podemos estimar as probabilidades para as 3 categorias, para cada valor de Dif, através das expressões em (3). Os resultados estão apresentados na Figura 3.

Figura 3: Ajuste do MRG aos dados de Futebol Série A



Fonte: Elaborado pelos autores

4. Resultados e Discussão

Podemos observar na Figura 3 que o modelo se adequou muito bem aos dados, que foram compostos de $38 \times 3 = 114$ pontos distribuídos em 3 curvas. A Soma dos Quadrados dos Resíduos (SQR) obtida com as estimativas em (4) foi 0,2509, um número consideravelmente baixo, ressaltando a boa qualidade do ajuste.

Tabela 3. Probabilidades de (D,E,V) pelo MRG

Dif	D	E	V
-19	0.463	0.281	0.256
-18	0.449	0.284	0.267
-17	0.435	0.287	0.279
-16	0.421	0.289	0.290
-15	0.407	0.291	0.302
-14	0.393	0.293	0.314
-13	0.380	0.294	0.327
-12	0.367	0.294	0.339
-11	0.353	0.295	0.352
-10	0.340	0.294	0.365
-9	0.328	0.294	0.379
-8	0.315	0.293	0.392
-7	0.303	0.291	0.406
-6	0.291	0.289	0.419
-5	0.280	0.287	0.433
-4	0.268	0.284	0.447
-3	0.257	0.281	0.462
-2	0.247	0.278	0.476
-1	0.236	0.274	0.490
0	0.226	0.270	0.504
1	0.216	0.265	0.519
2	0.207	0.261	0.533
3	0.197	0.256	0.547
4	0.189	0.250	0.561
5	0.180	0.245	0.575
6	0.172	0.239	0.589
7	0.164	0.234	0.603
8	0.156	0.228	0.616
9	0.149	0.222	0.630
10	0.142	0.216	0.643
11	0.135	0.209	0.656
12	0.128	0.203	0.668
13	0.122	0.197	0.681
14	0.116	0.191	0.693
15	0.110	0.184	0.705
16	0.105	0.178	0.717
17	0.100	0.172	0.728
18	0.095	0.166	0.739
19	0.090	0.160	0.750

Fonte: Elaborada pelos autores

No ponto central, com $Dif = 0$, considerando uma situação de times equivalentes, temos que a probabilidade de derrota é 0,226, de empate é 0,270 e de Vitória é 0,504. Caso não houvesse algum Efeito-Casa, as probabilidades de derrota e empate deveriam ser iguais, mas não são. Assim, podemos definir o Efeito-Casa como $PV(0) - PD(0)$, que neste caso resulta em 0,278, indicando o acréscimo em termos de probabilidade de ganho por jogar na própria sede. De forma alternativa, temos que $\frac{PV(0)}{PD(0)} = 2{,}23$, indicando que a probabilidade de ganhar em casa é mais que o dobro de perder, uma informação consideravelmente importante de forma geral.

5. Conclusões

Este artigo propõe uma nova metodologia com o uso do Modelo de Resposta Gradual (MRG) de Samejima, largamente empregado na área de Teoria da Resposta ao Item (TRI), para estimar probabilidades de Derrota, Empate e Vitória nas partidas de futebol do Campeonato Brasileiro Série A. Ela permitirá um cálculo fácil de probabilidades de resultados que pode ser facilmente implementado em planilhas e páginas da Internet. Um dos achados foi associado ao Efeito-Casa, indicando que a probabilidade de vitória em casa é mais que o dobro de perder, quando considera-se times equivalentes.

A mesma metodologia pode ser aplicada às categorias B, C e D do futebol, visando identificar se há diferenças consideráveis nas estimativas dos parâmetros do modelo, e Efeito-Casa, quando comparados aos resultados da Série A. Adaptações podem ser feitas para outras modalidades esportivas.

Alternativamente ao MRG, há modelos mais flexíveis, como o Modelo de Resposta Nominal (Bock, 1970), com mais parâmetros, mas que pode incorporar

Referências

1. Andrade, D.F., Tavares, H.R., Valle, R.C. (2000). Teoria da Resposta ao Item: Conceitos e Aplicações. Associação Brasileira de Estatística: São Paulo.
2. Azevedo, C. L. N. (2003) Métodos de Estimação na Teoria da Resposta ao Item. UFC.
3. Baker, F. B. (1992). Item Response Theory - Parameter Estimation Techniques. New York: Marcel Dekker, Inc.
4. Bates, D. M. and Chambers, J. M. (1992) *Nonlinear models.* Chapter 10 of *Statistical Models in S* eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

5. Darrell Bock, R. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29-51.
6. BOCK, R. D. and MISLEVY. Marginal maximum likelihood estimation of item parameters. An application of an EM algorithm. *Psychometrika*, 46, 443-459, 1981.
7. R Core Team (2024). R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.