GET00211 - Modelos Lineares 2

Rafael Erbisti

Instituto de Matemática e Estatística

Universidade Federal Fluminense

Aula 13

Introdução

- Se a variável resposta é categórica, com mais de duas categorias, então há duas abordagens de Modelos Lineares Generalizados.

- Uma é baseada na generalização dos modelos para dados com duas categorias para o caso de mais categorias (modelo politômico), aplicável a respostas nominais ou ordinais.

- Outra abordagem é modelar as frequências, ou contagens, para os diferentes níveis da covariável, como resposta de um modelo de Poisson (modelo log-linear).

Distribuição Multinomial

- Considere uma variável aleatória $\mathbf{Y}$ com J categorias.
- Sejam $\pi_1, \pi_2, \ldots, \pi_J$ as respectivas probabilidades, com

$$\pi_1 + \pi_2 + \cdots + \pi_J = 1.$$

- Se há n observações independentes que resultam em y_1 para a categoria 1, y_2 para a categoria 2, e assim por diante, então seja

$$\mathbf{y} = (y_1, y_2, \ldots, y_J)', \quad \text{com} \quad \sum_{j=1}^{J} y_j = n.$$

Distribuição Multinomial

- A distribuição multinomial é dada por:

$$f(\mathbf{y}; n) = \frac{n!}{y_1!\, y_2!\, \ldots\, y_J!}\, \pi_1^{y_1} \pi_2^{y_2} \ldots \pi_J^{y_J} \tag{1}$$

- Se $J = 2$, então $\pi_2 = 1 - \pi_1$ e $y_2 = n - y_1$, e vemos que a distribuição binomial é um caso particular da distribuição multinomial.

- Em geral, a distribuição multinomial não satisfaz os pré-requisitos para ser um membro da família exponencial.

Distribuição Multinomial

- Entretanto, a relação a seguir com a distribuição de Poisson garante que o uso de MLG é adequado.

- Sejam $Y_1, \ldots, Y_J$ variáveis aleatórias independentes com

$$Y_j \sim \text{Poisson}(\lambda_j).$$

- A função de probabilidade conjunta é:

$$f(\mathbf{y}) = \prod_{j=1}^{J} \frac{\lambda_j^{y_j} e^{-\lambda_j}}{y_j!}.$$

- Seja $n = Y_1 + Y_2 + \ldots + Y_J$. Então n é uma variável aleatória com distribuição:

$$n \sim \text{Poisson}(\lambda_1 + \lambda_2 + \ldots + \lambda_J).$$

Distribuição Multinomial

- Observe que o valor de n é completamente determinado pelos valores de $\mathbf{y} = (y_1, \ldots, y_J)$.
- Assim, o evento $(\mathbf{Y} = \mathbf{y}, N = n)$ só tem probabilidade não nula se

$$n = \sum_{j=1}^{J} y_j.$$

- Portanto, a conjunta é essencialmente a mesma função de probabilidade:

$$f(\mathbf{y}, n) = f(\mathbf{y}) \quad \text{para } n = \sum_{j} y_j.$$

- Isso permite escrever a condicional como:

$$f(\mathbf{y} \mid n) = \frac{f(\mathbf{y})}{f(n)}$$

Distribuição Multinomial

- Portanto, a distribuição condicional de $\mathbf{y}$ dado n é:

$$f(\mathbf{y} \mid n) = \frac{\prod_{j=1}^{J} \frac{\lambda_j^{y_j} e^{-\lambda_j}}{y_j!}}{\frac{(\lambda_1 + \cdots + \lambda_J)^n e^{-(\lambda_1 + \cdots + \lambda_J)}}{n!}}$$

- A expressão anterior pode ser simplificada para:

$$f(\mathbf{y} \mid n) = \frac{n!}{y_1! \cdots y_J!} \left(\frac{\lambda_1}{\sum_{k=1}^{J} \lambda_k} \right)^{y_1} \cdots \left(\frac{\lambda_J}{\sum_{k=1}^{J} \lambda_k} \right)^{y_J} \tag{2}$$

- Logo, $\mathbf{Y} \mid n \sim \text{Multinomial}\,(n, \pi_1, \ldots, \pi_J)$, com

$$\pi_j = \frac{\lambda_j}{\sum_{k=1}^{J} \lambda_k}$$

Distribuição Multinomial

- Se $\pi_j = \dfrac{\lambda_j}{\sum_{k=1}^{J} \lambda_k}$ para $j = 1, \ldots, J$, então a equação (1) é a mesma que a equação (2), e

$$\sum_{j=1}^{J} \pi_j = 1$$

- Portanto, a distribuição multinomial pode ser assumida como a distribuição conjunta de variáveis de Poisson, condicional ao fato de que sua soma seja n.

- Para a distribuição multinomial, pode-se mostrar que:

$$E(Y_j) = n\pi_j \quad \mathrm{V}(Y_j) = n\pi_j(1 - \pi_j) \qquad \mathrm{Cov}(Y_j, Y_k) = -n\pi_j\pi_k$$

Regressão Logística Nominal

- A regressão logística nominal é utilizada quando não há qualquer ordem natural entre as categorias da variável resposta.
- Uma categoria é arbitrariamente escolhida como a categoria de referência.
- Suponha que a referência seja a primeira categoria. Então, os logitos para as outras categorias são definidos como:

$$\text{logit}(\pi_j) = \log\left(\frac{\pi_j}{\pi_1}\right) = \mathbf{x}_j'\boldsymbol{\beta}_j, \quad j = 2, \ldots, J. \tag{3}$$

Regressão Logística Nominal

- As $(J-1)$ equações de logito são usadas simultaneamente para estimar os parâmetros $\boldsymbol{\beta}_j$.

- Uma vez obtidas as estimativas $\hat{\boldsymbol{\beta}}_j$, os preditores lineares

$$\mathbf{x}_j'\hat{\boldsymbol{\beta}}_j$$

  podem ser calculados.

- Da equação anterior, temos:

$$\hat{\pi}_j = \hat{\pi}_1 \exp\left(\mathbf{x}_j'\hat{\boldsymbol{\beta}}_j\right), \quad j = 2, \ldots, J.$$

Regressão Logística Nominal

- Como $\hat{\pi}_1 + \hat{\pi}_2 + \cdots + \hat{\pi}_J = 1$, temos:

$$\hat{\pi}_1 = \frac{1}{1 + \sum_{j=2}^{J} \exp\left(\mathbf{x}_j' \hat{\boldsymbol{\beta}}_j\right)}$$

- E, para $j = 2, \ldots, J$,

$$\hat{\pi}_j = \frac{\exp\left(\mathbf{x}_j' \hat{\boldsymbol{\beta}}_j\right)}{1 + \sum_{k=2}^{J} \exp\left(\mathbf{x}_k' \hat{\boldsymbol{\beta}}_k\right)}$$

Resíduos e Bondade de Ajuste

- Os resíduos do qui-quadrado de Pearson são dados por:

$$r_i = \frac{o_i - e_i}{\sqrt{e_i}},$$

onde o_i e e_i são, respectivamente, as frequências observadas e esperadas, para $i = 1, \ldots, N$, sendo N igual a J vezes o número de padrões distintos da covariável.

- Os resíduos podem ser utilizados para verificar a bondade de ajuste do modelo.

- As estatísticas para avaliação da bondade de ajuste são análogas às da regressão logística binomial.

Resíduos e Bondade de Ajuste

- Estatística Qui-quadrado de Pearson:

$$X^2 = \sum_{i=1}^{N} r_i^2,$$

onde $r_i = \dfrac{o_i - e_i}{\sqrt{e_i}}$ são os resíduos de Pearson.

- Deviance: definida em termos dos valores máximos da log-verossimilhança do modelo ajustado ($\hat{\beta}$) e do modelo saturado ($\hat{\beta}_{\text{max}}$):

$$D = 2\left[\ell(\hat{\beta}_{\text{max}}) - \ell(\hat{\beta})\right]$$

- Se o modelo se ajusta bem, então X^2 e D seguem, assintoticamente, uma distribuição qui-quadrado χ^2_{N-p}, onde p é o número de parâmetros estimados.

Razão de Chances

- Usualmente, é mais fácil interpretar os efeitos das variáveis explicativas em termos das razões de chances (odds ratios) do que diretamente pelos parâmetros β.

- Considere uma variável resposta com J categorias e uma variável explicativa x, que indica se o fator está presente ($x = 1$) ou ausente ($x = 0$).

- As razões de chances para a resposta j ($j = 2, \ldots, J$) em relação à categoria de referência ($j = 1$) são dadas por:

$$OR_j = \psi_j = \frac{\pi_j^{(p)}}{\pi_1^{(p)}} \Bigg/ \frac{\pi_j^{(a)}}{\pi_1^{(a)}}$$

onde $\pi_j^{(p)}$ e $\pi_j^{(a)}$ denotam, respectivamente, as probabilidades da categoria j quando o fator está presente ou ausente.

Razão de Chances

- Para o modelo:

$$\log\left(\frac{\pi_j}{\pi_1}\right) = \beta_{0j} + \beta_{1j}x, \quad j = 2, \ldots, J,$$

as chances são dadas por:

1. Quando o fator está ausente ($x = 0$):

$$\frac{\pi_j^{(a)}}{\pi_1^{(a)}} = \exp(\beta_{0j}),$$

2. Quando o fator está presente ($x = 1$):

$$\frac{\pi_j^{(p)}}{\pi_1^{(p)}} = \exp(\beta_{0j} + \beta_{1j}).$$

Razão de Chances

- Portanto, a razão de chances pode ser escrita como:

$$OR_j = \exp(\beta_{1j}),$$

que é estimado por:

$$\widehat{OR}_j = \exp(\hat{\beta}_{1j}).$$

- Se $\beta_{1j} = 0$, então $OR_j = 1$, o que corresponde a um "não efeito" da presença do fator sobre a categoria j em relação à categoria de referência.

Razão de Chances

- Por exemplo, intervalos de 95% de confiança para OR_j são calculados como:

$$\exp\left(\hat{\beta}_{1j} \pm 1{,}96 \times \text{s.e.}(\hat{\beta}_{1j})\right),$$

  onde s.e.$(\hat{\beta}_{1j})$ é o erro padrão da estimativa $\hat{\beta}_{1j}$.
- Intervalos de confiança que não incluem o valor unitário indicam que o parâmetro β_{1j} é significativamente diferente de zero, ou seja, há efeito da variável explicativa sobre a categoria j em relação à referência.
- Para a regressão logística nominal, as variáveis explicativas podem ser categóricas ou contínuas.
- A escolha da categoria de referência para a variável resposta afeta as estimativas dos parâmetros $\hat{\boldsymbol{\beta}}$.
- Mas não afeta as estimativas das probabilidades ajustadas $\hat{\pi}_j$ nem os valores ajustados do modelo.

Exemplo

- Num estudo sobre veículos, homens e mulheres que digiriam carros grandes, médios e pequenos foram entrevistados sobre suas preferências a respeito de carros.
- A amostra compreendeu 50 indivíduos em cada uma das 6 categorias (masculino/feminino e 3 faixas de idade).
- Pediu-se que eles ordenassem algumas características consideradas na compra de seus carros.
- A tabela a seguir apresenta as proporções observadas segundo o nível de importância dos itens ar condicionado e direção hidráulica, de acordo com o sexo e a idade do indivíduo.
- As categorias de resposta são: *"sem importância"*, *"importante"* ou *"muito importante"*.

Exemplo

Sexo	**Idade**	**Resposta**			Total
		Sem importância	Importante	Muito importante	
Mulheres	18–23	26 (58%)	12 (27%)	7 (16%)	45
	24–40	9 (20%)	21 (47%)	15 (33%)	45
	> 40	5 (8%)	14 (23%)	41 (68%)	60
Homens	18–23	40 (62%)	17 (26%)	8 (12%)	65
	24–40	17 (39%)	15 (34%)	12 (27%)	44
	> 40	8 (20%)	15 (37%)	18 (44%)	41
Total		105	94	101	300

Exemplo

- Escolhemos a categoria "sem importância" como a categoria de base.
- A tabela a seguir mostra o resultado para um modelo de regressão logística, assumindo como categoria de referência a resposta "sem importância" e o grupo de idade 18–23 anos.

$$\log\left(\frac{\pi_j}{\pi_1}\right) = \beta_{0j} + \beta_{1j}x_1 + \beta_{2j}x_2 + \beta_{3j}x_3, \quad j = 2, 3$$

Codificação das variáveis explicativas

$$x_1 = \begin{cases} 1, & \text{se homem} \\ 0, & \text{se mulher} \end{cases} \qquad x_2 = \begin{cases} 1, & \text{se idade entre 24 e 40 anos} \\ 0, & \text{caso contrário} \end{cases}$$

$$x_3 = \begin{cases} 1, & \text{se idade maior que 40 anos} \\ 0, & \text{caso contrário} \end{cases}$$

Exemplo

Parâmetro β	Estimativa b (erro padrão)	Razão de chances, $OR =$ (intervalo de confiança de 95%)	
$\log(\pi_2/\pi_1)$: importante vs. sem importância			
β_{02}: constante	−0,591 (0,284)		
β_{12}: homens	−0,388 (0,301)	0,68	(0,38, 1,22)
β_{22}: 24−40	1,128 (0,342)	3,09	(1,58, 6,04)
β_{32}: > 40	1,588 (0,403)	4,89	(2,22, 10,78)
$\log(\pi_3/\pi_1)$: muito importante vs. sem importância			
β_{03}: constante	−1,039 (0,331)		
β_{13}: homens	−0,813 (0,321)	0,44	(0,24, 0,83)
β_{23}: 24−40	1,478 (0,401)	4,38	(2,00, 9,62)
β_{33}: > 40	2,917 (0,423)	18,48	(8,07, 42,34)

Figura: Ajuste do modelo Multinomial.

Exemplo

- $\ell(\hat{\beta}_{\text{nulo}}; y) = -329{,}27$ (modelo contendo apenas β_{02} e β_{03})
- $\ell(\hat{\beta}; y) = -290{,}35$,
- Pseudo-$R^2 = \dfrac{(-329{,}27 + 290{,}35)}{-329{,}27} = 0{,}118$
- $AIC = -2(-290{,}35) + 16 = 596{,}7$
- Pelas estimativas, intervalos de confiança e razões de chances (OR), observa-se que a importância do ar-condicionado e da direção hidráulica cresce significativamente com a idade.
- Aparentemente, homens consideram essas características menos importantes.

Exemplo

- Para estimar as probabilidades $\hat{\pi}_1$, $\hat{\pi}_2$ e $\hat{\pi}_3$, consideramos o grupo de mulheres ($x_1 = 0$) e faixa etária 18–23 anos ($x_2 = 0$, $x_3 = 0$).
- Para esse grupo:

$$\log\left(\frac{\hat{\pi}_2}{\hat{\pi}_1}\right) = -0{,}591 \qquad \text{e} \qquad \log\left(\frac{\hat{\pi}_3}{\hat{\pi}_1}\right) = -1{,}039$$

- Além disso, sabemos que:

$$\hat{\pi}_1 + \hat{\pi}_2 + \hat{\pi}_3 = 1$$

- Portanto:

$$\hat{\pi}_1 = \frac{1}{1 + e^{-0{,}591} + e^{-1{,}039}} = 0{,}524, \quad \hat{\pi}_2 = 0{,}290, \quad \hat{\pi}_3 = 0{,}186$$

Exemplo

- Agora, considere homens ($x_1 = 1$) com idade acima de 40 anos ($x_2 = 0$, $x_3 = 1$).
- Nesse caso:

$$\log\left(\frac{\hat{\pi}_2}{\hat{\pi}_1}\right) = -0{,}591 - 0{,}388 + 1{,}588 = 0{,}609$$

$$\log\left(\frac{\hat{\pi}_3}{\hat{\pi}_1}\right) = -1{,}039 - 0{,}813 + 2{,}917 = 1{,}065$$

- Sabendo que $\hat{\pi}_1 + \hat{\pi}_2 + \hat{\pi}_3 = 1$, temos:

$$\hat{\pi}_1 = \frac{1}{1 + e^{0{,}609} + e^{1{,}065}} = 0{,}174, \quad \hat{\pi}_2 = 0{,}320, \quad \hat{\pi}_3 = 0{,}505$$

Exemplo: Modelo completo com interações

- O modelo completo ajustado inclui termos para idade, sexo e as interações entre sexo e idade.
- Ele possui 6 parâmetros para $j = 2$ (um intercepto, coeficientes para sexo, duas categorias de idade e duas interações sexo $\times$ idade) e mais 6 parâmetros para $j = 3$, resultando em um total de 12 parâmetros.
- Aqui, $\ell(\hat{\beta}_{\text{max}}; y) = -288{,}38$ e, portanto,

$$D = 2(-288{,}38 + 290{,}35) = 3{,}94$$

- Os graus de liberdade associados a essa deviance são: $gl = 12 - 8 = 4$.
- Como esperado, os valores das estatísticas de bondade de ajuste ($D = 3{,}94$ e $X^2 = 3{,}93$) são muito similares.
- Quando comparadas ao quantil da distribuição $\chi^2_{(4)}$, elas sugerem que o modelo ajustado fornece uma boa descrição dos dados.