Modelagem Bayesiana e Aplicações

Márcia D'Elia Branco

Universidade de São Paulo
Instituto de Matemática e Estatística
http:www.ime.usp.br/m̃branco

Modelos de Regressão 3

Modelo de Regressão multinomial

- Considere agora K possíveis categorias de respostas.
- x_i é o vetor de covariáveis associado a i-ésima unidade amostral.
- $y_{ij} = 1$ se o i-ésima unidade esta na categoria j e $y_{ik} = 0$ para $k \neq j$.
- $p_{ij} = P(y_{ij} = 1)$. Vários tipos de logitos podem ser definidos.
- Um dos mais populares é o logito categoria de referência

$$logitoR_{ij} = log\left(\frac{p_{ij}}{p_{i1}}\right)$$

Modelo de Regressão multinomial

- O modelo de regressão é dado por

$$logitoR_{ij} = \alpha_j + x_i^t \beta_j \quad j = 2, \ldots, K.$$

- Uma reta de regressão para cada uma das $K - 1$ categorias.
- As probabilidades são obtidas por

$$p_{ij} = \frac{e^{\eta_{ij}}}{1+e^{\eta_{ij}}} \quad j = 2, \ldots, K$$

$$p_{i1} = \frac{1}{1+e^{\eta_{ij}}}$$

- em que η_{ij} é o preditor linear.

Modelo de Regressão multinomial ordinal

- Suponha que existe uma ordem entre as categorias. Por exemplo, $C1 < C2 \cdots < C_K$.
- Neste caso, é possível considerar um estrutura latente contínua. Assim, como antes,

$$z_i = x_i^t \beta + \epsilon_i$$

- Em que ϵ_i tem uma distribuição escolhida de acordo com a função de ligação considerada (normal, logistica, skew-normal,...).
- A construção da resposta ordinal é dada por considerar $y_i = k$ se $\theta_{k-1} \leq z_i < \theta_k$ com $k = 1, \ldots, K$. $\theta_0 = -\infty$ e $\theta_K = \infty$.

A Figura a seguir foi copiada do livro do Agresti e ilustra a estrutura latente para ligação probito e K=4 categorias.

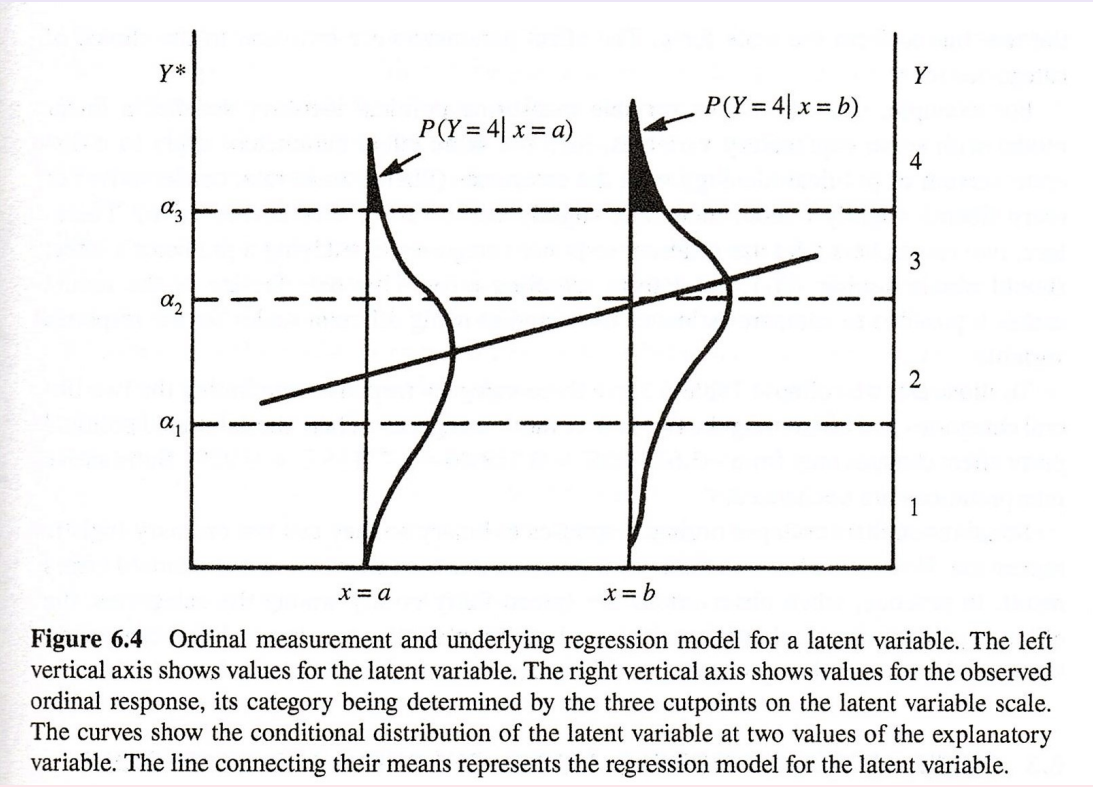

Figure 6.4 Ordinal measurement and underlying regression model for a latent variable. The left vertical axis shows values for the latent variable. The right vertical axis shows values for the observed ordinal response, its category being determined by the three cutpoints on the latent variable scale. The curves show the conditional distribution of the latent variable at two values of the explanatory variable. The line connecting their means represents the regression model for the latent variable.

Modelo de Regressão multinomial ordinal

$$P(y_i = k) = P(\theta_{k-1} \leq z_i < \theta_k) = P(\theta_{k-1} - x_i^t\beta \leq \epsilon_i < \theta_k - x_i^t\beta) =$$

$$F(\theta_k - x_i^t\beta) - F(\theta_{k-1} - x_i^t\beta) = \gamma_{i,k} - \gamma_{i,k-1}$$

em que F é a f.d.a. dos erros latentes.

Na notação introduzida, $\gamma_{i,k}$, representa a probabilidade acumulada associada a i-ésima observação dada por

$$\gamma_{i,k} = p_{i1} + p_{i2} + \cdots + p_{ik}$$

Se considerarmos F como a f.d.a. da logística, temos

$$\gamma_{i,k} = \frac{e^{\theta_k - x_i^t\beta}}{1 + \theta_k - x_i^t\beta}$$

Isolando o preditor linear, temos que

$$logitoA_{ik} = \theta_j - x_i^t \beta$$

em que:

$$logitoA_{ik} = \log\left(\frac{\gamma_{ik}}{1-\gamma_{ik}}\right) = \log\left(\frac{p_{i1} + p_{i2} + \cdots + p_{ik}}{p_{i,k+1} + p_{i,k+2} + \cdots + p_{iK}}\right)$$

$k = 1, \ldots, K-1$. Esses são denominados logitos acumulados.

Por questões de identificabilidade do modelo, o parâmetro intercepto β_0 não é considerado no vetor β.

Exemplo 1: Considere $K = 4$ e as seguintes categorias

C_1: discorda completamente
C_2: discorda parcialmente
C_3: concorda parcialmente
C_4: concorda completamente

Temos, 3 logitos:

$$logitoA_{i1} = \log\left(\frac{p_{i1}}{p_{i2}+p_{i3}+p_{i4}}\right)$$

$$logitoA_{i2} = \log\left(\frac{p_{i1}+p_{i2}}{p_{i3}+p_{i4}}\right)$$

$$logitoA_{i3} = \log\left(\frac{p_{i1}+p_{i2}+p_{i3}}{p_{i4}}\right)$$

Esses logitos podem ser interpretados, respectivamente, como:

- o logaritmo da chance de discordar completamente;
- o logaritmo da chance de discordar e
- o logaritmo do inverso da chance de concordar completamente.

Para completar o modelo, temos que definir distribuições *a priori.*

- Para o vetor dos preditores $\beta \sim N_p(m_0, V_0)$.
- Para os pontos de cortes θ_k precisamos estabelecer a ordenação $\theta_1 < \theta_2 < \cdots < \theta_{K-1}$.
- Uma maneira de incluir essa restrição é considerar

$$\theta_k = \theta_{k-1} + e^{\Delta_k} \quad \text{e} \quad \theta_1 = \Delta_1$$

Modelo de Regressão multinomial ordinal

- $(\Delta_1, \Delta_2, \ldots, \Delta_{K-1})$ é um vetor de hiperparâmetros em R^{K-1}
- Podemos atribuir para esse vetor uma distribuição normal.
- A vantagem de trabalhar com o modelo com estrutura latente é de poder monitorar os resíduos latentes $\epsilon_i = z_i - x_i^t \beta$.
- Os codigos BUGS para implementação deste modelo estão na página 134 do livro.
- Ver Exemplo 3.9.

Aplicação de Métodos de Seleção de Variáveis

- Doenças vasculares são altamente influenciadas por uma medida clinica denominada HSP. Valores elevados de HSP são indicativos de risco maior de doença.
- Foi considerada uma amostra de $n = 145$ pacientes que sofreram infarte ou AVC.
- O primeiro objetivo do estudo foi identificar as covariáveis que influenciam na medida HSB, com base na amostra.
- Foram consideradas inicialmente 13 covariáveis; sendo 4 qualitativas e 9 quantitativas.
- A variável resposta HSP apresenta uma forte assimetria e por isso o modelo normal foi descartado.

Aplicação de Métodos de Seleção de Variáveis

- Dois modelos foram considerados: Gama e LogNormal.
- Para o modelo Gama: $Y_i \sim Gamma(\nu, b_i)$ em que $\mu_i = \nu/b_i$.
- Foi considerada a função de ligação $g(\mu_i) = \log(\mu_i)$.
- No contexto de MLG, temos

$$g(\mu_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}.$$

- As distribuições a priori consideradas foram

$$\beta_j \sim N(0, 10^4) \quad \nu \sim Gamma(1, 10^{-4})$$

independentes.

As estatísticas resumos *a posteriori* são apresentadas na Tabela 10.12 do livro Paulino et al. (2018), reproduzida abaixo.

Tabela 10.12: Resumos de distribuições *a posteriori* para o modelo Gama.

variável	par.	média	dp	2.5%	50%	97.5%
	β_0	2.133	0.309	1.522	2.137	2.728
sexo	β_1	0.275	0.074	0.125	0.274	0.419
histfa	β_2	-0.112	0.047	-0.206	-0.111	-0.020
gluc	β_3	0.000	0.001	-0.001	0.000	0.001
ureia	β_4	-0.001	0.003	-0.007	-0.001	0.004
creat	β_5	0.123	0.043	0.041	0.122	0.208
ggt	β_6	-0.002	0.001	-0.003	-0.002	-0.000
falc	β_7	0.003	0.001	0.002	0.003	0.005
ldl	β_8	0.001	0.001	-0.000	0.001	0.003
hdl	β_9	-0.002	0.002	-0.005	-0.002	0.002
pasis	β_{10}	0.003	0.002	-0.002	0.003	0.008
padia	β_{11}	-0.010	0.004	-0.018	-0.010	-0.002
fumar1	β_{12}	0.213	0.071	0.068	0.213	0.352
fumar2	β_{13}	-0.009	0.058	-0.125	-0.008	0.102
	ν	15.897	2.092	12.080	15.790	20.270
	$\ln(\nu)$	2.757	0.132	2.492	2.759	3.009

Aplicação de Métodos de Seleção de Variáveis

- Os intervalos de credibilidade 0.95 que não contêm o zero, correspondem as covariáveis *sexo, histfa, creat, ggt, falc, padia* e *fumar1*.
- Os demais intervalos contêm o zero, usando esse critério um novo modelo poderia ser ajustado apenas com a covariáveis mencionadas acima.
- Reduzindo-se de 13 para 7 o número de preditores.
- O segundo método de seleção de variáveis considerado foi o SSVS (*Stochastic search variable selection*).
- Relembrando o método:

$$h(\beta_j \mid v_j) = (1 - v_j)N(0, c\tau_j^2) + v_j N(0, \tau_j^2) \quad 0 < c < 1$$

$$v_j \sim Be(r_j)$$

Aplicação de Métodos de Seleção de Variáveis

- Foram considerados $v_j = 0.5$ para todo j. Indiferença entre incluir ou não a covariável j.
- Para o valor de τ_j^2 foram consideradas as variâncias a posteriori quando ajustado o modelo completo.
- Finalmente, foi fixado $c = 0.10$.
- A Tabela 10.14 (Paulino et al , 2018) apresenta os cinco modelos com maiores probabilidades *a posteriori*.
- Na última linha da tabela temos as probabilidades obtidas para cada um dos modelos.
- Na última coluna da tabela temos as probabilidades de incluir cada uma das covariáveis.

Tabela 10.14: Método SSVS no modelo Gama.

	M_1	M_2	M_3	M_4	M_5	$P(\upsilon_j = 1\|\mathcal{D})$
sexo (x_1)	1	1	1	1	1	0.955
histfa (x_2)	1	1	1	1	1	0.741
gluc (x_3)	0	0	0	0	0	0.417
ureia (x_4)	0	1	0	0	0	0.406
creat (x_5)	1	1	1	1	1	0.989
ggt (x_6)	1	1	1	1	1	0.582
falc (x_7)	1	1	1	1	1	0.942
ldl (x_8)	1	1	1	0	1	0.540
hdl (x_9)	1	1	1	1	0	0.591
pasis (x_{10})	0	0	0	0	0	0.423
padia (x_{11})	1	1	1	1	1	0.860
fumar1 (x_{12})	1	1	1	1	1	0.990
fumar2 (x_{13})	0	0	1	0	0	0.417
$P(M_k\|\mathcal{D})$	0.015	0.011	0.011	0.010	0.010	

Aplicação de Métodos de Seleção de Variáveis

- Segundo o critério SSVS, o melhor modelo é o M_1 que possue 9 preditores.
- Note que o modelo escolhido anteriormente, olhando apenas os IC, não figura entre os 5 de maior probabilidade.
- O segundo ajuste considerado, foi o modelo LogNormal. Fazendo $Y_i \sim Ln(\mu_i, \nu)$ com $g(\mu_i) = \mu_i$.
- Similarmente, utilizando o método SSVS, foram selecionadas as 7 covariáveis: *sexo, histfa, creat, ggt, falc, padia* e *fumar1*.
- A Tabela 10.16 (Paulino et al , 2018) apresenta medidas de comparação de modelos.
- Os modelos Gama2 e LogNormal2 correspondem aos ajustados com as covariáveis selecionadas segundo SSVS para o melhor modelo Gama.

Tabela 10.16: Comparação dos modelos via medidas de desempenho preditivo.

	p	DIC	p_D	$WAIC_1$	p_{w_1}	$WAIC_2$	p_{w_2}
Gama0	14	603.00	15.03	601.72	13.80	606.45	16.16
Gama1	8	597.57	9.09	598.82	10.40	602.29	12.13
Gama2	10	**596.87**	11.08	**597.32**	11.53	**600.89**	13.32
LogNorm0	14	594.50	15.15	592.29	12.94	596.08	14.83
LogNorm1	8	**587.90**	9.03	**588.27**	9.36	**590.44**	10.45
LogNorm2	10	588.80	11.11	588.60	10.87	591.66	12.39

Aplicação de Métodos de Seleção de Variáveis

- Os critérios DIC e WAIC selecionam os mesmos modelos indicados pelo SSVS, para cada uma das distribuições.
- Os critérios também indicam uma leve preferência pelo *LogNorm1*.
- Além disso, esse modelo requer menos covariáveis.