4

Introdução à Regressão Logística

Parte 2 – Modelos Logitos

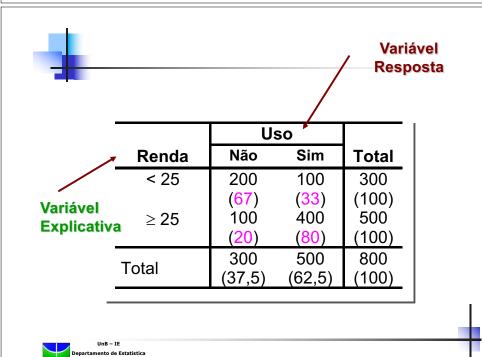
Unidade III

UnB – IE

Departamento de Estatística

Análise de Dados Categorizados

Maria Teresa Leão Costa





Modelos Logitos para Variáveis Qualitativas

Exemplo 1 - Uso de Cartão de Crédito

Em uma pesquisa de mercado para uma certa firma, deseja-se analisar os dados sobre uso de cartão de crédito e renda anual.

Coletou-se dados sobre Renda em 1000 u.m. e uso de cartão de crédito (sim / não)

Categorias de renda: < 25 000 u.m.;

≥ 25,000 u.m.;





Variável Explicativa Qualitativa Binária

 O uso de variáveis indicadoras ("dummy") permite a inclusão de variáveis qualitativas, denominadas fatores, no modelo.

$$\pi(x_i) = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}$$

sendo que: $X_i = 0$ se renda < 25

 $X_i = 1$. se renda ≥ 25





Resposta média logito:

$$\pi^*(\mathbf{x}_i) = \ln \left(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} \right) = \beta_o + \beta_1 \mathbf{x}_i$$

• Note que se $X_i = 0$ então:

$$\ln \left(\frac{\pi(0)}{1-\pi(0)}\right) = \beta_o + \beta_1 \times 0 = \beta_o \Rightarrow \frac{\pi(0)}{1-\pi(0)} = e^{\beta_o}$$

e se $X_i = 1$ então:

$$\ln \left(\frac{\pi(1)}{1-\pi(1)}\right) = \beta_o + \beta_1 \times 1 = \beta_o + \beta_1 \Rightarrow \frac{\pi(1)}{1-\pi(1)} = e^{\beta_o + \beta_1}$$





Model Fit Statistics				
Intercept an Criterion Only Covariate				
AIC	1060.501	886.311		
sc	1065.186	895.680		
-2 Log L	1058.501	882.311		

Number of Observations Read	4
Number of Observations Used	4
Sum of Frequencies Read	800
Sum of Frequencies Used	800

Response Profile				
Ordered Value	Total Frequency			
1	n	300		
2	s	500		

Probability modeled is cartao='s'

Class Level Information				
Class	Design Variables			
renda	<25	0		
	>=25	1		

Testing Global Null Hypothesis: BETA=0					
Test	Chi-Square	DF	Pr > ChiSo		
Likelihood Ratio	176.1903	1	<.0001		
Score	174.2222	1	<.0001		

Analysis of Maximum Likelihood Estimates						
Parameter DF Estimate Standard Wald Chi-Square Pr > ChiSq						
Intercept		1	-0.6931	0.1225	32.0302	<.0001
renda	>=25	1	2.0794	0.1658	157.2385	<.0001

$$\hat{\beta}_1 = 2,0794 \implies \hat{\theta} = e^{2,0794} = 7,9997$$





Consequentemente:

$$\frac{\left(\frac{\pi(1)}{1-\pi(1)}\right)}{\left(\frac{\pi(0)}{1-\pi(0)}\right)} = \frac{e^{\beta_o + \beta_1}}{e^{\beta_o}} = e^{\beta_1} = \theta$$

que mede o quanto mais provável (ou improvável) é o "sucesso" entre os com nível x=1 do que entre os com nível x=0.

 Cabe ressaltar que este resultado só vale quando a variável for codificada como 0 ou 1.





Efeito da codificação e o papel das variáveis dummy

X	Sucesso (Y=1)	Insucesso (Y=0)	odds	Logito ln(odds)
a	$\pi(a) = \frac{e^{\beta_0 + \beta_1 a}}{1 + e^{\beta_0 + \beta_1 a}}$	$1 - \pi(a) = \frac{1}{1 + e^{\beta_0 + \beta_1 a}}$	$\frac{\pi(a)}{1-\pi(a)} = e^{\beta_0 + \beta_1 a}$	$\beta_0 + \beta_1 a$
b	$\pi(b) = \frac{e^{\beta_0 + \beta_1 b}}{1 + e^{\beta_0 + \beta_1 b}}$	$1 - \pi(b) = \frac{1}{1 + e^{\beta_0 + \beta_1 b}}$	$\frac{\pi(b)}{1-\pi(b)} = e^{\beta_0 + \beta_1 b}$	$\beta_0 + \beta_1 b$

$$ln(\theta) = ln\left(\frac{\frac{\pi(a)}{1-\pi(a)}}{\frac{\pi(b)}{1-\pi(b)}}\right) = ln\left(\frac{\pi(a)}{1-\pi(a)}\right) - ln\left(\frac{\pi(b)}{1-\pi(b)}\right) =$$
$$= \beta_0 + \beta_1 a - (\beta_0 + \beta_1 b) = \beta_1 (a - b)$$

$$ln(\theta) = \beta_1(a-b) \implies \theta = e^{\beta_1(a-b)}$$





Métodos de Construção de Variáveis dummy

- *Método parcial célula de referência* atribui *0* ao nível a ser considerado como referência e 1 ao outro nível.
- *Método marginal* desvio da média atribui –1 a um dos níveis e 1 ao outro nível.





EXEMPLO 2 - Coração x Idade

Em uma pesquisa deseja-se estudar se idade é um fator de risco para doença coronariana.

Coletou-se dados para uma amostra de 100 pacientes sobre:

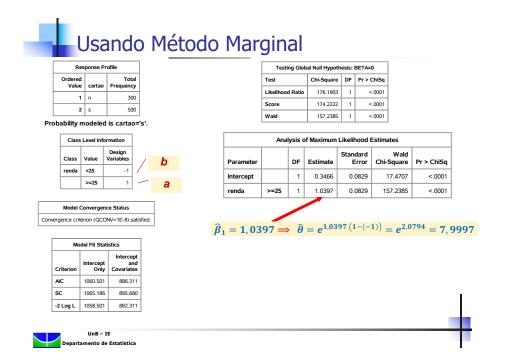
• Doença coronariana $(0 - n\tilde{a}o/1 - sim).$

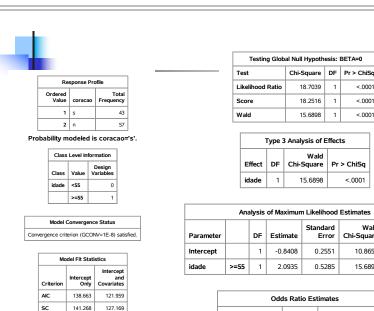
• Idade (< 55 anos / >=55 anos).

e os seguintes resultados foram obtidos:

ldade	Doença Coronariana		Total
	sim	não	
< 55 anos	22	51	73
>=55 anos	21	6	27
Total	43	57	100







136.663

	Odds Ratio Estimates				
Effect	:	Point Estimate			
idade	>=55 vs <55	8.114	2.880	22.861	

<.0001

<.0001

Chi-Square

15.6898

Pr > ChiSq

0.0010

<.0001



Variável Explicativa Qualitativa **Politômica**

- Em geral, se uma variável qualitativa tem *l* níveis possíveis, então 1-1 variáveis dummy ou de delineamento são necessárias.
- As l-1 variáveis dummy serão designadas como D_u e os coeficiente destas variáveis serão designados por β_{u} , u = 1, ... l - 1.

O modelo expresso em termos da função logito será:

$$logito(\pi) = ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \sum_{u=1}^{l-1} \beta_u D_u$$





EXEMPLO 3 - Coração x Raça Variável qualitativa politômica

Em uma pesquisa deseja-se estudar se existe associação entre raça e doença coronariana.

Coletou-se dados para uma amostra de 100 pacientes

- Doença coronariana $(0 - n\tilde{a}o/1 - sim).$
- Raça (branca / negra /hispânica/ outra)

e os resultados seguintes resultados foram obtidos:

Raça	Doen Corona	Total	
	sim	nao	
Branca	5	20	25
Negra	20	10	30
Hispânica	15	10	25
Outra	10	10	20
Total	50	50	100

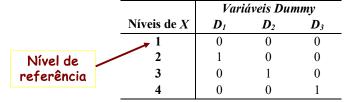




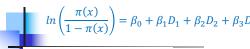
Métodos de Construção de Variáveis *dummy*

Método parcial – célula de referência –

Este método faz todas as variáveis dummy igual a zero para o nível considerado de referência e então aloca uma única variável *dummv* igual a 1 para cada um dos outros níveis de X.







Response Profile			
Ordered Value	coracao	Total Frequency	
1	s	50	
2	n	50	

Probability modeled is coracao='s'.

Class Level Information					
Class	Value	Design Variables			
raca	branca	-1 -1 -1			
	hispanic	1	0	0	
	negra	0	1	0	
	outra	0	0	1	

Model Convergence Status onvergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics				
Criterion	Intercept Only	Intercept and Covariates		
AIC	140.629	132.587		
sc	143.235	143.008		
-2 Log L	138.629	124.587		

n l	$(\underline{\pi(x)})$	$= \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3$
π	$(1-\pi(x))$	$-p_0 + p_1 D_1 + p_2 D_2 + p_3 D_3$

3	Testing Global Null Hypothesis: BETA=0						
J	Test	Chi-Square	DF	Pr > ChiSo			
	Likelihood Ratio	14.0420	3	0.0028			
	Score	13.3333	3	0.0040			
	Wald	11.7715	3	0.0082			

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
raca	3	11 7715	0.0082

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
ntercept		1	-0.0719	0.2189	0.1079	0.7425
aca	hispanic	1	0.4774	0.3623	1.7363	0.1876
aca	negra	1	0.7651	0.3506	4.7619	0.0291
aca	outra	1	0.0719	0.3846	0.0350	0.8517

Odds Ratio Estimates					
Effect		95% Wald fidence Limits			
aca hispanic vs branca	6.000	1.693 21.261			
aca negra vs branca	8.000	2.316	27.633		
aca outra vs branca	4.000	1.074	14.895		

П

