REGRESSÃO LOGÍSTICA

1. Introdução

Definimos variáveis categóricas como aquelas variáveis que podem ser mensurados usando apenas um número limitado de valores ou categorias. Esta definição distingue variáveis categóricas de variáveis contínuas, as quais, em princípio, podem assumir um número infinito de valores. Muitas variáveis de interesse para cientistas sociais são claramente categóricas, entre as quais podemos destacar raça, gênero, estado civil, emprego, nascimento, e morte.

Esse método é utilizado para estudar variáveis *dummys* que são aquelas que são compostas apenas por duas opções de eventos, como “sim” ou “não”. Por exemplo:

Seja Y uma variável aleatória dummy definida como:

$$Y = \begin{cases} 1 \text{ se a pessoa obteve crédito} \\ 0 \text{ se a pessoa não obteve crédito} \end{cases}$$

Onde cada Y_i tem distribuição de Bernoulli, cuja função de distribuição de probabilidade é dada por;

$$P(y \mid p) = p^{y}(1-p)^{1-y}$$

onde:

y identifica o evento ocorrido

p é a probabilidade de sucesso para a ocorrência do evento

Como se trata de uma seqüência de eventos com distribuição de Bernoulli, a soma do número de sucessos ou fracassos neste experimento terá distribuição Binomial de parâmetros n (número de observações) e p (probabilidade de sucesso). A função de distribuição de probabilidade da Binomial é dada por;

$$P(y \mid n, p) = \binom{n}{y} p^{y}(1-p)^{1-y}$$

A transformação logística pode ser interpretada como sendo o logaritmo da razão de probabilidades, sucesso versus fracasso, onde a regressão logística nos dará uma idéia do risco de uma pessoa obter crédito dado o efeito de algumas variáveis explicativas que serão introduzidas mais à frente.

A função de ligação deste modelo linear generalizado é dada pela seguinte equação:

$$\eta_i = \log\left(\frac{p_i}{1-p_i}\right) = \sum_{k=0}^{K} \beta_k x_{ik}$$

onde a probabilidade p_i é dada por:

$$p_i = \frac{\exp\left(\sum_{k=0}^{K} \beta_k x_{ik}\right)}{1 + \exp\left(\sum_{k=0}^{K} \beta_k x_{ik}\right)}$$

2. Exemplo

Os modelos utilizados aqui como exemplo têm como objetivo identificar as variáveis relacionadas com as características de interesse (variável resposta). Ao realizar o ajuste do modelo, deseja-se encontrar, e identificar, quais são os fatores importantes que melhor descrevem o comportamento/variação das características de interesse.

O modelo linear generalizado aqui utilizado é definido por uma distribuição de probabilidade para a variável resposta, um conjunto de variáveis independentes (fatores explicativos) que compõem o previsor linear do modelo, e uma função de ligação entre a média da variável resposta e o referido previsor linear.

Mobilidade Ocupacional e a utilização de Regressões logísticas.

Questiona se a tendência dos afro-brasileiros é prosperar – reduzindo as disparidades raciais – ou regredir, ampliando-as. Nas próximas seções pretendemos averiguar a mobilidade ocupacional, por raça, avaliando as chances de indivíduos com as mesmas características (sexo, educação, idade, etc.) exceto a sua cor, em um horizonte de cinco anos.

A tabela a seguir representa a transição das categorias de ocupação dos indivíduos entre 1991 e 1996.

Definição das variáveis de interesse:

OCUPA2:

Tabela 3.1 - Mobilidade ocupacional entre 1991 e 1996

Categoria de ocupação em 1991		Categoria de ocupação em 1996	
		Ocupado	Desempregado
	Total 1		
Ocupado	34685	28538	6147
Desempregado	19594	4712	14882

Fonte:

- ⇨ Ocupa2=1 se o entrevistado declarou estar desempregado em 1991 e ocupado em 1996;
- ⇨ Ocupa2=0 se o entrevistado declarou estar desempregado em 1991 e desempregado em 1996

.

Tabela 3.2 - Análise univariada : OCUPA2

Frequência das variávies explicativas segundo a condição de ocupação do entrevistado

	Total	Observações utilizadas	
		Total de "0"	Total de "1"
	2119	**220**	**1899**
Sindicalizado			
Não		212	1572
Sim		8	327
Grupos etários			
Entre 15 e 29 anos		117	761
Entre 30 e 44 anos		83	860
Entre 45 e 59 anos		17	253
60 anos ou mais		3	25
Região Metropolitana			
Rio de Janeiro		24	289
São Paulo		70	391
Porto Alegre		23	171
Belo Horizonte		31	323
Recife		33	344
Salvador		39	381

Fonte: PME/IBGE

Para selecionar o modelo utilizou-se a PROC GENMOD do SAS (maiores detalhes em www.sas.com). Os modelos finais foram selecionados passo a passo, após agrupamento de níveis dos fatores com base na estatística de Wald, incluindo-se em cada passo as interações que produziam maior decréscimo da Deviance, considerando o teste da razão.

Seleção do modelo para variável OCUPA2.

Os modelos finais foram selecionados passo a passo, após agrupamento de níveis dos fatores com base na estatística de Wald, incluindo-se em cada passo as interações que produziam maior decréscimo da Deviance, considerando o teste da razão. Nenhum das interações foram significativas.

Tabela 3.3 - Teste da Razão de Verossimilhança para o modelo final					
Tipo	Código	Deviance	G.L.	Qui-quadrado	P-valor
	INTERCEPT	101.9391	0	.	.
Sindicalizado ou associado	SINDA	65.4219	1	36.5172	0.0001
Região Metropolitana	REG	43.8721	5	21.5498	0.0006
Grupos etários	FXAGE	26.6021	3	17.27	0.0006

Interpretação das estimativas – Vantagens e Razão de vantagens - OCUPA2

Considerando as estimativas apresentadas na tabela 4.8, verifica-se que uma pessoa da Região Metropolitana do Rio de Janeiro apresenta uma vantagem de ***2.3*** de sair do desemprego do que uma pessoa em São Paulo.

A vantagem em favor da ocorrência do evento (estar desempregado em 1991 e ocupado em 1996) para os entrevistados que declaram não ser sindicalizado ou associado a algum órgão de classe é 85% menor do que declarou ser sindicalizada. Para as pessoas entre 30 e 44 anos a vantagem em favor da ocorrência do evento é 10% maior do que os outros grupos.

Estimativas dos Parâmetros para o modelo final

Parâmetro	Códigos	Descrição	G.L.	Estimativa	Erro Padrão	Qui-quadrado	P-valor	Vantagem
Intercepto			1	3.4481	0.7218	22.8227	0.0001	31.441
Sindicalizado ou associado	NSIND	Sindicalizado ou associado	1	-1.867	0.3678	25.7633	0.0001	0.155
	SIND	Não é sindicalizado	0	0	0	.	.	
Região Metropolitana	BA	Salvador	1	0.6778	0.2157	9.871	0.0017	1.970
	PE	Recife	1	0.7019	0.2271	9.552	0.002	2.018
	RS	Porto Alegre	1	0.2063	0.2626	0.6171	0.4321	1.229
	MG	Belo Horizonte	1	0.7334	0.2317	10.0162	0.0016	2.082
	RJ	Rio de Janeiro	1	0.834	0.2523	10.9271	0.0009	2.303
	SP	São Paulo	0	0	0	.	.	
Grupos etários	ID_15_29	Grupo etário - 15 a 29 anos	1	-0.3811	0.6251	0.3717	0.5421	0.683
	ID_30_44	Grupo etário - 30 a 44 anos	1	0.0989	0.6277	0.0248	0.8748	1.104
	ID_45_59	Grupo etário - 45 a 59 anos	1	0.5155	0.666	0.5991	0.4389	1.674
	ID_60_MA	Grupo etário - 60 anos ou mais	0	0	0	.	.	

3. Modelo Logit Multinomial[1]

Muitos estudos de relevância social são mensurados através de variáveis qualitativas não ordenadas. Por exemplo, sociólogos e economistas estão interessados na composição da força de trabalho (empregados, desempregados); cientistas políticos em afiliações partidárias (direita, esquerda); geógrafos e demógrafos nas regiões de residência (Nordeste, Norte, Sul, etc.).

É um dos muitos métodos utilizado para analisar variáveis de resposta categórica não ordenada (nominal) nas pesquisas de ciências sociais. Podemos citar algumas razões para esta popularidade: tal modelo é uma generalização do modelo *logit binomial*; é equivalente para o modelo log-linear com dados agrupados e; estão disponíveis no mercado de vários softwares estatísticos para o ajuste destes modelos.

Quando dizemos que uma variável é não ordenada, dizemos que cada categoria é única em comparação com outras categorias.

Para o resultado da variável (y) com J categorias (j=1, ..., J), vejamos a diferença da j-ésima (j>1) categoria com a primeira ou a categoria base, derivando a base logit para a j-ésima categoria.

$$B_j = \log\left[\frac{P(y=j)}{P(y=1)}\right] = \log\left(\frac{p_j}{p_1}\right), j = 2,..., J \longrightarrow (1)$$

Onde p_j e p_1 denotam as probabilidades da j-ésima e primeira categoria. A escolha do uso da primeira categoria como base foi arbitrária.
Alguma outra categoria poderia ser usada como base. Na transformação da estrutura, nós podemos retornar a base do logit especificado na Eq. (1) como função linear de x. Contanto, é necessário especificar a categoria de contraste (isto é j) como também a categoria base (1 neste caso) quando modelamos resultados qualitativos não ordenados. Existe J-1 bases não redundantes para resultados de variáveis com J categorias.

Agora consideramos o caso de termos apenas uma variável independente x com um número limitado de categorias (x=1,...I). Este caso é equivalente a tabela de contigência, cada valor de x (x=i), a base é:

$$\log\left[\frac{P(y=j/x=i)}{P(y=1/x=i)}\right] = \log\left[\frac{p_{ij}}{p_{i1}}\right] = B_{ij} \longrightarrow (2)$$

$$\log\left[\frac{F_{ij}}{Fi1}\right] = \log\left[\frac{f_{ij}}{f_{i1}}\right], \longrightarrow (3)$$

Considerando neste contexto temos especificado um modelo saturado, a estimação da Eq (2) pode ser obtida como:

[1] Esta seção baseia-se no livro Statistical Methods for Categorical Data Analysis – Daniel A Powers, Yu Xie – capítulo 7.

onde f_{ij} e F_{ij}, são as freqüências observada e esperada na i-ésima linha e j-ésima coluna para a classificação da tabela X x Y. Nós podemos facilmente rescrever o resultado na forma de Modelo Linear Generalizado:

$$B_{ij} = \sum_{i=1}^{I} \log\left(\frac{F_{ij}}{F_{i1}}\right).I(x=i) \longrightarrow (4)$$

onde I(.) é a função indicadora, I=1 se verdadeira, 0, caso contrário. Com variável dummy codificando e a primeira categoria como referência, Eq. (4) é usualmente escrita como:

$$B_{ij} = \alpha_j \sum_{i=1}^{I} \beta_{ij}.I(x=i), x>1, \longrightarrow (5)$$

onde α_j é a base para x=1, e β_{ij} é a diferença na base entre x=i e x=1, Nesse caso simples, αj e βij podem ser estimados separadamente para todo i e j. Estimações simultâneas resultarão num modelo equivalente neste caso. Para outros modelos do que o modelo saturado, separar e estimar simultaneamente em geral gera resultados diferentes.

Modelo Logit Multinomial padrão

Vejamos agora a uma situação mais geral com dados individuais e mudanças na notação dado que i agora represente o i-ésimo indivíduo. Seja y_i uma variável com resultados politômicos com categorias codificadas por 1, ..., J. Associando com cada categoria é uma probabilidade de resposta, (P_{i1}, P_{i2},...P_{iJ}) representam a chance do i-ésimo respondente numa categoria particular.

Como no caso de resultados binários, assumimos a presença de um vetor que mede características dos repondentes, xi (incluindo 1 como o primeiro elemento), como preditores das probabilidades respondente.

Utilizando a notação da função índice, a resposta para a probabilidade depende de transformações não lineares da função linear $X_i\beta_{ij} = \sum_{k=0} \beta_{jk}\, x_{ik}$, onde k é o número de preditores (na notação,o primeiro parâmetro B_0 é o termo de intercepto, o mesmo alfa da eq. 8). É importante notar que, os casos para modelo binomial logit, os parâmetros no modelo multinomial logit apresentam vários resultados categóricos.

O modelo multinomial logit pode ser visto como uma extensão do modelo binário logit, expresso pela eq. 2 e 3, situações em que o resultado das variáveis tem múltiplas categorias não ordenadas. Por exemplo, no caso de três categorias (J=3), nós podemos escrever as probabilidades:

onde $\beta 2$ e $\beta 3$ denotam os efeitos das covariáveis especificadas para a segunda e terceira categorias de resposta com a primeira categoria usada como referencia. Note que a

$$\Pr(y_i = 1/x_i) = P_{i1} = \frac{1}{1+\exp(x_i'\beta_{12})+\exp(x_i'\beta_{13})},$$

$$\Pr(y_i = 2/x_i) = P_{i1} = \frac{\exp(x_i'\beta_2)}{1+\exp(x_i'\beta_{22})+\exp(x_i'\beta_{23})}$$

$$\Pr(y_i = 3/x_i) = P_{i1} = \frac{\exp(x_i'\beta_3)}{1+\exp(x_i'\beta_{32})+\exp(x_i'\beta_{33})},$$

$$P_{i1} = \frac{\eta_{i1}}{\eta_{i1}+\eta_{i2}+\eta_{i3}},$$

$$P_{i2} = \frac{\eta_{i2}}{\eta_{i1}+\eta_{i2}+\eta_{i3}}, \longrightarrow (10)$$

$$P_{i3} = \frac{\eta_{i3}}{\eta_{i1}+\eta_{i2}+\eta_{i3}},$$

equação Pi1 é derivada do contraste entre a soma das três probabilidades que é 1. Isto é, $P_{i1}=1-(P_{i2}+P_{i3})$, onde yi = 1 define a base.

As probabilidades da equação acima podem ser expressas em termos da função exponencial dos termos lineares ηij=exp(x'iβj):

Estimação

A estimação é obtida iterativamente usando máxima verossimilhança. É coveniente definir um conjunto de J variáveis dummy: dij=1 se yi=j e 0 caso contrário. Este resultado em um e apenas um dij=1 para cada observação. O log da verossimilhança é:

$$\log L = \sum_{i=1}^{n} \sum_{j=1}^{J} d_{ij} \log Pij \longrightarrow (13)$$

Interpretando os resultados de um Modelo Logit Multinomial - Vantagem e Razão de vantagem

Uma importante parte do modelo multinomial somente como elas são em respostas binárias e modelos loglineares. Na estrutura modelo multinomial logit, a vantagem entre categorias j e 1 é dada por i simplesmente:

$$\frac{P_{ij}}{P_{i1}} = \frac{\eta_{ij}}{\eta_{i1}} = \exp(x_i^{'}\beta_j) \longrightarrow j = 2,...J \longrightarrow (14)$$

O log da vantagem, ou logit, está na função linear de x_i:

Mobilidade Ocupacional e a utilização do Logit Multinomial

O objetivo desta seção é verificar quais são os fatores explicativos a serem consiiderados para explicar a qualidade da mobilidade ocupacional. A partir dos resultados obtidos na seção 3 (Aplicação da Análise de Correspondência) identificamos três grupos, resta-nos agora investigar a mobilidade destes grupos entre 1991 e 1996.

Quadro 6.1 – Definição das variáveis de interesse

1991		**1996**	
	Grupo 1 em 1996	Grupo 2 em 1996	Grupo 3 em 1996
Grupo 1 em 1991	**3**	**1**	**2**
Grupo 2 em 1991	**1**	**3**	**2**
Grupo 3 em 1991	**1**	**2**	**3**

Definição das variáveis endógenas (de transição) analisadas:

GRUPO_B:

- ⇨ GRUPO_B=1 se (grupo 2 em 1991) e (grupo1 em 1996)
- ⇨ GRUPO_B=2 se (grupo 2 em 1991) e (grupo 3 em 1996)

⇨ GRUPO_B=3 se (grupo 2 em 1991) e (grupo 2 em 1996) SERÁ A BASE

Distribuição das pessoas de 20 anos e mais – grupos ocupacionais

6.1 - Distribuição das pessoas de 20 anos e mais - grupos ocupacionais

Grupos de posição na ocupação em 1991		Grupos de posição na ocupação em 1996		
	Total 2	Grupo 1	Grupo 2	Grupo 3
Total 1		1528	16167	10780
Grupo 1	939	537	96	306
Grupo 2	19128	409	14213	4506
Grupo 3	8408	582	1858	5968

Nota 1: Total 1 distribuição dos grupos ocupacionais em 1996

Nota 2: Total 2 distribuição dos grupos ocupacionais em 1991

Tabela 6.3 - Análise da variância - Teste da Razão de Máxima Verossimilhança
Modelo Logit Multinomial

Variável resposta: GRUPO_B

Número de níveis da var. resposta: 3

Frequência das observações utilizadas: 17825

Parâmetros	G.L.	Qui-quadrado	P-valor
INTERCEPT	2	479.98	0
SINDA	2	899.36	0
GRAU2	2	87.92	0
SERV	2	652.26	0
SANPGP	2	68.57	0
CURPRO	2	40.5	0
FX_COR	2	41.75	0
SEXO	2	96.67	0
CNECMAI	2	13.25	0.0013
CHEFE	2	12.08	0.0024
CLOBO91	2	10.29	0.0058

Fonte: PME/IBGE

Seleção do modelo

Análise das estimativas de Máxima Verossimilhança - GRUPO_B					
Parâmetro	Estimativa	Erro padrão	Qui-	P-valor	Vantagem
Grupo_B =					
INTERCEP	-4.6306	0.2195	445.12	0	0.0097
SINDA	-1.0057	0.1172	73.64	0	0.3658
GRAU2	1.0827	0.1201	81.32	0	2.9526
SERV	-0.9384	0.1105	72.12	0	0.3913
SANPGP	0.4682	0.1753	7.14	0.0076	1.5971
CURPRO	0.3864	0.1151	11.28	0.0008	1.4717
FX_COR	-0.7702	0.131	34.54	0	0.4629
SEXO	0.7052	0.1453	23.54	0	2.0243
CNECMA	0.4895	0.1347	13.2	0.0003	1.6315
CHEFE	0.4616	0.1356	11.58	0.0007	1.5866
CLOBO91	0.269	0.1073	6.28	0.0122	1.3087
Grupo_B=					
INTERCEP	-0.4176	0.056	55.69	0	1.5183
SINDA	-1.3904	0.0475	856.81	0	4.0165
GRAU2	-0.0873	0.0598	2.13	0.1443	1.0912
SERV	-0.9782	0.0396	609.89	0	2.6597
SANPGP	-0.325	0.0428	57.71	0	1.3840
CURPRO	-0.2573	0.0512	25.26	0	1.2934
FX_COR	0.0848	0.0388	4.78	0.0289	0.9187
SEXO	0.4221	0.0473	79.53	0	0.6557
CNECMA	0.00242	0.0402	0	0.9521	0.9976
CHEFE	0.0449	0.0448	1	0.3163	0.9561
CLOBO91	0.0891	0.0398	5.01	0.0252	0.9148

Fonte: PME/IBGE

Para selecionar o modelo utilizamos a PROC CATMOD do SAS (maiores detalhes ver www.sas.com). Os modelos finais foram selecionados passo a passo, após agrupamento de níveis dos fatores com base na estatística de Wald, incluindo-se em cada passo as interações que produziam maior decréscimo da Deviance, considerando o teste da razão. Apresentaremos todos os modelos selecionados Em seguida as análises das estimativas para os grupos: A, B e C.

Interpretação dos resultados do GRUPO_B

A partir dos resultados da tabela 6.6 encontramos, por exemplo, que a vantagem, na ocorrência do evento de um o afro-brasileiro transitar do grupo dos empregados com carteira assinada e funcionários públicos em 1991, para empregadores ou trabalhadores autônomos em 1996 é 0,4629 vezes (*coluna Vantagem*) ou 53% (*Exp(0,4629) – 1 x 100*) menor do que o grupo formado pelos brancos e amarelos. Com isso concluímos que a mobilidade ocupacional ascendente dos afro-brasileiros sofre uma espécie de discriminação em nosso mercado de trabalho.