AULA EXTRA
Análise de Regressão Logística

Ernesto F. L. Amaral

Avaliação de Políticas Públicas (DCP 046)

VARIÁVEL DEPENDENTE BINÁRIA

– O modelo de regressão logístico é utilizado quando a variável resposta é qualitativa com dois resultados possíveis.

– Probabilidade de sucesso = p

– Probabilidade de fracasso = 1 – p = q

– Chance = (prob. de sucesso) / (prob. de fracasso)

– Por exemplo, se a probabilidade de sucesso é 0,75, a chance é igual a:

$$p / (1 - p) = p / q = 0,75 / 0,25 = 3$$

RAZÃO DE CHANCES

– Razão de chances para variáveis dependentes binárias é a razão entre a chance de uma linha (ou coluna) de uma tabela 2x2, dividida pela chance da outra linha (ou coluna):

$$\frac{A}{B} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1 * (1-p_2)}{p_2 * (1-p_1)}$$

EXEMPLO DE CÁLCULO DE RAZÃO DE CHANCES

Sexo	Dilma	Serra	Total
Homem	52	39	91
Mulher	43	44	87
Total	**95**	**83**	**178**

- **Chance** de votar na Dilma entre homens:

$$p_1 / (1-p_1) = (52/91) / (39/91) = 0,57 / 0,43 = 1,33$$

- **Chance** de votar na Dilma entre mulheres:

$$p_2 / (1-p_2) = (43/87) / (44/87) = 0,49 / 0,51 = 0,96$$

- **Razão de chances** de votar na Dilma entre homens, em relação às mulheres:

$$[p_1 /(1- p_1)] / [p_2 /(1- p_2)] = 1,33 / 0,96 = 1,39$$

FUNÇÃO DE RESPOSTA
QUANTO VARIÁVEL DEPENDENTE É BINÁRIA

– Vamos considerar o modelo de regressão linear simples:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$Y_i = \begin{cases} 1 \\ 0 \end{cases}$$

– A resposta esperada é dada por:

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

– Na regressão logística, Y_i possui uma distribuição de probabilidade:

$$Y_i = 1 \rightarrow P(Y_i = 1) = \pi_i$$

$$Y_i = 0 \rightarrow P(Y_i = 0) = 1 - \pi_i$$

LOGITO

– O logito (*logit*) equivale ao logaritmo natural (base *e*) da chance:

$$logit(p) = log\left(\frac{p}{1-p}\right) = log(p) - log(1-p)$$

– A função logística é dada pelo logito-inverso (anti-logit) que nos permite transformar o logito em probabilidade:

$$p = \frac{exp(x)}{1 + exp(x)}$$

RAZÃO DE CHANCES (*ODDS RATIO*)

– Compara a chance de sucesso de um grupo em relação a outro grupo:

$$log(R) = log\left(\frac{p_1/(1-p_1)}{p_2/(1-p_2)}\right)$$

$$log(R) = log\left(\frac{p_1}{1-p_1}\right) - log\left(\frac{p_2}{1-p_2}\right)$$

$$log(R) = logit(p_1) - logit(p_2)$$

– Portanto, a diferença entre os logitos de duas probabilidades equivale ao logaritmo da razão de chances.

RAZÃO DE CHANCES (*ODDS RATIO*)

$$\frac{A}{B} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{exp(\beta_0 + \gamma)}{exp(\beta_0)} = \frac{exp(\beta_0) * exp(\gamma)}{exp(\beta_0)} = exp(\gamma)$$

– Razão de chance é dada pela expressão exp(γ): chance de sucesso no grupo A, em relação ao grupo B.

– Se exp(γ) for maior que uma unidade, chance de sucesso em A é maior que em B.
  – Ex.: exp(γ)=1,17, chance de sucesso em A é 1,17 vezes maior do que em B, ou seja, é 17% maior do que em B.

– Se exp(γ) for menor que uma unidade, chance de sucesso em A é menor que em B.
  – Ex.: exp(γ)=0,61, chance de sucesso em A é 0,61 vezes a chance de B, ou seja, é 39% menor do que em B.

DEFINIÇÃO DO VALOR ESPERADO

– Pela definição de valor esperado, obtemos:

$$E(Y_i) = \beta_0 + \beta_1 X_i = \pi_i$$

– Assim, a resposta média, quando a variável resposta é uma variável binária (1 ou 0), representa a probabilidade de $Y = 1$, para o nível da variável independente X_i.

REGRESSÃO LOGÍSTICA
COM UMA VARIÁVEL INDEPENDENTE

– Considerações teóricas e práticas sugerem que quando a variável resposta é binária, a forma da função resposta será frequentemente curvilínea.

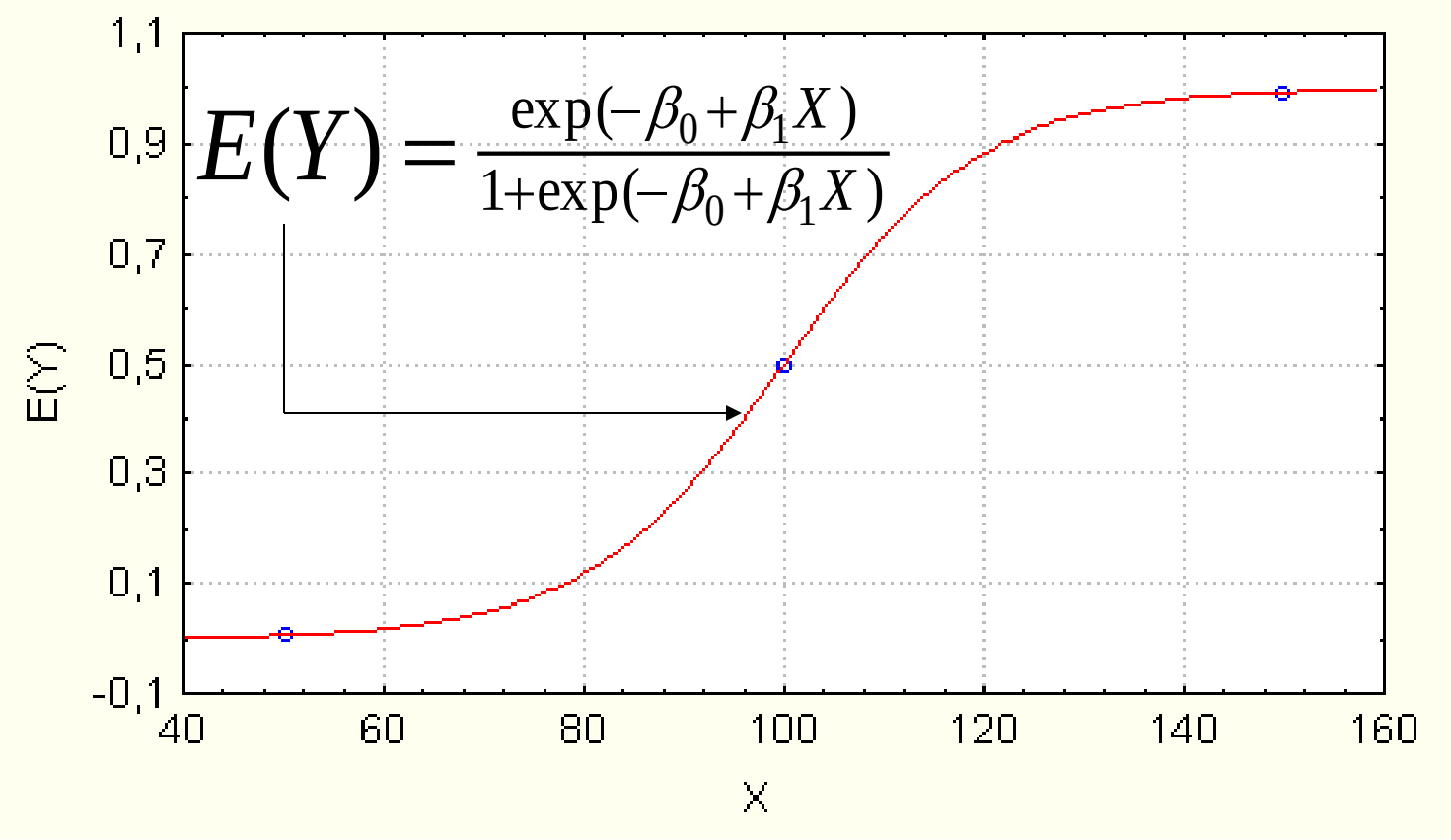
– As funções respostas (valores preditos) das figuras são denominadas funções logísticas, cuja expressão é:

$$E(Y) = \frac{\exp(\beta_0 + \beta_1 X)}{1+\exp(\beta_0 + \beta_1 X)}$$

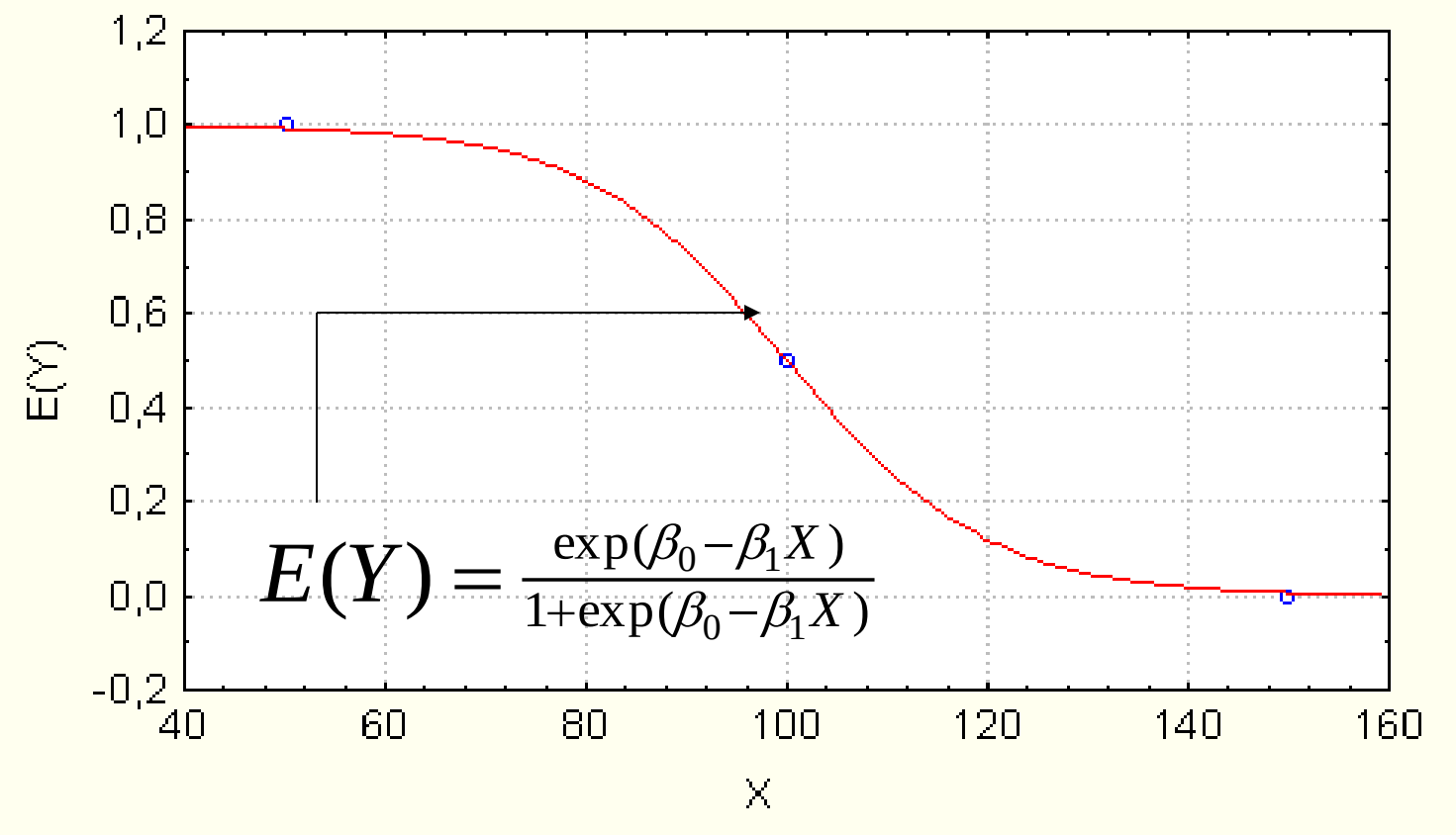
– Forma equivalente:

$$E(Y) = \left[1 + \exp\left(-\beta_0 - \beta_1 X\right)\right]^{-1}$$

VARIÁVEL DEPENDENTE ESTIMADA PELA VARIÁVEL INDEPENDENTE OBSERVADA



VARIÁVEL DEPENDENTE ESTIMADA PELA VARIÁVEL INDEPENDENTE OBSERVADA



REGRESSÃO LOGÍSTICA
COM MAIS DE UMA VARIÁVEL INDEPENDENTE

– Função com uma variável independente:

$$E(Y) = \frac{\exp(\beta_0 + \beta_1 X)}{1+\exp(\beta_0 + \beta_1 X)}$$

– Função com uma série de variáveis independentes:

$$E(Y) = \frac{\exp(\boldsymbol{\beta}'\mathbf{X})}{1+\exp(\boldsymbol{\beta}'\mathbf{X})}$$

– Uma forma equivalente é dada por:

$$E(Y) = (1+\exp(-\boldsymbol{\beta}'\mathbf{X}))^{-1}$$

EQUAÇÃO DE REGRESSÃO

– A parte linear da equação da regressão logística é usada para encontrar a probabilidade de estar em uma categoria, baseado na combinação de variáveis independentes.

$$\bar{Y}_i = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3}}$$

– Os coeficientes de regressão e seus erros padrões são estimados com métodos de máxima verossimilhança.

AJUSTANDO O MODELO

– A função log-verossimilhança estende-se diretamente para o modelo de regressão logística múltipla, dada por:

$$\log_e L(\beta) = \sum_{i=1}^{n} Y_i(\boldsymbol{\beta}'\mathbf{X}_i) - \sum_{i=1}^{n} \log_e(1+\exp(\boldsymbol{\beta}'\mathbf{X}_i))$$

– Métodos numéricos devem ser utilizados para encontrar os valores de $\beta_0, \beta_1,\ldots, \beta_{p-1}$ para maximizar a expressão.

– As estimativas de máxima verossimilhança serão denotadas por $b_0, b_1,\ldots,b_{p-1}$.

– A função resposta logística ajustada e os valores ajustados são dados por:

$$\hat{\pi} = \frac{\exp(\mathbf{b}'\mathbf{X})}{1+\exp(\mathbf{b}'\mathbf{X})} = (1+\exp(-\mathbf{b}'\mathbf{X}))^{-1}$$

$$\hat{\pi}_i = \frac{\exp(\mathbf{b}'\mathbf{X}_i)}{1+\exp(\mathbf{b}'\mathbf{X}_i)} = (1+\exp(-\mathbf{b}'\mathbf{X}_i))^{-1}$$

ESTIMADORES DE MÁXIMA VEROSSIMILHANÇA

– Não existe uma solução analítica para os valores β_0 e β_1 que maximizam a função de verossimilhança.

– Métodos numéricos são necessários para encontrar as estimativas de máxima verossimilhança, b_0 e b_1.

– Encontradas as estimativas b_0 e b_1, substitui-se esses valores para encontrar os valores ajustados.

– O valor ajustado para o *i-ésimo* valor é dado por:

$$\hat{\pi}_i = \frac{\exp(b_0 + b_1 X_i)}{1 + \exp(b_0 + b_1 X_i)}$$

– Se usarmos a transformação *logit*, a função é:

$$\hat{\pi} = \frac{\exp(b_0 + b_1 X)}{1 + \exp(b_0 + b_1 X)}$$

– A função de resposta ajustada é dada por:

$$\hat{\pi}' = b_0 + b_1 X \quad \text{onde:} \quad \hat{\pi}' = \log_e\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)$$

TESTE DE QUI-QUADRADO DA RAZÃO DE VEROSSIMILHANÇA

- Logaritmo da verossimilhança (*Log-likelihood*):

$$\log - likelihood = \sum_{i=1}^{N} [Y_i \ln(\widehat{Y}_i) + (1 - Y_i) \ln(1 - \widehat{Y}_i)]$$

- Modelos são comparados com uso dos logaritmos das verossimilhanças dos modelos:

X^2 = – 2 [(*log-likelihood* do modelo restrito)

– (*log likelihood* do modelo irrestrito)]

ou

X^2 = 2 [(*log-likelihood* do modelo irrestrito)

– (*log likelihood* do modelo restrito)]

- Modelos precisam ser aninhados para comparação, ou seja, todas variáveis independentes do menor modelo (restrito) devem estar incluídas no maior modelo (irrestrito).

MAIS TESTE DE QUI-QUADRADO

– O teste de qui-quadrado da razão da verossimilhança é igual ao ajuste do modelo restrito (–2*log. da verossimilhança do modelo anterior) menos o ajuste do modelo irrestrito (–2*log. da versossimilhança do modelo atual).

– O logaritmo da verossimilhança multiplicado por –2 é usado para testar hipóteses entre modelos aninhados, sendo que seu valor não tem um significado específico.

– Esta razão é testada em uma distribuição de qui-quadrado, levando em consideração a diferença entre os graus de liberdade (número de variáveis independentes do modelo irrestrito menos o número de variáveis independentes do modelo restrito).

– Se o teste de qui-quadrado é significante, é afirmado que o modelo irrestrito não pode ter redução de variáveis independentes, dado um nível de significância específico.

TESTE DE WALD

- Cada coeficiente é avaliado usando o teste de Wald, que é simplesmente um teste de escore z:

$$W_j = \frac{\beta_j}{EP_{\beta_j}}$$

- Os testes dos coeficientes são aproximadamente escores z, os quais são posteriormente elevados ao quadrado, fazendo com que esta estatística tenha distribuição de qui-quadrado.
- Esse teste é usado para avaliar a significância de cada coeficiente (β) no modelo.
- O teste de Wald é conhecido por ser conservador (aumenta o erro II).

ERROS TIPO I E TIPO II

- Ao testar H_0, chegamos a uma conclusão de rejeitá-la ou de deixar de rejeitá-la.
- Tais conclusões pode estar corretas ou erradas.

		Estado verdadeiro da natureza	
		A hipótese nula é verdadeira	A hipótese nula é falsa
Decisão	Decidimos rejeitar a hipótese nula.	**Erro tipo I** (rejeitar uma hipótese nula verdadeira) α	Decisão Correta
	Deixamos de rejeitar a hipótese nula	Decisão Correta	**Erro tipo II** (deixar de rejeitar uma hipótese nula falsa) β

- α: probabilidade de erro tipo I (probabilidade de rejeitar hipótese nula quando ela é verdadeira).
- β: probabilidade de erro tipo II (probabilidade de deixar de rejeitar hipótese nula quando ela é falsa).

PSEUDO R^2

- Há várias medidas de associação que pretendem servir como um R^2 na regressão logística.

- Porém, nenhuma destas medidas é realmente o R^2.

- A interpretação não é a mesma, mas eles podem ser vistos como uma aproximação da variação na variável dependente, devido à variação nas variáveis independentes.

- Para comparação de grau de ajuste entre modelos é mais apropriado fazer o teste de qui-quadrado da razão da verossimilhança.

MODELO LOGÍSTICO MULTINOMIAL

- É possível estimar uma regressão logística em que a variável dependente tem mais de duas categorias.

- Ou seja, o modelo logístico pode ser estendido quando a variável resposta qualitativa tem mais do que duas categorias.

- Por exemplo, posicionamento ideológico: esquerda, centro, direita.

- São geradas k – 1 equações, sendo k o número de categorias.

- As equações geram probabilidades para predizer se uma categoria está acima/abaixo da categoria de referência.

EXEMPLO DE MODELO LOGÍSTICO

IMPACTO DO BOLSA FAMÍLIA SOBRE ABANDONO ESCOLAR

– Banco de dados de Avaliação de Impacto do Programa Bolsa Família (AIBF) de 2005 do Ministério do Desenvolvimento Social e Combate à Fome (MDS).

– Modelos logísticos foram estimados para três grupos de domicílios, segundo limites máximos da renda domiciliar per capita:

1) R$50,00: população com piores condições sócio-econômicas.

2) R$100,00: limite oficial de renda definido para elegibilidade ao PBF.

3) R$200,00: garante representatividade amostral em todos grupos.

VARIÁVEL DEPENDENTE

- Variável dependente indica se a criança abandonou a escola entre 2004 e 2005:
  - No ano passado, frequentava escola ou creche?
  - Frequenta escola ou creche atualmente?
- Foi realizada análise multivariada, controlando as estimativas por características do domicílio, mãe e criança.

VARIÁVEIS INDEPENDENTES DE DOMICÍLIO

- Número de membros da família.
- Presença de idosos.
- Presença de rede geral de água.
- Iluminação elétrica.
- Serviço de coleta de lixo.
- Domicílio em zona urbana ou rural.
- Região de residência (Sul/Sudeste; Norte/Centro-Oeste; Nordeste).

VARIÁVEIS INDEPENDENTES DA MÃE

- Indicação se mãe é chefe do domicílio.
- Cor/raça.
- Anos de escolaridade.
- Idade.
- Residia há menos de 10 anos no município.
- Participação em organizações sociais.
- Horas de trabalho por semana.
- Tempo gasto em cuidados com a casa por dia.

DEMAIS VARIÁVEIS INDEPENDENTES

Variáveis independentes da criança:

– Idade da criança.

– Indicação se criança trabalha.

– Mãe reside no domicílio.

Beneficiário do Programa Bolsa Família:

– Indicação se criança reside em domicílio que recebe o benefício.

DESCRIÇÃO DA AMOSTRA

– Distribuição percentual de crianças por grupos de renda domiciliar per capita e recebimento do benefício.

Programa Bolsa Família	**Limite de renda domiciliar per capita**		
	R$50,00	**R$100,00**	**R$200,00**
Sim	68,39%	64,71%	59,75%
Não	31,61%	35,29%	40,25%
Nº casos (n)	3.312	6.761	9.232

Fonte: AIBF/MDS (2005).

DISTRIBUIÇÃO DA VARIÁVEL DEPENDENTE

– Percentual de crianças que abandonaram a escola entre 2004 e 2005 por grupo de renda e recebimento do benefício.

Programa Bolsa Família	**Limite de renda domiciliar per capita**		
	R$50,00	**R$100,00**	**R$200,00**
Sim	1,10%	1,42%	1,30%
Não	2,39%	1,97%	1,80%
Diferença	1,28%***	0,55%***	0,50***

***Significativo ao nível de confiança de 99%.
Fonte: AIBF/MDS (2005).

RAZÕES DE CHANCES DA CRIANÇA TER ABANDONADO A ESCOLA ENTRE 2004 E 2005

Variáveis independentes	R$50,00	R$100,00	R$200,00
Variáveis de domicílio			
Nº de membros da família	1,122	1,124***	1,108***
Idosos no domicílio	1,454	1,678	1,331
Rede de água	1,066	0,767	0,694*
Iluminação elétrica	1,270	1,106	1,293
Coleta de lixo	0,994	0,756	0,621**
Rural	ref.	ref.	ref.
Urbano	1,729	1,910*	2,309***
Sul/Sudeste	ref.	ref.	ref.
Norte/Centro-Oeste	2,536**	1,889**	1,630**
Nordeste	3,035**	2,248***	2,064***

RAZÕES DE CHANCES DA CRIANÇA TER ABANDONADO A ESCOLA ENTRE 2004 E 2005 (cont.)

Variáveis independentes	R$50,00	R$100,00	R$200,00
Variáveis da mãe			
Mãe é chefe do domicílio	1,974***	1,445*	1,508**
Preta/Parda	ref.	ref.	ref.
Branca	2,248**	2,029***	1,465**
0 anos de estudo	ref.	ref.	ref.
1-4 anos de estudo	1,267	1,195	1,135
5-8 anos de estudo	0,701	0,898	0,902
9+ anos de estudo	0,251*	0,440*	0,481*
0-24 anos	1,507	4,757***	4,534***
25-34 anos	ref.	ref.	ref.
35-49 anos	1,170	1,111	1,109
50+ anos	0,053***	0,532	0,645

RAZÕES DE CHANCES DA CRIANÇA TER ABANDONADO A ESCOLA ENTRE 2004 E 2005 (cont.)

Variáveis independentes	R$50,00	R$100,00	R$200,00
Variáveis da mãe			
<10 anos no município	1,325	1,411	1,838***
Participa org. social	0,731	0,643*	0,565***
0 hora/semana trabalho	ref.	ref.	ref.
1-20 horas/semana trabalho	0,257*	0,920	1,177
21-39 horas/semana trabalho	0,736	0,744	0,907
40+ horas/semana trabalho	0,904	1,790**	1,529*
0-2 hora/dia trab. casa	ref.	ref.	ref.
3-4 hora/dia trab. casa	2,975	1,089	0,854
5-6 hora/dia trab. casa	2,399	1,241	1,050
7+ hora/dia trab. casa	2,084	1,563	1,443

RAZÕES DE CHANCES DA CRIANÇA TER ABANDONADO A ESCOLA ENTRE 2004 E 2005 (cont.)

Variáveis independentes	R$50,00	R$100,00	R$200,00
Variáveis da criança			
Idade	1,174**	1,226***	1,194***
Criança trabalha	1,417	1,177	1,465
Mãe reside no domicílio	0,218***	0,455**	0,610*
Beneficiário do Programa Bolsa Família	**0,428*** **	**0,662** **	**0,666** **
Número de casos (crianças)	3.312	6.761	9.232

*Significativo ao nível de 90%; **Significativo ao nível de 95%; ***Significativo ao nível de 99%.
Fonte: AIBF/MDS (2005).