DOI: http://dx.doi.org/10.20396/san.v23i1.8647217

SEGURANÇA
alimentar e nutricional

Odds Ratio *versus* Razão de Prevalências ou Modelo de Lógite *versus* Regressão de Poisson

Rodolfo Hoffmann[1]

São analisadas as limitações e as vantagens do uso da Regressão de Poisson, em comparação com o modelo de lógite, quando o desfecho é uma variável binária. São discutidos exemplos nos quais todas as variáveis são binárias e também um exemplo com variável explanatória contínua. O modelo deve ser escolhido considerando as características do fenômeno analisado. A preferência pelo uso da razão de prevalências, e não da odds ratio, na análise dos resultados, não deve ser motivo decisivo na escolha do modelo de relação entre as variáveis.

Palavras-chave: odds ratio, razão de prevalências, modelo de lógite, regressão de Poisson.

Odds Ratio versus Prevalence Ratio or logistic versus Poisson regression

The limitations and advantages of using the Poisson Regression are analyzed, in comparison with a logistic regression, when the outcome is binary. Examples where all variables are binary and also an example with a continuous explanatory variable are discussed. The proper model should be chosen considering the characteristics of the problem being analyzed. The preference for the prevalence ratio rather than the odds ratio should not be decisive in choosing the model of the variables relationship.

Keywords: odds ratio, prevalence ratio, logit, Poisson regression.

INTRODUÇÃO

A odds ratio (OR) e a razão de prevalências (RP) são duas das várias maneiras de medir a associação entre duas variáveis quando a variável dependente é dicotômica (binária). Trata-se de duas medidas diferentes e, como será discutido adiante, não tem sentido dizer que a estimativa de uma é mais ou menos precisa que a estimativa da outra simplesmente comparando a amplitude dos respectivos intervalos de confiança.

Em um modelo de lógite, se determinada variável explanatória (um fator de risco) for uma variável contínua e aparecer em um único termo do modelo, a respectiva OR é determinada pelo valor do coeficiente dessa variável, sendo constante ao longo de toda a curva que relaciona a variável com o desfecho.

[1] Professor Sênior da ESALQ-USP com apoio do CNPq. O autor agradece a Angela Kageyama, Ana Lúcia Kassouf e Josimar Gonçalves de Jesus pelas críticas e sugestões a uma versão preliminar do artigo. Endereço para correspondência: Departamento de Economia, Administração e Sociologia – Escola Superior de Agricultura "Luiz de Queiroz" – Caixa Postal 09 – CEP 13418-900 – Piracicaba – SP – Brasil. *E-mail*: hoffmannr@usp.br

Por outro lado, a respectiva RP é diferente em cada ponto da curva.

Quando se utiliza a regressão de Poisson para estimar a mesma relação ocorre o oposto: a RP é determinada por um parâmetro, sendo a mesma em todos os pontos da curva, e a OR varia de ponto para ponto.

Difundiu-se a ideia de que, se o pesquisador está interessado em estimar a RP, é mais conveniente utilizar a regressão de Poisson, e não um modelo de lógite. Mas essa é uma maneira inapropriada de identificar o modelo a ser usado. Deve-se adotar um modelo que respeite as características essenciais do fenômeno analisado e que se ajuste bem aos dados observados. Se, por exemplo, houver razões para acreditar que a RP não é constante para o intervalo de valores analisado, não é apropriado usar um modelo que "impõe" uma RP constante. O cálculo de valores da OR e da RP pode ser feito com relativa facilidade tanto a partir de um modelo de lógite estimado como a partir de uma regressão de Poisson.

O modelo de lógite

A curva logística foi criada no século XIX para descrever o crescimento de populações e o desenvolvimento de reações químicas autocatalíticas, cabendo destacar os trabalhos de Pierre-François Verhulst publicados de 1838 a 1847 [1].

Sendo X o tempo e W a população, a função logística, com parâmetros α, β e θ, é

$$W = \frac{\theta}{1 + e^{-(\alpha+\beta X)}}$$

ou, definindo $Q = \alpha + \beta X$,

$$W = \frac{\theta}{1 + e^{-Q}} = \frac{\theta e^{Q}}{1 + e^{Q}} \qquad (1)$$

Com $\theta > 0$ e $\beta > 0$, a representação gráfica da variação de W em função de X é uma curva sigmoide que cresce continuamente entre duas assíntotas horizontais: o eixo das abscissas e uma linha paralela com ordenada θ.

Pode-se verificar que

$$\frac{dW}{dX} = \frac{\beta}{\theta} W(\theta - W),$$

mostrando que o ritmo de crescimento é proporcional ao próprio valor da função (W) e ao valor que falta para atingir o valor de saturação θ. Enquanto $W < \theta/2$, o ritmo de crescimento é crescente e a curva é convexa (vista de baixo). Para $W > \theta/2$ o ritmo de crescimento vai diminuindo e a curva é côncava. O ponto de inflexão ocorre quando

$$X = -\frac{\alpha}{\beta} \quad \text{e} \quad W = \frac{\theta}{2}$$

A curva logística tem ampla aplicação na análise do crescimento de seres vivos.

Se a variável dependente for uma probabilidade ou uma proporção, a assíntota superior é igual a 1 e a expressão da curva logística fica

$$P = \frac{1}{1 + e^{-Q}} = \frac{e^{Q}}{1 + e^{Q}} \qquad (2)$$

com

$$Q = \alpha + \beta X \qquad (3)$$

Entre 1944 e 1980 Joseph Berkson escreveu grande número de trabalhos defendendo o uso dessa função na análise estatística de ensaios biológicos como, por exemplo, a variação da proporção de insetos mortos em função da dose de um inseticida. A partir de (2) pode-se obter

$$\ln \frac{P}{1 - P} = Q, \qquad (4)$$

que Berkson denominou de *logit*, por analogia com o *probit*[2], baseado na função de distribuição normal e que

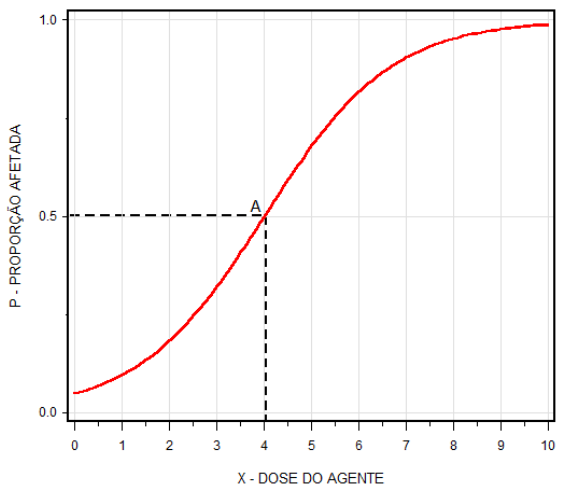
[2] De acordo com o Dicionário Brasileiro de Estatística [2], os termos para probit e logit em português são próbite e lógite.

era a técnica mais usual na análise estatística de ensaios biológicos [1].

No caso de ensaios para avaliar a toxicidade de uma substância usa-se a função (2) estimada para calcular a dose letal mediana. Sendo $\hat{\alpha}$ e $\hat{\beta}$ as estimativas dos parâmetros α e β, a dose letal mediana é dada por $-\hat{\alpha}/\hat{\beta}$.

A Figura 1 mostra o gráfico da função (2) com $Q = -3 + 0{,}75X$, destacando o ponto de inflexão (A), cuja abscissa é a dose letal mediana.

Figura 1. A função logística para uma proporção.



O uso do lógite se estendeu a muitos outros campos. A função é muito utilizada para analisar, por exemplo, o crescimento, no tempo, da proporção de produtores que adotam uma nova técnica ou como a renda afeta a proporção de famílias que possuem determinado bem de consumo durável.

Não há dificuldade de introduzir, no modelo de lógite, mais de uma variável explanatória, substituindo a expressão (3) por

$$Q = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_h X_h \qquad (5)$$

É importante observar que nas expressões (2) ou (4) o valor de *P* fica entre zero e 1 qualquer que seja o valor de Q. Com o modelo de lógite os resultados nunca estarão em conflito com os limites de variação de uma proporção (ou de uma probabilidade). O mesmo é válido para o próbite e para a transformação log-log complementar

$$Q = \ln[-\ln(1 - P)] \qquad (6)$$

Em qualquer dos três casos, a estimação dos parâmetros é feita preferencialmente pelo método da máxima verossimilhança, tendo em vista suas propriedades estatísticas. A variável dependente pode ser o resultado individual ($Y = 0$ ou $Y = 1$) ou a proporção de resultados favoráveis para cada combinação de valores das variáveis explanatórias.

Se, em um modelo de lógite, P_0 indicar o valor da probabilidade de ocorrência de certo desfecho para determinados valores das variáveis explanatórias e P_1 indicar o valor dessa probabilidade quando a variável explanatória X_j aumenta de uma unidade, verifica-se que

$$\text{OR} = \frac{\frac{P_1}{1-P_1}}{\frac{P_0}{1-P_0}} = e^{\beta_j} \qquad (7)$$

ou seja, a OR associada ao aumento de uma unidade em X_j depende apenas do parâmetro β_j. Se X_j for uma variável binária indicando a ausência ou presença de certo fator, e^{β_j} é a OR associada à presença do fator. A RP, dada por P_1/P_0, por outro lado, depende do valor inicial de X_j e do valor de todas as outras variáveis explanatórias do modelo.

A regressão de Poisson

Se, em um conjunto de *n* indivíduos, cada um pode ou não apresentar determinado desfecho, independentemente, com probabilidade *P*, o número de indivíduos *Y* que apresenta o desfecho tem distribuição binomial com parâmetros *n* e *P*, com média $\mu = nP$ e variância $\sigma^2 = nP(1 - P) = \mu(1 - P)$.

A distribuição de Poisson pode ser definida como o resultado obtido quando, em uma distribuição binomial, n cresce ilimitadamente $(n \to \infty)$, mantendo μ fixo. Com $\mu = np$ fixo e $n \to \infty$, P tende a zero e $1 - P$ tende a 1. Assim, no limite temos uma distribuição com $\mu = \sigma^2$, isto é, com variância igual a sua média. A distribuição de Poisson é apropriada para analisar a contagem $(Y = 0, 1, 2, \ldots)$ de eventos raros.

Um exemplo simples seria o número de clientes que entram em determinada loja por minuto. Note-se que nesse contexto nem é possível definir uma proporção de desfechos favoráveis ou não.

Na regressão de Poisson admite-se, usualmente, que o logaritmo da média da distribuição ($\ln\mu$) seja função de uma ou mais variáveis explanatórias.

Se usarmos a regressão de Poisson para modelar resultados de uma distribuição binomial, com $Y = 0$ ou $Y = 1$, estamos implicitamente admitindo que se trate de uma distribuição de Poisson na qual a probabilidade de resultado $Y \geq 2$ possa ser considerada desprezível. A rigor, é um uso inapropriado do modelo. O uso da regressão de Poisson para analisar resultados em que o desfecho é, por definição, uma variável binária ($Y = 0$ ou $Y = 1$) pode ser válido em casos especiais, mas é perigoso generalizar tal procedimento, como será visto adiante. Para uma variável binária com valores 0 e 1, a média é igual à proporção de valores iguais a 1 e, nesse caso o modelo da regressão de Poisson fica

$$\ln P = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_h X_h$$

Para esse caso especial de aplicação da regressão de Poisson é verdadeira uma expressão semelhante a (7), substituindo-se a OR pela RP:

$$\text{RP} = \frac{P_1}{P_0} = e^{\beta_j} \qquad (8)$$

Essa expressão mostra que a RP associada ao aumento de uma unidade em X_j depende apenas do parâmetro β_j. Tendo em vista esse fato, os defensores da regressão de Poisson argumentam que seu uso é preferível ao uso de um modelo de lógite se a estimação da RP é o principal objetivo da pesquisa.

No modelo de regressão de Poisson para resultados dicotômicos, a RP é determinada pelo valor de β e independe dos valores assumidos pelas variáveis explanatórias. A OR irá variar com esses valores. No modelo de lógite, por outro lado, é a OR que é constante e a RP varia com os valores assumidos pelas variáveis explanatórias.

Primeiro exemplo

Para discutir a conveniência de usar a regressão de Poisson, e não um modelo de lógite, serão analisados alguns exemplos. O primeiro é um exemplo artificial de Zou [3], que publicou em 2004 artigo defendendo o uso da regressão de Poisson para resposta binária[3].

Como mostra a Tabela 1, há dois estratos, cada um com 200 observações. Uma variável binária X_1 é utilizada para distinguir os dois estratos, com $X_1 = 0$ no estrato I e $X_1 = 1$ no estrato II. Outra variável binária indica se o indivíduo é $(X_2 = 1)$ ou não é $(X_2 = 0)$ exposto a determinado fator de risco. O desfecho (Y) também é uma variável binária.

Tabela 1. O exemplo artificial de Zou [3].

Estrato	Exposição a fator de risco	Número (n)	Número com		Prevalência (m/n)
			$Y = 1$ (m)	$Y = 0$	
I ($X_1 = 0$)	Sim ($X_2 = 1$)	140	112	28	0,8
	Não ($X_2 = 0$)	60	24	36	0,4
II ($X_1 = 1$)	Sim ($X_2 = 1$)	60	6	54	0,1
	Não ($X_2 = 0$)	140	7	133	0,05

[3] A fórmula para variância da estimativa do risco relativo apresentada na página 703 do artigo de Zou [3] é, na realidade, a expressão para a variância da estimativa de β. Tudo indica que esse artigo foi desenvolvido independentemente do trabalho de Barros e Hirakata [4], que também defendem o uso da regressão de Poisson com estimativas de variância robustas.

DOI: http://dx.doi.org/10.20396/san.v23i1.8647217

Tanto para indivíduos expostos como para indivíduos não expostos ao fator de risco, a proporção com desfecho $Y = 1$ no estrato I é 8 vezes maior do que no estrato II, fazendo com que a RP seja exatamente a mesma (RP = 2) nos dois estratos.

Como a RP é a mesma nos dois estratos, a regressão de Poisson, com X_1 e X_2 como variáveis explanatórias estima exatamente as prevalências observadas nas 4 situações.

As estimativas dos coeficientes de X_1 e X_2 (com os correspondentes intervalos de 95% de confiança entre parênteses) são –2,0794 (–2,6270 a –1,5318) e 0,6931 (0,3806 a 1,0057), respectivamente[4].

Ajustando aos dados um modelo de lógite com X_1 e X_2 como variáveis explanatórias, as prevalências estimadas são diferentes. Considerando a ordem em que as prevalências observadas são apresentadas na Tabela 1, os valores estimados são 0,7835 , 0,4385 , 0,1385 e 0,0335. Notando que a OR no estrato I é muito maior do que no estrato II (6 e 2,111), cabe introduzir no modelo de lógite um termo de interação X_1X_2. Dessa maneira o modelo de lógite produz estimativas das prevalências nas quatro situações que são idênticas às prevalências observadas.

No que se refere à reprodução das RP observadas, a regressão de Poisson e o modelo de lógite com o termo de interação X_1X_2 se mostram igualmente satisfatórios. A regressão de Poisson é mais simples, pois não é necessário incluir o termo de interação, mas isso se deve a uma característica especial desse exemplo numérico artificial. E deve-se ter em mente que o uso da regressão de Poisson implica fazer de conta que a distribuição de Poisson só produz defechos $Y = 0$ ou $Y = 1$.

Voltaremos a discutir essa questão depois de apresentar outro exemplo artificial.

Segundo exemplo

A partir da Tabela 1, alterando apenas o número de casos com desfechos $Y = 1$ e $Y = 0$, são obtidos os dados artificiais apresentados na Tabela 2.

Tabela 2. Segundo exemplo artificial.

Estrato	Exposição a fator de risco	Número (n)	Número com		Prevalência (m/n)
			$Y = 1$ (m)	$Y = 0$	
I ($X_1 = 0$)	Sim ($X_2 = 1$)	140	126	14	0,9
	Não ($X_2 = 0$)	60	30	30	0,5
II ($X_1 = 1$)	Sim ($X_2 = 1$)	60	30	30	0,5
	Não ($X_2 = 0$)	140	14	126	0,1

Agora a RP é bastante diferente nos dois estratos: é igual a 1,8 no estrato I e igual a 5,0 no estrato II. E os números foram escolhidos de maneira que a OR fosse a mesma nos dois estratos (OR = 9).

Graças à constância da OR, um modelo de lógite com apenas X_1 e X_2 como variáveis explanatórias produz estimativas das prevalências nas 4 situações que são idênticas às observadas. Entretanto, a regressão de Poisson sem termo de interação produz estimativas das prevalências bem diferentes. Considerando a ordem em que as quatro situações são apresentadas na Tabela 2, os valores estimados são 0,9515, 0,3797 e 0,1515. Basta introduzir a interação X_1X_2 como variável explanatória para que a regressão de Poisson produza estimativas das prevalências idênticas aos valores observados. A estimativa do coeficiente de X_1X_2 é 1,0217. O teste Z para a hipótese de nulidade desse coeficiente, com base em estimativa robusta da variância, é igual a 3,26, significativo ao nível de 1%, indicando que o modelo de

[4] São resultados obtidos utilizando o PROC GENMOD do SAS, com variâncias robustas obtidas por meio do comando REPEATED, como indicado por Zou [3] e Spiegelman e Hertzmark [5].

regressão de Poisson sem o termo de interação está claramente mal especificado.

Neste caso o modelo de lógite é o mais simples, dispensando a inclusão do termo de interação.

A regressão de Poisson sem o termo de interação produz uma estimativa da RP associada a X_2, com controle de X_1, igual a 2,5058, com intervalo de 95% de confiança de 1,8924 a 3,3180 (usando estimativa robusta da variância). Mas isso é uma descrição inapropriada do fenômeno, pois o efeito de X_2, medido por meio da RP, é *diferente* nos dois estratos: RP = 1,8 no estrato I e RP = 5 no estrato II, como comprova, apropriadamente, a regressão com interação, com todos os coeficientes fortemente significativos. E é interessante notar que, quando avaliado por meio da OR o efeito de X_2 é o mesmo nos dois estratos.

Um exemplo com variável explanatória contínua

Na Pesquisa Nacional por Amostra de Domicílios (PNAD) de 2013 o IBGE incluiu, pela terceira vez, a avaliação da Segurança Alimentar. Utilizando os microdados, agrupamos as pessoas em 9 estratos de renda domiciliar per capita (RDPC), como mostra a Tabela 3, separando os que foram classificados com tendo segurança alimentar $(Y = 1)$ dos com algum grau de insegurança alimentar (leve, moderada ou grave) $(Y = 0)$. Uma vez que o logaritmo da RDPC será usado como variável explanatória da probabilidade de segurança alimentar de cada pessoa, foi necessário desconsiderar observações sem informação sobre a insegurança alimentar ou sobre a RDPC, e os domicílios com RDPC nula. Também foram feitas outras depurações, de maneira a ficar com a mesma amostra de 107.772 domicílios usada em análises de lógite apresentadas em artigo anterior (Hoffmann, 2014) [6].

Na PNAD, para avaliar a probabilidade de uma pessoa ter ou não segurança alimentar, é necessário ponderar cada domicílio de amostra por um peso resultante da multiplicação do fator de expansão do domicílio pelo número de pessoas do domicílio. Isso dá origem a números da ordem de milhões. Se aquele peso for sempre dividido pela própria média, obtemos, utilizando esse peso modificado, um número total igual ao número de domicílios da amostra. Criamos, dessa maneira, uma amostra hipotética, como se cada domicílio da amostra tivesse apenas 1 pessoa.

A prevalência de segurança alimentar nessa amostra artificial, em cada estrato de RDPC, é exatamente a mesma que se obtém expandindo a amostra e utilizando o número de pessoas na população. Os números de pessoas apresentados na Tabela 3 são os dessa amostra artificial, criada para facilitar eventual reprodução das estimativas de modelos de lógite e regressões de Poisson discutidos adiante.

Note-se que, a partir do limite superior do primeiro estrato de RDPC, os demais limites são obtidos multiplicando o anterior por 2. Isso significa que, excluindo os dois estratos extremos, os demais têm a mesma amplitude em uma escala logarítmica.

Tabela 3. Pessoas e prevalência de segurança alimentar conforme estratos de RDPC, no Brasil. Amostra artificial criada a partir de dados da PNAD de 2013.

Estrato de RDPC (R$)	RDPC média	Pessoas no estrato	Pessoas com segurança alimentar	Prevalência de segurança alimentar
Mais de 0 a 80	51,3	3.017	904	0,300
Mais de 80 a 160	124,1	5.480	2.027	0,370
Mais de 160 a 320	239,3	17.755	9.627	0,542
Mais de 320 a 640	463,1	30.273	21.606	0,714
Mais de 640 a 1.280	877,8	31.432	26.879	0,855
Mais de 1.280 a 2.560	1.739,6	13.245	12.359	0,933
Mais de 2.560 a 5.120	3.495,4	4.681	4.545	0,971
Mais de 5.120 a 10.240	6.988,5	1.494	1.476	0,988
Mais de 10.240	15.764,5	395	395	1,000
Total	953,5	107.772	79.818	0,741

DOI: http://dx.doi.org/10.20396/san.v23i1.8647217

Sendo X o logaritmo neperiano da RDPC média em cada estrato, o lógite estimado é (entre parênteses estão as estimativas dos desvios padrões das estimativas dos parâmetros, obtidas pelo procedimento padrão, isto é, sem nenhuma correção da variância)

$$\hat{Q} = \underset{(0,0597)}{-5,9102} + \underset{(0,0098)}{1,1236X} \quad (9)$$

A correspondente proporção estimada é

$$\hat{P} = \frac{1}{1 + e^{-\hat{Q}}} \quad (10)$$

Uma vez que se trata de escolher o melhor modelo, também foi ajustado um modelo de próbite, verificando-se que ele se ajusta um pouco melhor do que o lógite quando se considera apenas o efeito linear de X. Mas quando se introduz o quadrado de X como variável explanatória (ver adiante), o lógite se ajusta melhor do que o próbite.

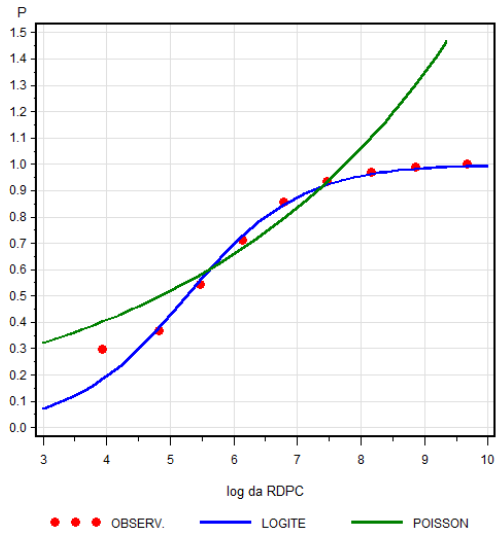
A regressão de Poisson é

$$\ln \hat{P} = \underset{(0,0247)}{-1,8516} + \underset{(0,0037)}{0,2388X} \quad (11)$$

A Figura 2 mostra os pontos observados e as curvas estimadas correspondentes às expressões (10) e (11). É notório que a regressão de Poisson se ajusta muito mal aos dados. A proporção estimada por essa regressão é maior do que 1 nos três últimos estratos, atingindo 1,58 no último. Isso mostra que é claramente inapropriado, nesse caso, impor um modelo que pressupõe constante a RP associada a um acréscimo de uma unidade na variável explanatória. No caso da equação (11), um acréscimo de uma unidade em X (que corresponde a multiplicar a renda por $e \cong 2,7183$), está associado a uma razão de prevalências

$$\text{RP} = e^{0,2388} = 1,27 \quad (12)$$

Figura 2. Proporção de pessoas com segurança alimentar em função de $X = \ln(\text{RDPC})$: pontos observados e curvas estimadas por lógite e por regressão de Poisson simples.



A Figura 2 mostra que o modelo de lógite se ajusta muito bem aos dados, com exceção do ponto referente ao estrato mais pobre. O curioso é que se pode argumentar que a prevalência de segurança alimentar prevista pelo modelo de lógite pode, neste caso, ser mais correta que o valor observado. É obvio que uma RDPC próxima de zero torna impossível ter segurança alimentar em uma economia mercantil, ou seja, a prevalência de segurança deve se aproximar de zero. O fato de o ponto observado apresentar ordenada mais alta talvez seja devido à maior subestimação da renda nesse estrato. Agricultores pobres, por exemplo, podem obter parte substancial da sua renda real da produção para autoconsumo, cujo valor não é medido na PNAD.

Uma maneira de dar mais flexibilidade ao modelo da regressão de Poisson, permitindo que a curva estimada se ajuste melhor aos pontos observados, é introduzir um termo em X^2. Agora a equação estimada com os dados da Tabela 3 fica (colocando entre parênteses as estimativas dos desvios padrões obtidas pelo procedimento padrão)

$$\ln \hat{P} = \underset{(0,1239)}{-4,9819} + \underset{(0,0371)}{1,1937X} - \underset{(0,0028)}{0,0714X^2} \quad (13)$$

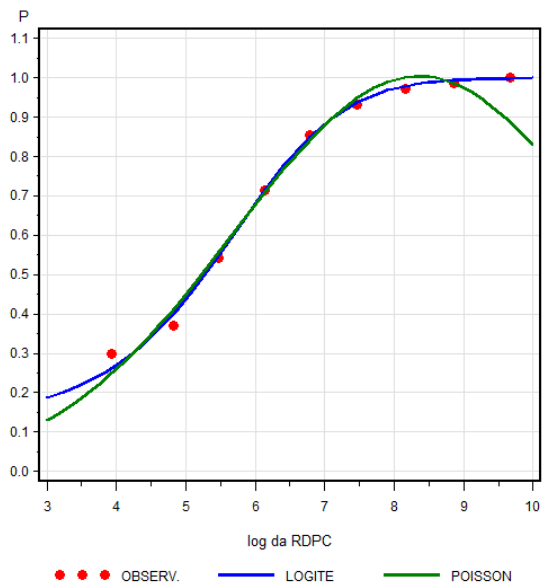
DOI: http://dx.doi.org/10.20396/san.v23i1.8647217

Com a adição de um termo em X^2 o lógite estimado fica

$$\hat{Q} = \underset{(0,2875)}{-1,2633} - \underset{(0,0979)}{0,4703X} + \underset{(0,0083)}{0,1343X^2} \quad (14)$$

A Figura 3 mostra as curvas correspondentes às equações (13) e (14). Observa-se que o ajustamento da regressão de Poisson se torna muito melhor que na Figura 2. Mas agora a curva é decrescente a partir do ponto de abcissa $X = 8,36$, que corresponde a uma RDPC perto de R$ 4.300. Isso não é razoável, pois não há nenhuma razão para prever essa queda da prevalência com o crescimento da RDPC acima daquele nível. E continua havendo probabilidade estimada acima de 1 ($\hat{P} = 1,002$) na abscissa correspondente ao sétimo estrato).

Figura 3. Proporção de pessoas com segurança alimentar em função de $X = \ln(\text{RDPC})$: pontos observados e curvas estimadas por lógite e por regressão de Poisson incluindo X^2.



A ocorrência de proporções estimadas acima de 1 na regressão de Poisson pode ser evitada, nesse exemplo, substituindo a variável contínua X por 8 variáveis binárias que permitem distinguir os 9 estratos de RDPC. Se as 8 variáveis binárias W_h, com $h = 1, 2, \ldots, 9$ forem definidas de maneira que $W_h = 1$ no h-ésimo estrato e $W_h = 0$ nos demais casos, essa regressão leva a estimar 8 coeficientes de regressão, cada um permitindo estimar a RP associada com a mudança de um estrato para o seguinte. Não cabe, entretanto, calcular uma única RP referente ao "efeito da RDPC".

Entre os modelos testados, o que se ajustou melhor foi o de lógite incluindo X^2. Mas mesmo nesse caso verifica-se que o teste para falta de ajustamento[5] é significativo. Trata-se de um qui-quadrado com 6 graus de liberdade que é igual a 83,6, implicando em um fator de heterogeneidade igual a 13,9. Corrigindo as estimativas das variâncias para a falta de ajustamento, as estimativas dos desvios padrões dos três parâmetros passam a ser, na ordem em que são apresentados na equação (14), 1,0733, 0,3656 e 0,0309. O teste para o coeficiente de X^2 continua indicando que ele é diferente de zero.

Quando a renda é um controle importante (por estar associada com muitas características das condições de vida da pessoa), o uso de apenas 3 ou 4 estratos de renda tende a ser insatisfatório. O pesquisador pode pensar que está controlando o efeito da renda, quando controla apenas parte do efeito, pois ainda há muita heterogeneidade de renda dentro dos estratos.

Quando se dispõe de dados individuais, o uso de uma variável explanatória contínua X permite aproveitar melhor a informação referente a cada pessoa. Entretanto, pode ser conveniente substituir a variável contínua por variáveis binárias para um número apropriado de estratos, pois se evita, assim, pressupor qualquer forma específica para a relação funcional entre X e o desfecho. Por outro lado, o uso da variável contínua torna possível usar a equação estimada para fazer interpolações ou extrapolações.

DISCUSSÃO

Ao analisar um conjunto de dados por meio de um modelo estatístico é fundamental especificar esse modelo da melhor maneira possível, respeitando as características essenciais do fenômeno, para que o modelo possa se ajustar bem aos dados. Especificar o modelo envolve tanto a escolha das variáveis (a variável dependente e as variáveis explanatórias) como da forma

[5] Esse teste só pode ser feito se houver repetições e o número de valores distintos da abcissa for maior do que o número de parâmetros da equação estimada. Ele pode ser obtido usando, no SAS, o PROC PROBIT com comando LACKFIT.

da função que as relaciona. A opção por determinada maneira de apresentar os resultados (destacando, por exemplo, os valores da RP ou da OR em cada caso) deve ser algo secundário.

Quando a variável dependente é binária, o uso do lógite, do próbite ou da transformação log-log complementar garantem que a proporção estimada ficará entre 0 e 1, o que não acontece com a regressão de Poisson, como mostra o exemplo apresentado na seção anterior e foi amplamente assinalado por Peterson e Deddens [7,8]. Esses autores já haviam se manifestado [9] contra o uso indiscriminado da regressão de Poisson ao comentarem uma nota de Spiegelman e Hertzmark [10]. As ponderações de Tian e Liu [11] sobre essa nota são muito apropriadas. Na resposta aos comentários, os autores da nota, em certo momento, apresentam, indevidamente, a questão da especificação correta do modelo como algo que poderia estar em contradição com o parâmetro de interesse do pesquisador. Deve-se ter em mente que não há nenhuma dificuldade em estimar a RP quando se usa o lógite, o próbite ou a transformação log-log complementar. É verdade, sim, que a RP não é constante em um modelo de lógite. Mas se o fenômeno analisado se caracteriza por uma RP variável, é enganador e inapropriado forçar o ajuste de uma regressão de Poisson com RP constante, como mostram tanto o exemplo apresentado na seção anterior como o segundo exemplo.

Às vezes se argumenta que um inconveniente do uso do lógite é que o pesquisador pode, indevidamente, considerar a OR uma estimativa da RP [12]. Sabe-se que a RP é aproximadamente igual à OR para eventos raros. Usar a OR como estimativa da RP quando o desfecho em análise tem prevalência alta é, simplesmente, um erro grosseiro de análise. Não faz sentido deixar de usar um modelo de lógite em situações em que ele é apropriado, devido à possibilidade de o resultado ser mal interpretado. Basta o pesquisador calcular e apresentar os valores da RP relevantes para o estudo.

Para comparar dois métodos de estimar um mesmo parâmetro podemos verificar qual é mais preciso comparando a amplitude dos respectivos intervalos de confiança. Uma amplitude maior significa que a variância do estimador é maior e, portanto, menos preciso. Mas não tem sentido afirmar que a estimativa de uma OR é menos precisa que a estimativa de uma RP porque o intervalo de confiança para a primeira é mais amplo [13]. Trata-se de dois parâmetros diferentes. Quando a RP é maior do que 1, a correspondente OR é maior do que a RP e, consequentemente, o intervalo de confiança para a OR deverá ser mais amplo, sem que se possa dizer que o estimador seja menos preciso. Barros e Hirakata [4] não cometem esse erro, mas talvez tenham contribuído para induzir outros ao erro quando colocam nas suas Tabelas 7, 8 e 9 estimativas de OR junto com estimativas de RP, sugerindo que pudessem ser comparadas como se fossem estimativas de um mesmo parâmetro.

Conforme analisado por Oliveira *et al.* [14] já em 1997, é possível obter intervalos de confiança para as estimativas de RP obtidas por meio de um modelo de lógite. Alega-se, entretanto, que o procedimento é muito complicado [12]. Mas isso poderia ser resolvido com a criação de comandos apropriados nos pacotes estatísticos. Mas, antes de tudo, é necessário respeitar a teoria e os fatos. Se a RP é variável, é inapropriada a ideia de resumir o resultado da pesquisa em uma única estimativa da RP e o respectivo intervalo de confiança.

Eventualmente, ao usar o lógite (o próbite ou a transformação log-log complementar), seja conveniente separar a estimativa da RP da avaliação da significância estatística dos vários efeitos relevantes para a pesquisa. A análise da significância estatística desses efeitos seria feita com base nas estimativas de parâmetros do modelo e dos testes de hipóteses apropriados.

Não devemos esquecer, ainda, que no caso de dados provenientes de estudos caso-controle a OR é uma medida apropriada, mas a RP não é, pois seu cálculo exige o uso de um total sem significado estatístico [15].

REFERÊNCIAS

[1] Cramer JS. The origins and development of the logit model. Versão atualizada do capítulo 9 – Logit models from economics and otheer fields. Cambridge: Cambridge University Press; 2003.

[2] Rodrigues MS. Dicionário brasileiro de estatística. 2. ed. Rio de Janeiro: IBGE; 1970.

[3] Zou G. A modified Poisson regression approach to prospective studies with binary data. Am. J. Epidemiol. 2004;159(7):702-706.

[4] Barros AJD, Hirakata VN. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. BMC Med. Res. Methodol. 2003;3:21.

[5] Spiegelman D, Hertzmark E. Easy SAS calculations for risk or prevalence ratios and diferences. Am. J. Epidemiol. 2005;162(3):199-200.

[6] Hoffmann R. Brasil, 2013: mais segurança alimentar. Segur. Aliment. Nutr. 2014;21(2):422-36.

[7] Petersen MR, Deddens JA. A comparison of two methods for estimating prevalence ratios. BMC Med. Res. Methodol. 2008;8:9. DOI: http://doi.org/10.1186/1471-2288-8-9.

[8] Deddens JA, Petersen MR. Approaches for estimating prevalence ratios. Occup. Environ. Med. 2008;65:501-506.

[9] Petersen MR, Deddens JA. RE: Easy SAS calculations for risk or prevalence rations and differences. Am. J. Epidemiol. 2006;163(12):1158-1159.

[10] Spiegelman D, Hertzmark E. The authors reply. Am. J. Epidemiol. 2006;163(12):1159-1161.

[11] Tian L, Liu K. RE: Easy SAS calculations for risk or prevalence rations and diferences. Am. J. Epidemiol. 2006;163(12):1157-1158.

[12] Coutinho LMS, Scazufca M, Menezes PR. Methods for estimating prevalence ratios in cross-sectional studies. Rev. Saúde Pública. 2008;42(6):1-6.

[13] Francisco PMSB, Donalisio MR, Barros MBA, Cesar CLG, Carandina L, Goldbaum M. Medidas de associação em estudo transversal com delineamento complexo: razão de chances e razão de prevalência. Rev. Bras. Epidemiol. 2008;11(3):347-55.

[14] Oliveira NF, Santana VS, Lopes AA. Razões de proporçõs e uso do método delta para intervalos de confiança em regressão logística. Rev. Saúde Pública. 1977;31(1):90-9.

[15] Kale PL, Costa AJL, Luiz RR. Medidas de efeito e medida de associação. In: Medronho RA. Epidemiologia. São Paulo: Editora Atheneu; 2002. p. 115-125.