Efeito do plano amostral em modelo logístico ordinal: uma análise do estado de saúde autorreferido de adultos no Brasil usando a *Pesquisa Nacional por Amostra de Domicílios* de 2008

Effect of sampling plan on ordinal logistic models: an analysis of self-rated health status among Brazilian adults based on the *National Household Sample Survey* (PNAD 2008)

José Rodrigo de Moraes [1]
Jessica Pronestino de Lima Moreira [2]
Ronir Raggio Luiz [2]

[1] *Instituto de Matemática e Estatística, Universidade Federal Fluminense, Niterói, Brasil.*
[2] *Instituto de Estudos em Saúde Coletiva, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil.*

Correspondência
J. R. Moraes
Departamento de Estatística, Instituto de Matemática e Estatística, Universidade Federal Fluminense.
Rua Mario Santos Braga s/n, 7º andar, Niterói, RJ 24020-140, Brasil.
jrodrigo78@est.uff.br

Abstract

Studies that draw on individual and environmental variables to explain differences in self-rated health status have increased gradually in Brazil, but are still limited in number. Due to time and cost issues, many studies use a complex sample design involving features (stratification, clustering, and different sample weights) that, when ignored, can influence odds ratios and standard errors in the statistical models. Using the National Household Sample Survey *(PNAD 2008), this paper assesses the impact on these measurements when some or all of these features are not taken into account in fitting ordinal logistic models to establish associations between adults' self-rated health and various individual and environmental factors. According to this study, failure to take these three features into account simultaneously led to changes in the magnitude of the odds ratio between better self-rated health and most of the factors, besides important underestimation of standard errors.*

Logistic Models; Morbidity; Sampling Studies

Introdução

A inferência estatística clássica pressupõe que os dados amostrais sejam obtidos usando amostragem aleatória simples com reposição, sendo considerados realizações de variáveis aleatórias independentes e identicamente distribuídas. Entretanto, a maioria dos inquéritos nacionais que levantam informações de saúde como, por exemplo, a *Pesquisa Nacional por Amostra de Domicílios* (PNAD 2008) e a *Pesquisa Nacional de Saúde do Escolar* (PeNSE 2009), realizadas pelo Instituto Brasileiro de Geografia e Estatística (IBGE), assim como a *Pesquisa Mundial de Saúde* (PMS 2003), realizada pela Organização Pan-Americana da Saúde (OPAS), utilizam planos de amostragem complexa. A PNAD, por exemplo, apresenta todas as características de um plano amostral complexo, como estratificação, conglomeração e probabilidades desiguais de seleção, além de ajustes dos pesos amostrais para calibração com totais conhecidos da população [1]. Os pesos ajustados (ou calibrados) correspondem ao produto dos pesos naturais do desenho (inversos das probabilidades de seleção em cada estágio) e um fator de ajuste calculado pela razão entre os totais populacionais estimados e conhecidos (ou projetados).

A análise estatística de dados de inquéritos amostrais complexos pode ser realizada tanto com fins descritivos ou analíticos usando diferentes pacotes estatísticos (SAS, SUDAAN,

Stata, R etc.), cuja escolha depende basicamente se o pacote opera num ambiente computacional familiar ao analista e se contém as técnicas estatísticas necessárias para executar a análise de interesse [1]. Cabe ressaltar ainda que para efetuar a análise de dados provenientes de inquéritos amostrais complexos é preciso incorporar as informações do plano amostral da pesquisa, informando os pesos das unidades da amostra e as informações estruturais da pesquisa, identificando o estrato e pelo menos as unidades primárias de amostragem. Segundo Pessoa & Silva [1] e Szwarcwald & Damacena [2] as estimativas pontuais são influenciadas por pesos distintos associados às unidades da amostra, enquanto as estimativas de variância são influenciadas pela estratificação, conglomeração e pesos amostrais.

Segundo Caiaffa et al. [3], o conceito de saúde deve considerar o cotidiano dos indivíduos vivendo nas cidades e as relações de interdependência que existem entre o indivíduo e o meio físico e social onde ele vive. Estudos que consideram variáveis individuais e contextuais do local de moradia para explicar as diferenças no estado de saúde autorreferido de indivíduos vêm paulatinamente crescendo no Brasil, mas ainda são escassos [4].

Apesar do estado de saúde autorreferido, frequentemente levantado em diferentes pesquisas amostrais, ser um indicador de natureza subjetiva é considerado uma *proxy* das avaliações objetivas de saúde [5] e um preditor de alta validade e sensibilidade da morbidade e mortalidade [6,7]. Embora seja um indicador policotômico ordinal, em muitas análises o estado de saúde autorreferido é geralmente dicotomizado [8,9,10,11]. Além disso, há estudos que ainda não explicitam ou não consideram o uso de todas as informações do desenho amostral (estrato, UPA e pesos amostrais) no ajuste do modelo estatístico adotado [11,12].

Usando os dados da PNAD 2008 [13,14] este artigo tem como objetivo fazer uma análise comparativa entre as medidas de associação (*odds ratio* – OR), os erros padrão associados às estimativas pontuais dos parâmetros do modelo e a significância estatística dessas estimativas obtidas a partir do ajuste de um modelo logístico ordinal, cujo propósito é estabelecer associação entre a autoavaliação do estado de saúde de adultos no Brasil e a área de localização dos domicílios em que residem (urbana e rural), sob três formas de ajuste: (1) pelo método de máxima pseudo-verossimilhança (MPV) considerando os pesos amostrais e as informações estruturais do plano amostral: essa forma de ajuste se refere a uma situação adequada de modelagem na qual o pesquisador reconhece os aspectos de complexidade do plano amostral da PNAD, como estratificação, conglomeração e probabilidades desiguais de seleção e os incorpora no ajuste do modelo; (2) pelo método de máxima verossimilhança (MV) considerando os pesos amostrais e desconsiderando as informações estruturais do plano amostral: essa forma de ajuste representa uma situação inadequada de modelagem na qual o pesquisador ignora ou negligencia que a amostra é complexa, levando em conta apenas as ponderações associadas às unidades da amostra da PNAD; e (3) pelo método de MV desconsiderando os pesos amostrais e as informações estruturais do plano amostral: essa forma de ajuste também indica uma situação inadequada de modelagem na qual o pesquisador pressupõe que a amostra da PNAD é uma amostra aleatória simples com reposição, situação na qual não existiria estratificação e conglomeração das unidades e os pesos seriam considerados constantes, e portanto desnecessários.

Material e métodos

PNAD

A PNAD é um inquérito amostral de abrangência nacional, realizado pelo IBGE, que em 2008 coletou informações de uma amostra probabilística de 150.591 domicílios e 391.868 pessoas. A partir da amostra da PNAD é possível obter estimativas representativas para o Brasil, grandes regiões, Unidades da Federação e nove regiões metropolitanas [Belém (Pará), Fortaleza (Ceará), Recife (Pernambuco), Salvador (Bahia), Belo Horizonte (Minas Gerais), Rio de Janeiro, São Paulo, Curitiba (Paraná) e Porto Alegre (Rio Grande do Sul)]. Embora a PNAD tenha como finalidade básica produzir informações para o estudo do desenvolvimento socioeconômico do país, a PNAD 2008 através de seu suplemento levantou ainda uma série de características de saúde sobre a população brasileira que compreendem: a autoavaliação do estado global de saúde, a morbidade autorreferida, a mobilidade física autorreferida e a necessidade de cuidados médicos e utilização de serviços.

No que se refere ao seu planejamento amostral, a PNAD é um estudo seccional (ou transversal) que utiliza um plano amostral complexo [13,14] que envolve estratificação, conglomeração, com probabilidades desiguais de seleção que resultam em pesos amostrais distintos para as unidades amostrais.

População de estudo

A população de estudo trata-se de um domínio de estimação composto por adultos de 20 anos ou mais de idade que residem em domicílios particulares permanentes (DPP), correspondendo a 162.213 registros na amostra da PNAD. Foram considerados na análise apenas os indivíduos que informaram seu próprio estado global de saúde, por entender que informações prestadas por terceiros (outras pessoas moradoras do domicílio, ou ainda outras pessoas não moradoras do domicílio) aumentariam a chance de ocorrer viés de informação ou classificação do desfecho de saúde utilizado neste artigo. Do total de adultos residentes em DPP, 37,2% tiveram seu estado de saúde informado por terceiros.

Modelagem estatística

Modelos de regressão ordinal [15,16] são utilizados quando o desfecho do modelo é policotômico ordinal e as variáveis explicativas numéricas ou categóricas. Neste artigo, foi ajustado um modelo de regressão logística ordinal, utilizando como desfecho do modelo a autoavaliação do estado de saúde com três níveis: muito bom/bom, regular, ruim/muito ruim.

A autoavaliação de saúde é um indicador obtido no nível individual que tem sido amplamente utilizado em pesquisas que visam estabelecer o estado de saúde do indivíduo, por ser de fácil mensuração e por permitir comparações internacionais [17]. A principal vantagem deste indicador é a sua forte associação com o estado real de saúde dos indivíduos, indicando que a autoavaliação da saúde pode ser utilizada como uma *proxy* das avaliações objetivas de saúde [5]. Além disso, a subjetividade não é uma limitação e sim um ponto forte deste indicador, tendo em vista que o processo saúde-doença é um processo contínuo, relativo e subjetivo, cuja percepção, cognição e interpretação variam no tempo e no espaço, em função de fatores individuais e contextuais [18].

Com o objetivo de estabelecer a associação entre a área de localização do domicílio (urbana e rural) e o desfecho de autoavaliação do estado geral de saúde do indivíduo, utilizou-se como controle um conjunto de 18 variáveis demográficas, socioeconômicas, comportamentais e de saúde que retratam características individuais e do ambiente intradomiciliar (domicílio) e extradomiciliar (setor): (1) sexo, (2) faixa etária, (3) cor/raça, (4) escolaridade, (5) situação de ocupação, (6) atividade física, (7) fumo, (8) morbidade autorreferida, (9) mobilidade física, (10) posse de plano de saúde, (11) consulta médica nos últimos 12 meses, (12) domicílio cadastrado no Programa Saúde da Família (PSF), (13) qualidade de moradia, (14) posse de bens básicos no domicílio, (15) condição de ocupação do domicílio, (16) renda domiciliar mensal *per capita*, (17) região de residência e (18) proporção de domicílios considerados adequados no setor (vizinhança) quanto aos serviços sociais básicos (água, esgoto, lixo e luz), às densidades habitacionais e aos padrões construtivos das moradias (material usado nas paredes e telhado).

No ajuste do modelo de regressão logística ordinal incorporando o plano amostral da pesquisa, foi utilizado o método de MPV descrito em Pessoa & Silva [1] através do módulo *complex samples* do pacote estatístico SPSS for Windows, versão 17.0 (SPSS Inc., Chicago, Estados Unidos). No ajuste do modelo multivariado foram testados (teste de Wald) primeiramente todos os efeitos principais das variáveis explicativas (18 variáveis de controle e área de localização do domicílio), a partir do qual se excluíram aquelas variáveis de controle que não apresentavam efeito estatisticamente significativo. Num segundo momento, foram adicionados e testados separadamente os efeitos de interação (duplas) entre cada uma das variáveis de controle selecionadas e a área de localização do domicílio (variável de interesse). Num terceiro momento, foi ajustado um modelo incluindo simultaneamente além da área de localização do domicílio, as variáveis de controle selecionadas e as variáveis de interação que apresentaram efeitos significativos.

A representação geral do modelo selecionado pelo método de MPV se encontra na equação abaixo:

$$logit\left[P\left(Y_k \le m\right)\right] = \tau_m - \left(\alpha_a + \beta_b + \gamma_c + \delta_d + \varpi_e + \xi_f + \varphi_g + \eta_h + \psi_i + \lambda_j + \phi_l + \varsigma_o + \pi_p + \rho_q + \vartheta_r + \theta X_k + (\alpha\gamma)_{ac} + (\alpha\varpi)_{ae} + (\alpha\phi)_{al} + (\alpha\pi)_{ar} + (\alpha\theta)_a X_k\right)$$

$m = 1,..., M\text{-}1$ e $k = 1,2, ..., n_{abcdefghijlopqr}$, onde:

Y_k é a autoavaliação do estado de saúde do k-ésimo adulto com $M = 3$ categorias. A autoavaliação do estado de saúde (desfecho observado) do k-ésimo indivíduo associa-se com a variável latente não observável contínua (Y_k^*) da seguinte forma, como mostrado abaixo:

$$Y_k = \begin{cases} 1 \rightarrow (muito\ ruim\ ou\ ruim) & se \quad Y_k^* < \gamma_1 \\ 2 \rightarrow \quad (regular) & se \quad \gamma_1 \le Y_k^* < \gamma_2 \\ 3 \rightarrow (muito\ bom\ ou\ bom) & se \quad Y_k^* \ge \gamma_2 \end{cases}$$

Os valores γ_1, e γ_2, denominados pontos de corte, são estritamente crescentes, isto é, $\gamma_1 < \gamma_2$. Estes pontos de corte permitem que as probabilidades do desfecho em cada classe difiram entre si.

α_a é o efeito principal do *a*-ésimo nível do fator área de localização do domicílio; $a = 1,2$;

β_b é efeito principal do *b*-ésimo nível do fator região de residência; $b = 1,2,3,4,5$;

γ_c é efeito principal do *c*-ésimo nível do fator sexo; $c = 1,2$;

δ_d é efeito principal do *d*-ésimo nível do fator faixa etária ; $d = 1,2,3,4,5$;

ϖ_e é efeito principal do *e*-ésimo nível do fator cor/raça; $e = 1,2$;

ξ_f é efeito principal do *f*-ésimo nível do fator escolaridade; $f = 1,2,3,4,5$;

φ_g é efeito principal do *g*-ésimo nível do fator renda domiciliar mensal *per capita*; $g = 1,2,3,4$;

η_h é efeito principal do *h*-ésimo nível do fator situação de ocupação; $h = 1,2$;

ψ_i é efeito principal do *i*-ésimo nível do fator atividade física; $i = 1,2,3$;

λ_j é efeito principal do *j*-ésimo nível do fator fumo; $j = 1,2,3,4$;

ϕ_l é efeito principal do *l*-ésimo nível do fator morbidade autorreferida; $l = 1,2$;

ζ_o é efeito principal do *o*-ésimo nível do fator mobilidade física autorreferida; $o = 1,2,3,4$;

π_p é efeito principal do *p*-ésimo nível do fator posse de plano de saúde; $p = 1,2$;

ρ_q é efeito principal do q-ésimo nível do fator consulta ao médico; $q = 1,2$;

ϑ_r é efeito principal do r-ésimo nível do fator posse de bens básicos; $r = 1,2$;

θ é o quanto varia o logito ao aumentar em uma unidade o percentual de domicílios adequados quanto a qualidade da moradia no setor;

X_k é o percentual de domicílios adequados quanto a qualidade da moradia no setor em que reside o *k*-ésimo indivíduo;

$(\alpha\gamma)_{ac}$ é o efeito de interação entre o *a*-ésimo nível do fator área de localização do domicílio e o *c*-ésimo nível do fator sexo;

$(\alpha\varpi)_{ae}$ é o efeito de interação entre o *a*-ésimo nível do fator área de localização do domicílio e o *e*-ésimo nível do fator cor/raça;

$(\alpha\phi)_{al}$ é o efeito de interação entre o *a*-ésimo nível do fator área de localização do domicílio e o *l*-ésimo nível do fator morbidade autorreferida;

$(\alpha\vartheta)_{ar}$ é o efeito de interação entre o *a*-ésimo nível do fator área de localização do domicílio e o *r*-ésimo nível do fator posse de bens básicos;

$(\alpha\theta)_a$ é o efeito de interação referente à variável percentual de domicílios adequados quanto à qualidade da moradia no setor e devido ao *a*-ésimo nível do fator área de localização do domicílio.

Resultados

Ao ajustar o modelo logístico ordinal considerando apenas os efeitos principais das variáveis explicativas, observou-se que as variáveis condição de ocupação do domicílio, qualidade da moradia e PSF não apresentaram efeito estatisticamente significativo.

Ao adicionar separadamente no modelo as interações entre cada uma das variáveis de controle selecionadas (15 variáveis) e a área de localização do domicílio (variável de interesse), observou-se que seis interações apresentaram efeito estatisticamente significativo: área*sexo, área*cor, área*morbidade autorreferida, área* plano de saúde, área*posse de bens básicos e área*percentual de domicílios adequados quanto à qualidade da moradia. Em seguida, foi ajustado um modelo incluindo além da área de localização do domicílio, as 15 variáveis de controle e as seis variáveis de interação mencionadas, mas a partir desse ajuste observou-se que a interação entre a área de localização do domicílio e o plano de saúde deixou de apresentar efeito significativo, sendo excluído do modelo.

Os resultados do ajuste do modelo (modelo selecionado) pelo método de MPV considerando todas as informações do plano amostral são apresentados na Tabela 1.

Ao comparar os resultados do modelo de regressão logística ordinal ajustado pelo método de MPV considerando todas as informações do plano amostral (Tabela 1) com os resultados do ajuste deste mesmo modelo pelo método de MV considerando as pesos amostrais e ignorando as informações estruturais do plano amostral (estrato e UPA), observa-se (Tabela 2) que as medidas de OR não sofrem qualquer alteração, uma vez que as estimativas pontuais dos parâmetros do modelo (efeitos principais e de interação) são influenciadas apenas pelos pesos amostrais. Entretanto, ao desconsiderar as informações estruturais do ajuste todas as estimativas dos erros padrão dos estimadores dos parâmetros do modelo ficam subestimadas, dando uma falsa ideia, de que as estimativas pontuais obtidas são mais precisas do que na realidade elas são, resultando em valores de p menores.

Ao analisar, por exemplo, a interação entre as variáveis área de localização do domicílio e sexo, conclui-se que tanto no ajuste do modelo por MPV quanto por MV_2, mulheres residentes em DPP localizados na área rural possuem uma chance 7,5% [1/OR = 1/(0,93*1*1) = 1/0,93 = 1,075)] maior de reportarem melhores estados de saúde do que mulheres residentes em DPP na área urbana. Todavia, no modelo ajustado por MV_2, essa associação é considerada

Tabela 1

Modelo logístico ordinal ajustado por máxima pseudo-verossimilhança (MPV) considerando todas as informações do plano amostral.

Características (variáveis)	**Modelo 1 ajustado por MPV considerando os pesos amostrais e as informações estruturais do plano amostral**		
	OR_{MPV}	**Erro padrão$_{MPV}$**	**Valor de p**
Área de localização do domicílio			
Urbano	0,93	0,059	0,226
Rural	1,00	-	-
Região de residência			
Norte	0,75	0,051	< 0,001
Nordeste	0,78	0,032	< 0,001
Sudeste	1,10	0,032	0,003
Sul	0,99	0,039	0,819
Centro-oeste	1,00	-	-
Sexo			
Masculino	0,96	0,033	0,187
Feminino	1,00	-	-
Faixa etária (anos)			
20-29	1,42	0,029	< 0,001
30-39	1,20	0,024	< 0,001
40-49	0,96	0,023	0,101
50-59	0,84	0,022	< 0,001
60 ou +	1,00	-	-
Cor/Raça			
Branca	1,02	0,039	0,618
Não branca	1,00	-	-
Escolaridade (anos)			
Sem instrução ou menos de 1	0,46	0,042	< 0,001
1-7	0,48	0,037	< 0,001
8-14	0,70	0,036	< 0,001
15 ou +	1,00	-	-
Sem declaração	0,47	0,167	< 0,001
Renda domiciliar mensal *per capita* (salários mínimos)			
Sem renda ou até 1	0,53	0,046	< 0,001
1-5	0,71	0,044	< 0,001
Mais de 5	1,00	-	-
Sem declaração	0,69	0,064	< 0,001
Situação de ocupação			
Ocupada	1,19	0,017	< 0,001
Não ocupada	1,00	-	-
Atividade física			
Pratica	1,36	0,019	< 0,001
Não pratica	1,00	-	-
Sem declaração	0,50	0,050	< 0,001
Fumo			
Fumante	0,86	0,021	< 0,001
Ex-fumante	0,88	0,019	< 0,001
Nunca fumou	1,00	-	-
Sem declaração	0,95	0,020	0,009

(continua)

Tabela 1 (continuação)

Características (variáveis)	Modelo 1 ajustado por MPV considerando os pesos amostrais e as informações estruturais do plano amostral		
	OR_{MPV}	Erro padrão$_{MPV}$	Valor de p
Morbidade autorreferida			
Pelo menos uma doença crônica	0,33	0,037	< 0,001
Nenhuma doença crônica	1,00	-	-
Mobilidade física autorreferida			
Muita limitação	0,13	0,031	< 0,001
Limitação	0,18	0,023	< 0,001
Pouca limitação	0,42	0,021	< 0,001
Sem limitação	1,00	-	-
Posse de plano de saúde			
Sim	1,31	0,021	< 0,001
Não	1,00	-	-
Consulta médica			
Sim	0,55	0,020	< 0,001
Não	1,00	-	-
Posse de bens básicos			
Tem todos os quatro bens básicos	1,02	0,039	0,540
Não tem pelo menos um bem básico	1,00	-	-
% domicílios adequados	0,98	0,116	0,864
Área*Sexo			
Urbana*Masculino	0,86	0,035	< 0,001
Área*Cor/Raça			
Urbana*Branca	1,17	0,042	< 0,001
Área*Morbidade autorreferida			
Urbana*Pelo menos uma doença	0,83	0,041	< 0,001
Área*Posse de bens básicos			
Urbana*Todos os bens	1,17	0,044	< 0,001
Área*% domicílios adequados			
Urbana*% domicílios adequados	1,40	0,119	0,005

Nota: modelo 1: γ_1 = -5,961 e γ_2 = -3,152.

estatisticamente significante ao nível de 10%, enquanto que no modelo ajustado por MPV tal associação não apresenta significância estatística.

Ao ajustar o modelo (modelo 3) de regressão logística ordinal pelo método de MV_3 desconsiderando os pesos amostrais e as informações estruturais do plano amostral (estrato e UPA) (Tabela 3), observa-se, para quase todos os fatores, que as medidas de OR sofrem alterações. Embora, de modo geral, o sentido das associações se mantenha, a magnitude dessas associações é de extrema importância, inclusive na análise das interações entre as variáveis. Além disso, é possível observar que os erros padrão dos estimadores dos parâmetros do modelo ficam bem subestimados. Ainda considerando, para fins de exemplificação, a interação entre as variáveis área de localização do domicílio e sexo, observa-se que quando o modelo é ajustado por MV_3, mulheres que moram em DPP localizados na área rural têm 12,4% [1/OR = 1/(0,89*1*1) = 1/0,89 = 1,124] de chance de reportarem melhores estados de saúde, comparativamente às mulheres residentes em DPP na área urbana. Tal associação é mais forte, além de ser estatisticamente significante ao nível de 1% (OR = 1,124; valor de p = 0,002), do que a associação encontrada ao ajustar o modelo por MPV (OR = 1,075; valor de p = 0,226).

Quando se ajusta o modelo por MV_3, apesar de as estimativas pontuais dos efeitos principais e de interação sofrerem aparentemente pequenas modificações, as medidas de associação calculadas para efetuar a análise das interações entre determinados níveis de variáveis podem ser bem distintas. Por exemplo, no modelo ajustado por MPV, adultos brancos que residem em

Tabela 2

Modelo logístico ordinal ajustado por máxima verossimilhança (MV) considerando apenas os pesos amostrais.

Características (variáveis)	Modelo 2 ajustado por MV considerando apenas os pesos amostrais			$\frac{EP_{MPV}}{EP_{MV2}}$
	OR_{MV2} #	Erro padrão$_{MV2}$	Valor de p	
Área de localização do domicílio				
Urbano	0,93	0,039	0,065	1,51
Rural	1,00	-	-	-
Região de residência				
Norte	0,75	0,032	< 0,001	1,59
Nordeste	0,78	0,026	< 0,001	1,23
Sudeste	1,10	0,026	< 0,001	1,23
Sul	0,99	0,029	0,761	1,34
Centro-oeste	1,00	-	-	-
Sexo				
Masculino	0,96	0,029	0,135	1,14
Feminino	1,00	-	-	-
Faixa etária (anos)				
20-29	1,42	0,025	< 0,001	1,16
30-39	1,20	0,023	< 0,001	1,04
40-49	0,96	0,02	0,069	1,15
50-59	0,84	0,019	< 0,001	1,16
60 ou +	1,00	-	-	-
Cor/Raça				
Branca	1,02	0,03	0,518	1,3
Não branca	1,00	-	-	-
Escolaridade (anos)				
Sem instrução ou menos de 1 ano	0,46	0,037	< 0,001	1,14
1-7	0,48	0,034	< 0,001	1,09
8-14	0,70	0,033	< 0,001	1,09
15 ou +	1,00	-	-	-
Sem declaração	0,47	0,154	< 0,001	1,08
Renda domiciliar mensal *per capita* (salários mínimos)				
Sem renda ou até 1	0,53	0,04	< 0,001	1,15
1-5	0,71	0,039	< 0,001	1,13
Mais de 5	1,00	-	-	-
Sem declaração	0,69	0,053	< 0,001	1,21
Situação de ocupação				
Ocupada	1,19	0,014	< 0,001	1,21
Não ocupada	1,00	-	-	-
Atividade física				
Pratica	1,36	0,017	< 0,001	1,12
Não pratica	1,00	-	-	-
Sem declaração	0,50	0,037	< 0,001	1,35
Fumo				
Fumante	0,86	0,018	< 0,001	1,17
Ex-fumante	0,88	0,017	< 0,001	1,12
Nunca fumou	1,00	-	-	-
Sem declaração	0,95	0,018	0,004	1,11

(continua)

Tabela 2 (continuação)

Características (variáveis)	Modelo 2 ajustado por MV considerando apenas os pesos amostrais			$\frac{EP_{MPV}}{EP_{MV2}}$
	OR_{MV2} #	Erro padrão$_{MV2}$	Valor de p	
Morbidade autorreferida				
Pelo menos uma doença crônica	0,33	0,029	< 0,001	1,28
Nenhuma doença crônica	1,00	-	-	-
Mobilidade física				
Muita limitação	0,13	0,025	< 0,001	1,24
Limitação	0,18	0,018	< 0,001	1,28
Pouca limitação	0,42	0,017	< 0,001	1,24
Sem limitação	1,00	-	-	-
Posse de plano de saúde				
Sim	1,31	0,017	< 0,001	1,24
Não	1,00	-	-	-
Consulta ao médico				
Sim	0,55	0,017	< 0,001	1,18
Não	1,00	-	-	-
Posse de bens básicos				
Tem todos os quatro bens básicos	1,02	0,03	0,431	1,3
Não tem ao menos um bem básico	1,00	-	-	-
% domicílios adequados	0,98	0,081	0,806	1,43
Área*Sexo				
Urbana*Masculino	0,86	0,032	< 0,001	1,09
Área*Cor/Raça				
Urbana*Branca	1,17	0,033	< 0,001	1,27
Área*Morbidade autorreferida				
Urbana*Pelo menos uma doença	0,83	0,032	< 0,001	1,28
Área*Posse de bens básicos				
Urbana*Todos os bens	1,17	0,035	< 0,001	1,26
Área*% domicílios adequados				
Urbana*% domicílios adequados	1,40	0,084	< 0,001	1,42

As razões de chance são iguais as obtidas no ajuste do modelo 1 por máxima pseudo-verossimilhança (MPV).

Nota: modelo 2: γ_1 = -5,961 e γ_2 = -3,152 e modelo 3: γ_1 = -5,971 e γ_2 = -3,158

DPP localizados na área urbana possuem uma chance 11% (OR = 0,93*1,02*1,17 = 1,110) maior de autorreferirem melhores níveis de saúde do que adultos não brancos residentes em DPP na área rural. Enquanto que no modelo ajustado por MV_3, a medida de OR obtida na comparação desses mesmos grupos é de apenas 5,7% (OR = 0,89*0,99*1,20 = 1,057).

Discussões

Os resultados encontrados neste artigo indicam o grande impacto de não se considerar simultaneamente as três informações do plano amostral (peso, estrato e UPA) da PNAD ao ajustar um modelo de regressão logística ordinal para estabelecer a associação entre a área de localização do domicílio (urbana, rural) e o estado de saúde autorreferido de adultos no Brasil, considerando um conjunto de variáveis individuais e contextuais. Este impacto é traduzido por alterações nas magnitudes das medidas de razões de chance (OR) associadas à quase totalidade dos fatores considerados e na grande subestimação das medidas de precisão (erros padrão dos estimadores dos parâmetros do modelo).

Neste artigo também se verificou que mesmo ajustando o modelo por MV informando as ponderações referentes aos adultos da amostra (modelo 2), sem contudo considerar as informações estruturais do plano amostral da PNAD, a subestimação dos erros padrão continua expressiva apesar de as medidas de associações (OR)

Tabela 3

Modelo logístico ordinal ajustado por máxima verossimilhança (MV) desconsiderando os pesos amostrais e as informações estruturais do plano amostral.

Características (variáveis)	Modelo 3 ajustado desconsiderando os pesos amostrais e as informações estruturais do plano amostral			$\frac{OR_{MPV}}{OR_{MV3}}$	$\frac{EP_{MPV}}{EP_{MV3}}$
	OR_{MV3}	Erro padrão$_{MV3}$	Valor de p		
Área de localização do domicílio					
Urbano	0,89	0,038	0,002	1,04	1,55
Rural	1,00	-	-	-	-
Região de residência					
Norte	0,78	0,026	< 0,001	0,96	1,96
Nordeste	0,80	0,022	< 0,001	0,98	1,45
Sudeste	1,12	0,023	< 0,001	0,98	1,39
Sul	1,06	0,026	0,02	0,93	1,50
Centro-oeste	1,00	-	-	-	-
Sexo					
Masculino	0,94	0,029	0,048	1,02	1,14
Feminino	1,00	-	-	-	-
Faixa etária (anos)					
20-29	1,40	0,025	< 0,001	1,01	1,19
30-39	1,19	0,022	< 0,001	1,01	1,09
40-49	0,97	0,021	0,135	0,99	1,10
50-59	0,84	0,02	< 0,001	1	1,10
60 ou +	1,00	-	-	-	-
Cor/Raça					
Branca	0,99	0,03	0,836	1,03	1,3
Não branca	1,00	-	-	-	-
Escolaridade (anos)					
Sem instrução ou menos de 1 ano	0,45	0,037	< 0,001	1,02	1,14
1-7	0,47	0,034	< 0,001	1,02	1,09
8-14	0,68	0,032	< 0,001	1,03	1,13
15 ou +	1,00	-	-	-	-
Sem declaração	0,50	0,147	< 0,001	0,94	1,14
Renda domiciliar mensal *per capita* (salários mínimos)					
Sem renda ou até 1	0,54	0,04	< 0,001	0,98	1,15
1-5	0,71	0,038	< 0,001	1,00	1,16
Mais de 5	1,00	-	-	-	-
Sem declaração	0,68	0,053	< 0,001	1,01	1,21
Situação de ocupação					
Ocupada	1,17	0,014	< 0,001	1,02	1,21
Não ocupada	1,00	-	-	-	-
Atividade física					
Pratica	1,36	0,016	< 0,001	1,00	1,19
Não pratica	1,00	-	-	-	-
Sem declaração	0,52	0,037	< 0,001	0,96	1,35
Fumo					
Fumante	0,85	0,018	< 0,001	1,01	1,17
Ex-fumante	0,87	0,017	< 0,001	1,01	1,12
Nunca fumou	1,00	-	-	-	-
Sem declaração	0,95	0,018	0,008	1,00	1,11

(continua)

Tabela 3 (continuação)

Características (variáveis)	Modelo 3 ajustado desconsiderando os pesos amostrais e as informações estruturais do plano amostral			$\frac{OR_{MPV}}{OR_{MV3}}$	$\frac{EP_{MPV}}{EP_{MV3}}$
	OR_{MV3}	Erro padrão$_{MV3}$	Valor de p		
Morbidade autorreferida					
Pelo menos uma doença crônica	0,33	0,03	< 0,001	1,00	1,23
Nenhuma doença crônica	1,00	-	-	-	-
Mobilidade física					
Muita limitação	0,14	0,025	< 0,001	0,93	1,24
Limitação	0,19	0,018	< 0,001	0,95	1,28
Pouca limitação	0,42	0,017	< 0,001	1,00	1,24
Sem limitação	1,00	-	-	-	-
Posse de plano de saúde					
Sim	1,33	0,018	< 0,001	0,98	1,17
Não	1,00	-	-	-	-
Consulta ao médico					
Sim	0,56	0,016	< 0,001	0,98	1,25
Não	1,00	-	-	-	-
Posse de bens básicos					
Tem todos os quatro bens básicos	1,02	0,03	0,416	1,00	1,30
Não tem ao menos um bem básico	1,00	-	-	-	-
% domicílios adequados	0,96	0,085	0,639	1,02	1,36
Área*Sexo					
Urbana*Masculino	0,89	0,032	0,001	0,97	1,09
Área*Cor/Raça					
Urbana*Branca	1,20	0,033	< 0,001	0,98	1,27
Área*Morbidade autorreferida					
Urbana*Pelo menos uma doença	0,86	0,032	< 0,001	0,97	1,28
Área*Posse de bens básicos					
Urbana*Todos os bens	1,18	0,035	< 0,001	0,99	1,26
Área*% domicílios adequados					
Urbana*% domicílios adequados	1,38	0,087	< 0,001	1,01	1,37

estarem corretas, isto é, serem equivalentes as obtidas ao ajustar o modelo pelo método de MPV.

Outros estudos foram realizados com o intuito de avaliar o impacto do plano amostral no ajuste de diferentes modelos estatísticos usando dados de pesquisas amostrais complexas, entre eles pode-se citar o trabalho de Leite [19] em que foi ajustado um modelo multinomial logístico usando os dados da PNAD 1999 e o de Pessoa & Silva [1] que ajustaram um modelo logístico binário com os dados da PNAD 1990. O presente artigo também identificou a necessidade de consideração do planejamento amostral na análise de dados com fins analíticos, mas no contexto de ajuste de um modelo de regressão logística ordinal usando os dados da PNAD 2008. As diferenças observadas nas magnitudes das razões de chance (modelo 3) quanto nos erros padrão dos estimadores (modelos 2 e 3) revelam a importância de se considerar todas as informações do plano amostral (pesos amostrais e informações estruturais) na análise de dados epidemiológicos. Caso contrário, pode comprometer a análise de confundimento e de interação entre as variáveis de extrema importância no campo da Epidemiologia, que por sua vez pode comprometer as conclusões do estudo.

Este estudo apresenta limitação com relação ao processo de calibração dos pesos. Embora o processo de calibração adicione uma nova fonte de incerteza nas estimativas, neste artigo foram

consideradas apenas duas fontes de incerteza: o modelo de superpopulação e o plano amostral adotado para a seleção da amostra da PNAD. No presente artigo foi utilizado o método de MPV, que se baseia na modelagem de superpopulação, para o ajuste de modelo estatístico que leva em conta estratificação, conglomeração e pesos calibrados que são os pesos gravados no banco de dados da PNAD. Segundo Silva et al. [14] este método proporciona estimativas consistentes para a amostra completa da PNAD ou mesmo no caso de domínios de estudo com tamanhos amostrais suficientemente grandes.

Devido à estrutura hierárquica bem definida dos dados da PNAD, onde os indivíduos estão agrupados em unidades domiciliares, que por sua vez estão grupadas em setores censitários, poderia ser realizado um estudo futuro usando modelo multinível [20] (ou hierárquico) visando a avaliar, por exemplo, o impacto de se ignorar tal estrutura nas estimativas pontuais (ou medidas de OR) dos parâmetros do modelo e de suas medidas de precisão. Entretanto, cabe mencionar a maior dificuldade envolvida no ajuste desse tipo de modelo usando grandes bases de dados oriundos de amostragem complexa como a PNAD, devido à necessidade de incorporação não só das informações típicas do desenho amostral da pesquisa (estrato, UPA e peso amostral), mas também da estrutura hierárquica dos dados [21].

Resumo

Estudos que consideram variáveis individuais e ambientais para explicar as diferenças no estado de saúde autorreferido de indivíduos vêm paulatinamente crescendo no Brasil, mas ainda são escassos. Por razões de tempo e custo, muitas pesquisas utilizam planos amostrais complexos que envolvem aspectos (estratificação, conglomeração e pesos amostrais distintos) que quando ignorados podem influenciar as medidas de razões de chance e as medidas de precisão das estimativas dos parâmetros de modelos estatísticos. Usando a Pesquisa Nacional por Amostra de Domicílios *(PNAD 2008), este artigo avalia o impacto nessas medidas quando não se consideram alguns ou todos os aspectos ao ajustar um modelo logístico ordinal para estabelecer a associação entre o estado de saúde autorreferido de adultos e um conjunto de fatores individuais e ambientais. Observou-se que quando não se considera os três aspectos simultaneamente, ocorrem alterações nas magnitudes das medidas de razões de chance do adulto autorreferir melhor estado de saúde associadas à maioria dos fatores, além de grande subestimação dos erros padrões.*

Modelos Logísticos; Morbidade; Amostragem

Colaboradores

J. R. Moraes, J. P. L. Moreira e R. R. Luiz participaram da concepção e projeto, análise e interpretação dos dados, redação e revisão crítica relevante do conteúdo intelectual, e aprovação final da versão a ser publicada.

Agradecimentos

Este projeto foi parcialmente financiado com recurso da Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ, processos nº.: E-26/100.682/2007 e E-26/101.506/2010).

Referências

1. Pessoa DGC, Silva PL. Análise de dados amostrais complexos. São Paulo: Associação Brasileira de Estatística; 1998.
2. Szwarcwald CL, Damacena GN. Amostras complexas em inquéritos populacionais: planejamento e implicações na análise estatística dos dados. Rev Bras Epidemiol 2008; 11 Suppl 1:38-45.
3. Caiaffa WT, Ferreira FR, Ferreira AD, Oliveira CDL, Camargos VP, Proietti FA. Saúde urbana: "a cidade é uma estranha senhora, que hoje sorri e amanhã te devora". Ciênc Saúde Coletiva 2008; 13:1785-96.
4. Santos SM, Chor D, Werneck GL, Coutinho ESF. Associação entre fatores contextuais e auto-avaliação de saúde: uma revisão sistemática de estudos multinível. Cad Saúde Pública 2007; 23:2533-54.
5. Camargos MCS, Rodrigues RN, Machado CJ. Expectativa de vida saudável para idosos brasileiros, 2003. Ciênc Saúde Coletiva 2009; 14:1903-9.
6. Idler EL, Benyamini Y. Self-rated health and mortality: a review of twenty-seven community studies. J Health Social Behav 1997; 38:21-37.
7. Bailis DS, Segall A, Chipperfield JG. Two views of selfrated general health status. Soc Sci Med 2003; 56:203-17.
8. Barros MBA, Zanchetta LM, Moura EC, Malta DC. Auto-avaliação da saúde e fatores associados, Brasil, 2006. Rev Saúde Pública 2009; 43:27-37.
9. Hofelmann DA, Blank N. Auto-avaliação de saúde entre trabalhadores de uma indústria no sul do Brasil. Rev Saúde Pública 2007; 41:777-87.
10. Peres MA, Masiero AV, Longo GZ, Rocha GC, Matos IB, Najnie K, et al. Auto-avaliação da saúde em adultos no Sul do Brasil. Rev Saúde Pública 2010; 44:901-11.
11. Sousa TF, Silva KS, Garcia LMT, Del Duca GF, Oliveira ES, Nahas MV. Autoavaliação de saúde e fatores associados em adolescentes do Estado de Santa Catarina, Brasil. Rev Paul Pediatr 2010; 28:333-9.
12. Romero DE, Sousa Júnior PRB. Determinantes da auto-avaliação da saúde entre adultos e idosos: uma perspectiva de gênero da inter-relação com as doenças crônicas e as limitações funcionais autorreferidas. In: Anais do XIV Encontro Nacional de Estudos Populacionais. Belo Horizonte: Associação Brasileira de Estudos Populacionais; 2004. p. 1-11.
13. Instituto Brasileiro de Geografia e Estatística. Pesquisa Nacional por Amostra de Domicílios 2008. Notas metodológicas: pesquisa básica, pesquisa especial de tabagismo e pesquisas suplementares de saúde e acesso à internet e posse de telefone móvel celular para uso pessoal: Brasil. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística; 2010.
14. Silva PLDN, Pessoa DGC, Lila MF. Análise estatística de dados da PNAD: incorporando a estrutura do plano amostral. Ciênc Saúde Coletiva 2002; 7:659-70.
15. Abreu MNS, Siqueira AL, Caiaffa WT. Regressão logística ordinal em estudos epidemiológicos. Rev Saúde Pública 2009; 43:183-94.
16. Marôco J. Análise estatística com o PASW Statistics (ex-SPSS). Pêro Pinheiro: ReportNumber; 2010.
17. Theme Filha MM, Szwarcwald CL, Souza Júnior PRB. Medidas de morbidade referida e inter-relações com dimensões de saúde. Rev Saúde Pública 2008; 42:73-81.
18. Nogueira H. Os lugares e a saúde. Coimbra: Imprensa da Universidade de Coimbra; 2008.
19. Leite PGPG. Análise da situação ocupacional de crianças e adolescentes nas regiões Sudeste e Nordeste do Brasil utilizando informações da PNAD 1999 [Dissertação de Mestrado]. Rio de Janeiro: Escola Nacional de Ciências Estatísticas; 2001.
20. Diez-Roux AV. Bringing context back into epidemiology: variables and fallacies in multilevel analysis. Am J Public Health 1998; 88:216-22.
21. Scott LT, Ronald HH. Analysis of large-scale secondary data in higher education research: potential perils associated with complex sampling designs. Research in Higher Education 2001; 42: 517-40.

Recebido em 25/Jul/2011
Versão final reaprtesentada em 15/Dez/2011
Aprovado em 16/Jan/2012