Alternatives in modeling of body mass index as a continuous response variable and relevance of residual analysis

Alternativas para modelagem do índice de massa corporal como variável contínua e relevância da análise de resíduos

Maria de Jesus Mendes da Fonseca [1]
Valeska Lima Andreozzi [1]
Eduardo Faerstein [2]
Dora Chor [1]
Marília Sá Carvalho [1]

[1] Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz, Rio de Janeiro, Brasil.
[2] Instituto de Medicina Social, Universidade do Estado do Rio de janeiro, Rio de Janeiro, Brasil.

Correspondence
M. J. M. Fonseca
Departamento de Epidemiologia e Métodos Quantitativos em Saúde, Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz.
Rua Leopoldo Bulhões 1480, Rio de Janeiro, RJ 21041-210, Brasil.
mariajmf@ensp.fiocruz.br

Abstract

This article presents alternatives for modeling body mass index (BMI) as a continuous variable and the role of residual analysis. We sought strategies for the application of generalized linear models with appropriate statistical adjustment and easy interpretation of results. The analysis included 2,060 participants in Phase 1 of a longitudinal study (Pró-Saúde Study) with complete data on weight, height, age, race, family income, and schooling. In our study, the residual analysis of models estimated by maximum likelihood methods yielded inadequate adjustment. The transformed response variable resulted in a good fit but did not lead to estimates with straightforward interpretation. The best alternative was to apply quasi-likelihood as the estimation method, presenting a better adjustment and constant variance. In epidemiological data modeling, researchers should always take trade-offs into account between adequate statistical techniques and interpretability of results.

Body Mass Index; Linear Models; Statistical Data Interpretation

Introduction

Obesity is one of the main contemporary public health problems in societies with different levels of socioeconomic development [1,2].

In most studies, obesity is investigated through body mass index (BMI), analyzed as a discrete variable. Although there is controversy on the advantages and disadvantages of these categories [3], the cut-off points recommended by the World Health Organization (WHO) are widely used [1].

Discretization of continuous variables is widely used, for clinical and epidemiological reasons (e.g., classification of risk brackets, definition of diagnostic criteria, adjustment of biological thresholds) as well as statistical ones (e.g., greater ease in the interpretation of measures of association).

However, the analysis of BMI as a continuous variable allows studying the association between the covariables and the entire distribution of the BMI, without losing the information imposed by its discretization, given that this procedure treats observations grouped in the same stratum as identical [4]. Besides an increase in the statistical power in the analyses is expected [5].

Multiple linear regression models have been used when the response variable is continuous. However, such models assume homogeneity of variance, in this variable, which is not always true. In addition, this approach is not appropri-

ate when the response does not display normal distribution, whether in the shape or the limit of the possible values, as is the case with BMI, which only assumes positive values. In such cases, the usual alternative is transformation of the variable (e.g., logarithmic, inverse functions) before modeling, to linearize the relations between the response variable and the independent variables. Although this strategy is adequate from the statistical point of view (as to the premises of homogeneity and linearity), it may lead to difficulties in interpreting the results, since the parameters will be estimated based on the transformed variable. Some transformations, like the inverse, restrict the interpretation of the parameters, since it is not possible to return the beta coefficient's estimated value to an effect on the original variable. This is because the model's estimate is made for the variable's expected value, which does not present the necessary commutative property for this reparametrization. It would thus be necessary to interpret the effect of schooling on the inverse of BMI (m^2/kg).

With the development of theory and methods for fitting generalized linear models, those problems can be adequately treated in data analysis [6]. When the distribution of the response variable belongs to the exponential family, the relationship between its expected value and the independent predictors can be linearized using a link function. It thus becomes possible to interpret the effects of the independent variables on the expected value of the response variable in the original scale. In addition, the use of alternative estimation methods (other than maximum likelihood) makes possible to fit approximately normal distributions with overdispersion or heterogeneity of variance.

This article discusses strategies for the application of generalized linear models that are adequate from the point of view of statistical fit and, interpretability of results, analysing the residuals to verify the models' assumptions. These strategies are exemplified by modeling that aimed to estimate associations between income, schooling, and BMI, the detailed discussion of which is presented in another article [7].

Methods

This study included participants from phase 1 of the Pró-Saúde Study, employees of a university in Rio de Janeiro, Brazil. Data were collected on the 4,030 subjects from August to October 1999, using a self-completed questionnaire and measuring weight and height. These analyses included only the data for 2,060 women with complete information on weight, age, race, income, and schooling (with 6.2% of missing data). The independent variables were: (a) per capita household income in number of the minimum wage (MW), classified in three categories, < 3 MW, 3-6 MW, > 6 MW; (b) schooling, categorized in complete primary school or less, complete secondary school, and complete university or greater; (c) age, analyzed as a continuous variable; (d) race, reported by subjects themselves based on the classification of the Instituto Brasileiro de Geografia e Estatística [IBGE; Brazilian Institute of Geography and Statistics]; the analyses included only those self-classified as white, mixed-race, or black. The response variable was the BMI.

The analysis was conducted by way of generalized linear models with estimation by the methods of maximum likelihood and quasi-likelihood [6]. The first model assumed a Gaussian distribution of BMI and used the maximum likelihood estimation model. Through Box-Cox transformation, the inverse of BMI was identified as the best transformation of this variable and was thus used in model 2, assuming normal distribution, identity link function, and also estimated by maximum likelihood. In the next step (model 3), a gamma distribution was assumed for BMI – this distribution is continuous and only assumes positive values – still with maximum likelihood estimation and link function equal to identity. Finally, model 4 assumed a normal distribution allowing for overdispersion, and thus estimated by the quasi-likelihood method.

The model fit was evaluated via graphic analysis of the residuals, shown as a detrended quantile-quantile graph for normal distribution (qq-plot) [8] and standardized residuals graphs, based on the deviance function (deviance residual) [6]. To be considered normal, the points representing the residuals in the detrended qq-plot graph must be distributed on the line y = 0. The assumption of homogeneous (constant) variance is assessed by the standardized residuals graph versus the adjusted values. This graph should show a cloud of dots around the horizontal straight line (y = 0), without any defined pattern. The data were analyzed using the free software R, version 2.4.1 (R Development Core Team, Vienna, Austria).

The research protocols were approved by the Institutional Review Board of the university where the research was conducted.

Results

In models 1, 3, and 4 (Table 1), the inverse gradients of the association between schooling and BMI are clearly interpretable. In women

with less than complete primary schooling, the mean estimated BMI was 1.9kg/m² higher than in women with a university education or greater (reference category). In model 2, one can also infer an inverse gradient in the association between schooling and BMI. However, since the response variable used in this model is the inverse of BMI, the value of the coefficients are not very elucidative.

In analyzing the fit of the four models, we observed that the residuals displayed a parabolic curvature in the detrended qq-plot in relation to the line y = 0, with the points diverging at the extremes, mainly on the right side of the segment (Figures 1a and 1b). However, this divergence was greater for model 1 (Figure 1a). Model 2 (Figure 1b) displayed the smallest curvature, and its extreme points are all within the confidence interval. The quasi-likelihood estimation method used in model 4 (Figure 1d) produced a smaller curvature and greater approximation of the extreme points in comparison to model 1 and to a lesser extent to model 3 (Figure 1c), thus showing a better fit considering the premise of normality in the residuals.

The graphs of standardized deviance residuals versus values adjusted by models 2, 3, and 4 (Figures 2b, 2c, and 2d) did not display a clear points, and the observations showed homogeneous variance. However, in model 1 (Figure 2a) it is possible to observe that residuals are not regularly dispersed around the zero line, but increasing as the fitted values increase.

Discussion

In modeling continuous response variables, estimation of coefficients by the maximum likelihood method is widely used. However, in our study the residual analysis suggested an inadequate fit, where normality and homogeneity of variance presuposes were violated, as shown by the sharp bend in the detrended qq-plot and increased amplitude of the standardized residuals, following the adjusted values. Another widely used alternative is transformation of the response variable, used in model 2. Although it displayed an adequate fit, it was not easily interpretable – what is the meaning of the inverse of BMI? –, thus limiting the comparison of these results with findings from other studies.

Despite the better fit with the adoption of the gamma distribution for BMI, the diagnosis of model 3 was still not satisfactory, since the residuals still displayed a curvature, mainly at the extreme values of the distribution.

The best alternative was to change the parameter estimation method to quasi-likelihood (model 4), specifying variance equal to three times the mean. The fit was better, and the variance remained constant. The quasi-likelihood method has several advantages: it is not necessary to specify the distribution of the response variable; interpretation of the coefficients is direct; and the fit for data overdispersion is achieved. The disadvantage is the difficulty in identifying the most adequate variance function for the ob-

Table 1

Estimates of coefficients from regression models and their 95% confidence intervals (95%CI) to evaluate the association between schooling and income and body mass index (BMI) in women. Pró-Saúde Study, Rio de Janeiro, Brazil, 1999.

	Model 1	Model 2	Model 3	Model 4
Schooling *				
≤ complete primary	1.94 (1.30; 2.59) ***	-0.0025 (-0.0034; -0.0016) ***	1.92 (1.25; 2.58) ***	1.91 (1.23; 2.58) ***
Complete secondary	1.43 (0.95; 1.92) ***	-0.0020 (-0.0026; -0.0013) ***	1.41 (0.94; 1.89) ***	1.40 (0.93; 1.87) ***
Income (number of minimum wage) **				
< 3	-0.13 (-0.67; 0.39)	0.0001 (-0.0006; 0.0009)	-0.07 (-0.59; 0.45)	-0.03 (-0.55; 0.48)
3-6	-0.19 (-0.70; 0.33)	0.0002 (-0.0005; 0.0009)	-0.09 (-0.60; 0.41)	-0.04 (-0.54; 0.45)

Model 1: E(BMI) = schooling + income + age + race, normal distribution, estimated by maximum likelihood; model 2: E(1/BMI) = schooling + income + age + race, normal distribution, estimated by maximum likelihood; model 3: 1/E(BMI) = schooling + income + age + race, gamma distribution, estimated by maximum likelihood; model 4: E(BMI) = schooling + income + age + race, estimated by quasi-likelihood, variance function equal to mean cubed.

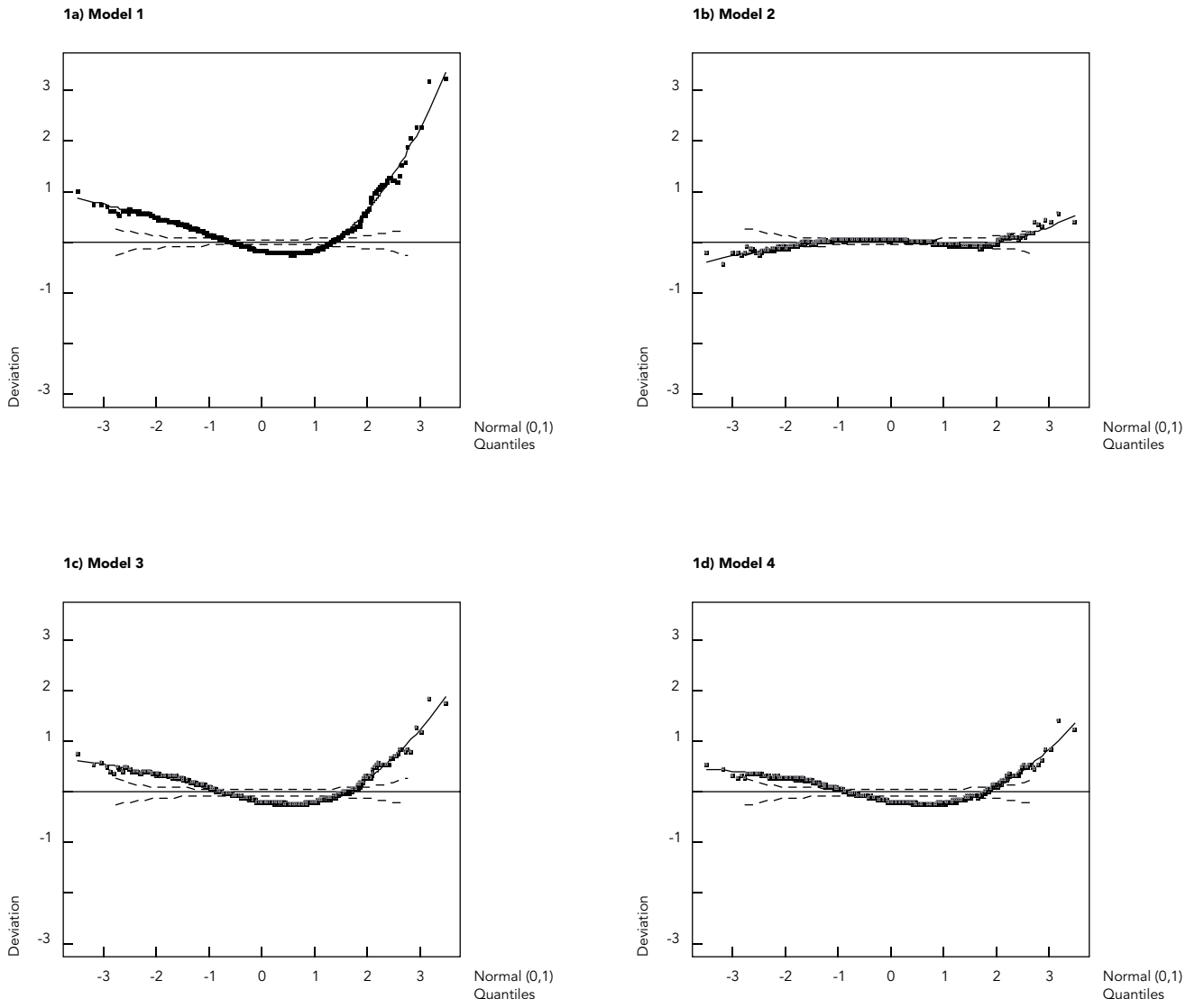
* Reference category: university diploma or greater;

** Reference category: > 6 number of minimum wage;

*** $p < 0.000$.

Figura 1

Normal graph of probabilities for standardized deviance residuals (detrended qq-plot).



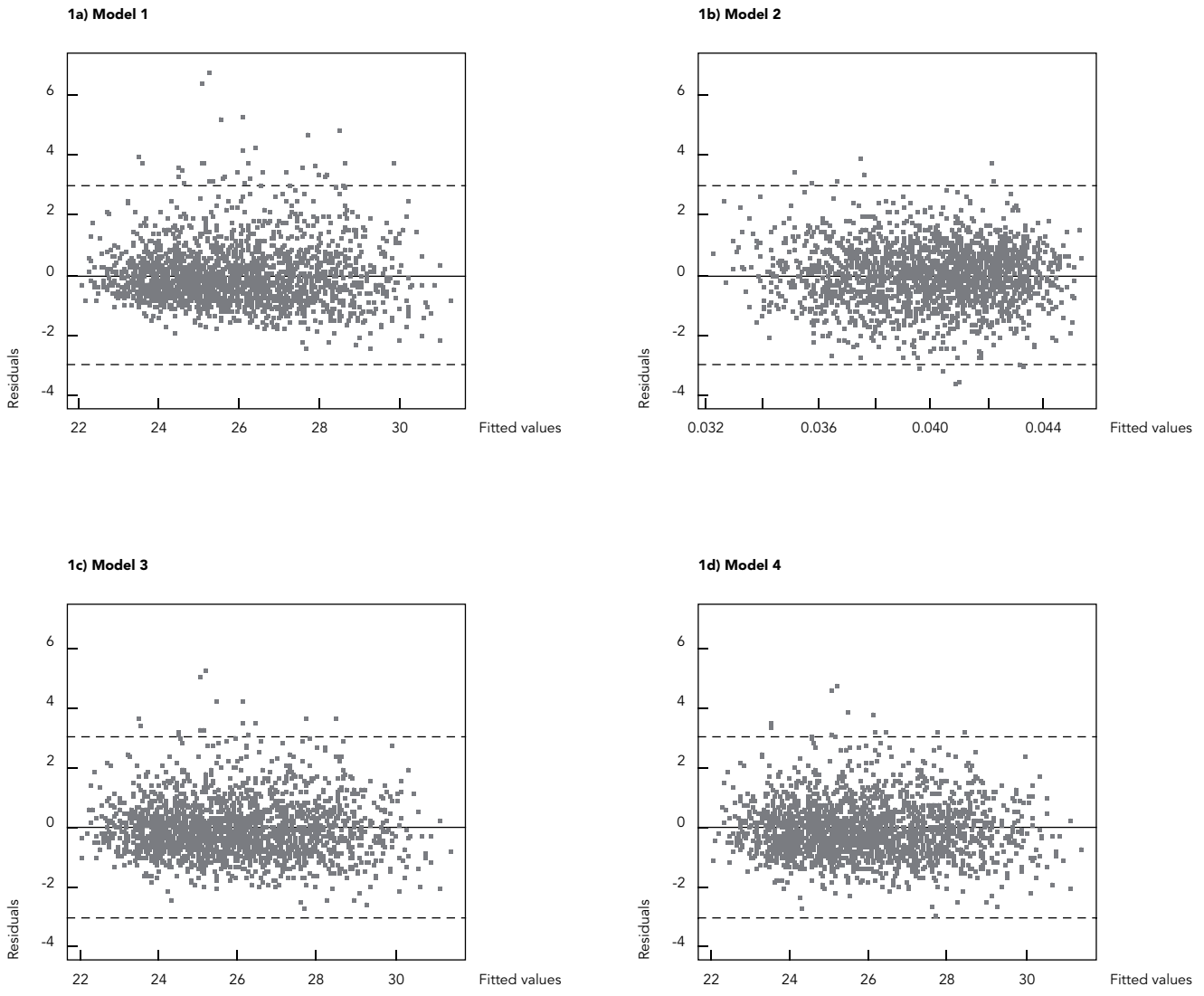
Model 1: normal; model 2: inverse; model 3: gamma; model 4: estimated by quasi-likelihood, variance function equal to mean cubed.

served data. In our case, the choice was based on the fit obtained in the analysis of residuals in the models.

There are dozens of statistical alternatives for the various data modeling problems, but the standardized procedures used in the epidemiological literature are not always the most adequate for the observed data. In the analyses presented here, it was possible to exercise some modeling aspects that have received little emphasis in statistics courses and little use by epidemiologists. Finally, it is up to the researcher to weigh the relations between statistical techniques, the objectives, and the need to interpret the results.

Figura 2

Standardized deviance residuals and adjusted values.



Model 1: normal; model 2: inverse; model 3: gamma; model 4: estimated by quasi-likelihood, variance function equal to mean cubed.

Resumo

Neste artigo, discutem-se alternativas de modelagem do índice de massa corporal (IMC), analisado como variável contínua, e a análise de resíduos. Buscaram-se estratégias de aplicação dos modelos lineares generalizados adequadas tanto do ponto de vista do ajuste estatístico quanto da facilidade de interpretação dos resultados. Nestas análises, foram incluídos dados relativos a 2.060 participantes da Fase 1 de estudo longitudinal (Estudo Pró-Saúde), com informação completa de peso, estatura, idade, raça/cor, renda familiar e escolaridade. Em nosso estudo, a análise de resíduos dos modelos estimados pelo método da máxima verossimilhança, amplamente utilizado, não possibilitou ajuste adequado dos modelos aos dados. A transformação da variável resposta, apesar de resultar em um bom ajuste, não conduziu a estimativas de fácil interpretação. Considerou-se como melhor alternativa a mudança do método de estimação para quase-verossimilhança. Assim, melhor ajuste foi alcançado e a variância permaneceu constante. Na modelagem de dados epidemiológicos, cabe aos pesquisadores buscarem o melhor equilíbrio entre a aplicação adequada de técnicas estatísticas e a facilidade de interpretação dos dados.

Índice de Massa Corporal; Modelos Lineares; Interpretação Estatística de Dados

Contributors

All the authors participated in the literature review, data analysis, and discussion of the results.

References

1. World Health Organization. Report of a WHO Consultation on Obesity. Obesity: preventing and managing the global epidemic. Geneva: World Health Organization; 1998.
2. Wadden TA, Brownell KD, Foster GD. Obesity: responding to the global epidemic. J Consult Clin Psychol 2002; 70:510-25.
3. Seidell JC, Flegal KM. Assessing obesity: classification and epidemiology. Br Med Bull 1997; 53: 238-52.
4. Cumsille F, Bangdiwala SJ. Categorizing variables in the statistical analysis of data: consequences for interpreting the results. Rev Panam Salud Pública 2000; 8:348-54.
5. Moser BK, Coombs LP. Odds ratios for a continuous outcome variable without dichotomizing. Stat Med 2004; 23:1843-60.
6. McCullagh P, Nelder JA. Generalized linear models. London/New York: Chapman & Hall; 1989.
7. Fonseca MJM, Faerstein E, Chor D, Lopes CS, Andreozzi VL. Associações entre escolaridade, renda e índice de massa corporal em funcionários de uma universidade no Rio de Janeiro, Brasil: Estudo Pró-Saúde. Cad Saúde Pública 2006; 22:2359-67.
8. Buuren SV, Fredriks M. Worm plot: a simple diagnostic device for modeling growth reference curves. Stat Med 2001; 20:1259-77.

Submitted on 03/Aug/2006
Final version resubmitted on 10/Jul/2007
Approved on 24/Jul/2007