Índice de Massa Corporal e sua relação com variáveis nutricionais e sócio-econômicas: um exemplo de uso de regressão linear para um grupo de adultos brasileiros

Body mass index and its relationship to nutritional and socioeconomic variables: a linear regression approach to a Brazilian adult sub-population

Mauricio Teixeira Leite de Vasconcellos [1,2]
Margareth Crisóstomo Portela [3]

[1] *Departamento de Metodologia, Diretoria de Pesquisas, Fundação Instituto Brasileiro de Geografia e Estatística. Av. República do Chile 500, 10º andar, Rio de Janeiro, RJ 20031-170, Brasil. mtlv@ibge.gov.br*
[2] *Centro de Estudos da Saúde do Trabalhador e Ecologia Humana, Escola Nacional de Saúde Pública, Fundação Oswaldo Cruz. Rua Leopoldo Bulhões 1480, Rio de Janeiro, RJ 21041-210, Brasil.*
[3] *Departamento de Administração e Planejamento em Saúde, Escola Nacional de Saúde Pública, Fundação Oswaldo Cruz. Rua Leopoldo Bulhões 1480, Rio de Janeiro, RJ 21041-210, Brasil. mportela@ensp.fiocruz.br*

Abstract *This paper focuses on the relationship between body mass index (BMI) and family energy intake, occupational energy expenditure, per capita family expenditure, sex, age, and left arm circumference for a group of Brazilian adults randomly selected among those interviewed for a survey on food consumption and family budgets, called the National Family Expenditure Survey. The authors discuss linear regression methodological issues related to treatment of outliers and influential cases, multicollinearity, model specification, heteroscedasticity, as well as the use of two-level variables derived from samples with complex design. The results indicate that the model is not affected by outliers and that there are no significant specification errors. They also show a significant linear relationship between BMI and the variables listed above. Although the hypothesis tests indicate significant heteroscedasticity, its corrections did not significantly change the model's parameters, probably due to the sample size (14,000 adults), making hypothesis tests more rigorous than desired.*

Key words *Body Mass Index; Linear Models; Eating*

Resumo *O artigo estuda, para um grupo de adultos brasileiros selecionados aleatoriamente dentre os entrevistados no ENDEF, a relação entre o índice de massa corporal (IMC) e as seguintes variáveis: ingestão de energia média na família; gasto energético para atividades laborais; despesa corrente per capita da família, sexo, idade e perímetro braquial. Também discute problemas inerentes ao uso de regressão linear no estabelecimento da relação entre variáveis de mais de um nível de observação (família e pessoa) obtidas por amostras complexas, e técnicas para o diagnóstico e tratamento da influência de pontos extremos, de multicolinearidade, de erros de especificação do modelo e de violação da pressuposição de homocedasticidade. Os resultados mostram que o modelo não é afetado por pontos extremos nem por multicolinearidade, não havendo indicação de erro de especificação. Testes aplicados indicam, no entanto, a presença de heterocedasticidade, cuja correção não acarreta modificação significativa nos parâmetros do modelo estimado. O artigo ressalta que o tamanho da amostra (cerca de 14.000 pessoas) faz com que os testes sejam mais rigorosos do que o desejado e conclui que há associação linear significativa entre o IMC e as variáveis indicadas.*

Palavras-chave *Índice de Massa Corporal; Modelos Lineares; Ingestão de Alimentos*

Introdução

O Índice de Massa Corporal (IMC), ou índice de Quételet (Quételet, 1869), definido como a razão entre a massa corporal (kg) e o quadrado da estatura (m^2), já vinha sendo utilizado como indicador de obesidade em adultos, quando seu uso para avaliação do grau de deficiência crônica de energia em adultos foi proposto por James et al. (1988). Estes autores definiram deficiência crônica de energia e, combinando valores do IMC e de nível de atividade física (expresso em múltiplos da taxa do metabolismo basal diária), estabeleceram pontos de cortes para três graus dessa deficiência. Posteriormente, foi sugerida uma simplificação na avaliação do grau de deficiência crônica de energia em adultos, baseada apenas no IMC (Ferro-Luzzi et al., 1992). Desde então, diversos estudos foram realizados, utilizando esses pontos de corte, e diversas discussões metodológicas foram travadas sobre o assunto. De fato, o IMC é uma medida grosseira da massa corporal que não leva em conta a composição corporal (massa gorda e massa livre de gordura), cuja facilidade de cálculo e comprovada associação com a morbi-mortalidade tem justificado sua utilização como indicador de estado nutricional de adultos em estudos epidemiológicos (Anjos, 1992).

Diante dessas evidências, em 1992, a Organização das Nações Unidas para Agricultura e Alimentação (*Food and Agriculture Organization of the United Nations* – FAO) e o Grupo Consultivo Internacional para Dietética de Energia (*International Dietary Energy Consultative Group* – IDECG*)* organizaram um seminário sobre o significado funcional dos índices de massa corporal baixos, a fim de examinar e discutir a relação entre o IMC e diversas variáveis nutricionais, demográficas, econômicas e sociais e o IMC (IDECG, 1994; Shetty & James, 1994). Nesta ocasião, os resultados divulgados mostraram a relação entre o IMC e as variáveis de ingestão energética familiar e gasto energético ocupacional por meio da análise de correspondências, não tendo sido possível, no entanto, estabelecer esta relação por meio de modelos lineares em função de diversos aspectos de método que são tratados neste artigo: as variáveis utilizadas combinam dois níveis de observação distintos, família e pessoa, e foram obtidas pelo emprego de uma amostra complexa; não foi introduzida uma variável que indicasse o estoque de energia acumulado na massa corporal individual, e não foram adequadamente tratadas as violações aos pressupostos básicos do método de mínimos quadrados.

Este artigo objetiva, portanto, estudar estes aspectos mencionados e estabelecer, por meio do método de regressão linear, a relação entre o IMC e variáveis sobre ingestão energética por adulto-dia da família, gasto energético para atividades físicas ocupacionais, despesa corrente per capita da família, sexo, idade e perímetro braquial, para um conjunto de adultos brasileiros.

Material e métodos

Os dados utilizados neste trabalho derivam de uma subamostra de 13.996 adultos investigados pelo Estudo Nacional da Despesa Familiar (ENDEF), uma pesquisa domiciliar, de abrangência nacional, realizada pela Fundação Instituto Brasileiro de Geografia e Estatística (IBGE), em 1974/1975, sobre consumo alimentar e orçamentos familiares. O ENDEF utilizou uma amostra autoponderada por estrato geográfico e selecionada em quatro estágios: municípios, setores censitários, subsetores e domicílios, com cerca de 55.000 famílias e mais de 123.000 adultos (vinte anos ou mais). A autoponderação foi posteriormente abandonada, quando a amostra foi expandida por um fator de razão entre as dimensões da população e da amostra, controlada pelo tamanho do domicílio, a fim de assegurar a dimensão da população brasileira nos estratos geográficos (Vasconcellos, 1983a). Dessa forma, os dados utilizados neste trabalho provêm de uma amostra domiciliar, selecionada com um desenho complexo, que combina estratificação e conglomeração na seleção e pós-estratificação na expansão, o que merece reflexão face aos pressupostos da modelagem de inferência estatística clássica que se pretende utilizar.

Inferência clássica e amostras complexas

A inferência estatística clássica pressupõe que os valores observados são realizações de variáveis aleatórias independentes e identicamente distribuídas (IID). Com base nesta hipótese de variáveis IID, as técnicas de regressão linear e os testes e diagnósticos em regressão foram desenvolvidos, encontrando-se programados em um grande número de sistemas estatísticos. Ocorre porém, que o único desenho de amostragem que assegura que a amostra reproduza essa distribuição comum, a distribuição da população, é o de amostragem aleatória simples com reposição, o que não é o caso da amostra do ENDEF. De fato, a modelagem de amostragem probabilística largamente empregada pe-

los institutos oficiais de estatística, objetiva a inferência de parâmetros de uma população finita, de forma operacional e com o menor custo possível. Isso implica a utilização de desenhos de amostragem com probabilidades desiguais de seleção, conglomeração, estratificação e tratamentos para não-respostas, que resultam, de um modo geral, em fatores de expansão ou pesos amostrais de alta variabilidade e dificultam a aplicação das técnicas tradicionais da modelagem da inferência estatística clássica.

Pessoa & Nascimento-Silva (1998) fizeram uma revisão pormenorizada dessa questão e, a partir de exemplos extraídos de pesquisas do IBGE, apresentaram alternativas de tratamento usando a modelagem de superpopulação, que combina a aleatorização proveniente do desenho da amostra com a aleatorização da modelagem clássica. Para verificar o efeito da aleatorização do desenho da amostra dentro desta abordagem, é necessário dispor de informações sobre as unidades de seleção, como foram estratificadas e quais as suas probabilidades de seleção. Como a informação sobre a formação das unidades primárias de seleção e sua estratificação na amostra do ENDEF não se encontram disponíveis em meio magnético (Vasconcellos, 1983a) foi impossível utilizar a modelagem de superpopulação.

Diante desta impossibilidade e como o desenho de amostra utilizado no ENDEF não permite que o método dos mínimos quadrados seja aplicado diretamente sobre os dados amostrais, sendo necessário aplicar alguma forma de correção antes da modelagem, optou-se por selecionar uma subamostra autoponderada de adultos como forma de eliminar a necessidade de ponderação dos dados amostrais e, simultaneamente, assegurar que as estimativas pontuais dos parâmetros da população brasileira seriam não-tendenciosas. Em conseqüência, o nível de inferência ficou limitado ao conjunto de adultos estudados.

Seleção da subamostra de adultos

A subamostra de adultos é uma subamostra de domicílios da amostra do ENDEF, na qual apenas um adulto foi selecionado por domicílio, a fim de evitar que as variáveis relativas à família, como a ingestão energética por adulto-dia da família, fossem associadas a diferentes valores do IMC (variável em nível de pessoa), gerando uma covariação que, para a variável em questão, só poderia ser controlada por uma variável que explicasse a distribuição intrafamiliar de alimentos, que não existe na pesquisa ENDEF.

Desta forma, a seleção da subamostra transformou-se na seleção de domicílios que tivessem pelo menos um morador de vinte anos completos de idade ou mais com: (a) medidas antropométricas de massa corporal, estatura e perímetro braquial esquerdo observadas (eliminados 3.726 adultos); (b) medidas antropométricas não alteradas em função de gestação ou lactação (eliminadas 2.261 gestantes adultas e 3.473 lactantes adultas); (c) participação no consumo dos alimentos pesados no domicílio que permitisse uma boa associação com a ingestão por adulto-dia da família, ou seja com mais de 75% de presença às refeições de seu ritmo alimentar (eliminados 18.614 adultos) e (d) dados sobre o consumo familiar de alimentos observados por pelo menos cinco dias (eliminados 2.838 adultos de famílias sem alimentação em casa; 1.788 adultos de famílias com menos de cinco dias de pesquisa e 1.108 adultos de famílias classificadas como em estado de miséria constatada). Com a aplicação destes critérios foram eliminados 33.808 dos 123.466 adultos na amostra do ENDEF, restando 89.658 adultos que correspondem ao universo a partir do qual foi selecionada a subamostra.

Para a seleção da subamostra de adultos foi definida uma nova estratificação, combinando os valores das variáveis estrato geográfico, sexo, classe de idade e de tamanho do domicílio, as mesmas utilizadas na pós-estratificação da amostra (Vasconcellos, 1983a). Foi arbitrado um tamanho inicial de 14.000 adultos que foi alocado entre os estratos de subamostragem proporcionalmente à estimativa de adultos do estrato na população brasileira, que após os necessários arredondamentos, gerou uma subamostra de 13.996 adultos. Para a determinação do adulto a ser selecionado em cada domicílio foi aplicado o algoritmo de Hájek (1960), que consiste em associar, a cada registro, um número aleatório uniformemente distribuído no intervalo (0,1), classificar os registros em ordem crescente do número aleatório e selecionar, neste caso, o primeiro registro, ou seja, o do adulto ao qual foi associado o menor número aleatório. Em seguida, os domicílios foram alocados ao estrato de subamostragem do adulto selecionado e foram selecionados por meio do algoritmo de Hájek, aplicado em cada estrato de modo independente.

Variáveis do modelo

A relação entre a ingestão energética e a massa corporal e, portanto, o índice de massa corporal, não é assunto novo. A dificuldade observada nesta aplicação, no entanto, deriva do fato

de não se conhecer a ingestão energética de cada adulto, e sim uma média de ingestão na família que, apesar de normalizada pela presença individual às refeições (François, 1970) e pelas diferenças de composição familiar por sexo e idade, através de uma escala de adulto equivalente (Vasconcellos, 1983b), ainda pressupõe que a distribuição intrafamiliar de alimentos é proporcional aos requerimentos de energia que devem ser satisfeitos pelo consumo observado em casa, além de ser referida ao momento (semana) de pesquisa. O mesmo problema aplica-se à despesa corrente per capita, que é uma média na família, apesar de poder ser olhada como um indicador das condições de vida e da qualidade da alimentação do adulto.

Os efeitos sobre a massa corporal do aumento ou diminuição do gasto energético em função das atividades físicas também já foram estabelecidos (Durnin & Passmore, 1967). O problema nesta aplicação é o desconhecimento da quantidade de energia gasta em todas as atividades físicas do adulto, pois a variável disponível considera apenas o gasto energético ocupacional e supõe que as horas restantes do dia (24h do dia – 8h de sono – horas trabalhadas) são dedicadas a atividades físicas de dispêndio energético moderadamente ativo (FAO/WHO, 1973). Assim, o que distingue o gasto energético de atividades físicas entre os adultos é apenas a sua atividade laboral.

Apesar dessas limitações, as relações do IMC com a ingestão energética por adulto-dia e com a despesa corrente per capita, já foram estabelecidas por meio da análise de correspondências a partir dos dados do ENDEF (Vasconcellos, 1994). Apesar da forma funcional dessas relações não ter sido objeto do artigo citado, figuras apresentadas indicam que a relação entre o IMC e a idade tem um comportamento parabólico: o IMC aumenta, em média, com a idade até cerca de cinqüenta anos, quando começa a diminuir. Este comportamento, observado para a população brasileira em 1974/1975 (dados do ENDEF), apresenta claras diferenças por sexo e nível de renda (Vasconcellos, 1994).

O grande problema de qualquer modelo envolvendo a ingestão, gasto para atividades físicas, despesa, idade e sexo, reside no fato dessas variáveis não fornecerem qualquer informação sobre o estoque de energia armazenada e o passado nutricional do adulto. A variável de massa corporal (em kg) traz a informação sobre o estoque atual de energia e a de estatura fornece uma indicação da situação nutricional passada (na fase de crescimento), mas não há sentido em sua utilização pois são as variáveis utilizadas no cálculo do IMC. Assim, o perímetro braquial surge como a melhor alternativa disponível para introduzir a informação de estoque de energia do adulto.

Detecção de pontos extremos e influentes

Os pontos extremos (*outliers*) são pontos que se afastam da mediana e que afetam o valor da média de uma ou mais variáveis explicativas. São considerados influentes os pontos que afetam a linha de regressão de tal forma que sua retirada altere significativamente as estimativas. Assim, os pontos extremos devem ser identificados para avaliar seu grau de influência sobre as estimativas, através de uma análise de sensibilidade, que consiste em repetir o procedimento de regressão eliminando os pontos extremos para verificar qual a sua influência sobre os parâmetros estimados. Não havendo modificações significativas, aceita-se que os pontos extremos não são pontos influentes, ou seja, não afetam o modelo.

Nesta aplicação, no entanto, optou-se por executar um outro tipo de análise de sensibilidade. Decidiu-se verificar a influência dos pontos extremos substituindo-os por outros, também selecionados aleatoriamente da amostra do ENDEF, para estimar novamente os parâmetros do modelo. Esse procedimento de substituição foi repetido várias vezes, e os parâmetros do modelo estimado na última iteração foram comparados aos do modelo inicial para avaliar a influência dos pontos extremos. O procedimento de substituição dos pontos extremos respeitou o desenho de seleção da subamostra de adultos, ou seja, o adulto substituto foi o próximo adulto a ser selecionado no estrato, isto é, aquele que recebeu o menor número aleatório dentre os não-selecionados. Na última repetição do procedimento de substituição, no entanto, um dos estratos não tinha mais adulto para ser selecionado pois todos já haviam sido selecionados na subamostra ou em substituição. Para contornar essa impossibilidade, optou-se por excluir os pontos extremos e estimar o último modelo, que foi então, comparado ao primeiro, para avaliar o nível de influência dos pontos extremos.

Diante do tamanho da subamostra, as técnicas gráficas de detecção de pontos extremos não foram utilizadas, sendo aplicados os métodos baseados na diagonal da matriz chapéu, nos resíduos *studentizados*, nos "DFITS" e na estatística D de Cook (Bollen & Jackman, 1990).

A matriz chapéu, representada por $H = \{h_{ij}\}$, corresponde à matriz que (pré) multiplica o vetor de valores observados para gerar o vetor de

valores preditos, ou seja, $\hat{Y} = HY = X(X'X)^{-1}X'Y$. Os elementos da diagonal da matriz H, h_{ii}, indicam a influência de y_i sobre o valor estimado $\hat{y}_i$. Além disso, como o traço (soma dos valores da diagonal principal) de H é igual a p, o número de parâmetros do modelo, espera-se uma influência média da ordem de p/n, onde n é o número de observações. Pode-se mostrar que quanto maior h_{ii}, maior será a distância do valor observado em relação à média de todas as variáveis explicativas. Assim, observações que tenham h_{ii} maior que duas vezes a sua média, $2p/n$, devem ser alvo de exame. De fato, Bollen & Jackman (1990) sugerem $2p/n$ como limite inferior (ou conservador) e $3p/n$ como limite superior para detenção de pontos de "alavancagem" (*leverage*).

Apesar de os resíduos terem, por hipótese do método dos mínimos quadrados, a mesma variância ($E(ee') = \sigma^2 I$), na prática os resíduos não são homocedásticos e sua variância varia em função de σ^2 e de h_{ii}, ou seja, a variância do resíduo da *i-ésima* observação pode ser escrito como $var(e_i) = \sigma^2(1 - h_{ii})$. Assim, quanto maior h_{ii}, tanto menor será a variância do resíduo da observação *i*. Dentre as formas de comparação dos resíduos, Belsey et al. (1980) propuseram a padronização pelo desvio-padrão dos resíduos estimados sem a observação corrente, o que se convencionou chamar de "resíduo *studentizado*", que é definido por

$$e_i^* = e_i / \sqrt{s^2(i)(1 - h_{ii})},$$

onde $s^2(i)$ é a variância dos resíduos estimada sem a observação *i*.

Os resíduos *studentizados* têm a mesma variância (= 1) e distribuição aproximadamente *t* de Student (Gosset, 1908) com (n - p - 1) graus de liberdade. Para identificação de observações com resíduos *studentizados* significantes, recai-se no problema de testes simultâneos e não-independentes (os resíduos não são independentes entre si), sendo o valor crítico da distribuição *t* associado ao valor α / n, para testes unilaterais, e ao valor $\alpha / 2n$, para testes bilaterais.

Como uma observação com resíduo pequeno pode, no entanto, ser um ponto influente na estimação dos parâmetros de regressão, a análise de pontos extremos e influentes não pode ser baseada apenas nos resíduos *studentizados*, sendo necessário considerar também os efeitos de alavancagem, o que é feito por duas medidas: os DFITS e a estatística D de Cook. O $DFITS_i$ é definido por:

$$DFITS_i = \left(\sqrt{h_{ii} / (1 - h_{ii})}\right)\left(e_i / \sqrt{s^2(i)(1 - h_{ii})}\right)$$

ou $DFITS_i = (\hat{Y}_i - X_i b_{(i)}) / \sqrt{s^2(i)\, h_{ii}}$.

Na segunda expressão, o numerador é o valor predito menos o valor que seria predito se a observação *i* fosse eliminada, enquanto o denominador é desvio-padrão do valor ajustado com a variância do resíduo estimada por $s^2(i)$. Assim, o $DFITS_i$ pode ser visto como uma medida da alteração no valor predito da *i-ésima* observação quando os seus valores observados são excluídos dos cálculos. Bollen & Jackman (1990) indicam dois pontos de corte para o módulo dos DFITS: $2\sqrt{p/n}$ como limite mais conservador e $\sqrt{p}$ como limite superior.

A estatística D de Cook (1977), também é definida de mais de uma forma:

$$D_i = \left(1/p\right)\left(h_{ii}/(1 - h_{ii})\right)\left(e_i^2/s^2(1 - h_{ii})\right)$$

ou $D_i = \left(1/p\right)\left(DFITS_i\,[s(i)/s]\right)^2$

A segunda expressão, ao relacionar D_i com $DFITS_i$, mostra que D_i está em uma escala diferente, que tem uma métrica da distribuição F com p e (n-p) graus de liberdade, o que não significa que D_i tenha distribuição F. Por analogia aos pontos de corte sugeridos para os DFITS, Bollen & Jackman (1990) sugerem 4/n como ponto de corte conservador e 1 como limite superior.

Tanto D_i quanto $DFITS_i$ medem alterações em todos os parâmetros de regressão decorrentes da exclusão da observação *i*. Uma medida que indica a influência da retirada de uma observação sobre apenas um parâmetro, por exemplo, β_j, é o $DFBETAS_{ij}$.

Na aplicação em questão, os DFBETAS não foram utilizados e para as outras quatro estatísticas foram utilizados os pontos de corte superiores, ou seja: para a diagonal da matriz H, o valor 3p/n = 0,002144; para o resíduo *studentizado*, o valor $t_{\alpha/2n;(n-p-1)gl} = 4{,}64$; para os DFITS, o valor $\sqrt{p} = 3{,}16628$ e para a estatística D de Cook, o valor 1.

Verificação do grau de multicolinearidade

O termo colinearidade refere-se à existência de uma relação linear perfeita entre algumas das variáveis explicativas. Multicolinearidade refere-se à existência de mais de uma relação linear envolvendo algumas ou todas as variáveis explicativas. O método dos mínimos quadrados pressupõe que não há multicolinearidade entre as variáveis explicativas, uma vez que uma relação linear perfeita entre elas implicaria que seus coeficientes de regressão seriam indeterminados e teriam erro-padrão infinito.

Na prática, no entanto, a multicolinearidade perfeita é rara e o problema passa a ser de grau e não de existência. Quanto maior for o grau de multicolinearidade, maior serão os erros-padrão dos coeficientes de regressão e menor será a precisão em sua estimativa, apesar dos estimadores de mínimos quadrados continuarem a ser os melhores estimadores lineares não-tendenciosos.

Nesta aplicação, o grau de multicolinearidade foi avaliado através da tolerância de cada variável, definida da seguinte forma: $TOL_j = 1 - R_j^2$, onde R_j^2 é o coeficiente de determinação da regressão de X_j pelas demais variáveis explicativas. Se a variável X_j tem uma relação linear boa com as demais variáveis explicativas, $R_j^2 \rightarrow 1$ e $TOL_j \rightarrow 0$. Nesse caso, VIF_j, o fator de inflação da variância, é igual a $1/TOL_j$ e tende a infinito.

Testes sobre especificação do modelo

Os erros de especificação do modelo podem ser decorrentes de: (1) omissão de uma variável relevante, (2) inclusão de variáveis desnecessárias ou redundantes, (3) escolha de uma forma funcional equivocada, ou (4) erros de medida nas variáveis envolvidas no modelo.

O teste RESET (*Regression Specification Error Test*/Teste de Erro de Especificação de Regressão) proposto por Ramsey (1969) é o mais utilizado para testar a hipótese de que o modelo está corretamente especificado. Seu esquema geral consiste em estimar o modelo com novas variáveis e comparar o valor do R^2 novo com o do original através da seguinte estatística de teste:

$$F_{obs} = \frac{(R^2_{novo} - R^2_{original})/nº \text{ de variáveis novas}}{(1 - R^2_{novo})/(N - p_{novo})},$$

onde F_{obs} tem distribuição F de Snedecor com os graus de liberdade indicados na expressão acima.

Duas variantes desse teste foram aplicadas. Na primeira, foram introduzidos, como variáveis explicativas do modelo, o quadrado e o cubo do valor predito, enquanto na segunda, foram introduzidas as potências de segundo, terceiro e quarto graus das variáveis explicativas e todas as interações de primeira ordem.

Testes para homocedasticidade

O método dos mínimos quadrados pressupõe homocedasticidade, ou seja, que os erros aleatórios da função de regressão da população, μ_i, tenham variância igual ou, em termos mais formais, que $E[\mu_i - E(\mu_i)]^2 = E(\mu_i^2) = \sigma^2$, para i = 1, 2, ..., n.

Esta pressuposição de homocedasticidade é fundamental para a construção dos intervalos de confiança e para os testes de hipóteses, e sem ela não se pode assegurar que o método dos mínimos quadrados produza os melhores estimadores lineares não-tendenciosos. De fato, na presença de heterocedasticidade, os estimadores de mínimos quadrados dos parâmetros do modelo continuam a ser lineares e não-tendenciosos, mas os estimadores das variâncias dos parâmetros são tendenciosos.

A análise gráfica dos resíduos é um elemento importante para identificar a existência de uma relação entre os resíduos da regressão e cada variável explicativa e, caso exista, sua forma. Nesta aplicação, no entanto, o número de pontos não permite que se identifique se há alguma relação entre os resíduos e as variáveis explicativas, sendo utilizados os testes estatísticos de Park, da correlação de Spearman e de Breusch-Pagan.

Park (1966) formalizou a análise de resíduos em um teste que pressupõe que a variância dos erros aleatórios, σ_i^2, é uma função das variáveis explicativas do tipo $\sigma_i^2 = \sigma^2 X_{2i}^{\beta 2} X_{3i}^{\beta 3} X_{pi}^{\beta p} e^{vi}$. Usando os quadrados dos resíduos observados, e_i^2, como aproximações dos σ_i^2, e aplicando logaritmo à expressão, Park obteve o seguinte modelo $log e_i^2 = \alpha + \beta_2 log X_{2i} + \beta_3 log X_{3i} + ... + \beta_p log X_{pi} + v_i$ e concluiu que se algum β fosse significativo, aceitava-se a existência de relação entre os resíduos observados e a variável correspondente, rejeitando-se, portanto, a hipótese de homocedasticidade. Não havendo β significativo, não há indicação para rejeição da hipótese de homocedasticidade.

A correlação de Spearman é definida em função das diferenças entres os postos (*rank*), os números de ordem associados do menor ao maior valor de cada variável. Assim, se os postos dos resíduos coincidirem com os de uma variável explicativa, a soma das suas diferenças será nula e a correlação de Spearman será igual a 1. O teste baseado nesta correlação, consiste em calcular a correlação de Spearman entre o valor absoluto dos resíduos e de cada variável explicativa e testar, para cada uma delas, a hipótese nula de sua correlação com os resíduos, usando a seguinte estatística

$$t_{obs} = r_s \sqrt{(n - 2)} / \sqrt{1 - r_s^2},$$

onde $t_{obs} \approx t_{(n - 2)g.l.}$ Se alguma correlação for significativa (isto é, se $t_{obs} > t_\alpha$ ou p-valor $< \alpha$) rejeita-se a hipótese de homocedasticidade.

O teste de Breusch-Pagan (Gujarati, 1988) pressupõe que existe uma relação linear entre o quadrado dos resíduos padronizados, p_i, e as

variáveis explicativas do modelo, que pode ser representada por:

$$p_i = \beta_1 + \beta_2 X_2 + ... + \beta_p X_p + v_i$$

onde, $p_i = e_i^2 / \hat{\sigma}_{MV}^2 = e_i^2 / (\sum e_i^2 / n)$, sendo σ_{MV}^2 o estimador de máximo verossimilhança de σ^2.

Após estimar os parâmetros do modelo acima pelo método dos mínimos quadrados, calcula-se o estimador θ, que é igual à metade da soma de quadrados explicada pelo modelo. Na hipótese de homocedasticidade, o estimador θ tem distribuição aproximadamente χ^2 com (p-1) graus de liberdade e, portanto, se $\theta > \chi^2_{p-1}$, rejeita-se a hipótese de homocedasticidade, por ser significativo o grau de explicação linear das variáveis explicativas sobre os quadrados de resíduos padronizados.

Correção da heterocedasticidade

Sendo detectada a violação do pressuposto de homocedasticidade, deve-se proceder a algum tipo de correção no modelo para tornar os resíduos homocedásticos e poder fazer inferências e testes. Gujarati (1988) apresenta diversos tipos de transformações que podem ser aplicadas quando σ_i^2 é conhecido ou não, o que é mais freqüente.

A solução adotada nesta aplicação, foi o uso dos mínimos quadrados ponderados pelo inverso dos quadrados dos resíduos estimados, $\hat{e}_i^2$, obtidos pelo uso dos mínimos quadrados com o modelo abaixo:

$$e_i^2 = \beta_1 + \beta_2 X_2 + ... + \beta_p X_p + v_i$$

onde e_i^2 é o quadrado do resíduo de mínimos quadrados da i-ésima observação do modelo original. Assim, o método dos mínimos quadrados simples foi aplicado ao modelo original para obtenção de e_i. Os quadrados desses resíduos entraram no modelo acima para estimar $\hat{e}_i^2$. Os pesos foram calculados como $w_i = 1/\hat{e}_i^2$, se $\hat{e}_i^2 \geq 0{,}001$ ou $w_i = 1/e_i^2$, caso contrário.

Resultados

Nos resultados apresentados nesta seção, as variáveis básicas e do modelo final são referidas pelos seguintes rótulos: IMC, índice de massa corporal; CIRC_BRE, circunferência do braço esquerdo (mm); CIRCBRE2, quadrado da circunferência do braço; IDADE, idade em anos completos; ENER_ADU, ingestão energética por adulto-dia; ATV_FIS, energia para atividades físicas diárias; DCP_COR, despesa corrente *per capita*; LOGENER, logaritmo da ingestão de energia por adulto-dia; LOGATVF, logaritmo da energia para atividades físicas; MULHER, indicadora de sexo feminino; M_CIRCBR, interação entre MULHER e CIRC_BRE; M_DCPCOR, interação entre MULHER e DCP_COR; RURAL, indicadora de área rural; FATOR, fator de expansão da subamostra; INTERCEP, coeficiente linear da regressão; PREDITO2, quadrado do valor predito; e PREDITO3, cubo do valor predito. As variáveis adicionais de potências são referidas pelos seguintes rótulos: CIRCBRE3, cubo da circunferência do braço; CIRCBRE4, circunferência do braço à quarta; IDADE2, quadrado da idade; IDADE3, cubo da idade; IDADE4, idade à quarta potência; DCP_COR2, quadrado da despesa corrente *per capita*; DCP_COR3, cubo da despesa corrente per capita e DCP_COR4, despesa corrente *per capita* à quarta potência. Outras variáveis criadas para testes são: LOGRES2, logaritmo do quadrado do resíduo; LLOGENER, logaritmo de LOGENER; LLOGATVF, logaritmo de LOGATVF e RES_PAD2, quadrado dos resíduos padronizados.

Tanto o cadastro de seleção (amostra do ENDEF) quanto a subamostra de 13.996 adultos permitem, desde que corrigidos os respectivos fatores de expansão, gerar estimativas das médias das variáveis do ENDEF usadas nesta aplicação. Assim, colocando entre parêntesis a estimativa da média obtida com o cadastro de seleção, foram estimadas as seguintes médias a partir da subamostra de adultos: IMC 22,27 (22,27); CIRC_BRE 266,92 (270,22); IDADE 39,75 (40,51); ENER_ADU 2806,40 (2756,20); ATV_FIS 409,76 (382,22); DCP_COR 4719,41 (5071,63) e MULHER 0,52 (0,53). As correlações de Pearson entre o IMC e as variáveis básicas na subamostra indicam: CIRC_BRE 0,872 ($p < 0{,}0001$); IDADE 0,140 ($p < 0{,}0001$); ENER_ADU 0,104 ($p < 0{,}0001$); ATV_FIS -0,145 ($p < 0{,}0001$); DCP_COR 0,167 ($p < 0{,}0001$) e MULHER 0,010 ($p < 0{,}0001$).

A primeira especificação do modelo, denominado "modelo inicial para o IMC" (Tabela 1), não incluía uma variável indicativa do estoque de energia acumulado pelo adulto. Em conseqüência suas variáveis independentes explicavam apenas 11,7% da variação linear do IMC ($R^2 = 0{,}1174$).

Com a introdução da variável de circunferência do braço esquerdo como indicadora do estoque de energia do adulto, foi feita uma segunda especificação, denominada "segundo modelo para o IMC" (Tabela 1), cuja explicação linear do IMC passou para 84,1% ($R^2 = 0{,}8408$). Este modelo, no entanto, não resistiu aos testes

Tabela 1

Parâmetros dos modelos para o Índice de Massa Corporal (IMC).

Variável	Modelo inicial para o IMC			Segundo modelo para o IMC		
	β	Erro padrão Ep (β)	H_0: β = 0 Prob > |t|	β	Erro padrão Ep (β)	H_0: β = 0 Prob > |t|
INTERCEP	16,746733	0,25368375	0,0001	0,975461	0,17823410	0,0001
CIRC_BRE				0,075813	0,00057621	0,0001
IDADE	0,198846	0,00893649	0,0001	0,000406	0,00387693	0,9166
IDADE2	-0,002044	0,00009659	0,0001	0,000209	0,00004200	0,0001
ENER_ADU	0,000447	0,00005427	0,0001	0,000088	0,00002313	0,0001
ATV_FIS	-0,000483	0,00012594	0,0001	-0,000692	0,00005350	0,0001
DCP_COR	0,000079	0,00000711	0,0001	0,000024	0,00000307	0,0001
MULHER	-1,251432	0,26785795	0,0001	-4,105411	0,21404505	0,0001
M_CIRCBR				0,021549	0,00071543	0,0001
M_IDADE	0,027818	0,00340921	0,0001	0,001454	0,00145183	0,3167
M_ENER	0,000350	0,00007352	0,0001	-0,000052	0,00003137	0,0989
M_ATVF	-0,000604	0,00027057	0,0255	-0,000588	0,00011492	0,0001
M_DCPCOR	-0,000055	0,00000890	0,0001	-0,000043	0,00000384	0,0001
RURAL	-0,866926	0,06075148	0,0001	0,036077	0,02606586	0,1664

Variável	Modelo final para o IMC			Final, após substituição de *outliers*		
	β	Erro padrão Ep (β)	H_0: β = 0 Prob > |t|	β	Erro padrão Ep (β)	H_0: β = 0 Prob > |t|
INTERCEP	18,340182	0,61728870	0,0001	18,183155	0,62485205	0,0001
CIRC_BRE	-0,048560	0,00375538	0,0001	-0,047477	0,00382102	0,0001
CIRCBRE2	0,000224	0,00000671	0,0001	0,000222	0,00000683	0,0001
IDADE	0,016622	0,00070827	0,0001	0,016631	0,00070853	0,0001
LOGENER	0,240355	0,04201976	0,0001	0,240685	0,04207233	0,0001
LOGATVF	-0,415923	0,02220208	0,0001	-0,415170	0,02230542	0,0001
DCP_COR	0,000019	0,00000287	0,0001	0,000020	0,00000315	0,0001
MULHER	-5,146741	0,18475139	0,0001	-5,132724	0,18519688	0,0001
M_CIRCBR	0,023550	0,00068815	0,0001	0,023533	0,00069131	0,0001
M_DCPCOR	-0,000035	0,00000357	0,0001	-0,000037	0,00000386	0,0001

IMC = índice de massa corporal; CIRC_BRE = circunferência do braço esquerdo (mm);
CIRCBRE2 = quadrado da circunferência do braço; IDADE = idade em anos completos;
ENER_ADU = ingestão energética por adulto-dia; ATV_FIS = energia para atividades físicas diárias;
DCP_COR = despesa corrente *per capita*; LOGENER = logaritmo da ingestão de energia por adulto-dia;
LOGATVF = logaritmo da energia para atividades físicas; MULHER = indicadora de sexo feminino;
M_CIRCBR = interação entre MULHER e CIRC_BRE; M_DCPCOR = interação entre MULHER e DCP_COR;
RURAL = indicadora de área rural; INTERCEP = coeficiente linear da regressão;
IDADE2 = quadrado da idade.

de especificação. Diversas formas polinomiais e logarítmicas foram então analisadas e, com apoio do teste RESET, fixou-se em um modelo final que combinava potências, logaritmos e interações, apresentado na Tabela 1. Nesse modelo final, as variáveis independentes conseguiram explicar 85,3% da variação linear do IMC ($R^2 = 0,8528$).

A análise dos resíduos do modelo final para detecção de pontos extremos foi feita usando os limites superiores de cada um dos quatro critérios descritos na seção anterior. Foram detectados 403 pontos extremos, representando 2,9% dos casos, todos com alto poder de alavancagem da função ajustada à medida que correspondiam a valores da diagonal da matriz H superiores ao limite adotado (3p/n). De acordo com os três outros critérios – resíduo *studentizado*, DFITS e D de Cook – não havia ponto extremo na subamostra para o modelo adotado.

Os pontos extremos encontrados foram substituídos de acordo com o procedimento de substituição descrito, o modelo foi ajustado para o novo conjunto de adultos e as técnicas descritas para detecção de pontos extremos foram reaplicadas. Isto foi repetido nove vezes,

até que não foi possível substituir um dos três pontos extremos, já que no estrato de subamostragem não havia mais adulto disponível para participar como substituto. Ao longo das nove repetições, todos os pontos extremos foram detectados por seu poder de alavancagem, não sendo constatada variação importante no coeficiente de determinação: a primeira repetição teve 183 pontos extremos e $R^2 = 0,8528$; a segunda 128 pontos e $R^2 = 0,8527$, a terceira 52 pontos e $R^2 = 0,8526$, a quarta 42 pontos e $R^2 = 0,8526$, a quinta 24 pontos e $R^2 = 0,8525$, a sexta 12 pontos e $R^2 = 0,8524$, a sétima 7 pontos e $R^2 = 0,8523$, a oitava 4 pontos e $R^2 = 0,8522$ e a nona 3 pontos e $R^2 = 0,8521$.

Os três pontos extremos da nona repetição foram então eliminados, o modelo foi novamente ajustado e, após a correção dos valores limites para alavancagem e resíduo *studentizado*, que dependem de n (o número de observações), foram reaplicadas as técnicas de detecção de pontos extremos. Nenhum ponto extremo foi identificado e os resultados constam da Tabela 1 com o rótulo Final, após substituição dos *outliers*. A comparação do modelo final para o IMC com o Final, após substituição de outliers, mostra que os parâmetros estimados para o modelo sem pontos extremos estão contidos nos intervalos de confiança dos parâmetros do modelo final. Assim, apesar de seu poder de alavancagem, os pontos extremos não têm influência significativa sobre os parâmetros do modelo final.

A tolerância (TOL) e o fator de inflação da variância (VIF) de cada variável do modelo final, elementos básicos para avaliação do grau de multicolinearidade, foram: CIRC_BRE com TOL = 0,00731445 e VIF = 136,71566681, CIRCBRE2 com TOL = 0,00761370 e VIF = 131,34215133, IDADE com TOL = 0,92598084 e VIF = 1,07993595, LOGENER com TOL = 0,94901164 e VIF = 1,05372785, LOGATVF com TOL = 0,53502612 e VIF = 1,86906764, DCP_COR com TOL = 0,34978224 e VIF = 2,85892158, MULHER com TOL = 0,01251112 e VIF = 79,92890795, M_CIRCBR com TOL = 0,01262542 e VIF = 79,20527603 e M_DCPCOR com TOL = 0,29298881 e VIF = 3,41309965.

Os testes RESET, para verificar a hipótese de inexistência de erros de especificação do modelo final, foram baseados em dois modelos diferentes. No primeiro, foram incluídas duas variáveis, o quadrado e o cubo do valor predito pelo modelo final, como indica a Tabela 2. Essas duas variáveis não foram significativas ao percentual de 5% e o valor do coeficiente de determinação ($R^2 = 0,8529$) não foi significativamente maior do que o do modelo final.

Tabela 2

Primeiro modelo do teste RESET.

Variável	β	Erro padrão Ep (β)	H_0: β = 0 Prob > |t|
INTERCEP	15,301430	4,45381100	0,0003
CIRC_BRE	-0,098216	0,02727673	0,0003
CIRCBRE2	0,000251	0,00008949	0,0051
IDADE	0,008116	0,00493471	0,1001
LOGENER	0,115070	0,08182505	0,1597
LOGATVF	-0,206534	0,12447948	0,0971
DCP_COR	0,000010557	0,00000645	0,1018
MULHER	-3,868245	1,73297462	0,0256
M_CIRCBR	0,016957	0,00779878	0,0297
M_DCPCOR	-0,000018696	0,00001104	0,0903
PREDITO2	0,050461	0,10726715	0,6380
PREDITO3	-0,000885	0,00630498	0,1602

CIRC_BRE = circunferência do braço esquerdo (mm);
CIRCBRE2 = quadrado da circunferência do braço;
IDADE = idade em anos completos; DCP_COR = despesa corrente *per capita*;
LOGENER = logaritmo da ingestão de energia por adulto-dia;
LOGATVF = logaritmo da energia para atividades físicas;
MULHER = indicadora de sexo feminino; M_CIRCBR = interação entre MULHER e CIRC_BRE; M_DCPCOR = interação entre MULHER e DCP_COR;
INTERCEP = coeficiente linear da regressão;
PREDITO2 = quadrado do valor predito; PREDITO3 = cubo do valor predito.

A estatística *F* do teste RESET foi igual a

$$F = \frac{(0,8528693 - 0,8528137) / (12 - 10)}{(1 - 0,8528693) / (13996 - 12)} = 2,64224$$

que é menor que $F(2;13984;0,05) = 3,00$, não havendo motivo para rejeitar a hipótese nula de que o modelo não tem erros de especificação.

No segundo modelo para o teste RESET, foram introduzidas trinta variáveis relativas a: potências das variáveis não-logarítmicas do modelo final (CIRCBR3, CIRCBR4, IDADE2, IDADE3, IDADE4, DCP_COR2, DCP_COR3 e DCP_COR4) e, todas as interações de primeira ordem entre as variáveis do modelo final e as de potências introduzidas. O aumento do valor do coeficiente de determinação também não foi significativo nesse caso e a estatística *F* foi igual a

$$F = \frac{(0,8532283 - 0,8528137) / (40 - 10)}{(1 - 0,8532283) / (13996 - 40)} = 1,314095,$$

que comparada a $F(30;13956;0,05) = 1,46$, não rejeita a hipótese nula de que o modelo não tem erro de especificação.

O primeiro teste usado para verificar a pressuposição de homocedasticidade foi o teste de Park. A regressão do logaritmo do quadrado do resíduo pelos logaritmos das variáveis explica-

tivas do modelo final indicou um $R^2 = 0,0017$ e dois parâmetros significativos ($\alpha = 0,05$) como mostra a Tabela 3. Assim, foi rejeitada a hipótese de homocedasticidade dos erros aleatórios da função de regressão.

O teste baseado nas correlações de Spearman entre os resíduos e as variáveis explicativas, também conduziu à rejeição da hipótese de homocedasticidade, pois havia correlação significativa entre a circunferência do braço e os resíduos do modelo final (Tabela 3).

A regressão entre os quadrados dos resíduos padronizados e as variáveis explicativas do modelo final indica que a soma de quadrados explicada (SQE) por esse modelo, corresponde a 186,59313 de um total de 22746,80403. Assim, o teste de Breusch-Pagan conduz à rejeição da hipótese de homocedasticidade (SQE/2 = 93,297 é maior que $\chi^2_{0,05;9gl} = 16,919$).

Para correção da heterocedasticidade foram calculados os pesos baseados no inverso do valor predito para o resíduo. O método dos mínimos quadrados ponderados por esses pesos foi utilizado para estimar novamente o modelo final. Com o uso da ponderação, a explicação linear das variáveis independentes sobre o IMC reduziu-se de 85,28% para 82,40%. Além disso, os resultados obtidos (Tabela 4) mostraram que as diferenças nos valores dos parâmetros entre o modelo final não-ponderado e o ponderado não foram significativas a 5%.

Para assegurar que o processo adotado eliminou os vestígios de heterocedasticidade, os mesmos três testes (Park, Spearman e Breusch-Pagan) foram aplicados. Na regressão proposta por Park, observou-se que nenhum parâmetro foi significativo (Tabela 3), o que indica que o teste de Park não rejeita a hipótese nula de ho-

Tabela 3

Resultados relativos aos testes de heterocedasticidade.

Variável	**Regressão de Park** $loge_i^2 = \alpha + \beta_2 logX_{2i} + \beta_3 logX_{3i} + ...\beta_p logX_{pi} + v_i$					
	Modelo final			**Modelo final ponderado**		
	β	Erro padrão Ep (β)	H_0: β = 0 Prob > |t|	β	Erro padrão Ep (β)	H_0: β = 0 Prob > |t|
INTERCEP	2,032424	1,99277139	0,3078	1,588836	1,99406891	0,4256
LOGCIRCB	0,112484	0,34043183	0,7411	0,260612	0,34065348	0,4443
LOGIDADE	**-0,145737**	0,06389710	**0,0226**	-0,111008	0,06393871	0,0826
LLOGENER	0,647075	0,76879075	0,4000	0,594053	0,76929132	0,4400
LLOGATVF	**-0,886806**	0,30674876	**0,0038**	-1,074179	0,65529722	0,0859
LOGDESP	-0,035460	0,04373207	0,4175	-0,030590	0,04376055	0,4845
LOGMULH	0,705483	0,57883018	0,2229	0,710024	0,57920706	0,2203
LOGMCIRC	-0,406789	0,42459550	0,3380	-0,358338	0,42487196	0,3990
LOGMDESP	-0,091290	0,05527643	0,0987	-0,137381	0,08704615	0,1145

Variáveis	**Correlações de Spearman**			
	Modelo final		**Modelo final ponderado**	
	Correlação (ρ)	H_0: ρ =0; Prob > |t|	Correlação (ρ)	H_0: ρ =0; Prob > |t|
CIRC_BRE	**0,03424**	**0,0001**	0,01606	0,0574
IDADE	-0,01377	0,1032	-0,00960	0,2563
LOGENER	-0,00289	0,7326	-0,00384	0,6500
LOGATVF	0,00250	0,7673	-0,00086	0,9194
DCP_COR	-0,01511	0,0739	-0,01613	0,0563
MULHER	-0,00517	0,5406	-0,00479	0,5708
M_CIRCBR	0,00296	0,7258	0,00442	0,6015
M_DCPCOR	-0,01368	0,1057	-0,01444	0,0875

CIRC_BRE = circunferência do braço esquerdo (mm); IDADE = idade em anos completos;
DCP_COR = despesa corrente *per capita*; LOGENER = logaritmo da ingestão de energia por adulto-dia;
LOGATVF = logaritmo da energia para atividades físicas; MULHER = indicadora de sexo feminino;
M_CIRCBR = interação entre MULHER e CIRC_BRE; M_DCPCOR = interação entre MULHER e DCP_COR;
INTERCEP = coeficiente linear da regressão; LLOGENER = logaritmo de LOGENER;
LLOGATVF = logaritmo de LOGATVF.

mocedasticidade com nível de significância de 5%. De forma análoga, o teste baseado nas correlações de Spearman entre os resíduos e as variáveis explicativas do modelo final ponderado não fornece indicação para rejeição da hipótese de homocedasticidade, dado que nenhuma das correlações foi significativa a 5% (Tabela 3).

Da mesma forma que nos dois testes anteriores, o de Breusch-Pagan não conduz à rejeição da hipótese de homocedasticidade com nível de significância de 5%, pois SQE/2 = 28,71094/2 = 14,35547 é menor que $\chi^2_{0,05;9gl}$ = 16,919.

Tabela 4

Modelo final ponderado.

Variável	β	Erro padrão Ep (β)	H_0: β = 0 Prob > |t|
INTERCEP	18,499860	0,64975313	0,0001
CIRC_BRE	-0,049742	0,00412424	0,0001
CIRCBRE2	0,000226	0,00000755	0,0001
IDADE	0,015927	0,00068638	0,0001
LOGENER	0,223709	0,04132399	0,0001
LOGATVF	-0,390054	0,02185741	0,0001
DCP_COR	0,000018	0,00000308	0,0001
MULHER	-5,102292	0,18999970	0,0001
M_CIRCBR	0,023329	0,00072233	0,0001
M_DCPCOR	-0,000028	0,00000330	0,0001

CIRC_BRE = circunferência do braço esquerdo (mm);
CIRCBRE2 = quadrado da circunferência do braço;
IDADE = idade em anos completos; DCP_COR = despesa corrente *per capita*;
LOGENER = logaritmo da ingestão de energia por adulto-dia;
LOGATVF = logaritmo da energia para atividades físicas;
MULHER = indicadora de sexo feminino; M_CIRCBR = interação entre MULHER e CIRC_BRE; M_DCPCOR = interação entre MULHER e DCP_COR;
INTERCEP = coeficiente linear da regressão.

Discussão

A subamostra selecionada foi comparada com a amostra do ENDEF por meio de estimativas das médias das variáveis básicas, mostrando pequena variação dos valores, o que indica que a subamostra utilizada não apresenta tendência em relação às variáveis consideradas. As correlações de Pearson entre o IMC e as variáveis básicas do modelo são todas significativas. A correlação entre o IMC e a circunferência do braço esquerdo justifica o uso desta variável como indicador do estoque atual de energia acumulada na massa corporal dos adultos. As associações entre IMC e as variáveis de ingestão de energia, despesa per capita, idade e sexo, apesar das correlações lineares baixas com o IMC, existem e foram justificadas, para toda a população de 18 anos ou mais em 1975 (Vasconcellos, 1994). Assim, há uma base empírica que sustenta o uso da subamostra de adultos e a escolha das variáveis incluídas na modelagem.

A forma funcional da associação entre o IMC e essas variáveis, no entanto, não é indicada por resultados prévios, exceto para a idade. Apesar da indicação de um comportamento parabólico entre o IMC e a idade (Vasconcellos, 1994), o modelo final não validou essa indicação, sobretudo porque esse comportamento já era levado ao modelo pela variável de circunferência do braço, que é um indicador da massa corporal e, portanto, varia com a idade e seu quadrado. A questão ligada à necessidade de incluir na modelagem um indicador do estoque de energia do adulto fica evidente pela comparação do modelo inicial para o IMC com os demais modelos apresentados na Tabela 1.

A análise da influência dos pontos extremos sobre o modelo final mostrou que a substituição de adultos que poderiam ter alto poder de alavancagem da curva de regressão não influiu, de forma significativa, sobre os resultados obtidos a partir da subamostra originalmente selecionada. Esse fato pode ser visto como um indicador de que o modelo representa um grupo maior de adultos brasileiros, mas a impossibilidade de incluir, nesta análise, a aleatorização decorrente da modelagem amostral do ENDEF não autoriza que as inferências transcendam à subamostra utilizada.

De fato, a subamostra usada tem como função maior definir um conjunto de adultos selecionados de forma aleatória (e não escolhidos por algum critério determinístico) e que respeitou as distribuições da população brasileira por sexo, classe de idade, estrato geográfico e classe de tamanho de domicílio. A autoponderação decorreu, portanto, desse desejo de manter na subamostra as mesmas proporções de adultos observadas na população por estrato de subamostragem.

Voltando ao modelo final estimado, observou-se que os fatores de inflação da variância (VIF) eram baixos para todas as variáveis, exceto naquelas para as quais era esperado e aceitável valores maiores, como as polinomiais (CIRC_BRE e CIRCBRE2), a *dummy* MULHER e a interação entre mulher e circunferência do braço esquerdo.

Os testes para erros de especificação do modelo final não conduziram à rejeição da hipótese de inexistência de erros, enquanto os três testes para a hipótese de homocedasticidade foram significativos a 5%. Feita a correção da heterocedasticidade pelo método dos mínimos

quadrados ponderados, constatou-se que as novas estimativas dos parâmetros (Tabela 4) não diferiram significativamente das estimativas originais (Tabela 1) e que a correção de heterocedasticidade foi efetiva, na medida em que os mesmos três testes foram reaplicados e não conduziram à rejeição da hipótese de que os resíduos eram homocedásticos.

Aceitar que existe heterocedasticidade significativa, que uma vez corrigida, não altera significativamente os resultados é uma situação no mínimo curiosa. De fato, pode ser uma indicação de que os testes aplicados foram excessivamente rigorosos, pois em amostras grandes, qualquer pequena variação tem alta probabilidade de ser considerada significativa por testes desenvolvidos para lidar com pequenas amostras. No teste de Spearman (Tabela 3), por exemplo, uma correlação de 0,034 foi significativamente diferente de 0, apesar de ser uma correlação muito pequena.

Cabe ainda ressaltar que a modelagem realizada não tem qualquer uso prático para fins de predição, pois quem dispõe de dados de pesquisas de consumo alimentar do tipo usado certamente possui informação sobre a massa corporal e a estatura das pessoas e pode, portanto, calcular o IMC diretamente dos dados. Assim, o modelo estimado tem sentido para mostrar a existência, o grau e a forma funcional da associação entre o IMC e as demais variáveis utilizadas.

Referências

ANJOS, L. A., 1992. Índice de massa corporal (massa corporal.estatura^{-2}) como indicador do estado nutricional de adultos: Uma revisão da literatura. *Revista de Saúde Pública*, 6:431-436.

BELSLEY, D. A.; KUH, E. & WELSCH, R. E., 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley.

BOLLEN, K. A. & JACKMAN, R. W., 1990. Regression diagnostics: An expository treatment of outliers and influential cases. In: *Modern Methods of Data Analysis* (J. Fox & J. S. Long, ed.), pp. 257-291, Newbury Park: Sage Publications.

COOK, R. D., 1977. Detection of influential observations in linear regression. *Technometrics*, 19:15-18.

DURNIN, J. V. G. A. & PASSMORE, R., 1967. *Energy, Work and Leisure*. London: Heinemann Education Books.

FAO (Food and Agriculture Organization of the United Nations)/WHO (World Health Organization), 1973. *Energy and Protein Requirements: Report of a Joint FAO/WHO ad hoc Expert Consultation*. Nutrition Meeting Report Series 52. Rome: FAO.

FERRO-LUZZI, A.; SETTE, S.; FRANKLIN, M. & JAMES, W. P. T., 1992. A simplified approach to assessing adult chronic energy deficiency. *European Journal of Clinical Nutrition*, 46:173-186.

FRANÇOIS, P. J., 1970. Food consumption surveys: Study of a general formula for the estimation of per caput, household and group consumption. *FAO Nutrition Newsletter* 8:35-58.

GOSSET, W. S., 1908. The probable error of a mean. *Biometrika*, 6:1-25.

GUJARATI, D. N., 1988. *Basic Econometrics*. 2nd Ed. New York: McGraw-Hill.

HÁJEK, J., 1960. Limiting distribution in simple random sampling from a finite population. *Publications of the Mathematics Institute of the Hungarian Academy of Science*, 5:361-374.

IDECG (International Dietary Energy Consultative Group), 1994. Functional significance of low body mass index. *European Journal of Clinical Nutrition*, 48(Sup. 3):S1-S202.

JAMES, W. P. T.; FERRO-LUZZI, A. & WATERLOW, J. C., 1988. Definition of chronic energy deficiency in adults. *European Journal of Clinical Nutrition*, 42(Sup. 2):S969-S981.

PARK, R. E., 1966. Estimation with heteroscedastic error terms. *Econometrica*, 34:888-896.

PESSOA, D. G. C. & NASCIMENTO-SILVA, P. L., 1998. *Análise de Dados Amostrais Complexos*. São Paulo: Associação Brasileira de Estatística.

QUÉTELET, A., 1869. *Physique Sociale ou Essai sur le Développement des Facultés de l'Homme*. Bruxelles: C. Muquardt.

RAMSEY, J. B., 1969. Tests for specification errors in classical linear least squares regression analysis. *Journal of the Royal Statistical Society, Series B*, 31: 350-371.

SHETTY, P. S. & JAMES, W. P. T., 1994. *Body Mass Index: A Measure of Chronic Energy Deficiency in Adults*. Food and Nutrition Paper, 56. Rome: Food and Agricultural Organization of the United Nations.

VASCONCELLOS, M. T. L., 1983a. *Objetivos, Descrição e Metodologia Usada no ENDEF*. Rio de Janeiro: Fundação Instituto Brasileiro de Geografia e Estatística.

VASCONCELLOS, M. T. L., 1983b. *Núcleo do Banco de Informações ENDEF*. Rio de Janeiro: Fundação Instituto Brasileiro de Geografia e Estatística.

VASCONCELLOS, M. T. L., 1994. Body mass index: Its relationship with food consumption and socioeconomic variables in Brazil. *European Journal of Clinical Nutrition*, 48(Sup. 3):S115-S123.