UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM EPIDEMIOLOGIA

DISSERTAÇÃO DE MESTRADO

Métodos para estimar prevalências ajustadas

Natália Bordin Barbieri

Orientador: Prof. Dr. Álvaro Vigo

Porto Alegre, fevereiro de 2016

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM EPIDEMIOLOGIA

DISSERTAÇÃO DE MESTRADO

Métodos para estimar prevalências ajustadas

Natália Bordin Barbieri

Orientador: Prof. Dr. Álvaro Vigo

A apresentação desta dissertação é exigência do Programa de Pós-graduação em Epidemiologia, Universidade Federal do Rio Grande do Sul, para obtenção do título de Mestre.

Porto Alegre, Brasil.
2016

BANCA EXAMINADORA

Prof.ª Dra. Vivian Cristine Luft
Faculdade de Medicina
PPG em Epidemiologia
Universidade Federal do Rio Grande do Sul

Prof.ª Dra. Suzi Alves Camey
Instituto de Matemática e Estatística
PPG em Epidemiologia
Universidade Federal do Rio Grande do Sul

Prof. Dr. Cleber Bisognin
Instituto de Matemática e Estatística
Departamento de Estatística
Universidade Federal do Rio Grande do Sul

AGRADECIMENTOS

Agradeço aos meus pais, Jeferson e Marfisa, e aos meus irmãos, Renata e Tomás, que sempre me apoiaram, incentivaram e oportunizaram para que eu pudesse chegar até aqui. Eles são a base de tudo.

Ao Lázaro Ribeiro Luz pelo companheirismo e amor em nossas jornadas.

Ao Professor Álvaro Vigo, meu orientador, pelo exemplo de pessoa e profissional. Se um dia eu conseguir chegar perto de onde ele chegou, a caminhada valeu a pena.

Aos Professores Maria Inês Schmidt e Bruce Duncan, pelas oportunidades ao longo do tempo. É uma honra poder trabalhar com vocês.

Ao Professor Loyd Chambless pela ajuda e discussão sobre os métodos utilizados.

Aos professores Cleber Bisognin, Suzi Camey e Vivian Luft, pela oportunidade de tê-los como banca neste trabalho.

A Paula Sientchkovski pelo companheirismo desde os tempos da graduação.

E a querida grande Equipe de Estatística ELSA-Brasil, que oportuniza a cada dia o crescimento pessoal e profissional, e que fazem a palavra "equipe" valer muito a pena.

SUMÁRIO

ABREVIATURAS E SIGLAS

ARIC	The Atherosclerosis Risk in Communities Study
ELSA-Brasil	Estudo Longitudinal da Saúde do Adulto
ICHD-3	International Classification of Headache Disorders

RESUMO

Objetivo: Apresentar e discutir métodos para estimar prevalências ajustadas em pesquisas clínicas e epidemiológicas, bem como desenvolver rotinas computacionais em SAS e R.

Métodos: No contexto de estudo transversal, foi simulada uma amostra de 2.000 observações independentes, considerando o desfecho dicotômico diabetes, sexo como a variável de exposição e idade como variável de ajuste. As estimativas de prevalências ajustadas (IC 95%) foram estimadas pelos métodos de predição condicional e marginal, utilizando as rotinas desenvolvidas em SAS e R. O método Delta foi usado para construir os intervalos de confiança. Os resultados foram comparados com aqueles do SUDAAN (SAS-Callable), Stata e a macro %ADJ_PROP (SAS).

Resultados: No exemplo simulado, 68,2% são do sexo feminino e a idade média (DP) foi 57,6 (5,0) anos, sendo 54,2 (3,9) anos em homens e 59,2 (4,6) anos em mulheres. A estimativa da prevalência global do desfecho foi de 25,3% (IC 95%:23,4-27,3); sendo 13,8% (IC 95%:11,7-16,7) e 30,7% (IC 95%:28,3-33,2), respectivamente para homens e mulheres. As estimativas de prevalências ajustadas por idade, por meio do método de predição condicional, foram de 19,6% (IC 95%:16,2-23,6) para homens, e 23,6% (IC 95%:21,2-26,1) para mulheres. Pelo método de predição marginal, as estimativas foram de 22,4% (IC 95%:18,7-26,5) para homens, e 26,3% (IC 95%:24,1-28,6) para mulheres.

Conclusão: A discrepância entre as estimativas não ajustadas é devida ao confundimento pela idade. Estimativas livres de confundimento podem ser obtidas por meio das prevalências ajustadas pela idade. No entanto, a estimativa pelo método de predição condicional não engloba a prevalência global. Em virtude disso, o método de predição marginal é, geralmente, mais adequado. A rotina desenvolvida na versão para R é uma alternativa aos softwares comerciais.

Palavras-chave: Prevalências ajustadas, Método de predição condicional, Método de predição marginal, Método Delta.

ABSTRACT

Objective: To present and discuss methods to estimate adjusted prevalences for clinical and epidemiological research, and develop computational routines in SAS and R.

Methods: In the context of cross-sectional study, it was simulated a sample of 2,000 independent observations, considering the dichotomous outcome diabetes, sex as the exposure variable and age as an adjustment variable. Adjusted prevalences were estimated by the conditional and marginal methods, using routines developed in SAS and R. Confidence intervals were constructed using the Delta method. The results were compared with those of the SUDAAN (SAS-callable), Stata and macro %ADJ_PROP (SAS).

Results: In simulated example, 68.2% are female and the mean (SD) age was 57.6 (5.00) years old, being that 54.2 (3.94) years for men and 59.2 (4.60) years in women. The estimated global prevalence of outcome was 25.3% (CI 95%: 23.4-27.3) and 13.8% (CI 95%: 11.7-16.7) and 30.7% (CI 95%: 28.3-33.2), respectively for men and women. Estimates of adjusted prevalence for age, through the conditional method, were 19.6% (CI 95%: 16.2-23.6) for men, and 23.6% (CI 95%: 21,2-26.1) for women. For marginal method, the estimates were 22.4% (CI 95%: 18.7-26.5) for men and 26.3% (CI 95%: 24.1-28.6) for women.

Conclusion: The observed discrepancy in estimates by sex, unadjusted, can be attributed to confounding due to difference in age distribution between sexes. Comparable estimates (without confounding) of the prevalences can be obtained through prevalence adjusted for age. However, the estimate for the conditional method does not comprise the global prevalence. As a result, the marginal method is in general more suitable. The developed routines can be useful for estimating adjusted prevalences, particularly the R version (an alternative to commercial software).

Keywords: Adjusted Prevalence, Conditional prediction method, Marginal prediction method, Delta method.

1. APRESENTAÇÃO

Este trabalho consiste na dissertação de mestrado intitulada **“Métodos para estimar prevalências ajustadas”,** apresentada ao Programa de Pós-Graduação em Epidemiologia da Universidade Federal do Rio Grande do Sul, em 23 de fevereiro de 2016. O trabalho é apresentado em quatro partes, na ordem que segue:

1. Introdução, Revisão da Literatura e Objetivo.
2. Exemplo
3. Artigo
4. Conclusões e Considerações finais.

Documentos de apoio estão apresentados no apêndice.

2. INTRODUÇÃO

Estimativas de prevalências são reportadas com grande frequência em estudos clínicos ou epidemiológicos. Para que essas estimativas sejam comparáveis em diferentes populações, é necessário levar em conta os desequilíbrios nas distribuições de variáveis que possam alterar a frequência do evento. Em outras palavras, na presença de confundimento, estimativas não ajustadas podem não ser adequadas, e podem dificultar a compreensão do fenômeno.

Reportar prevalências ajustadas possibilita ao leitor um julgamento da frequência do evento livre de possíveis confundimentos, uma vez que a magnitude da associação entre desfecho e o fator em estudo pode variar entre os diferentes níveis do fator. Estratificação e modelos de ajuste multivariável são ferramentas comuns utilizadas para controlar os efeitos de confundimento (Szklo e Nieto, 2007).

Em uma revisão de programas de análise de dados foi constatado que somente o SUDAAN disponibiliza, de forma direta, estimativas de prevalências ajustadas (e intervalos de confiança) por meio dos métodos marginal e condicional. O Stata fornece apenas estimativas pelo método condicional. Para os programas SAS e R, é necessário o uso de rotinas específicas (macros ou funções).

Essa dissertação apresenta uma revisão dos diferentes métodos para estimar prevalências ajustadas, bem como aspectos computacionais disponíveis. Também são apresentadas rotinas computacionais em SAS e R, desenvolvidas para esta finalidade. Um exemplo é utilizado para explorar e discutir os métodos estatísticos e computacionais.

3. REVISÃO DA LITERATURA

Nas seções a seguir foram detalhados os métodos de estimação de prevalências ajustadas e também as rotinas computacionais disponíveis.

3.1 Estimação de prevalências ajustadas

Em estudos clínicos ou epidemiológicos muitas vezes é necessário estimar prevalências de um evento/doença. Para que essas estimativas sejam comparáveis em diferentes populações, é necessário levar em conta os desequilíbrios das distribuições das variáveis/fatores importantes que possam alterar a frequência da doença. Reportar prevalências ajustadas possibilita ao leitor um julgamento da frequência do evento livre de possíveis confundimentos, uma vez que a magnitude da associação entre desfecho e o fator em estudo pode variar entre os diferentes níveis do fator. Estratificação e modelos de ajuste multivariável são ferramentas comuns utilizadas para controlar os efeitos de confundimento (Szklo e Nieto, 2007).

O uso de prevalências ajustadas é frequente na literatura. Como exemplo, An (2015) utilizou prevalências ajustadas para descrever o diagnóstico de diabetes e leitura de rótulos nutricionais entre adultos norte-americanos para suas escolhas alimentares. Para obter as prevalências ajustadas, foi utilizado o modelo de regressão logística, porém não foi especificado o método utilizado.

Shon *et al*. (2015) também utilizaram prevalências ajustadas para comparar e analisar a prevalência de infecção pelo vírus da hepatite C por região na República da Coréia durante os anos de 2005-2012. O método utilizado para a prevalência ajustada foi o método direto de padronização (*direct standardization method*), utilizando como referência para o ajuste de idade a população de 2010. As prevalências foram ajustadas para sexo, idade e região, e foi possível identificar regiões onde a doença tinha uma maior prevalência.

Lebedeva *et al*. (2015) estimaram a prevalência de transtornos de dor de cabeça primária, diagnosticados de acordo com a ICHD-3 beta (*International Classification of*

Headache Disorders), em três grupos sociais diferentes. Para obter as prevalências ajustadas, utilizaram o método direto de padronização.

Esses são alguns exemplos que ilustram o uso de prevalências ajustadas, tendo sido estimadas por diferentes métodos. Entre as abordagens utilizadas estão padronização direta (*direct adjustment*), padronização indireta (*indirect adjustment*) e modelos multivariáveis.

A padronização direta utiliza as taxas observadas na população em estudo (que pode ser afetada por instabilidade, ocasionada por tamanho amostral pequeno), enquanto que a padronização indireta utiliza taxas de uma população de referência, e que por isso tem sua aplicação limitada a essa comparação de referência. A vantagem da padronização direta é que as taxas estimadas podem ser comparadas entre grupos (por exemplo, comparar o grupo feminino *versus* masculino), enquanto a padronização indireta permite comparar cada grupo estimado apenas com sua taxa utilizada da população de referência - por exemplo, comparar a taxa do sexo masculino estimada com a taxa do sexo masculino da população de referência (Woodward, 2014).

Em estudos ecológicos, o método de padronização direta é preferível uma vez que a informação existente é sobre a população. Sempre que tivermos dados no nível individual, é preferível utilizar ajuste multivariável. (Szklo e Nieto, 2007).

Wilcosky e Chambless (1985) afirmam que as principais vantagens do método de ajuste direto são a simplicidade computacional e as poucas suposições estatísticas. Entretanto, com a maior disponibilidade e diversidade de recursos computacionais, esses aspectos atualmente não são mais relevantes, além de ter como desvantagem a necessidade de categorização de variáveis quantitativas. Os autores também discutiram os métodos de predição condicional ("*conditional prediction method*") e marginal ("*marginal prediction method*") para estimar prevalências ajustadas. Esses métodos utilizam o modelo de regressão logística multivariável e são mais convenientes para testes de interações e de diferenças entre grupos, permitem explorar a natureza da relação funcional entre as variáveis de controle e o desfecho, e não exigem a

categorização de variáveis quantitativas. Lane e Nelder (1982) também apresentaram a abordagem de ajuste multivariável em modelos lineares generalizados para estimar proporções ajustadas por meio dos métodos de predição marginal e condicional.

Os métodos propostos por Wilcosky e Chambless (1985) têm sido bastante utilizados. Para estimar a associação entre incidência da doença arterial coronariana e espessura da parede da artéria carótida, Chambless *et al*. (1997) apresentaram prevalências dos principais fatores de risco, ajustadas por idade, centro de investigação e raça. No estudo sobre lipoproteína fosfolipase A_2 associada e alta sensibilidade da proteína c-reativa, esse método foi utilizado por Nambi *et al*. (2009) para melhorar a estratificação de risco de acidente vascular isquêmico, tendo sido estimadas prevalências de características (como hipertensão, diabetes, e uso de aspirina, estatina, etc.) na linha de base, ajustadas por idade, raça e sexo.

Chor *et al. (2015)* utilizaram o método de predição marginal para estimar a prevalência de pressão arterial elevada, ajustado por idade e sexo, com dados do ELSA-Brasil.

Ohira *et al.* (2010) apresentaram prevalências de diferentes variáveis de linha de base do estudo *ARIC (The Atherosclerosis Risk in Communities Study),* ajustadas por idade e raça, em participantes com e sem risco de tromboembolismo venoso. Não foi explicitado se foi utilizado o método de predição marginal ou condicional.

Para estudar fatores associados a patologias nas cordas vocais em professores, Souza *et al. (2011),* utilizaram o método de predição condicional para obter prevalências ajustadas para cada grupo de interesse.

Epstein *et al*. (2013a, 2013b) utilizaram o método de predição marginal para estimar a prevalência de uso de medicamentos antipsicóticos e anticonvulsantes e de analgésicos opióides durante a gestação, ao longo do tempo. As prevalências em cada ano, expressadas por mil gestações, foram ajustadas para diferentes variáveis maternas.

Os métodos de predição condicional e marginal são descritos nas seções 3.1.1 e

3.1.2, respectivamente.

3.1.1 Método de Predição Condicional

Para descrever estes métodos será considerado o contexto de um estudo transversal em que a variável dependente (Y) representa presença $(Y=1)$ ou ausência $(Y=0)$ de diabetes, a variável dicotômica (x_1) representa sexo (1=masculino, 0=feminino), e a variável x_2 representa idade (quantitativa). Naturalmente, pode ser generalizado para mais preditores. Dada uma amostra aleatória com n indivíduos, obtida por meio de um estudo transversal, deseja-se estimar a prevalência do desfecho para cada categoria da exposição, ajustada pela variável x_2. O modelo logístico definido na equação [1] pode ser usado para essa finalidade, ou seja,

$$\log\frac{P[Y=1|\mathbf{x}]}{1-P[Y=1|\mathbf{x}]}=\beta_0+\beta_1x_1+\beta_2x_2 \qquad [1]$$

em que $\mathbf{x}=(x_1,x_2)^t$ representa o vetor de covariáveis do modelo. O método de estimação de máxima verossimilhança pode ser usado para estimar os parâmetros do modelo, obtendo-se o modelo estimado descrito a seguir

$$\log\frac{P[Y=1|\mathbf{x}]}{1-P[Y=1|\mathbf{x}]}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2 \qquad [2].$$

Detalhes sobre o método de estimação podem ser obtidos em Mcculagh e Nelder (1989). As probabilidades estimadas pelo modelo em [2] são obtidas por

$$\hat{p}_{ij}(x_{1i},x_{2i})=\hat{P}[Y=1|\,x_{1i}=j,x_{2i}]=\frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1\times j+\hat{\beta}_2x_{2i})}} \qquad [3]$$

em que $j=0,1$ e $i=1,2,...,n$.

Para obter estimativas de prevalências ajustadas pelo método de predição condicional ($\hat{p}_{Ci}$), a média amostral do preditor x_2, denotada por $\bar{x}_2$, é usada na equação [3] para cada categoria da variável x_1, como descrito nas equações [4] e [5]. Detalhes do método podem ser encontrados no manual do programa SUDAAN (SUDAAN Language Manual, 2012).

$$\hat{p}_{C_0}(\mathbf{x}) = \hat{P}[Y=1 \mid x_1 = 0, x_2 = \bar{x}_2] = \frac{1}{1+e^{-(\hat{\beta}_0 + \hat{\beta}_1 \times 0 + \hat{\beta}_2 \bar{x}_2)}} = \frac{1}{1+e^{-(\hat{\beta}_0 + \hat{\beta}_2 \bar{x}_2)}}, \quad [4]$$

e

$$\hat{p}_{C_1}(\mathbf{x}) = \hat{P}[Y=1 \mid x_1 = 1, x_2 = \bar{x}_2] = \frac{1}{1+e^{-(\hat{\beta}_0 + \hat{\beta}_1 \times 1 + \hat{\beta}_2 \bar{x}_2)}} = \frac{1}{1+e^{-(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 \bar{x}_2)}}. \quad [5]$$

3.1.2 Método de Predição Marginal

Adotando a notação da seção anterior, as prevalências ajustadas pelo método de predição marginal do desfecho para as categorias da exposição são estimadas pelas equações [6] e [7]:

$$\hat{p}_{M_0}(\mathbf{x}) = \hat{P}[Y=1 \mid x_1 = 0, x_{2i}] = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1+e^{-(\hat{\beta}_0 + \hat{\beta}_1 \times 0 + \hat{\beta}_2 x_{2i})}} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1+e^{-(\hat{\beta}_0 + \hat{\beta}_2 x_{2i})}}, \quad [6]$$

$$\hat{p}_{M_1}(\mathbf{x}) = \hat{P}[Y=1 \mid x_1 = 1, x_{2i}] = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1+e^{-(\hat{\beta}_0 + \hat{\beta}_1 \times 1 + \hat{\beta}_2 x_{2i})}} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1+e^{-(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 x_{2i})}}. \quad [7]$$

As estimativas do intervalo de confiança para as prevalências ajustadas dependem das variâncias e covariâncias das estimativas dos parâmetros do modelo estimado. Aspectos essenciais do processo de construção desses intervalos de confiança são descritos na próxima seção.

3.1.3 Estimação por intervalo utilizando método Delta

As probabilidades estimadas pelo modelo são funções dos estimadores de máxima verossimilhança, podendo ser reescritas como

$$h(\hat{\boldsymbol{\beta}}) = \hat{p}(\mathbf{x}) = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2)}}, \quad [8]$$

em que $\mathbf{x} = (x_1, x_2)^t$ e $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)^t$ representa os estimadores de máxima verossimilhança no modelo definido em [2]. A variância de $h(\hat{\boldsymbol{\beta}})$ também é função dos estimadores de máxima verossimilhança, não sendo possível estimá-la de forma analítica. Nestas situações, o método Delta é um procedimento geral e flexível para construção de intervalos de confiança. Este método está extensamente detalhado na literatura, podendo-se destacar Xu e Long (2005) para sua aplicação na predição de probabilidades.

A função $h(\hat{\boldsymbol{\beta}})$ pode ser expandida em série de Taylor até o termo de primeira ordem, mostrada abaixo, que geralmente produz estimativas relativamente acuradas para a variância:

$$h(\hat{\boldsymbol{\beta}}) \approx h(\boldsymbol{\beta}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^t \, h'(\boldsymbol{\beta}), \quad [9]$$

em que $h'(\boldsymbol{\beta}) = \frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$. Assim,

$$\sqrt{n}\left[h(\hat{\boldsymbol{\beta}}) - h(\beta)\right] \approx \sqrt{n}\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^t \frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right], \quad [10]$$

tal que,

$$h(\hat{\boldsymbol{\beta}}) \rightarrow N\left(h(\boldsymbol{\beta}), \frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^t} Var(\hat{\boldsymbol{\beta}}) \frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right). \quad [11]$$

Para estimar a variância de $h(\hat{\boldsymbol{\beta}})$, as derivadas parciais $\frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^t}$ são avaliadas em $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, tal que

$$Var\left(h(\hat{\boldsymbol{\beta}})\right) = \left(\frac{\partial h(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}^t}\right) Var(\hat{\boldsymbol{\beta}}) \left(\frac{\partial h(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}}\right) = \left(\nabla h(\hat{\boldsymbol{\beta}})\right)^t \hat{\boldsymbol{\Sigma}} \left(\nabla h(\hat{\boldsymbol{\beta}})\right), \quad [12]$$

em que $\hat{\Sigma}$ é a matriz de estimativas de variâncias e covariâncias dos parâmetros do modelo. As derivadas parciais de primeira ordem da equação [9] são mostradas abaixo:

$$\frac{\partial h(\hat{\boldsymbol{\beta}})}{\partial \beta_0} = \frac{e^{-(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2)}}{\left[1+e^{-(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2)}\right]^2} = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2)}} \times \frac{e^{-(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2)}}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2)}} = \hat{p}(\mathbf{x}) \times (1-\hat{p}(\mathbf{x})). \quad [13]$$

Similarmente,

$$\frac{\partial h(\hat{\boldsymbol{\beta}})}{\partial \beta_1} = x_1 \times \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2)}} \times \frac{e^{-(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2)}}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2)}} = x_1 \times \hat{p}(\mathbf{x}) \times (1-\hat{p}(\mathbf{x})) \quad [14]$$

e

$$\frac{\partial h(\hat{\boldsymbol{\beta}})}{\partial \beta_2} = x_2 \times \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2)}} \times \frac{e^{-(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2)}}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2)}} = x_2 \times \hat{p}(\mathbf{x}) \times (1-\hat{p}(\mathbf{x})). \quad [15]$$

Assim, $\nabla h(\hat{\boldsymbol{\beta}})$ descrito na equação [12] é definido como

$$\nabla h(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2)}} \times \frac{e^{-(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2)}}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2)}} \\ x_1 \times \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2)}} \times \frac{e^{-(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2)}}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2)}} \\ x_2 \times \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2)}} \times \frac{e^{-(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2)}}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2)}} \end{bmatrix} = \begin{bmatrix} 1 \times \hat{p}(\mathbf{x}) \times (1-\hat{p}(\mathbf{x})) \\ x_1 \times \hat{p}(\mathbf{x}) \times (1-\hat{p}(\mathbf{x})) \\ x_2 \times \hat{p}(\mathbf{x}) \times (1-\hat{p}(\mathbf{x})) \end{bmatrix}. \quad [16]$$

No contexto deste trabalho deseja-se obter as estimativas de prevalência para cada categoria do preditor x_1, ajustado por x_2. Assim, para cada um dos métodos de predição, condicional e marginal, o termo $\hat{p}(\mathbf{x})$ deve ser substituído de forma adequada (Xu e Long, 2005).

Para o método de predição condicional, quando $x_1 = 0$

$$\hat{p}_{C_0}(\mathbf{x}) = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1\times 0+\hat{\beta}_2\bar{x}_2)}}, \quad [17]$$

e quando, $x_1 = 1$,

$$\hat{p}_{C_1}(\mathbf{x}) = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1\times 1+\hat{\beta}_2\bar{x}_2)}}. \quad [18]$$

De maneira similar, para o método de predição marginal, para $x_1 = 0$ (não exposto),

$$\hat{p}_{M_0}(\mathbf{x}) = \frac{1}{n}\times\sum_{i=1}^{n}\frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1\times 0+\hat{\beta}_2 x_{2i})}}, \quad [19]$$

enquanto que, para $x_1 = 1$ (exposto),

$$\hat{p}_{M_1}(\mathbf{x}) = \frac{1}{n}\times\sum_{i=1}^{n}\frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1\times 1+\hat{\beta}_2 x_{2i})}}. \quad [20]$$

O método Delta, descrito nessa seção, foi implementado em rotinas computacionais SAS e R para obter as variâncias de $h(\hat{\boldsymbol{\beta}})$ e, assim, estimar os intervalos de confiança para prevalências ajustadas. Alguns programas, como por exemplo, o SUDAAN, calculam estimativas de prevalências ajustadas pelos métodos de predição marginal e condicional e intervalos de confiança. A seção 3.2 descreve os métodos disponíveis, além do SUDAAN, para os programas SAS, Stata e R.

3.2 Aspectos computacionais

Rotinas computacionais (macros ou funções) para estimar prevalências ajustadas por meio de modelos multivariáveis estão disponíveis em programas usuais, tais como SAS, SUDAAN, R e STATA. Nesta seção é apresentada uma breve descrição das rotinas computacionais destes programas.

3.2.1 SAS - Macro *%ADJ_PROP*

O programa SAS não disponibiliza, de forma direta, procedimentos para estimar prevalências ajustadas e respectivos intervalos de confiança. Para esta finalidade, Dingyi Zhao (1985) criou a macro %ADJ_PROP escrita na linguagem SAS, originalmente para a versão 6.1, utilizando os métodos de predição marginal e condicional. Esta macro utiliza basicamente os comandos *PROC LOGISTIC* para a obtenção de prevalências ajustadas e o procedimento PROC *IML* para a obtenção dos respectivos intervalos de confiança. A macro *%ADJ_PROP* pode ser utilizada tanto para o cálculo de prevalências ajustadas, por meio de regressão logística, bem como para o cálculo de médias ajustadas, por meio de regressão linear. Por ter sido criada em uma versão antiga do programa SAS, foram realizadas modificações para que fosse possível obter as prevalências ajustadas.

Em 1998, Dingyi Zhao também criou à macro %ADJWPROP, para obter prevalências ajustadas no contexto de dados de planos amostrais com ponderação, mas que também pode ser utilizada para prevalências ajustadas sem o uso de pesos. Catellier (1998) realizou modificações na macro %ADJWPROP, renomeando-a para %ADJ. Face à documentação incompleta as macros %ADJWPROP e %ADJ não serão abordadas no trabalho.

3.2.2 SUDAAN (SAS-Callable): CONDMARG e PREDMARG

O SUDAAN versão SAS-Callable, é um programa para análise estatística de dados executado em conjunto com o programa SAS. No procedimento PROC

LOGISTIC do SUDAAN (*alias* PROC RLOGIST) estão disponíveis os comandos CONDMARG e PREDMARG, por meio dos quais é possível obter estimativas de prevalências ajustadas e intervalos de confiança para os métodos condicional e marginal, respectivamente. Neste trabalho os resultados do programa SUDAAN serão considerados como referência para as comparações com os resultados dos demais programas.

3.2.3 Stata - Função *margins*

Williams (2012) apresentou e discutiu diferentes métodos para estimação de prevalências ajustadas e efeitos marginais por meio do programa Stata, com destaque para o comando *margins* disponível a partir da versão 11. Com o comando *margins* é possível obter prevalências ajustadas e seus respectivos intervalos de confiança por meio do método condicional. O Stata não disponibiliza, de forma direta, estimativas de prevalências ajustadas para o método marginal.

3.2.4 R – Função *margins*

Thomas Leeper (2014) adaptou no programa R os mesmos procedimentos disponíveis para o Stata, apresentados por Williams (2012). Sendo assim, é possível obter apenas as prevalências ajustadas e seus respectivos intervalos de confiança por meio do método de predição condicional. Esta abordagem não será utilizada neste trabalho.

Rotinas computacionais para SAS e R foram desenvolvidas para estimar prevalências ajustadas por meio dos métodos descritos nas seções 3.1.1 e 3.1.2. Os respectivos intervalos de confiança foram derivados utilizando o método Delta, descrito na seção 3.1.3. Estes métodos, bem como aqueles descritos na revisão de literatura, serão mostrados em detalhes por meio de um exemplo.

4. OBJETIVOS

Objetivo geral

Apresentar e discutir métodos para estimar prevalências ajustadas em pesquisas clínicas e epidemiológicas e desenvolver rotinas computacionais em SAS e R para os métodos de predição marginal e condicional.

Objetivos específicos

a) Caracterizar a importância de utilizar prevalências ajustadas na presença de confundimento.
b) Revisar os métodos existentes para obtenção de prevalências ajustadas.
c) Revisar aspectos computacionais disponíveis para obtenção de prevalências ajustadas.
d) Desenvolver rotinas computacionais em SAS e R para estimar prevalências ajustadas e respectivos intervalos de confiança.

5. REFERÊNCIAS BIBLIOGRÁFICAS

An R. Diabetes diagnosis and nutrition facts label use among US adults, 2005–2010. Public Health Nutr. 2015 outubro 20: 1-8.

Chambless LE, Heiss G, Folsom AR, Rosamond W, Szklo M, Sharrett AR, et al. Association of coronary heart disease incidence with carotid arterial wall thickness and major risk factors: the Atherosclerosis Risk in Communities (ARIC) Study, 1987-1993. Am J Epidemiol. 1997 setembro 15; 146(6): 483-94.

Chor D, Ribeiro ALP, Carvalho MS, BB Duncan, Lotufo PA, Nobre AA, *et al*. Prevalence, Awareness, Treatment and Influence of Socioeconomic Variables on Control of High Blood Pressure: Results of the ELSA-Brasil Study. PLOS ONE. 2015 junho 23; 10(6): e0127382.

Epstein RA, Bobo WV, Martin PR, Morrow JA, Wang W, Chandrasekhar R, *et al*. Increasing pregnancy-related use of prescribed opioid analgesics. Ann Epidemiol. 2013 agosto. 23(8): 498-503.

Epstein RA, Bobo WV, Shelton RC, Arbogast PG, Morrow JA, Wang W, *et al.* Increasing use of atypical antipsychotics and anticonvulsants during pregnancy. Pharmacoepidem Dr S. 2013 julho. 22(7): 794-801.

Lane PW, Nelder JA. Analysis of Covariance and Standardization as Instances of Prediction. Biometrics. 1982; 38(3):613-21.

Lebedeva ER, Kobzeva NR, Gilev D, Olesen J. Prevalence of primary headache disorders diagnosed according to ICHD-3 beta in three different social groups. Cephalalgia. 2015 outubro 6.

Leeper T. Margins. Disponível de: https://github.com/leeper/margins. Acessado setembro 20, 2015.

McCullagh P, Nelder JA. Generalized Linear Models. 2th ed. 1989.

Nambi V, Hoogeveen RC, Chambless L, Hu Y, Bang H, Coresh J, et al. Lipoprotein-associated phospholipase A2 and high-sensitivity C-reactive protein improve the stratification of ischemic stroke risk in the Atherosclerosis Risk in Communities (ARIC) study. Stroke. 2009 fevereiro; 40(2): 376-81.

Ohira T, Folsom AR, Cushman M, White RH, Hannan PJ, Rosamond WD. Reproductive History, Hormone Replacement, and Incidence of Venous Thromboembolism: The Longitudinal Investigation of Thromboembolism Etiology. Br J Haematol. 2010 maio; 149(4): 606-612

Research Triangle Institute. SUDAAN Language Manual, Volumes 1 and 2, Release 11. 2012.

Schoenbach VJ, Rosamond WD. Understanding the Fundamentals of Epidemiology - an envolving text. 2000, Fall Edition.

Shon H-S, Choi HY, Kim JR, Ryu SY, Lee Y-J, Lee MJ*, et al*. Comparison and analysis of the prevalence of hepatitis C virus infection by region in the Republic of Korea during 2005-2012. Clin Mol Hepatol. 2015 setembro; 21(3): 249-56.

Smith AK, Cenzer IS, John Boscardin W, Ritchie CS, Wallhagen ML, Covinsky KE. Increase in Disability Prevalence Before Hip Fracture. J Am Geriatr Soc. 2015 outubro; 63(10): 2029-35.

Souza CL, Carvalho FM, Araújo TM, Reis EJF, Lima VCM, Porto LA. Factors associated with vocal fold pathologies in teachers. Rev Saúde Pública. 2011; 45(5).

Szklo M, Nieto FJ. Epidemiology Beyond the Basics. 2th ed. 2007.

Wilcosky TC, Chambless LE. A comparision of Direct Adjustment and Regression Adjustment of Epidemiologic Measures. J Chron Dis. 1985; 38: 849-856.

Williams R. Using margins command. The Stata Journal. 2012; 12(2): 308-331.

Woodward, M. Epidemiology Study Design and Data Analysis. 3th ed. 2014.

Xu J, Long JS. Using the Delta Method to Construct Confidence Intervals for Predicted Probabilities, Rates, and Discrete Changes Indiana University. 2005 agosto 22.

Zhao D. Logistic Regression Adjustment of Proportions and its Macro Procedure Disponível de: http://www2.sas.com/proceedings/sugi22/POSTERS/PAPER227.PDF. Acessado 7 de setembro de 2015.

6. CONCLUSÕES E CONSIDERAÇÕES FINAIS

Prevalências ajustadas são frequentemente reportadas em estudos epidemiológicos e são importantes para o leitor fazer uma análise do fenômeno livre de confundimento. Atualmente os métodos de estimação marginal e condicional são as abordagens mais utilizadas.

As rotinas SAS e R desenvolvidas no trabalho podem ser úteis para estimar prevalências ajustadas, produzindo resultados muito similares aos do SUDAAN. Em especial, a versão para o R é uma alternativa aos *softwares* comerciais.

As estimativas pontuais das rotinas desenvolvidas para o SAS e R são idênticas aos do SUDAAN, e as discrepâncias entre as estimativas por intervalo não são relevantes na prática.

No método condicional, a média ponderada (proporcional às frações amostrais) das prevalências ajustadas para cada categoria pode ser diferente da proporção global observada. No exemplo, esta média ponderada foi 22,3% (68,2×23,58 + 31,8×19,61), muito distante da prevalência global 25,3%. Esta é uma desvantagem do método condicional, que muitas vezes pode confundir o leitor. O método marginal não é suscetível a este problema, exceto talvez por arredondamentos no cálculo da média ponderada, que no exemplo foi igual a 25,0% (68,2×26,27 + 31,8×22,40).

Os resultados relativamente discrepantes gerados macro %ADJ_PROP pelo método marginal podem ser explicados pela diferença dos métodos utilizados. O método condicional utiliza uma constante de aproximação k, produzindo resultados similares aos do SUDAAN. Outra desvantagem é a falta de documentação dos códigos, dificultando sua compreensão para descrição dos métodos ou eventuais modificações.

As rotinas computacionais ainda precisam ser estendidas para possibilitar a incorporação de número maior de variáveis de exposição e de controle, de forma

automática. Outra possibilidade é estender o estudo e as rotinas para o modelo log-binomial, comparando as estimativas de prevalências ajustadas com aquelas obtidas pelo modelo logístico.

Em suma, prevalências ajustadas são importantes para uma análise mais realista do fenômeno em estudo. A discussão apresentada é importante para entender os métodos para estimação de prevalências ajustadas, e as rotinas computacionais desenvolvidas podem ser úteis para sua utilização.