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

DISSERTAÇÃO DE MESTRADO

Equações de Estimação Generalizadas (GEE): Aplicação em estudo sobre mortalidade neonatal em gemelares de Porto Alegre, RS (1995-2007).

Marilyn Agranonik

Orientador: Profa. Dra. Suzi Alves Camey

Porto Alegre, Dezembro de 2009.

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

EQUAÇÕES DE ESTIMAÇÃO GENERALIZADAS (GEE): APLICAÇÃO EM ESTUDO SOBRE MORTALIDADE NEONATAL EM GEMELARES DE PORTO ALEGRE, RS (1995-2007).

Marilyn Agranonik

Orientadora: Profa. Dra. Suzi Alves Camey

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.
2009

A277e Agranonik, Marilyn

Equações de estimação generalizadas (GEE) : aplicação em estudo sobre mortalidade neonatal em gemelares de Porto Alegre, RS (1995-2007) / Marilyn Agranonik ; orient. Suzi Alves Camey. – 2009.
110 f. : il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Faculdade de Medicina. Programa de Pós-Graduação em Epidemiologia. Porto Alegre, BR-RS, 2009.

1. Mortalidade neonatal 2. Gêmeos 3. Porto Alegre (RS) 4. Epidemiologia

5. Modelos estatísticos I. Camey, Suzi Alves II. Título.

NLM: WA 900

Catalogação Biblioteca FAMED/HCPA

BANCA EXAMINADORA

Prof. Dr. Álvaro Vigo, Programa de Pós Graduação em Epidemiologia, Universidade Federal do Rio Grande do Sul.

Prof. Dr. Clécio Homrich da Silva, Departamento de Pediatria e Puericultura, Universidade Federal do Rio Grande do Sul.

Profa. Dra. Luciana Neves Nunes, Departamento de Estatística, Universidade Federal do Rio Grande do Sul.

AGRADECIMENTOS

Agradeço a minha orientadora, Profª Suzi Alves Camey, pelo incentivo e oportunidade de aprendizado.

Ao professor Marcelo Goldani pelos conselhos, ensinamentos e incentivo à minha participação em pesquisas desde quedo iniciei como bolsista na graduação.

Aos membros da banca examinadora, professores Álvaro Vigo, Clécio Homrich da Silva e Luciana Neves Nunes, pelas importantes sugestões e contribuições para o meu trabalho.

Ao Programa de Pós-graduação em Epidemiologia pela oportunidade oferecida.

À CAPES pelo suporte financeiro que foi fundamental ao longo do curso.

À Coordenação Geral de Vigilância Sanitária da Secretaria de Municipal de Saúde de Porto Alegre.

Aos meus amigos, colegas do curso e colegas do Núcleo de Estudos sobre Saúde da Criança e do Adolescente - UFRGS por todo apoio, incentivo e pelos momentos de descontração. Agradeço a todos aqueles que de alguma forma contribuíram para construção deste trabalho.

Em especial, agradeço a meus pais, pelo carinho, apoio e incentivo e por terem me ensinado o valor do estudo.

SUMÁRIO

ABREVIATURAS E SIGLAS

AIC: *Akaike's Information Criterion* (critério de informação de Akaike)

AR(1): Auto-regressivo de primeira ordem

CIC: *Correlation Information Criterion* (Critério de Informação de Correlação)

DN: Declaração de Nascimento

DO: Declaração de Óbito

GEE: *Generalized Estimating Equations* (Equações de Estimação Generalizadas)

GLM: *Generalized Linear Models* (Modelos Lineares Generalizados)

MCAR: *Missing Completly at Random*

QIC: *Quasi-likelihood under the Independence model Criterion* (critério de quasi-verossimilhança sob o modelo de independência)

RN: recém nascido

SINASC: Sistema de Informação sobre Nascidos Vivos

SIM: Sistema de Informação sobre Mortalidade

RESUMO

A gravidez múltipla, como a de gêmeos e trigêmeos, é um exemplo de conglomerado natural no qual as respostas dos fetos são interdependentes ou agregadas. Ou seja, em estudos com gêmeos e trigêmeos é esperado que exista correlação entre os dados dos irmãos. Desse modo, os modelos de regressão tradicionais, como os GLM, podem levar à inferências incorretas, uma vez que a suposição de independência entre os sujeitos não é mais satisfeita. Para solucionar este problema, Zeger e Liang (1986) propuseram uma classe de Equações de Estimação Generalizadas (GEE), semelhante aos GLM, porém, incluindo uma estrutura de correlação de trabalho nas estimativas dos parâmetros do modelo.

Ainda hoje, poucos estudos utilizam esta metodologia. Considerando que a taxa de mortalidade infantil é maior em gemelares do que para os demais e a tendência de aumento da taxa de gemelaridade, existe uma preocupação crescente para um aumento do risco de morte precoce para gêmeos e trigêmeos quando comparados aos não gemelares. Este trabalho busca apresentar a metodologia do GEE, através de uma aplicação na análise de dados de mortalidade neonatal em gemelares. Foram utilizados dados de gêmeos e trigêmeos provenientes do SIM e do SINASC, nos quais todas as crianças que constituem o par ou o trio nasceram vivas em Porto Alegre, com peso superior a 500g entre 1995 a 2007. Verificou-se a associação de fatores perinatais, como peso ao nascer e índice de Apgar, com a mortalidade neonatal. Comparando os resultados obtidos no GEE com os do GLM foram encontradas pequenas diferenças nas estimativas pontuais dos parâmetros do modelo. Entretanto, ao comparar os erros padrões, as diferenças foram maiores, interferindo na significância de uma das variáveis (tipo de hospital). Maiores diferenças entre os modelos não foram encontradas, provavelmente porque o tamanho da amostra utilizado era grande. Desse modo, recomenda-se a utilização do GEE quando houver agrupamento de indivíduos, já que este modelo considera a correlação entre sujeitos do mesmo grupo e está implementado nos programas estatísticos.

ABSTRACT

Multiple births such as twins and triplets are a natural cluster in which the responses of the fetuses are interdependent. That is, in multiple births studies correlation can exists between siblings data. Therefore, traditional regression models, such as Generalized Linear Models (GLM), can lead to incorrect inferences because the assumption of independence among the subjects no longer exists. To solve this problem, Zeger and Liang (1986) proposed a class of Generalized Estimation Equation (GEE), similar to GLM, however, including a working correlation structure to estimate the regression parameters.

Even today, few studies use this methodology. Considering the higth rates of infant mortality in multiple births when compared to singles and the trend of increased multiple births rate, there is a concern for an increased risk of early death for twins and triplets compared to singletons. This study present GEE through an application in neonatal mortality in twins and triplets. Data from twins and triplets were obtained from SIM and SINASC, considering only clusters were all children were live births and had more than 500g in Porto Alegre from 1995 to 2007.

There was association of perinatal factors, such as birth weight and Apgar score, with neonatal mortality. Comparing the results from GEE and GLM, small differences were found in model parameters estimates. However, when comparing the standard errors, the differences were larger, interfering in the significance of a variable (type of hospital). Major differences between the models were not found, probably because the sample size used was large. Thus, it is recommended the use of GEE when there is clustered data, since this model considers the correlation between subjects within the group and is implemented in statistical programs.

LISTA DE QUADROS E TABELAS

Quadros e tabelas da dissertação

Quadros do artigo

LISTA DE FIGURAS

1. APRESENTAÇÃO

Este trabalho consiste na dissertação de mestrado intitulada "Equações de Estimação Generalizadas (GEE): Aplicação em estudo sobre mortalidade infantil em Gemelares de Porto Alegre, RS (1995-2007)", apresentada ao Programa de Pós-Graduação em Epidemiologia da Universidade Federal do Rio Grande do Sul, em 15 de dezembro de 2009. O trabalho é apresentado em três partes, na ordem que segue:

1. Introdução, Revisão da Literatura e Objetivos
2. Artigo(s)
3. Conclusões e Considerações Finais.

Documentos de apoio estão apresentados nos anexos: Projeto de Pesquisa (anexo A), aprovação pelo Comitê de Ética e Pesquisa (anexo B), Formulário da Declaração de Nascido Vivo (anexo C), Formulário da Declaração de Óbito (anexo D), comandos utilizados no SPSS, versão 16.0 (anexo E), comandos utilizados no R (anexo F) e comparação entre coeficientes e erros padrões estimados através de GEE e GLM (anexo G).

2. INTRODUÇÃO

Quando se deseja estudar a relação entre uma variável resposta (desfecho) e variáveis independentes (exposições), técnicas de modelagem são utilizadas, nas quais se incluem os modelos de regressão. Através destes modelos é possível avaliar, por exemplo, fatores de risco para mortalidade infantil.

Uma das principais suposições dos modelos de regressão tradicionais, como os Modelos Lineares Generalizados (GLM), é a suposição de independência entre os sujeitos observados. No caso do modelo para mortalidade infantil, isso significa supor que o conhecimento a respeito da ocorrência de óbito em uma criança não fornece nenhuma informação a respeito do estado de outra criança nesse estudo. Entretanto, caso a amostra estudada contenha irmãos, é razoável supor que esta hipótese não esteja correta. Ao avaliar resultados provenientes de gemelares (gêmeos, trigêmeos, ou de ordem superior) verifica-se que fetos de uma mesma gestação, expostos às mesmas características maternas e a condições semelhantes no útero, apresentam respostas mais semelhantes do que os de gestações diferentes [1]. Ou seja, as observações de indivíduos que não pertençam à mesma família são independentes, entretanto as de irmãos não são.

Esta questão da dependência de observações pode ocorrer sempre que for possível identificar agrupamentos entre os indivíduos estudados. Além de pertencerem à mesma família, também pode ocorrer correlação entre alunos de uma mesma escola, ou pacientes de um mesmo hospital [2]. É possível também ocorrer correlação entre observações realizadas em um mesmo indivíduo ao longo do tempo, como ocorre muitas vezes em estudos longitudinais. Nesse caso, cada indivíduo pode ser considerado como um grupo de medidas repetidas [2].

Em todas as situações mencionadas acima é razoável esperar que as respostas observadas dentro de um grupo sejam mais semelhantes do que aquelas observadas entre grupos. Por isso, para avaliar a relação entre os fatores de risco e o desfecho estudado é necessário considerar a dependência entre as observações do mesmo grupo. E, desse modo, não é possível utilizar os modelos tradicionais de regressão, que supõe independência entre os indivíduos observados.

Atualmente existem pelo menos duas abordagens adequadas para a análise de dados agrupados. As principais são as Equações de Estimação Generalizadas (*Generalized Estimating Esquations* - GEE) e os Modelos de Efeitos Mistos. O método de GEE foi proposto por Zeger e Liang [3] e Liang e Zeger [4] com o objetivo de estimar parâmetros de regressão especialmente quando os dados estão correlacionados. Os autores basearam-se nos GLM's, incluindo uma estrutura de correlação de "trabalho" ("*working" correlation matrix*) entre as observações para a obtenção de estimativas consistentes e não viciadas. No modelo de efeitos aleatórios, proposto por Laird & Ware [5], os coeficientes de regressão podem ser diferentes entre indivíduos, considerando a heterogeneidade existente entre eles. Esse modelo tem dois componentes: um intra-indivíduo (uma mudança longitudinal intra-indivíduo é descrita pelo modelo de regressão com um intercepto e inclinação populacional) e outro entre indivíduos (variação no intercepto e inclinação individual) [6].

A principal diferença entre estes métodos está no fato do GEE avaliar a relação entre a variável resposta e as variáveis preditoras em um contexto populacional, e não individual, enquanto o modelo de efeitos aleatórios tem como foco o indivíduo. Desse modo, quando se tem interesse em avaliar diversas medidas de um mesmo indivíduo ao longo do tempo, e avaliar seu crescimento individual, é mais indicado utilizar um modelo de efeitos aleatórios. E, quando se estiver interessado em estudos epidemiológicos, por exemplo, com o objetivo de

se estudar a diferença na resposta média populacional entre dois grupos com diferentes fatores de risco, o GEE é o método mais recomendado [7].

Apesar da existência destes modelos e de eles estarem implementados em diversos programas estatísticos, como SPSS, STATA, SAS e R, ainda hoje é pouco comum encontrar artigos, especialmente no Brasil, que utilizem a modelagem adequada para dados correlacionados. Considerando o crescente número de estudos epidemiológicos envolvendo observações correlacionadas, seja em estudos longitudinais ou em estudos envolvendo dados agrupados e os problemas que podem ocorrer com a utilização da análise inadequada, este trabalho tem por objetivo apresentar a metodologia GEE, através de uma aplicação na análise de dados de mortalidade neonatal em gemelares (gêmeos, trigêmeos ou de ordem superior).

3. REVISÃO DE LITERATURA

3.1 MODELOS LINEARES GENERALIZADOS

Em diferentes áreas de pesquisa, incluindo a área da saúde, é freqüente a situação em que se deseja estudar o comportamento de uma variável resposta em relação a uma ou mais variáveis independentes. As variáveis independentes, também chamadas de preditoras ou explicativas, são responsáveis por explicar a variabilidade da variável resposta, ou dependente. Para esses casos, técnicas de modelagem são utilizadas, nas quais se incluem os modelos de regressão.

Inicialmente os modelos de regressão foram desenvolvidos considerando a variável resposta com distribuição normal. McCullagh e Nelder [8] sintetizaram o modelo linear clássico considerando um vetor $\boldsymbol{y}$ de n observações independentes, $\boldsymbol{y} = (y_1,...,y_n)'$, que representa a variável resposta e uma matriz $\boldsymbol{X}$ de p variáveis preditoras. Neste modelo, supõe- se que $\boldsymbol{y}$ segue distribuição normal com média, $\mathrm{E}(\boldsymbol{y})$, e variância, $\mathrm{V}(\boldsymbol{y})$, dados por:

$$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{\mu} = \boldsymbol{X\beta} \quad \text{e}$$

$$\mathrm{V}(\boldsymbol{y}) = \boldsymbol{\sigma}^2,$$

onde $\boldsymbol{\mu}$ é um vetor $n \times 1$ de médias, $\boldsymbol{X}$ é uma matriz $n \times p$ de variáveis independentes, $\boldsymbol{\beta}$ é um vetor $p \times 1$ de parâmetros e $\boldsymbol{\sigma^2}$ é uma matriz diagonal $n \times n$ de variâncias.

Em 1972, Nelder e Wedderburn [9] estenderam esse modelo para todos os membros da família exponencial, criando os Modelos Lineares Generalizados (*Generalized Linear Models* - GLM). Algumas distribuições de probabilidade que pertencem à família exponencial são: normal, gama, Poisson e binomial. Uma importante característica dos GLM's é a suposição de independência, ou pelo menos de não correlação, entre observações.

3.1.1 Formulação do Modelo

Para formular um GLM é necessário escolher:

(1) Uma distribuição de probabilidade para a variável resposta, que deve pertencer à família exponencial de distribuições;

(2) As variáveis preditoras, que podem ser quantitativas e/ou qualitativas e

(3) Uma função de ligação que irá relacionar as componentes aleatória e sistemática do modelo. (Ver secção 3.1.2).

Para melhorar a escolha da referida distribuição de probabilidade é aconselhável examinar os dados para observar algumas características, tais como: assimetria, natureza discreta ou contínua, intervalo de variação, etc.

É importante salientar que os termos que compõe a estrutura linear do modelo podem ser de natureza quantitativa, qualitativa ou mista, e devem dar uma contribuição significativa na explicação da variável resposta.

3.1.2 Componentes de um GLM

De forma geral, a estrutura de um GLM é formada por três partes:

(1) Componente aleatória: composta de uma variável resposta $\boldsymbol{y}$ com n observações independentes, um vetor de médias $\boldsymbol{\mu}$ e uma distribuição de probabilidade pertencente à família exponencial.

(2) Componente sistemática: composta por variáveis explicativas $X_1,...,X_p$ e pelos parâmetros desconhecidos.

(3) Função de ligação: função monotônica diferenciável que relaciona as duas componentes anteriores.

3.1.2.1 Componente Aleatória

Cada componente de $\boldsymbol{y}$ segue uma mesma distribuição da família exponencial, ou seja, a função densidade de y_i é dada por:

$$f(y_i;\theta;\phi) = \exp\left\{\frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i,\phi)\right\}, \qquad (1)$$

onde a(.), b(.) e c(.) são funções conhecidas; $\phi > 0$ é denominado parâmetro de dispersão e θ é denominado parâmetro canônico que caracteriza a distribuição. Se ϕ é conhecido, a equação (1) representa a família de densidades exponenciais uniparamétricas indexada por θ.

Exemplo: Distribuição Normal.

$$f(y_i;\theta,\phi) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y_i-\mu)^2}{2\sigma^2}\right\} = \exp\left\{\frac{y_i\mu_i - \mu^2/2}{\sigma^2} - \frac{1}{2}\left(\frac{y_i^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right\}$$

onde $\theta = \mu, \quad \phi = \sigma^2, \quad a(\phi) = \phi, \quad b(\theta) = \frac{\theta^2}{2} \quad e \quad c(y,\phi) = -\frac{1}{2}\left\{\frac{y^2}{\phi} + log(2\pi\phi)\right\}$.

O quadro 1 apresenta características de algumas distribuições da família exponencial.

Quadro 1: Características de algumas distribuições da família exponencial.

Modelo	θ	ϕ	$a(\phi)$	$b(\theta)$	$c(y,\phi)$	Ligação canônica $\theta(\mu)$
Normal: $N(\mu,\sigma^2)$	μ	σ^2	σ^2	$\frac{\theta^2}{2}$	$-\frac{1}{2}\left\{\frac{y^2}{\phi} + log(2\pi\phi)\right\}$	Identidade: $\eta = \mu$
Binomial: $\frac{B(m,\mu)}{m}$	$log\left(\frac{\mu}{1-\mu}\right)$	m	$\frac{1}{m}$	$log(1+e^{\theta})$	$log\binom{m}{my}$	logit: $\eta = log\left[\frac{\mu}{1-\mu}\right]$
Poisson: $P(\mu)$	$log\,\mu$	1	1	$exp(\theta)$	$-log\ y!$	log: $\eta = log\,\mu$

3.1.2.2 Componente sistemática e função de ligação

Considere a estrutura linear de um modelo de regressão

$$\boldsymbol{\eta} = \boldsymbol{X\beta},$$

onde $\boldsymbol{\eta} = (\eta_1,...,\eta_n)'$, $\boldsymbol{\beta} = (\beta_1,...,\beta_p)'$ e $\boldsymbol{X}$ é a matriz de variáveis independentes $n \times p$ ($p < n$) conhecida de posto p. A função linear $\boldsymbol{\eta}$ dos parâmetros desconhecidos $\boldsymbol{\beta}$ é chamada de preditor linear. Como $X_i \in \Re^p$ e $\beta_i \in \Re$, então cada componente de $\eta_i \in \Re$.

Através de uma função de ligação, $g(.)$, adequada pode-se escrever a esperança da variável resposta, $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{y})$, em função das variáveis explicativas, $\boldsymbol{X}$. Ou seja, para cada distribuição da família exponencial existe uma função $g(.)$, com contradomínio na reta real, tal que:

$$g(\mu_i) = \eta_i, \qquad i = 1,...,n$$

Se cada componente de $\boldsymbol{y}$ segue uma distribuição normal, então $\mu_i \in \Re$ e como $\eta_i \in \Re$, a função de ligação do tipo identidade $(\eta = \mu)$ é plausível para modelar dados normais. Se cada componente de $\boldsymbol{y}$ tem distribuição Poisson, então $\mu_i > 0$, e, portanto, a função de ligação adequada é a logarítmica $(\eta = \log \mu)$, pois esta tem domínio positivo e contradomínio na reta real. No caso de cada componente de $\boldsymbol{y}$ assumir a distribuição binomial, então $0 < \mu_i < 1$. Logo, existe a restrição de que o domínio da função de ligação esteja no intervalo $(0;1)$. As três principais funções que garantem esta restrição são:

(1) *Logit (ou logística)*: $\eta = \log\left[\dfrac{\mu}{1-\mu}\right]$

(2) *Probit*: $\eta = \Phi^{-1}(\mu)$, onde Φ^{-1} é a função de distribuição acumulada da normal reduzida.

(3) *Complemento log-log:* $\eta = \log[-\log(1-\mu)]$.

Cada uma das distribuições apresentadas no quadro 1 tem uma função de ligação especial para qual existe uma estatística suficiente com igual dimensão de $\boldsymbol{\beta}$ associada ao preditor linear $\boldsymbol{\eta} = \boldsymbol{X\beta}$. Essas ligações são denominadas ligações canônicas e ocorrem quando $\theta = \eta$, onde θ é o parâmetro canônico definido em (1) e apresentado no quadro 1.

3.1.3 Estimação

Após escolher um determinado modelo, é necessário estimar seus parâmetros e avaliar a precisão das estimativas. No caso dos GLM's, os parâmetros podem ser estimados através de diversos métodos, como o qui-quadrado mínimo, o Bayesiano e a estimação-M [10]. O último inclui o método da máxima verossimilhança, onde os estimadores possuem propriedades como consistência e eficiência assintótica.

Neste trabalho será apresentada somente a estimação pelo método.da máxima verossimilhança. Para obter as estimativas dos parâmetros, deve-se maximizar a função de verossimilhança, ou a função de log-verossimilhança, em relação aos parâmetros, supondo fixos os dados observados. Assim, considerando $f(\, y_i ; \beta\,)$ a função densidade para y_i dado o parâmetro β, cuja forma é conhecida, mas o parâmetro β é desconhecido, a função de log-verossimilhança para a i-ésima observação é definida por:

$$l(\, \beta; y_i\,) = log\, f(\, y_i ; \beta\,).$$

A log-verossimilhança do vetor de observações independentes $(y_1,...,y_n)$ é a soma das contribuições individuais, assim

$$l(\beta; \mathrm{y}) = l(\beta) = \sum_{i=1}^{n} \log f(y_i; \beta).$$

Nelder e Wedderburn [9] desenvolveram um algoritmo para estimação dos parâmetros β através da máxima verossimilhança, baseado em um método muito semelhante ao de

Newton-Raphson, conhecido como Método de Escore de Fisher. Este método consiste em resolver o sistema

$$U(\beta_j) = \frac{\partial l(\beta)}{\partial \beta_j} = 0\,,\ j = 1,..,p\,.$$

onde $U(\beta)$ é conhecida como função escore e $l(\beta)$ é a log-verossimilhança de β. Além disso, utiliza a matriz de informação de Fisher

$$K = \left\{ -E\left(\frac{\partial^2 l(\beta)}{\partial \beta_j \partial \beta_s} \right) \right\} = -E\left(\frac{\partial U(\beta)}{\partial \beta_j} \right),\ j = 1,..,p \text{ e } s = 1,..,p\,.$$

A partir daqui, os índices j e s serão omitidos para simplificar a notação.

Expandindo a função escore em série de Taylor até primeira ordem, obtém-se:

$$U\left(\beta^{(m+1)}\right) = U\left(\beta^{(m)}\right) + \frac{\partial U(\beta)^m}{\partial \beta}\left[\beta^{(m+1)} - \beta^{(m)}\right] = 0$$

ou $$\beta^{(m+1)} = \beta^{(m)} - \left[\frac{\partial U(\beta)^m}{\partial \beta}\right]^{-1} U\left(\beta^{(m)}\right),$$

onde o índice (m) significa o valor da m-ésima iteração. Este é o método de Newton-Raphson para o cálculo iterativo da estimativa de máxima verossimilhança $\hat{\beta}$ de β.

O método escore de Fisher é obtido pela substituição de $-\frac{\partial U(\beta)}{\partial \beta}$ pelo seu valor esperado K.

3.2 MODELOS PARA DADOS CORRELACIONADOS

Em muitas situações, apesar dos sujeitos estudados serem independentes, a informação sobre uma determinada variável é coletada repetidas vezes ao longo do tempo, tornando as observações correlacionadas. É possível também que os sujeitos dividam características em

comum (por exemplo, estudantes de uma mesma escola, pacientes de um mesmo hospital, pessoas que trabalham em um mesmo local, irmãos,...) e, portanto, não podem ser considerados independentes. Neste caso pode haver uma estrutura natural de correlação entre os sujeitos.

O primeiro caso é conhecido como medidas repetidas e, o segundo, como dados agrupados (*clustered data*). A correlação, nesses casos, pode ocorrer já que as observações feitas em um mesmo indivíduo (estudos longitudinais) ou em pessoas de um mesmo grupo (dados agrupados) tendem a ser mais semelhantes do que observações de indivíduos diferentes ou de grupos diferentes [2].

Os modelos tradicionais de regressão têm uso limitado em estudos longitudinais ou de dados agrupados devido à suposição de independência entre os sujeitos. Este é o caso dos GLM's [8, 9]. Apesar deste ser um método poderoso e flexível, se for utilizado para dados correlacionados, é provável a obtenção de distorções nas estimativas dos parâmetros e de seus erros padrões, levando a inferências estatísticas incorretas [3, 4, 11, 12].

Quando a variável resposta tem distribuição aproximadamente normal, pode-se contar com vários métodos estatísticos para dados correlacionados. Rao [13], Grizzle & Allen [14], e Hui [15] apresentaram métodos baseados em curvas de crescimento para modelar observações realizadas em um mesmo sujeito. Fearn [16] discutiu uma abordagem bayesiana para modelos de curvas de crescimento. Harville [17] e Laird & Ware [5] desenvolveram modelos de efeitos aleatórios nos quais assume-se que as observações repetidas de cada sujeito compartilham um mesmo componente aleatório. Azzalini [18] apresentou modelos nos quais assume-se uma estrutura auto-regressiva para o erro. Ware [19] apresentou uma revisão geral sobre modelos lineares para dados longitudinais gaussianos.

Para dados com distribuição não normal e correlacionados, existem pelo menos duas abordagens estatísticas: as Equações de Estimação Generalizadas (*Generalized Estimating*

Equations - GEE) [3] e os modelos de efeitos aleatórios (um caso especial de modelos mistos ou de modelos multiníveis [5]). Estas técnicas, inicialmente desenvolvidas para variáveis resposta com distribuição normal, foram estendidas para variáveis com outras distribuições [20-22].

O método de GEE foi proposto por Zeger e Liang [3] e Liang e Zeger [4] com o objetivo de estimar parâmetros de regressão especialmente quando os dados estão correlacionados. Os autores basearam-se nos GLM's, incluindo uma estrutura de correlação de trabalho entre as observações para a obtenção de estimativas consistentes e não viciadas.

No modelo de efeitos aleatórios proposto por Laird & Ware [5] os coeficientes de regressão podem ser diferentes entre indivíduos, considerando a heterogeneidade existente entre eles. Stiratelli, Laird, & Ware [21], Anderson & Aitkin [23], e Gilmour, Anderson, & Rae [24] apresentam aplicações deste modelo para dados binomiais.

Quando a variável resposta é de natureza contínua, há pouca diferença nos resultados apresentados por esses dois métodos [7]. Entretanto, se a variável resposta for dicotômica, eles podem apresentar resultados bem divergentes. Neste caso, Twisk [7] aconselha a utilização do GEE quando o interesse for avaliar a relação entre a variável resposta e as variáveis preditoras em um contexto populacional, e não individual, e do modelo de efeitos aleatórios se o foco for no indivíduo.

Neste trabalho será utilizada a notação para dados agrupados na definição do GEE. Em estudos com famílias, o grupo é cada família e os indivíduos são os membros da família.

3.3 EQUAÇÕES DE ESTIMAÇÃO GENERALIZADAS

As Equações de Estimação Generalizadas (*Generalized Estimating Equations - GEE*) [3, 4] foram desenvolvidas para produzir estimativas mais eficientes e não viciadas para os parâmetros do modelo de regressão quando se lida com dados correlacionados, pois considera a estrutura de correlação entre as observações. GEE é uma extensão dos GLM, sendo que não é necessário assumir que a variável resposta seja da família exponencial. Assume-se, entretanto, que a média e a variância estão caracterizadas como em um GLM.

O GEE estima coeficientes de regressão e erros padrões com distribuições amostrais assintoticamente normais [3]. Pode ser utilizado para testar efeitos principais e interações, permitindo avaliar variáveis independentes categóricas ou contínuas.

Este método deve ser utilizado quando o objetivo da análise estatística é descrever a esperança da variável resposta em função de um conjunto de covariáveis considerando a correlação entre as observações. Assim, Liang e Zeger [4] especificaram a esperança da variável resposta como uma função linear das covariáveis, assumiram a variância como uma função conhecida da média e definiram uma matriz de correlação de trabalho (*working correlation matrix*). Essas equações são extensões das utilizadas no método de quasi-verossimilhança [25], definido na secção 3.3.2 (ver equação 5).

Inicialmente Zeger e Liang [3, 4] introduziram o conceito de GEE voltado para estimação somente da média, no qual é necessário especificar corretamente apenas a estrutura do modelo de regressão, tratando os parâmetros de correlação como parâmetros de perturbação (*nuisance parameters*). Posteriormente, Prentice [26] descreve um segundo tipo de GEE, conhecido por GEE2, no qual a estimação da média e da correlação ocorrem simultaneamente e, nesse caso, torna-se necessário especificar corretamente a estrutura de correlação, além do modelo para a média. Zorn [27] adverte que se utilize o GEE2 somente

quando a estrutura de correlação de trabalho correta for conhecida, caso contrário os parâmetros estimados através do GEE2 podem não ser consistentes. Qu, Lindsay, e Li [28] propuseram um método diferente para melhorar a eficiência com base em funções de inferência quadráticas. Os autores mostram que a sua abordagem, com a escolha adequada dos escores para as funções de inferência quadráticas, é mais eficiente do que o GEE quando a matriz de correlação de trabalho não está bem especificada. No entanto, esta abordagem não é implementada nos programas estatísticos padrões.

No presente trabalho será apresentada somente a metodologia proposta por Zeger e Liang [3, 4].

3.3.1 Formulação do GEE

Considere n grupos de indivíduos semelhantes, onde y_{ij} é a variável resposta de interesse para o j-ésimo indivíduo do i-ésimo grupo e X_{ij} é um vetor $p \times 1$ de covariáveis para o j-ésimo indivíduo do i-ésimo grupo, $i = 1,...,n$ e $j = 1,...,m_i$. O valor de m pode variar de grupo para grupo. Define-se, para o i-ésimo grupo, o vetor $m_i \times 1$ de respostas, $\boldsymbol{y}_i = (y_{i1},...,y_{im_i})'$ e a matriz de covariáveis $m_i \times p$, $\mathrm{X}_i = (X_{i1},...,X_{im_i})'$. Para se escrever as Equações de Estimação Generalizadas supõe-se que:

1 - A relação entre a média da variável resposta, $\boldsymbol{\mu}_i$, e as variáveis explicativas $\boldsymbol{X}$, pode ser expressa sob forma linear através de uma função de ligação conhecida, g. Esta função é tal que:

$$g(\boldsymbol{\mu}_i) = \boldsymbol{X}_i'\boldsymbol{\beta}, \qquad (1)$$

onde $\boldsymbol{\beta}$ é o vetor de p parâmetros.

2 - A variância da variável resposta pode ser expressa por uma função conhecida da média desta variável, ou seja,

$$\mathrm{V_i} = f(\boldsymbol{\mu}_i)/\phi\,, \tag{2}$$

onde ϕ é o parâmetro de dispersão definido como na família exponencial.

Nota-se que os autores definem a relação entre a média da variável resposta e as variáveis explicativas (equação 1) e a relação entre variância e a média da variável resposta (equação 2) da mesma maneira que em um GLM.

Liang e Zeger [4] definem a estimativa de $\boldsymbol{\beta}$ como sendo a solução do sistema equações diferenciais quasi-escore dado a seguir:

$$U_k(\boldsymbol{\beta}) = \sum_{i=1}^{n} D_i \, \mathrm{V}_i^{-1} \, S_i = 0 \qquad\qquad k = 1,..., p. \tag{3}$$

onde, $D_i = \partial\mu_i / \partial\beta_k$ e $S_i = (y_i - \mu_i)$.

Para utilizar essas equações para dados correlacionados, Liang e Zeger especificaram uma matriz de correlação de trabalho incorporada no termo de variância da equação (2). Considerando $R_\mathrm{i}(\alpha)$ tal matriz, com dimensão $m_i \times m_i$ para cada $\boldsymbol{y_i}$, onde α é um vetor que caracteriza completamente $R_\mathrm{i}(\alpha)$, a equação (2) torna-se uma matriz de covariância para o i-ésimo grupo:

$$V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2} / \phi\,, \tag{4}$$

onde A_i é uma matriz diagonal $m_i \times m_i$, com $f(\mu_i)$ como elementos da diagonal principal e ϕ é o parâmetro de escala para distribuições da família exponencial. Note que o número de observações e a matriz de correlação podem diferir de grupo para grupo. Porém, é possível assumir que $R_\mathrm{i}(\alpha)$ é completamente especificado pelo vetor de parâmetros desconhecidos α, que é o mesmo para todos os grupos [3]. Assim, será utilizado $R(\alpha)$ para denotar qualquer matriz de correlação de trabalho.

Quando $m_i = 1$, ou no caso de haver independência, o estimador do GEE equivale ao do GLM. É possível perceber que o GEE é uma extensão do GLM e, portanto, a interpretação dos parâmetros estimados é semelhante a dos GLM.

É importante ressaltar que no GEE, apesar de observações pertencentes a um mesmo grupo possam estar correlacionadas, supõe-se que observações em grupos diferentes sejam independentes.

3.3.2 Função de Quasi-verossimilhança

A função de quasi-verossimilhança foi proposta por Wedderburn em 1974 [25] e posteriormente reexaminada por McCullagh e Nelder em 1983 [8]. Esta metodologia necessita de poucas suposições sobre a distribuição da variável resposta e é de grande utilidade quando se deseja obter estimadores dos parâmetros dos modelos de regressão, porém não se conhece a forma da distribuição conjunta das observações.

A função de quasi-verossimilhança pode ser utilizada para estimação de forma semelhante à função de verossimilhança. Sua grande vantagem é necessitar apenas da especificação da relação entre a média e a variância das observações, enquanto a verossimilhança necessita a especificação da forma de distribuição das observações.

Dependendo da relação entre a média e variância especificada, a função de quasi-verossimilhança pode se tornar uma função de verossimilhança conhecida. Para um membro da família exponencial uniparamétrica, a função de log-verossimilhança é a mesma que a de quasi-verossimilhança e pertencer a esta família é a suposição mais fraca que pode ser feita sobre a distribuição.

Para definir a função de quasi-verossimilhança, considera-se y_i, $i = 1,...,n,$ observações independentes com médias μ_i e variâncias $V(\mu_i)$, onde V é uma função conhecida. Suponha que cada observação μ_i é uma função conhecida de um conjunto de

parâmetros $\beta_1,...,\beta_p$. Então para cada observação a função de quasi-verossimilhança $Q(y_i;\mu_i)$ é definida como:

$$\frac{\partial Q(y_i,\mu_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{V(\mu_i)}. \tag{5}$$

O quadro 2 apresenta a função de quasi-verossimilhança para algumas distribuições da família exponencial.

Quadro 2: Função de quasi-verossimilhança para algumas distribuições da família exponencial.

Modelo	$\mu(\theta)$	$V(\mu)$	Ligação canônica	$Q(y;\mu)$	Limites
Normal: $N(\mu,\sigma^2)$	θ	1	Identidade: $\eta = \mu$	$-(y-\mu)^2/2$	-
Binomial: $\frac{B(m,\mu)}{m}$	$\frac{e^\theta}{1+e^\theta}$	$\mu(1-\mu)$	Logit: $\eta = \log\left[\frac{\mu}{1-\mu}\right]$	$y\log\left[\frac{\mu}{1-\mu}\right] + \log(1-\mu)$	$0 < \mu > 1$ $0 \le y \le 1$
Poisson: $P(\mu)$	$exp(\theta)$	μ	Log: $\eta = log\,\mu$	$y\log\mu - \mu$	$\mu > 0$ $y \ge 0$

Em seu artigo, Wedderburn [25] demonstrou que a log quasi-verossimilhança tem as seguintes propriedades, semelhantes as da função de log verossimilhança, ou seja,

(i) $$E\left(\frac{\partial Q}{\partial \mu}\right) = 0$$

(ii) $$E\left(\frac{\partial Q}{\partial \beta_i}\right) = 0$$

(iii) $$E\left(\frac{\partial Q}{\partial \mu}\right)^2 = -E\left(\frac{\partial^2 Q}{\partial \mu^2}\right) = \frac{1}{V(\mu)}$$

(iv) $$E\left(\frac{\partial Q \partial Q}{\partial \beta_i \partial \beta_j}\right) = -E\left(\frac{\partial^2 Q}{\partial \beta_i \partial \beta_j}\right) = \frac{1}{V(\mu)}\frac{\partial \mu}{\partial \beta_i}\frac{\partial \mu}{\partial \beta_j}.$$

3.3.3 Especificação da matriz de correlação de trabalho

Nesta secção serão definidas as possíveis estruturas da matriz de correlação de trabalho. Como $R(\alpha)$ representa a correlação entre as observações de um mesmo grupo, ajustada pelas covariáveis presentes no modelo, os valores que α pode assumir estão no intervalo $[-1;+1]$ e a dimensão dessa matriz é determinada pelo número de observações feitas em cada grupo. Dentre as possíveis estruturas de correlação, destacam-se a permutável, na qual considera-se que a correlação entre as observações dos indivíduos de um mesmo grupo é a mesma; a não estruturada, para a qual assume-se que entre cada observação dentro do grupo há um valor de correlação diferente; a auto regressiva de primeira ordem, quando supõe-se que as medidas dentro do grupo têm uma relação auto-regressiva de primeira ordem, usualmente utilizada quando os dados estão correlacionados ao longo do tempo e, no caso de independência entre as observações, utiliza-se a estrutura independente. No quadro 3, são apresentadas as possíveis estruturas para essa matriz, fixando $m = 4$.

Especificar matriz de correlação de trabalho de forma correta aumenta a eficiência das estimativas dos parâmetros do modelo [29], o que é particularmente importante quando a correlação entre as respostas for alta. Liang e Zeger [4] afirmam que o modelo é robusto a erros na especificação na estrutura de correlação porque as estimativas dos parâmetros de regressão permanecem consistentes e ressaltam que a eficiência ganha pela especificação exata da estrutura de correlação é geralmente pequena. Entretanto, esta afirmação só é válida quando não há dados perdidos (*missing*) ou quando é possível assumir que eles são completamente aleatórios (MCAR). Além disso, Fitzmaurice [29] adverte que, caso a matriz de correlação de trabalho especificada não incorpore toda a informação sobre a correlação entre as medidas de um mesmo grupo, pode-se obter estimadores ineficientes. Desse modo, torna-se importante escolher a estrutura de correlação mais apropriada para a análise.

Quadro 3: Definição, exemplo e número de parâmetros estimados para cada tipo de estrutura de correlação de trabalho.

Estrutura	Definição	Exemplo $(m=4)$	Número de parâmetros
Independente	$Corr(Y_{ij}, Y_{ik}) = \begin{cases} 1, se\ j = k \\ 0, se\ j \neq k \end{cases}$	$R(\alpha) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$	0
Permutável	$Corr(Y_{ij}, Y_{ik}) = \begin{cases} 1, se\ j = k \\ \alpha, se\ j \neq k \end{cases}$	$R(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha & \alpha \\ \alpha & 1 & \alpha & \alpha \\ \alpha & \alpha & 1 & \alpha \\ \alpha & \alpha & \alpha & 1 \end{pmatrix}$	1
AR(1)	$Corr(Y_{ij}, Y_{i\,j+t}) = \alpha^t,$ $t = 0,1,2,3$	$R(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 \\ \alpha & 1 & \alpha & \alpha^2 \\ \alpha^2 & \alpha & 1 & \alpha \\ \alpha^3 & \alpha^2 & \alpha & 1 \end{pmatrix}$	1
M-dependente	$Corr(Y_{ij}, Y_{i\,j+t}) = \begin{cases} 1, se\ t = 0 \\ \alpha_t, se\ t = 1,2,...,M \\ 0, se\ t > M \end{cases}.$	$R(\alpha) = \begin{pmatrix} 1 & \alpha_1 & \alpha_2 & 0 \\ \alpha_1 & 1 & \alpha_1 & \alpha_2 \\ \alpha_2 & \alpha_1 & 1 & \alpha_1 \\ 0 & \alpha_2 & \alpha_1 & 1 \end{pmatrix}$ $M = 2$	$0 < M < m-1$
Não estruturada	$Corr(Y_{ij}, Y_{ik}) = \begin{cases} 1, se\ j = k \\ \alpha_{jk}, se\ j \neq k \end{cases}$	$R(\alpha) = \begin{pmatrix} 1 & \alpha_1 & \alpha_2 & \alpha_3 \\ \alpha_1 & 1 & \alpha_4 & \alpha_5 \\ \alpha_2 & \alpha_4 & 1 & \alpha_6 \\ \alpha_3 & \alpha_5 & \alpha_6 & 1 \end{pmatrix}$	$m(m-1)/2$

3.3.4 Critérios para seleção da estrutura de correlação de trabalho

Diversos autores aconselham avaliar a natureza dos dados para escolher a matriz de correlação mais adequada [30-33]. Estes autores fazem as seguintes recomendações: se o número de medidas no grupo é pequeno e os dados são balanceados e completos (todos os grupos com o mesmo número de indivíduos), utilizar a matriz não estruturada; se as observações são coletadas ao longo do tempo, então deve-se utilizar uma estrutura que considere a correlação em função do tempo (M-dependente, ou auto-regressiva); se as

observações estão agrupadas (ou seja, sem ordem lógica), deve-se utilizar a estrutura permutável e se o número de grupos for pequeno, os autores indicam a estrutura independente, com estimador robusto para variância, como a melhor escolha [32, 33].

Quando os dados se enquadram em mais de uma das situações citadas anteriormente, gerando dúvida sobre qual estrutura utilizar, pode-se optar por um critério estatístico para selecionar a estrutura de correlação mais adequada. Este critério é semelhante ao utilizado para selecionar as covariáveis que irão compor o modelo.

Pan [34] propôs um método de seleção de estrutura de correlação para GEE, semelhante ao AIC, mas considerando o fato de que as medidas possam ser correlacionadas. Este critério foi denominado critério de quasi-verossimilhança sob o modelo de independência (*Quasi-likelihood under the Independence model Criterion* - QIC) e para a matriz de correlação R é definido da seguinte forma:

$$\mathrm{QIC(R)} = -2\mathrm{Q}(\hat{\beta}(\mathrm{R}); \mathrm{I}, \mathcal{D}) + 2\mathrm{traço}(\hat{\Omega}_{\mathrm{I}}\hat{\mathrm{V}}_{\mathrm{R}})$$

onde Q é a quasi-verossimilhança, $\hat{\boldsymbol{\beta}}(R)$ é o vetor de estimadores de quasi-verossimilhança sob o modelo candidato com matriz de correlação R, I é a matriz identidade, $\mathcal{D}$ são os dados observados, $\hat{\Omega}_I = \dfrac{-\partial^2 Q(\beta, I, \mathcal{D})}{\partial\beta\partial\beta'} \mid \beta = \hat{\beta}$ e $\hat{V}(R)$ é o estimador de covariâncias robusto obtido através do modelo contendo a matriz de correlação R.

O QIC é calculado a partir da comparação de um modelo com uma determinada estrutura de correlação de trabalho com aquele gerado utilizando a estrutura independente. Os valores obtidos de QIC podem ser utilizados para comparar as diferentes estruturas de correlação. Do mesmo modo que para o AIC, quanto menor o valor do QIC, melhor o modelo. Algumas vezes ocorre de os valores do QIC não serem necessariamente muito diferentes. Nestes casos, Ballinger [31] recomenda que seja escolhido o modelo mais adequado segundo a teoria.

Mais recentemente, Hin e Wang [35] propuseram uma modificação do QIC, o Critério de Informação de Correlação (*Correlation Information Criterion – CIC*) para aperfeiçoar seu desempenho na escolha da estrutura de correlação de trabalho. Os autores implementaram o CIC no R (anexo F). Disponível também em http://www.mat.ufrgs.br/~camey/GEE_CIC)

3.3.5 Critérios para seleção das variáveis preditoras

Para selecionar o melhor conjunto de variáveis preditoras em um GLM utiliza-se o critério de informação de Akaike *(Akaike's Information Criterion* - AIC), que é baseado na máxima verossimilhança. Entretanto, no GEE em vez da verossimilhança é usada a quasi-verossimilhança, e, portanto, o AIC não pode ser utilizado.

Para solucionar este problema, Pan [34] propôs um método de seleção de modelo para GEE, uma modificação do QIC, o QIC_C, uma versão corrigida que penaliza a complexidade do modelo (isto é, recompensa a parcimônia).

$$QIC_C = -2Q(g^{-1}(X\,\boldsymbol{\beta}(R))) + 2p$$

onde Q é a quasi-verossimilhança calculada sob o modelo de independência, g^{-1} é a função de ligação inversa do modelo, X é a matriz de covariáveis, $\hat{\boldsymbol{\beta}}(R)$ é o vetor de estimadores de quasi-verossimilhança sob o modelo candidato com matriz de correlação R e $\mathrm{p} = \text{traço}(\hat{\Omega}_I \hat{V}_R)$.

Do mesmo modo que para o AIC e o QIC, quanto menor o valor do QIC_C, melhor o modelo.

3.3.6 Estimação

3.3.6.1 Estimação dos parâmetros do modelo

Para obter $\hat{\boldsymbol{\beta}}$, Liang e Zeger [4] sugerem um processo iterativo baseado no escore de Fisher modificado e nas estimativas de α e ϕ, obtidas através do método dos momentos (apresentadas nas secções 3.3.6.3 e 3.3.6.4, respectivamente). Neste processo, alterna-se entre estimar $\boldsymbol{\beta}$ para valores fixos de $\hat{\phi}$ e $\hat{\alpha}$ e estimar (ϕ,α) valores fixos de $\hat{\boldsymbol{\beta}}$.

O algoritmo apresentado a seguir pode ser utilizado para obter as estimativas de $\boldsymbol{\beta}$ através do GEE:

i. Calcular a estimativa inicial de $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_r$, através do modelo GLM, assumindo independência;

ii. A partir da estimativa de $\boldsymbol{\beta}$, calcular os resíduos padronizados, dados por:

$$\boldsymbol{r_{ij}} = (y_{ij} - \hat{\mu}_{ij}) / \sqrt{[\hat{V}_B]_{jj}}\,, \tag{6}$$

onde $\hat{V}_B = \sum_{i=1}^{n} D_i' V_i^{-1} D_i$.

Os resíduos padronizados são, então, utilizados para produzir estimativas consistentes para α e ϕ considerando a suposta estrutura de R;

iii. Calcular uma estimativa para covariância, através da equação (4), ou seja, $V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2} / \phi$;

iv. Atualizar $\hat{\boldsymbol{\beta}}_r$:

$$\hat{\boldsymbol{\beta}}_{r+1} = \hat{\boldsymbol{\beta}}_r + \left[\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} \boldsymbol{V}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right]^{-1} \left[\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)\right]; \tag{7}$$

v. Repetir os passos 2 a 4 até obter convergência.

3.3.6.2 Estimação de $V(\hat{\beta})$

Existem duas maneiras de estimar a variância de $\hat{\boldsymbol{\beta}}$, $V[\hat{\boldsymbol{\beta}}]$. O método mais simples é utilizar um estimador baseado no modelo (*model based ou naive estimator*), que é consistente quando o modelo para média e a estrutura de correlação, $R(\alpha)$, forem corretamente especificados. Como geralmente não se conhece a estrutura de correlação correta, é mais indicado utilizar um estimador empírico, também conhecido por estimador robusto ou estimador 'sanduíche'.

O estimador baseado no modelo utiliza a matriz de informação observada sob a suposição de uma determinada correlação e é definido, para o i-ésimo grupo, por:

$$\begin{aligned}\left[V_B\right]_i &= D'_i V_i^{-1} D_i \\ &= X'_i \mathrm{A}_i (\mathrm{A}_i^{1/2} \mathrm{R}_i(\alpha) \mathrm{A}_i^{1/2})^{-1} A_i X'_i\end{aligned} . \tag{8}$$

O estimador robusto ou estimador 'sanduíche' agrega à equação (8) uma matriz de informação, M_i, que utiliza resíduos empíricos, C_i, para estimar a matriz de covariância intra-grupo. A matriz M_i é dada por:

$$\begin{aligned}M_i = D'_i V_i^{-1} C_i V_i^{-1} D_i &= \\ &= X'_i A_i (A_i^{1/2} R_i(\alpha) A_i^{1/2})^{-1} C_i (A_i^{1/2} R_i(\alpha) A_i^{1/2})^{-1} A_i X'_i\end{aligned} \tag{9}$$

onde $C_i = (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)'$.

O estimador robusto é obtido avaliando todas as expressões sob o estimador $\hat{\boldsymbol{\beta}}$ e os respectivos valores de covariáveis, isto é,

$$V(\hat{\beta}) = \left(\sum_{i=1}^{n} \left[\hat{V}_B\right]_i\right)^{-1} \left(\sum_{i=1}^{n} \hat{M}_i\right) \left(\sum_{i=1}^{n} \left[\hat{V}_B\right]_i\right)^{-1} . \tag{10}$$

Este estimador é um estimador consistente da matriz de covariância de $\hat{\beta}$, mesmo quando a estrutura de correlação de trabalho não estiver bem especificada. Entretanto, é importante ressaltar que, por ele ser um estimador assintótico, suas propriedades são

garantidas somente quando o número de grupos é grande. Quando este número for pequeno (<20) o estimador de variância baseado no modelo pode apresentar propriedades melhores [26], mesmo se a especificação da matriz de covariância de trabalho estiver errada. Isto ocorre já que o estimador de variância robusto é assintoticamente não viciado, mas pode se tornar altamente viciado quando o número de grupos é pequeno.

3.3.6.3 Estimação da matriz de correlação de trabalho

Nos casos da estrutura fixa e da independente não é necessário estimar os parâmetros da matriz de correlação de trabalho. Os estimadores de α para cada uma das demais estruturas de correlação de trabalho envolvem os resíduos de Pearson, e_{ij}, e o parâmetro de dispersão ϕ e são apresentados no quadro 4.

Quadro 4: Estimadores para α de acordo com o tipo de estrutura de correlação de trabalho.

Estrutura	**Estimador para α**
Permutável	$\hat{\alpha} = \frac{1}{(N^* - p)\phi} \sum_{i=1}^{n} \sum_{j \neq k} e_{ij} e_{ik}$, onde $N^* = \sum_{i=1}^{n} m_i(m_i - 1)$
AR(1)	$\hat{\alpha}_t = \frac{1}{(K_1 - p)\phi} \sum_{i=1}^{n} \sum_{j \leq m_i - 1} e_{ij} e_{i,j+1}$, onde $K_1 = \sum_{i=1}^{n} (m_i - 1)$
M-dependente	$\hat{\alpha}_t = \frac{1}{(K_t - p)\phi} \sum_{i=1}^{n} \sum_{j < m_i - t} e_{ij} e_{i,j+t}$, onde $K_t = \sum_{i=1}^{n} (m_i - t)$
Não estruturada	$\hat{\alpha}_{jk} = \frac{1}{(K - p)\phi} \sum_{i=1}^{n} e_{ij} e_{ik}$

Onde $e_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\{[\hat{V}_i]_{jj}\}^{1/2}}$ são os resíduos de Pearson e $[\hat{V}_i]_{jj}$ é o j-ésimo elemento da diagonal da matriz $\hat{V}_i$.

3.3.6.4 Estimação do Parâmetro de dispersão

A estimativa para o parâmetro de dispersão ϕ é obtida através de

$$\hat{\phi} = \frac{1}{N-p} \sum_{i=1}^{n} \sum_{j=1}^{m_i} e_{ij}^2 ,$$

onde $N = \sum_{i=1}^{n} m_i$ é o número total de observações e p é o número de parâmetros de regressão.

3.3.7 Teste de hipóteses para os parâmetros do modelo

Usualmente, para testar a hipótese de que os coeficientes estimados pelo modelo são iguais a zero, é utilizada a estatística de Wald proposta por Rotnitzky e Jewell [36], definida por

$$T_W = K(\hat{\gamma} - \gamma_0)'(\hat{V}_R)^{-1}(\hat{\gamma} - \gamma_0).$$

onde a matriz de variâncias $\hat{V}_R$ é uma estimativa de variância que incorpora a estrutura de correlação dentro dos grupos. Esta estatística tem distribuição qui-quadrado com graus de liberdade igual para o número de parâmetros que são testados. Pode ser utilizada para testar a significância de um só parâmetro ou vários parâmetros.

Em casos nos quais a matriz de variâncias robusta não possa ser invertida, quando houver menos covariáveis do que observações por grupo, um teste de Wald de trabalho está disponível e é calculado usando o inverso da matriz de variância baseada no modelo [32].

3.3.8 R^2 marginal

Devido ao fato de não haver independência entre as observações, os resíduos também não são independentes. Logo, métodos baseados na verossimilhança e as medidas de ajuste modelo de regressão linear precisam ser adaptados. Zheng [37] introduziu uma extensão do R^2 para GEE em modelos com variável resposta contínua, binária, ou de contagem,

denominado R^2 *marginal*. É necessário calcular os valores preditos pelo modelo para obter o valor do R_m^2. Estes valores são comparados com os valores observados e divididos pela soma de quadrados dos desvios das observações em relação a média da variável de resposta. O R_m^2 é dado por:

$$R_m^2 = 1 - \frac{\sum_{j=1}^{m}\sum_{i=1}^{m}\left(y_{ij} - \hat{y}_{ij}\right)^2}{\sum_{j=1}^{m}\sum_{i=1}^{m}\left(y_{ij} - \overline{y}\right)^2},$$

onde $\overline{y} = \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n} y_{it}$ é a média marginal sobre todos os grupos.

O R_m^2, da mesma maneira como o R^2, é interpretado como o quanto da variância da variável resposta é explicada pela variabilidade do modelo ajustado [32]. Ele tem as mesmas propriedades que o R^2 para modelos de regressão, com a exceção de poder assumir valores inferiores a zero quando o modelo ajustado produz estimativas piores do que o modelo somente com o intercepto. Quando $m = 1$, $R_m^2 = R^2$. Zheng [37] enfatiza que a matriz de covariância do GEE não é explicitamente incluída no cálculo desta estatística. Atualmente o R_m^2 não está disponível em programas estatísticos.

3.3.9 Técnicas de diagnóstico para GEE

Após escolher as variáveis que compõe o modelo, deve-se verificar se o modelo é eficiente para descrever a relação entre as variáveis preditoras e a variável resposta. Para identificar a ocorrência de observações atípicas, são utilizadas as técnicas de diagnóstico. Estas técnicas também verificam se as suposições do modelo estão bem satisfeitas, se há presença de *outliers* e se o modelo está bem ajustado de acordo com as suas covariáveis.

Como mencionado anteriormente, o fato de não haver independência entre as observações faz com que os resíduos também não sejam independentes e, por isso, é necessário adaptar as técnicas de diagnóstico utilizadas nos GLM para o GEE.

3.3.9.1 Análise de resíduos

Uma das técnicas de diagnóstico mais utilizadas para modelos de regressão é a análise de resíduos. Um resíduo pode ser definido como a distância entre o valor estimado e seu correspondente valor observado da variável dependente [38]. O principal objetivo da análise de resíduos é identificar casos para os quais as estimativas do modelo se distanciem muito dos valores observados.

Os resíduos de Pearson no GEE são definidos do seguinte modo:

$$e_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\{[\hat{V}_i]_{jj}\}^{1/2}},$$

onde $[\hat{V}_i]_{jj}$ é o j-ésimo elemento da diagonal da matriz $\hat{V}_i$.

Chang [39] sugere avaliar o gráfico de dispersão dos resíduos *versus* cada tempo de seguimento (em estudos longitudinais) conjuntamente com o resultado do teste de Wald-Wolfowitz, definido a seguir.

3.3.9.2 Teste não paramétrico para aleatoriedade dos resíduos

Chang [39] indica a utilização do teste de Wald-Wolfowitz (*Wald-Wolfowitz run test*) para auxiliar na detecção de padrões de não-aleatoriedade nos resíduo. O teste codifica os resíduos em positivos (+) ou negativos (-). A seqüência de códigos é, então, analisada e é calculado o número total de repetições para cada um dos dois códigos. Este cálculo não considera o tamanho da repetição. Uma repetição é definida como uma seqüência de sinais iguais. Por exemplo, a seqüência + + – – + – + + – – – contém um total de quatro repetições.

Para executar o teste, considere T o número de repetições em uma seqüência de n_p resíduos positivos e de n_n resíduos negativos. Sob a hipótese nula de que os sinais dos resíduos estão distribuídos em uma seqüência aleatória, tem-se que a esperança de T é dada por:

$$E(T) = \frac{2n_p n_n}{n_p + n_n} + 1.$$

E sua variância é dada por:

$$V(T) = \frac{2n_p n_n (2n_p n_n - n_p - n_n)}{(n_p + n_n)^2 (n_p + n_n - 1)}.$$

A estatística do teste é $Z = \frac{T - E(T)}{\sqrt{V(T)}}$, que segue distribuição normal padrão sob a hipótese nula. Para valores absolutos altos de Z rejeita-se a hipótese de que os resíduos estão em seqüência aleatória, sugerindo que o modelo deve ser modificado a fim de melhor refletir a estrutura subjacente dos dados. Como este teste depende da ordem na qual estão dispostos os resíduos, Hardin e Hilbe [32] sugerem ordenar os resíduos pelo número de identificação dos grupos e pela ordem das observações dentro do grupo. Os autores ressaltam que o resultado obtido através deste teste não varia muito para estruturas de correlação diferentes se o modelo incluir as variáveis necessárias. Desse modo, eles sugerem a utilização deste teste para avaliar a adequação do modelo quanto às covariáveis incluídas, e o QIC para avaliar a adequação do modelo quanto à estrutura de correlação.

Este teste pode ser obtido no SPSS, utilizando o *run test*, através do qual avalia todos os indivíduos estudados ou através do teste de Wald-Wolfowitz, no qual é possível comparar dois grupos.

3.3.9.3 Outras técnicas de diagnóstico para GEE

Outras técnicas de diagnóstico descritas para os GLM foram adaptadas para o GEE por Preisser e Qaqish em 1996 [40]. Os autores apresentaram técnicas de diagnóstico baseadas na distância de Cook [41], DFBETAS e DFFITAS [42] para GEE, com o objetivo de medir a influência de um subconjunto de observações, tanto sobre os parâmetros estimados, como sobre os valores do preditor linear. Estas técnicas trabalham com a exclusão de uma observação (*observation-deletion diagnostic*) ou de um conjunto de observações, geralmente um grupo inteiro (*cluster deletion diagnostic*), para avaliar seu impacto nas estimativas dos parâmetros do modelo.

No site do John Preisser (http://www.bios.unc.edu/~jpreisse/personal/software.htm) está disponível um macro do SAS baseado nas técnicas de diagnóstico para GEE propostas por Preisser & Qaqish [40].

3.4 SISTEMAS DE INFORMAÇÃO EM SAÚDE

Através de Sistemas de Informação em Saúde é possível obter informações acerca da saúde da população. Com os dados disponíveis nestes sistemas, é possível acompanhar a evolução populacional do país proporcionando subsídios para implementar políticas públicas e monitoramento do exercício da cidadania [43].

As informações sobre nascidos vivos e sobre mortalidade são importantes para o planejamento e a avaliação das ações de saúde da criança no Sistema Único de Saúde (SUS), pois são usados no cálculo de vários indicadores de saúde, entre os quais os coeficientes de mortalidade infantil e materna. O Ministério da Saúde gerencia, entre outros, o Sistema de Informações sobre Mortalidade (SIM) [44], que possui informações sobre os óbitos ocorridas

no País; e o Sistema de Informações sobre Nascidos Vivos (SINASC) [45], que oferece informações sobre os nascimentos registrados no País.

3.4.1 Sistema de Informação sobre Mortalidade (SIM)

O Sistema de Informação sobre Mortalidade (SIM) foi criado pelo Ministério da Saúde em 1975 para a consolidação regular de dados nacionais sobre mortalidade. O SIM proporciona a produção de estatísticas de mortalidade e a construção dos principais indicadores de saúde. A análise dessas informações permite estudos não apenas do ponto de vista estatístico e epidemiológico, mas também sócio-demográfico.

O SIM é gerenciado pela Secretaria de Vigilância em Saúde (SVS/MS) e utiliza a Declaração de Óbito (DO) como instrumento padronizado de coleta de dados. O formulário da DO (Anexo D) possui três vias: a primeira é encaminhada à secretaria municipal de saúde e a partir dela são armazenados os dados do SIM; a segunda é entregue à família, que deve levá-la ao cartório para o registro de óbito; a terceira fica arquivada no prontuário do serviço de saúde onde ocorreu o óbito [46].

Após o encaminhamento à secretaria municipal de saúde, as DOs são codificadas e transcritas para um sistema informatizado. O Centro Nacional de Epidemiologia (Cenepi - Funasa) consolida os dados e os disponibiliza através do Departamento de Informática do SUS (DATASUS) via Internet ou em CD-ROM.

Quando ocorre óbito de crianças com idade inferior a um ano, o SIM fornece informações sobre características maternas (idade, escolaridade, número de filhos vivos tidos anteriormente), do indivíduo (data de ocorrência do óbito, peso ao nascer, sexo, índice de Apgar medido no primeiro e no quinto minuto de vida), da gestação (duração da gestação, tipo de gravidez – única, gêmeos, trigêmeos ou ordem superior - e número de consultas de pré-natal) e geográficas (local e estabelecimento da ocorrência do óbito; endereço/bairro de

residência da criança; endereço/bairro de residência da mãe (quando óbito fetal), além do tipo de óbito (fetal ou não fetal), causa de óbito (utilizando a 10ª Revisão da Classificação Internacional de Doenças - CID-10, a partir de 1996 e anteriormente, a 9ª Revisão - CID-9)).

Mais detalhes sobre o preenchimento da DO e o funcionamento do SIM podem ser obtidos nos manuais de preenchimento e procedimento editados pela Secretaria de Vigilância em Saúde, disponíveis nas secretarias estaduais e municipais de saúde e também no site www.saude.gov.br/svs.

3.4.2 Sistema de Informação sobre o Nascido Vivo (SINASC)

O Ministério da Saúde implantou o Sistema de Informações sobre Nascidos Vivos (SINASC) em 1990 com o objetivo de reunir informações epidemiológicas referentes aos nascimentos ocorridos em todo território nacional. Sua implantação ocorreu de forma lenta e gradual em todas as unidades da Federação.

No município de Porto Alegre, a implantação do SINASC pela Secretaria Municipal de Saúde de Porto Alegre ocorreu no ano de 1993. Atualmente, ele é processado pela Equipe de Informação em Saúde na Coordenação Geral de Vigilância Sanitária (EIS/CGVS) que anualmente atualiza o sistema. Seu principal instrumento é a declaração de nascido vivo (DN).

É importante ressaltar que, desde a implantação do SINASC, foi adotada uma definição única para nascido vivo, sendo ela: a expulsão ou a extração completa de um produto da concepção do corpo materno, independentemente da duração da gestação, o qual, depois da separação do corpo materno, respire ou dê qualquer outro sinal de vida, tais como: batimento cardíaco, pulsação do cordão umbilical ou movimentos efetivos dos músculos da contração voluntária, estando cortado ou não o cordão umbilical e estando ou não desprendida a placenta [43].

O formulário da DN (Anexo C) possui três vias: a primeira deve ser encaminhada ou recolhida pela secretaria municipal de saúde; a segunda, entregue à família, que a levará ao cartório para o pertinente registro de nascimento; a terceira deve ficar arquivada no prontuário do serviço de saúde responsável pelo parto [46].

O preenchimento da DN deve ocorrer logo após o nascimento, no serviço onde ocorreu o parto, por um profissional de saúde adequadamente treinado. No caso de partos domiciliares com assistência médica, a DN deve ser preenchida por um profissional de saúde que encaminhará sua primeira via para a obtenção da certidão de nascimento no Cartório de Registro Civil (que reterá o documento). Se o parto foi domiciliar, assistido por parteira tradicional, esta deverá informar tal fato ao serviço de saúde ao qual está vinculada – o qual preencherá a DN e distribuirá as três vias conforme o processo anteriormente descrito [46].

Mais detalhes sobre o preenchimento da DN e o funcionamento do SINASC podem ser obtidos nos manuais de preenchimento e procedimento editados pela Secretaria de Vigilância em Saúde, disponíveis nas secretarias estaduais e municipais de saúde e também no site www.saude.gov.br/svs.

A implementação do SINASC tornou possível, em nível populacional, a caracterização dos nascidos vivos do ponto de vista demográfico e epidemiológico, a partir de dados secundários. O SINASC fornece informações sobre características maternas (idade, escolaridade, número de filhos vivos tidos anteriormente), do recém nascido (data de ocorrência do nascimento, peso ao nascer, sexo, índice da Apgar medido no primeiro e no quinto minuto de vida), da gestação (duração da gestação, tipo de gravidez - gêmeos, trigêmeos ou ordem superior - e número de consultas pré-natal), do parto e geográficas (local - hospital, outros estabelecimentos de saúde, domicílio, outros - e estabelecimento da ocorrência do parto).

Como muitas informações presentes na DN também estão registradas na DO é possível a obtenção de coeficientes específicos de mortalidade infantil, necessários para análises mais minuciosas, na área de saúde materno-infantil [47]. Para que os resultados obtidos possam ser considerados confiáveis, é necessário avaliar as limitações do sistema, identificando o quão fidedignas e representativas são as informações coletadas e disponibilizadas. Neste sentido, é importante ressaltar estudos que têm procurado avaliar a eficácia do SINASC em coletar os nascimentos ocorridos, bem como a qualidade de seus registros [48, 49]. Nesse sentido, os pesquisadores buscam avaliar a cobertura obtida pelo sistema, quantificar o sub-registro e verificar o percentual da informação ignorada [47, 50]. Szwarcwald *et al* [51] verificaram que a Região Norte é a que possui as maiores deficiências, com 63% dos municípios com notificação inadequada (35% da população da região), seguida da Nordeste (29% da população). Já na Região Sul, somente 1% da população apresenta grande precariedade dos dados de óbitos. A qualidade dos dados registrados no SIM tem melhorado nos últimos anos, e sua cobertura tem sido bem próxima de 100% nas regiões Centro-Oeste, Sudeste e Sul do país [52].

No município de Porto Alegre, através da Equipe de Informação em Saúde da Coordenação Geral de Vigilância Sanitária (EIS/CGVS) da Secretaria Municipal de Saúde, existe uma excepcional qualidade dos dados do SINASC e do SIM [53]. Nesse sentido, cada Declaração de Óbito (DO) é avaliada e investigada em relação a sua causa de óbito correlacionando-se com outros sistemas de informação: SINASC e Sistema de Informação de Atenção Básica (SIAB), além da pesquisa de informações em prontuários hospitalares, quando necessário.

3.5 ENCADEAMENTO DE ARQUIVOS

A utilização simultânea do SIM e do SINASC permite o estudo da mortalidade infantil e de seus componentes segundo variáveis comuns à DN e à DO. Para isso, é necessário relacionar os registros destes dois bancos de dados.

A técnica de encadeamento de arquivos (*linkage*) pode ser utilizada tanto para agregar dados de um mesmo indivíduo provenientes de duas bases de dados distintas, como para identificar registros duplicados em um mesmo banco de dados. Através de um identificador único ou de algumas variáveis em comum é possível identificar indivíduos ou registros que fazem parte de dois bancos de dados [54], e, então, fazer o encadeamento destes arquivos. Uma das principais vantagens do encadeamento de arquivos é possibilitar a realização de estudos analíticos longitudinais com baixo custo.

Existem dois tipos de encadeamento de arquivos: o determinístico, baseado na concordância exata, e o probabilístico, no qual são utilizados modelos estatísticos para classificar pares de registros e através dos quais é possível mensurar o grau de concordância entre dois registros em bancos de dados distintos [55]. Espera-se que registros pertencentes ao mesmo indivíduo tenham grau de concordância maior em um conjunto de variáveis, quando comparados a registros que pertençam a diferentes indivíduos.

Quando cada indivíduo pode ser identificado nos dois bancos através de um campo identificador único (por exemplo: CPF, número de cartão de saúde), utiliza-se o método determinístico. Na ausência deste identificador, o relacionamento pode ser executado empregando-se o método probabilístico. Este último baseia-se na utilização conjunta de campos comuns presentes em ambos os bancos de dados (por exemplo: nome, data de nascimento, sexo), com o objetivo de identificar, através de modelos estatísticos, o quanto é provável que um par de registros se refira a um mesmo indivíduo [55-57].

Através desta técnica, Almeida e Mello Jorge [54] associaram os dados do SINASC com os dados de óbitos do Sistema de Informações sobre Mortalidade (SIM) do município de Santo André, em São Paulo, para avaliar a possibilidade do uso do relacionamento de registros para o estudo da mortalidade neonatal. As autoras ressaltaram a viabilidade do uso da técnica, mas chamam atenção para a necessidade de uma maior exatidão no preenchimento da DN, uma vez que detectaram a presença de Declaração de Óbito de crianças, sem que antes houvesse sido emitida a Declaração de Nascido Vivo. Machado e Hill [58] também utilizaram o relacionamento de registros para associar, de forma probabilística, os dados de nascimentos e óbitos da cidade de São Paulo para a coorte de 1998, com o objetivo de analisar os determinantes da mortalidade infantil.

O encadeamento de bases de dados vem sendo crescentemente utilizado para a monitorização de desfechos em estudos de coorte [54, 58]. Também pode ser utilizado na vigilância epidemiológica [59] e na melhoria da qualidade e quantidade de dados disponíveis em estudos que empregam fontes de dados secundários [60]. Instituições e pesquisadores nacionais e internacionais têm desenvolvido e aperfeiçoado estratégias de encadeamento de arquivos [61-63].

Apesar das vantagens da utilização do encadeamento de arquivos, é importante ressaltar que muitas vezes os dados secundários utilizados não possuem qualidade muito elevada. É importante avaliar a qualidade dos dados de identificação e sócio-demográficos, como, por exemplo, nome da mãe, data de nascimento e sexo. Essas variáveis são imprescindíveis para o bom funcionamento do *linkage* quando não há um número de identificação nos dois sistemas ou quando este se encontra duplicado. Em uma revisão sobre relacionamento de bases de dados [64] foram encontrados artigos que relatavam problemas de qualidade dos dados, como erros ortográficos e de digitação; dados como data de nascimento e sexo sem informação e pacientes com mesmo nome e mesmo número de prontuário, porém,

datas de nascimento muito diferentes. Fernandes [65] indica que o uso do SINASC é mais eficiente do que a coleta de informação no cartório para obtenção de dados que não estão preenchidos no SIM.

Neste trabalho, a fim de unificar os dados do SIM e do SINASC e identificar os grupos de irmãos, serão utilizadas as duas abordagens de encadeamento de arquivos. O método determinístico será utilizado para relacionar (quando possível) os registros de cada banco pelo número da DN. O método probabilístico será utilizado em dois momentos: (1) para relacionar cada óbito do SIM com o correspondente nascimento no SINASC utilizaram-se três variáveis (nome da mãe, data de nascimento e sexo), sendo posteriormente avaliado se o peso era o mesmo para evitar troca de informações entre irmãos e (2) para identificar grupos de irmãos, foram utilizadas duas variáveis (nome da mãe e data de nascimento). A escolha das variáveis foi baseada no estudo de Quantin *et al* [66]. Os autores avaliaram uma série de variáveis com o objetivo de distinguir quais seriam as melhores a serem utilizadas como identificadores no encadeamento probabilístico e elegeram como identificadores mais apropriados o nome da mãe e data de nascimento.

3.6 MORTALIDADE INFANTIL

A mortalidade infantil tem sido freqüentemente apontada como indicador sensível da qualidade de vida de uma população ([67] *apud* [68], [69]), determinada em sua dimensão mais ampla pelas condições sociais, econômicas e culturais dos indivíduos e da comunidade a qual pertencem.

A Taxa de Mortalidade Infantil (TMI) é uma estimativa do risco de morte a que está exposta uma população de nascidos vivos em uma determinada área e período, antes de

completar o primeiro ano de vida. É considerado um indicador sensível das condições de vida e saúde de uma comunidade [67]. É calculado da seguinte forma:

$$\text{TMI} = \frac{\text{número de óbitos de menores de 1 ano em determinada área e período}}{\text{número de nascidos vivos na mesma área e período}} \times 1.000.$$

Para uma melhor avaliação da mortalidade infantil é possível dividir a taxa de mortalidade infantil em dois componentes, de acordo com a idade na qual ocorreu o óbito:

a) **Taxa de Mortalidade Neonatal (ou precoce):** expressa a proporção de óbitos de crianças nascidas vivas com idade entre 0 e 27 dias (inclusive) em relação ao total de nascidos vivos em uma determinada área e período;

$$\text{TMN} = \frac{\text{número de óbitos de crianças de 0 a 27 dias em determinada área e período}}{\text{número de nascidos vivos na mesma área e período}} \times 1.000.$$

b) **Taxa de Mortalidade Pós-Neonatal (ou tardia):** expressa a proporção de óbitos em crianças nascidas vivas com idade entre 28 e 364 dias (inclusive) em relação ao total de nascidos vivos em uma determinada área e período.

$$\text{TMPN} = \frac{\text{número de óbitos de crianças de 28 a 364 dias em determinada área e período}}{\text{número de nascidos vivos na mesma área e período}} \times 1.000.$$

O risco de morte varia ao longo do primeiro ano de vida, principalmente quando se consideram as causas de óbito e seus fatores determinantes. Por isso, é importante analisar esses dois coeficientes separadamente, uma vez que no período neonatal predominam as causas de

óbito ligadas a problemas da gestação e do parto (afecções perinatais e anomalias congênitas). Medronho *et al* [68] apontam como fatores de grande importância na determinação da mortalidade infantil neonatal a cobertura e a qualidade da assistência pré-natal e perinatal. Já no período pós-neonatal, prevalecem as causas de óbito relacionadas ao meio ambiente e às condições de vida e de acesso aos serviços de saúde (doenças infecciosas, pneumonias, diarréia, por exemplo). Rouquayrol *et al* [70] avaliam que ao se comparar diferentes países, verifica-se que quanto melhor o nível de saúde, menor a proporção de óbitos pós-neonatais. Também está demonstrado que, para uma mesma região ou país, ao se organizar uma série histórica dos índices de mortalidade infantil, desdobrados em seus componentes neo e pós-neonatal, existe uma tendência de aumento progressivo da proporção de óbitos neonatais, cujas causas são de controle mais difícil e complexo. Dessa forma, nos países desenvolvidos, onde a mortalidade infantil é baixa e problemas relacionados ao meio ambiente já se encontram quase totalmente resolvidos, o componente neonatal predomina, enquanto em muitos países pobres ainda prevalece o componente pós-neonatal [70].

Particularmente no Brasil, a redução da mortalidade infantil ainda é um desafio. Apesar da tendência mundial e nacional de declínio do componente pós-neonatal [71], os índices continuam elevados, pois sua redução encontra obstáculos no componente neonatal, o que pode estar refletindo as desigualdades sociais, a cobertura e a qualidade da assistência à saúde. Em 1990, a proporção de óbitos neonatais ainda era menor do que a de pós neonatais nas regiões Norte e Nordeste do País. Já no ano 2000 pelo menos 60% dos óbitos infantis ocorreram no período neonatal em todas as regiões brasileiras [70].

3.7 GESTAÇÕES MÚLTIPLAS E MORTALIDADE

A gestação múltipla é definida pela existência de mais de um feto durante a gravidez. Esta gestação pode ter como desfecho dois (gêmeos), três (trigêmeos) ou ainda, um número superior de recém-nascidos.

Estudos com gêmeos sempre foram considerados de grande valor no aprendizado da etiologia de doenças, especialmente por possibilitarem a separação de efeitos ambientais e genéticos. Segundo Carlin e colaboradores [11], irmãos gêmeos são de especial interesse por serem indivíduos naturalmente pareados, com os quais é possível realizar análises controladas por um grande número de confundidores compartilhados por eles.

Historicamente, a gestação múltipla tem sido relacionada com o aumento do risco da morbidade e mortalidade no período neonatal e também por um subseqüente atraso no crescimento e desenvolvimento infantil [72, 73] *apud* [74]. A taxa de nascimentos múltiplos tem aumentado nos últimos anos, tanto em países desenvolvidos [75-78] como em países em desenvolvimento [79]. Um estudo realizado na Inglaterra [78] aponta para um aumento de aproximadamente 25% nas taxas de gêmeos entre 1980 e 1993, mostrando ainda que as taxas de trigêmeos ou de ordem superior dobraram. Nos Estados Unidos, ente 1980 e 1997, foi observado um aumento mais elevado nessas taxas, sendo superior a 50% para gêmeos e em torno de 400% para trigêmeos [76]. Em Porto alegre, foi observado um aumento de aproximadamente 30% nas taxas de nascimentos múltiplos (de 1,95% em 1994 foi para 2,53% em 2005) [79]. Esse aumento tem sido atribuído principalmente a dois fatores: (1) uso da estimulação ovariana e da fertilização *in vitro*; e (2) aumento da idade materna.[80-82].

Apesar de representarem apenas 1% a 2% de todos os nascimentos, os nascimentos múltiplos estão associados ao nascimento pré-termo, ao baixo peso ao nascer e a maiores índices de morbidade e mortalidade perinatal [78] e neonatal [83]. Dentre as complicações

obstétricas associadas com a gestação gemelar estão o aumento da incidência de hipertensão induzida pela gravidez, a hemorragia anteparto (antepartum), o parto prematuro e a necessidade de cesárea. Problemas neonatais associados à gemelaridade incluem baixo peso ao nascer e aumento da prevalência de malformações congênitas [78].

Estudos sugerem que a taxa de mortalidade infantil é maior para gestação múltipla (gêmeos ou trigêmeos ou de ordem superior) do que em uma gestação única [78, 84]. Além disso, verifica-se um aumento substancial na incidência de morbidades e mortalidade na medida em que aumenta o número de fetos no útero [85]. Porém, este aumento na incidência de mortalidade e morbidade neonatal ocorre devido, principalmente, às complicações associadas com o nascimento de pré-termo destas crianças [85]. Estes autores mostraram que o nascimento pré-termo é um fator de risco importante para desfechos neonatais e que a idade gestacional é inversamente proporcional ao número de fetos por gestação. Martin e Park [86] mostraram que 90% dos trigêmeos nascem pré-termos, e que trigêmeos e nascimentos de ordem superior tem 12 vezes a chance de morrer durante o primeiro ano de vida quando comparados aos nascimentos únicos. Huang verificou que a chance de óbito em um trigêmeo aumenta com o aumento do número de irmãos mortos na mesma gestação [1].

Pelas suas repercussões na morbimortalidade infantil, os nascimentos múltiplos tornaram-se um tema importante na área da saúde pública. Fetos de uma mesma gestação estão sujeitos a condições semelhantes no útero e são afetados mais ou menos igualmente pelas mesmas características maternas. Isso faz com eles sejam mais semelhantes do que os fetos de gestações diferentes [87] e, conseqüentemente, suas respostas são susceptíveis a estarem correlacionadas uma com a outra. Em estudos com gêmeos, a probabilidade de um resultado negativo, tal como a morte neonatal e perinatal, por um dos gêmeos aumenta se o co-gêmeo também apresentou esse resultado [1]. Ananth e Preisser [88] sugerem que uma gravidez múltipla como a de gêmeos e trigêmeos seja um conglomerado natural em que as

respostas dos fetos são interdependentes ou agregadas. Desse modo, não é possível utilizar as metodologias tradicionais de regressão, que supõe independência entre os indivíduos observados.

Considerando esta tendência de aumento da taxa de gemelaridade e seu impacto na mortalidade infantil e seus componentes, faz-se necessário à utilização de uma metodologia adequada para avaliar desfechos em gemelares.

4. OBJETIVO

Apresentar a metodologia de Equações de Estimação Generalizadas, através de uma aplicação na análise de dados de mortalidade neonatal em gemelares.

5. REFERÊNCIAS BIBLIOGRÁFICAS

[1] Huang JS, Lu SE, Ananth CV. The clustering of neonatal deaths in triplet pregnancies: application of response conditional multivariate logistic regression modes. J Clin Epidemiol. 2003;56(12):1202-9.

[2] Fitzmaurice G. Clustered data. Nutrition. 2001;17(6):487-8.

[3] Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics. 1986;42(1):121-30.

[4] Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73(1):13-22.

[5] Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics. 1982;38(4):963-74.

[6] Fausto MA, Carneiro M, Antunes CMF, Pinto JA, Colosimo EA. O modelo de regressão linear misto para dados longitudinais: uma aplicação na análise de dados antropométricos desbalanceados. Cad Saúde Pública. 2008;24(3):513-24.

[7] Twisk JW. Longitudinal data analysis. A comparison between generalized estimating equations and random coefficient analysis. Eur J Epidemiol. 2004;19(8):769-76.

[8] McCullagh P. Generalized Linear Models: Chapman and Hall 1983.

[9] Nelder JA, Wedderburn RWM. Generalized linear models. J R Stat Soc (Series A). 1972;135(3):370-84.

[10] Cordeiro GM, Demétrio CGB. Modelos Lineares Generalizados. Santa Maria: 12o SEAGRO 2007.

[11] Carlin JB, Gurrin LC, Sterne JA, Morley R, Dwyer T. Regression models for twin studies: a critical review. Int J Epidemiol. 2005;34(5):1089-99.

[12] Cannon MJ, Warner L, Taddei JA, Kleinbaum DG. What can go wrong when you assume that correlated data are independent: an illustration from the evaluation of a childhood health intervention in Brazil. Stat Med. 2001;20(9-10):1461 - 7.

[13] Rao CR. The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. Biometrika. 1965;52(3):447-58.

[14] Grizzle JE, Allen DM. Analysis of growth and dose response curves. Biometrics. 1969;25(2):357-81.

[15] Hui SL. Curve fitting for repeated measurements made at irregular time points. Biometrics. 1984;40(3):691-7.

[16] Fearn T. A Bayesian approach to growth curves. Biometrika. 1975;62(1):89-100.

[17] Harville DA. Maximum likelihood approaches to variance component estimation and to related problems. J Am Stat Assoc. 1977;72(358):320-38.

[18] Azzalini A. Estimation and hypothesis testing for collections of autoregressive time series. Biometrika. 1984;71(1):85-90.

[19] Ware JH. Linear models for the analysis of serial measurements in longitudinal studies. Am Stat. 1985;39(2):95- 101.

[20] Vonesh EF, Carter RL. Mixed effect nonlinear regression for unbalanced repeated measures. Biometrics. 1992;48(1):1-17.

[21] Stiratelli R, Laird N, Ware JH. Random-effects models for serial observations with binary response. Biometrics. 1984;40(4):961-71.

[22] Lipsitz SR, Laird NM, Harrington DP. Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. Biometrika. 1991;78(1):153-60.

[23] Anderson DA, Aitkin M. Variance component models with binary response: Interviewer variability. J R Stat Soc (Series B). 1985;47(2):203-10.

[24] Gilmour AR, Anderson RD, Rae AL. The analysis of binomial data by a generalized linear mixed model. Biometrika. 1985;72(3):593-9.

[25] Wedderburn RWM. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika. 1974;61(3):439-47.

[26] Prentice R. Correlated binary regression with covariates specific to each binary observation. Biometrics 1988;44(4):1033-48.

[27] Zorn CJW. Generalized estimating equation models for correlated data: A review with applications. Am J Pol Sci. 2001;45(2):470-90.

[28] Qu A, Lindsay BG, Li B. Improving generalized estimating equations using quadratic inference functions. Biometrika. 2000;87(4):823-36.

[29] Fitzmaurice GM. A caveat concerning independence estimating equations with multivariate binary data. Biometrics. 1995;51(1):309-17.

[30] Diggle PJ, Heagerty P, Liang K-Y & Zeger SL. Analysis of longitudinal data. 2nd ed ed: Oxford, UK: Oxford University Press 2002.

[31] Ballinger GA. Using Generalized Estimating Equations for Longitudinal Data Analysis. Organizational Research Methods. 2004;7(2):127-50.

[32] Hardin JW, Hilbe JM. Generalized estimating equations: Chapman and Hall / CRC Press 2003.

[33] Horton NJ, Lipsitz SR. Review of software to fit Generalized Estimating Equation regression models. Am Stat. 1999;53(2):160-9.

[34] Pan W. Information criterion in generalized estimating equations. Biometrics. 2001;57(1):120-5.

[35] Hin L-Y, Wang Y-G. Working-correlation-structure identification in generalized estimating equations. Stat Med. 2009;28(4):642–58.

[36] Rotnitzky A, Jewell NP. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. Biometrika. 1990;77(3):485-97.

[37] Zheng B. Summarizing the goodness of fit of generalized linear models for longitudinal data. Stat Med. 2000;19(10):1265-75.

[38] Cox DR, Snell EJ. A general definition of residuals. J R Stat Soc (Series B) 1968;30 (2):248-75.

[39] Chang YC. Residuals analysis of the generalized linear models for longitudinal data. Stat Med. 2000;19(10):1277-93.

[40] Preisser JS, Qaqish BF. Deletion diagnostics for generalized estimating equations. Biometrika. 1996;83(3):551-62.

[41] Cook RD. Deletion of Influential Observations in Linear Regression. Technometrics. 1977;19(1):15-8.

[42] Belsley DA, Kuh E, welsh RE. Regression diagnostics: identifying influential data sources of collinearity. New York 1980.

[43] IBGE. Notas técnicas. Disponível em http://www.ibge.gov.br/home/estatistica/populacao/registrocivil/2008/notastecnicas.pdf. [Acessado em novembro de 2009]

[44] Ministério_da_Saúde. Manual de procedimentos do Sistema de Informações sobre Mortalidade. Brasília: Ministério da Saúde; 2001.

[45] Ministério_da_Saúde. Manual de procedimentos do Sistema de Informações sobre Nascidos Vivos Brasília: Ministério da Saúde; 2001.

[46] Mello-Jorge MHP, Laurenti R, Gotlieb SLD. Análise da qualidade das estatísticas vitais brasileiras: a experiência de implantação do SIM e do SINASC. Ciência & Saúde Coletiva. 2007;12(3):643-54.

[47] Mello Jorge MHP, Gotlieb SLD, Soboll M, Almeida MF, Latorre M. Avaliação do Sistema de Informação sobre Nascidos Vivos e o uso de seus dados em epidemiologia e estatísticas de saúde. Rev Saude Publica. 1993;27 (suppl 6):1-46.

[48] Silva AAM, Ribeiro VS, Borba Junior AF, Coimbra LC, Silva RA. Avaliação da qualidade dos dados do Sistema de Informações sobre Nascidos Vivos em 1997-1998. Rev Saude Publica. 2001;35(6):508-14.

[49] Mello Jorge MHP, Gotlieb SLD, Oliveira H. O Sistema de Informação sobre Nascidos Vivos: primeira avaliação dos dados brasileiros. Inf Epidemiol SUS. 1996;5:15-48.

[50] Silva RI, Theme Filha MM, Noronha CP. Sistema de informação sobre nascidos vivos na cidade do Rio de Janeiro, 1993/1996. Inf Epidemiol SUS. 1997;6(2):33-48.

[51] Szwarcwald C, Leal M, Andrade C, Souza Jr. P. Estimação da mortalidade infantil no Brasil: o que dizem as informações sobre óbitos e nascimentos do Ministério da Saúde? Cad Saúde Pública. 2002;18(6):1725-36.

[52] Brasil. Ministério da Saúde. Secretaria Executiva. Datasus. Indicadores e Dados Básicos: Brasil 2005 - IDB 2005. Disponível em http://tabnet.datasus.gov.br/cgi/idb2007/matriz.htm#cober. [Acessado em 21 de junho de 2009].

[53] Shimakura SE, Carvalho MS, Aerts DRGC, Flores R. Distribuição espacial do risco: modelagem da mortalidade infantil em Porto Alegre, Rio Grande do Sul, Brasil. Cad Saúde Pública. 2001;17(5):1251-61.

[54] Almeida MF, Mello-Jorge MHP. O uso da técnica de "Linkage" de sistemas de informação em estudos de coorte sobre mortalidade neonatal. Rev Saude Publica 1996;30(2):141-7.

[55] Jaro MA. Probabilistic linkage of large public health data files. Stat Med. 1995;14(5-7):491-8.

[56] Fellegi I, Sunter A. A theory for record linkage. J Am Stat Assoc. 1969;64(328):1183-210.

[57] Newcombe HB, Kennedy JM, Axford SJ, James AP. Automatic Linkage of Vital Records. Science. 1959;130(3381):954-9.

[58] Machado CJ, Hill K. Determinantes da mortalidade neonatal e pós-neonatal no município de São Paulo. Rev Bras Epilemiol. 2003;6(4):345-58.

[59] Lucena FFA, Fonseca MGP, Sousa AIA, Coeli CM. O relacionamento de bancos de dados na implementação da vigilância da Aids. Relacionamento de dados e vigilância da Aids. Cad Saúde Colet. 2006;14(2):305-12.

[60] Teixeira CL, Klein CH, Bloch KV, Coeli CM. Reclassificação dos grupos de causas prováveis dos óbitos de causa mal definida, com base nas Autorizações de Internação Hospitalar no Sistema Único de Saúde, Estado do Rio de Janeiro, Brasil. Cad Saúde Pública. 2006;22(6):1315-24.

[61] Jaro MA. Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. J Am Stat Assoc. 1989;84:414-20.

[62] Winkler WE. Advanced Methods for Record Linkage. Technical Report Washington, DC: Statistical Research Division, US Bureau of the Census; 1994 Disponível em http://wwwcensusgov/srd/www/byyearhtml.

[63] Portela M, Schramm J, Pepe V, Noronha M, Pinto C, Cianeli M. Algoritmo para a composição de dados por internação a partir do sistema de informações hospitalares do sistema único de saúde (SIH/SUS) - Composição de dados por internação a partir do SIH/SUS. . Cad Saúde Pública. 1997;13(3):771-4.

[64] Coeli CM, Camargo-Jr KR. Relacionamento de Bases de Dados em Saúde. Cad Saúde Colet. 2006;14(2):197-224.

[65] Fernandes D. Concatenamento de informações sobre óbitos e nascimentos: uma experiência metodológica do Distrito Federal - 1989-1991. Belo Horizonte: Tese (Doutorado em Demografia) - Faculdade de Ciências Econômicas, UFMG 1997.

[66] Quantin C, Binquet C, Bourquard K, Pattisina R, Gouyon-Cornet B, Ferdynus C, et al. Which are the best identifiers for record linkage? Med Inform Internet Med. 2004;29(3-4):221-7.

[67] UNICEF - Fundação das Nações Unidas para Infância. Situação mundial da infância. . Brasília 1989.

[68] Medronho R, Bloch K, Luiz R, Werneck G. Epidemiologia. 2a ed. São Paulo 2009.

[69] Aertz DRGC. Investigação dos óbitos perinatais e infantis: seu uso no planejamento de políticas públicas de saúde. J Pediatr. 1997;73(6):364-6.

[70] Rouquayrol MZ, Almeida-Filho N, (organizadores). Epidemiologia e Saúde. 6a. ed. Rio de Janeiro 2006.

[71] Brasil. A mortalidade perinatal e neonatal no Brasil. Brasília (DF): Ministério da Saúde. Unicef 1998.

[72] Wenstrom KD, Gall SA. Incidence, morbidity and mortality, and diagnosis of twin gestation. Clin Perinatol. 1988;15(1):1-11.

[73] Leonard CH, Piechuch RE, Ballard RA, Cooper BAB. Outcome of Very Low Birth Weight Infants: Multiple Gestation Versus Singletons. Pediatrics. 1994;93(4):611-5.

[74] Homrich da Silva C. Baixo Peso ao Nascer e Gemelaridade no Município de Porto Alegre (Brasil): Um Novo Desafio. Porto Alegre, RS: Tese (Doutorado em Pediatria) - UFRGS 2006.

[75] Millar WJ, Wadhera S, Nimrod C. Multiple births: trends and patterns in Canada, 1974–90. Health Reports 1992;4(2):223-50.

[76] Martin J, Kung H, Mathews T, Hoyert D, Strobino D, Guyer B. Annual summary of vital statistics: 2006. Pediatrics. 2008;121(4):788-801.

[77] Dunn A, Macfarlane A. Recent trends in the incidence of multiple births and associated mortality in England and Wales. Arch Dis Fetal Neonatal 1996;75(1):9-10.

[78] Doyle P. The outcome of multiple pregnancy. Hum Reprod. 1996;11 (Suppl 4):110-7.

[79] Homrich da Silva C, Goldani MZ, de Moura Silva AA, Agranonik M, Bettiol H, Barbieri MA, et al. The rise of multiple births in Brazil. Acta Paediatr. 2008;97(8):1019-23.

[80] Luke B. The changing pattern of multiple births in the United States: Maternal and infant characteristics, 1973 and 1990. Obstet Gynecol. 1994;84(1):101-6.

[81] Jewel S, Yip R. Increasing trends in plural births in the United States. Obstet Gynecol. 1995;85(2):229-32.

[82] Angel JL, Kalter CS, Morales WJ, Rasmussen C, Caron L. Aggressive perinatal care for higher order multiple gestations: Does good perinatal outcome justify aggressive assisted reproductive techniques? Am J Obstet Gynecol 1999;181(1):253-9.

[83] Kaufman GE, Malone FD, Harvey-Wilkes KB, Chelmow D, Penzias AS, D'Alton ME. Neonatal morbidity and mortality associated with triplet pregnancy. Obstet Gynecol. 1998;91(3):342-8.

[84] Ferguson W. Perinatal mortality in multiple pregnancy. A review of perinatal deaths from 1609 multiple gestations. Obstet Gynecol. 1964;23 861-70.

[85] Luke B, Keith L. The contribution of singletons, twins and triplets to low birth weight, infant mortality and handicap in the United States. J Reprod Med. 1992;37(8):661-6.

[86] Martin J, Park M. Trends in twin and triplet births: 1980-97. Natl Vital Stat Rep. 1999;47(24):1-16.

[87] Ananth CV, Platt RW, Savitz DA. Regression models for clustered binary responses: implications of ignoring the intracluster correlation in an analysis of perinatal mortality in twin gestations. Ann Epidemiol. 2005;15(4):293-301.

[88] Ananth CV, Preisser JS. Bivariate logistic regression: modelling the association of small for gestational age births in twin gestations. Stat Med. 1999;18(15):2011-23.

[89] Twisk JWR. Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide: Cambridge University Press 2003.

6. ARTIGO

Equações de Estimação Generalizadas (GEE): aplicação em estudo sobre mortalidade neonatal em gemelares de Porto Alegre, RS (1995-2007)

Generalized Estimating Equations (GEE): an application on multiple births mortality in Porto Alegre, Brazil (1995-2007)

Marilyn Agranonik, Mestranda em Epidemiologia, UFRGS;

Marcelo Zubaran Goldani, UFRGS

Suzi Alves Camey, UFRGS

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL (UFRGS)

A ser enviado aos Cadernos de Saúde Pública.

Equações de Estimação Generalizadas (GEE): aplicação em estudo sobre mortalidade neonatal em gemelares de Porto Alegre, RS (1995-2007).

Marilyn Agranonik[1], Marcelo Zubaran Goldani[2], Suzi Alves Camey[1,3]

[1]Programa de Pós-Graduação em Epidemiologia, Universidade Federal do Rio Grande do Sul.
[2]Departamento de Pediatria e Puericultura da Faculdade de Medicina, Universidade Federal do Rio Grande do Sul.
[3]Departamento de Estatística, Universidade Federal do Rio Grande do Sul.

Running title: **GEE: aplicação em estudo sobre mortalidade neonatal em gemelares.**

Correspondence author: Marilyn Agranonik
Email: m_agrano@yahoo.com.br
Rua Ramiro Barcelos 2350, Porto Alegre, RS, Brasil.
CEP 90035-003

Resumo

Em estudos com gêmeos e trigêmeos é possível existir correlação entre os dados dos irmãos. Desse modo, modelos de regressão tradicionais podem levar a inferências incorretas, uma vez que a suposição de independência entre os sujeitos não é mais satisfeita. Para solucionar este problema, Zeger e Liang (1986) propuseram uma classe de Equações de Estimação Generalizadas (GEE), semelhante aos GLM, porém, incluindo uma estrutura de correlação de trabalho nas estimativas dos parâmetros do modelo. Ainda hoje, poucos estudos utilizam esta metodologia. Este artigo apresenta o GEE, através de uma aplicação na análise da mortalidade neonatal em gemelares. Verificou-se que o baixo peso ao nascer e o índice de Apgar associaram-se à mortalidade neonatal. Comparando os resultados obtidos no GEE com os do GLM foram encontradas pequenas diferenças nas estimativas pontuais dos parâmetros do modelo. Contudo, ao comparar erros padrões, as diferenças foram maiores, interferindo na significância de uma das variáveis. Desse modo, recomenda-se usar o GEE quando houver agrupamento de indivíduos.

Palavras chave: dados correlacionados, Equações de Estimação Generalizadas, GEE, mortalidade neonatal, gemelares.

Abstract

When studying twins and triplets it is possible that correlation exists between siblings' data. Thus, traditional regression models can lead to incorrect inferences, since the independence assumption among subjects is no longer satisfied. To solve this problem, Zeger and Liang (1986) proposed a class of Generalized Estimation Equations (GEE), similar to the GLM, however, including a working correlation structure in the parameters estimates. There are few studies using this methodology yet. This article presents GEE through an application in neonatal mortality analysis in multiple births. Perinatal factors, such as low birth weight and Apgar scores were associated with neonatal mortality. Comparing the results obtained by GEE and GLM, small differences were found in model parameters estimates. However, when comparing SEs, the differences were larger, interfering with the significance of a variable. Therefore, we recomend using GEE when working with correlated data.

Key words: correlated data, Generalized Estimation Equations, neonatal mortality, multiple births.

Introdução

Em estudos epidemiológicos, quando uma informação sobre uma determinada variável é coletada repetidas vezes ao longo do tempo, apesar dos sujeitos estudados serem independentes, suas observações podem estar correlacionadas. É possível também que os sujeitos dividam características em comum (por exemplo, estudantes de uma mesma escola, pacientes de um mesmo hospital, pessoas que trabalham em um mesmo local, irmãos,...) e, portanto, não podem ser considerados independentes. Nesta situação pode haver correlação entre os sujeitos. Segundo Carlin e colaboradores[1], irmãos gêmeos são de especial interesse em pesquisas por serem indivíduos naturalmente pareados, com os quais é possível realizar análises controladas por um grande número de confundidores compartilhados por eles. Além disso, possibilitam a separação de efeitos ambientais e genéticos. Entretanto, assim como nos demais estudos com dados agrupados, é possível existir correlação entre os dados dos irmãos. A correlação, nesses casos, pode ocorrer já que as observações feitas em um mesmo indivíduo (estudos longitudinais) ou em pessoas de um mesmo grupo (dados agrupados) tendem a ser mais semelhantes do que observações de indivíduos diferentes ou de grupos diferentes[2]. É o que ocorre com indivíduos nascidos de gravidez múltipla, como a de gêmeos e trigêmeos, na qual as respostas dos fetos são interdependentes ou agregadas, podendo, essa gravidez, ser considerada um conglomerado natural[3].

Para análise de estudos longitudinais ou de dados agrupados, os modelos tradicionais de regressão têm uso limitado, devido à suposição de independência entre os sujeitos. Este é o caso dos Modelos Lineares Generalizados (GLM)[4,5]. Apesar de este ser um método poderoso e flexível, se for utilizado para dados correlacionados, é provável a obtenção de distorções nas estimativas dos parâmetros e de seus erros padrões, levando a inferências estatísticas incorretas[1,6].

Uma alternativa é a utilização da Análise de Variância (ANOVA) de medidas repetidas para avaliar mudanças em um desfecho contínuo ao longo do tempo e comparar estas mudanças entre grupos. Entretanto, além deste método ser utilizado somente para desfechos com distribuição normal, não permite ajuste para exposições que se modifiquem ao longo do tempo, além de necessitar de balanceamento em relação ao número de repetições.

Para dados não normais e correlacionados, as principais abordagens estatísticas são: as Equações de Estimação Generalizadas (*Generalized Estimating Equations - GEE*)[6,7] e os modelos de efeitos aleatórios (um caso especial de modelos mistos ou de modelos multiníveis[8]). Estas técnicas, inicialmente desenvolvidas para variáveis resposta com distribuição normal, foram estendidas para variáveis com outras distribuições[9,10].

Zeger e Liang[6,7] propuseram, no final dos anos 80, uma classe de Equações de Estimação Generalizadas (*Generalized Estimating Equations* - GEE) para estimar parâmetros de regressão quando se trabalha com dados correlacionados. Este método foi desenvolvido para produzir estimativas mais eficientes e não viciadas para os parâmetros do modelo de regressão nesta situação, pois considera uma estrutura de correlação entre as observações. No modelo de efeitos aleatórios proposto por Laird & Ware[8] os coeficientes de regressão podem ser diferentes entre indivíduos, considerando a heterogeneidade existente entre eles.

A principal diferença entre estes métodos está no fato do GEE avaliar a relação entre a variável resposta e as variáveis preditoras em um contexto populacional, e não individual, enquanto o modelo de efeitos aleatórios tem como foco o indivíduo. Desse modo, quando se tem interesse em avaliar diversas medidas de um mesmo indivíduo, ao longo do tempo, e avaliar seu crescimento individual, é mais indicado utilizar um modelo de efeitos aleatórios. E, quando se estiver interessado em estudos epidemiológicos, por exemplo, com o objetivo de se estudar a diferença na resposta média populacional entre dois grupos com diferentes fatores de risco, o GEE é o método mais recomendado[11].

Ainda hoje é pouco comum encontrar artigos, especialmente no Brasil, que utilizem a modelagem apropriada quando estão presentes observações correlacionadas. Considerando o crescente número de estudos epidemiológicos envolvendo observações correlacionadas, seja em estudos longitudinais ou em estudos envolvendo dados agrupados, e os problemas que podem ocorrer com a utilização da análise inadequada, este artigo tem por objetivo apresentar a metodologia GEE, através de uma aplicação na análise de dados de mortalidade neonatal em gemelares (gêmeos, trigêmeos ou de ordem superior).

Material e métodos

Foi realizado um estudo de coorte retrospectivo. Nas análises foram utilizadas apenas informações de nascimentos gemelares, nos quais todas as crianças que constituem o par ou o trio nasceram vivas em Porto Alegre no período de 1995 a 2007. Essas informações foram obtidas através de dados do Sistema de Informações de Nascidos Vivos (SINASC), desenvolvido através de informações da Declaração de Nascimento (DN) e do Sistema de Informações de Mortalidade (SIM), desenvolvido por intermédio de informações da Declaração de Óbito (DO) e fornecidos pela secretaria municipal de saúde de Porto Alegre. A utilização simultânea desses dois sistemas de informação permite o estudo da mortalidade infantil e de seus componentes segundo variáveis comuns à DN e à DO.

Para as análises, os bancos SIM e SINASC foram unificados e recodificados. A construção do banco ocorreu em duas etapas: primeiro foram relacionados os registros de nascimento e óbito através do número da DN; quando esta informação não estava disponível foi utilizando o programa *Link Plus* versão 9.0[12] para relacionar estes bancos através do nome da mãe, do peso e da data de nascimento. Em uma segunda etapa os irmãos foram relacionados através do nome da mãe e da data de nascimento. Os casos identificados nesta

segunda etapa receberam um número de identificação para ser utilizado nas análises. Foram excluídas crianças para as quais não foi possível encontrar pelo menos um irmão ou grupos nos quais pelo menos um irmão apresentou peso ao nascer inferior a 500g ou pelo menos um irmão era natimorto. Com essas exclusões, permaneceram no banco de dados apenas duas gestações de quadrigêmeos, portanto, optamos por excluí-las das análises.

O desfecho avaliado foi o óbito neonatal (óbito ocorrido no período de 0 a 27 dias após o nascimento) e foram avaliadas como possíveis fatores de risco variáveis sócio-demográficas maternas (idade em anos e escolaridade - inferior e superior ou igual a 8 anos de estudo), de assistência pré e perinatais (duração da gestação - inferior e igual ou superior a 37 semanas, tipo de parto - vaginal ou cesariana , número de consultas pré-natal - inferior ou igual a 6 consultas e superior a 6 consultas e tipo de hospital - público, privado ou misto) e informações individuais e coletivas dos recém-nascidos (peso ao nascer da criança (em gramas) e peso total dos irmãos (em gramas), sexo por grupo - todos do sexo masculino/ todos do sexo feminino/ pelo menos um do sexo masculino e um do feminino, e Índice de Apgar no 5º minuto).

A qualidade dos dados registrados no SIM tem melhorado nos últimos anos, e sua cobertura tem sido bem próxima de 100% nas regiões Centro-Oeste, Sudeste e Sul do país[13]. Entretanto, ainda persistem problemas como dados faltantes e sub-registros. Mello Jorge e colaboradores[14] sugerem avaliar o percentual de dados faltantes como uma forma de verificar a qualidade dos dados. No presente estudo, todas as variáveis apresentaram esse percentual inferior a 1%. Para solucionar o problema de dados faltantes foi realizada imputação destes dados. Devido ao número reduzido de dados faltantes, o ganho com imputação múltipla seria muito pequeno, por isso, optamos por utilizar a imputação simples.

Formulação das Equações de Estimação Generalizadas

GEE é uma extensão dos GLM, que incorpora uma estrutura de dependência entre indivíduos de um mesmo grupo. Além disso, do mesmo modo que os GLM, permite a utilização de variáveis dependentes pertencentes à família exponencial que não sejam normalmente distribuídas (por exemplo, Poisson, Gama, Binomial Negativa). Ou seja, ela pode ser utilizada para modelar desfechos dicotômicos, de contagens ou intervalares.

Para definição do GEE, considere n gestações múltiplas, com m crianças por gestação, sendo que o valor de m pode variar de gestação para gestação. Definimos y_{ij} como a variável resposta de interesse para a j-ésima criança da i-ésima gestação e X_{ij} é um vetor $p \times 1$ de covariáveis para a j-ésima criança da i-ésima gestação, $i = 1,...,n$ e $j = 1,...,m_i$. Define-se, para a i-ésima gestação, o vetor $m_i \times 1$ de respostas, $\mathrm{y_i} = (y_{i1},...,y_{im_i})'$ e a matriz de covariáveis $m_i \times p$, $\mathrm{X_i} = (X_{i1},...,X_{im_i})'$.

Devido à ausência de distribuições multivariadas conhecidas, quando saímos do contexto de distribuições normais, utilizamos a quasi-verossimilhança para estimação dos parâmetros. Na quasi-verossimilhança, ao invés de especificar a distribuição do desfecho, é necessário apenas especificar a relação entre a média do desfecho e as covariáveis e a média do desfecho e sua variância. Portanto, para se escrever as equações de estimação generalizadas supõe-se que:

1 - A relação entre a média da variável resposta, $\boldsymbol{\mu_i}$, e as variáveis explicativas $\boldsymbol{X_i}$, pode ser expressa sob forma linear através de uma função de ligação conhecida, g, ou seja,

$$g(\mu_i) = X_i'\beta, \tag{1}$$

onde $\boldsymbol{\beta}$ é o vetor de p parâmetros.

2 - A variância da variável resposta pode ser expressa por uma função conhecida, f, da média desta variável, ou seja,

$$\mathrm{V_i} = f(\boldsymbol{\mu}_i)/\phi, \tag{2}$$

onde ϕ é o parâmetro de dispersão definido como na família exponencial. O quadro 1 apresenta características de algumas distribuições da família exponencial. Mais informações sobre a família exponencial podem ser encontradas em [3].

Liang e Zeger[6] definem a estimativa de $\boldsymbol{\beta}$ como sendo a solução do sistema de equações diferenciais quasi-escore dado a seguir:

$$\boldsymbol{U}_k(\beta) = \sum_{i=1}^{n} D_i V_i^{-1} S_i = 0 \qquad k = 1,..., p, \qquad (3)$$

onde, $D_i = \partial\mu_i / \partial\beta_k$ e $S_i = (y_i - \mu_i)$.

Para utilizar essas equações para dados correlacionados, Liang e Zeger [7] especificaram uma matriz de correlação de trabalho incorporada no termo de variância da equação (2). Considerando que $R_i(\alpha)$ seja tal matriz, com dimensão $m_i \times m_i$ para cada $\boldsymbol{y}_i$ onde α é um vetor que caracteriza completamente $R_i(\alpha)$ a equação (2) torna-se uma matriz de covariância para a i-ésima gestação:

$$V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2} / \phi, \qquad (4)$$

onde A_i é uma matriz diagonal $m_i \times m_i$ com $f(\mu_i)$ como elementos da diagonal e ϕ é o parâmetro de escala para distribuições da família exponencial. Note que o número de observações e a matriz de correlação podem diferir de grupo para grupo. Porém, é possível assumir que $R_i(\alpha)$ é completamente especificado pelo vetor de parâmetros desconhecidos α, que é o mesmo para todos os grupos [6]. Assim, será utilizado $R(\alpha)$ para denotar qualquer matriz de correlação de trabalho.

Quando $m_i = 1$, ou no caso de haver independência, o estimador dos parâmetros do GEE equivale ao do GLM.

A matriz de correlação de trabalho representa a correlação entre os indivíduos de uma mesma gestação para a variável desfecho ajustada pelas covariáveis presentes no modelo.

Desse modo, os valores que α pode assumir estão no intervalo [-1; +1]. A dimensão dessa matriz é determinada pelo número de indivíduos provenientes de uma mesma gestação. É possível especificar diferentes estruturas para essa matriz. Estas estruturas, bem como sua definição são apresentadas no quadro 2.

Usualmente, a escolha da melhor estrutura de correlação é baseada na natureza dos dados e na teoria, visto que há estruturas mais adequadas para situações específicas (quadro 2). Especificar esta matriz de forma correta aumenta a eficiência das estimativas dos parâmetros do modelo[15]. Liang e Zeger[7] afirmam que o modelo é robusto a erros na especificação na estrutura de correlação porque as estimativas dos parâmetros de regressão permanecem consistentes e ressaltam que a eficiência ganha pela especificação exata da estrutura de correlação é geralmente pequena. Entretanto, Fitzmaurice[15] adverte que é possível obter estimadores ineficientes quando a matriz de correlação de trabalho especificada não incorpora toda a informação sobre a correlação entre as medidas de um mesmo cluster.

Quando há dúvida quanto a qual estrutura de correlação utilizar, é possível recorrer ao critério proposto por Pan[16], o critério de quasi-verossimilhança sob o modelo de independência (*Quasi-likelihood under the Independence model Criterion* - QIC). O QIC é calculado a partir da comparação de um modelo com uma determinada estrutura de correlação de trabalho com aquele gerado utilizando a estrutura independente. Os valores obtidos de QIC podem ser utilizados para comparar as diferentes estruturas de correlação. Algumas vezes ocorre de os valores de QIC não serem necessariamente muito diferentes, tornando difícil a escolha através deste critério. Para solucionar este problema, Hin e Wang[17] propuseram o Critério de Informação de Correlação (*Correlation information criterion* - CIC), com o objetivo de aperfeiçoar o desempenho do QIC na escolha da estrutura de correlação de trabalho.

GEEs estimam coeficientes de regressão e erros padrões com distribuições amostrais assintoticamente normais[7]. Podem ser utilizados para testar efeitos principais e interações, e podem ser usados para avaliar variáveis independentes qualitativas ou quantitativas.

O $\boldsymbol{\beta}$ é estimado através de um processo iterativo, no qual alterna-se entre estimar $\boldsymbol{\beta}$ para valores fixos de $\hat{\phi}$ e $\hat{\alpha}$ e estimar (ϕ, α) para valores fixos de $\hat{\boldsymbol{\beta}}$ até se obter uma convergência nos valores estimados. Quanto à variância do estimador de $\boldsymbol{\beta}$, na maioria das vezes deve-se escolher um método robusto para estimá-la, com exceção de situações em que o tamanho da amostra é pequeno, visto que estes estimadores têm propriedades assintóticas, ou seja, sua qualidade depende de grandes amostras.

Ananth *et al*[18] mostram que os coeficientes estimados através de GEE podem ser interpretados do mesmo modo que os coeficientes estimados em um estudo transversal através de um GLM.

Para estimar o risco relativo de óbito neonatal foi utilizado GEE com a função de ligação log e a distribuição de Poisson com variância robusta incorporando a estrutura de correlação entre observações. A escolha da matriz de correlação foi baseada na natureza dos dados. A ordem de nascimento não foi considerada biologicamente importante, porque 70% dos nascimentos foram por cesariana. Assim, foi escolhida a estrutura de correlação permutável. Os grupos foram considerados independentes. Para avaliar associações entre as variáveis preditoras e o desfecho foi utilizado o teste de Wald modificado proposto por Rotnitzky e Jewell[19]. A adequação do modelo foi avaliada através da análise de resíduos.

As análises estatísticas foram realizadas no SPSS (*Statistical Package for Social Sciences*) versão 16.0[20].

Esta pesquisa foi aprovada pelo Comitê de Ética em Pesquisa da Secretaria Municipal de Saúde (nº do projeto: 001.046108.08.4).

Resultados

Foram avaliados 2.754 pares de gêmeos e 71 grupos de trigêmeos. A taxa de mortalidade infantil no primeiro grupo foi de 39,6‰ (218/5508) e no segundo de 51,6‰ (11/213). Os gêmeos apresentaram taxa de mortalidade neonatal de 29,4‰ e de mortalidade pós-neonatal de 10,2‰. Para os trigêmeos estas taxas foram 37,5‰ e 14,1‰. A distribuição dos óbitos neonatais para gemelares de acordo com características maternas e do recém nascido, de assistência pré e perinatais é apresentada na tabela 1. Houve maior prevalência de óbitos neonatais entre indivíduos cujas mães tinham menos de 18 anos, escolaridade inferior a oito anos de estudo, haviam realizado no máximo seis consultas pré-natal e cujo parto foi normal, pré-termo e realizado em hospital público. Grupos nos quais todos os irmãos eram do sexo masculino também apresentaram maior prevalência de óbitos neonatais. Os RNs que vieram a óbito apresentavam em média peso ao nascer e Apgar no 5º minuto inferior aos RNs que sobreviveram ao período neonatal.

A tabela 2 apresenta os resultados para a estimativa obtida através do GEE para o risco relativo ajustado e não ajustado para óbitos neonatais. Na análise não ajustada, todas as variáveis, exceto a idade materna, apresentaram associação estatisticamente significativa com o óbito neonatal. Após o ajuste, permaneceram significativos apenas o peso ao nascimento, o peso total do grupo e o Índice de Apgar no 5º minuto. Para estas três variáveis verificou-se que quanto maior seu valor, maior a proteção contra o óbito neonatal. Através do RR estimado, verifica-se que um aumento de 100g no peso do RN oferece uma proteção de 14% (IC-95%: 5%; 21%) para o óbito neonatal.

Para o modelo ajustado foram calculados os resíduos de Pearson. A figura 1 apresenta os resíduos de Pearson *versus* o número do RN. Na figura 1a não se observa padrão distinto dos resíduos conforme o número do RN, com apenas 0,9% dos resíduos superiores a 2,0. Entretanto, estratificando esta análise por óbito neonatal (figura 1b), verifica-se que entre os

óbitos, 30% dos resíduos estão acima de 2,0, enquanto entre os não óbitos não há nenhum resíduo superior a este valor. Assim como na análise de resíduos, o teste de aleatoriedade de resíduos não foi significativo na análise geral (P = 0,587), mas mostrou comportamento diferenciado entre resíduos relacionados a óbitos e não óbitos (P < 0,001).

Discussão

Neste artigo foi apresentada a metodologia do GEE, através de uma aplicação na análise de dados de mortalidade neonatal em gemelares. Os resultados encontrados na análise da mortalidade neonatal são semelhantes aos encontrados em outros estudos que evidenciam fatores perinatais, como peso ao nascer e índice de Apgar, influenciando fortemente na mortalidade neonatal [21].

A metodologia utilizada neste estudo é de interesse para análise de dados correlacionados, por possuir as mesmas propriedades de um GLM, incorporando ainda no modelo uma estrutura para ajuste da correlação existente entre as observações. Além disso, por ser semelhante em sua forma a um GLM, seus resultados podem ser interpretados da mesma maneira[18]. Atualmente, a metodologia GEE já está implementada nos principais programas para análise estatística, como SPSS, SAS, R e STATA[22]. Entretanto, ressaltamos que é possível existirem pequenas variações entre os resultados apresentados por estes programas, já que possuem diferentes processos iterativos[23].

Foi realizada uma comparação dos resultados obtidos no GEE com os obtidos através de regressão de Poisson com variância robusta (resultados disponíveis em http://www.mat.ufrgs.br/~camey/GEE_GLM). Foram encontradas pequenas diferenças nas estimativas pontuais dos parâmetros do modelo. Contudo, ao comparar os erros padrões, as diferenças foram maiores, interferindo na significância de uma variável (tipo de hospital), como sugerido por outros autores[1, 6, 16, 19]. Maiores diferenças entre os modelos não foram encontradas, provavelmente porque o tamanho da amostra utilizado era grande. Mais estudos

precisam ser realizados para avaliar o impacto do GEE em amostras menores. Carlin *et al* [5] ressaltam que deve-se utilizar o GEE por ser mais eficiente do que o modelo tradicional na estimação dos efeitos das covariáveis com valores diferentes dentro de um agrupamento.

Os pontos fortes de estudo são: boa qualidade dos dados secundários, a completude e boa definição das variáveis. Por utilizar dados secundários, este estudo apresenta algumas limitações. Não foi possível relacionar com o respectivo registro de nascimento 0,5% dos óbitos ocorridos no período estudado. A identificação dos gemelares em alguns casos estava incorreta, tendo sido incluídos nas análises 31 gemelares do SIM identificados como não gemelares no SINASC. Além disso, 2,6% (8,3% óbitos e 2,4% não óbitos) dos indivíduos identificados como gemelares foram excluídos das análises por não terem sido pareados.

Foram observados valores maiores de resíduos associados ao óbito neonatal. É provável que isto tenha ocorrido devido à falta de variáveis que expliquem melhor a ocorrência de óbito. Por exemplo, Anath *et al*[19] verificaram que a ocorrência de malformações congênitas e complicações obstétricas aumentam a ocorrência de mortalidade perinatal. Estas variáveis não estão disponíveis no SINASC em todo período estudado e, portanto, não puderam ser consideradas nas análises. Além disso, grande parte das variáveis era apresentada com categorias já definidas no SINASC e SIM, impossibilitando a exploração de sua forma quantitativa ou com outras categorizações.

Em conclusão, a utilização do GEE para essa análise apresentou resultados consistentes e diferentes do GLM, demonstrando a necessidade de sua aplicação quando analisa-se dados correlacionados. Desse modo, recomenda-se o seu uso sempre que houver agrupamento de indivíduos, já que este modelo considera a correlação entre os sujeitos do mesmo grupo e está implementado nos programas estatísticos.

Referências

1. Carlin JB, Gurrin LC, Sterne JA, Morley R, Dwyer T. Regression models for twin studies: a critical review. Int J Epidemiol. 2005 Oct;34(5):1089-99.

2. Fitzmaurice G. Clustered data. Nutrition. 2001 Jun;17(6):487-8.

3. Ananth CV, Preisser JS. Bivariate logistic regression: modelling the association of small for gestational age births in twin gestations. Stat Med. 1999 Aug;18(15):2011-23.

4. McCullagh P, Nelder JA. Generalized Linear Models. London: Chapman and Hall 1983.

5. Nelder JA, Wedderburn RWM. Generalized linear models. J R Stat Soc (Series A). 1972;135(3):370-84.

6. Zeger S, Liang K. Longitudinal data analysis for discrete and continuous outcomes. Biometrics. 1986 Mar;42(1):121-30.

7. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986 Abr;73(1):13-22.

8. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics. 1982;Dec 38(4):963-74.

9. Lipsitz SR, Laird NM, Harrington DP. Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. Biometrika. 1991;Mar 78(1):153-60.

10. Vonesh EF, Carter RL. Mixed effect nonlinear regression for unbalanced repeated measures. Biometrics. 1992;Mar 48 (1):1-17.

11. Twisk JW. Longitudinal data analysis. A comparison between generalized estimating equations and random coefficient analysis. Eur J Epidemiol. 2004 Ago;19(8):769-76.

12. Disponível em http://www.saude.sc.gov.br/download/LinkPlus/index.htm [acessado em Junho/2009].

13. Brasil. Ministério da Saúde. Secretaria Executiva. Datasus. Indicadores e Dados Básicos: Brasil 2005 - IDB 2005. Disponível em http://tabnet.datasus.gov.br/cgi/idb2007/matriz.htm#cober. [Acessado em 21 de junho de 2009].

14. Mello Jorge MHP, Gotlieb SLD, Oliveira H. O Sistema de Informação sobre Nascidos Vivos: primeira avaliação dos dados brasileiros. IESUS. 1996;5(2):15-48.

15. Fitzmaurice GM. A caveat concerning independence estimating equations with multivariate binary data. Biometrics. 1995 Mar;51(1):309-17.

16. Pan W. Akaike's information criterion in generalized estimating equations. Biometrics. 2001;Mar 57(1):120-5.

17. Hin L-Y, Wang Y-G. Working-correlation-structure identification in generalized estimating equations. Stat Med. 2009 Feb;28(4):642–58.

18. Ananth CV, Platt RW, Savitz DA. Regression models for clustered binary responses: implications of ignoring the intracluster correlation in an analysis of perinatal mortality in twin gestations. Ann Epidemiol. 2005 Apr;15(4):293-301.

19. Rotnitzky A, Jewell NP. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. Biometrika. 1990; Sep 77(3):485-97.

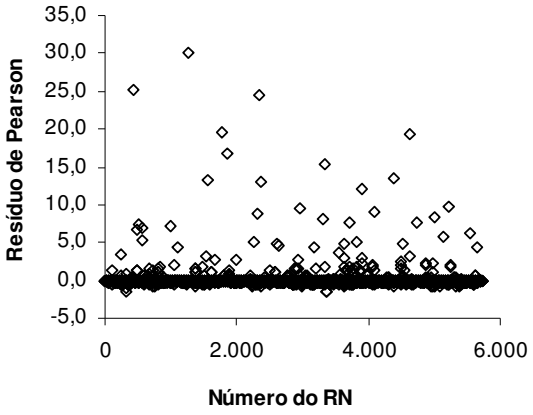
20. SPSS Inc 1989-2007. Chicago, Illinois.

21. Machado CJ, Hill K. Determinantes da mortalidade neonatal e pós-neonatal no município de São Paulo. Rev Bras Epilemiologia. 2003;6(4):345-58.

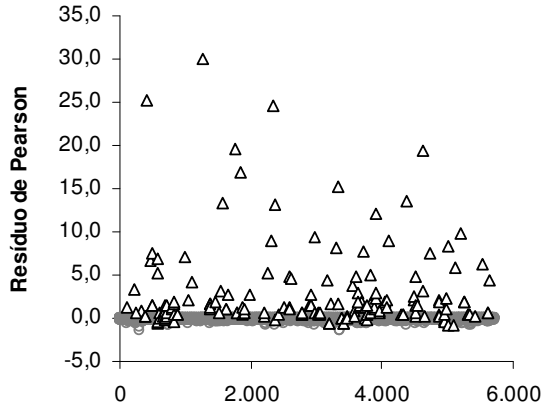
22. Horton NJ, Lipsitz SR. Review of software to fit Generalized Estimating Equation regression models. Am Stat. 1999 May;53(2):160-9.

23. Twisk JWR. Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide. Cambridge University Press ed 2003.

Figura



(a)



(b)

Figura 1: Resíduos de Pearson *versus* número do RN (a). Resíduos de Pearson *versus* número do RN segundo óbito (Δ) e não óbito (o) (b).

Quadros

Quadro 1: Características de algumas distribuições da família exponencial.

Modelo	Forma na família exponencial	Ligação canônica	ϕ
Normal: $N(\mu,\sigma^2)$	$exp\left\{\frac{1}{\sigma^2}\left(y\mu-\frac{\mu^2}{2}\right)-\frac{1}{2}log\left(2\pi\sigma^2\right)-\frac{y}{2\sigma^2}\right\}$	Identidade: $\eta=\mu$	σ^2
Binomial: $\frac{B(m,\mu)}{m}$	$\binom{m}{x}exp\left\{x\,log\left[\frac{\mu}{1-\mu}\right]+m\,log(1-\mu)\right\}$	logit: $\eta=log\left[\frac{\mu}{1-\mu}\right]$	m
Poisson: $P(\mu)$	$\frac{1}{x!}exp(x\,log\,\mu-\mu)$	log: $\eta=log\,\mu$	1

Quadro 2: Definição e exemplo para cada tipo de estrutura de correlação de trabalho.

Estrutura	**Definição**	**Exemplo** $(m=3)$
Independente	Utilizada no caso de independência entre as observações.	$R(\alpha)=\begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}$
Permutável	Considera-se que a correlação entre as observações dos indivíduos de um mesmo grupo é a mesma.	$R(\alpha)=\begin{pmatrix}1 & \alpha & \alpha\\ \alpha & 1 & \alpha\\ \alpha & \alpha & 1\end{pmatrix}$
AR(1)	Supõe-se que as medidas dentro do grupo têm uma relação auto-regressiva de primeira ordem, usualmente utilizada quando os dados estão correlacionados ao longo do tempo.	$R(\alpha)=\begin{pmatrix}1 & \alpha & \alpha^2\\ \alpha & 1 & \alpha\\ \alpha^2 & \alpha & 1\end{pmatrix}$
M-dependente	Assume-se que as correlações a t medidas de distância são iguais, que as correlações a $t+1$ medidas de distância são iguais, e assim por diante de $t=1,...,m$. Assume também que medidas muito distantes $(>m)$ não são correlacionadas.	$R(\alpha)=\begin{pmatrix}1 & \alpha & 0\\ \alpha & 1 & \alpha\\ 0 & \alpha & 1\end{pmatrix}$ $M=1$
Não estruturada	Assume-se que entre cada observação dentro do grupo há um valor de correlação diferente.	$R(\alpha)=\begin{pmatrix}1 & \alpha_1 & \alpha_2\\ \alpha_1 & 1 & \alpha_3\\ \alpha_2 & \alpha_3 & 1\end{pmatrix}$

Tabelas

Tabela 1. Distribuição dos óbitos neonatais em gemelares de acordo com características maternas, do recém nascido e de assistência pré e perinatais, Porto Alegre, 1995-2007.

	Óbito neonatal		Não óbito	
	n = 170	‰	n = 5551	‰
Escolaridade materna				
0 a 7 anos	101	39,2	2474	960,8
≥ 8 anos	69	21,9	3077	978,1
Idade materna				
≤17 anos	17	59,4	269	940,6
18 a 34	121	28,7	4089	971,3
35 ou mais	32	26,1	1193	973,9
Número de consultas pré natal				
0 a 6	121	56,0	2040	944,0
Mais de 6	49	13,8	3511	986,2
Tipo de hospital				
Privado	18	12,8	1384	987,2
Misto	25	25,4	961	974,6
Público	127	38,5	3172	961,5
Tipo de parto				
Cesáreo	90	22,2	3962	977,8
Normal	80	47,9	1589	952,1
Idade gestacional (em semanas)				
≤ 36	157	52,0	2864	948,0
37 ou mais	13	4,8	2687	995,2
Sexo				
Todos do sexo feminino	45	22,4	1960	977,6
Todos do sexo masculino	76	39,0	1873	961,0
Pelo menos 2 diferentes	49	27,7	1718	972,3
	média	DP	Média	DP
Índice de Apgar no 5º minuto	5,9	2,70	9,0	0,95
Peso do indivíduo	1064,7	589,68	2299,7	537,81
Peso do grupo	2224,4	1151,23	4660,6	1006,52

Tabela 2: Risco relativo (RR) bruto e ajustado estimado através de GEE para óbito neonatal em gemelares, Porto Alegre, 1995-2007.

	RR bruto	IC-95%		P	RR ajustado	IC-95%		P
Escolaridade materna								
0 a 7 anos	1,77	(1,21;	2,57)	0,003	1,23	(0,86;	1,77)	0,265
≥ 8 anos	1,00	-			1,00	-		
Idade materna (em anos)	0,97	(0,94;	1,00)	0,06	1,02	(0,99;	1,04)	0,159
Número de consultas pré natal								
0 a 6	4,05	(2,75;	5,97)	<0,001	0,99	(0,67;	1,45)	0,949
Mais de 6	1,00	-			1,00	-		
Tipo de hospital								
Público	2,95	(1,66;	5,23)	<0,001	1,69	(0,98;	2,93)	0,059
Misto	1,93	(0,96;	3,89)	0,067	1,55	(0,88;	2,74)	0,132
Privado	1,00	-			1,00	-		
Tipo de parto								
Normal	2,16	(1,51;	3,08)	<0,001	1,05	(0,78;	1,42)	0,763
Cesáreo	1,00	-			1,00	-		
Idade gestacional (em semanas)								
≤ 36	10,44	(6,02;	18,12)	<0,001	0,86	(0,44;	1,69)	0,665
37 ou mais	1,00	-			1,00	-		
Sexo								
Todos do sexo feminino	1,00	-			1,00	-		
Todos do sexo masculino	1,77	(1,13;	2,79)	0,013	1,32	(0,93;	1,86)	0,118
Pelo menos 2 diferentes	1,26	(0,77;	2,05)	0,363	1,22	(0,86;	1,74)	0,267
Índice de Apgar no 5º minuto	0,62	(0,59;	0,66)	<0,001	0,86	(0,80;	0,93)	<0,001
Peso indivíduo* (em 100g)	0,74	(0,71;	0,76)	<0,001	0,86	(0,79;	0,95)	0,002
Peso do grupo* (em 100g)	0,86	(0,85;	0,87)	<0,001	0,94	(0,90;	0,99)	0,01

RR: risco relativo; IC: Intervalo de Confiança; *RR para o aumento de 100g no peso.

Apêndice

A seguir são apresentados os comandos utilizados no SPSS, R, STATA e SAS. Mais informações sobre a utilização do GEE nestes programas pode ser obtida em[22]. Para obter os resultados da tabela 2 foi utilizado o SPSS versão 16.0.

* Variáveis utilizadas:
neonatal: óbito neonatal
idade_mae: idade materna
esc_mae: escolaridade materna
dur_gest: duração da gestação
parto: tipo de parto
pre_natal: número de consultas pré-natal
hospital: tipo de hospital
peso: peso ao nascer da criança
peso_total: peso total dos irmãos
sexo: sexo
apgar5: índice de Apgar no 5º minuto
dn_par: identificador único para cada gestação
ordem: ordem de nascimento (1,2,3)

Programa no SPSS

```
* Generalized Estimating Equations.
GENLIN neonatal BY esc_mae pre_natal hospital parto dur_gest sexo
(ORDER=DESCENDING) WITH idade_mae apgar5 peso peso_total
 /MODEL esc_mae pre_natal hospital parto dur_gest sexo idade_mae apgar5 peso peso_total
INTERCEPT=YES DISTRIBUTION=POISSON LINK=LOG
/CRITERIA METHOD=FISHER(1) SCALE=1 MAXITERATIONS=100
MAXSTEPHALVING=5 PCONVERGE=1E-006(ABSOLUTE)
SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95 LIKELIHOOD=FULL
```

```
/REPEATED SUBJECT=dn_par WITHINSUBJECT=ORDEM SORT=YES
CORRTYPE=EXCHANGEABLE ADJUSTCORR=YES COVB=ROBUST
MAXITERATIONS=100 PCONVERGE=1e-006(ABSOLUTE) UPDATECORR=1
/MISSING CLASSMISSING=EXCLUDE
/PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION
(EXPONENTIATED) WORKINGCORR.
```

Programa no R

```
library(foreign)

library(gee)

gee.exch<-gee(neonatal ~ I(esc_mae) + I(pre_natal) + I(hospital) + I(parto) + I(dur_gest) +
I(sexo) + idade_mae + apgar5 + peso +  peso_total, id=dn_par, data=x, family =poisson,
corstr="exchangeable", scale.fix = TRUE, scale.value = 1)

summary(gee.exch)
```

Programa no STATA

```
xi: xtgee neonatal i.esc_mae i.pre_natal i.hospital i.parto i.dur_gest i.sexo idade_mae apgar5
peso peso_total, fam(poisson) i(dn_par) robust corr(exch)
```

Programa no SAS (adaptado de Horton e Lipsitz [22])

```
proc genmod data = gee;
        class dn_par;
        model  neonatal =  esc_mae pre_natal hospital parto dur_gest sexo idade_mae apgar5
                peso peso_total / dist = poisson;
         repeated subject = dn_par / type = exch corrw within=setting;
          make ‘classlevels’ noprint;
          make ‘geercov’ out=rcov noprint;
run;
```

7. CONCLUSÕES E CONSIDERAÇÕES FINAIS

Dados correlacionados podem ocorrer em diversas situações. Exemplos incluem estudos longitudinais, nos quais os indivíduos possuem observações medidas repetidamente ao longo do tempo, estudos com observações de vários membros da mesma família, e os estudos com mais de um resultado para cada pessoa, tais como estudos oftalmológicos nos quais os dois olhos são medidos. Em todas essas situações, a análise correta requer que se considere a correlação existente entre as observações.

Nesta dissertação foi apresentada a metodologia GEE, que inclui uma estrutura de correlação de trabalho entre as observações para a obtenção de estimativas consistentes e não viciadas. Em sua utilização na análise de dados de gemelares, esta metodologia mostrou-se adequada visto que foi possível identificar diferenças entre o modelo estimado através dela e o estimado do modo tradicional (via GLM).

Liang e Zeger (1986) e Lipsitz *et al*. (1994) apresentam as vantagens da utilização do GEE para dados correlacionados. O GEE oferece uma razoável eficiência estatística e, quando μ_i está corretamente especificada, as estimativas dos parâmetros, $\hat{\beta}$, são também consistentes (Liang e Zeger, 1986). O GEE permite a utilização de uma grande variedade de estruturas de correlação entre os grupos. Além disso, assim como o GLM, o GEE pode ser aplicado a variáveis resposta com diferentes distribuições, além da gaussiana, como a Poisson e a binomial. Diferentemente de um modelo como a ANOVA, o GEE utiliza toda informação disponível de cada indivíduo, permitindo ainda a presença tanto de variáveis que podem apresentar valores diferentes entre os membros do mesmo grupo, como de variáveis constantes dentro do grupo. Além disso, por ser semelhante em sua forma a um GLM, seus resultados podem ser interpretados da mesma maneira [87].

Atualmente, a metodologia GEE já está implementada nos principais programas para análise estatística, como SPSS, SAS, R e STATA [33]. Entretanto, ressaltamos que é possível existirem pequenas variações entre os resultados apresentados por estes programas, já que possuem diferentes processos iterativos [89].

Desse modo, salientamos a importância da utilização do GEE sempre que houver agrupamento de indivíduos, já que este modelo considera a correlação entre os sujeitos do mesmo grupo é de simples interpretação e está implementado nos principais programas estatísticos.

8. ANEXOS

ANEXO A: PROJETO DE PESQUISA

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Faculdade de Medicina
Programa de Pós-Graduação em Epidemiologia

Utilização da Metodologia de Equações de Estimação Generalizadas para dados correlacionados: uma aplicação a estudo em gemelares

Marilyn Agranonik
Orientadora: Prof[a] Dra. Suzi Alves Camey

Porto Alegre, 25 de novembro de 2008.

Introdução

Para definir a análise estatística a ser aplicada em um conjunto de dados é importante conhecer a relação entre os sujeitos e entre as observações. Em muitas situações, na área da saúde, os sujeitos estudados são independentes. Entretanto, é coletada informação sobre a mesma variável repetidas vezes ao longo do tempo, tornando as observações correlacionadas. A situação inversa também ocorre, ou seja, os sujeitos dividem características em comum e, portanto, não podem ser considerados independentes. Neste caso pode haver correlação entre os sujeitos.

O primeiro caso é conhecido como medidas repetidas e o segundo como dados agrupados (clustered data). A correlação nesses casos pode ocorrer já que as observações feitas em um mesmo indivíduo (estudos longitudinais) ou em pessoas de um mesmo grupo (dados agrupados) tendem a ser mais semelhantes do que observações de indivíduos diferentes ou de grupos diferentes[7].

Frequentemente se deseja estudar o comportamento de uma variável resposta em relação a variáveis independentes. Para esses casos, técnicas de modelagem são utilizadas, nas quais se incluem os modelos de regressão. Os modelos tradicionais de regressão têm uso limitado em estudos longitudinais ou de dados agrupados devido à suposição de independência entre os sujeitos. Este é o caso dos Modelos Lineares Generalizados (GLM)[14,15], Apesar deste ser um método poderoso e flexível, se for utilizado para dados correlacionados, é provável a obtenção de distorções nas estimativas dos parâmetros e de seus erros padrões, levando a inferências estatísticas incorretas[5,13,17].

Existem pelo menos duas abordagens estatísticas para esse tipo de problema: as Equações de Estimação Generalizadas e os modelos multiníveis (um caso especial de modelos

mistos ou de modelos hierárquicos). Como no conjunto de dados utilizados neste trabalho há um único nível de agrupamento optamos por aplicar a primeira metodologia.

Zeger e Liang[17] e Liang e Zeger[13] propuseram uma classe de Equações de Estimação Generalizadas (*Generalized Estimating Equations - GEE*) para estimar parâmetros de regressão. Este método foi desenvolvido para produzir estimativas mais eficientes e não viciadas para os parâmetros do modelo de regressão quando se lida com dados correlacionados, pois considera a estrutura de correlação entre as observações. GEE é uma extensão dos Modelos Lineares Generalizados, o qual permite utilizar variáveis dependentes pertencentes à família exponencial que não sejam normalmente distribuídas (por exemplo, Poisson, Gama, Binomial Negativa).

Este método deve ser utilizado quando o objetivo da análise estatística é descrever a esperança da variável resposta em função de um conjunto de covariáveis, considerando a correlação entre as observações. Essas equações são extensões das utilizadas no método de quasi-verossimilhança[18]. Para definir uma verossimilhança é necessário especificar a forma de distribuição das observações, contudo, para a função de quasi-verossimilhança é necessário especificar apenas a relação entre a média e a variância das observações. A escolha do método de quasi-verossimilhança é o que permite uma distribuição não gaussiana dos dados. Assim, Liang e Zeger[13] especificaram a média da variável resposta como uma função linear das covariáveis, assumiram a variância como uma função conhecida da média e definiram a matriz de correlação de trabalho (*working correlation matrix*). Este modelo é apresentado da seguinte forma:

Considere que, para n gestações múltiplas, y_{ij} seja a variável resposta e X_{ij} um vetor $p \times 1$ de covariáveis para o j-ésimo nascimento da i-ésima gestação, $i = 1,...,n$ e $j = 1,...,m_i$. O valor de m pode variar de gestação para gestação, sendo os valores mais comuns, $m=2$ para gêmeos e $m=3$ para trigêmeos. Define-se, para i-ésima gestação, o vetor $m_i \times 1$ de respostas,

$\mathrm{y_i} = (y_{i1},...,y_{im_i})'$ e a matriz de covariáveis $m_i \times p$, $\mathrm{X_i} = (X_{i1},...,X_{im_i})'$. Para se escrever as equações de estimação generalizadas supõe-se que:

1 - A relação entre a esperança de variável resposta, $\boldsymbol{\mu_i}$, e as variáveis explicativas $\boldsymbol{X_i}$, pode ser expressa sob forma linear através de uma função de ligação conhecida, g. Esta função é dada por:

$$g(\mu_i) = X_i'\beta, \qquad (1)$$

onde $\boldsymbol{\beta}$ é o vetor de p parâmetros.

2 - A variância da variável resposta pode ser expressa por uma função conhecida da média desta variável da seguinte forma:

$$\mathrm{V_i} = f(\boldsymbol{\mu}_i)/\phi. \qquad (2)$$

Seguindo uma notação semelhante a Liang e Zeger (1986), a estimativa de $\boldsymbol{\beta}$ é a solução do sistema equações diferenciais quasi-escore dado a seguir:

$$\boldsymbol{U_k}(\beta) = \sum_{i=1}^{n} D_i V_i^{-1} S_i = 0 \qquad k = 1,..., p, \qquad (3)$$

onde, $D_i = \partial\mu_i/\partial\beta_k$ e $S_i = (y_i - \mu_i)$.

Para utilizar essas equações para dados correlacionados Liang e Zeger (1986) especificaram uma matriz de correlação de trabalho incorporada no termo de variância da equação (2). Considerando que $R_i(\alpha)$ seja a matriz de correlação $m_i \times m_i$ para cada $\boldsymbol{y}_i$ onde α é um vetor que completamente caracteriza $R_i(\alpha)$ a Equação (2) torna-se uma matriz de covariância para o i-ésimo grupo:

$$V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2} / \phi \qquad (4)$$

onde A_i é uma matriz diagonal $m_i \times m_i$ com $f(\mu_i)$ como elementos da diagonal e ϕ é o parâmetro de escala para distribuições da família exponencial. Note que o número de observações e a matriz de correlação podem diferir de grupo para grupo. Porém, é possível

assumir que $R_i(\alpha)$ é completamente especificado pelo vetor de parâmetros desconhecidos α, que é o mesmo para todos os grupos.

Quando $m_i = 1$, ou no caso de haver independência, o estimador do GEE equivale ao do GLM.

A matriz de correlação de trabalho representa a correlação entre as medidas feitas em um mesmo sujeito ou em sujeitos de um mesmo grupo. Seu tamanho é determinado pelo número de medidas feitas no sujeito (ou pelo número de sujeitos no grupo). É possível especificar uma das seguintes estruturas para essa matriz:

- **Independente:** Assume-se que as observações de um mesmo indivíduo (ou de indivíduos de um mesmo grupo) não são correlacionadas.
- **Permutável:** Assume-se que a correlação entre as observações de indivíduos de um mesmo grupo é a mesma. Esta matriz de correlação é utilizada quando não é possível estabelecer uma ordem lógica entre as medidas repetidas.
- **Auto-regressiva de 1ª ordem – AR(1):** Assume-se que as medidas repetidas têm uma relação auto-regressiva de primeira ordem. Ou seja, a correlação entre dois elementos quaisquer é igual a r para elementos distantes a uma posição, r^2 para elementos distantes a duas posições, e assim por diante. Desse modo, as correlações diminuem à medida que os elementos se afastam da diagonal principal da matriz. Esta matriz é a mais indicada para estudos com medidas repetidas ao longo do tempo[4], quando se assume que as correlações tornam-se mais fracas ao longo do tempo.
- **M-dependente:** Assume-se que as correlações a t medidas de distância são iguais, que as correlações a $t+1$ medidas de distância são iguais, e assim por diante de $t = 1,...,t = m$. Assume também que medidas muito distantes $(> m)$ não são correlacionadas.

- **Não estruturada:** É o caso mais geral, onde se assume que entre cada medida repetida há um valor de correlação diferente.

A especificação da forma correta da matriz de correlação aumenta a eficiência das estimativas dos parâmetros do modelo[4], o que é particularmente importante quando a correlação entre as respostas for alta. Porém, o modelo é robusto a erros na especificação na estrutura de correlação porque estimativas dos parâmetros de regressão permanecem consistentes; portanto, a eficiência ganha pela especificação exata da estrutura de correlação é geralmente pequena[13].

GEEs estimam coeficientes de regressão e erros padrões com distribuições amostrais assintoticamente normais[13]. Podem ser utilizados para testar efeitos principais e interações, e podem ser usados para avaliar variáveis independentes categóricas ou contínuas.

Justificativa

A mortalidade infantil tem sido freqüentemente apontada como indicador sensível da qualidade de vida de uma população[2], determinada em sua dimensão mais ampla pelas condições sociais, econômicas e culturais dos indivíduos e da comunidade a que pertencem.

A taxa de nascimentos múltiplos aumentou aproximadamente 30% nos últimos anos (de 1.95% em 1994 foi para 2.53% em 2005)[8]. Estudos sugerem que a taxa de mortalidade infantil é maior para gestação múltipla (gêmeos ou trigêmeos) do que em uma gestação única[6,19]. Considerando esta tendência de aumento da taxa de gemelaridade, existe uma preocupação crescente para um aumento do risco de morte precoce para gêmeos e trigêmeos quando comparados aos nascimentos únicos. Em comparação à gestação única, o excesso de risco para mortalidade tem sido atribuído à curta duração gestacional e maior freqüência de restrição de crescimento fetal[1,10-12]. Estas inferências, derivadas de abordagens convencionais,

assumem que os resultados das gestações múltiplas são independentes[9]. No entanto, é sabido que fetos de uma mesma gestação são mais semelhantes do que os de gestações diferentes e, portanto, as suas respostas são susceptíveis a serem correlacionadas uma com a outra. Em estudos com gêmeos, a probabilidade de um resultado negativo, tal como a morte neonatal e perinatal, por um dos gêmeos foi fortemente aumentada se o co-gêmeo também apresentou esse resultado[3]. Ananth e Preisser sugerem que uma gravidez múltipla como a de gêmeos e trigêmeos seja um conglomerado natural em que as respostas dos fetos são interdependentes ou agregadas. Desse modo, torna-se necessário utilizar um modelo apropriado para dados correlacionados.

Objetivos

Comparar as diferentes estruturas de correlação do GEE a fim de escolher o modelo mais adequado para estudar fatores de risco para mortalidade infantil em crianças nascidas de gravidez múltipla (gêmeos, trigêmeos ou de ordem superior).

Identificar principais fatores de risco para mortalidade infantil em crianças nascidas de gravidez múltipla (gêmeos, trigêmeos ou de ordem superior).

Metodologia

Serão utilizadas nas análises informações de todas as crianças nascidas de gravidez múltipla (gêmeos, trigêmeos ou de ordem superior) ocorridos em Porto Alegre no período de 1995 a 2007. Essas informações serão obtidas através de dados do Sistema de Informações de Nascidos Vivos (SINASC), desenvolvido através de informações da Declaração de Nascimento (DN) e do Sistema de Informações de Mortalidade (SIM), obtido por intermédio das informações da Declaração de Óbito (DO) do município de Porto Alegre no período estudado. Os bancos SIM e SINASC serão unificados e recodificados. Um banco de dados

secundário será elaborado com objetivo de unificar a codificação das varáveis existentes no SINASC e SIM. Com utilização de um algoritmo específico serão unificados os bancos de dados do SINASC e SIM através, inicialmente, do nome e da mãe, da data de nascimento e do número da Declaração de Nascido Vivo.

Para as análises, serão utilizadas as seguintes variáveis:

1 - *Variável dependente*

- Natimortalidade (óbito fetal).
- Mortalidade.

2 - *Variáveis independentes:*

- Sócio-Demográficas:

  Idade, Escolaridade, número de filhos vivos e número de filhos mortos da mãe.

- Geográficas:

  Local e estabelecimento da ocorrência do parto e/ou do óbito.

- Assistência pré e perinatais:

  Tipo de gravidez (gêmeos, trigêmeos ou ordem superior), Duração da gestação, Tipo de parto, Número de consultas pré-natal.

- Informações do recém-nascido:

  Ordem de nascimento (a partir da data e hora da nascimento), Peso ao nascer, Sexo, Índice de Apgar.

Na escolha do melhor modelo será utilizado o critério de quasi-verossimilhança sob o modelo de independência *(Quasi Likelihood under Independence Model Criterion - QIC)* proposto por Pan[16]. O autor propôs o QIC como uma modificação do AIC para ser utilizado no GEE, na qual o valor da função de verossimilhança obtido pelo AIC é substituído por um

valor da função de quasi-verossimilhança, supondo que $R_i(\square) = I$ e o ajuste apropriado é feito para o termo de penalidade. Do mesmo modo que para o AIC, quanto menor o valor do QIC e do QICc, melhor o modelo. O QIC é usado para escolher a melhor estrutura de correlação de trabalho e o QICc é usado para escolher o melhor subconjunto de preditores.

Para o procedimento de linkagem probabilística será utilizado o programa link plus versão 9.0. As análises estatísticas serão realizadas no SPSS (Statistical Package for Social Sciences) versão 16.0 e STATA versão 9.0.

Riscos

Trata-se de estudo de risco mínimo utilizando dados secundários do Banco de dados do Sistema de Informações de Nascidos Vivos (SINASC), do período de 1995 a 2007, desenvolvido através de informações obtidas da Declaração de Nascimento (DN) e do Sistema de Informações de Mortalidade (SIM), do período de 1995 a 2007, obtido por intermédio das informações da Declaração de Óbito (DO) do município de Porto Alegre no período estudado.

Benefícios

O estudo trará esclarecimentos sobre o melhor método estatístico para os determinantes de nascimento e de morte para recém nascidos oriundos de gestações múltiplas (gêmeos, trigêmeos).

O estudo identificara fatores de risco para morte nestes recém nascidos.

Aspectos éticos

Trata-se de um estudo observacional, aonde não haverá divulgação da identidade dos participantes.

O projeto foi aprovado pelo Comitê de Ética em Pesquisa da Secretaria Municipal de Saúde, no dia 16 de setembro de 2008, mediante protocolo de nº 001.046108.08.4. Para o acesso e a utilização das informações constantes nos bancos de dados (SINASC e SIM) foi preenchido e assinado um "Termo de Compromisso para Utilização de Dados" junto à Equipe de Informação em Saúde da Coordenação Geral de Vigilância Sanitária (EIS/CGVS) da Secretaria Municipal de Saúde.

De acordo com a resolução 196/96 do Conselho Nacional de Saúde, o presente projeto não apresenta risco para seres humanos e conflito de interesses.

Cronograma

Tarefa	**2008**							**2009**												**2010**
	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J
Revisão da literatura	x	x	x	x	x	x	x	x	x	x	x	x	x							
Apresentação do anteprojeto	x	x																		
Encaminhamento do projeto ao Comitê de Ética			x	x																
Preparação do banco de dados*					x	x	x	x												
Análise dos dados*								x	x	x	x									
Redação da dissertação e artigo	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x			
Defesa preliminar																		x		
Correções																			x	
Sessão pública																				x

*A preparação e análise do banco de dados se iniciará somente após a aprovação do projeto pelo Comitê de Ética.

Orçamento

Atividades e Equipamentos	**Reais (R$)**
Material de Escritório (folhas e tinta para impressão)	350,00
Pen drive 4GB	70,00
Preparação de material para eventos (pôsteres)	150,00
Revisão da Literatura (BIREME e fotocópias)	100,00
Livro didático	210,00
Total	880,00

Os custos deste projeto serão de responsabilidade do pesquisador.

Referências bibliográficas

1 - Alexander GR, Kogan M, Martin J, Papiernik E. What are the fetal growth patterns of singletons, twins, and triplets in the United States? *Clin Obstet Gynecol* 1998; 41:115–25.

2 - Aertz DRGC. Investigação dos óbitos perinatais e infantis: seu uso no planejamento de políticas públicas de saúde. *J Pediatr* 1997; 73: 364-6.

3 - Ananth CV, Preisser JS. Bivariate logistic regression: modelling the association of small for gestational age births in twin gestations. *Stat Med* 1999; 18:2011–23.

4 - Ballinger GA. Using Generalized Estimating Equations for Longitudinal Data Analysis. *Organizational Research Methods* 2004; 7(2):127-150.

5 - Carlin JB, Gurrin LC, Sterne JAC, Morley R, Dwyer T. Regression models for twin studies: a critical review. *Int J Epidemiol* 2005; 34(5):1089-99.

6 - Ferguson WF. Perinatal mortality in multiple pregnancy. A review of perinatal deaths from 1609 multiple gestations. *Obstetrics Gynecology* 1964; 23:854.

7 - Fitzmaurice GM. Clustered data. *Nutrition* 2001; 17: 487- 488.

8 - Homrich da Silva C, Goldani MZ, Silva AAM, Agranonik M, Bettiol H, Barbieri MA, Rona R . The rise of multiple births in Brazil. *Acta Paediatrica* 2008; 96:1019-1023.

9 - Huang JS, Lu SE, Ananth CV. The clustering of neonatal deaths in triplet pregnancies: application of response conditional multivariate logistic regression modes. *Journal of Clinical Epidemiology* 2003; 56:1202–1209.

10 - Imaizumi Y. Infant mortality rates in single, twin and triplet births, and influencing factors in Japan, 1995–98. *Paediatr Perinat Epidemiol* 2001; 15:346-51.

11 - Kaufman GE, Malone FD,Harvey-Wilkes K, Chelmow D, Penzias AS, D'Alton ME. Neonatal morbidity and mortality associated with triplet pregnancy. *Obstet Gynecol* 1998; 91:342-8.

12 - Kiely JL. The epidemiology of perinatal mortality in multiple births. *Bull NY Acad Med* 1990; 66:618-37.

13 - Liang K-Y & Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; 73: 13-22.

14 - McCullagh, P & Nelder, JA. Generalized linear models (2nd ed.). *London: Chapman and Hal.* 1989.

15 - Nelder JA & Wedderburn RWM. Generalized linear models. *Journal of the Royal Statistical Society*, Series A 1972, 135:370-384.

16 - Pan W. Akaike's information criterion in generalized estimating equations. Biometrics 2001; 57:120-125.

17 - Zeger SL & Liang K-Y. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986; 42 (1): 121-130.

18 - Wedderburn RWM. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* 1974; 61: 439-447.

19 - Verduzco RD, Rosario R, Rigarro H. Hyaline membrane disease in twins: a 7-year review with a study on zygosity. *American Journal of Obstetrics Gynecology* 1976; 125:668–671.

ANEXO B: APROVAÇÃO PELO COMITÊ DA ÉTICA E PESQUISA

Prefeitura Municipal de Porto Alegre
Secretaria Municipal de Saúde
Comitê de Ética em Pesquisa

PARECER CONSUBSTANCIADO

Pesquisador (a) Responsável: Suzy Camey
Equipe executora:
Registro do CEP: 272 **Processo** Nº. 001.046108.08.4
Instituição onde será desenvolvido: Secretaria Municipal de Saúde – SINASC
Utilização: PRONTUÁRIOS
Situação: **APROVADO**

O Comitê de Ética em Pesquisa da Secretaria Municipal de Saúde de Porto Alegre analisou o processo Nº.001.046108.08.4, referente ao projeto de pesquisa: "**Utilização da metodologia de equações de estimação generalizadas para dados correlacionados: uma aplicação a estudo em gemelares**", tendo como pesquisador responsável Suzy Camey, cujo objetivo é "Geral: Verificar qual o modelo mais adequado para estudar fatores de risco para mortalidade infantil em gemelares, no período de 1995 a 2007".

Assim, o projeto preenche os requisitos fundamentais das resoluções. O Comitê de Ética em Pesquisa segue os preceitos das resoluções CNS 196/96, 251/97 e 292/99, sobre as Diretrizes e Normas Regulamentadoras de Pesquisa Envolvendo Seres Humanos, do Conselho Nacional de Saúde / Conselho Nacional de Ética em Pesquisa / Agência nacional de Vigilância Sanitária. Em conformidade com os requisitos éticos, classificamos o presente protocolo como **APROVADO**.

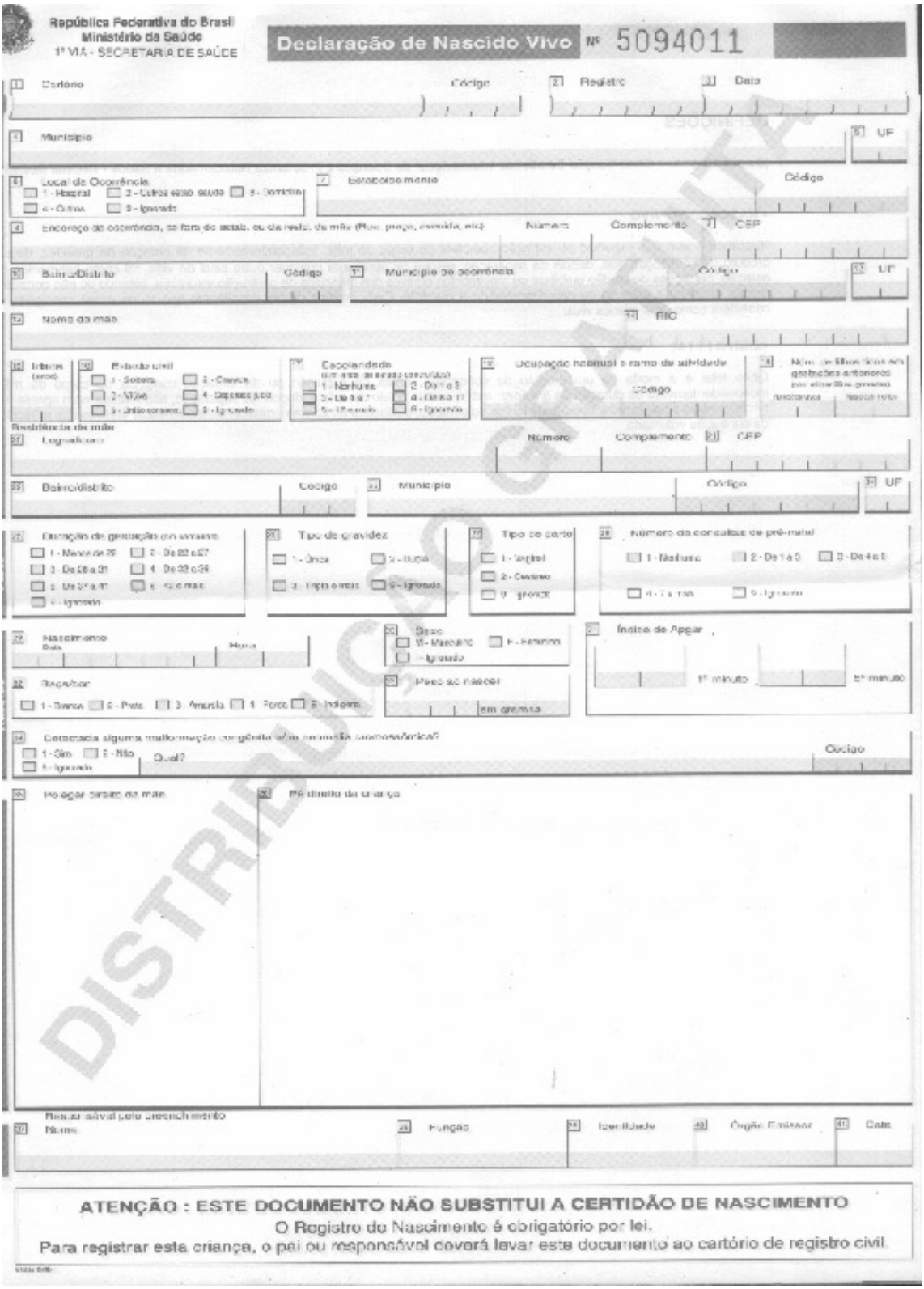
O Comitê de Ética em Pesquisa, solicita que :

1. Enviar primeiro relatório parcial em seis meses a contar desta data;
2. Informar imediatamente relatório sobre qualquer evento adverso ocorrido;
3. Comunicar qualquer alteração no projeto e no TCLE
4. Após o término desta pesquisa, o pesquisador responsável deverá apresentar os resultados junto à equipe da unidade a qual fez a coleta de dados e/ou entrevista, inclusive para o Conselho Local da Unidade de Saúde.

Porto Alegre, 16/09/08

Elen Maria Borba
Coordenadora do CEP

ANEXO C: FORMULÁRIO DA DECLARAÇÃO DE NASCIDO VIVO



República Federativa do Brasil
Ministério da Saúde
1ª VIA - SECRETARIA DE SAÚDE

Declaração de Nascido Vivo Nº 5094011

1 Cartório | Código | 2 Registro | 3 Data

4 Município | 5 UF

6 Local da Ocorrência
1 - Hospital | 2 - Outros estab. saúde | 3 - Domicílio
4 - Outros | 5 - Ignorado

7 Estabelecimento | Código

8 Endereço da ocorrência, se fora do estab. ou da resid. da mãe (Rua, praça, avenida, etc.) | Número | Complemento | 9 CEP

10 Bairro/Distrito | Código | 11 Município da ocorrência | Código | 12 UF

13 Nome da mãe | 14 NIC

15 Idade (anos) | 16 Estado civil
1 - Solteira | 2 - Casada
3 - Viúva | 4 - Separada jud.
5 - União consensual | 9 - Ignorado

17 Escolaridade (em anos de estudo concluídos)
1 - Nenhuma | 2 - De 1 a 3
3 - De 4 a 7 | 4 - De 8 a 11
5 - 12 e mais | 9 - Ignorado

18 Ocupação habitual e ramo de atividade | Código

19 Número de filhos tidos em gestações anteriores (por nascidos vivos) | Nascidos vivos | Nascidos mortos

Residência da mãe

20 Logradouro | Número | Complemento | 21 CEP

22 Bairro/distrito | Código | 23 Município | Código | 24 UF

25 Duração da gestação em semanas
1 - Menos de 22 | 2 - De 22 a 27
3 - De 28 a 31 | 4 - De 32 a 36
5 - De 37 a 41 | 6 - 42 e mais
9 - Ignorado

26 Tipo de gravidez
1 - Única | 2 - Dupla
3 - Tripla e mais | 9 - Ignorado

27 Tipo de parto
1 - Vaginal
2 - Cesáreo
9 - Ignorado

28 Número de consultas de pré-natal
1 - Nenhuma | 2 - De 1 a 3 | 3 - De 4 a 6
4 - 7 a 9 | 5 - 10 e mais | 9 - Ignorado

29 Nascimento
Data | Hora

30 Sexo
M - Masculino | F - Feminino
I - Ignorado

31 Índice de Apgar
1º minuto | 5º minuto

32 Raça/cor
1 - Branca | 2 - Preta | 3 - Amarela | 4 - Parda | 5 - Indígena

33 Peso ao nascer
em gramas

34 Constatada alguma malformação congênita ou anomalia cromossômica?
1 - Sim | 2 - Não | 9 - Ignorado
Qual? | Código

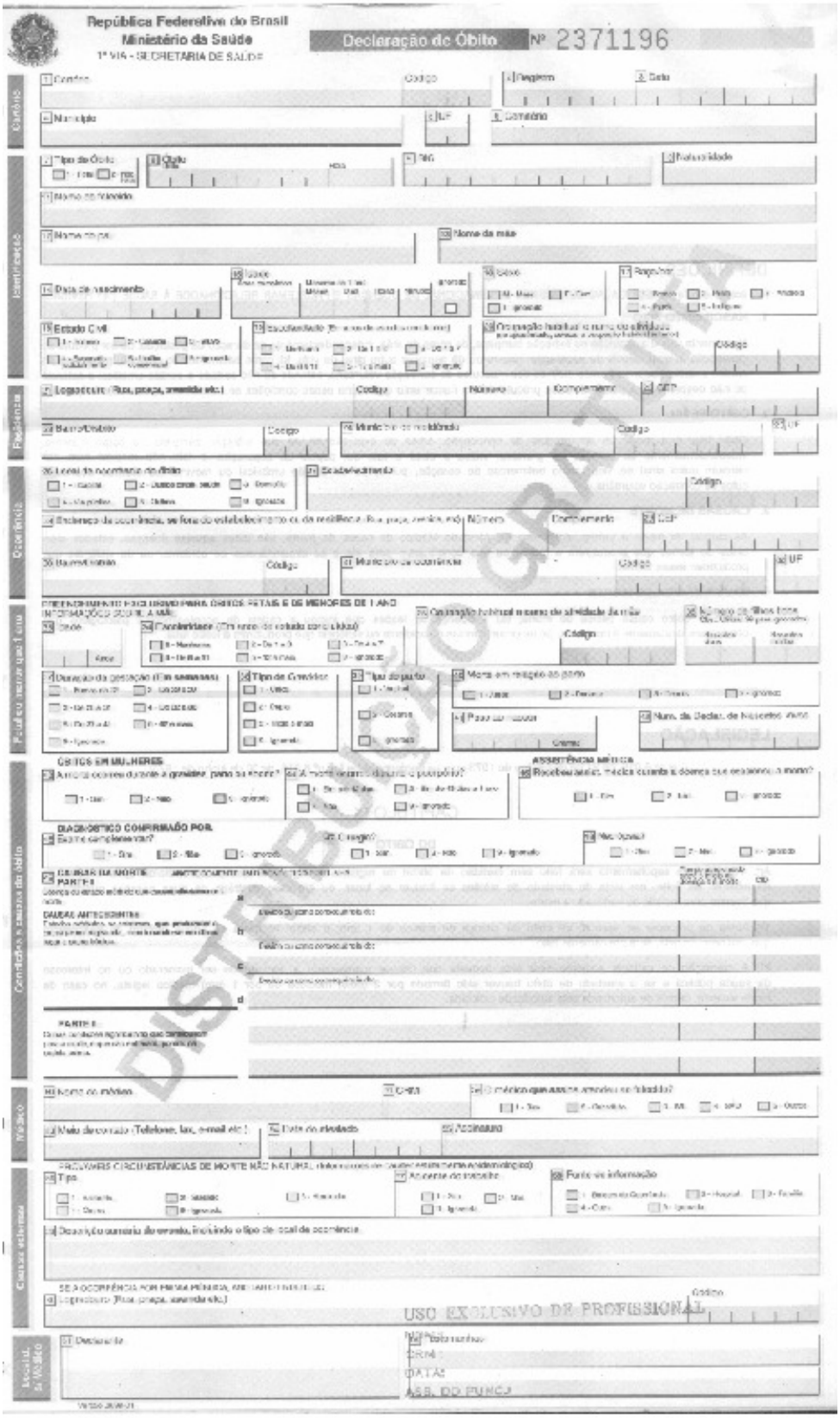
35 Nome completo da mãe | 36 Pré-natal da criança

37 Pessoa obrigada pelo preenchimento
Nome | 38 Função | 39 Identidade | 40 Órgão Emissor | 41 Data

ATENÇÃO : ESTE DOCUMENTO NÃO SUBSTITUI A CERTIDÃO DE NASCIMENTO
O Registro de Nascimento é obrigatório por lei.
Para registrar esta criança, o pai ou responsável deverá levar este documento ao cartório de registro civil.

DISTRIBUIÇÃO GRATUITA

ANEXO D: FORMULÁRIO DA DECLARAÇÃO DE ÓBITO



República Federativa do Brasil
Ministério da Saúde
1ª VIA - SECRETARIA DE SAÚDE

Declaração de Óbito Nº 2371196

Cartório
Identificação
Residência
Ocorrência
Fetal ou menor que 1 ano
Condições e causas do óbito
Médico
Causas externas
Responsável pelo preenchimento

DISTRIBUIÇÃO GRATUITA

USO EXCLUSIVO DE PROFISSIONAL

ANEXO E: COMANDOS UTILIZADOS NO SPSS, VERSÃO 16.0

A seguir são apresentados os comandos utilizados no SPSS, R, e STATA. Para obter os resultados da tabela 2 foi utilizado o SPSS versão 16.0.

* Variáveis utilizadas:

neonatal: óbito neonatal

idade_mae: idade materna

esc_mae: escolaridade materna

dur_gest: duração da gestação

parto: tipo de parto

pre_natal: número de consultas pré-natal

hospital: tipo de hospital

peso: peso ao nascer da criança

peso_total: peso total dos gemelares

sexo: sexo

apgar5: índice de Apgar no 5º minuto

dn_par: identificador único para cada gestação

ordem: ordem de nascimento (1,2,3)

* Generalized Estimating Equations.

```
GENLIN neonatal BY esc_mae pre_natal hospital parto dur_gest sexo
(ORDER=DESCENDING) WITH idade_mae apgar5 peso peso_total
 /MODEL esc_mae pre_natal hospital parto dur_gest sexo idade_mae apgar5 peso peso_total
INTERCEPT=YES DISTRIBUTION=POISSON LINK=LOG
/CRITERIA METHOD=FISHER(1) SCALE=1 MAXITERATIONS=100
MAXSTEPHALVING=5 PCONVERGE=1E-006(ABSOLUTE)
SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95 LIKELIHOOD=FULL
/REPEATED SUBJECT=dn_par WITHINSUBJECT=ORDEM SORT=YES
CORRTYPE=EXCHANGEABLE ADJUSTCORR=YES COVB=ROBUST
MAXITERATIONS=100 PCONVERGE=1e-006(ABSOLUTE) UPDATECORR=1
/MISSING CLASSMISSING=EXCLUDE
/PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION (EXPONENTIATED)
WORKINGCORR.
```

ANEXO F: CALCULO DO CIC NO R

Variáveis utilizadas.
neonatal: óbito neonatal
peso: peso ao nascer da criança
apgar5: índice de Apgar no 5º minuto
dn_par: identificador único para cada gestação
ordem: ordem de nascimento (1,2,3)

```
library(foreign)
library(gee)
x<-read.spss("E:/bd.sav", use.value.labels=TRUE, max.value.labels=Inf,
to.data.frame=TRUE)
ind<-gee(neonatal ~ apgar5 +, id=dn_par, data=x, family =poisson, corstr=" independence")
ex<-gee(neonatal ~ apgar5 + peso, id=dn_par, data=x, family =poisson,
corstr="exchangeable")
hessian.ind<-solve(ind$naive)
robust.ex<-ex$robust                    # Obtém o estimador robusto para matriz de covariância
                                        considerando a estrutura de correlação permutável
cic<-sum(diag(hessian%*%robust.ex))     # Calcula o valor do CIC para a estrutura de
                                        correlação permutável
```

ANEXO G: Comparação entre coeficientes e erros padrões estimados através de GEE e GLM.

	GEE				GLM				Δ(%) = (GEE-GLM)/GLM		
	b	EP	RR	P	b	EP	RR	P	b	EP	RR
Escolaridade materna < 7 anos	0,21	0,19	1,23	0,265	0,21	0,16	1,23	0,182	-0,5	19,3	-0,1
Idade materna	0,02	0,01	1,02	0,159	0,02	0,01	1,02	0,095	0,1	18,6	0,0
Nº de consultas pré natal < 6	-0,01	0,20	0,99	0,949	-0,01	0,19	0,99	0,950	6,6	5,7	-0,1
Hospital Público	0,53	0,28	1,70	0,059	0,53	0,26	1,71	0,041	-1,1	7,0	-0,6
Hospital Misto	0,44	0,29	1,55	0,132	0,44	0,28	1,56	0,115	-1,2	3,4	-0,5
Parto Normal	0,05	0,15	1,05	0,763	0,05	0,13	1,05	0,730	1,2	15,7	0,1
Idade gestacional < 36 semanas	-0,15	0,34	0,86	0,665	-0,15	0,35	0,86	0,674	1,0	-1,7	-0,1
Sexo: Pelo menos 2 diferentes	0,20	0,18	1,22	0,267	0,20	0,17	1,22	0,224	-0,5	9,0	-0,1
Todos do sexo masculino	0,27	0,18	1,32	0,118	0,28	0,15	1,32	0,074	-0,7	13,7	-0,2
Índice de Apgar no 5º minuto	-0,15	0,04	0,86	0,000	-0,15	0,03	0,86	0,000	0,0	16,0	0,0
Peso indivíduo* (em 100 g)	-0,15	0,05	0,86	0,002	-0,15	0,05	0,86	0,002	-1,0	-3,4	0,1
Peso do grupo* (em 100g)	-0,06	0,02	0,94	0,010	-0,06	0,02	0,94	0,013	1,3	-2,9	-0,1