Modelos Marginais para Respostas Agrupadas Correlacionadas

Antônio Arthur Silva de Lima
Romulo Barros de Freitas

Departamento de Estatística e Matemática Aplicada
Universidade Federal do Ceará

Fortaleza, 23 de setembro de 2024

Conteúdo

Introdução

- Modelagem de uma variável resposta pertencente a um grupo (*cluster*)

Introdução

- Modelagem de uma variável resposta pertencente a um grupo (*cluster*)
- Leva em consideração a correlação inerente ao grupo

Introdução

- Modelagem de uma variável resposta pertencente a um grupo (*cluster*)
- Leva em consideração a correlação inerente ao grupo
- Modelos **Multivariados** *vs* Modelos **Marginais**

Introdução

- Modelagem de uma variável resposta pertencente a um grupo (*cluster*)
- Leva em consideração a correlação inerente ao grupo
- Modelos **Multivariados** *vs* Modelos **Marginais**
- *Generalized Estimating Equations*

Introdução

- Modelagem de uma variável resposta pertencente a um grupo (*cluster*)
- Leva em consideração a correlação inerente ao grupo
- Modelos **Multivariados** *vs* Modelos **Marginais**
- *Generalized Estimating Equations*

Exemplos: estudos longitudinais, estudos clínicos, pesquisas de satisfação, dentre outros.

Modelo de Regressão Clássico

O modelo de regressão linear múltiplo com p variáveis explicativas é dado por

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

Suposições:

Modelo de Regressão Clássico

O modelo de regressão linear múltiplo com p variáveis explicativas é dado por

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

Suposições:

- $\boldsymbol{X}$ é a uma matriz de contantes e de posto completo;

Modelo de Regressão Clássico

O modelo de regressão linear múltiplo com p variáveis explicativas é dado por

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

Suposições:

- $\boldsymbol{X}$ é a uma matriz de contantes e de posto completo;
- $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{\sigma^2})$;

Modelo de Regressão Clássico

O modelo de regressão linear múltiplo com p variáveis explicativas é dado por

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

Suposições:

- $\boldsymbol{X}$ é a uma matriz de contantes e de posto completo;
- $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{\sigma^2})$;
- Homocedasticidade;

Modelo de Regressão Clássico

O modelo de regressão linear múltiplo com p variáveis explicativas é dado por

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon} \tag{1}$$

Suposições:

- $\boldsymbol{X}$ é a uma matriz de contantes e de posto completo;
- $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{\sigma^2})$;
- Homocedasticidade;
- $\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{X}\beta, \boldsymbol{\sigma^2 I}_n)$

Modelo de Regressão Clássico

O modelo de regressão linear múltiplo com p variáveis explicativas é dado por

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon} \tag{1}$$

Suposições:

- $\boldsymbol{X}$ é a uma matriz de contantes e de posto completo;
- $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{\sigma^2})$;
- Homocedasticidade;
- $\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{X}\beta, \boldsymbol{\sigma^2 I}_n)$
- $\boldsymbol{Cov}(\epsilon_i, \epsilon_j) = 0$

Modelos Lineares Generalizados

No modelo de regressão generalizado temos um **componente aleatório** representado por um conjunto de variáveis aleatórias **independentes** $Y_1, \cdots, Y_n$ provenientes de uma mesma distribuição que faz parte de da **família exponencial** com médias $\mu_1, \cdots, \mu_n$, ou seja,

$$\mathbf{E}(Y_i) = \mu_i, i = 1, \cdots, n,$$

Modelos Lineares Generalizados

No modelo de regressão generalizado temos um **componente aleatório** representado por um conjunto de variáveis aleatórias **independentes** $Y_1, \cdots, Y_n$ provenientes de uma mesma distribuição que faz parte de da **família exponencial** com médias $\mu_1, \cdots, \mu_n$, ou seja,

$$\mathbf{E}(Y_i) = \mu_i, i = 1, \cdots, n,$$

um **componente sistemático** dado por

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j = \boldsymbol{x}_i^\top \boldsymbol{\beta} \quad \text{ou} \quad \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta},$$

Modelos Lineares Generalizados

e uma **função de ligação** que liga o componente aleatório ao componente sistemático, ou seja, relaciona a média ao preditor linear

$$\eta_i = g(\mu_i)$$

Problemática: como modelar dados correlacionados que violam a suposição de independência?

Exemplos

1. Em um estudo de pesquisa ocular de Baltimore, mais de 5000 pessoas com 40 anos ou mais receberam um exame visual como parte de um estudo de prevalência baseado na população de distúrbios oculares [1].
    - Objetivo: identificar variáveis demográficas que estão associadas à perda de visão. Os dados estão disponíveis para ambos os olhos de todos os sujeitos. Um único modelo de regressão expressando a deficiência visual em termos de variáveis demográficas aborda os objetivos científicos. No entanto, os dois olhos da mesma pessoa provavelmente não são independentes, porque muitas causas de deficiência são binoculares. Essa associação deve ser considerada.

Método *GEE*

Quais os passos para se utilizar o método?

1. Definir um modelo marginal para a média da resposta, $(E[Y_i])$, e escolher $g(\mu)$

Método *GEE*

Quais os passos para se utilizar o método?

1. Definir um modelo marginal para a média da resposta, $(E[Y_i])$, e escolher $g(\mu)$
2. Supor como $Var[Y_i]$ depende de $E[Y_i]$

Método *GEE*

Quais os passos para se utilizar o método?

1. Definir um modelo marginal para a média da resposta, $(E[Y_i])$, e escolher $g(\mu)$
2. Supor como $Var[Y_i]$ depende de $E[Y_i]$
3. Propor um comportamento da correlação no grupo

Método *GEE*

Quais os passos para se utilizar o método?

1. Definir um modelo marginal para a média da resposta, $(E[Y_i])$, e escolher $g(\mu)$
2. Supor como $Var[Y_i]$ depende de $E[Y_i]$
3. Propor um comportamento da correlação no grupo
4. Estimar através do modelo marginal e da estrutura da correlação supostos

Método *GEE*

Quais os passos para se utilizar o método?

1. Definir um modelo marginal para a média da resposta, $(E[Y_i])$, e escolher $g(\mu)$
2. Supor como $Var[Y_i]$ depende de $E[Y_i]$
3. Propor um comportamento da correlação no grupo
4. Estimar através do modelo marginal e da estrutura da correlação supostos
5. Ajustar os erros padrão e obter estimativas mais robustas

- Com uma única observação para cada sujeito ($r_i = 1$), um modelo linear generalizado (McCullagh & Nelder, 1983) pode ser aplicado.

- Com uma única observação para cada sujeito ($r_i = 1$), um modelo linear generalizado (McCullagh & Nelder, 1983) pode ser aplicado.
- Com observações repetidas, a **correlação** para um determinado sujeito deve ser levada em consideração.

- Com uma única observação para cada sujeito ($r_i = 1$), um modelo linear generalizado (McCullagh & Nelder, 1983) pode ser aplicado.
- Com observações repetidas, a **correlação** para um determinado sujeito deve ser levada em consideração.
- Uma dificuldade na análise de dados longitudinais **não-Gaussianos** é a falta de uma classe rica de modelos, como o Normal Multivariado, para a distribuição conjunta de y_{it}, $t = 1, 2, \cdots, r_i$. Assim, métodos de verossimilhança não estão disponíveis, exceto nos poucos casos mencionados acima.

- Com uma única observação para cada sujeito ($r_i = 1$), um modelo linear generalizado (McCullagh & Nelder, 1983) pode ser aplicado.
- Com observações repetidas, a **correlação** para um determinado sujeito deve ser levada em consideração.
- Uma dificuldade na análise de dados longitudinais **não-Gaussianos** é a falta de uma classe rica de modelos, como o Normal Multivariado, para a distribuição conjunta de y_{it}, $t = 1, 2, \cdots, r_i$. Assim, métodos de verossimilhança não estão disponíveis, exceto nos poucos casos mencionados acima.
- Para contornar o problema é introduzindo uma estrutura de correlação na função escore, produzindo um novo sistema de esquações para estimar β.

Suposições para a estimação dos parâmetros de um modelo de regressão

Modelo Linear Clássico	**Modelo Linear Generalizado**	**GEE**
i.i.d.	*i.i.d.*	**não**
$y \sim \mathcal{N}(\mu; \sigma^2)$	$y \sim Fam.Expo.$	$y \sim V(\mu), \phi, R(\rho)$
$\mu = \boldsymbol{x}^\top \boldsymbol{\beta}$	$g(\mu) = \boldsymbol{x}^\top \boldsymbol{\beta}$	$g(\mu) = \boldsymbol{x}^\top \boldsymbol{\beta}$

Definição

Seja $\boldsymbol{Y}_i = (y_{i1}, \cdots, y_{ir_i})^\top$ o vetor resposta multivariado para a i-ésima unidade experimental, $i = 1, \cdots, n$, e assumindo que apenas a distribuição marginal de Y_{it} é conhecida e expressa da seguinte maneira:

$$f(y; \theta_{it}, \phi) = \exp\{\phi[y_{it}\theta_{it} - b(\theta_{it})] + c(y_{it}; \phi)\}, \qquad (2)$$

em que

$$\mathbb{E}(Y_{it}) = b'(\theta_{it}) = \mu_{it},$$

$$\mathbb{V}ar(Y_{it}) = \phi^{-1} V(\mu_{it}),$$

$$V(\mu_{it}) = b''(\theta_{it}) = \frac{d\mu_{it}}{d\theta_{it}},$$

$$\phi > 0.$$

Podemos então definir um modelo linear generalizado para cada instante t acrescentando uma função de ligação

$$g(\mu_{it}) = \eta_{it}, \tag{3}$$

em que $\eta_{it} = \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}$ é o preditor linear, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^{\top}$ é um vetor de parâmetros desconhecidos a serem estimados, e $\boldsymbol{x}_{it} = (x_{it1}, \cdots, x_{itp})^{\top}$ representa a matriz de delineamento para a i-ésima unidade experimental no tempo.

Função Escore e Matriz de Informação

A **função escore** e a **matriz de informação** para β, ignorando a estrutura de correlação intragrupo, ficam respectivamente como

$$\boldsymbol{U}_{\boldsymbol{\beta}} = \phi \sum_{i=1}^{n} \boldsymbol{D}_i^{\top} \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i), \tag{4}$$

$$\boldsymbol{K}_{\beta\beta} = \phi \sum_{i=1}^{n} \boldsymbol{D}_i^{\top} \boldsymbol{V}_i \boldsymbol{D}_i, \tag{5}$$

onde $\boldsymbol{D}_i = \boldsymbol{W}_i^{1/2} \boldsymbol{V}_i^{1/2} \boldsymbol{X}_i$.

Substituindo $\boldsymbol{D}_i$ na expressão da **função escore**, temos que

$$\begin{aligned}
\boldsymbol{U}_{\boldsymbol{\beta}} &= \phi \sum_{i=1}^{n} (\boldsymbol{W}_i^{1/2} \boldsymbol{V}_i^{1/2} \boldsymbol{X}_i)^\top \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) \\
&= \phi \sum_{i=1}^{n} \boldsymbol{X}_i^\top (\boldsymbol{V}_i^{1/2})^\top (\boldsymbol{W}_i^{1/2})^\top \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) \\
&= \phi \sum_{i=1}^{n} \boldsymbol{X}_i^\top \boldsymbol{V}_i^{1/2} (\boldsymbol{W}_i^{1/2})^\top \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) \\
&= \phi \sum_{i=1}^{n} \boldsymbol{X}_i^\top \boldsymbol{W}_i^{1/2} \boldsymbol{V}_i^{-1/2} (\boldsymbol{y}_i - \boldsymbol{\mu}_i).
\end{aligned}$$

O processo é análogo para a **matriz de informação**.

- $\boldsymbol{D}_i = \boldsymbol{W}_i^{1/2} \boldsymbol{V}_i^{1/2} \boldsymbol{X}_i$
- $\boldsymbol{X}_i$ é uma matriz r_i x p de linhas $\boldsymbol{x}_{it}^\top$
- $\boldsymbol{W}_i = diag\{w_{i1}, \cdots, w_{ir_i}\}$, $w_{it} = (d\mu_{it}/d\eta_{it})^2 / V_{it}$
- $\boldsymbol{V}_i = diag\{V_{i1}, \cdots, V_{ir_i}\}$
- $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{ir_i})^\top$
- $\boldsymbol{\mu}_i = (\mu_{i1}, \cdots, \mu_{ir_i})^\top$

Quando há **ligação canônica** a função escore e a matriz de informação de Fisher ficam dadas, respectivamente, por:

$$\boldsymbol{U}_\beta = \phi \sum_{i=1}^{n} \boldsymbol{X}_i^\top (\boldsymbol{y}_i - \boldsymbol{\mu}_i) \tag{6}$$

$$\boldsymbol{K}_{\beta\beta} = \phi \sum_{i=1}^{n} \boldsymbol{X}_i^\top \boldsymbol{V}_i \boldsymbol{X}_i \tag{7}$$

Exemplo: Vamos supor inicialmente que os dados não são correlacionados e que a matriz de correlação correspondente ao i-ésimo grupo é denotado por $\boldsymbol{R}_i$. Portanto, teremos $\boldsymbol{R}_i = \boldsymbol{I}_{r_i}$.

Exemplo: Vamos supor inicialmente que os dados não são correlacionados e que a matriz de correlação correspondente ao i-ésimo grupo é denotado por $\boldsymbol{R}_i$. Portanto, teremos $\boldsymbol{R}_i = \boldsymbol{I}_{r_i}$. A matriz de variância-covariância para Y_i, por definição é dada por

$$\mathbb{V}ar(Y_i) = \phi^{-1} \boldsymbol{V}_i^{1/2} \boldsymbol{R}_i \boldsymbol{V}_i^{1/2}.$$

Equações de Estimação Generalizadas (EEGs)

Em 1986 Kung-Yee Liang e Scott L. Zeger publicaram o artigo *Longitudinal data analysis using generalized linear models*, onde propuseram uma matriz de correlação dada por $\boldsymbol{R}_i(\boldsymbol{\rho})$, em que $\boldsymbol{\rho} = (\rho_1, \cdots, \rho_p)^\top$ é um **vetor de parâmetros de pertubação** que não depende de $\boldsymbol{\beta}$.

GEE

Matriz de variância-covariância

Se a verdadeira correlação entre os elementos de $\boldsymbol{Y}_i$ for dada por $\boldsymbol{R}_i(\boldsymbol{\rho})$, temos que a matriz de variância-covariância pode ser escrita como

$$\boldsymbol{\Omega}_i = \phi^{-1}\boldsymbol{V}_i^{1/2}\boldsymbol{R}_i(\boldsymbol{\rho})V_i^{1/2}. \tag{8}$$

$\boldsymbol{R}_i(\boldsymbol{\rho})$ é uma matriz $r_i x r_i$ que depende do número finito de parâmetros $\boldsymbol{\rho} = (\rho_1, \cdots, \rho_q)^\top$, sendo denominada matriz trabalho.

GEE

Para estimarmos $\boldsymbol{\beta}$ devemos resolver o seguinte sistema de equações:

$$\boldsymbol{S}_{\beta}(\hat{\boldsymbol{\beta}}_G) = 0,$$

$$\boldsymbol{S}_{\beta} = \sum_{i=1}^{n} \boldsymbol{D}_i^{\top} \boldsymbol{\Omega}_i (\boldsymbol{y}_i - \boldsymbol{\mu}_i). \tag{9}$$

$\boldsymbol{S}_{\beta}(\boldsymbol{\beta})$ também depende de ϕ e $\boldsymbol{\rho} = (\rho_1, \cdots, \rho_q)^{\top}$ que são estimados separadamente de $\boldsymbol{\beta}$.

Estimação

O processo iterativo para a estimação de $\boldsymbol{\beta}$, que é uma modificação do método escore de Fisher, é dado por:

$$\boldsymbol{\beta}_G^{(m+1)} = \boldsymbol{\beta}_G^{(m)} + \left\{ \sum_{i=1}^{n} \boldsymbol{D}_i^{(m)^\top} \boldsymbol{\Omega}_i^{(m)} \boldsymbol{D}_i^{(m)} \right\}^{-1} \times$$

$$\left[\sum_{i=1}^{n} \boldsymbol{D}_i^{(m)^\top} \boldsymbol{\Omega}_i^{-(m)} \left(\boldsymbol{y}_i - \boldsymbol{\mu}_i^{(m)} \right) \right]$$

Definição

Supondo que $\hat{\boldsymbol{\rho}}$ e $\hat{\phi}$ são estimadores consistentes de $\boldsymbol{\rho}$ e ϕ, respectivamente, temos que

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\boldsymbol{G}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}) \tag{10}$$

em que

$$\boldsymbol{\Sigma} = \lim_{n\to\infty} \left[n \left(\sum_{i=1}^{n} \boldsymbol{D}_i^\top \boldsymbol{\Omega}_i^{-1} \boldsymbol{D}_i \right)^{-1} \times \left\{ \sum_{i=1}^{n} \boldsymbol{D}_i^\top \boldsymbol{\Omega}_i^{-1} \mathbf{Var}(\boldsymbol{Y}_i) \boldsymbol{\Omega}_i^{-1} \boldsymbol{D}_i \right\} \right.$$
$$\left. \times \left(\sum_{i=1}^{n} \boldsymbol{D}_i^\top \boldsymbol{\Omega}_i^{-1} \boldsymbol{D}_i \right)^{-1} \right]$$

Definição

Se a matriz de correlação $\boldsymbol{R}_i(\boldsymbol{\rho})$ é definida corretamente, um estimador consistente para $\mathbf{Var}(\hat{\boldsymbol{\beta}}_G)$ é dado por $\boldsymbol{H}^{-1}(\hat{\boldsymbol{\beta}}_G)$, em que

$$\boldsymbol{H}(\hat{\boldsymbol{\beta}}_G) = \sum_{i=1}^{n} (\hat{\boldsymbol{D}}_i^{\top} \hat{\boldsymbol{\Omega}}_i^{-1} \hat{\boldsymbol{D}}_i), \tag{11}$$

com $\hat{\boldsymbol{D}}_i$ sendo avaliado em $\hat{\boldsymbol{\beta}}_G$ e $\hat{\boldsymbol{\Omega}}_i$ avaliado em $(\hat{\phi}, \hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\beta}}_G)$

Definição

Se a matriz trabalho $\boldsymbol{R}_i(\boldsymbol{\rho})$ é definida incorretamente $\boldsymbol{H}_1^{-1}(\hat{\boldsymbol{\beta}}_G)$ pode ser inconsistente. Um estimador robusto para $\mathbf{Var}(\hat{\boldsymbol{\beta}}_G)$, sugerido por Liang e Zeger (1986), é expresso por

$$\hat{\boldsymbol{V}}_G = \hat{\boldsymbol{H}}_1^{-1}(\hat{\boldsymbol{\beta}}_G)\boldsymbol{H}_2(\hat{\boldsymbol{\beta}}_G)\boldsymbol{H}_1^{-1}(\hat{\boldsymbol{\beta}}_G) \tag{12}$$

em que $\boldsymbol{H}_2(\hat{\boldsymbol{\beta}}_G) = \sum_{i=1}^{n}\{\hat{\boldsymbol{D}}_i^{\top}\hat{\boldsymbol{\Omega}}_i^{-1}(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)^{\top}\hat{\boldsymbol{\Omega}}^{-1}\hat{\boldsymbol{D}}_i\}$. O estimador $\hat{\boldsymbol{V}}_G$ é consistente mesmo se a matriz trabalho for definida incorretamente.

Matriz de correlação não estruturada

Quando a matriz de correlação $\boldsymbol{R}_i$ é não estruturada teremos $r_i(r_i - 1)/2$ parâmetros a serem estimados. Denotando $\boldsymbol{R}_i = \{R_{ijj'}\}$, o j, j'-ésimo elemento de $\boldsymbol{R}_i$ poderá ser estimado por

$$\hat{R}_{jj'} = \frac{\phi}{n} \sum_{i=1}^{n} \frac{(y_{ij} - \hat{\mu}_{ij})}{\sqrt{\hat{V}_{ij}}} \frac{(y_{ij'} - \hat{\mu}_{ij'})}{\sqrt{\hat{V}_{ij'}}}. \tag{13}$$

Matriz de correlação simétrica ou permutável

Nesse caso, assumimos $\boldsymbol{R}_i = \boldsymbol{R}_i(\rho)$, em que o (j, j')-ésimo elemento de $\boldsymbol{R}_i$ é dado por $R_{ijj'} = 1$, para $j = j'$, e $R_{ijj'} = \rho$, para $j \neq j'$. Um estimador consistente para ρ é dado por:

$$\hat{\rho} = \frac{\phi}{n} \sum_{i=1}^{n} \frac{1}{r_i(r_i - 1)} \sum_{j=1}^{r_i} \sum_{j'=1, j' \neq j}^{r_i} \frac{(y_{ij} - \hat{\mu}_{ij})}{\sqrt{\hat{V}_{ij}}} \frac{(y_{ij'} - \hat{\mu}_{ij'})}{\sqrt{\hat{V}_{ij'}}}. \qquad (14)$$

Autoregressiva

Matriz de correlação autorregressiva

Assumimos $\boldsymbol{R}_i = \boldsymbol{R}_i(\rho)$, em que o (j, j')-ésimo elemento de $\boldsymbol{R}_i$ é dado por $R_{ijj'} = 1$, para $j = j'$, e $R_{ijj'} = \rho^{|j-j'|}$, para $j \neq j'$. Um estimador consistente para ρ é dado por:

$$\hat{\rho} = \frac{\phi}{n} \sum_{i=1}^{n} \frac{1}{(r_i - 1)} \sum_{j=1}^{r_i - 1} \frac{(y_{ij} - \hat{\mu}_{ij})}{\sqrt{\hat{V}_{ij}}} \frac{(y_{i(j+1)} - \hat{\mu}_{i(j+1)})}{\sqrt{\hat{V}_{i(j+1)}}}. \qquad (15)$$

1. **Identidade:**

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

1. **Identidade:**

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

2. **Simétrica ou permutável:**

$$\begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

1. **Autoregressiva:**

$$\begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}$$

1. **Autoregressiva:**

$$\begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}$$

2. **Não estruturada:**

$$\begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix}$$

Parâmetro de dispersão

Estimação de ϕ

O parâmetro ϕ^{-1} pode ser estimado consistentemente por

$$\hat{\phi}^{-1} = \frac{1}{N-p} \sum_{i=1}^{n} \sum_{j=1}^{r_i} \frac{(y_{ij} - \hat{\mu}_{ij})^2}{\hat{V}_{ij}}, \tag{16}$$

Em que $N = \sum_{i=1}^{n} r_i$. Testes de hipóteses para $\boldsymbol{\beta}$ ou para subconjuntos de $\boldsymbol{\beta}$ podem ser desenvolvidos através de estatísticas tipo **Wald** com a matriz de variância-covariância estimada $\hat{\boldsymbol{V}}_G$.

Métodos de Diagnóstico

Resíduos

Em EEGs os **resíduos de Pearson** são calculados da seguinte maneira

$$\hat{r}_{P_{ij}} = \frac{\boldsymbol{e}_{ij}^{\top}\hat{\boldsymbol{A}}^{1/2}(\hat{\boldsymbol{V}}_i\hat{\boldsymbol{W}}_i)^{-1}(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)}{\sqrt{1-\hat{h}_{ijj}}} \tag{17}$$

para $i = 1, \cdots, n$ e $j = 1, \cdots, r_i$, em que $\boldsymbol{A}_i^{1/2} = \phi \boldsymbol{W}_i^{1/2}\boldsymbol{R}_i^{-1}\boldsymbol{W}_i^{1/2}$ é uma matriz de dimensão r_ixr_i, $\boldsymbol{e}_{ij}^{\top}$ é um vetor de dimensão 1xr_i de zeros com 1 na j-ésima posição.

Métodos de Diagnóstico

Alavancagem

Duas medidas de **alavancagem** são usualmente aplicadas em EEGs. Medidas de alavancagem referente ao j-ésimo indivíduo do i-ésimo grupo, dada por $\hat{h}_{ijj}$ e medida de alavancagem referente ao i-ésimo grupo, definida por

$$\hat{h}_i = \frac{1}{r_i} \sum_{j=1}^{r_i} h_{ijj}. \tag{18}$$

Métodos de Diagnóstico

Observações influentes

Uma versão aproximada da **distância de Cook** para avaliar o impacto da eliminação individual das observações na estimativa de $\hat{\beta}_G$ é dada por

$$LD_{ij} = \frac{\hat{h}_{ijj}}{(1 - \hat{h}_{ijj})} \hat{r}^2_{P_{ij}}. \tag{19}$$

Métodos de Diagnóstico

Critério de seleção de modelos

Uma proposta de critério de seleção de modelos em EEGs é dada por:

$$QIC = -2Q(\hat{\beta}_G) + 2\text{tr}(\hat{\boldsymbol{V}}_G \hat{\boldsymbol{H}}_{1I}), \tag{20}$$

em que $\hat{\beta}_G$ é a estimativa de quase-verossimilhança para uma matriz específica de correlação $\boldsymbol{R}_i(\rho)$ e $\boldsymbol{H}_{1I}$ é a matriz $\boldsymbol{H}_1$ avaliada sob a estrutura de independência.

Dados Ausentes

- Em um estudo longitudinal, alguns sujeitos podem abandonar o estudo antes de seu término, talvez por se mudarem para outra cidade ou por algum motivo que os faça não querer mais participar.
- Uma análise estatística que exclui os dados dos sujeitos para os quais há dados ausentes pode resultar na perda de muitas informações e em erros padrão maiores.
- Analisar apenas os dados observados, como se não houvesse dados ausentes, pode resultar em estimativas de parâmetros viesadas.

Dados Ausentes

Missing Completely at Random (**MCAR**)

- Se a probabilidade de que uma resposta esteja ausente for a mesma para todos os sujeitos, independentemente dos valores das variáveis explicativas, os dados são *missing completely at random* (**MCAR**).

Dados Ausentes

Missing Completely at Random (**MCAR**)

- Se a probabilidade de que uma resposta esteja ausente for a mesma para todos os sujeitos, independentemente dos valores das variáveis explicativas, os dados são *missing completely at random* (**MCAR**).

Missing at Random (**MAR**)

- Se a probabilidade de que a resposta esteja ausente variar de acordo com o tempo, mas não variar de acordo com a resposta para sujeitos com os mesmos valores das variáveis explicativas, então os dados não são MCAR, mas são *missing at random* (**MAR**).

Distribuição Binomial

Seja Y_{it} uma variável aleatória com distribuição binomial. Podemos expressá-la na forma da família exponencial:

$$
\begin{aligned}
f(y;\pi_{it},\phi) &= \binom{n_{it}}{y_{it}} \pi_{it}^{y_{it}} (1-\pi_{it})^{(n_{it}-y_{it})} \\
&= \exp\left\{ \log\binom{n_{it}}{y_{it}} + y_{it}\log(\pi_{it}) + (n_{it}-y_{it})\log(1-\pi_{it}) \right\} \\
&= \exp\left\{ \log\binom{n_{it}}{y_{it}} + y_{it}\log(\pi_{it}) + n_{it}\log(1-\pi_{it}) - y_{it}\log(1-\pi_{it}) \right\} \\
&= \exp\left\{ y_{it}\log\left(\frac{\pi_{it}}{1-\pi_{it}}\right) + \log\binom{n_{it}}{y_{it}} + n_{it}\log(1-\pi_{it}) \right\} \\
&= \exp\left\{ y_{it}\log\left(\frac{\pi_{it}}{1-\pi_{it}}\right) + n_{it}\log(1-\pi_{it}) + \log\binom{n_{it}}{y_{it}} \right\}.
\end{aligned}
$$

Portanto,

$$\phi = 1.$$

$$\begin{aligned}
\theta_{it} &= \log\left(\frac{\pi_{it}}{1-\pi_{it}}\right) \\
&\Rightarrow \exp(\theta_{it}) = \frac{\pi_{it}}{1-\pi_{it}} \\
&\Rightarrow \exp(\theta_{it}) - \pi_{it}\exp(\theta_{it}) = \pi_{it} \\
&\Rightarrow \exp(\theta_{it}) = \pi_{it} + \pi_{it}\exp(\theta_{it}) \\
&\Rightarrow \exp(\theta_{it}) = \pi_{it}(1+\exp(\theta_{it})) \\
&\Rightarrow \pi_{it} = \frac{\exp(\theta_{it})}{1+\exp(\theta_{it})}.
\end{aligned}$$

$$c(y_{it};\phi) = \log\binom{n_{it}}{y_{it}}.$$

$$
\begin{aligned}
b(\theta_{it}) &= -n\log(1-\pi_{it}) \\
&\Rightarrow -n\log\left\{1-\frac{\exp(\theta_{it})}{1+\exp(\theta_{it})}\right\} \\
&\Rightarrow -n\log\left\{\frac{1+\exp(\theta_{it})-\exp(\theta_{it})}{1+\exp(\theta_{it})}\right\} \\
&\Rightarrow -n\log\left\{\frac{1}{1+\exp(\theta_{it})}\right\} \\
&\Rightarrow -n\{\log(1)-\log(1+\exp(\theta_{it}))\} \\
&\Rightarrow n\log(1+\exp(\theta_{it})).
\end{aligned}
$$

Esperança

$$\begin{aligned}\mathbb{E}(Y_{it}) = b'(\theta_{it}) &= [n_{it}\log(1+\exp(\theta_{it}))]' \\ &= n_{it}\frac{1}{1+\exp(\theta_{it})}\exp(\theta_{it}) \\ &= n_{it}\pi_{it}.\end{aligned}$$

$$\begin{aligned}\mu_{it} &= n_{it}\pi_{it} \\ \Rightarrow \pi_{it} &= \frac{\mu_{it}}{n_{it}}.\end{aligned}$$

Variância

$$\mathbb{V}ar(Y_{it}) = \frac{b''(\theta)}{\phi} = V(\mu_{it}) = n_{it}\pi_{it}(1-\pi_{it}).$$

$$\Rightarrow n_{it}\frac{\mu_{it}}{n_{it}}\left(1-\frac{\mu_{it}}{n}\right)$$

$$\Rightarrow \mu_{it}\left(\frac{n_{it}}{n_{it}}-\frac{\mu_{it}}{n_{it}}\right)$$

$$\Rightarrow \mu_{it}\left(\frac{n_{it}-\mu_{it}}{n_{it}}\right)$$

$$\Rightarrow \frac{\mu_{it}}{n_{it}}(n_{it}-\mu_{it}).$$

Ligação Canônica

Usando a função de ligação canônica, conhecida como logit, temos que

$$\theta_{it} = \log\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \log\left(\frac{\mu_{it}/n_{it}}{1-\mu_{it}/n_{it}}\right) = \log\left(\frac{\mu_{it}}{n_{it}-\mu_{it}}\right).$$

$$\eta_{it} = g(\pi_{it}) = \log\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \boldsymbol{X}_{it}\boldsymbol{\beta}.$$

$$\eta_{it} = g(\pi_{it}) = \log\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \boldsymbol{X}_{it}\boldsymbol{\beta}.$$

$$\eta_{it} = g\left(\frac{\mu_{it}}{n_{it}}\right) = \log\left(\frac{\mu_{it}}{n_{it}-\mu_{it}}\right) = \boldsymbol{X}_{it}\boldsymbol{\beta}.$$

$$v_i = \frac{d\mu_{it}}{d\theta_{it}} = \mu_{it}\left(\frac{n_{it} - \mu_{it}}{n_{it}}\right).$$

$$v_i = \frac{d\mu_{it}}{d\theta_{it}} = \mu_{it}\left(\frac{n_{it} - \mu_{it}}{n_{it}}\right).$$

$$w_i = \left(\frac{d\mu_{it}}{d\eta_{it}}\right)^2 \frac{1}{V(\mu_{it})} = \frac{1}{V(\mu_{it})[g'(\mu_{it})]^2} = \frac{\mu_{it}}{n_{it}}(n_{it} - \mu_{it}).$$

Aplicação

Um exemplo de aplicação do ajuste do método *GEE* é apresentado no seguinte link: ▸ Link.

Referências I

TIELSCH, J. M. Blindness and Visual Impairment in an American Urban Population: The Baltimore Eye Survey. *Archives of Ophthalmology*, v. 108, n. 2, p. 286, fev. 1990. ISSN 0003-9950. Disponível em: <http://archopht.jamanetwork.com/article.aspx?doi=10.1001/archop

AGRESTI, A. *An introduction to categorical data analysis*. Third edition. Hoboken, NJ: John Wiley & Sons, 2019. (Wiley series in probability and statistics). ISBN 9781119405276 9781119405283.

Referências II

- LIANG, K.-Y.; ZEGER, S. L. Longitudinal data analysis using generalized linear models. *Biometrika*, v. 73, n. 1, p. 13–22, 04 1986. ISSN 0006-3444. Disponível em: <https://doi.org/10.1093/biomet/73.1.13>.
- GUEORGUIEVA, R. V.; AGRESTI, A. A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association*, Taylor & Francis, v. 96, n. 455, p. 1102–1112, 2001.

Referências III

- INTRODUCTION to Generalized Estimating Equations — STAT 504. https://online.stat.psu.edu/stat504/lesson/12/12.1. (Acesso em 18/05/2024).
- LIANG, K. Y.; ZEGER, S. L. Regression Analysis for Correlated Data. *Annual Review of Public Health*, v. 14, n. 1, p. 43–68, maio 1993. ISSN 0163-7525, 1545-2093. Disponível em: <https://www.annualreviews.org/doi/10.1146/annurev.pu.14.050193
- DOWNLOAD R-4.3.1 for Windows. The R-project for statistical computing. Disponível em: <https://cran.r-project.org/bin/windows/base/>.

Referências IV

- HALEKOH, U.; HøJSGAARD, S.; YAN, J. The *R* Package **geepack** for Generalized Estimating Equations. *Journal of Statistical Software*, v. 15, n. 2, 2006. ISSN 1548-7660. Disponível em: <http://www.jstatsoft.org/v15/i02/>.
- CAREY, V. J. et al. *gee: Generalized Estimation Equation Solver*. 2024. Disponível em: <https://cran.r-project.org/web/packages/gee/index.html>.

Referências V

DAVERN, M. et al. *General Social Survey 1972-2024. [Machine-readable data file]*. Chicago: NORC, 2024. Principal Investigator, Michael Davern; Co-Principal Investigators, Rene Bautista, Jeremy Freese, Pamela Herd, and Stephen L. Morgan. Sponsored by National Science Foundation. NORC ed. Chicago: NORC, 2024: NORC at the University of Chicago [producer and distributor]. Data accessed from the GSS Data Explorer website at gssdataexplorer.norc.org. Disponível em: <https://gssdataexplorer.norc.org>.