Análise de Dados Longitudinais
Modelo Marginal - GEE

Enrico A. Colosimo/UFMG
http://www.est.ufmg.br/~enricoc/

Modelos Marginais para Dados Longitudinais

1. Modelar a resposta média $E(Y)$.
2. Modelar a Estrutura de Variância-Covariância $Var(Y_i), i = 1, \ldots N$.
3. Assumir uma distribuição (normal) para a resposta contínua (**dispensável**).

Dois caminhos:

1. Assumir resposta normal: usar MQG ou MV (usual ou restrita).
2. Não assumir distribuição para a resposta: usar GEE: ”Generalized Estimation Equations”(Equações de Estimação Generalizadas).

Modelo Marginal

1. Utilizar o Método de Máxima Verossimilhança (usual ou restrito) para estimar β e também os componentes de variância. Ou seja, os parâmetros da média e também da estrutura escolhida de covariância .

2. Sem especificar distribuiçao para a resposta: Investigar qual é o impacto ao utilizarmos W ao invés de V. Ideia de GEE. (W deve ser aquela mais adequada para modelar a estrutura de covariância dos dados.)

Estimador de Máxima Verossimilhança

Encontrar simultaneamente o estimador da média (β) e o estimador para os componentes de variância (σ^2, α).
Seja

$$Y_i \sim N_n(X_i\beta, \sigma^2 V_0(\alpha))$$

$$f(y_i|\beta, \sigma^2, \alpha, X_i) = \frac{1}{(2\pi)^{n/2}|V_0|^{1/2}(\sigma^2)^{n/2}} exp\left\{-\frac{1}{2\sigma^2}Q_i\right\}$$

observe que β é estimado a partir de:

$$Q_i = (y_i - X_i\beta)' V_0^{-1}(y_i - X_i\beta)$$

Flexibilizar Suposições - GEE

1. Investigar a possibilidade de não especificar distribuiçao para Y.
2. Investigar qual é o impacto ao utilizarmos erradamente W ao invés da estrutura correta V.
3. Princípio das Equações de Estimação Generalizadas (GEE).

Supor que ao invés de $Var(Y) = V$ *foi utilizada erradamente* W*.* Ou seja,

$$\hat{\beta}_W = (X'W^{-1}X)^{-1}X'W^{-1}Y$$

Pergunta: Qual é o impacto na estimação de β se utilizarmos W ao invés de V?

Isto é,

- Qual é o vício de $\widehat{\beta}_W$?
- Qual é a $Var(\widehat{\beta}_W)$?

1. **Vício:**

$$\begin{aligned} E(\hat{\beta}_W) &= (X'W^{-1}X)^{-1}X'W^{-1}E(Y) \\ &= (X'W^{-1}X)^{-1}X'W^{-1}X\beta \\ &= \beta \end{aligned}$$

2. **Variância:**

$$\begin{aligned} Var(\hat{\beta}_W) &= Var\left[(X'W^{-1}X)^{-1}X'W^{-1}Y\right] \\ &= (X'W^{-1}X)^{-1}X'W^{-1}Var(Y)\left[W^{-1}X(X'W^{-1}X)^{-1}\right] \\ &= (X'W^{-1}X)^{-1}X'W^{-1}VW^{-1}X(X'W^{-1}X)^{-1} \end{aligned}$$

Pergunta: O que acontece ao especificarmos um W errado?

- $\hat{\beta}_W$ é não-viciado para qualquer especificação de W;
- Por exemplo, se $W = \sigma^2 I_{Nn}$

$$Var(\hat{\beta}_I) = (X'X)^{-1}X'VX(X'X)^{-1}$$

Observações:

- $\hat{\beta}_I = (X'X)^{-1}X'Y$ (Estimador de Mínimos Quadrados Ordinários) é não viciado.
- No entanto,

$$Var(\widehat{\beta}_{ols}) = \sigma^2(X'X)^{-1},$$

é viciada.

Pergunta Quanto $Var(\hat{\beta}_I)$ é diferente de $Var(\hat{\beta}_{MQG})$?

Ou seja, quanto

$$Var(\hat{\beta}_I) = (X'X)^{-1}X'VX(X'X)^{-1}$$

é diferente de

$$Var(\hat{\beta}_{MQG}) = (X'V^{-1}X)^{-1}$$

Resposta Na maioria da vezes estes estimadores são bem próximos.

Exemplo (Diggle et al., p. 59):

$$\begin{aligned} N &= 10 \\ k &= 5 \qquad (t = -2, -1, 0, 1, 2) \end{aligned}$$

$W = \sigma^2 I_{50}$

$$V_0 = [(1-\rho) I_5 + \rho 1_5 1_5']$$

e $V_i = \sigma^2 V_0$. O Modelo: $Y_{ij} = \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij}$

$$X_{50,2} = \begin{pmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & 2 \end{pmatrix}$$

Fazendo as contas:

$$X'X = \begin{pmatrix} 50 & 0 \\ 0 & 100 \end{pmatrix}$$

e

$$X'VX = \begin{pmatrix} 50(1+4\rho) & 0 \\ 0 & 100(1-\rho) \end{pmatrix}.$$

Desta forma,

$$Var(\hat{\beta}_I) = \sigma^2(X'X)^{-1}X'VX(X'X)^{-1} = \sigma^2 \begin{pmatrix} 0.02(1+4\rho) & 0 \\ 0 & 0.01(1-\rho) \end{pmatrix}.$$

Continuando as contas:

$$V_0^{-1} = (1-\rho)^{-1}\rho((1-\rho)(1+4\rho))^{-1}1_5 1_5'$$

e

$$X'V^{-1}X = \begin{pmatrix} 50(1+4\rho)^{-1} & 0 \\ 0 & 100(1-\rho)^{-1} \end{pmatrix}.$$

Desta forma,

$$Var(\hat{\beta}_{MQG}) = \sigma^2(X'V^{-1}X)^{-1} = \sigma^2 \begin{pmatrix} 0.02(1+4\rho) & 0 \\ 0 & 0.01(1-\rho) \end{pmatrix}$$

Ou seja , neste caso $Var(\hat{\beta}_I) = Var(\hat{\beta}_{MQG})$

Observação: Em várias situações a $Var(\hat{\beta}_I)$ é um estimador razoável para $Var(\hat{\beta}_{MQG})$.

Resumo

Assumindo o estimador de Mínimos Quadrados Ordinário $W = I_{Nk}$:

$$\hat{\beta}_I = (X'X)^{-1}X'Y$$

$$E(\hat{\beta}_I) = \beta$$

e sua variância fica usualmente bem estimada por:

$$Var(\hat{\beta}_I) = (X'X)^{-1}X'VX(X'X)^{-1}$$

Precisamos de um estimador consistente para V!!

Estimador Consistente de V

$$\hat{V}_{0i} = (Y_i - X_i\hat{\beta})(Y_i - X_i\hat{\beta})'$$

$$\hat{V} = \begin{bmatrix} \hat{V}_{01} & 0 & \cdots & 0 \\ 0 & \hat{V}_{02} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{V}_{0N} \end{bmatrix}_{Nn \times Nn}$$

Obs. O parâmetro σ^2 foi absorvido em V.

Equações de Estimação Generalizadas (GEE)

Proposto por Liang e Zeger (1986) para dados correlacionados.

Requer apenas a especificação correta da estrutura de média das variáveis respostas, sem fazer qualquer suposição distribucional.

Especificamos:

1. $E(Y_i) = X_i\beta = \mu_i$, e
2. matriz de correlação "de trabalho" das medidas repetidas, R_i, em que:

$$Var(Y_i) = W_i = A_i^{1/2} R_i A_i^{1/2}$$

W é a especificação mais próxima de V, correta e desconhecida $Var(Y)$.

GEE gera estimadores consistentes e assintoticamente normais para β, mesmo com má especificação R_i.

O Estimador GEE - Motivação

Uma motivação para o enfoque GEE vem dos estimadores de MQG que minimiza a função objetivo:

$$\sum_{i=1}^{N}(Y_i - X_i\beta)' V_i^{-1}(Y_i - X_i\beta).$$

O estimador de β, específico para o modelo linear, é a solução de

$$\sum_{i=1}^{N} X_i' V_i^{-1}(Y_i - X_i\beta) = 0,$$

que produz, resolvendo para β,

$$\hat{\beta} = \left(\sum_{i=1}^{N} X_i' V_i^{-1} X_i\right)^{-1} \left(\sum_{i=1}^{N} X_i' V_i^{-1} Y_i\right).$$

O Estimador GEE

O estimador GEE para β é dado por:

$$\sum_{i=1}^{N} X_i' W_i^{-1}(\alpha)(y_i - X_i\beta) = 0,$$

em que α são os componentes de variância.
Usualmente tomamos:

$$W_i(\alpha) = A_i^{1/2} R_i(\alpha) A_i^{1/2}$$

em que A_i é uma matriz diagonal com elementos $Var(Y_{ij})$ e $R_i(\alpha) = Cor(Y_{ij}, Y_{ik})$ (matriz de trabalho) é matriz de correlação.

Formas de Correlação de Trabalho R_i

- *independência*,
  $\Rightarrow$ dados longitudinais não correlacionados.
- *simetria composta*,
  $\Rightarrow$ mesma correlação para todos componentes.
- *AR1*,
  $\Rightarrow$ válida para medidas igualmente espaçadas no tempo;
- *não estruturada* estima todas as $n(n-1)/2$ correlações de R.
- Outras: banded, toeplitz, etc.

Variância do Estimador

1. *Naive* ou “baseada no modelo” - Viciada

$$\widehat{Var}(\hat{\beta}) = \left(\sum_{i=1}^{N} X_i' W_i(\hat{\alpha})^{-1} X_i \right)^{-1}.$$

2. *Robusta* ou “empírica”ou Sanduíche

$$\widehat{Var}(\hat{\beta}) = M_0^{-1} M_1 M_0^{-1},$$

em que

$$\begin{aligned} M_0 &= \sum_{i=1}^{N} X_i' W_i(\hat{\alpha})^{-1} X_i, \\ M_1 &= \sum_{i=1}^{N} X_i' W_i(\hat{\alpha})^{-1} (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)' W_i(\hat{\alpha})^{-1} X_i. \end{aligned}$$

Método de Estimação: GEE - Passos

1. Escolher $R(\alpha)$: matriz de trabalho e usualmente assumimos $A = \sigma^2 I_n$ (homocedasticidade).

2. Dado estimativas para α e σ, obtemos $\hat{W}$ e:

$$\hat{\beta} = (X' \hat{W}^{-1} X)^{-1} X' \hat{W}^{-1}) Y$$

   Obs. Inicializar o processo iterativo com $R(\alpha) = I_n$.

3. Encontrar os resíduos: $e_{ij} = Y_{ij} - X_{ij}\widehat{\beta}$. A partir dos resíduos é possível estimar

$$\widehat{\sigma}^2 = \frac{\sum_i \sum_j e_{ij}^2}{nN}$$

   e também os outros componentes de variância ϕ. Retornar ao passo 2 até a convergência.

Método de Estimação: GEE - Passos

Após a convergência estimar $Var(Y_i)$ e obter $\widehat{Var}(\hat{\beta})$:

$$\begin{aligned}\widehat{Var}(\hat{\beta}) &= (X'\widehat{W}^{-1}X)^{-1}X'\widehat{W}^{-1}\widehat{Var}(Y)\hat{V}^{-1}X(X'\hat{V}^{-1}X)^{-1} \\ &= \hat{M}_0^{-1}\hat{M}_1\hat{M}_0^{-1}\end{aligned}$$

Obs. A estimativa de ϕ é baseada nos resíduos. Por exemplo, em um desenho balanceado, a forma não estruturada é estimada por:

$$\hat{\alpha}_{jk} = \frac{1}{\hat{\sigma}N}\sum_{i=1}^{N} e_{ij}e_{ik}$$

GEE - Observações

1. Este estimador de $Var(\widehat{\beta})$ é chamado de estimador sanduíche (M_0^{-1} é o pão e M_1 é a carne)
2. Se tomarmos $V = I_{Nn}$, temos

$$\widehat{Var}(\hat{\beta}_I) = (X'X)^{-1}X'\widehat{Var}(Y_i)X(X'X)^{-1}$$

3. Se tomarmos $W = V = Var(Y)$,

$$\hat{Var}(\hat{\beta}_V) = \widehat{M}_0^{-1}$$

GEE - Características e Limitações

1. Vantagens/Características
   - $\widehat{\beta}$ é consistente mesmo que $Var(Y)$ for incorretamente especificada.
   - Não é necessário especificar uma distribuição para Y_i.
   - $Var(\widehat{\beta})$ é adequadamente estimada pelo estimador sanduíche.
2. Limitações
   - Desenho desbalanceado é uma restrição para a estimação usando GEE, especialmente para o estimador sanduíche.
   - A robustez do estimador sanduíche é uma propriedade assintótica.
   - A matriz de trabalho W_i deve ser especificada o mais próximo possível de $Var(Y_i)$ para obter eficiência/precisão para a estimação de β.

GEE - Características e Limitações

3. Continuação: Limitações
   - O estimador GEE, $\widehat{\beta}$ fica viciado na presença de dados perdidos se a matriz de trabalho não for corretamente especificada e o mecanismo de perda não for MCAR.
   - Na maioria dos softwares σ^2 é tomado como sendo invariante no tempo. Ou seja, $\sigma_1^2 = \sigma_2^2 = \cdots, \sigma_n^2 = \sigma^2$. Este fato é restritivo para analisar respostas contínuas.
4. O GEE apresenta, em geral, resultados semelhantes aos EMV e aos estimadores no modelo de efeitos aleatórios. A interpretação dos parâmetros é a mesma para todos os enfoques sob o modelo linear. O GEE é mais utliizado para modelos não-lineares, por sua interpretação populacional.

Exemplo: Chumbo em Crianças - GEE

- Modelo Não-Estruturado para a média (intercepto comum): (R: y factor(tempo)*factor(grupo)).
- Comparando estruturas para $Var(W_i)$, obtemos o mesmo ajuste para as quatro estruturas.
- Estimativas para média e erro-padrão para os coeficientes que comparam os grupos nos quatro tempos.

Coeficiente	GEE: Independente		GEE: Simetria Composta		GLS: Não Estruturada	
	Est.	EP	Est.	EP	Est.	EP
Linha base	−0,268	0,994	−0,268	0,994	−0,268	1,004
1a semana	11,406	1,109	11,406	1,109	11,406	1,120
4a semana	8,824	1,141	8,824	1,141	8,824	1,152
6a semana	3,152	1,244	3,152	1,244	3,152	1,257

Crianças - Transmissão Vertical - GEE

- Modelo quadrático para a média com termos de interação.
- Algumas formas para a $Var(W_i)$: exponencial, simetria composta.
- Modelo para média com 9 termos (interceptos diferentes)
- Resultados para os quatro termos da interação.

Coeficiente	Independente		Simetria Composta		GLS: Estrut. Expon.	
	Est.	EP	Est.	EP		
Idade:grupo	−0,164	0,059	−0,142	0,057	−0,160	0,057
Idade2:grupo	0,020	0,011	0,018	0,008	0,017	0,008
Idade:sexo	0,046	0,050	0,100	0,047	0,166	0,052
Idade2:sexo	−0,014	0,009	−0,015	0,007	−0,020	0,008

Modelos Marginais: GEE/EMVR

1. Características:
   - GEE e EMVR são similares (mesma eficiência) com dados completos.
   - A única condição para GEE produzir inferências válidas é a estrutura da média estar corretamente especificada.
   - Especificando corretamente a estrutura de variância-covariância ganha-se em eficiência no processo inferencial.
   - Na presença de dados faltantes (MAR e NMAR), o GEE não produz inferências válidas. Por outro lado, o EMVR produz inferências válidas nesta condição (somente MAR)se a distribuição normal for corretamente especificada para a resposta.

2 Limitações:

- Dados longitudinais desbalanceados. Somente a estrutura de variância-covariância

$$Cor(Y_{ij}, Y_{il}) = \rho^{|t_{ij} - t_{il}|}$$

é possível ser especificada sob desbalanceamento. Disponível no pacote gls do R, além da simetria composta e independente.

- Falta de flexibilidade do GEE no R na especificação da estrutura de variância da resposta.