MAT02035 - Modelos para dados correlacionados

Modelos marginais e Equações de Estimação Generalizadas

Rodrigo Citton P. dos Reis

citton.padilha@ufrgs.br

Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Departamento de Estatística

Porto Alegre, 2019

Introdução

- A premissa básica dos modelos marginais é fazer inferências sobre as médias populacionais.
- O termo marginal é usado aqui para enfatizar que a resposta média modelada é condicional apenas para covariáveis e não para outras respostas ou efeitos aleatórios.
- Uma característica dos modelos marginais é que os modelos para a média e a "associação dentro do indivíduo" (por exemplo, covariância) são especificados separadamente.

Notação

- Y_{ij} denota a variável de resposta para o i-ésimo indivíduo na j-ésima ocasião.
- Y_{ij} pode ser contínuo, binário ou uma contagem.
- Assumimos que existem n_i medições repetidas no i-ésimo indivíduo e cada Y_{ij} é observado no tempo t_{ij}.
- Associado a cada resposta, Y_{ij}, há um vetor $p \times 1$ de covariáveis, X_{ij}.
- As covariáveis podem ser invariantes no tempo (por exemplo, grupo de tratamento) ou variar no tempo (por exemplo, tempo desde a linha de base).

Características dos modelos marginais

- O foco dos modelos marginais está nas inferências sobre as médias populacionais.
- A média marginal, $\mu_{ij} = \mathrm{E}\left(Y_{ij}|X_{ij}\right)$, de cada resposta é modelada em função das covariáveis.
- Especificamente, os modelos marginais têm as três partes a seguir especificadas.

Características dos modelos marginais

1. A média marginal da resposta, μ_{ij}, depende das covariáveis através de uma função de ligação conhecida

$$g(\mu_{ij}) = \beta_1 X_{1ij} + \beta_2 X_{2ij} + \ldots + \beta_p X_{pij}$$

2. A variância marginal de Y_{ij} depende da média marginal de acordo com

$$\mathrm{Var}\left(Y_{ij}|X_{ij}\right) = \phi v(\mu_{ij}),$$

em que $v(\mu_{ij})$ é uma "função de variância" conhecida e ϕ é um parâmetro de escala que pode precisar ser estimado.

Nota: Para resposta contínua, pode permitir $\mathrm{Var}\left(Y_{ij}|X_{ij}\right) = \phi_j v(\mu_{ij})$.

3. A "associação intra-indivídual" entre as respostas é uma função das médias e de parâmetros adicionais, digamos α, que também precisam ser estimados.

Características dos modelos marginais

- Por exemplo, quando α representa correlações dois-a-dois entre respostas, as covariâncias entre as respostas dependem de $\mu_{ij}(\beta)$, ϕ e α:

$$\begin{aligned}\text{Cov}\,(Y_{ij}, Y_{ik}) &= dp(Y_{ij})\text{Corr}\,(Y_{ij}, Y_{ik})dp(Y_{ik}) \\ &= \sqrt{\phi v(\mu_{ij})}\text{Corr}\,(Y_{ij}, Y_{ik})\sqrt{\phi v(\mu_{ik})},\end{aligned}$$

em que $dp(Y_{ij})$ é o desvio padrão de Y_{ij}.

Medidas de Associação para Respostas Binárias

- Com respostas binárias, as correlações não são a melhor opção para modelar a associação, porque são limitadas pelas probabilidades marginais.
- Por exemplo, se $\mathrm{E}(Y_1) = \Pr(Y_1 = 1) = 0.2$ e $\mathrm{E}(Y_2) = \Pr(Y_2 = 1) = 0.8$, então $\mathrm{Corr}(Y_1, Y_2) < 0.25$.
- As correlações devem satisfazer certas desigualdades lineares determinadas pelas probabilidades marginais.
- É provável que essas restrições causem dificuldades na modelagem paramétrica da associação.

Medidas de Associação para Respostas Binárias

- Com respostas binárias, o *odds ratio* é uma medida natural de associação entre um par de respostas.
- O *odds ratio* para qualquer par de respostas binárias, Y_j e Y_k, é definido como

$$\mathrm{OR}(Y_j, Y_k) = \frac{\Pr(Y_j = 1, Y_k = 1)\Pr(Y_j = 0, Y_k = 0)}{\Pr(Y_j = 1, Y_k = 0)\Pr(Y_j = 0, Y_k = 1)}.$$

- Observe que as restrições no *odds ratio* são muito menos restritivas do que na correlação.
  - Com a resposta binária, pode-se modelar a associação dentro do indivíduo em termos de razão de chances (*odds ratio*), em vez de correlações.

Exemplos de modelos marginais

Exemplo 1. Respostas contínuas

1. $\mu_{ij} = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \ldots + \beta_p X_{ijp}$ (ou seja, regressão linear).
2. $\text{Var}\,(Y_{ij}|Xij) = \phi_j$ (ou seja, variância heterogênea, mas nenhuma dependência da variância em relação à média).
3. $\text{Corr}\,(Y_{ij}, Y_{ik}) = \alpha^{|k-j|}$ $(0 \leq \alpha \leq 1)$ (ou seja, correlação autoregressiva).

Exemplos de modelos marginais

Exemplo 2. Respostas binárias

1. $\text{logit}\,(\mu_{ij}) = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \ldots + \beta_p X_{ijp}$ (ou seja, regressão logística).
2. $\text{Var}\,(Y_{ij}|Xij) = \mu_{ij}(1-\mu_{ij})$ (variância Bernoulli).
3. $\text{OR}(Y_{ij}, Y_{ik}) = \alpha_{jk}$ (ou seja, *odds ratio* não estruturado), em que

$$\text{OR}(Y_j, Y_k) = \frac{\Pr(Y_j = 1, Y_k = 1)\Pr(Y_j = 0, Y_k = 0)}{\Pr(Y_j = 1, Y_k = 0)\Pr(Y_j = 0, Y_k = 1)}.$$

Exemplos de modelos marginais

Exemplo 3. Dados de contagem

1. $\log(\mu_{ij}) = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \ldots + \beta_p X_{ijp}$ (ou seja, regressão de Poisson).
2. $\mathrm{Var}\,(Y_{ij}|Xij) = \phi\mu_{ij}$ (ou seja, variância extra-Poisson, ou "sobredispersão" quando $\phi > 1$).
3. $\mathrm{Corr}\,(Y_{ij}, Y_{ik}) = \alpha$ (ou seja, correlação de simetria composta).

Interpretação dos parâmetros marginais do modelo

- Os parâmetros de regressão, β, têm interpretações na "média da população" (em que "média" é sobre todos os indivíduos dentro dos subgrupos da população):
  - Descreve o efeito das covariáveis nas respostas médias.
  - contrasta as médias nas subpopulações que compartilham valores de covariáveis comuns
- **Modelos marginais são mais úteis para inferências em nível populacional.**
- Os parâmetros de regressão são diretamente estimados a partir dos dados.
- A natureza ou magnitude da associação dentro do indivíduo (por exemplo, correlação) não altera a interpretação de β.

Interpretação dos parâmetros marginais do modelo

- Por exemplo, considere o seguinte modelo logístico,

$$\text{logit}\left(\mu_{ij}\right) = \text{logit}\left(\mathrm{E}\left[Y_{ij}|X_{ij}\right]\right) = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \ldots + \beta_p X_{ijp}.$$

- Cada elemento mede a mudança nas log-odds de uma resposta "positiva" por mudança de unidade na respectiva covariável, para subpopulações definidas por valores covariáveis fixos e conhecidos.
- A interpretação de qualquer componente de β, digamos β_k, é em termos de mudanças na resposta média transformada (ou "média da população") para uma alteração de uma unidade na covariável correspondente, digamos X_{ijk}.

Interpretação dos parâmetros marginais do modelo

- Quando X_{ijk} assume um valor x, o log-odds de uma resposta positiva é,

$$\log\left[\frac{\Pr(Y_{ij}=1|X_{ij1},\ldots,X_{ijk}=x,\ldots,X_{ijp})}{\Pr(Y_{ij}=0|X_{ij1},\ldots,X_{ijk}=x,\ldots,X_{ijp})}\right]=\beta_1X_{ij1}+\ldots+\beta_kx+\ldots+\beta_pX_{ijp}.$$

- Similarmente, quando X_{ijk} assume algum valor $x+1$,

$$\log\left[\frac{\Pr(Y_{ij}=1|X_{ij1},\ldots,X_{ijk}=x+1,\ldots,X_{ijp})}{\Pr(Y_{ij}=0|X_{ij1},\ldots,X_{ijk}=x+1,\ldots,X_{ijp})}\right]=\beta_1X_{ij1}+\ldots+\beta_k(x+1)+\ldots+\beta_pX_{ijp}.$$

- β_k é a mudança na log-odds para subgrupos da população de estudo (definadas por quaisquer valores fixos $X_{ij1},\ldots,X_{ij(k-1)},X_{ij(k+1)},\ldots,X_{ijp}$).

Inferência estatística para modelos marginais

- Máxima verossimilhança (MV)?
  - Infelizmente, com dados de resposta discretos, não existe um análogo simples da distribuição normal multivariada.
- Na ausência de uma função de verossimilhança "conveniente" para dados discretos, não existe uma abordagem unificada baseada em probabilidade para modelos marginais.
- Abordagem alternativa para estimação: **Equações de Estimação Generalizadas (GEE)**.

Equações de Estimação Generalizadas

- Evita fazer suposições distribucionais sobre Y_i completamente.
- Potenciais vantagens:
  - O pesquisador não precisa se preocupar com o fato de a distribuição de Y_i se aproximar de alguma distribuição multivariada.
  - Isso evita a necessidade de especificar modelos para as associações complexas (momentos de ordem superior) entre as respostas.
- Isso leva a um método de estimação, conhecido como equações de estimação generalizadas (GEE), fácil de implementar.

Equações de Estimação Generalizadas

- A abordagem GEE tornou-se um método extremamente popular para analisar dados longitudinais discretos.
- Esta fornece uma abordagem flexível para modelar a média e a estrutura de associação intra-indivíduo em pares.
- Ele pode lidar com delineamentos inerentemente desbalanceados e com a perda de dados com facilidade (embora faça suposições fortes sobre o mecanismo de perda de dados).
- A abordagem GEE é computacionalmente direta e foi implementada *softwares* estatístico amplamente disponíveis.

Equações de Estimação Generalizadas

- O estimador GEE de β soluciona as seguintes **equações de estimação generalizadas**

$$\sum_{i=1}^{N} D' V_i^{-1}(y_i - \mu_i) = 0,$$

em que V_i é a chamada matriz de covariância de **"trabalho"**.

- Por matriz de covariância de trabalho, queremos dizer que V_i se aproxima da verdadeira matriz de covariância subjacente para Y_i.
- Ou seja, $V_i \approx \text{Cov}(Y_i)$, reconhecendo que $V_i \neq \text{Cov}(Y_i)$, a menos que os modelos para as variâncias e as associações intra-indivíduo estejam corretos.
- $D_i = \partial\mu_i/\partial\beta$ é a matriz "derivativa" (de μ_i em relação aos componentes de β).

Equações de Estimação Generalizadas

- Portanto, as equações de estimação generalizadas dependem de β e α.
- Como as equações de estimação generalizadas dependem de ambos, é necessário um **procedimento iterativo** de estimação em dois estágios:
  1. Dadas as estimativas atuais de α e ϕ, é obtida uma estimativa de β como a solução para as "equações de estimação generalizadas";
  2. Dada a estimativa atual de β, estimativas α e ϕ são obtidas com base nos resíduos padronizados,

$$r_{ij} = (Y_{ij} - \mu_{ij})/v(\hat{\mu}_{ij})^{1/2}.$$

Equações de Estimação Generalizadas

- Por exemplo, ϕ pode ser estimado por

$$\frac{1}{Nn - p} \sum_{i=1}^{N} \sum_{j=1}^{n} r_{ij}^2$$

- Os parâmetros de correlação, α, podem ser estimados de maneira similar.
- Por exemplo, correlações não-estruturadas, $\alpha_{jk} = \text{Corr}\,(Y_{ij}, Y_{ik})$, pode ser estiamda por

$$\hat{\alpha} = \frac{1}{N - p} \hat{\phi}^{-1} \sum_{i=1}^{N} r_{ij} r_{ik}$$

- Finalmente, no procedimento de estimação em duas etapas, iteramos entre as etapas **(1)** e **(2)** até que a convergência seja alcançada.

Propriedades dos estimadores GEE

- $\hat{\beta}$, a solução para as equações de estimação generalizadas, possui as seguintes propriedades:
  1. $\hat{\beta}$ é um estimador consistente de β.
  2. Em amostras grandes, $\hat{\beta}$ tem distribuição normal multivariada.
  3. $\text{Cov}(\hat{\beta}) = B^{-1}MB^{-1}$, em que

$$B = \sum_{i=1}^{N} D_i' V_i^{-1} D_i,$$

e

$$M = \sum_{i=1}^{N} D_i' V_i^{-1} \text{Cov}(Y_i) V_i^{-1} D_i.$$

Propriedades dos estimadores GEE

- B e M podem ser estimados substituindo α, ϕ, e β por suas estimativas, e substituindo $\text{Cov}\,(Y_i)$ por $(Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)'$.

Estimador sanduíche

- **Nota:** podemos usar esse **estimador de variância empírico** ou denominado **"sanduíche"**, mesmo quando a covariância foi mal especificada.

Resumindo

Os estimadores GEE têm as seguintes propriedades atraentes:

1. Em muitos casos, $\hat{\beta}$ é quase tão eficiente quanto ao EMV. Por exemplo, o GEE tem a mesma forma que as equações de verossimilhança para o modelo normal multivariado e também alguns modelos para dados discretos.
2. $\hat{\beta}$ é consistente mesmo que a covariância de Y_i tenha sido mal especificada.
3. Erros padrão para $\hat{\beta}$ podem ser obtidos usando o **estimador empírico** ou denominado **"sanduíche"**.

Avisos

- **Próxima aula (03/12):** Modelos marginais (GEE) - exemplos.
- **Para casa:** ler o Capítulo 12 e 13 do livro "**Applied Longitudinal Analysis**".
  - Caso ainda não tenha lido, leia também os Caps. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 e 11.
- **Para casa:** veja o *help* do pacote `geepack` do `R`.

Bons estudos!