MAT02035 - Modelos para dados correlacionados

Modelos lineares generalizados: uma breve revisão

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, 2022

Modelos lineares generalizados

Modelos lineares generalizados

- **Modelos lineares generalizados** (MLG) são uma classe de modelos de regressão; eles incluem o modelo de regressão linear padrão, mas também muitos outros modelos importantes:
  - Regressão linear para dados contínuos
  - Regressão logística para dados binários
  - Modelos de regressão log-linear / Poisson para dados de contagem
- Modelos lineares generalizados estendem os métodos de análise de regressão a configurações nas quais a variável resposta pode ser categórica.

Notação

- Assuma N realizações independentes de uma única variável resposta Y_i sejam observadas.
- Associado a cada resposta Y_i, existe um vetor $p \times 1$ de covariáveis, $X_{i1}, \ldots, X_{ip}$.
- **Objetivo:** o interesse principal está em relacionar a média de Y_i, $\mu_i = \mathrm{E}\left(Y_i | X_{i1}, \ldots, X_{ip}\right)$, às covariáveis.

Modelos lineares generalizados

Em modelos lineares generalizados:

1. Assume-se que distribuição da variável resposta, Y_i, pertence a família de distribuições conhecida como **família exponencial**.

- Fazem parte da família exponencial, **entre outras** , os modelos:
  - Normal;
  - Gama;
  - Bernoulli/Binomial;
  - Poisson.

Modelos lineares generalizados

2. Um **componente sistemático** que especifica os efeitos das covariáveis na média da distribuição de Y_i

$$\eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} = \sum_{k=1}^{p} \beta_k X_{ik}.$$

3. A transformação da média da resposta, μ_i, tem uma relação linear com as covariáveis por meio de uma **função de ligação** apropriada:

$$g(\mu_i) = \eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip},$$

em que a função de ligação $g(\cdot)$ é uma função conhecida, por exemplo, $\log(\mu_i)$.

A família exponencial

- Todas as distribuições que pertencem a família exponencial podem ser expressas como

$$f(y_i; \theta_i, \phi) = \exp[\{y_i\theta_i - a(\theta_i)\}/\phi + b(y_i, \phi)],$$

em que $a(\cdot)$ e $b(\cdot)$ são funções específicas que distinguem um membro da família de outro.

- A família exponencial expressa desta forma tem θ_i como um parâmetro de locação ("canônico") e ϕ como um parâmetro de escala (ou dispersão).
- **Exercício:** Identifique as distribuições Normal, Bernoulli e Poisson como membros da família exponencial.

Média e variância das distribuições na família exponencial

- Distribuições na família exponencial compartilham algumas propriedades estatísticas comuns.
  - Por exemplo, $\mathrm{E}\left(Y_i\right) = \mu_i = \frac{\partial a(\theta_i)}{\partial \theta}$ e $\mathrm{Var}\left(Y_i\right) = \phi \frac{\partial^2 a(\theta_i)}{\partial \theta^2}$.
- Assim, a variância de Y_i pode ser expressa em termos de

$$\mathrm{Var}\left(Y_i\right) = \phi v(\mu_i),$$

em que o parâmetro de escala $\phi > 0$.

- A **função de variância**, $v(\mu_i)$, descreve como a variância da resposta está funcionalmente relacionada μ_i, a média de Y_i.

A função de ligação

- A função de ligação aplica uma transformação à média e, em seguida, vincula as covariáveis à média transformada,

$$g(\mu_i) = \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}$$

em que a função de ligação $g(\cdot)$ é uma função conhecida, por exemplo, $\log(\mu_i)$.

- Isso implica que é a resposta média transformada que muda linearmente com as mudanças nos valores das covariáveis.

Modelos lineares generalizados

Funções de ligação canônicas e de variância para as distribuições normais, Bernoulli e Poisson.

Distribuição	**Função de variância**	**Ligação canônica**
Normal	$v(\mu) = 1$	Identidade: $\mu = \eta$
Bernoulli	$v(\mu) = \mu(1 - \mu)$	Logit: $\log\left[\frac{\mu}{1-\mu}\right] = \eta$
Poisson	$v(\mu) = \mu$	Log: $\log(\mu) = \eta$

em que $\eta = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$.

Extensões de modelos lineares generalizados para dados longitudinais

Extensões de MLG para dados longitudinais

- Quando a variável resposta é categórica (por exemplo, dados binários e de contagem), modelos lineares generalizados (por exemplo, regressão logística) podem ser estendidos para lidar com as respostas correlacionadas.
- No entanto, transformações não lineares da resposta média (por exemplo, logit) levantam questões adicionais relativas à interpretação dos coeficientes de regressão.
- Como veremos, modelos diferentes para dados longitudinais discretos têm objetivos de inferência um tanto diferentes.

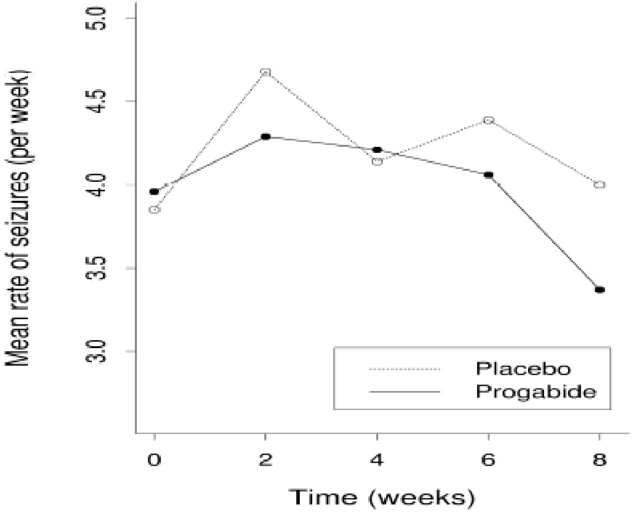
Exemplo: Tratamento oral da infecção das unhas dos pés

- Estudo aleatorizado, duplo-cego, grupo paralelo, multicêntrico de 294 pacientes comparando 2 tratamentos orais (denotados A e B) para infecção nas unhas dos pés.
- **Variável resposta:** variável binária indicando presença de onicólise (separação da placa ungueal do leito ungueal).
- Pacientes avaliados quanto ao grau de onicólise (separação da placa ungueal do leito ungueal) na linha de base (semana 0) e nas semanas 4, 8, 12, 24, 36 e 48.
- Interesse está na taxa de declínio da proporção de pacientes com onicólise ao longo do tempo e os efeitos do tratamento nessa taxa.

Exemplo: Ensaio clínico de progabida anti-epiléptica

- Estudo aleatorizado, controlado por placebo, do tratamento de crises epilépticas com progabida.
- Os pacientes foram aleatorizados para tratamento com progabida ou placebo, além da terapia padrão.
- **Variável resposta:** Contagem do número de convulsões
- Cronograma de medição: medição da linha de base durante 8 semanas antes da randomização. Quatro medições durante intervalos consecutivos de duas semanas.
- Tamanho da amostra: 28 epiléticos com placebo; 31 epiléticos em progabide

Exemplo: Ensaio clínico de progabida anti-epiléptica



MLG para dados longitudinais

- Em seguida, focamos em várias abordagens distintas para analisar respostas longitudinais.
- Essas abordagens podem ser consideradas extensões de modelos lineares generalizados para dados correlacionados.
- A ênfase principal será em dados de resposta discreta, por exemplo, dados de contagem ou respostas binárias.

MLG para dados longitudinais

- **Nota:** nos modelos lineares (efeitos mistos) para respostas contínuas, a interpretação dos coeficientes de regressão é independente da correlação entre as respostas.
- Com dados de resposta discreta, esse não é mais o caso.
- Com modelos não lineares para dados discretos, diferentes abordagens para contabilizar a correlação levam a modelos com coeficientes de regressão com interpretações distintas.
  - Voltaremos a esta questão importante no decorrer do curso.
- No restante desta aula, examinaremos brevemente três extensões principais de modelos lineares generalizados.

MLG para dados longitudinais

- Suponha que $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in})'$ é um vetor de respostas correlacionadas do i-ésimo indivíduo.
- Para analisar esses dados correlacionados, precisamos especificar ou pelo menos fazer suposições sobre a distribuição multivariada ou conjunta,

$$f(Y_{i1}, Y_{i2}, \ldots, Y_{in}).$$

- A maneira pela qual a distribuição multivariada é especificada produz três abordagens analíticas distintas:
  1. Modelos marginais;
  2. Modelos de efeitos mistos;
  3. Modelos de transição.

Modelos marginais

- Esta abordagem especifica a distribuição marginal em cada momento:

$$f(Y_{ij}) \quad \text{para} \quad j = 1, 2, \ldots, n.$$

juntamente com algumas suposições sobre a estrutura de covariância das observações.

- 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).

Ilustração

- Considere o **estudo tratamento oral da infecção das unhas dos pés**.
- Estudo aleatorizado, duplo-cego, grupo paralelo, multicêntrico, de 294 pacientes comparando 2 tratamentos orais (denotados A e B) para infecção das unhas dos pés.
- **Variável resposta:** variável binária que indica presença de onicólise (separação da placa ungueal do leito ungueal).
- Pacientes avaliados quanto ao grau de onicólise (separação da placa ungueal do leito ungueal) na linha de base (semana 0) e nas semanas 4, 8, 12, 24, 36 e 48.
- O interesse encontra-se na taxa de declínio da proporção de pacientes com onicólise ao longo do tempo e nos efeitos do tratamento nessa taxa.

Ilustração

- Suponha que a probabilidade marginal de onicólise siga um modelo logístico,

$$\text{logit}\left\{\Pr(Y_{ij} = 1)\right\} = \beta_1 + \beta_2 \text{Mês}_{ij} + \beta_3 \text{Trt}_i + \beta_4 (\text{Trt}_i \times \text{Mês}_{ij}),$$

em que $\text{Trt} = 1$ se o grupo de tratamento B e 0, caso contrário.

- Este é um exemplo de um modelo marginal.
- Note, no entanto, que **a estrutura de covariância ainda precisa ser especificada**.

Modelos de efeitos mistos

- Outra possibilidade é supor que um subconjunto dos parâmetros de regressão no modelo linear generalizado varie de indivíduo para indivíduo.
- Especificamente, poderíamos assumir que os dados de um único indivíduo são observações independentes com uma distribuição pertencente à família exponencial, mas que os coeficientes de regressão podem variar de indivíduo para indivíduo.
- Ou seja, **condicional** aos efeitos aleatórios, supõe-se que as respostas para um único indivíduo sejam observações independentes de uma distribuição pertencente à família exponencial.

Ilustração

- Considere o **estudo tratamento oral da infecção das unhas dos pés**.
- Suponha, por exemplo, que a probabilidade de onicólise para os participantes do estudo seja descrita por um modelo logístico, mas que o risco para um indivíduo dependa de seu "nível de resposta aleatória" **latente** (talvez determinado ambiental e geneticamente).
- Podemos considerar um modelo em que

$$\text{logit}\{\Pr(Y_{ij} = 1)\} = \beta_1 + \beta_2 \text{Mês}_{ij} + \beta_3 \text{Trt}_i + \beta_4 (\text{Trt}_i \times \text{Mês}_{ij}) + b_i.$$

- Observe que esse modelo também requer especificação da distribuição de efeitos aleatórios, $F(b_i)$.
- Este é um exemplo de um **modelo linear generalizado de efeitos mistos**.

Modelos de transição (Markov)

- Finalmente, outra abordagem é expressar a distribuição conjunta como uma série de distribuições condicionais,

$$f(Y_{i1}, Y_{i2}, \ldots, Y_{in}) = f(Y_{i1}) \times f(Y_{i2}|Y_{i1}) \times \ldots \times f(Y_{in}|Y_{i1}, \ldots, Y_{i,n-1}).$$

- Isso é conhecido como **modelo de transição** (ou modelo para as transições) porque representa a distribuição de probabilidade em cada ponto do tempo como condicional ao passado.
- Isso fornece uma representação completa da distribuição conjunta.

Ilustração

- Considere o **estudo tratamento oral da infecção das unhas dos pés**.
- Poderíamos escrever o modelo de probabilidade como

$$
\begin{aligned}
f(Y_{i1}, Y_{i2}, \ldots, Y_{in}|X_i) &= f(Y_{i1}|X_i) \times f(Y_{i2}|Y_{i1}, X_i) \times f(Y_{i3}|Y_{i1}, Y_{i2}, X_i) \\
&\quad \times \ldots \times f(Y_{i7}|Y_{i1}, Y_{i2}, \ldots, Y_{i6}, X_i).
\end{aligned}
$$

- Ou seja, a probabilidade de onicólise no tempo 2 é modelada condicionalmente à presença/ausência de onicólise no tempo 1 e assim por diante.

Ilustração

- Por exemplo, um modelo logístico de **"primeira ordem"**, permitindo dependência apenas da resposta anterior, é fornecido por

$$\text{logit}\left\{\Pr(Y_{ij} = 1 | Y_{i,j-1})\right\} = \beta_1 + \beta_2 \text{Mês}_{ij} + \beta_3 \text{Trt}_i + \beta_4 (\text{Trt}_i \times \text{Mês}_{ij}) + \beta_5 Y_{i,j-1}.$$

Em resumo

- Discutimos as principais características dos modelos lineares generalizados.
- Descrevemos brevemente três extensões principais de modelos lineares generalizados para dados longitudinais:
  1. Modelos marginais;
  2. Modelos de efeitos mistos;
  3. Modelos de transição.
- No restante do curso, focaremos em Modelos Marginais.

Em resumo

- Em geral, os modelos de transição são um pouco menos úteis para modelar efeitos de covariáveis.
- Especificamente, inferências de um modelo de transição podem ser potencialmente enganosas se um tratamento ou exposição alterar o risco ao longo do período de acompanhamento.
- Nesse caso, o risco condicional, dado o histórico anterior do resultado, é alterado de maneira menos nitidamente.

Avisos

- **Próxima aula:** Modelos marginais (GEE).
- **Para casa:** ler o Capítulo 11 do livro "**Applied Longitudinal Analysis**" (em particular a Seção 11.7).
  - Caso ainda não tenha lido, leia também os Caps. 1, 2, 3, 4, 5, 6, 7, 8, 9 e 10.
  - Veja o *help* da função `glm` do `R`; rode os exemplos apresentados no *help* da função.

Bons estudos!