MAT02035 - Modelos para dados correlacionados

Estimação e inferência estatística

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

Introdução

- Até agora, nossa discussão de modelos para dados longitudinais tem sido muito geral, sem menção de métodos para estimar os coeficientes de regressão ou a covariância entre as medidas repetidas.
- **Relembrando:** o modelo de regressão linear geral para o vetor de resposta média

$$\mathrm{E}\left(Y_i|X_i\right) = X_i\beta.$$

- Assumimos que o vetor de respostas, Y_i, tem uma distribuição condicional que é **normal multivariada** com matriz de covariância

$$\mathrm{Cov}\left(Y_i|X_i\right) = \Sigma_i = \Sigma_i(\theta).$$

- Note que θ é um vetor $q \times 1$ de **parâmetros de covariância**.

Introdução

- Dados balanceados ($n_i = n$), em que covariância "não-estruturada" é assumida, temos n variâncias e $\frac{n(n-1)}{2}$ covariâncias como elementos do vetor θ.
- Se a covariância é assumida ter um padrão de "simetria composta", então $q = 2$ e θ tem dois elementos.
- Nesta aula, consideramos uma estrutura para estimativa dos parâmetros desconhecidos, β e θ (ou equivalentemente, Σ_i).

Estimação: máxima verossimilhança

Estimação: máxima verossimilhança

- Dado que foram feitas suposições completas sobre a distribuição do vetor de respostas, Y_i, uma abordagem muito geral de estimativa é o **método da máxima verossimilhança** (MV).
  - A ideia fundamental por trás da estimativa de MV é realmente bastante simples e é transmitida por seu nome: use como estimativas de β e θ os valores que são mais prováveis (ou mais "verossímeis") para os dados que foram realmente observados.
  - As estimativas de verossimilhança máxima de β e θ são aqueles valores de β e θ que maximizam a probabilidade conjunta das variáveis resposta **avaliadas em seus valores observados**.

Estimação: máxima verossimilhança

- A probabilidade das variáveis de resposta avaliadas no conjunto fixo de valores observados e consideradas como funções de β e $\Sigma_i(\theta)$ é conhecida como **função de verossimilhança**.
  - Assim, a estimativa de β e θ prossegue **maximizando** a função de probabilidade.
- Os valores de β e θ que maximizam a função de verossimilhança são chamados de **estimativas de máxima verossimilhança** de β e θ, e geralmente são indicados $\hat{\beta}$ e $\hat{\theta}$ (ou $\widehat{\Sigma}_i$, $\Sigma_i(\hat{\theta})$).

Observações independentes

- Suponha que os dados surjam de uma série de estudos transversais que são repetidos em n ocasiões diferentes.
- Em cada ocasião, os dados são obtidos em uma amostra de N indivíduos.
  - Aqui é razoável supor que as observações sejam **independentes** umas das outras, uma vez que cada indivíduo é medido em apenas uma ocasião.
- Além disso, para facilitar a exposição, assumimos que a variância é constante, digamos a σ^2.
- A resposta média está relacionada às covariáveis através do seguinte modelo de regressão linear:

$$\mathrm{E}\left(Y_{ij}|X_{ij}\right) = X_{ij}'\beta.$$

Observações independentes

- Para obter estimativas de máxima verossimilhança de β, devemos encontrar os valores dos parâmetros de regressão que maximizem a função de densidade de probabilidade normal conjunta de todas as observações, avaliados nos valores observados da resposta e considerados como uma função de β (e σ^2).
- **Lembrando** que a função de densidade de probabilidade normal (ou gaussiana) univariada para Y_{ij}, dado X_{ij}, pode ser expressa como

$$f(y_{ij}) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2}(y_{ij} - \mu_{ij})^2/\sigma^2\right\},\ -\infty < y_{ij} < \infty.$$

Observações independentes

- Quando todas as respostas são independentes umas das outras, a função de verossimilhança é simplesmente o produto das funções individuais de densidade de probabilidade normal univariada para Y_{ij} dado X_{ij},

$$\prod_{i=1}^{N} \prod_{j=1}^{n} f(y_{ij}).$$

Observações independentes

- É mais comum trabalhar com a função **log-verossimilhança**, que envolverá somas, em vez de produtos, das funções individuais de densidade de probabilidade normal univariada para Y_{ij}.
  - Observe que maximizar a verossimilhança é equivalente a maximizar o logaritmo da verossimilhança; o último é indicado por l.
- Portanto, o objetivo é maximizar

$$l = \log\left\{\prod_{i=1}^{N}\prod_{j=1}^{n} f(y_{ij})\right\} = -\frac{K}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{n}(y_{ij} - X'_{ij}\beta)^2/\sigma^2,$$

avaliado nos valores numéricos observados dos dados, em relação aos parâmetros de regressão, β. - Aqui $K = n \times N$, o **número total de observações**.

Observações independentes

- Observe que β não aparece no primeiro termo da log-verossimilhança.
  - Como resultado, esse termo pode ser ignorado ao maximizar a log-verossimilhança em relação a β.
- Além disso, como o segundo termo tem um sinal negativo, maximizar a log-verossimilhança em relação a β é equivalente a minimizar a seguinte função:

$$\sum_{i=1}^{N} \sum_{j=1}^{n} (y_{ij} - X'_{ij}\beta)^2.$$

Observações independentes

- Maximizar ou minimizar uma função é um problema matemático comum que pode ser resolvido usando o cálculo.
- Especificamente, a estimativa de máxima verossimilhança de β pode ser obtida igualando a derivada da log-verossimilhança, frequentemente chamada de **função escore**, a zero e encontrando a solução para a equação resultante.
- No entanto, no exemplo considerado aqui, não há necessidade real de recorrer ao cálculo.
- A obtenção da estimativa de máxima verossimilhança de β é equivalente a encontrar a estimativa de **mínimos quadrados ordinários** (MQO) de β, ou seja, o valor de β que minimiza a soma dos quadrados dos resíduos.

Observações independentes

- Usando a notação vetorial, a solução dos mínimos quadrados pode ser escrita como

$$\hat{\beta} = \left\{ \sum_{i=1}^{N} \sum_{j=1}^{n} (X_{ij} X_{ij}') \right\}^{-1} \sum_{i=1}^{N} \sum_{j=1}^{n} (X_{ij} y_{ij}).$$

- Essa estimativa de mínimos quadrados é o valor produzido por qualquer *software* estatístico padrão para regressão linear (por exemplo, `PROC GLM` ou `PROC REG` no `SAS`, a função `lm` no `R` e o comando `regress` no `Stata`).
  - **Exercício:** considere o modelo dos dados de nível de chumbo no sangue; compare as estimativas através de uma implementação sua de $\hat{\beta}$ com a função `lm`.
- Além disso, note que até agora apenas focamos na estimativa de β, ignorando a estimativa de σ^2; a seguir, também consideramos a estimativa da matriz de covariância.

Observações correlacionadas

- Quando há n_i medidas repetidas no mesmo indivíduo, **não se pode assumir que essas medidas repetidas são independentes**.
  - Como resultado, precisamos considerar a função de densidade de probabilidade conjunta para o **vetor de medidas repetidas**.
- Observe, no entanto, que os vetores de medidas repetidas são assumidos como **independentes** uns dos outros.
  - Assim, a função log-verossimilhança, l, pode ser expressa como uma **soma das funções multivariadas individuais** da densidade de probabilidade normal para Y_i dado X_i.

Observações correlacionadas

- Primeiro assumimos que Σ_i (ou θ) é conhecido (e, portanto, não precisa ser estimado); depois, relaxaremos essa suposição muito irrealista.
- Dado que $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})'$ é assumido como tendo uma distribuição condicional que é normal multivariada, devemos maximizar a seguinte função de log-verossimilhança:

$$l = -\frac{K}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{N}\log|\Sigma_i| - \frac{1}{2}\left\{\sum_{i=1}^{N}(y_i - X_i\beta)'\Sigma_i^{-1}(y_i - X_i\beta)\right\},$$

- $K = \sum_{i=1}^{N} n_i$ é o **número total de observações**.

Observações correlacionadas

- Note que β não aparece nos dois primeiros termos na log-verossimilhança.
  - Esses dois termos podem ser ignorados ao maximizar a log-verossimilhança em relação a β.
- Além disso, como o terceiro termo tem um sinal negativo, maximizar a log-verossimilhança em relação a β é equivalente a minimizar

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

Observações correlacionadas

- O estimador de β que minimiza essa expressão é conhecido como estimador de **mínimos quadrados generalizados** (MQG) de β e pode ser expresso como

$$\hat{\beta} = \left\{ \sum_{i=1}^{N} \sum_{j=1}^{n} (X_i' \Sigma_i^{-1} X_i) \right\}^{-1} \sum_{i=1}^{N} \sum_{j=1}^{n} (X_i' \Sigma_i^{-1} y_i).$$

- Veja a função `gls` do pacote `nlme` do `R`.

Propriedades do estimador MQG (Σ_i conhecida)

- A primeira propriedade muito notável é que, para **qualquer escolha** de Σ_i, a estimativa MQG de β é **não viesada**. Ou seja,

$$\mathrm{E}\,[\hat{\beta}] = \beta.$$

- Além disso, **em amostras grandes** (ou assintoticamente), pode se mostrar que a distribuição amostral de $\hat{\beta}$ é uma distribuição normal multivariada com média β e covariância

$$\mathrm{Cov}\,(\hat{\beta}) = \left\{ \sum_{i=1}^{N} \sum_{j=1}^{n} (X_i' \Sigma_i^{-1} X_i) \right\}^{-1}.$$

Propriedades do estimador MQG (Σ_i conhecida)

- Isso é verdade exatamente quando Y_i tem uma distribuição condicional que é multivariada normal e verdadeira em **amostras grandes**, mesmo quando a distribuição condicional de Y_i não é normal multivariada.
  - Por "grandes amostras", entendemos que o tamanho da amostra, N, aumenta quando o número de medidas repetidas e parâmetros do modelo permanece fixo.
- Observe também que se Σ_i for assumido como uma matriz diagonal, com variância constante σ^2 ao longo da diagonal, o estimador MQG reduz para o estimador de mínimos quadrados ordinários considerado mais cedo.
  - **Exercício:** demonstre este resultado.
- Finalmente, embora o estimador MQG de β seja não viesado para qualquer escolha de Σ_i, pode ser mostrado que o estimador MQG mais eficiente de β (ou seja, o estimador com menor variância ou maior precisão) é aquele que usa o valor verdadeiro de Σ_i.
  - **Pergunta:** o que isto quer dizer?

MQG (Σ_i desconhecida)

- Vamos abordar o caso que geralmente ocorre na prática: não conhecemos Σ_i (ou θ).
  - Neste caso precisamos estimar Σ_i (θ) a partir dos dados disponíveis.
- A estimativa da máxima verossimilhança de θ prossegue da mesma maneira que a estimativa de β, maximizando a log-verossimilhança em relação a θ.
- Especificamente, a estimativa de probabilidade máxima de θ pode ser obtida igualando a derivada da log-verossimilhança em relação a θ (função escore) a zero e encontrando a solução para a equação resultante.
- Entretanto, em geral, essa **equação é não linear** e não é possível escrever expressões simples e de forma fechada para o estimador de MV de θ.
  - A estimativa de MV deve ser encontrada resolvendo-se essa equação usando uma técnica **iterativa**.

MQG (Σ_i desconhecida)

- Felizmente, algoritmos de computador foram desenvolvidos para encontrar a solução.
- Uma vez obtida a estimativa de MV de θ, simplesmente substituímos a estimativa de $\Sigma_i(\theta)$, digamos $\widehat{\Sigma}_i = \Sigma_i(\hat{\theta})$, no estimador de mínimos quadrados generalizados de β para obter a estimativa de MV de β:

$$\hat{\beta} = \left\{ \sum_{i=1}^{N} \sum_{j=1}^{n} (X_i' \widehat{\Sigma}_i^{-1} X_i) \right\}^{-1} \sum_{i=1}^{N} \sum_{j=1}^{n} (X_i' \widehat{\Sigma}_i^{-1} y_i).$$

Propriedades do estimador MQG (Σ_i conhecida)

- Curiosamente, **em amostras grandes** (ou assintoticamente), o estimador resultante de β que substitui a estimativa de MV de Σ_i tem todas as **mesmas propriedades** de quando Σ_i é realmente conhecido.
- Assim, em termos de propriedades da distribuição amostral de $\hat{\beta}$, não há penalidade por realmente ter que estimar Σ_i a partir dos dados longitudinais em questão.
- No entanto, por mais reconfortante que esse resultado possa parecer, deve-se ter em mente que esta é uma propriedade de grande amostra (ou seja, quando N se aproxima do infinito) de $\hat{\beta}$.
  - Com tamanhos de amostra da magnitude frequentemente encontrados em muitas áreas de aplicação, pode-se esperar que as propriedades da distribuição amostral de $\hat{\beta}$ sejam adversamente influenciadas pela estimativa de um número muito grande de parâmetros de covariância.

Questões de dados ausentes

Questões de dados ausentes

Embora a maioria dos estudos longitudinais seja projetada para coletar dados de cada indivíduo da amostra a cada momento do acompanhamento, muitos estudos têm algumas observações ausentes.

Dados ausentes têm três implicações importantes para a análise longitudinal:

1. O conjunto de dados é necessariamente desbalanceado ao longo do tempo, pois nem todos os indivíduos têm o mesmo número de medições repetidas em um conjunto comum de ocasiões.
   - Como resultado, os métodos de análise precisam ser capazes de lidar com os dados desequilibrados sem precisar descartar dados de indivíduos com dados ausentes.

Questões de dados ausentes

2. Haverá perda de informações e redução na precisão com que mudanças na resposta média ao longo do tempo pode ser estimado.
    - Essa redução na precisão está diretamente relacionada à quantidade de dados ausentes e também será influenciada em certa medida pela maneira como a análise lida com os dados ausentes.
    - Por exemplo, usar apenas os casos completos (ou seja, aqueles indivíduos sem dados ausentes) geralmente será o método menos eficiente.
3. A validade de qualquer método de análise exigirá que certas suposições sobre os motivos de qualquer falta, geralmente chamadas de **mecanismo de perda de dados**, são sustentáveis.
    - Consequentemente, quando faltam dados, devemos considerar cuidadosamente os motivos da falta.

Mecanismo de perda de dados

- O mecanismo de perda de dados pode ser pensado como **um modelo** que descreve a probabilidade de uma resposta ser observada ou ausente em qualquer ocasião.
- Fazemos uma distinção importante entre mecanismos de dados ausentes que são referidos como **ausentes completamente ao acaso** (*missing completely at random* - MCAR) e **ausentes ao acaso** (*missing at random* - MAR).
- A distinção entre esses dois mecanismos determina a adequação da estimativa de máxima verossimilhança sob o pressuposto de uma distribuição normal multivariada para as respostas e o MQG sem exigir suposições sobre o formato da distribuição.

MCAR

- Diz-se que os dados são MCAR quando a probabilidade de falta de respostas não está relacionada aos valores específicos que, em princípio, deveriam ter sido obtidos (as respostas ausentes) ou ao conjunto de respostas observadas.
- Ou seja, dados longitudinais são MCAR quando falta em Y_i é simplesmente o resultado de um mecanismo de chance que não depende de componentes observados ou não observados de Y_i.
- A característica essencial do MCAR é que os dados observados podem ser considerados uma amostra aleatória dos dados completos.
  - Como resultado, os momentos (por exemplo, médias, variâncias e covariâncias) e, de fato, a distribuição dos dados observados não diferem dos momentos correspondentes ou da distribuição dos dados completos.

MCAR

- Qualquer método de análise que produza inferências válidas na ausência de dados ausentes também produzirá inferências válidas quando os dados ausentes forem MCAR e a análise for baseada em todos os dados disponíveis, ou mesmo quando estiver restrito aos "completadores" (ou seja, aqueles sem dados ausentes).
- Dado que estimativas válidas das médias, variâncias e covariâncias podem ser obtidas, o MQG fornece estimativas válidas de β sem exigir nenhuma premissa de distribuição para Y_i.
- O estimador MQG de β é válido, desde que o modelo para a resposta média tenha sido especificado corretamente; não requer nenhuma suposição sobre a distribuição conjunta das respostas longitudinais.

MCAR

- O estimador de MV de β, pressupondo que as respostas tenham uma distribuição normal multivariada, também é o estimador MQG (com a estimativa MV de $\Sigma_i(\theta)$, por exemplo, $\widehat{\Sigma}_i = \Sigma_i(\hat{\theta})$, substituída).
- Assim, nessa configuração, os estimadores MV e MQG têm exatamente as mesmas propriedades, independentemente da verdadeira distribuição de Y_i.

MAR

- Ao contrário do MCAR, diz-se que os dados são MAR quando a probabilidade de falta de respostas depende do conjunto de respostas observadas, mas não está relacionada aos valores ausentes específicos que, em princípio, deveriam ter sido obtidos.
- Em outras palavras, se os indivíduos são estratificados com base em valores semelhantes para as respostas observadas, a falta é simplesmente o resultado de um mecanismo de chance que não depende dos valores das respostas não observadas.
- No entanto, como o mecanismo de falta agora depende das respostas observadas, a distribuição de Y_i em cada um dos estratos distintos definidos pelos padrões de falta não é o mesmo que a distribuição de Y_i na população alvo.
- Isso tem consequências importantes para a análise.

MAR

- Uma é que uma análise restrita aos "completadores" não é válida.
  - Em outras palavras, os "completadores" são uma amostra tendenciosa da população-alvo.
- Além disso, a distribuição dos componentes observados de Y_i, em cada um dos estratos distintos definidos pelos padrões de falta, não coincide com a distribuição dos mesmos componentes de Y_i na população alvo.
  - Portanto, as médias amostrais, variâncias e covariâncias com base nos "completadores" ou nos dados disponíveis são estimativas tendenciosas dos parâmetros correspondentes na população-alvo.

MAR

- Como resultado, o MQG não fornece mais estimativas válidas de β sem fazer suposições corretas sobre a distribuição conjunta das respostas longitudinais.
- Por outro lado, a estimativa de MV de β é válida quando os dados são MAR, desde que a distribuição normal multivariada foi especificada corretamente.
  - Isso requer a especificação correta não apenas do modelo para a resposta média, mas também do modelo para a covariância entre as respostas.
- Em certo sentido, a estimativa de MV permite que os valores ausentes sejam validamente "previstos" ou "imputados" usando os dados observados e um modelo correto para a distribuição conjunta das respostas.

Exercícios

Exercícios

1. (Re)fazer as provas dos resultados discutidos na aula de hoje.
2. Utilize os dados do estudo dos níveis de chumbo no sangue (TLC).
    - Proponha um modelo de regressão linear para a média com base nas questões de pesquisa.
    - Encontre as estimativas para os coeficientes de regressão do modelo proposto (use a função `gls` do pacote `nlme` do `R`; explore diferentes especificações da covariância - veja a documentação de `corClasses`).
    - **Com base nas estimativas**, faça a exposição das suas conclusões.

Avisos

- **Para casa:** ler o Capítulo 4 do livro "**Applied Longitudinal Analysis**".
  - Caso ainda não tenha lido, leia também os Caps. 1, 2 e 3.
  - Ver também as referências com respeito à derivadas de vetores, matrizes, etc.
- **Próxima aula:** Estimação e inferência estatística.

Bons estudos!