MAT02035 - Modelos para dados correlacionados

Visão geral de modelos lineares para dados longitudinais

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, 2021

Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

Introdução

Introdução

- Neste primeiro momento, o nosso foco será exclusivamente nos **modelos lineares** para **dados longitudinais** com **variáveis resposta contínuas** e **com distribuições aproximadamente simétricas**, **sem caudas excessivamente longas** (**ou assimetria**) **ou outliers**.
- Estes modelos fornecem as bases para modelos mais gerais para dados longitudinais quando a variável de resposta é discreta ou é uma contagem.
- Nesta aula apresentamos algumas notações de vetores e matrizes e apresentamos um modelo de regressão linear geral para dados longitudinais.

Introdução

- Nas próximas duas aulas:
  1. Apresentamos uma ampla visão geral de diferentes abordagens para modelar a resposta média ao longo do tempo e para contabilizar a correlação entre medidas repetidas no mesmo indivíduo.
  2. Consideramos alguns métodos descritivos elementares para explorar dados longitudinais, especialmente tendências na resposta média ao longo do tempo.
  3. Concluímos nossa discussão com uma pesquisa histórica de alguns dos primeiros desenvolvimentos em métodos para analisar dados de medidas longitudinais e repetidas.

Introdução

- Devemos enfatizar desde o início que os métodos estatísticos apresentados nesta primeira parte usam a suposição de que as respostas longitudinais têm uma **distribuição normal multivariada *aproximada*** para derivar estimativas e testes estatísticos, mas não exigem isso.

Normalidade $\rightsquigarrow$ Máxima verossimilhança $\rightsquigarrow$ Estimação intervalar $\rightsquigarrow$ Testes de hipóteses $\rightsquigarrow$ Avaliação da adequabilidade dos modelos.

Notação

Notação

- Assumimos que uma amostra de N indivíduos são medidos repetidamente ao longo do tempo.
- Denotamos Y_{ij} a variável resposta do i-ésimo indivíduo na j-ésima ocasião de medição.
- Como mencionado anteriormente, os indivíduos podem não ter o mesmo número de medidas e podem não ser medidos nas mesmas ocasiões.

- Para tal, utilizamos n_i para representar o número de medidas repetidas e t_{ij} os tempos de medida do i-ésimo indivíduo.
  - Se n é o número de **ocasiões planejadas** do estudo, então $n_i \leq n$.

O vetor de respostas

- É conveniente **agrupar** n_i medidas repetidas da variável resposta do i-ésimo indivíduo em um vetor $n_i \times 1$

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix}, \quad i = 1, \ldots, N.$$

O vetor de repostas

- Presume-se que os vetores de respostas Y_i, para os N indivíduos, sejam **independentes** um do outro.

Observe, no entanto, que embora os vetores de respostas obtidas em diferentes indivíduos possam geralmente ser considerados independentes uns dos outros (por exemplo, não se espera que medidas repetidas de um resultado de saúde para um paciente em um estudo clínico prevejam ou influenciem os resultados de saúde para outro paciente no mesmo estudo), as medidas repetidas sobre o mesmo indivíduo não são enfaticamente consideradas observações independentes.

Notação em dados balanceados

- Quando o número de medidas repetidas é o mesmo para todos os indivíduos do estudo (e não há dados ausentes), não é necessário incluir o índice i em n_i (já que $n_i = n$ para $i = 1, \ldots, N$).
- Da mesma forma, se as medidas repetidas forem observadas no mesmo conjunto de ocasiões, não é necessário incluir o índice i em t_{ij} (já que $t_{ij} = t_j$ para $i = 1, \ldots, N$).

Vetor de covariáveis

- Associado a cada resposta, Y_{ij}, há um vetor $p \times 1$ de covariáveis

$$X_{ij} = \begin{pmatrix} X_{ij1} \\ X_{ij2} \\ \vdots \\ X_{ijp} \end{pmatrix}, \; i = 1, \ldots, N, j = 1, \ldots, n_i.$$

- Observe que X_{ij} é um vetor de covariáveis associadas a Y_{ij}, a variável de resposta para o i-ésimo indivíduo na j-ésima ocasião.
- As p linhas de X_{ij} correspondem a **diferentes covariáveis**.

Vetor de covariáveis

- Existe um vetor correspondente de covariáveis associado a cada uma das n_i medidas repetidas no i-ésimo indivíduo.
  - X_{i1} é o vetor $p \times 1$ cujos elementos são os valores das covariáveis associadas à variável de resposta do i-ésimo indivíduo na 1ª ocasião de medição;
  - X_{i2} é o vetor $p \times 1$ cujos elementos são os valores das covariáveis associadas à variável de resposta do i-ésimo indivíduo na 2ª ocasião de medição e assim por diante.

Vetor de covariáveis

- O vetor X_{ij} pode incluir dois tipos principais de covariáveis: **(i)** covariáveis cujos valores não mudam ao longo da duração do estudo e **(ii)** covariáveis cujos valores mudam ao longo do tempo.

i. Exemplos do primeiro caso incluem tratamentos experimentais fixos.
ii. Exemplos do segundo caso incluem o tempo desde a linha de base, idade, o status atual do tabagismo e as exposições ambientais.

Vetor de covariáveis

- No primeiro caso, **os mesmos valores das covariáveis são replicados nas linhas** correspondentes de X_{ij} para $j = 1, \ldots, n_i$.
- Já para no segundo caso, os valores obtidos pelas covariáveis podem variar ao longo do tempo (para pelo menos alguns indivíduos) e os valores nas linhas correspondentes de X_{ij} podem ser diferentes a cada ocasião da medição.

Matriz de covariáveis

- Podemos **agrupar** os vetores de covariáveis em matrizes $n_i \times p$ de covariáveis:

$$X_i = \begin{pmatrix} X'_{i1} \\ X'_{i2} \\ \vdots \\ X'_{in_i} \end{pmatrix} = \begin{pmatrix} X_{i11} & X_{i12} & \cdots & X_{i1p} \\ X_{i21} & X_{i22} & \cdots & X_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{in_i1} & X_{in_i2} & \cdots & X_{in_ip} \end{pmatrix}, \ i = 1, \ldots, N.$$

- As linhas de X_i correspondem às covariáveis associadas às respostas nas n_i diferentes ocasiões de medição;
- As colunas de X_i correspondem às p covariáveis distintas.

Modelo de regressão linear

- Consideramos um modelo de regressão linear para alterações na resposta média ao longo do tempo e para relacionar estas mudanças às covariáveis,

$$Y_{ij} = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \ldots + \beta_p X_{ijp} + e_{ij}, j = 1, \ldots, n_i; \qquad (1)$$

em que $\beta_1, \ldots, \beta_p$ são **coeficientes de regressão desconhecidos** relacionando a média de Y_{ij} às suas correspondentes covariáveis.

- Este modelo descreve como as respostas em cada ocasião são relacionadas com as covariáveis.

Modelo de regressão linear

- Ou seja, há n_i equações de regressão separadas para a variável resposta em cada uma das n_i ocasiões

$$\begin{array}{rcccccl} Y_{i1} & = & \beta_1 X_{i11} + \beta_2 X_{i12} + \ldots + \beta_p X_{i1p} + e_{i1} & = & X_{i1}'\beta + e_{i1}, \\ Y_{i2} & = & \beta_1 X_{i21} + \beta_2 X_{i22} + \ldots + \beta_p X_{i2p} + e_{i2} & = & X_{i2}'\beta + e_{i2}, \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ Y_{in_i} & = & \beta_1 X_{in_i1} + \beta_2 X_{in_i2} + \ldots + \beta_p X_{in_ip} + e_{in_i} & = & X_{in_i}'\beta + e_{in_i}, \end{array} \tag{2}$$

em que $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ é um vetor $p \times 1$.

Modelo de regressão linear

- No modelo da Equação (1) os e_{ij} são erros aleatórios, com **média zero**, representando desvios das respostas a partir de suas respectivas médias preditas

$$\mathrm{E}(Y_{ij}|X_{ij}) = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \ldots + \beta_p X_{ijp}.$$

- Tipicamente, mas não sempre, $X_{ij1} = 1$ para todo i e j, e então β_1 é o termo de **intercepto** do modelo.
  - Não utilizaremos β_0 nem α.

Modelo de regressão linear

- Por fim, usando notação de vetor e matriz, o modelo de regressão dado pelas Equações (1) ou (2) pode ser expresso de uma forma ainda mais compacta,

$$Y_i = X_i\beta + e_i, \tag{3}$$

em que $e_i = (e_{i1}, e_{i2}, \ldots, e_{in_i})'$ é um vetor $n_i \times 1$ de **erros aleatórios**.

- O modelo de regressão dado pela Equação (3) é simplesmente uma representação abreviada para

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix} = \begin{pmatrix} X_{i11} & X_{i12} & \cdots & X_{i1p} \\ X_{i21} & X_{i22} & \cdots & X_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{in_i1} & X_{in_i2} & \cdots & X_{in_ip} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{in_i} \end{pmatrix}.$$

Exemplo: Tratamento de crianças expostas ao chumbo (estudo TLC)

Estudo TLC

- Lembre-se de que no **estudo sobre tratamento de crianças expostas ao chumbo**, há 100 participantes do estudo que têm níveis de chumbo no sangue medidos no mesmo conjunto de quatro ocasiões: linha de base (ou semana 0), semana 1, semana 4 e semana 6.
- Como todos os indivíduos tem o mesmo número de medidas repetidas observadas no mesmo conjunto de ocasiões, o índice i pode ser retirado de n_i e t_{ij}.
  - Ou seja, $n_1 = n_2 = \ldots = n_N = n$ e da mesma forma $t_{1j} = t_{2j} = \ldots = t_{Nj} = t_j$ para $j = 1, \ldots, 4$.
- No estudo TLC, o vetor de resposta tem comprimento 4 $(n = 4)$ e todos os indivíduos são medidos no mesmo conjunto de ocasiões: $t_1 = 0, t_2 = 1, t_3 = 4$ **e** $t_4 = 6$.

Estudo TLC

- Suponha que seja interessante ajustar um modelo à resposta média que pressupõe que o nível médio de chumbo no sangue mude linearmente ao longo do tempo, mas a uma taxa que pode ser diferente para os dois grupos de tratamento.
- Em particular, podemos querer ajustar um modelo em que **os dois grupos de tratamento tenham o mesmo intercepto** (ou sejam, a mesma resposta média na linha de base), mas inclinações diferentes.
  - Isso pode ser representado no seguinte modelo de regressão

$$\begin{aligned} Y_{ij} &= \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} + e_{ij} \\ &= X'_{ij}\beta + e_{ij}, \end{aligned}$$

em que $X_{ij1} = 1$ para todo i e j **(β_1 é um termo de intercepto)**.

Estudo TLC

- A segunda covariável, $X_{ij2} = t_j$, representa a semana em que o nível de chumbo no sangue foi obtido.
- Por fim, $X_{ij3} = t_j \times$ Grupo$_i$, em que Grupo$_i = 1$ se o i-ésimo indivíduo é designado ao grupo succimer e Grupo$_i = 0$ se o i-ésimo indivíduo é designado ao grupo placebo.

Estudo TLC

- Essa codificação de X_{ij2} e X_{ij3} permite que as inclinações do tempo sejam diferentes para os dois grupos de tratamento.
- As três covariáveis podem ser agrupadas em um vetor 3×1 das covariáveis X_{ij}.
- Assim, para crianças do grupo placebo

$$\mathrm{E}(Y_{ij}|X_{ij}) = \beta_1 + \beta_2 t_j.$$

- β_1 representa o nível de chumbo no sangue médio na linha de base (semana 0);
- β_2 tem interpretação como uma mudança no nível médio de chumbo no sangue (em $\mu g/dL$) por semana.

Estudo TLC

- Similarmente para as crianças no grupo succimer

$$\mathrm{E}(Y_{ij}|X_{ij}) = \beta_1 + (\beta_2 + \beta_3)t_j.$$

- β_1 representa o nível de chumbo no sangue médio na linha de base (assumido ser o mesmo como no grupo placebo, pois o ensaio aleatorizou indivíduos para dois grupos);
- $\beta_2 + \beta_3$ tem interpretação como uma mudança no nível médio de chumbo no sangue (em $\mu g/dL$) por semana.

Assim, se os dois grupos de tratamentos diferem em suas taxas de declínio nos níveis de chumbo no sangue, então $\beta_3 \neq 0$.

Estudo TLC

- Os parâmetros de regressão têm interpretações úteis que se relacionam diretamente com questões de interesse científico.
- Além disso, hipóteses de interesse podem ser expressas em termos da ausência de certos parâmetros de regressão.
- Por exemplo, a hipótese de que os dois tratamentos são igualmente eficazes na redução dos níveis de chumbo no sangue corresponde a uma hipótese que $\beta_3 = 0$.

Estudo TLC

- Os valores das respostas para os indivíduos 79 e 8 são apresentados

$$y_{79} = \begin{pmatrix} 30.8 \\ 26.9 \\ 25.8 \\ 23.8 \end{pmatrix} \text{ e } y_{8} = \begin{pmatrix} 26.5 \\ 14.8 \\ 19.5 \\ 21.0 \end{pmatrix}.$$

Estudo TLC

- Associados aos vetores de repostas, temos as matrizes de covariáveis

$$X_{79} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 4 & 0 \\ 1 & 6 & 0 \end{pmatrix} \text{ e } X_8 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 4 & 4 \\ 1 & 6 & 6 \end{pmatrix}.$$

Estudo TLC

- O modelo para a média dos níveis de chumbo no sangue pode ser representado

$$\mathrm{E}(Y_i|X_i) = X_i\beta,$$

$$\mathrm{E}(Y_i|X_i) = \begin{pmatrix} \mathrm{E}(Y_{i1}|X_{i1}) \\ \mathrm{E}(Y_{i2}|X_{i2}) \\ \mathrm{E}(Y_{i3}|X_{i3}) \\ \mathrm{E}(Y_{i4}|X_{i4}) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 4 & 0 \\ 1 & 6 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_1 + \beta_2 \\ \beta_1 + 4\beta_2 \\ \beta_1 + 6\beta_2 \end{pmatrix}$$

para crianças no grupo placebo, e

$$\mathrm{E}(Y_i|X_i) = \begin{pmatrix} \mathrm{E}(Y_{i1}|X_{i1}) \\ \mathrm{E}(Y_{i2}|X_{i2}) \\ \mathrm{E}(Y_{i3}|X_{i3}) \\ \mathrm{E}(Y_{i4}|X_{i4}) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 4 & 4 \\ 1 & 6 & 6 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_1 + (\beta_2 + \beta_3) \\ \beta_1 + 4(\beta_2 + \beta_3) \\ \beta_1 + 6(\beta_2 + \beta_3) \end{pmatrix}$$

para crianças no grupo succimer.

Suposições distribucionais

Suposições distribucionais

- Até agora, as únicas suposições feitas diziam respeito a padrões de mudança na resposta média ao longo do tempo e sua relação com covariáveis.
- Especificamente, dado que o vetor de erros aleatórios, e_i, é assumido como tendo média zero, o modelo de regressão dado por (3) implica que

$$\mathrm{E}(Y_i|X_i) = \mu_i = X_i\beta, \tag{4}$$

em que $\mu_i = (\mu_{i1}, \ldots, \mu_{in_i})'$ é o vetor $n_i \times 1$ de médias condicionais para o i-ésimo indivíduo, com $\mu_{ij} = \mathrm{E}(Y_{ij}|X_i) = \mathrm{E}(Y_{ij}|X_{ij})$.

Componente sistemático e distribuição dos erros

- Em seguida, consideramos as suposições distribucionais relativas ao vetor de erros aleatórios, e_i.
- O vetor de resposta Y_i em (3) é assumido como sendo composto por dois componentes:
  1. um "componente sistemático", $X_i\beta$;
  2. um "componente aleatório", e_i.
- A variabilidade aleatória de Y_i decorre da adição de e_i.
  - Isso implica que suposições feitas sobre a forma da distribuição dos erros aleatórios se traduzem em suposições sobre a forma da **distribuição condicional** de Y_i dado X_i.
  - Podemos quase indistintamente nos referir à distribuição dos erros, e_i, ou das respostas, Y_i; suas respectivas distribuições diferem apenas em termos de mudança de locação.

Distribuição condicional da resposta

- Y_i, o vetor de respostas contínuas, é assumido como tendo uma distribuição condicional que é **normal multivariada**, com vetor de resposta médio

$$\mathrm{E}(Y_i|X_i) = \mu_i = X_i\beta, \text{ e matriz de covariância } \Sigma_i = \mathrm{Cov}(Y_i|X_i).$$

Normal multivariada

- É completamente especificada pelo vetor de médias, μ_i, e a matriz de covariância,Σ_i.
- Pode ser considerada o análogo multivariado da distribuição normal univariada.
  - Se Y_i tem uma distribuição condicional que é normal multivariada, então cada um de seus componentes, Y_{ij}, tem uma distribuição normal univariada correspondente, com média condicional μ_{ij} e variância condicional σ_j^2.

A matriz de covariância

- Lembre que, embora as observações de diferentes indivíduos sejam consideradas independentes umas das outras, as medidas repetidas do mesmo indivíduo não são consideradas independentes.
  - Essa falta de independência é capturada pelos elementos fora da diagonal da matriz de covariância Σ_i.
- A matriz de covariância foi indexada por i, e isso permite, em princípio, que a matriz de covariância dependa das covariáveis, X_i (por exemplo, nos tempos das medidas repetidas).
- Quando $n_i = n$ e $t_{ij} = t_j$ $(i = 1, \ldots, N)$, e onde não há dependência da matriz de covariância nas covariáveis, podemos descartar o índice i e simplesmente denotar a matriz de covariância por Σ.
  - Análogo à suposição de homogeneidade de variância na regressão linear para uma resposta univariada, ou seja, para o vetor de respostas, supõe-se que haja homogeneidade de covariância.

A matriz de covariância

- No entanto, quando os indivíduos têm números desiguais de medidas repetidas e/ou quando as medidas repetidas são obtidas em ocasiões diferentes, a matriz de covariância normalmente dependerá do número e do tempo das medidas.
  - Em princípio, a covariância também pode depender de outras covariáveis além do tempo; por exemplo, a covariância pode depender do grupo de tratamento.
  - No entanto, na prática, esse tipo de dependência da covariância nas covariáveis raramente é assumido.

A distribuição normal multivariada

Distribuições de probabilidade

- A base formal para muitos métodos estatísticos é uma distribuição de probabilidade assumida para a variável resposta.
- Em termos gerais, uma distribuição de probabilidade descreve a frequência relativa de ocorrência de valores particulares da variável resposta.
- Em particular, a função de densidade de probabilidade para Y, denotada por $f(y)$, descreve a frequência relativa de ocorrência de valores particulares de Y.
- Antes de descrever algumas das propriedades da distribuição normal multivariada, primeiro revisamos a distribuição normal univariada.

A distribuição normal univariada

- Considere uma única resposta univariada de um estudo longitudinal em uma ocasião particular, digamos Y_{ij}.
- Assumimos que a média de Y_{ij} está relacionada com as covariáveis pelo seguinte modelo de regressão linear:

$$Y_{ij} = X'_{ij}\beta + e_{ij},$$

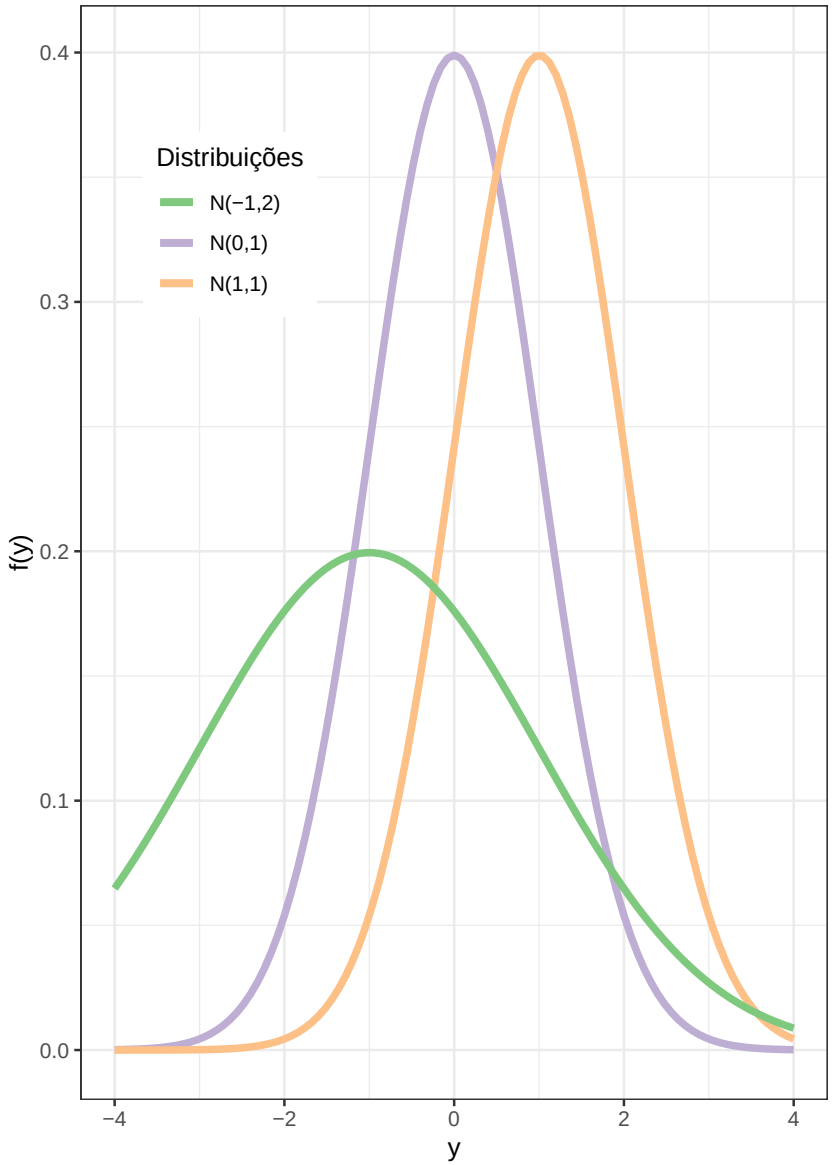
em que $e_{ij} \sim N(0, \sigma_j^2)$.

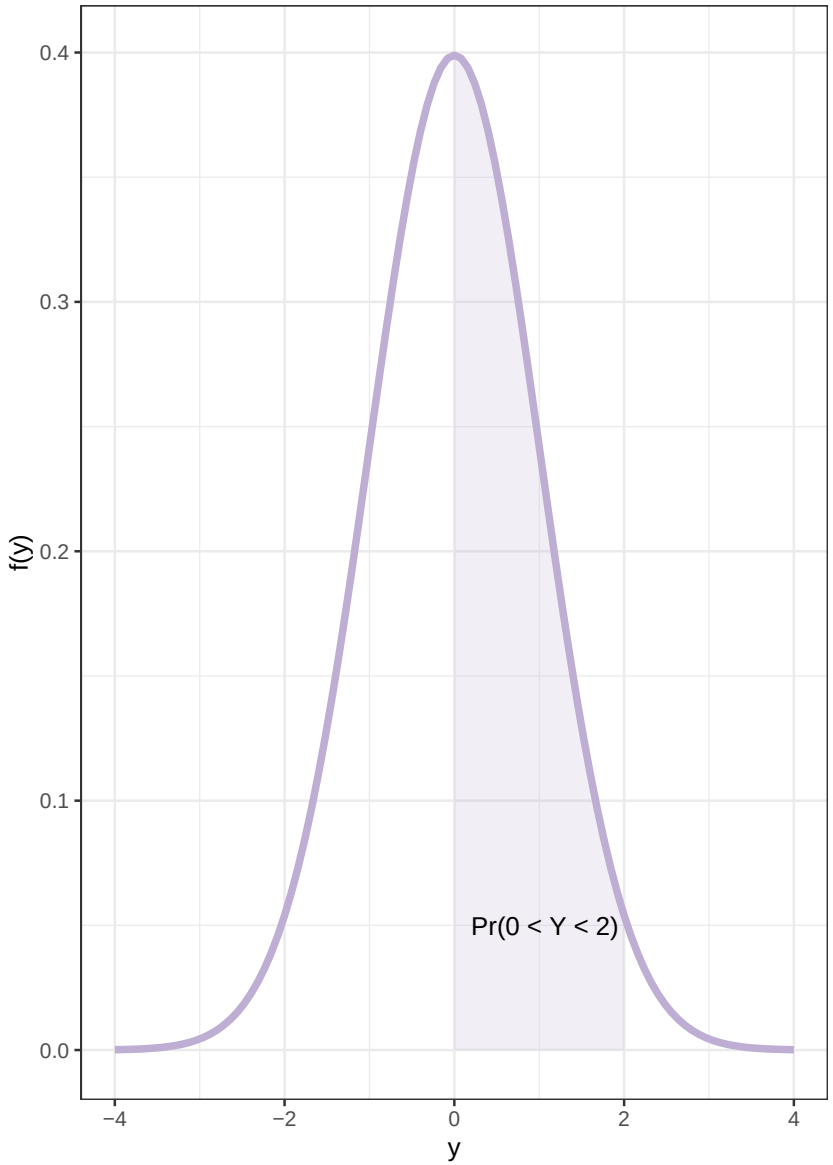
- Isto implica que a distribuição condicional de Y_{ij} (dado as covariáveis) também é normal, porém com média $\mu_{ij} = X'_{ij}\beta$ (e variância constante σ_j^2). Ou seja,

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

em que $-\infty < y < \infty$.

A distribuição normal univariada





A distribuição normal univariada

- Observe também que a expressão para a densidade de probabilidade normal depende em grande medida de

$$\frac{(y_{ij} - \mu_{ij})^2}{\sigma_j^2} = (y_{ij} - \mu_{ij})(\sigma_j^2)^{-1}(y_{ij} - \mu_{ij}).$$

- Esta é a distância ao quadrado entre y_{ij} e μ_{ij}, mas expressa em unidades de desvio padrão.
  - Assim, pode ser interpretado como a distância padronizada de y_{ij} para sua média condicional, em relação à variabilidade ou dispersão de valores em torno da média condicional, μ_{ij}.

A distribuição normal multivariada

- No contexto de um estudo longitudinal, com n_i medidas repetidas no i-ésimo indivíduo, temos um vetor de respostas e precisamos considerar sua **distribuição de probabilidade conjunta**.
- Enquanto uma função de densidade de probabilidade univariada descreve a probabilidade ou frequência relativa de ocorrência de valores particulares de uma única variável aleatória, uma função de densidade de probabilidade conjunta descreve a probabilidade ou frequência relativa com a qual o vetor de respostas assume um determinado conjunto de valores.

A distribuição normal multivariada

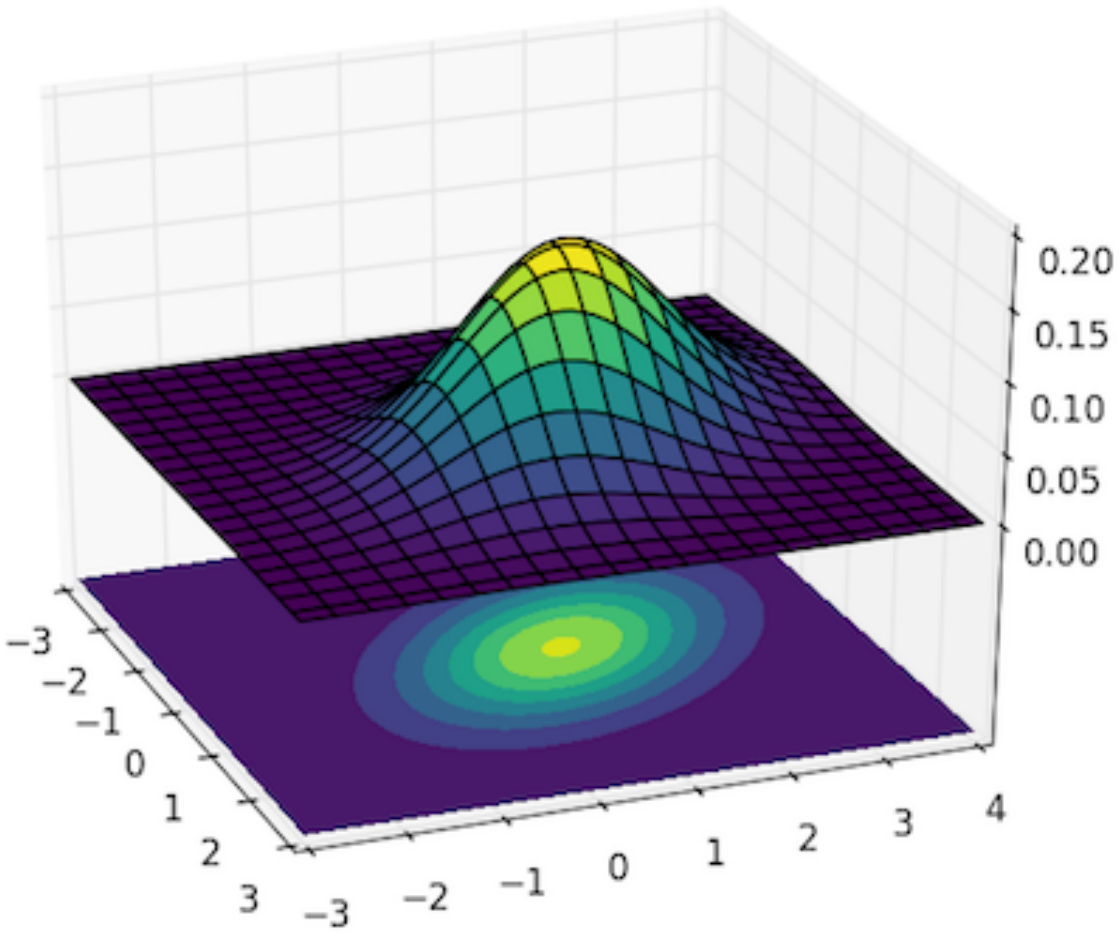
- A distribuição normal multivariada é uma generalização natural da distribuição normal univariada.
- A função densidade de probabilidade conjunta normal multivariada para Y_i dado X_i pode ser expressa como

$$\begin{aligned} f(y_i) &= f(y_{i1}, y_{i2}, \ldots, y_{in_i}) \\ &= (2\pi)^{-n_i/2} |\Sigma_i|^{-1/2} \exp\left\{ -\frac{1}{2}(y_i - \mu_i)' \Sigma_i^{-1} (y_i - \mu_i) \right\}, \end{aligned}$$

em que $-\infty < y_{ij} < \infty$ para $j = 1, \ldots, n_i$,
$\mu_i = \mathrm{E}(Y_i|X_i) = (\mu_{i1}, \ldots, \mu_{in_i})'$, $\Sigma_i = \mathrm{Cov}(Y_i|X_i)$ e $|\Sigma_i|$ denota o **determinante** de Σ_i.

A distribuicão normal multivariada

Exemplo de uma normal bivariada



A distribuição normal multivariada

- A distribuição normal multivariada também e completamente determinada pelo vetor de respostas médias, μ_i, e pela matriz de covariância Σ_i.
- O determinante de Σ_i também é conhecido como a **variância generalizada**.
  - Resume as características salientes da variação expressa por Σ_i em um único número.
- A falta de independência entre as medidas repetidas do vetor y_i é contemplada pelos elementos fora da diagonal principal da matriz de covariância Σ_i.

A distribuição normal multivariada

Semelhança entre $f(y_{ij})$ e $f(y_i)$

Em certo sentido, a função de densidade de probabilidade conjunta normal multivariada simplesmente substitui a expressão para a distância padronizada entre y_{ij} e μ_{ij},

$$(y_{ij} - \mu_{ij})(\sigma_j^2)^{-1}(y_{ij} - \mu_{ij}),$$

com um análogo multivariado para a distância padronizada do vetor y_i para μ_i,

$$(y_i - \mu_i)'\Sigma_i^{-1}(y_i - \mu_i),$$

em que Σ_i^{-1} denota a inversa da matriz Σ_i.

Comentários

- A suposição de normalidade multivariada é muito mais difícil de ser verificada a partir dos dados.
- Talvez a avaliação mais útil da validade da suposição de normalidade multivariada seja através do uso de gráficos.
  - Histogramas e diagramas de caixa (*boxplot*) dos **resíduos em cada ocasião** podem ser usados para **detectar grandes desvios de e_{ij} da normalidade univariada**.
  - Lembrando que $e_i \sim N(0, \Sigma_i)$ implica $e_{ij} \sim N(0, \sigma_j^2), j = 1, \ldots, n_i$; mas o contrário não é verdadeiro.

Comentários

- Outra propriedade da distribuição normal multivariada para Y_i dado X_i é que a associação entre qualquer par de respostas é linear.
- Consequentemente, se a distribuição condicional de Y_i é normal multivariada, então gráficos de dispersão dos **resíduos** em todos os pares de ocasiões possíveis **não devem fornecer nenhuma evidência de desvios discerníveis de uma tendência linear** entre os pares de variáveis.
  - Mais uma vez, uma ressalva desta técnica gráfica simples é que ela não pode ser usada para estabelecer que a distribuição condicional de Y_i é normal multivariada; ele só pode fornecer evidências de desvios discerníveis da normalidade multivariada.

Comentários

- A suposição de normalidade multivariada não é crucial para **estimação** e validade das inferências com respeito a β quando os dados são completos (sem ausência de dados).
  - Além disso, essa propriedade se estende à configuração de dados incompletos se os dados observados puderem ser considerados como uma amostra aleatória dos dados completos.
- Os desvios da normalidade, a menos que sejam muito extremos (por exemplo, dados de resposta altamente assimétricos), não são tão críticos.
- No cenário de dados longitudinais existem resultados muito semelhantes, que sugerem que são as suposições sobre a dependência entre os erros e as suposições sobre as variâncias e covariâncias que têm maior impacto na inferência estatística.
  - Desvios da normalidade multivariada, a menos que sejam muito extremas, não são tão críticos.
- Na prática, não se espera que os dados longitudinais tenham uma distribuição conjunta que seja **exatamente** normal multivariada.
  - A distribuição normal multivariada é adotada como uma **aproximação**, mas possui muitas propriedades estatísticas convenientes.

Avisos

- **Para casa:** ler o Capítulo 3 do livro "**Applied Longitudinal Analysis**". Caso ainda não tenha lido, leia também os Caps. 1 e 2.
- **Próxima aula:** Métodos de análise descritiva para dados longitudinais.

Bons estudos!