

Análise de Dados Longitudinais

Análise Exploratória

Enrico A. Colosimo/UFGM

<http://www.est.ufmg.br/~enricoc/>

1 Estrutura da Média

- Medidas Resumo;
- Perfis individuais.
- Perfil típico: Técnicas de suavização.

2 Estrutura de Variância e Covariância

- Perfis individuais.
- Gráficos de dispersão.
- Correlograma/variograma.

Explorando Dados Longitudinais

- 1 Explore o máximo possível os dados brutos antes de utilizar medidas resumo.
- 2 Focalize nos objetivos do estudo.
- 3 Identifique padrões transversal e longitudinal.
- 4 Identifique valores influentes/atípicos.

Explorando a Média

- 1 Comportamento temporal é linear?
- 2 Existe diferença entre grupos?
- 3 Existe indicação de interação entre grupo e tempo?

- ❶ Sem Estruturar (tempo discreto/ignora escala temporal contínua)
 - Apropriado para desenhos balanceados com poucos tempos.
 - Mais robusto pois assume os tempos como discreto.
- ❷ Estruturando a Forma Funcional da Média
 - Descreve o comportamento longitudinal da média (linear, polinomial, exponencial, etc).
 - Usualmente utiliza menos parâmetros que o não estruturado.
 - Sujeito a vício.

- 1 Splines
- 2 Núcleo Estimador (Kernel)
- 3 ****Lowess**** (Núcleo Robusto)

- 1 Modelos Polinomiais por Partes
- 2 Caso mais simples: linear com um único knot: dois componentes lineares.

$$E(Y_{ij}) = \beta_1 + \beta_2 t_{ij} + \beta_3 (t_{ij} - t^*)_+$$

em que $(t_{ij} - t^*)_+ = t_{ij} - t^*$, se $t_{ij} \geq t^*$ e 0, caso contrário.

- 3 Extensões: vários knots e estrutura polinomial.
- 4 Dificuldade: definir o **número** (importante) e localização dos knots.
- 5 Utilizado também para estruturar a forma funcional da média (caso simples).

Núcleo Estimador

- 1 Estimador local em uma janela.
- 2 Tomar uma média em uma janela de possíveis valores.

$$Y_t = \mu(t) + \epsilon_t.$$

- 3 Média ponderada: 1, dentro da janela e 0, fora da janela.
Extensão é tomar um núcleo, tipo gaussiano, que diminui o peso a medida que afastamos do valor t .
- 4 Dificuldade: definir o tamanho da janela.
- 5 Limitações: sensível a pontos influentes e problema de fronteira.

Alisamento - Lowess

- 1 LOWESS: LOcally WEighted Sum of Square.
- 2 É uma regressão linear/quadrática local ponderada.
- 3 Peso diminui à medida que a distância aumenta.
- 4 Valor do lowess no tempo t é igual ao valor predito da regressão \hat{Y}_t .
- 5 Largura da janela indica a fração dos dados a ser utilizada na regressão.
- 6 Quanto maior a janela, maior o alisamento.
- 7 Tamanho usual da janela $\alpha = (\lambda + 1)/n$, λ é o grau do polinômio.
- 8 Método de alisamento menos sensível a outliers e a problemas de fronteira (outliers são subponderados).

Explorando a Estrutura de Covariância

- 1 Um modelo para a estrutura de covariância deve ser escolhido com base no modelo para a média.
- 2 Utilizar os resíduos para remover o efeito de covariáveis.
- 3 Uma forma alternativa é utilizar variável resposta padronizada por tempo. Ou seja,

$$Z_{ij} = (Y_{ij} - \bar{Y}_j) / s_j.$$

Somente válido para estudos balanceados.

- 4 Matriz de correlação empírica para os resíduos: indica uma possível estrutura de covariância.

Explorando a Estrutura de Covariância

- 1 **IMPORTANTE:** existe uma interdependência entre a resposta média e a estrutura de covariância.
- 2 **Fato:** variâncias e covariâncias são inflacionadas quando a estrutura para a média é incorretamente especificada.
- 3 **Razão:** covariância entre pares de resíduos ($Y_{ij} - \mu_{ij}$) e ($Y_{il} - \mu_{il}$) depende do modelo para a média.
- 4 **Modelo** para a covariância deve então ser escolhido com base no modelo para a média.
- 5 **A decisão** deve ser feita com base nos resíduos.

Modelos para a Estrutura de Covariância

- 1 Não Estruturado: somente é adequada para desenhos balanceados com poucos tempos. Número de parâmetros: $(k + 1)k/2$, k é número de observações por unidade.
- 2 Estruturando a Covariância: simetria composta, AR(1), etc. Usualmente adequada para desenhos balanceados com poucos tempos.
- 3 Modelos de Efeitos Aleatórios.

Propostas de Modelagem para a Estrutura de Covariância

- 1 Utilizar o modelo "maximal" para a média e testar diferentes estruturas para a covariância.
- 2 Existem algumas estruturas disponíveis para desenhos balanceados com poucos tempos de observação.
- 3 Estrutura Exponencial para desenhos desbalanceados.

$$\text{Cov}(Y_{ij}, Y_{il}) = \sigma^2 \rho^{|t_{ij} - t_{il}|}, \quad 0 < \rho < 1.$$

Variograma

- 1 Útil para delineamentos desbalanceados.
- 2 Uma extensão da função de auto-correlação. Ou seja,

$$\text{Cor}(Y_{ij}, Y_{ij-u}), \quad u = 1, \dots, k$$

- 3 O variograma é definido como:

$$\gamma(u) = 0.5E\{(Y(t) - Y(t-u))^2\} = \sigma^2(1 - \rho(u)), \quad u \geq 0.$$

- 4 a função de auto-correlação

$$\rho(u) = 1 - \frac{\gamma(u)}{\sigma^2}$$

Ou seja, aumentando o variograma, diminui a correlação no tempo.

- 5 Construção: fazer o gráfico alisado de $v_{ijk} = 0.5(r_{ij} - r_{ik})^2$ vs $u_{ijk} = t_{ij} - t_{ik}$.

Exemplos

1 Níveis de chumbo no sangue (FLW, 2011)

- balanceado;
- poucos tempos (4).

2 Transmissão Vertical do HIV: Comprimento de recém-nascidos

- desbalanceado;
- muitos tempos.

Exemplo: Níveis de chumbo no sangue (FLW, 2011)

1 Descrição

- Tratamento de Crianças Expostas ao chumbo
- Estudo clínico aleatorizado para placebo e um tratamento em criança com níveis de chumbo no sangue entre 20-44 micrograms/dL.
- Quatro medidas repetidas de níveis de chumbo na linha de base semana 0, semana 1, semana 4 e semana 6
- 100 crianças aleatoriamente alocadas entre tratamento e placebo.

2 Banco de Dados Largo com as seguintes colunas:

- ID,
- Grupo,
- Week0,
- Week1,
- Week4,
- Week6.

Tabela : Estatísticas descritivas do Nível de chumbo, por semana.

Semana	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
0	19.70	22.05	25.60	26.41	29.60	41.10
1	2.80	12.38	20.60	19.09	25.02	40.80
4	3.00	15.25	19.70	19.79	24.60	40.40
6	4.10	18.25	21.25	22.20	25.60	63.90

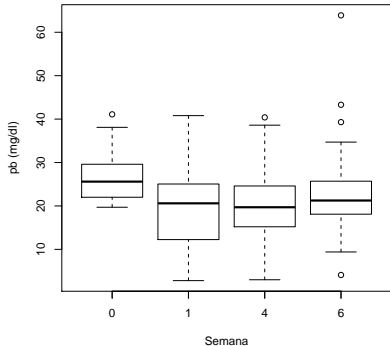
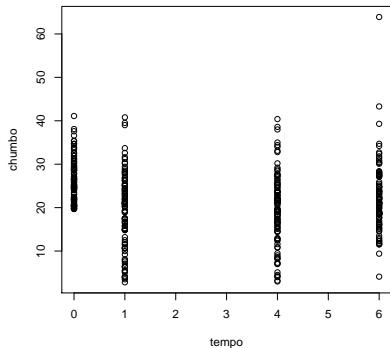


Gráfico Alternativo do Nível de chumbo, por Semana.



Obs. Estes gráficos devem ser avaliados com cuidado pois desconsideram os grupos.

Tabela : Estatísticas descritivas do Nível de chumbo, somente por Grupo.

Grupo	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
A	2.8	12.38	19.15	19.08	24.6	63.9
P	13.5	20.67	23.90	24.66	27.9	43.3

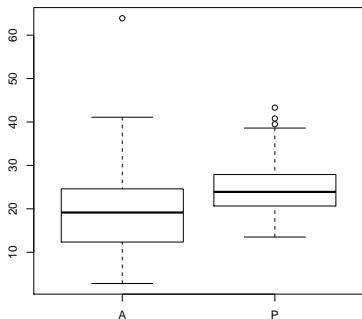
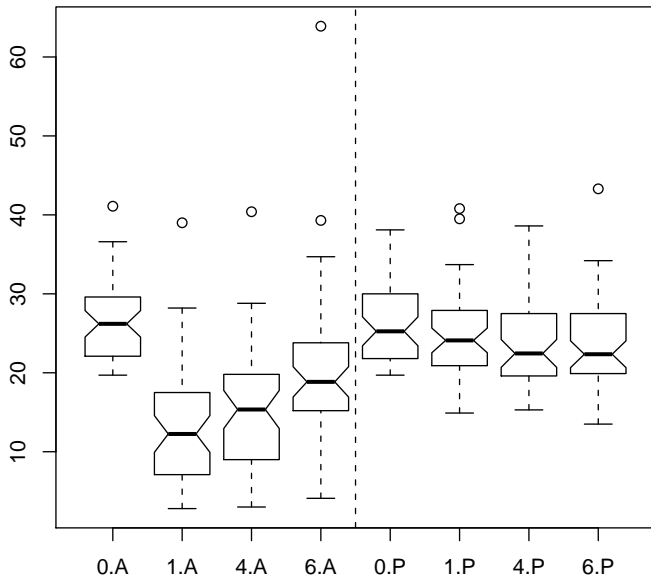
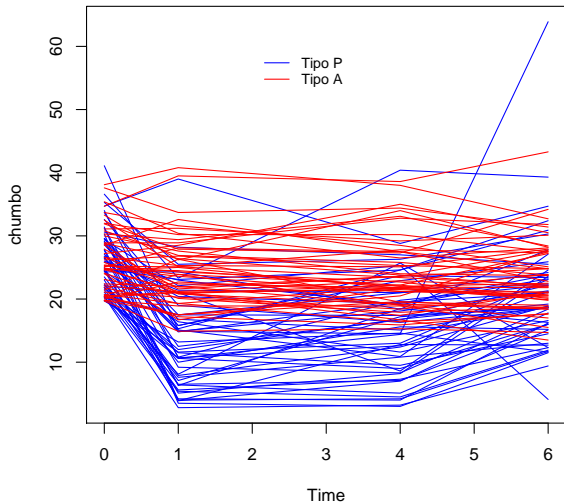


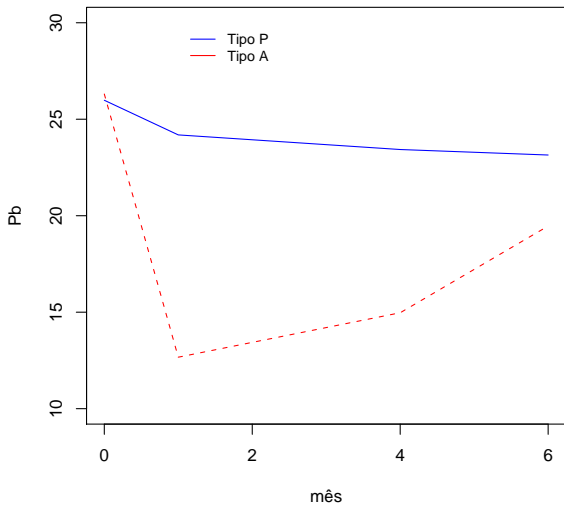
Tabela : Estatísticas descritivas do Nível de chumbo, por Grupo e Semana.

	Semana	Min	1º Quartil	Mediana	Média	3º Quartil	Max
Grupo A	0	19.7	22.13	26.20	26.54	29.55	41.1
	1	2.8	7.23	12.25	13.52	17.50	39.0
	4	3.0	9.13	15.35	15.51	19.73	40.4
	6	4.1	15.40	18.85	20.76	23.75	63.9
Grupo P	0	19.7	21.88	25.25	26.27	29.73	38.1
	1	14.9	20.92	24.10	24.66	27.82	40.8
	4	15.3	19.83	22.45	24.07	27.45	38.6
	6	13.5	19.95	22.35	23.65	27.50	43.3

Dispersão do chumbo por grupo







- Das tabelas e figura anteriores, observa-se que, para o grupo de crianças que tomaram Placebo, os níveis de chumbo praticamente permanecem inalterados ao longo do estudo.
- Já para o grupo de crianças submetidas ao Agente, os níveis de chumbo se reduzem bastante da semana 0 para a semana 1, e a partir daí têm um ligeiro aumento até a semana 4 e a semana 6.
- Excetuando-se os níveis obtidos na baseline (semana 0), que foram praticamente idênticos para ambos os grupos, os níveis de chumbo do grupo Tratamento são inferiores aos níveis do grupo Placebo;

Propostas de Modelo para a Média

1 Não Estruturada: quatro tempos discretos.

- Modelo 1

$$\begin{aligned} E(Y_{ij}) &= \beta_1 + \beta_2 I(t_j = 2) + \beta_3 I(t_j = 3) + \beta_4 I(t_j = 4) \\ &+ \beta_5 I(\text{grupo}_i = A) + \beta_6 I(t_j = 2) * I(\text{grupo}_i = A) \\ &+ \beta_7 I(t_j = 3) * I(\text{grupo}_i = A) + \beta_8 I(t_j = 4) * I(\text{grupo}_i = A) \end{aligned}$$

- Modelo 2

$$\begin{aligned} E(Y_{ij}) &= \beta_1 I(t_j = 1) + \beta_2 I(t_j = 2) + \beta_3 I(t_j = 3) + \beta_4 I(t_j = 4) \\ &+ \beta_5 I(t_j = 1) * I(\text{grupo}_i = A) + \beta_6 I(t_j = 2) * I(\text{grupo}_i = A) \\ &+ \beta_7 I(t_j = 3) * I(\text{grupo}_i = A) + \beta_8 I(t_j = 4) * I(\text{grupo}_i = A) \end{aligned}$$

Propostas de Modelo para a Média

2 Estruturada: Spline com knot no tempo = 1 semana:

$$\begin{aligned} E(Y_{ij}) &= \beta_1 + \beta_2 t_j + \beta_3 (t_j - 1)_+ + \beta_4 I(\text{grupo}_i = A) t_j \\ &+ \beta_5 I(\text{grupo}_i = A) (t_j - 1)_+ \end{aligned}$$

- Grupo Placebo

$$E(Y_{ij}) = \beta_1 + \beta_2 t_j + \beta_3 (t_j - 1)_+$$

- Grupo Tratado

$$E(Y_{ij}) = \beta_1 + (\beta_2 + \beta_4) t_j + (\beta_3 + \beta_5) (t_j - 1)_+.$$

- Interesse: $H_0 : \beta_4 = \beta_5 = 0$ (não existe interação grupo*tempo).

Tratando os valores da linha de base.

Em um estudo como este, aleatorizado, as medidas de linha de base (tempo 0) são obtidas antes de aplicação das intervenções. Neste caso, existem formas de tratar a linha de base.

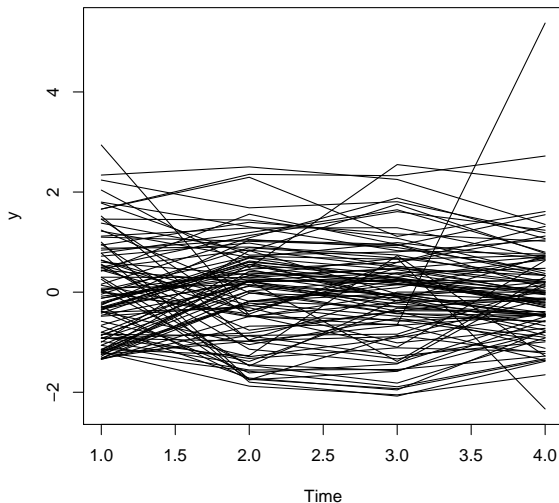
- Manter as respostas, sem fazer menção a diferença entre grupos.
- Manter as respostas e assumir médias iguais na linha de base (assumida nos modelos anteriores).
- Subtrair a resposta de linha de base das demais e analisar as diferenças.
- Usar a linha de base como covariável na análise dos dados.

Explorando a Estrutura de Correlação (Ignorando efeito de tratamento)- Usando dados brutos - Usualmente não é adequado.

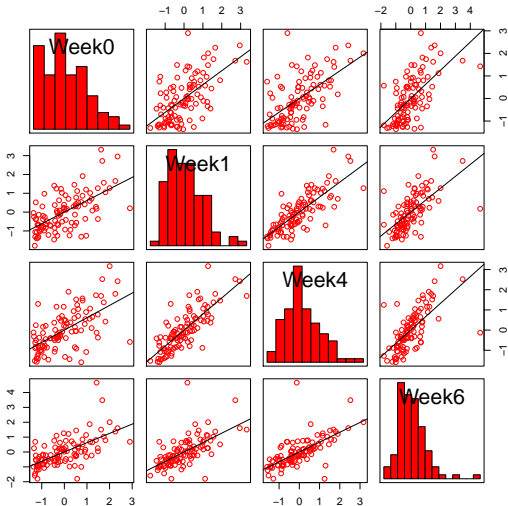
	Week0	Week1	Week4	Week6
Week0	1.00			
Week1	0.42	1.00		
Week4	0.47	0.84	1.00	
Week6	0.56	0.56	0.58	1.00

Perfil Padronizado - Ignorando Estrato por Tratamiento

Perfil Individual Padronizado



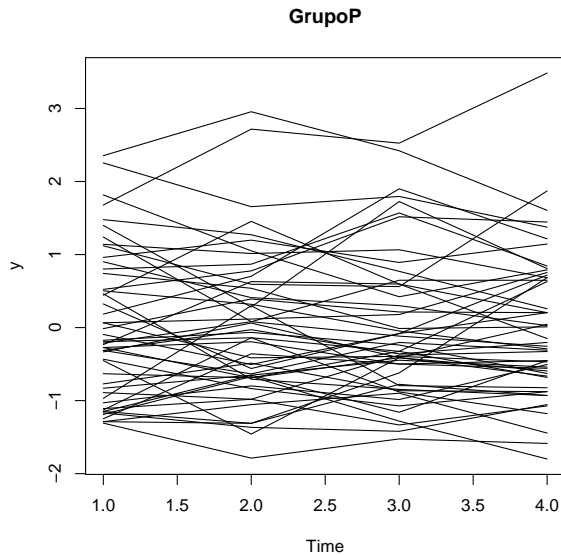
- Explorando a Estrutura de Dependência dos Tempos (Níveis de chumbo padronizados e Ignorando Efeito de Tratamento):



Matriz de correlação entre os Níveis de chumbo padronizados por tempo (Ignorando Efeito de Tratamento):

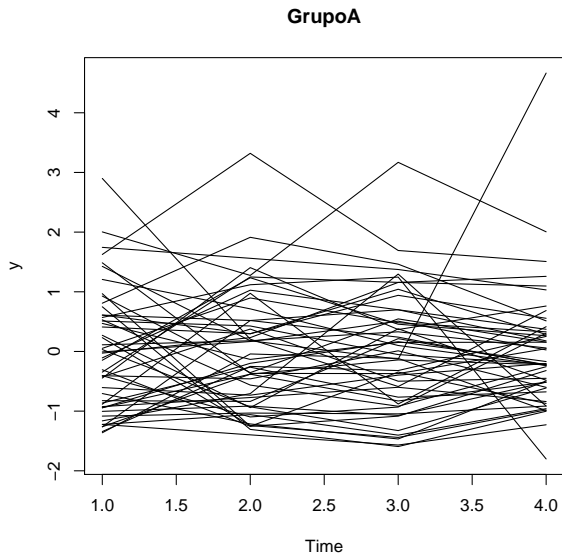
	Week0	Week1	Week4	Week6
Week0	1.00			
Week1	0.62	1.00		
Week4	0.61	0.80	1.00	
Week6	0.63	0.63	0.66	1.00

GRUPO - PLACEBO



Matriz de correlação para Grupo Placebo:

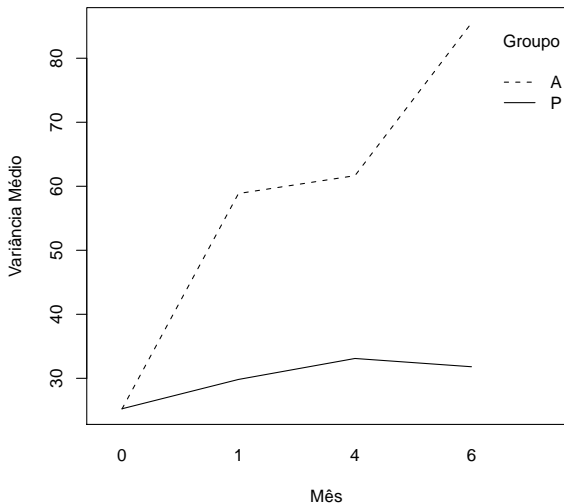
	Week0	Week1	Week4	Week6
Week0	1.00			
Week1	0.83	1.00		
Week4	0.84	0.86	1.00	
Week6	0.76	0.76	0.87	1.00



Matriz de correlação para Grupo A:

	Week0	Week1	Week4	Week6
Week0	1.00			
Week1	0.40	1.00		
Week4	0.38	0.73	1.00	
Week6	0.50	0.50	0.45	1.00

- Explorando a Estrutura de Variância:



Matriz de Correlação para os Resíduos:

- Modelo com todos os termos (tempo + grupo + interação: 8 parâmetros);
- Ajuste de Mínimos Quadrados Ordinários.

	Week0	Week1	Week4	Week6
Week0	1.00			
Week1	0.57	1.00		
Week4	0.57	0.78	1.00	
Week6	0.58	0.58	0.58	1.00

Matriz de Covariância para os Resíduos:

- Modelo com todos os termos (tempo + grupo + interação: 8 parâmetros);
- Ajuste de Mínimos Quadrados Ordinários.

	Week0	Week1	Week4	Week6
Week0	25			
Week1	19	44		
Week4	20	35	47	
Week6	22	29	30	58

Propostas de Modelo para a Estrutura de Covariância

- 1 Possivelmente explorar a possibilidade de utilizar a estrutura simples de simetria composta. No entanto, existe uma indicação de heterocedasticidade.
- 2 Ajustar o modelo "maximal" que consiste nos seguintes termos: (1) intercepto; (2) todos efeitos principais (1 + 3 parâmetros) e (3) interação tempo e grupo (3 parâmetros). Iniciar com a forma não estruturada (10 parâmetros) e testar as possíveis formas estruturadas.

Estudo Longitudinal Desbalanceado: Avaliação longitudinal do crescimento de lactentes nascidos de mães infectadas com o HIV-1.

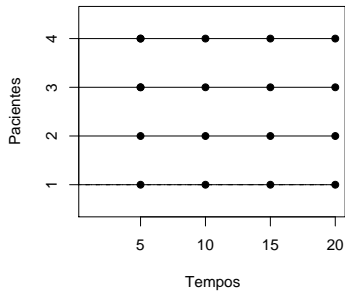
- Comparar longitudinalmente altura de lactentes infectados e não-infectados nascidos de mães infectadas pelo HIV.
- Uma coorte aberta acompanhada no ambulatório de AIDS pediátrica do Hospital das Clínicas da Universidade Federal de Minas Gerais.
- Período: 1995 a 2003.
- Inclusão: primeiros três meses de vida.
- Grupos: (1) não-infectados: 97; (2) infectados: 42.
- Controlado por sexo.

Estrutura Longitudinal

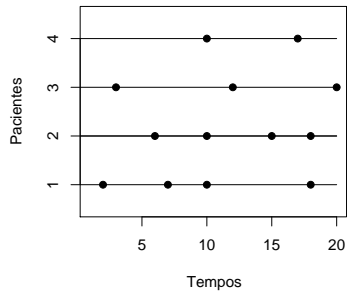
- Visitas regulares ao pediatra.
- Planejado para acompanhamento de 18 meses.
- Tempo: idade da criança.
- Tempo mediano de acompanhamento foi 15 meses (7 a 18).
- Número total de medidas: Não-infectados: 907; Infectados: 411.
- Número médio de visitas por criança: 9,5.
- Delineamento não-balanceado.

Delineamentos Balanceados vs Não-Balanceados

Estrutura Balanceada



Estrutura Desbalanceada



Delineamentos Balanceados vs Não-Balanceados

1 Estrutura Balanceada:

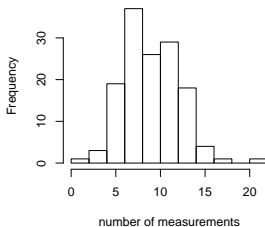
- todos os indivíduos avaliados (medidos) exatamente nos mesmos tempos;
- apresenta uma série de vantagens;
- pode-se utilizar todas as técnicas estatísticas disponíveis;
- pode ser difícil de executar.

2 Estrutura Desbalanceada:

- os indivíduos são avaliados em tempos diferentes;
- análise fica restrita a algumas técnicas;
- situação comum em estudos clínicos.

Delineamentos Balanceados vs Não-Balanceados

- 1 A Situação em questão é tipicamente desbalanceada;
- 2 Exemplo de três crianças do banco de dados
 - Criança 1: (0,0; 1,3; 4,0; 7,9; 8,8; 8,9; 17,6) meses
 - Criança 2: (0,9; 3,0; 4,1; 4,8; 5,2; 6,4; 7,5; 9,1; 10,0; 13,3; 17,2)
 - Criança 3: (2,6; 3,2; 3,7; 4,9; 5,8; 7,8; 9,7; 12,2)
- 3 Uma criança foi medida somente uma vez (ao nascimento) e outra apresenta 22 medidas.
- 4 Histograma do número de medidas para as 139 crianças.



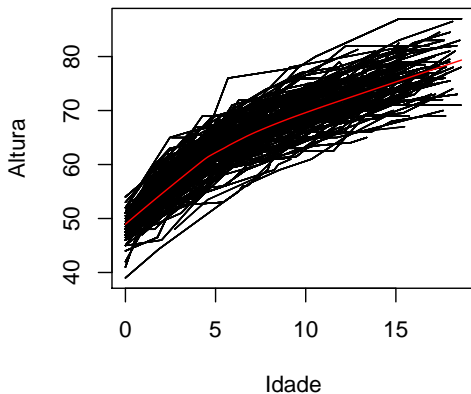
Comparação na Linha de Base

Comparação dos Grupos ao Nascimento

- 1 Comprimento dos dois grupos ao nascimento (média/desvio-padrão):
 - Infectados: 48,8 cm (2,9 cm).
 - Não-infectados: 48,7 cm (1,4 cm).
- 2 Distribuição de Sexo entre os grupos:
 - Infectados: 48% meninas.
 - Não-infectados: 45% meninas.

Perfis individuais e suavizado

Perfis das Crianças



Perfis médio por grupo

Gráfico para os Grupos

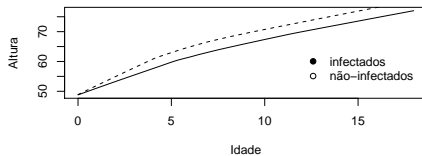
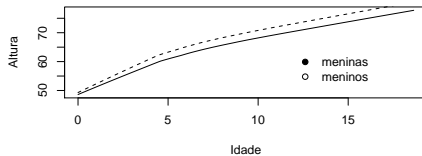


Gráfico para Meninos e Meninas



Modelo para a Média Populacional

- O comportamento da altura não é linear ao longo dos 18 meses de acompanhamento.
- Existe uma aparente indicação de mudança de velocidade de crescimento em torno de 5 meses.
- Os dois grupos, assim como os dois sexos, aparentemente têm valores semelhantes de altura ao nascimento, em torno de 50 cm.
- Existe uma indicação de interação entre grupo e idade, assim como sexo e idade.

Modelo para a Média Populacional

1 Modelo Polinomial:

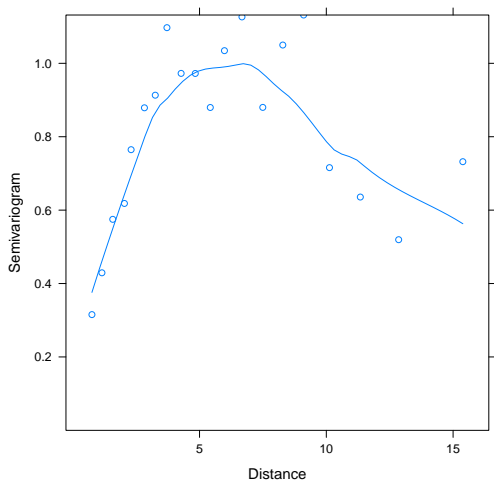
$$\begin{aligned}
 E(Y_{ij}) &= \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 t_{ij} * \text{sexo}_i + \beta_4 t_{ij} * \text{grupo}_i \\
 &+ \beta_5 t_{ij}^2 * \text{sexo}_i + \beta_6 t_{ij}^2 * \text{grupo}_i
 \end{aligned}$$

e

2 Modelo Segmentado (com knot em t=5 meses):

$$\begin{aligned}
 E(Y_{ij}) &= \beta_0 + \beta_1 t_{ij} + \beta_2 (t_{ij} - 5)_+ + \beta_3 t_{ij} * \text{sexo}_i + \beta_4 t_{ij} * \text{grupo}_i \\
 &+ \beta_5 (t_{ij} - 5)_+ * \text{sexo}_i + \beta_6 (t_{ij} - 5)_+ * \text{grupo}_i
 \end{aligned}$$

Variograma - Resíduos (para o modelo de efeitos aleatórios)



Modelo para a Estrutura de Covariância

- 1 $Var(Y_i) = W_i$
- 2 Como especificar W_i para estruturas desbalanceadas?
- 3 Tomar emprestado de Estatística Espacial (Cressie, 1991; Pinheiro e Bates, 2000):

$$Cov(y_{ij}, y_{ik}) = \sigma^2 g(\phi, |t_{ij} - t_{ik}|)$$

Algumas formas para $g(\phi, |t_{ij} - t_{ik}|)$:

- Exponential: $g(\phi, |t_{ij} - t_{ik}|) = \exp(-\phi |t_{ij} - t_{ik}|)$;
- AR contínuo: $g(\phi, |t_{ij} - t_{ik}|) = \phi^{|t_{ij} - t_{ik}|}$;
- Gaussiano: $g(\phi, |t_{ij} - t_{ik}|) = \exp(-(\phi |t_{ij} - t_{ik}|)^2)$;
- Esférica: $g(\phi, |t_{ij} - t_{ik}|) = (1 - 1.5(\phi |t_{ij} - t_{ik}|) + 0.5(\phi |t_{ij} - t_{ik}|)^3)$.