

Análise de Dados Longitudinais

Análise Exploratória

Enrico A. Colosimo/UFMG

<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/Correlação

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

Explorando Dados Longitudinais

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

Explorando a Média

- 1 Comportamento da resposta é linear em termos temporais?
- 2 Qual é a forma funcional de covariáveis tempo-dependentes?
- 3 Existe diferença (descritiva) entre grupos?
- 4 Existe indicação de interação entre grupo e tempo?

Explorando a Média

- 1 Sem Estruturar (tempo discreto/ignora escala temporal contínua)
 - Adequado para desenhos balanceados com poucos tempos.
 - Mais robusto pois assume os tempos como discretos.
- 2 Estruturando a Forma Funcional da Média
 - Descreve o comportamento longitudinal da resposta média (linear, polinomial, exponencial, etc).
 - Usualmente utiliza menos parâmetros que o não estruturado.
 - Sujeito a viés. O modelo funcional será sempre uma aproximação para o "verdadeiro" comportamento temporal.

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

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

$\mu(t)$ vai ser o valor do estimador no tempo 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 a 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).

Alisamento - Na Prática....

- 1 De forma adequada, os métodos nos passam mensagem descritiva similar para o comportamento da resposta média.
- 2 Forma adequada significa:
 - número de knots "ótimo" para o splines.
 - tamanho de janela "ótima" para o kernel.
- 3 O LOWESS é bastante utilizado por não ter problema de fronteira.
 - Veja Cleveland (1979), Harrell (2015).
 - Fácil de ser calculado.
 - Facilmente disponível no R.

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: $(n + 1)n/2$, n 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,

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

- 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)

Descrição do Estudo.

- 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.
- Objetivos:
 - Comparar os dois grupos;
 - Avaliar o comportamento temporal dos dois grupos.

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

Banco de Dados Largo com as seguintes colunas:

- ID,
- Groupo,
- 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

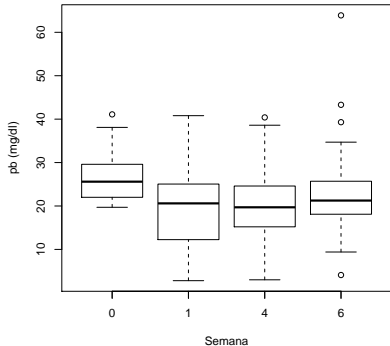
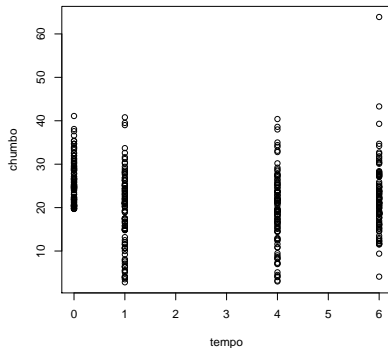


Gráfico Alternativo do Nível de chumbo, por Semana.



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

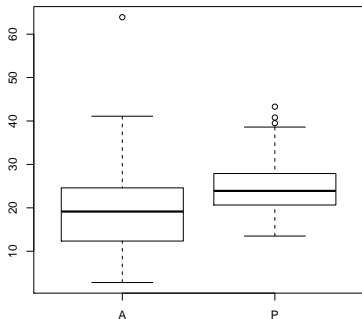
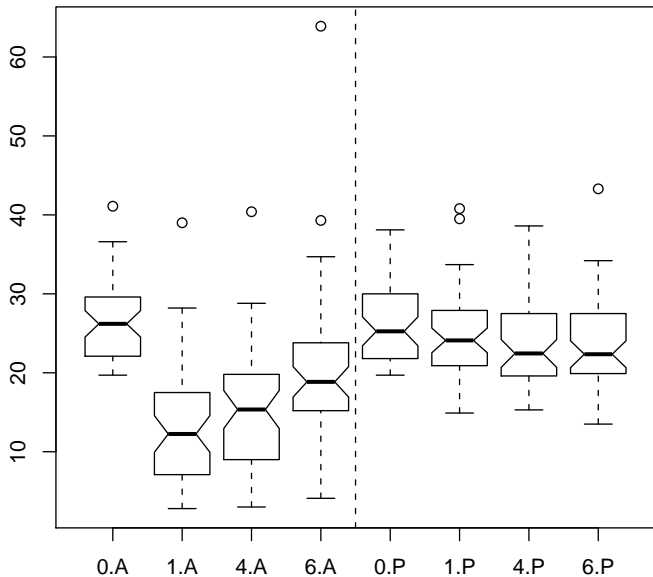
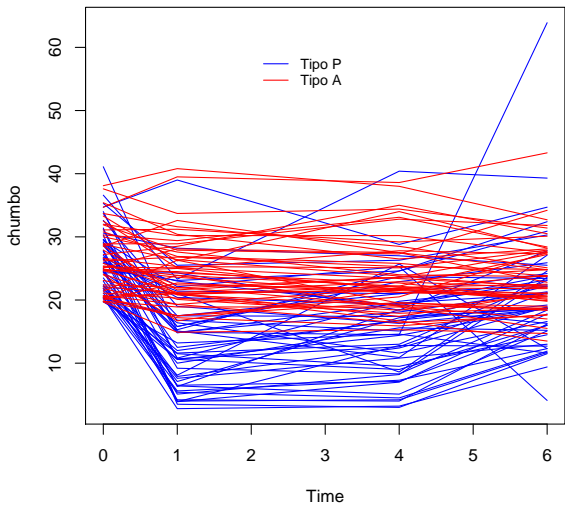


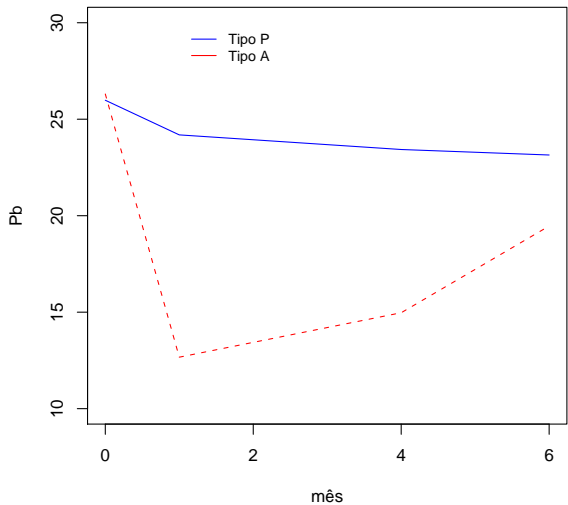
Tabela: Estatísticas descritivas do Nível de chumbo, por Grupo e Semana.

	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







Impressões Iniciais

- 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 (intercepto comum)

$$\begin{aligned} E(Y_{ij}) &= \beta_1 + \beta_2 I(t_j = 1) + \beta_3 I(t_j = 4) + \beta_4 I(t_j = 6) \\ &+ \beta_5 I(\text{grupo}_i = A) + \beta_6 I(t_j = 1) * I(\text{grupo}_i = A) \\ &+ \beta_7 I(t_j = 4) * I(\text{grupo}_i = A) + \beta_8 I(t_j = 6) * I(\text{grupo}_i = A) \end{aligned}$$

- Modelo 2

$$\begin{aligned} E(Y_{ij}) &= \beta_1 I(t_j = 0) + \beta_2 I(t_j = 1) + \beta_3 I(t_j = 4) + \beta_4 I(t_j = 6) \\ &+ \beta_5 I(t_j = 0) * I(\text{grupo}_i = A) + \beta_6 I(t_j = 1) * I(\text{grupo}_i = A) \\ &+ \beta_7 I(t_j = 4) * I(\text{grupo}_i = A) + \beta_8 I(t_j = 6) * I(\text{grupo}_i = A) \end{aligned}$$

Propostas de Modelo para a Média

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

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

- 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 (FLW, p. 128)

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

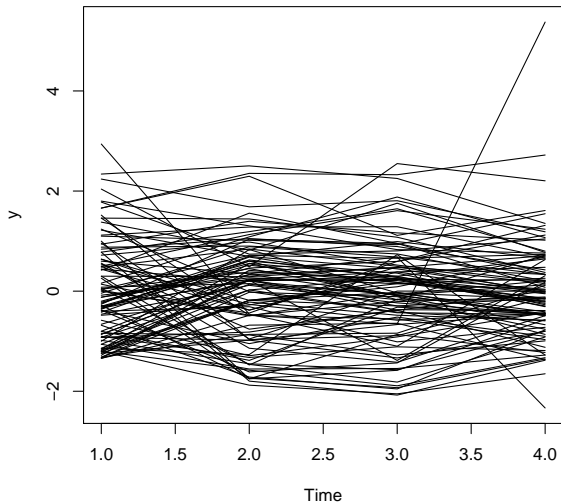
- Manter as respostas, sem fazer menção a diferença entre grupos.
- Manter as respostas e assumir médias iguais na linha de base (assumida nos modelos anteriores).
- Subtrair a resposta de linha de base das demais e analisar as diferenças.
- Usar a linha de base como covariável na análise dos dados.

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

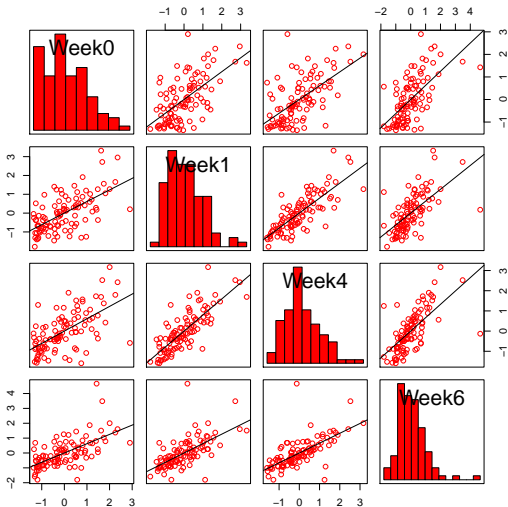
	Week0	Week1	Week4	Week6
Week0	1.00			
Week1	0.42	1.00		
Week4	0.47	0.84	1.00	
Week6	0.56	0.56	0.58	1.00

Perfil Padronizado - Ignorando Estrato por Tratamiento

Perfil Individual Padronizado



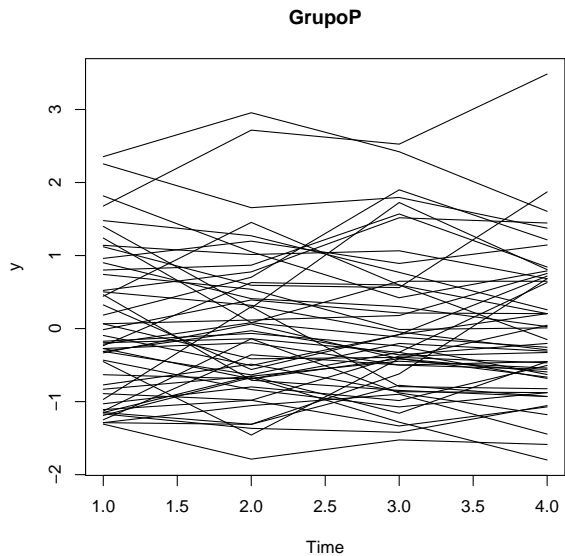
- Explorando a Estrutura de Dependência dos Tempos (Níveis de chumbo padronizados e Ignorando Efeito de Tratamento):



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

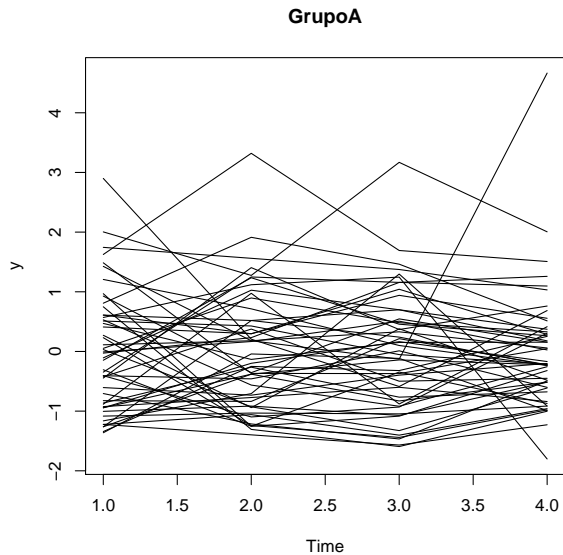
	Week0	Week1	Week4	Week6
Week0	1.00			
Week1	0.62	1.00		
Week4	0.61	0.80	1.00	
Week6	0.63	0.63	0.66	1.00

GRUPO - PLACEBO



Matriz de correlação para Grupo Placebo:

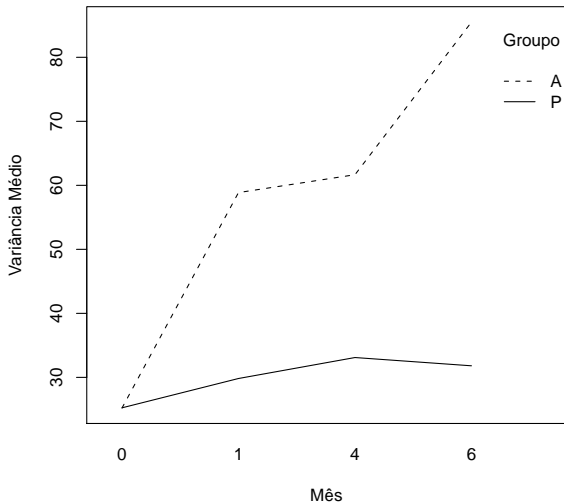
	Week0	Week1	Week4	Week6
Week0	1.00			
Week1	0.83	1.00		
Week4	0.84	0.86	1.00	
Week6	0.76	0.76	0.87	1.00



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.

Avaliando a Estrutura de Covariância

Utilizando o Teste da RVR.

- 1 Valor da Log-verossimilhança restrita para os modelos concorrentes.
 - Modelo Maximal para a média (8 parâmetros) e o Não estruturado heterocedástico para a Variância-Covariância (10 parâmetros): $\hat{l}_r = -1208.04$.
 - Modelo Maximal para a média (8 parâmetros) e o Simetria composta heterocedástico para a Variância-Covariância (5 parâmetros): $\hat{l}_r = -1216.98$.
 - Modelo Maximal para a média (8 parâmetros) e o Não estruturado homocedástico para a Variância-Covariância (7 parâmetros): $\hat{l}_r = -1220.82$.

2 TRVR

- Não estruturado heterocedástico vs Simetria composta heterocedástico
TRVR $2(-1208.04 + 1216.98) = 17.88$, $p < 0.001$.
- Não estruturado heterocedástico vs Não estruturado homocedástico
TRVR $2(-1208.04 + 1220.82) = 25.56$, $p < 0.001$.

Considerações Finais - Estrutura de Variância-Covariância

- 1 O modelo final com estrutura de variância heterocedástica e não estruturada parece ser o mais indicado para a análise destes dados.
- 2 O termo de efeito principal de grupos pode ser eliminado a partir da constatação que o estudo é experimental (clínico aleatorizado).
- 3 O gráfico de resíduos vs ajustados pode ser útil para validarmos o modelo ajustado.

Avaliando a Estrutura da Média

1 Duas Propostas

- Modelo 1: Não Estruturada (tempos discretos): 8 parâmetros.
- Modelo 2: Duas retas com knot na semana 1 (tempo contínuo): 5 parâmetros.
- FLW (2011) também testaram o modelo quadrático (sem sucesso, ver p. 160, Table 6.5).

2 Os modelos não são encaixados.

3 Qual modelo utilizar?

- Principal: responder a pergunta de interesse.
- Usar critérios estatísticos: AIC, BIC, etc.
 - Modelo 1: AIC= 2452.1; BIC= 2523.6 (menor, melhor)
 - Modelo 2: AIC= 2467.5; BIC= 2527.1 (menor, melhor)

Resultados Modelo Não Estruturado

Saída do R

Coefficients:

	Value	Std.Error	t-value	p-value
(Intercept)	26.272	0.7102939	36.98751	0.0000
factor(tempo)1	-1.612	0.7919197	-2.03556	0.0425
factor(tempo)4	-2.202	0.8149099	-2.70214	0.0072
factor(tempo)6	-2.626	0.8885239	-2.95546	0.0033
factor(Groupo)A	0.268	1.0045072	0.26680	0.7898
factor(tempo)1:factor(Groupo)A	-11.406	1.1199436	-10.18444	0.0000
factor(tempo)4:factor(Groupo)A	-8.824	1.1524566	-7.65669	0.0000
factor(tempo)6:factor(Groupo)A	-3.152	1.2565625	-2.50843	0.0125

- FLW (2011) interpreta o outro modelo (p.157-160).
- EP para outros contrastes de interesse tem ser obtidos pela $\widehat{\text{Var}}(\hat{\beta})$.

Resultados Modelo Não Estruturado

Supondo igualdade de médias na linha de base

- Os termos de ordem inferior devem sempre aparecer no modelo na presença daqueles de ordem superior.
- Ou seja, ao incluirmos a interação entre grupo e tempo, os efeitos principais de grupo e tempo devem aparecer na estrutura de regressão. Isso foi feito na saída do R do slide anterior.
- Uma exceção pode ser aberta, ao assumirmos, em estudos experimentais, igualdade de média na linha de base. Desta forma, excluimos o efeito principal de grupo. Observe que o mesmo foi não significativo na saída anterior. Veja FLW (2011, p. 128-132).
- A vantagem deste ajuste é a parcimônia, ao reduzirmos um parâmetro do modelo.

Resultados Modelo Não Estruturado

Supondo igualdade de médias na linha de base

Coefficients:

	Value	Std.Error	t-value	p-value
(Intercept)	26.406000	0.4998905	52.82357	0.0000
factor(tempo) 1	-1.644501	0.7824043	-2.10186	0.0362
factor(tempo) 4	-2.231356	0.8073816	-2.76369	0.0060
factor(tempo) 6	-2.642065	0.8864611	-2.98046	0.0031
factor(tempo) 1:factor(Grpo) A	-11.340999	1.0931205	-10.37488	0.0000
factor(tempo) 4:factor(Grpo) A	-8.765289	1.1312579	-7.74827	0.0000
factor(tempo) 6:factor(Groupo) A	-3.119870	1.2507769	-2.49435	0.0130

- Tempo = 0 (linha de base): $E(Y_{ij}|Grupo = P) = E(Y_{ij}|Grupo = A) = \beta_1$
Estimativa: $26,4 \pm 1,96 * 0.5$ (25, 5; 27, 3)
- Tempo = 1: $E(Y_{ij}|Grupo = A) = E(Y_{ij}|Grupo = P) = (\beta_1 + \beta_2 + \beta_5) - (\beta_1 + \beta_2) = \beta_5$
Estimativa: $-11,3 \pm 1,96 * 1.09$ (-13, 5; -9, 1)
- Houve uma redução da resposta do tempo =1 para o 6 no grupo =P (efeito placebo)?
Grupo = P: $E(Y_{ij}|Tempo = 1) = E(Y_{ij}|Tempo = 6) = \beta_2 - \beta_4$. Estimativa:
 $-1,64 + 2.64 = 1.0$ e a $\widehat{Var}(\hat{\beta}_2 - \hat{\beta}_4) = \widehat{Var}(\hat{\beta}_2) + \widehat{Var}(\hat{\beta}_4) - 2\widehat{Cov}(\hat{\beta}_2, \hat{\beta}_4) = 0.6^2$,
IC (-0, 17; 2, 17), não existe evidência de efeito placebo.

Transmissão Vertical - HIV

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

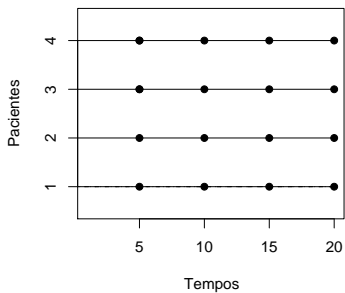
- Comparar longitudinalmente altura de lactentes infectados e não-infectados nascidos de mães infectadas pelo HIV.
- Uma coorte aberta acompanhada no ambulatório de AIDS pediátrica do Hospital das Clínicas da Universidade Federal de Minas Gerais.
- Período: 1995 a 2003.
- Inclusão: primeiros três meses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

Estrutura Longitudinal

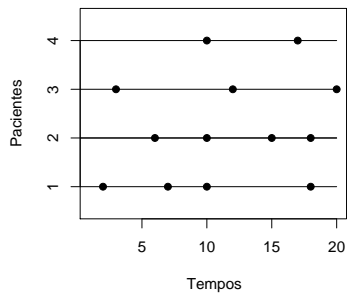
- Visitas regulares ao pediatra.
- Planejado para acompanhamento de 18 meses.
- Tempo: idade da criança.
- Tempo mediano de acompanhamento foi 15 meses (7 a 18).
- Número total de medidas: Não-infectados: 907; Infectados: 411.
- Número médio de visitas por criança: 9,5.
- Delineamento não-balanceado.

Delineamentos Balanceados vs Não-Balanceados

Estrutura Balanceada



Estrutura Desbalanceada



Delineamentos Balanceados vs Não-Balanceados

1 Estrutura Balanceada:

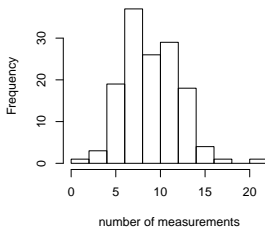
- todos os indivíduos avaliados (medidos) exatamente nos mesmos tempos;
- apresenta uma série de vantagens;
- pode-se utilizar todas as técnicas estatísticas disponíveis;
- pode ser difícil de executar.

2 Estrutura Desbalanceada:

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

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.

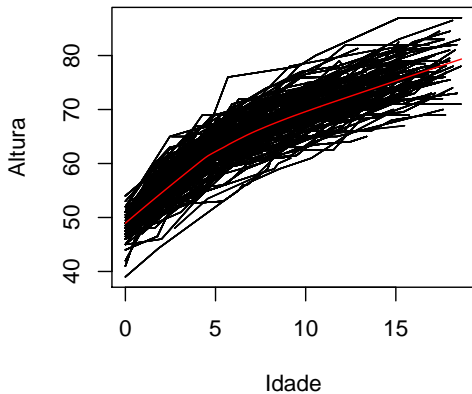


Comparação dos Grupos ao Nascimento

- 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

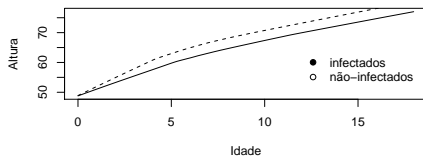
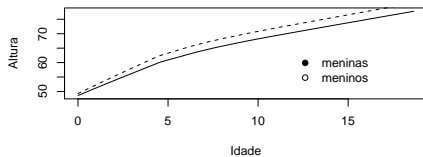


Gráfico para Meninos e Meninas



Modelo para a Média Populacional

- O comportamento da altura não parece ser 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 meninos e as meninas, 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.

Propostas de Modelos para a Média

1 Modelo Polinomial:

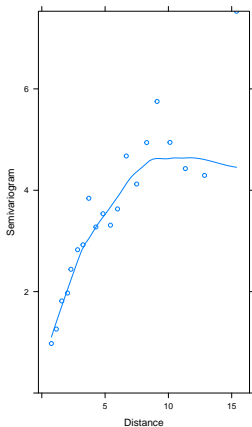
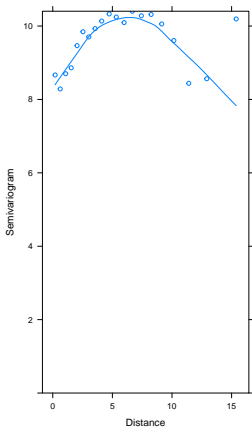
$$\begin{aligned} E(Y_{ij}) &= \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 \text{sexo}_i + \beta_4 \text{grupo}_i + \beta_5 (t_{ij} * \text{sexo}_i) \\ &+ \beta_6 (t_{ij} * \text{grupo}_i) + \beta_7 (t_{ij}^2 * \text{sexo}_i) + \beta_8 (t_{ij}^2 * \text{grupo}_i) \end{aligned}$$

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

$$\begin{aligned} E(Y_{ij}) &= \beta_0 + \beta_1 t_{ij} + \beta_2 (t_{ij} - 5)_+ + \beta_3 \text{sexo}_i + \beta_4 \text{grupo}_i + \beta_5 t_{ij} * \text{sexo}_i \\ &+ \beta_6 t_{ij} * \text{grupo}_i + \beta_7 (t_{ij} - 5)_+ * \text{sexo}_i + \beta_8 (t_{ij} - 5)_+ * \text{grupo}_i \end{aligned}$$

Nas estruturas consideradas, principalmente na polinomial, é indicado centrar a idade na média para evitar colinearidade.

Variograma - Resíduos (Independente e Sim. Composta)



Estrutura Independente (esquerda) e Simetria Composta (direita)

Indicação: correlação decresce com o aumento da diferença de idade.

Modelo para a Estrutura de Covariância

- 1 $Var(Y_i) = W_i$
- 2 Como especificar W_i para estruturas desbalanceadas?
Simetria composta é mais simples!!!
- 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: $g(\phi, |t_{ij} - t_{ik}|) = (1 - 1.5(\phi|t_{ij} - t_{ik}|) + 0.5(\phi|t_{ij} - t_{ik}|)^3)$.

Crianças - Transmissão Vertical

- Dois modelos para a média.
- Quatro modelos para a estrutura de correlação.
- Modelos não são encaixados.
- Resultados para o AIC e BIC (menor, melhor).

Est. Cor.	Modelo 2 Inclinações		Modelo Polinomial	
	AIC	BIC	AIC	BIC
Exponencial	5615	5672	5672	5729
AR1	5724	5780	5772	5829
Gaussiano	6732	6790	6828	6885
Esférica	5677	5734	6044	6101

Resultados Modelo Duas Inclinações

Estrutura de Correlação Exponencial

Parameter estimate(s):

range

6.081969

Coefficients:

	Value	Std.Error	t-value	p-value
(Intercept)	48.80520	0.4445630	109.78243	0.0000
Idade	2.90258	0.0986994	29.40825	0.0000
Age	-1.68790	0.1310430	-12.88050	0.0000
sexo	0.14506	0.5534911	0.26209	0.7933
status	-0.82771	0.6127635	-1.35078	0.1770
Idade:status	-0.37536	0.1384563	-2.71104	0.0068
Age:status	0.33165	0.1814328	1.82795	0.0678
Idade:sexo	0.35338	0.1231529	2.86944	0.0042
Age:sexo	-0.31721	0.1631197	-1.94466	0.0520

Residual standard error: 3.221604

Análise de Resíduos e Diagnóstico

Pontos Principais:

- A análise de dados longitudinais não fica completa sem a examinação dos resíduos. Ou seja, a verificação das suposições impostas ao modelo e ao processo de inferência.
- As ferramentas usuais de análise de resíduos para a regressão convencional (com observações independentes) podem ser estendidas para a estrutura longitudinal.

Suposições dos Modelos

- Estrutura da média: forma analítica, linearidade dos β 's.
- Normalidade (resposta e efeitos aleatórios).
- Estrutura de Variância-Covariância: Homocedasticidade e correlação das medidas do mesmo indivíduo.

Suposições dos Modelos

- Defina o vetor de resíduos para cada indivíduo

$$r_i = Y_i - X_i \hat{\beta}, \quad i = 1, \dots, N,$$

que é um estimador para o vetor de erros

$$\epsilon_i = Y_i - X_i \beta, \quad i = 1, \dots, N.$$

- Tratando-se de dados longitudinais, sabemos que os componentes do vetor de resíduos r_i são correlacionados e não necessariamente têm variância constante.

Utilidade dos Resíduos r_j

Gráficos:

- Gráfico de r_{ij} vs \hat{Y}_{ij} : é útil para identificar alguma tendência sistemática (por exemplo, presença de curvatura) e presença de pontos extremos ("outliers"). O modelo corretamente especificado não deve apresentar nenhuma tendência neste gráfico.
- Limitação: este gráfico não tem necessariamente uma largura constante. Ou seja, cuidado ao interpretar este gráfico com relação a homocedasticidade.
- Gráfico de r_{ij} vs t_{ij} : é também útil para identificar violação da homocedasticidade.

Solução: Examinar resíduos transformados

- Há muitas possibilidades para transformar os resíduos.
- A transformação deve ser realizada de forma que os resíduos “imitem” aqueles da regressão linear padrão.
- Os resíduos r_i^* definidos a seguir são não-correlacionados e têm variância unitária:

$$r_i^* = L_i^{-1} r_i,$$

em que L_i é a matriz triangular superior resultante da decomposição de Cholesky da matriz de covariâncias estimada $\widehat{Var}(Y_i)$, ou seja, $\widehat{Var}(Y_i) = L_i L_i'$.

Resíduos transformados

- Podemos aplicar a mesma transformação ao vetor de valores preditos \hat{Y}_i , ao vetor da variável resposta Y_i e à matriz de covariáveis \mathbf{X}_i :

$$\hat{Y}_i^* = L_i^{-1} \hat{Y}_i$$

$$Y_i^* = L_i^{-1} Y_i$$

$$\mathbf{X}_i^* = \hat{L}_i^{-1} \mathbf{X}_i$$

e então todos os diagnósticos de resíduos usuais para a regressão linear padrão podem ser aplicados para r_i^* .

Gráficos de Adequação

- Gráfico de dispersão dos resíduos transformados r_{ij}^* versus os valores preditos transformados \hat{Y}_{ij}^* : não deve apresentar nenhum padrão sistemático para um modelo corretamente especificado. Ou seja, deve apresentar um padrão aleatório em torno de uma média zero. Útil para verificar homocedasticidade.
- Gráfico de dispersão dos resíduos transformados r_{ij}^* versus covariáveis transformadas X_{ij}^* (em especial, idade ou tempo): verificar padrões de mudança na resposta média ao longo do tempo;
- QQ-plot de r_i^* : verificar normalidade e identificar outliers.

Semi-variograma

- O semi-variograma, denotado por $\gamma(h_{ijk})$, é dado por:

$$\gamma(h_{ijk}) = \frac{1}{2}E(r_{ij} - r_{ik})^2,$$

em que $h_{ijk} = t_{ij} - t_{ik}$.

- O semi-variograma pode ser utilizado como uma ferramenta para verificar a adequação do modelo selecionado para a estrutura de covariância dos dados.

Semi-variograma

- Como os resíduos têm média zero, o semi-variograma pode ser reescrito como:

$$\begin{aligned}\gamma(h_{ijk}) &= \frac{1}{2}E(r_{ij} - r_{ik})^2 \\ &= \frac{1}{2}E(r_{ij}^2 + r_{ik}^2 - 2r_{ij}r_{ik}) \\ &= \frac{1}{2}\text{Var}(r_{ij}) + \frac{1}{2}\text{Var}(r_{ik}) - \text{Cov}(r_{ij}, r_{ik}).\end{aligned}$$

- Quando o semivariograma é aplicado aos resíduos transformados, r_{ij}^* , a seguinte simplificação é obtida:

$$\gamma(h_{ijk}) = \frac{1}{2}(1) + \frac{1}{2}(1) - 0 = 1.$$

Semi-variograma

- Logo, se o modelo é corretamente especificado para a matriz de covariâncias, o gráfico do semi-variograma amostral $\hat{\gamma}(h_{ijk})$ dos resíduos transformados versus h_{ijk} deveria flutuar aleatoriamente em torno de uma linha horizontal centrada em 1.
- O semi-variograma é muito sensível a outliers.

Crianças - Transmissão Vertical: Análise de Resíduos

