Análise descritiva para dados longitudinais

Prof. Caio Azevedo

Introdução

- Os dados longitudinais podem ser classificados como
  - Em relação às condições de avaliação: regulares - quando o intervalo entre duas medidas consecutivas é constante; irregulares - caso contrário (determinação da escolha da estrutura de covariância).
  - Quanto ao planejamento: balanceados - quando as medidas são obtidas nos mesmos instantes de avaliação em todas as unidades amostrais; não balanceados - caso contrário (estimação dos parâmetros).
  - Em relação as observações: completos - quando não há observações perdidas ou dados omissos (por alguma razão); incompletos - caso contrário (estimação dos parâmetros).

Exemplo 1: Concentração de bilirrubina em recem-nascidos saudáveis

- Os dados correspondem a um estudo realizado na Escola Paulista de Medicina (UNIFESP), em que foi medida a concentração de bilirrubina (μ mol/L) em 89 recém-nascidos a termo (gestação entre 37 e 42 semanas) saudáveis em aleitamento materno durante 1, 2, 3, 4, 5, 6, 8, 10 e 12 dias após o nascimento.
- O objetivo era explicar a variação da concentração de bilirrubina em função da idade.

Exemplo 1: cont.

- A bilirrubina é uma substância amarelada encontrada na bile, que permanece no plasma sanguíneo até ser eliminada na urina. Quanto mais bilirrubina eliminada na urina, mais amarela ela se torna. Excesso de bilirrubina (hiperbilirrubinemia) pode indicar problemas no fígado, baço, nos rins ou na vesícula biliar.
- Estudo irregular, balanceado e completo (89 observações para cada condição de avaliação e 9 por indivíduo).

Banco de dados (multivariado)

RN	Dia								
	1	2	3	4	5	6	8	10	12
1	2,70	0,40	0,00	0,50	0,60	0,00	0,00	0,50	0,80
2	4,50	5,50	3,90	2,70	2,90	2,00	1,50	1,30	1,70
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
87	3,60	6,60	9,90	8,80	11,50	12,00	12,00	11,30	9,70
88	3,60	3,70	2,80	2,00	1,50	0,00	1,20	1,60	0,50
89	2,60	1,40	1,30	1,00	1,60	0,40	0,00	0,30	0,00

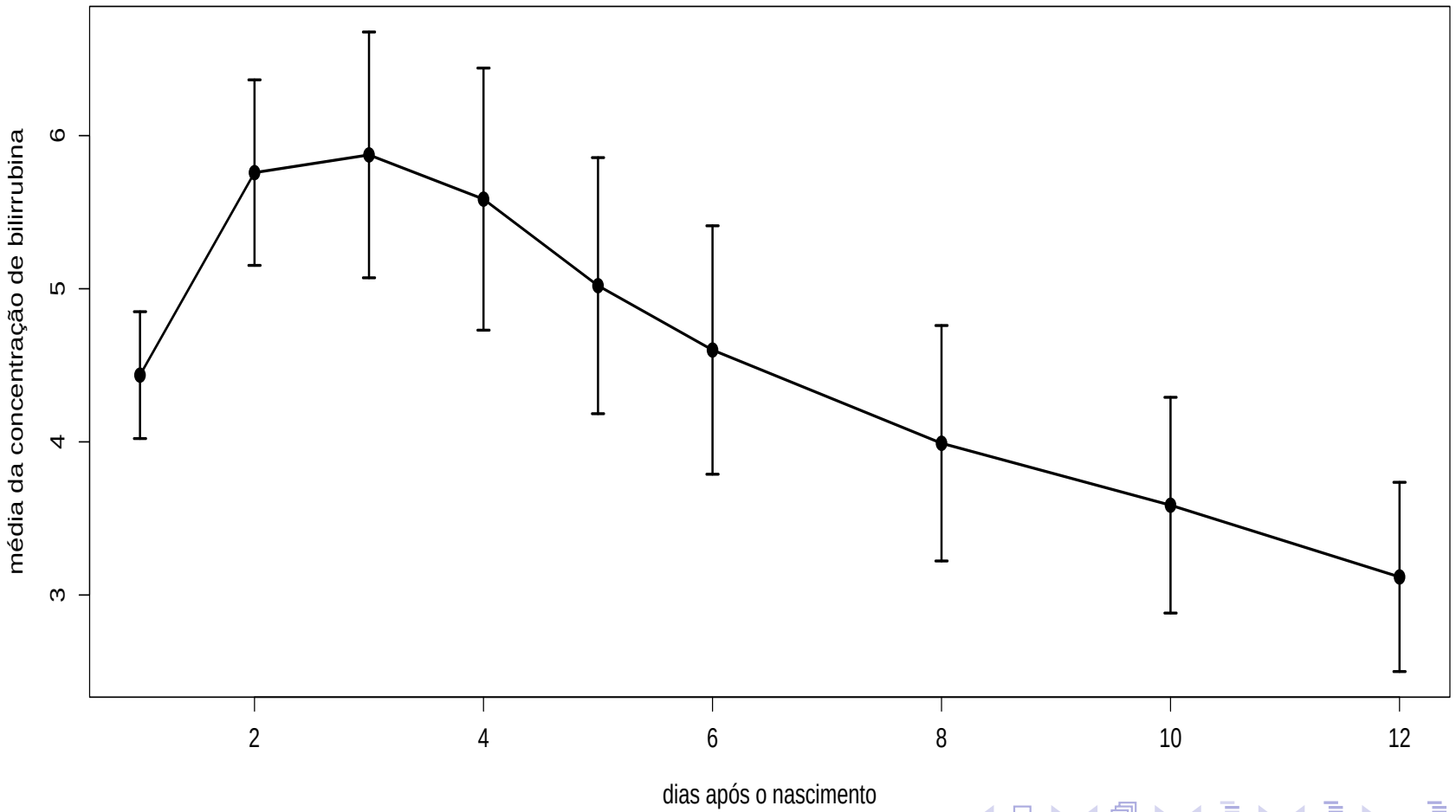
Banco de dados (longitudinal)

RN	Dia	Bilirrubina
1	1	2,70
1	2	0,40
⋮	⋮	⋮
1	12	0,80
⋮	⋮	⋮
89	1	2,60
89	2	1,40
⋮	⋮	⋮
89	12	0,60

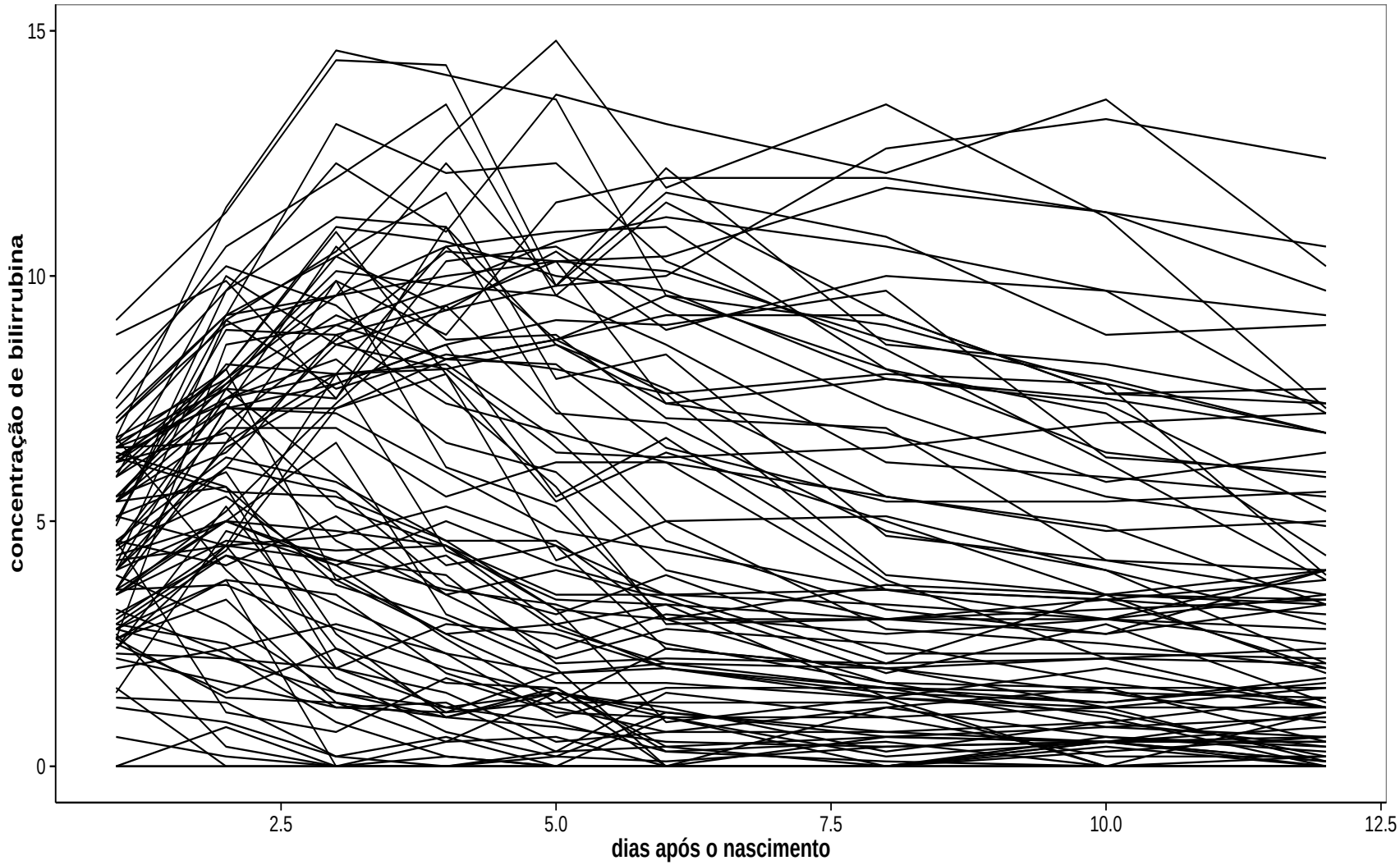
Medidas resumo

Dia	Média	DP	Var.	CV(%)	Min.	Med.	Máximo	n
1	4,44	1,99	3,98	44,95	0,00	4,50	9,10	89
2	5,76	2,92	8,50	50,64	0,00	6,10	11,40	89
3	5,87	3,86	14,92	65,77	0,00	5,80	14,60	89
4	5,59	4,12	16,97	73,76	0,00	4,60	14,30	89
5	5,02	4,02	16,20	80,17	0,00	4,10	14,80	89
6	4,60	3,90	15,25	84,89	0,00	3,30	13,10	89
8	3,99	3,70	13,70	92,73	0,00	2,80	13,50	89
10	3,59	3,39	11,50	94,56	0,00	2,70	13,60	89
12	3,12	2,97	8,85	95,39	0,00	2,10	12,40	89

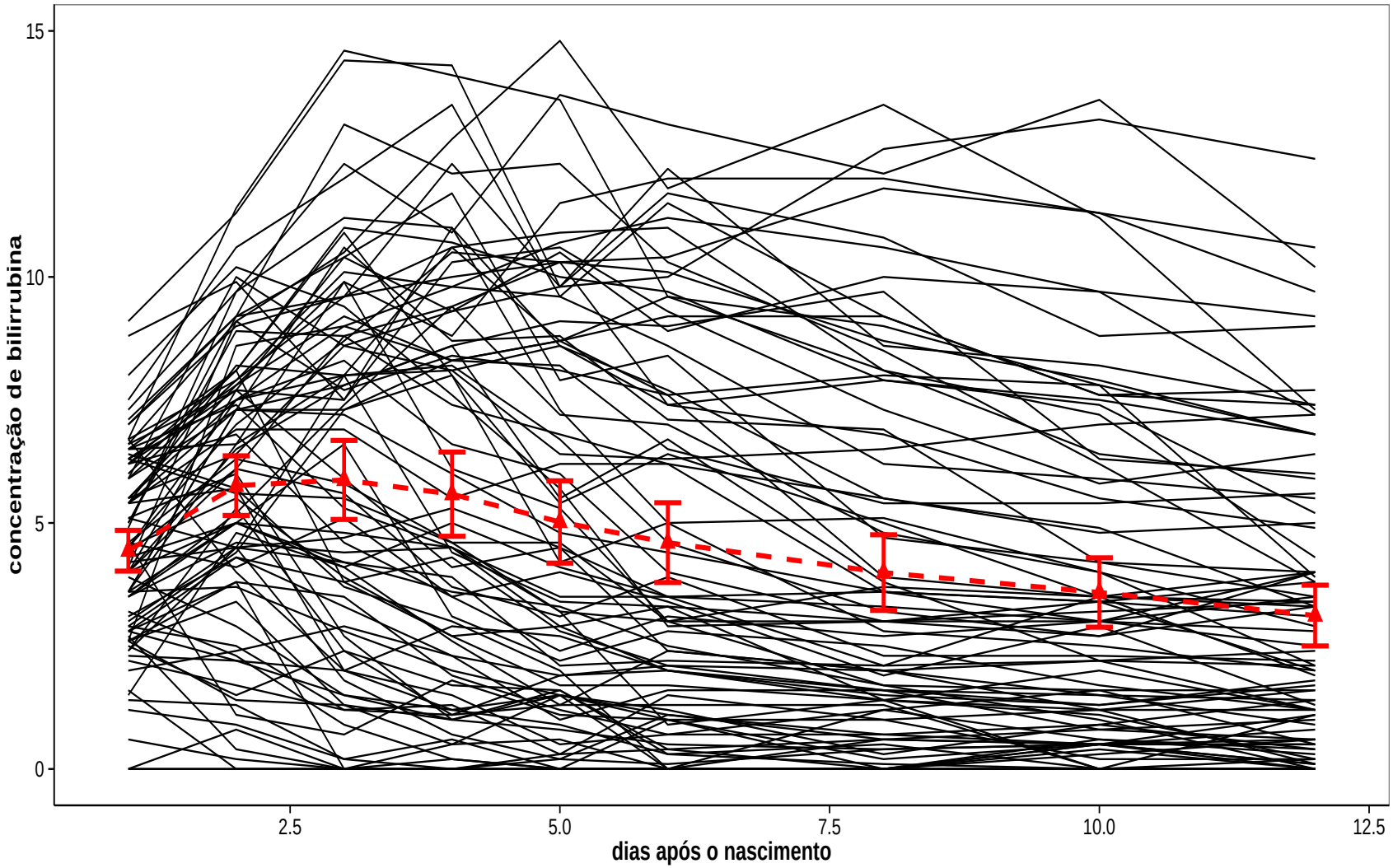
Perfil médio



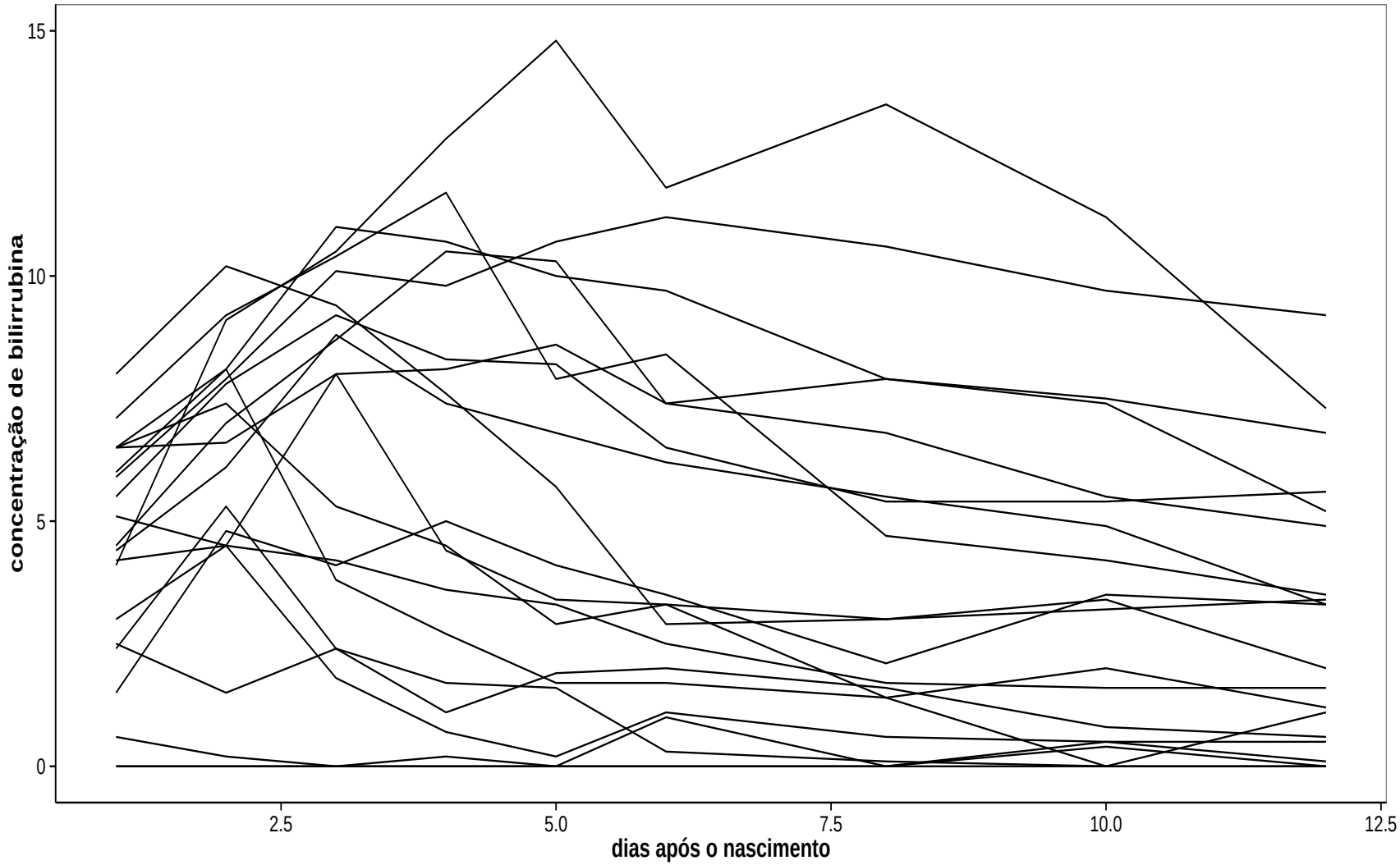
Perfis individuais: amostra completa



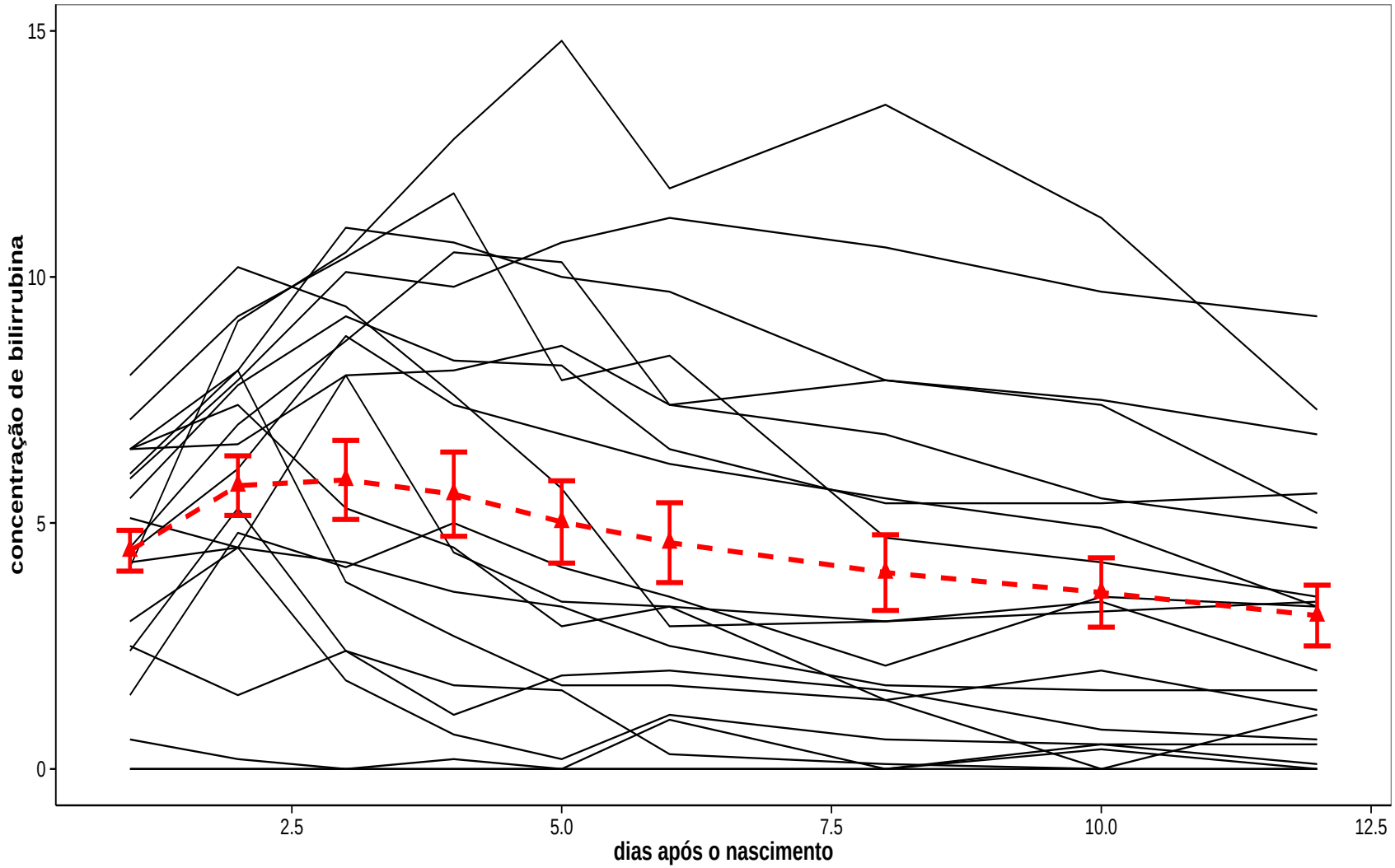
Perfis individuais: amostra completa e perfil médio



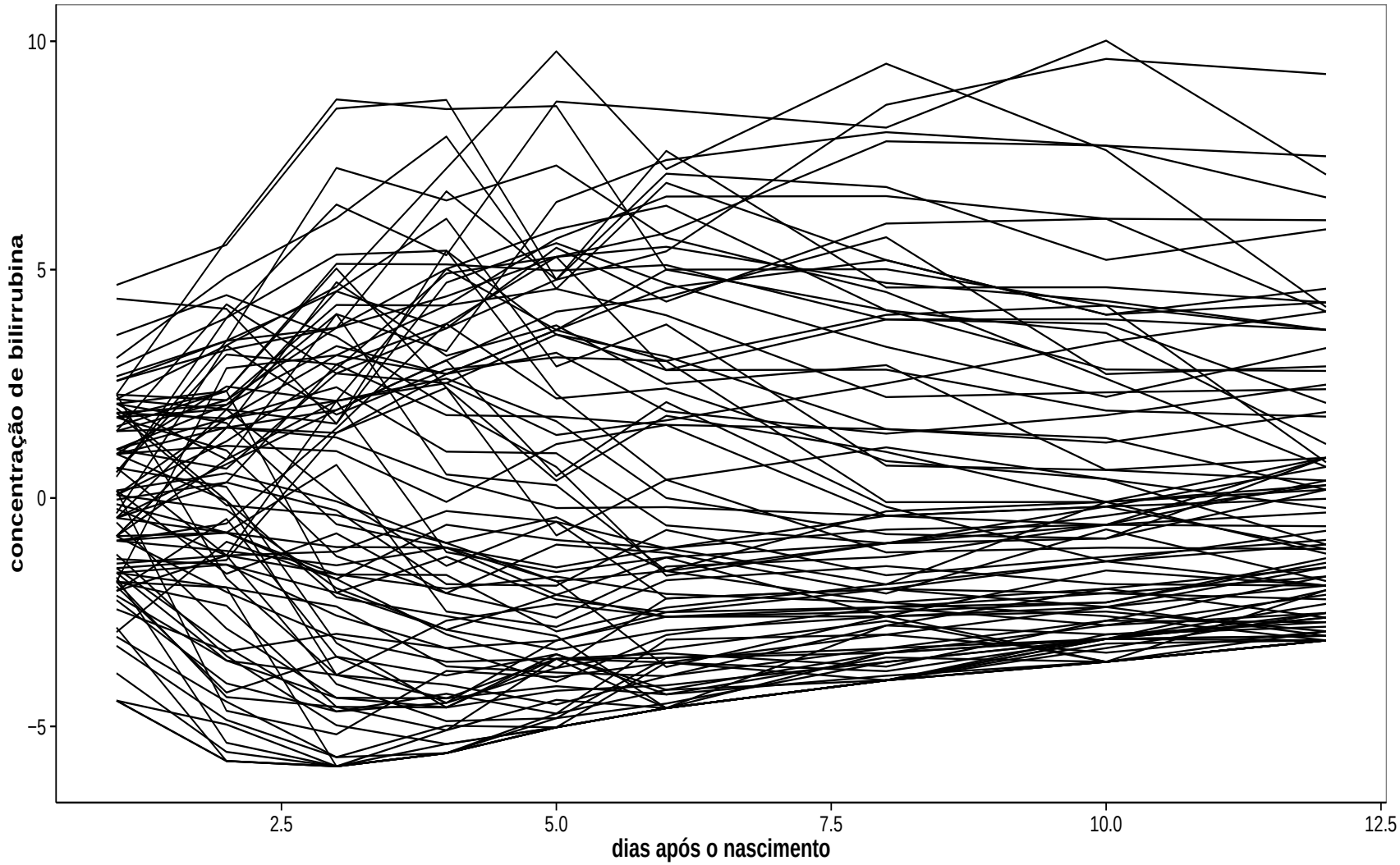
Perfis individuais: 20 RN selecionados aleatoriamente



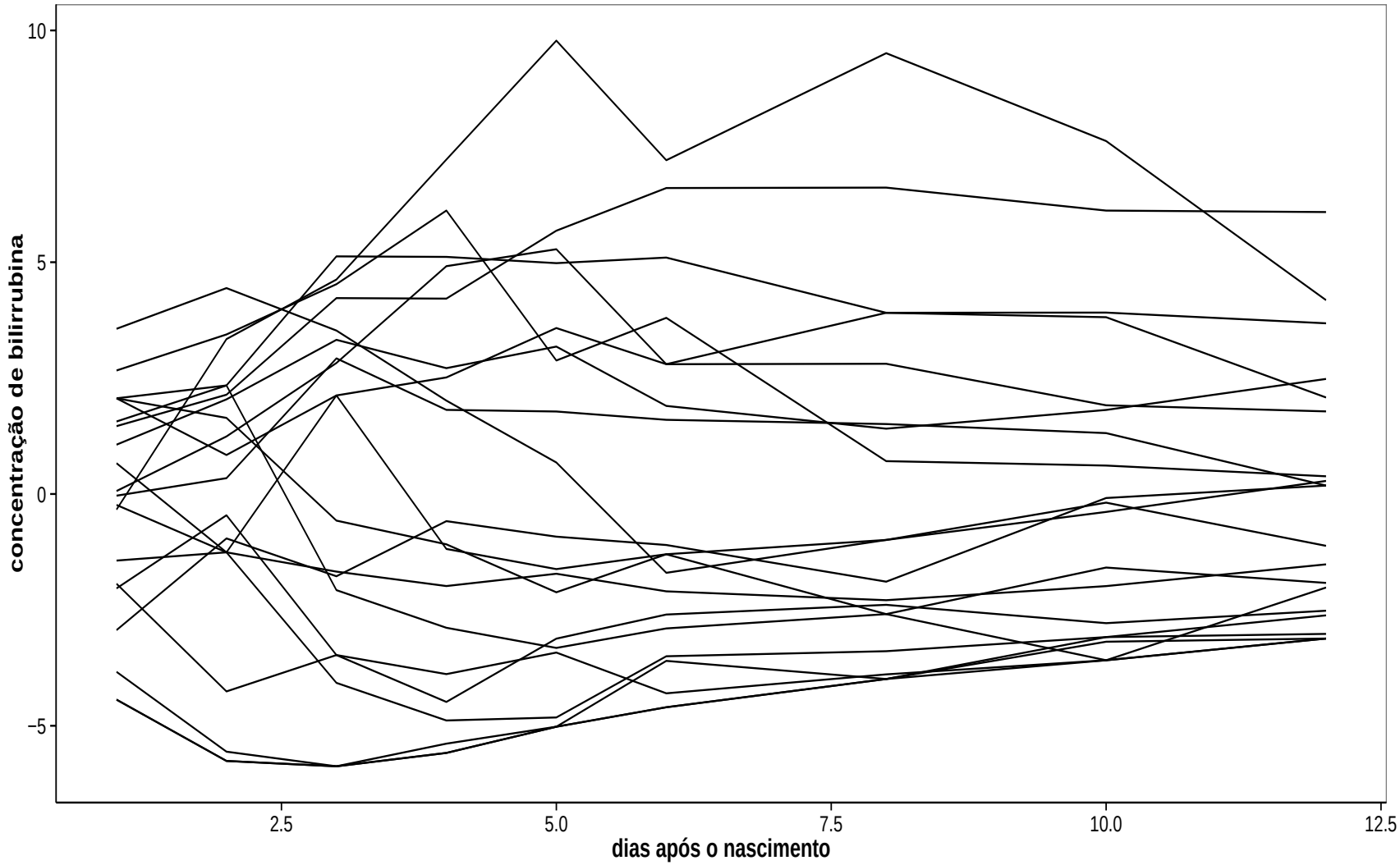
Cont.: 20 RN selecionados aleatoriamente e perfil médio



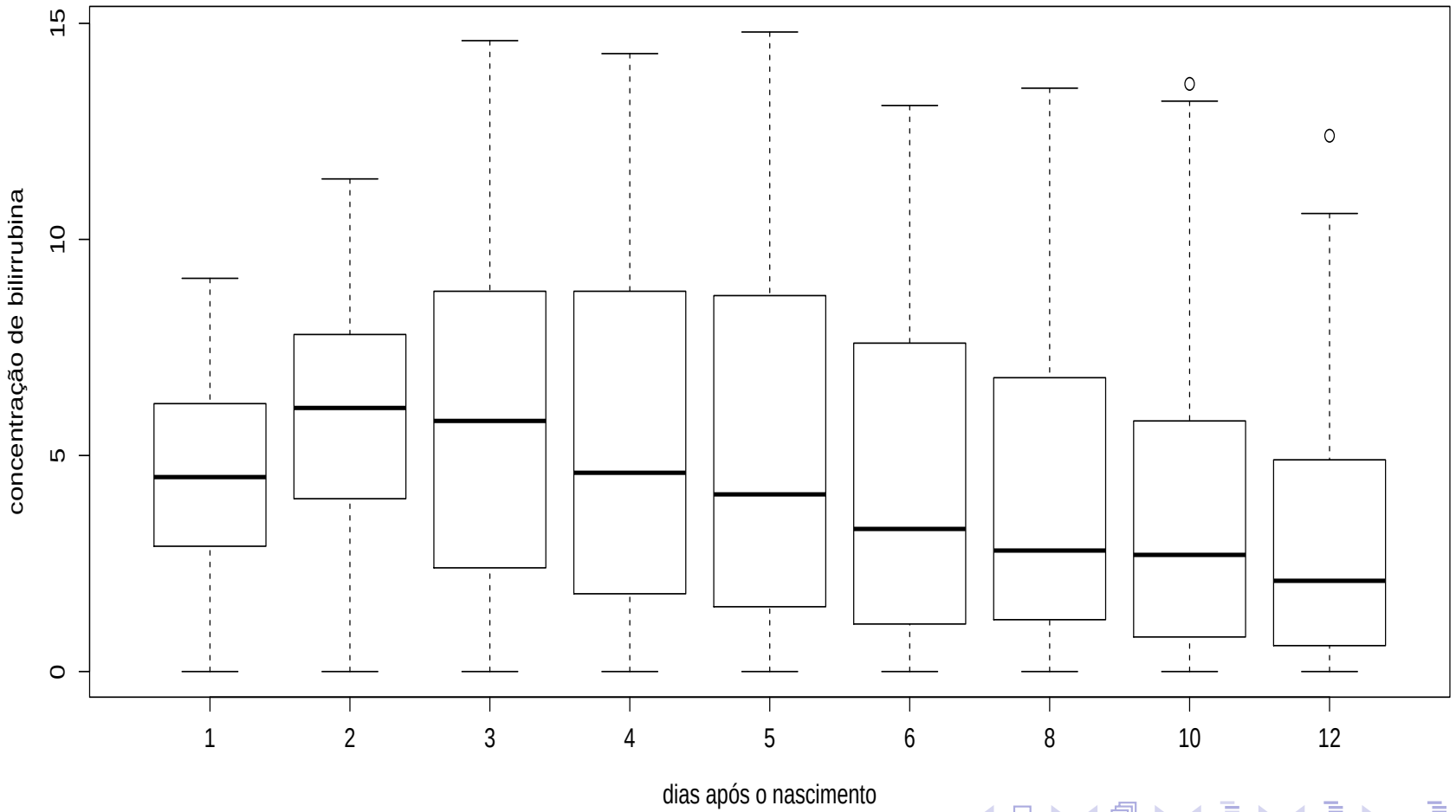
Perfis individuais centrados: amostra completa



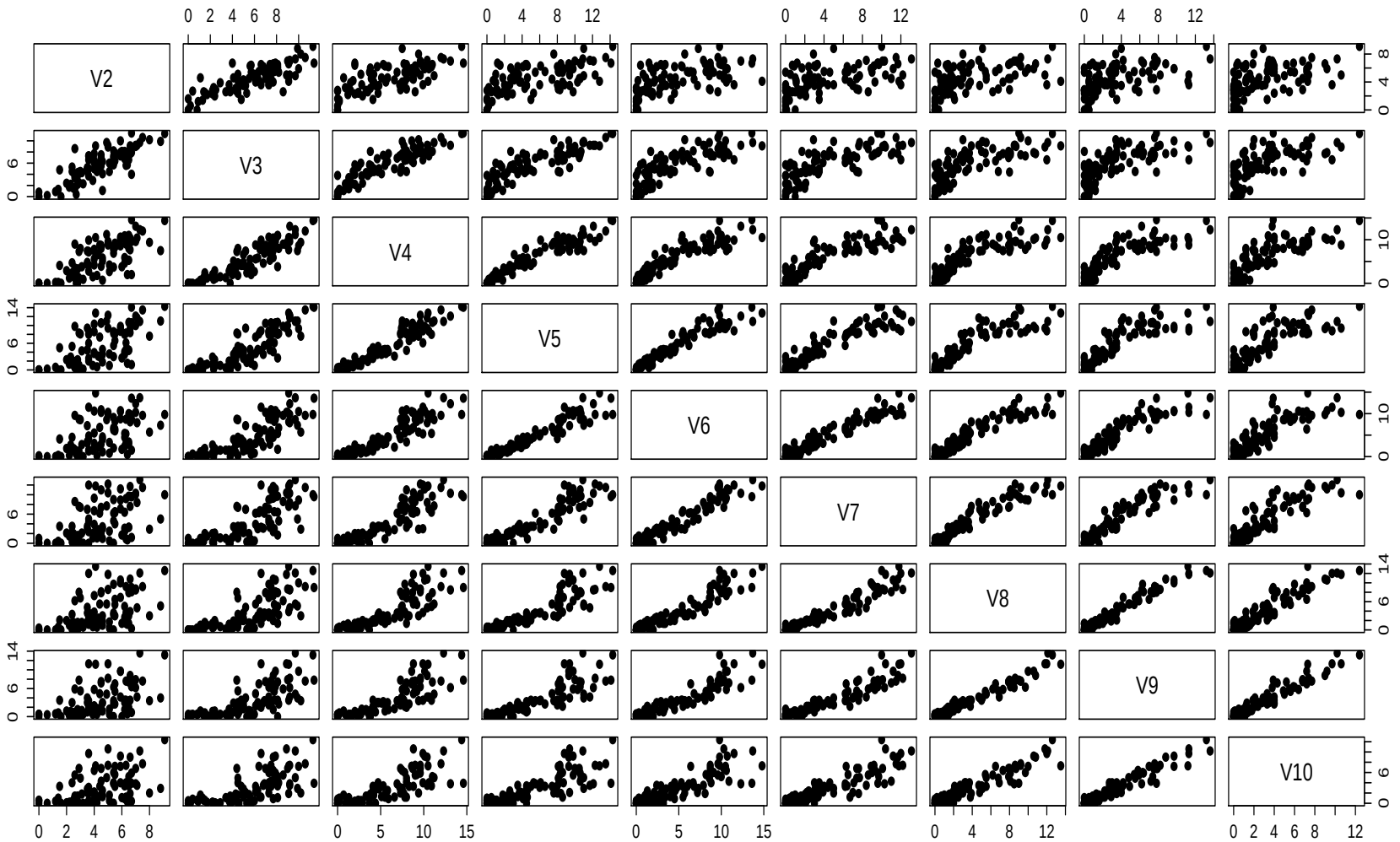
Cont.: 20 RN selecionados aleatoriamente



Box plot


concentração de bilirrubina
0
5
10
15
1
2
3
4
5
6
8
10
12
dias após o nascimento

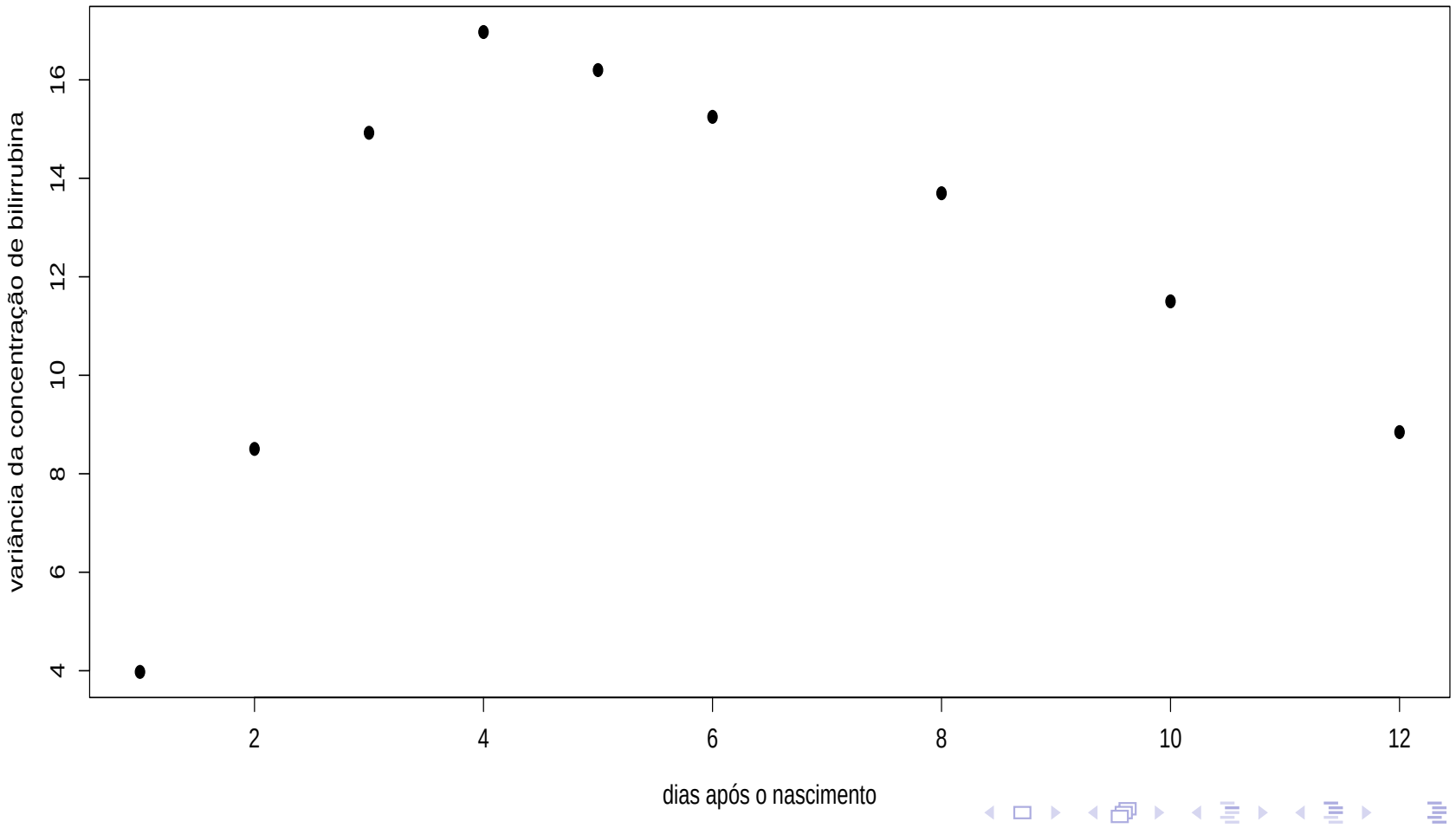
Matriz de diagramas de dispersão



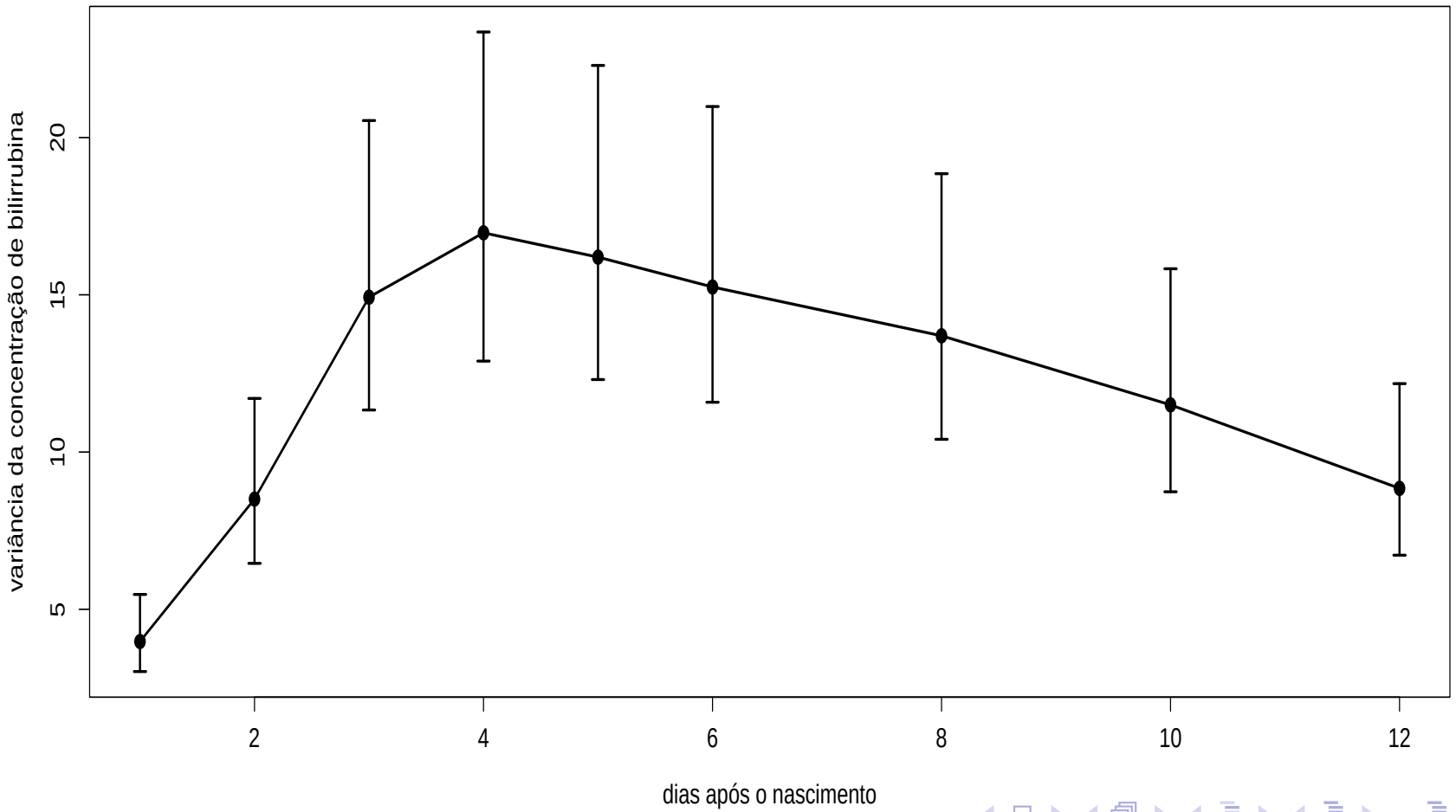
Variâncias (diagonal), correlações (acima) e covariâncias (abaixo)

	Dia								
Dia	1	2	3	4	5	6	8	10	12
1	3,98	0,82	0,71	0,63	0,55	0,52	0,51	0,50	0,48
2	4,76	8,51	0,90	0,86	0,79	0,76	0,72	0,70	0,68
3	5,47	10,09	14,93	0,95	0,91	0,88	0,85	0,82	0,78
4	5,16	10,32	15,06	16,97	0,95	0,93	0,88	0,85	0,81
5	4,42	9,32	14,09	15,71	16,20	0,96	0,94	0,91	0,85
6	4,05	8,62	13,27	14,92	15,06	15,25	0,96	0,93	0,88
8	3,78	7,79	12,11	13,49	14,01	13,81	13,70	0,98	0,94
10	3,37	6,95	10,79	11,94	12,47	12,28	12,25	11,50	0,96
12	2,87	5,88	9,02	9,90	10,16	10,26	10,31	9,68	8,85

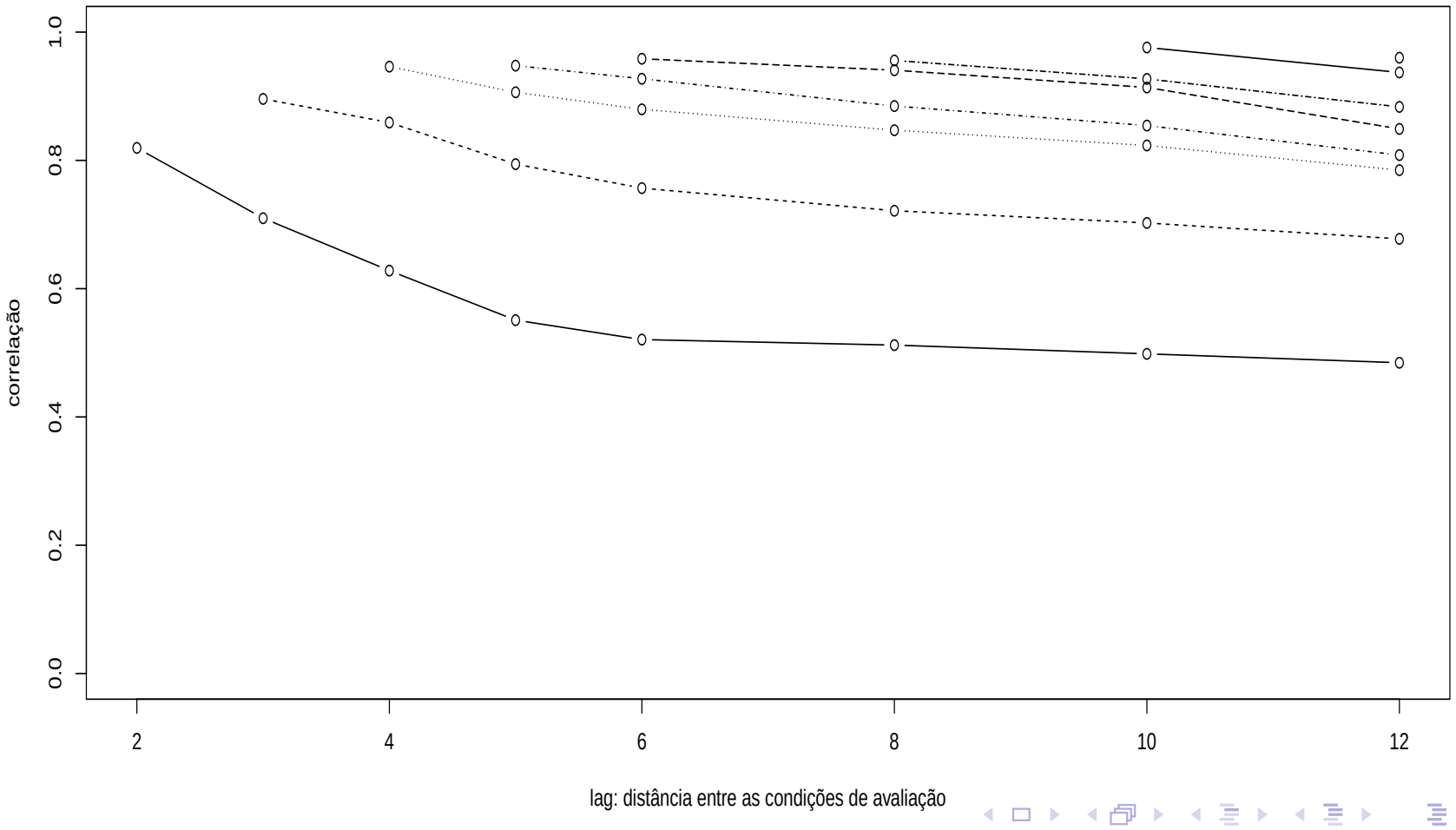
Variâncias em cada condição



Variâncias em cada condição com intervalos de confiança



Gráficos dos perfis das linhas da matriz de correlações



Procedimento para se gerar o gráfico de envelopes (quantil-quantil)

1) Simule n variáveis aleatórias ind. de interesse ($N(0,1)$ ou $F_{(p,n-p)}$). Repita este processo m vezes.
2) Ao final teremos uma matriz com valores simulados dessas variáveis aleatórias, digamos V_{ij}, i=1,...,n, (tamanho da amostra) j=1,...,m (réplica).

$$\mathbf{V} = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1m} \\ v_{21} & v_{22} & \cdots & v_{v2m} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nm} \end{bmatrix}$$

Cont.

3) Dentro de cada amostra, ordena-se, de modo crescente, os valores simulados, obtendo-se $v_{(i)j}$ (estatísticas de ordem):

$$\mathbf{V}^* = \begin{bmatrix} v_{(1)1} & v_{(1)2} & \cdots & v_{(1)m} \\ v_{(2)1} & v_{(2)2} & \cdots & v_{(2)m} \\ \vdots & \vdots & \ddots & \vdots \\ v_{(n)1} & v_{(n)2} & \cdots & v_{(n)m} \end{bmatrix}$$

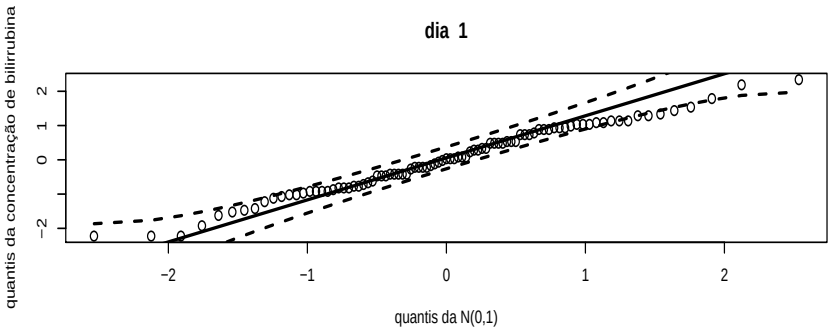
4) Obtem-se os limites $v_{(i)I} = \min\limits_{1 \leq j \leq m} v_{(i)j}$ e $v_{(i)S} = \max\limits_{1 \leq j \leq m} v_{(i)j}$, $i = 1, 2, ..., n$.

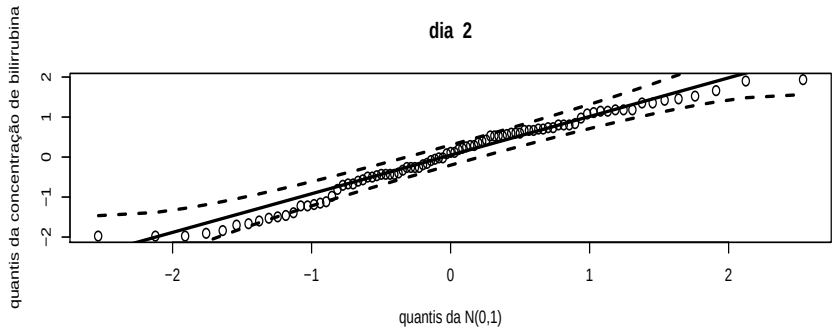
Cont.

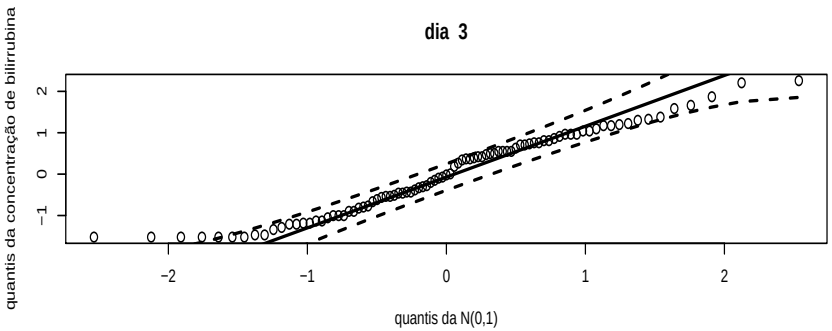
5) Na prática considera-se $v_{(i)I} = \frac{v_{(i)(2)} + v_{(i)(3)}}{2}$ e $v_{(i)S} = \frac{v_{(i)(m-2)} + v_{i(m-1)}}{2}$ (para se gerar limites de confiança), em que $v_{(i)(r)}$ é a r-ésima estatística de ordem dentro de cada linha, $i = 1, 2, ..., n$.

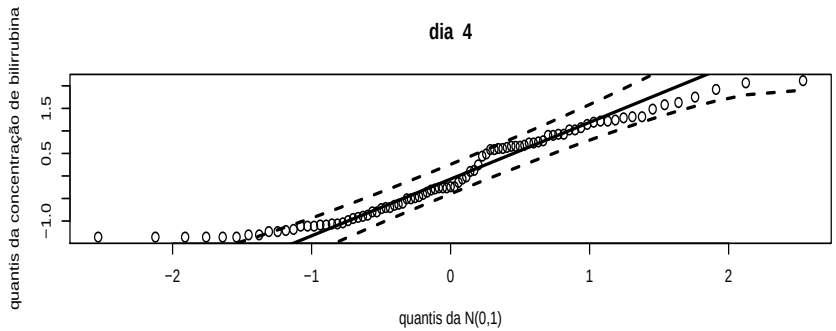
- Além disso, consideramos como a linha de referência $v_{(i)} = \frac{1}{m} \sum_{j=1}^{m} v_{(i)j}, i = 1, 2, ..., n$.

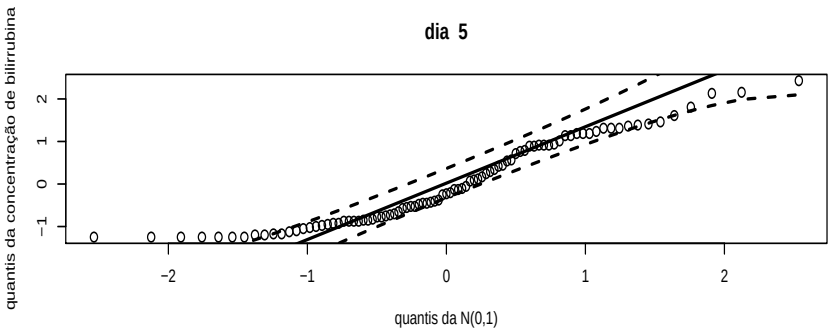
Gráficos de Envelope



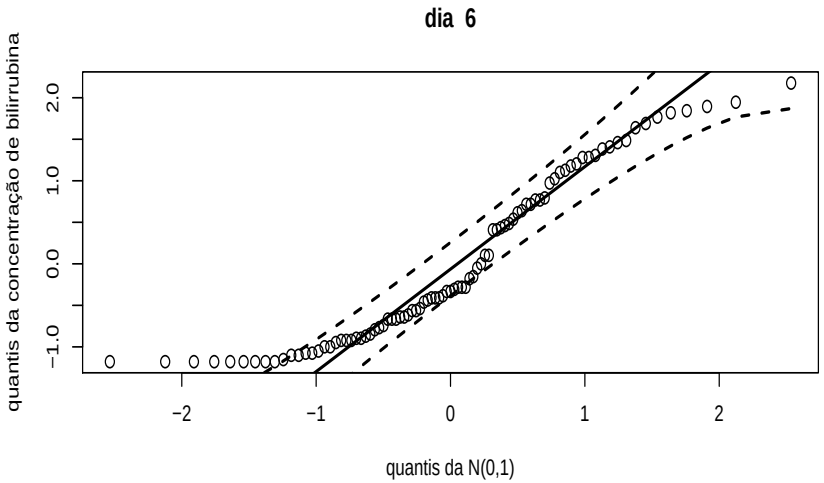


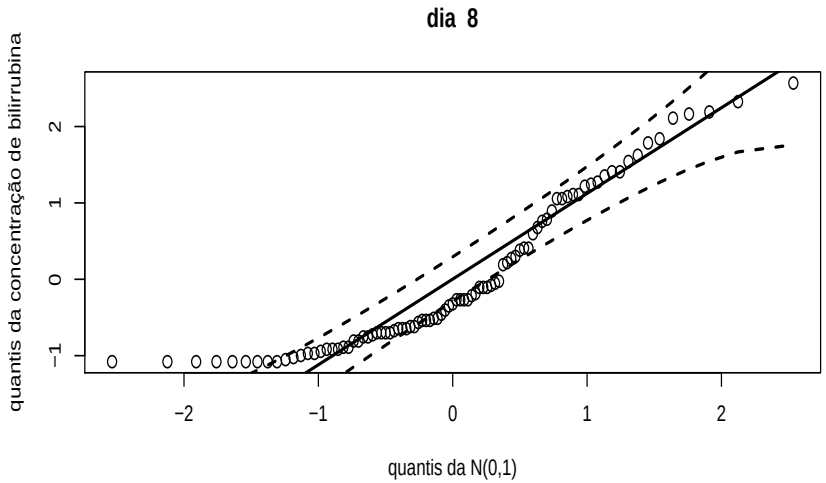


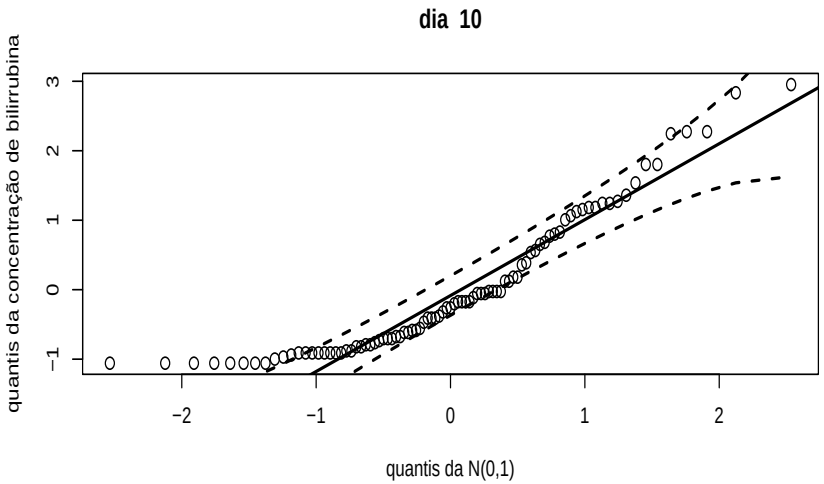


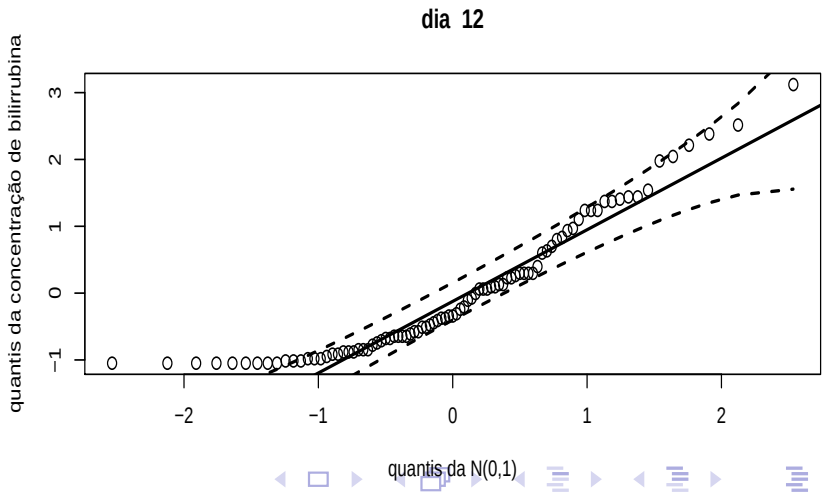


Gráficos de Envelope (continuação)









Variograma

- Para um processo estocástico estacionário $\{Y(t), t \in \mathcal{R}\}$, ou seja, $\mathcal{E}(Y(t)) = \mathcal{E}(Y(t-u))$ e $\mathcal{V}(Y) = \mathcal{V}(Y(t-u))$, $\forall t \in \mathcal{R}$ e $u \in \mathcal{R}^+$, o variograma é definido como:

$$g(u) = \frac{1}{2}\mathcal{E}\left[(Y(t) - Y(t-u))^2\right]$$

- Definindo, $\gamma(u) = Cov(Y(t), Y(t-u))$, temos que $g(u) = \gamma(0) - \gamma(u) = \sigma^2(1 - \rho(u))$. (exercício)
- Para estimar o variograma é útil considerar as observações padronizadas $\Delta_{ij} = \frac{y_{ij} - \overline{y}_j}{s_j}$ (para se obter estacionariedade).

Variograma (Cont.)

- Os pontos componentes do variograma amostral são calculados a partir de duas observações da mesma unidade amostral como $v_{ijk} = \frac{1}{2}(\Delta_{ij} - \Delta_{ik})^2$.
- Plota-se v_{ijk} em função de $u_{ijk} = |t_{ij} - t_{ik}|$ (distância entre as condições de avaliação, também conhecido como "lag").
- Estima-se σ^2 através de

$$\widehat{\sigma}^2 = \frac{1}{2Nk} \sum_{i \neq l} \sum_{j,k} \frac{1}{2} (\Delta_{ij} - \Delta_{lk})^2 = \frac{1}{2Nk} \sum_{i \neq l} \sum_{j,k} v_{ijkl},$$

Variograma (Cont.)

- em que k é a quantidade de termos de $\sum_{j,k}$, $\upsilon_{ijlk} = \frac{1}{2}(\Delta_{ij} - \Delta_{lk})^2$ e N é o número de pares de observações obtidas em unidades experimentais diferentes.
- Como o processo é estacionário e sob independência entre as observações de diferentes indivíduos, temos que $\widehat{\sigma}^2$ é um estimador não viciado de σ^2.
- Como $\rho(u) = 1 - \frac{g(u)}{\sigma^2}$, quanto mais próximos de $\widehat{\sigma}^2$ forem os valores de $\widehat{g}(u)$, menor o valor da correlação para a defasagem u.

Variograma