Modelos de regressão para dados correlacionados

Cibele Russo

cibele@icmc.usp.br

ICMC USP

Mini-curso oferecido no
Workshop on Probabilistic and Statistical Methods

28 a 30 de janeiro de 2013

Conteúdo da aula

1. **Modelos não lineares**

2. **Modelos não lineares com efeitos mistos:**

    1. Onde introduzir efeitos aleatórios?

    2. Qual o resultado da introdução dos efeitos aleatórios?

    3. Estimação no modelo não linear com efeitos mistos

Referências: Pinheiro and Bates (2000) 'Mixed-effects Models in S and S-PLUS', Springer.

Russo, Paula and Aoki (2009): Influence diagnostics in nonlinear mixed-effects elliptical models. Computational Statistics & Data Analysis 53(12): 4143–4156.

Modelos não lineares

Grande parte da teoria de modelos de regressão está desenvolvida para modelos lineares.

Em muitos casos, lineariza-se o problema e um modelo linear é ajustado.

Problema: os parâmetros em modelos lineares têm sua interpretação, como nos modelos lineares. Porém, ao linearizar os dados, os parâmetros perdem sua interpretação, que muitas vezes é o principal foco do estudo.

Modelos não lineares

Um modelo não linear para observações independentes é dado por

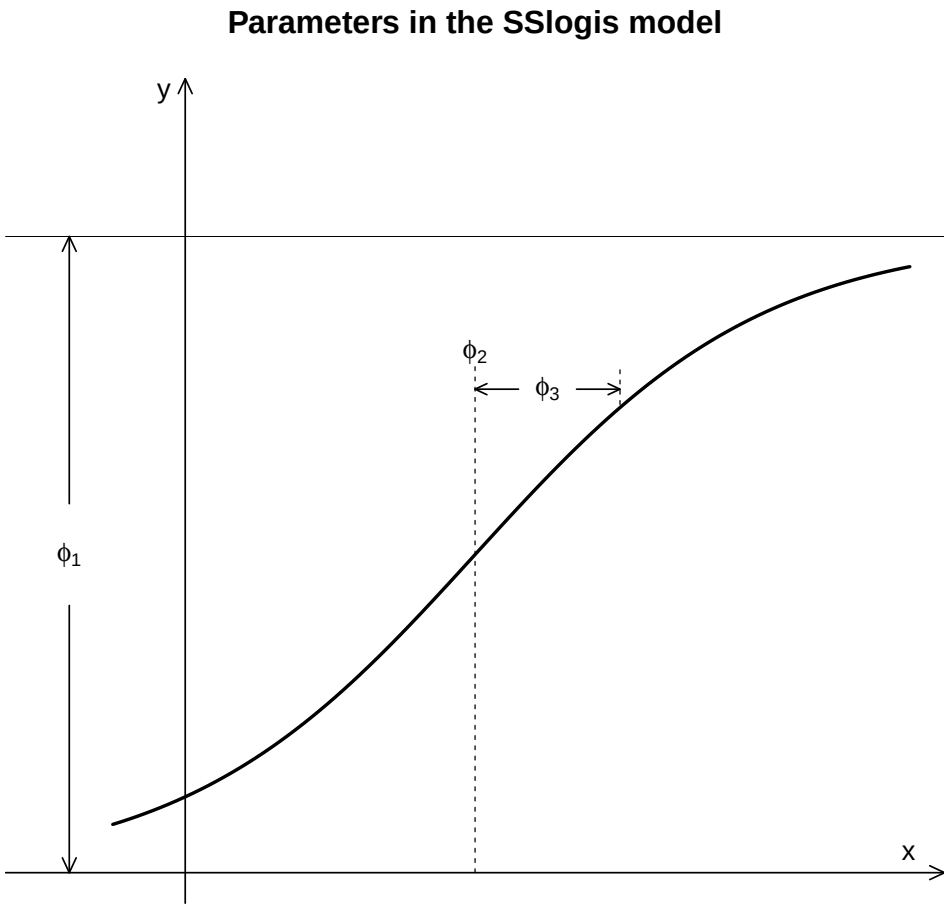
$$Y_i = \eta(\boldsymbol{\beta}, \mathbf{X}_i) + \epsilon_i,$$

com

- Y_i é a variável resposta (v. aleatória).
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ vetor de parâmetros desconhecidos,
- η é uma função de $\boldsymbol{\beta}$ e $\mathbf{X}_i$, não linear em $\boldsymbol{\beta}$,
- $\mathbf{X}_i$ é vetor de variáveis explicativas com valores conhecidos (v. não aleatória),
- ϵ_i é o erro aleatório,
- Suposição usual: $\epsilon_i \overset{i.i.d.}{\sim} \mathrm{N}(\mathbf{0}, \sigma^2)$.

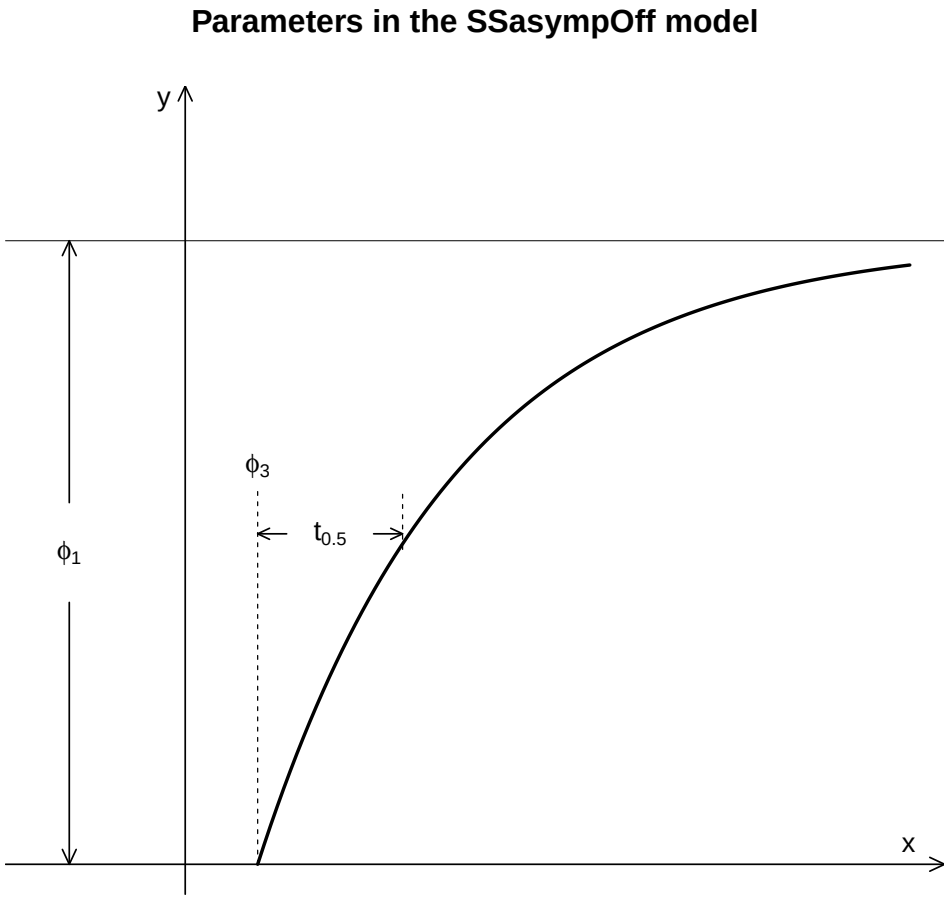
Exemplo: Modelo de crescimento logístico

```
> ?SSlogis
> example(SSlogis)
```



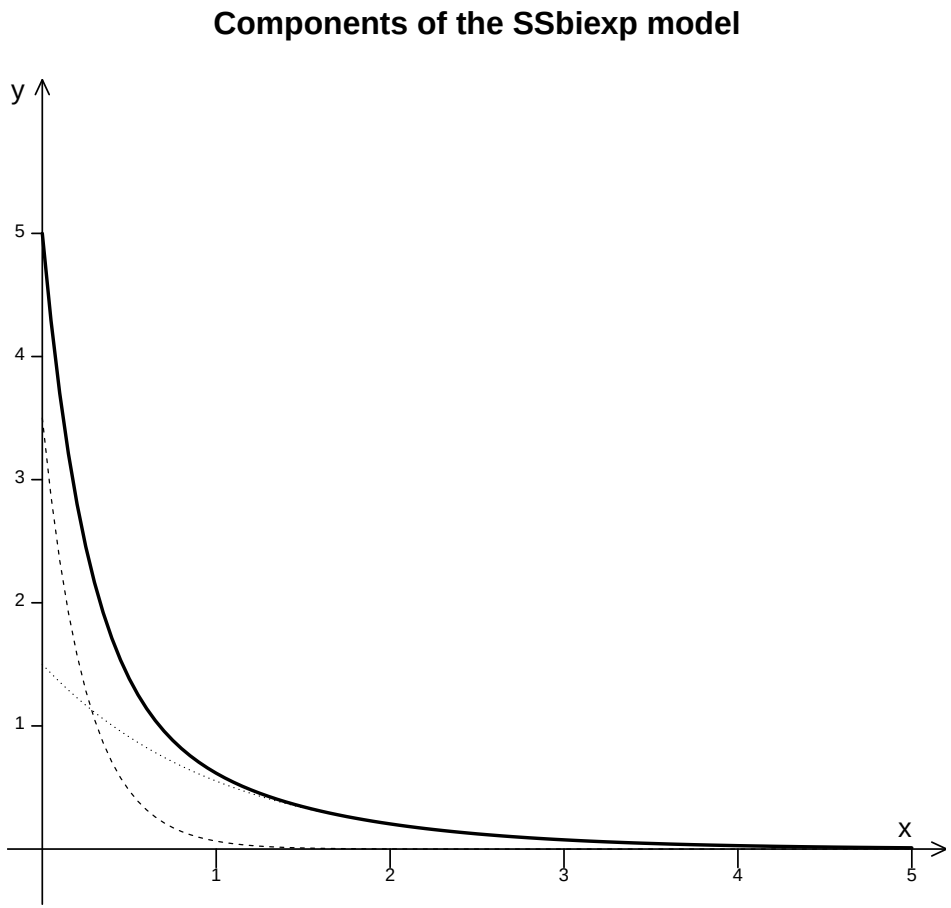
Exemplo: Modelo assintótico com um offset

```
> ?SSasympOff
> example(SSasympOff)
```



Exemplo: Modelo biexponencial

```
> ?SSbiexp
> example(SSbiexp)
```



Modelos não lineares

Mesmo sem adicionar efeitos aleatórios, já é difícil estimar os parâmetros no modelo não linear

Em geral, recorre-se a métodos iterativos como

- Algoritmo de Gauss-Newton
- Algoritmo de Newton-Raphson
- Método Escore de Fisher.

Mesmo assim, já foi mostrado que as estimativas dos parâmetros obtidas iterativamente são bastante precisas para os parâmetros de interesse. Muitas vezes, problemas não lineares envolvem dados correlacionados.

Motivação: Dados de hemodiálise

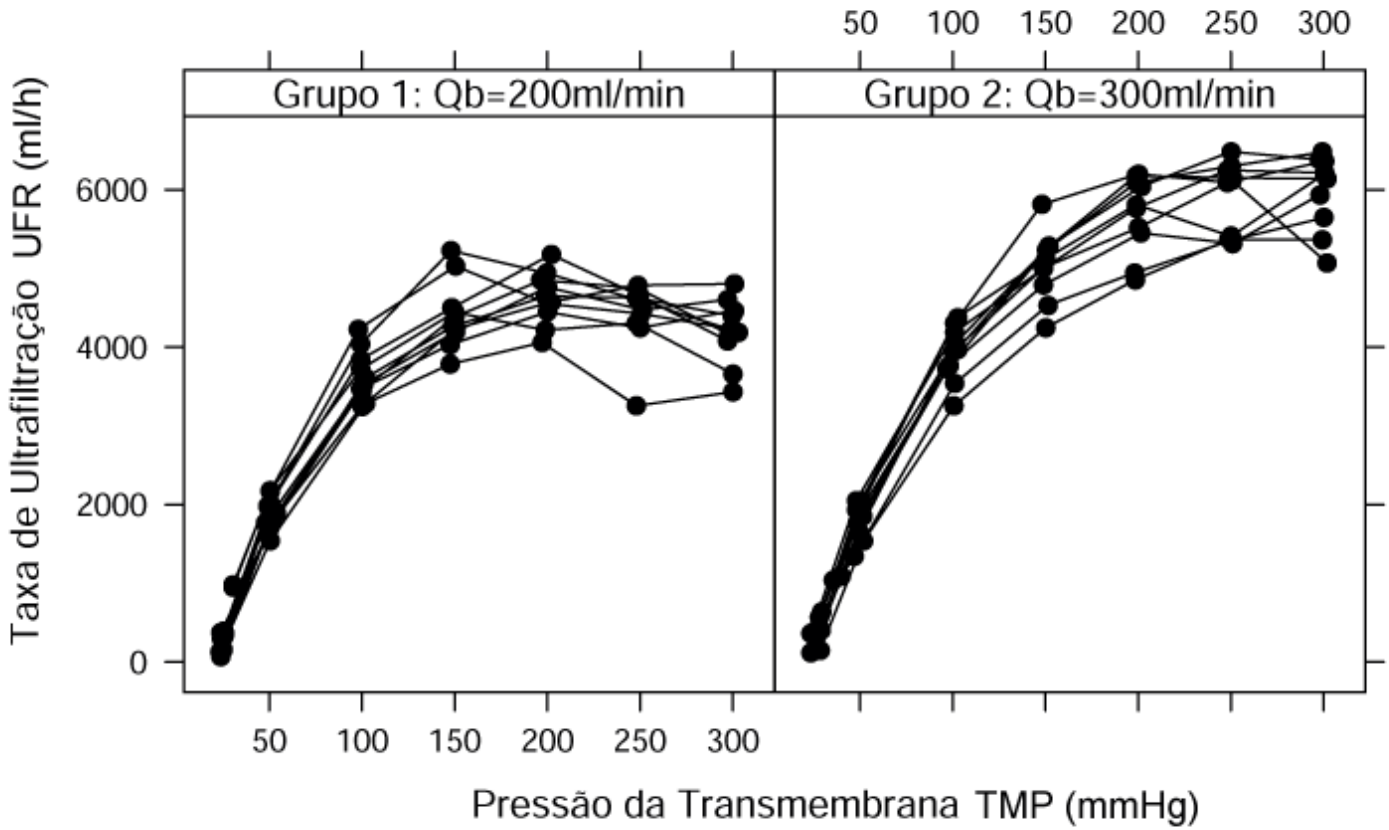
Problema cinético de hemodiálise (Vonesh & Carter, 1992)

- UFR: Taxa de ultrafiltração (ml/h)
- TMP: Pressão da transmembrana (mmHg)

$$\mathrm{E(UFR)} = \beta_0\{1 - \exp[-\beta_1(\mathrm{TMP} - \beta_2)]\}$$

- β_0: UFR máxima que pode ser alcançada
- β_1: taxa de transporte de permeabilização hidráulica e
- β_2: TMP requerida para contrabalancear a pressão oncótica do paciente.

Motivação: Dados de hemodiálise



Dados de hemodiálise

Calculando a covariância amostral

```
> attach(Dialyzer)
> cov(matrix(rate,nrow=20,byrow=T))
> detach(Dialyzer)
```

```
> cov(matrix(Dialyzer$rate, nrow=20, byrow=T))
           [,1]      [,2]      [,3]      [,4]      [,5]       [,6]       [,7]
[1,] 10.6706226  1.771906  5.003768  4.512624 -1.149745  0.7237066   2.897446
[2,]  1.7719058  4.888669  4.901217  4.996895  2.131893  1.2687543  -1.508709
[3,]  5.0037679  4.901217 12.507406 15.930639 15.365545 15.3622924  16.156541
[4,]  4.5126237  4.996895 15.930639 28.056529 30.137767 34.8690316  36.294999
[5,] -1.1497453  2.131893 15.365545 30.137767 46.319127 54.3120780  61.037006
[6,]  0.7237066  1.268754 15.362292 34.869032 54.312078 76.4763602  79.380041
[7,]  2.8974458 -1.508709 16.156541 36.294999 61.037006 79.3800414 101.656087
>
```

Dados de hemodiálise

Calculando a correlação amostral

```
> library(nlme)
> attach(Dialyzer)
> cor(matrix(rate,nrow=20,byrow=T))
> detach(Dialyzer)
```

```
> cor(matrix(Dialyzer$rate, nrow=20, byrow=T))
            [,1]       [,2]      [,3]      [,4]       [,5]       [,6]       [,7]
[1,]  1.00000000  0.24532987 0.4331304 0.2608058 -0.0517162 0.02533402  0.08797388
[2,]  0.24532987  1.00000000 0.6267942 0.4266659  0.1416739 0.06561734 -0.06767737
[3,]  0.43313039  0.62679423 1.0000000 0.8504183  0.6383872 0.49671674  0.45310417
[4,]  0.26080581  0.42666587 0.8504183 1.0000000  0.8360147 0.75276481  0.67961526
[5,] -0.05171620  0.14167394 0.6383872 0.8360147  1.0000000 0.91254151  0.88950081
[6,]  0.02533402  0.06561734 0.4967167 0.7527648  0.9125415 1.00000000  0.90028709
[7,]  0.08797388 -0.06767737 0.4531042 0.6796153  0.8895008 0.90028709  1.00000000
>
```

Motivação: Concentração de indomethacin

Cinética do antiinflamatório indomethacin (Bocheng & Xuping, 2001)

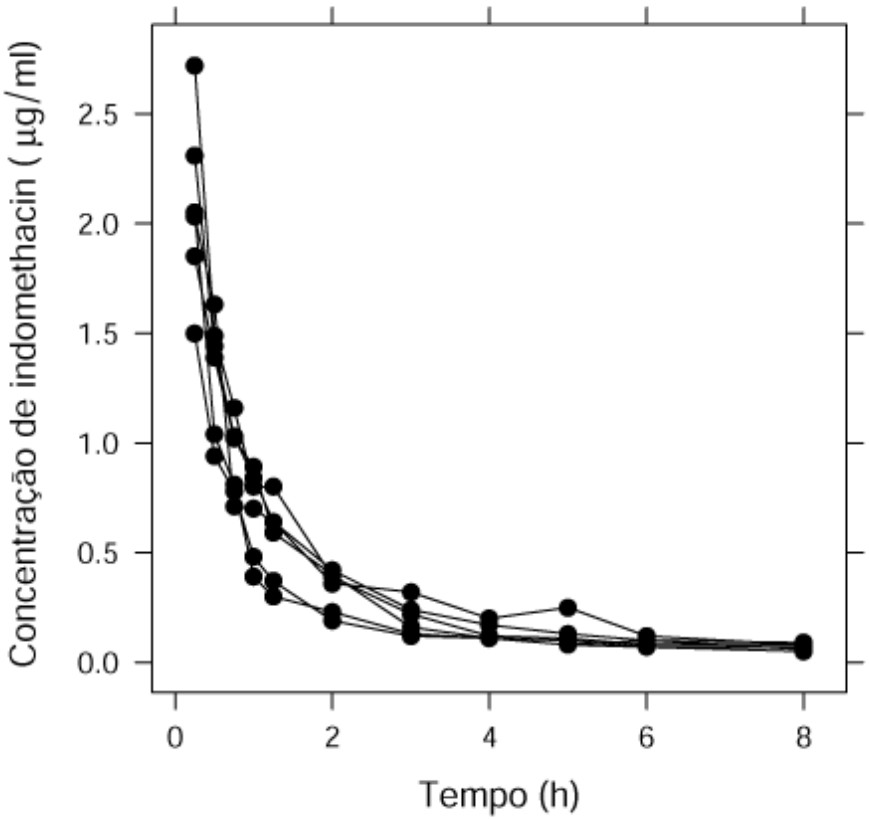
- Y: Concentração de indomethacin (μg/ml)
- T: Tempo (h)

$$\mathrm{E}(Y) = e^{\beta_1} \exp(-e^{\beta_2} T) + e^{\beta_3} \exp(-e^{\beta_4} T).$$

(função biexponencial, Pinheiro & Bates, 2000)

- e^{β_1} e e^{β_3}: coeficientes de uma combinação linear entre as duas parcelas $\exp(-e^{\beta_2} T)$ e $\exp(-e^{\beta_4} T)$
- β_2 e β_4: logaritmos de taxas de decaimento.

Motivação: Concentração de indomethacin



Motivação: Concentração de indomethacin

Exercício: Calcule a covariância e correlação amostrais para os dados de indomethacin.

```
> library(nlme)
> attach(Indometh)
> cov(matrix(conc,nrow=6,byrow=T))
> cor(matrix(conc,nrow=6,byrow=T))
> detach(Indometh)
```

Motivação: Concentração de theophylline

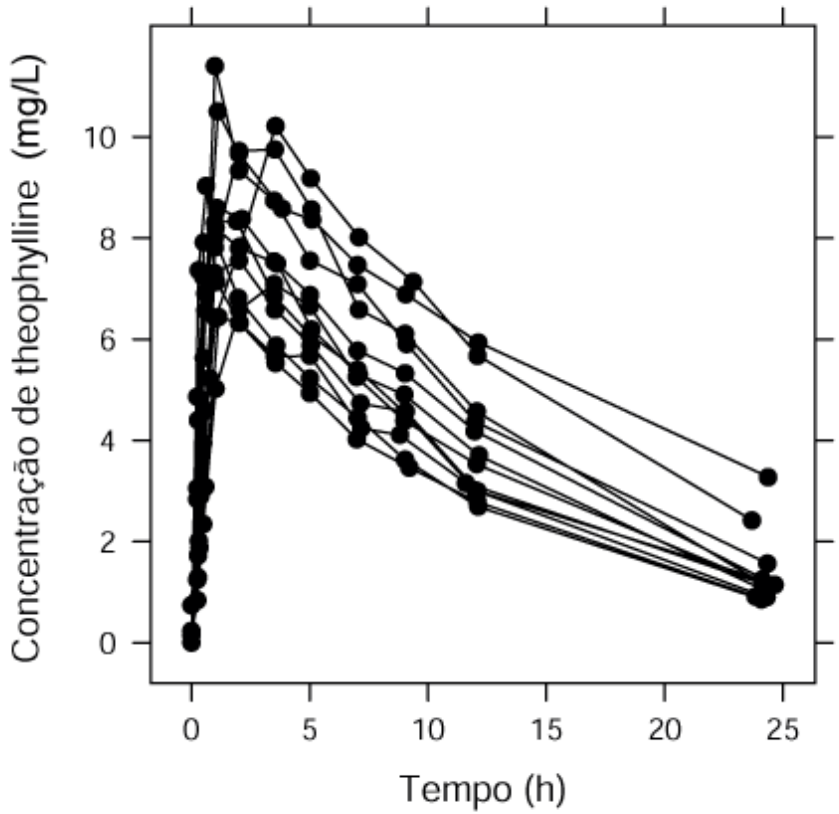
Cinética do agente antiasmático theophylline (Pinheiro & Bates 2000)

- Y: Concentração de theophylline (mg/L)
- T: Tempo (h)
- D: Dose aplicada (mg/kg)

$$\mathrm{E}(Y) = D\exp(lK_e + lK_a - lC_l)\frac{[\exp(-e^{lK_e}T) - \exp(-e^{lK_a}T)]}{e^{(lK_a)} - e^{lK_e}}$$

- lK_a: logaritmo da taxa de absorção da substância,
- lK_e: logaritmo da taxa de eliminação da substância e
- lC_l: o logaritmo da depuração plasmática.

Motivação: Concentração de theophylline



Motivação: Concentração de theophylline

Exercício: Calcule a covariância e correlação amostrais para os dados de theophylline.

```
> library(nlme)
> attach(Theoph)
> cov(matrix(conc,nrow=12,byrow=T))
> cor(matrix(conc,nrow=12,byrow=T))
> detach(Theoph)
```

Motivação: Crescimento de plantas de soja

Crescimento de plantas (Pinheiro & Bates, 2000)

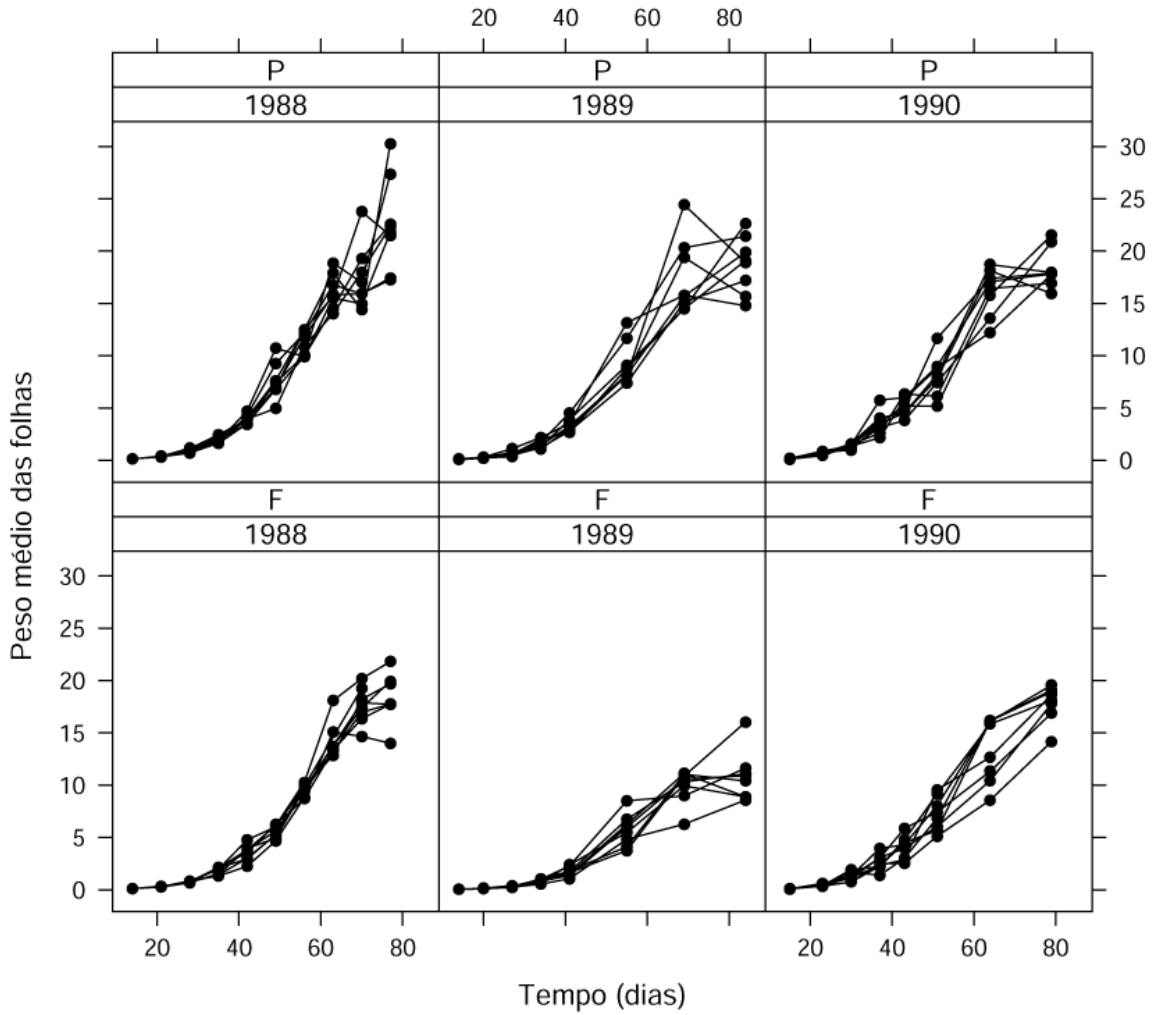
- Y: peso da folha (g)
- T: tempo (dias)

$$\mathrm{E}(Y) = \frac{\beta_1}{1 + \exp\left\{-\left[\frac{(T-\beta_2)}{\beta_3}\right]\right\}}$$

(modelo de crescimento logístico de três parâmetros)

- β_1 representa **o peso assintótico das folhas das plantas**,
- β_2 é o tempo em que a **folha** atinge a metade **de seu peso assintótico**
- β_3 representa o tempo decorrido entre os instantes em que a **folha** tem metade e $1/(1 + e^{-1}) \simeq 3/4$ de **seu peso**.

Motivação: Crescimento de plantas de soja



Motivação: Crescimento de plantas de soja

Novidade nos dados de plantas de soja
(muito comum em modelos mistos):

Dados desbalanceados
(grupos de observações com tamanhos diferentes).

Modelo não linear com efeitos mistos

A forma mais comum de definir um modelo não linear (multivariado) com efeitos mistos é considerar que

$$\begin{cases} \mathbf{Y}_i &= \mathbf{g}(\phi_i, \mathbf{X}_i) + \epsilon_i, \ i = 1, \ldots, n, \\ \phi_i &= \mathbf{A}_i\beta + \mathbf{b}_i, \end{cases}$$

supondo que

$$\mathbf{b}_i \sim N_q(\mathbf{0}, D)$$
$$\epsilon_i \sim N_{m_i}(\mathbf{0}, \sigma^2 I)$$
$$\mathbf{b}_1, \ldots, \mathbf{b}_n, \epsilon_1, \ldots, \epsilon_n \text{ são independentes}$$

Dificuldade: Não é possível obter o modelo marginal em forma fechada nesse caso.

Modelos não lineares com efeitos aleatórios

Outra formulação, mais geral, é dada por

$$\mathbf{Y}_i = \boldsymbol{\eta}(\boldsymbol{\beta}, \mathbf{b}_i, \mathbf{x}_i) + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n,$$

em que $\mathbf{b}_i \sim \mathrm{N}_r(\mathbf{0}, \mathbf{D})$ e $\boldsymbol{\epsilon}_i \sim \mathrm{N}_{m_i}(\mathbf{0}, \sigma^2 \mathbf{I}_{m_i})$.

Dificuldade: a distribuição marginal de $\mathbf{Y}_i$ não tem forma fechada, sendo necessário o uso de métodos de integração numérica, como métodos de quadratura por exemplo, para a obtenção de formas aproximadas para as marginais.

Modelos não lineares com efeitos aleatórios

Consideramos o modelo analisado por Russo et al (2009) (no contexto de distribuições elípticas)

$$\mathbf{Y}_i = \boldsymbol{\eta}(\boldsymbol{\beta}, \mathbf{x}_i) + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n,$$

- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ vetor de parâmetros desconhecidos,
- $\boldsymbol{\eta}(\boldsymbol{\beta}, \mathbf{x}_i) = (f(\boldsymbol{\beta}, \mathbf{x}_{i1}), \ldots, f(\boldsymbol{\beta}, \mathbf{x}_{im_i}))^\top$ função não linear de $\boldsymbol{\beta}$,
- $\mathbf{x}_i$ vetor de variáveis explicativas,
- $\mathbf{Z}_i$ matriz constante conhecida,
- $\mathbf{b}_i = (b_{i1}, \ldots, b_{ir})^\top$ vetor de efeitos aleatórios não observados.

Modelos não lineares com efeitos aleatórios

Suposições:

$$\begin{bmatrix} \mathbf{Y}_i \\ \mathbf{b}_i \end{bmatrix} \sim \mathrm{N}_{m_i+r} \left\{ \begin{pmatrix} \boldsymbol{\eta}(\boldsymbol{\beta}, \mathbf{x}_i) \\ \mathbf{0} \end{pmatrix}, \begin{bmatrix} \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^\top + \sigma^2 \mathbf{I}_{m_i} & \mathbf{Z}_i \mathbf{D} \\ \mathbf{D} \mathbf{Z}_i^\top & \mathbf{D} \end{bmatrix} \right\},$$

- $V_i = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^\top + \sigma^2 \mathbf{I}_{m_i}$ é a matriz de variâncias e covariâncias $\mathrm{Var}(\mathbf{Y}_i)$,
- $\mathbf{D} = \mathbf{D}(\boldsymbol{\tau})$ é a matriz de variâncias e covariâncias $\mathrm{Var}(\mathbf{b}_i)$
- $\mathbf{Z}_i \mathbf{D}$é a matriz de variâncias e covariâncias$\mathrm{Cov}(\mathbf{Y}_i, \mathbf{b}_i)$.

Modelo marginal: $\mathbf{Y}_i \sim \mathrm{N}_{m_i}(\boldsymbol{\eta}(\boldsymbol{\beta}, \mathbf{x}_i); V_i)$

Modelos não lineares com efeitos aleatórios

Possíveis expressões para $\mathbf{Z}_i$ $(i = 1, \ldots, n)$:

(i) $\mathbf{Z}_i = \mathbf{1}$,

(ii) $\mathbf{Z}_i = (\mathbf{1}, \mathbf{x}_i)$,

(iii) $\mathbf{Z}_i = (\mathbf{1}, \mathbf{x}_i, \mathbf{x}_i^2)$, e

(iv) modelos mistos pseudo não lineares (PNMEM), em que PNMEM$_{\beta_0\beta_1\beta_2}$ indica que

$$\mathbf{Z}_i = \left[\frac{\partial \boldsymbol{\eta}(\boldsymbol{\beta}, \mathbf{x}_i)}{\partial \beta_0}, \frac{\partial \boldsymbol{\eta}(\boldsymbol{\beta}, \mathbf{x}_i)}{\partial \beta_1}, \frac{\partial \boldsymbol{\eta}(\boldsymbol{\beta}, \mathbf{x}_i)}{\partial \beta_2}\right]\Bigg|_{\beta=\tilde{\beta}},$$

em que $\tilde{\boldsymbol{\beta}}$ é a estimativa de mínimos quadrados de $\boldsymbol{\beta}$.

Modelo não linear com efeitos mistos

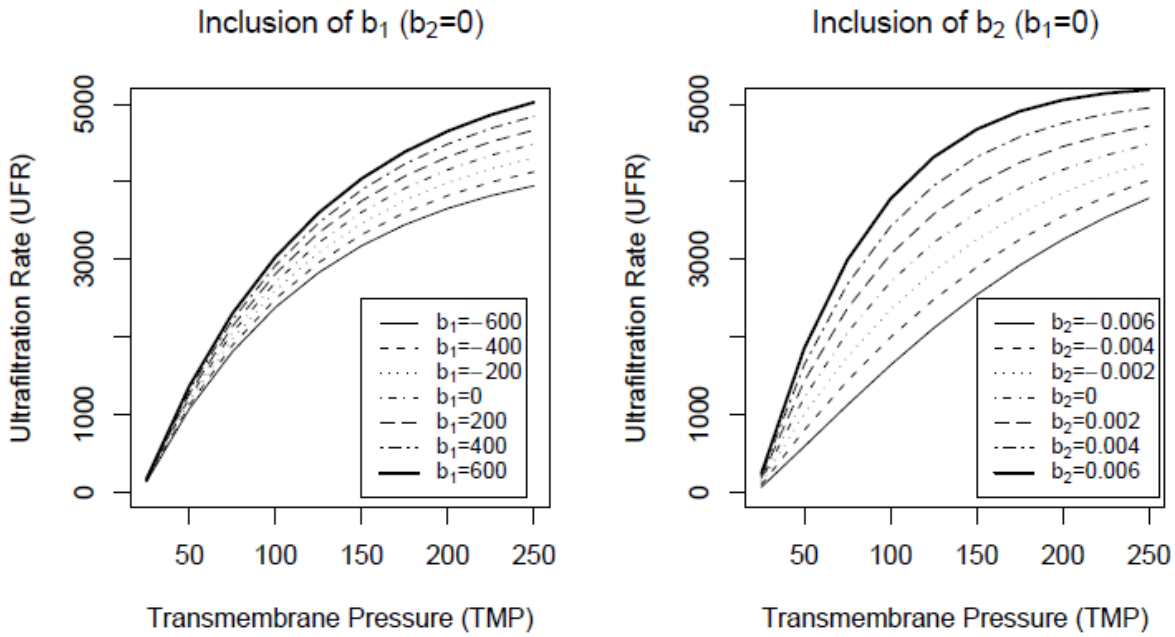
Alternativa: Considerar algoritmos que linearizam o modelo mais geral.

Modelo não linear com efeitos mistos

Importante: Na maioria das vezes, não é interessante introduzir efeito aleatório no intercepto, por exemplo. Os efeitos adicionados nos parâmetros referentes aos efeitos fixos levam a diferentes perfis dependendo da não-linearidade do problema. Mesmo assim, alguns autores adicionam efeito no intercepto sem fazer a devida interpretação.

Interpretando os efeitos aleatórios em modelos não lineares

Dados de hemodiálise

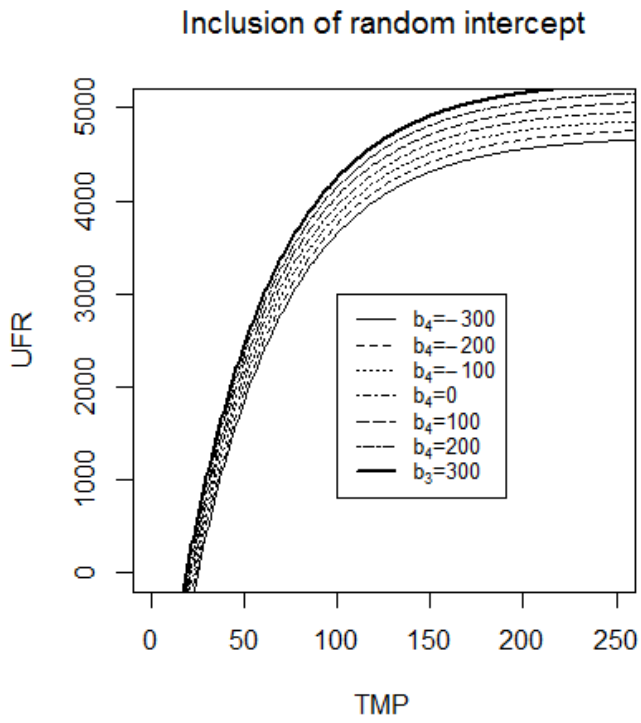


Efeitos aleatórios adicionados nos parâmetros de efeitos fixos, dentro da função não linear.

(ver dados originais - slide 10)

Interpretando os efeitos aleatórios em modelos não lineares

Dados de hemodiálise

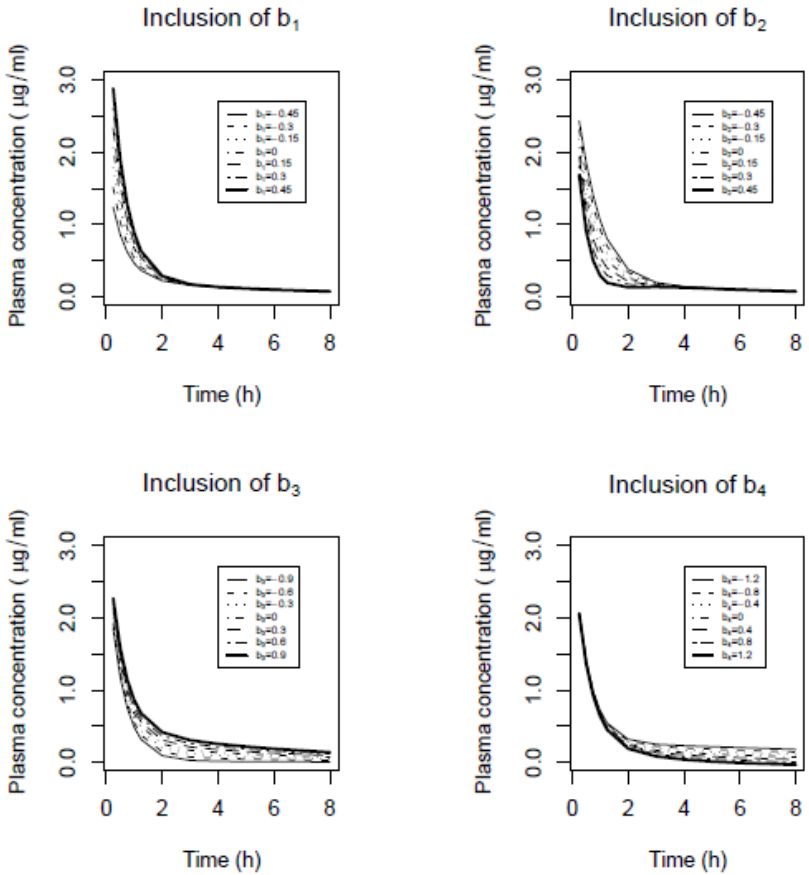


Resultado da inclusão de intercepto aleatório.

(ver dados originais - slide 10)

Interpretando os efeitos aleatórios em modelos não lineares

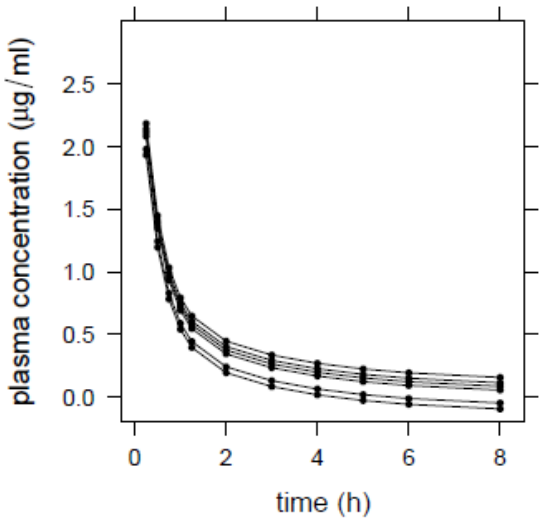
Dados de concentração de indomethacin



Efeitos aleatórios adicionados nos parâmetros de efeitos fixos, dentro da função não linear.

Interpretando os efeitos aleatórios em modelos não lineares

Dados de concentração de indomethacin

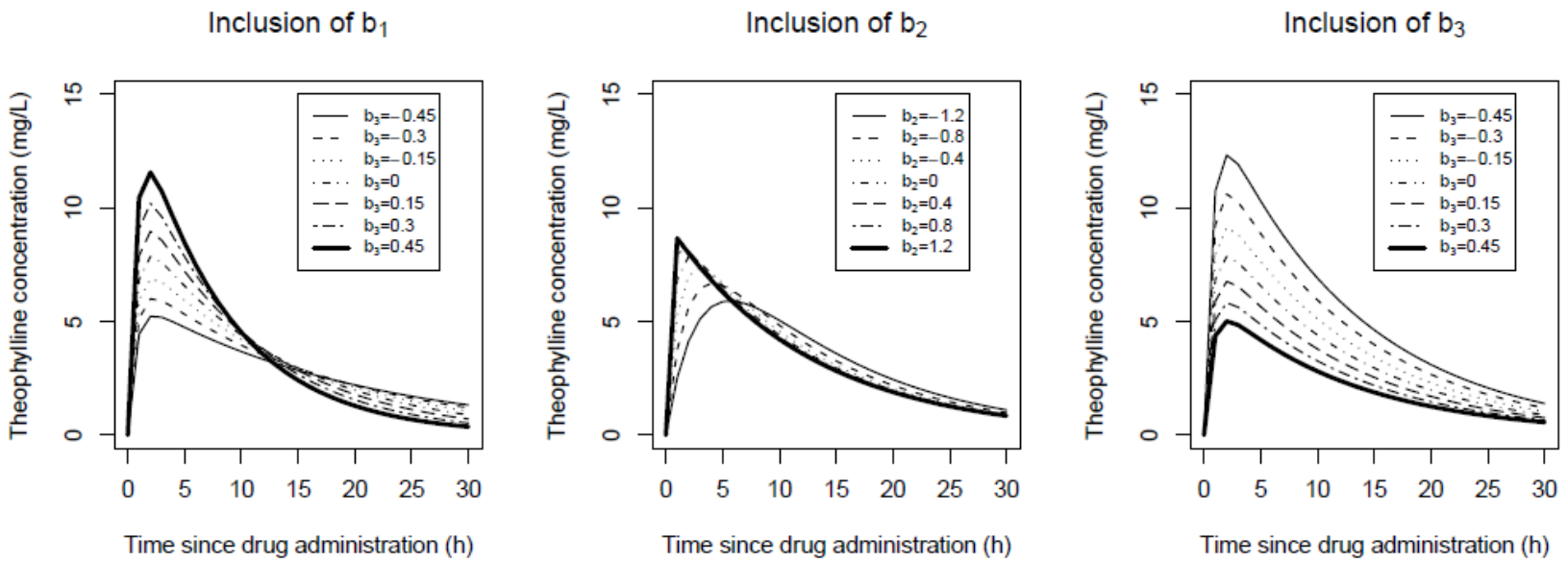


Resultado da inclusão de intercepto aleatório.

(ver dados originais - slide 14)

Interpretando os efeitos aleatórios em modelos não lineares

Dados de concentração de theophylline

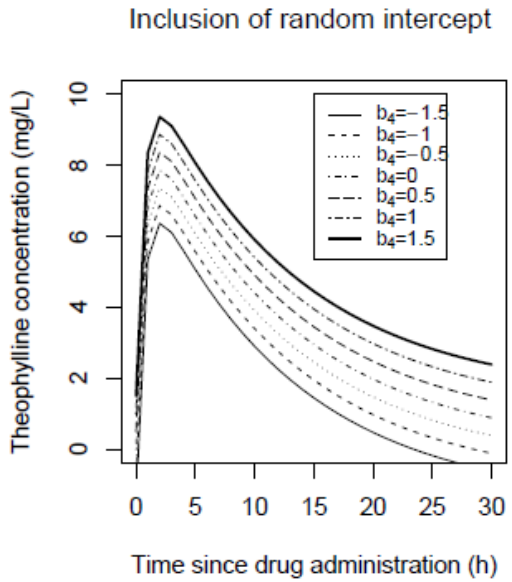


Efeitos aleatórios adicionados nos parâmetros de efeitos fixos, dentro da função não linear.

(ver dados originais - slide 17)

Interpretando os efeitos aleatórios em modelos não lineares

Dados de concentração de theophylline

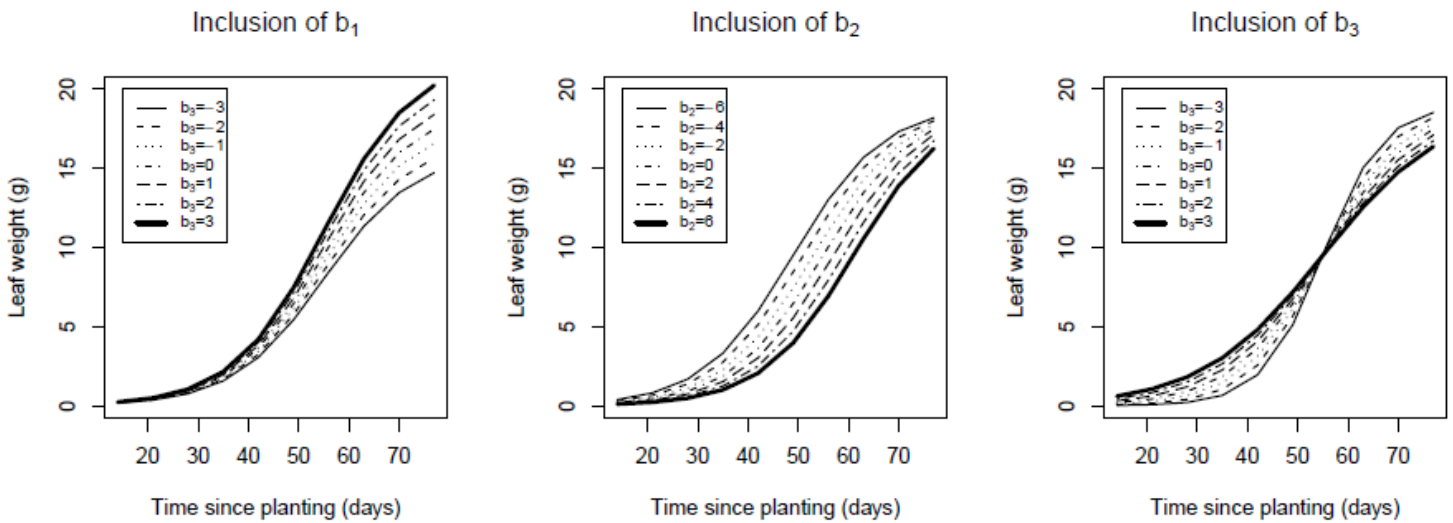


Resultado da inclusão de intercepto aleatório.

(ver dados originais - slide 17)

Interpretando os efeitos aleatórios em modelos não lineares

Dados de crescimento de plantas de soja

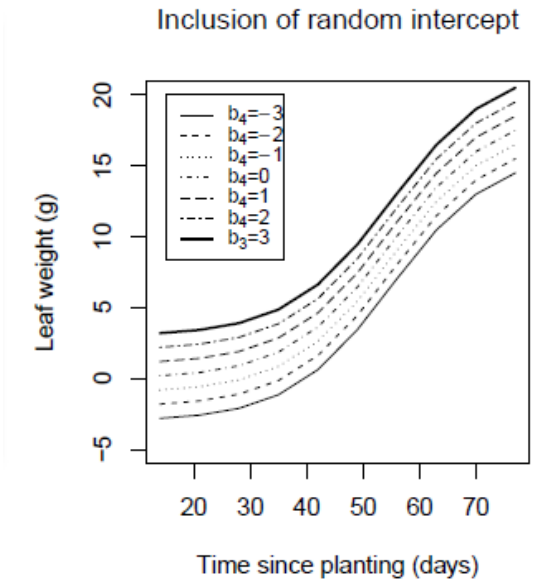


Efeitos aleatórios adicionados nos parâmetros de efeitos fixos, dentro da função não linear.

(ver dados originais - slide 20)

Interpretando os efeitos aleatórios em modelos não lineares

Dados de crescimento de plantas de soja



Resultado da inclusão de intercepto aleatório.

(ver dados originais - slide 20)

Comandos em R:

Ver arquivo Aula4Comandos.r

Próxima aula:

Análise de diagnóstico e seleção de modelos