MAT02035 - Modelos para dados correlacionados

Métodos de análise descritiva para dados longitudinais: uma introdução

Rodrigo Citton Padilha dos Reis
citton.padilha@ufrgs.br

Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Departamento de Estatística

Porto Alegre, 2023

Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

Questões preliminares

Carregando os dados

```
# install.packages("haven")
library(haven)

chumbo <- read_dta(file = "tlc.dta")
```

Remodelando os dados

- Os dados de um estudo são frequentemente inseridos em um formato **largo** (*wide*), em que cada linha é um local/indivíduo/paciente e as medidas repetidas e covariáveis observadas dispostas em colunas.

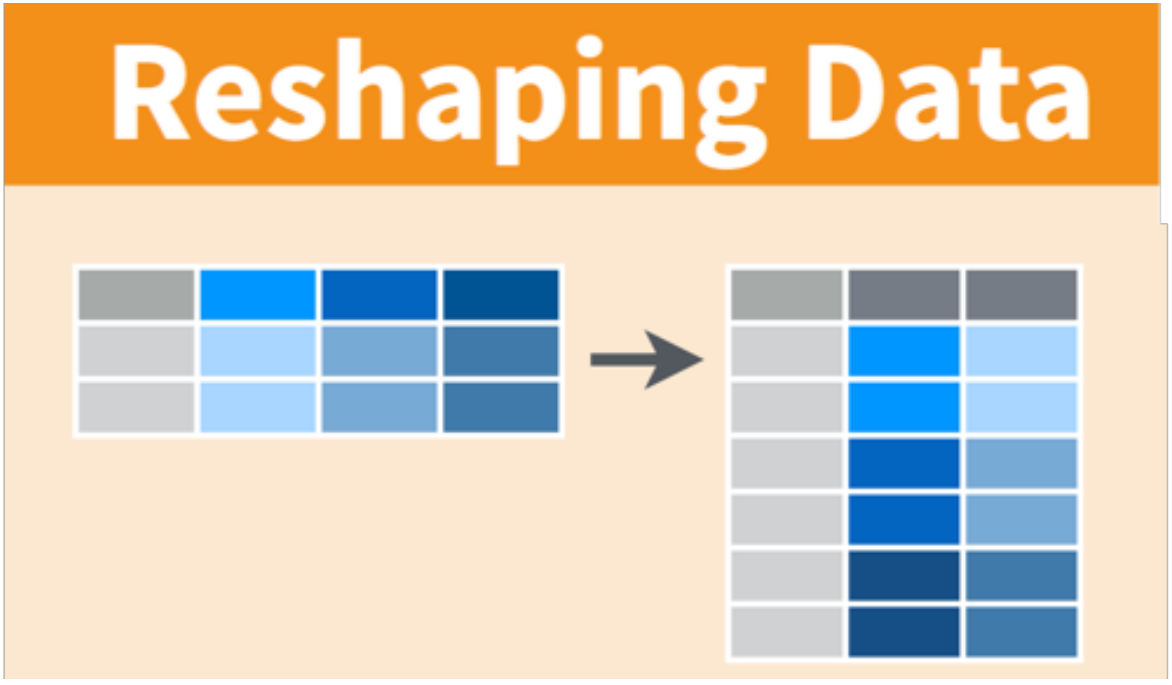
```
head(chumbo)
```

```
## # A tibble: 6 x 6
##      id trt              y0    y1    y4    y6
##   <dbl> <dbl+lbl>    <dbl> <dbl> <dbl> <dbl>
## 1     1 0 [Placebo]   30.8 26.9  25.8  23.8
## 2     2 1 [Succimer]  26.5 14.8  19.5  21
## 3     3 1 [Succimer]  25.8 23    19.1  23.2
## 4     4 0 [Placebo]   24.7 24.5  22    22.5
## 5     5 1 [Succimer]  20.4  2.80  3.20  9.40
## 6     6 1 [Succimer]  20.4  5.40  4.5  11.9
```

Remodelando os dados

- O formato **largo** é intuitivo para entrada de dados, mas nem tanto para análise de dados.
- Nas análises de dados em geral, mas principalmente no caso de dados longitudinais, os dados devem ser remodelados.
- Um formato mais apropriado para a maioria das análises que iremos realizar é o formato **longo** (*long*).
  - As medidas repetidas são empilhadas em uma única coluna.
  - A coluna id, e demais covariáveis fixas no tempo, repetem o seu valor.
  - Uma nova coluna que indexa as ocasiões, ou com os valores dos tempos de medição, é criada.

Remodelando os dados



Remodelando os dados

De “largo” para “longo”

```
# install.packages("tidyr")
library(tidyr)

chumbo.longo <- gather(data = chumbo,
                       key = "tempo",
                       value = "chumbo", -id, -trt)
```

Remodelando os dados

```
head(chumbo.longo)
```

```
## # A tibble: 6 x 4
##      id trt           tempo chumbo
##   <dbl> <dbl+lbl>     <chr>  <dbl>
## 1     1 0 [Placebo]   y0      30.8
## 2     2 1 [Succimer]  y0      26.5
## 3     3 1 [Succimer]  y0      25.8
## 4     4 0 [Placebo]   y0      24.7
## 5     5 1 [Succimer]  y0      20.4
## 6     6 1 [Succimer]  y0      20.4
```

Remodelando os dados

Pacotes para remodelar dados

Os pacotes `reshape` e `reshape2` foram substituídos pelo pacote `tidyr`.
Para maiores detalhes das funções do pacote, consulte:

- https://tidyr.tidyverse.org/
- https://github.com/rstudio/cheatsheets/blob/master/data-import.pdf

Transformando os dados

```
chumbo.longo$tempo <- as.numeric(
  as.character(factor(chumbo.longo$tempo,
                      labels = c(0, 1, 4, 6))))

chumbo.longo$trt <- factor(chumbo.longo$trt,
                           labels = c("Placebo", "Succimer"))
```

Transformando os dados

```
head(chumbo.longo)
```

```
## # A tibble: 6 x 4
##      id trt       tempo chumbo
##   <dbl> <fct>     <dbl>  <dbl>
## 1     1 Placebo       0   30.8
## 2     2 Succimer      0   26.5
## 3     3 Succimer      0   25.8
## 4     4 Placebo       0   24.7
## 5     5 Succimer      0   20.4
## 6     6 Succimer      0   20.4
```

Time plot

Time plot

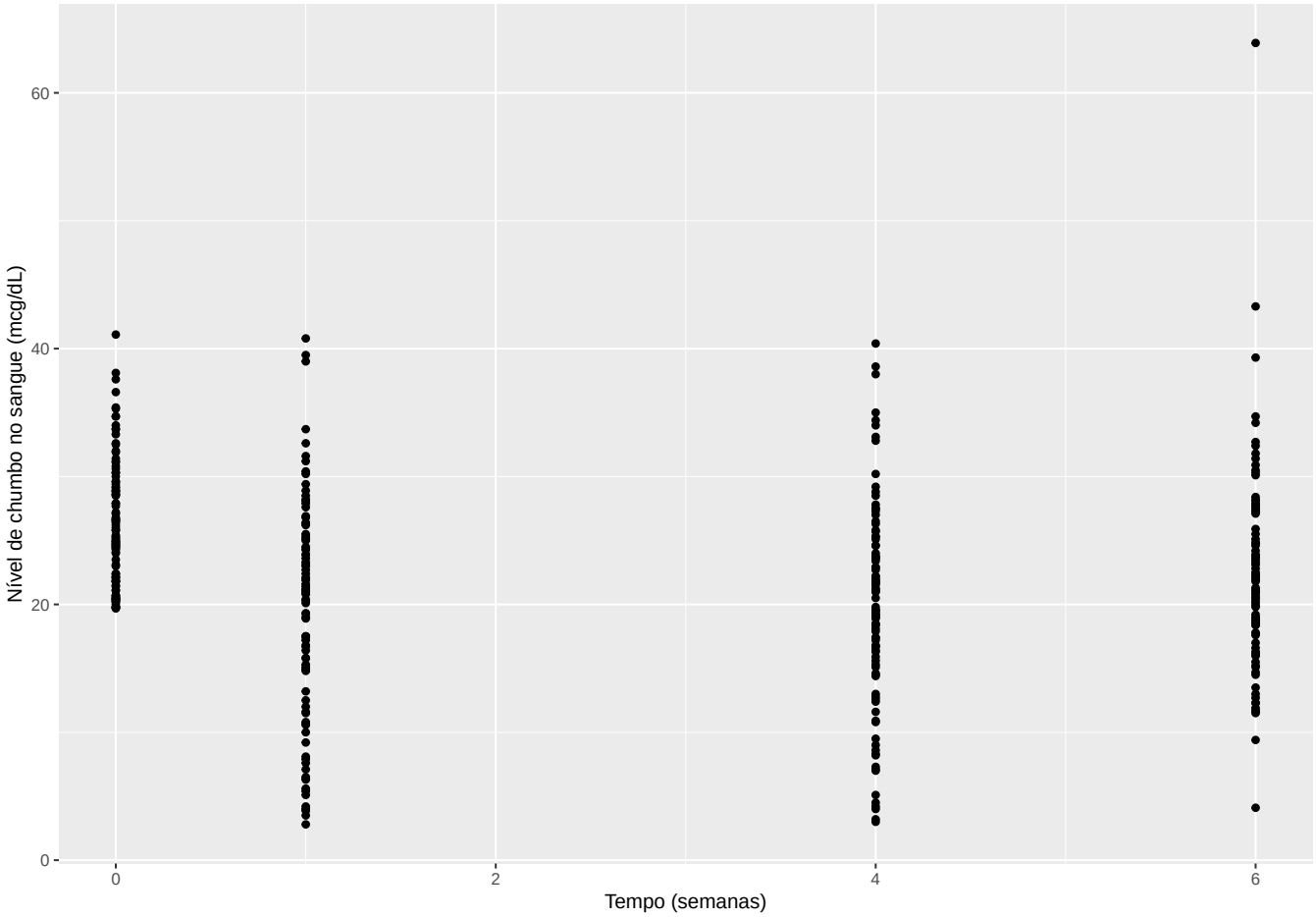
- Apresenta os valores da variável resposta no eixo vertical e os tempos das medidas no eixo horizontal, com medidas repetidas sucessivas no mesmo indivíduo unidas por linhas retas.
- É utilizado para observar as trajetórias individuais, tendências, variabilidade entre indivíduos e dentro de indivíduos.

Diagrama de dispersão

```
# install.packages("ggplot2")
library(ggplot2)

p <- ggplot(data = chumbo.longo,
            mapping = aes(x = tempo, y = chumbo)) +
  geom_point() +
  labs(x = "Tempo (semanas)",
       y = "Nível de chumbo no sangue (mcg/dL)")
p
```

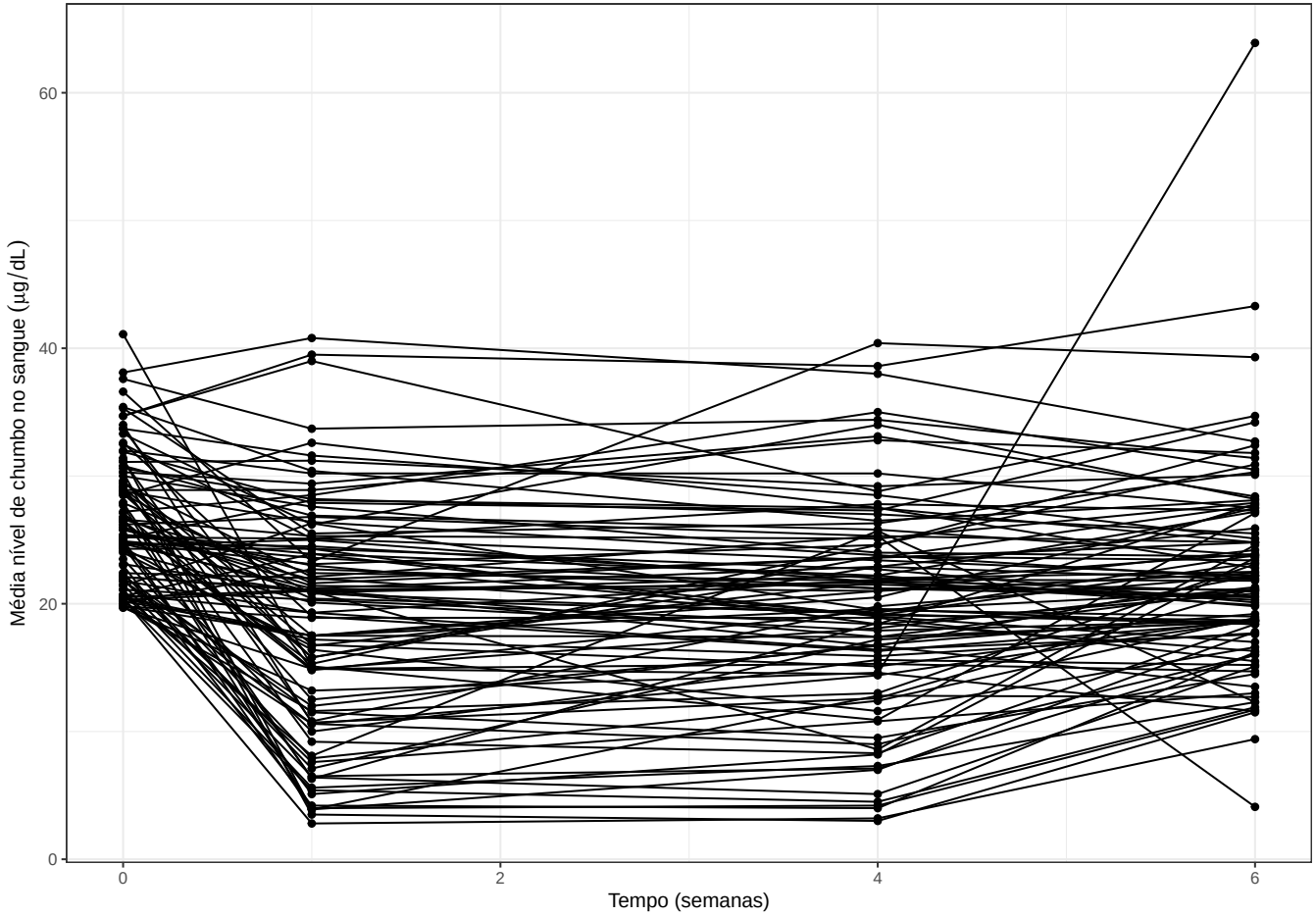
Diagrama de dispersão



Perfis individuais (spaghetti)

```
p <- ggplot(data = chumbo.longo,
            mapping = aes(x = tempo, y = chumbo,
                          group = id)) +
  geom_point() + geom_line() +
  labs(x = "Tempo (semanas)",
       y = expression("Média nível de chumbo no sangue"~(mu*g/dL))) +
  theme_bw()
p
```

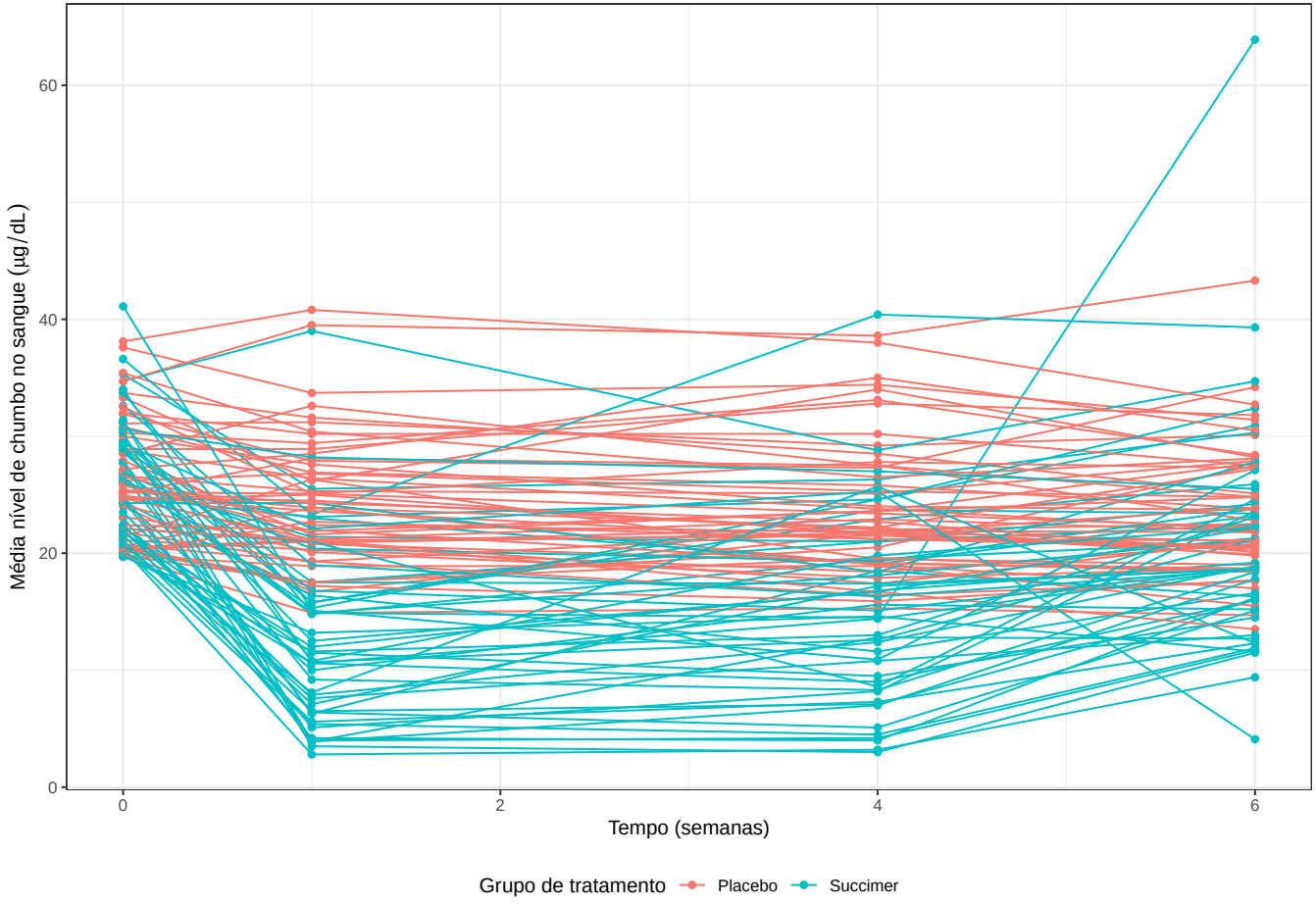
Perfis individuais (spaghetti)



Perfis individuais (spaghetti)

```
p <- ggplot(data = chumbo.longo,
            mapping = aes(x = tempo, y = chumbo,
                          group = id, colour = trt)) +
  geom_point() + geom_line() +
  labs(x = "Tempo (semanas)",
       y = expression("Média nível de chumbo no sangue"~(mu*g/dL)),
       colour = "Grupo de tratamento") +
  theme_bw() + theme(legend.position = "bottom")
p
```

Perfis individuais (spaghetti)



Perfis individuais (spaghetti)

- Com gráficos de perfis para dados longitudinais balanceados, pode ser difícil discernir o "sinal" (isto é, a tendência na resposta média ao longo do tempo) do "ruído" e as fontes de variabilidade não são distinguíveis.
- Em geral, é mais informativo exibir um gráfico de **perfis de médias**, com pontos sucessivos no gráfico unidos por linhas retas.

Perfis de médias

"Pré-processamento"

```
# install.packages("dplyr")
library(dplyr)

chumbo.resumo <- chumbo.longo %>%
  group_by(trt, tempo) %>%
  summarise(chumbo.m = mean(chumbo))
```

Perfis de médias

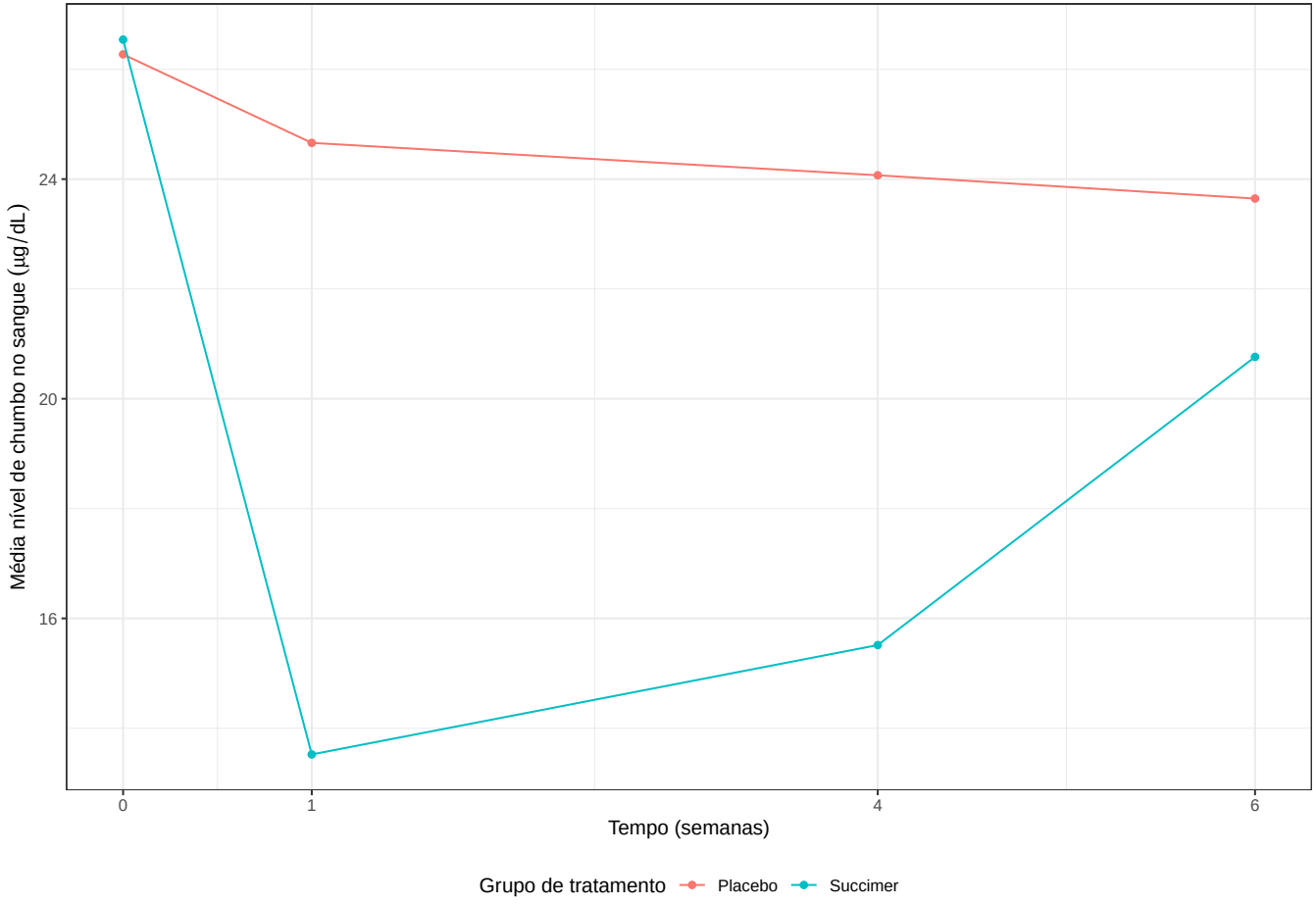
```
chumbo.resumo
```

```
## # A tibble: 8 x 3
## # Groups:   trt [2]
##   trt       tempo chumbo.m
##   <fct>     <dbl>    <dbl>
## 1 Placebo       0     26.3
## 2 Placebo       1     24.7
## 3 Placebo       4     24.1
## 4 Placebo       6     23.6
## 5 Succimer      0     26.5
## 6 Succimer      1     13.5
## 7 Succimer      4     15.5
## 8 Succimer      6     20.8
```

Perfis de médias

```
p <- ggplot(data = chumbo.resumo,
            mapping = aes(x = tempo, y = chumbo.m,
                          colour = trt)) +
  geom_point() + geom_line() +
  scale_x_continuous(breaks = c(0, 1, 4, 6)) +
  labs(x = "Tempo (semanas)",
       y = expression("Média nível de chumbo no sangue"~(mu*g/dL)),
       colour = "Grupo de tratamento") +
  theme_bw() + theme(legend.position = "bottom")
p
```

Perfis de médias



Perfis de médias com barras de erros

- Os gráficos de perfis médios também podem ser aprimorados incluindo **barras de erro padrão** para a resposta média em cada ocasião.

"Pré-processamento"

```
chumbo.resumo <- chumbo.longo %>%
  group_by(trt, tempo) %>%
  summarise(chumbo.m = mean(chumbo),
            dp = sd(chumbo), n = n()) %>%
  mutate(ep = dp/sqrt(n))
```

Perfis de médias com barras de erros

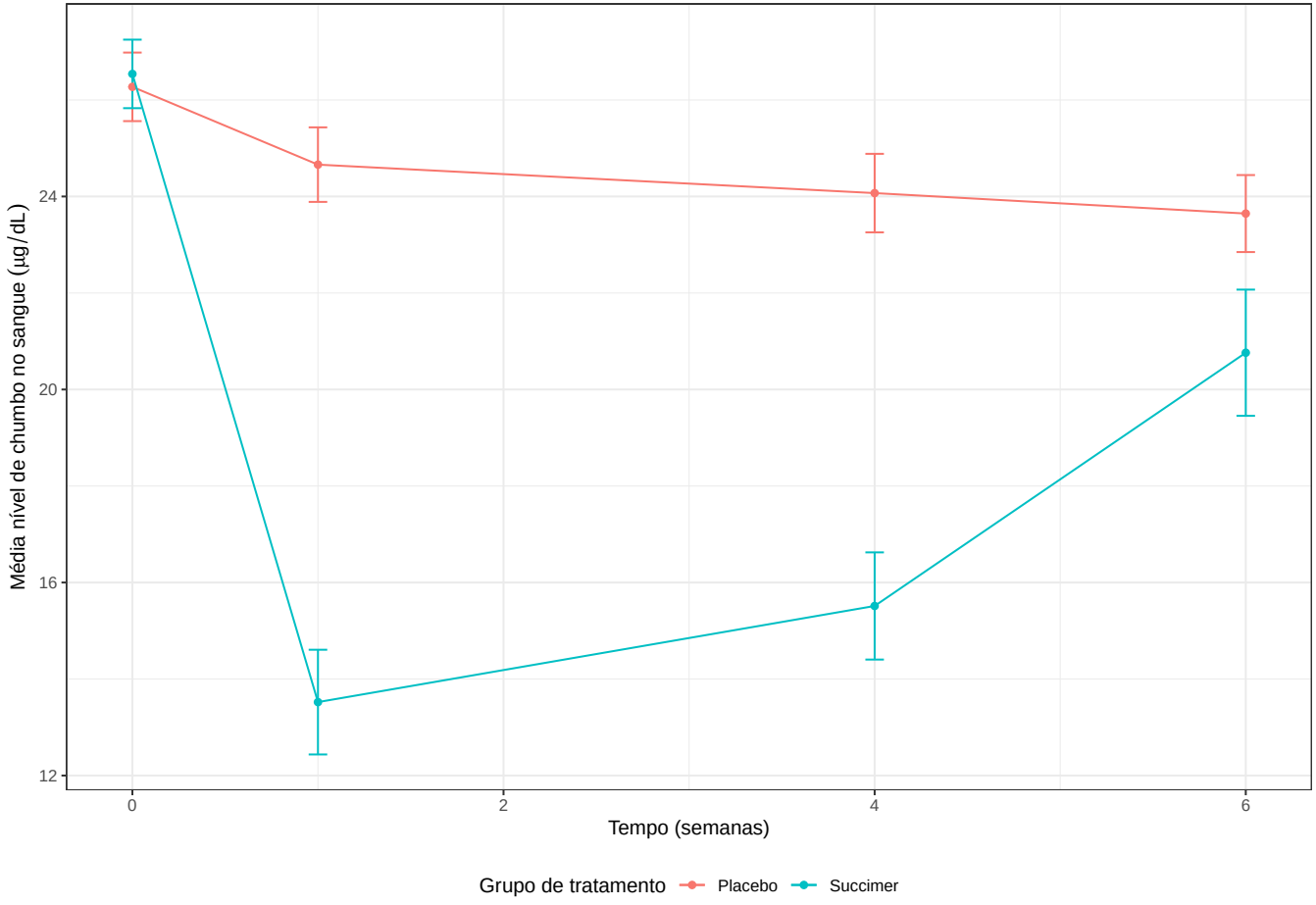
```
chumbo.resumo
```

```
## # A tibble: 8 x 6
## # Groups:   trt [2]
##   trt       tempo chumbo.m    dp     n    ep
##   <fct>     <dbl>    <dbl> <dbl> <int> <dbl>
## 1 Placebo       0     26.3  5.02    50 0.711
## 2 Placebo       1     24.7  5.46    50 0.772
## 3 Placebo       4     24.1  5.75    50 0.814
## 4 Placebo       6     23.6  5.64    50 0.798
## 5 Succimer      0     26.5  5.02    50 0.710
## 6 Succimer      1     13.5  7.67    50 1.09
## 7 Succimer      4     15.5  7.85    50 1.11
## 8 Succimer      6     20.8  9.25    50 1.31
```

Perfis de médias com barras de erros

```
p <- ggplot(data = chumbo.resumo,
            mapping = aes(x = tempo, y = chumbo.m,
                          colour = trt)) +
  geom_errorbar(aes(ymin = chumbo.m - ep,
                    ymax = chumbo.m + ep),
                width = .1) +
  geom_point() + geom_line() +
  labs(x = "Tempo (semanas)",
       y = expression("Média nível de chumbo no sangue"~(mu*g/dL)),
       colour = "Grupo de tratamento") +
  theme_bw() + theme(legend.position = "bottom")
p
```

Perfis de médias com barras de erros



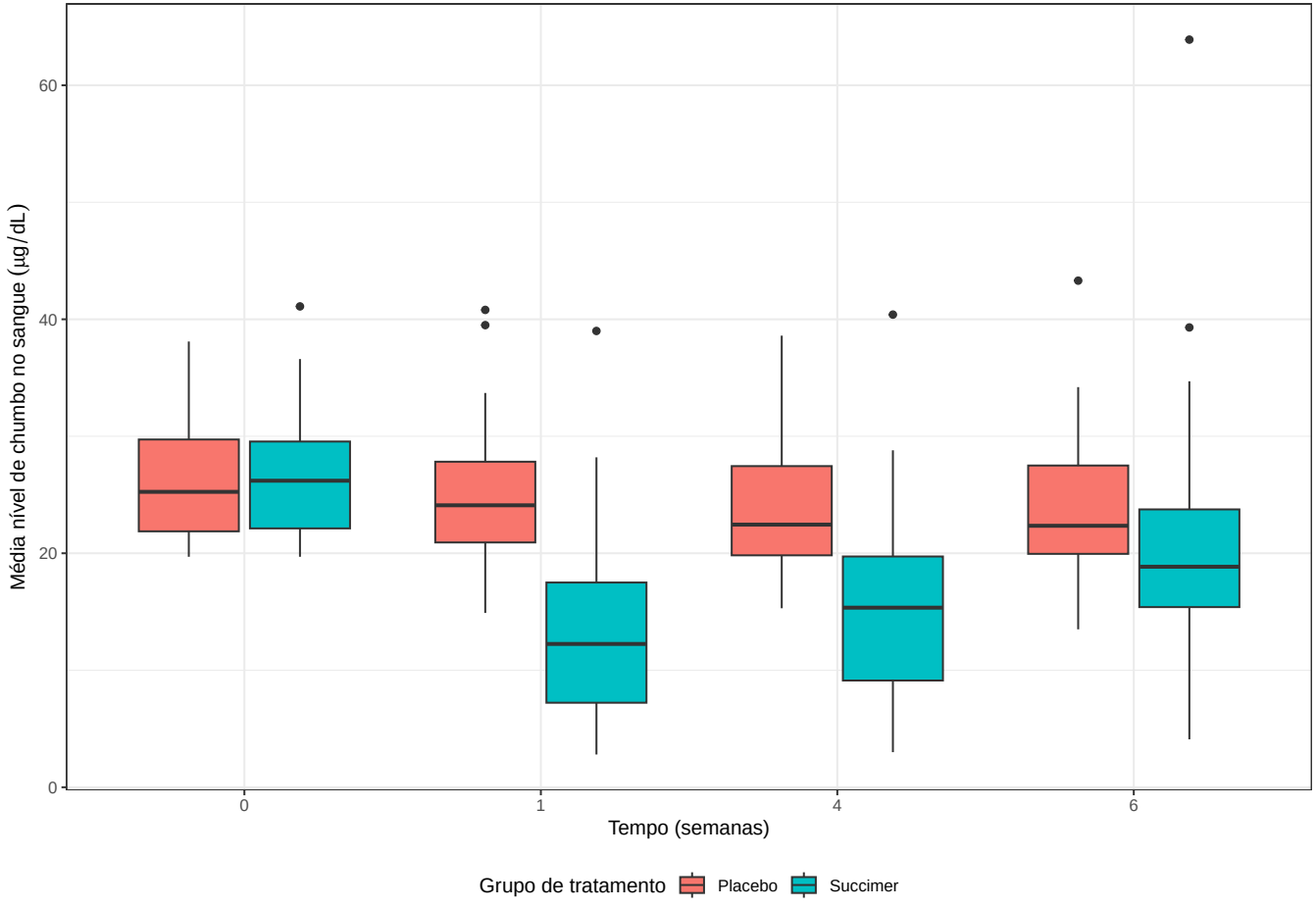
Perfis de médias com barras de erros

- Uma alternativa: o *boxplot*.

```
chumbo.longo$tempof <- factor(chumbo.longo$tempo)

p <- ggplot(data = chumbo.longo,
            mapping = aes(x = tempof, y = chumbo,
                          fill = trt)) +
  geom_boxplot() +
  labs(x = "Tempo (semanas)",
       y = expression("Média nível de chumbo no sangue"~(mu*g/dL)),
       fill = "Grupo de tratamento") +
  theme_bw() + theme(legend.position = "bottom")
p
```

Perfis de médias com barras de erros



Dados desbalanceados

Um exemplo

- Dados longitudinais sobre o crescimento da função pulmonar em crianças e adolescentes do *Six Cities Study of Air Pollution and Health*.
- Os dados são de uma coorte de 300 meninas em idade escolar que moram em Topeka, Kansas, que, na maioria dos casos, estavam matriculadas na primeira ou segunda série (com idades entre seis e sete anos).
- As meninas foram medidas anualmente até a formatura do ensino médio (aproximadamente aos dezoito anos) ou perda de acompanhamento, e cada menina forneceu no mínimo uma e no máximo 12 observações.
- A cada exame, as medidas da função pulmonar foram obtidas a partir da espirometria simples.
- Uma medida amplamente utilizada, calculada a partir da espirometria simples, é o volume de ar expirado no primeiro segundo da manobra, FEV_1.

Carregando os dados

```
fev <- read_dta(file = "fev1.dta")

# cria variável log(fev1/ht)
fev$fev1 <- exp(fev$logfev1)
fev$logfh <- log(fev$fev1/fev$ht)

# uma observação atípica (?)
fev <- fev[- which(fev$logfev1/fev$ht < -0.5), ]
```

Carregando os dados

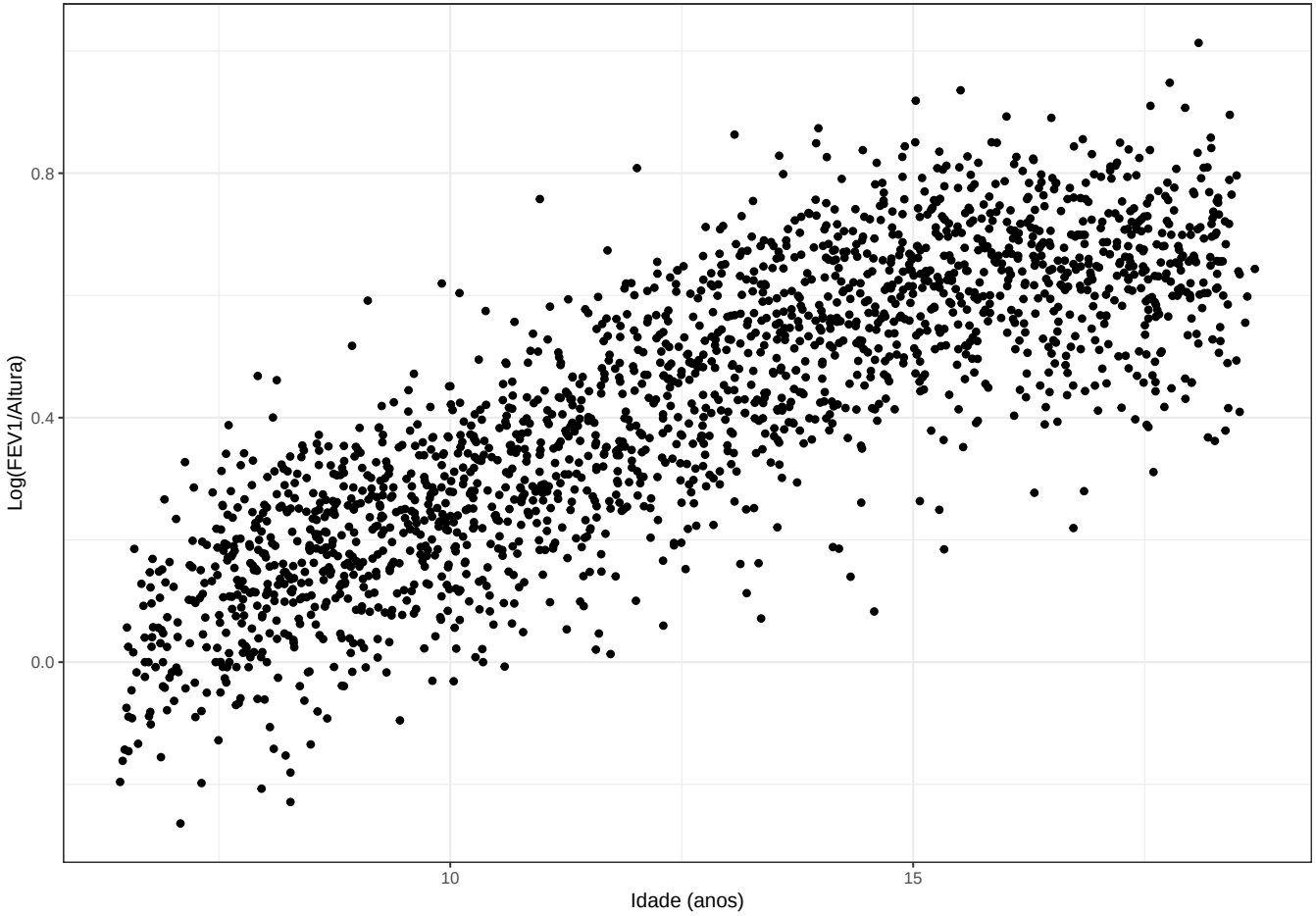
```
head(fev, 16)
```

```
## # A tibble: 16 x 8
##       id    ht   age baseht baseage logfev1  fev1  logfh
##    <dbl> <dbl> <dbl>  <dbl>   <dbl>   <dbl> <dbl>  <dbl>
##  1     1  1.20  9.34   1.20    9.34   0.215  1.24 0.0328
##  2     1  1.28 10.4    1.20    9.34   0.372  1.45 0.125
##  3     1  1.33 11.5    1.20    9.34   0.489  1.63 0.203
##  4     1  1.42 12.5    1.20    9.34   0.751  2.12 0.401
##  5     1  1.48 13.4    1.20    9.34   0.833  2.30 0.441
##  6     1  1.5  15.5    1.20    9.34   0.892  2.44 0.487
##  7     1  1.52 16.4    1.20    9.34   0.871  2.39 0.453
##  8     2  1.13  6.59   1.13    6.59   0.307  1.36 0.185
##  9     2  1.19  7.65   1.13    6.59   0.351  1.42 0.177
## 10     2  1.49 12.7    1.13    6.59   0.756  2.13 0.357
## 11     2  1.53 13.8    1.13    6.59   0.867  2.38 0.442
## 12     2  1.55 14.7    1.13    6.59   1.05   2.85 0.609
## 13     2  1.56 15.8    1.13    6.59   1.15   3.17 0.709
## 14     2  1.57 16.7    1.13    6.59   0.924  2.52 0.473
## 15     2  1.57 17.6    1.13    6.59   1.13   3.11 0.684
## 16     3  1.18  6.91   1.18    6.91   0.432  1.54 0.266
```

Diagrama de dispersão

```
p <- ggplot(data = fev,
            mapping = aes(x = age, y = logfh)) +
  geom_point() +
  labs(x = "Idade (anos)",
       y = "Log(FEV1/Altura)") +
  theme_bw()
p
```

Diagrama de dispersão

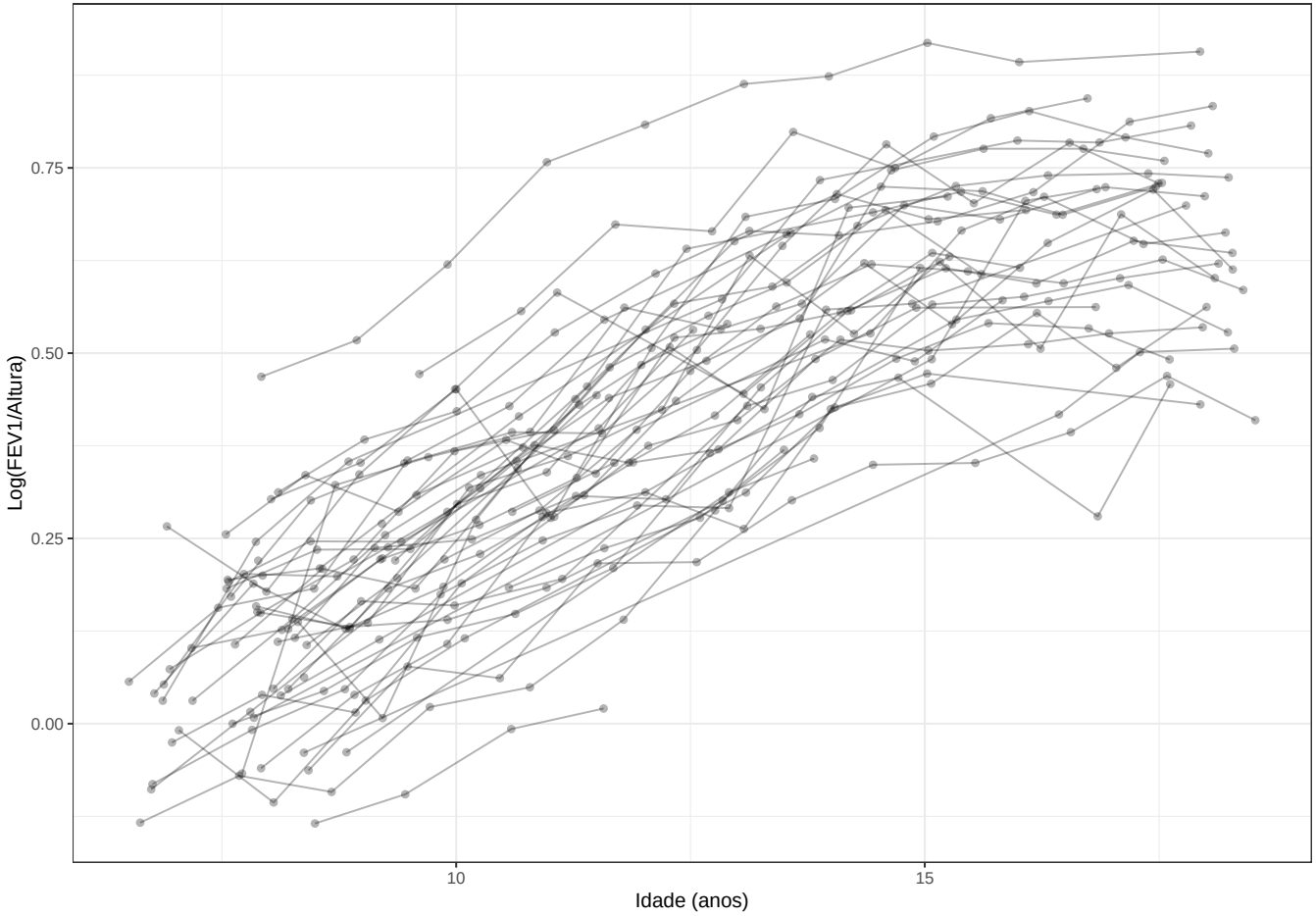


Perfis individuais

```
set.seed(5)
id.s <- sample(x = unique(fev$id), size = 50,
               replace = FALSE)

p <- ggplot(data = fev[which(fev$id %in% id.s),],
            mapping = aes(x = age, y = logfh,
                          group = id)) +
  geom_point(alpha = 0.3) +
  geom_line(alpha = 0.3) +
  labs(x = "Idade (anos)",
       y = "Log(FEV1/Altura)") +
  theme_bw()
p
```

Perfis individuais



Perfis de médias

- Como cada indivíduo não é medido na mesma idade, a construção de gráficos da resposta média em função da idade pode representar dificuldades devido à escassez de dados em qualquer idade específica.
- Nos casos em que as ocasiões de medição são diferentes, é útil produzir um gráfico **"suavizado"** da tendência média de resposta ao longo do tempo.
- Tal gráfico pode ser obtido usando uma variedade de abordagens chamadas de **"técnicas de suavização"**.
- Muitas dessas técnicas de suavização abordam a estimativa da resposta média a qualquer momento, considerando não apenas as observações naquela ocasião, mas também as observações "vizinhas".
  - Ou seja, a média estimada é baseada em observações feitas antes, no e depois do momento de interesse.
  - A resposta média em qualquer momento, digamos t, é considerada uma média ponderada das observações em alguma vizinhança em torno de t.

Suavização

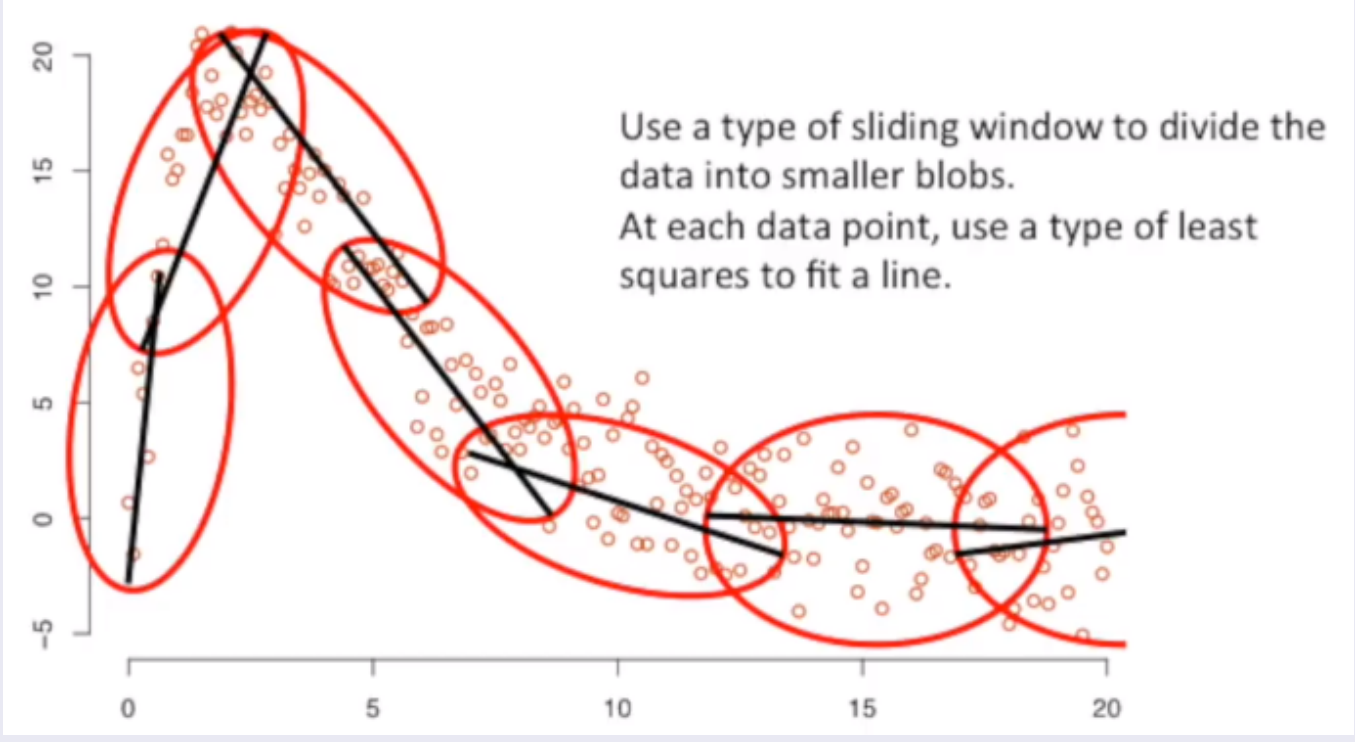
- Para dados longitudinais que são balanceados e completos (sem dados ausentes), a média móvel no tempo t, denotado S_t, é dada por

$$S_t = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=-k}^{k} w_j y_{i,t+j}, \ \ t = k+1, \ldots, n-k, \sum_{j=-k}^{k} w_j = 1.$$

- Quando os dados longitudinais são espaçados irregularmente e desbalanceados ao longo do tempo, outros métodos de regressão não paramétricos podem ser usados para estimar a tendência de resposta média ao longo do tempo.
- Um método popular disponível na maioria dos pacotes de software estatístico padrão é a regressão localmente ponderada ou ***lowess***.

Suavização

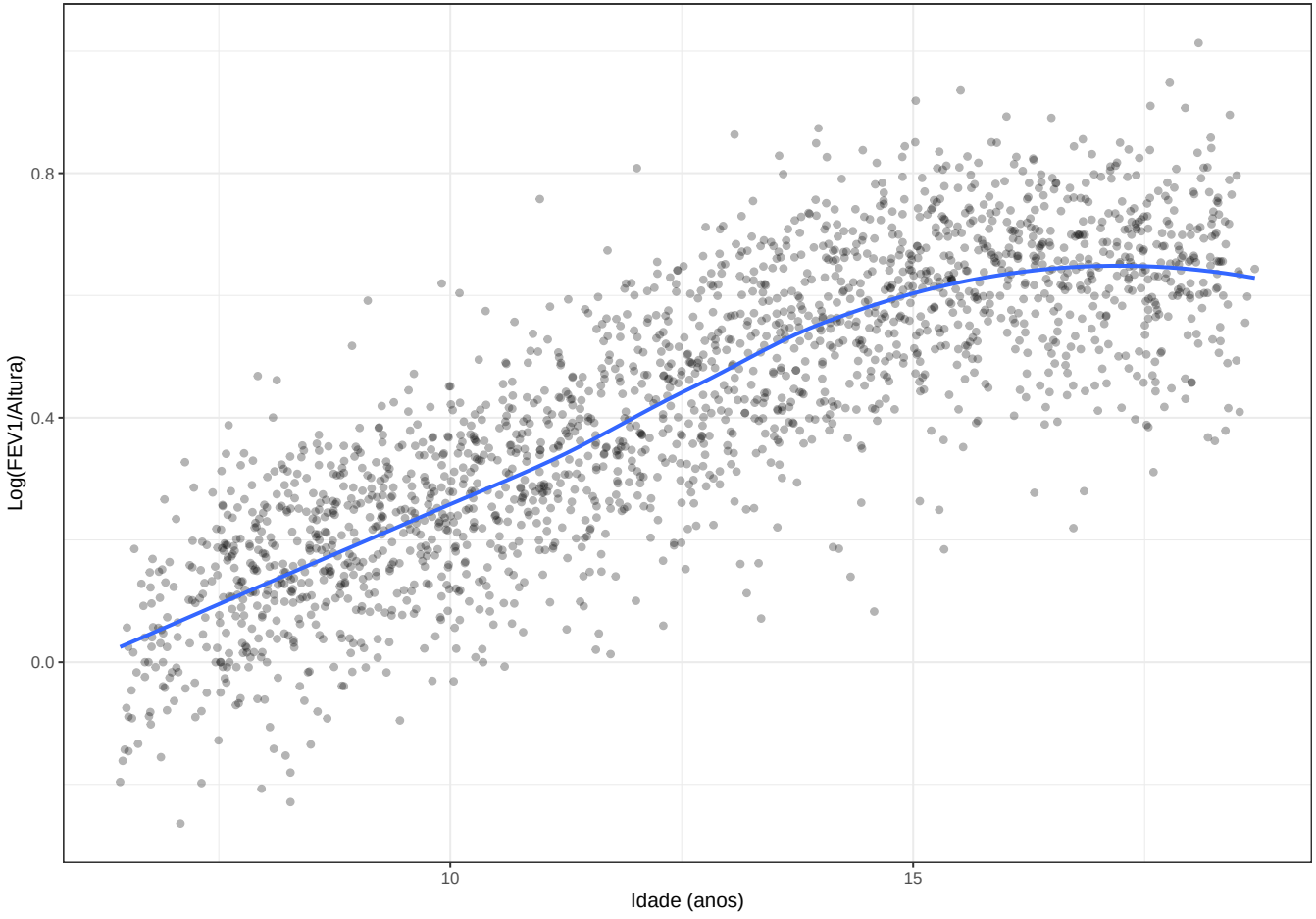
Lowess ou loess explicado



Suavização: lowess

```
p <- ggplot(data = fev,
            mapping = aes(x = age, y = logfh)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "loess", se = FALSE) +
  labs(x = "Idade (anos)",
       y = "Log(FEV1/Altura)") +
  theme_bw()
p
```

Suavização: lowess



Estrutura de correlação (succimer)

Estrutura de correlação (succimer)

```
chumbo.succimer <- chumbo %>%
  filter(trt == 1) %>%
  select(y0, y1, y4, y6) %>%
  mutate(y0 = as.numeric(y0),
         y1 = as.numeric(y1),
         y4 = as.numeric(y4),
         y6 = as.numeric(y6))
```

Estrutura de correlação (succimer)

```
chumbo.succimer
```

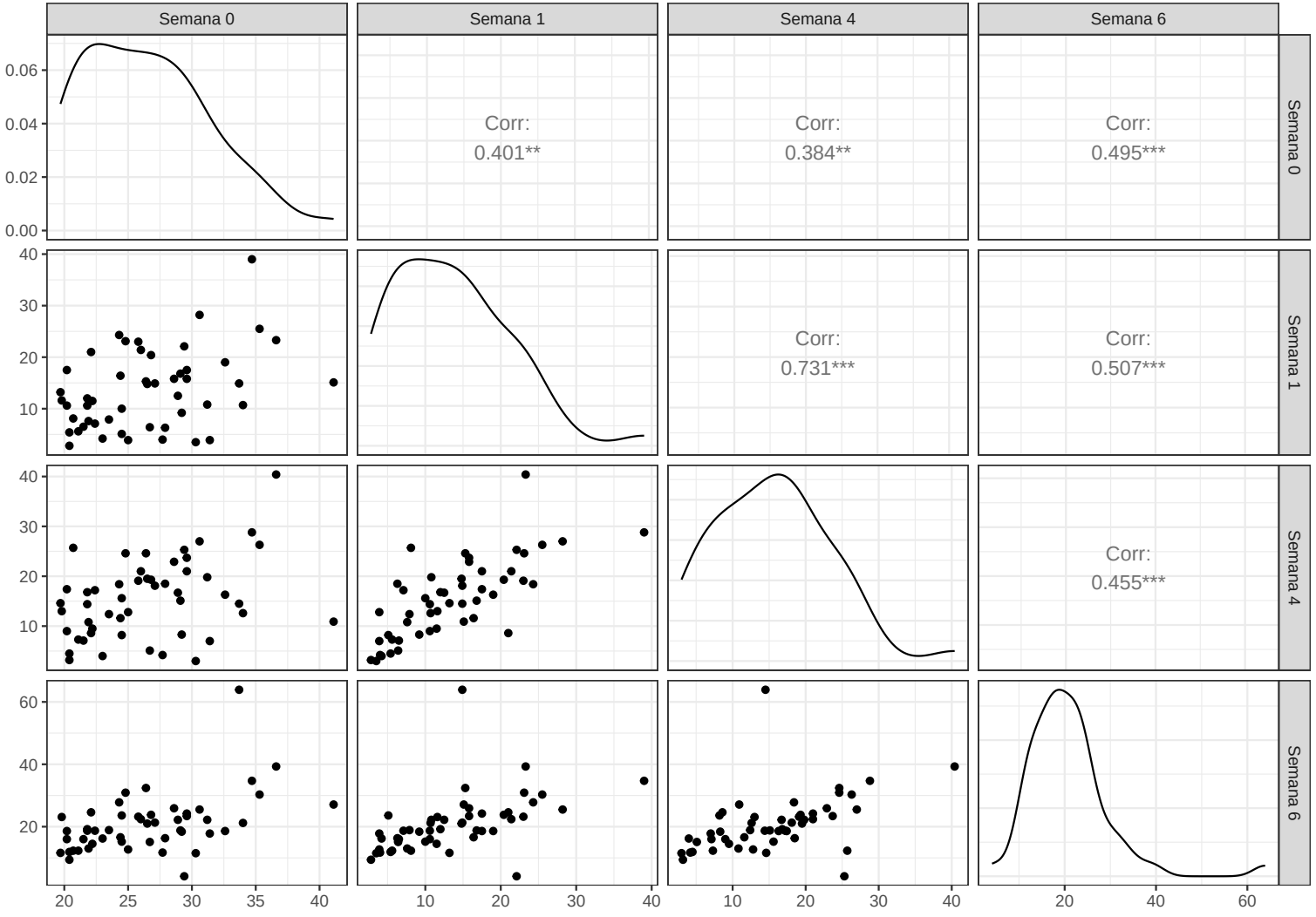
```
## # A tibble: 50 x 4
##       y0    y1    y4    y6
##    <dbl> <dbl> <dbl> <dbl>
##  1  26.5 14.8  19.5  21
##  2  25.8 23    19.1  23.2
##  3  20.4  2.80  3.20  9.40
##  4  20.4  5.40  4.5  11.9
##  5  24.8 23.1  24.6  30.9
##  6  27.9  6.30 18.5  16.3
##  7  35.3 25.5  26.3  30.3
##  8  28.6 15.8  22.9  25.9
##  9  29.6 15.8  23.7  23.4
## 10  21.5  6.5   7.10 16
## # i 40 more rows
```

Estrutura de correlação (succimer)

```
library(GGally)

p <- ggpairs(chumbo.succimer,
             columnLabels = paste("Semana", c(0, 1, 4, 6))) +
  theme_bw()
p
```

Estrutura de correlação (succimer)



Avisos

- **Para casa:** ler o Capítulo 3 do livro "**Applied Longitudinal Analysis**". Caso ainda não tenha lido, leia também os Caps. 1 e 2.
- Uma introdução ao pacote `ggplot2` pode ser vista em https://datathon-ufrgs.github.io/Pintando_e_Bordando_no_R/
- **Próxima aula:** Considerações a respeito da modelagem da média e da covariância, e abordagem histórica dos métodos de análise de medidas repetidas.

Bons estudos!