Modelos de Dados Longitudinais

Introdução ; Representação Gráfica

Inês Sousa

Departamento de Matemática

Universidade do Minho

1.º semestre

Bibliografia

* Diggle P.J., Heagerty P., Liang K-Y. and Zeger S.L. (2002), *Analysis of Longitudinal Data*, Oxford

* Verbeke G. and Molenberghs G. (2000), *Linear Mixed Models for Longitudinal Data*, Springer

* Bates J. and Pinheiro D. (2002), *Mixed Effects Models in S and S-Plus*, Springer

Objectivos Finais do Curso

- Identificar um conjunto de dados longitudinais.
- Decidir aplicar métodods estatísticos para dados longitudinais.
- Manipular bases de dados longitudinais em `R`.
- Apresentar dados longitudinais visualmente através de gráficos.
- Fazer uma análise exploratória que apoie o modelo longitudinal usado, tanto para a estrutura de correlação, bem como para o valor médio da população.

Objectivos Finais do Curso (cont.)

- Caracterizar um modelo linear para dados longitudinais.
- Estimar parâmetros do modelos longitudinal por máxima verosimilhança.
- Fazer inferência sobre os parâmetros do modelo longitudinal e identificar qual o melhor método a utilizar.
- Utilizar métodos de diagnóstico para o modelo longitudinal aplicado.
- Interpretar os resultados obtidos da análise estatística no contexto do problema.

Introdução

O que são **Dados Longitudinais**?

- Dados Longitudinais são gerados por medidas repetidas ao longo do tempo em diferentes indivíduos.
- Assuminos sempre independência entre os indivíduos.
- Análise de Dados Longitudinais combina técnicas de análise multivariada e análise de séries temporais.
- Série Temporal é uma única série longa, Longitudinais são várias séries mais curtas.
- Medidas de diferentes indivíduos são independentes, impossibilidade de usar clássica análise multivariada

Estudos Longitudinais / Medidas Repetidas

- **Desenhos Clássicos** (Cross-sectional): *única* medida em *uma* variável em cada indivíduo
- **Desenhos Multivariados**: *única* medida em *mais do que uma* variáveis em cada indivíduo
- **Desenhos Medidas Repetidas**: *múltipas* medidas em *uma* variável em cada indivíduo
- **Desenhos Longitudinais**: Desenhos de medidas repetidas ao longo do *tempo*

  Um conjunto de medidas repetidas só é longitudinal se for ao longo do TEMPO
- **Multivariate Longitudinal Designs**

Exemplo: Crescimento Humano

- **Cross-sectional** (OU TRANSVERSAL): medir altura de vários indivíduos com diferentes idades. (Mais ráido)

- **Multivariate**: medir altura e peso de vários indivíduos com diferentes idade.

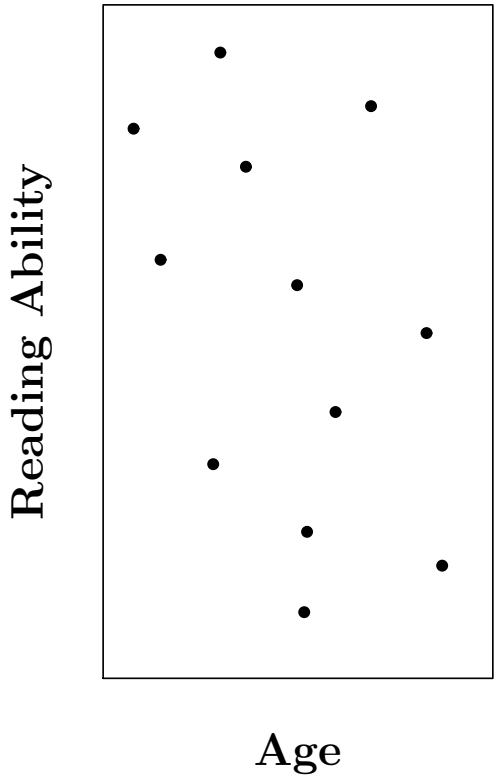
- **Longitudinal**: seleccionar vários indivíduos com a mesma idade, e medir altura (tb peso) numa sequência de tempos.

Exemplo: Medidas Repetidas não necessariamente Longitudinais

- **Oftalmologia**: medidas em ambos os olhos em cada indivíduo.
- **Experimentos em animais**: tratamentos aplicados a conjuntos de animais, e não animais únicos.
- **Experimentos em Educação**: aplicar políticas a diferentes escolas/turmas, e obter respostas para vários alunos por turma/escola
- **Gémeos**: Tratamentos em gémeos

Estudos Longitudinais

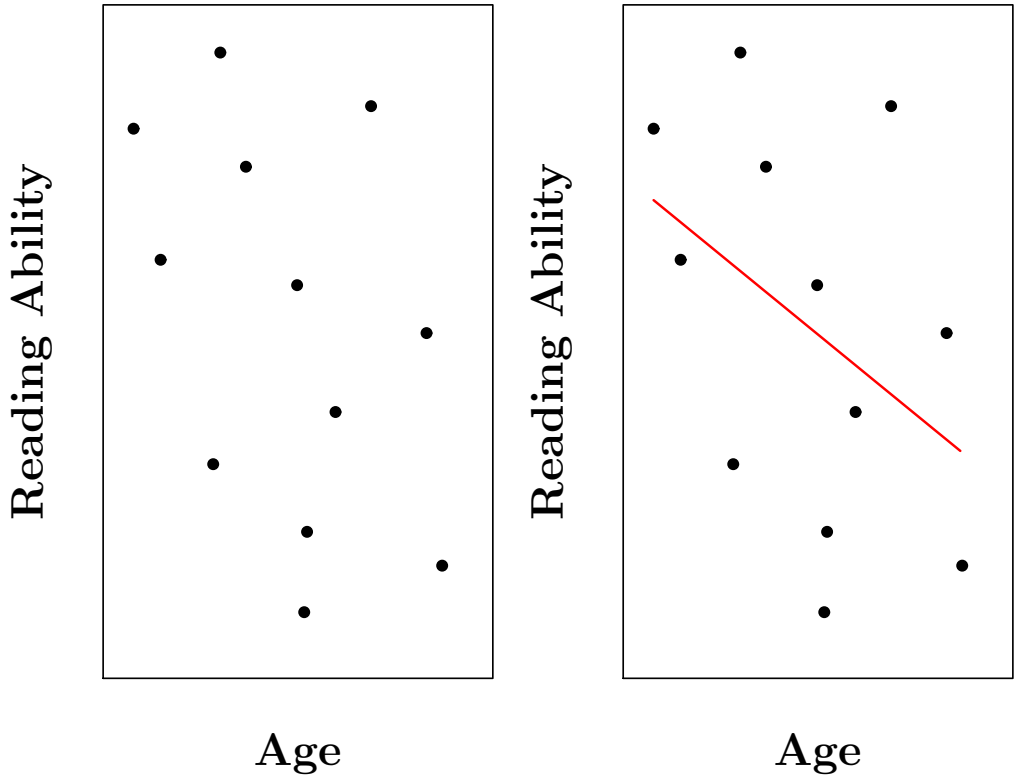
Permitem distinguir: efeitos *do cohort* e efeitos *da idade/tempo*



in *Diggle et al (2002)*

Estudos Longitudinais

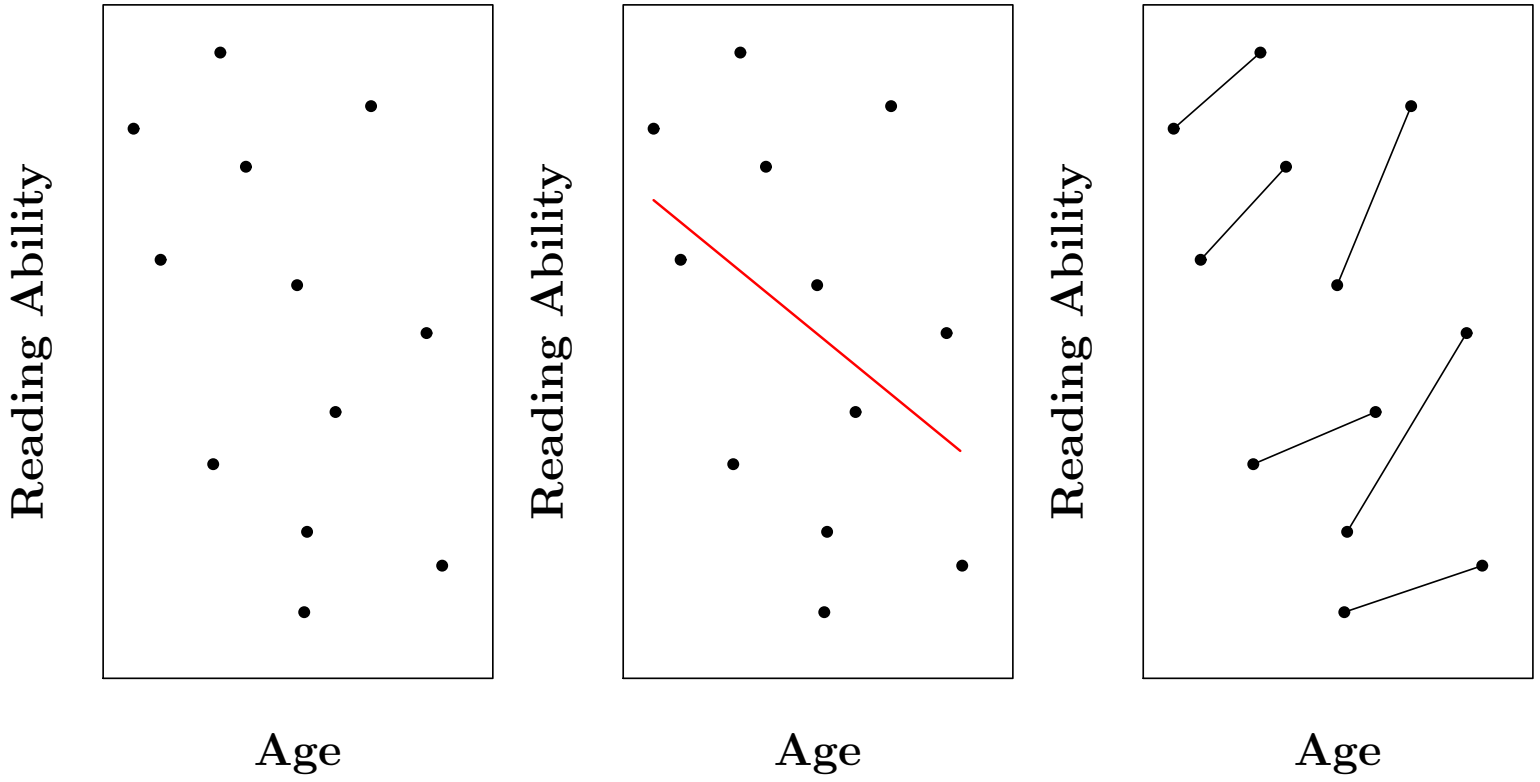
Permitem distinguir: efeitos *do cohort* e efeitos *da idade/tempo*



in *Diggle et al (2002)*

Estudos Longitudinais

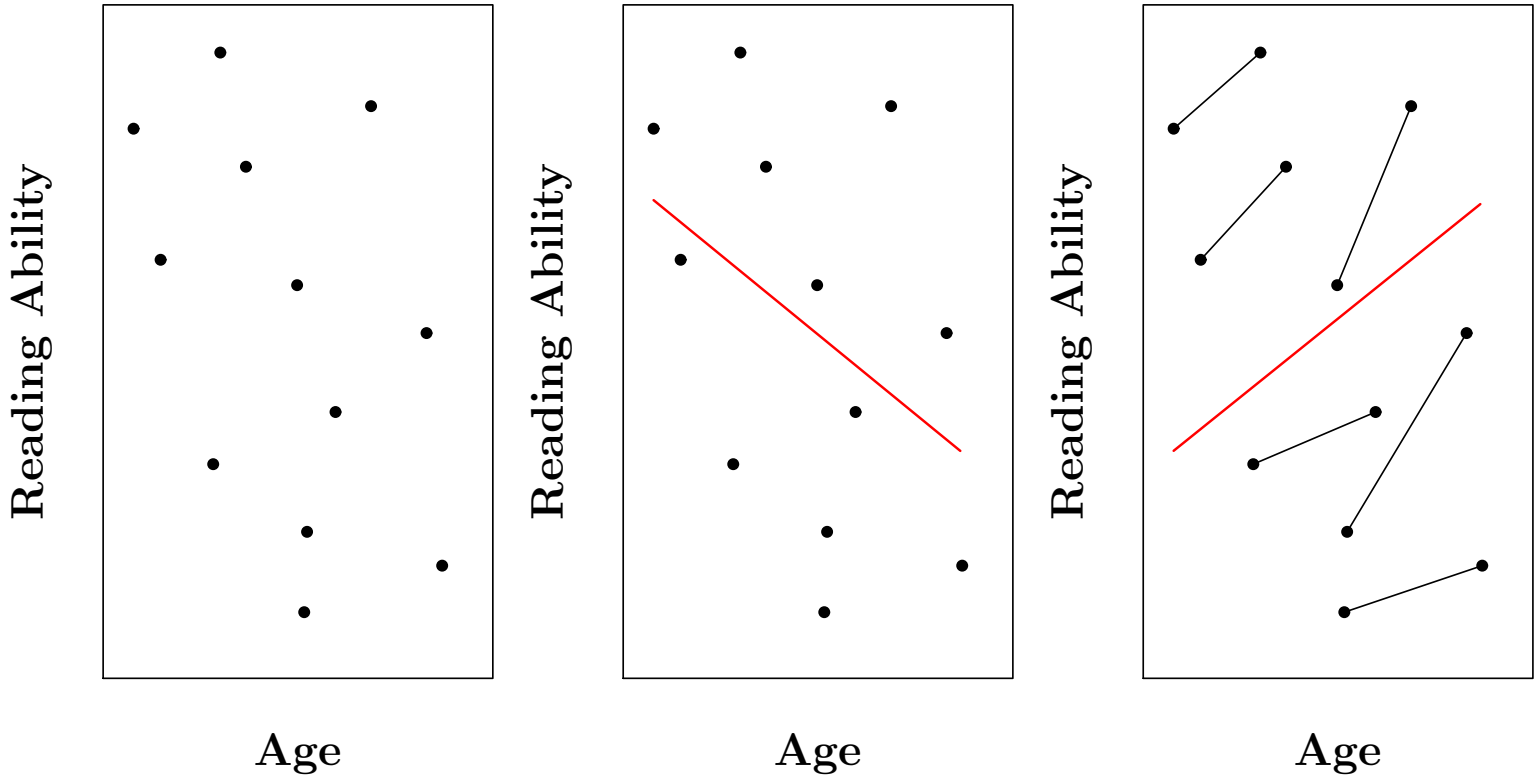
Permitem distinguir: efeitos *do cohort* e efeitos *da idade/tempo*



in *Diggle et al (2002)*

Estudos Longitudinais

Permitem distinguir: efeitos *do cohort* e efeitos *da idade/tempo*

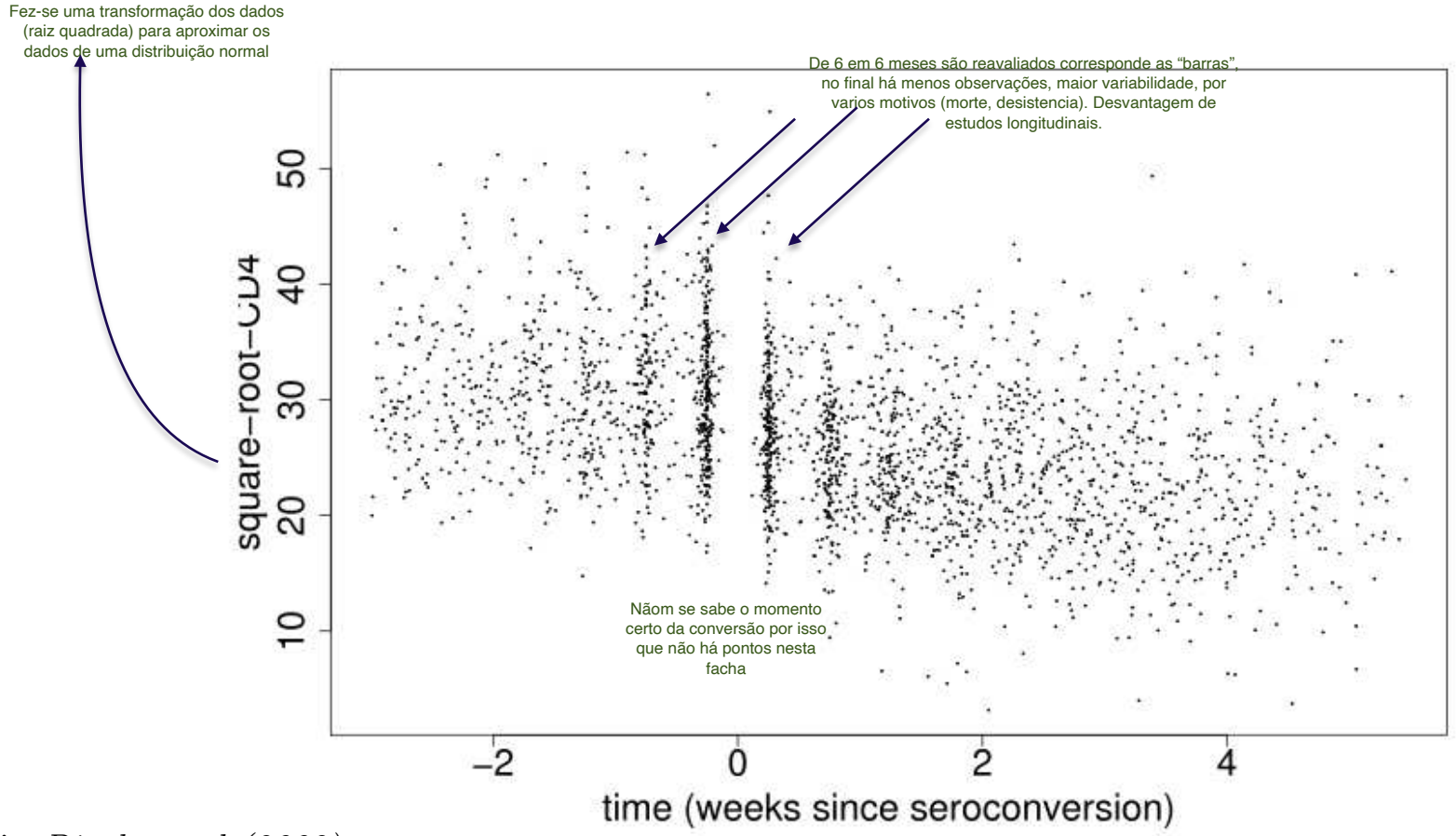


in *Diggle et al (2002)*

Exemplo 1: Número de células CD4+

Cohort de 369 HIV seropositivos, medidas do número de células CD4+ a aproximadamente intervalos de 6-meses.

Número de medidas repetidas, e tempos de medidas, são diferentes para cada indivíduo.



in *Diggle et al (2002)*

0 - momento de conversão, tempos negativo a pessoa é sero positiva, tempos positivos. a pessoa passa a ser doente de SIDA

Exemplo 1: Número de células CD4+

Objectivos da análise longitudinal destes dados:

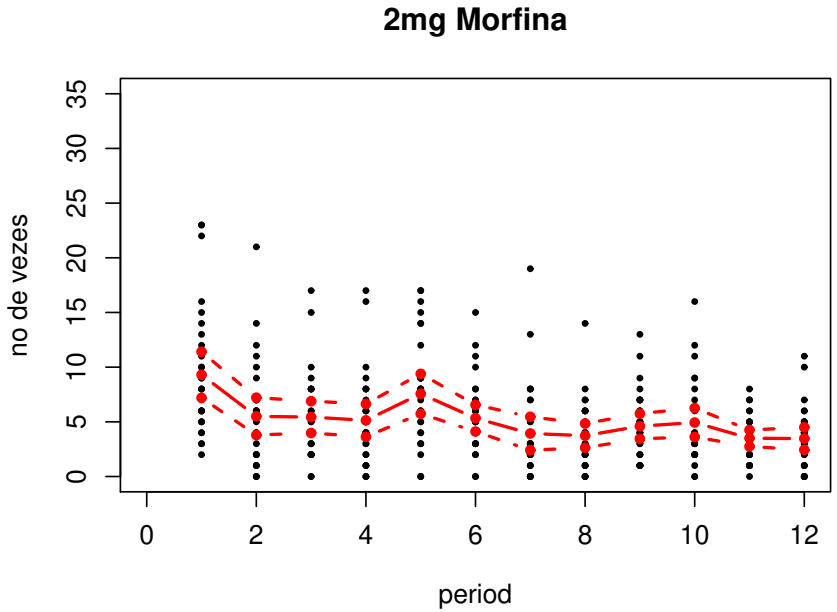
- estimar o tempo médio até ao decréscimo de células CD4+
- estimar a progressão para sujeitos individuais, considerando o erro de medida na determinação de céluas CD4+
- caracterizar o grau de heterogeneidade entre indivíduos, no grau de progressão
- identificar factores de predição para mudanças nos valores das células CD4+

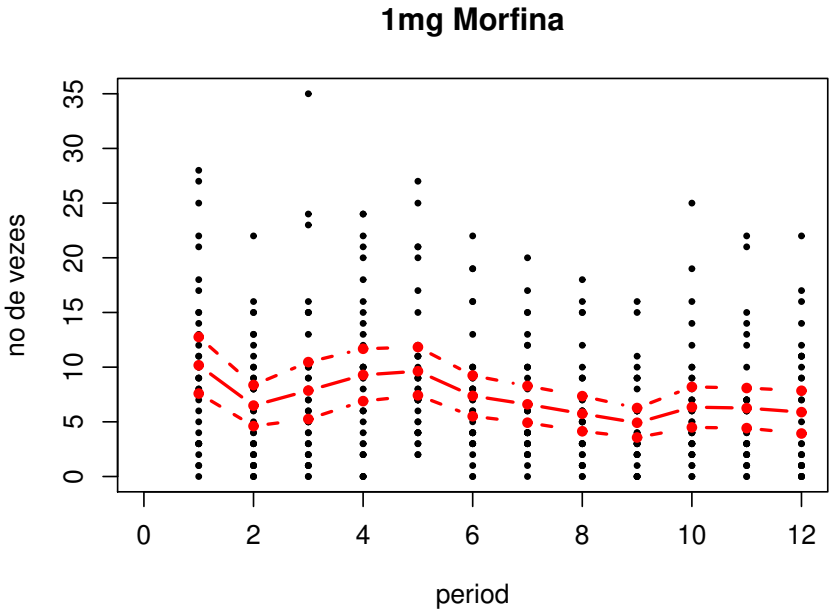
Exemplo 2: Ensaio Clínico Anestesia (PCA)

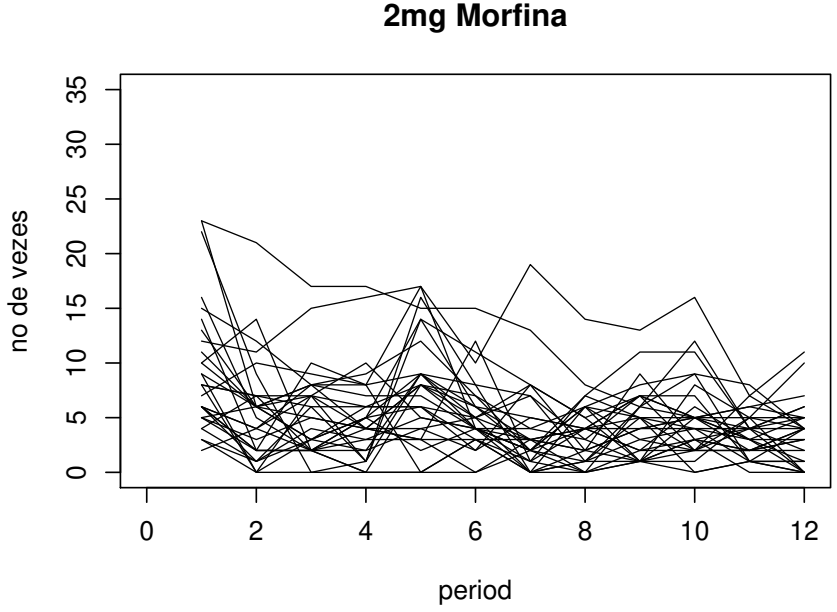
Estudo caso-controlo

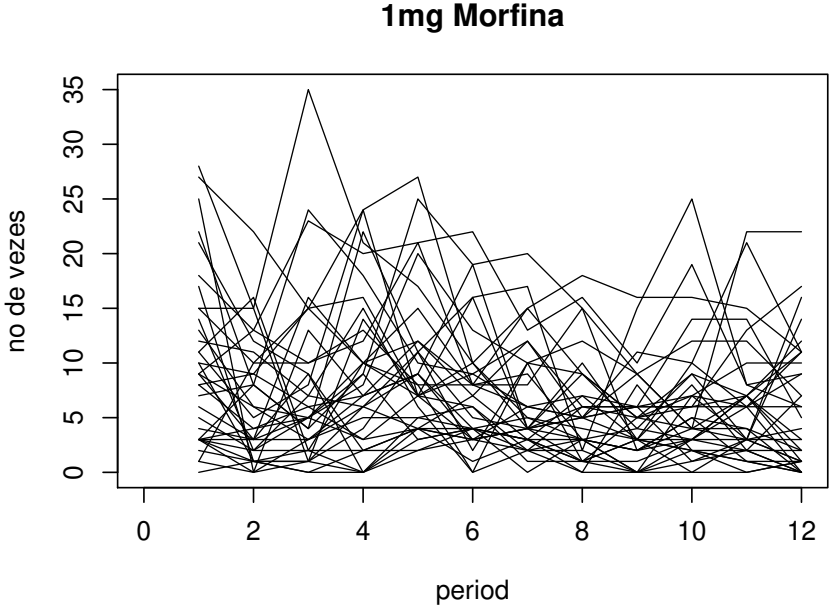
- Dados de um ensaio clínico para o uso de uma anestesia controlada pelo doente.
- A droga pode ser tomada apenas em intervalos de tempo controladas.
- Dados: corresponde ao número de vezes que o paciente auto-administra a droga num intervalo de 4 horas, durante 2 dias.
- Dados para 65 pacientes.
- Há dois grupos de tratamento:
  - 2mg morfina, com 8 minutos de espera
  - 1mg morfina, com 4 minutos de espera

Exemplo 2: Ensaio Clínico Anestesia (PCA)









Maior variabilidade pq podem carregar mais vezes

Exemplo 3: Tratamento Esquizofrenia (PANSS)

- Ensaio clínico aleatorizado, para terapias à base de drogas
- Três tratamentos:
  - haloperidol (standard)
  - placebo
  - risperidone (novo)

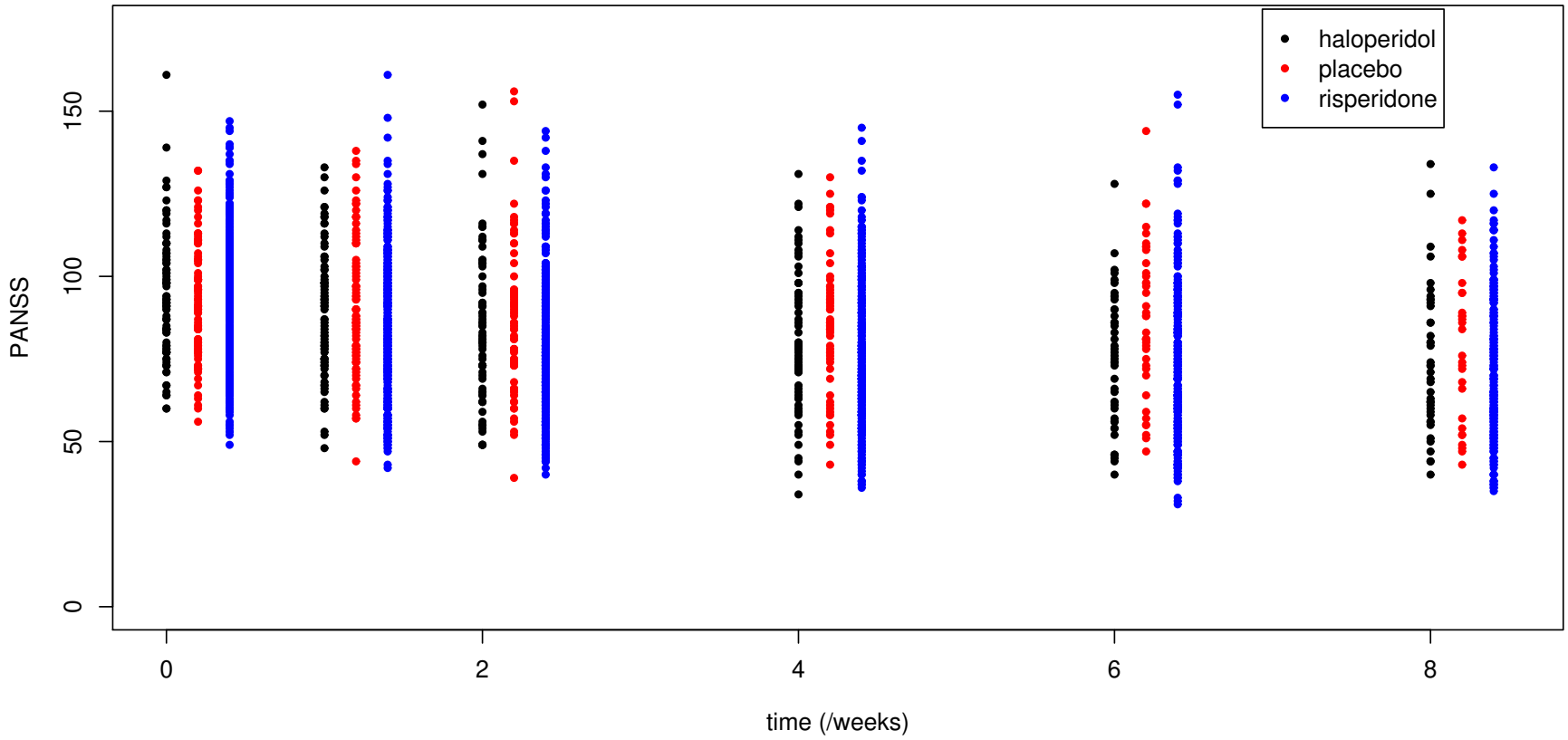
O medicamento a testar tem que ter um n maior

- Dropout por causa de "resposta inadequada ao tratamento"

Treatment	Número de não dropouts na semana					
	0	1	2	4	6	8
haloperidol	85	83	74	64	46	41
placebo	88	86	70	56	40	29
risperidone	345	340	307	276	229	199
total	518	509	451	396	315	269

O nº do final do tratamento também é imporatnte

Exemplo 3: Tratamento Esquizofrenia (PANSS)



Neste gráfico não conseguimos ver a progressão,
isto tb tem a ver com o tamanho do gráfico

O que estes exemplos têm em comum

- Há observações repetidas em cada unidade experimental.
- Unidades podem ser assumidas independentes - réplicas séries temporais.
- Múltiplas respostas em cada unidade possivelmente correlacionads.

  As observações são independentes umas das outras mas correlacionados entre si (nos diferentes momentos de medição/observação)
- Os objectivos podem ser formulados como problemas de regressão.
- A escolha do modelo estatístico depende do tipo de variável resposta.

Vantagens dos Métodos Longitudinais

- Economia em número de indivíduos; cada sujeito serve como o seu próprio controlo;

São necessárias menos pessoas, pq temos mais observações em diferentes momentos. Cada pessoa funciona como controlo de si próprio ao longo tempo.

- Variação entre indivíduos excluída do erro;
- Fornece estimadores mais eficientes do que desenhos cross-sectional/transversais, com o mesmo número e padrão dos dados observados;
- Permite separar *efeito de idade* (mudanças ao longo do tempo em cada indivíduo) de efeitos de cohort/grupo (diferenças entre indivíduos em baseline) $\Rightarrow$ estudos transversais não permite fazer;

Conseguimos saber a média populacional, mas também a media individual e quanto cada individuo se afasta da media de grupo

- Permite-nos obter informação sobre mudanças ao nível individual.

Vantagens dos Métodos Longitudinais

- Mais flexível em desenhos de investigação;
  - sem necessidade de todos terem os dados recolhidos nos mesmos momentos - $\neq$ tempos de medida, tempo pode ser contínuo.
  - sem necessidade de todos serem medidos o mesmo número de vezes - $\neq$ número de medidas.
- Identificar padrões temporais nos dados;
  - A variável resposta aumenta/diminui/estabiliza ao longo do tempo?
  - É o padrão geral linear ou não linear?
  - Há evidência de pontos de mudança de efeitos?
- Incluir preditores que variam com o tempo;
- Incluir efeitos de interação com o tempo
  - Testar se os efeitos dos preditores variam ao longo do tempo.

Objectivos Científicos

Filosofia Pragmática: Métodos de análise devem ter em consideração os objectivos científicos do estudo.

All models are wrong, but some models are useful, G.E.P. Box

O que queremos?

- compreender cientificamente vs. descrever empiricamente?
- focar ao nível individual vs. ao nível da população?
- comportamento em média vs. variação à volta da média?

Objectivos Científicos (cont.)

Example: Política de redução de fumadores

Perspectiva de Saúde Pública - "De que forma as políticas irão afectar a saúde da comunidade" (população em média)

Perspectiva Clínica - "De que forma as políticas irão afectar a saúde do **meu** paciente?" (ao nível do indivíduo)

Correlação e porque importa ...

- Diferentes medidas, no tempo, no mesmo indivíduo estão tipicamente correlacionados

- Esta correlação deve ser reconhecida no processo de inferência

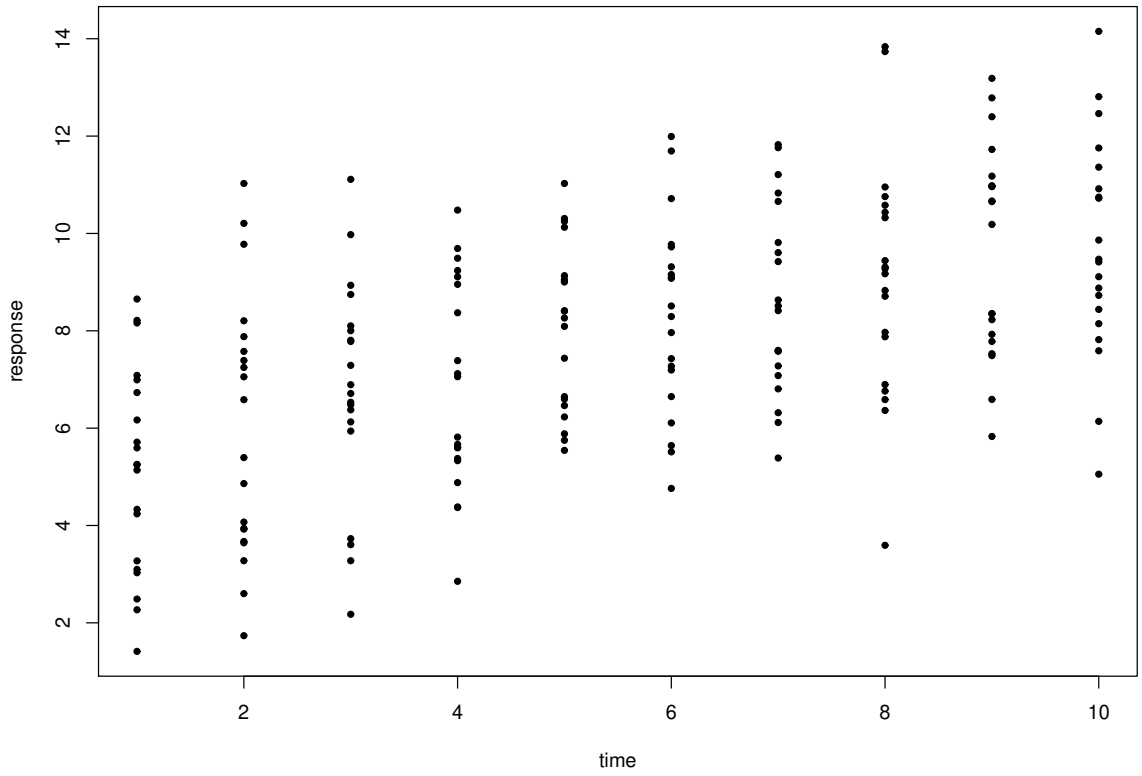
Notação

Letras Gregas representam parâmetros e
Letras Romanas são quantidades observadas

- $f(y;\theta)$ genérico para função densidade ou distribuição de probabilidade, com θ o vector de parâmetros
- $i = 1, ..., m$ número de indivíduos
- $j = 1, ..., n_i$ número de observações no indivíduo i
- $t_{i1}, ..., t_{in_i}$ tempos das observações no indivíduo i
- $y_{i1}, ..., y_{in_i}$ as observações/realizações
- $\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{in_i}$ vectores de covariáveis
- $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{in_i})$ n_i-vector de respostas variáveis aleatórias
- μ_i, V_i valor esperado e matrix variância de $\boldsymbol{Y}_i$
- $N = \sum_i n_i$ total número de observações

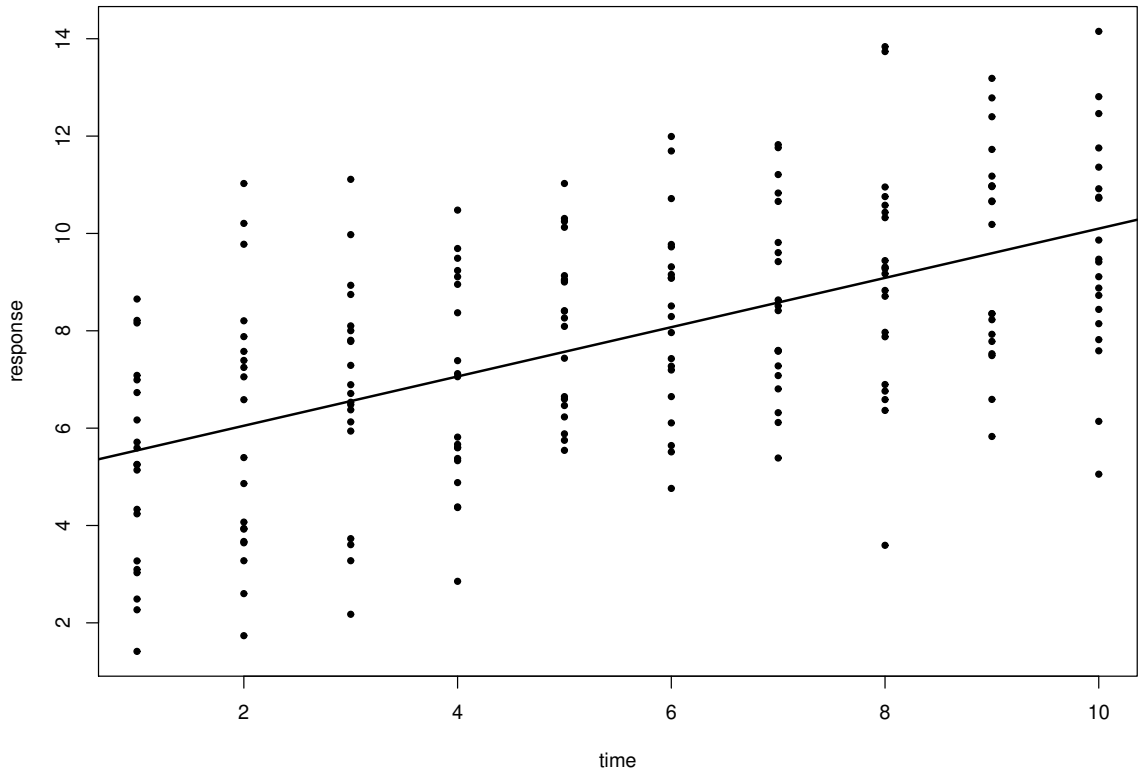
Correlação pode ou não importar ...

$$Y_{it} = \alpha + \beta * t + \epsilon_{it} \qquad i = 1, ..., m = 20 \quad t = 1, ..., n = 10$$



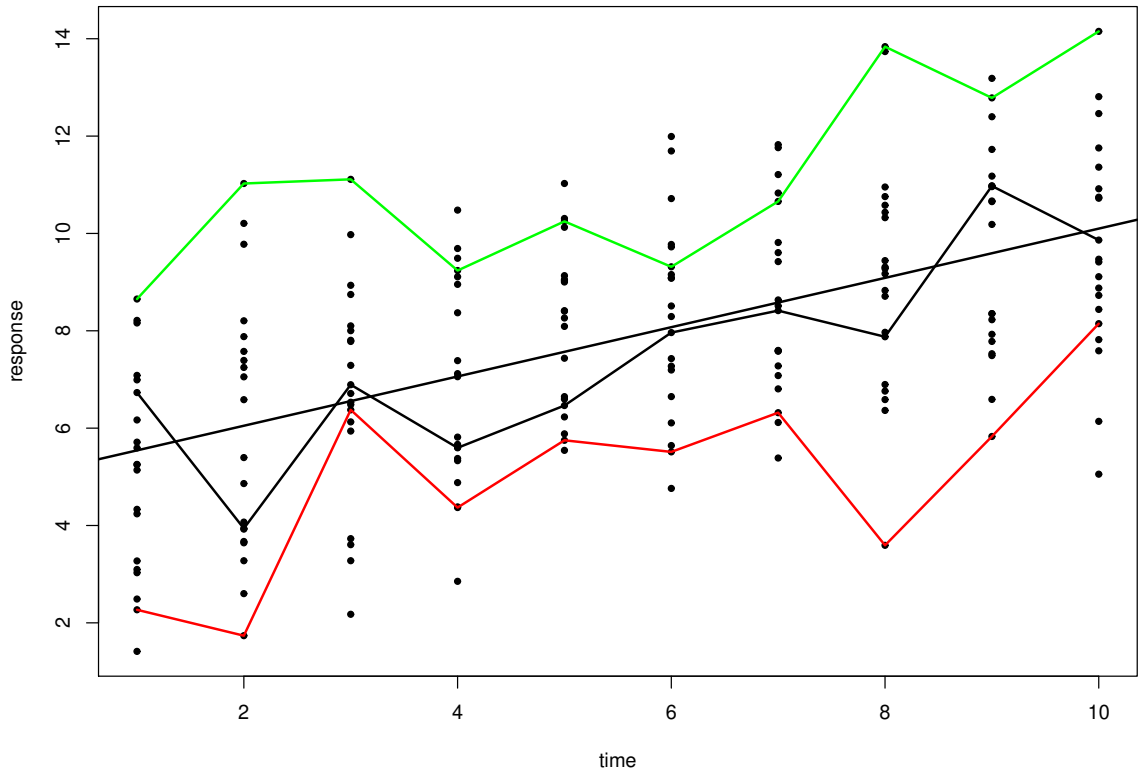
Correlação pode ou não importar ...

$$Y_{it} = \alpha + \beta * t + \epsilon_{it} \qquad i = 1, ..., m = 20 \quad t = 1, ..., n = 10$$



Correlação pode ou não importar ...

$$Y_{it} = \alpha + \beta * t + \epsilon_{it} \qquad i = 1, ..., m = 20 \quad t = 1, ..., n = 10$$



Correlação pode ou não importar ...

$$Y_{it} = \alpha + \beta * t + \epsilon_{it} \qquad i = 1, ..., m = 20 \quad t = 1, ..., n = 10$$

Estimadores pontuais para os parâmetros, e erros padrões

	ignorando correlação		reconhecendo correlação	
	estimador	s.e.	estimador	s.e.
$\alpha = 5$	5.037	0.330	5.037	0.424
$\beta = 0.5$	0.506	0.053	0.506	0.035

Estudos Transversais vs Longitudinais

$$\text{Modelo de Regressão} \quad Y_i = X_i\beta + \epsilon_i$$

- Cross-sectional $(n_i = 1)$

$$Y_{i1} = \beta_c x_{i1} + \epsilon_{i1}, \quad i = 1, ..., m$$

  - β_c representa a diferença em média de Y entre duas sub-populações que diferem por uma unidade em x.

- Observações repetidas extensão do modelo

$$Y_{ij} = \beta_c x_{i1} + \beta_L(x_{ij} - x_{i1}) + \epsilon_{ij}, \quad j = 1, ..., n_i, i = 1, ..., m$$

$$(Y_{ij} - Y_{i1}) = \beta_L(x_{ij} - x_{i1}) + \epsilon_{ij} - \epsilon_{i1}$$

  - β_L reprensenta a mudança esperada em Y ao longo do tempo por unidade de mudança em x para um indivíduo específico.

Estudos Transversais vs Longitudinais

- Em CS a base é a comparação de indivíduos com um valor particular de x em relação a outros indivíduos com diferentes valores.
- Em LDA cada indivíduo é o seu próprio controle. β_L é estimado pela comparação de medidas em dois tempos de um mesmo indivíduo assumindo que x muda ao longo do tempo.
- Em LDA podemos distinguir o grau de variação de Y *ao longo do tempo* para um indivíduo, da variação de Y *entre indivíduos*.

Estudos Transversais vs Longitudinais - exemplo

- Suponhamos que queremos estimar o estado imunológico de um homem pelo seu nível de células $CD4+$.
- Em CS, informação é "emprestada" de outros indivíduos para ultrapassar erro de medida. Mas, ao fazer a média de todas as pessoas, ignoramos a diferença natural em $CD4+$ entre indivíduos.
- Em LDA, informação é "emprestada" das medidas ao longo do tempo nos indivíduos de interesse , bem como da variabilidade entre indivíduos.
- pequena variabilidade entre indivíduos, o estimador é fiável usando apenas informação de outros indivíduos como no CS caso.
- alta variabilidade entre eindivíduos, a informação dos outros indivíduos não é fiável, é necessario considerar variabilidade entre indivíduos.

Desenhos Balanceados e Não Balanceados

Com $Y_{ij} = j^{th}$ medida no indivíduo i e
$t_{ij} =$ tempo em que foi medido Y_{ij}

- Desenho Balanceado: $t_{ij} = t_j$ para todos os indivíduos i

- Desenho Não Balanceado: todos os tempos podem ser diferentes

Nota: Um desenho balanceado pode gerar uma base de dados não balanceada.

Missing Data (Dados Omissos/Faltantes)

- dados faltantes na sequência temporal **(intermittent missing)**: faltar a uma consulta ; instrumento não está a funcionar.

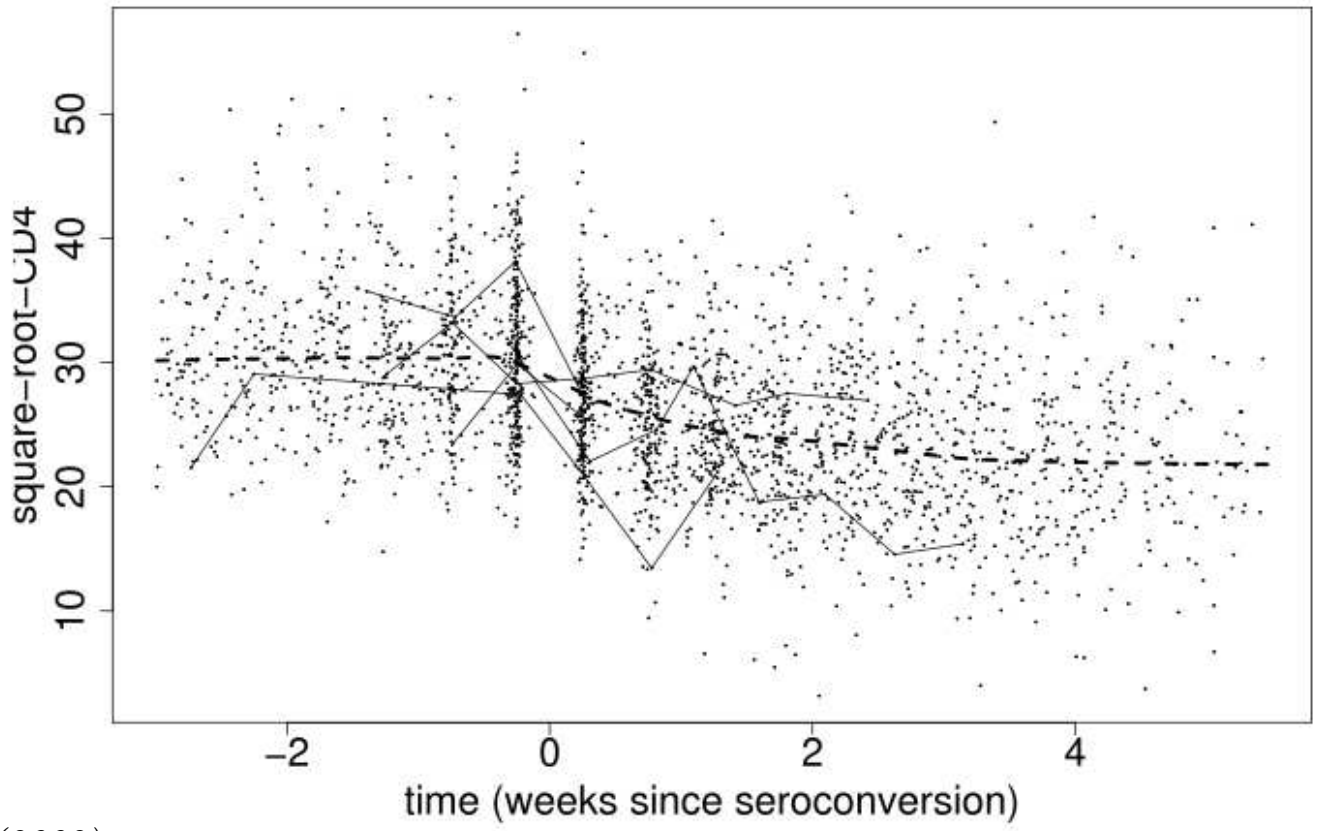
- perder indivíduos em determinado ponto do estudo **(loss to follow-up)**: mudança de endereço ; paciente não quer participar, mas por razões não relacionadas com o tópico de estudo.

- sair do estudo **(dropout)**: morrer; tratamento não está a ajudar.

Exemplo 1: Número de células CD4+

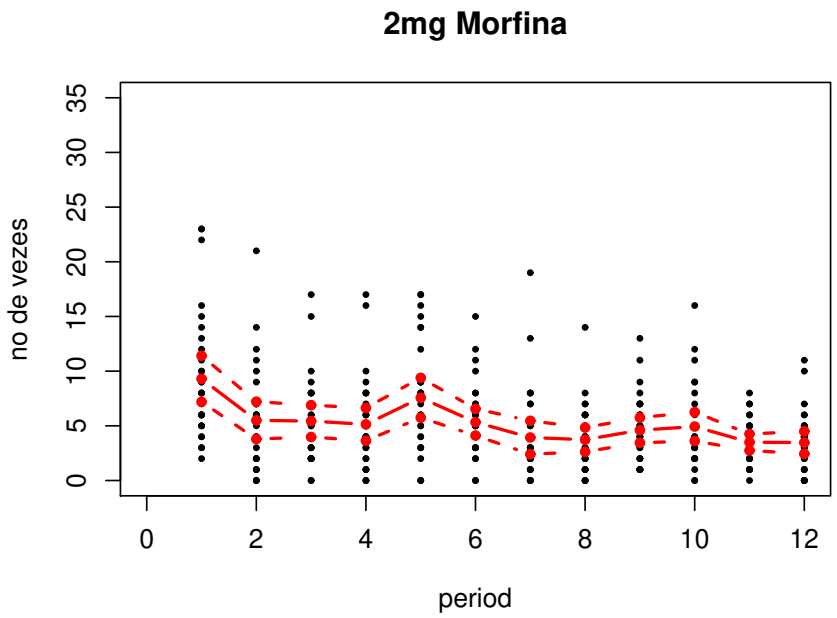
Cohort de 369 HIV seropositivos, medidas do número de células CD4+ a aproximadamente intervalos de 6-meses.

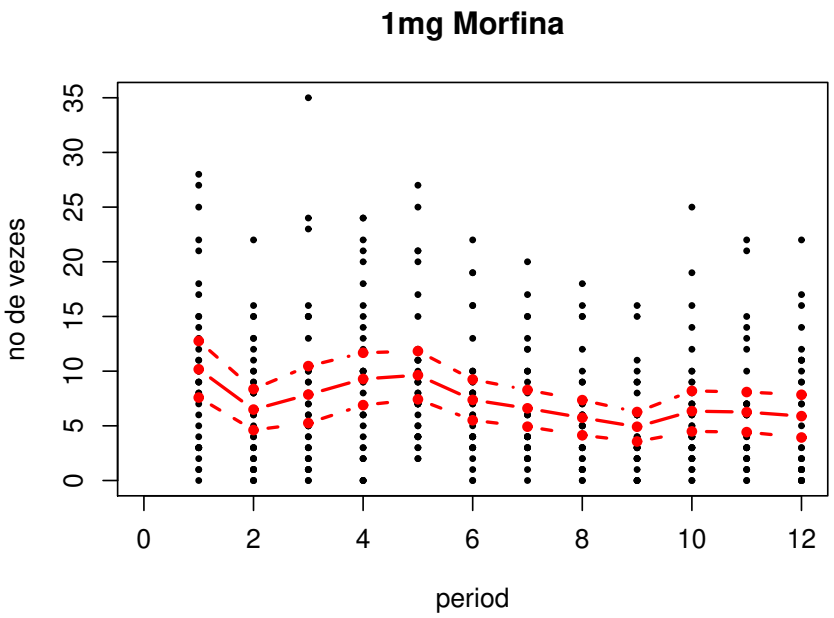
Número de medidas repetidas, e tempos de medidas, são diferentes para cada indivíduo.

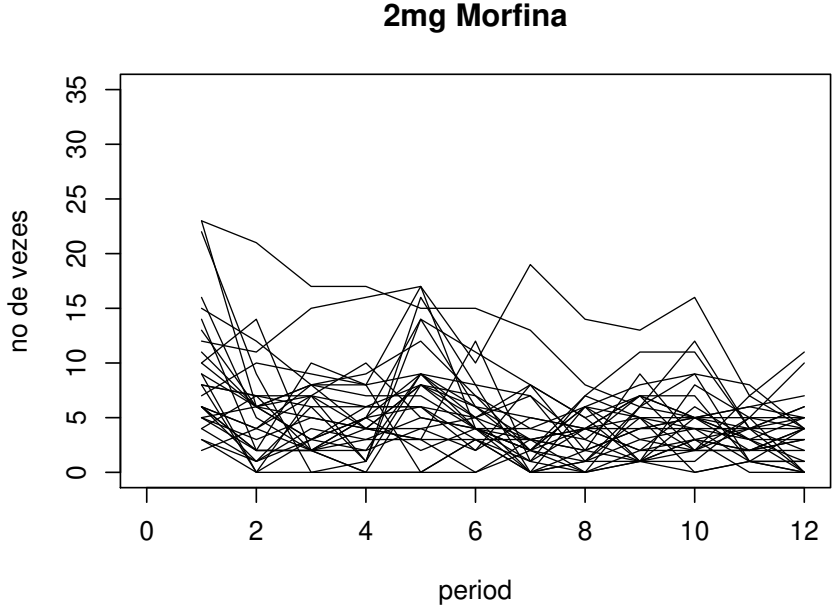


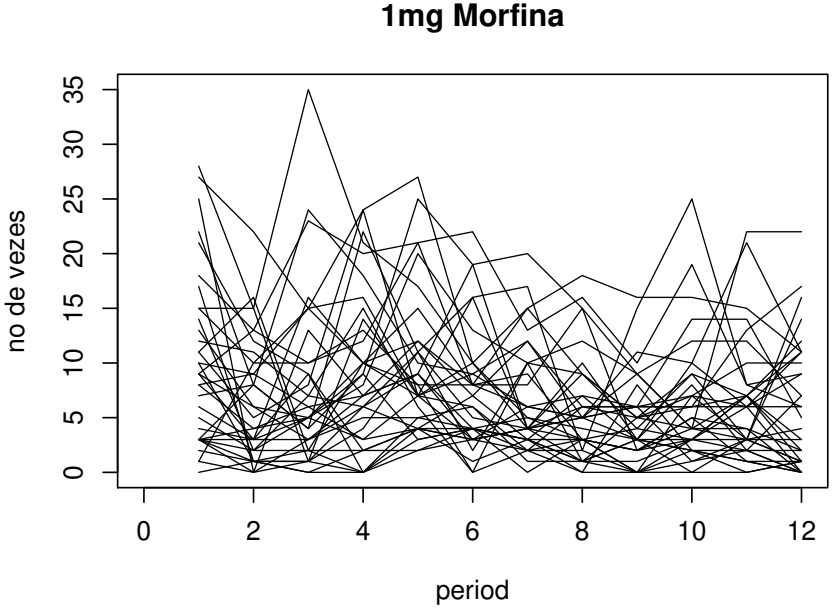
in *Diggle et al (2002)*

Exemplo 2: Ensaio Clínico Anestesia (PCA)

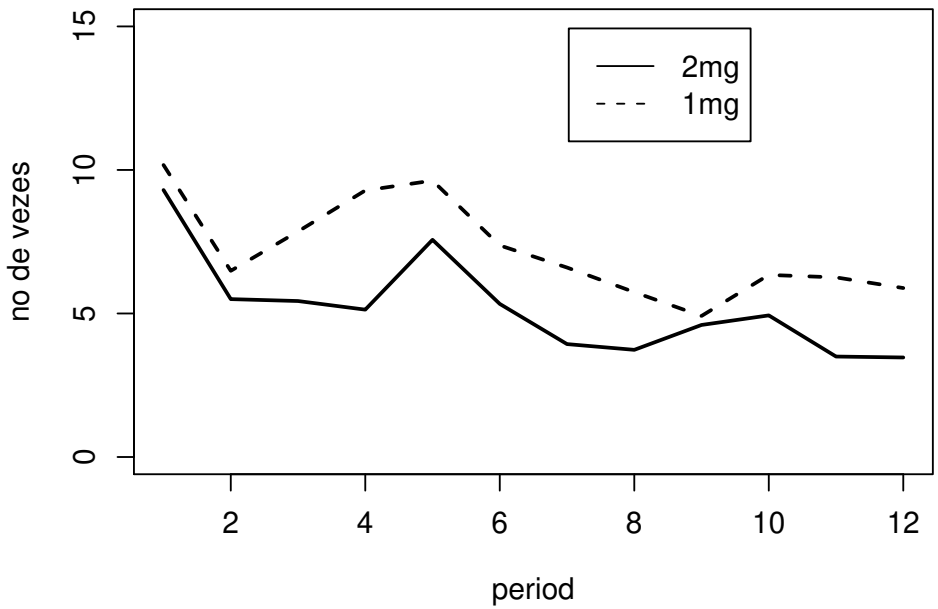


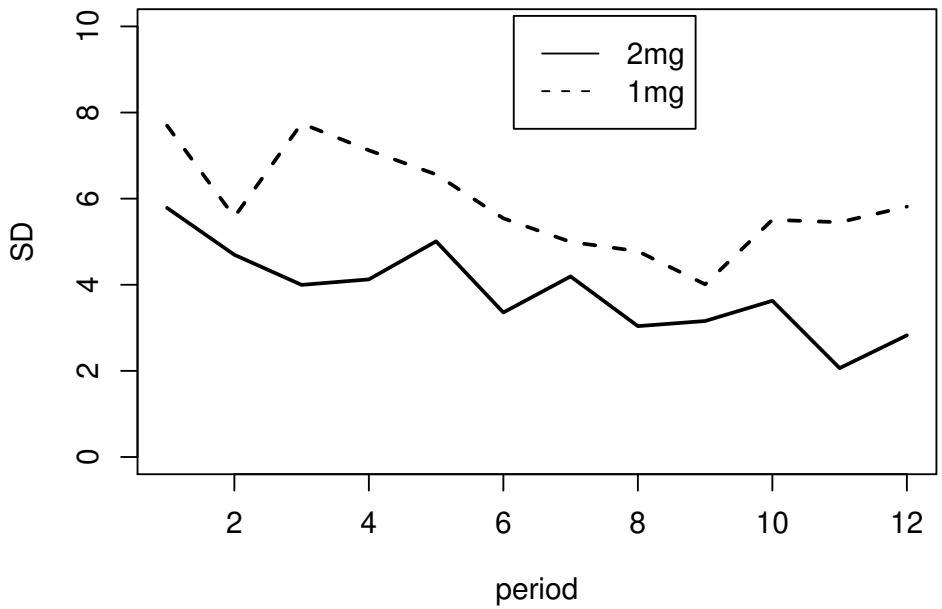




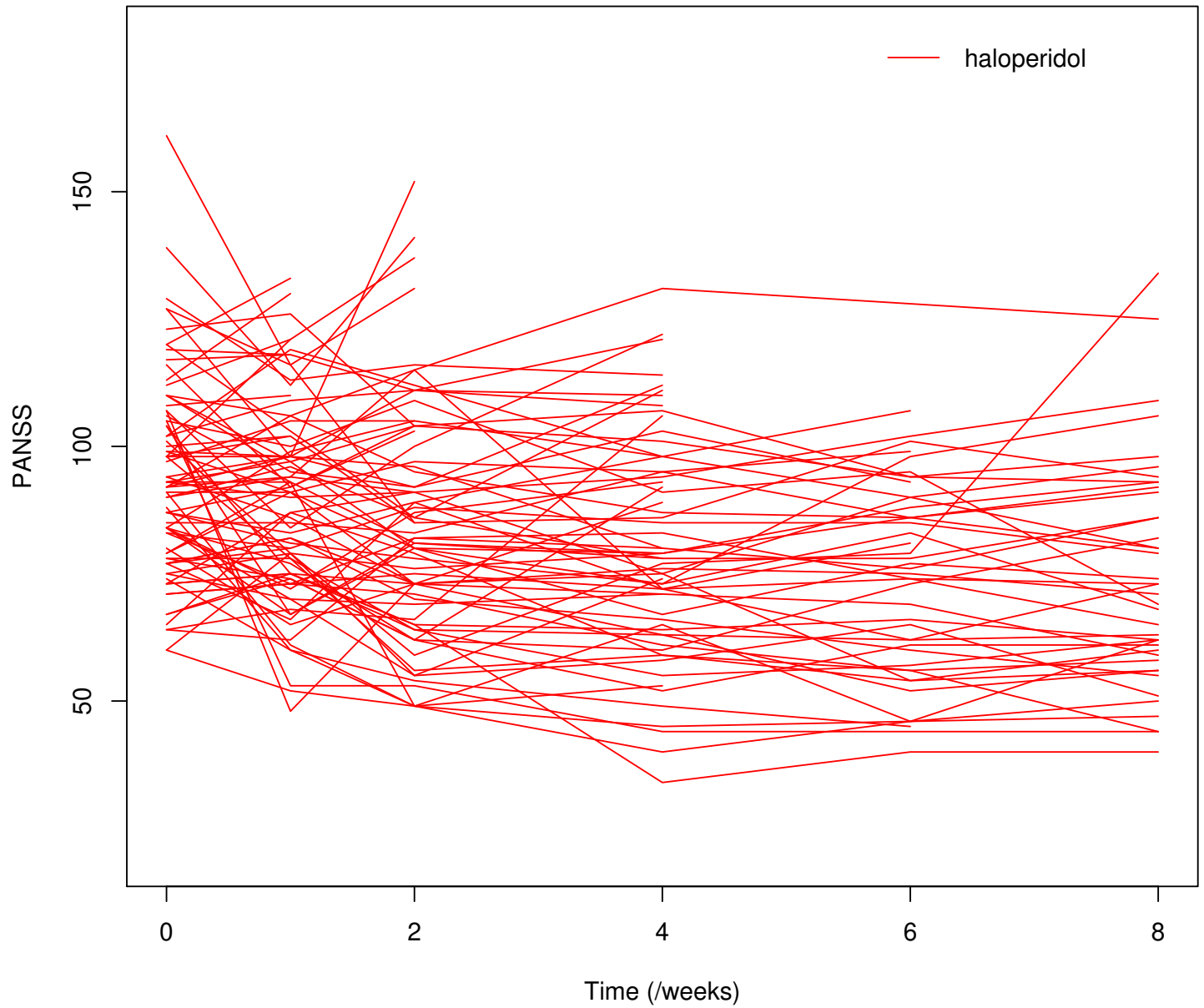


Exemplo 2: Ensaio Clínico Anestesia (PCA)

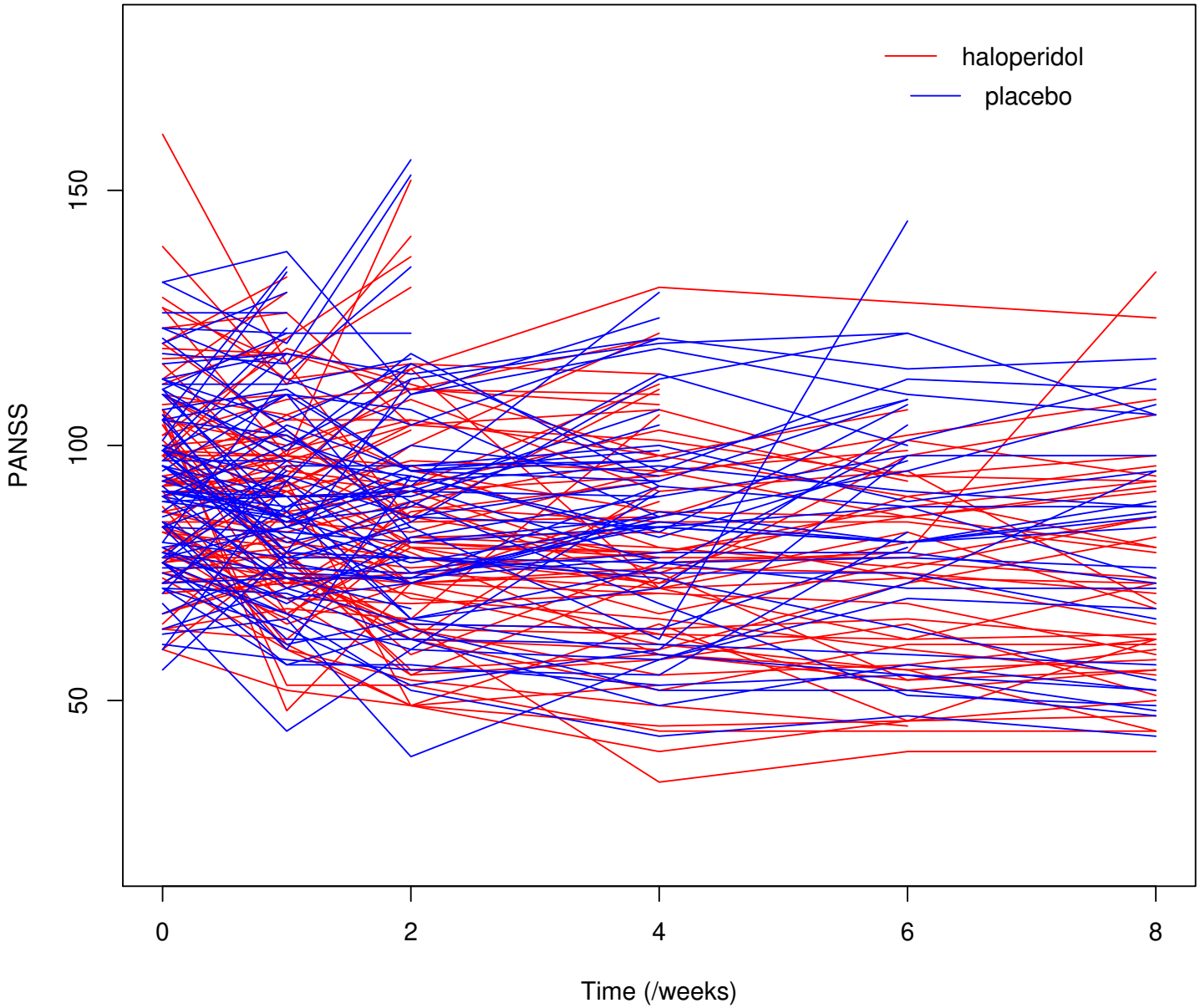




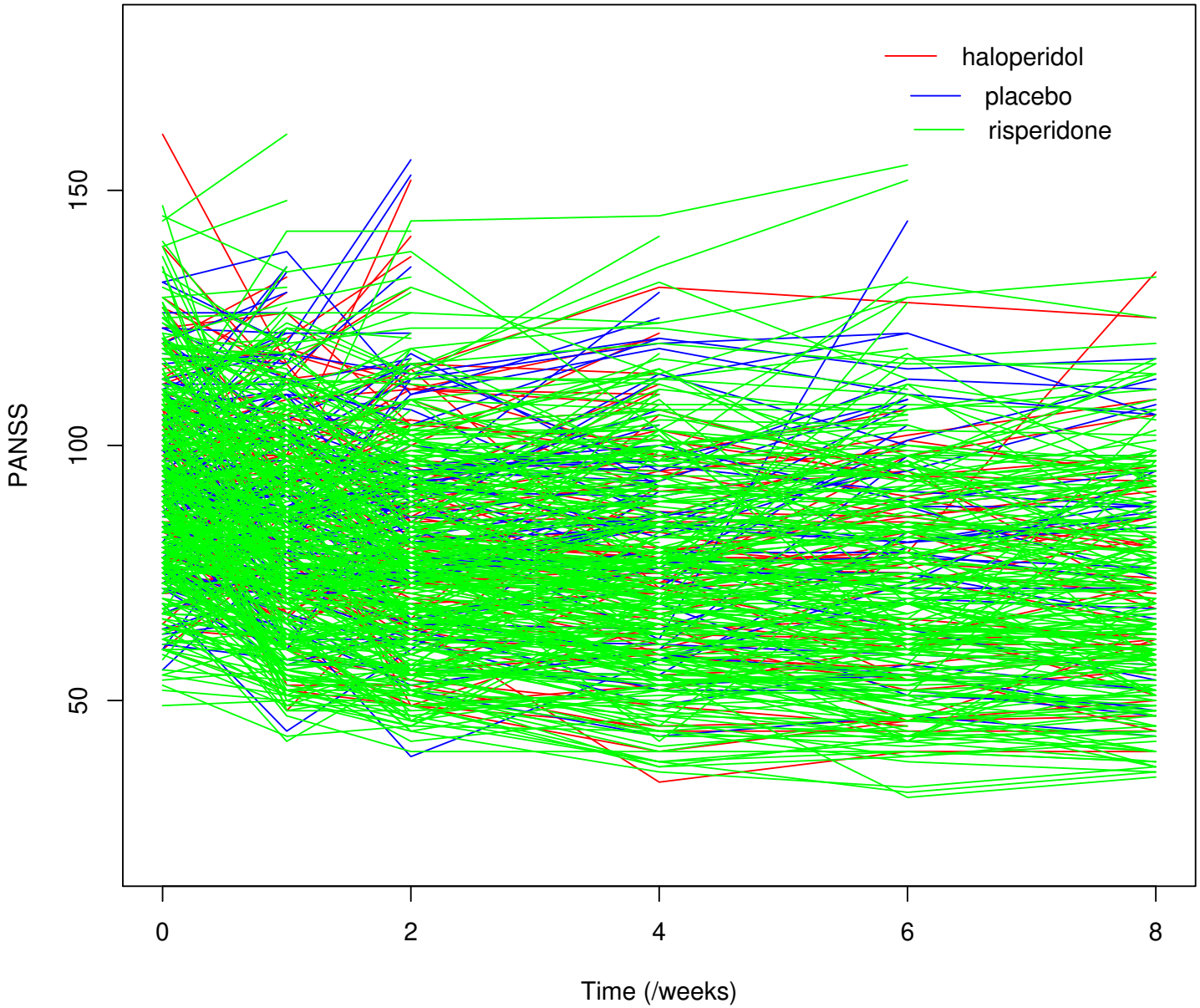
Exemplo 3: Tratamento Esquizofrenia (PANSS)



Exemplo 3: Tratamento Esquizofrenia (PANSS)

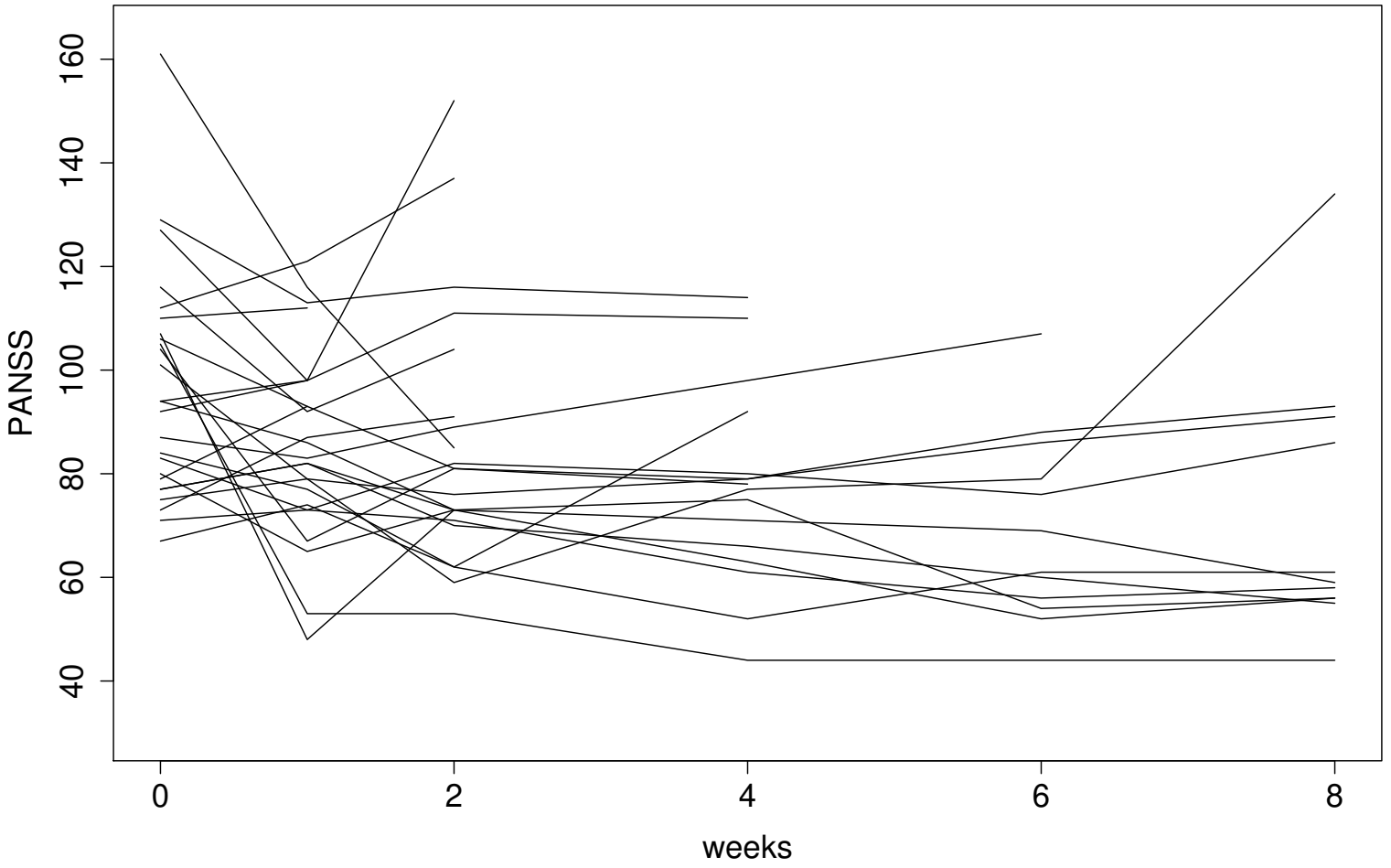


Exemplo 3: Tratamento Esquizofrenia (PANSS)

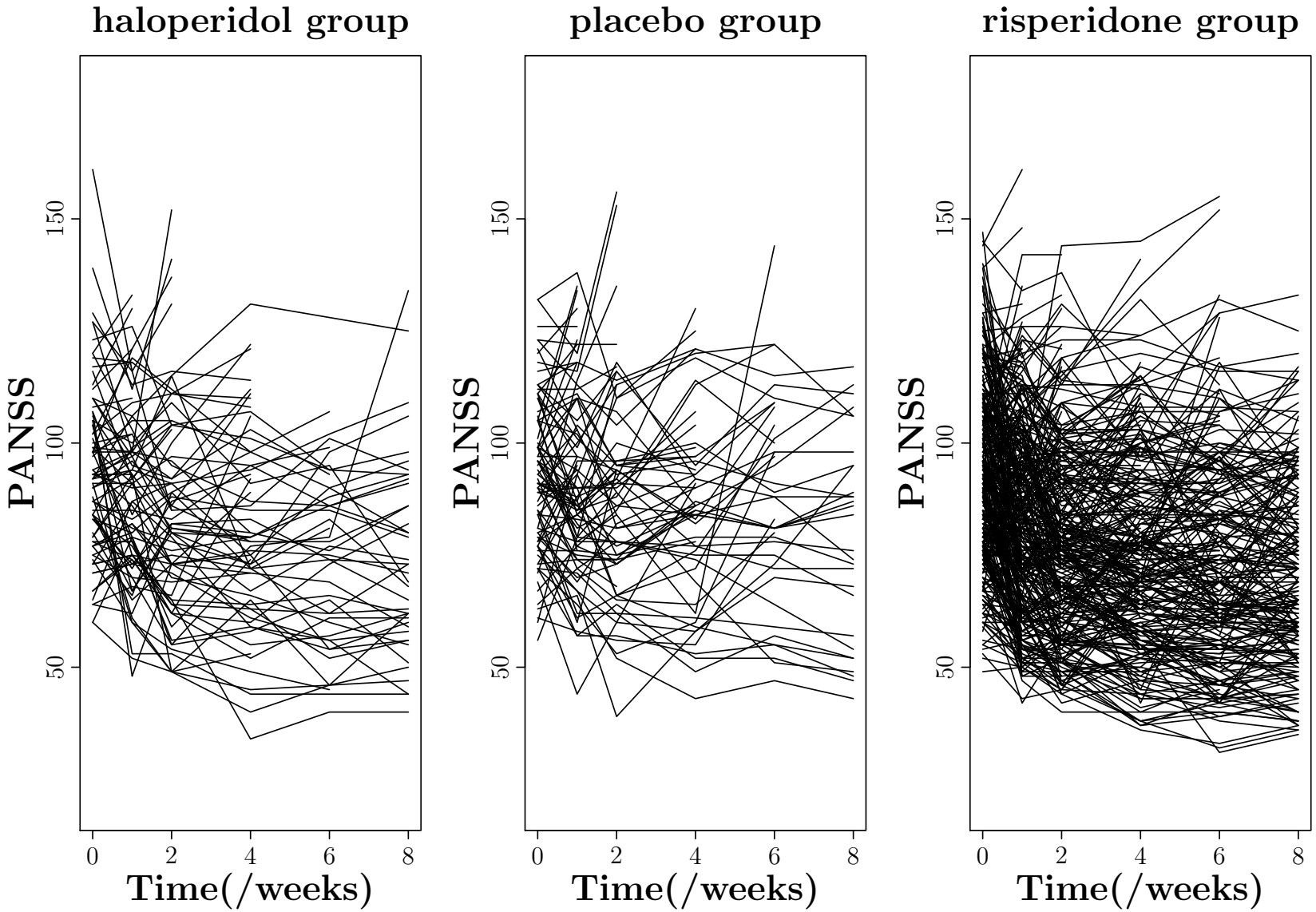


Example 3: Tratamento Esquizofrenia (PANSS)

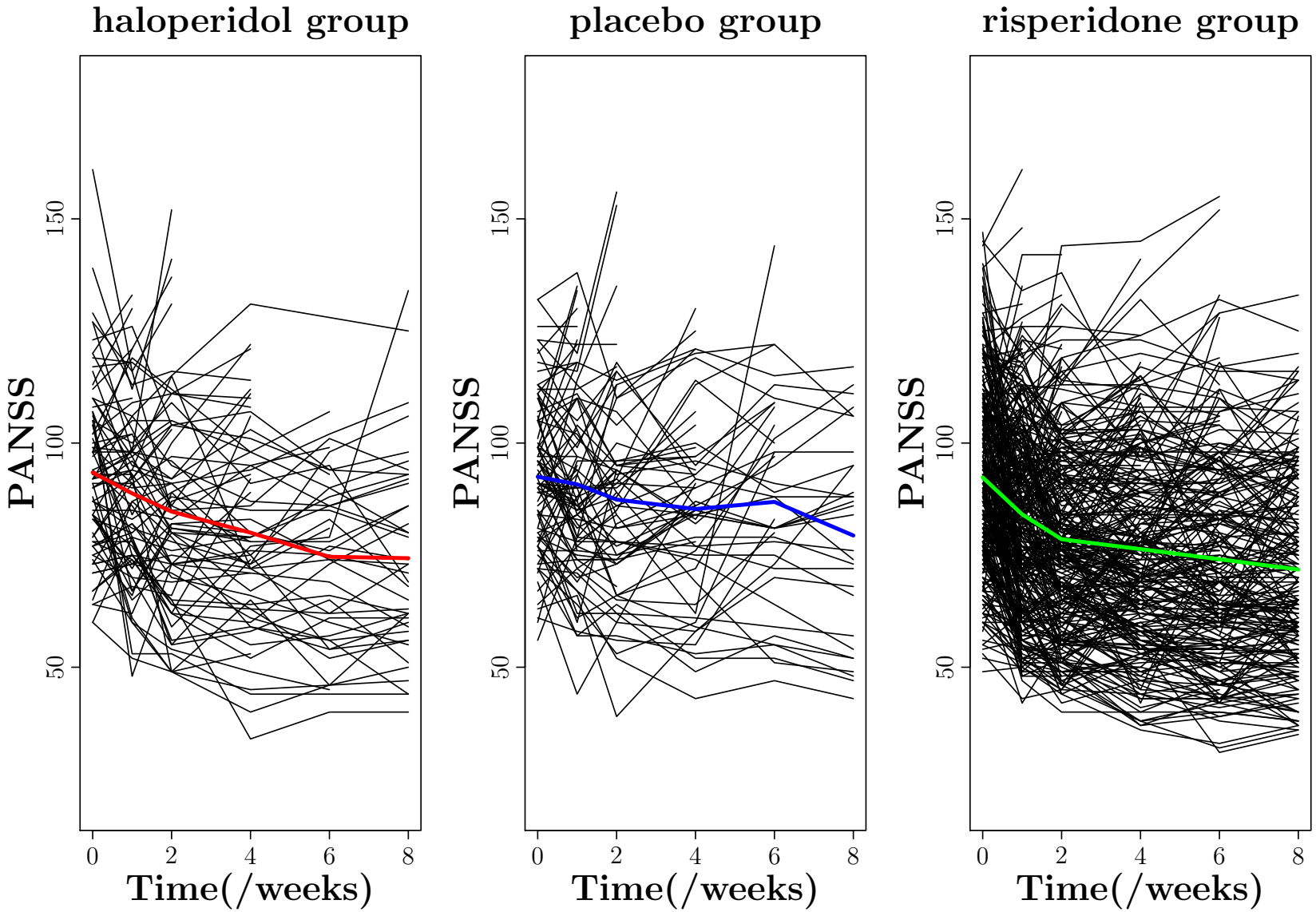
Uma amostra aleatória de progressões individuais



Um gráfico bem melhor - spaghetti plot!



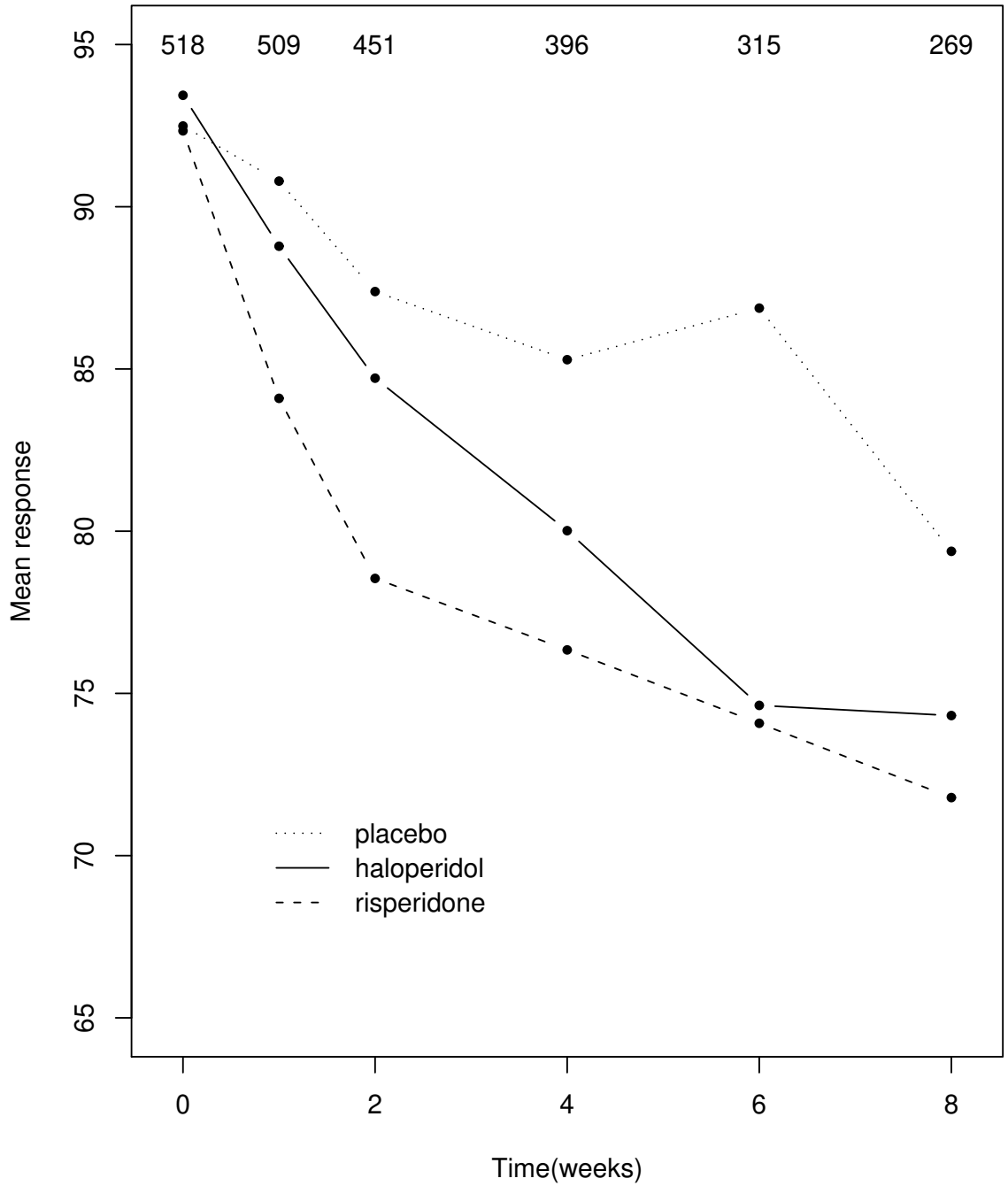
Um gráfico bem melhor - spaghetti plot!



Qual o melhor gráfico a utilizar?

- Gráfico de pontos mostra um decréscimo com o tempo.
- Mas a informação ao nível individual está perdida.
- Plot de linhas é confuso, muito cheio, e não é possível distinguir o efeito dos diferentes tratamentos.
- **Mas**, mostra um decréscimo no tempo, bem como uma tendência para pessoas com picos sairem dos estudo logo de seguida (dropout).
- O gráfico com as médias mostra o efeito do tratamento bastante claro.
- **Mas** não mostra variabilidade da população em torno da média.
- Como seria possível corrigir?

Exemplo 3: Tratamento Esquizofrenia (PANSS)



Example 3 - Tratamento Esquizofrenia(`mental`)

Tabela Sumário - placebo

Week	Mean	Variance	Correlation					
0	55.44	109.15	1.00	0.7	0.49	0.48	0.59	0.43
1	56.96	163.91	.	1.00	0.77	0.63	0.60	0.60
2	53.61	190.14	.	.	1.00	0.76	0.53	0.44
4	52.40	138.25	.	.	.	1.00	0.55	0.65
6	55.43	175.26	.	.	.	.	1.00	0.90
8	52.56	172.13	.	.	.	.	.	1.00

Example 3 - Tratamento Esquizofrenia(`mental`)

Tabela Sumário - haloperidol

Week	Mean	Variance	Correlation					
0	56.60	150.16	1.00	0.64	0.55	0.56	0.45	0.40
1	53.18	149.82	.	1.00	0.75	0.70	0.75	0.66
2	51.18	174.52	.	.	1.00	0.83	0.87	0.75
4	50.75	178.19	.	.	.	1.00	0.91	0.83
6	45.93	173.99	.	.	.	.	1.00	0.92
8	44.64	156.99	.	.	.	.	.	1.00

Example 3 - Tratamento Esquizofrenia(`mental`)

Tabela Sumário - risperidone

Week	Mean	Variance	Correlation					
0	55.28	141.39	1.00	0.51	0.38	0.38	0.27	0.21
1	49.04	119.18	.	1.00	0.80	0.65	0.64	0.60
2	46.84	126.91	.	.	1.00	0.77	0.74	0.68
4	43.42	147.33	.	.	.	1.00	0.88	0.74
6	41.88	247.17	.	.	.	.	1.00	0.80
8	43.44	188.41	.	.	.	.	.	1.00

Example 3 - Tratamento Esquizofrenia(`mental`)

Tabela Sumário - Mais do que um tratamento?

- tabelas separadas para cada grupo de tratamento
- procurar semelhanças e diferenças

Tabela Sumário - Que covariáveis são significativas?

- usar resíduos do modelo ajustado por "ordinary least squares"

Como representar dados não balanceados

Como representar a média para os dados CD4

- Em cada tempo calcular a média. Não ocorrem 2 ou mais medidas exactamente ao mesmo tempo.

- Útil para dados **Não Balanceados**

- *Agrupar os dados*: estimar a média em cada tempo, como a média empírica de todas as observações que ocorrem numa janela (eg. aos 6 meses, numa janela de 5 a 7 meses)

$$\hat{\mu}(t) = \text{average}\,\{y_{ij} : |t_{ij} - t| < h/2\}$$

Como representar dados não balanceados

Como representar a média para CD4 (Kernel Smoothing)

- Função Kernel $k(.)$ (pdf simétrica)

- Com uma bandwidth h

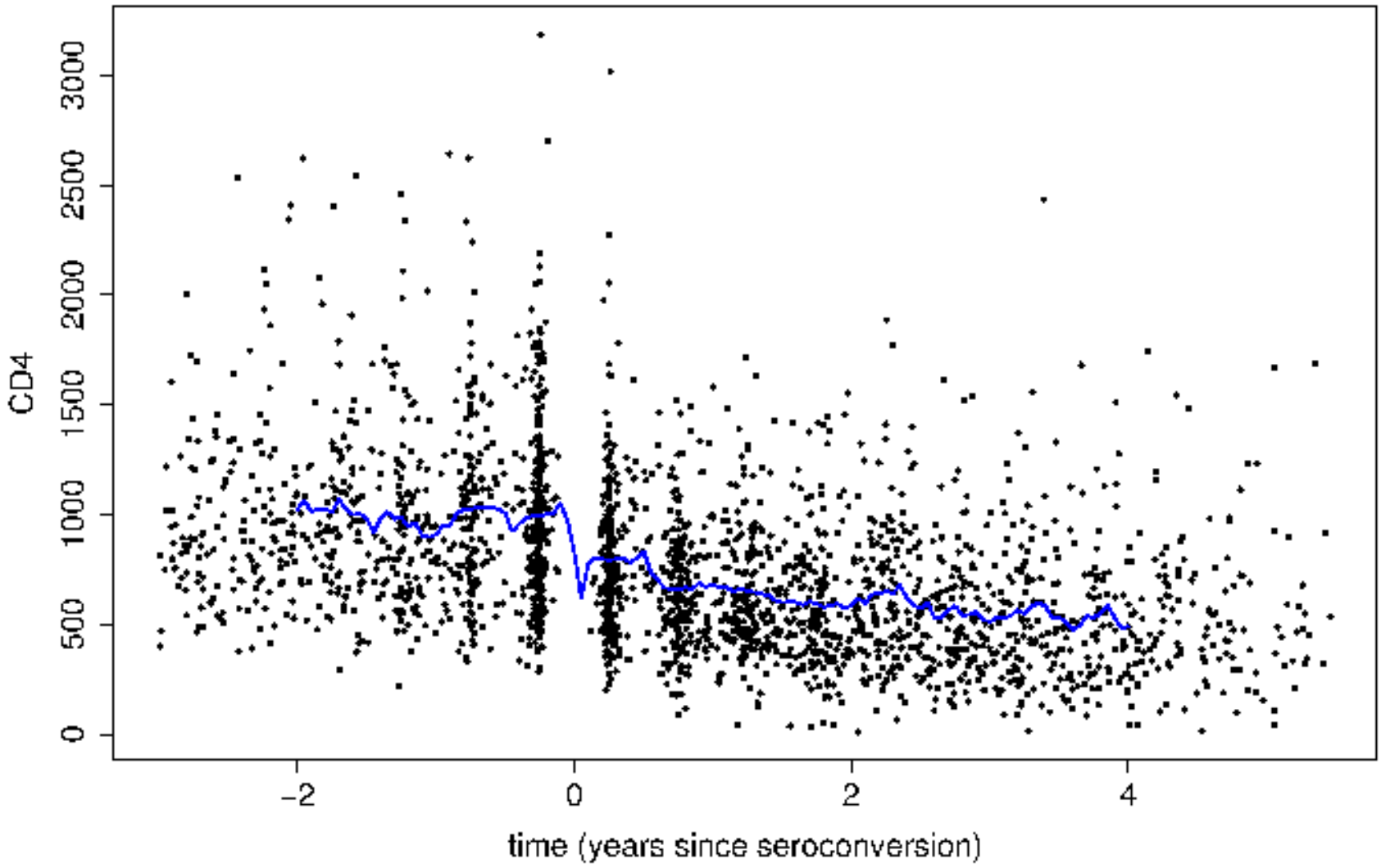
$$\hat{\mu}(t) = \frac{\sum y_{ij}\ k\left\{(t_{ij} - t)/h\right\}}{\sum k\left\{(t_{ij} - t)/h\right\}}$$

Como representar a média para CD4 (Smoothing Spline)

- *Smoothing Spline*: encontrar uma "curva suave", equilibrando minimização dos erros e suavidade
  - pequena penalidade (h) $\Rightarrow$ uma linha que interpola os dados
  - grande penalidade (h) $\Rightarrow$ linha suave que não se ajusta tão bem aos dados
  - penalidade automática (h) $\Rightarrow$ cross-validation, de cada vez retirar um ponto dos dados e encontrar o parâmetro h que minimiza o erro de perdição
  - Hastie & Tibshirani (1990) *Generalized Additive Models* ou Wood (2006) *Generalized Additive Models : An Introduction to* `R`
  - `smooth.spline()`; `lowess()`; `loess()`

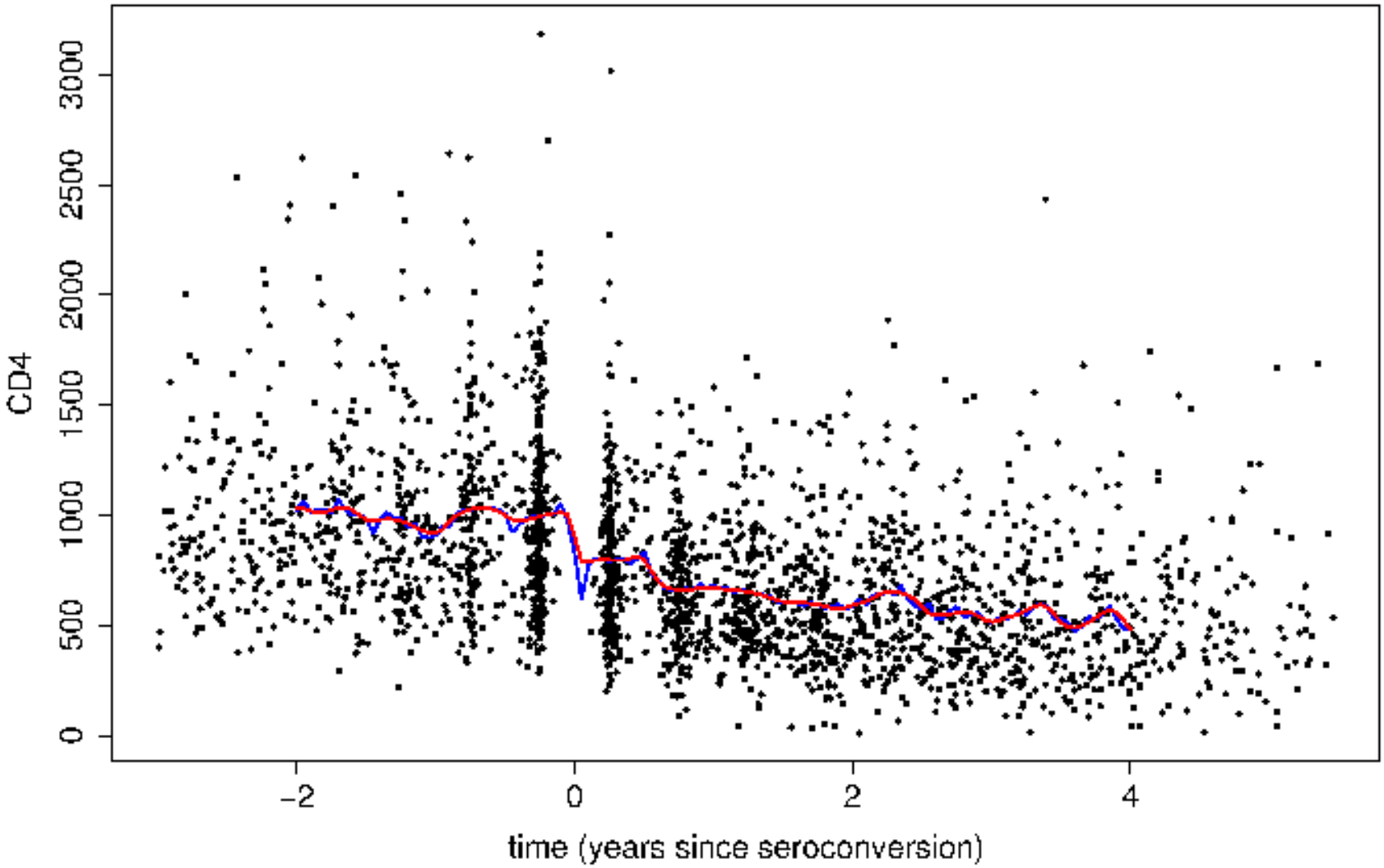
Dados CD4 - Smoothing

Dados e Uniform Kernel



Dados CD4 - Smoothing

Dados, Uniform e Gaussian Kernel



Dados CD4 - Smoothing

Dados, Gaussian Kernels com bandas larga e pequena