

CE-092: Extensões de Modelos de Regressão

1. Considere o modelo “básico” de regressão linear simples relacionando uma variável resposta Y e uma preditora (contínua) x :

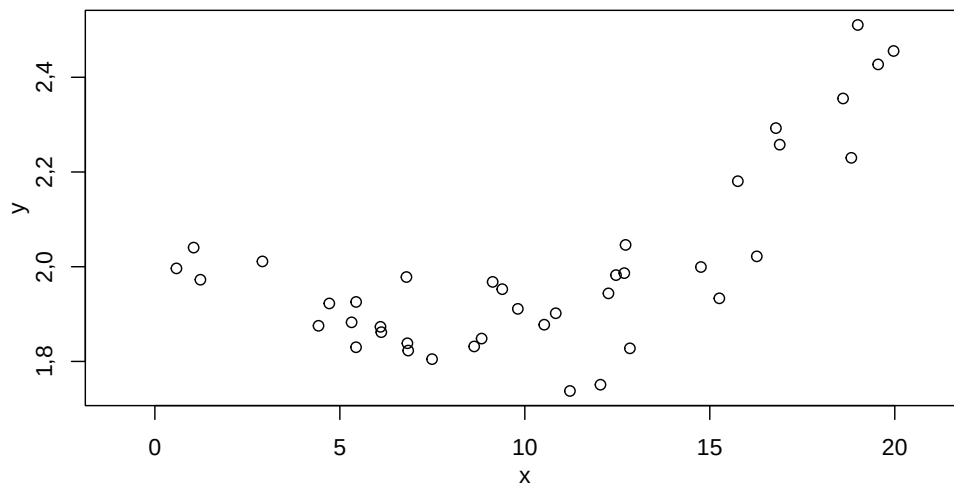
$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Explique de que forma cada um dos modelos/métodos a seguir estende o modelo acima. (Utilize notação adequada para melhor explicação.)

- Regressão com transformação na variável resposta.
- Regressão polinomial.
- Regressão segmentada linear.
- Regressão por *splines* cúbicos.
- Suavização por *splines*.
- Suavização por *kernel*.
- Regressão Poisson.
- Regressão não linear.
- Árvore de regressão.
- Regressão quantílica.
- Regressão heterocedástica.

Regressão Segmentada, Polinomial e Suavização

2. Seja o conjunto de dados mostrado no gráfico a seguir onde a variável resposta Y deve ser modelada a partir de uma variável explicativa x .



Deseja-se comparar ajustes de regressão segmentada com possíveis pontos de corte em $x = 5$ e $x = 11$ e utilizar critérios estatísticos para adotar um modelo com um ou dois pontos de corte. Alguns resultados dos ajustes são fornecidos a seguir.

```
fit1a
```

```
Call:
```

```
lm(formula = y ~ x + x5, data = df)
```

```
Coefficients:
```

(Intercept)	x	x5
2,0871	-0,0631	0,0978

```
fit1b
```

```
Call:
```

```
lm(formula = y ~ x + x11, data = df)
```

```
Coefficients:
```

(Intercept)	x	x11
1,9998	-0,0165	0,0845

```
fit2
```

```
Call:
```

```
lm(formula = y ~ x + x5 + x11, data = df)
```

```
Coefficients:
```

(Intercept)	x	x5	x11
2,0383	-0,0285	0,0173	0,0774

```
c(logLik(fit1a),logLik(fit1b),logLik(fit2))
```

```
[1] 29,68 44,72 45,18
```

- (a) Faça gráfico(s) esboçando o ajuste de cada um dos modelos considerados.
- (b) Mostre como são definidas as variáveis $x5$ e $x11$.
- (c) Escreva a expressão da $E[Y]$ de cada um dos modelos ajustados.
- (d) Quantos parâmetros de média cada modelo possui.
- (e) Segundo o objetivo mencionado, qual o modelo deve ser adotado? Justifique sua resposta.
- (f) Forneça a equação de cada reta do modelo `fit1b` (arredondar coeficientes para três casas decimais).
- (g) O modelo acima pode ser escrito na forma matricial $Y = X\beta + \epsilon$. Forneça a matriz X do modelo.

3. Ainda no contexto da questão anterior, considera-se agora os ajustes fornecidos pelos comandos a seguir.

```
fit3 <- lm(y ~ poly(x, 3, raw=TRUE), data=df)
fit3bs <- lm(y ~ poly(x, 3, raw=TRUE) + ifelse(x<10, 0, I(x-10)^3), data=df)
fit3ns <- lm(y ~ splines::ns(x, knots=c(10)), data=df)
c(logLik(fit3),logLik(fit3bs),logLik(fit3ns))
```

- (a) Quais foram os modelos ajustados? (descreva e declare utilizando notação)
- (b) Quantos parâmetros (na estrutura de média) cada um deles possui?
- (c) Como voce explicaria as diferenças entre os três modelos?
- (d) Entre estes três modelos qual voce adotaria para os dados em questão? Justifique.

4. Discorra sobre os itens a seguir.

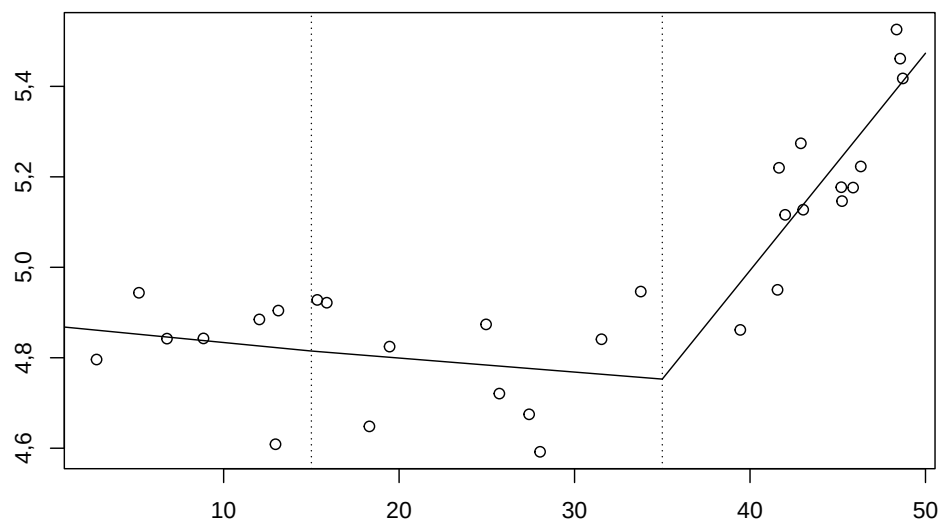
- (a) Explique o que são “função base” (*basis functions*) indicando quais são nas duas questões anteriores.
- (b) Explique em que sentido *regressão por splines*, *suavização por splines* e suavização por polinômios locais (**lowess**) generalizam o modelo de regressão linear simples. Aponte a(s) diferença(s) entre os métodos. Forneça os passos de cada algoritmo.
- (c) Qual a função do kernel em algoritmos de suavização que utilizam este elemento?
- (d) Uma das dificuldades relacionadas ao ajuste de modelos não lineares, além da própria escolha do modelo, é a proposição de valores iniciais para os parâmetros, necessários para métodos iterativos de estimação. Discuta sobre de que maneiras pode-se tornar esse processo mais efetivo.
- (e) Em modelos de regressão em geral, não é aconselhável fazer previsões considerando valores fora do intervalo observado das variáveis explicativas. No entanto, existe uma preocupação menor com relação a isso quando se utiliza modelos de regressão não linear. Por que?

5. Considere um ajuste de um modelo de regressão por *splines* cúbicos com um ponto de corte. Tal modelo teria então para cada parte (antes e depois do ponto de corte) uma equação do tipo:

$$E[Y] = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 ,$$

com valores dos coeficientes diferentes em cada parte e portanto oito coeficientes. Entretanto, deseja-se que o modelo forneça uma curva que seja contínua e suave no ponto de corte.

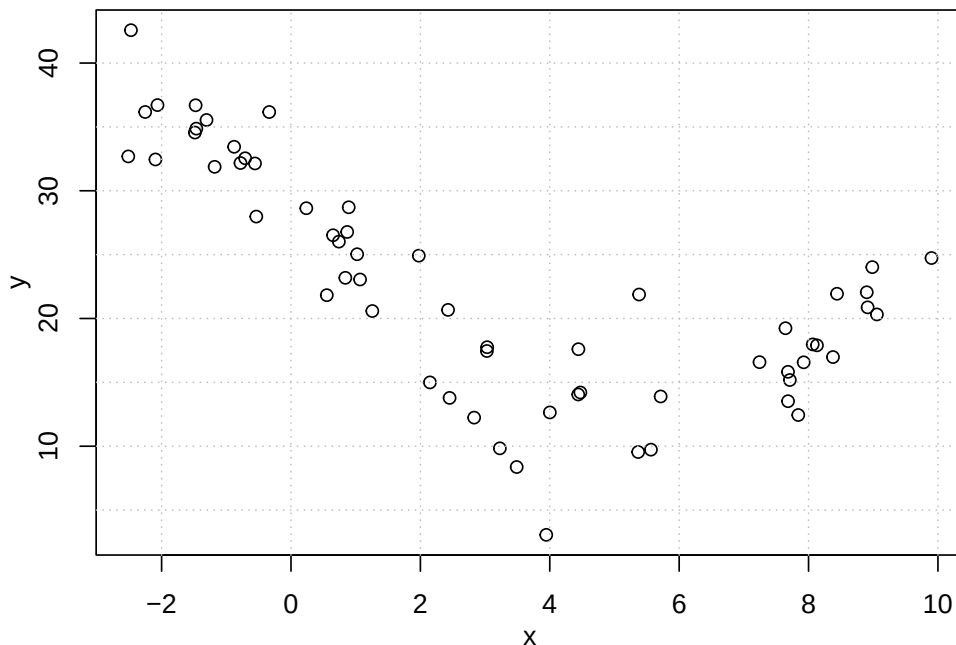
- (a) Quantos parâmetros teria então tal modelo?
 - (b) Deduza as relações destes parâmetros com os oito coeficientes do modelo irrestrito.
 - (c) Qual expressão poderia ser utilizada para ajustar tal modelo ?
6. Considere o ajuste de um modelo de regressão por partes com dois pontos de corte. O modelo pode ser descontínuo no primeiro ponto de corte, porém deve necessariamente ser contínuo e suave no segundo ponto de corte. Quantos graus de liberdade teria tal modelo? Escreva a expressão do modelo.
7. Ajustou-se o modelo segmentado com dois pontos de corte a um conjunto de dados conforme mostrado no gráfico a seguir. Também é mostrada a estimativa dos coeficientes. Obtenha a equação de cada reta ajustada (utilize valores arredondados dos coeficientes).



```
Call:
lm(formula = y ~ x + x15 + x35, data = df)

Coefficients:
(Intercept)          x          x15          x35
  4,871348    -0,003766    0,000664    0,051131
```

8. Nas questões a seguir considere que deseja-se usar algum método e/ou modelo para explicar o comportamento (da média) de uma variável resposta Y a partir de valores de uma covariável (preditora) x .
- Considere que deseja ajustar um modelo de regressão linear por partes para estes dados, fixando dois pontos de corte que são considerados conhecidos.
 - Qual a expressão do modelo e o algoritmo de estimação se o modelo admitir descontinuidades nos pontos de corte? Quantos parâmetros possui este modelo?
 - Qual a expressão do modelo e o algoritmo de estimação se o modelo exigir continuidade nos pontos de corte? Quantos parâmetros possui este modelo? Mostra como a equação da reta de cada parte pode ser obtida.
 - Escreva os passos do algoritmo para ajuste do modelo por polinômios locais.
 - Se voce tiver como opções adotar um modelo linear (regressão linear simples) ou um modelo não-linear para estes dados, que análises, comparações formais e considerações voce adotaria para decidir qual deles adotar como o modelo escolhido?
 - Suponha agora que voce possui mais de um modelo não linear a ser considerado para o ajuste. Qual(ais) critério(s) voce adotaria para escolha do modelo?
9. Foram observados dados de uma variável resposta Y como função da covariável denotada por x que estão apresentados no gráfico a seguir.



Considere o modelo não linear descrito por

$$f(x) = \beta_0 + \beta_1 x + \beta_2(x - \beta_3)I(x > \beta_3),$$

em que β_i com $i \in \{0, 1, 2, 3\}$ são os parâmetros e $I()$ é a função indicadora.

Dado o gráfico, o modelo não linear e a interpretação dos parâmetros, encontre valores iniciais para os parâmetros razoáveis com sua definição e dados fornecidos. Justifique a lógica por trás da obtenção de cada um dos valores.

Regressão Não Linear

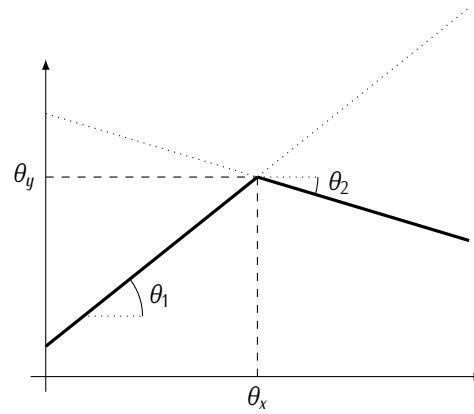
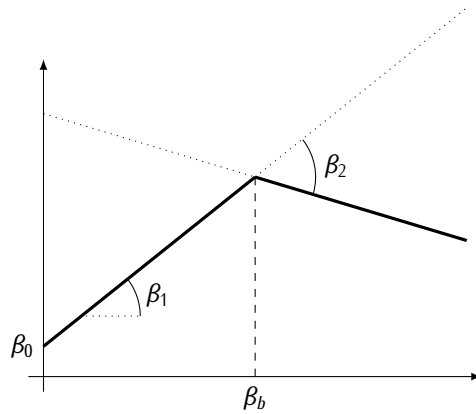
10. Em algumas situações, o pesquisador possui um valor razoável para o parâmetro ponto de mudança de regime (β_b) o que faz com que este, portanto, não seja estimado. Isso acontece quando conhece-se o ponto de mudança de regime (e.g. mudança de uma nova lei sobre investimentos na série financeira, momento de troca de fornecedor de matéria prima na taxa de itens defeituoso, data de início de uma dieta em um programa de hipertrofia, etc.). Quando β_b tiver que ser estimado concomitantemente com os demais, tem-se um modelo não linear. Caso contrário o modelo é linear. Mostre que o modelo é não linear quando β_b é desconhecido.

$$f(x) = \begin{cases} \beta_0 + \beta_1 x & \text{se } x \leq \beta_b \\ \beta_0 + \beta_1 x + \beta_2(x - \beta_b) & \text{se } x > \beta_b \end{cases}$$

11. Uma das técnicas relacionadas ao ajuste de modelos não lineares é a necessidade de fornecer valores iniciais para os parâmetros.
- Por que há esta necessidade nos modelos não lineares mas não nos modelos lineares?
 - A número de valores iniciais a serem fornecidos será sempre igual ao número de parâmetros? Justifique.
 - Se você fosse fazer um ajuste de modelos não lineares, qual ou quais opções consideraria para definir os valores iniciais?
12. Considere duas parametrizações do modelo bilinear apresentadas na figura a seguir. Na parametrização (1) os parâmetros representam o intercepto (β_0), a taxa no primeiro segmento (β_1), o ponto de mudança de regime (β_b) e a diferença de taxa entre o primeiro e segundo segmentos (β_2).

$$(1) f(x) = \beta_0 + \beta_1 x + \beta_2(x - \beta_b)I(x > \beta_b)$$

$$(2) f(x) = ???$$



- (a) Descreva o significado dos parâmetros da segunda parametrização (2) indicados na figura.
- (b) O modelo ajustado com a primeira parametrização (1) teve as estimativas $\hat{\beta}_0 = 0.1$, $\hat{\beta}_1 = 1.1$, $\hat{\beta}_2 = -2.1$ e $\hat{\beta}_b = 2$. Qual é o valor para as estimativas dos parâmetros na segunda parametrização?
- (c) Obtenha a expressão do modelo escrito sob a segunda parametrização.

13. Deseja-se ajustar um modelo de regressão linear simples

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

aos dados da tabela a seguir.

Descreva o procedimento a ser utilizado no ajuste do modelo.

X	0.4	1.2	1.8	1.9	2.0	6.8	7.6	8.3	8.7	9.3	10.7	11.3	13.0	13.4	14.2
Y	0.8	< 1.8	[1,5; 20]	2.4	> 2	2.9	3.6	< 4	[2,5; 4]	> 4.5	< 4	3.2	4.1	[4; 5]	< 5

14. No modelo a seguir aponte quais são (a) lineares, (b) parcialmente lineares e (c) completamente não lineares.

(a) $\theta_a(1 - \exp\{-\theta_c \cdot x\})$;

(b) $\frac{\theta_a \cdot x}{\theta_v + x}$;

(c) $\begin{cases} \theta_0 + \theta_1 \cdot x & , x \leq x_b; \\ \theta_0 + \theta_1 \cdot x_b & , x > x_b \end{cases}$;

(d) $\begin{cases} \theta_0 + \theta_1 \cdot x & , x \leq \theta_b; \\ \theta_0 + \theta_1 \cdot \theta_b + \theta_2(x - \theta_b) & , x > \theta_b \end{cases}$;

(e) $x(\theta_0 + \theta_1 \cdot x)^{-1/\theta_2}$;

15. Mostre como o modelo não-linear a seguir pode ser escrito como um modelo linearizável.

$$Y_i = \frac{\alpha x_i}{x_i + \beta}.$$

GAM e GAMLSS

- Os modelos aditivos generalizados (**gam**'s) têm se tornado cada vez mais utilizados em análises estatísticas. Tomando o modelo de regressão linear simples como ponto de referência inicial, comente sobre as principais características e motivações para uso dos **gam**'s. Descreva pelo menos três aspectos em que generalizam tal modelo de referência inicial. Além disto, comente sobre alguma característica ou desvantagem em relação aos modelos de regressão múltipla e quando devem ser adotados em substituição a estes.

Modelos de Efeitos Aleatórios

2. Como voce explicaria o que é um “modelo de efeitos aleatórios”? Cite ao menos três situações nas quais estes modelos se apresentam como uma opção razoável de análise. Em que sentido o modelo de efeitos aleatórios “estende” o modelo de regressão usual?
3. Escreva a forma geral, com suposições usuais, do modelo de efeitos aleatórios para resposta gaussiana. Mostre como é obtida a expressão da distribuição marginal das respostas e como fica a expressão da função de verossimilhança neste caso.
4. Considere um modelo possivelmente adequado para modelar uma resposta de contagem Y que registra o número mensal de movimentações em conta bancária, como função de uma covariável x que representa o saldo médio no mês. Mostre a estrutura de um modelo com apenas um efeito aleatório para estes dados. Explique cada termo do modelo e sua interpretação no contexto do problema. Escreva a função de verossimilhança para este modelo e comente como pode se obter as estimativas dos parâmetros. Forneça os passos do algoritmo que pode ser usado para simular deste modelo.
5. Ao considerar um modelo de regressão linear simples para um conjunto de dados verificou-se que o pressuposto de homocedasticidade não era adequado. Descreva ao menos quatro diferentes abordagens que um estatístico poderia considerar adotar, explicando suas características e comentando em que situações poderiam (ou não) ser adotadas.
6. O artigo discutido no curso de Singer et al. utiliza um conjunto de dados com quatro observações tomadas em cada um de cinco locais de coleta. Os dados referem-se então somente ao local de coleta e valores resposta (poluente) e são mostrados a seguir.

local	1	1	1	1	2	2	2	2	3	3	3	3
Poluente	170,6	154,7	136,4	153,1	68,0	66,4	70,3	71,1	151,5	138,0	128,4	118,1
local	4	4	4	4	5	5	5	5				
Poluente	153,9	149,1	147,5	103,3	83,9	101,4	117,3	114,1				

É adotado um modelo de efeitos aleatórios na análise.

- (a) Forneça a expressão do modelo adotado, definindo adequadamente os termos.
- (b) Explique por qual motivo o modelo de efeitos aleatórios é adequado para este problema ao invés do modelo de efeitos fixos.
- (c) Quantos e quais são os parâmetros do modelo? E os efeitos aleatórios?
- (d) O método da ANOVA foi adotado no artigo para estimação dos parâmetros. Explique o funcionamento do método listando os passos para obtenção das estimativas dos parâmetros do modelo.
- (e) Foi discutido o uso da verossimilhança como método alternativo de estimação. Forneça a expressão da verossimilhança do modelo e explique passo a passo como seriam obtidas as estimativas neste caso.
- (f) No modelo de efeitos aleatórios, uma vez obtidas as estimativas dos parâmetros, os efeitos podem ser “preditos”. Explique como é obtida tal predição. Quantos efeitos aleatórios seriam preditos neste caso?
- (g) Os modelos de efeitos aleatórios gaussianos pode ser denotados de forma geral como

$$Y = X\beta + F\gamma + \epsilon,$$

em que Y é o vetor de respostas, X é a matriz de delineamento dos efeitos fixos e β o respectivo vetor de parâmetros, F matriz de delineamento dos efeitos aleatórios e γ é o vetor de efeitos aleatórios. Forneça os termos deste modelo para o exemplo do artigo indicando a dimensão dos vetores e matrizes.

7. Suponha que no contexto do artigo de Singer et al. a variável resposta fosse um dado de contagem e assumida uma distribuição Poisson para resposta.

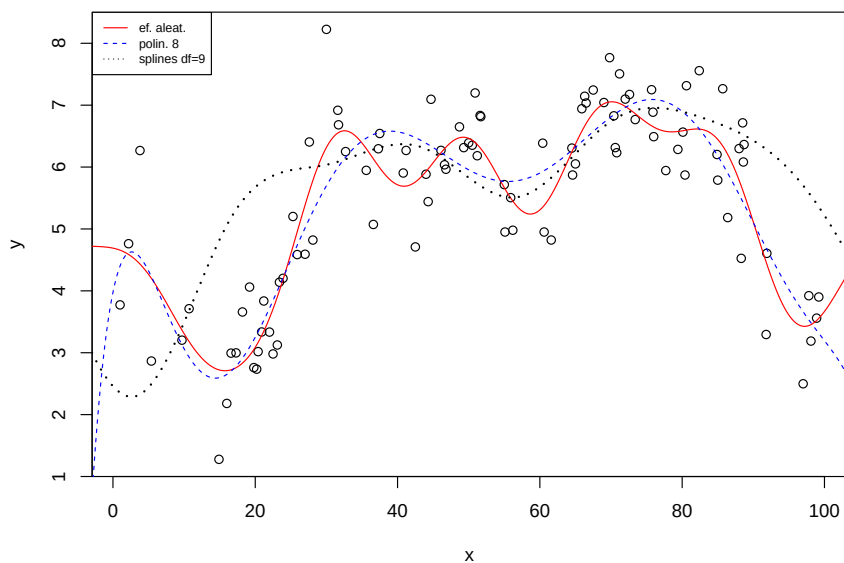
- (a) Como ficaria então a especificação do modelo?
- (b) Quais seriam os parâmetros do modelo?
- (c) Como os parâmetros poderiam ser estimados?
- (d) Como seria a expressão de função de verossimilhança?
- (e) Em comparação com o exemplo anterior, haveria alguma diferença no cálculo da predição dos efeitos aleatórios? Se positivo, qual seria?

8. Os itens a seguir se referem ao artigo de Singer *et al.* referenciado no curso.

- (a) Escreva a especificação do modelo considerado no artigo na forma de notação geral (matricial) vista em aulas para modelos de efeitos aleatórios Gaussianos. Forneça as matrizes e suas dimensões.
- (b) Carregue os dados do artigo e reproduza análise dos dados utilizando algum pacote/função do R. Mostre o comando e as estimativas dos parâmetros fixos e aleatórios.
- (c) Mostre comandos para simular dados utilizando a mesma estrutura e estimativas obtidas do modelo para os dados do artigo.
- (d) Defina a partir de sua notação o coeficiente de correlação intraclasse e explique seu significado.

9. Considere os ajustes por um modelo de efeitos aleatórios, polinômio de oitavo grau e *splines* a um conjunto de dados mostrados na figura a seguir. Comente:

- (a) Como seria a especificação do modelo em cada caso?
- (b) Em que a especificação do modelo de efeitos aleatórios difere da dos demais modelos?
- (c) Quantos e quais são os efeitos aleatórios e parâmetros do modelo.
- (d) Como seriam e quais seriam as dimensões das matrizes e vetores do modelo? (considere a notação geral dada nas questões anteriores do modelo de efeitos aleatórios).



10. Considere um estudo longitudinal no qual três indivíduos foram observados em cinco tempos. Há interesse em ajustar uma reta para descrever a resposta de cada um dos indivíduo ao longo do tempo. Considere que deseja-se ajustar o modelo gaussiano de efeitos aleatórios que tem como expressão geral

$$Y = X\beta + F\gamma + \epsilon,$$

com termos definidos em uma questão anterior. Forneça a expressão de cada termo do modelo, incluindo a dimensão das matrizes e vetores envolvidos. Indique quantos e quais são os efeitos aleatórios e parâmetros do modelo. Faça ainda um esboço de um gráfico que mostraria os dados e o ajuste adotado neste caso,

11. Considere um estudo longitudinal no qual quatro indivíduos foram observados em 3, 6, 4 e 3 instantes de tempo, respectivamente. Há interesse em ajustar regressões distintas para descrever a resposta de cada um dos indivíduo como uma função linear do tempo de observação. Opta-se por ajustar o modelo gaussiano de efeitos aleatórios que tem como expressão geral

$$Y = X\beta + F\gamma + \epsilon,$$

com descrição dos termos conforme discutido no curso.

- (a) Discuta as opções para inclusão dos efeitos aleatórios no modelo (efeitos no intercepto e/ou nas inclinações). Explique a diferença entre as opções e como é o comportamento esperado das observações sob cada modelo.
 - (b) Forneça a expressão escalar e matricial dos modelos, descrevendo cada termo e incluindo a dimensão das matrizes e vetores envolvidos. Indique quantos e quais são os efeitos aleatórios e parâmetros do modelo.
 - (c) Faça ainda um esboço de um gráfico que mostraria os dados e o ajuste adotado neste caso,
12. Considere um contexto semelhante ao item anterior porém agora em um estudo real com 50 pacientes no qual mediu-se a concentração de um elemento medido em um exame. Foram tomadas medidas de cada um dos pacientes que foram observados oito vezes ao longo do tempo. Neste estudo deseja-se estimar uma curva logística (sigmoide)

$$f(x) = \frac{\beta_0}{1 + e^{-\beta_1(x-\beta_2)}}$$

para cada paciente. Os parâmetros têm a seguinte interpretação: β_0 é o valor máximo da curva, β_1 é a declividade da curva e β_2 é o ponto médio da curva sigmoide.

- (a) Como voce classificaria este modelo, considerando as opções vistas no curso? Justifique.
- (b) O modelo poderia ser escrito na forma do modelo da questão anterior? Justifique.
- (c) Justifique porque seria possível e razoável considerar efeitos aleatórios neste caso.
- (d) Como os efeitos aleatórios poderiam ser incluídos? O que representariam? Quantos efeitos aleatórios seriam então definidos.