Planejamento e Análise Estatística de Experimentos com um único fator: análise de dados de experimentos completamente aleatorizados (Parte 2)

Prof. Caio Azevedo

Identificação das médias (grupos de médias) diferentes

- Uma vez que a ANOVA indica a existência de diferenças entre as médias surge o interesse em identificar os padrões dessas diferenças.
- Todas as médias são diferentes entre si? Existem grupos de médias cujos elementos são iguais entre si mas diferentes entre grupos?
- Qual a melhor forma de responder as perguntas de interesse (melhor forma de implementar as comparações)?

Contrastes

- As hipóteses relativas à igualdades de médias podem postas como

$$H_0 : D_{(q\times k)}\boldsymbol{\mu}_{(k\times 1)} = \mathbf{0}_{(q\times 1)} \text{ vs } H_1 : D_{(q\times k)}\boldsymbol{\mu}_{(k\times 1)} \neq \mathbf{0}_{(q\times 1)}$$

  em geral, $q \leq k$ e D é uma matriz de posto linha completo $=$ q.

- Em geral, as linhas da matriz D são vetores chamados de **contrastes**.
- Contraste: vetor cuja soma de seus elementos é igual à 0.

Exemplos

- Voltando ao exemplo 2 (solventes).
- Suponha que queremos testar :
  - $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$, ou seja, queremos testar

    $H_1 : \mu_1 - \mu_2 = 0$ vs $H_1 : \mu_1 - \mu_2 \neq 0$.
  - Neste caso, em termos da estrutura $D\boldsymbol{\mu}$ temos que as a hipóteses podem ser escritas como:

$$H_0 : D_{(1\times 5)}\boldsymbol{\mu} = 0 \text{ vs } H_1 : D_{(1\times 5)}\boldsymbol{\mu} \neq 0,$$

em que

$$D = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \end{bmatrix}; \boldsymbol{\mu}' = \begin{bmatrix} \mu_1 & \mu_2 & \mu_3 & \mu_4 & \mu_5 \end{bmatrix}$$

Exemplos (cont.)

- Suponha que queremos testar a veracidade de :

$$H_0 : \begin{cases} \mu_1 - \mu_2 = 0, \text{ e} \\ \mu_1 - \mu_3 = 0 \end{cases}$$

- Neste caso, em termos da estrutura $D\boldsymbol{\mu}$ temos que as a hipóteses podem ser escritas como:

$$H_0 : D_{(2\times 5)}\boldsymbol{\mu} = \mathbf{0}_{(2\times 1)} \text{ vs } H_1 : D_{(2\times 5)}\boldsymbol{\mu} \neq \mathbf{0}_{(2\times 1)},$$

em que

$$D = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \end{bmatrix}; \boldsymbol{\mu}' = \begin{bmatrix} \mu_1 & \mu_2 & \mu_3 & \mu_4 & \mu_5 \end{bmatrix}$$

Exemplos (cont.)

- Em termos da parametrização casela de referência, temos que as hipóteses anteriores devem ser reescritas, respectivamente, como:
  - $H_0 : \alpha_2 = 0$ vs $H_1 : \alpha_2 \neq 0$
  - $$H_0 : \begin{cases} \alpha_2 = 0, \text{ e} \\ \alpha_3 = 0 \end{cases} \quad \text{vs } H_1 : \text{pelo menos uma diferença}$$

Exemplos (cont.)

- Dessa forma, poderíamos reescrever as hipóteses (na parametrização CR), em termos matriciais, ou seja:

$$H_0 : C_{(r\times p)}\beta_{(p\times 1)} = \mathbf{0}_{(r\times 1)} \text{ vs } H_1 : C_{(r\times p)}\beta_{(p\times 1)} \neq \mathbf{0}_{(r\times 1)}$$

em geral, $r \leq p$ e C é uma matriz de posto linha completo = r.

Exemplos (cont.)

- Lembremos que $\boldsymbol{\beta}' = \begin{bmatrix} \mu & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 \end{bmatrix}$.
- Dessa forma, teríamos, para cada uma das hipóteses anteriories, que:
  - $C = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \end{bmatrix}$
  - $C = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$
- Como testar as hipóteses $H_0 : C\boldsymbol{\beta} = \mathbf{0}$?

Estatística do Teste

- Lembremos que $\widehat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2 \left(X'X\right)^{-1})$.
- Assim, $C\widehat{\boldsymbol{\beta}} \sim N_p(C\boldsymbol{\beta}, \sigma^2 C \left(X'X\right)^{-1} C')$.
- Portanto, sob H_0, $Q^* = \frac{1}{\sigma^2}\left(C\widehat{\boldsymbol{\beta}}\right)' \left(C \left(X'X\right)^{-1} C'\right)^{-1} \left(C\widehat{\boldsymbol{\beta}}\right) \sim \chi^2_{(r)}$.
- Além disso, $SQR/\sigma^2 = (n-k)\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{(n-k)}$.
- Logo *sob* H_0, $Q = \frac{1}{r\widehat{\sigma}^2}\left(C\widehat{\boldsymbol{\beta}}\right)' \left(C \left(X'X\right)^{-1} C'\right)^{-1} \left(C\widehat{\boldsymbol{\beta}}\right) \sim F_{(r,n-k)}$.
- pvalor $P(F > q|H_0)$, em que $F \sim F_{(r,n-k)}$ e q é o valor observado de Q.

Aplicação no exemplo 2

- Lembrando os grupos : grupo 1(E50), grupo 2(E70), grupo 3(EAW), grupo 4(M1M), grupo 5(MAW)
- Considere as hipóteses (H_0)
  - $H_{01} : \mu_1 = \mu_2$.
  - $H_{02} : \begin{cases} \mu_1 - \mu_2 = 0, \text{ e} \\ \mu_1 - \mu_3 = 0 \end{cases}$
  - $H_{03} : \mu_1 = \mu_3$.
  - $H_{04} : \frac{\mu_1+\mu_2+\mu_3}{3} = \frac{\mu_4+\mu_5}{2}$.

Continuação: em termos das parametrização CR

- Considere as hipóteses (H_0)
  - $H_{01} : \alpha_2 = 0.$
  - $H_{02} : \begin{cases} \alpha_2 = 0, \text{ e} \\ \alpha_3 = 0 \end{cases}$
  - $H_{03} : \alpha_3 = 0.$
  - $H_{04} : 2\alpha_2 + 2\alpha_3 - 3\alpha_4 - 3\alpha_5 = 0.$

Estatísticas (valores p)

- Resultados:
  - H_{01} : $18,47 (< 0,0001)$.
  - H_{02} : $9,35 (< 0,0014)$
  - H_{03} : $2,98 (0,0998)$.
  - H_{04} : $581,90 (< 0,001)$.

Descrição do Exemplo 3

- Tem-se o interesse em se saber se a quantidade de fósforo existente (administrada) no solo afeta a produção de milho (de uma certa variedade).
- Fator: quantidade de fósforo, k $= 5$ níves, $n_i = 4, i = 1, 2, 3, 4$ repetições por tratamento (quantidade de fósforo administrada).
- Procedimento: 20 porções de terras, chamadas de parcelas, (em condições semelhantes) foram consideradas e cada uma delas recebeu uma determinada quantidade de fósforo, de modo aleatório (completamente casualizado).

Descrição do Exemplo 3 (cont.)

- Resposta: produção de milho (kg/parcela).
- Experimento balanceado.
- Fator é quantitativo.

Conjunto de dados

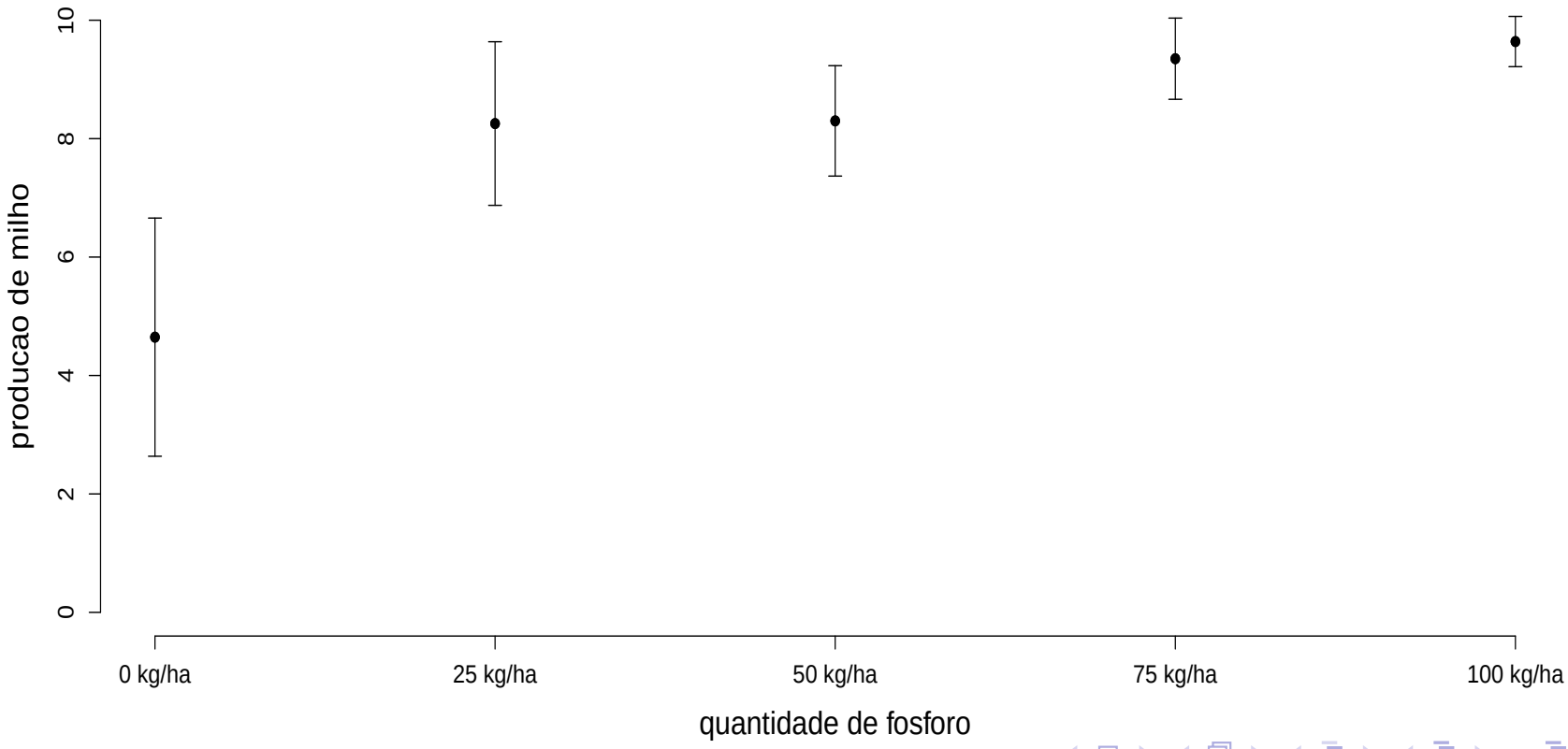
Quantidade de Fósforo	**Produção**			
	1	2	3	4
0 kg/ha	2,38	6,77	3,50	5,94
25 kg/ha	6,15	8,78	8,99	9,10
50 kg/ha	9,07	8,73	6,92	8,48
75 kg/ha	9,55	8,95	10,24	8,66
100 kg/ha	9,14	10,17	9,75	9,50

Análise descritiva

Não há sentido em construir box-plots ou histogramas.

Fósforo	**Medida descritiva**					
	Média	**DP**	**Var.**	**CV%**	**Mínimo**	**Máximo**
0 kg/ha	4,65	2,05	4,21	44,15	2,38	6,77
25 kg/ha	8,26	1,41	1,99	17,07	6,15	9,10
50 kg/ha	8,30	0,96	0,90	11,46	6,92	9,07
75 kg/ha	9,35	0,70	0,49	7,48	8,66	10,24
100 kg/ha	9,64	0,43	0,19	4,49	9,14	10,17

Gráfico de perfis médios



Modelo (casela de referência)

$$Y_{ij} = \mu + \alpha_i + \xi_{ij}, i = 1, 2, 3, 4, 5$$

(grupos); $j = 1, ..., 4$(unidades experimentais)

- Erros $\xi_{ij} \overset{i.i.d}{\sim} N(0, \sigma^2)$, μ, α_i não aleatório.
- $\mathcal{E}_{\xi_{ij}}(Y_{ij}) = \mu_i, \mathcal{V}_{\xi_{ij}}(Y_{ij}) = \sigma^2$.
- $\mu + \alpha_i$: média populacional relacionada ao i-ésimo fator, $\alpha_1 = 0$.
- $Y_{ij} \overset{ind.}{\sim} N(\mu + \alpha_i, \sigma^2)$.

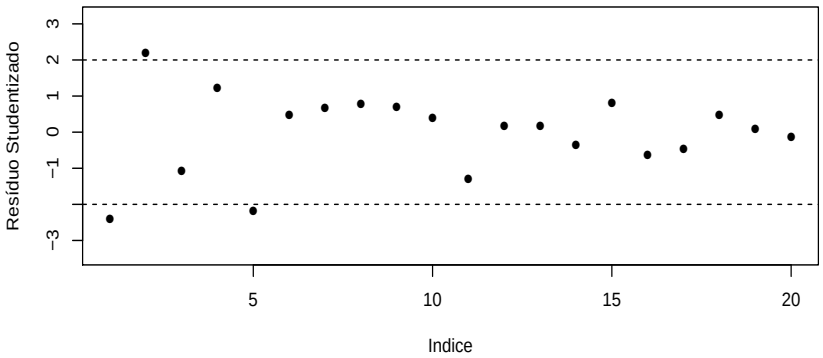
Suposições do modelo

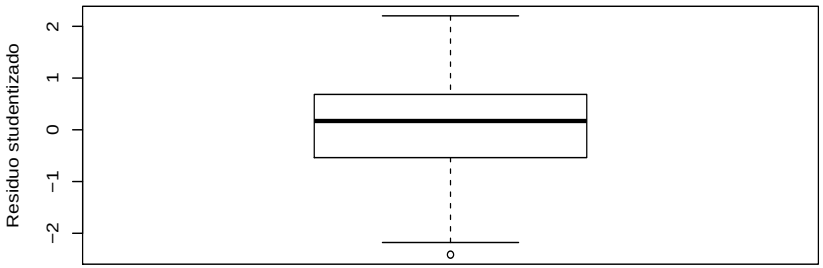
- Homocedasticidade.
- Normalidade dos erros.
- Independência entre as observações.

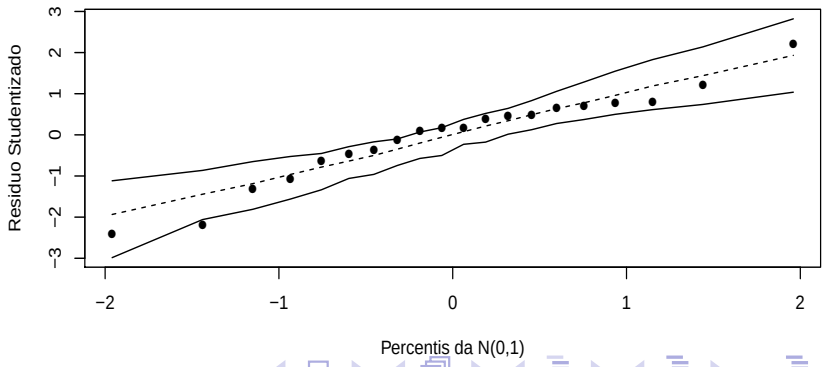
Testes para homocedasticidade

- Teste de Bartlett : 6,82 (0,1460).
- Teste de Levene : 2,12 (0,1275).
- Hipótese de homocedasticidade parace não ser desprezível (cautela).

Análise de resíduos







Comentários

- As suposições do modelo parece que não estão sendo satisfeitas pelo conjunto de dados.
- Ausência de homocedasticiades, normalidade e dependência.
- Uma alternativa: modelos de regressão com distribuição positiva e assimétrica para a variável resposta, que permita variâncias diferentes entre os grupos e com diferentes coeficientes de variação.
- Distribuições positivas: família gama (mãs não a tradicional), família normal inversa, família Weibull, família lognormal, família Birbaun-Saunders, normal assimétrica (apesar de ter suporte na reta).

Comentários

- Contudo: o comportamento dos resíduos sugerem que uma distribuição assimétrica negativa pode ser apropriada (normal assimétrica).
- Atenção: o modelo de regressão normal linear não é adequado para analisar os dados em questão.
- Contudo, seguiremos com ele por questões pedagógicas.

Tabela ANOVA

FV	SQ	GL	QM	Estatística F	pvalor
Nível de fósforo	63,60	4	15,90	10,22	0,0003
Resíduo	23,34	15	$<$ 1,56		
Total	86,94	19			

Rejeita-se H_0.

Estimativas dos parâmetros do modelo

Parâmetro	Estimativa	EP	IC(95%)	Estat. t	pvalor
μ (0)	4,65	0,62	[3,42;5,87]	7,45	<0,0001
α_2 (25)	3,61	0,88	[1,88;5,34]	4,09	0,0010
α_3 (50)	3,65	0,88	[1,92;5,38]	4,14	0,0010
α_4 (75)	4,70	0,88	[2,97;6,43]	5,33	<0,0001
α_5 (100)	4,99	0,88	[3,26;6,72]	5,33	<0,0001

Parâmetros α significativos (diferentes de zero). Porém, parece não haver diferença entre alguns deles.

Continuação: em termos das parametrização CR

- Grupos : 1 ($0kg/ha$), 2 ($25kg/ha$), 3 ($50kg/ha$), 4 ($75kg/ha$), ($100kg/ha$).
- Hipóteses de interesse (H_0)
  - $H_{01} : \mu_2 = \mu_3 = \mu_4 = \mu_5$.
  - $H_{02} : \mu_2 = \mu_3$.
  - $H_{03} : \mu_4 = \mu_5$.
  - $H_{04} : \frac{\mu_2+\mu_3}{2} = \frac{\mu_4+\mu_5}{2}$.
- Como ficam as hipóteses acima em termos da estrutura $C\boldsymbol{\beta} = \mathbf{0}$?

Estatísticas (valores p)

- Resultados:
  - H_{01} : $11,31(0,3087)$.
  - H_{02} :$< 0,01(0,9600)$
  - H_{03} : $0,11(0,7469)$.
  - H_{04} : $3,81(0,0699)$.

Comentários

- Os resultados indicam que a média de produção de milho na ausência de fósforo é menor do que com qualquer outra quantidade.
- Além disso, há uma equivalência na produção de milho entre as quantidades $25kg/ha$ e $50kg/ha$, como também entre as quantidades $75kg/ha$ e $100kg/ha$.
- Contudo, a significância em relação à um possível crescimento na produção de milho, quando se passa da quantidade de fósforo entre $25kg/ha - 50kg/ha$ para $75kg/ha - 100kg/ha$ é marginal.

Comentários cont.

- Partindo da premissa de que a igualdade estrita nunca é verdade, que o modelo não se ajustou bem e que o número de observações por tratamento é, aparentemente, pequeno, vamos considerar que o crescimento é significativo.

Modelo reduzido (casela de referência)

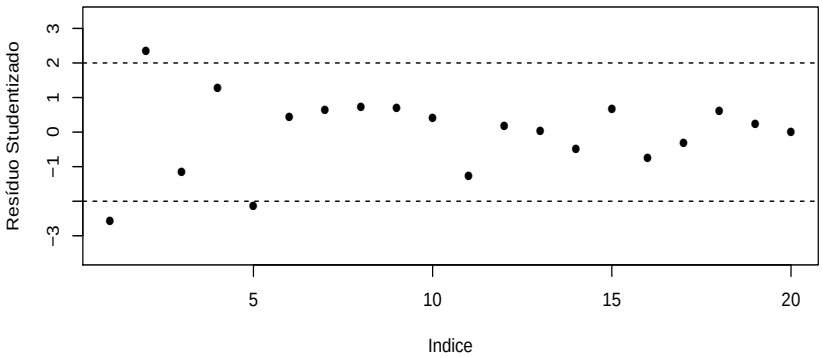
$$Y_{ij} = \mu + \alpha_i + \xi_{ij}, i = 1, 2, .., 5$$

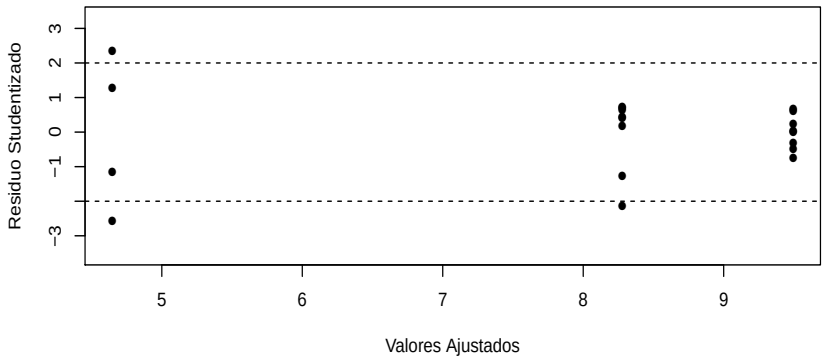
(grupos); $j = 1, ..., 5$(unidades experimentais)

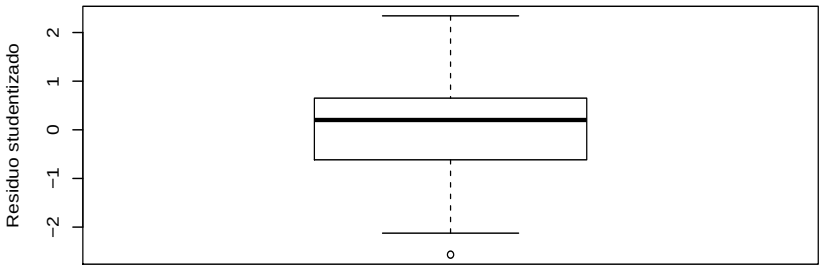
- Erros $\xi_{ij} \overset{i.i.d}{\sim} N(0, \sigma^2)$, μ, α_i não aleatório.
- $\mathcal{E}_{\xi_{ij}}(Y_{ij}) = \mu_i, \mathcal{V}_{\xi_{ij}}(Y_{ij}) = \sigma^2$.
- $\mu + \alpha_i$: média populacional relacionada ao i-ésimo fator,

  $\alpha_1 = 0, \alpha_2 = \alpha_3$ e $\alpha_4 = \alpha_5$.
- $Y_{ij} \overset{ind.}{\sim} N(0, \sigma^2)$.

Análise de resíduos







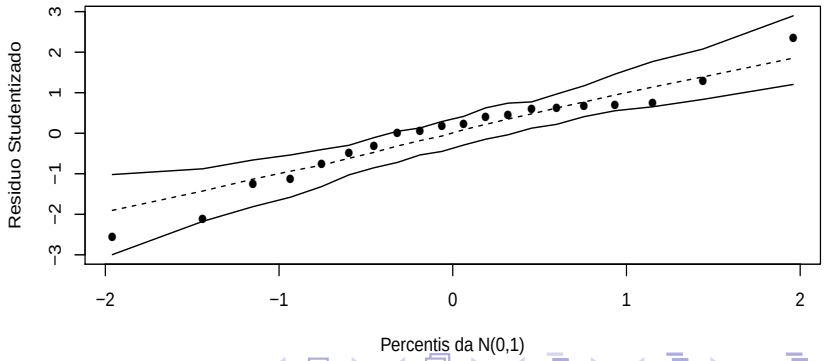


Tabela ANOVA do modelo reduzido

FV	SQ	GL	QM	Estatística F	pvalor
Nível de fósforo	63,42	2	31,230	22,93	$< 0{,}0001$
Resíduo	23,51	17	$< 1{,}38$		
Total	86,93	19			

Rejeita-se H_0.

Estimativas dos parâmetros do modelo

Parâmetro	Estimativa	EP	IC(95%)	Estat. t	pvalor
μ (0)	4,65	0,59	[3,39;5,90]	7,90	$<$ 0,0001
α_{23} (25/50)	3,63	0,71	[2,22;5,04]	5,04	0,0001
α_{45} (75/100)	4,84	0,72	[3,44;6,26]	6,73	$<$ 0,0001

Estimativas finais das médias

Nível de fósforo	Estimativa	EP	IC(95%)
(0)	4,65	0,59	[3,49;5,80]
(25/50)	8,28	0,42	[7,46;9,09]
(75/100)	9,50	0,42	[8,68;10,31]

- Há um aumento significativo ao se adicionar o mínimo de fósforo considerado em relação à não usá-lo.
- Aumento significativo a cada acréscimo de 50 kg/ha de fósforo.
- Aumento na produtividade apresenta um limite superior (possível decaimento?).

Utilização de modelos de regressão

- Relacionar o fator com a resposta levando em consideração a natureza quantitativa do fato.
- Modelo de regressão (linear e não-linear).

Modelo linear 1: reta

$$Y_i = \beta_0 + \beta_1 x_i + \xi_i, i = 1, 2, ..., 20$$

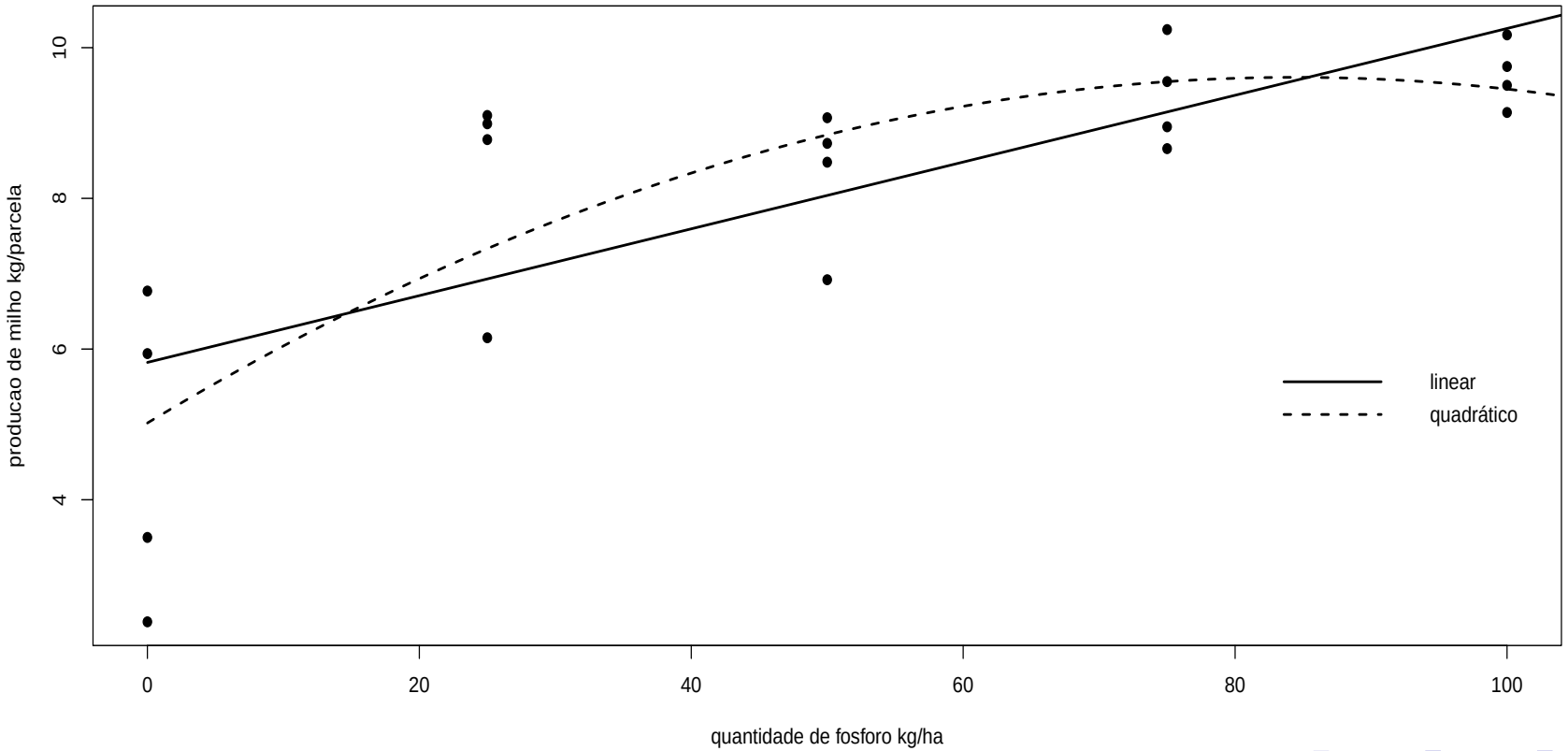
- x_i : quantidade de fósforo ministrada a i-ésima parcela.
- β_0 : valor esperado (média) da produção de milho quando a quantidade de fósforo aplicada é igual à 0.
- β_1 : incremento no valor esperado da produção de milho quando a quantidade de fósforo aplicada aumenta em uma unidade.
- $\xi_i \overset{iid}{\sim} N(0, \sigma^2)$.

Modelo linear 2: parábola

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \xi_i, i = 1, 2, ..., 20$$

- x_i : quantidade de fósforo ministrada a i-ésima parcela.
- β_0 : valor esperado (média) da produção de milho quando a quantidade de fósforo aplicada é igual à 0.
- A interpretação isolada dos parâmetros β_1 e β_2 é complicada mas, podemos dizer que $\frac{-\beta_1}{2\beta_2}$ é o máximo (ou mínimo) do valor esperado da produção de milho.
- $\xi_i \overset{iid}{\sim} N(0, \sigma^2)$.

Ajuste dos modelos de regressão



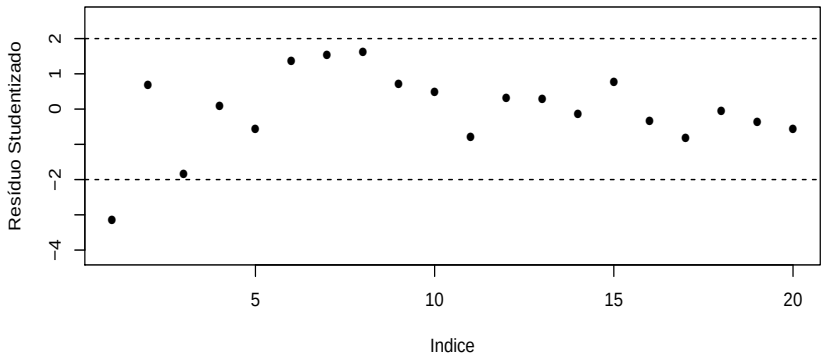
Estimativa da quantidade de fósforo que retorna a máxima produtividade

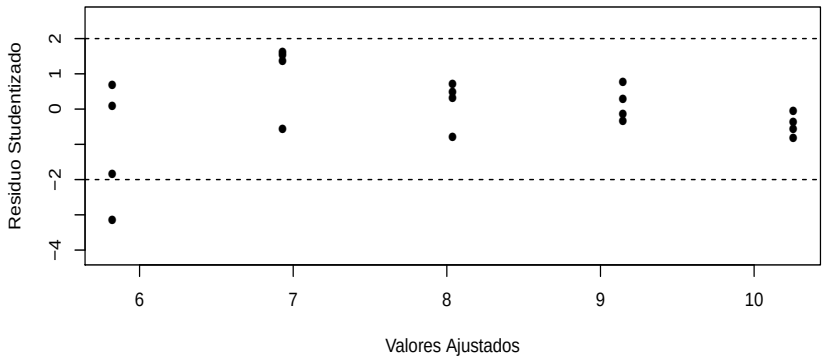
- $\theta = -\frac{\beta_1}{2\beta_2}$.
- Estimador pontual $\widehat{\theta} = -\frac{\widehat{\beta}_1}{2\widehat{\beta}_2}$.
- Método Delta: sob as suposições, temos que $\widehat{\theta} \approx N(\theta, \sigma^2_\theta)$, em que

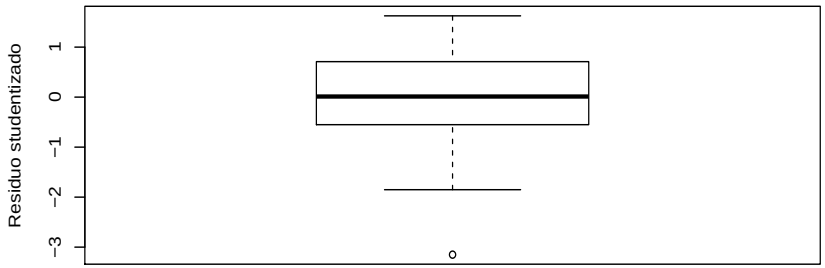
$$\sigma^2_\theta = \sigma^2 (\mathbf{\Delta})' (X'X)^{-1} (\mathbf{\Delta})$$

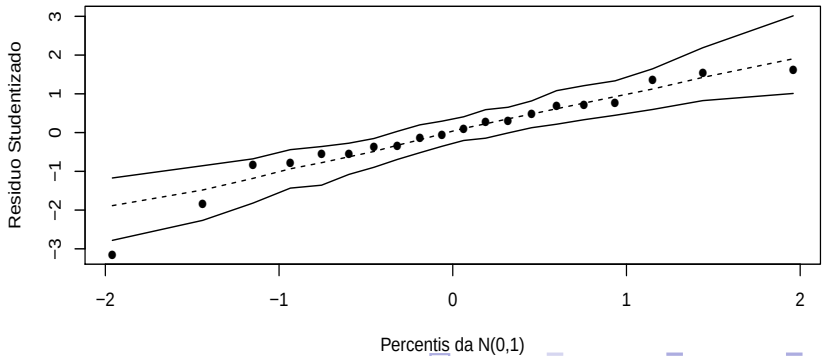
- $\mathbf{\Delta} = \begin{bmatrix} 0 & -\frac{1}{2\beta_2} & \frac{\beta_1}{2\beta_2^2} \end{bmatrix}'$.
- Em nosso exemplo $\widetilde{\theta} = 84,45$, $IC(95\%; \theta) = [52,68; 116,2]$

Análise de resíduos: modelo linear (reta)









Análise de resíduos: modelo linear (parábola)