Análise de dados e testes do tipo $\mathbf{C}\beta$ via modelos lineares

Prof. Caio Azevedo

Exemplo 1: considerando as etiologias cardíacas

$$Y_{ij} = \beta_{0i} + \beta_{1i} x_{ij} + \xi_{ij}, i = 1, ..., ; j = 1, ..., n_i$$

- Etiologias = CH ($i = 1$), ID ($i = 2$), IS ($i = 3$), C: ($i = 4$).
- $\xi_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
- $(\beta_{01}, \beta_{02}, \beta_{03}, \beta_{04}, \beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}, \sigma^2)$ parâmetros desconhecidos.
- x_{ij}: carga à que o paciente j que apresenta a etiologia cardíaca i foi submetido (conhecido e não aleatório).
- Parte sistemática: $\mathcal{E}(Y_{ij}) = \beta_{0i} + \beta_{1i} x_{ij}$.
- Parte aleatória: ξ_{ij}.
- O modelo acima implica que $Y_{ij} \overset{ind.}{\sim} N(\beta_{0i} + \beta_{1i} x_{ij}, \sigma^2)$.

Análise no R

- Ao ajustarmos o modelo anterior no R, ele fornece a seguinte "Tabela ANOVA":

FV	GL	SQ	QM	Estatística F	p-valor
(1)?	4	13749,95	3437,49	1015,73	<0,0001
(2)?	4	473,30	118,33	34,96	<0,0001
Resíduos	116	392,57	3,38		

- Que hipóteses estão sendo testadas em cada linha da tabela acima?
- (1)?: $H_0 : \beta_{01} = \beta_{02} = \beta_{03} = \beta_{04} = 0$ vs H_1 : há pelo menos uma diferença?
- (2)?: $H_0 : \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = 0$ vs H_1 : há pelo menos uma diferença?

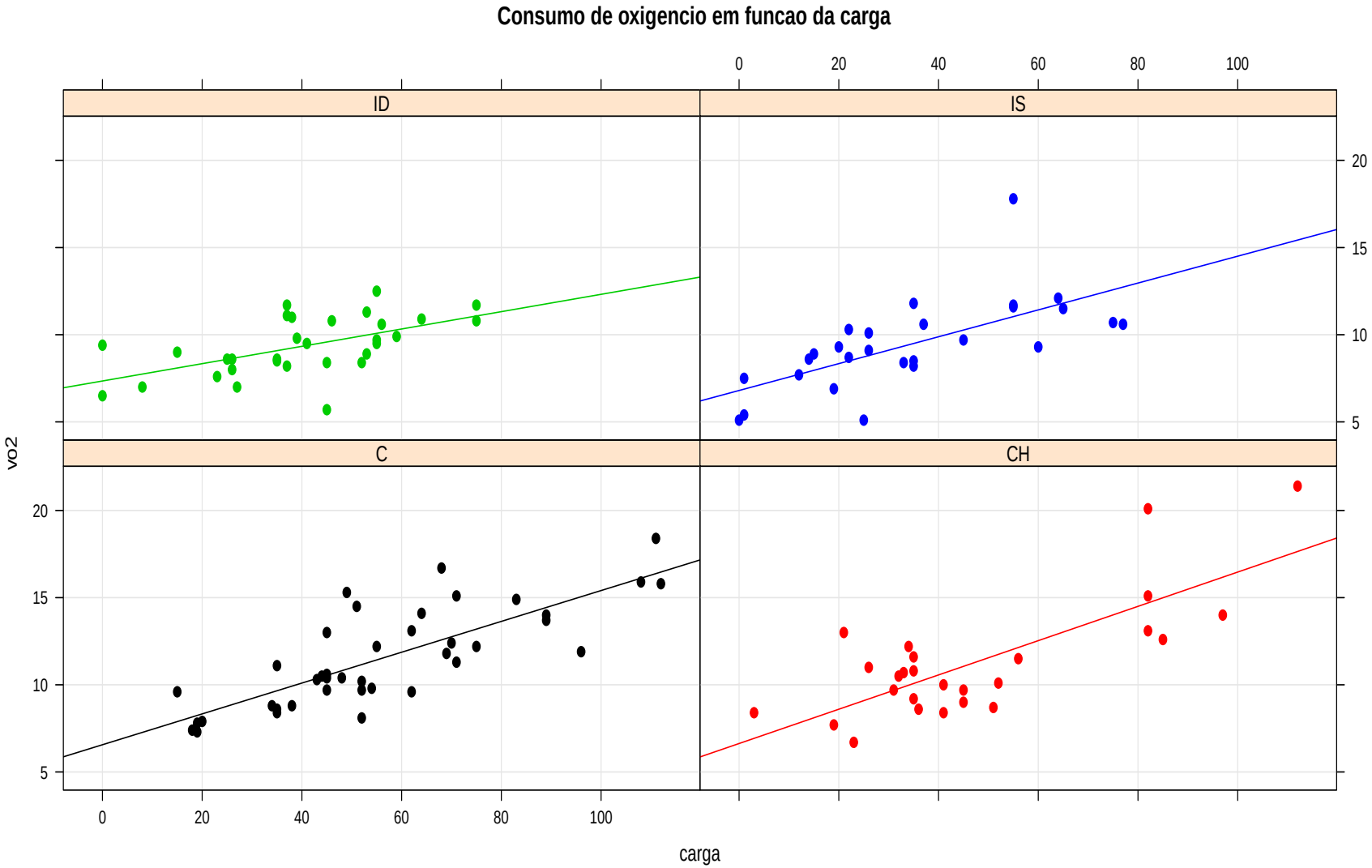
Análise no R

- Para responder à estas perguntas, precisamos saber como as somas de quadrados foram calculadas (matricialmente, de preferência) e estudar suas propriedades.
- Sugestões:
  - Note que $SQ(1) + SQ(2) = SQT - SQR = \mathbf{Y}' \left(\mathbf{H} - n^{-1}\mathbf{J} \right) \mathbf{Y}$.
  - Utilizar o mesmo raciocínio considerado em modelos ANOVA?
- Exercício: obter as expressões matriciais das somas de quadrados em questão.

Estimativas dos parâmetros

Parâmetro	Estimativa	EP	Estat. t	IC(95%)	p-valor
$\beta_{01}(C)$	6,56	0,71	9,18	[5,16 ; 7,96]	<0,0001
$\beta_{02}(CH)$	6,63	0,75	8,88	[5,17 ; 8,10]	<0,0001
$\beta_{03}(ID)$	7,35	0,78	9,45	[5,82 ; 8,87]	<0,0001
$\beta_{04}(IS)$	6,80	0,66	10,33	[5,51 ; 8,09]	<0,0001
$\beta_{11}(C)$	0,09	0,01	7,62	[0,07 ; 0,11]	<0,0001
$\beta_{12}(CH)$	0,10	0,01	7,14	[0,07 ; 0,13]	<0,0001
$\beta_{13}(ID)$	0,05	0,02	2,82	[0,02 ; 0,08]	0,0056
$\beta_{14}(IS)$	0,08	0,02	4,78	[0,05 ; 0,11]	<0,0001

O consumo de oxigênio dos pacientes para carga 0 parecem ser semelhantes entre os grupos. O aumento no consumo parecer ser menor que os demais, para pacientes idiopáticos e igual para os outros três tipos.


Consumo de oxigencio em funcao da carga
ID
IS
C
CH
vo2
carga
0
20
40
60
80
100
5
10
15
20

- Temos interesse em saber os consumos de oxigênio, para pacientes submetidos à uma carga nula, são os mesmos entre os grupos. Ou seja, desejamos testar se:

$$H_0 : \beta_{01} = \beta_{02} = \beta_{03} = \beta_{04} \; vs \; H_1 : \text{há pelo menos uma diferença} \quad (1)$$

- Temos interesse em saber se os aumentos no consumo de oxigênio, são todos nulos entre os grupos. Ou seja, desejamos testar:

$$H_0 : \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = 0 \; vs \; H_1 : \text{há pelo menos uma diferença} \quad (2)$$

- Em sendo não nulos, temos interesse em saber se os aumentos no consumo de oxigênio, são os mesmos entre os grupos. Ou seja, desejamos testar:

$$H_0 : \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} \; vs \; H_1 : \text{há pelo menos uma diferença} \quad (3)$$

- Ao se detectar a existência de pelo menos uma diferença (rejeitar H_0), devemos identificar os padrões dela (comparações dois a dois, por exemplo, sempre procedendo-se com cautela).
- Em geral, a grande maioria das hipóteses de interesse, podem ser descritas como:

$$H_0 : \mathbf{C}_{(q\times p)}\boldsymbol{\beta}_{(p\times 1)} = \mathbf{0} \; vs \; H_1 : \mathbf{C}_{(q\times p)}\boldsymbol{\beta}_{(p\times 1)} \neq \mathbf{0} \tag{4}$$

em que, via de regra, $q \leq p$.

- Como podemos testar as hipóteses acima?

- Lembremos que $\boldsymbol{\beta} = (\beta_{01}, \beta_{02}, \beta_{03}, \beta_{04}, \beta_{11}, \beta_{21}, \beta_{31}, \beta_{41})'$
- A hipótese (nula) (1), pode ser escrita como:

$$H_0 : \begin{cases} \beta_{01} - \beta_{02} = 0 \\ \beta_{01} - \beta_{03} = 0 \\ \beta_{01} - \beta_{04} = 0 \end{cases} \Leftrightarrow H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0},$$

em que

$$\mathbf{C} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

- A hipótese (nula) (2), pode ser escrita como:

$$H_0 : \begin{cases} \beta_{11} = 0 \\ \beta_{12} = 0 \\ \beta_{13} = 0 \\ \beta_{14} = 0 \end{cases} \Leftrightarrow H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0},$$

em que

$$\mathbf{C} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{(4\times 4)} & \mathbf{I}_4 \end{bmatrix}$$

- A hipótese (nula) (3), pode ser escrita como:

$$H_0 : \begin{cases} \beta_{11} - \beta_{12} = 0 \\ \beta_{11} - \beta_{13} = 0 \\ \beta_{11} - \beta_{14} = 0 \end{cases} \Leftrightarrow H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0},$$

em que

$$\mathbf{C} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \end{bmatrix}$$

Construção da Estatística do Teste

- Sabemos que:

$$\widehat{\boldsymbol{\theta}} = \mathbf{C}\widehat{\boldsymbol{\beta}} \sim N_q(\mathbf{C}\boldsymbol{\beta}, \sigma^2\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}').$$

- Como $\widehat{\boldsymbol{\beta}} \perp \widehat{\sigma}^2$, então $\mathbf{C}\widehat{\boldsymbol{\beta}} \perp \widehat{\sigma}^2$, em que

$$\widehat{\sigma}^2 = \frac{1}{n-p}(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}) = \frac{1}{n-p}\mathbf{Y}'\left(\mathbf{I}-\mathbf{H}\right)\mathbf{Y} = \frac{SQR}{n-p} = QMR$$

- Portanto, sob $H_0(\mathbf{C}\boldsymbol{\beta} = \mathbf{0})$ e usando alguns resultados de distribuições de formas quadráticas (provar), temos que

$$Q^* = \frac{1}{\sigma^2}\left(\mathbf{C}\widehat{\boldsymbol{\beta}}\right)'\left(\mathbf{C}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{C}'\right)^{-1}\left(\mathbf{C}\widehat{\boldsymbol{\beta}}\right) \sim \chi^2_{(q)}$$

Cont.

- Além disso, sabemos que $(n-p)\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{(n-p)}$.
- Portanto, pelos resultados acima, temos, sob H_0, que:

$$F = \frac{Q^*/q}{\widehat{\sigma}^2/\sigma^2} = \frac{1}{q\widehat{\sigma}^2}\left(\mathbf{C}\widehat{\boldsymbol{\beta}}\right)'\left(\mathbf{C}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{C}'\right)^{-1}\left(\mathbf{C}\widehat{\boldsymbol{\beta}}\right) \sim F_{(q,n-p)}$$

- $p - valor = P(F > f|H_0)$, em que f é o valor calculado da estatística definida acima, e $F \sim F_{(q,n-p)}$.
- Sob H_1, $F \sim F_{\left[q,n-p,\delta=\frac{1}{\sigma^2}\left((\mathbf{C}\boldsymbol{\beta})'\left(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'\right)^{-1}(\mathbf{C}\boldsymbol{\beta})\right)\right]}$.

Voltando ao exemplo

- Para o teste de nulidade simultânea de todos os interceptos, temos (estatística (p-valor)): 89,95 ($< 0{,}0001$) $\neq$ 1015,73 (ANOVA).
- Para o teste de nulidade simultânea de todos os incrementos, temos (estatística (p-valor)): 34,96 ($< 0{,}001$) $=$ 34,96 (ANOVA).
- Para o teste de igualdade simultânea de todos os interceptos, temos (estatística (p-valor)): 0,22 (0,8842).
- Para o teste de igualdade simultânea de todos os incrementos, temos (estatística (p-valor)): 1,72 (0,1666).

- À rigor, após (ou mesmo antes) de ajustar o modelo, devemos verificar se as hipóteses se verificam (homocedasticidade, ausência de correlação e normalidade dos erros). Faremos isso mais adiante.
- Devemos ajustar um modelo reduzido que contemple apenas uma intercepto e um incremento (comuns à todos os grupos).

Exemplo 1: modelo reduzido

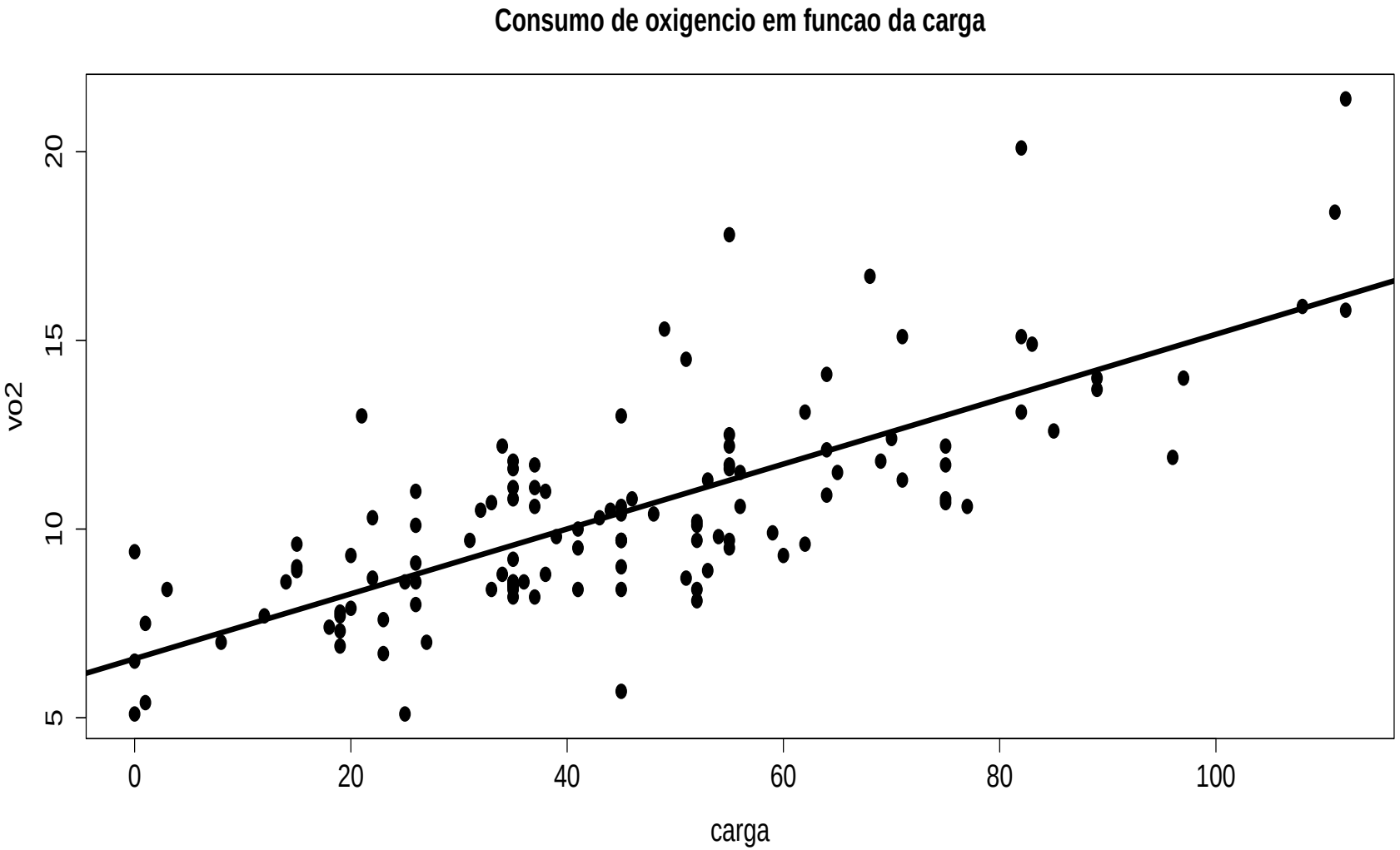
$$Y_i = \beta_0 + \beta_1 x_i + \xi_i, i = 1, ..., 124$$

- $\xi_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
- $(\beta_0, \beta_1, \sigma^2)$ parâmetros desconhecidos.
- x_i: carga à que o paciente i foi submetido (conhecido e não aleatório).
- Parte sistemática: $\mathcal{E}(Y_i) = \beta_0 + \beta_1 x_i$.
- Parte aleatória: ξ_i.
- O modelo acima implica que $Y_i \overset{ind.}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$

Exemplo 1: modelo 1

Parâmetro	Estimativa	EP	Estat. t	IC(95%)	p-valor
β_0	6,56	0,36	18,43	[5,87 ; 7,26]	<0,0001
β_1	0,09	0,01	12,52	[0,07 ; 0,10]	<0,0001

Os dois parâmetros são diferentes de 0. A carga influencia positivamente o consumo de oxigênio. O consumo de oxigênio para pacientes submetidos à carga 0 tende a se apresentar entre 5,87 e 7,26. Também, para o modelo reduzido, devemos verificar se as suposições são satisfeitas.


Consumo de oxigencio em funcao da carga
vo2
carga
5
10
15
20
0
20
40
60
80
100

Exemplo 2: Modelo (casela de referência)

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

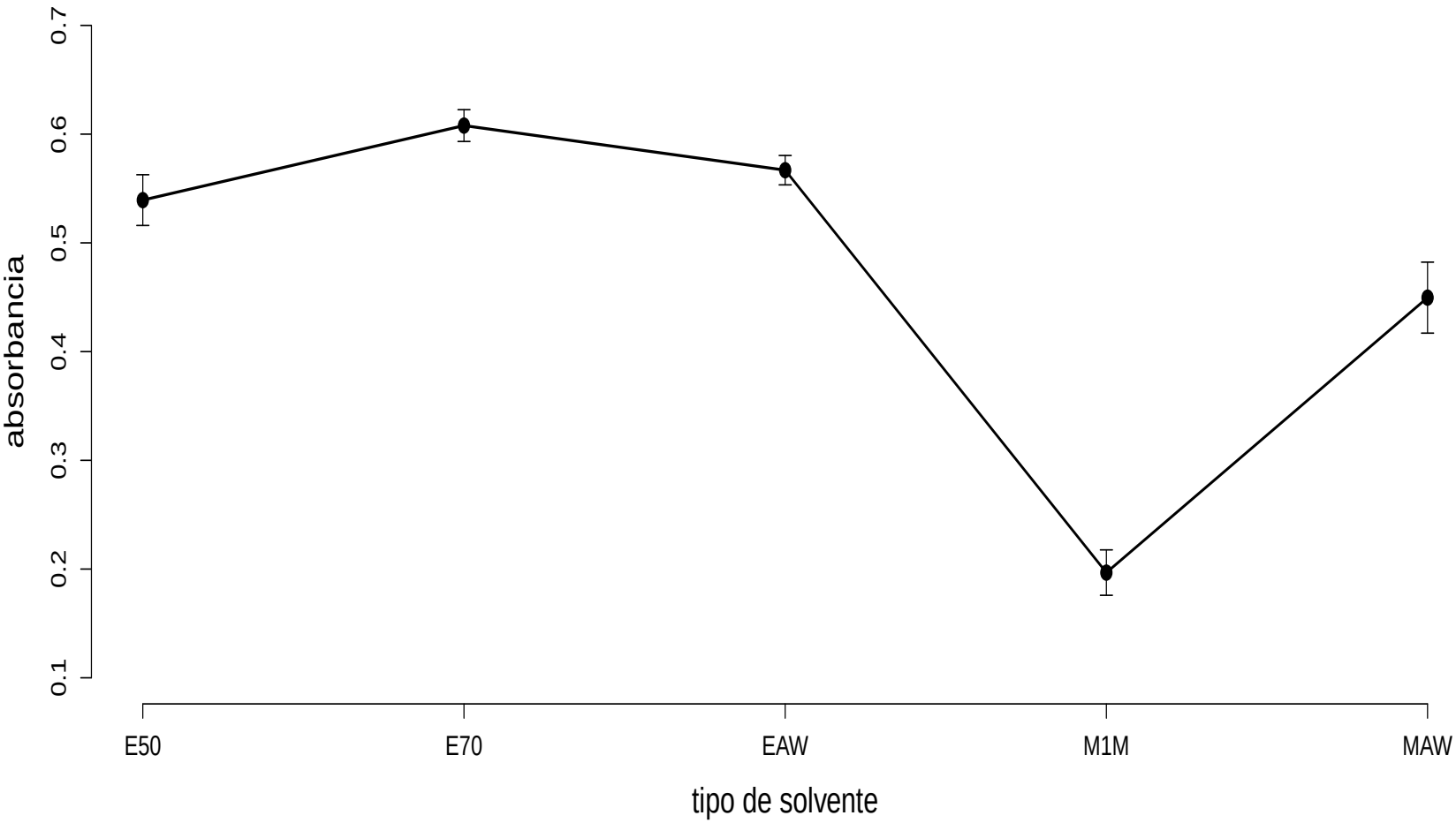
(grupos); $j = 1, ..., 5$(unidades experimentais)

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

Análise descritiva

Não há sentido em construir box-plots ou histogramas (poucas ob observações por grupo).

Solvente	**Medida descritiva**					
	Média	**DP**	**Var.**	**CV%**	**Mínimo**	**Máximo**
E50	0,539	0,026	0,0007	4,937	0,510	0,562
E70	0,608	0,017	0,0003	2,744	0,583	0,629
EAW	0,567	0,015	0,0002	2,717	0,544	0,586
M1M	0,197	0,024	0,0006	12,107	0,165	0,225
MAW	0,450	0,037	0,0014	8,283	0,409	0,501


absorbancia
0.1
0.2
0.3
0.4
0.5
0.6
0.7
E50
E70
EAW
M1M
MAW
tipo de solvente

Teste de Bartlett para igualdade de variâncias

- $H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$ vs $H_1 : \sigma_i^2 \neq \sigma_j^2$ para pelo menos um $i \neq j$
- Estatística do teste:

$$Q_B = \frac{q}{c},$$

em que

$$q = (n-k)\ln S_p^2 - \sum_{i=1}^{k}(n_i - 1)\ln S_i^2, S_p^2 = QMR = \frac{\sum_{i=1}^{k}(n_i-1)S_i^2}{n-k}$$

$$c = 1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k}(n_i-1)^{-1} - (n-k)^{-1}\right]$$

- Sob H_0, $Q_B \approx \chi^2_{(k-1)}$. Rejeita-se H_0 quando $P(Q_B > q_B|H_0) < \alpha$, q_B valor calculado e α e o nível de significância.

Teste de Levene para igualdade de variâncias

- $H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$ vs $H_1 : \sigma_i^2 \neq \sigma_j^2$ para pelo menos um $i \neq j$
- Estatística do teste:

$$Q_L = \frac{(n-k)\sum_{i=1}^{k} n_i \left(Z_{i.} - Z_{..}\right)^2}{(k-1)\sum_{i=1}^{k}\sum_{i=1}^{n_i} \left(Z_{ij} - Z_{i.}\right)^2},$$

em que

$$Z_{ij} = |Y_{ij} - \overline{Y}_{i.}|; Z_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} Z_{ij}; Z_{..} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Z_{ij}$$

- Sob H_0, $Q_L \approx F_{(k-1,n-k)}$. Rejeita-se H_0 quando

$P(Q_L > q_L|H_0) < \alpha$, q_L valor calculado e α e o nível de significância.

Testes para homocedasticidade

- Teste de Bartlett : 3,772 (0,4378).
- Teste de Levene : 0,696 (0,6033).
- Hipótese de homocedasticidade parace razoável.

Tabela ANOVA

FV	SQ	GL	QM	Estatística F	pvalor
Solvente	0,541	4	0,135	212,81	$<$ 0,0001
Resíduo	0,012	20	$<$ 0,001		
Total	0.553	24			

Rejeita-se H_0.

Estimativas dos parâmetros do modelo

Parâmetro	Estimativa	EP	IC(95%)	Estat. t	pvalor
μ (E50)	0,539	0,011	[0,517; 0,561]	47,826	$<$ 0,0001
α_2 (E70)	0,069	0,0160	[0,037 ; 0,010]	4,298	0,0003
α_3 (EAW)	0,028	0,0160	[-0,004 ; 0,059]	1,726	0,0998
α_4 (M1M)	-0,343	0,0160	[-0,374; -0,311]	-21,481	$<$ 0,0001
α_5 (MAW)	-0,090	0,0160	[-0,121 ;-0,058]	-5,624	$<$ 0,0001

Parâmetro α_3 não significativo. Isto sugere uma possível equivalência entre os solventes E50 e EAW.

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)
  - $H_{01}: \begin{cases} \mu_1 - \mu_2 = 0, \text{ e} \\ \mu_1 - \mu_3 = 0 \end{cases}$
  - $H_{02}: \mu_1 = \mu_2$.
  - $H_{03}: \mu_1 = \mu_3$.
  - $H_{04}: \frac{\mu_1+\mu_2+\mu_3}{3} = \frac{\mu_4+\mu_5}{2}$.
  - $\mathbf{H_{05}}: \mu_{\mathbf{3}} = \mu_{\mathbf{5}}$.

Continuação: em termos das parametrização CR

- Considere as hipóteses (H_0)
  - $H_{01}: \begin{cases} \alpha_2 = 0, \text{ e} \\ \alpha_3 = 0 \end{cases}$
  - $H_{02}: \alpha_2 = 0.$
  - $H_{03}: \alpha_3 = 0.$
  - $H_{04}: 2\alpha_2 + 2\alpha_3 - 3\alpha_4 - 3\alpha_5 = 0.$
  - $\mathbf{H_{05}}: \alpha_{\mathbf{3}} = \alpha_{\mathbf{5}}.$

Estatísticas (valores p)

- Resultados:
  - H_{01} : $9,35 (0,0014)$
  - H_{02} : $18,47 (< 0,0001)$.
  - H_{03} : $2,98 (0,0998)$.
  - H_{04} : $581,90 (< 0,0001)$.
  - H_{05} : $54,02 (< 0,0001)$.
- É importante tentar controlar o nível de significância global de todas as hipóteses testadas. Aconselha-se a utilizar $\alpha^* = \frac{\alpha}{k}$, em que α é o nível de significância adotado na tabela ANOVA (se for o caso) ou algum valor pré-fixado de interesse e k o número de hipóteses testadas.

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,

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

Estimativas dos parâmetros do modelo

Parâmetro	Estimativa	EP	IC(95%)	Estat. t	pvalor
μ (E50/EAW)	0,553	0,008	[0,537;0,569]	66,310	$<$ 0,0001
α_2 (E70)	0,055	0,0114	[0,026;0,083]	3,792	0,0011
α_4 (M1M)	-0,356	0,0114	[-0,385;-0,328]	-24,665	$<$ 0,0001
α_5 (MAW)	-0,103	0,0114	[-0,132;-0,075]	-7,161	$<$ 0,0001

Todos os incrementos α são significativos e todos parecem distintos entre si.

Estimativas finais das médias

Solvente	Estimativa	EP	IC(95%)
E50/EAW	0,553	0,008	[0,537;0,569]
E70	0,607	0,012	[0,584;0,631]
M1M	0,197	0,012	[0,173;0,220]
MAW	0,450	0,012	[0,426;0,472]

- Melhor solvente: E70.
- Pior solvente: M1M.
- Os soventes E50 e EAW são equivalentes.

Gráficos de perfis ajustados