A Validação do Modelo (análise dos resíduos)

TODA a inferência feita até aqui admitiu a validade do Modelo Linear, e em particular, dos pressupostos relativos aos erros aleatórios: Normais, de média zero, variância homogénea e independentes.

Uma análise de regressão não fica completa sem que haja uma validação dos pressupostos do modelo.

A validação dos pressupostos relativos aos erros aleatórios (que são desconhecidos) faz-se através dos seus preditores, os resíduos.

A análise de Resíduos e outros diagnósticos

Uma análise de regressão linear não fica completa sem o estudo dos resíduos e de alguns outros diagnósticos.

O modelo linear admite que $\varepsilon_i \frown \mathcal{N}(0, \sigma^2) \quad \forall i = 1, ..., n.$

Sob o modelo linear, os resíduos têm a seguinte distribuição:

$$E_i \frown \mathcal{N}\left(0, \sigma^2(1 - h_{ii})\right) \qquad \forall i = 1, ..., n\,,$$

sendo h_{ij} o i-ésimo elemento diagonal da matriz $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ de projecção ortogonal sobre o subespaço $\mathscr{C}(\mathbf{X})$.

Este resultado demonstra-se mais facilmente considerando o vector dos resíduos, $\vec{\mathbf{E}} = \vec{\mathbf{Y}} - \hat{\vec{\mathbf{Y}}} = \vec{\mathbf{Y}} - \mathbf{H}\vec{\mathbf{Y}} = (\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}$.

Propriedades dos Resíduos sob o modelo linear

Teorema (Distribuição dos Resíduos no Modelo Linear)

Dado o Modelo Linear, tem-se:

$$\vec{\mathbf{E}} \frown \mathcal{N}_n\left(\vec{\mathbf{0}}, \sigma^2(\mathbf{I}_n - \mathbf{H})\right) \qquad \text{sendo} \quad \vec{\mathbf{E}} = (\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}} \,.$$

Como no Modelo Linear $\vec{\mathbf{Y}} \frown \mathcal{N}(\mathbf{X}\vec{\boldsymbol{\beta}}, \sigma^2\mathbf{I}_n)$, o vector dos resíduos $\vec{\mathbf{E}} = (\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}$, tem distribuição Multinormal em sentido generalizado

O vector esperado de $\vec{\mathbf{E}}$ resulta das propriedades do acetato 125:

- $E[\vec{\mathbf{E}}] = E[(\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}] = (\mathbf{I}_n - \mathbf{H})E[\vec{\mathbf{Y}}] = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\vec{\boldsymbol{\beta}} = \vec{\mathbf{0}}$,
  pois $\mathbf{X}\vec{\boldsymbol{\beta}} \in \mathscr{C}(\mathbf{X})$, logo permanece invariante sob a projecção: $\mathbf{H}\mathbf{X}\vec{\boldsymbol{\beta}} = \mathbf{X}\vec{\boldsymbol{\beta}}$.
- Pelas propriedades do acetato 126 e o facto de $\mathbf{H}$ ser simétrica ($\mathbf{H}^t = \mathbf{H}$) e idempotente ($\mathbf{H}\mathbf{H} = \mathbf{H}$), tem-se:
  $V[\vec{\mathbf{E}}] = V[(\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}] = (\mathbf{I}_n - \mathbf{H})\,V[\vec{\mathbf{Y}}](\mathbf{I}_n - \mathbf{H})^t = \sigma^2 \cdot (\mathbf{I}_n - \mathbf{H})$.

Propriedades dos Resíduos no Modelo Linear (cont.)

Nota: Embora no modelo RL os erros aleatórios sejam independentes, os resíduos não são variáveis aleatórias independentes, pois as covariâncias entre resíduos diferentes são (em geral), não nulas:

$$cov(E_i, E_j) = -\sigma^2 \cdot h_{ij} \ , \quad \text{se } i \neq j \ ,$$

onde h_{ij} indica o elemento da linha i e coluna j da matriz $\mathbf{H}$.

Se $\vec{\mathbf{E}} \frown \mathcal{N}_n\left(\vec{\mathbf{0}}, \sigma^2(\mathbf{I}_n - \mathbf{H})\right)$, então cada resíduo tem distribuição:

$$E_i \frown \mathcal{N}\left(0 \ , \ \sigma^2(1 - h_{ii})\right) \ ,$$

onde h_{ii} é o i-ésimo elemento diagonal de $\mathbf{H}$ e

$$\frac{E_i}{\sqrt{\sigma^2(1 - h_{ii})}} \frown \mathcal{N}(0 \ , \ 1) \ .$$

Resíduos habituais : $E_i = Y_i - \hat{Y}_i$;

Resíduos (internamente) estandardizados : $R_i = \frac{E_i}{\sqrt{QMRE \cdot (1-h_{ii})}}$.

Resíduos Studentizados (ou externamente estandardizados):

$$T_i = \frac{E_i}{\sqrt{QMRE_{[-i]} \cdot (1-h_{ii})}}$$

sendo $QMRE_{[-i]}$ o valor de $QMRE$ resultante de um ajustamento da Regressão excluíndo a i-ésima observação (associada ao resíduo E_i).

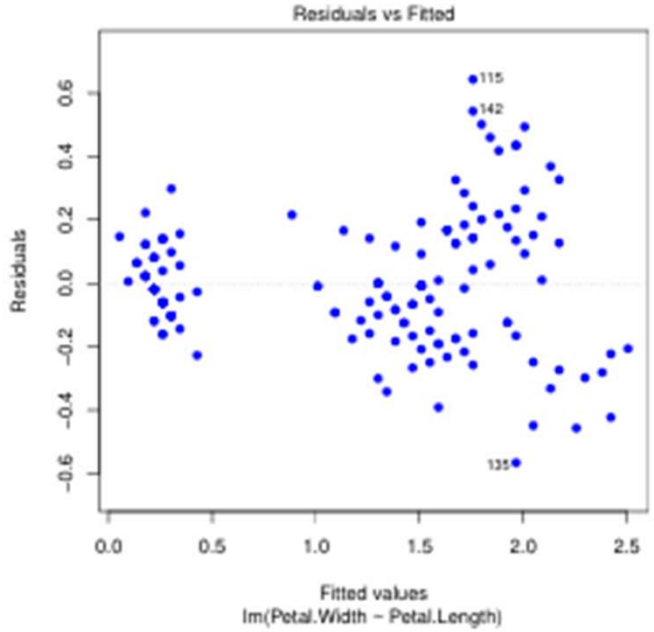
Como $\frac{E_i}{\sqrt{\sigma^2(1-h_{ii})}} \frown \mathcal{N}(0,1)$, definem-se resíduos normalizados:

Para grandes amostras, os R_i são aproximadamente $\mathcal{N}(0,1)$.

Nas regressões lineares, avalia-se a validade dos pressupostos do modelo através de gráficos de resíduos. Não se efectuam testes de Normalidade, já que os resíduos não são (em geral) independentes.

Validação do modelo: (1) Gráficos de resíduos vs. $\hat{Y}_i$

Gráfico indispensável: Resíduos (usuais) vs. Valores ajustados de Y.



- Os resíduos devem estar aproximadamente numa banda horizontal em torno de zero.
- Não deve existir qualquer padrão aparente. Sendo válido o Modelo RL, $cor(E_i, \hat{Y}_i) = 0$.

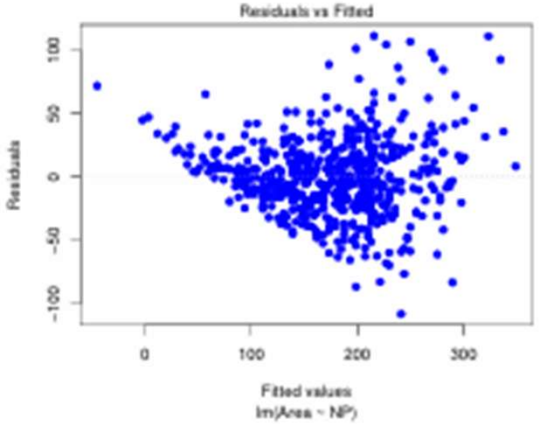
Possíveis padrões indicativos de problemas

Num gráfico de E_i vs. $\hat{Y}_i$ podem surgir padrões problemáticos:

Curvatura na disposição dos resíduos Indica violação da hipótese de linearidade entre *y* e os preditores.

Gráfico em forma de funil Indica violação da hipótese de homogeneidade de variâncias.

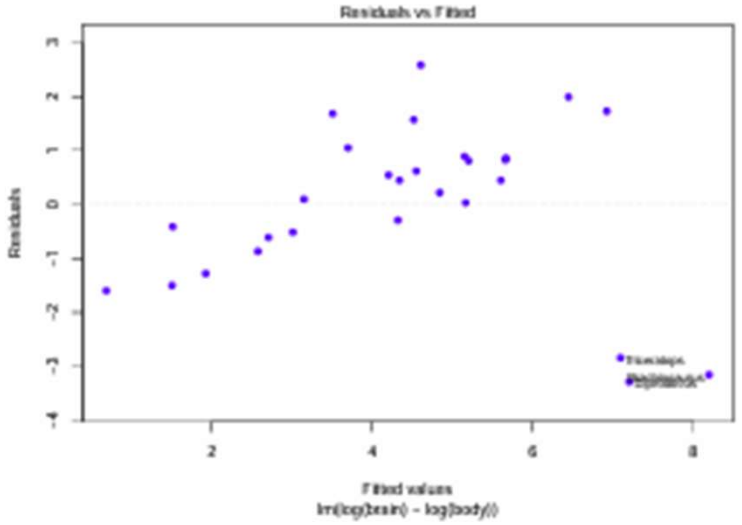
Um ou mais resíduos muito destacados Indica a existência de observações atípicas.



Um exemplo de resíduos em forma de funil, e sugerindo alguma curvatura na relação entre as duas variáveis

Padrões indicativos de problemas (cont.)

Um ou mais resíduos muito destacados e/ou banda oblíqua: Indica possíveis observações atípicas.



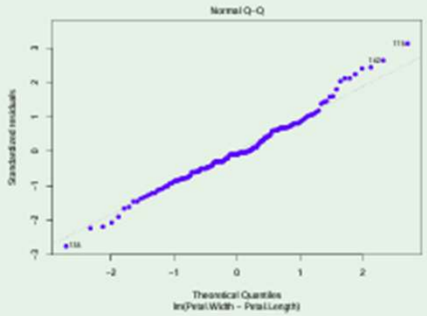
Validação do modelo: (2) Gráficos para avaliar a Normalidade

para grandes amostras os resíduos estandardizados $R_i = \frac{E_i}{\sqrt{QMRE \cdot (1-h_{ii})}}$, devem ser aproximadamente $\mathcal{N}(0,1)$

O pressuposto de erros aleatórios Normais pode ser validado com:

- um qq-plot que confronte os quantis empíricos dos n resíduos standardizados, com os quantis teóricos numa $\mathcal{N}(0,1)$.

Um qq-plot concordante com a hipótese de Normalidade dos erros aleatórios deverá apresentar colinearidade aproximada. O exemplo seguinte sugere algum desvio à Normalidade para os resíduos mais extremos.



Este *qq-plot* sugere algum desvio para os resíduos mais extremos, mas não em quantidade ou de forma suficientemente severa para pôr em dúvida o pressuposto da Normalidade dos erros aleatórios.

Validação do modelo: (3) Gráficos para avaliar independência

Dependência entre erros aleatórios pode surgir como resultado de:

- correlação cronológica;
- correlação espacial.

Nesse caso, pode ser útil inspeccionar gráficos de resíduos vs. ordem de observação ou distribuição espacial dos resíduos, para verificar se existem padrões que sugiram falta de independência. Nesse caso, modelos alternativos para series temporais ou dados espaciais podem ser necessários.

Validação do modelo: (4) Gráficos de resíduos vs. preditores

A presença de não-linearidade em gráficos de resíduos vs. preditores individuais pode sugerir a necessidade de transformações desses preditores.

Observações atípicas

Outras ferramentas de diagnóstico visam identificar observações individuais que merecem ulterior análise.

Observações atípicas (*outliers* in English). Conceito sem definição rigorosa, procura designar observações que se distanciam da relação linear de fundo entre Y e as variáveis preditoras.

Muitas vezes surgem associadas a resíduos grandes (em módulo). Em particular, e como os resíduos Studentizados têm distribuição aproximadamente $\mathcal{N}(0,1)$ para n grande, observações para as quais $|R_i| > 3$ (ou $|T_i| > 3$) podem ser classificadas como atípicas.

Mas por vezes, observações distantes da tendência geral podem afectar o próprio ajustamento do modelo, e não serem facilmente identificáveis a partir dos seus resíduos.

As chamadas "observações alavanca"

Define-se o valor do efeito alavanca (*leverage*) da i-ésima observação como sendo o i-ésimo valor diagonal da matriz $\mathbf{H}$: $h_{ii} = \mathbf{H}_{(i,i)}$.

Como $\vec{\hat{\mathbf{Y}}} = \mathbf{H}\vec{\mathbf{Y}}$, tem-se $\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$ (cada valor ajustado é combinação linear dos valores observados). O efeito alavanca h_{ii} é a ponderação associada a y_i na definição do valor ajustado correspondente, $\hat{y}_i$. Não deveria ser excessivo.

Observações alavanca (*leverage points*) são observações com h_{ii} elevado, que tendem a "atrair" a hipersuperfície ajustada numa regressão.

Como $V[E_i] = \sigma^2 (1 - h_{ii})$, se h_{ii} é elevado, a variância do resíduo E_i é baixa e o resíduo tende a estar perto da sua média (zero). Ou seja, a superfície ajustada tende a passar próximo desse ponto.

Observações alavanca (cont.)

Verifica-se, para qualquer observação:

$$\frac{1}{n} \leq h_{ii} \leq 1 \, .$$

O valor médio das observações alavanca numa regressão linear é a razão entre o no. de parâmetros e o no. de observações:

$$\overline{h} = \frac{p+1}{n} \, ,$$

Logo, quanto mais observações, menor o efeito alavanca médio.

Observações alavanca (cont.)

Observações com um efeito alavanca elevado podem, ou não, estar dispostas com a mesma tendência de fundo que as restantes observações, i.e., podem, ou não, ser atípicas (outliers).

Efeito alavanca numa regressão linear Simples

Numa regressão linear simples, tem-se

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{(n-1) \cdot s_x^2} \ .$$

Assim, numa RLS, o efeito alavanca da observação i depende do valor x_i em relação à média $\overline{x}$: quanto maior $(x_i - \overline{x})^2$, maior h_{ii}. O maior efeito alavanca tem de pertencer a uma das duas observações mais extrema em x.

Numa regressão linear múltipla, os maiores efeitos alavanca correspondem às observações em que os valores dos preditores estão mais afastados do vector das médias dos preditores.

Observações influentes

Observações influentes são observações que, se retiradas da análise, gerariam variações assinaláveis no conjunto dos valores ajustados de Y e nos parâmetros ajustados, b_j.

Medida de influência frequente é a distância de Cook, definida como:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{[-i]_j})^2}{(p+1) \cdot QMRE},$$

sendo $\hat{y}_{[-i]_j}$ o valor ajustado da observação i, obtido estimando os β_js sem a observação i. Expressão equivalente é:

$$D_i = R_i^2 \cdot \left(\frac{h_{ii}}{1 - h_{ii}} \right) \cdot \frac{1}{p+1}$$

Quanto maior D_i, maior é a influência da i-ésima observação.

É frequente considerar $D_i > 0.5$ como limiar de observação influente.

Uma prevenção

Observações atípicas, influentes ou alavanca não são o mesmo conceito, embora possam estar relacionados.

$$D_i = R_i^2 \cdot \left(\frac{h_{ii}}{1 - h_{ii}} \right) \cdot \frac{1}{p+1}$$

R_i^2 grande e h_{ii} grande $\Rightarrow$ D_i grande (observação influente)

R_i^2 pequeno e h_{ii} pequeno $\Rightarrow$ D_i pequeno (observação não influente)

R_i^2 grande e h_{ii} pequeno (ou viceversa) – D_i pode, ou não, ser grande

(Se obs. i é, ou não, influente depende da grandeza relativa de R_i^2 e h_{ii})

Estes diagnósticos servem sobretudo para identificar observações que merecem maior atenção e consideração.

ESTUDOS DOS RESÍDUOS

```
proc reg data=porcos plots=diagnostics;
model Musculo = Area Gordurasubcut Peso Rendimento Gordurarenalpel Comprimento
LarguraAnca /clb covb xpx R CLI CLM R P ;
output out=diagnostics r=r student=int_r rstudent=ext_r h=leverage cookd=cooksd
p=predicted ;
RUN;
proc print data=diagnostics;
run;
```

Resíduos usuais, Ei

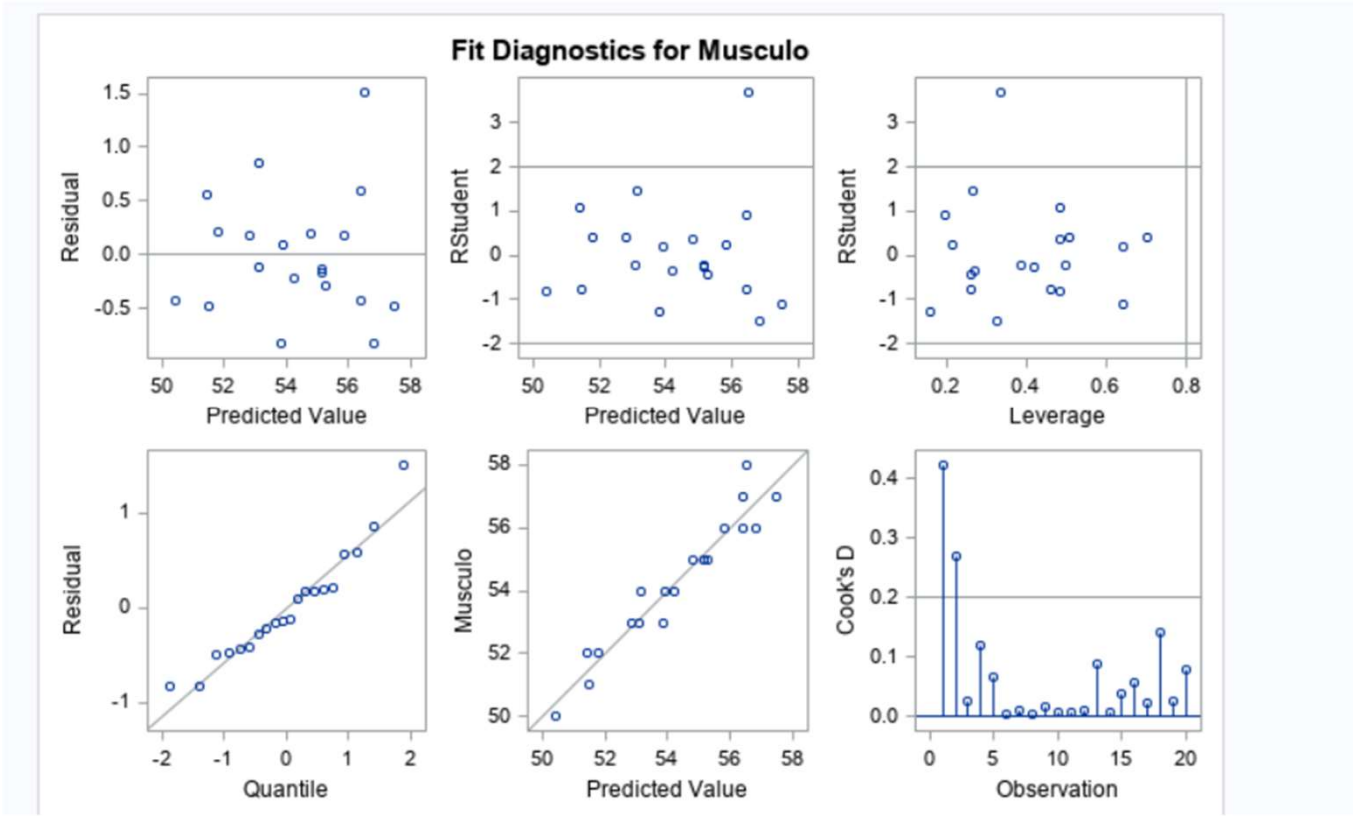
Resíduos standardizados, Ri

Resíduos studentizados, Ti

Efeitos alavanca, hii

Distâncias de Cook, Di

Sum of Residuals	0
Sum of Squared Residuals	6.30036
Predicted Residual SS (PRESS)	10.26395



Exercícios

a) Estude os gráficos de resíduos e outros diagnósticos

b) Os resultados do ajustamento do modelo com todos os preditores apresentam-se seguidamente:

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	54.11487	9.27061	5.84	<.0001	33.91594	74.31380
Area	1	0.06200	0.70162	0.09	0.9310	-1.46670	1.59070
Gordurasubcut	1	-0.93861	0.36030	-2.61	0.0230	-1.72363	-0.15359
Peso	1	0.24489	0.26196	0.93	0.3683	-0.32587	0.81565
Rendimento	1	0.00623	0.08323	0.07	0.9416	-0.17511	0.18756
Gordurarenalpel	1	-0.01436	0.00714	-2.01	0.0673	-0.02991	0.00119
Comprimento	1	0.01774	0.04832	0.37	0.7199	-0.08755	0.12302
LarguraAnca	1	0.11974	0.06255	1.91	0.0797	-0.01654	0.25602

The REG Procedure
Model: MODEL1
Dependent Variable: Musculo

Number of Observations Read	20
Number of Observations Used	20

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	77.98252	11.14036	21.50	<.0001
Error	12	6.21748	0.51812		
Corrected Total	19	84.20000			

Root MSE	0.71981	R-Square	0.9262
Dependent Mean	54.30000	Adj R-Sq	0.8831
Coeff Var	1.32561		

Output Statistics											
								Resíduos usuais	Resíduos standardizados		Distâncias de Cook, Di
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual	Std Error Residual	Student Residual	Cook's D
1	58	56.4895	0.4179	55.5789	57.4001	54.6760	58.3030	1.5105	0.586	2.577	0.422
2	57	57.4744	0.5758	56.2198	58.7290	55.4660	59.4828	-0.4744	0.432	-1.098	0.268
3	57	56.4076	0.3174	55.7161	57.0992	54.6936	58.1217	0.5924	0.646	0.917	0.025
4	56	56.8284	0.4102	55.9347	57.7221	55.0233	58.6335	-0.8284	0.592	-1.401	0.118
5	56	56.4186	0.4894	55.3524	57.4849	54.5222	58.3151	-0.4186	0.528	-0.793	0.068
6	56	55.8263	0.3354	55.0955	56.5572	54.0961	57.5566	0.1737	0.637	0.273	0.003
7	55	55.2852	0.3695	54.4800	56.0903	53.5222	57.0481	-0.2852	0.618	-0.462	0.010
8	55	55.1387	0.4487	54.1611	56.1163	53.2906	56.9867	-0.1387	0.563	-0.246	0.005
9	55	54.8051	0.5010	53.7134	55.8967	52.8942	56.7159	0.1949	0.517	0.377	0.017
10	55	55.1592	0.4663	54.1431	56.1752	53.2905	57.0278	-0.1592	0.548	-0.290	0.008
11	54	54.2200	0.3731	53.4070	55.0329	52.4535	55.9864	-0.2200	0.616	-0.357	0.006
12	54	53.9125	0.5763	52.6568	55.1681	51.9034	55.9215	0.0875	0.431	0.203	0.009
13	54	53.1392	0.3714	52.3301	53.9484	51.3745	54.9040	0.8608	0.617	1.396	0.088
14	53	53.1109	0.5091	52.0015	54.2202	51.1899	55.0319	-0.1109	0.509	-0.218	0.006
15	53	53.8286	0.2863	53.2048	54.4524	52.1407	55.5164	-0.8286	0.660	-1.255	0.037
16	53	52.8268	0.6023	51.5145	54.1391	50.7818	54.8718	0.1732	0.394	0.439	0.056
17	52	51.7865	0.5118	50.6715	52.9015	49.8622	53.7108	0.2135	0.506	0.422	0.023
18	52	51.4349	0.5011	50.3432	52.5266	49.5240	53.3458	0.5651	0.517	1.093	0.140
19	51	51.4835	0.3672	50.6833	52.2836	49.7228	53.2441	-0.4835	0.619	-0.781	0.027
20	50	50.4242	0.5004	49.3340	51.5144	48.5142	52.3342	-0.4242	0.517	-0.820	0.079

Exercícios

bi) Mostre que o valor do resíduo usual associado à primeira observação é $e_1 = 1.5105$

bii) Mostre que o valor do resíduo standardizado associado à primeira observação é $R_1 = 2.577$

The SAS System

Obs	Area	Gordurasubcut	Peso	Rendimento	Gordurarenalpel	Comprimento	LarguraAnca	Musculo	predicted	Ei r	Ri int_r	Di cooksd	hii leverage	Ti ext_r
1	8.8	3.5	13.9	50	200	72	25	58	56.4895	1.51050	2.57743	0.42232	0.33713	3.69342
2	9.2	3.0	15.0	47	170	68	24	57	57.4744	-0.47437	-1.09829	0.26799	0.63994	-1.10874
3	8.6	3.5	13.4	48	180	73	23	57	56.4076	0.59236	0.91688	0.02536	0.19443	0.91031
4	8.7	4.0	14.2	48	150	74	25	56	56.8284	-0.82843	-1.40054	0.11790	0.32471	-1.46607
5	8.5	3.5	13.0	51	160	69	22	56	56.4186	-0.41865	-0.79311	0.06758	0.46223	-0.78007
6	8.2	4.0	14.8	49	190	70	21	56	55.8263	0.17368	0.27271	0.00258	0.21717	0.26191
7	8.0	4.5	12.8	46	210	71	27	55	55.2852	-0.28516	-0.46164	0.00953	0.26356	-0.44596
8	7.9	5.0	15.1	48	200	76	23	55	55.1387	-0.13867	-0.24637	0.00482	0.38853	-0.23648
9	7.6	4.5	13.6	47	190	65	20	55	54.8051	0.19494	0.37718	0.01671	0.48447	0.36328
10	7.5	5.0	14.1	50	210	66	28	55	55.1592	-0.15915	-0.29025	0.00762	0.41970	-0.27888
11	7.6	4.5	13.7	49	250	65	22	54	54.2200	-0.21996	-0.35732	0.00586	0.26867	-0.34395
12	7.4	4.0	12.2	48	280	74	21	54	53.9125	0.08754	0.20298	0.00920	0.64101	0.19468
13	7.3	6.0	14.7	49	230	68	20	54	53.1392	0.86076	1.39598	0.08837	0.26620	1.46038
14	7.0	5.5	13.1	52	280	62	26	53	53.1109	-0.11086	-0.21788	0.00594	0.50031	-0.20902
15	7.5	5.0	14.0	50	250	72	21	53	53.8286	-0.82858	-1.25462	0.03698	0.15820	-1.28869
16	6.8	6.0	12.9	43	300	75	29	53	52.8268	0.17321	0.43946	0.05638	0.70019	0.42418
17	6.5	6.5	14.2	53	310	71	23	52	51.7865	0.21348	0.42173	0.02272	0.50547	0.40680
18	6.8	7.0	12.8	45	260	65	22	52	51.4349	0.56507	1.09344	0.14049	0.48455	1.10329
19	7.0	6.5	13.5	47	290	68	20	51	51.4835	-0.48349	-0.78098	0.02683	0.26029	-0.76749
20	6.8	7.0	12.9	48	330	69	21	50	50.4242	-0.42420	-0.81980	0.07856	0.48323	-0.80785

Exercícios

biii) Mostre que o erro padrão do resíduo associado à primeira observação é 0.586

biv) Mostre que o valor da distância de Cook da primeira observação é $D_1 = 0.42232$

Algumas transformações de variáveis

Por vezes, é possível tornear violações às hipóteses de Normalidade dos erros aleatórios ou homogeneidade de variâncias através de transformações de variáveis. Por exemplo,

$$\begin{array}{lll} \text{Se } var(\varepsilon_i) \propto E[Y_i] & \text{então} & Y \longrightarrow \sqrt{Y} \\ \text{Se } var(\varepsilon_i) \propto (E[Y_i])^2 & \text{então} & Y \longrightarrow \ln Y \\ \text{Se } var(\varepsilon_i) \propto (E[Y_i])^4 & \text{então} & Y \longrightarrow 1/Y \end{array}$$

são propostas usuais para estabilizar as variâncias.

Os exemplos acima são casos particulares da família Box-Cox de transformações:

$$Y \longrightarrow \begin{cases} \frac{Y^\lambda - 1}{\lambda} & , \lambda \neq 0 \\ \ln(Y) & , \lambda = 0 \end{cases}$$

Prevenções sobre transformações

Mas a utilização de transformações de variáveis, sobretudo quando afecta a variável resposta, deve ser feita com cautela.

- Uma transformação de variáveis muda também a relação de base entre as variáveis originais;
- Uma transformação que "corrija" um problema (e.g., variâncias heterogéneas) pode gerar outro (e.g., não-normalidade);
- Existe o perigo de usar transformações que resolvam o problema duma amostra específica, mas não tenham qualquer generalidade.

Advertências finais

1. Podem surgir problemas associados à (quase) multicolinearidade das variáveis preditoras, ou seja, ao facto das colunas da matriz $\mathbf{X}$ serem (quase) linearmente dependentes:

- podem existir problemas numéricos no cálculo de $(\mathbf{X}^t\mathbf{X})^{-1}$, logo no ajustamento do modelo e na estimação dos parâmetros;
- podem existir variâncias muito grandes de alguns $\hat{\beta}_i$s, o que significa muita instabilidade na inferência.

Multicolinearidade reflecte redundância de informação nos preditores. É possível eliminá-la excluíndo da análise uma ou várias variáveis preditoras que sejam responsáveis pela (quase) dependência linear dos preditores.

Advertências finais (cont.)

2. Não se deve confundir a existência de uma relação linear entre preditores X_1, X_2, ..., X_p e uma variável resposta Y, com uma relação de causa e efeito.

Pode existir uma relação de causa e efeito. Mas pode também verificar-se:

- Uma relação de variação conjunta, mas não de tipo causal (como por exemplo, em muitos conjuntos de dados morfométricos). Por vezes, preditores e variável resposta são todos efeito de causas comuns subjacentes.
- Uma relação espúria, de coincidência numérica.

Uma relação causal só pode ser afirmada com base em teoria própria do fenómeno sob estudo, e não com base na relação linear estabelecida estatisticamente.

Conceitos fundamentais e principais tipos de delineamento experimental

Antes de escolher o delineamento experimental de uma experiência é importante saber claramente qual o objectivo do estudo (quais a(s) hipóteses(s) a testar)

Alguns termos importantes:

Factor: uma variável categórica (qualitativa);

Níveis do factor: as diferentes categorias do factor, ou seja, diferentes situações experimentais a serem testadas e onde se efectuam observações da variável resposta. **Conhecidos como Tratamentos** (diferentes variedades, raças, regimes alimentares, etc.).

Unidade experimental (indivíduo, parcela de terreno, vaso, placa de *petri*, etc.)**:** as n observações da variável resposta correspondem a n diferentes unidades experimentais.

A agricultura tem uma longa tradição no desenvolvimento de delineamentos experimentais para estabelecer ensaios de campo rigorosos. Muitos dos princípios importantes do delineamento experimental foram desenvolvidos por R.A. Fisher nas décadas de 1920 e 1930.

Princípios gerais a seguir:

Repetição

A repetição de observações independentes é essencial para estimar a variância dos erros aleatórios; permite a diminuição dos erros de amostragem e outros; a precisão de um ensaio aumenta com o número de repetições.

Casualização (aleatorização)

ou seja, aleatoriedade na escolha das unidades experimentais e na associação que lhes é feita de um dado nível do factor.

- Condição fundamental para a validade de todo o processo indutivo (para se poder trabalhar com a teoria de probabilidades) e para evitar enviesamentos (mesmo inconscientes).
- É essencial para validar a estimativa a variância dos erros aleatórios.
- Há um paralelismo rigoroso entre o tipo de casualização adoptado e o modelo de análise dos dados.

Controlo da heterogeneidade entre unidades experimentais

→ Diferentes tipos de casualização/diferentes tipos de delineamento experimental

Delineamento experimentais

Alguns exemplos

Delineamento experimental totalmente casualizado (CRD)

Exemplo:

Exemplo de uma disposição de 5 tratamentos (**A**,**B**,**C**,**D**,**E**) em 4 repetições, em condições homogéneas.
1 factor em estudo, com 5 níveis.

D	B	A	E
C	E	D	B
A	C	E	A
D	B	A	C
B	C	E	D

Cada quadrado representa a unidade experimental (um animal, uma planta, uma parcela de terreno com várias plantas, um vaso, uma placa de Petri, etc.).

Uma casualização possível (exemplo de um delineamento totalmente casualizado (CRD)

Nota: Repetições e pseudo-repetições

Convém distinguir entre repetições e pseudo-repetições. Por pseudo-repetições entende-se medições que são feitas na mesma unidade experimental. Por exemplo: pesos de um animal ao longo do tempo; pesos de vários frutos de uma planta; medições do comprimento das folhas de uma planta; observações em plantas diferentes no mesmo vaso, na mesma parcela de terreno; observações na mesma amostra no laboratório, etc.. Estes exemplos são repetições que não são independentes.

Mas a existência de pseudo-repetições é essencial para a redução da variabilidade entre observações independentes. De facto, substituindo cada grupo de pseudo-repetições por uma única observação média pode-se diminuir a variabilidade entre diferentes observações médias (que são independentes).

A	A	A	A
B	B	B	B
C	C	C	C
D	D	D	D
E	E	E	E

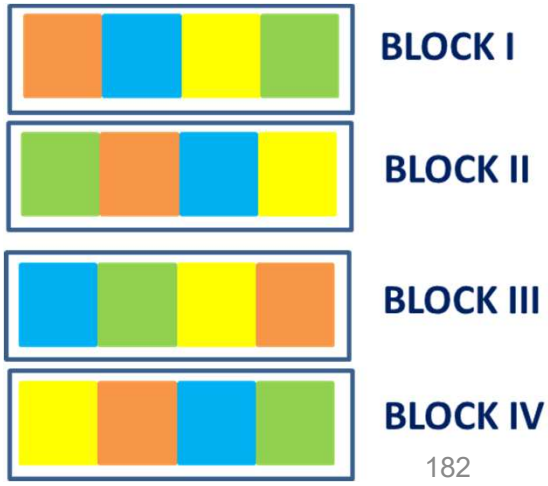
Uma disposição sistemática

A mesma letra repetida são pseudo-repetições

Delineamento em blocos completos casualizados (RCBD) desenvolvido por *Fisher* (1935) é dos delineamentos mais simples para o controlo da heterogeneidade (nomeadamente, da variação espacial em ensaios agronómicos).

- O ensaio é dividido em unidades homogéneas, de forma a controlar a variação no campo (ou na estufa, no laboratório, etc).
- Os níveis do factor a ser estudado (designados por tratamentos), são aleatoriamente atribuídos às unidades experimentais dentro do bloco (cada tratamento aparece uma única vez em cada bloco).
- O número de blocos é o número de repetições.

Exemplo: cada linha representa 1 bloco. Existem 4 blocos (I-IV) and 4 tratamentos (diferentes cores)



2 factores em estudo. Por exemplo, factor variedade (ou fertilizante, etc.) (4 níveis) e factor bloco (4 níveis).

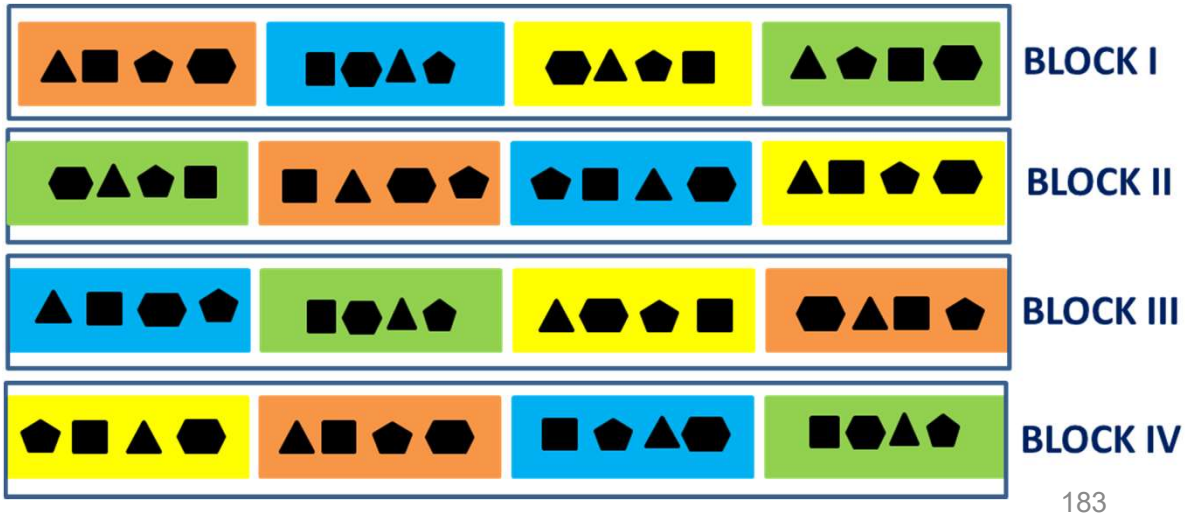
Bloco – é um factor para controlar uma fonte de variação que se sabe que existe (diferentes tipos de terrenos (pastagens diferentes), diferentes disponibilidades de água, operações culturais, diferentes temperaturas de um estábulo, diferentes luminosidades numa estufa, diferentes operadores de um aparelho, diferentes provadores, etc.).

Delineamento em parcelas divididas (*split-plot design em RCB*)

- O emsaio é dividido em unidades homogéneas, de forma a controlar a variação no campo, estufa, etc.
- O número de blocos é o número de repetições.
- Existem dois outros fatores em estudo (o objectivo do estudo). As unidades experimentais são organizadas em blocos, cada bloco com grandes parcelas (*whole plots)*, tantas quantas o número de níveis do factor A (factor principal), e cada grande parcela é depois dividida em pequenas parcelas (*split plots* ou *subplots*), às quais são aleatoriamente atribuídos os níveis do factor B.

Exemplo:
Cada linha representa 1 bloco.
Existem 4 blocos (I-IV) cada um com 4 tratamentos principais (cores, níveis do factor A; por exemplo, diferentes variedades) divididos em 4 outros tratamentos (*subplo*t, símbolos, níveis do factor B; por exemplo, sistemas de condução).

3 factores em estudo e várias interações



Delineamento em Quadrado Latino (LSD) (*Latin Square Design*)

Útil quando padrões de heterogeneidade estão associados a 2 factores com o mesmo número de níveis e configuração quadrada (chamados Linha e Coluna).

- ✓ O número de tratamentos (o número de níveis do factor que representa o objectivo do estudo) tem de ser igual ao número de linhas e de colunas
- ✓ A alocação dos tratamentos é feita de forma a que cada tratamento apareça exactamente uma única vez em cada linha e uma única vez em cada coluna.

Por exemplo: 4 linhas (L1, L2, L3, L4), 4 colunas (C1, C2, C3, C4) e 4 tratamentos (A,B,C,D)
3 factores em estudo: factor variedade (ou, fertilizante, etc.) (4 níveis), factor Linha (4 níveis) e Factor Coluna (4 níveis).

	C1	C2	C3	C4
L1	A	C	B	D
L2	B	D	C	A
L3	C	A	D	B
L4	D	B	A	C

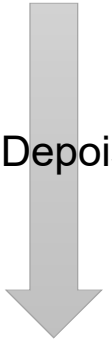
Mas, quando o número de tratamentos é elevado (nº de níveis de 1 factor), torna-se difícil garantir a homogeneidade dentro do bloco!

Surgiram depois os delineamentos pertencentes à classe dos

Delineamentos em Blocos Incompletos,

direcionados para o controlo da heterogeneidade (nomeadamente, espacial em ensaios agrícolas) em ensaios com maior número de tratamentos.

E muitos outros delineamentos experimentais…



Modelos de análise de dados adequados
(devem traduzir o delineamento experimental adotado)

Para os delineamentos mais simples e mais clássicos, serão apresentados os correspondentes modelos de "Análise de Variância"

Gerar delineamentos experimentais no SAS, Proc Plan
https://support.sas.com/documentation/onlinedoc/stat/141/plan.pdf

Delineamento totalmente casualizado (CRD)
Exemplo: 4 tratamentos, 3 repetições

```
proc plan seed=333569;
factors unit=12;
treatments Treatment=12 cyclic (1 1 1 2 2 2 3 3 3 4 4 4) ;
output out=CRDPlan;
run;
proc print data=CRDPlan;
run;
```

The PLAN Procedure

Plot Factors			
Factor	Select	Levels	Order
unit	12	12	Random

Treatment Factors				
Factor	Select	Levels	Order	Initial Block / Increment
Treatment	12	12	Cyclic	(1 1 1 2 2 2 3 3 3 4 4 4) / 1

unit												Treatment											
10	7	12	2	5	3	4	6	11	9	1	8	1	1	1	2	2	2	3	3	3	4	4	4

The SAS System

Obs	unit	Treatment
1	10	1
2	7	1
3	12	1
4	2	2
5	5	2
6	3	2
7	4	3
8	6	3
9	11	3
10	9	4
11	1	4
12	8	4

Gerar delineamentos experimentais no SAS, Proc Plan
https://support.sas.com/documentation/onlinedoc/stat/141/plan.pdf

Delineamento em blocos completos casualizados (RCBD)
Exemplo: 5 tratamentos, 6 blocos completos

```
 proc plan ordered seed=78390;
 factors blocks=6 cell=5;
 treatments t=5 random;
 output out=rcdb;
 run;
proc print data=rcdb;
run;
```

The PLAN Procedure

Plot Factors			
Factor	Select	Levels	Order
blocks	6	6	Ordered
cell	5	5	Ordered

Treatment Factors			
Factor	Select	Levels	Order
t	5	5	Random

blocks	cell					t				
1	1	2	3	4	5	2	3	1	4	5
2	1	2	3	4	5	4	2	3	1	5
3	1	2	3	4	5	2	1	4	5	3
4	1	2	3	4	5	1	2	4	5	3
5	1	2	3	4	5	5	4	3	1	2
6	1	2	3	4	5	3	1	5	2	4

The SAS System

Obs	blocks	cell	t
1	1	1	2
2	1	2	3
3	1	3	1
4	1	4	4
5	1	5	5
6	2	1	4
7	2	2	2
8	2	3	3
9	2	4	1
10	2	5	5
11	3	1	2
12	3	2	1
13	3	3	4
14	3	4	5
15	3	5	3
16	4	1	1

...

Gerar delineamentos experimentais no SAS, Proc Plan
https://support.sas.com/documentation/onlinedoc/stat/141/plan.pdf

Delineamento em Quadrado Latino (LSD)
Exemplo: 4 tratamentos

```
proc plan seed=37430;
factors Row=4 ordered Col=4 ordered / noprint;
treatments Tmt=4 cyclic;
output out=LatinSquare
Row cvals=('Day 1' 'Day 2' 'Day 3' 'Day 4') random
Col cvals=('Lab 1' 'Lab 2' 'Lab 3' 'Lab 4') random
Tmt nvals=( 0 100 250 450) random;
quit;
proc print data=LatinSquare noobs;
run;
```

The SAS System

Row	Col	Tmt
Day 3	Lab 4	250
Day 3	Lab 2	0
Day 3	Lab 1	100
Day 3	Lab 3	450
Day 4	Lab 4	0
Day 4	Lab 2	100
Day 4	Lab 1	450
Day 4	Lab 3	250
Day 2	Lab 4	100
Day 2	Lab 2	450
Day 2	Lab 1	250
Day 2	Lab 3	0
Day 1	Lab 4	450
Day 1	Lab 2	250
Day 1	Lab 1	0
Day 1	Lab 3	100

Gerar delineamentos experimentais no SAS, Proc Plan
https://support.sas.com/documentation/onlinedoc/stat/141/plan.pdf

Delineamento em parcelas divididas (Split-Plot)

Exemplo: delineamento em parcelas divididas em blocos completos casualizados - 5 blocos completos cada um com 2 tratamentos principais (grandes parcelas, 2 níveis do factor A), divididos em 3 outros tratamentos (pequenas parcelas, 3 níveis do factor B)

```
title 'Split Plot Design';
proc plan seed=37277;
factors Block=5 ordered a=2 b3;
run;
```

The PLAN Procedure

Factor	Select	Levels	Order
Block	5	5	Ordered
a	2	2	Random
b	3	3	Random

Block	a	b		
1	2	3	2	1
	1	3	1	2
2	1	2	3	1
	2	2	3	1
3	1	2	3	1
	2	2	1	3
4	2	2	1	3
	1	3	2	1
5	2	3	1	2
	1	3	1	2

Block	a	b
1	2	3
1	2	2
1	2	1
1	1	3
1	1	1
1	1	2
2	1	2
2	1	3
2	1	1
2	2	2
2	2	3
2	2	1
3	1	2
3	1	3
3	1	1
3	2	2
3	2	1
3	2	3
4	2	2

...