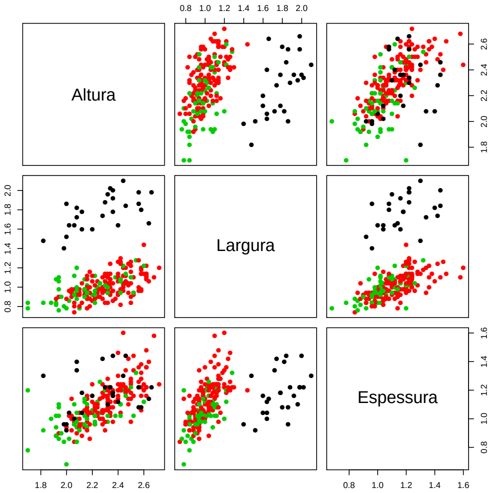
INSTITUTO SUPERIOR DE AGRONOMIA

Modelos Matemáticos e Aplicações (2020-21)

19 Abril, 2021 **Teste – Modelo Linear** Duração: 2h30

I [16 valores]

Um estudo sobre alho envolveu a medição (em cm) da `Altura`, `Largura` e `Espessura` de dentes de alho. A cor do tegumento (pele) dos dentes foi classificada numa de três categorias (codificadas como cores 3, 4 ou 5). Em baixo são dadas as médias e variâncias de cada variável numérica observada, para todos os dentes, bem como para os dentes com cada cor de tegumento. Também são dados os coeficientes de correlação linear (para todos os dentes), bem como as nuvens de pontos de cada par de variáveis numéricas, usando cores diferentes para cada cor de tegumento.



Tegumento: cor 3 cor 4 cor 5

Médias

dados	dimensão	`Altura`	`Largura`	`Espessura`
Todos	223	2.282870	1.095605	1.112242
Cor 3	26	2.273077	1.777692	1.178462
Cor 4	143	2.330070	1.016084	1.131958
Cor 5	54	2.162593	0.977778	1.028148

Variâncias

dados	`Altura`	`Largura`	`Espessura`
Todos	0.04117731	0.07903465	0.02178864
Cor 3	0.05315015	0.03255446	0.02281354
Cor 4	0.03005563	0.01532540	0.02028487
Cor 5	0.04608372	0.01519497	0.01575122

Correlações (todos os dados)

```
            Altura   Largura Espessura
Altura    1.0000000 0.2546324 0.6402415
Largura   0.2546324 1.0000000 0.4229474
Espessura 0.6402415 0.4229474 1.0000000
```

1. Um primeiro modelo considerou apenas as observações com a cor 4 do tegumento. A espessura do dente foi modelada com uma regressão linear sobre a altura e largura, com os seguintes resultados:

```
Call:
lm(formula = Espessura ~ Altura + Largura, data = alho.cordentes[alho.cordentes$Cor==4, ])
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.31984    0.10028  -3.190  0.00176
Altura       0.46123    0.05141   8.973 1.70e-15
Largura      0.37113    0.07199   5.155 8.46e-07
---
Residual standard error: 0.08846 on 140 degrees of freedom
Multiple R-squared:  0.6197,  Adjusted R-squared:  0.6143
F-statistic: ??? on ??? and ??? DF,  p-value: < 2.2e-16
```

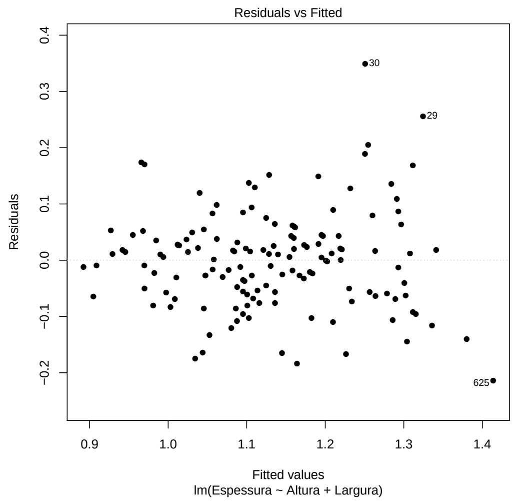
   (a) Escreva a equação do plano em $\mathbb{R}^3$ ajustado por esta regressão.

   (b) Qual é o valor ajustado para a variável resposta `Espessura`, quando ambos os preditores tomam o respectivo valor médio dos dados usados para ajustar o modelo (cor 4 do tegumento)?

   (c) Discuta e teste ($\alpha = 0.05$) a qualidade global do ajustamento do modelo.

   (d) Para uma dada altura de dente, calcule um intervalo a 95% de confiança para a variação média na sua espessura, correspondente a aumentar a largura em 1 mm (0.1 cm).

(e) Em baixo encontra a nuvem de pontos dos resíduos (usuais) sobre os valores ajustados da variável resposta. Comente-a.



(f) A `Espessura` observada para a observação designada 30, no topo da nuvem de pontos acima, é 1.60. O seu resíduo *estandardizado* é 3.9924. *Sem efectuar quaisquer cálculos*, o que pode afirmar sobre o valor ajustado desta observação?

2. Considere agora modelos ANOVA duma das variáveis numéricas sobre o factor cor do tegumento.

   (a) *Sem efectuar quaisquer cálculos* diga, justificando, qual a variável resposta que espera tenha os efeitos de cor de tegumento mais significativos.

   (b) Para a variável resposta que escolheu na alínea anterior, construa a tabela-resumo e discuta os seus resultados. Que conclusões pode extrair sobre a relação entre a variável resposta escolhida e a cor do tegumento?

3. Decidiu-se modelar a `Largura` dos dentes em função da sua `Altura`, mas admitindo diferentes equações de regressão linear para cada cor de tegumento.

   (a) Um primeiro modelo admite que possa existir uma recta de regressão totalmente diferente para cada cor de tegumento. Eis os resultados do ajustamento deste modelo:

```
Call: lm(formula = Largura ~ Altura * Cor, data = alho.cordentes)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.69794    0.21405   3.261  0.00129
Altura       0.47502    0.09370   5.069 8.54e-07
Cor4        -0.60257    0.24645  -2.445  0.01528
Cor5        -0.47481    0.26148  -1.816  0.07077
Altura:Cor4 -0.07987    0.10730  -0.744  0.45745
Altura:Cor5 -0.12606    0.11644  -1.083  0.28016
---
Residual standard error: 0.108 on 217 degrees of freedom
Multiple R-squared:  0.8557,  Adjusted R-squared:  0.8524
F-statistic: 257.4 on 5 and 217 DF,  p-value: < 2.2e-16
```

   i. Escreva a equação da recta de regressão ajustada para os dentes com a cor 4 do tegumento.

   ii. Teste se os declives das rectas ajustadas para as cores 3 e 5 do tegumento são significativamente diferentes.

   (b) Um segundo modelo admitiu que as rectas de regressão das três cores de tegumento são paralelas, mas podendo ter diferentes ordenadas na origem. Eis os resultados:

```
Call: lm(formula = Largura ~ Altura + Cor, data = alho.cordentes)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.88138    0.08898   9.905   <2e-16
Altura       0.39432    0.03802  10.370   <2e-16
Cor4        -0.78408    0.02309 -33.962   <2e-16
Cor5        -0.75635    0.02608 -29.006   <2e-16
---
Residual standard error: 0.1078 on 219 degrees of freedom
Multiple R-squared:  0.8549,  Adjusted R-squared:  0.8529
F-statistic: 430.2 on 3 and 219 DF,  p-value: < 2.2e-16
```

Descreva em pormenor o teste para determinar se este modelo tem um ajustamento significativamente pior que o modelo da alínea anterior. Quais as implicações práticas do resultado que obteve?

(c) Qual o coeficiente de determinação duma recta de regressão únca de `Largura` sobre `Altura`, independentemente da cor do tegumento? Esse valor é significativamente diferente de zero? Quais as conclusões práticas e a lição geral resultantes desse valor?

II [4 valores]

1. Sabe-se que no Modelo Linear, a distribuição de probabilidades do vector de valores da variável resposta é $\vec{\mathbf{Y}} \frown \mathcal{N}_n(\mathbf{X}\vec{\boldsymbol{\beta}}, \sigma^2\mathbf{I}_n)$. Mostre que o vector $\frac{1}{\sigma}(\vec{\mathbf{Y}}-\mathbf{X}\vec{\boldsymbol{\beta}})$ tem a distribuição Multinormal estandardizada, $\mathcal{N}_n(\vec{\mathbf{0}}, \mathbf{I}_n)$.

2. Seja $\mathbf{X}_c$ a matriz do modelo dum Modelo Linear com p preditores e $\mathbf{X}_s$ a matriz do modelo dum seu submodelo, com apenas $k < p$ preditores. Sejam $\mathbf{H}_c$ e $\mathbf{H}_s$ as correspondentes matrizes de projecção ortogonal (*hat matrices*).

   (a) Mostre que o espaço das colunas de $\mathbf{X}_s$ está contido no espaço das colunas de $\mathbf{X}_c$, ou seja, que qualquer vector $\vec{\mathbf{y}} \in \mathcal{C}(\mathbf{X}_s)$ também pertence a $\mathcal{C}(\mathbf{X}_c)$.

   (b) Mostre que tem de se verificar $\mathbf{H}_c\mathbf{H}_s = \mathbf{H}_s$, e também $\mathbf{H}_s = \mathbf{H}_s\mathbf{H}_c$. Interprete estes resultados.

   (c) A aplicação do Teorema de Cochran para justificar o teste F parcial que compara o modelo completo e o submodelo exige que, na decomposição $\mathbf{I}_n = (\mathbf{I}_n - \mathbf{H}_c) + (\mathbf{H}_c - \mathbf{H}_s) + (\mathbf{H}_s - \mathbf{P}_{\vec{\mathbf{1}}_n}) + \mathbf{P}_{\vec{\mathbf{1}}_n}$, as duas primeiras parcelas no lado direito da igualdade sejam matrizes de projecção ortogonal que, quando multiplicadas, originem uma matriz de zeros, $\mathbf{0}_n$.

      i. Sabendo que as matrizes de projecção ortogonal sobre subespaços de $\mathbb{R}^n$ são as matrizes simétricas e idempotentes de dimensão $n \times n$ mostre que, quer $(\mathbf{I}_n - \mathbf{H}_c)$, quer $(\mathbf{H}_c - \mathbf{H}_s)$ são matrizes de projecção ortogonal. **Nota:** Pode admitir que ambas as *hat matrices* são matrizes de projecção ortogonal.

      ii. Mostre que $(\mathbf{I}_n - \mathbf{H}_c)(\mathbf{H}_c - \mathbf{H}_s) = \mathbf{0}_n$.