INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2016-17

9 de Janeiro de 2017 Primeira Chamada de Exame Duração: 3h30

I [2,5 valores]

Pretende-se avaliar o comportamento de duas espécies de borboletas e de pássaros que delas se alimentam. A primeira espécie de borboletas tem coloração amarela e listas pretas, e é indesejada pelos pássaros, presume-se que devido a mau sabor. A segunda espécie, mais apetecível, tem uma aparência variável. Por vezes aparece com coloração amarela e listas, o que a torna difícil de distinguir da primeira espécie. Outras vezes surge com cor amarela mas sem listas, ficando menos parecida com a primeira espécie. Pode ainda surgir com cor branca, o que a torna bastante diferente das borboletas da primeira espécie. Uma investigadora hipotiza que a segunda espécie mimetiza a primeira espécie, como mecanismo de defesa que visa confundir os pássaros. Decide-se estudar o assunto, começando por identificar a frequência relativa, na população da segunda espécie, de cada tipo de aparência. Depois, identifica-se a categoria que pertencem 45 borboletas da segunda espécie que são comidas por pássaros. Os resultados obtidos são indicados na seguinte tabela.

	amarelo com listas	amarelo sem listas	branco
Frequência na população	0.33	0.54	0.13
Borboletas comidas	7	30	8

1. Indique um teste baseado na estatística de Pearson para avaliar se cada uma das 3 categorias de borboletas da segunda espécie é comida em proporção idêntica à da sua frequência populacional. Em particular, indique as hipóteses nula e alternativa, interpretando-as, a estatística do teste e sua distribuição assintótica sob H_0, bem como a região crítica do teste.

2. Efectue o teste (com um nível de significância 0.05), tendo o cuidado de verificar se é legítimo admitir a validade da distribuição assintótica. Comente a sua conclusão.

3. Tendo em conta toda a informação disponível, diga se considera os dados compatíveis com a hipótese avançada pela investigadora.

II [8,5 valores]

Em estudos de fisiologia da videira é importante quantificar os carotenóides. Contudo, a sua medição envolve a utilização de solventes mais caros do que os usados na mediação das clorofilas e de metabolitos antioxidantes. Pretende-se estudar a modelação do teor de carotenóides (variável `Carot`, em μmol ml^{-1}), a partir das concentrações de clorofila a (variável `Chla`, em μmol ml^{-1}), clorofila b (variável `Chlb`, em μmol ml^{-1}), glutationa oxidada (variável `GSSG`, em μmol g^{-1}), glutationa reduzida (variável `GSH`, em μmol g^{-1}) e ácido ascórbico (variável `AsA`). Dispõem-se de medições destes compostos em folhas de 36 plantas da casta Trincadeira, cujas médias, desvios padrão, mínimos, máximos, e matriz de correlações são dados de seguida:

	$\overline{x}$	s	min	max
Carot	82.8300	55.9640	-0.158	182.915
Chla	179.700	115.0790	2.909	385.012
Chlb	194.800	139.2254	4.347	532.908
GSSG	97.650	71.5406	9.034	303.103
GSH	268.40	170.2865	14.975	661.843
AsA	1.2620	0.4083	0.583	2.175

	Carot	Chla	Chlb	GSSG	GSH	AsA
Carot	1.000	0.989	0.883	-0.014	0.274	0.304
Chla	0.989	1.000	0.915	-0.042	0.358	0.292
Chlb	0.883	0.915	1.000	0.016	0.276	0.230
GSSG	-0.014	-0.042	0.016	1.000	0.141	0.108
GSH	0.274	0.358	0.276	0.141	1.000	0.228
AsA	0.304	0.292	0.230	0.108	0.228	1.000

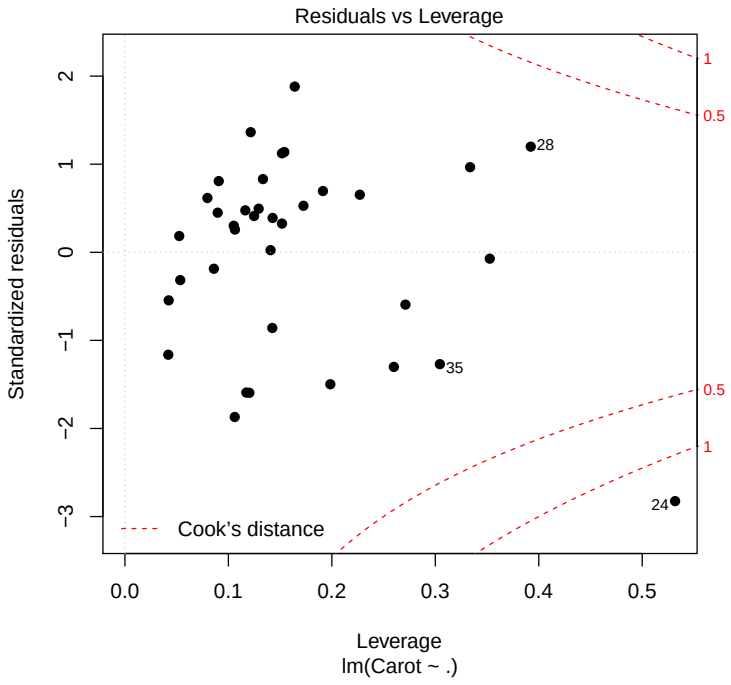
1. Foi inicialmente ajustado um modelo de regressão linear múltipla utilizando a totalidade dos preditores disponíveis, obtendo-se os seguintes resultados:

```
Call: lm(formula = Carot ~ ., data = carot)
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) -4.291443    3.122114   -1.375   0.17946
Chla         0.581162    0.020504   28.344   < 2e-16
Chlb        -0.073829    0.016167   -4.567  7.90e-05
GSSG         0.042434    0.012783    3.320   0.00238
GSH         -0.037575    0.005736   -6.551  3.02e-07
AsA          2.360861    2.305387    1.024   0.31399
---
Residual standard error: ??? on 30 degrees of freedom
Multiple R-squared:  0.9925,   Adjusted R-squared:  0.9913
F-statistic: 798.8 on 5 and 30 DF,  p-value: < 2.2e-16
```

   (a) Calcule, justificando, uma estimativa da variância σ^2 dos erros aleatórios do modelo.

   (b) Um algoritmo de exclusão sequencial, baseado no Critério de Informação de Akaike (AIC), seleccionou um submodelo final com quatro variáveis preditoras e com $AIC = 123.64$. Diga, justificando, qual foi o preditor excluído. Qual o valor do Quadrado Médio residual no submodelo escolhido? Comente.

   (c) Descreva e comente o seguinte gráfico. Tenha também em conta os valores da videira 24, indicados ao lado.




```
> t(carot[24,])
                24
Carot 103.829
Chla  249.572
Chlb  101.334
GSSG    9.034
GSH   661.843
AsA     2.175
```

2. Na tentativa de obter um submodelo ainda mais parcimonioso, um investigador ajustou um modelo só com dois preditores: `Chla` e `GSH`. O ajustamento do referido submodelo produziu os seguintes resultados.

```
Call: lm(formula = Carot ~ Chla + GSH, data = carot)
```

```
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  1.590835    2.562321    0.621  0.538959
Chla         0.496737    0.011120   44.672   < 2e-16
GSH         -0.029991    0.007515   -3.991  0.000345
---
Residual standard error: 7.069 on 33 degrees of freedom
Multiple R-squared:  0.985,Adjusted R-squared:  0.984
F-statistic:  1080 on 2 and 33 DF,  p-value: < 2.2e-16
```

Teste formalmente ($\alpha = 0.05$) se este submodelo difere significativamente do modelo completo. Comente, tendo em conta os valores dos coeficientes de determinação de cada modelo.

3. Considere agora modelos de regressão linear simples para modelar o teor de carotenóides.

   (a) Qual o melhor preditor, de entre os cinco preditores estudados? Justifique e indique a proporção de variabilidade dos teores observados de carotenóides que é explicado pelo modelo que escolheu.

   (b) Diga, justificando brevemente, se esta regressão linear simples se ajusta significativamente pior ($\alpha = 0.05$) do que o modelo com dois preditores ajustado anteriormente.

   (c) Calcule a variância $s_{\hat{y}}^2$ dos teores de carotenóides ajustados pelo modelo que escolheu.

   (d) Construa um intervalo a 95% de confiança para o declive da recta de regressão do modelo escolhido. Interprete o resultado.

III [4,5 valores]

No âmbito dum estudo sobre uma variedade de arroz, pretende-se avaliar o teor final de zinco que é assimilado no grão (variável `Zn`, em mg kg^{-1} de matéria seca), para cada um de três diferentes níveis de adubação foliar (0, 300 e 600 em mg kg^{-1} de solo). A experiência foi realizada com base em grãos de dois tipos (polido e integral). Para cada combinação de tipo de grão e nível de adubação foliar existem 8 observações. As médias de cada um desses grupos de 8 observações são dados na tabela seguinte:

	0	300	600
grão polido	3.464	3.591	5.131
grão integral	9.934	10.313	10.664

1. Indique o delineamento experimental utilizado e descreva pormenorizadamente o modelo ANOVA adequado à experiência.

2. Um investigador efectuou uma ANOVA, tendo obtido a seguinte tabela-resumo:

```
               Df  Sum Sq  Mean Sq  F value
Adubacao       ??   12.78     6.39    2.715
Grao           ??  467.44       ??       ??
Adubacao:Grao  ??    3.14     1.57       ??
Residuals      ??      ??     2.35
```

   (a) Complete a tabela-resumo da ANOVA, indicando como obtém os oito valores em falta

(b) Que tipo de efeitos devem ser considerados significativos ao nível $\alpha = 0.10$? Responda de forma pormenorizada num caso, e de forma sucinta no(s) restante(s).

(c) Que pares de combinações de tipos de grão com níveis de adubação têm médias significativamente diferentes ao abrigo da teoria de Tukey ($\alpha = 0.05$)?

(d) Construa a tabela-resumo resultante de ajustar, aos mesmos dados, um modelo ANOVA que apenas preveja a existência de efeitos de adubação. Qual a conclusão que se teria num teste F à existência desse tipo de efeitos ($\alpha = 0.10$)? Comente.

IV [4,5 valores]

1. Considere um modelo de regressão linear múltipla com p variáveis preditoras, ajustado com base em n observações.

   (a) Descreva pormenorizadamente o modelo, *usando a notação vectorial/matricial*.

   (b) Mostre que o vector de estimadores dos parâmetros do modelo, $\vec{\hat{\boldsymbol{\beta}}}$, também se pode escrever como $\vec{\hat{\boldsymbol{\beta}}} = \vec{\boldsymbol{\beta}} + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\boldsymbol{\epsilon}}$

   (c) Deduza *a partir da expressão da alínea anterior*, o vector esperado e a matriz de covariâncias do vector dos estimadores, $\vec{\hat{\boldsymbol{\beta}}}$, ao abrigo do modelo de regressão linear múltipla.

2. Considere os coeficientes de determinação usual (R^2) e modificado (R^2_{mod}), no contexto duma regressão linear múltipla com p variáveis preditoras , ajustada com base em n observações.

   (a) Mostre que se verifica a relação $R^2_{mod} = 1 - (1 - R^2)\frac{n-1}{n-(p+1)}$.

   (b) Mostre que a estatística do teste F de ajustamento global do modelo se pode escrever apenas à custa de R^2 e R^2_{mod}, verificando-se $F_{calc} = \frac{R^2}{R^2 - R^2_{mod}}$.

   (c) Mostre que o coeficiente de determinação modificado é negativo quando $R^2 < \frac{p}{n-1}$. Comente as implicações desta condição para a estatística do teste F de ajustamento global.