

INSTITUTO SUPERIOR DE AGRONOMIA
ESTATÍSTICA E DELINEAMENTO – 2020-21
Segundo Teste

11 Janeiro 2021

Duração: 2h30

Aviso: Justifique convenientemente as suas respostas.

I [7 valores]

Um estudo sobre pastagens visou modelar a quantidade de matéria seca (variável **Produtividade**, em kg/ha). Dispunham-se de algumas potenciais variáveis preditoras de três tipos: composição química dos solos, composição das pastagens e um grande número de índices de vegetação, obtidos a partir de imagens de satélite. Foram obtidos valores de todas as variáveis em 47 diferentes locais no Alentejo.

1. Um primeiro modelo de regressão linear múltipla produziu os seguintes resultados:

Residual standard error: 927.3 on 25 degrees of freedom
Multiple R-squared: 0.586, Adjusted R-squared: 0.2382
F-statistic: ???? on 21 and 25 DF, p-value: ????

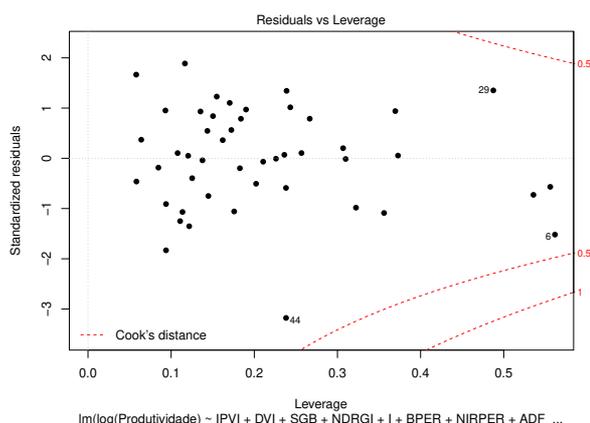
- (a) Indique o número de variáveis preditoras usadas e o número de parâmetros do modelo.
 - (b) Efectue um teste de ajustamento global e discuta a qualidade do ajustamento do modelo, tendo também em conta os valores dos R^2 usual e modificado.
2. Foi seguidamente ajustado um modelo com transformação logarítmica da variável resposta **Produtividade** (sem alteração dos preditores), tendo-se obtido os seguintes resultados.

Residual standard error: 0.3396 on 25 degrees of freedom
Multiple R-squared: 0.7199, Adjusted R-squared: 0.4846
F-statistic: 3.06 on 21 and 25 DF, p-value: 0.004255

- (a) Comente brevemente os resultados do novo modelo. Comente a validade da seguinte afirmação: *“este modelo explica quase 72% da variabilidade das produtividades observadas”*.
- (b) Um algoritmo de exclusão sequencial produziu um submodelo com os seguintes resultados.

Residual standard error: 0.2978 on 37 degrees of freedom
Multiple R-squared: 0.6814, Adjusted R-squared: 0.6039
F-statistic: 8.791 on 9 and 37 DF, p-value: 6.413e-07

- i. Como se explica que, comparando com o modelo completo, haja uma subida considerável no valor do R^2 modificado, enquanto o R^2 usual diminui?
- ii. Teste formalmente se o ajustamento deste submodelo difere significativamente do ajustamento do modelo completo indicado no início desta pergunta 2.
- iii. Descreva e comente o seguinte gráfico relativo ao submodelo agora ajustado. Sabe-se que a observação com a maior distância de Cook é a observação 44. Calcule, com base na informação disponível no gráfico, um valor *aproximado* dessa distância de Cook e comente-o.



3. Foi decidido ajustar um novo submodelo, apenas com 5 preditores, todos índices de vegetação observáveis a partir de imagens de satélite. Eis alguns resultados.

```
Call: lm(formula=log(Produtividade) ~ DVI + SGB + NDRGI + I + NIRPER, data=pastagens)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	12.534430	1.752840	7.151	1.01e-08
DVI	0.051758	0.013432	3.853	0.000402
SGB	0.020533	0.008181	2.510	0.016123
NDRGI	5.486398	1.826882	3.003	0.004538
I	-0.013159	0.003919	-3.358	0.001706
NIRPER	-8.275832	2.769283	-2.988	0.004721

Residual standard error: 0.3323 on 41 degrees of freedom
 Multiple R-squared: 0.5603, Adjusted R-squared: 0.5067
 F-statistic: 10.45 on 5 and 41 DF, p-value: 1.635e-06

Teste ($\alpha=0.01$) se é possível afirmar que, sendo os restantes índices constantes, um aumento do índice NIRPER (*Near-infrared percentage*) está associado a uma diminuição da log-produtividade.

II [6 valores]

Ensaio com a casta Vital foram realizados para determinar se o seu rendimento difere entre localidades e entre terrenos onde se pretende efectuar a produção. Foram ensaiados três diferentes locais: Bombarral, Cadaval e Caldas da Rainha. Em cada local foram utilizados oito terrenos, tendo os terrenos características de solos diferentes, quer em cada localidade, quer entre localidades. Cada terreno foi dividido em sete parcelas, tendo sido medido o rendimento em kg/planta obtido com a casta em cada parcela. A média dos 168 rendimentos observados foi 3.60162 kg/planta e a respectiva variância foi 3.08414 (kg/planta)². O rendimento médio obtido no Bombarral foi 2.280, no Cadaval 4.361, e nas Caldas 4.164. Os rendimentos médios em cada um dos 24 terrenos foram os seguintes:

	terrenos							
Bombarral	1.823	2.080	2.270	2.694	2.030	2.696	2.331	2.313
Cadaval	5.772	3.994	3.718	4.617	5.357	4.048	3.639	3.742
Caldas	1.791	3.064	3.596	5.652	4.891	5.054	5.538	3.729

1. Identifique o delineamento experimental utilizado e descreva em pormenor o modelo ANOVA adequado.
2. Sabendo que o Quadrado Médio Residual é 1.673 e que o valor da estatística F associada ao teste a efeitos de terreno é 3.592, complete a tabela de síntese do modelo ANOVA, indicando como obtém os valores em falta.
3. Teste se há efeitos significativos de localidade e comente a sua conclusão.
4. Foi importante ter utilizado diferentes terrenos em cada localidade? Responda, justificando brevemente através dum teste de hipóteses adequado.
5. O maior rendimento médio foi observado no primeiro terreno do Cadaval. Este rendimento é significativamente diferente dos rendimentos de quais outros terrenos, ao nível $\alpha=0.05$?

III [2,5 valores]

Foi realizado um estudo de afinidade à enxertia (viabilidade) de videiras da casta Azal, uma casta da região de Vinhos Verdes. Foram usados três diferentes porta-enxertos, designados 1103P, R110 e 196/17. O experimentador efectuou 1754 enxertos com o porta-enxertos 1103P; 1787 com o R110; e 1747 com o 196/17. Estas plantas foram para estratificação durante 3 semanas após o que foram plantadas em viveiro. Passado algum tempo, fez-se a contagem do número de plantas viáveis (comercializáveis). Eis os resultados obtidos.

Porta-enxerto	Plantas viáveis	Plantas não viáveis	Total
1103P	1226	528	1754
R110	1295	492	1787
196/17	1357	390	1747
Total	3878	1410	5288

Pretende-se saber se é possível afirmar que os porta-enxertos têm igual desempenho no que respeita à afinidade à enxertia (viabilidade).

1. Identifique o tipo de Teste de Hipóteses a efectuar e descreva as Hipóteses em confronto.
2. Considera que existem observações suficientes para utilizar a distribuição assintótica da respectiva estatística de teste?
3. Calcule a contribuição para o valor da estatística de teste das plantas não viáveis no porta-enxertos 196/17.
4. Sabendo que o valor calculado da estatística de teste é $X_{calc}^2 = 28.123$, qual a sua conclusão ao nível de significância $\alpha = 0.05$?

IV [4,5 valores]

1. Considere o modelo ANOVA a um factor, com k níveis e n_i observações da variável resposta Y no nível i do factor.

- (a) Caracterize a matriz do modelo \mathbf{X} .
 - (b) Mostre que para qualquer vector do espaço das colunas $\mathcal{C}(\mathbf{X})$, os elementos nas posições correspondentes a observações dum mesmo nível do factor são todos iguais.
2. Considere uma regressão linear múltipla, com p preditores e ajustada com base em n observações.
- (a) Escreva o Modelo de regressão linear múltipla *usando notação vectorial/matricial*.
 - (b) Deduza a distribuição do vector aleatório das observações da variável resposta, $\vec{\mathbf{Y}}$, ao abrigo do modelo.
3. Considere uma generalização do modelo de regressão linear múltipla, em que a matriz de covariâncias do vector de erros aleatórios seja uma matriz simétrica *genérica* Σ , admitida conhecida.
- (a) A que corresponde admitir que a matriz Σ *não* é uma matriz diagonal (ou seja, que haja elementos não nulos fora da diagonal principal)?
 - (b) Sabendo que a distribuição de probabilidades do vector $\vec{\mathbf{Y}}$ é agora $\mathcal{N}_n(\mathbf{X}\vec{\beta}, \Sigma)$, deduza a distribuição de probabilidades do vector usual de estimadores, $\vec{\hat{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{Y}}$. Os estimadores são centrados?