INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO EXPERIMENTAL – 2024-25

13 de janeiro de 2025 Primeira Chamada de Exame Duração: 3h

I [13 valores]

Um estudo sobre a variedade de macieira "**Bravo** de Esmolfe", realizado um ano após a transplantação das árvores para o terreno, visa encontrar um modelo para prever o número médio de frutos à colheita (variável `nfrColh`) a partir de outras 4 variáveis: o diâmetro das árvores 10 cm acima do ponto de enxertia (variável `diametro`, em cm), a altura da árvore (variável `altura`, em m), o número médio de frutos contados em Junho (variável `nfrJun`) e o número médio de frutos contados em Setembro (variável `nfrSet`). Dispõem-se de valores médios de todas as variáveis para 66 parcelas plantados ao acaso num ensaio.

1. Um dos técnicos que realizou este estudo decidiu ajustar o modelo de regressão linear múltipla do número médio de frutos à colheita (variável `nfrColh`) sobre as restantes variáveis observadas. Eis os resultados obtidos com o ajustamento desse modelo, bem como a matriz de (co-) variâncias estimadas dos estimadores dos parâmetros do modelo:

```
> macieirabravo1.lm<-lm(nfrColh~diametro+altura+nfrJun+nfrSet, data=macieirabravo)
> summary(macieirabravo1.lm)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.33619    0.97581   1.369    0.176
diametro     0.01684    0.05833   0.289    0.774
altura      -1.22301    0.77971  -1.569    0.122
nfrJun      -0.08156    0.16230  -0.502    0.617
nfrSet       0.85605    0.19427    A       4.33e-05
---

Residual standard error: 0.5034 on B degrees of freedom
Multiple R-squared:  0.9496,Adjusted R-squared:  0.9463
F-statistic: C on 4 and 61 DF,  p-value: < 2.2e-16


> round(vcov(macieirabravo1.lm), 3)
            (Intercept) diametro altura nfrJun nfrSet
(Intercept)       0.952   -0.001 -0.526  0.022 -0.019
diametro         -0.001    0.003 -0.032  0.001 -0.002
altura           -0.526   -0.032    D   -0.024  0.031
nfrJun            0.022    0.001 -0.024  0.026 -0.031
nfrSet           -0.019   -0.002  0.031 -0.031  0.038
>
```

   (a) No *output* dos comandos *summary* e *vcov* do R identifique e calcule, justificando, os valores em falta A, B, C e D .

   (b) Ao nível de significância $\alpha = 0.05$, o modelo ajustado difere significativamente do modelo nulo? Justifique a sua resposta.

   (c) Dado o Modelo de Regressão Linear Múltipla, e sabendo que o vector de estimadores dos parâmetros é dado por $\vec{\hat{\boldsymbol{\beta}}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\vec{\mathbf{Y}}$, prove que a matriz de (co-)variâncias do vector de estimadores é dada por $V[\vec{\hat{\boldsymbol{\beta}}}] = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$.

   (d) Após a aplicação do algoritmo de exclusão sequencial com base em testes T aos parâmetros do modelo (para $\alpha = 0.10$), chegou-se ao seguinte submdelo:

```
> macieirabravo2.lm<-lm(nfrColh~altura+nfrSet, data=macieirabravo)
> summary(macieirabravo2.lm)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.41493    0.95357   1.484   0.1428
altura      -1.10790    0.54136  -2.047   0.0449
nfrSet       0.76444    0.02365  32.322   <2e-16
---

Residual standard error: 0.4968 on 63 degrees of freedom
Multiple R-squared:  0.9493,Adjusted R-squared:  0.9477
F-statistic: 589.8 on 2 and 63 DF,  p-value: < 2.2e-16
```

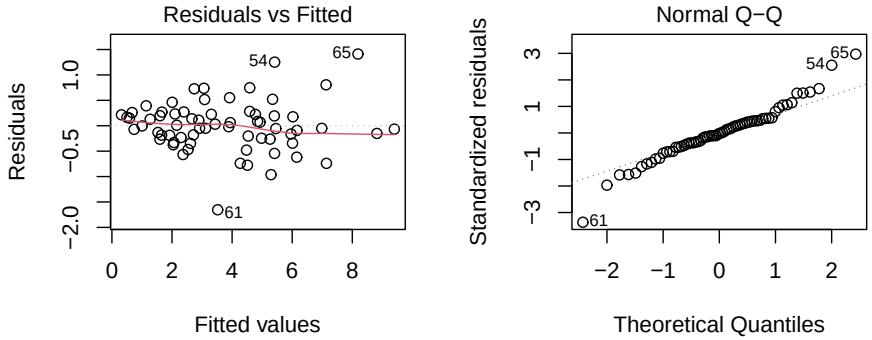
i. Qual foi a ordem de exclusão das duas variáveis que saíram do modelo? Explique e justifique a sua resposta.

ii. A qualidade do ajustamento do modelo completo e do seu submodelo difere significativamente ($\alpha = 0.05$)? Descreva, em pormenor, o teste efectuado para responder à questão.

iii. Interprete o significado da estimativa do coeficiente associado ao preditor `altura`.

iv. Calcule um intervalo a 95% de confiança para o coeficiente associado ao preditor `altura`. Explique o seu significado.

v. Alguns dos gráficos dos resíduos obtidos com o ajustamento deste submodelo com dois preditores apresentam-se seguidamente:



A. Analise os 2 gráficos e discuta as suas implicações para o estudo efetuado.

B. Calcule o valor exacto do resíduo usual associado à observação 61 (valores das variáveis na observação 61: $altura_{61} = 1.917$, $nfrSet_{61} = 5.533$, $nfrColh_{61} = 1.867$).

2. No estudo efetuado, as árvores das 66 parcelas são de diferentes regiões, estando 22 parcelas plantadas com material originário de cada uma de 3 regiões de origem: Trás-os-Montes (`TM`), Beira Interior (`BI`) e Beira Litoral (`BL`). Assim, o técnico, considerando válidos os pressupostos do modelo e olhando para o resultado do ajustamento do submodelo ao nível $\alpha = 0.01$, decidiu ajustar um modelo de Análise de Covariância à totalidade das $n = 66$ observações, tendo como variável resposta o número médio de frutos à colheita (variável `nfrColh`), variável preditora numérica o número médio de frutos contados em Setembro (variável `nfrSet`) e variável preditora categórica a `origem` (factor, com 3 níveis). Eis os resultados do ajustamento deste modelo:

```
> macieirabravo3.lm<-lm(nfrColh~nfrSet*origem, data=macieirabravo)
> summary(macieirabravo3.lm)
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.16111    0.22461  -0.717   0.4760
nfrSet          0.64392    0.04845  13.290   <2e-16
```

```
origemBL         -0.29445    0.34268  -0.859   0.3936
origemTM         -0.70381    0.36783  -1.913   0.0605
nfrSet:origemBL   0.10240    0.06123   1.672   0.0996
nfrSet:origemTM   0.15256    0.06316   2.416   0.0188
---

Residual standard error: 0.4987 on 60 degrees of freedom
Multiple R-squared:  0.9513,Adjusted R-squared:  0.9473
F-statistic: 234.6 on 5 and 60 DF,  p-value: < 2.2e-16
```

(a) Interprete as estimativas dos parâmetros do modelo e escreva a equação da recta ajustada referente a cada região de origem de "**Bravo** de Esmolfe".

(b) Os declives das rectas de regressão para as origens Trás-os-Montes (TM) e Beira Interior (BI) são significativamente diferentes ao nível $\alpha = 0.05$? Justifique a sua resposta com o teste de hipóteses adequado, indicando todos os seus passos.

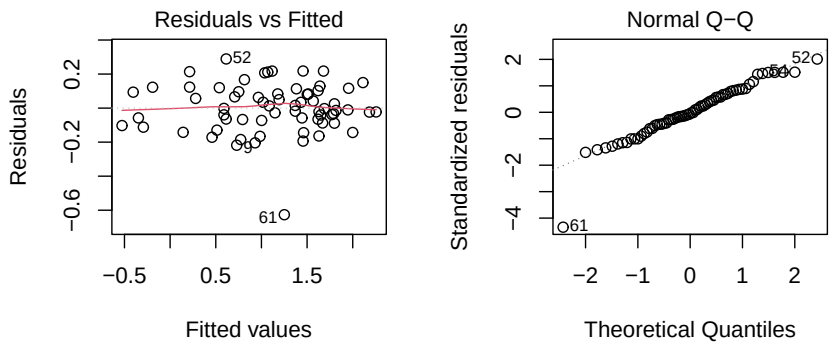
3. Um outro técnico que também analisou os resultados dos modelos ajustados, sugeriu ajustar uma regressão linear simples para prever o logaritmo do número médio de frutos à colheita a partir do logaritmo do número médio de frutos contados em Setembro. O resultado do ajustamento foi o seguinte:

```
> macieirabravo22.lm<-lm(log(nfrColh)~log(nfrSet), data=macieirabravo)
> summary(macieirabravo22.lm)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.73860    0.05327  -13.87   <2e-16
log(nfrSet)  1.16293    0.03143   37.00   <2e-16
---
Residual standard error: 0.1454 on 64 degrees of freedom
Multiple R-squared:  0.9553,Adjusted R-squared:  0.9546
F-statistic:  1369 on 1 and 64 DF,  p-value: < 2.2e-16
```

(a) Interprete o valor do coeficiente de determinação R^2. Qual o valor do coeficiente de correlação entre as variáveis transformadas?

(b) Deduza a relação não linear entre o número médio de frutos à colheita e o número médio de frutos contados em Setembro que corresponde à regressão linear ajustada. Escreva a equação da curva ajustada.

(c) Com o ajustamento deste modelo obtiveram-se os gráficos dos residuos apresentados seguidamente. Considera que o técnico fez uma boa opção? Justifique sucintamente a sua resposta.



II [7 valores]

1. No estudo sobre a variedade de macieira "**Bravo** de Esmolfe"anteriormente descrito foram estudadas no mesmo local as 3 regiões de origem das árvores plantadas nas parcelas, designadas `TM`, `BI` e `BL`. No ensaio realizado, foram aleatoriamente associadas 22 parcelas a cada região de origem e avaliou-se o número médio de frutos à colheita em cada parcela. A média amostral de todas as observações e as médias amostrais por região de origem são indicadas seguidadamente:

```
> model.tables(macieirabravo.aov, type="means")
Tables of means
Grand mean

3.727788

 origem
origem
   BI    BL    TM
2.468 4.248 4.467
```

   (a) Descreva o delineamento experimental adotado para o ensaio e, em pormenor, o modelo ANOVA adequado ao problema sob estudo.

   (b) Seguidamente apresenta-se a tabela ANOVA relativa ao estudo efectuado:

```
> summary(macieirabravo.aov)
            Df Sum Sq  Mean Sq F value  Pr(>F)
origem       2   A     26.441    C     0.00258
Residuals    B 253.83   4.029
```

   Complete a tabela, indicando como obtém cada um dos valores em falta A, B e C.

   (c) Pode afirmar-se que o número médio de frutos à colheita é igual para todas as regiões de origem? Formalize e efectue o teste F adequado ao problema, ao nível $\alpha = 0.05$.

   (d) Quais as estimativas dos parâmetros do modelo?

   (e) O número médio de frutos à colheita em parcelas com origem Beira Litoral é igual ao número médio de frutos à colheita em parcelas com origem Beira Interior? Justifique a sua resposta com o teste de Tukey ($\alpha = 0.05$)

   (f) Prove que, no contexto da ANOVA descrita neste exercício, o resíduo da observação Y_{ij} é dado pela sua diferença em relação à média amostral de nível

$$E_{ij} \;=\; Y_{ij} - \overline{Y}_{i\cdot}$$

   .

   (g) Prove que no contexto da ANOVA descrita neste exercício, o Quadrado Médio Residual é a média (simples) das k variâncias de nível S_i^2 da variável resposta Y:

$$QMRE \;=\; \frac{1}{k}\sum_{i=1}^{k} S_i^2$$

2. O técnico que realizou este estudo não ficou satisfeito quando analisou os gráficos dos resíduos para validação dos pressupostos do modelo. Dada a natureza dos dados, adoptou de seguida uma abordagem não paramétrica. A soma das ordenações do número médio de frutos à colheita das 22 parcelas avaliadas para cada uma das 3 regiões de origem (R_i) foi a seguinte: $R_{BI} = 474$, $R_{BL} = 854$, $R_{TM} = 883$.
   Descreva em pormenor o teste não paramétrico que neste contexto deverá ser realizado. O que conclui sobre o estudo efectuado ao nível $\alpha = 0.05$?
   ($\chi^2_{0.05(1)} = 3.841$; $\chi^2_{0.05(2)} = 5.991$; $\chi^2_{0.05(3)} = 7.815$; $\chi^2_{0.05(4)} = 9.488$).

3. Ainda neste estudo, pretendeu-se determinar o conteúdo de vitamina C na polpa da maçã. Para tal, ao acaso, separam-se 2 frutos de cada uma das 22 parcelas de cada região de origem, tendo sido levados para o laboratório.

   (a) Indique as unidades experimentais do ensaio.

   (b) Esta experiência tem pseudo-repetições?

   (c) Quantas repetições existem por região de origem?