INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2019-20

13 de Janeiro de 2020 Primeira Chamada de EXAME Duração: 3h30

I [2,5 valores]

Há interesse em estudar se, para uma determinada variedade de alho, é possível admitir que o número de dentes por bolbo segue uma distribuição de Poisson. Recolheram-se ao acaso 348 bolbos e, para cada um, foi registado o número de dentes. Os resultados obtidos são indicados na seguinte tabela.

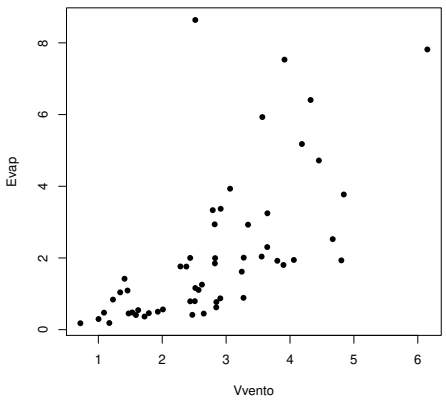
No. dentes	≤ 5	6	7	8	9	10	11	12	13	14	15	16	17	18	≥ 19
No. bolbos	7	4	10	24	28	63	63	47	43	28	9	5	8	4	5

1. Efectue um teste à hipótese de que o número de dentes por bolbo segue uma distribuição Poisson, com parâmetro $\lambda = 11$, sabendo que o valor calculado da estatística na tabela acima é 59.454 e admitindo válida a distribuição assintótica da estatística do teste.
2. Estude a validade da distribuição assintótica da estatística de teste.
3. Calcule a parcela da estatística do teste correspondente à categoria ≤ 5 dentes por bolbo.

II [8,5 valores]

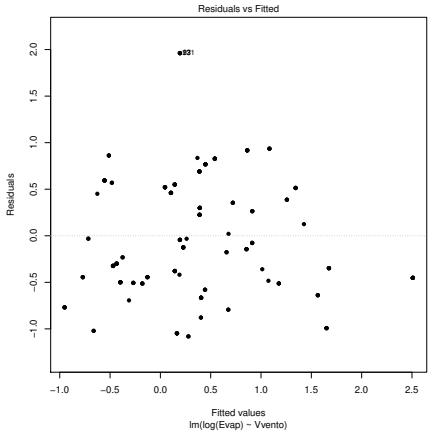
Um estudo sobre o balanço hídrico em oliveiras envolveu medições de várias variáveis meteorológicas durante 55 chuvadas: precipitação (`Pg`, em mm); duração da chuvada (`dur`, em horas); intensidade média da precipitação (`int`, em mm h^{-1}); taxa máxima de evaporação (`Evap`, em mm h^{-1}); e velocidade média do vento (`Vvento`, em m s^{-1}). Para cada chuvada, foram escolhidas algumas árvores, de forma aleatória e independente, e mediu-se o escorrimento médio ao longo do tronco de árvores (`Stemflow`, em mm), bem como características morfológicas médias das árvores: altura total (`alt`, em m); altura do tronco (`alt.tronco`, em m); perímetro do tronco (`P.tronco`, em m); área da copa (`Area.copa`, em m^2).

1. Um primeiro modelo visa relacionar a taxa máxima de evaporação (`Evap`) com a velocidade do vento (`Vvento`).
   (a) Eis a nuvem de pontos de `Evap` sobre `Vvento`. Indique duas características da nuvem de pontos que sugerem ser preferível usar transformações de variáveis antes de ajustar uma regressão linear.



(b) Foi ajustada uma regressão linear simples dos logaritmos de `Evap` sobre `Vvento`, com os resultados indicados em baixo à esquerda. À direita encontra-se o respectivo gráfico dos resíduos usuais *vs.* valores ajustados da variável resposta.

```
Call: lm(formula = log(Evap) ~ Vvento, data = StemFV)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.46881    0.22851  -6.428 3.75e-08
Vvento       0.63961    0.07607   8.409 2.50e-11
---
Residual standard error: 0.6406 on ??? degrees freedom
Multiple R-squared: 0.5716, Adjusted R-squared: 0.5635
F-statistic: ??? on ??? and ??? DF, p-value: 2.503e-11
```



i. Discuta, com base no gráfico, se a transformação usada permitiu ultrapassar os dois problemas que indicou na alínea anterior.

ii. Discuta a qualidade de ajustamento do modelo e, em particular, efectue o teste de ajustamento global. Comente.

iii. A observação $i=51$ é a que tem o maior resíduo, de valor $e_{51}=2.01484$. A velocidade do vento registada nessa observação foi 2.5175 $m\,s^{-1}$. Diga qual a taxa máxima de evaporação correspondente.

iv. Qual a equação da curva resultante do modelo ajustado, na nuvem de pontos nas escalas originais das variáveis. Qual o valor estimado da taxa de variação relativa da variável `Evap`?

2. Seguidamente, modelou-se o logaritmo natural do escorrimento ao longo do tronco (`log(Stemflow)`) à custa do logaritmo natural da precipitação (`log(Pg)`) e das restantes variáveis (não transformadas) disponíveis. Eis os resultados obtidos:

```
Call: lm(formula = log(Stemflow) ~ log(Pg) + int + dur + Evap + Vvento +
    alt.arv + alt.tronco + P.tronco + Area.Copa, data = StemFV)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -12.79092    2.34568  -5.453 2.01e-06
log(Pg)       1.89779    0.25387   7.475 2.02e-09
int          -0.02079    0.12293  -0.169 0.866430
dur          -0.01614    0.01738  -0.929 0.358028
Evap         -0.51082    0.12087  -4.226 0.000115
Vvento        0.92802    0.22048   4.209 0.000121
alt.arv       2.51834    1.22131   2.062 0.045011
alt.tronco    0.94257    2.02048   0.467 0.643102
P.tronco     -4.50441    1.01435  -4.441 5.77e-05
Area.Copa    -0.38015    0.27904  -1.362 0.179870
---
Residual standard error: 1.195 on 45 degrees of freedom
Multiple R-squared: 0.8429, Adjusted R-squared: 0.8115
F-statistic: 26.83 on 9 and 45 DF,  p-value: 2.991e-15
```

(a) Calcule e interprete o intervalo a 95% de confiança para a variação esperada na variável resposta, por cada metro quadrado adicional na área da copa, mantendo fixas as restantes variáveis. Será que se pode afirmar que esta variável preditora é dispensável no modelo?

(b) Comente a seguinte afirmação: "*sendo tudo o resto igual, são diferentes os aumentos médios que ocorrem no logaritmo do escorrimento ao longo do tronco por cada metro adicional na altura da árvore, ou por cada metro adicional na altura do tronco*". Baseie a sua resposta num teste de hipóteses adequado, sabendo que a estimativa da covariância entre os estimadores dos parâmetros destas duas variáveis é -1.537745.

(c) Com base num algoritmo de exclusão sequêncial, diga se é possível simplificar este modelo sem perda significativa de qualidade. Independentemente da sua resposta, qual seria o melhor valor de R^2 que se pode obter num submodelo resultante da exclusão dum único preditor?

(d) Foi ajustado um submodelo, resultante de excluir quatro preditores (`int`, `dur`, `alt.tronco` e `Area.Copa`), cujo coeficiente de determinação foi $R^2 = 0.8127$. Este valor deve ser considerado significativamente inferior ao do modelo com 9 preditores, inicialmente ajustado?

(e) Calcule o valor do coeficiente de determinação modificado para o submodelo considerado na alínea anterior. Comente.

III [5 valores]

Num estudo envolvendo genótipos da casta Alvarinho, em Monção, foi estudo o rendimento (variável `rend`, em kg/planta) de 6 genótipos (ou clones), em 4 ambientes. Em cada ambiente foram aleatoriamente associadas 9 parcelas a cada um dos genótipos. Em baixo indicam-se os rendimentos médios obtidos em cada situação experimental. A média e a variância amostrais da totalidade dos rendimentos foram, respectivamente, 3.505421 kg/planta e 8.756217 $(\text{kg/planta})^2$.

	clone					
ambiente	AI1	AI2	AI3	AI4	AI5	AI6
Amb1	1.292	1.776	1.496	1.871	1.614	1.334
Amb2	6.319	6.001	7.567	8.055	5.823	6.158
Amb3	5.303	7.031	4.341	7.026	3.879	4.098
Amb4	0.518	0.647	0.478	0.592	0.449	0.463

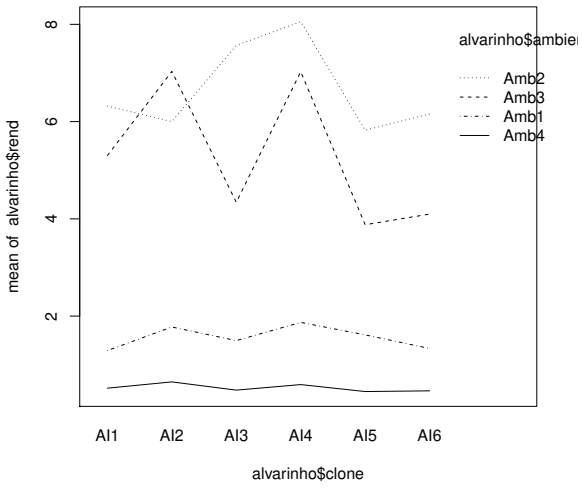
1. Diga, justificando, qual o delineamento experimental utilizado e descreva pormenorizadamente o modelo ANOVA mais adequado.

2. Construa a tabela de síntese da ANOVA que indicou, sabendo que a estimativa da variância dos erros aleatórios é 1.8728; que a Soma de Quadrados associada aos efeitos de genótipo é 53.5437; e que o valor calculado da estatística do teste aos efeitos de ambiente é 247.150.

3. Que tipos de efeitos podem ser considerados significativos? Descreva em pormenor um dos testes e de forma mais sucinta o(s) restante(s).

4. Pode afirmar-se que os rendimentos obtidos no Ambiente 2 com qualquer dos genótipos ensaiados são significativamente diferentes dos rendimentos obtidos no Ambiente 1, com qualquer desses genótipos? Justifique formalmente.

5. Descreva e comente o gráfico seguinte à luz da informação disponível.



IV [4 valores]

1. Considere uma regressão linear simples, ajustada com base em n pares de observações $\{(x_i, y_i)\}_{i=1}^{n}$, e admita válido o respectivo Modelo. Deduza a distribuição do estimador do declive β_1 da recta populacional. Pode admitir conhecido que $\sum_{i=1}^{n} c_i\, x_i = 1$.

2. Considere uma regressão linear múltipla com p preditores e ajustada com base em n observações.

   (a) Descreva em pormenor o modelo de regressão linear múltipla, usando notação matricial/vectorial.

   (b) Descreva o triângulo rectângulo no espaço das variáveis, $\mathbb{R}^n$, cujos lados estão directamente associados à Fórmula Fundamental da Regressão.

   (c) Mostre que o vector $\vec{\boldsymbol{\epsilon}}$ dos erros aleatórios do modelo e o vector dos resíduos, $\vec{\mathbf{E}} = \vec{\mathbf{Y}} - \vec{\hat{\mathbf{Y}}}$, estão relacionados pela seguinte equação: $\vec{\mathbf{E}} = (\mathbf{I} - \mathbf{H})\,\vec{\boldsymbol{\epsilon}}$. Utilize esta equação para mostrar que a Soma de Quadrados Residual não pode exceder a Soma de Quadrados dos erros aleatórios, $\vec{\boldsymbol{\epsilon}}^{\,t}\vec{\boldsymbol{\epsilon}}$.

   (d) Utilize a equação da alínea anterior para deduzir a distribuição de probabilidades do vector dos resíduos, $\vec{\mathbf{E}}$, ao abrigo do modelo.

   (e) Sabendo que se verifica também $\vec{\mathbf{E}} = (\mathbf{I} - \mathbf{H})\vec{\mathbf{Y}}$, utilize a equação da alínea 2c para mostrar que em produtos de matrizes e vectores *não* se verifica a lei do anulamento do produto.