INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2014/15

28 de Janeiro de 2015 SEGUNDA CHAMADA de EXAME Duração: 3h30

I [2,5 valores]

Num estudo sobre colónias bacterianas em placas de Petri, pretende-se saber se é admissível considerar que as contagens do número de colónias por área de $1\,cm^2$ segue uma distribuição de Poisson, a que corresponde uma distribuição espacial aleatória. Foi utilizada uma placa com área total de $100\,cm^2$, considerando-se existir uma grelha de 100 quadrados contíguos de $1\,cm^2$ cada. Em cada quadrado da grelha foi contado o número de colónias de bactérias, tendo sido obtidos os seguintes resultados.

No. de colónias	0	1	2	3	4	5	6	7	8	9
No. de quadrados	27	50	11	6	3	1	1	1	0	0

1. Qual o número médio de colónias por quadrado de $1\,cm^2$?
2. Efectue um teste baseado na estatística de Pearson para testar a hipótese de o número de colónias por quadrado da grelha seguir uma distribuição de Poisson, para os dados em questão. Considere um nível de signficância $\alpha = 0.05$. Discuta as suas conclusões.

II [9 valores]

Num estudo sobre resolução espacial de imagens de satélite, determinou-se a resolução espacial das imagens (variável resposta `qres`, em metros) em função do ângulo zenital de observação (variável `zen`, em graus), do ângulo azimutal (variável `az`, em graus) sobre um alvo no terreno com orientações entre 55 e 80 graus (variável `pos`), em datas distintas, tendo-se obtido um total de 372 observações. Eis alguns indicadores relativos a cada variável, bem como a respectiva matriz de correlações.

```
    qres            zen             az               pos
 Min. : 69.46   Min. :   2.809  Min. :-108.97    Min. :55.0
 Mean :370.43   Mean : 30.379   Mean :   10.76   Mean :67.5
 Max. :781.88   Max. : 55.598   Max. :   82.23   Max. :80.0
 Var. :    ??   Var. :239.04292 Var. :6920.05009 Var. :73.11321

        qres    zen       az    pos
qres  1.0000 0.8173 -0.0136 0.1715
zen   0.8173 1.0000  0.0820 0.0000
az   -0.0136 0.0820  1.0000 0.0000
pos   0.1715 0.0000  0.0000 1.0000
```

1. O ajustamento duma regressão linear múltipla aos dados originou os seguintes resultados:

```
lm(formula = qres ~ zen + az + pos, data = dados)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.87563   29.61571  -0.232 0.81654
zen          6.78759    0.23450  28.945 < 2e-16
az          -0.12423    0.04358  -2.850 0.00461
pos          2.55470    0.42259   6.045 3.66e-09
```

```
---
Residual standard error: 69.6 on 368 degrees of freedom
Multiple R-squared: 0.7038, Adjusted R-squared: 0.7014
F-statistic: 291.5 on 3 and 368 DF, p-value: < 2.2e-16
```

(a) Interprete o valor do coeficiente de determinação obtido para o ajustamento.

(b) Estime o valor esperado da resolução das imagens quando o ângulo zenital de observação é 20 graus, o ângulo azimutal é -100 e o alvo está na posição correspondente a 60 graus. Indique as unidades na sua resposta.

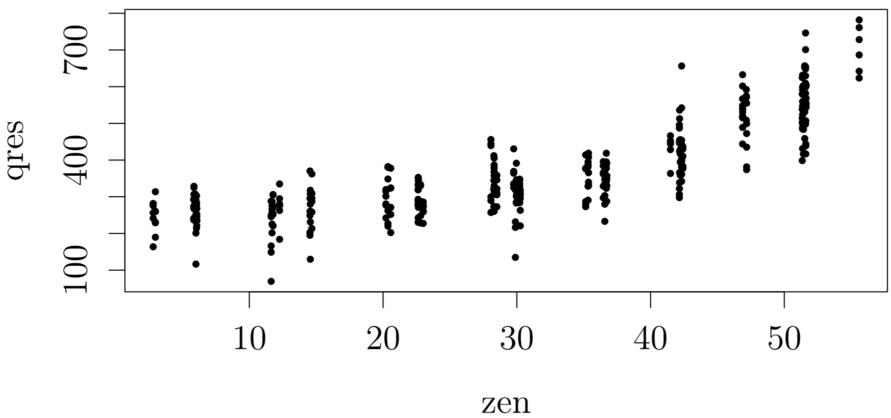
(c) Mostre que a variância amostral das 372 observações da variável resposta é aproximadamente 16223.

(d) Teste ($\alpha = 0.05$) se o valor de R^2 deste modelo é significativamente diferente do que poderia obter com o melhor modelo de regressão linear simples para estes dados. Comente.

(e) Independentemente da sua resposta na alínea anterior, calcule a recta de regressão linear simples de `qres` sobre o preditor `zen`, indicando como obtém essa equação.

(f) O declive da recta de regressão obtida na alínea anterior é significativamente diferente de zero? Justifique formalmente, através dum teste de hipóteses adequado, com nível de significância 0.05.

2. Um investigador que usa um produto do mesmo satélite para o qual apenas está disponível a informação sobre o ângulo de observação zenital `zen` pretende um modelo para a resolução espacial que dependa unicamente dessa variável. A nuvem de pontos relacionando `qres` e `zen` para as 372 observações é dada na seguinte figura:



Dada a relação observada acima foi ajustado um polinómio de terceiro grau aos dados, tendo-se obtido os seguintes resultados:

```
Call: lm(formula = qres ~ zen + I(zen^2) + I(zen^3), data = dados)

               Estimate Std. Error t value Pr(>|t|)
(Intercept)  240.759993  16.430240  14.653 < 2e-16
zen            3.426980   2.464975   1.390 0.165287
I(zen^2)      -0.144179   0.098816  -1.459 0.145399
```

```
I(zen^3)       0.003929   0.001131   3.475 0.000572
---
Residual standard error: 58.71 on 368 degrees of freedom
Multiple R-squared: 0.7893, Adjusted R-squared: 0.7875
F-statistic: 459.4 on 3 and 368 DF, p-value: < 2.2e-16
```

(a) É possível comparar os valores de R^2 neste modelo polinomial e no modelo de regressão linear múltipla ajustado inicialmente? Justifique.

(b) Sabendo que no ajustamento deste modelo polinomial a uma dada observação correspondem valores $\texttt{qres}_i$=656.2878 e $\texttt{zen}_i$=42.31931, um resíduo $e_i = 230.9676$ e um valor de efeito alavanca 0.008574429, mostre que:

  i. o resíduo (internamente) estandardizado dessa observação é aproximadamente 4;

  ii. o valor da distância de Cook associado a essa observação é aproximadamente 0.035.

(c) Comente os valores da alínea anterior relativamente ao efeito dessa observação no ajustamento.

(d) Seria preferível um modelo quadrático? Justifique brevemente a sua resposta.

III [4,5 valores]

Após uma fase de selecção inicial de genótipos da casta Alvarinho, em que se procurou identificar alguns genótipos que assegurassem bons rendimentos em diferentes situações ambientais (a chamada estabilidade ambiental), decidiu-se proceder a um estudo final em Monção para comparar os rendimentos (variável `rend`, em kg/planta) de quatro genótipos, em seis diferentes anos (de 1994 a 1999). Eis as médias relativas a este estudo:

```
Tables of means
Grand mean       ano                                   genotipo
4.534083         A1994 A1995 A1996 A1997 A1998 A1999   AI0122 AI1011 AI1025 AI1050
                 0.766 3.718 4.610 5.382 2.911 9.818    3.836  4.734  4.884  4.682
 ano:genotipo
      genotipo
ano     AI0122 AI1011 AI1025 AI1050
  A1994  0.413  0.798  0.965  0.887
  A1995  3.207  3.340  4.682  3.645
  A1996  3.911  4.888  4.543  5.095
  A1997  4.964  5.283  5.784  5.495
  A1998  2.286  3.111  3.140  3.109
  A1999  8.237 10.987 10.190  9.857
```

Foi ajustado um modelo ANOVA, tendo sido obtidos os seguintes resultados:

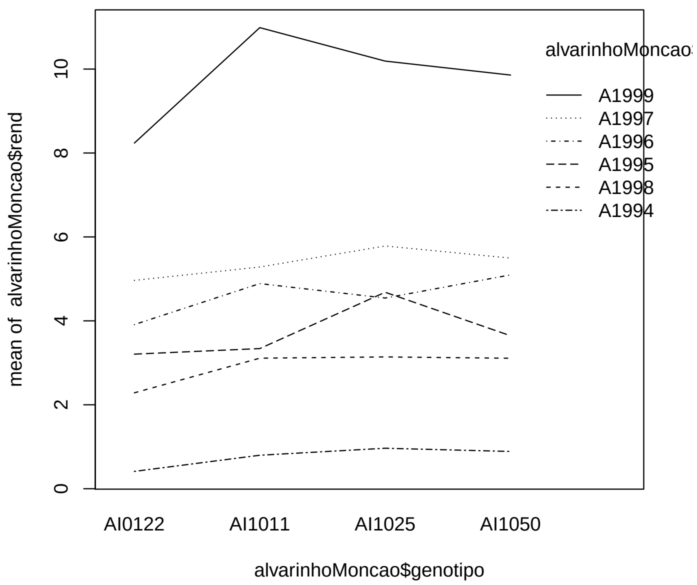
```
              Df Sum Sq Mean Sq F value  Pr(>F)
ano            5 1661.1   332.2 132.007 < 2e-16
genotipo       3   36.2    12.1   4.800 0.00301
ano:genotipo  15   28.7     1.9   0.759 0.72116
Residuals    192  483.2     2.5
```

1. Identifique o tipo de delineamento experimental utilizado. Sabendo que o delineamento era equilibrado, indique, justificando, quantas repetições foram observadas em cada situação experimental.

2. Descreva detalhadamente o modelo ANOVA utilizado.

3. Identifique o(s) tipo(s) de efeitos previsto(s) no modelo que considera significativos. Caso exista mais do que um tipo de efeitos no modelo, justifique em pormenor a sua conclusão para um tipo de efeitos, e de forma mais resumida no(s) restante(s).

4. Discuta o seguinte gráfico, tendo em conta toda a informação já disponível.



5. Descreva as consequências de se ignorar, neste estudo, a existência de datas diferentes de observação. Em particular, re-calcule a tabela-resumo da ANOVA para este novo contexto e diga qual seria a sua conclusão sobre a existência de efeitos de genótipos. Comente.

IV [4 valores]

Considere um modelo de regressão linear múltipla com p variáveis preditoras e ajustado com base em n conjuntos de observações da variável resposta Y e dos p preditores (X_1 a X_p). Admitindo conhecidas as propriedades de vectores esperados e matrizes de variâncias-covariâncias de vectores aleatórios, bem como as propriedades da distribuição Multinormal, responda às seguintes questões.

1. Descreva de forma pormenorizada o modelo de regressão linear múltipla utilizando a notação matricial/vectorial.

2. Deduza, a partir do modelo referido na alínea anterior, que o vector das n observações da variável resposta tem distribuição $\mathbf{Y} \cap \mathcal{N}(\,\mathbf{X}\boldsymbol{\beta}\,,\,\sigma^2\mathbf{I}_n\,)$, sendo $\mathbf{I}_n$ a matriz identidade $n \times n$.

3. Sabendo que o vector $\hat{\boldsymbol{\beta}}$ dos estimadores do modelo é dado por $\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$, mostre que:

   (a) $\hat{\boldsymbol{\beta}}$ é um vector de estimadores centrados;

   (b) a matriz de variâncias-covariâncias do vector de estimadores é dada por $V[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$.