INSTITUTO SUPERIOR DE AGRONOMIA

Modelos Matemáticos e Aplicações (2019-20)

27 de Abril, 2020 — **Teste – Modelo Linear** — Duração: 2h30

I [17 valores]

Um estudo da Secção de Produção Animal do ISA incidiu sobre a digestibilidade de alimentos em leitões. A eficiência do processo digestivo é medida através de Coeficientes de Utilização Digestiva (CUDs). Foram recolhidas 48 observações de CUDs de fibra neutro-detergente (`NDF`), fibra ácido-detergente (`ADF`) e hemi-celuloses (`HC`). Foi usado um delineamento factorial equilibrado, com dois factores de dois níveis cada, que correspondiam a alimentos contendo dois tipos de fibra (factor `Fibra`) e a presença ou ausência de suplementos de enzimas digestivas (factor `Enzimas`) no alimento dos leitões.

Eis alguns indicadores e as correlações entre as variáveis numéricas, calculados para a totalidade das observações:

	NDF	ADF	HC
mínimo	0.15	-0.06	0.30
média	0.52708	0.38188	0.60146
máximo	0.68	0.62	0.76
variância	0.01569770	0.01963684	0.01344676

```
        NDF     ADF      HC
NDF 1.00000 0.85831 0.97445
ADF 0.85831 1.00000 0.72678
HC  0.97445 0.72678 1.00000
```

Eis as médias e variâncias do CUD da fibra neutro-detergente (`NDF`) *em diferentes contextos*:

```
      Fibra                  Enzima                  Fibra:Enzima (médias)  Fibra:Enzima (variâncias)
              1         2          1          2             Enzima                 Enzima
Média    0.6042    0.4500     0.4967     0.5575     Fibra 1      2          Fibra 1          2
Var.  0.00286015 0.0168174  0.0229797 0.00716739        1 0.5975 0.6108          1 0.00443864 0.00144470
                                                         2 0.3958 0.5042          2 0.02142652 0.00733561
```

1. Pretende-se estudar se o CUD de `NDF` difere consoante o tipo de fibra e/ou a presença de suplemento de enzimas digestivos.

   (a) Identifique o modelo ANOVA mais adequado ao estudo pretendido, descrevendo-o em pormenor.

   (b) Indique a natureza da matriz do modelo $\mathbf{X}$ associada ao modelo que indicou para este estudo.

   (c) Construa a tabela-resumo correspondente ao modelo que indicou. Admita que a Soma de Quadrados associada aos efeitos de Fibra é 0.2852.

   (d) Efectue os testes necessários para indicar se os tipos de efeitos previstos no modelo devem ser, ou não, considerados significativos ao nível $\alpha = 0.01$. Comente.
   Nota: Descreva um teste em pormenor, e apenas discuta as conclusões do(s) restante(s).

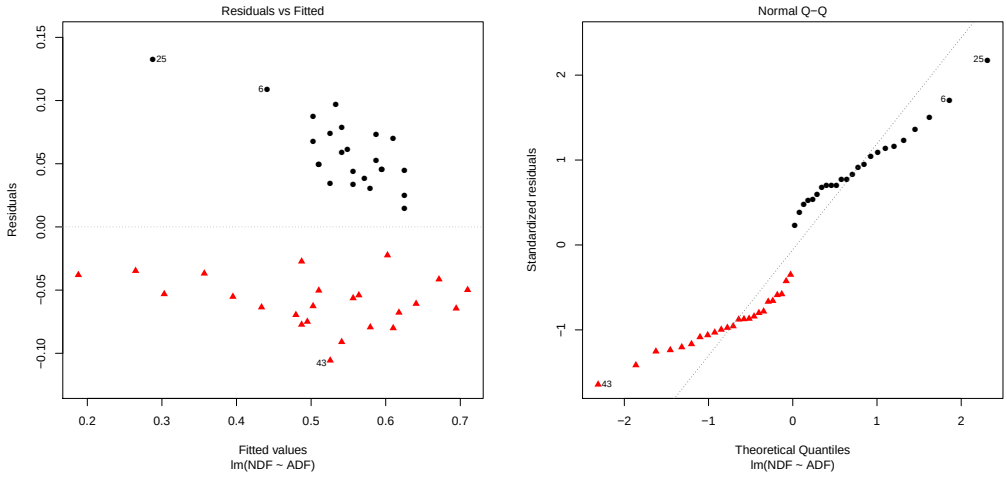
   (e) Foi ajustado com o comando `lm` o modelo linear abaixo indicado. Com base nos testes t indicados na listagem, o que pode dizer sobre a relação entre a média populacional da célula de referência e a média populacional nas restantes situações experimentais?

```
Call:  lm(formula = NDF ~ Fibra * Enzima, data = leitoes)
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.59750    0.02687  22.240  < 2e-16
Fibra2          -0.20167    0.03799  -5.308 3.47e-06
Enzima2          0.01333    0.03799   0.351    0.727
Fibra2:Enzima2   0.09500    0.05373   1.768    0.084
---
Residual standard error: 0.09307 on 44 degrees of freedom
Multiple R-squared: 0.4835,  Adjusted R-squared: 0.4482
F-statistic: 13.73 on 3 and 44 DF,  p-value: 1.867e-06
```

2. Foi agora considerado um modelo de regressão linear simples de `NDF` sobre `ADF`, usando as 48 observações.

(a) Calcule a recta de regressão ajustada. Determine e comente o respectivo coeficiente de determinação.

(b) Eis dois graficos de resíduos, onde foram usados símbolos/cores diferentes para indicar os resíduos associados a cada tipo de fibra. Comente os gráficos.



3. Foi seguidamente ajustado um modelo ANCOVA, estudando a relação linear entre a variável resposta `NDF` e o preditor numérico `ADF`, mas admitindo a possibilidade de serem necessárias rectas de regressão diferentes para os dois tipos de fibra usados nos alimentos (foi ignorada a utilização de enzimas digestivos, tratando-se todas as observações como repetições). Eis os resultados:

```
Call:  lm(formula = NDF ~ ADF * Fibra, data = leitoes)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.38537    0.01772  21.749  < 2e-16
ADF          0.53968    0.04264  12.658 2.89e-16
Fibra2      -0.19954    0.01991 -10.021 6.26e-13
ADF:Fibra2   0.19752    0.04839   4.082 0.000185
---
Residual standard error: 0.0191 on 44 degrees of freedom
Multiple R-squared: 0.9783,  Adjusted R-squared: 0.9768
F-statistic: 659.8 on 3 and 44 DF,  p-value: < 2.2e-16
```

(a) Escreva as equações das rectas de regressão de `NDF` sobre `ADF` para cada tipo de fibra. Com base na informação apresentada, considera que há justificação para usar rectas distintas nas duas situações?

(b) Calcule e interprete o intervalo a 95% de confiança para o declive da recta relativa à fibra 1.

(c) Utilize um teste F para determinar se o modelo de recta única entre `NDF` e `ADF`, calculado no ponto anterior, tem um ajustamento significativamente diferente do modelo ANCOVA agora ajustado. Comente.

4. Foi finalmente ajustado um modelo de regressão linear da variável `NDF` sobre dois preditores: `ADF` e `HC`. Eis os resultados:

```
Call: lm(formula = NDF ~ ADF + HC, data = leitoes)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.064533   0.005325  -12.12 9.08e-16
ADF          0.284442   0.010068   28.25  < 2e-16
HC           0.803040   0.012167   66.00  < 2e-16
---
Residual standard error: 0.006644 on 45 degrees of freedom
Multiple R-squared: 0.9973,  Adjusted R-squared: 0.9972
F-statistic:  8335 on 2 and 45 DF,  p-value: < 2.2e-16
```

(a) Usando a representação alternativa dos dados no espaço das variáveis ($\mathbb{R}^n$), descreva e comente o significado do valor $R^2 = 0.9973$.

(b) Como se pode explicar o elevadíssimo valor do coeficiente de determinação deste modelo, se não tem como preditor o factor `Fibra` que se sabe ter efeitos significativos sobre o CUD de `NDF`?

(c) Quer-se comparar o modelo deste ponto com os modelos ANOVA, ANCOVA e de regressão linear simples com recta única ajustados acima. Diga, justificando *e sem fazer qualquer conta*:

i. Com que modelos é que o valor $R^2 = 0.9973$ pode ser comparado?

ii. Quais modelos podem ser comparados com o modelo deste ponto através dum teste F parcial?

iii. Entre o modelo deste ponto e o modelo ANCOVA, qual terá o melhor valor do Critério de Informação de Akaike (AIC)?

II [3 valores]

Considere uma regressão linear múltipla com p preditores e n observações, matriz do modelo $\mathbf{X}$, e equação $\vec{\mathbf{Y}} = \mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\epsilon}}$. Modifique o Modelo Linear, admitindo agora que a distribuição do vector dos erros aleatórios seja $\vec{\boldsymbol{\epsilon}} \frown \mathcal{N}_n(\vec{\mathbf{0}}, \Sigma)$, com uma matriz Σ conhecida (não aleatória).

1. Determine a distribuição de probabilidades do vector $\vec{\mathbf{Y}}$ das observações.

2. Considere um novo vector de estimadores dos parâmetros, dado por $\hat{\vec{\boldsymbol{\beta}}}_* = (\mathbf{X}^t\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^t\Sigma^{-1}\vec{\mathbf{Y}}$. Diga justificando, qual a sua distribuição de probabilidades. Trata-se dum vector de estimadores centrado?

3. Diga, justificando, qual a expressão de um intervalo de confiança para um parâmetro individual β_j.

4. Qual a implicação para o modelo de Σ ser uma matriz *não* diagonal?