INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2020-21

11 Janeiro 2021 **Segundo TESTE** **Uma resolução possível**

I

1. O primeiro modelo de regressão linear múltipla.

   (a) Sabemos que os graus de liberdade da estatística do teste F global são p (número de variáveis preditoras) e $n-(p+1)$. Na listagem do enunciado constata-se que o número de variáveis preditoras é $p=21$. Há $p+1=22$ parâmetros β_j no modelo (um para cada preditor e ainda o β_0).

   (b) Eis o teste de ajustamento global:

   Hipóteses: $H_0 : \mathcal{R}^2 = 0 \quad vs. \quad H_1 : \mathcal{R}^2 > 0.$

   Estatística do Teste: $F = \frac{QMR}{QMRE} = \frac{n-(p+1)}{p} \frac{R^2}{1-R^2} \underset{H_0}{\frown} F_{(p,n-(p+1))}$, sob H_0.

   Nível de significância: $\alpha = 0.05$.

   Região Crítica: Unilateral direita. Rejeitar H_0 se $F_{calc} > f_{0.05[21,25]}$. Vamos usar o valor tabelado mais próximo, $f_{0.05[20,25]}=2.01$.

   Conclusões: No enunciado não é dado o valor calculado da estatística, mas como vimos, $F_{calc} = \frac{n-(p+1)}{p} \cdot \frac{R^2}{1-R^2} = \frac{25}{21} \times \frac{0.586}{1-0.586} = 1.68507 < 2.01$. Logo, não se rejeita H_0: o modelo ajustado não difere significativamente do Modelo Nulo, pelo que não haverá sustentação para recomendar a sua utilização. Este resultado não é surpreendente. O valor do coeficiente de determinação usual é baixo ($R^2=0.586$), mas mesmo esse valor foi obtido à custa dum número muito elevado de variáveis preditoras ($p=21$) face ao número de observações ($n=47$), facto que se traduz num valor bastante inferior (e muito baixo) do coeficiente de determinação modificado ($R^2_{mod}=0.2382$). Esta grande diferença entre as duas variantes do coeficiente de determinação vai gerar um valor bastante baixo da estatística F.

   Recordar: Como se viu no Exercício RLM 22, verifica-se sempre $F_{calc} = \frac{R^2}{R^2-R^2_{mod}}$.

2. Modelo RLM com os mesmo 21 preditores, mas com a logaritmização da variável resposta.

   (a) No novo modelo, o teste F de ajustamento global rejeita a Hipótese Nula, uma vez que o valor de prova (*p-value*) $p = 0.004255$ é inferior aos níveis de significância com que usualmente se trabalha. Assim, neste caso podemos falar dum modelo significativamente diferente do modelo Nulo (o modelo em que as log-produtividades não são explicadas por qualquer variável preditora). O valor $R^2 = 0.7199$ significa que este modelo explica quase 72% da variabilidade nas *log*-produtividades, e não nas produtividades como refere erradamente a frase do enunciado. Aliás, este valor de R^2 *não* é directamente comparável ao valor do modelo do ponto 1, uma vez que se referem a proporções de variabilidade explicada de *quantidades diferentes*.

   (b) Tem-se agora um submodelo com apenas $k = 9$ variáveis preditoras, de entre as $p = 21$ usadas no modelo completo.

      i. O valor do coeficiente de determinação de todo e qualquer *submodelo* nunca pode ser maior que o R^2 usual do modelo completo, e por isso a diminuição do valor de R^2 seria expectável. No entanto, esse tipo de relação de ordem não tem de se verificar no que respeita ao coeficiente de determinação modificado, como foi salientado nas aulas. O R^2_{mod} pode ser maior no submodelo quando a perda de capacidade explicativa (isto é,

a redução no R^2 usual) do submodelo é compensada pela maior parcimónia (redução no número de preditores). No nosso caso, o submodelo tem menos de metade dos preditores do modelo completo ($k = 9$ em vez de $p = 21$), o que afasta bastante o número de preditores em relação ao número de observações n e essa redução mais do que compensa a relativamente pequena perda de capacidade explicativa ($R^2_s = 0.6814$ em vez de $R^2_c = 0.7199$). Concretizando, e tendo em conta a relação vista nas aulas (e no Exercício RLM 22) entre as duas variantes do coeficiente de determinação, tem-se $R^2_{mod} = 1-(1-R^2)\,\frac{n-1}{n-(p+1)}$ (no submodelo $n-(k+1)$). Ora, o factor $\frac{n-1}{n-(p+1)}$ que no R^2 modificado penaliza a proporção da variabilidade *não* explicada pelo modelo, é $46/25 = 1.84$ no caso do modelo completo, mas apenas $46/37 = 1.243243$ no submodelo. Assim, uma proporção de variabilidade inexplicada semelhante (cerca de 30% em ambos os casos) é sujeita a um acréscimo cerca de $1,5$ vezes maior no caso do modelo completo, que fica assim com um R^2_{mod} menor.

ii. Eis o teste F parcial comparando o ajustamento do modelo e do submodelo:

Hipóteses: $H_0 : \mathcal{R}^2_c = \mathcal{R}^2_s$ vs. $H_1 : \mathcal{R}^2_c > \mathcal{R}^2_s$

Estatística do teste $F = \frac{n-(p+1)}{p-k} \cdot \frac{R^2_c - R^2_s}{1-R^2_c} \;\frown\; F_{[p-k,n-(p+1)]}$, se H_0 verdade.

Nível de significância: $\alpha = 0.05$

Região Crítica: Unilateral direita. Rejeita-se H_0 se

$F_{calc} > f_{\alpha[p-k,n-(p+1)]} = f_{0.05(12,25)} = 2.16$.

Conclusões: Tem-se $F_{calc} = \frac{25}{12} \times \frac{0.7199-0.6814}{1-0.7199} = 0.2863561$. Assim, não se rejeita H_0. Não se pode concluir que os dois modelos tenham qualidade de ajustamento significativamente diferente.

iii. O gráfico do enunciado tem, no eixo horizontal, os valores do efeito alavanca (h_{ii}) e, no eixo vertical, valores de resíduos (internamente) estandardizados, R_i. Da observação do gráfico constata-se que os resíduos estandardizados estão quase todos no intervalo $[-3, 3]$, não havendo razões para falar em observações muito distanciadas do hiperplano ajustado pela regressão. Mesmo a observação com o maior (em módulo) resíduo estandardizado, a observação 44, tem um valor de R_i próximo de -3. No que respeita aos efeitos alavanca, há quatro observações com efeitos alavanca superiores, ou muito próximos a 0.5, o que é de assinalar num ajustamento baseado em 47 observações, e para o qual o valor médio é dado por $\overline{h} = \frac{p+1}{n} = \frac{10}{47} = 0.212766$. Como se sabe, os valores do efeito alavanca estão necessariamente compreendidos entre $\frac{1}{n} = 0.0212766$ e 1, com valores grandes a indiciar que a hipersuperficie ajustada está 'forçada' a passar próximo do ponto.

De acordo com o enunciado, a maior distância de Cook não corresponde a uma das quatro observações com maior efeito alavanca. A distância de Cook é uma medida da influência duma observação, ou seja, do impacte que teria a exclusão dessa observação sobre o hiperplano ajustado. Tendo em conta a relação (dada no formulário) entre os valores indicados nos eixos do gráfico e as distâncias de Cook, ou seja $D_i = R^2_i \cdot \frac{h_{ii}}{1-h_{ii}} \cdot \frac{1}{p+1}$, é possível calcular um valor aproximado dessa maior das distâncias de Cook. Verifica-se a partir do gráfico que para a observação 44 o resíduo estandardizado é $R_{44} \approx -3$, enquanto que o seu efeito alavanca é aproximadamente $h_{44,44} \approx 0.25$. Assim, $D_{44} \approx (-3)^2 \cdot \frac{0.25}{1-0.25} \cdot \frac{1}{10} = 0.3$. Trata-se dum valor relativamente elevado, mas ainda aquém do limiar 0.5. Tratando-se do maior valor de distância de Cook, pode afirmar-se que nenhuma observação é excessivamente influente, ou seja, que a exclusão duma única observação não modificaria excessivamente o hiperplano ajustado. **Nota:** O valor exacto é $D_{44} = 0.3158$.

3. O enunciado pede um teste para saber se é possível afirmar que $\beta_5 < 0$, sendo β_5 o coeficiente

do preditor NIRPER. De facto, a mesma lógica que conduz a interpretar β_5 como o aumento no valor médio de Y^* associada a aumentar o preditor X_5 em uma unidade leva à conclusão que aumentar X_5 em $c > 0$ unidades leva a que a diferença entre os correspondentes valores esperados de Y^* (mantendo os outros preditores constantes) seja:

$$\begin{array}{rcl} E[Y^*|X_5 = x_5 + c] & = & \cancel{\beta_0} + \cancel{\beta_1 x_1} + \cancel{\beta_2 x_2} + \cancel{\beta_3 x_3} + \cancel{\beta_4 x_4} + \overbrace{\beta_5 (x_5 + c)}^{=\cancel{\beta_5 x_5} + c\beta_5} \\ - \quad E[Y^*|X_5 = x_5] & = & \cancel{\beta_0} + \cancel{\beta_1 x_1} + \cancel{\beta_2 x_2} + \cancel{\beta_3 x_3} + \cancel{\beta_4 x_4} + \cancel{\beta_5 x_5} \\ \hline & = & c\,\beta_5 \end{array}$$

Logo, haver uma diminuição em $E[Y^*]$ para qualquer aumento em X_5 (com o resto fixo) corresponde a $\beta_5 < 0$ (já que $c > 0$). Vamos formular as hipoteses como $H_0 : \beta_5 \geq 0$ e $H_1 : \beta_5 < 0$. **Nota:** A estimativa amostral de β_5 é negativa ($b_5 = -8.275832$), pelo que colocar como Hipótese Nula que β_5 seja negativa nunca poderia resultar numa rejeição. Ao colocar essa hipótese em H_1, estamos a exigir-lhe o ónus da prova, ou seja, estamos a perguntar se essa estimativa é 'suficientemente negativa' para poder concluir que o valor populacional β_5 é negativo.

Vejamos os restantes passos nesse teste de hipóteses.

Estatística do Teste: $T = \frac{\hat{\beta}_5 - \overbrace{\beta_{5|H_0}}^{=0}}{\hat{\sigma}_{\hat{\beta}_5}} \frown t_{n-(p+1)}$, sob H_0.

Nível de significância: $\alpha = \mathrm{P}[$ Erro do tipo I $] = \mathrm{P}[$ Rej. $H_0 \mid H_0$ verdade $] = 0.01$.

Região Crítica: (Unilateral esquerda) Rejeitar H_0 se $T_{calc} < -t_{0.01(41)} \approx -2.42$ (considerando os valores tabelados $t_{0.01(40)} = 2.42326$ e $t_{0.01(50)} = 2.40327$).

Conclusões: O valor da estatística de teste consta da listagem no enunciado, uma vez que o valor charneira entre H_0 e H_1 é 0, mas o valor de prova (*p-value*) ao lado *não* é utilizável, já que diz respeito a um teste de região crítica bilateral, correspondente a $H_1 : \beta_5 \neq 0$. Tem-se $T_{calc} = \frac{b_5 - 0}{\hat{\sigma}_{\hat{\beta}_5}} = -2.988 < -2.42$. Rejeita-se H_0 (ao nível $\alpha = 0.01$) e conclui-se por $H_1 : \beta_5 < 0$, ou seja, pode sustentar-se a afirmação feita no enunciado (ao nível $\alpha = 0.01$).

II

1. A variável resposta rendimento pode ser explicada por dois factores, o factor localidade (com $a = 3$ níveis) e o factor terreno. Uma vez que os terrenos são todos diferentes entre si, não existe possibilidade de relacionar cada terreno dum local com um dado terreno de outro local. Assim, o delineamento é de natureza *hierarquizada*, com o factor terreno subordinado ao factor local (a correcta identificação dum terreno exige a identificação prévia dum local). Em cada local existem os mesmos $b_i = 8$ $(i = 1, 2, 3)$ níveis do factor terreno, para um total de $\sum_{i=1}^{3} b_i = 24$ terrenos, ou seja, situações experimentais. O delineamento é equilibrado porque em cada um desses 24 terrenos existe igual quantidade de observações: $n_{ij} = n_c = 7$, para um total de $n = 24 \times 7 = 168$ observações. Eis o modelo ANOVA correspondente a este delineamento:

   Equação do Modelo: $Y_{ijk} = \mu_{11} + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}$, onde $i = 1, 2, 3$ indica localidade; para cada localidade i, $j = 1, ..., 8$ indica terrenos; $k = 1, ..., 7$ repetição (dentro da situação experimental); Y_{ijk} indica o rendimento da k-ésima repetição no terreno j da localidade i; ϵ_{ijk} é o correspondente erro aleatório. Com as restrições $\alpha_1 = 0$ e $\beta_{1(i)} = 0$ para qualquer i, a constante aditiva comum a todas as observações, μ_{11}, representa o rendimento médio

populacional no primeiro terreno do primeiro local (Bombarral, por ordem alfabética); α_i indica o efeito associado ao local i; e $\beta_{j(i)}$ indica o efeito associado ao j-ésimo terreno da localidade i.

Distribuição dos erros: $\epsilon_{ijk} \frown \mathcal{N}(0, \sigma^2)$, para qualquer i, j, k.

Independência dos erros: $\{\epsilon_{ijk}\}_{i,j,k}$ são variáveis aleatórias independentes.

2. Havendo dois tipos de efeitos (do factor dominante A, localidade, e do factor subordinado B, terreno) o quadro de síntese terá três linhas (uma para cada tipo de efeito, e ainda a linha associada à variabilidade residual), sem contar com a linha correspondente à variabilidade total. Há dois valores dados no enunciado: o Quadrado Médio Residual, QMRE=1.673 e $F_{B(A)} = 3.592$. Os graus de liberdade são: $a-1=2$ (Factor A); $\sum_{i=1}^{a}(b_i-1)=21$ (Factor B) e $n-\sum_{i=1}^{a} b_i = 168-24 = 144$ (Residual). Assim, tem-se $QMB(A) = F_{B(A)} \times QMRE = 6.009416$. Logo, $SQB(A) = QMB(A) \times \left[\sum_{i=1}^{a}(b_i-1)\right] = 6.009416 \times 21 = 126.1977$. Por outro lado, $SQRE = \left(n-\sum_{i=1}^{a} b_i\right) \times QMRE = 144 \times 1.673 = 240.912$. Existem duas formas de calcular SQA: uma é pela fórmula dada no formulário. Outra (mais fácil) resulta de saber que $SQT = SQA + SQB(A) + SQRE$. Como sabemos que $SQT = (n-1)\, s_y^2 = 167 \times 3.08414 = 515.0514$, usando o valor de s_y^2 dado no enunciado. Logo, $SQA = SQT - [SQB(A) + SQRE] = 515.0514 - (126.1977 + 240.912) = 147.9417$. O respectivo Quadrado Médio é $QMA = \frac{SQA}{a-1} = \frac{147.9417}{2} = 73.97085$. Finalmente, a estatística do teste aos efeitos do factor dominante (localidade) é $F^A_{calc} = \frac{QMA}{QMRE} = \frac{73.97085}{1.673} = 44.21449$. Eis a tabela-resumo:

Fontes de Variação	gl	Somas de Quadrados	Quadrados Médios	F_{calc}
Localidades (Factor A)	2	147.9417	73.97085	44.21449
Terrenos (Factor B(A))	21	126.1977	6.009416	3.592
Residual	144	240.912	1.673	—
Total	167	515.0514	–	–

3. No teste aos efeitos de localidade, as hipóteses são $H_0 : \alpha_2 = \alpha_3 = 0$ e $H_1 : \alpha_2 \neq 0$ e/ou $\alpha_3 \neq 0$. A estatística de teste é $F^A = \frac{QMA}{QMRE} \frown F_{[a-1,\, n-\sum_{i=1}^{a} b_i]}$, sob H_0. A regra de rejeição ao nível de significância $\alpha = 0.05$ é rejeitar H_0 se $F_{calc} > f_{0.05(2,144)} \approx 3.05$. Como $F^A_{calc} = 44.21449$, há uma clara rejeição de H_0, ou seja, conclui-se pela existência de efeitos das localidades sobre o rendimento. Esta conclusão não surpreende, tendo em conta que o rendimento médio amostral no Bombarral é cerca de metade dos rendimentos médios observados nos dois outros locais.

4. No teste aos efeitos de terreno, a Hipótese Nula $H_0 : \beta_{j(i)} = 0$ para todos os terrenos (sendo H_1 que existe i, j tal que $\beta_{j(i)} \neq 0$) é rejeitada. O limiar da região crítica é aproximadamente $f_{0.05(21,144)} \approx 1.66$ (usando o valor tabelado para os 20 e 120 graus de liberdade). O valor calculado da estatística é $F^{B(A)} = 3.592$ Assim, conclui-se que a variabilidade de rendimentos ao longo dos terrenos é significativa, mesmo tendo em conta os efeitos entre localidades que já foram identificados na alínea anterior. Esta conclusão também não surpreende, se observarmos as variações de rendimentos dentro duma mesma localidade que, em especial no caso do Cadaval e Caldas da Rainha, são notórias. Assim, foi importante considerar este segundo factor no delineamento.

5. Duas médias populacionais de rendimento, em dois diferentes terrenos (de qualquer localidade) podem ser consideradas diferentes (ou seja, rejeita-se $\mu_{ij} = \mu_{i'j'}$ a favor de $\mu_{ij} \neq \mu_{i'j'}$) se se verificar a desigualdade $|\overline{y}_{ij.} - \overline{y}_{i'j'.}| > q_{\alpha(\sum_i b_i,\, n-\sum_i b_i)} \sqrt{\frac{QMRE}{n_c}}$. Tem-se $\sqrt{\frac{QMRE}{n_c}} = \sqrt{\frac{1.673}{7}} =$

0.4888763. Usando o nível global de significância $\alpha = 0.05$, tem-se que o quantil da distribuição Tukey (com parâmetros $\sum_{i=1}^{a} b_i = 24$ e $n - \sum_{i=1}^{a} b_i = 144$) é $q_{0.05(24,144)} \approx q_{0.05(20,120)} = 5.13$. Logo, o termo de comparação (limiar de significância) é dado por $5.13 \times 0.4888763 = 2.507935$. O maior rendimento médio amostral no Cadaval é 5.772. Logo, qualquer rendimento médio amostral de terreno inferior a $5.772 - 2.507935 = 3.264065$ deve ser considerado significativamente diferente (ao nível $\alpha = 0.05$). É o caso de *todos* os rendimentos médios observados no Bombarral, e dos dois primeiros terrenos das Caldas da Rainha. Este resultado é coerente com a ideia que é sobretudo o Bombarral que foi responsável pela rejeiçao de H_0 no teste aos efeitos de localidade.

III

1. Uma vez que foram fixados pelo experimentador os totais de linha (ou seja, o número de enxertos com cada tipo de porta-enxertos), trata-se dum *teste de homogeneidade*. O que se pretende é ver se é admissível pensar que a probabilidade dum enxerto viável é igual, nos três porta-enxertos, o que implicaria que a probabilidade do enxerto não ser viável, em cada porta-enxertos, é também igual. Designando por $\pi_{j|i}$ estas probabilidades condicionais de resultado ($j = 1, 2$), dado o porta-enxertos ($i = 1, 2, 3$), escrevemos a Hipótese Nula (que corresponde à homogeneidade) como:

$$H_0 : \pi_{1|1} = \pi_{1|2} = \pi_{1|3} \qquad \text{e} \qquad \pi_{2|1} = \pi_{2|2} = \pi_{2|3}$$

   A Hipótese Alternativa H_1 corresponde a que pelo menos uma das igualdades referidas não se verifique, isto é, que existem dois porta-enxertos diferentes, i e i', tais que $\pi_{j|i} \neq \pi_{j|i'}$ ($j = 1, 2$).

2. Pede-se para verificar a validade das condições de Cochran, que permitem admitir a validade da distribuição assintótica da estatística de Pearson. Essas condições incidem sobre os valores esperados (neste contexto, os valores esperados *estimados*), e consistem em exigir que nenhum $\hat{E}_{ij}$ seja inferior a 1 e não mais do que um quinto dos valores esperados estimados sejam inferiores a 5. O número muito elevado de observações ($N = 5288$) faz suspeitar que as condições se verificam. Mas confirmaremos, calculando o menor de todos os valores $\hat{E}_{ij} = \frac{N_{i.} \times N_{.j}}{N}$, que está associado a tomar a célula (i, j) da linha i e coluna j com os menores totais marginais. Trata-se da linha $i = 3$, com $N_{3.} = 1747$, e coluna $j = 2$, com $N_{.2} = 1410$. Tem-se $\hat{E}_{32} = \frac{N_{3.} \times N_{.2}}{N} = \frac{1747 \times 1410}{5288} = 465.8226$. Trata-se dum valor muito superior a 5, pelo que as condições de Cochran estão verificadas de forma muito folgada, não havendo qualquer reserva na utilização da distribuição assintótica da estatística, que é $\chi^2_{(a-1)(b-1)}$.

3. A estatística de Pearson é da forma $\sum_{i=1}^{3} \sum_{j=1}^{2} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$. A parcela pedida no enunciado corresponde a $i = 3$ e $j = 2$. O valor observado nesta parcela é $O_{32} = 390$. O correspondente valor esperado estimado foi calculado na alínea anterior: $\hat{E}_{32} = 465.8226$. Assim, a contribuição desta parcela para o valor final da estatística do teste é dada por: $\frac{(O_{32} - \hat{E}_{32})^2}{\hat{E}_{32}} = \frac{(390 - 465.8226)^2}{465.8226} = 12.34175$.

4. O valor da estatística de teste dado no enunciado tem de ser comparado com a fronteira da região crítica que sabemos ser unilateral direita, sendo a distribuição assintótica da estatística do teste, sob H_0, uma $\chi^2_{(a-1)(b-1)}$, onde $a = 3$ indica o número de linhas da tabela e $b = 2$ o respectivo número de colunas. Assim, usando $\alpha = 0.05$, rejeitar-se-á H_0 se $X^2_{calc} > \chi^2_{0.05(2)} = 5.991$. Ora, $X^2_{calc} = 28.123$, pelo que se rejeita H_0 (só a parcela pedida na alínea anterior já assegura a rejeição). Conclui-se que não há igual afinidade à enxertia dos três porta-enxertos.

IV

1. O modelo ANOVA a um factor, com k níveis, tem equação para cada observação Y_{ij} dada por $Y_{ij} = \mu_1 + \alpha_i + \epsilon_{ij}$ onde $i = 1, ..., k$ indica o nível do factor; $j = 1, ..., n_i$ indica cada uma das observações da variável resposta Y no nível i do factor; e, com a restrição $\alpha_1 = 0$, a constante aditiva μ_1 é a média populacional do primeiro nível do factor.

   (a) A matriz do modelo $\mathbf{X}$ tem n linhas (tantas quantas as observações) e k colunas (tantas quantos os parâmetros do modelo). Como em qualquer outro Modelo Linear estudado nesta disciplina, a primeira coluna de $\mathbf{X}$ é uma coluna de n uns, associada à parcela constante da equação do modelo, ou seja a μ_1. As restantes $k-1$ colunas de $\mathbf{X}$, que estão associadas aos restantes $k-1$ parâmetros do modelo (os α_i, após a introdução da restrição $\alpha_1 = 0$), são as colunas indicatrizes de pertença a cada um dos $k-1$ níveis do factor para os quais $i > 1$. Ou seja, a segunda coluna de $\mathbf{X}$ é o vector $\vec{\mathcal{I}}_2$, em que os únicos elementos não nulos (todos de valor 1) estão nas n_2 posições correspondentes às observações efectuadas no nível $i = 2$ do factor; a terceira coluna é o vector $\vec{\mathcal{I}}_3$, com uns nas n_3 posições correspondentes às observações efectuadas no nível $i = 3$ do factor (zero noutras posições); e por aí fora, como indicado em baixo.

$$\mathbf{X} = \left[\begin{array}{ccccc}
1 & 0 & 0 & \dots & 0 \\
1 & 0 & 0 & \dots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 0 & \dots & 0 \\
1 & 0 & 0 & \dots & 0 \\
-- & -- & -- & -- & -- \\
1 & 1 & 0 & \dots & 0 \\
1 & 1 & 0 & \dots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 1 & 0 & \dots & 0 \\
1 & 1 & 0 & \dots & 0 \\
-- & -- & -- & -- & -- \\
1 & 0 & 1 & \dots & 0 \\
1 & 0 & 1 & \dots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 1 & \dots & 0 \\
1 & 0 & 1 & \dots & 0 \\
-- & -- & -- & -- & -- \\
1 & 0 & 0 & \dots & 1 \\
1 & 0 & 0 & \dots & 1 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 0 & \dots & 1 \\
1 & 0 & 0 & \dots & 1 \\
\uparrow & \uparrow & \uparrow & & \uparrow \\
\vec{\mathbf{1}}_n & \vec{\boldsymbol{\mathcal{I}}}_2 & \vec{\boldsymbol{\mathcal{I}}}_3 & \dots & \vec{\boldsymbol{\mathcal{I}}}_a
\end{array}\right]$$

   (b) Os vectores do espaço das colunas $\mathcal{C}(\mathbf{X})$ são, por definição, combinações lineares das colunas de $\mathbf{X}$, ou seja, vectores da forma $a_1 \vec{\mathbf{1}}_n + a_2 \vec{\boldsymbol{\mathcal{I}}}_2 + a_3 \vec{\boldsymbol{\mathcal{I}}}_3 + ... + a_k \vec{\boldsymbol{\mathcal{I}}}_k$. Esses vectores resultam de multiplicar cada coluna de $\mathbf{X}$ pelo respectivo coeficiente a_i, e depois somar. A propriedade indicada no enunciado resulta do facto de as observações dum mesmo nível do factor terem, nas colunas de $\mathbf{X}$, sempre os mesmos valores 1 ou 0. De facto, os elementos das indicatrizes $\vec{\boldsymbol{\mathcal{I}}}_2$, ... , $\vec{\boldsymbol{\mathcal{I}}}_a$ são todos nulos para as n_1 observações efectuadas no nível $i = 1$ do factor e apenas o vector $\vec{\mathbf{1}}_n$ tem valores não nulos, todos iguais a 1. Logo, para essas n_1 observações do nível $i = 1$, a combinação linear vai produzir o valor a_1. As n_2 observações do segundo nível do factor, nas indicatrizes de nível $i > 2$ ($\vec{\boldsymbol{\mathcal{I}}}_3$, ... , $\vec{\boldsymbol{\mathcal{I}}}_a$) tomam valor nulo, mas na indicatriz $\vec{\boldsymbol{\mathcal{I}}}_2$ do nível $i = 2$ tomam valor 1, tal como no vector $\vec{\mathbf{1}}_n$. Assim, para essas n_2 observações, o valor da combinação linear será $a_1 + a_2$. Um raciocínio análogo faz com que

as observações do nível $i > 2$ do factor tenham valor $a_1 + a_i$, como indicado em baixo.

$$a_1\,\vec{\mathbf{1}}_n + a_2\vec{\boldsymbol{\mathcal{I}}}_2 + a_3\vec{\boldsymbol{\mathcal{I}}}_3 + ... + a_k\vec{\boldsymbol{\mathcal{I}}}_k = \begin{bmatrix} a_1 \\ ... \\ a_1 \\ \hline a_1 + a_2 \\ ... \\ a_1 + a_2 \\ \hline a_1 + a_3 \\ ... \\ a_1 + a_3 \\ \hline (...) \\ \hline a_1 + a_k \\ ... \\ a_1 + a_k \end{bmatrix}$$

2. (a) O modelo RLM em notação vectorial é constituido pela equação do modelo, e pela indicação dos pressupostos exigidos ao vector dos erros aleatórios. Mais concretamente,

   i. $\vec{\mathbf{Y}} = \mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\epsilon}}$ (equação do modelo).

   ii. $\vec{\boldsymbol{\epsilon}} \frown \mathcal{N}_n(\vec{\mathbf{0}}\,,\, \sigma^2\,\mathbf{I}_n)$ (pressupostos sobre os erros aleatórios).

   onde:

   - $\vec{\mathbf{Y}} = (Y_1, ..., Y_n)^t$ é o vector aleatório das n observações da variável resposta;
   - $\mathbf{X}$ é a matriz do modelo (não aleatória) de dimensões $n \times (p+1)$ cujas colunas são dadas por uma primeira coluna de uns, associada a constante aditiva do modelo (β_0) e por p colunas adicionais, cada uma das quais contém as n observações de cada variável preditora;
   - $\vec{\boldsymbol{\beta}} = (\beta_0, \beta_1, ..., \beta_p)^t$ é o vector (não aleatório) dos $p+1$ parâmetros do modelo:;
   - $\vec{\boldsymbol{\epsilon}} = (\epsilon_1, ..., \epsilon_n)^t$ é o vector aleatório dos n erros aleatórios;
   - $\mathbf{I}_n$ é a matriz identidade de dimensão $n \times n$;
   - σ^2 é uma constante, que corresponde à variância comum de todos os erros aleatórios.

   (b) A equação do modelo é $\vec{\mathbf{Y}} = \mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\epsilon}}$, sendo $\vec{\boldsymbol{\epsilon}}$ um vector aleatório com distribuição Multinormal e o produto $\mathbf{X}\vec{\boldsymbol{\beta}}$ um vector não aleatório. Assim, o vector aleatório $\vec{\mathbf{Y}}$ é uma soma $\vec{\mathbf{a}} + \vec{\mathbf{W}}$, com $\vec{\mathbf{a}}$ constante e $\vec{\mathbf{W}}$ Multinormal. Como se viu no estudo das propriedades da Multinormal, isso implica que $\vec{\mathbf{Y}}$ também tem distribuição Multinormal. Para calcular os seus parâmetros, utilizamos as propriedades dos vectores esperados e matrizes de (co-)variâncias para somas do tipo $\vec{\mathbf{a}} + \vec{\mathbf{W}}$. Tem-se:

$$\begin{aligned} E[\vec{\mathbf{Y}}] &= E[\underbrace{\mathbf{X}\vec{\boldsymbol{\beta}}}_{=\vec{\mathbf{a}}} + \underbrace{\vec{\boldsymbol{\epsilon}}}_{=\vec{\mathbf{W}}}] = \mathbf{X}\vec{\boldsymbol{\beta}} + \underbrace{E[\vec{\boldsymbol{\epsilon}}]}_{=\vec{\mathbf{0}}} = \mathbf{X}\vec{\boldsymbol{\beta}}\ . \\ V[\vec{\mathbf{Y}}] &= V[\underbrace{\mathbf{X}\vec{\boldsymbol{\beta}}}_{=\vec{\mathbf{a}}} + \underbrace{\vec{\boldsymbol{\epsilon}}}_{=\vec{\mathbf{W}}}] = V[\vec{\boldsymbol{\epsilon}}] = \sigma^2\mathbf{I}_n\ . \end{aligned}$$

   Logo, $\vec{\mathbf{Y}} \frown \mathcal{N}_n(\ \mathbf{X}\vec{\boldsymbol{\beta}}\ ,\ \sigma^2\mathbf{I}_n\ )$.

3. O modelo RLM usual é modificado apenas no facto de se ter $V[\vec{\boldsymbol{\epsilon}}] = \Sigma$ (matriz simétrica conhecida, logo não aleatória), ou seja, tem-se $\vec{\boldsymbol{\epsilon}} \frown \mathcal{N}_n(\ \vec{\mathbf{0}}\ ,\ \Sigma\ )$.

   (a) Os elementos da matriz Σ são as (co-)variâncias entre cada par de erros aleatórios (elementos ϵ_i do vector $\vec{\boldsymbol{\epsilon}}$). Em particular, os elementos não diagonais de Σ são $\Sigma_{ij} = Cov[\epsilon_i, \epsilon_j]$,

com $i \neq j$. Se a matriz Σ não é diagonal, pelo menos uma destas covariâncias é diferente de zero, o que implica que ϵ_i e ϵ_j não são independentes. Logo, ao admitir-se esta generalização de $V[\vec{\boldsymbol{\epsilon}}]$, deixa-se cair a exigência de erros aleatórios independentes.

(b) O enunciado garante que se tem agora $\vec{\mathbf{Y}} \frown \mathcal{N}_n(\, \mathbf{X}\vec{\boldsymbol{\beta}}\,, \Sigma\,)$. Por outro lado, o vector de estimadores é $\hat{\vec{\boldsymbol{\beta}}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{Y}}$, um produto duma matriz não aleatória ($\mathbf{B} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$) com um vector aleatório Multinormal ($\vec{\mathbf{W}} = \vec{\mathbf{Y}}$). As propriedades da distribuição Multinormal garantem que esse produto preserva a Multinormalidade. Vamos calcular os respectivos parâmetros, utilizando as propriedades de vectores esperados e matrizes de (co-)variâncias, bem como propriedades de matrizes disponíveis no formulário:

$$E[\hat{\vec{\boldsymbol{\beta}}}] \;=\; E[\underbrace{(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t}_{=\mathbf{B}}\,\underbrace{\vec{\mathbf{Y}}}_{=\vec{\mathbf{W}}}] \;=\; (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\,\underbrace{E[\vec{\mathbf{Y}}]}_{=\mathbf{X}\vec{\boldsymbol{\beta}}} \;=\; \underbrace{(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}}_{=\mathbf{I}}\,\vec{\boldsymbol{\beta}} \;=\; \vec{\boldsymbol{\beta}}.$$

$$\begin{aligned} V[\hat{\vec{\boldsymbol{\beta}}}] \;&=\; V[\underbrace{(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t}_{=\mathbf{B}}\,\underbrace{\vec{\mathbf{Y}}}_{=\vec{\mathbf{W}}}] \;=\; (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\cdot V[\vec{\mathbf{Y}}]\cdot[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t]^t \\ &=\; (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\cdot\Sigma\cdot[\mathbf{X}^t]^t\,[(\mathbf{X}^t\mathbf{X})^{-1}]^t \;=\; (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\Sigma\mathbf{X}[(\mathbf{X}^t\mathbf{X})^t]^{-1} \\ &=\; (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\Sigma\mathbf{X}\left[(\mathbf{X}^t[\mathbf{X}^t]^t\right]^{-1} \;=\; (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\Sigma\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\,. \end{aligned}$$

Esta última expressão não pode, em geral, ser ulteriormente simplificada, uma vez que *a multiplicação de matrizes não é comutativa*. Em particular, a matriz genérica Σ não pode, em geral, sair da posição onde se encontra. Logo, $\hat{\vec{\boldsymbol{\beta}}} \frown \mathcal{N}_n(\,\vec{\boldsymbol{\beta}}\,,\; (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\Sigma\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\,)$. Os estimadores são centrados, porque $E[\hat{\vec{\boldsymbol{\beta}}}] = \vec{\boldsymbol{\beta}}$ significa que, para todo o $i = 0, ..., p$, se tem $E[\hat{\beta}_i] = \beta_i$ (e um estimador diz-se centrado quando o seu valor esperado é o parâmetro que estima).