INSTITUTO SUPERIOR DE AGRONOMIA

**ESTATÍSTICA E DELINEAMENTO**

13 Janeiro 2020 **Segundo Teste 2019-20** **Uma resolução possível**

## I

1. Há dois aspectos a referir. Por um lado, o Coeficiente de Determinação $R^2 = 0.8429$ indica que a relação linear entre `log(Stemflow)` e os preditores explica $84,29\%$ da variância observada nos logaritmos dos escorrimentos. Trata-se de um bom ajustamento. Como seria de esperar, é um valor suficientemente elevado para ser considerado significativamente diferente de zero, como se confirma através do teste de ajustamento global do modelo:

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

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

   **Nível de significância:** $\alpha = 0.05$.

   **Região Crítica:** Unilateral direita. Rejeitar $H_0$ se $F_{calc} > f_{0.05[9,45]}$. Este limiar está entre os valores tabelados $f_{0.05[9,40]} = 2.12$ e $f_{0.05[9,60]} = 2.04$. No que se segue utilizaremos o valor aproximado $f_{0.05[9,45]} \approx 2.10$.

   **Conclusões:** No enunciado não é dado o valor calculado da estatística, mas como se viu, $F_{calc} = \frac{n-(p+1)}{p} \frac{R^2}{1-R^2} = \frac{45}{9} \times \frac{0.8429}{1-0.8429} = 26.82686 \gg 2.10$. A rejeição de $H_0$ é clara, pelo que o modelo ajustado é muito significativamente diferente do Modelo Nulo (o correspondente valor de prova é $p = 2.991 \times 10^{-15}$).

2. É pedido um intervalo a $(1-\alpha)\times 100\% = 95\%$ de confiança para $\beta_9$, o coeficiente da última variável preditora (`Area.Copa`). Sabemos que este intervalo de confiança tem como ponto central o valor estimado $b_9 = -0.38015$ e semi-amplitude dada pelo produto do quantil de ordem $1-\frac{\alpha}{2} = 0.975$, numa distribuição $t_{n-(p+1)}$, ou seja, $t_{0.025(45)} \approx 2.015$; e do erro padrão do estimador de $\beta_9$, ou seja, $\hat{\sigma}_{\hat{\beta}_9} = 0.27904$ (dado no enunciado). Assim, tem-se 95% de confiança na veracidade da afirmação de que, na população, a variação esperada no log-escorrimento, por cada $m^2$ adicional na área da copa da árvore (e mantendo fixas as restantes variáveis preditoras) é um dos valores do seguinte intervalo:

   $$\left] b_9 - t_{0.025(45)} \times \hat{\sigma}_{\hat{\beta}_9} \ , \ b_9 + t_{0.025(45)} \times \hat{\sigma}_{\hat{\beta}_9} \right[ \quad = \quad ] -0.9424156 \ , \ 0.1821156 [ \ .$$

   Uma vez que o intervalo contém o valor zero, não se pode rejeitar que $\beta_9 = 0$. A ser esse o verdadeiro valor de $\beta_9$, a variável preditora `Area.Copa` não contribuiria em nada para a modelação dos valores de log-escorrimento. Assim, a variável `Area.Copa` é dispensável do modelo, sem perda significativa na qualidade do ajustamento.

3. O aumento médio no log-escorrimento (variável resposta) associado a aumentar uma unidade a sexta variável preditora (`alt.arv`), mantendo o resto igual, é dado pelo coeficiente $\beta_6$. O correspondente aumento médio no log-escorrimento associado a aumentar uma unidade o séptimo preditor (`alt.tronco`), mantendo o resto igual, é dado pelo coeficiente $\beta_7$. O enunciado pergunta se é possível rejeitar a igualdade destes dois coeficientes, o que pode ser feito através dum teste bilateral a $H_0 : \beta_6 - \beta_7 = 0$ ($\beta_6 - \beta_7$ é uma combinação linear dos parâmetros).

   **Hipóteses:** $H_0 : \beta_6 - \beta_7 = 0$ vs. $H_1 : \beta_6 - \beta_7 \neq 0$

**Estatística do teste** $T = \frac{(\hat{\beta}_6 - \hat{\beta}_7) - (\beta_6 - \beta_7)|_{H_0}}{\hat{\sigma}_{\hat{\beta}_6 - \hat{\beta}_7}} \underset{\cdot}{\sim} t_{n-(p+1)}$, sob $H_0$.

**Nível de significância:** $\alpha = 0.05$

**Região Crítica:** Bilateral. Rejeita-se $H_0$ se $|T_{calc}| > t_{\frac{\alpha}{2}(n-(p+1)} = t_{0.025(45)} \approx 2.015$.

**Conclusões:** No nosso caso, o numerador da estatística calculada reduz-se a $b_6 - b_7 = 2.51834 - 0.94257 = 1.57577$. No que respeita ao denominador, e tratando-se do erro padrão duma diferença de dois estimadores, podemos começar por calcular a variância estimada dessa diferença:

$$\widehat{V[\hat{\beta}_6 - \hat{\beta}_7]} = \widehat{V[\hat{\beta}_6]} + \widehat{V[\hat{\beta}_7]} - 2\,\widehat{Cov[\hat{\beta}_6, \hat{\beta}_7]} = \hat{\sigma}^2_{\hat{\beta}_6} + \hat{\sigma}^2_{\hat{\beta}_7} - 2\,\widehat{Cov[\hat{\beta}_6, \hat{\beta}_7]} \ .$$

Os dois primeiros valores são os quadrados dos erros padrões de $\hat{\beta}_6$ e $\hat{\beta}_7$, que constam da listagem produzida aquando do ajustamento do modelo. A terceira e última parcela é dada no enunciado. Logo, $\widehat{V[\hat{\beta}_6 - \hat{\beta}_7]} = 1.22131^2 + 2.02048^2 - 2 \times (-1.537745) = 8.649428$. Logo, a estatística calculada tem valor $T_{calc} = \frac{1.57577}{\sqrt{8.649428}} = 0.535796$. Assim, não se rejeita $H_0$, sendo de admitir a igualdade de $\beta_6$ e $\beta_7$.

4. Procedeu-se a ajustar o submodelo resultante de excluir os quatro preditores cujo *p-value* nos testes bilaterais a $H_0 : \beta_j = 0$ excedia $p = 0.05$ (ou seja, um submodelo com apenas $k = 5$ preditores). Como é dito no enunciado, o Coeficiente de Determinação resultante foi $R^2_s = 0.8127$. É pedido um teste $F$ parcial para verificar se se deve considerar que este submodelo tem uma qualidade de ajustamento significativamente mais baixa que a do modelo completo original.

   **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** $F = \frac{n-(p+1)}{p-k} \cdot \frac{R^2_c - R^2_s}{1 - R^2_c} \underset{\cdot}{\sim} F_{[p-k, n-(p+1)]}$, se $H_0$ verdade.

   **Nível de significância:** $\alpha = 0.05$

   **Região Crítica:** Unilateral direita. Rejeita-se $H_0$ se $F_{calc} > f_{\alpha[p-k, n-(p+1)]} = f_{0.05(4,45)} \approx 2.59$.

   **Conclusões:** Tem-se $F_{calc} = \frac{45}{4} \times \frac{0.8429 - 0.8127}{1 - 0.8429} = 2.1626$. Assim, não se rejeita $H_0$. Não se pode concluir que os dois modelos tenham qualidade de ajustamento significativamente diferente.

   **Nota:** *Não* se podia garantir isto a partir do facto de, na alínea anterior, os *p-values* individuais de cada um dos preditores excluídos serem superiores a 0.05.

5. Por definição, o Coeficiente de Determinação usual é dado por $R^2 = \frac{SQR}{SQT} = \frac{SQT - SQRE}{SQT} = 1 - \frac{SQRE}{SQT}$. No formulário pode verificar-se que o Coeficiente de Determinação modificado é dado por $R^2_{mod} = 1 - \frac{QMRE}{QMT}$. Uma vez que se trata dum submodelo com $k$ preditores, tem-se: $QMRE = \frac{SQRE}{n-(k+1)}$ e $QMT = \frac{SQT}{n-1}$. Logo, tem-se:

$$R^2_{mod} = 1 - \frac{n-1}{n-(k+1)} \cdot \frac{SQRE}{SQT} = 1 - \frac{n-1}{n-(k+1)} \cdot (1 - R^2) \ .$$

Sabemos que, neste submodelo, $R^2_s = 0.8127$, $k = 5$ e $n = 55$. Logo, $R^2_{mod} = 1 - \frac{54}{49} \times (1 - 0.8127) = 0.7936$. O facto de a diferença entre os valores das duas variantes do Coeficiente de Determinação ser menor no submodelo (cerca de 0.02) do que no modelo completo (cerca de 0.03) reflecte o facto de um valor semelhante do Coeficiente de Determinação usual no submodelo ser obtido com cerca de metade dos preditores.

## II

1. A variável resposta (numérica) é o rendimento (em kg/planta). Há dois factores explicativos: `ambiente` (factor A, com $a=4$ níveis) e `clone` (factor B, com $b=6$ níveis). Havendo rendimentos observados para as $ab = 24$ possíveis combinações de ambientes e genótipos (clones), tem-se um *delineamento factorial*. Em cada uma das $ab = 24$ situações experimentais (associações ambiente/genótipo) existe um número idêntico de $n_c = 9$ repetições, pelo que se trata dum *delineamento equilibrado*, com um total de $n = ab\,n_c = 24 \times 9 = 216$ observações de rendimento. O facto de haver repetições nas células (situações experimentais) significa que é possível ajustar um modelo com efeitos de interacção, ou seja, que se vai considerar o seguinte modelo ANOVA:

   **Equação do Modelo:** $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$,
   com as restrições $\alpha_1 = 0$; $\beta_1 = 0$; $(\alpha\beta)_{ij} = 0$ se $i = 1$ e/ou $j = 1$, e onde:

   - $i = 1, 2, 3, 4$ indica o nível do Factor `ambiente` a que corresponde uma observação;
   - $j = 1, 2, 3, 4, 5, 6$ indica o genótipo (factor `clone`) a que corrersponde uma observação;
   - $k = 1, 2, ..., 9$ indica a repetição no seio da célula $(i, j)$;
   - $Y_{ijk}$ indica o rendimento da $k$-ésima repetição na célula $(i, j)$, sendo $\epsilon_{ijk}$ o correspondente erro aleatório;
   - $\mu_{11}$ representa a média populacional na situação experimental $(1, 1)$, ou seja, do primeiro genótipo no primeiro ambiente;
   - $\alpha_i$ indica o efeito principal associado ao ambiente $i$;
   - $\beta_j$ indica o efeito principal associado ao genótipo $j$;
   - $(\alpha\beta)_{ij}$ indica o efeito de interacção na célula $(i, j)$.

   **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 três tipos de efeitos (principais do factor ambiente, principais do factor clone e de interacção) o quadro de síntese terá quatro linhas (uma para cada tipo de efeito, e ainda a linha associada à variabilidade residual), sem contar com a linha correspondente à variabilidade Total. Nesta tabela, há três valores directamente referidos no enunciado: o Quadrado Médio Residual, QMRE=1.8728, SQB=53.5437 e $F^A_{calc} = \frac{QMA}{QMRE} = 247.150$. Assim, tem-se $QMA = QMRE \times F^A_{calc} = 1.8728 \times 247.150 = 462.8625$. Os graus de liberdade são dados por: $a-1 = 3$ para o factor A; $b-1 = 5$ para o factor B; $(a-1)(b-1) = 15$ para os efeitos de interacção; e $n–ab = 216-24 = 192$ para a variabilidade residual. Logo, o Quadrado Médio associado ao Factor B tem valor $QMB = \frac{SQB}{b-1} = 10.70874$, de onde decorre que o valor calculado da estatística $F^B$ para o teste à existência de efeitos principais do Factor B tem valor $F^B_{calc} = \frac{QMB}{QMRE} = 5.718037$. A Soma de Quadrados Residual tem valor $SQRE = QMRE \times (n - ab) = 1.8728 \times 192 = 359.5776$ e a Soma de Quadrados associada ao Factor A tem valor $SQA = QMA \times (a - 1) = 462.8625 \times 3 = 1388.588$. Falta apenas calcular as três quantidades em falta na linha de valores correspondentes aos efeitos de interacção. Pode calcular-se a Soma de Quadrados associada à interacção a partir das fórmulas $SQT = (n-1)\,s^2_y = 215 \times 8.756217 = 1882.587$ e $SQAB = SQT - (SQA + SQB + SQRE) = 1882.587 - (1388.588 + 53.5437 + 359.5776) = 80.8777$. Deste valor decorre que o respectivo Quadrado Médio é $QMAB = \frac{SQAB}{(a-1)(b-1)} = 5.391847$ e, finalmente, $F^{AB}_{calc} = \frac{QMAB}{QMRE} = 2.87903$. Juntando todos estes valores tem-se a seguinte tabela:

| Fontes de Variação | gl | Somas de Quadrados | Quadrados Médios | $F_{calc}$ |
|---|---|---|---|---|
| Factor Ambiente | 3 | 1388.588 | 462.8625 | 247.150 |
| Factor Genótipo | 5 | 53.5437 | 10.70874 | 5.718037 |
| Interacção | 15 | 80.8777 | 5.391847 | 2.87903 |
| Residual | 192 | 359.5776 | 1.8728 | — |

3. Neste modelo há três testes $F$ de interesse, um para cada tipo de efeito. Nos três casos, as hipóteses nulas correspondem à inexistência do tipo de efeitos referido (e a Hipótese alternativa à sua negação), e nos três casos a estatística do teste corresponde à razão entre o Quadrado Médio do tipo de efeito a ser testado e o Quadrado Médio Residual. As distribuições das estatísticas, sob as respectivas Hipóteses Nulas, são sempre $F$, com graus de liberdade associados aos Quadrados Médios que definem a estatística. Vejamos em pormenor o teste aos efeitos de interacção:

   **Hipóteses:** $H_0 : (\alpha\beta)_{ij} = 0 \ , \ \forall \ i,j \quad$ vs. $\quad H_1 : \exists\, i,j$ tal que $(\alpha\beta)_{ij} \neq 0$.

   **Estatística do Teste:** $F^{AB} = \frac{QMAB}{QMRE} \ \frown \ F_{[\,(a-1)(b-1)\,,\,n-ab\,]}$, sob $H_0$.

   **Nível de significância:** $\alpha = 0.05$.

   **Região Crítica:** (Unilateral direita) Rejeitar $H_0$ se $F_{calc} > f_{0.05(15,192)} \approx 1.75$.

   **Conclusões:** Como $F^{AB}_{calc} = 2.87903 > 1.75$, rejeita-se $H_0$. Pode concluir-se que existem efeitos de interacção significativos,ao nível $\alpha = 0.05$.

   No teste aos efeitos principais de ambiente, a Hipótese Nula $H_0 : \alpha_i = 0$, para todos os ambientes $i$, é rejeitada de forma clara. De facto, o limiar da Região Crítica será agora $f_{0.05(3,192)} \approx 2.68$, enquanto que $F^{A}_{calc} = 247.150$. Também se rejeita (embora de forma menos enfática) a Hipótese Nula da inexistência de efeitos principais do Factor B ($H_0 : \beta_j = 0$, para todos os genótipos $j$), já que a fronteira da Região Crítica é $f_{0.05(5,192)} \approx 2.29$, e $F^{B}_{calc} = 5.718037$. Assim, os três tipos de efeitos são significativos, ao nível de significância $\alpha = 0.05$, embora os efeitos de ambiente sejam bastantes mais pronunciados que os restantes tipos de efeitos.

4. Utilizamos testes de Tukey para comparar as médias de célula. A diferença de duas médias amostrais de célula pode ser considerada significativa se se verificar a desigualdade $|\overline{y}_{ij.} - \overline{y}_{i'j'.}| > \tau_{\alpha(ab,n-ab)}\sqrt{\frac{QMRE}{n_c}}$. No cálculo do termo de comparação tem-se $\sqrt{\frac{QMRE}{n_c}} = \sqrt{\frac{1.8728}{9}} = 0.4561676$. O quantil de ordem $1-\alpha$ da distribuição de Tukey pode ser obtido (aproximadamente) nas tabelas de Tukey. Usando o nível global de significância $\alpha = 0.05$, tem-se $\tau_{0.05(24,192)} \approx 5.01$ (usando o valor tabelado correspondente aos parâmetros 20 e $\infty$). Logo, o limiar (aproximado) de significância é dado por $5.01 \times 0.4561676 = 2.2854$. Uma inspecção preliminar das médias de célula constantes do enunciado permite constatar que as diferenças entre as médias de célula do Ambiente 2 e do Ambiente 1 são sempre superiores a esse limiar. Para confirmar, basta considerar a menor média amostral de célula no Ambiente 2, que é $\overline{y}_{25.} = 5.823$, e a maior média associada ao Ambiente 1, que é $\overline{y}_{14.} = 1.871$. A diferença entre essas duas médias amostrais de célula é $5.823 - 1.871 = 3.952 > 2.2854$. Qualquer outro par de médias (uma de cada ambiente) tem diferenças ainda maiores, logo mais significativas. A afirmação do enunciado é correcta.

5. Trata-se dum gráfico de interacção, com o Factor B (`clone`) associado ao eixo horizontal. Por cima do marcador de cada clone encontram-se 4 pontos, a alturas correspondentes às quatro médias amostrais de célula associadas a esse clone (um ponto por cada ambiente, ou seja, cada nível do Factor A). Segmentos de recta unem os pontos que correspondem a um mesmo ambiente (a legenda do gráfico indica o tipo de traço usado para construir os segmentos). O gráfico torna evidente que as células com os ambientes 2 e 3 têm, globalmente falando, médias amostrais

bastante maiores que as dos ambientes 1 e 4. Estas diferenças justificam que os efeitos de ambiente sejam claramente significativos, como se verificou no respectivo teste $F$. Os efeitos de interacção (que foram considerados significativos ao nível $\alpha = 0.05$) correspondem à falta de paralelismo entre as "curvas" associadas aos ambientes de maiores rendimentos (2 e 3), e entre estas e as dos ambientes de menores rendimentos (1 e 4). Note-se em particular a forma como, no ambiente 3, o rendimento associado ao clone `AI2` é visivelmente superior aos rendimentos dos clones `AI1` e `AI3`, enquanto que no ambiente 2 se passa precisamente o contrário. Ao mesmo tempo, pouca diferença existe entre os rendimentos de todos os clones nos ambientes 1 e 4 (que são consistentemente baixos). Naturalmente, são os testes que nos indicam se estas diferenças sugeridas pelo gráfico devem, ou não, ser consideradas significativas.

## III

1. No primeiro passo dum algoritmo de exclusão sequencial baseado nos testes bilaterais $t$ à hipótese nula $H_0 : \beta_j = 0$, a variável a excluir será a que está associada ao valor $t_{calc}$ mais próximo de zero. Sabemos também que o quadrado do valor $t_{calc}$ é igualmente o valor da estatística do teste $F$ parcial comparando o modelo completo e o submodelo que resulta da exclusão do preditor $x_j$. Logo, a exclusão corresponderá à variável para a qual seja mínimo o valor:

$$(t_{calc})^2 = F_{calc} = \frac{n-(p+1)}{p-k} \cdot \frac{R_c^2 - R_s^2}{1-R_c^2} \ ,$$

sendo $k = p-1$ igual para todos os submodelos considerados no primeiro passo do algoritmo. Uma vez que $n$ e $R_c^2$ também são sempre iguais, o menor valor de $F_{calc}$ terá de corresponder ao submodelo em que $R_s^2$ esteja mais próximo de $R_c^2$, ou seja, corresponderá ao maior valor do Coeficiente de Determinação $R_s^2$, de entre todos os submodelos com $p-1$ preditores.

No caso de se usar a variante do algoritmo baseada no AIC, o submodelo de $k = p-1$ preditores escolhido no primeiro passo será aquele para o qual o AIC seja menor. Ora, por definição,

$$AIC = n \ln\left(\frac{SQRE_k}{n}\right) + 2(k+1) \ .$$

De novo, em todos os submodelos considerados no primeiro passo do algoritmo, tem-se $k = p-1$, logo a segunda parcela do AIC terá valor idêntico. Apenas pode diferir o valor do AIC como resultado da primeira parcela. O logaritmo é uma função crescente, logo o menor AIC corresponde ao submodelo com menor valor de $SQRE_k$. Mas como $R_s^2 = 1 - \frac{SQRE_k}{SQT}$, o menor valor de $SQRE_k$ está associado ao maior valor de $R_s^2$. Por isso, também na variante do algoritmo baseada no AIC o submodelo escolhido no primeiro passo corresponde ao submodelo com maior valor do Coeficiente de Determinação $R_s^2$.

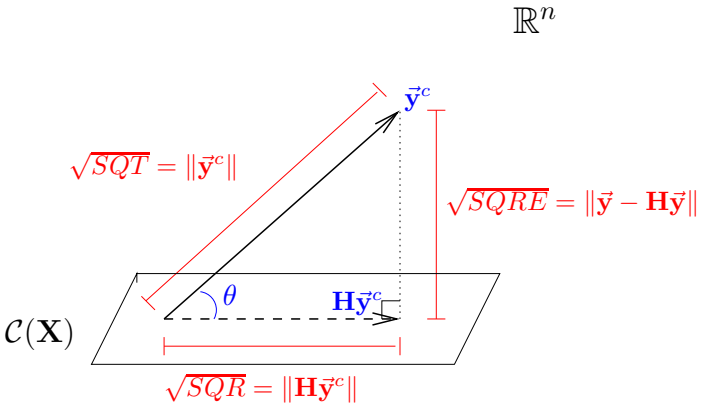
2. (a) Seja $\vec{\mathbf{Y}}$ o vector aleatório com as $n$ observações da variável resposta, e $\vec{\boldsymbol{\epsilon}}$ o vector aleatório dos correspondentes erros aleatórios. Seja $\mathbf{X}_{n\times(p+1)}$ a matriz (não aleatória) do modelo, cuja primeira coluna é constituída por $n$ uns, e cujas colunas seguintes contêm as $n$ observações de cada uma das $p$ variáveis preditoras. Seja $\vec{\boldsymbol{\beta}}$ o vector (não aleatório) constituído pelos $p+1$ parâmetros do modelo: $\vec{\boldsymbol{\beta}} = (\beta_0, \beta_1, ..., \beta_p)^t$. O Modelo de Regressão Linear Múltipla admite os seguintes pressupostos:

   - Equação do Modelo: $\vec{\mathbf{Y}} = \mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\epsilon}}$;
   - Pressupostos sobre os erros aleatórios: $\vec{\boldsymbol{\epsilon}} \sim \mathcal{N}_n(\vec{\mathbf{0}}\,,\, \sigma^2\,\mathbf{I}_n)$.

   A equação do Modelo corresponde à relação linear de fundo entre os preditores e a variável resposta. Os erros aleatórios representam a variabilidade em torno dessa relação linear,

admitindo-se a Multinormalidade, independência e variâncias homogéneas no segundo pressuposto do Modelo.

(b) O triângulo ao qual se aplica o Teorema de Pitágoras tem como hipotenusa o vector centrado das observações de $Y$, o vector $\vec{\mathbf{y}}^c$, cujo elemento genérico é $y_i - \bar{y}$. Um dos catetos é a projecção ortogonal de $\vec{\mathbf{y}}^c$ sobre o subespaço gerado pelas colunas da matriz $\mathbf{X}$, o subespaço $\mathcal{C}(\mathbf{X})$. Essa projecção é o vector $\mathbf{H}\vec{\mathbf{y}}^c$, onde a matriz $\mathbf{H}$ é a matriz de projecção ortogonal no referido subespaço, dada por $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$.




O outro cateto é dado pela diferença $\vec{\mathbf{y}}^c - \mathbf{H}\vec{\mathbf{y}}^c$. Sabemos que o comprimento da hipotenusa é $\|\vec{\mathbf{y}}^c\| = \sqrt{SQT}$. O do cateto que reside em $\mathcal{C}(\mathbf{X})$ é $\|\mathbf{H}\vec{\mathbf{y}}^c\| = \sqrt{SQR}$. O outro cateto tem comprimento $\|\vec{\mathbf{y}}^c - \mathbf{H}\vec{\mathbf{y}}^c\| = \sqrt{SQRE}$ (que é também igual a $\|\vec{\mathbf{y}} - \mathbf{H}\vec{\mathbf{y}}\|$). Assim, o Teorema de Pitágoras garante que $SQT = SQR + SQRE$ (ver a figura).

(c) Tem-se, tendo em conta a equação do modelo e o facto de $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$; e $\hat{\vec{\mathbf{Y}}} = \mathbf{H}\vec{\mathbf{Y}}$:

$$\begin{aligned}\vec{\mathbf{E}} &= \vec{\mathbf{Y}} - \hat{\vec{\mathbf{Y}}} = \vec{\mathbf{Y}} - \mathbf{H}\vec{\mathbf{Y}} = (\mathbf{I}-\mathbf{H})\vec{\mathbf{Y}} = (\mathbf{I}-\mathbf{H})(\mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\epsilon}}) = \mathbf{X}\vec{\boldsymbol{\beta}} - \mathbf{H}\mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\epsilon}} - \mathbf{H}\vec{\boldsymbol{\epsilon}} \\ &= \mathbf{X}\vec{\boldsymbol{\beta}} - \mathbf{X}\underbrace{(\mathbf{X}^t\mathbf{X})^{-1}(\mathbf{X}^t\mathbf{X})}_{=\mathbf{I}}\vec{\boldsymbol{\beta}} + (\mathbf{I}-\mathbf{H})\vec{\boldsymbol{\epsilon}} = \cancel{\mathbf{X}\vec{\boldsymbol{\beta}}} - \cancel{\mathbf{X}\vec{\boldsymbol{\beta}}} + (\mathbf{I}-\mathbf{H})\vec{\boldsymbol{\epsilon}} = (\mathbf{I}-\mathbf{H})\vec{\boldsymbol{\epsilon}}\,.\end{aligned}$$

A Soma de Quadrados dos Resíduos é dada por $SQRE = \|\vec{\mathbf{E}}\|^2 = \vec{\mathbf{E}}^t\vec{\mathbf{E}}$. Tendo em conta a alínea anterior, as propriedades de matrizes e ainda o facto das matrizes $\mathbf{I}$ e $\mathbf{H}$ serem simétricas e idempotentes, tem-se:

$$\begin{aligned}SQRE &= \vec{\mathbf{E}}^t\vec{\mathbf{E}} = [(\mathbf{I}-\mathbf{H})\vec{\boldsymbol{\epsilon}}]^t(\mathbf{I}-\mathbf{H})\vec{\boldsymbol{\epsilon}} = \vec{\boldsymbol{\epsilon}}^t(\mathbf{I}-\mathbf{H})^t(\mathbf{I}-\mathbf{H})\vec{\boldsymbol{\epsilon}} = \vec{\boldsymbol{\epsilon}}^t(\mathbf{I}^t-\mathbf{H}^t)(\mathbf{I}-\mathbf{H})\vec{\boldsymbol{\epsilon}} \\ &= \vec{\boldsymbol{\epsilon}}^t(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})\vec{\boldsymbol{\epsilon}} = \vec{\boldsymbol{\epsilon}}^t(\mathbf{I}-\mathbf{H}-\mathbf{H}+\underbrace{\mathbf{H}\mathbf{H}}_{=\mathbf{H}})\vec{\boldsymbol{\epsilon}} = \vec{\boldsymbol{\epsilon}}^t(\mathbf{I}-\mathbf{H})\vec{\boldsymbol{\epsilon}} = \vec{\boldsymbol{\epsilon}}^t\vec{\boldsymbol{\epsilon}} - \vec{\boldsymbol{\epsilon}}^t\mathbf{H}\vec{\boldsymbol{\epsilon}}\,.\end{aligned}$$

Ora a primeira parcela na expressão final ($\vec{\boldsymbol{\epsilon}}^t\vec{\boldsymbol{\epsilon}}$) é a soma de quadrados dos erros aleatórios. A segunda parcela não pode ser negativa, já que é uma norma (comprimento) ao quadrado: $\vec{\boldsymbol{\epsilon}}^t\mathbf{H}\vec{\boldsymbol{\epsilon}} = \vec{\boldsymbol{\epsilon}}^t\mathbf{H}\mathbf{H}\vec{\boldsymbol{\epsilon}} = \vec{\boldsymbol{\epsilon}}^t\mathbf{H}^t\mathbf{H}\vec{\boldsymbol{\epsilon}} = (\mathbf{H}\vec{\boldsymbol{\epsilon}})^t(\mathbf{H}\vec{\boldsymbol{\epsilon}}) = \|\mathbf{H}\vec{\boldsymbol{\epsilon}}\|^2$. Logo, tem de ter-se $SQRE = \vec{\mathbf{E}}^t\vec{\mathbf{E}} \leq \vec{\boldsymbol{\epsilon}}^t\vec{\boldsymbol{\epsilon}}$.

(d) O Modelo de Regressão Linear exige que $\vec{\boldsymbol{\epsilon}} \sim \mathcal{N}_n(\vec{\mathbf{0}}, \sigma^2\,\mathbf{I}_n)$. Como se viu acima, o vector dos resíduos obtém-se multiplicando, à esquerda, o vector dos erros aleatórios $\vec{\boldsymbol{\epsilon}}$ por $(\mathbf{I}-\mathbf{H})$. Sabemos que esse tipo de transformação dum vector Multinormal não afecta a Multinormalidade. Fica por saber quais os dois parâmetros (vector esperado e matriz de (co-)variâncias) associados a $\vec{\mathbf{E}} = (\mathbf{I}-\mathbf{H})\vec{\boldsymbol{\epsilon}}$. Pelas propriedades dos vectores esperados e matrizes de (co-)variâncias, tem-se:

$$E[\vec{\mathbf{E}}] = E[(\mathbf{I}-\mathbf{H})\vec{\boldsymbol{\epsilon}}] = (\mathbf{I}-\mathbf{H})E[\vec{\boldsymbol{\epsilon}}] = (\mathbf{I}-\mathbf{H})\vec{\mathbf{0}} = \vec{\mathbf{0}}.$$

e

$$\begin{aligned}V[\vec{\mathbf{E}}] &= V[(\mathbf{I}-\mathbf{H})\vec{\boldsymbol{\epsilon}}] = (\mathbf{I}-\mathbf{H})V[\vec{\boldsymbol{\epsilon}}](\mathbf{I}-\mathbf{H})^t = (\mathbf{I}-\mathbf{H})\cdot\sigma^2\mathbf{I}\cdot(\mathbf{I}^t-\mathbf{H}^t) = \sigma^2(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H}) \\ &= \sigma^2(\mathbf{I}-\mathbf{H}-\mathbf{H}+\underbrace{\mathbf{H}\mathbf{H}}_{=\mathbf{H}}) = \sigma^2(\mathbf{I}-\mathbf{H})\,.\end{aligned}$$

Assim, tem-se $\vec{\mathbf{E}} \frown \mathcal{N}_n\left(\vec{\mathbf{0}}, \sigma^2(\mathbf{I}-\mathbf{H})\right)$.