INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO

15 de Novembro, 2019 **PRIMEIRO TESTE 2019-20** **Uma resolução possível**

I

Os dados correspondem a uma tabela de contingências com $a=2$ linhas (correspondentes aos porta-enxertos) e $b=2$ colunas (correspondentes aos resultados). É pedido um teste de homogeneidade, para determinar se as enxertias ensaiadas em cada um dos porta-enxertos se distribuem de forma análoga pelos dois possíveis resultados ("vivos", na coluna 1, e "mortos" na coluna 2). Havendo homogeneidade, a probabilidade do resultado "vivos" (1) é igual em cada porta-enxertos, o mesmo sucedendo com a probabilidade do resultado "mortos" (2). Eis os passos do teste de hipótese pedido:

Hipóteses: $H_0 : \pi_{1|99R} = \pi_{1|110R} [= \pi_{.1}]$ e $\pi_{2|99R} = \pi_{2|110R} [= \pi_{.2}]$
vs. H_1 : pelo menos uma das igualdades de H_0 não se verifica.

Estatística do Teste: A estatística de Pearson, é dada por $X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}}$. A distribuição assintótica desta estatística, caso seja verdade H_0, é $\chi^2_{(a-1)(b-1)}$.

Nível de Significância Não sendo explicitado no enunciado, pode-se escolher
$\alpha = \mathrm{P}[\text{ Erro de tipo I }] = \mathrm{P}[\text{ Rejeitar } H_0 | H_0 \text{ verdade }] = 0.05$.

Região Crítica: (Unilateral direita) Para um nível de significância $\alpha = 0.05$, a regra de rejeição consiste em rejeitar H_0 se $\chi^2_{\text{calc}} > \chi^2_{0.05(1)} = 3.84146$.

Conclusões Para calcular o valor da estatística do teste são necessários os valores esperados estimados. Utilizamos a fórmula $\hat{E}_{ij} = \frac{N_{i.} \times N_{.j}}{N}$, sendo $N_{1.} = 1885$ o total marginal do porta-enxertos `99R`; $N_{2.} = 1972$ o total marginal do porta-enxertos `110R`; $N_{.1} = 463 + 690 = 1153$ o total marginal do resultado "vivas"; e $N_{.2} = 1422 + 1282 = 2704$ o total marginal do resultado "mortas". O total global das observações é $N = 1153 + 2704 = 3857$. Logo, os quatro valores esperados estimados são:

$$\hat{E}_{11} = \frac{1885 \times 1153}{3857} = 563.4962 \quad ; \quad \hat{E}_{12} = \frac{1885 \times 2704}{3857} = 1321.504$$
$$\hat{E}_{21} = \frac{1972 \times 1153}{3857} = 589.5038 \quad ; \quad \hat{E}_{22} = \frac{1972 \times 2704}{3857} = 1382.496$$

Note-se que todos os valores esperados estimados são muito superiores a 5, pelo que o Critério de Cochran é amplamente respeitado, não havendo hesitação na utilização da distribuição assintótica da estatística do teste que, no nosso caso, será χ^2_1.

O valor calculado da estatística de teste é

$$X^2_{calc} = \frac{(463 - 563.4962)^2}{563.4962} + \frac{(1422 - 1321.504)^2}{1321.504} + \frac{(690 - 589.5038)^2}{589.5038} + \frac{(1282 - 1382.496)^2}{1382.496} = 50.0027 \, .$$

Assim, tem-se uma clara rejeição de H_0, concluindo-se (ao nível $\alpha = 0.05$) pela inexistência de homogeneidade no comportamento dos dois porta-enxertos. O valor calculado da estatística de teste é (muito) superior ao quantil $1-\alpha$ da distribuição χ^2_1 para o menor nível de significância tabelado, $\alpha = 0.001$, pelo que concluir-se-ia pela rejeição mesmo a esse nível α. Assim, o valor de prova tem de ser (muito) inferior a 0.001.

II

1. A melhor variável preditora (x) é a variável `compDedo`, por ser a variável mais fortemente correlacionada com a variável resposta (y) `compAsa` (tendo-se a correlação $r_{xy}=0.830$).

   (a) O declive da recta ajustada é dado por $b_1 = \frac{cov_{xy}}{s_x^2} = r_{xy}\,\frac{s_y}{s_x} = 0.830\,\frac{41.0781}{8.7377} = 3.902036348$. A ordenada na origem é dada por $b_0 = \overline{y} - b_1\,\overline{x} = 94.2814 - (3.902036348)(21.6434) = 9.828066501$. Logo, com arredondamentos, a equação da recta de regressão ajustada é $y = 9.8281 + 3.9020\,x$. O Coeficiente de Determinação correspondente é $R^2 = 0.830^2 = 0.6889$. Assim, verifica-se uma relação crescente entre comprimento do dedo médio da pata e comprimento da asa, em que para cada mm adicional com comprimento do dedo corresponde, em média, um aumento de 3.9020 mm no comprimento da asa. Esta recta de regressão ajustada explica quase 69% da variabilidade observada nos comprimentos das asas entre espécies.

   (b) A Soma de Quadrados Total é dada por $SQT = (n-1)\,s_y^2 = 128 \times 41.0781^2 = 215\,988.5184$. A Soma de Quadrados da Regressão pode ser calculada a partir da definição do coeficiente de determinação, $R^2 = \frac{SQR}{SQT}$, tendo-se $SQR = R^2\,SQT = 0.6889 \times 215\,988.5184 = 148\,794.4903$. Finalmente, a Soma de Quadrados Residual resulta da fórmula fundamental da regressão, $SQT = SQR + SQRE$, sendo $SQRE = SQT - SQR = 67\,194.02811$.

   (c) Pede-se um intervalo de confiança para o valor esperado de Y (comprimento da asa) quando o preditor (comprimento do dedo) toma o valor $x=15$, ou seja, para $\mu_{Y|x=15}$. A expressão geral do intervalo a $(1-\alpha)\times 100\%$ de confiança é análoga à dos intervalos de predição, dada no formulário, mas sem a parcela "1+", ou seja, é:

   $$\left]\ (b_0+b_1x) - t_{\frac{\alpha}{2};n-2}\cdot\sqrt{QMRE\cdot\left[\frac{1}{n}+\frac{(x-\overline{x})^2}{(n-1)\,s_x^2}\right]}\ ,\ (b_0+b_1x) + t_{\frac{\alpha}{2};n-2}\cdot\sqrt{QMRE\cdot\left[\frac{1}{n}+\frac{(x-\overline{x})^2}{(n-1)\,s_x^2}\right]}\ \right[\ .$$

   Já conhecemos os valores $b_0 = 9.828066501$; $b_1 = 3.902036348$; $n = 129$, $\overline{x} = 21.6434$; $s_x^2 = (8.7377)^2 = 76.34740129$; $\sqrt{QMRE} = \sqrt{\frac{SQRE}{n-2}} = 23.00188765$, utilizando o valor de $SQRE$ obtido na alínea anterior. Logo, $\hat{\sigma}_{\hat{\mu}_{y|x=15}} = \sqrt{QMRE\cdot\left[\frac{1}{n}+\frac{(x-\overline{x})^2}{(n-1)\,s_x^2}\right]} = 2.547730133$. Finalmente, pela leitura das tabelas, $t_{0.025(127)} \approx 1.98$. Usando estes valores, temos o intervalo (a 95% de confiança)] 63.3141 , 73.4031 [.

   Assim, espécies com comprimento do dedo médio da pata igual a 15 mm terão, em média, comprimentos de asa entre 63.3141 mm e 73.4031 mm. Do ponto de vista gráfico, podemos afirmar com 95% de confiança que a recta populacional atravessa o intervalo que, na vertical sobre $x=15$, contém os valores de y pertencentes ao intervalo.

2. (a) O gráfico da esquerda tem no eixo vertical os valores dos resíduos usuais (e_i) e no eixo horizontal os valores ajustados de y ($\hat{y}_i$). Caso a nuvem de pontos neste gráfico se disperse essencialmente numa banda horizontal em torno do valor médio dos resíduos, $\overline{e}=0$, não haverá razões para duvidar dos pressupostos de linearidade, nem de variâncias homogéneas dos erros aleatórios. No nosso caso é visível um padrão em forma de funil, que sugere heterogeneidade nas variâncias dos erros aleatórios, violando-se assim um dos pressupostos do modelo linear.

   No gráfico da direita temos os resíduos estandardizados (R_i) no eixo vertical. É visível que a algumas observações correspondem resíduos estandardizados bastante grandes, próximos (em valor absoluto) a 4. É o caso, em particular, das duas observações no canto inferior direito assinaladas como `E078` e `E079`, bem como pelo menos uma observação (sem legenda)

no topo esquerdo do gráfico. Trata-se de observações bastante afastadas da recta ajustada, que mereceriam ulterior atenção. No eixo horizontal deste segundo gráfico temos os valores do efeito alavanca (h_{ii}). Sabemos que estes valores têm de estar entre $\frac{1}{n} = \frac{1}{129} = 0.00775$ e 1, com valor médio $\overline{h} = \frac{2}{n} = 0.0155038$. Apesar de haver algumas observações com efeito alavanca cerca de 4 vezes maior que o valor médio $\overline{h}$, nenhum desses valores está remotamente próximo do valor máximo 1, pelo que não existem efeitos alavanca dignos de registo. Finalmente, são visíveis nos cantos à direita as isolinhas de distâncias de Cook $D_i = 0.5$. As distâncias de Cook assinalam a influência de cada observação, ou seja, o impacto que a sua exclusão teria sobre o ajustamento da recta. A maiores distâncias de Cook corresponde maior influência, sendo a condição $D_i > 0.5$ usada frequentemente para identificar observações com influência excessiva. No nosso caso não existe qualquer observação com distância de Cook superior a 0.5, mas 3 observações têm uma influência considerável: as observações legendadas como `E078`, `E079` e `E001`.

(b) Um valor aproximado da distância de Cook pode ser obtido a partir da fórmula $D_i = R_i^2 \cdot \frac{h_{ii}}{1-h_{ii}} \cdot \frac{1}{2}$, lendo-se nos eixos os valores aproximados de R_i e h_{ii}. Para a observação `E079` tem-se $h_{E079,E079} \approx 0.06$ e $R_{E079} \approx -3.5$. Logo, $D_{E079} \approx (-3.5)^2 \times \frac{0.06}{1-0.06} \times 0.5 = 0.3909574 \approx 0.4$.

3. Neste ponto considera-se o modelo linear entre `log(compAsa)` e `log(peso)`.

(a) Há dois aspectos a referir. Por um lado o 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: $F = \frac{QMR}{QMRE} = (n-2)\frac{R^2}{1-R^2} \underset{H_0}{\sim} F_{(1,n-2)}$, sob H_0.

Nível de significância: $\alpha = 0.05$.

Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05[1,127]}$, pelas tabelas, é um valor próximo do valor tabelado $f_{0.05[1,120]} = 3.92$.

Conclusões: No enunciado é dado o valor calculado da estatística, $F_{calc} = 672$. A rejeição de H_0 é muito clara, pelo que o modelo ajustado é muito significativamente diferente do Modelo Nulo (o correspondente valor de prova é inferior à precisão de máquina, ou seja, $p < 2.2 \times 10^{-16}$).

Este resultado era previsível, dado o valor relativamente elevado do Coeficiente de Determinação, $R^2 = 0.8411$, que indica que quase 85% da variabilidade da variável resposta é explicada pela regressão. Mas deve ter-se em atenção que a variável resposta é o *logaritmo* dos comprimentos da asa, razão pela qual a frase constante no enunuciado não é correcta. O que se pode afirmar é que quase 85% da variância *dos logaritmos* do comprimento da asa é explicada pela regressão.

(b) Nas escalas originais de x e y, a regressão linear agora ajustada corresponde a uma curva potência:

$$\ln(y) = b_0 + b_1 \ln(x) \quad \Leftrightarrow \quad y = \mathrm{e}^{b_0 + b_1 \ln(x)} \quad \Leftrightarrow \quad y = \mathrm{e}^{b_0}\mathrm{e}^{\ln(x^{b_1})} \quad \Leftrightarrow \quad y = \mathrm{e}^{b_0} x^{b_1} \ .$$

No nosso caso concreto, a curva potência ajustada (que é a curva visível no gráfico do enunciado) tem equação $y = \mathrm{e}^{3.39058} x^{0.32275} = 29.68316\, x^{0.32275}$.

(c) Na alínea anterior viu-se que, na amostra, o comprimento da asa é proporcional à potência $b_1 = 0.32275$ do peso do pássaro, que é um valor muito próximo de $\frac{1}{3}$. Assim, é legítimo perguntar se é admissível a regra simples de que o comprimento da asa é proporcional à potência $\frac{1}{3} = 0.333333$ do peso. Tal pergunta corresponde a um teste de hipóteses a $H_0 : \beta_1 = \frac{1}{3}$, contra a hipótese alternativa complementar.

Hipóteses: $H_0 : \beta_1 = \frac{1}{3} \quad vs. \quad H_1 : \beta_1 \neq \frac{1}{3}$.

Estatística do Teste: É dada por $T = \frac{\hat{\beta}_1 - \beta_{1|H_0}}{\hat{\sigma}_{\hat{\beta}_1}}$, com distribuição t_{n-2} caso o valor de β_1 tenha o valor da Hipótese Nula (ou seja, se $\beta_{1|H_0} = \frac{1}{3}$).

Nível de Significância Defina-se $\alpha = 0.05$.

Região Crítica: (Bilateral) A regra de rejeição consiste em rejeitar H_0 no caso de $|T_{\text{calc}}| > t_{0.025(127)} \approx 1.980$.

Conclusões O valor calculado da estatística é $T_{calc} = \frac{0.32275 - 0.33333}{0.01245} = -0.8497992$. Este valor encontra-se fora da região de rejeição, pelo que não se rejeita H_0 (ao nível $\alpha = 0.05$). Assim, a informação existente é compatível com a hipótese formulada no enunciado.

Sabemos ainda que uma relação potência surge de admitir que ambas as variáveis (x e y, nas escalas originais) são função dum terceiro variável (como o tempo t), e que as respectivas taxas de variação relativas são proporcionais. Mais concretamente, surge de admitir que se tem $\frac{y'(t)}{y(t)} = d\,\frac{x'(t)}{x(t)}$. Sabemos ainda que a constante de proporcionalidade d corresponde à potência b_1. Assim, temos que a taxa de variação relativa do comprimento da asa é cerca de um terço da taxa de variação relativa do peso.

(d) A observação no canto superior direito do gráfico é a observação a que corresponde o maior peso e também o maior comprimento de asa. Estes valores máximos constam do enunciado, e são respectivamente: $x_{max} = 500.0$ e $y_{max} = 255.0$. Na regressão linear ajustada, o resíduo corresponde à diferença entre *o logaritmo* deste valor observado de y, $y^* = \ln(255) = 5.541264$, e o valor que, através da recta ajustada, corresponde a y quando *o logaritmo* de x toma o valor $x^* = \ln(500) = 6.214608$. Tem-se assim que esse residuo (que será designado por e_{cs}, onde `cs` indica "canto superior") é dado por $e_{cs} = y^*_{cs} - \widehat{y^*}_{cs} = \ln(y_{cs}) - (3.39058 + 0.32275\,\ln(x_{cs})) = 5.541264 - (3.39058 + 0.32275 \times 6.214608) = 5.541264 - 5.396345 = 0.144919$. No entanto a distância na vertical entre o ponto e a curva, nas escala original de y (mm) **não** é obtida exponenciando este valor. Pode ser calculada como a diferença entre $y_{max} = 255$ e a exponencial do valor ajustado $\hat{y}^*_{cs} = 5.396345$, ou seja: $y_{max} - \mathrm{e}^{\hat{y}^*_{cs}} = 255 - \mathrm{e}^{5.396345} = 255 - 220.5987 = 34.40135$.

III

1. Estamos no contexto do Modelo de Regressão Linear Simples.

(a) Pede-se para deduzir a distribuição de $\hat{\mu}_{Y|x} = \hat{\beta}_0 + \hat{\beta}_1\,x$. A primeira constatação é que sendo $\hat{\beta}_0$ e $\hat{\beta}_1$ ambas combinações lineares dos $Y_i = \beta_0 + \beta_1\,x_i + \epsilon_i$, também $\hat{\mu}_{Y|x}$ é uma combinação linear dos Y_i. Sabemos que, dado o Modelo de Regressão Linear Simples, os $\{Y_i\}_{i=1}^n$ são um conjunto de variáveis aleatórias Normais (porque são uma transformação linear das Normais ϵ_i) e independentes (porque os erros aleatórios ϵ_i são independentes). Como qualquer combinação linear de Normais independentes tem distribuição Normal, sai imediatamente que $\hat{\mu}_{Y|x}$ tem uma distribuição Normal. Falta determinar os respectivos parâmetros. Com base nas propriedades dos valores esperados, variâncias e covariâncias, sabendo que $\hat{\beta}_0$ e $\hat{\beta}_1$ são estimadores centrados, e usando as fórmulas das variâncias e

covariância desses estimadores, que constam do formulário, tem-se:

$$\begin{aligned}
E[\hat{\mu}_{Y|x}] &= E[\hat{\beta}_0 + \hat{\beta}_1\, x] = E[\hat{\beta}_0] + E[\hat{\beta}_1\, x] = \beta_0 + x\, E[\hat{\beta}_1] = \beta_0 + \beta_1\, x \\
V[\hat{\mu}_{Y|x}] &= V[\hat{\beta}_0 + \hat{\beta}_1\, x] = V[\hat{\beta}_0] + V[\hat{\beta}_1\, x] + 2Cov[\hat{\beta}_0, \hat{\beta}_1\, x] = V[\hat{\beta}_0] + x^2\, V[\hat{\beta}_1] + 2x\, Cov[\hat{\beta}_0, \hat{\beta}_1] \\
&= \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{(n-1)\, s_x^2}\right] + x^2\, \frac{\sigma^2}{(n-1)\, s_x^2} + 2x\, \frac{-\overline{x}\sigma^2}{(n-1)\, s_x^2} \\
&= \sigma^2 \left[\frac{1}{n} + \frac{x^2 + \overline{x}^2 - 2x\overline{x}}{(n-1)\, s_x^2}\right] = \sigma^2 \left[\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)\, s_x^2}\right] .
\end{aligned}$$

Logo, $\hat{\mu}_{Y|x} \frown \mathcal{N}\left(\ \beta_0 + \beta_1\, x\ ,\ \sigma^2 \left[\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)\, s_x^2}\right]\right)$.

(b) As semi-amplitudes de cada um dos intervalos são dadas pelo produto do quantl da distribuição t-Student, $t_{\frac{\alpha}{2}(n-2)}$ e do erro padrão de cada estimador, $\hat{\sigma}_{\hat{\beta}_0}$ e $\hat{\sigma}_{\hat{\beta}_1}$. Usando o mesmo grau de confiança nos dois casos, são apenas estes erros padrão que distinguem as semi-amplitudes. Tomando a razão dos quadrados desses erros padrão (para não ter de trabalhar com as raízes quadradas), tem-se:

$$\frac{\hat{\sigma}^2_{\hat{\beta}_1}}{\hat{\sigma}^2_{\hat{\beta}_0}} = \frac{\frac{\cancel{QMRE}}{(n-1)\, s_x^2}}{\cancel{QMRE}\left[\frac{1}{n} + \frac{\overline{x}^2}{(n-1)\, s_x^2}\right]} = \frac{\frac{1}{\cancel{(n-1)\, s_x^2}}}{\frac{1}{n\cdot\cancel{(n-1)\, s_x^2}}\left[(n-1)\, s_x^2 + n\, \overline{x}^2\right]} = \frac{n}{(n-1)\, s_x^2 + n\, \overline{x}^2} = \frac{n}{\sum_{i=1}^{n} x_i^2} ,$$

já que a variância amostral de x se pode escrever como $(n-1)\, s_x^2 = \sum_{i=1}^{n} x_i^2 - n\, \overline{x}^2$. Assim, o intervalo de confiança para β_1 terá menor amplitude que o de β_0 se e só se $n < \sum_{i=1}^{n} x_i^2$. Esta condição depende das unidades de medida de x e, por exemplo, uma conversão dessas unidades de metros para centimetros aumentaria o valor de $\sum_{i=1}^{n} x_i^2$ em $100^2 = 10\,000$ vezes. Assim, pode acontecer que para uma mesma amostra, o intervalo de confiança de β_1 seja maior ou menor de que o intervalo de confiança para β_0, consoante as unidades de medida usadas na medição de x. Esta constatação não deve surpreender, uma vez que o valor do próprio parâmetro β_1 depende das unidades de medida de x (já que é expresso no quociente das unidades de medida de y sobre as unidades de medida de x).

2. Agora no contexto duma regressão linear múltipla descritiva.

(a) Por definição, a matriz do modelo $\mathbf{X}$ é a matriz de dimensões $n \times (p+1)$, cuja primeira coluna é constituída pelo vector $\vec{\mathbf{1}}_n$ (cujos n elementos são todos 1) e cada uma das restantes p colunas, $\vec{\mathbf{x}}_j$ ($j = 1, 2, ..., p$) contém os n valores observados da j-ésima variáveis preditora. Também por definição, o espaço das colunas da matriz $\mathbf{X}$ é o conjunto de todas as possíveis combinações lineares das colunas de $\mathbf{X}$, ou seja, de $a_0\, \vec{\mathbf{1}}_n + a_1 \vec{\mathbf{x}}_1 + a_2\, \vec{\mathbf{x}}_2 + ... + a_p \vec{\mathbf{x}}_p$, para qualquer conjunto de coeficientes $a_0, a_1, a_2, ..., a_p$. Este conjunto é um subespaço de $\mathbb{R}^n$, representado por $\mathcal{C}(\mathbf{X})$.

(b) A média $\overline{y}$ dos valores observados de Y é dada por $\overline{y} = \frac{1}{n} \vec{\mathbf{1}}_n^t \vec{\mathbf{y}}$, onde $\vec{\mathbf{y}}$ indica o vector das n observações da variável resposta. De facto, o produto interno $\vec{\mathbf{1}}_n^t \vec{\mathbf{y}}$ devolve a soma dos elementos do vector dessas observações, $\vec{\mathbf{y}}$.

Ora, sabemos que o vector dos valores ajustados é dado por $\vec{\hat{\mathbf{y}}} = \mathbf{H}\vec{\mathbf{y}}$. Assim, a média dos n valores ajustados, que podemos representar por $\overline{\hat{y}} = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i$, pode ser calculado tomando o produto interno do vector $\vec{\mathbf{1}}_n$ de n uns com o vector $\vec{\hat{\mathbf{y}}}$, uma vez que esse produto interno devolve a soma dos elementos de $\vec{\hat{\mathbf{y}}}$. Assim, a média dos valores ajustados

é $\overline{\hat{y}} = \frac{1}{n}\vec{1}_n^t\hat{\vec{y}} = \frac{1}{n}\vec{1}_n^t\mathbf{H}\vec{y} = \frac{1}{n}(\mathbf{H}^t\vec{1}_n)^t\vec{y} = \frac{1}{n}(\mathbf{H}\vec{1}_n)^t\vec{y} = \frac{1}{n}\vec{1}_n^t\vec{y}$, uma vez que $\mathbf{H}$ é simétrica ($\mathbf{H}^t = \mathbf{H}$) e $\mathbf{H}\vec{1}_n = \vec{1}_n$, já que que o vector $\vec{1}_n$ pertence ao subespaço das colunas de $\mathbf{X}$, logo fica invariante quando projectado nesse subespaço. Mas a expressão final obtida, $\frac{1}{n}\vec{1}_n^t\vec{y}$, é a média $\overline{y}$ dos valores observados de y. Assim, também na regressão linear múltipla, valores observados de y e correspondentes valores ajustados partilham o mesmo valor médio.

(c) O centro de gravidade da nuvem dos n pontos correspondentes às observações é, por definição, o ponto cujas coordenadas são as médias das $p+1$ variáveis observadas, ou seja, o ponto $(\overline{x}_1, \overline{x}_2, ..., \overline{x}_p, \overline{y})$. Ora, na representação tradicional (no espaço $\mathbb{R}^{p+1}$), o hiperplano de mínimos quadrados tem equação $y = b_0 + b_1\,x_1 + b_2\,x_2 + ... + b_p\,x_p$. Os valores ajustados de y, $\hat{y}_i$, obtêm-se substituindo os correspondentes valores dos preditores nessa equação, ou seja, $\hat{y}_i = b_0 + b_1\,x_{1(i)} + b_2\,x_{2(i)} + ... + b_p\,x_{p(i)}$. Logo, a média dos n valores ajustados é dada por

$$\begin{aligned}
\overline{\hat{y}} \quad = \quad \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i \quad &= \quad \frac{1}{n}\sum_{i=1}^{n}(b_0 + b_1\,x_{1(i)} + b_2\,x_{2(i)} + ... + b_p\,x_{p(i)}) \\
&= \quad \frac{1}{n}\sum_{i=1}^{n}b_0 + \frac{1}{n}\sum_{i=1}^{n}b_1\,x_{1(i)} + \frac{1}{n}\sum_{i=1}^{n}b_2\,x_{2(i)} + ... + \frac{1}{n}\sum_{i=1}^{n}b_p\,x_{p(i)} \\
&= \quad \frac{1}{\cancel{n}}\times\cancel{n}\,b_0 + b_1\,\frac{1}{n}\sum_{i=1}^{n}x_{1(i)} + b_2\,\frac{1}{n}\sum_{i=1}^{n}x_{2(i)} + ... + b_p\,\frac{1}{n}\sum_{i=1}^{n}x_{p(i)} \\
\Leftrightarrow \quad \overline{y} \quad &= \quad b_0 + b_1\,\overline{x}_1 + b_2\,\overline{x}_2 + ... + b_p\,\overline{x}_p
\end{aligned}$$

Esta última equação (que resulta da igualdade $\overline{\hat{y}} = \overline{y}$ provada na alínea anterior), significa que o ponto $(\overline{x}_1, \overline{x}_2, ..., \overline{x}_p, \overline{y})$ satisfaz a equação, ou seja, o centro de gravidade é um dos pontos do hiperplano ajustado.