INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO

5 de Novembro, 2021 **PRIMEIRO TESTE 2020-21** (40%) **Uma resolução possível**

I

1. Regressão linear simples de rendimento sobre peso dos sarmentos.

   (a) O declive da recta ajustada é dado por $b_1 = \frac{cov_{xy}}{s_x^2} = r_{xy}\,\frac{s_y}{s_x} = 0.9146154 \times \frac{0.5528155}{0.3978292} = 1.270931$. Este valor indica a variação média estimada no rendimento, em kg/planta, por um hectograma/planta adicional no peso dos sarmentos. A ordenada na origem é dada por $b_0 = \overline{y} - b_1\,\overline{x} = 3.368 - 1.270931 \times 2.425 = 0.2859923$ kg/planta. Assim, a equação da recta estimada é $y = 0.286 + 1.271\,x$.

   (b) A correlação entre preditor e variável resposta é $r_{xy} = 0.9146154$. A proporção de variabilidade dos rendimentos observados que pode ser explicada por essa regressão é $R^2 = (r_{xy})^2 = 0.9146154^2 = 0.8365213$. Assim, cerca de 84% da variabilidade observada nos rendimentos será explicada pela recta de regressão sobre o peso dos sarmentos. Este valor é significativamente diferente de zero, como se comprova pelo teste F de ajustamento global:

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

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

   Nível de significância: $\alpha = 0.05$.

   Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05[1,32]}$. Vamos aproximar esta fronteira da Região Crítica pelo valor tabelado $f_{0.05[1,30]} \approx 4.17$.

   Conclusões: O valor calculado da estatística é $F_{calc} = 32 \times \frac{0.8365213}{1-0.8365213} = 163.7442$. A rejeição de H_0 é claríssima, logo o modelo ajustado é muito significativamente diferente do Modelo Nulo, como era de esperar, dado o valor elevado de R^2.

   (c) A Soma de Quadrados Total obtém-se pela fórmula $SQT = (n-1) \times s_y^2 = 33 \times (0.5528155)^2 = 10.08496$. A Soma de Quadrados da Regressão pode obter-se a partir da definição do Coeficiente de Determinação, $R^2 = \frac{SQR}{SQT}$, uma vez que este é conhecido. Assim, tem-se $SQR = R^2 \times SQT = 0.8365213 \times 10.08496 = 8.436287$. A última das três Somas de Quadrados obtém-se, a partir da fórmula fundamental da regressão linear, como $SQRE = SQT - SQR = 10.08496 - 8.436287 = 1.648673$.

   (d) Aumentar em dez gramas o peso por planta dos sarmentos corresponde a aumentar esse peso do preditor em 0.1 hectogramas. Sabemos que a cada aumento em c unidades da variável preditora corresponde uma variação média estimada na variável resposta de $c \times b_1$ unidades. Como $c = 0.1$ e $b_1 = 1.270931$, tem-se uma variação média estimada no rendimento de 0.1270931 kg/planta por cada dez gramas adicionais no peso/planta dos sarmentos.

   (e) O enunciado pergunta se é admissível considerar que $\beta_1 > 1$. Vamos efectuar um teste de hipóteses que exija o ónus da prova a esta hipótese, ou seja, que a coloque como Hipótese Alternativa:

   Hipóteses: $H_0 : \beta_1 \leq 1 \quad vs. \quad H_1 : \beta_1 > 1$.

   Estatística do Teste: $T = \frac{\hat{\beta}_1 - \beta_{1|H_0}}{\hat{\sigma}_{\hat{\beta}_1}} \frown t_{n-2}$, sob H_0.

   Nível de significância: $\alpha = 0.05$.

Região Crítica: Unilateral direita. A Regra de Rejeição é rejeitar H_0 se $T_{calc} > t_{0.05[32]}$. Este último valor está entre os valores tabelados $t_{0.05[30]} = 1.69726$ e $t_{0.05[40]} = 1.68385$.

Conclusões: O valor calculado da estatística na nossa amostra é $t_{calc} = \frac{b_1 - 1}{\hat{\sigma}_{\hat{\beta}_1}}$. Sabe-se que $b_1 = 1.270931$; $n - 1 = 33$, $s_x^2 = 0.3978292^2 = 0.1582681$; e que $QMRE = \frac{SQRE}{n-2} = \frac{1.648673}{32} = 0.05152103$. Logo, $\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{QMRE}{(n-1)\, s_x^2}} = \sqrt{\frac{0.05152103}{33 \times 0.1582681}} = \sqrt{0.009864548} = 0.09932044$. Assim, $t_{calc} = \frac{1.270931 - 1}{0.09932044} = 2.727847 > 1.69726$. Assim, rejeita-se H_0 em favor de $H_1 : \beta_1 > 1$, ao nível de significância $\alpha = 0.05$. O declive da recta ajustada é significativamente maior que 1.

(f) O gráfico mostrado tem, no eixo vertical, os valores dos resíduos (usuais) e_i e no eixo horizontal os valores ajustados $\hat{y}_i$. Neste gráfico pode estudar-se a validade dos pressupostos de linearidade de fundo e de homogeneidade das variâncias dos erros aleatórios, em cujo caso os pontos se devem dispersar essencialmente numa banda horizontal em torno do valor zero (que é o valor médio dos resíduos). No nosso caso, é visível uma tendência para a dispersão dos resíduos ir aumentando à medida que se caminha da parte esquerda para a parte direita do gráfico (o chamado 'efeito funil'). Essa tendência indicia a violação do pressuposto de variâncias constantes dos erros. Não havendo curvilinearidade aparente na nuvem de pontos, o pressuposto duma tendência de fundo linear é admissível. Finalmente, nenhuma observação tem um resíduo muito maior que os restantes.

2. A regressão com todas as variáveis logaritmizadas.

(a) A afirmação não é correcta. Uma vez que neste modelo a variável resposta foi logaritmizada, o valor de R^2 indica a proporção da variabilidade dos *log*-rendimentos que é explicada pela regressão, e não a proporção da variabilidade dos rendimentos, como no modelo da pergunta 1. Os dois valores não são directamente comparáveis.

(b) Uma vez que ambas as variáveis nesta regressão linear foram logaritmizadas, a relação não linear entra as variáveis originais correspondente ao ajustamento agora feito é uma relação potência. Concretamente, e exponenciando a recta ajustada, tem-se:

$$\begin{aligned} \ln(y) &= 0.41536 + 0.90085 \cdot \ln(x) \\ \Leftrightarrow \quad y &= \mathrm{e}^{0.41536 + 0.90085 \cdot \ln(x)} = \mathrm{e}^{0.41536} \cdot \mathrm{e}^{0.90085 \cdot \ln(x)} \\ \Leftrightarrow \quad y &= 1.514916 \cdot \mathrm{e}^{\ln(x^{0.90085})} = 1.514916 \cdot x^{0.90085} \end{aligned}$$

(c) Tendo em conta a curva potência ajustada, que estima uma curva potência populacional com equação $y = \mathrm{e}^{\beta_0}\, x^{\beta_1}$, y e x serão directamente populacionais se $\beta_1 = 1$. Vamos construir um intervalo a 95% de confiança para β_1 a fim de validar essa hipótese. Sabemos que a forma geral do IC a $(1-\alpha) \times 100\%$ de confiança é: $]\, b_1 - t_{\frac{\alpha}{2}(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_1}\ ,\ b_1 + t_{\frac{\alpha}{2}(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_1}\, [$. No enunciado pode ver-se que $b_1 = 0.90085$ e $\hat{\sigma}_{\hat{\beta}_1} = 0.06572$. Como $t_{0.025(32)} \approx 2.04227$, tem-se o intervalo de confiança a 95% (aproximadamente) dado por: $]\ 0.7666\ ,\ 1.0351\ [$. Assim, e embora por pouco, o IC contém o valor 1. Logo, a proporcionalidade entre rendimento e peso dos sarmentos não pode ser excluída, a 95% de confiança.

(d) Neste gráfico, os valores do eixo vertical correspondem aos resíduos estandardizados, R_i, enquanto que os valores no eixo horizontal correspondem aos valores do efeito alavanca, h_{ii}. As curvas que são visíveis nos cantos direitos correspondem às isolinhas de distâncias de Cook iguais a 0.5 (e resultam de, na equação $D_i = R_i^2 \frac{h_{ii}}{1 - h_{ii}} \frac{1}{2}$, igualar $D_i = 0.5$). Constata-se que nenhuma observação tem distância de Cook maior que esse limiar de guarda, o que indica que não observações particularmente influentes, ou seja, observações que, caso fossem

(individualmente) retiradas do conjunto de dados, provocassem grandes alterações na recta ajustada. Igualmente, nenhuma observação tem resíduos estandardizados superiores (em valor absoluto) a 3, pelo que não existem observações afastadas de forma anómala da recta ajustada. Quanto ao efeito alavanca h_{ii}, mede o que poderíamos designar a 'força de atracção' dum ponto sobre a recta, já que a variância dos resíduos (usuais) é $V[E_i] = \sigma^2\,(1-h_{ii})$, pelo que valores grandes de h_{ii} (que está sempre contido no intervalo $\left[\frac{1}{n}, 1\right]$) têm pequena variabilidade em torno do seu valor médio zero, ou seja, correspondem a pontos que têm de estar próximos da recta. A observação mais à direita no gráfico é a observação com maior efeito alavanca. Uma vez que na regressão linear simples se tem $h_{ii} = \frac{1}{n} + \frac{(x_i^* - \overline{x^*})^2}{(n-1)\,s_{x^*}^2}$, sabemos que esse maior efeito alavanca tem de corresponder à observação em que o preditor log-peso dos sarmentos toma o valor mais afastado da média $\overline{x^*} = 0.8724938$, de entre todas as observações. Para identificar qual essa observação, registamos que tem de ser uma de duas observações: ou a observação com o menor valor de x_i (logo de $x_i^* = \ln(x_i)$), que no enunciado se verifica ser 1.680, com $x_i^* = \ln(1.680) = 0.5187938$; ou a observação com o maior valor de x_i, que é 3.150, e para a qual $x_i^* = \ln(3.150) = 1.147402$. O valor, de entre estes dois, mais afastado da média 0.8724938 é o primeiro ($0.8724938 - 0.5187938 = 0.3537$). Logo, a observação a que corresponde esse ponto mais à direita no gráfico é a observação com o menor valor observado de peso dos sarmentos.

II

1. Modelo de Regressão Linear Simples.

   (a) O modelo de regressão linear simples é ajustado com base em n pares de observações $\{(x_i, Y_i)\}_{i=1}^n$, sendo Y_i variáveis aleatórias que indicam as observações da variável resposta e x_i os correspondentes valores da variável preditora, que se admite serem fixados pelo experimentador (não aleatórios). Admite-se ainda que: (i) $Y_i = \beta_0 + \beta_1\, x_i + \epsilon_i$ é a equação do modelo; (ii) $\epsilon_i \frown \mathcal{N}(0, \sigma^2)$, para todo o i; e (iii) $\{\epsilon_i\}_{i=1}^n$ são variáveis aleatórias independentes. As variáveis aleatórias Y_i são assim transformações lineares dos erros aleatórios (ϵ_i), obtidas somando parcelas não aleatórias ($\beta_0 + \beta_1\, x_i$). Como os erros aleatórios têm distribuição Normal, as propriedades da distribuição Normal garantem que cada Y_i também terá distribuição Normal (transformações lineares de Normais também têm distribuição Normal). O valor médio e variância de cada Y_i podem calcular-se a partir das propriedades operatórias desses indicadores. Tem-se:

   $$E[Y_i] = E[\beta_0 + \beta_1\, x_i + \epsilon_i] = \beta_0 + \beta_1\, x_i + \underbrace{E[\epsilon_i]}_{=0} = \beta_0 + \beta_1\, x_i$$

   De forma análoga,

   $$V[Y_i] = V[\beta_0 + \beta_1\, x_i + \epsilon_i] = V[\epsilon_i] = \sigma^2\ .$$

   Logo, $Y_i \frown \mathcal{N}(\,\beta_0 + \beta_1\, x_i\,,\ \sigma^2\,)$. Embora as observações tenham todas distribuição Normal e variâncias homogéneas, não são identicamente distribuídas, pois o seu valor médio varia consoante o correspondente valor do preditor, x_i.

   (b) Pede-se para deduzir a fórmula (que consta do formulário) para $V[\hat{\beta}_1]$. Sabemos pelo formulário que $\hat{\beta}_1 = \sum_{i=1}^{n} c_i\, Y_i$, com $c_i = \frac{x_i - \overline{x}}{(n-1)\,s_x^2}$. Uma vez que, dado o Modelo, as observações Y_i (que são transformações lineares dos erros aleatórios ϵ_i) são independentes, o somatório pode passar para fora da variância. Também as constantes multiplicativas c_i podem passar,

elevadas ao quadrado, para fora da variância (mas não para fora do somatório!). Assim, e tendo em conta que $V[Y_i]=\sigma^2$, $\forall\, i$, vem:

$$\begin{aligned} V[\hat{\beta}_1] &= V\left[\sum_{i=1}^{n} c_i\, Y_i\right] = \sum_{i=1}^{n} c_i^2\, V[Y_i] = \sum_{i=1}^{n} c_i^2\, \sigma^2 = \sigma^2 \sum_{i=1}^{n} c_i^2 = \sigma^2 \sum_{i=1}^{n} \left[\frac{x_i - \overline{x}}{(n-1)\, s_x^2}\right]^2 \\ &= \frac{\sigma^2}{[(n-1)\, s_x^2]^2} \underbrace{\sum_{i=1}^{n} (x_i - \overline{x})^2}_{=(n-1)\, s_x^2} = \sigma^2 \frac{\cancel{(n-1)\, s_x^2}}{[(n-1)\, s_x^2]^{\cancel{2}}} = \frac{\sigma^2}{(n-1)\, s_x^2}, \end{aligned}$$

como indicado no formulário.

2. (a) Numa regressão linear simples, o efeito alavanca tem a expressão dada no formulário: $h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{(n-1)\, s_x^2}$. Apenas o numerador da segunda parcela varia de observação para observação (depende de i), e o menor valor que pode tomar será zero, quando $x_i = \overline{x}$. Assim, o efeito alavanca alcança o seu menor valor possível ($\frac{1}{n}$) numa observação com valor x_i do preditor igual à média dos valores observados de x. Não havendo uma observação que satisfaça essa condição, será a observação com o valor x_i mais próxima de $\overline{x}$ que terá o menor valor do efeito alavanca.

   (b) Pela expressão do efeito alavanca acima indicada, o valor médio dos h_{ii} é dada por:

$$\begin{aligned} \overline{h} &= \frac{1}{n} \sum_{i=1}^{n} h_{ii} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{n} + \frac{(x_i - \overline{x})^2}{(n-1)\, s_x^2}\right] = \frac{1}{n} \underbrace{\sum_{i=1}^{n} \frac{1}{n}}_{=1} + \frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{(n-1)\, s_x^2} \\ &= \frac{1}{n} + \frac{1}{n} \cdot \frac{1}{(n-1)\, s_x^2} \cdot \underbrace{\sum_{i=1}^{n} (x_i - \overline{x})^2}_{=(n-1)\, s_x^2} = \frac{1}{n} + \frac{1}{n} \cdot \frac{\cancel{(n-1)\, s_x^2}}{\cancel{(n-1)\, s_x^2}} = \frac{2}{n}\ . \end{aligned}$$

   (c) A variância dos erros aleatórios é, de acordo com o Modelo, σ^2. Esta variância é estimada pelo Quadrado Médio Residual, $QMRE$. Tendo em conta que a distância de Cook é dada por $D_i = R_i^2 \cdot \frac{h_{ii}}{1-h_{ii}} \cdot \frac{1}{2}$ e que o i-ésimo resíduo estandardizado é dado por $R_i = \frac{E_i}{\sqrt{QMRE \cdot (1-h_{ii})}}$, tem-se:

$$D_i = \frac{E_i^2}{QMRE\,(1-h_{ii})} \cdot \frac{h_{ii}}{1-h_{ii}} \cdot \frac{1}{2} = \frac{E_i^2 \cdot h_{ii}}{2\, QMRE\,(1-h_{ii})^2} \quad \Leftrightarrow \quad QMRE = \frac{E_i^2\, h_{ii}}{2\, D_i\,(1-h_{ii})^2}\ ,$$

   que é a expressão pedida no enunciado.

   (d) A contribuição da i-ésima observação para a Soma de Quadrados dos Resíduos é o quadrado do seu resíduo, ou seja E_i^2. Assim, a sua contribuição *relativa* é $\frac{E_i^2}{SQRE}$. Ora, o valor médio do efeito alavanca é, como se viu na alínea b), $\overline{h} = \frac{2}{n}$. Caso $h_{ii} = \overline{h} = \frac{2}{n}$, a expressão da alínea anterior fica:

$$\begin{aligned} QMRE &= \frac{SQRE}{n-2} = \frac{E_i^2 \cdot \frac{2}{n}}{2\, D_i\,(1-\frac{2}{n})^2} = \frac{E_i^2 \cdot \frac{\cancel{2}}{\cancel{n}}}{\cancel{2}\, D_i\, \frac{(n-2)^2}{n^{\cancel{2}}}} = \frac{E_i^2}{D_i\, \frac{(n-2)^2}{n}} \\ \Leftrightarrow \quad \frac{E_i^2}{SQRE} &= D_i\, \frac{(n-2)^{\cancel{2}}}{n \cdot \cancel{(n-2)}} = D_i \cdot \frac{n-2}{n} < D_i\ , \end{aligned}$$

   que é a desigualdade que se pedia para provar.