INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO

2 de Novembro, 2018 **PRIMEIRO TESTE 2018-19** **Uma resolução possível**

I

É pedido para estudar o ajustamento dos dados à distribuição de Poisson com parâmetro $\lambda = 7$, distribuição que a qualquer número inteiro não negativo $i \in \mathbb{N}_0$ associa a probabilidade $P[X=i] = \mathrm{e}^{-7}\frac{7^i}{i!}$. No nosso caso, X é a variável aleatória que conta o número de gomos florais por árvore.

1. Não é possível usar o teste de hipóteses baseado na estatística de Pearson com a tabela indicada no enunciado. A dimensão da amostra não é suficiente para garantir a validade da distribuição χ^2 assintótica da estatística X^2 de Pearson. De facto, o número esperado de observações correspondente à categoria de zero gomos é dado por $E_0 = N \times P[X=0]$, onde $N=80$ é a dimensão da amostra e, sob a hipótese de que $X \cap Po(7)$, tem-se $P[X=0] = \mathrm{e}^{-7}\,\frac{7^0}{0!} = 0.000911882$ (por convenção, $0! = 1$). Logo, $E_0 = 0.07295056 \ll 1$. Ao existirem categorias com valor esperado inferior a 1, viola-se o Critério de Cochran para a admissibilidade da distribuição assintótica. Este critério exige que em nenhuma classe haja contagens inferiores a 1, e em não mais do que 20% das classes haja contagens inferiores a 5.

2. O agrupamento da tabela indicado no enunciado resulta na nova tabela a seguir indicada.

No. de gomos	≤ 3	4	5	6	7	8	9	10	≥ 11
No. de árvores	13	6	15	4	8	7	6	4	17

   Nesta tabela há agora $k=9$ categorias. O enunciado afirma que podemos admitir a validade da distribuição assintótica, que neste caso será χ^2 com $k-1=8$ graus de liberdade (não foi necessário estimar qualquer parâmetro).

   (a) Eis o teste de hipóteses pedido:

   Hipóteses: $H_0 : X \cap Po(7) \quad vs. \quad H_1 : X \not\cap Po(7)$.

   Estatística do Teste: A estatística de Pearson é $X^2 = \sum_{i=1}^{9} \frac{(O_i - E_i)^2}{E_i}$. A distribuição assintótica desta estatística, caso seja verdade H_0, é χ^2_8.

   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 $\chi^2_{\text{calc}} > \chi^2_{0.05(8)} = 15.5073$.

   Conclusões No enunciado é dado o valor $X^2_{calc} = 28.122$. Logo, rejeita-se H_0, a hipótese de que X tenha uma distribuição de Poisson com parâmetro $\lambda = 7$, ao nível $\alpha = 0.05$.

   (b) É pedida a parcela da estatística de teste correspondente à última categoria, ou seja, à categoria de 11 ou mais gomos florais. Essa parcela é dada por $\frac{(O_{\geq 11} - \hat{E}_{\geq 11})^2}{\hat{E}_{\geq 11}}$, onde $O_{\geq 11} = 17$ e $E_{\geq 11} = N \times P[X \geq 11] = 80 \times (1 - P[X \leq 10]) = 80 \times (1 - 0.901) = 7.92$ (o valor $P[X \leq 10] = 0.901$ é dado directamente nas tabelas da distribuição Poisson). Logo, a parcela pedida é $\frac{(O_{\geq 11} - \hat{E}_{\geq 11})^2}{\hat{E}_{\geq 11}} = \frac{(17-7.92)^2}{7.92} = 10.4099$. Trata-se de um valor elevado, que

corresponde a mais de um terço do valor final da estatística. Assim, esta categoria de 11 ou mais gomos contribui de forma importante para a rejeição da hipótese de que X tenha uma distribuição $Po(7)$.

(c) A frase não é verdadeira. O teste efectuado rejeitou a hipótese de que X tenha distribuição $Po(7)$, mas não exclui a possibilidade de a distribuição ser Poisson com outro valor do parâmetro λ.

II

1. A Soma de Quadrados Total é dada por $SQT = (n-1)\, s_y^2 = 179 \times 14.26665 = 2553.73$. A Soma de Quadrados da Regressão pode ser calculada a partir da definição do coeficiente de determinação, $R^2 = \frac{SQR}{SQT}$, tendo-se $SQR = R^2 \times SQT = 0.5826 \times 2553.73 = 1487.803$. Finalmente, a Soma de Quadrados Residual resulta da fórmula fundamental da regressão, $SQT = SQR + SQRE$, sendo $SQRE = SQT - SQR = 2553.73 - 1487.803 = 1065.927$. **Nota:** Seria também possível obter $SQRE$ a partir da sua relação com $QMRE = \frac{SQRE}{n-2}$, cuja raíz quadrada é dada no enunciado. Nesse caso, obter-se-ia $SQRE = \sqrt{QMRE}^2\,(n-2) = (2.447)^2 \times 178 = 1065.83$. A diferença nos dois valores de $SQRE$ resulta de erros de arredondamento.

2. Há dois aspectos a referir. Por um lado, o coeficiente de determinação, $R^2 = 0.5826$, que indica que pouco mais de 58% da variância dos pesos de uva observados nas 180 parcelas é explicada pela regressão (um valor modesto). Por outro lado, deve efectuar-se o teste F de ajustamento global do modelo, para saber se ele é significativamente diferente do Modelo Nulo (de equação $Y = \beta_0 + \epsilon$) que não permite explicar nenhuma da variabilidade de Y a partir da sua relação linear com X. Eis este teste:

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

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

   Nível de significância: $\alpha = 0.05$.

   Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05[1,178]}$ que, pelas tabelas, é um valor entre os valores tabelados 3.84 e 3.92.

   Conclusões: No enunciado está omisso o valor calculado da estatística F, mas esse valor é calculável a partir de R^2, sendo $F_{calc} = 178 \times \frac{0.5826}{1-0.5826} = 248.4494$. A rejeição de H_0 é muito clara, pelo que o modelo ajustado, embora não tenha um valor muito elevado de R^2, é muito significativamente diferente do Modelo Nulo.

3. O aumento médio do peso total das uvas (variável resposta Y), associado a um cacho adicional (um aumento de uma unidade no valor do preditor X) é o significado do declive da recta. Logo é pedido um teste a β_1. É preciso cautela ao formular as hipóteses, uma vez que o declive β_1 tem as unidades de medida de Y sobre as unidades de medida de X, o que neste caso significa kg/cacho. O valor de $200g$ referido no enunciado corresponde a $0.2kg$, pelo que a hipótese indicada é $\beta_1 < 0.2$. É explicitamente dito no enunciado para *não* dar o benefício da dúvida a esta hipótese, ou seja, para não a colocar como Hipótese Nula. O valor estimado de β_1 ($b_1 = 0.17058$) aponta no sentido de $\beta_1 < 0.2$, mas falta saber se essa indicação é significativa. Assim, tem-se o seguinte teste:

   Hipóteses: $H_0 : \beta_1 \geq 0.2 \quad vs. \quad H_1 : \beta_1 < 0.2.$

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 seja o valor fronteira da Hipótese Nula, $\beta_{1|H_0} = 0.2$.

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

Região Crítica: (Unilateral esquerda) A regra de rejeição consiste em rejeitar H_0 se $t_{\text{calc}} < -t_{0.05(178)} \approx -1.65$.

Conclusões Para calcular o valor da estatística, será necessário primeiro calcular o valor do erro padrão de $\hat{\beta}_1$, omisso no enunciado. Pode recorrer-se à fórmula $\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{QMRE}{(n-1)\, s_x^2}} = \frac{2.447}{\sqrt{179 \times 285.6494100}} = 0.01082158$ (esta fórmula corresponde à raíz quadrada da variância de $\hat{\beta}_1$, dada no formulário, substituindo o valor desconhecido σ^2 pela sua estimativa $QMRE$). Logo, $t_{calc} = \frac{0.17058 - 0.2}{0.01082158} = -2.718641$. Este valor encontra-se na região de rejeição, pelo que se rejeita H_0 (ao nível $\alpha = 0.05$), ou seja, pode considerar-se a estimativa amostral $b_1 = 0.17058$ como sendo significativamente inferior a 0.2.

Nota: Altermativamente, o erro padrão de $\hat{\beta}_1$ pode ser calculado a partir da estimativa b_1 e do valor da estatística T_{calc} associada ao teste à hipótese $\beta_1 = 0$, ambos disponíveis no enunciado. De facto, $T_{calc} = \frac{b_1 - 0}{\hat{\sigma}_{\hat{\beta}_1}}$, pelo que $\hat{\sigma}_{\hat{\beta}_1} = \frac{b_1}{T_{calc}} = \frac{0.17058}{15.76} = 0.0108236$. A diferença, a partir da sexta casa decimal, nos dois valores para $\hat{\sigma}_{\hat{\beta}_1}$ resulta de erros de arredondamento.

4. Pede-se um intervalo de confiança para o valor esperado de Y (peso) quando o preditor toma o valor $x = 50$, o que equivale a pedir um intervalo de confiança para a ordenada na recta de regressão populacional, associada à abcissa $x = 50$. 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+":

$$\left] (b_0 + b_1 x) - t_{\frac{\alpha}{2};n-2} \cdot \sqrt{QMRE \cdot \left[\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)\, s_x^2}\right]} \;,\; (b_0 + b_1 x) + t_{\frac{\alpha}{2};n-2} \cdot \sqrt{QMRE \cdot \left[\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)\, s_x^2}\right]} \right[.$$

Conhecemos os valores $b_0 = 0.61660$, $b_1 = 0.17058$, $\sqrt{QMRE} = 2.447$, $n = 180$, $\overline{x} = 42.75556$, $s_x^2 = 285.64941$. Pela leitura das tabelas, $t_{0.025(178)} \approx 1.96$. Usando estes valores temos, para $x = 50$ cachos, o intervalo a 95% de confiança para o peso total esperado das uvas $]\, 8.756kg \;,\; 9.535kg \,[$. Do ponto de vista gráfico, podemos afirmar com 95% de confiança que a recta populacional atravessa o intervalo vertical que, por cima de $x = 50$, contém os valores de y pertencentes ao intervalo.

5. Neste ponto considera-se o modelo linear entre `log(pesototal)` (y^*) e `log(ncachos)` (x^*).

   (a) Não esquecer que ambas as variáveis foram log-transformadas, pelo que se tem uma relação linear entre $y^* = \ln(y)$ e $x^* = \ln(x)$. Logo, o resíduo associado à observação 114 é $e_{114} = y^*_{114} - \widehat{y^*}_{114} = \ln(y_{114}) - (-1.46223 + 0.93036 \ln(x_{114})) == \ln(4.6) - (-1.46223 + 0.93036 \ln(63)) = 1.526056 - (2.392377) = -0.8663205$. A Soma de Quadrados dos Resíduos pode ser calculada a partir da raíz quadrada de $QMRE$ (disponível no enunciado), ou seja: $SQRE = (\sqrt{QMRE})^2\,(n-2) = 0.3015^2 \times 178 = 16.1806$. Assim, a observação 114 contribui com uma proporção $\frac{e_{114}^2}{SQRE} = \frac{(-0.8663205)^2}{16.1806} = 0.0463834$, um pouco menos de 5% do valor de $SQRE$. Trata-se duma proporção considerável, tendo em conta que a média dos resíduos ao quadrado é próxima do Quadrado Médio Residual: $\frac{SQRE}{n} = \frac{n-2}{n}\, QMRE$. No nosso caso tem-se $QMRE = 0.3015^2 = 0.09090225$, pelo que $e_{114}^2 = 0.7505112$ é cerca de oito vezes maior que o valor médio dos resíduos ao quadrado. Uma vez que o modelo ajusta valores de peso log-transformados, o peso das uvas (em kg) ajustado pelo modelo é dado por $\mathrm{e}^{2.392377} = 10.93946$ kg.

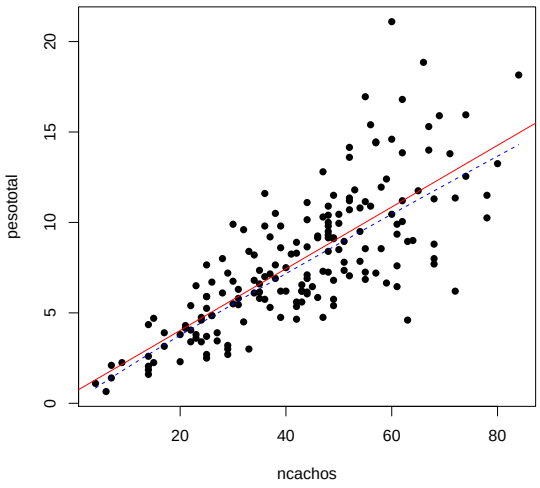
(b) A relação linear entre *os logaritmos* das duas variáveis observadas corresponde a admitir que a relação entre as variáveis originais x e y é de tipo potência. De facto, admitir a linearidade entre $y^* = \ln(y)$ e $x^* = \ln(x)$ corresponde a ter:

$$\begin{aligned} \ln(y) = b_0 + b_1 \ln(x) &\Leftrightarrow e^{\ln(y)} = e^{b_0 + b_1 \ln(x)} \\ &\Leftrightarrow y = \underbrace{e^{b_0}}_{=a} e^{b_1 \ln(x)} = a\, e^{ln\, x^{b_1}} = a\, x^{b_1} . \end{aligned}$$

Assim, a curva potência ajustada à relação original entre `pesototal` (y) e `ncachos` (x) é dada por $y = e^{-1.46223}\, x^{0.93036} = 0.231719\, x^{0.93036}$. Como sabemos das aulas teóricas, este tipo de relação potência corresponde a admitir que ambas as variáveis são funções duma terceira variável t (que, neste contexto se poderia supôr ser o número de gomos deixados à poda) e que as respectivas taxas de variação relativas são proporcionais, sendo a constante de proporcionalidade dada pelo declive da recta na relação linear entre as variáveis logaritmizadas ou, de forma equivalente, pela potência na relação entre as variáveis originais. Assim, a relação que se admite existir na população entre as taxas de variação relativas das variáveis y e x é $\frac{y'(t)}{y(t)} = \beta_1 \frac{x'(t)}{x(t)}$. A relação estimada, com $b_1 = 0.93036$, é aproximadamente uma relação de igualdade entre as duas taxas de variação relativas. Um teste à hipótese $\beta_1 = 1$ responde à pergunta feita no enunciado. Trata-se dum teste com região crítica bilateral que (ao nível $\alpha = 0.05$) tem fronteiras em $\pm t_{0.025(178)} \approx 1.96$. O valor calculado da estatística é $T_{calc} = \frac{0.93036 - 1}{0.04414} = -1.577707$ que não pertence à região crítica. Logo, é admissível que $\beta_1 = 1$, e portanto que as duas taxas de variação relativas sejam iguais.

6. O gráfico da esquerda, correspondente ao modelo sem transformações logaritmicas, revela um claro efeito de funil, que lança dúvidas sobre a validade do pressuposto de variâncias homogéneas exigido no modelo de regressão linear simples. Não existem, nesse mesmo gráfico, nem indicações de curvilinearidade na relação de fundo, nem indicação da presença de observações atípicas. O gráfico da direita, que está associado ao modelo com transformações logarítmicas das variáveis, é um típico gráfico "bom", em que não há qualquer indício de violação dos pressupostos do modelo RLS: os pontos dispersam-se, sem padrões aparentes, numa banda horizontal em torno de zero. Registe-se que o efeito funil desapareceu com a transformação logarítmica das duas variáveis, pelo que o pressuposto de variâncias constantes é agora sustentado.

Nota adicional: Embora não fosse pedido, nem possível de verificar no contexto do Teste, verifica-se que a transformação logarítmica das duas variáveis produz uma relação potência que, para a gama de valores de X (`ncachos`) observados, é quase linear. No gráfico seguinte pode ver-se a recta ajustada pelo modelo inicial (traçada com uma linha contínua) e a curva potência (com linha tracejada) ajustada com base no modelo linear entre as variáveis logaritmizadas. Assim, constata-se que a dupla logaritmização estabilizou a variância de Y, sem destruir uma boa relação de fundo entre peso das uvas e número de cachos.



7. Como se viu nas aulas, as bandas de predição não são rectas, mas sim curvas. De facto, a fórmula dos intervalos de predição (que consta do formulário) indica que a amplitude dos intervalos depende da distância ao quadrado entre o valor de x usado e a média dos valores de x observados, ou seja, do numerador da última parcela debaixo da raíz quadrada, $(x-\overline{x})^2$. O valor de x para o qual se obtem o intervalo de predição de menor amplitude tem de ser $x=\overline{x}$, que anula essa última parcela. Esse facto exclui a possibilidade de as bandas de predição do modelo nas escalas originais de x e y serem as bandas em traço contínuo, que abrem em forma de funil, e para as quais os intervalos de menor amplitude correspondem aos valores mais pequenos de x (e não ao valor médio). Assim, as bandas de predição para o modelo original (sem transformações logarítmicas) têm de ser as curvas a tracejado. É visível que essas bandas sobrestimam a variabilidade das observações individuais de Y para os menores valores de x observados, e ao mesmo tempo parecem subestimar essa variabilidade na zona dos maiores valores de x. Esse facto é o reflexo de as bandas serem construídas admitindo que a variância dos erros aleatórios é constante, pressuposto que a forma em funil da nuvem de pontos desmente. As bandas associadas ao modelo que passou pelas transformações logarítmicas (em traço contínuo) acompanham melhor a variabilidade das observações individuais, uma vez que foram construídas na escala duplamente logarítmica, onde (como se viu na alínea anterior) o pressuposto de variâncias homogéneas parece inteiramente legítimo. A abertura gradual dessas bandas no gráfico do enunciado é o resultado da exponenciação que foi necessário efectuar nos intervalos de predição do modelo linear em $\ln(y)$ e $\ln(x)$, de forma a obter intervalos na escala original dos y.

III

1. O Modelo de regressão linear simples, em contexto inferencial, afirma que: (i) as observações Y_i da variável resposta são dadas por uma relação linear com X, mais um erro aleatório ($Y_i = \beta_0 + \beta_1\, x_i + \epsilon_i$); (ii) esses erros aleatórios são Normais, de média zero e variância constante ($\epsilon_i \cap \mathcal{N}(0, \sigma^2)$, para todo o i); e (iii) os erros aleatórios são variáveis aleatórias independentes.

   (a) As variáveis aleatórias Y_i são transformações lineares dos erros aleatórios ϵ_i (resultantes de a cada erro somar a constante $\beta_0 + \beta_1\, x_i$). Ora, transformações lineares preservam a Normalidade, pelo que os Y_i têm distribuição Normal. Falta determinar os dois parâmetros

dessa Normal. Tendo em conta que constantes aditivas saltam para fora dos valores esperados, tem-se $E[Y_i]=E[\beta_0+\beta_1\,x_i+\epsilon_i]=\beta_0+\beta_1\,x_i+\underbrace{E[\epsilon_i]}_{=0}=\beta_0+\beta_1\,x_i$. Tendo em conta que constantes aditivas não afectam a variância, tem-se $V[Y_i]=V[\beta_0+\beta_1\,x_i+\epsilon_i]=V[\epsilon_i]=\sigma^2$. Logo, para qualquer observação i, tem-se $Y_i \cap \mathcal{N}(\beta_0+\beta_1\,x_i\,,\,\sigma^2)$.

(b) Como consta do formulário, $\hat{\beta}_1=\sum_{i=1}^{n} c_i\,Y_i$, com $c_i=\frac{x_i-\overline{x}}{(n-1)\,s_x^2}$. Ora, tendo em conta que a variância duma soma de variáveis aleatórias independentes (como são os Y_i) é a soma das variâncias; que constantes multiplicativas saltam para fora das variâncias *ao quadrado*; e que a distribuição dos Y_i deduzida na alínea anterior; tem-se:

$$V[\hat{\beta}_1] \;=\; \sum_{i=1}^{n} V[c_i\,Y_i] \;=\; \sum_{i=1}^{n} c_i^2\,\underbrace{V[Y_i]}_{=\sigma^2} \;=\; \sigma^2 \sum_{i=1}^{n} c_i^2 \;=\; \frac{\sigma^2}{[(n-1)\,s_x^2]^2}\,\underbrace{\sum_{i=1}^{n}(x_i-\overline{x})^2}_{=(n-1)\,s_x^2} \;=\; \frac{\sigma^2}{(n-1)\,s_x^2}$$

(c) Pretende-se mostrar que $E[\hat{\beta}_0]=\beta_0$. Por definição, $\hat{\beta}_0=\overline{Y}-\hat{\beta}_1\,\overline{x}$. Tendo em conta que: a esperança duma soma de variáveis aleatórias é a soma das esperanças; constantes multiplicativas saltam para fora da esperança; $\hat{\beta}_1$ é um estimador centrado; e a distribuição dos Y_i acima deduzida; tem-se:

$$\begin{aligned} E[\hat{\beta}_0] \;&=\; E[\overline{Y}]-E[\hat{\beta}_1\,\overline{x}] \;=\; E\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right]-\overline{x}E[\hat{\beta}_1] \;=\; \frac{1}{n}\sum_{i=1}^{n} E[Y_i]-\beta_1\,\overline{x} \\ &=\; \frac{1}{n}\sum_{i=1}^{n}(\beta_0+\beta_1\,x_i)-\beta_1\,\overline{x} \;=\; \frac{1}{n}\sum_{i=1}^{n}\beta_0+\beta_1\,\frac{1}{n}\sum_{i=1}^{n} x_i-\beta_1\,\overline{x} \;=\; \beta_0+\beta_1\overline{x}-\beta_1\overline{x} \;=\; \beta_0\;. \end{aligned}$$

(d) Sabemos que:

$$\begin{aligned} & F_{calc}=(n-2)\frac{R^2}{1-R^2} && \Leftrightarrow && \frac{1-R^2}{R^2}=\frac{(n-2)}{F_{calc}} \\ \Leftrightarrow\quad & \frac{1}{R^2}-1=\frac{(n-2)}{F_{calc}} && \Leftrightarrow && \frac{1}{R^2}=1+\frac{(n-2)}{F_{calc}}=\frac{F_{calc}+(n-2)}{F_{calc}} \\ & && \Leftrightarrow && R^2=\frac{F_{calc}}{F_{calc}+(n-2)} \end{aligned}$$

Para se ter $R^2=\frac{F_{calc}}{F_{calc}+(n-2)}=0.9$, é preciso que $9[F_{calc}+(n-2)]=10\,F_{calc}$, ou seja que $F_{calc}=9(n-2)$. Assim, um valor tão elevado de R^2 está associado a um valor calculado da estatística do teste F de ajustamento global de cerca de 9 vezes o tamanho da amostra.

2. A equação diferencial no enunciado diz que a taxa absoluta de variação de y (ou seja, a derivada $y'(x)$) é proporcional ao quadrado da razão entre y e x. Re-arrumando essa equação, primitivando em ordem a x e designando a constante de primitivação por k, tem-se:

$$\begin{aligned} \frac{y'(x)}{y^2(x)} \;=\; \frac{c}{x^2} \quad &\Leftrightarrow \quad P_x\left(\frac{y'(x)}{y^2(x)}\right) \;=\; P_x\left(\frac{c}{x^2}\right) \quad \Leftrightarrow \quad \frac{-1}{y(x)} \;=\; \frac{-c}{x}+k \;=\; \frac{-c+kx}{x} \\ &\Leftrightarrow \quad y(x) \;=\; \frac{x}{c-kx}\;, \end{aligned}$$

que é a equação de Michaelis-Menten (com $d=-k$).