

Estatística e Análise de Dados em Zootecnia

Aulas Teóricas

Elsa Gonçalves
Secção de Matemática (DCEB)
Instituto Superior de Agronomia (ULisboa)
2026-27

Pressupostos

Admite-se que houve frequência numa disciplina introdutória de Estatística no primeiro ciclo (semelhante à existente no ISA) e que são conhecidos:

- principais indicadores descritivos (média, variância, covariância, coeficiente de correlação linear, etc.) e suas propriedades;
- conceitos básicos de probabilidades;
- variáveis aleatórias e sua caracterização;
- principais distribuições de probabilidades (Normal, t-Student, χ^2 , F, etc.);
- conceitos de intervalos de confiança e testes de hipóteses.

Programa

Trata-se de uma UC de aprofundamento, que procura relacionar uma variável de interesse com outras variáveis.

O programa da UC consiste no estudo do principal modelo estatístico: o Modelo Linear, que inclui como casos particulares:

- Regressão Linear (Simples e Múltipla);
- Regressão Polinomial;
- Análise de Variância (ANOVA), de efeitos fixos e de efeitos aleatórios;
- Análise de Covariância (ANCOVA).

Delineamento Experimental

Modelo Linear

Elsa Gonçalves

(Adaptado, Cadima, J. (2021). O Modelo Linear. ISA, ULisboa)

Modelação de relações entre variáveis

Importância central da recolha de informação (dados).

Nas disciplinas introdutórias de Estatística aprende-se a trabalhar com dados relativos a uma variável.

Nesta disciplina: relações (modelos) entre duas ou mais variáveis.

Variáveis podem ser:

- numéricas (medições, rendimentos, contagens, etc.) ou categóricas (factores) (espécies, locais, tratamentos, etc.);
- foco de interesse (variável resposta) ou auxiliares para explicar uma variável resposta (variável preditora ou explicativa).

Modelos determinísticos e modelos estatísticos

Uma relação (modelo) entre duas ou mais variáveis pode ser:

- essencialmente exacta (como na Mecânica: $F = ma$).
Trata-se de **modelos determinísticos**.

Ou

- apenas uma tendência de fundo, sabendo-se que existe variabilidade das observações em torno dessa tendência de fundo. Trata-se de **modelos estatísticos** ou **probabilísticos**.

Modelação Estatística

Objectivo (informal): Descrever a **relação de fundo** entre

- uma **variável resposta** (ou **dependente**) y ; e
- uma ou mais **variáveis preditoras** (**variáveis explicativas** ou **independentes**), $x_1, x_2, \dots, x_p$.

Informação: A identificação da relação de fundo é feita com base em n observações do conjunto de variáveis envolvidas na relação.

Vamos inicialmente considerar o contexto de **um único preditor numérico**, para modelar **uma única variável resposta numérica**.

Motivamos a discussão com **dois exemplos**.

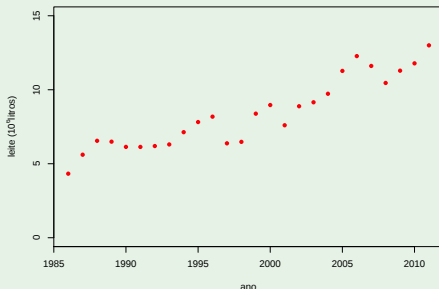
O Modelo Linear

- O **Modelo Linear** é um **caso particular** de modelação estatística;
- **engloba um grande número de modelos específicos:**
Regressão Linear (Simples e Múltipla) , Regressão Polinomial, Análise de Variância, Análise de Covariância;
- é o **mais completo e bem estudado tipo de modelo**;
- serve de **base para numerosas extensões**
(Regressão não linear, Modelos Lineares Generalizados, Modelos Lineares Mistos, etc.).

Exemplo 1

Produção de leite de cabra em Portugal, 1986 a 2011 (INE)

Produção (y) vs. Anos (x), $n = 26$ pares de valores, $\{(x_i, y_i)\}_{i=1}^{26}$.



Existe uma **tendência de fundo** e é **aproximadamente linear**.

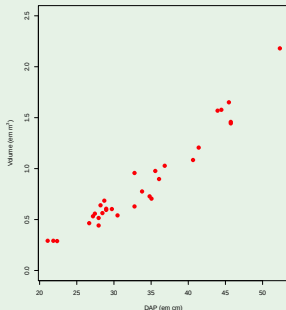
O **coeficiente de correlação linear** é $r_{xy} = 0.9348$.

Qual a “melhor” equação de recta, $y = b_0 + b_1 x$, para descrever as n observações (e que critério de “melhor”)?

Exemplo 2 - relação linear

Volume de tronco vs. DAP em cerejeiras

DAP (Diâmetro à altura do peito, variável x) e Volume de troncos (y) de cerejeiras. Existem $n = 31$ pares de medições: $\{(x_i, y_i)\}_{i=1}^{31}$.



A tendência de fundo é aproximadamente linear. O coeficiente de correlação linear é $r_{xy} = 0.9671$. Mas os $n = 31$ pares de observações são apenas uma amostra aleatória duma população mais vasta. Interessa o contexto inferencial: o que se pode dizer sobre a recta populacional $y = \beta_0 + \beta_1 x$?

Regressão Linear - Abordagem Descritiva

Regressão Linear Simples - contexto descritivo

Revisão: Estudado nas disciplinas introdutórias de Estatística.

Se n pares de observações $\{(x_i, y_i)\}_{i=1}^n$ têm relação linear de fundo, a recta de regressão de y sobre x define-se como:

Recta de Regressão Linear de y sobre x

$$y = b_0 + b_1 x$$

com

$$\text{Declive} \quad b_1 = \frac{\text{COV}_{xy}}{s_x^2}$$

$$\text{Ordenada na origem} \quad b_0 = \bar{y} - b_1 \bar{x}$$

sendo

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{cov}_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) .$$

Regressão Linear Simples - contexto descritivo

Exemplo das cerejeiras

$n = 31$ pares de medições, $\{(x_i, y_i)\}_{i=1}^{31}$.

DAP (x) e Volume de troncos (y) de cerejeiras.

$$\text{cov}_{xy} = 3.5881929$$

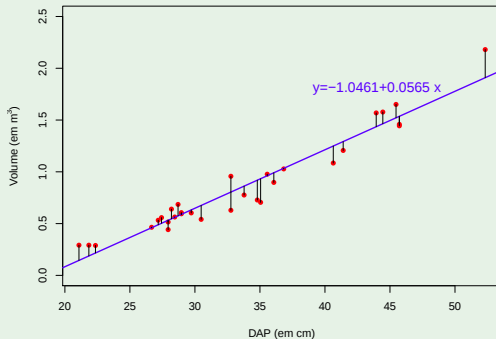
$$s_x^2 = 63.5348018$$

$$\bar{x} = 33.6509032$$

$$\bar{y} = 0.8543468$$

$$b_1 = \frac{\text{cov}_{xy}}{s_x^2} = 0.056476$$

$$b_0 = \bar{y} - b_1 \bar{x} = -1.046122$$



Como se chegou à equação da recta?

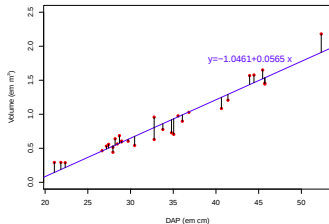
Valores ajustados e Resíduos

Dada uma recta, valores de y podem ser previstos a partir de valores de x , obtendo-se os “valores de y ajustados pela recta”, $\hat{y}_i$:

$$\hat{y}_i = b_0 + b_1 x_i .$$

Os **resíduos** são as diferenças entre os valores de y **observados** e **ajustados** ou seja, são as diferenças **na vertical** entre pontos e recta ajustada:

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i) ,$$



O Critério de Mínimos Quadrados

Critério: minimizar a Soma de Quadrados dos Resíduos

$$SQRE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n [y_i - (b_0 + b_1 x_i)]^2 .$$

Determinar b_0 e b_1 que minimizam $SQRE$ é um problema de minimizar uma função ($SQRE$) de duas variáveis (aqui chamadas b_0 e b_1).

Ver páginas 9 a 11 dos apontamentos teóricos disponíveis na página da UC.

Regressão Linear Simples - contexto descritivo

O critério de minimizar Soma de Quadrados dos Resíduos tem, subjacente, um pressuposto:

O papel das 2 variáveis, x e y , não é simétrico.

y – **variável resposta** (“dependente”)

- **variável que se deseja modelar**, prever a partir da variável x .

x – **variável preditora** (“independente”)

- **variável com base na qual se pretende tirar conclusões sobre y .**

Regressão Linear Simples - contexto descritivo

O i -ésimo resíduo é o desvio (com sinal) da observação y_i face à sua previsão a partir da recta:

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$$

Interpretação do Critério de Mínimos Quadrados

Minimizar a soma de quadrados dos resíduos corresponde a minimizar a soma de quadrados dos “erros de previsão”.

O critério tem subjacente a preocupação de prever o melhor possível a variável y , a partir da sua relação com o preditor x .

Revisão: Propriedades dos parâmetros da recta

Propriedades dos parâmetros da recta de regressão

- A ordenada na origem b_0 :
 - ▶ é o valor de y (na recta) associado a $x = 0$;
 - ▶ tem unidades de medida iguais às de y .
- O declive b_1 :
 - ▶ é a variação (**média**) de y associada a um aumento de uma unidade em x ;
 - ▶ tem unidades de medida iguais a $\frac{\text{unidades de } y}{\text{unidades de } x}$.

Exemplo das cerejeiras

$$b_1 = 0.05648 \frac{m^3}{cm}$$

por cada cm a mais no DAP, o volume do tronco aumenta, **em média**, $0.05648 m^3$.

Revisão: Propriedades da recta de regressão

Propriedades da recta de regressão

- A recta de regressão passa sempre no centro de gravidade da nuvem de pontos, isto é, no ponto $(\bar{x}, \bar{y})$, como é evidente a partir da fórmula para a ordenada na origem:

$$b_0 = \bar{y} - b_1 \bar{x} \quad \Leftrightarrow \quad \bar{y} = b_0 + b_1 \bar{x} .$$

- $\bar{y}$ é simultaneamente a média dos y_i observados e dos $\hat{y}_i$ ajustados.
- Embora não tenha sido explicitamente exigido, a média dos resíduos e_i é nula, ou seja, $\bar{e} = 0$.

Ver página 13 dos apontamentos teóricos disponíveis na página da UC.

Revisão: RLS - As três Somas de Quadrados

Recordar: $s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$ a variância amostral das observações y_i .

Soma de Quadrados Total (SQT)

$$\text{SQ Total} \quad SQT = \sum_{i=1}^n (y_i - \bar{y})^2 = (n-1) s_y^2$$

Tem-se: $s_{\hat{y}}^2 = \frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ a variância amostral dos $\hat{y}_i$ ajustados.

Soma de Quadrados da Regressão (SQR)

$$\text{SQ Regressão} \quad SQR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = (n-1) s_{\hat{y}}^2$$

Soma de Quadrados Residual (SQRE) - já dado

$$\text{SQ Residual} \quad SQRE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = (n-1) s_e^2$$

Revisão: RLS - Fórmula fundamental e R^2

Fórmula Fundamental da Regressão

Prova-se a seguinte **Fórmula Fundamental** (ver Exercício RLS 5):

$$SQT = SQR + SQRE \quad \Leftrightarrow \quad s_y^2 = s_{\hat{y}}^2 + s_e^2$$

Definição: Coeficiente de Determinação

$$R^2 = \frac{SQR}{SQT} = \frac{s_{\hat{y}}^2}{s_y^2}, \quad (s_y^2 \neq 0)$$

R^2 mede a **proporção da variabilidade total da variável resposta Y que é explicada pela regressão**. Quanto maior, melhor.

Propriedades do Coeficiente de Determinação

Propriedades de $R^2 = \frac{SQR}{SQT}$ (Ver páginas 15 a 17 dos apontamentos teóricos disponíveis na página da UC.)

- $0 \leq R^2 \leq 1$ (Todas as SQs são não negativas e $SQT = SQR + SQRE$)

- $R^2 = 1$ se, e só se, os n pontos são colineares. (“ideal”)

$$(SQT = SQR \Leftrightarrow SQRE = \sum_{i=1}^n e_i^2 = 0 \Rightarrow e_i = 0, \text{ para todo } i.)$$

Logo, todos os resíduos são nulos: os pontos estão todos em cima da recta.)

- $R^2 = 0$ se, e só se, a recta de regressão for horizontal. (“inútil”)

($SQR = 0 \Leftrightarrow SQRE = SQT$. Toda a variabilidade de y é residual.

$SQR = 0$ implica $\hat{y}_i = \bar{y}$, para todo o i . A recta é $y = \bar{y} \Leftrightarrow b_1 = 0$)

- Numa regressão linear **simples**, R^2 é o quadrado do coeficiente de correlação linear entre x e y :

$$R^2 = r_{xy}^2 = \left(\frac{COV_{xy}}{S_x S_y} \right)^2 \quad \text{se } S_x \neq 0 \text{ e } S_y \neq 0$$

Algumas ideias prévias sobre modelação

- Todos os modelos são apenas **aproximações** da realidade.
- Pode haver mais do que um modelo adequado a uma relação. Um dado modelo pode ser melhor num aspecto, mas pior noutro.
- O **princípio da parcimónia** na modelação: de entre os modelos considerados **adequados**, é preferível o **mais simples**.
- Os modelos **estatísticos** apenas descrevem **tendência de fundo**: há **variação** das observações em torno da tendência de fundo.
- Num modelo estatístico **não há necessariamente uma relação de causa e efeito entre variável resposta e preditores**. Há apenas **associação**. A eventual existência de uma relação de causa e efeito só pode ser **justificada por argumentos extra-estatísticos**.

Transformações linearizantes

Nalguns casos, a relação de fundo entre x e y é não-linear, mas pode ser linearizada caso se proceda a transformações numa ou em ambas as variáveis.

Tais transformações podem permitir utilizar a Regressão Linear Simples, apesar de a relação original ser não-linear.

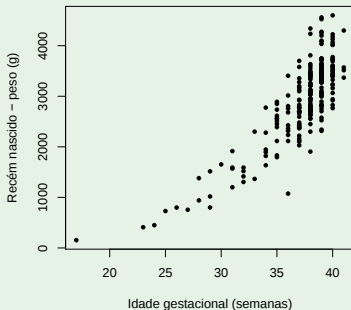
Vamos ver alguns exemplos particularmente frequentes de relações não-lineares que são linearizáveis através de transformações da variável resposta e, nalguns casos, também do preditor.

Exemplo 3 - Uma relação não linear

Peso de bebés à nascença

$n = 251$ pares de observações

Idade gestacional (x) e peso de bebé à nascença y , $\{(x_i, y_i)\}_{i=1}^{251}$.



A tendência de fundo é **não-linear**: $y = f(x)$.

Exemplo 3 (cont.)

Neste caso, há uma **questão adicional**:

- Qual a **forma da relação** (qual a natureza da função f)?
 - ▶ f exponencial ($y = c e^{dx}$)?
 - ▶ f função potência ($y = c x^d$)?
 - ▶ outra?

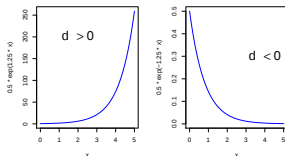
Além das perguntas análogas ao caso linear:

- Como determinar os “melhores” **parâmetros c e d** ?
- E, se os dados forem amostra aleatória, **o que se pode dizer sobre os respectivos parâmetros populacionais?**

A **Regressão Não Linear** **não** faz parte do programa da disciplina. Mas **transformações linearizantes** de uma ou ambas as variáveis podem criar uma relação linear, que permita usar o Modelo Linear.

Relação exponencial

Relação exponencial : $y = ce^{dx}$
($y > 0$; $c > 0$)



Transformação : Logaritmizando, obtém-se:

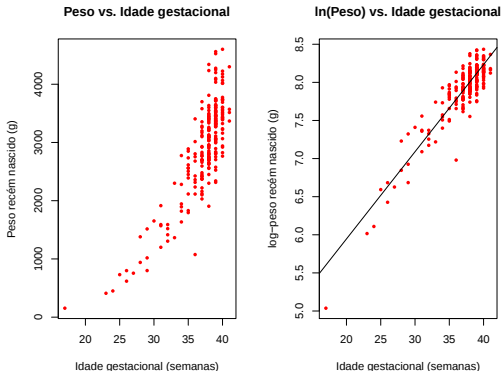
$$\begin{aligned}\ln(y) &= \ln(c) + dx \\ \Leftrightarrow y^* &= b_0 + b_1x\end{aligned}$$

que é uma **relação linear entre $y^* = \ln(Y)$ e x** , com declive $b_1 = d$ e ordenada na origem $b_0 = \ln(c)$.

O sinal do declive da recta indica se a relação exponencial original é crescente ($b_1 > 0$) ou decrescente ($b_1 < 0$).

Uma linearização no Exemplo 3

O gráfico de **log-pesos** dos recém-nascidos contra idade gestacional produz uma **relação de fundo linear**:



Esta linearização da relação significa que **a relação original (peso vs. idade gestacional)** pode ser considerada exponencial.

Ainda a relação exponencial

Uma relação exponencial resulta de admitir que y é função de x e que a **taxa de variação de y** , ou seja, a derivada $y'(x)$, **é proporcional a y** :

$$y'(x) = d \cdot y(x) ,$$

isto é, que a **taxa de variação relativa de y** é constante:

$$\frac{y'(x)}{y(x)} = d .$$

Primitivando (em ordem a x), tem-se:

$$\ln(y(x)) = \underbrace{d}_{=b_1} x + \underbrace{C}_{=b_0} \quad \Leftrightarrow \quad y(x) = e^C e^{dx} .$$

Repare-se que o declive b_1 da recta é o valor (constante) d da taxa de **variação relativa de y** . A constante de primitivação C é a ordenada na origem da recta: $C = b_0$.

Modelo exponencial de crescimento populacional

Um modelo exponencial é frequentemente usado para descrever o **crescimento de populações**, numa fase inicial onde não se faz ainda sentir a escassez de recursos limitantes.

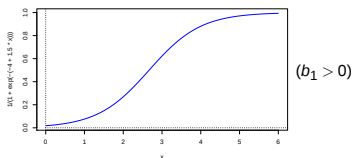
Mas nenhum crescimento populacional exponencial é sustentável a longo prazo.

Em **1838 Verhulst** propôs uma **modelo de crescimento populacional alternativo**, prevendo os efeitos resultantes da escassez de recursos: o **modelo logístico**.

Considera-se aqui uma versão simplificada (com 2 parâmetros) desse modelo. Pode pensar-se que a variável **y mede a dimensão duma população, relativa a um máximo possível**, sendo assim uma proporção.

Relação Logística (com 2 parâmetros)

Relação Logística : $y = \frac{1}{1+e^{-(c+dx)}}$



Transformação : Como $y \in]0,1[$, tem-se uma relação linear entre a transformação *logit* de Y , i.e., $y^* = \ln\left(\frac{y}{1-y}\right)$, e x :

$$\Rightarrow 1-y = \frac{e^{-(c+dx)}}{1+e^{-(c+dx)}}$$

$$\Rightarrow \frac{y}{1-y} = \frac{1}{e^{-(c+dx)}} = e^{c+dx}$$

$$\Rightarrow \underbrace{\ln\left(\frac{y}{1-y}\right)}_{=y^*} = \underbrace{c}_{=b_0} + \underbrace{d}_{=b_1}x$$

Ainda a Logística

A relação logística resulta de admitir que y é função de x e que a taxa de variação relativa de y diminui com o aumento de y :

$$\frac{y'(x)}{y(x)} = d \cdot [1 - y(x)] .$$

De facto, a expressão anterior equivale a:

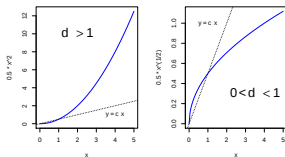
$$\frac{y'(x)}{y(x) \cdot (1 - y(x))} = d \quad \Leftrightarrow \quad \frac{y'(x)}{1 - y(x)} + \frac{y'(x)}{y(x)} = d$$

Primitivando (em ordem a x), tem-se:

$$\begin{aligned} -\ln(1 - y(x)) + \ln y(x) &= dx + C \\ \Leftrightarrow \ln\left(\frac{y}{1 - y}\right) &= b_1 x + b_0 . \end{aligned}$$

Relação potência ou alométrica

Relação potência : $y = c x^d$
($x, y > 0$; $c, d > 0$)



Transformação : Logaritmizando, obtém-se:

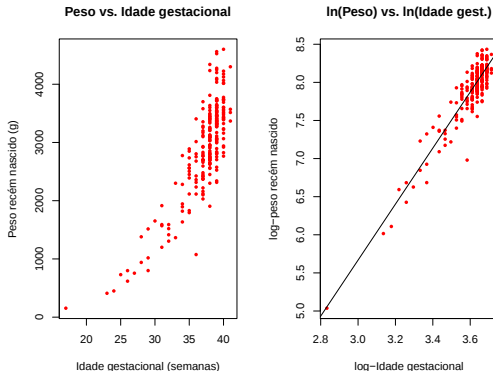
$$\begin{aligned}\ln(y) &= \ln(c) + d \ln(x) \\ \Leftrightarrow y^* &= b_0 + b_1 x^*\end{aligned}$$

que é uma **relação linear** entre $y^* = \ln(y)$ e $x^* = \ln(x)$.

O declive b_1 da recta é o expoente d na relação potência original.
Mas $b_0 = \ln(c)$.

Outra linearização no Exemplo 3

O gráfico de **log-pesos** dos recém-nascidos contra **log-idade gestacional** produz outra **relação de fundo linear**:



Esta linearização significa que a **relação original** (peso vs. idade gestacional) **também** pode ser considerada uma **relação potência**.

Ainda a relação potência

Uma Equação Diferencial da potência

Uma relação potência resulta de admitir que y é função de x e a taxa de variação relativa de y , i.e., a razão $\frac{y'(x)}{y(x)}$, é inversamente proporcional a x :

$$\frac{y'(x)}{y(x)} = \frac{d}{x}.$$

Primitivando (em ordem a x), tem-se (pois $y > 0$ e $x > 0$):

$$\underbrace{\ln|y(x)|}_{=y^*} = \underbrace{d}_{=b_1} \underbrace{\ln|x|}_{=x^*} + \underbrace{K}_{=b_0} \quad \Leftrightarrow \quad y(x) = e^{K+\ln(x^d)} \quad \Leftrightarrow \quad y(x) = e^K x^d.$$

O declive b_1 da recta é a constante de proporcionalidade d .

A constante de primitivação K é a ordenada na origem da recta: $K = b_0$.

O contexto alométrico da relação potência

A Equação Diferencial da alometria

Outra forma de obter uma relação potência, muito usada nos estudos alométricos, resulta de admitir que y e x são ambas funções duma terceira variável t (ou seja, $y(t)$ e $x(t)$) e que as taxas de variação relativas de y e x são proporcionais:

$$\frac{y'(t)}{y(t)} = d \cdot \frac{x'(t)}{x(t)}.$$

Primitivando (em ordem a t) tem-se:

$$\ln y = d \ln x + K$$

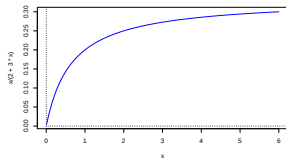
e exponenciando,

$$y = e^{d \ln x + K} = e^{d \ln x} \cdot e^K = e^{\ln x^d} \cdot \underbrace{e^K}_{=c} \Leftrightarrow y = c x^d.$$

Os estudos de alometria comparam a dimensão de partes diferentes dum organismo. A isometria corresponde ao valor $d = 1$.

Relação Michaelis-Menten

Relação Michaelis-Menten : $y = \frac{x}{c+dx}$



Transformação :

Tomando recíprocos, obtém-se uma **relação linear entre**
 $y^* = \frac{1}{y}$ e $x^* = \frac{1}{x}$:

$$\frac{1}{y} = \frac{c}{x} + d \quad \Leftrightarrow \quad y^* = b_0 + b_1 x^*,$$

com $b_0 = d$ e $b_1 = c$.

Relação Michaelis-Menten (cont.)

- A relação Michaelis-Menten é muito utilizada no estudo de reacções enzimáticas, relacionando a taxa da reacção com a concentração do substrato.
- Em modelos agronómicos de rendimento é conhecido como modelo Shinozaki-Kira, com y o rendimento total e x a densidade duma cultura ou povoamento.
- Nas pescas é conhecido como modelo Beverton-Holt: y é recrutamento e x a dimensão do manancial (stock) de progenitores.
- Resulta de admitir que a taxa de variação de y é proporcional ao quadrado da razão entre y e x :

$$y'(x) = c \left(\frac{y(x)}{x} \right)^2 .$$

Advertência sobre transformações linearizantes

A regressão linear simples não modela directamente relações não lineares entre x e y . Pode modelar uma relação linear entre as variáveis transformadas.

Transformações da variável-resposta y têm um impacto grande no ajustamento: a escala dos resíduos é alterada.

Nota: Linearizar, obter os parâmetros b_0 e b_1 da recta e depois desfazer a transformação linearizante não produz os mesmos parâmetros ajustados que resultariam de minimizar a soma de quadrados dos resíduos directamente na relação não linear. Esta última abordagem corresponde a efectuar uma regressão não linear, metodologia não englobada nesta disciplina.