

Modelo Linear

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.

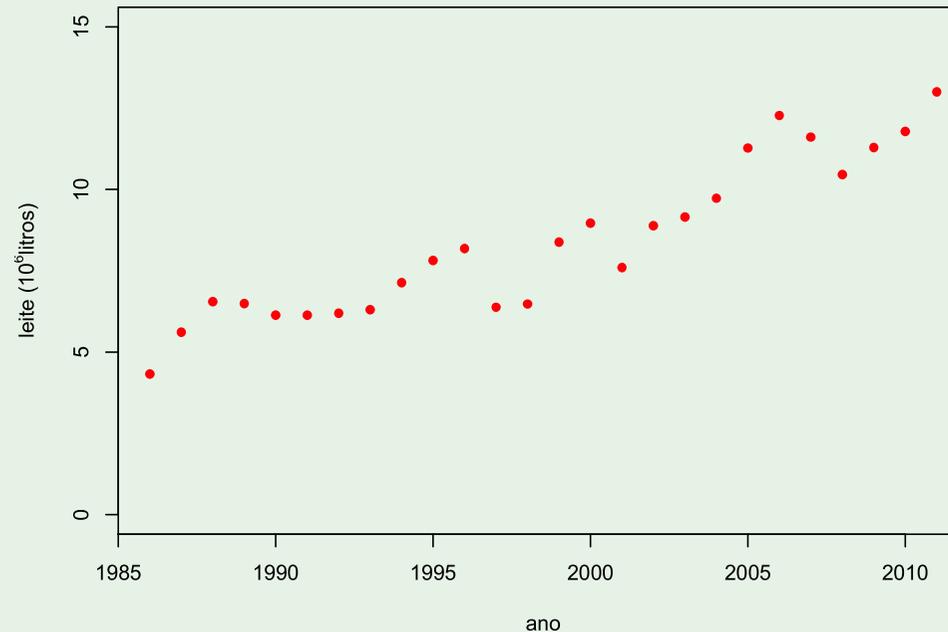
Vamos inicialmente considerar o contexto de **um único preditor numérico**, para modelar **uma única variável resposta numérica**.

Motivamos a discussão com **dois exemplos**.

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**.

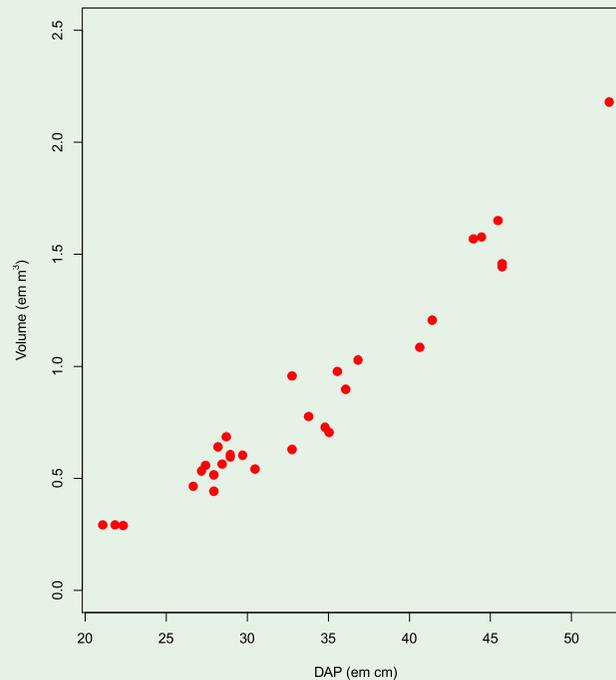
O coeficiente de correlação linear é $r_{xy} = 0.9348$.

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 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.

$$cov_{xy} = 3.5881929$$

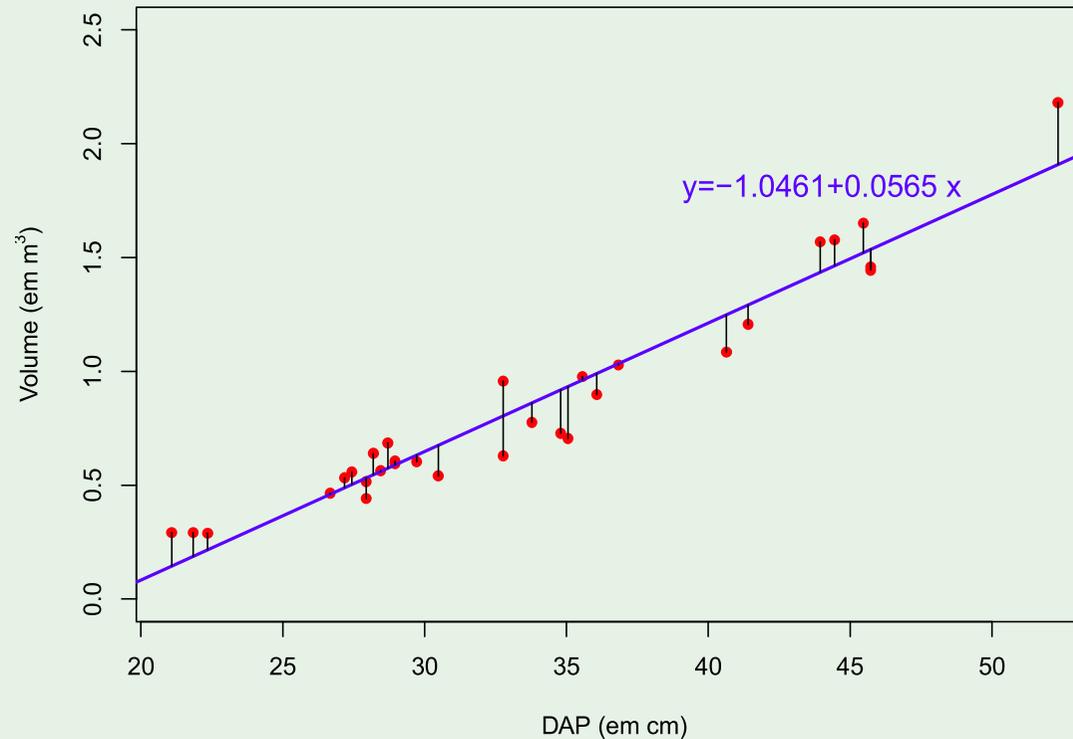
$$s_x^2 = 63.5348018$$

$$\bar{x} = 33.6509032$$

$$\bar{y} = 0.8543468$$

$$b_1 = \frac{cov_{xy}}{s_x^2} = 0.056476$$

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



Como se chegou à equação da recta?

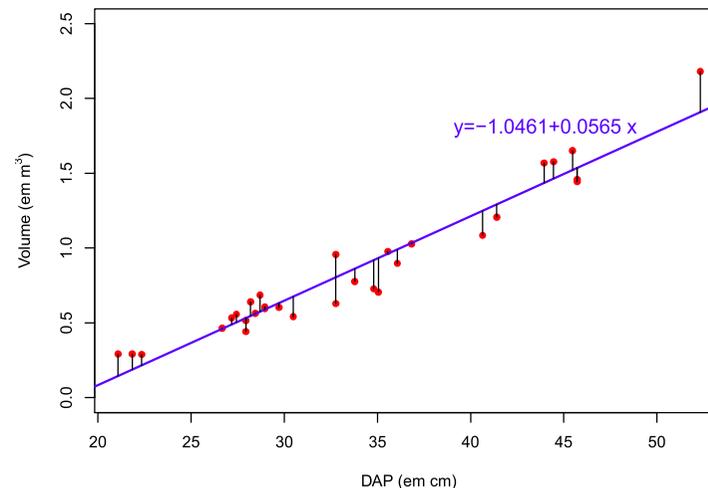
Valores ajustados e Resíduos

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).

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 .

Regressão Linear Simples - contexto descritivo no R

As regressões lineares são ajustadas no R usando o comando `lm` (as iniciais de `linear model`).

A função `lm` tem dois argumentos fundamentais:

- `formula` – identifica a **variável resposta** e as **variáveis preditoras**; numa RL simples da variável y sobre o preditor x , é da forma: $y \sim x$.
- `data` – indica o nome da *data frame* contendo os dados.

Comando R para a RLS do exemplo das cerejeiras

```
> lm( Volume ~ DAP , data=cerejeiras )
```

```
Call: lm(formula = Volume ~ DAP, data = cerejeiras)
```

```
Coefficients:
```

```
(Intercept)          DAP  
-1.04612         0.05648    <- valores ajustados de b0 e b1
```

Comandos R para o estudo da regressão

Vejamos **alguns comandos do R** úteis para estudar uma regressão.

Começemos por guardar a regressão do exemplo das cerejeiras:

```
> cerejeiras.lm <- lm(Volume ~ DAP , data=cerejeiras )
```

- **fitted** devolve os valores ajustados $\hat{y}_i = b_0 + b_1 x_i$:

```
> fitted(cerejeiras.lm)
```

```
      1      2      3      4      5      6      7      8      9     10  
0.1445051 0.1875398 0.2162296 0.4600931 0.4887829 0.5031278 0.5318176 0.5318176 0.5461625 0.5605074  
     11     12     13     14     15     16     17     18     19     20  
0.5748523 0.5891972 0.5891972 0.6322320 0.6752667 0.8043709 0.8043709 0.8617505 0.9191301 0.9334750  
[...]
```

Comandos R (cont.)

- `residuals` devolve os resíduos $e_i = y_i - \hat{y}_i$:

```
> residuals(cerejeiras.lm)
```

```
      1      2      3      4      5      6      7      8  
0.147158427 0.104123704 0.072602203 0.004303217 0.043573833 0.054714087 -0.090074800 -0.016450998  
      9     10     11     12     13     14     15     16  
0.093798219 0.002997825 0.110415357 0.005456540 0.016783279 -0.029083129 -0.134414916 -0.175736863  
[...]
```

A Soma dos Quadrados dos Resíduos, *SQRE*, pode ser calculada por:

```
> sum(residuals(cerejeiras.lm)^2)
```

```
[1] 0.4204087
```

SQRE tem unidades de medida: o quadrado das unidades de y .

Comandos R para a regressão (cont.)

- `predict` – ajusta uma regressão a novas observações, dadas numa *data frame* com nomes de preditores iguais aos do ajustamento.

```
> novos <- data.frame( DAP=c(25, 50) )  
> predict( cerejeiras.lm , new=novos )
```

```
      1      2  
0.3657781 1.7776785
```

O valor \hat{y} ajustado pela recta, para $x = 25$, é (arredondamentos aparte):

$$\begin{aligned}\hat{y} &= b_0 + b_1 x \\ \Leftrightarrow &= -1.04612 + 0.05648 \times 25 .\end{aligned}$$

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. (Ver Exercício RLS 5).
- Embora não tenha sido explicitamente exigido, a média dos resíduos e_i é nula, ou seja, $\bar{e} = 0$. (Ver Exercício RLS 5).

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 $R^2 = \frac{SQR}{SQT}$

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

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

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

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 (Ver Exercício RLS 6):

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

Exemplo das cerejeiras

O coeficiente de determinação R^2 obtem-se aplicando o comando **summary** a uma **regressão ajustada**. Surge com a designação **Multiple R-Squared**:

```
> summary(cerejeiras.lm)
```

```
Call: lm(formula = Volume ~ DAP, data = cerejeiras)
```

```
[...]
```

```
Residual standard error: 0.1204 on 29 degrees of freedom
```

```
Multiple R-squared: 0.9353, Adjusted R-squared: 0.9331
```

```
F-statistic: 419.4 on 1 and 29 DF, p-value: < 2.2e-16
```

O valor de R^2 (com maior precisão) pode ser obtido da seguinte forma:

```
> summary(cerejeiras.lm)$r.sq
```

```
[1] 0.9353199
```

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.

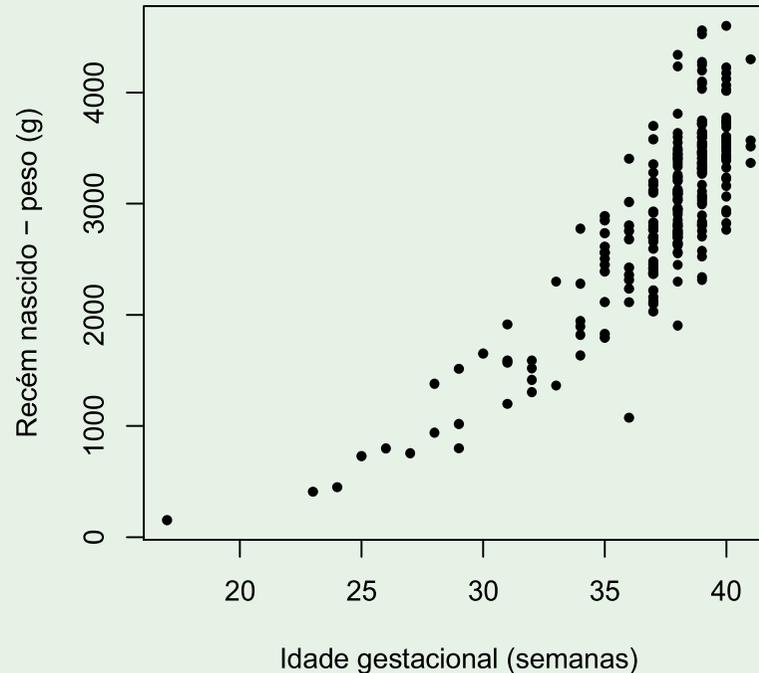
Vamos ver dois exemplos particularmente frequentes de relações não-lineares que são linearizáveis através de transformações da variável resposta e, num caso, também do preditor.

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?**

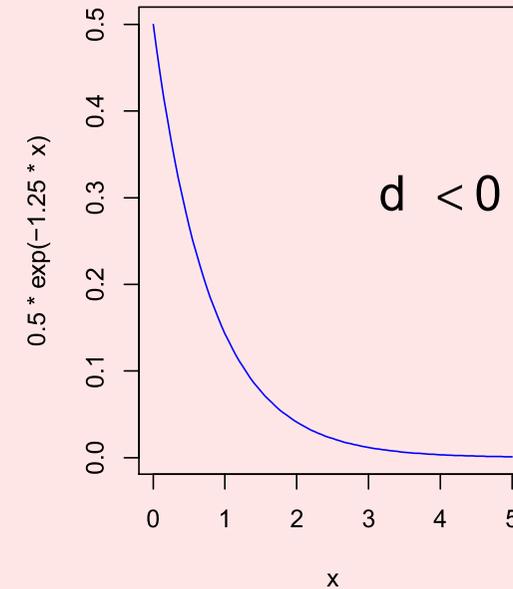
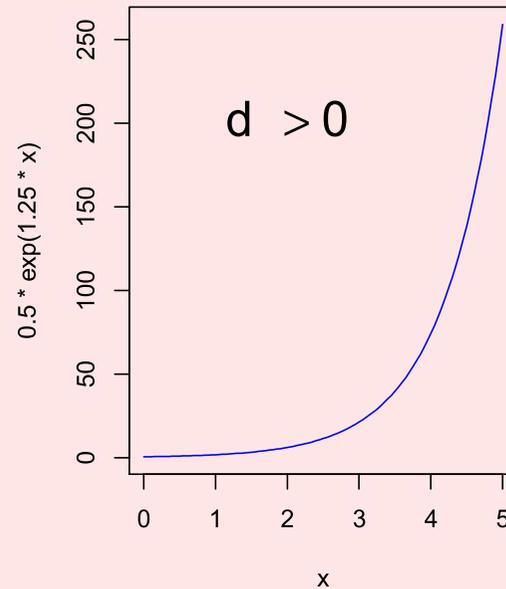
A **Regressão Não Linear** **não** faz parte do programa da disciplina. Mas **transformações linearizantes** de uma ou ambas as variáveis podem criar uma relação linear, que permita usar o Modelo Linear.

Relação exponencial

Relação exponencial

$$y = c e^{d x}$$

$$(y > 0 \quad ; \quad c > 0)$$



Transformação linearizante: $y^* = \ln(y)$ e $x^* = x$

A linearização da relação exponencial

Logaritmizando a equação da exponencial, obtém-se uma **relação linear** entre $y^* = \ln(Y)$ e x :

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

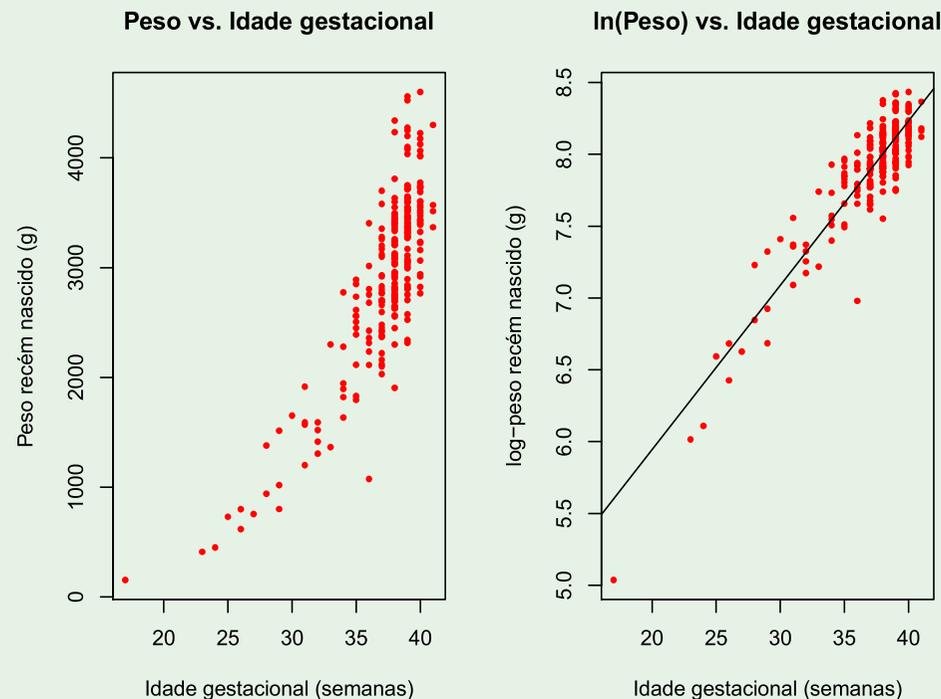
com **declive** $b_1 = d$ e **ordenada na origem** $b_0 = \ln(c)$.

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

Exemplo 3: peso de bebés à nascença

Uma linearização no peso dos bebés

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



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

Ainda a relação exponencial

Equação Diferencial da 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 ($P \frac{f'}{f} = \ln|f|$), tem-se (já que $y > 0$):

$$\ln|y(x)| = dx + K \quad \Leftrightarrow \quad y(x) = e^{K+dx} \quad \Leftrightarrow \quad y(x) = e^K e^{dx} = c e^{dx} .$$

O declive b_1 da recta é o valor constante d da taxa de variação relativa de y .

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.

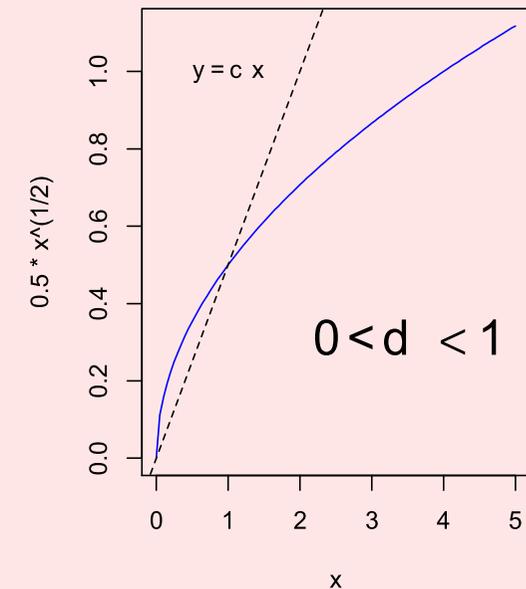
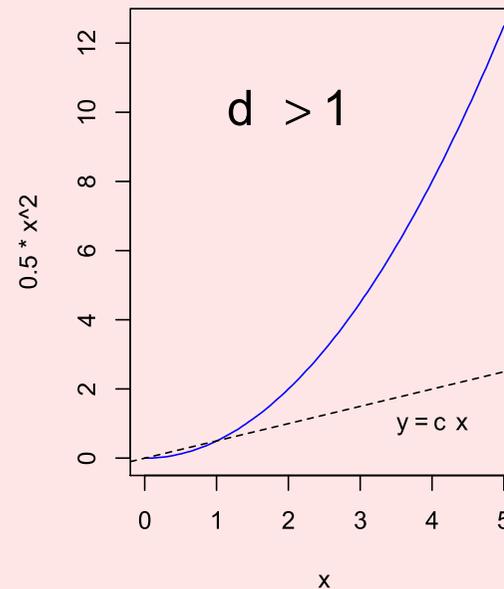
Em **1838 Verhulst** propôs uma **modelo de crescimento populacional alternativo**, prevendo os efeitos resultantes da escassez de recursos: o **modelo logístico** (não leccionado).

Relação potência ou alométrica

Relação potência

$$y = cX^d$$

$$(x, y > 0 \ ; \ c > 0)$$



Transformação linearizante: $y^* = \ln(y)$ e $x^* = \ln(x)$.

A linearização dum relação potência

Logaritmizando, obtém-se:

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

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

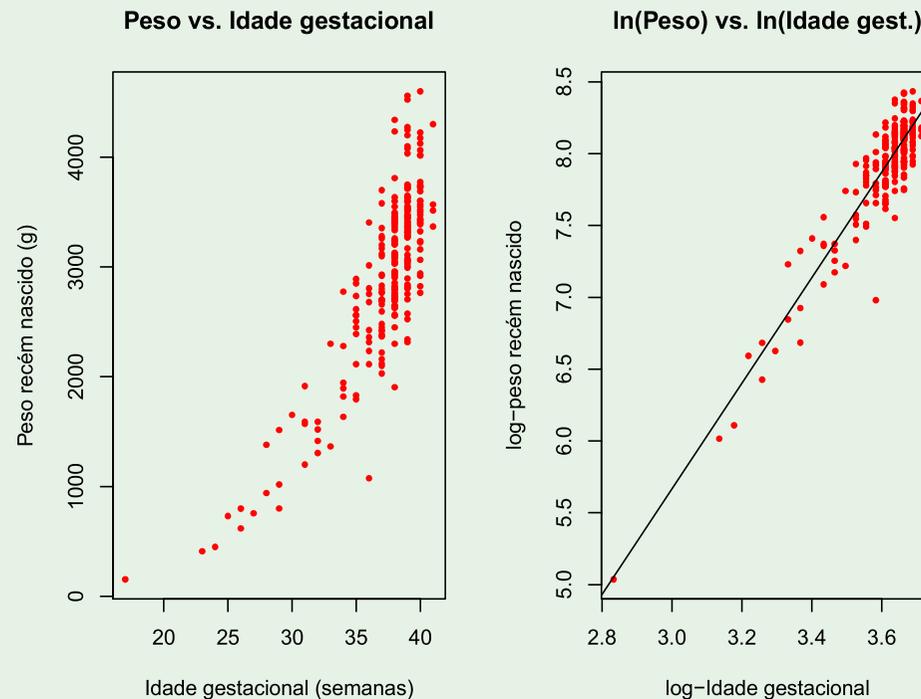
O declive b_1 da recta é o expoente d na relação potência.

A ordenada na origem é $b_0 = \ln(c)$, ou seja, $c = e^{b_0}$.

Outra linearização no Exemplo 3

Outra linearização dos pesos dos bebês

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 $d = 1$.

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.

Conceitos que dependem da escala de Y , como $SQRE$ e R^2 , **não são directamente comparáveis**, antes e após uma transformação da variável resposta.

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, através duma **regressão não linear** (não estudada nesta disciplina).