Regressão - um pouco de história

O critério de mínimos quadrados surge no início do Século XIX, associado ao trabalho do francês Legendre, motivado pelo problema de conciliar diferentes observações geodésicas e astronómicas que se sabia estarem afectadas por erros de observação.

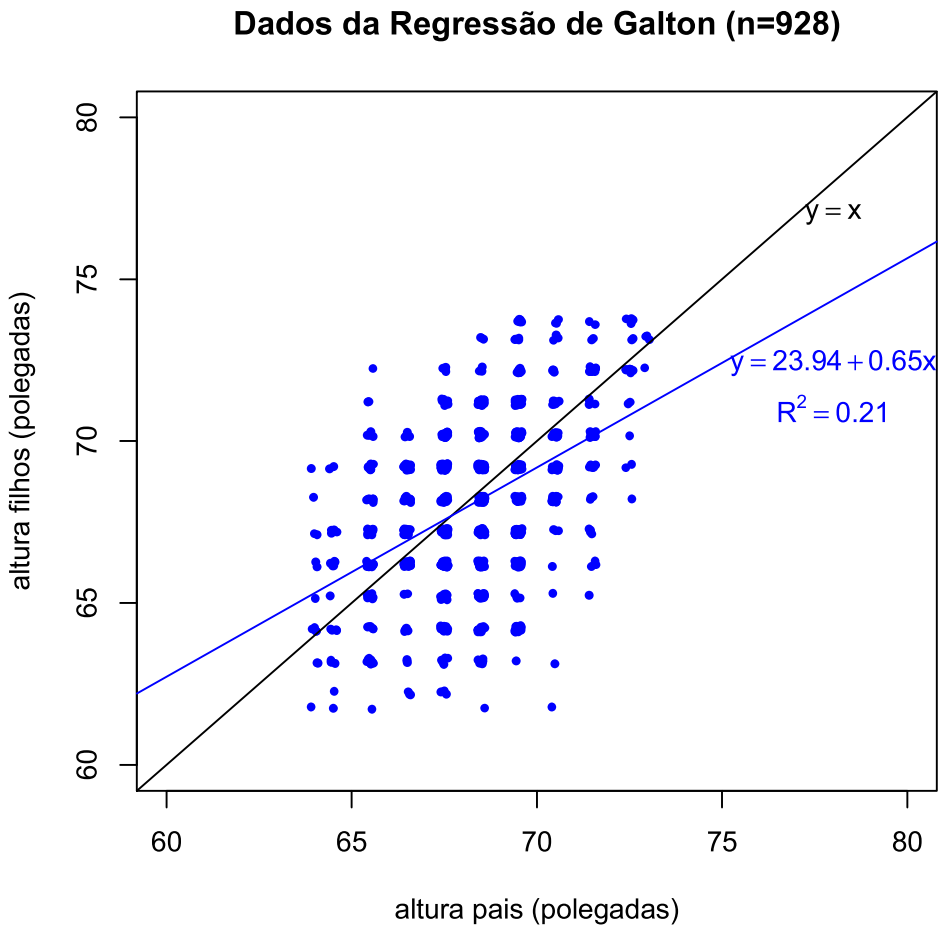
A designação Regressão tem origem num estudo de Francis Galton (1886), relacionando a altura de $n = 928$ jovens adultos com a altura (média) dos pais. Galton inventou a designação eugenia, conceito que era considerado respeitável até às primeiras décadas do Século XX.

Galton constatou que pais com alturas acima da média tinham tendência a ter filhos com altura acima da média - mas menos que os pais (análogo para os abaixo da média).

Galton chamou ao seu artigo *Regression towards mediocrity in hereditary stature*. A expressão regressão ficou associada ao método devido a esta acaso histórico.

Um pouco de história (cont.)

Curiosamente, o exemplo de Galton tem um valor muito baixo do Coeficiente de Determinação.



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.

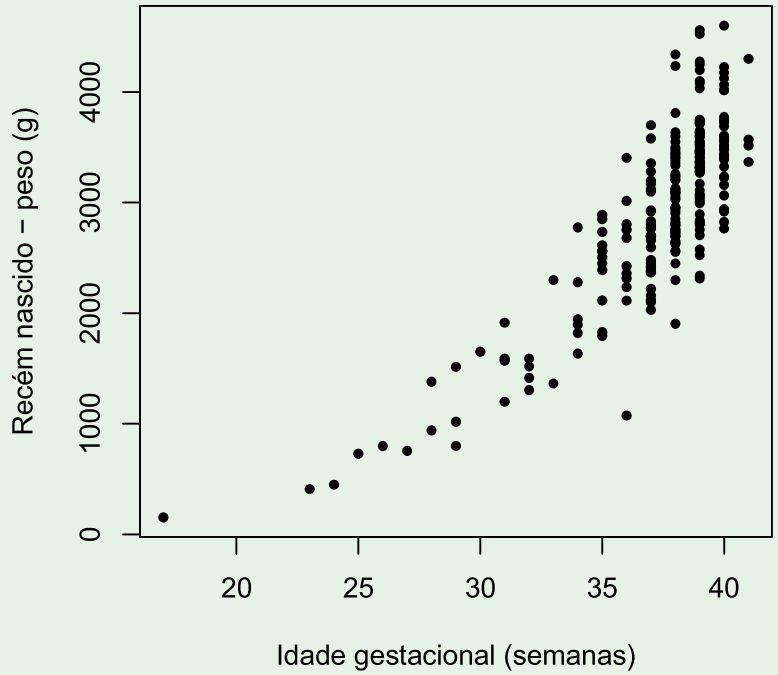
Vamos ver dois exemplos particularmente frequentes de relações não-lineares que são linearizáveis através de transformações da variável resposta e, num caso, também do preditor.

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 ($y = c\,\mathrm{e}^{dx}$)?
  - 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?

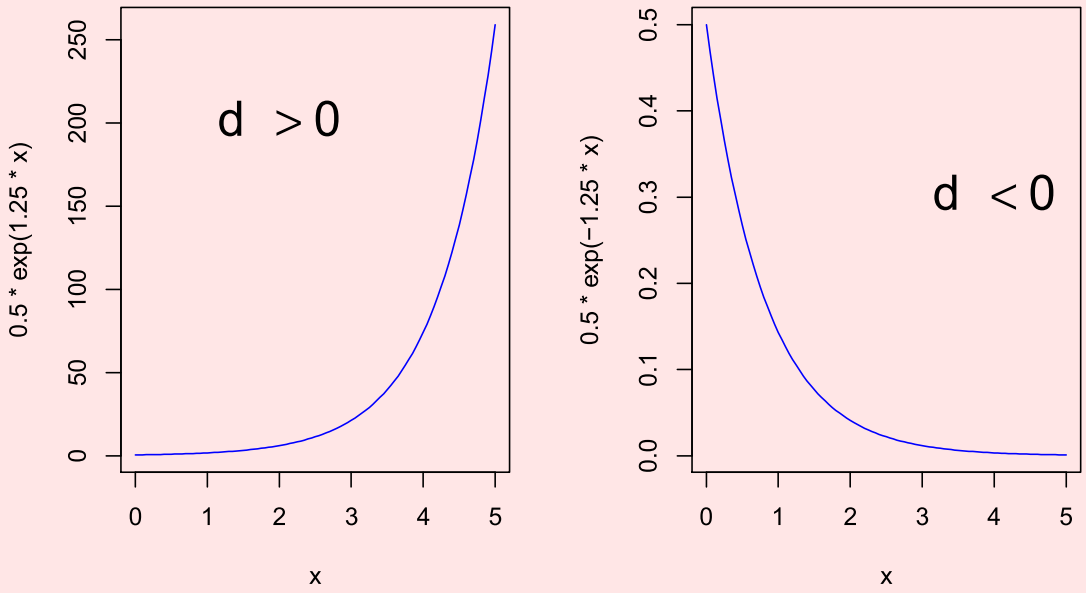
A Regressão Não Linear não faz parte do programa da disciplina. Mas transformações linearizantes de uma ou ambas as variáveis podem criar uma relação linear, que permita usar o Modelo Linear.

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:

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

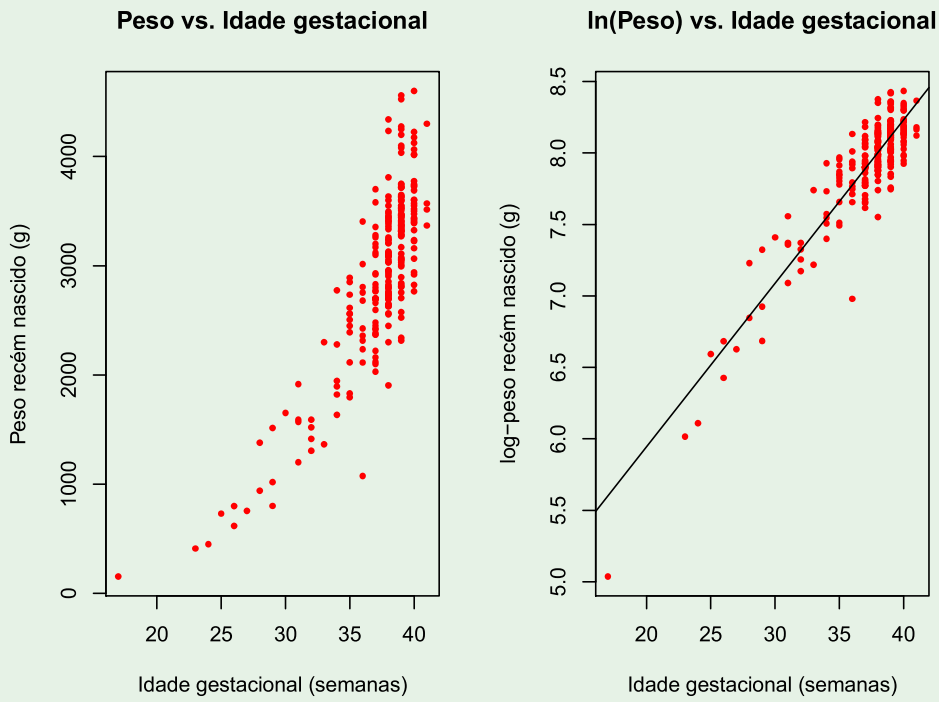
com declive $b_1 = d$ e ordenada na origem $b_0 = \ln(c)$.

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

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

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.

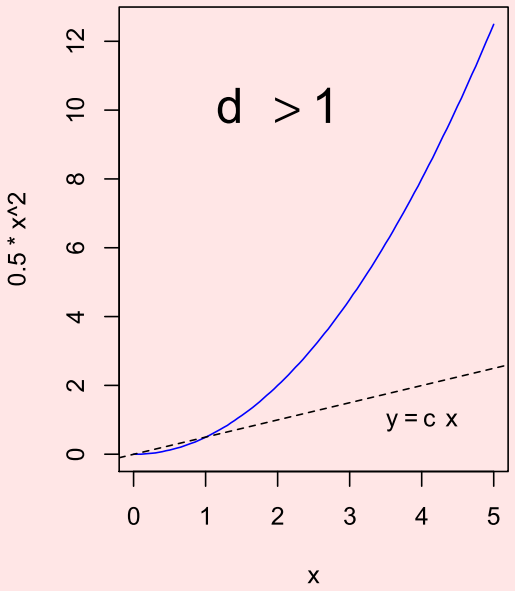
Em 1838 Verhulst propôs uma modelo de crescimento populacional alternativo, prevendo os efeitos resultantes da escassez de recursos: o modelo logístico (não leccionado).

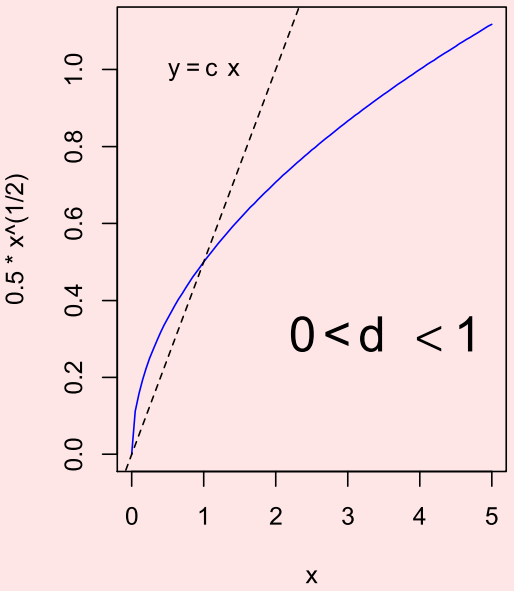
Relação potência ou alométrica

Relação potência

$$y = c\,x^{d}$$

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





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

A linearização duma 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 $x^* = \ln(x)$.

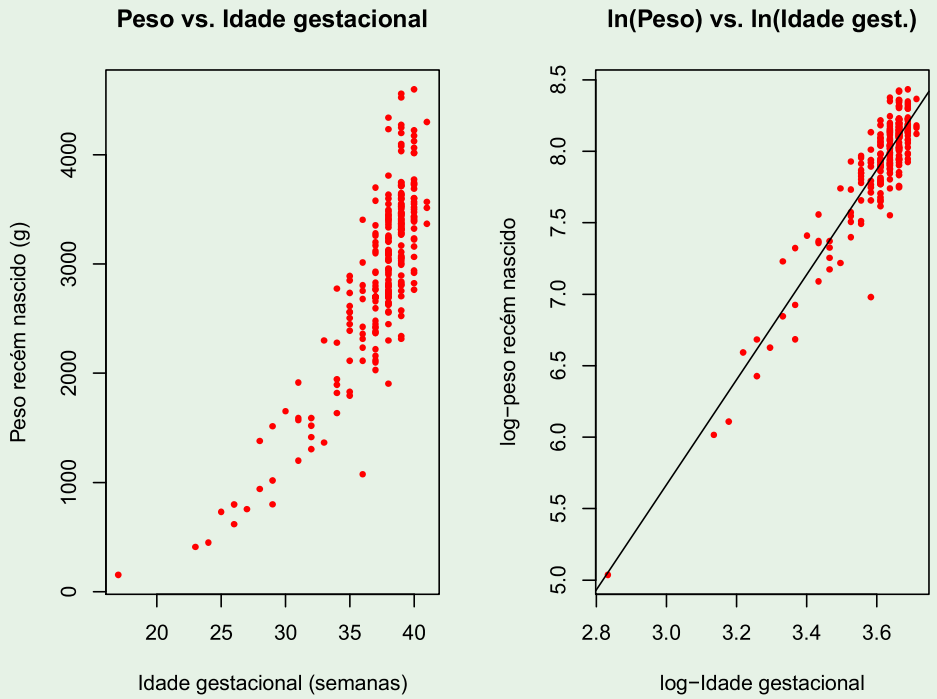
O declive b_1 da recta é o expoente d na relação potência.

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 bebés

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,

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

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).