Modelos Gaussianos com covariáveis

Paulo Inácio Prado e João L.F. Batista

BIE5781 - Pós-Graduação em Ecologia USP

Novembro de 2021

Objetivo da Aula

Os objetivos dessa aula são:

1. Apresentar o conceito de modelos com parâmetros que são funções de outras medidas;
2. Apresentar os modelos deste tipo para a distribuição Gaussiana:
3. Mostrar os comandos básicos no R para ajustar e avaliar esses modelos;
4. Mostrar que alguns deles são modelos de regressão linear e não-linear;

Onde estamos

Até agora vimos:

- Modelos de várias distribuições com parâmetros constantes:

$$
\begin{aligned}
Y &\sim N(\mu = \alpha_0,\ \sigma = \beta_0) \\
Y &\sim P(\lambda = \alpha_0) \\
&\dots
\end{aligned}
$$

Para onde vamos

Hoje veremos modelos de distribuições Gaussianas cujos parâmetros são funções de outra medidas X_i:

- Modelos lineares:

$$Y \sim N(\mu = \alpha_0 + \alpha_1 X_1,\ \sigma = \beta_0)$$

- Modelos não-lineares:

$$Y \sim N(\mu = \alpha_0 X_1^{\alpha_1},\ \sigma = \beta_0)$$

- Modelos com σ também variável:

$$
\begin{aligned}
Y &\sim N(\mu = \alpha_0 + \alpha_1 X_1,\ \sigma = e^{\beta_0 + \beta_1 X_1}) \\
Y &\sim N(\mu = \alpha_0 X_1^{\alpha_1},\ \sigma = e^{\beta_0 + \beta_1 X_1})
\end{aligned}
$$

Ou seja

Uma expressão geral para modelos Gaussianos

$$Y \sim N(\mu = f(X_i), \sigma = g(X_i))$$

Em palavras

Y é uma variável aleatória, cujos valores têm probabilidades definidas pela função de densidade Gaussiana, cujos parâmetros μ e σ são funções de variáveis X_i.

Conceitos

Y: variável dependente, ou variável resposta

X_i: variáveis independentes ou preditoras

Modelos lineares

- Com uma preditora

$$Y \sim N(\mu = \alpha_0 + \alpha_1 X_1,\ \sigma = \beta_0)$$

- Com muitas preditoras

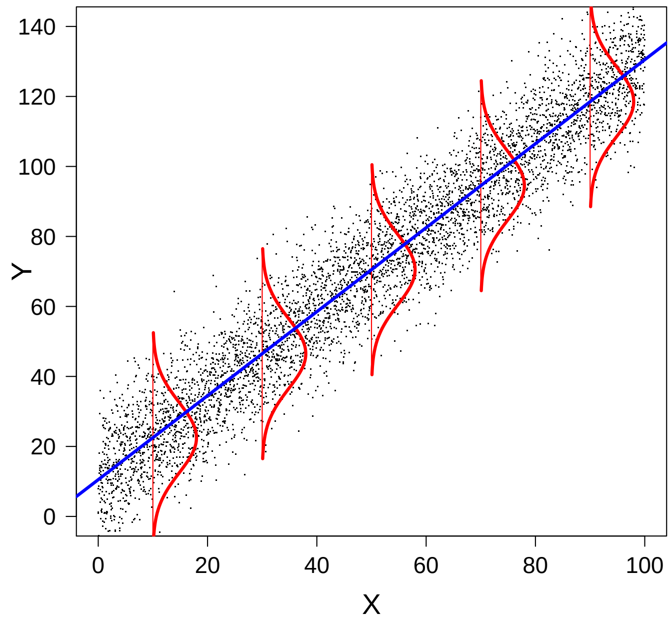
$$Y \sim N(\mu = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2, + \ldots + \alpha_i X_i,\ \sigma = \beta_0)$$

- Generalizando

$$Y \sim N(\mu = \alpha_0 + \sum \alpha_i X_i,\ \sigma = \beta_0)$$

- Chamamos $[\alpha_0 + \sum \alpha_i X_i]$ de **preditor linear** (detalhes em breve).

Modelo Linear Gaussiano com uma preditora e σ constante



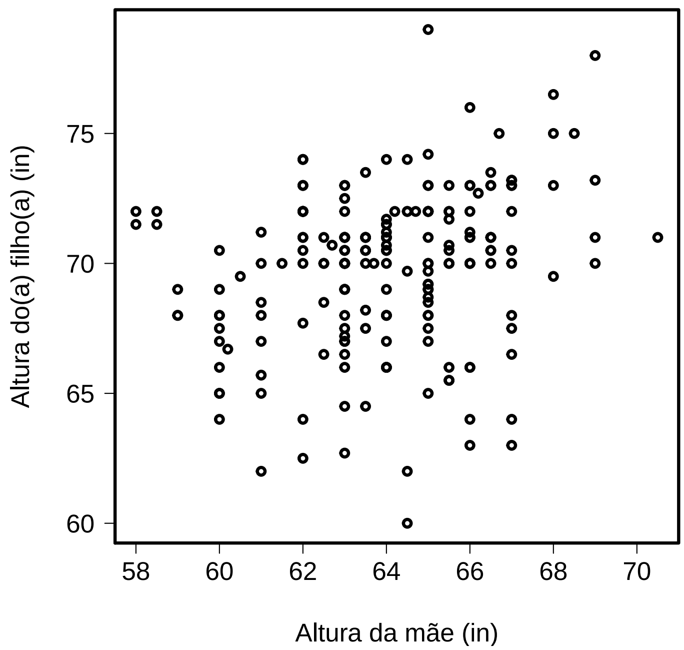
Um exemplo de modelo linear com uma preditora

- Altura de filhos e de seus pais (Galton 1885 [1])
- Preditora: altura da mãe (polegadas)
- Resposta: altura de um filho ou filha (polegadas)
- Tamanho da amostra = 197

```
> head(galton1, 5)
   Family Father Mother Gender Height Kids
1       1   78.5   67.0      M   73.2    4
5       2   75.5   66.5      M   73.5    4
9       3   75.0   64.0      M   71.0    2
11      4   75.0   64.0      M   70.5    5
16      5   75.0   58.5      M   72.0    6
```

[1] https://www.randomservices.org/random/data/Galton.html

Um exemplo de modelo linear com uma preditora



Modelo linear com uma variável: Ajuste no R

Função de log-verossimilhança negativa

```
> galton.LL1 <- function(a0, a1, b0){
+     PL <- a0 + a1*galton1$Mother
+     -sum( dnorm(galton1$Height, mean = PL, sd = b0 ,
+                  log=T))}
```

Ajuste numérico do modelo no R

```
> galton.m1 <- mle2(galton.LL1,
+                     start = list(a0=0, a1=1, b0=5))
```

Modelo linear com uma variável: Ajuste no R

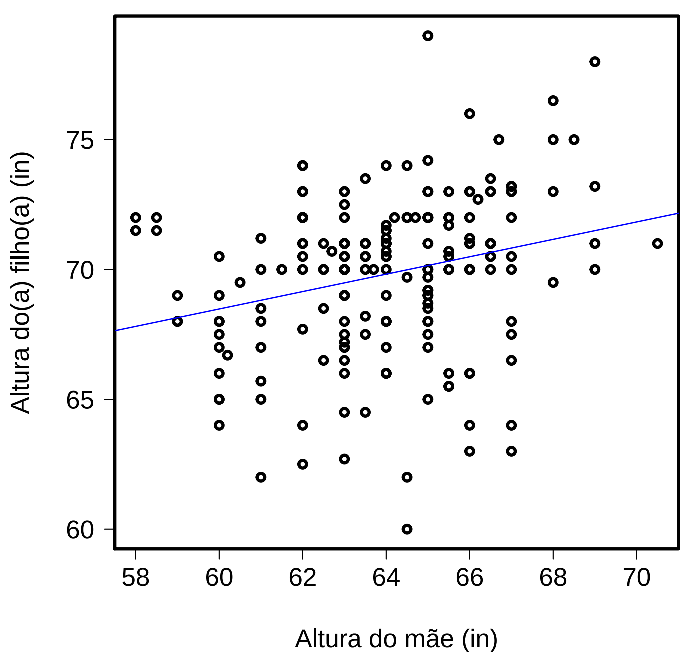
MLEs

```
> coef(galton.m1)

         a0          a1          b0
48.3657449   0.3352066   2.9485752
```

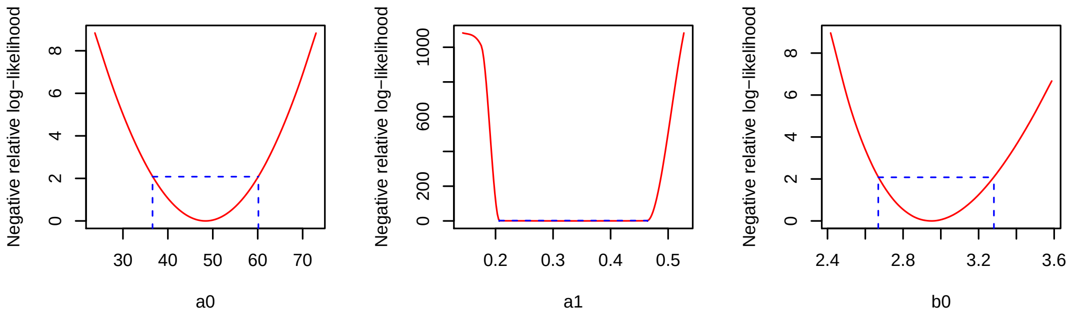
Log-verossimilhança

```
> logLik(galton.m1)

'log Lik.' -492.5617 (df=3)
```

Inspeção visual do ajuste



Intervalos de plausibilidade



Regressão não-linear simples no R – Ajuste

Otimização

```
> galton.LL1 <- function(a0, a1, b0){
+     PL <- a0 + a1*galton1$Mother
+     -sum( dnorm(galton1$Height, mean = PL, sd = b0 ,
+                 log=T))}
```

```
> galton.m1 <- mle2(galton.LL1,
+                   start = list(a0=0, a1=1, b0=5))
```

Regressão linear simples no R – Ajuste

Mínimos quadrados

```
> galton.lm1 <- lm(Height ~ Mother, data = galton1)
```

Regressão linear simples no R – MLEs

Otimização

```
> coef(galton.m1)

        a0          a1          b0 
48.3657449   0.3352066   2.9485752 
```

Mínimos quadrados

```
> coef(galton.lm1)

(Intercept)      Mother 
 48.3699889   0.3351405 
```

Regressão linear simples no R- Log-verossimilhança

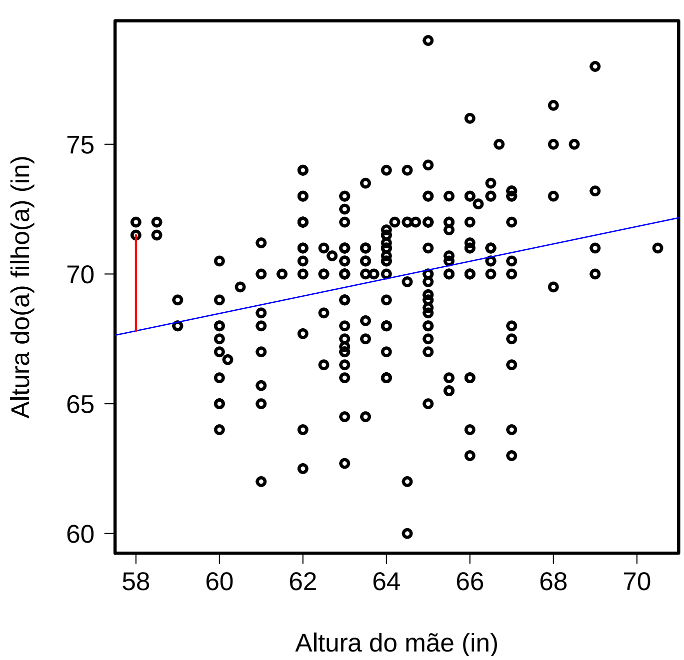
Otimização

```
> logLik(galton.m1)

'log Lik.' -492.5617 (df=3)
```

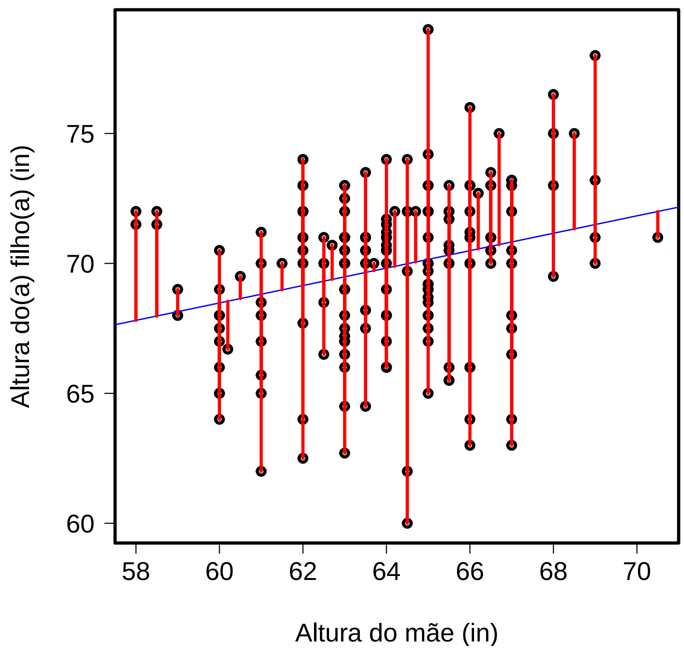
Mínimos quadrados

```
>          logLik(galton.lm1)

'log Lik.' -492.5617 (df=3)
```

Quadrados mínimos



Quadrados mínimos



Duas maneiras de interpretar uma regressão linear

Resposta como variável Gaussiana

$$Y \sim N(\mu = \alpha_0 + \sum \alpha_i X_i,\ \sigma = \beta_0)$$

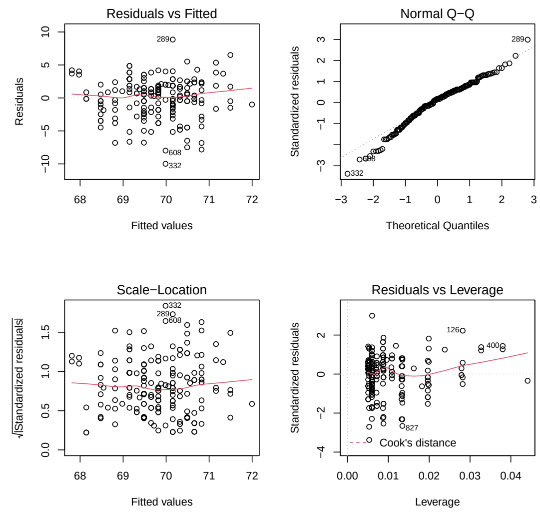
Resposta como medida + erro

$$Y = \alpha_0 + \sum \alpha_i X_i + \epsilon,$$

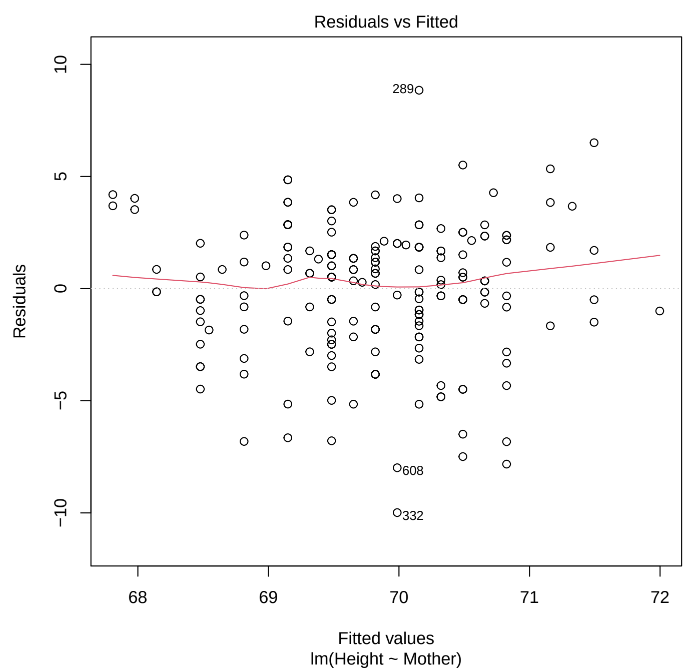
$$\epsilon \sim N(\mu = 0,\ \sigma = \beta_0)$$

Análise de resíduos

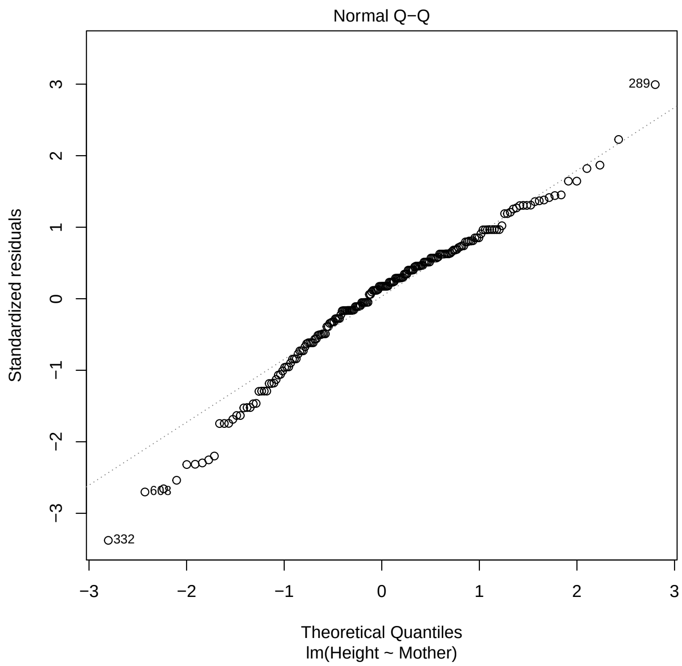
```
> par(mfrow=c(2,2))
> plot(galton.lm1)
> par(mfrow=c(1,1))
```



Análise de resíduos: têm média zero e variância constante?



Análise de resíduos: têm distribuição normal?



Preditores lineares

Todos que podem ser escritos na forma $\Theta = \sum \alpha_i X_i$, sendo Θ uma resposta.

exemplos

- $\alpha_0 + \alpha_1 X_1$, sendo $X_0 = 1$
- $\alpha_0 + \alpha_1 X_1 + \alpha_2 X_1^2$, pois podemos chamar X_1^2 de X_2
- $\alpha_0 + \alpha_1 X_1 + \alpha_2 \ln(X_1)$, pois podemos chamar $\ln X_1$ de X_2
- $\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_1 X_2$, pois podemos chamar $X_1 X_2$ de X_3
- ...

Modelos lineares múltiplos aditivos

- O preditor é a soma de variáveis independentes isoladas, cada uma multiplicada por seu coeficiente
- A soma **não** inclui variáveis independentes que são produtos de outras presentes no modelo:

$$\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_i X_i$$

- Se a uma variável independente X_i dobra, seu efeito no preditor vai aumentar em $2 \times \alpha_i$, não importando o valor das outras variáveis.
- Ou seja, os efeitos das preditoras são **aditivos**. Não são afetados pelo valor das demais variáveis.

Modelos lineares múltiplos com interações

- O preditor inclui variáveis que são produtos de pelos menos outras duas presentes no modelo, por exemplo

$$\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_1 X_2$$

- Se a uma variável independent X_i dobra, seu efeito no preditor vai aumentar em
$$2 \times \alpha_i + 2 \times \alpha_3 X_2.$$
- Ou seja, o efeito de uma preditora depende do valor de outra(s). Há uma **interação** estatística.

Exemplo de preditor linear aditivo

Duas preditoras:

X_1 : altura da mãe, contínua

X_2 : Gênero binário (Masculino = 0, Feminino = 1)

Preditor:

expressão : $\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2$

masculino : $\alpha_0 + \alpha_1 X_1 + \alpha_2 \times 0 \;=\; \alpha_0 + \alpha_1 X_1$

feminino : $\alpha_0 + \alpha_1 X_1 + \alpha_2 \times 1 \;=\; [\alpha_0 + \alpha_2] + \alpha_1 X_1$

Exemplo de preditor linear com uma interação

Duas preditoras:

X_1 : altura da mãe, contínua

X_2 : Gênero binário (Masculino = 0, Feminino = 1)

Preditor:

expressão : $\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_1 X_2$

masculino : $\alpha_0 + \alpha_1 X_1 + \alpha_2 \times 0 \; + \alpha_3 X_1 \times 0 = \; \alpha_0 + \alpha_1 X_1$

feminino :

$$\alpha_0 + \alpha_1 X_1 + \alpha_2 \times 1 + \alpha_3 X_1 \times 1 = \; [\alpha_0 + \alpha_2] + [\alpha_1 + \alpha_3] X_1$$

Gráfico do exemplo aditivo

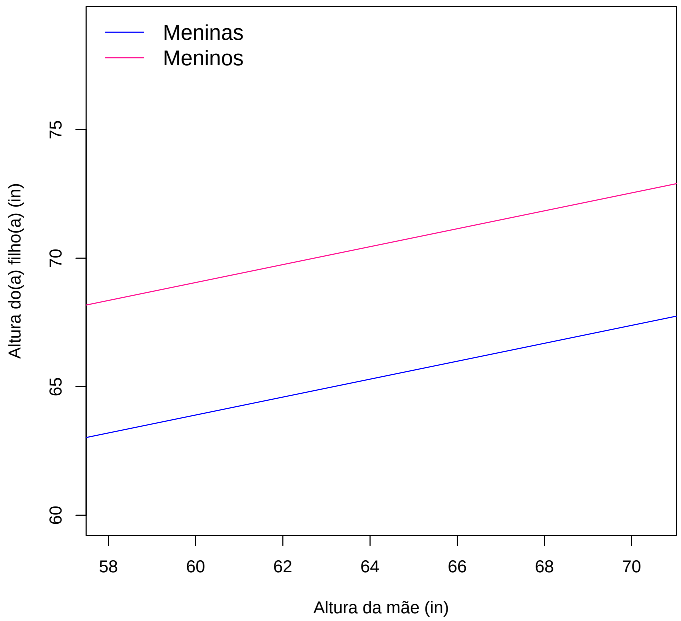
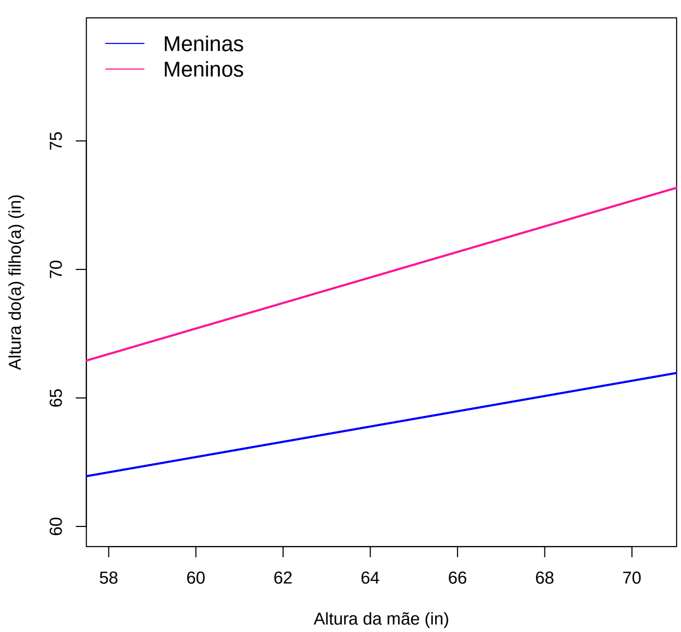


Gráfico do exemplo com interação



Preditores não lineares

Todos que não podem ser reduzidos à forma $\Theta = \sum \alpha_i X_i$.

exemplos

- $\alpha_0 X_1^{\alpha_1}$
- $\alpha_0 e^{\alpha_1 X_1}$
- $\frac{\alpha_1 X_1}{1+\alpha_1 \alpha_2 X_1}$
- E muitas, muitas mais.

Regressão não-linear

Resposta como variável Gaussiana

$$Y \sim N(\mu = f(X_i),\ \sigma = \beta_0)$$

Resposta como medida + erro

$$Y = f(X_i) + \epsilon,$$

$$\epsilon \sim N(\mu = 0,\ \sigma = \beta_0)$$

Um exemplo de modelo não-linear

- N de espécies de plantas endêmicas em áreas costeiras da Califórnia (Jhonson *et al* 1968 [2])
- Preditora: área (mi^2)
- Resposta: número de espécies registradas na área
- Tamanho da amostra = 9

```
> head(johnson, 4)
                         area Nspp
Tiburon Peninsula          5.9  370
San Francisco             45.0  640
Santa Barbara area       110.0  680
Santa Monica Mountains 320.0  640
```

[2] http://math.hws.edu/~mitchell/SpeciesArea/index.html

Modelo não-linear: Ajuste no R

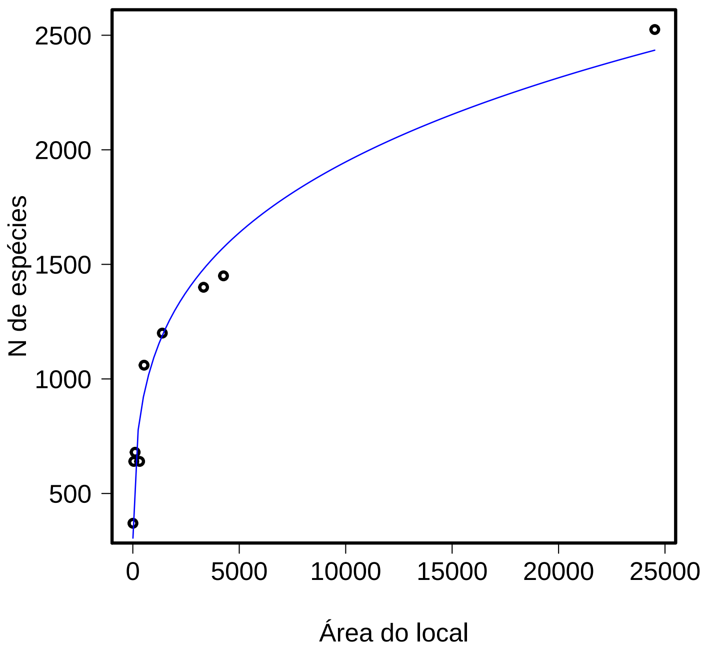
Função de log-verossimilhança negativa

```
> johnson.LL <- function(a0, a1, b0){
+     PL <- a0*johnson$area^a1
+     -sum( dnorm(johnson$Nspp, mean = PL, sd = b0 ,
+                 log=T))}
```

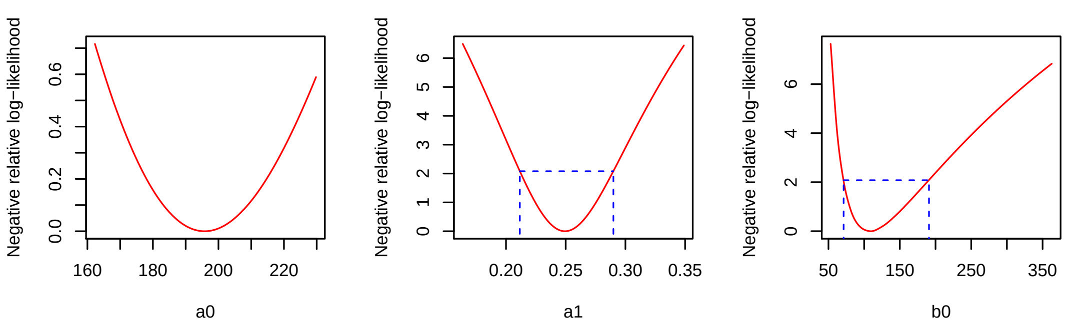
Ajuste numérico do modelo no R

```
> johnson.m1 <- mle2(johnson.LL,
+                    start = list(a0=200, a1=0.25, b0=10)
+                    )
```

Inspeção visual do ajuste



Intervalos de plausibilidade



Regressão não-linear no R – Ajuste

Mínimos quadrados

```
> johnson.nls <- nls(Nspp ~ a0*area^a1, data = johnson)

Warning in nls(Nspp ~ a0 * area^a1, data = johnson):
No starting values specified for some parameters.
Initializing 'a0', 'a1' to '1.'.
Consider specifying 'start' or using a selfStart model
```

Regressão não-linear no R – MLEs

Otimização

```
> coef(johnson.m1)

         a0          a1          b0
196.0414936   0.2492566 107.7050628
```

Mínimos quadrados

```
> coef(johnson.nls)

         a0          a1
195.7665566   0.2494165
```

Regressão não-linear no R – Verossimilhança

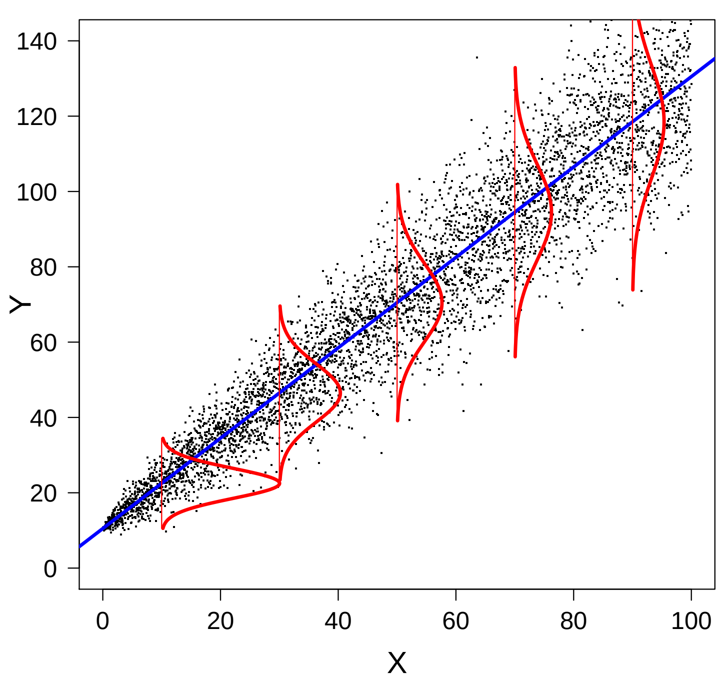
Otimização

```
> logLik(johnson.m1)

'log Lik.' -54.87376 (df=3)
```

Mínimos quadrados

```
> logLik(johnson.nls)

'log Lik.' -54.8737 (df=3)
```

Heteroscedasticidade



Um modelo linear com σ variável

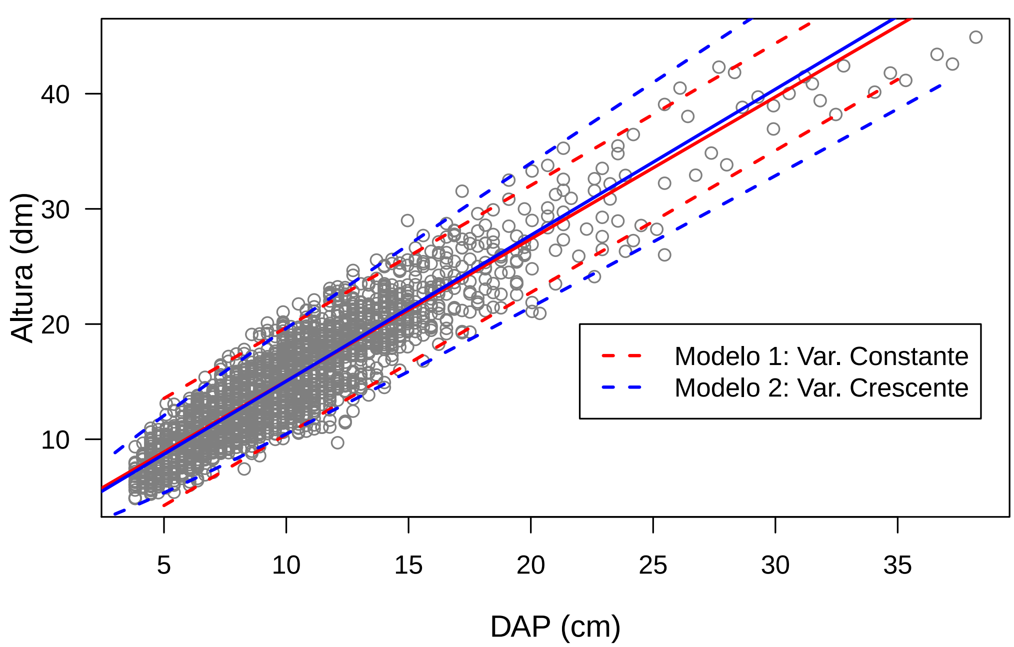
Expressão simbólica

$$Y \sim N(\mu = \alpha_0 + \alpha_1 X_1,\ \sigma = e^{\beta_0 + \beta_1 X_1})$$

Função de verossimilhança no R

```
> LL3 <- function(a0, a1, b0, b1){
+     LP.mu <- a0 + a1*X
+     LP.sig <- exp(b0 + b1*X)
+     -sum(dnorm(Y, mean = LP.mu, sd = LP.sig,
+                start = list(a0 = 0, a1= 1, b0 = 0.1, b1 = -1),
+                log = TRUE))
+ }
```

Relação Altura-DAP em *E.grandis*: Comparação Gráfica



Modelos Gaussianos: Resumo

1. Nos modelos Gaussianos podemos ter duas funções preditivas:
   - uma para média,
   - outra da o desvio padrão
2. Nos modelo de Regressão <u>Linear</u> Clássicos:
   - $\Rightarrow$ a média é uma função *linear*.
   - $\Rightarrow$ o desvio padrão é constante.
3. Nos modelo de Regressão <u>Não-Linear</u> Clássicos:
   - $\Rightarrow$ a média é uma função *não-linear*.
   - $\Rightarrow$ o desvio padrão é constante.
4. O contexto dos modelos clássicos pode ser ampliado:
   - $\Rightarrow$ Acrescentando uma função preditiva para o desvio padrão.
   - $\Rightarrow$ Outras formas? Sim! *muitas outras*!