



# Introdução ao uso do software R

Fernando de Pol Mayer<sup>1</sup>   Rodrigo Sant'Ana<sup>2</sup>

<sup>1</sup>Laboratório de Estatística Ambiental (LEA)  
Instituto de Matemática, Estatística e Física (IMEF)  
Universidade Federal do Rio Grande (FURG)  
fernando.mayer@furg.br

<sup>2</sup>Instituto Albatroz  
oc.rodigosantana@gmail.com



# Sumário

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

- 1 Distribuições de probabilidade
- 2 Inferência
- 3 Regressão e correlação
  - Regressão
    - Estimação dos parâmetros
  - Correlação
- 4 Análise de Variância
- 5 Modelos Lineares Generalizados



# Sumário

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

- 1 Distribuições de probabilidade
- 2 Inferência
- 3 Regressão e correlação
  - Regressão
    - Estimação dos parâmetros
  - Correlação
- 4 Análise de Variância
- 5 Modelos Lineares Generalizados



# Distribuições de probabilidade

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

A maioria das distribuições de probabilidade tradicionais estão implementadas no R, e podem ser utilizadas para substituir as tabelas estatísticas tradicionais. Existem 4 itens fundamentais que podem ser calculados para cada distribuição:

- $d^*$  Calcula a densidade de probabilidade ou probabilidade pontual
- $p^*$  Calcula a função de probabilidade acumulada
- $q^*$  Calcula o quantil correspondente a uma dada probabilidade
- $r^*$  Gera números aleatórios (ou “pseudo-aleatórios”)



# Distribuições de probabilidade

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

As distribuições de probabilidade mais comuns são:

Distribuição	Nome no R	Parâmetros
Binomial	*binom	size, prob
$\chi^2$	*chisq	df
Normal	*norm	mean, sd
Poisson	*pois	lambda
t	*t	df
Uniforme	*unif	min, max



# Distribuições de probabilidade

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

Alguns exemplos:

```
# valores críticos de z com alfa = 0,05 (bilateral)  
qnorm(0.025)
```

```
[1] -1.96
```

```
qnorm(0.975)
```

```
[1] 1.96
```

```
# valores críticos de t com diferentes G.L.  
qt(0.025, df = 9)
```

```
[1] -2.2622
```

```
qt(0.025, df = 900)
```

```
[1] -1.9626
```



# Distribuições de probabilidade

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

Intervalos de confiança: suponha uma amostra de  $n = 5$ , com  $\bar{x} = 83$  e  $s = 12$ . Um intervalo de 95% de confiança ( $\alpha = 0.05$ ) para  $\mu$  pode ser calculado como:

```
## Dados
xbarra <- 83
desvio <- 12
n <- 5

## Erro padrão
erro <- desvio/sqrt(n)

## Média - erro
xbarra + erro * qt(0.025, df = n)

[1] 69.205

## Média + erro
xbarra + erro * qt(0.975, df = n)

[1] 96.795
```



# Sumário

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

- 1 Distribuições de probabilidade
- 2 Inferência
- 3 Regressão e correlação
  - Regressão
    - Estimação dos parâmetros
  - Correlação
- 4 Análise de Variância
- 5 Modelos Lineares Generalizados





# Base de dados

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

```
dados <- read.table("../dados/crabs.csv", header = T,  
                     sep = ";", dec = ",")
```

```
str(dados)
```

```
'data.frame': 156 obs. of 7 variables:
```

```
$ especie: Factor w/ 2 levels "azul","laranja": 1 1 1 1 1 1 1 1 1 1 ...
```

```
$ sexo : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
```

```
$ FL : num 8.1 8.8 9.2 9.6 10.8 11.6 11.8 12.3 12.6 12.8 ...
```

```
$ RW : num 6.7 7.7 7.8 7.9 9 9.1 10.5 11 10 10.9 ...
```

```
$ CL : num 16.1 18.1 19 20.1 23 24.5 25.2 26.8 27.7 27.4 ...
```

```
$ CW : num 19 20.8 22.4 23.1 26.5 28.4 29.3 31.5 31.7 31.8 ...
```

```
$ BD : num 7 7.4 7.7 8.2 9.8 10.4 10.3 11.4 11.4 11 ...
```



# Testes de hipótese

## Teste-t para uma amostra

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

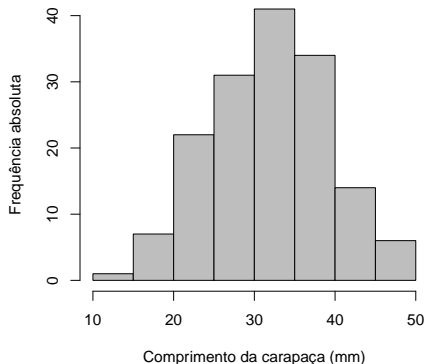
Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

```
hist(dados$CL, main = "", ylab = "Frequência absoluta",  
      xlab = "Comprimento da carapaça (mm)", col = "grey")
```





# Testes de hipótese

## Teste-t para uma amostra

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

Procedimentos gerais para um teste de hipótese

- (1) Definir a hipótese nula ( $H_0$ ) e a alternativa ( $H_1$ )
- (2) Definir um nível de **significância**  $\alpha$  (ex.:  $\alpha = 0,05$ ), que irá determinar o nível de **confiança**  $100(1 - \alpha)\%$  do teste
- (3) Determinar a **região de rejeição** com base no nível de significância  $\rightarrow t_{crit}$
- (4) Calcular a **estatística de teste**, sob a hipótese nula

$$t_{calc} = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$

- (5) Rejeitar a hipótese nula se a estatística de teste calculada estiver dentro da região de rejeição ( $t_{calc} > t_{crit}$ )
  - Alternativamente, calcula-se o p-valor, que é a probabilidade de se obter um valor de  $t$  igual ou maior do que  $t_{calc}$



# Testes de hipótese

## Teste-t para uma amostra

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

- Testar a hipótese de que a média ( $\mu$ ) de CL é igual a 30 mm (com 95% de confiança)
- As hipóteses são

$$H_0 : \mu = 30$$

$$H_1 : \mu \neq 30$$



# Testes de hipótese

## Teste-t para uma amostra

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

```
t.test(dados$CL, mu = 30, alternative = "two.sided",  
       conf.level = 0.95)
```

One Sample t-test

data: dados\$CL

t = 3.4627, df = 155, p-value = 0.0006913

alternative hypothesis: true mean is not equal to 30

95 percent confidence interval:

30.861 33.147

sample estimates:

mean of x

32.004



# Testes de hipótese

## Teste-t para uma amostra

### Fazendo manualmente

```
## Dados
```

```
xbarra <- mean(dados$CL)
```

```
mu0 <- 30
```

```
dp <- sd(dados$CL)
```

```
n <- nrow(dados)
```

```
# t calculado
```

```
(tcalc <- (xbarra - mu0)/(dp/sqrt(n)))
```

```
[1] 3.4627
```

```
# t critico (não é apresentado no resultado)
```

```
qt(0.025, df = n - 1, lower.tail = FALSE)
```

```
[1] 1.9754
```

```
# valor p (multiplicado por 2 pois o teste é bilateral)
```

```
pt(tcalc, df = n - 1, lower.tail = FALSE) * 2
```

```
[1] 0.00069135
```

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs



# Testes de hipótese

## Teste-t para uma amostra

**Detalhe:** O teste pode ser armazenado em um objeto para futuras referências

```
teste <- t.test(dados$CL, mu = 30, alternative = "two.sided",  
               conf.level = 0.95)
```

```
names(teste)
```

```
[1] "statistic"      "parameter"      "p.value"        "conf.int"  
[5] "estimate"       "null.value"     "alternative"     "method"  
[9] "data.name"
```

```
teste$statistic
```

```
      t  
3.4627
```

```
teste$p.value
```

```
[1] 0.00069135
```



# Testes de hipótese

## Teste-t para uma amostra

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

- Testar a hipótese de que a média ( $\mu$ ) de CL é menor do que 30 mm (com 95% de confiança)
- As hipóteses são

$$H_0 : \mu \leq 30$$

$$H_1 : \mu > 30$$





# Testes de hipótese

## Teste-t para uma amostra

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

```
t.test(dados$CL, mu = 30, alternative = "greater",  
       conf.level = 0.95)
```

One Sample t-test

data: dados\$CL

t = 3.4627, df = 155, p-value = 0.0003457

alternative hypothesis: true mean is greater than 30

95 percent confidence interval:

31.046      Inf

sample estimates:

mean of x

32.004



# Testes de hipótese

## Teste-t para uma amostra

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

- Testar a hipótese de que a média ( $\mu$ ) de CL é maior do que 30 mm (com 95% de confiança)
- As hipóteses são

$$H_0 : \mu \geq 30$$

$$H_1 : \mu < 30$$



# Testes de hipótese

## Teste-t para uma amostra

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

```
t.test(dados$CL, mu = 30, alternative = "less",  
       conf.level = 0.95)
```

One Sample t-test

data: dados\$CL

t = 3.4627, df = 155, p-value = 0.9997

alternative hypothesis: true mean is less than 30

95 percent confidence interval:

-Inf 32.961

sample estimates:

mean of x

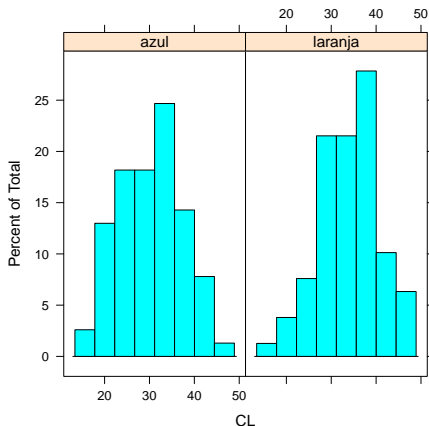
32.004



# Testes de hipótese

## Teste-t para duas amostras

```
require(lattice) # pacote para gráficos avançados  
histogram(~CL | especie, data = dados)
```





# Testes de hipótese

## Teste-t para duas amostras

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

```
with(dados, tapply(CL, especie, summary))
```

\$azul

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
14.7	24.6	30.1	29.9	34.5	47.1

\$laranja

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
16.7	29.4	34.5	34.1	39.2	47.6

Existem evidências de que uma espécie é maior do que a outra?



# Testes de hipótese

## Teste-t para duas amostras

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

- Testar a hipótese de que a **diferença** entre a média de CL da espécie azul ( $\mu_A$ ) e a média de CL da espécie laranja ( $\mu_L$ ) é igual a 0 (zero) (com 95% de confiança)
- As hipóteses são

$$H_0 : \mu_A - \mu_L = 0 \quad \Rightarrow \quad \mu_A = \mu_L$$

$$H_1 : \mu_A - \mu_L \neq 0 \quad \Rightarrow \quad \mu_A \neq \mu_L$$



# Testes de hipótese

## Teste-t para duas amostras

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

```
t.test(CL ~ especie, data = dados, mu = 0,  
       alternative = "two.sided", conf.level = 0.95)
```

Welch Two Sample t-test

data: CL by especie

t = -3.7935, df = 152.73, p-value = 0.0002135

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-6.4116 -2.0204

sample estimates:

mean in group azul	mean in group laranja
29.869	34.085



# Testes de hipótese

## Teste-t para duas amostras

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

- Testar a hipótese de que a **diferença** entre a média de CL da espécie azul ( $\mu_A$ ) e a média de CL da espécie laranja ( $\mu_L$ ) é **menor** do que 0 (zero) (com 95% de confiança)
- Em outras palavras: “O CL médio é menor para a espécie azul?”
- As hipóteses são

$$H_0 : \mu_A - \mu_L \leq 0 \quad \Rightarrow \quad \mu_A \leq \mu_L$$

$$H_1 : \mu_A - \mu_L > 0 \quad \Rightarrow \quad \mu_A > \mu_L$$





# Testes de hipótese

## Teste-t para duas amostras

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

```
t.test(CL ~ especie, data = dados, mu = 0,  
       alternative = "greater", conf.level = 0.95)
```

Welch Two Sample t-test

data: CL by especie

t = -3.7935, df = 152.73, p-value = 0.9999

alternative hypothesis: true difference in means is greater than

95 percent confidence interval:

-6.0552      Inf

sample estimates:

mean in group azul mean in group laranja

29.869

34.085

Como você faria para calcular a diferença observada das médias de CL entre as duas espécies?



# Exercícios

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

Com base no objeto dados:

- (1) Faça um histograma de CW
- (2) Com base no histograma, construa uma hipótese para a média de CW
  - (a) Teste a igualdade dessa hipótese
  - (b) Teste uma desigualdade dessa hipóteseEm ambos os casos use um nível de confiança de 90%, e escreva uma frase com a sua conclusão.
- (3) Faça um histograma de CW para cada sexo
- (4) Com base nesses histogramas, construa uma hipótese para a diferença média de CW entre os sexos
  - (a) Teste a igualdade dessa hipótese
  - (b) Teste uma desigualdade dessa hipóteseEm ambos os casos use um nível de confiança de 90%, e escreva uma frase com a sua conclusão.



# Sumário

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

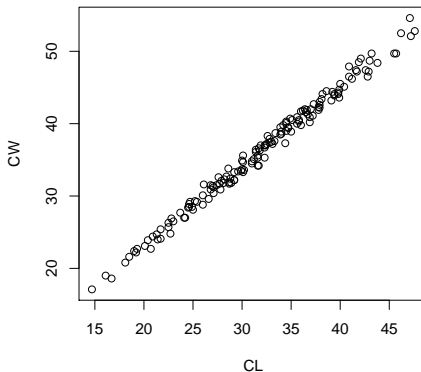
- 1 Distribuições de probabilidade
- 2 Inferência
- 3 Regressão e correlação
  - Regressão
    - Estimação dos parâmetros
  - Correlação
- 4 Análise de Variância
- 5 Modelos Lineares Generalizados



# Regressão e correlação

Vamos analisar a relação que existe entre CL e CW

```
plot(CW ~ CL, data = dados)
```





# Regressão e correlação

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade  
Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

Um **modelo linear** entre duas variáveis  $X$  e  $Y$ , é definido matematicamente como uma equação com dois parâmetros desconhecidos,

$$Y = \beta_0 + \beta_1 X$$

A **análise de regressão** é a técnica estatística que analisa as relações existentes entre uma única variável **dependente**, e uma ou mais variáveis **independentes**

O objetivo é estudar as relações entre as variáveis, a partir de um **modelo matemático**, permitindo **estimar** o valor de uma variável a partir da outra

- Exemplo: sabendo a altura podemos determinar o peso de uma pessoa, se conhecemos os parâmetros do modelo anterior



# Regressão linear

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade  
Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

O problema da análise de regressão consiste em definir a **forma** de relação existente entre as variáveis.

Por exemplo, podemos ter as seguintes relações

$$Y = \beta_0 + \beta_1 X \quad \text{linear}$$

$$Y = \beta_0 X^{\beta_1} \quad \text{potência}$$

$$Y = \beta_0 e^{\beta_1 X} \quad \text{exponencial}$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 \quad \text{polinomial}$$

Em todos os casos, a variável **dependente** é  $Y$ , aquela que será **predita** a partir da relação e da variável **independente**  $X$



# Sumário

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

- 1 Distribuições de probabilidade
- 2 Inferência
- 3 Regressão e correlação
  - Regressão
    - Estimação dos parâmetros
  - Correlação
- 4 Análise de Variância
- 5 Modelos Lineares Generalizados



# Regressão linear

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

Em uma **análise de regressão linear** consideraremos apenas as variáveis que possuem uma **relação linear** entre si.

Uma análise de regressão linear **múltipla** pode associar  $k$  variáveis independentes ( $X$ ) para “explicar” uma única variável dependente ( $Y$ ),

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + e$$

Uma análise de regressão linear **simples** associa uma única variável independente ( $X$ ) com uma variável dependente ( $Y$ ),

$$Y = \beta_0 + \beta_1 X + e$$





# Regressão linear

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

Assim, dados  $n$  pares de valores,  $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ , se for admitido que  $Y$  é função linear de  $X$ , pode-se estabelecer uma regressão linear simples, cujo modelo estatístico é

$$Y_i = \beta_0 + \beta_1 X_i + e_i, \quad i = 1, 2, \dots, n$$

onde:

- $Y$  é a variável **resposta** (ou **dependente**)
- $X$  é a variável **explicativa** (ou **independente**)
- $\beta_0$  é o **intercepto** da reta (valor de  $Y$  quando  $X = 0$ )
- $\beta_1$  é o **coeficiente angular** da reta (**efeito** de  $X$  sobre  $Y$ )
- $e_i \sim N(0, \sigma^2)$  é o **erro**, ou **desvio**, ou **resíduo**

O problema agora consiste em **estimar** os parâmetros  $\beta_0$  e  $\beta_1$ .



# Regressão linear

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade  
Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

## Interpretação dos parâmetros:

$\beta_0$  representa o ponto onde a reta corta o eixo  $Y$  (na maioria das vezes não possui interpretação prática)

$\beta_1$  representa a variabilidade em  $Y$  causada pelo aumento de uma unidade em  $X$ . Além disso,

- $\beta_1 > 0$  mostra que com o aumento de  $X$ , também há um aumento em  $Y$
- $\beta_1 = 0$  mostra que **não há efeito** de  $X$  sobre  $Y$
- $\beta_1 < 0$  mostra que com a aumento de  $X$ , há uma diminuição em  $Y$



# Estimação dos parâmetros

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

Como através de uma amostra obtemos uma estimativa da verdadeira equação de regressão, denominamos

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

ou seja,  $\hat{Y}_i$  é o valor **estimado** de  $Y_i$ , através das **estimativas** de  $\beta_0$  e  $\beta_1$ , que chamaremos de  $\hat{\beta}_0$  e  $\hat{\beta}_1$ .

Para cada valor de  $Y_i$ , temos um valor  $\hat{Y}_i$  estimado pela equação de regressão,

$$Y_i = \hat{Y}_i + e_i$$



# Estimação dos parâmetros

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

Portanto, o erro (ou desvio) de cada observação em relação ao modelo adotado será

$$e_i = Y_i - \hat{Y}_i$$

$$e_i = Y_i - (\beta_0 + \beta_1 X_i)$$

Devemos então adotar um modelo cujos parâmetros  $\beta_0$  e  $\beta_1$ , tornem esse diferença a menor possível.

Isso equivale a **minimizar a soma de quadrados dos resíduos (SQR)**, ou do erro,

$$SQR = \sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 X_i)]^2$$



# Estimação dos parâmetros

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade  
Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

O método de minimizar a soma de quadrados dos resíduos é denominado de **método dos mínimos quadrados**.

Para se encontrar o ponto mínimo de uma função, temos que obter as derivadas parciais em relação a cada parâmetro,

$$\frac{\partial SQR}{\partial \beta_0} = 2 \sum_{i=1}^n [Y_i - \beta_0 - \beta_1 X_i](-1)$$

$$\frac{\partial SQR}{\partial \beta_1} = 2 \sum_{i=1}^n [Y_i - \beta_0 - \beta_1 X_i](-X_i)$$

e igualar os resultados a zero

$$\hat{\beta}_0 = \frac{\partial SQR}{\partial \beta_0} = 0 \quad \text{e} \quad \hat{\beta}_1 = \frac{\partial SQR}{\partial \beta_1} = 0$$



# Estimação dos parâmetros

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão

Estimação

Correlação

ANOVA

MLGs

Dessa forma, chegamos às **estimativas de mínimos quadrados** para os parâmetros  $\beta_0$  e  $\beta_1$ :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i - \frac{\sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n}}{\sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n}}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

onde

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i \quad \text{e} \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$



# Regressão

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação

Correlação

ANOVA

MLGs

## Ajustando um modelo linear no R

```
mod <- lm(CW ~ CL, data = dados)
mod
```

Call:

```
lm(formula = CW ~ CL, data = dados)
```

Coefficients:

(Intercept)	CL
1.19	1.10



# Regressão

## Sumário

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação

Correlação

ANOVA

MLGs

**summary**(mod)

Call:

```
lm(formula = CW ~ CL, data = dados)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.776	-0.570	0.110	0.463	1.827

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	1.1869	0.2853	4.16	5.3e-05	***
CL	1.0975	0.0087	126.17	< 2e-16	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.783 on 154 degrees of freedom

Multiple R-squared: 0.99, Adjusted R-squared: 0.99

F-statistic: 1.59e+04 on 1 and 154 DF, p-value: <2e-16





# Regressão

## Ajuste gráfico

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

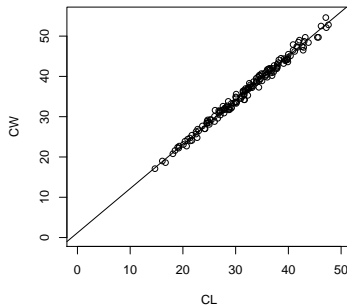
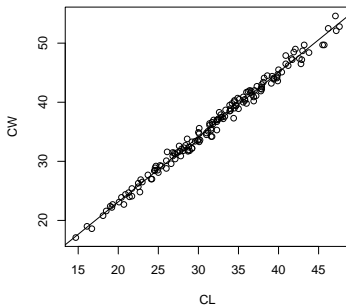
Regressão  
Estimação

Correlação

ANOVA

MLGs

```
plot(CW ~ CL, data = dados)  
abline(mod)  
plot(CW ~ CL, data = dados, xlim = c(0,50), ylim = c(0,55))  
abline(mod)
```





# Regressão

## Análise dos resíduos

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

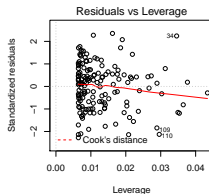
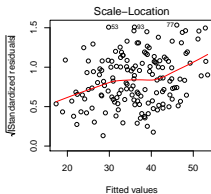
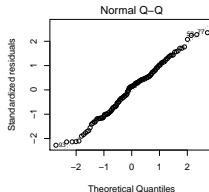
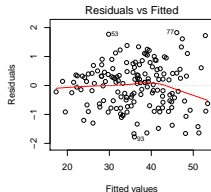
Regressão  
Estimação

Correlação

ANOVA

MLGs

```
par(mfrow = c(2,2))  
plot(mod)  
par(mfrow = c(1,1))
```





# Regressão

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação

Correlação

ANOVA

MLGs

Acessando os componentes do objeto mod:

**names**(mod)

```
[1] "coefficients" "residuals" "effects"  
[4] "rank"         "fitted.values" "assign"  
[7] "qr"          "df.residual" "xlevels"  
[10] "call"         "terms"       "model"
```

**names**(summary(mod))

```
[1] "call"         "terms"       "residuals"  
[4] "coefficients" "aliased"     "sigma"  
[7] "df"          "r.squared"   "adj.r.squared"  
[10] "fstatistic"   "cov.unscaled"
```



# Sumário

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

- 1 Distribuições de probabilidade
- 2 Inferência
- 3 Regressão e correlação
  - Regressão
    - Estimação dos parâmetros
  - Correlação
- 4 Análise de Variância
- 5 Modelos Lineares Generalizados



# Correlação

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação

Correlação

ANOVA

MLGs

Até agora o interesse estava em estudar qual a influência de uma V.A.  $X$  sobre uma V.A.  $Y$ , por meio de uma **relação linear**.

Assim, em uma análise de regressão é indispensável identificar qual variável é dependente.

Na **análise de correlação** isto não é necessário, pois queremos estudar o **grau de relacionamento** entre as variáveis  $X$  e  $Y$ , ou seja, uma medida de **covariabilidade** entre elas.

A correlação é considerada como uma medida de **influência mútua** entre variáveis, por isso não é necessário especificar quem influencia e quem é influenciado.



# Correlação

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

O **grau de relação** entre duas variáveis pode ser medido através do **coeficiente de correlação linear** ( $r$ ), dado por

$$r = \frac{\sum_{i=1}^n X_i Y_i - \frac{\sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n}}{\sqrt{\sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n}} \cdot \sqrt{\sum_{i=1}^n Y_i^2 - \frac{(\sum_{i=1}^n Y_i)^2}{n}}}$$

onde

$$-1 \leq r \leq 1$$

Portanto,

- $r = 1$  correlação **positiva** perfeita entre as variáveis
- $r = 0$  **não há** correlação entre as variáveis
- $r = -1$  correlação **negativa** perfeita entre as variáveis



# Correlação

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

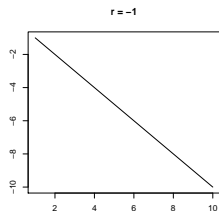
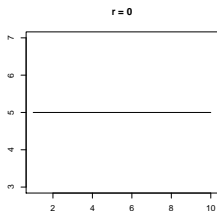
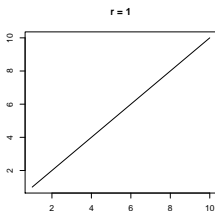
Regressão e  
correlação

Regressão  
Estimação

Correlação

ANOVA

MLGs





# Correlação

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação

Correlação

ANOVA

MLGs

O **coeficiente de determinação** ( $r^2$ ) é o quadrado do coeficiente de correlação, por consequência

$$0 \leq r^2 \leq 1$$

O  $r^2$  nos dá a **porcentagem de variação em  $Y$  que pode ser explicada pela variável independente  $X$ .**

Quanto mais próximo de 1, maior é a explicação da variável  $Y$  pela variável  $X$ .





# Correlação

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

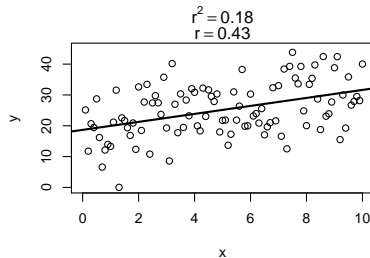
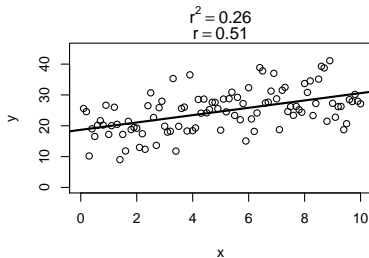
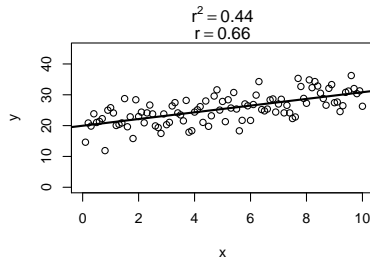
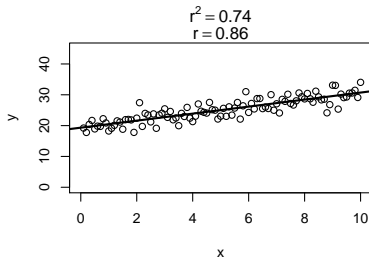
Regressão e  
correlação

Regressão  
Estimação

Correlação

ANOVA

MLGs





# Exercícios

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação

Correlação

ANOVA

MLGs

Com as colunas BD e CL do objeto dados

- (1) Faça um gráfico da relação entre estas variáveis
- (2) Faça um teste de correlação
- (3) Ajuste um modelo linear
  - (a) Veja o sumário
  - (b) Ajuste a linha do modelo no gráfico
  - (c) Verifique os resíduos

Qual sua conclusão?

- Existe correlação significativa? De que tipo (positiva, negativa)?
- O modelo linear descreve bem a relação entre estas duas variáveis (verifique com o valor de  $Pr(>|t|)$  e do  $R^2$ )
- O modelos foi bem ajustado aos dados (observe os resíduos)



# Sumário

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

- 1 Distribuições de probabilidade
- 2 Inferência
- 3 Regressão e correlação
  - Regressão
    - Estimação dos parâmetros
  - Correlação
- 4 **Análise de Variância**
- 5 Modelos Lineares Generalizados



# Análise de Variância

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade  
Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

Definição:  $y_{ij}$  representa a observação  $j$  do grupo  $i$ ;  $\bar{y}_i$  é a média do grupo  $i$ ;  $\bar{y}$  é a média geral de todas as observações. As observações podem ser decompostas em

$$y_{ij} = \bar{y} + (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$$

que corresponde ao modelo

$$y_{ij} = \theta + \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2)$$

A hipótese a ser testada de que todos os grupos são iguais (*i.e* médias iguais) implica que todos os  $\mu_i$  são iguais:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_n$$

$$H_1 : \text{pelo menos um } \mu_i \text{ é diferente dos demais}$$



# Análise de Variância

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

Voltando ao exemplo da diferença de CL entre as duas espécies:  
 $\bar{y}_A = 29.9$  e  $\bar{y}_L = 34.1$

```
with(dados, tapply(CL, especie, summary))
```

\$azul

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
14.7	24.6	30.1	29.9	34.5	47.1

\$laranja

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
16.7	29.4	34.5	34.1	39.2	47.6

Média geral  $\bar{y} = 32$

```
mean(dados$CL)
```

```
[1] 32.004
```



# Análise de Variância

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

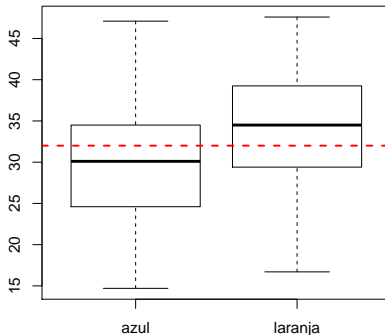
Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

```
boxplot(CL ~ especie, data = dados)  
abline(h = mean(dados$CL), lty = 2, col = "red", lwd = 2)
```





# Análise de Variância

## Geometricamente

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

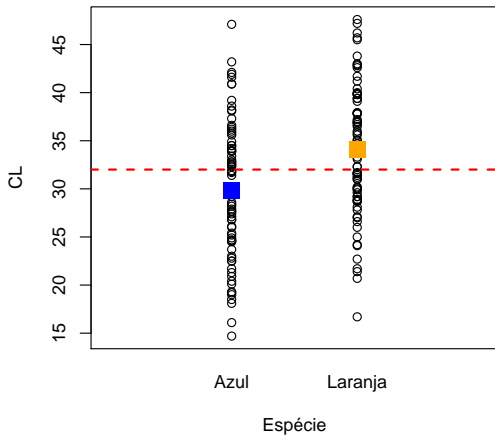
Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs





# Análise de Variância

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

Podemos ajustar um modelo linear entre CL e espécie

```
mod <- lm(CL ~ especie, data = dados)
summary(mod)
```

Call:

```
lm(formula = CL ~ especie, data = dados)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-17.385	-5.019	0.273	5.019	17.231

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	29.87	0.79	37.8	< 2e-16 ***
especieleranja	4.22	1.11	3.8	0.00021 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.93 on 154 degrees of freedom

Multiple R-squared: 0.0856, Adjusted R-squared: 0.0797

F-statistic: 14.4 on 1 and 154 DF, p-value: 0.00021





# Análise de Variância

## Ajustando o modelo

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

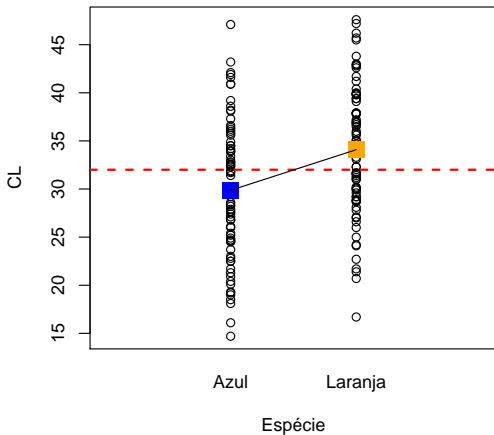
Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs





# Análise de Variância

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

Você lembra do teste-t feito anteriormente?

```
teste <- t.test(CL ~ especie, data = dados, mu = 0,  
                 alternative = "two.sided", conf.level = 0.95)
```

teste

Welch Two Sample t-test

data: CL by especie

t = -3.7935, df = 152.73, p-value = 0.0002135

alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:

-6.4116 -2.0204

sample estimates:

mean in group azul	mean in group laranja
29.869	34.085



# Análise de Variância

Notou a relação?

```
summary(mod)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	29.869	0.7902	37.7990	8.1924e-80
especiellaranja	4.216	1.1104	3.7968	2.1042e-04

```
teste$p.value
```

```
[1] 2.1352e-04
```

```
teste$estimate
```

mean in group azul	mean in group laranja
29.869	34.085

```
diff(teste$estimate)
```

```
mean in group laranja  
4.216
```



# Análise de Variância

A ANOVA vai testar apenas a hipótese inicial

$$H_0 : \mu_A = \mu_L$$

$$H_1 : \mu_A \neq \mu_L$$

```
anova(mod)
```

```
Analysis of Variance Table
```

```
Response: CL
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
especie	1	693	693	14.4	0.00021 ***
Residuals	154	7404	48		

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Aqui a única conclusão é de que os  $\mu_i$  não são iguais (mas você não sabe quanto e nem quais!)



# Análise de Variância

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

Se olharmos apenas o resultado da ANOVA, podemos prosseguir com a análise fazendo um teste *a posteriori* para verificarmos quais são os grupos que diferem entre si. Um deles é o teste de Tukey

```
mod.anova <- aov(CL ~ especie, data = dados)
TukeyHSD(mod.anova)
```

```
Tukey multiple comparisons of means
 95% family-wise confidence level
```

```
Fit: aov(formula = CL ~ especie, data = dados)
```

```
$especie
```

	diff	lwr	upr	p adj
laranja-azul	4.216	2.0224	6.4096	0.00021



# Análise de Variância

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

Porque então fazer uma ANOVA???

- Quando formos comparar a média de mais de 2 grupos
- Não é possível fazer um teste-t para mais de 2 grupos
- Por exemplo, com 3 grupos (A, B, C) teríamos que fazer 3 comparações (A:B, A:C, B:C)
  - Com um nível de confiança de 95% ( $\alpha = 0.05$ ) para cada teste, os 3 testes teriam um nível de confiança  $(1 - \alpha)^3$
  - Portanto  $(1 - 0.05)^3 = (0.95)^3 = 0.85$
  - Isso implica que quanto mais comparações forem feitas, menor será seu nível de confiança no resultado dos testes.



# Sumário

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

- 1 Distribuições de probabilidade
- 2 Inferência
- 3 Regressão e correlação
  - Regressão
    - Estimação dos parâmetros
  - Correlação
- 4 Análise de Variância
- 5 Modelos Lineares Generalizados



# Modelos Lineares Generalizados

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

Nelder e Wedderburn (1972) mostraram que uma série de técnicas estatísticas podem ser formuladas de forma unificada, como uma classe de modelos de regressão. A essa teoria, uma extensão dos modelos clássicos de regressão, deram o nome de **Modelos Lineares Generalizados**.

$\text{Teste-t} \subset \text{ANOVA} \subset \text{ANCOVA}^* \subset \text{ML} \subset \text{ML-MULT}^* \subset \text{MLG}$

- Teste-t: compara uma ou duas médias
- ANOVA: compara 2 ou mais médias (fator)
- ANCOVA: compara 2 ou mais médias (fator) + variáveis numéricas
- ML: regressão de  $y$  (numérico) em função de um único  $x$  (numérico ou fator)
- ML-MULT: regressão de  $y$  (numérico) em função de mais de um  $x$  (numéricos ou fatores)
- MLG: Similar ao ML-MULT, mas estende o modelo para que  $y$  possa ser um fator ou ter uma distribuição diferente da normal.





# Modelos Lineares Generalizados

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

Os MLGs são formados por três componentes:

**Componente aleatório:** a variável resposta do modelo, com distribuição pertencente à família de distribuições exponencial.

**Componente sistemático:** as variáveis explicativas, que entram na forma de uma estrutura linear.

**Função de ligação:** função que liga os componentes aleatório e sistemático.



# Modelos Lineares Generalizados

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

De maneira geral, os MLGs descrevem a relação entre a variável resposta  $y_i$  ( $i = 1, \dots, n$ ) através de preditores  $x_i$ . A média de  $y_i$  condicionada aos preditores  $x_i$  é

$$E(y_i|x_i) = \mu_i$$

e existe uma transformação de  $\mu_i$  de forma que

$$g(\mu_i) = x_i^T \beta$$

onde  $g(\cdot)$  é uma função de ligação conhecida, e  $\beta$  é o vetor de parâmetros a ser estimado.



# Modelos Lineares Generalizados

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

Distribuições da família exponencial e funções de ligação ( $P = \text{link padrão}$ )

Link	binomial	poisson	negative binomial	Gamma	gaussian	inverse gaussian
logit	P					
probit	•					
cloglog	•					
identity		•	•	•	P	
inverse				P		
log		P	P	•		
$1/\mu^2$						P
sqrt		•	•			



# Modelos Lineares Generalizados

Para ajustar um MLG usamos a função `glm()`

```
mod.glm <- glm(CL ~ especie, data = dados,  
               family = gaussian(link = "identity"))  
summary(mod.glm)
```

Call:

```
glm(formula = CL ~ especie, family = gaussian(link = "identity"),  
    data = dados)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-17.385	-5.019	0.273	5.019	17.231

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	29.87	0.79	37.8	< 2e-16 ***
especiellaranja	4.22	1.11	3.8	0.00021 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 48.08)



# Modelos Lineares Generalizados

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

Quando existe mais de uma variável resposta ( $y$ )? **Métodos multivariados!**



# Exercícios

Módulo III  
Inferência e  
Modelagem

CBO 2014

Distribuições  
de  
probabilidade

Inferência

Regressão e  
correlação

Regressão  
Estimação  
Correlação

ANOVA

MLGs

Com o objeto dados

- (1) Faça um boxplot de CW por sexo
  - (2) Faça um teste-t para testar se existe diferença entre as médias de CW para machos e fêmeas
  - (3) Ajuste um modelo linear para testar essa mesma hipótese
  - (4) Faça uma ANOVA e o teste de Tukey
- Qual sua conclusão?