Modelos Lineares Generalizados (MLGs)

Fernando de Pol Mayer

Laboratório de Estatística e Geoinformação (LEG)
Departamento de Estatística (DEST)
Universidade Federal do Paraná (UFPR)

Sumário

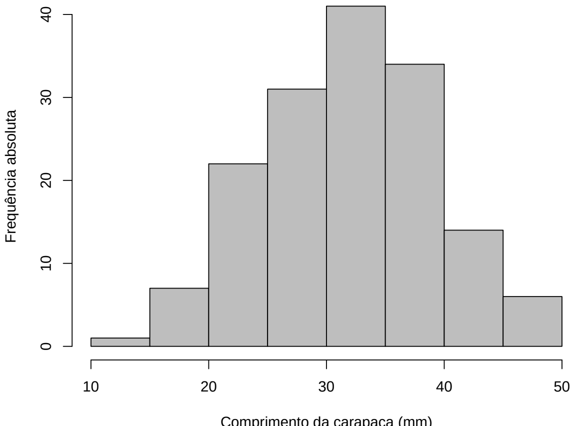
Plano de aula

Base de dados

```
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 31.5 ...
 $ BD     : num  7 7.4 7.7 8.2 9.8 10.4 10.3 11.4 11.4 11 ...
```

Plano de aula

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



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** α (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) Calcula a **estatística de teste**, sob a hipótese nula

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

(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}

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

$$H_0 : \mu = 30$$
$$H_1 : \mu \neq 30$$

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] 3.462731

# t critico (não é apresentado no resultado da função do R)
qt(0.025, df = n - 1, lower.tail = FALSE)

[1] 1.975387

# valor p (multiplicado por 2 pois o teste é bilateral)
pt(tcalc, df = n - 1, lower.tail = FALSE) * 2

[1] 0.000691346
```

```r
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:
 30.86071 33.14698
sample estimates:
mean of x
 32.00385
```

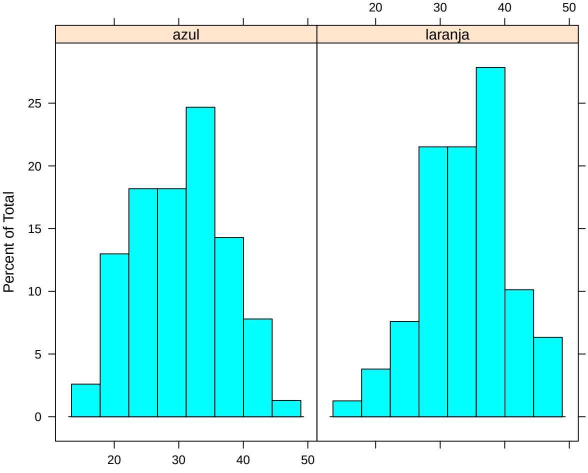
Detalhe: O teste pode ser armazenado em um objeto para futuras referências

```
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
3.462731

teste$p.value

[1] 0.000691346
```

Testes de hipótese

Teste-t para duas amostras

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



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

$azul
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  14.70   24.60   30.10   29.87   34.50   47.10

$laranja
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.70   29.40   34.50   34.08   39.25   47.60
```

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

Testes de hipótese

Teste-t para duas amostras

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

$$
\begin{aligned}
H_0 &: \mu_A - \mu_L = 0 \quad \Rightarrow \quad \mu_A = \mu_L \\
H_1 &: \mu_A - \mu_L \neq 0 \quad \Rightarrow \quad \mu_A \neq \mu_L
\end{aligned}
$$

Testes de hipótese

Teste-t para duas amostras

```
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:
 -6.411592 -2.020366
sample estimates:
   mean in group azul mean in group laranja
             29.86883              34.08481
```

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

Exercícios

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 hipótese

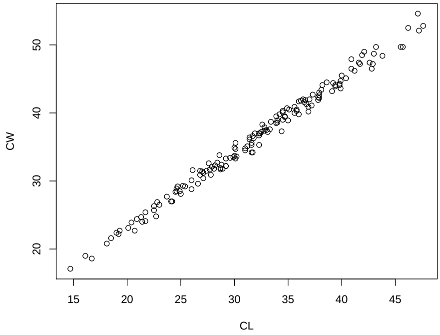
    Em ambos os casos use um nível de confiança de 90%, e escreva uma frase com a sua conclusão.
(3) Faça um histograma de CW para cada sexo
(4) Com base nesses histogramas, construa uma hipótese para a diferença média de CW entre os sexos
    (a) Teste a igualdade dessa hipótese
    (b) Teste uma desigualdade dessa hipótese

    Em ambos os casos use um nível de confiança de 90%, e escreva uma frase com a sua conclusão.

Plano de aula

Vamos analisar a relação que existe entre CL e CW

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



Regressão e correlação

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

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

Por exemplo, podemos ter as seguintes relações

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X && \text{linear} \\
Y &= \beta_0 X^{\beta_1} && \text{potência} \\
Y &= \beta_0 e^{\beta_1 X} && \text{exponencial} \\
Y &= \beta_0 + \beta_1 X + \beta_2 X^2 && \text{polinomial}
\end{aligned}
$$

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

Plano de aula

Regressão linear

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

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

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

onde:

- Y é a variável **resposta** (ou **dependente**)
- X é a variável **explicativa** (ou **independente**)
- β_0 é o **intercepto** da reta (valor de Y quando $X = 0$)
- β_1 é o **coeficiente angular** da reta (**efeito** de X sobre Y)
- $e_i \sim \mathrm{N}(0, \sigma^2)$ é o **erro**, ou **desvio**, ou **resíduo**

O problema agora consiste em **estimar** os parâmetros β_0 e β_1.

Regressão linear

Interpretação dos parâmetros:

β_0 representa o ponto onde a reta corta o eixo Y (na maioria das vezes não possui interpretação prática)

β_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

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 β_0 e β_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

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

$$
\begin{aligned}
e_i &= Y_i - \hat{Y}_i \\
e_i &= Y_i - (\beta_0 + \beta_1 X_i)
\end{aligned}
$$

Devemos então adotar um modelo cujos parâmetros β_0 e β_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

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

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

Dessa forma, chegamos às **estimativas de mínimos quadrados** para os parâmetros β_0 e β_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

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

Regressão

Ajustando um modelo linear no R

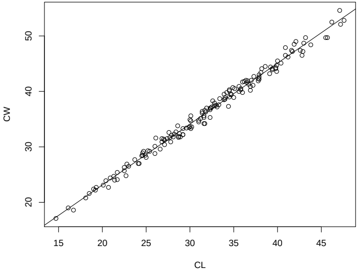
```r
mod <- lm(CW ~ CL, data = dados)
mod


Call:
lm(formula = CW ~ CL, data = dados)

Coefficients:
(Intercept)           CL
      1.187        1.097
```

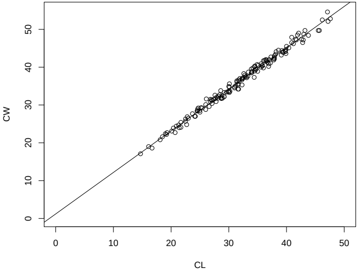
```
summary(mod)


Call:
lm(formula = CW ~ CL, data = dados)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7762 -0.5699  0.1098  0.4629  1.8273

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.186950   0.285340    4.16 5.28e-05 ***
CL          1.097451   0.008698  126.17  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7827 on 154 degrees of freedom
Multiple R-squared:  0.9904,Adjusted R-squared:  0.9904
F-statistic: 1.592e+04 on 1 and 154 DF,  p-value: < 2.2e-16
```

```
anova(mod)

Analysis of Variance Table

Response: CW
           Df Sum Sq Mean Sq F value    Pr(>F)
CL          1 9752.6  9752.6   15919 < 2.2e-16 ***
Residuals 154   94.3     0.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regressão
Ajuste gráfico

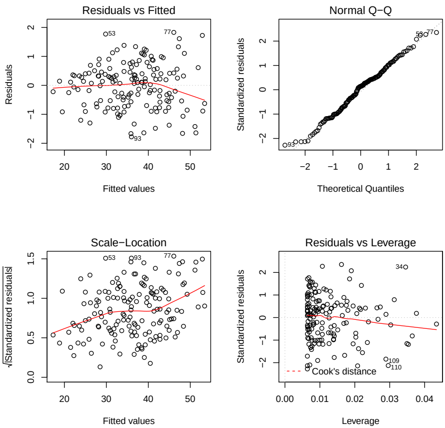
```r
plot(CW ~ CL, data = dados)
abline(mod)
plot(CW ~ CL, data = dados, xlim = c(0,50), ylim = c(0,55))
abline(mod)
```





Regressão
Análise dos resíduos

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



Regressão

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"

names(anova(mod))

[1] "Df"      "Sum Sq"  "Mean Sq" "F value" "Pr(>F)"
```

Regressão

Veja que o `Residual standard error: 0.7827` é o estimador do desvio-padrão residual $\hat{\sigma}_e^2 = \frac{\text{SQRes}}{n-2}$, ou seja,

```
sqrt(anova(mod)$Sum[2]/anova(mod)$Df[2])

[1] 0.7827079
```

e que `F-statistic: 1.592e+04` (15920) é o mesmo valor de

```
anova(mod)$F[1]

[1] 15919.11
```

que testa a mesma hipótese da ANOVA. De fato, o valor de t^2 para β_1 no sumário do modelo é

```
summary(mod)$coef[2,3]^2

[1] 15919.11
```

Plano de aula

Correlação

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

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

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

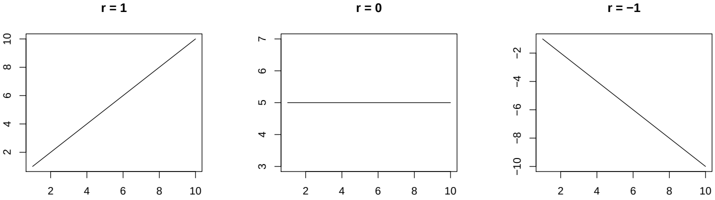
onde

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

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



Correlação

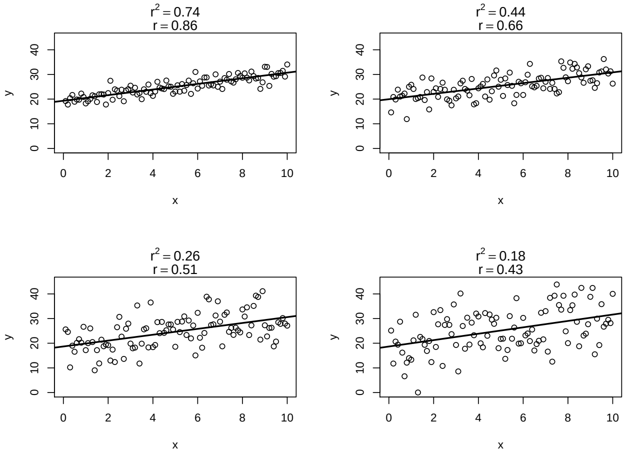
O **coeficiente de determinação** (r^2) é o quadrado do coeficiente de correlação, por consequência

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

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


$r^2 = 0.74$
$r = 0.86$
$r^2 = 0.44$
$r = 0.66$
$r^2 = 0.26$
$r = 0.51$
$r^2 = 0.18$
$r = 0.43$
y
x
0
10
20
30
40
0
2
4
6
8
10

Exercícios

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)

Plano de aula

Análise de Variância

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

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

que corresponde ao modelo

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

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

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_n$$
$$H_1 : \text{pelo menos um } \mu_i \text{ é diferente dos demais}$$

Análise de Variância

Voltando ao exemplo da diferença de CL entre as duas espécies:
$\bar{y}_A = 29.87$ e $\bar{y}_L = 34.08$

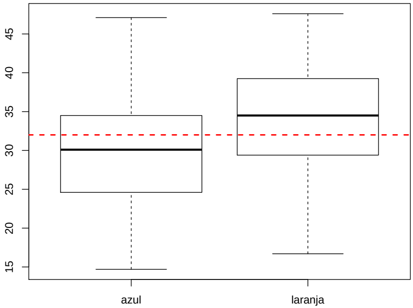
```
with(dados, tapply(CL, especie, summary))

$azul
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  14.70   24.60   30.10   29.87   34.50   47.10

$laranja
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.70   29.40   34.50   34.08   39.25   47.60
```

Média geral $\bar{y} = 32$

```
mean(dados$CL)

[1] 32.00385
```

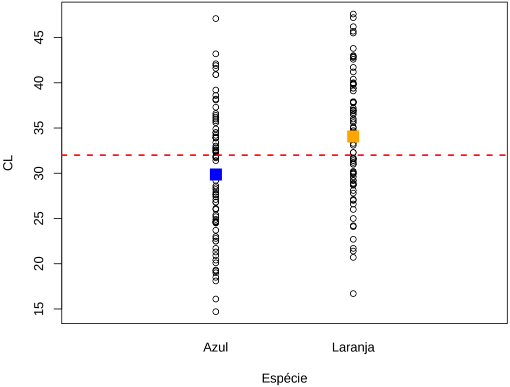
Análise de Variância

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



Análise de Variância

Geometricamente



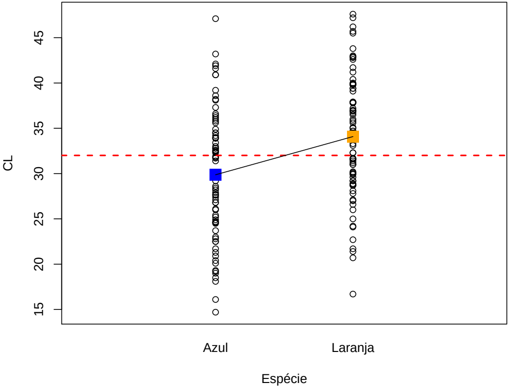
Análise de Variância

Podemos ajustar um modelo linear entre CL e espécie

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

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

Residuals:
     Min       1Q   Median       3Q      Max
-17.3848  -5.0188   0.2732   5.0192  17.2312

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     29.8688     0.7902  37.799  < 2e-16 ***
especielaranja   4.2160     1.1104   3.797  0.00021 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.934 on 154 degrees of freedom
Multiple R-squared:  0.08559,Adjusted R-squared:  0.07966
F-statistic: 14.42 on 1 and 154 DF,  p-value: 0.0002104
```

Ajustando o modelo



Análise de Variância

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:
 -6.411592 -2.020366
sample estimates:
   mean in group azul mean in group laranja
             29.86883              34.08481
```

Notou a relação?

```
summary(mod)$coefficients

                 Estimate Std. Error  t value     Pr(>|t|)
(Intercept)     29.868831  0.7902012 37.79902 8.192364e-80
especielaranja   4.215979  1.1104178  3.79675 2.104221e-04

teste$p.value

[1] 2.135202e-04

teste$estimate

   mean in group azul mean in group laranja
             29.86883              34.08481

unname(diff(teste$estimate))

[1] 4.215979
```

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  693.1  693.09  14.415 0.0002104 ***
Residuals 154 7404.3   48.08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

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 4.215979 2.022362 6.409596 0.0002104
```

Análise de Variância

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.

Plano de aula

Modelos Lineares Generalizados

Todos os modelos anteriores podem ser classificados como um caso particular de uma **família de modelos** mais geral, denominada **Modelos Lineares Generalizados** (MLGs):

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

- 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 extende o modelo para que Y possa ser um fator ou ter uma distribuição diferente da normal.

A seleção de modelos é uma parte importante de toda pesquisa, envolve a procura de um modelo o mais simples possível, que descreva bem os dados observados.

Na maior parte das situações pode-se pensar na variável resposta (Y) consistindo de duas partes distintas:

1. Um **componente sistemático**, que é estabelecido durante o planejamento do experimento, resultando em modelos de regressão, ANOVA ou ANCOVA.
2. Um **componente aleatório**, que é estabelecido assim que são definidas as medidas a serem feitas, que podem ser contínuas ou discretas, exigindo o ajuste de distribuições diferentes.

Matematicamente, e assumindo o modelo clássico de regressão, temos:

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{e}$$

onde:

- Y o vetor de dimensão $n \times 1$ da variável **resposta**
- $\boldsymbol{\mu} = \mathrm{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ o **componente sistemático**
- $\mathbf{X}$ é a **matriz do modelo**, de dimensão $n \times p$
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ o vetor de parâmetros
- $\mathbf{e} = (e_1, \ldots, e_n)^T$ o **componente aleatório** com $e_i \sim \mathrm{N}(0, \sigma^2)$

Em muitos casos, porém, essa estrutura aditiva entre o componente sistemático e o componente aleatório não é satisfeita.

Além disso:

- Não há razão para se restringir à estrutura simples dada por $\boldsymbol{\mu} = \mathrm{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ para o componente sistemático
- Nem sempre a distribuição normal é adequada para o componente aleatório
- Nem sempre a suposição de homogeneidade de variâncias é atendida (e em muitos casos não deve ser mesmo)

Modelos Lineares Generalizados

Nelder e Wedderburn (1972) propuseram uma teoria unificadora da modelagem estatística, a que deram o nome de **Modelos Lineares Generalizados (MLG)**, como uma extensão dos modelos lineares clássicos.

Na realidade, eles mostraram que uma série de técnicas comumente estudadas separadamente podem ser reunidas sob o nome de Modelos Lineares Generalizados.

Os desenvolvimentos que levaram a esta visão geral da modelagem estatística, remontam a mais de um século.

Eles mostraram, então, que a maioria dos problemas estatísticos, que surgem nas áreas de oceanografia, agricultura, ecologia, economia, etc. podem ser formulados, de uma **maneira unificada**, como **modelos de regressão**.

Os MLGs possuem uma estrutura similar à dos modelos lineares clássicos, e podem ser usados quando se tem uma única variável aleatória Y, e associado a ela um conjunto de variáveis explicativas $X_1, \ldots, X_p$

Para uma amostra de n observações $(y_i, \mathbf{x}_i)$ em que $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ é o vetor coluna de variáveis explicativas, o modelo linear generalizado envolve os três componentes:

1) **Componente aleatório**: variável resposta do modelo, representado por um conjunto de variáveis aleatórias independentes $Y_1, \ldots, Y_n$ provenientes de uma **mesma distribuição** que faz parte da **família exponencial** com médias $\mu_1, \ldots, \mu_n$, ou seja,

$$\mathrm{E}(Y_i) = \mu_i, \qquad i = 1, \ldots, n$$

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

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$
$$= \sum_{j=1}^{p} \beta_j x_{ij}$$

Ou, em forma matricial

$$\boldsymbol{\eta} = \mathbf{x}_i^T \boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}$$

sendo $X = (x_1, \ldots, x_n)^T$ a matriz do modelo, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ o vetor de parâmetros, e $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$ o preditor linear.

3) **Função de ligação**: função que liga os componentes aleatório e sistemático. O modelo liga μ_i a η_i através de

$$\eta_i = g(\mu_i)$$

onde $g(\cdot)$ é uma função monótona e diferenciável. Portanto, $g(\cdot)$ liga $\mathrm{E}(Y_i)$ com as variáveis explicativas através de

$$g(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij} \qquad i = 1, \ldots, n$$

Família exponencial

A **família exponencial** é uma forma geral de definição de algumas distribuições de probabilidade. A função (densidade) de probabilidade desta família é:

$$f(y_i; \theta_i) = a(\theta_i) b(y_i) \exp\left[y_i Q(\theta_i)\right]$$

Diversas distribuições importantes como: normal, binomial e Poisson fazem parte desta família (*i.e.* são casos particulares).

O termo $Q(\theta)$ é chamado de **parâmetro natural**.

Se a função de ligação for $Q(\theta)$, ou seja,

$$g(\mu_i) = Q(\theta) = \sum_{j=1}^{p} \beta_j x_{ij}$$

ela é chamada de **função de ligação canônica**.

Família exponencial

Exemplo: Binomial/Bernoulli

A distribuição Bernoulli é um caso particular de uma distribuição binomial com $n = 1$, e especifica as probabilidades $P(Y = 1) = \pi$ e $P(Y = 0) = 1 - \pi$, e $\mathrm{E}(Y) = \pi$. Na famíla exponencial:

$$
\begin{aligned}
f(y;\pi) &= \pi^y(1-\pi)^{1-y} \\
&= (1-\pi)\exp\left[y\log\frac{\pi}{1-\pi}\right] \\
&= a(\theta)b(y)\exp\left[yQ(\theta)\right]
\end{aligned}
$$

Portanto, com $\theta = \mu$, $a(\pi) = 1 - \pi$, $b(y) = 1$, $Q(\pi) = \log[\frac{\pi}{1-\pi}]$, a **função de ligação canônica** é chamada *logit*, e

$$
\mathrm{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \sum_{j=1}^{p}\beta_j x_{ij}
$$

é chamada de *regressão logística*.

Família exponencial

Exemplo: Poisson

A distribuição de Poisson é comumente utilizada para modelar dados de contagem. Seja Y uma contagem, e $\mu = \mathrm{E}(Y)$, a função densidade de probabilidade na família exponencial fica:

$$\begin{aligned} f(y;\mu) &= \frac{e^{-\mu}\mu^{y}}{y!} \\ &= \exp(-\mu)\left(\frac{1}{y!}\right)\exp(y\log\mu) \\ &= a(\theta)b(y)\exp[yQ(\theta)] \end{aligned}$$

Portanto, com $\theta = \mu$, $a(\mu) = \exp(-\mu)$, $b(y) = 1/y!$, $Q(\mu) = \log\mu$, a **função de ligação canônica** é o log, e

$$g(\mu) = \log\mu = \sum_{j=1}^{p}\beta_j x_{ij}$$

que é chamado de *modelo loglinear de Poisson*.

A classe de MLGs inclui também modelos para variáveis respostas contínuas.

A distribuição normal faz parte da família exponencial que inclui um **parâmetro de dispersão**, e seu parâmetro natural é a média. Portanto,

$$g(\mu) = \mu$$

e um modelo de regressão linear simples é um MLG com função de ligação **identidade**.

Funções de ligação e tipos de modelo

Componente aleatório	Link		Componente sistemático	Modelo
Normal	Identidade	μ	Contínuo	Regressão linear
Normal	Identidade	μ	Categórico	ANOVA
Normal	Identidade	μ	Ambos	ANCOVA
Binomial	Logit	$\log_e\left(\frac{\mu}{1-\mu}\right)$	Ambos	Regressão logística
Poisson	Log	$\log_e \mu$	Ambos	Loglinear
Multinomial	Logit gen.		Ambos	Multinomial

Funções de ligação no R

Distribuições da família exponencial e funções de ligação (P = link canônico)

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`		•	•			

Um método tradicional de análise de dados consistia em transformar Y, para que a variável resposta ficasse com distribuição normal e variância constante.

Em MLGs, a escolha de uma função de ligação **não** é relacionada com a escolha do componente aleatório.

Se uma função de ligação é capaz de linearizar a relação entre a média e os preditores, então **não é necessário** que ela também estabilize a variância ou produza normalidade.

Isso está relacionado com o processo de ajuste do modelo, que maximiza a verossimilhança para a distribuição de Y, que não é mais restrita à normal.

Modelos Lineares Generalizados

MLGs fornecem uma **teoria unificada de modelagem**, que compreende os modelos mais importantes para variáveis contínuas e discretas.

A estimativa dos parâmetros em MLGs é realizada através de um algoritmo que usa uma versão ponderada dos mínimos quadrados, *Iteratively Reweighted Least Squares* (IRLS).

A razão de restringir os MLGs à família exponencial para Y é porque este mesmo algoritmo se aplica à todos os membros dessa família, para qualquer escolha de função de ligação.

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
-17.3848   -5.0188    0.2732    5.0192   17.2312

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     29.8688     0.7902  37.799  < 2e-16 ***
especielaranja   4.2160     1.1104   3.797  0.00021 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 48.08018)
```

Deviance

Resíduos e diagnósticos

Quando existe mais de uma variável resposta (Y)?

- Métodos multivariados (restritos à normalidade)
- McGLM (*Multivariate covariance Generalized Linear Models*) (Bonat e Jorgensen, 2016)

Exercícios

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?

Plano de aula

Referências

- Agresti, A. **Categorical data analysis**. John Wiley & Sons. 2002.
- Fox, J; Weisberg, S. **An R companion to applied regression**. Sage. 2011.