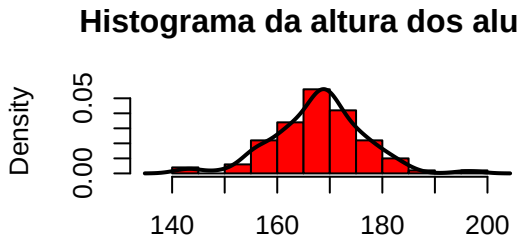
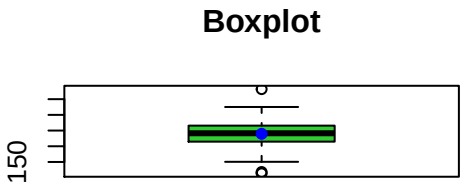
finalExam_2018

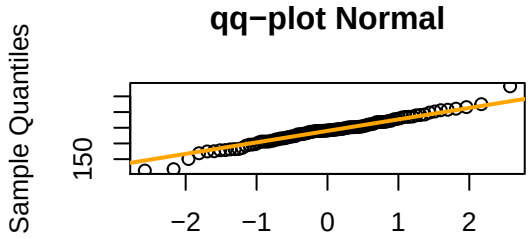
J M Fernandes - Estudante- Lucas

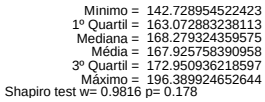
5/1/2018

Pergunta 1: Distribuição Normal









Pergunta 2: Qual é a probabilidade de selecionar um indivíduo ao acaso de uma amostra de 100 indivíduos com

média = 167.9

desvio padrão= 8.5

a) com a altura maior que 186.6

b) com a altura menor que 157.7

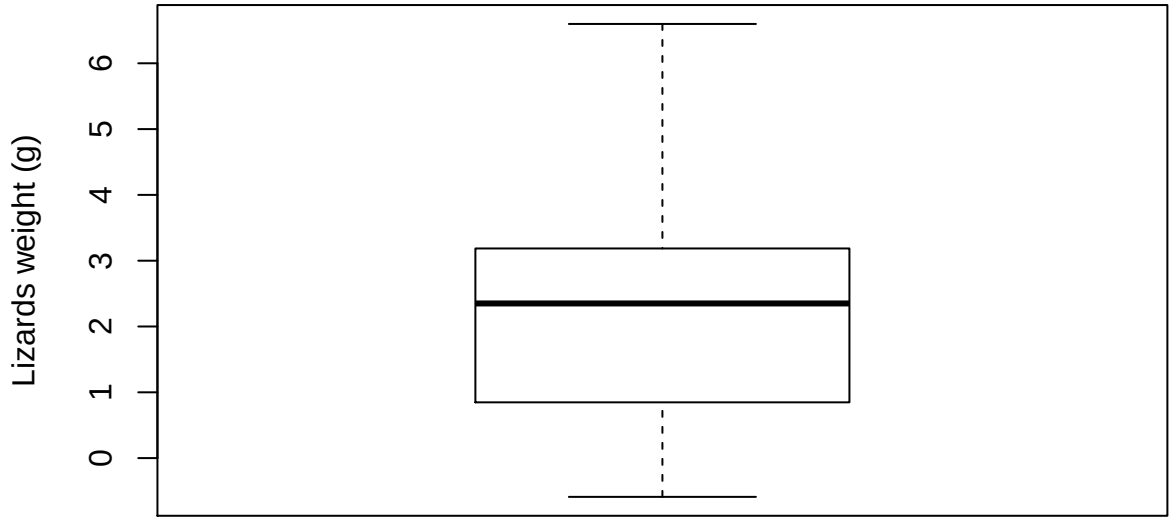
c) com a altura entre 159.4 e 176.4

d) Dada a análise exploratória apresentada acima (gráficos e sumário) poderíamos afirmar que os dados amostrais da altura de 100 alunos pertencem a uma distribuição normal? Justifique a sua resposta.

Pergunta 3: t-test para uma amostra

Exemplo do mundo real: imagine que você tenha o peso de 40 lagartos coletados em sua pesquisa e queira compará-lo com os pesos médios conhecidos disponíveis na literatura científica. Além disso, esperamos que nossos lagartos sejam mais leves que os da literatura, porque nossos dados vêm de uma área com escassez de alimentos. Nossa hipótese nula é que a média não é menor que 2,5

```r
y<-rnorm(40,2.6,1.9)

boxplot(y, ylab = "Lizards weight (g)")
```




```r
summary(y)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.5889  0.8791  2.3504  2.3055  3.1710  6.5978
```

```r
t.test(y, mu = 2.5, alt = "less", conf = 0.95)  # mean = 2.5, alternative hypothesis 1 sided; we get a
```

```
##
##  One Sample t-test
##
## data:  y
## t = -0.66453, df = 39, p-value = 0.2551
## alternative hypothesis: true mean is less than 2.5
## 95 percent confidence interval:
##      -Inf 2.798589
## sample estimates:
## mean of x
##  2.305537
```

Escreva a sua conclusão sobre o peso dos lagartos com base teste estatístico realizado.

Pergunta 4: t-test para comparar médias entre 2 amostras

O tempo de espera em caixas de super mercados da cidade foi medido e um test-t foi aplicado para comparar a média entre as duas amostras. A hipótese foi formulada ao nível de 5% de probabilidade.

```r
## data
n <- 40 + sample(1:12, 2) * 3
Waiting <- rnorm(sum(n), sd = sample(30:40, 1)/10) + rep(sample(30:80, 2)/10, n)
Waiting[Waiting < 0] <- 0
dat <- data.frame(
  Waiting = Waiting,
```

```
  Supermarket = factor(rep(1:2, c(n[1], n[2])), levels = 1:2, labels = c("Sparag", "Consumo"))
)

## questions/answer
questions <- character(5)
solutions <- logical(5)
explanations <- character(5)

t.test(Waiting ~ Supermarket, data = dat, var.equal = TRUE,alternative = "two.sided")
```

```
##
##  Two Sample t-test
##
## data:  Waiting by Supermarket
## t = 4.9884, df = 114, p-value = 2.2e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.662291 3.852182
## sample estimates:
##  mean in group Sparag mean in group Consumo
##              6.757328              4.000091
```

Teste a hipótese

II. $H_0 : \mu_1 - \mu_2 = 0$ $H_A : \mu_1 - \mu_2 \neq 0$

Escreva a sua conclusão sobre o tempo médio de espera com base no teste estatístico realizado.

Pergunta 5- Intervalo de Confiança

Question (6)

The daily expenses of summer tourists in Vienna are analyzed. A survey with 105 tourists is conducted. This shows that the tourists spend on average 121.2 EUR. The sample variance s^2_{n-1} is equal to 140.5.

Determine a 95% confidence interval for the average daily expenses (in EUR) of a tourist.

Answerlist

- What is the lower confidence bound?
- What is the upper confidence bound?

Question (7)

A machine fills milk into 200ml packages. It is suspected that the machine is not working correctly and that the amount of milk filled differs from the setpoint $\mu_0 = 200$. A sample of 149 packages filled by the machine are collected. The sample mean $\bar{y}$ is equal to 194 and the sample variance s^2_{n-1} is equal to 144.44.

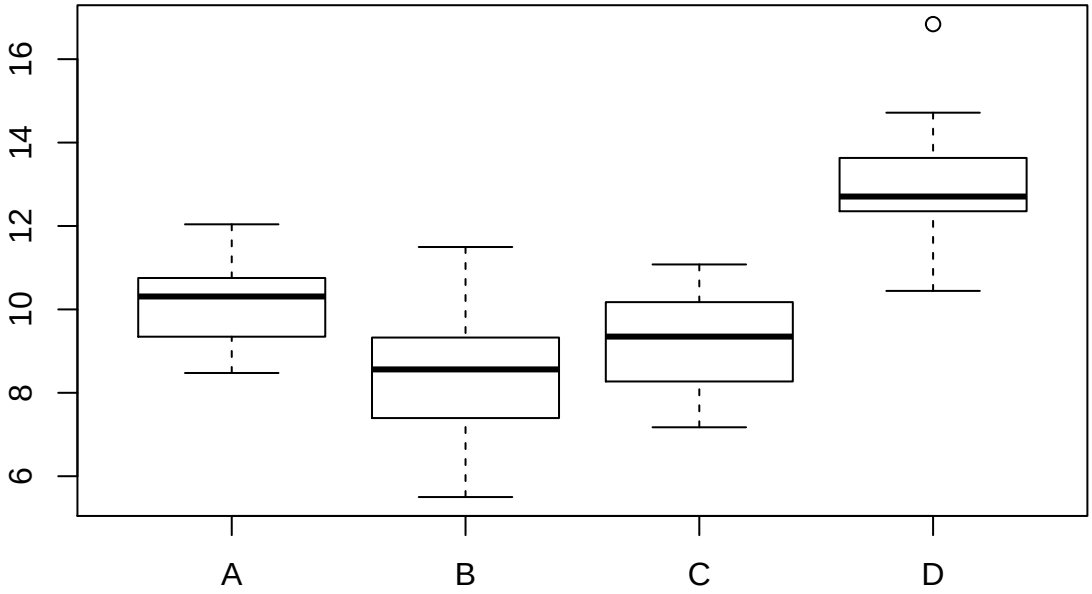
Test the hypothesis that the amount filled corresponds on average to the setpoint. What is the value of the t-test statistic?

Answerlist

- -11.067
- -6.094
- -9.088
- 21.690
- -9.479

Write a conclusion in context of the problem.

Pergunta 8- Anova 1-Fator



Estes dados são de um pesquisador que mediu a produção em 40 fungos selecionados aleatoriamente em 4 diferentes tipos de habitat. Os tipos de habitat foram definidos com base nas principais espécies de árvores que ocorrem dentro de um buffer de 10 metros ao redor dos fungos amostrados. Com base nessas informações, qual variável é depedente e qual é a independente?

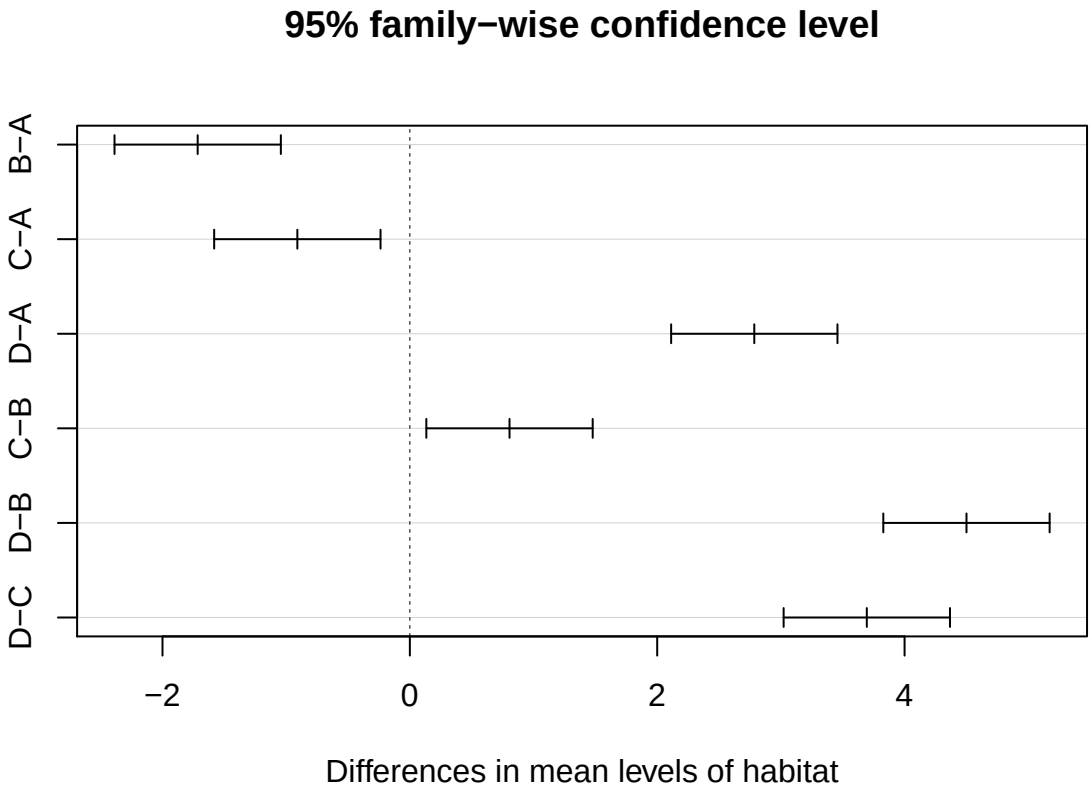
I. Let μ_A be the true mean Let μ_B be the true mean Let μ_C be the true mean Let μ_D be the true mean

II. $H_0 : \mu_A = \mu_B = \mu_C = \mu_D$ H_A : At least one group mean is different.

```
# Quadro da análise de variância
m1<-aov(y~habitat,data=df)
summary(m1)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## habitat       3  460.8  153.61   114.5 <2e-16 ***
## Residuals   156  209.2    1.34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Comparação de médias segundo o teste de Tukey a 5%
plot(TukeyHSD(m1))
```



Descreva resumidamente seus resultados finais.

Pergunta 9- Anova 2-Fatores

O responsável por um viveiro de mudas de eucalipto estava interessado em saber qual seria o melhor recipiente para crescer as mudas de eucalipto. No viveiro comercializava diferentes espécies de eucalipto. Foi realizado um experimento que teve como variável de resposta a altura das mudas em centímetros após um determinado período de tempo.

Os dados foram enviados aos alunos da disciplina de Análise de Dados e Inferência Estatística que propuseram uma análise variância. Foi usado o programa R para rodar a análise. O código está documentado com os passos da análise de variância.

Após rodar o código proposto você deve fazer um relatório para o responsável pelo viveiro de mudas de eucalipto fazendo a recomendação sobre os recipientes que deve ser escolhidos para crescer as mudas de eucalipto.

```r
knitr::opts_chunk$set(echo = TRUE)


# O experimento fatorial descrito em Banzato & kronka (1989) comparou o crescimento de mudas de
# eucalipto considerando como fatores diferentes tipos de recipientes e espécies.
# http://www.leg.ufpr.br/~paulojus/embrapa/Rembrapa/Rembrapase26.html

# A seguir deve-se ler ("importar") os dados para R com o comando read.table(): Se voce não tiver restr
# de acesso (firewall, etc) pode importar o arquivo diretamente fornecendo a URL (endereço web) do arqu
```

```r
ex04 <- read.table('http://mosaico.upf.br/~mauricio/stat/aulas/exemplo04.txt',header = T)
head(ex04)
```

```
##   rec esp resp
## 1  r1  e1 26.2
## 2  r1  e1 26.0
## 3  r1  e1 25.0
## 4  r1  e1 25.4
## 5  r1  e2 24.8
## 6  r1  e2 24.6
```

```r
#Inicialmente vamos obter um resumo de nosso conjunto de dados usando a função summary().
#Note que para os fatores são exibidos o número de dados em cada nível do fator.
#Já para a variável numérica são mostrados algumas medidas estatísticas.
summary(ex04)
```

```
##  rec    esp          resp
##  r1:8   e1:12   Min.   :18.60
##  r2:8   e2:12   1st Qu.:19.75
##  r3:8           Median :23.70
##                 Mean   :22.97
##                 3rd Qu.:25.48
##                 Max.   :26.70
```

```r
#Vamos explorar um pouco mais os dados calculando as médias para cada nível
# de cada fator e também para as combinações dos níveis dos fatores.
ex04.mr <- with(ex04, tapply(resp, rec, mean))
ex04.mr
```

```
##      r1      r2      r3
## 25.4875 22.7250 20.6875
```

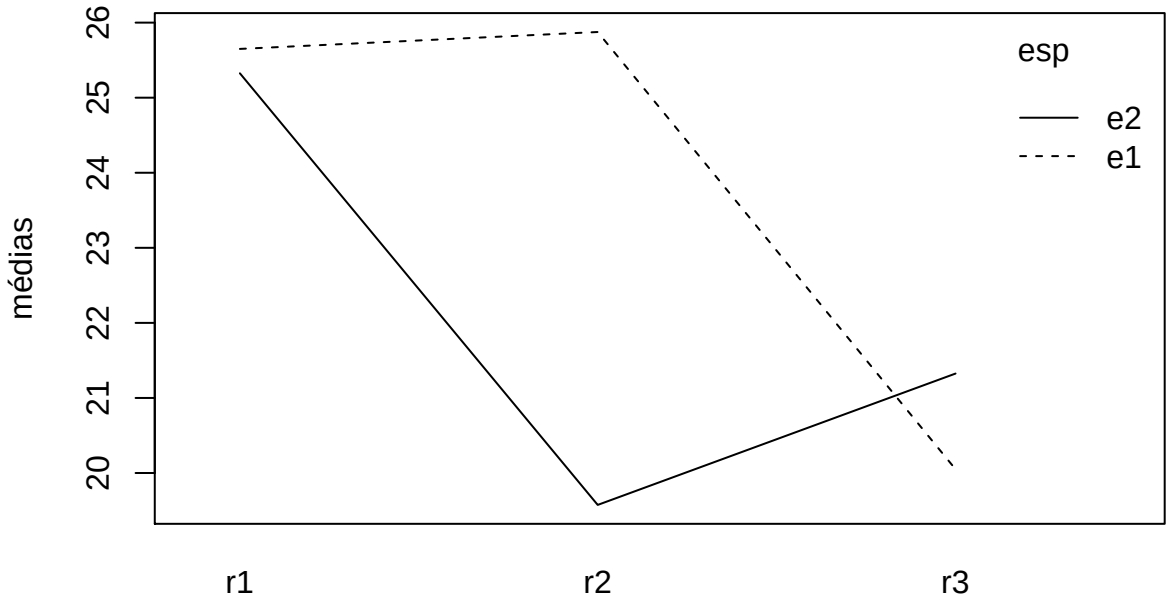
```r
ex04.me <- with(ex04, tapply(resp, esp, mean))
ex04.me
```

```
##       e1       e2
## 23.85833 22.07500
```

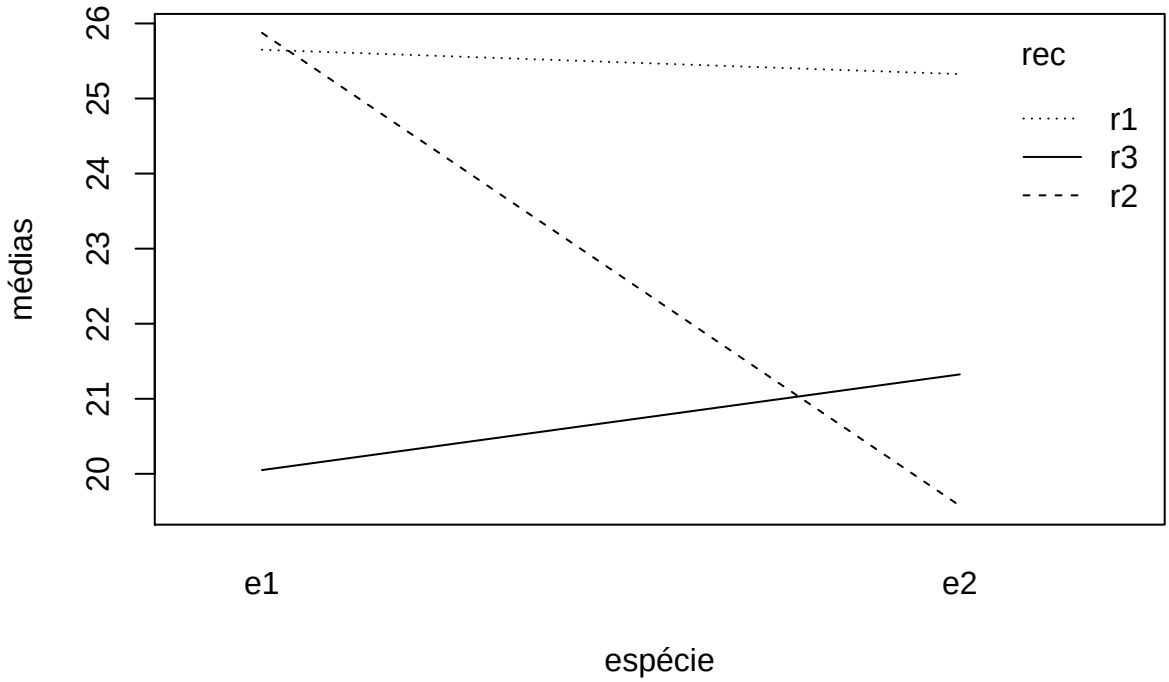
```r
ex04.m <- with(ex04, tapply(resp, list(rec, esp), mean))
ex04.m
```

```
##        e1     e2
## r1 25.650 25.325
## r2 25.875 19.575
## r3 20.050 21.325
```

```r
# Em experimentos fatoriais é importante verificar se existe interação entre os fatores.
# Inicialmente vamos fazer isto graficamente e mais a frente faremos um teste formal para presença de i
# Os comandos a seguir são usados para produzir os gráficos

with(ex04, interaction.plot(rec, esp, resp, ylab = "médias", xlab = "recipiente", xpd = F))
```




```
with(ex04, interaction.plot(esp, rec, resp, ylab = "médias", xlab = "espécie", xpd = F))
```




```
# Seguindo o modelo adequado, o análise de variância para este experimento inteiramente casualizado
# em esquema fatorial pode ser obtida com as funções aov() ("analysis of variance")
ex04.av <- aov(resp ~ rec * esp, data = ex04)
summary(ex04.av)
```

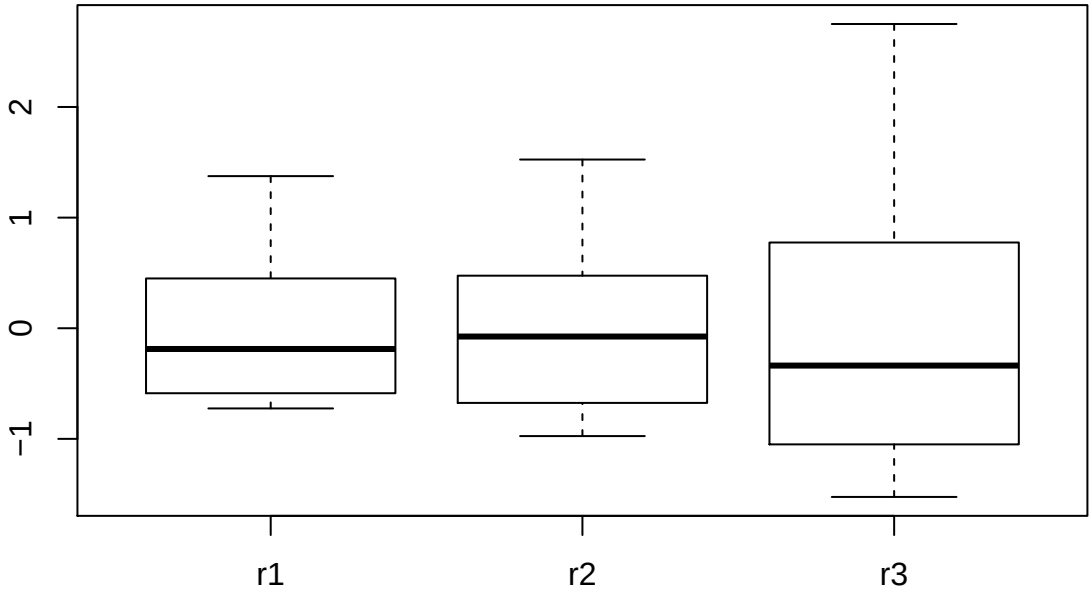
```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## rec          2  92.86   46.43   36.20 4.92e-07 ***
## esp          1  19.08   19.08   14.88  0.00116 **
## rec:esp      2  63.76   31.88   24.85 6.64e-06 ***
```

```
## Residuals   18  23.09    1.28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

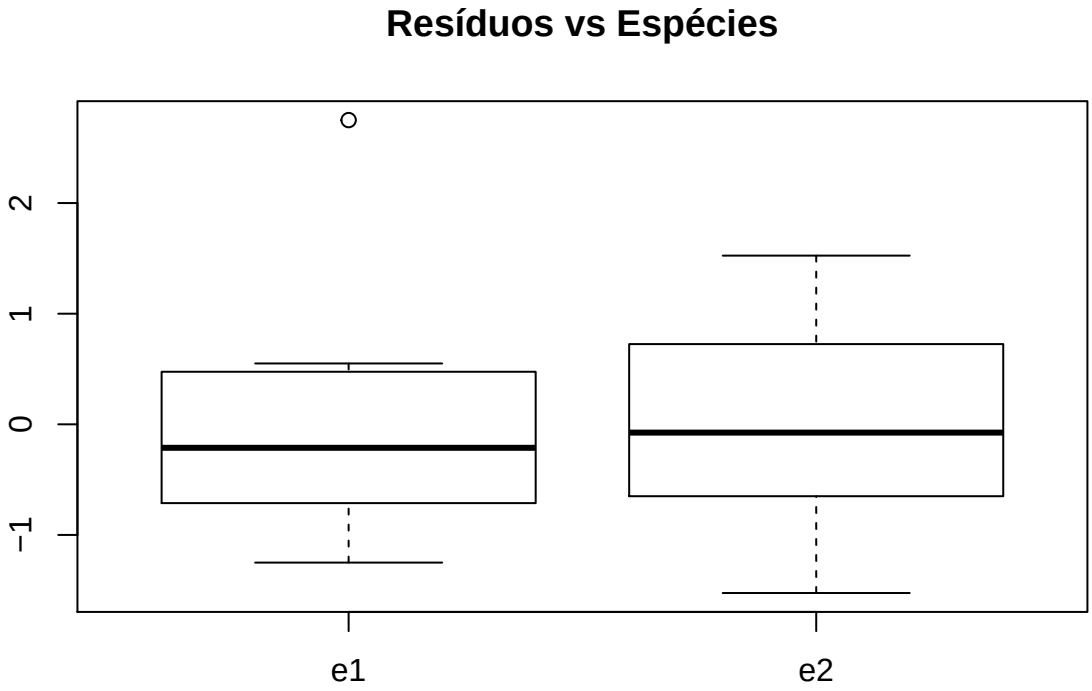
```
#O objeto ex04.av guarda todos os resultados da análise e pode ser explorado por diversos comandos.
model.tables(ex04.av, type = "means")
```

```
## Tables of means
## Grand mean
##
## 22.96667
##
##  rec
## rec
##     r1     r2     r3
## 25.487 22.725 20.687
##
##  esp
## esp
##     e1     e2
## 23.858 22.075
##
##  rec:esp
##     esp
## rec  e1     e2
##   r1 25.650 25.325
##   r2 25.875 19.575
##   r3 20.050 21.325
```

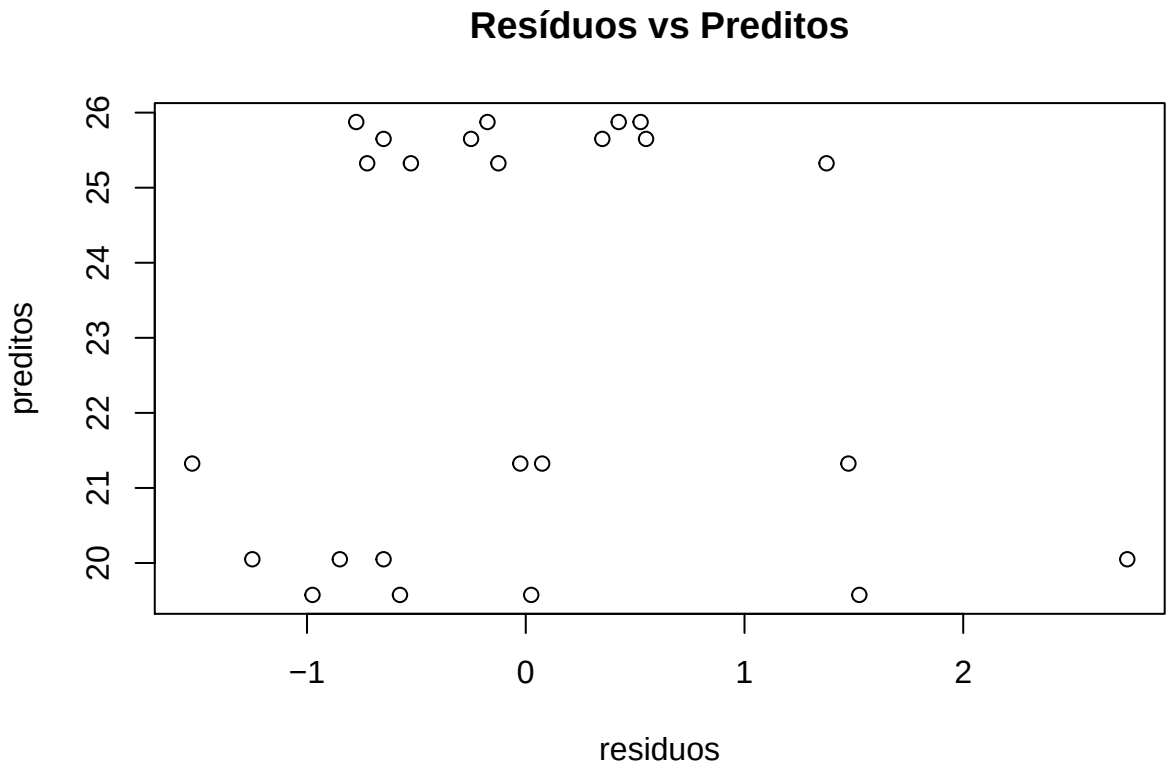
```
# A análise de resíduos é útil para verificar os pressupostos do modelo.
# Usando o mecanismos de classes, o comando plot(ex04.av) aplicado sobre o
# objeto que contém o ajuste do modelo produz uma figura com quatro gráficos
# básicos para análise dos resíduos conforme mostrado na Figura


residuos <- resid(ex04.av)
plot(ex04$rec, residuos)
title("Resíduos vs Recipientes")
```



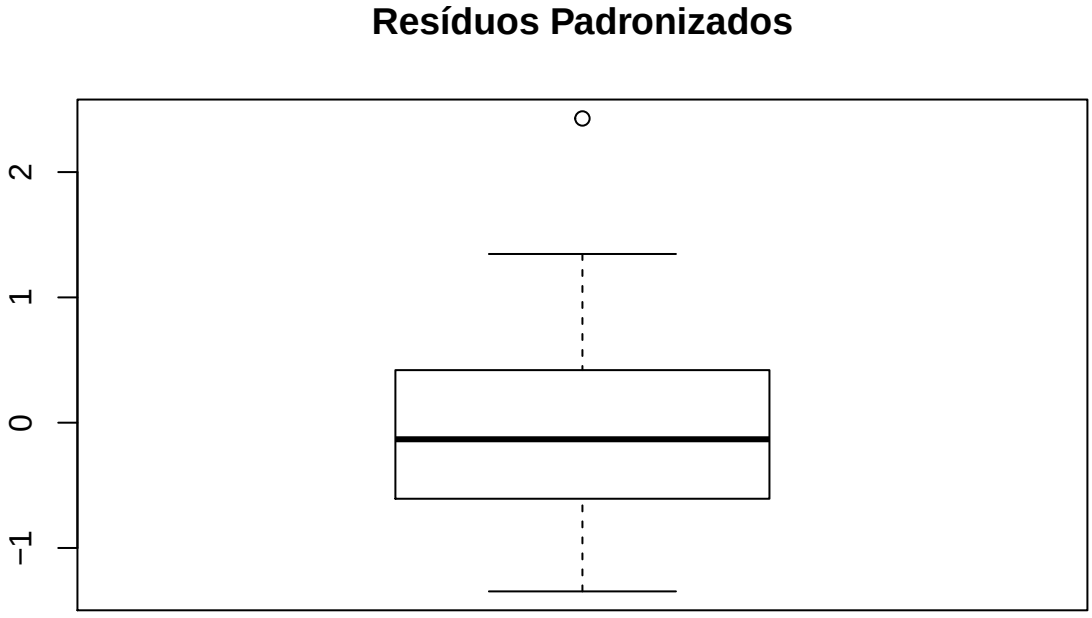

```
plot(ex04$esp, residuos)
title("Resíduos vs Espécies")
```



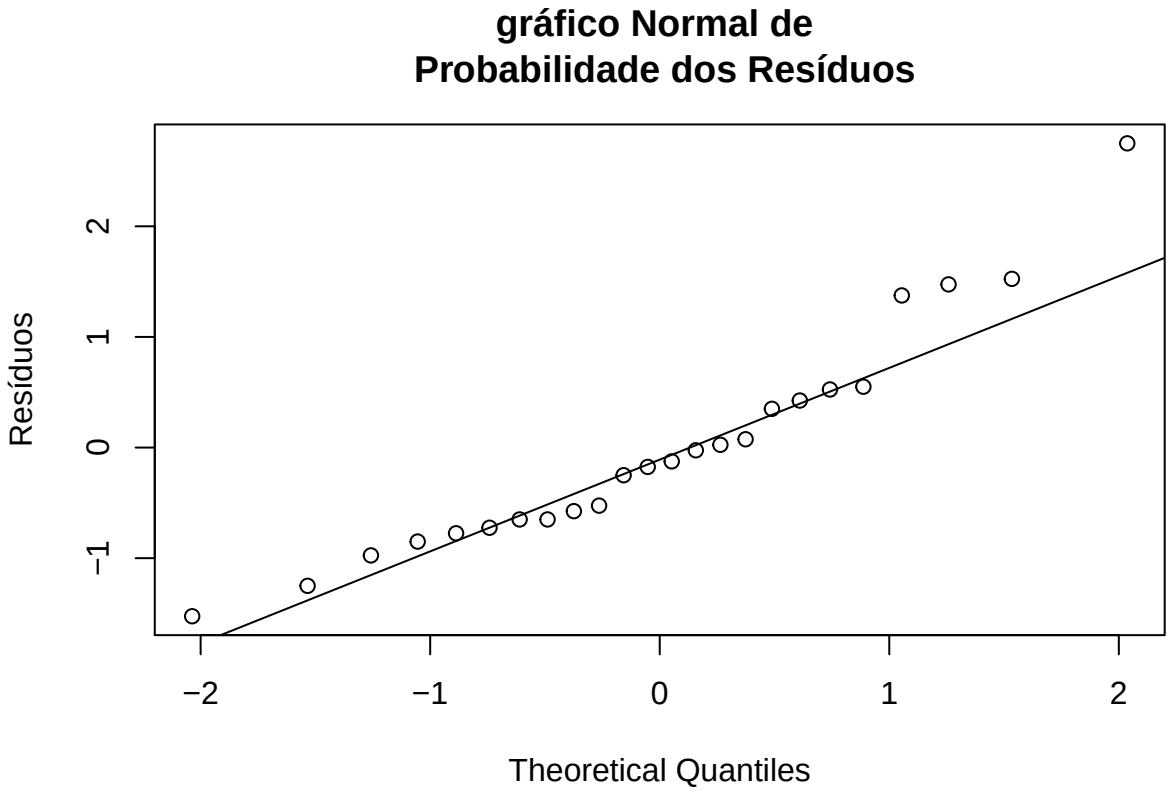

```
preditos <- fitted(ex04.av)
plot(residuos, preditos)
title("Resíduos vs Preditos")
```




```
s2 <- sum(residuos^2)/ex04.av$df.res
respad <- residuos/sqrt(s2)
boxplot(respad)
title("Resíduos Padronizados")
```




```
qqnorm(residuos, ylab = "Resíduos", main = NULL)
qqline(residuos)
title("gráfico Normal de \n Probabilidade dos Resíduos")
```




```
# Além da análise gráfica de resíduos existem alguns testes já programados em funções.
# Como exemplo vejamos o teste de Shapiro-Wilks para testar a normalidade dos Resíduos.
shapiro.test(residuos)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuos
## W = 0.9293, p-value = 0.09402
```

```
# Quando a interação entre os fatores é significativa pode-se adotar como estratégia de análise o desdo
#  graus de liberdade de um fator dentro de cada nível do outro fator. Uma forma de obter tal desdobram
# seria reajustar o modelo utilizando a notação / que indica efeitos aninhados. Desta forma podemos des
# de espécie dentro de cada recipiente e vice versa conforme mostrado a seguir.
ex04.avr <- aov(resp ~ rec/esp, data = ex04)
summary(ex04.avr, split = list('rec:esp' = list(r1 = 1, r2 = 2, r3 = 3)))
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## rec            2  92.86   46.43  36.195 4.92e-07 ***
## rec:esp        3  82.84   27.61  21.527 3.51e-06 ***
##   rec:esp: r1  1   0.21    0.21   0.165    0.690
##   rec:esp: r2  1  79.38   79.38  61.881 3.11e-07 ***
##   rec:esp: r3  1   3.25    3.25   2.535    0.129
## Residuals     18  23.09    1.28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

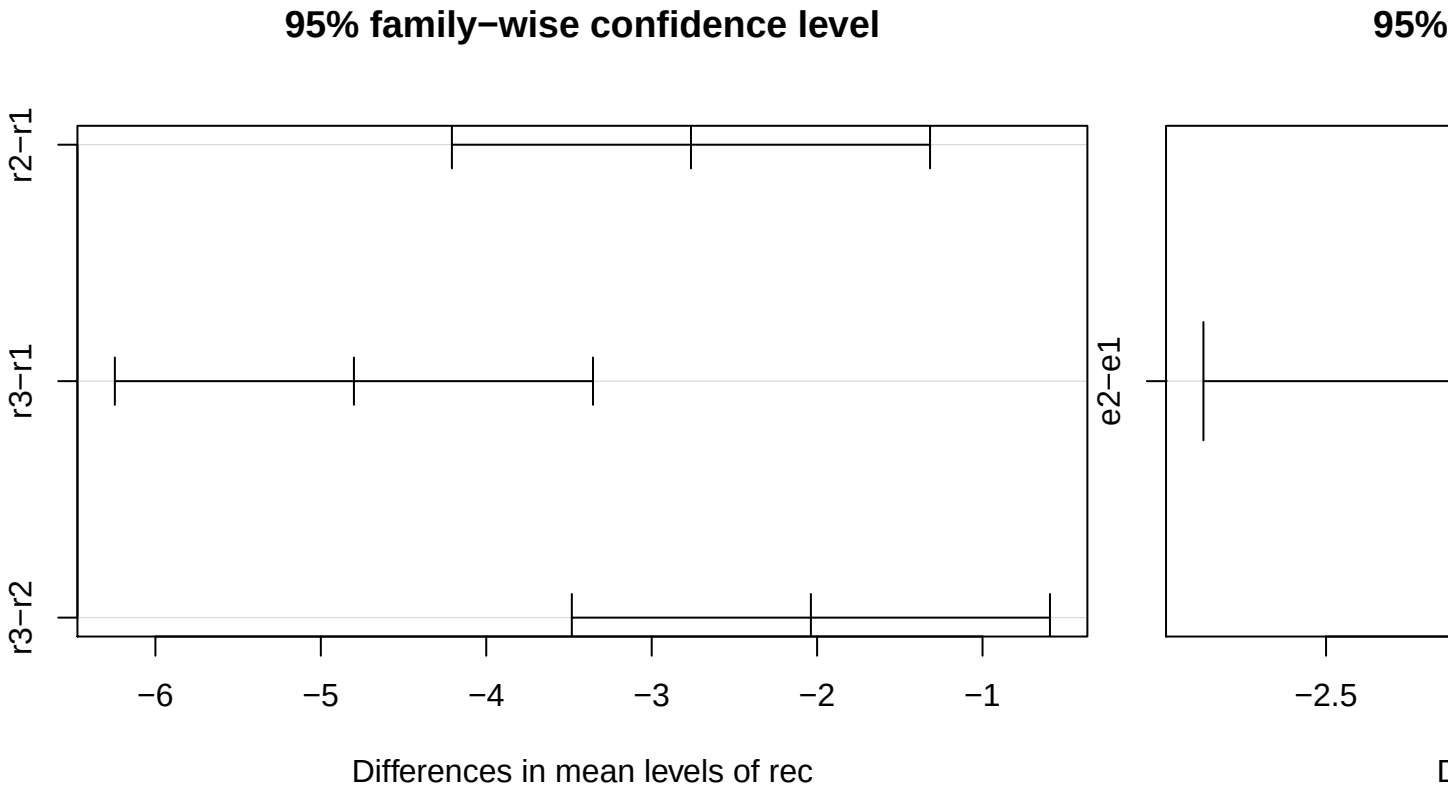
```r
ex04.ave <- aov(resp ~ esp/rec, data = ex04)
summary(ex04.ave, split = list('esp:rec' = list(e1 = c(1, 3), e2 = c(2, 4))))
```

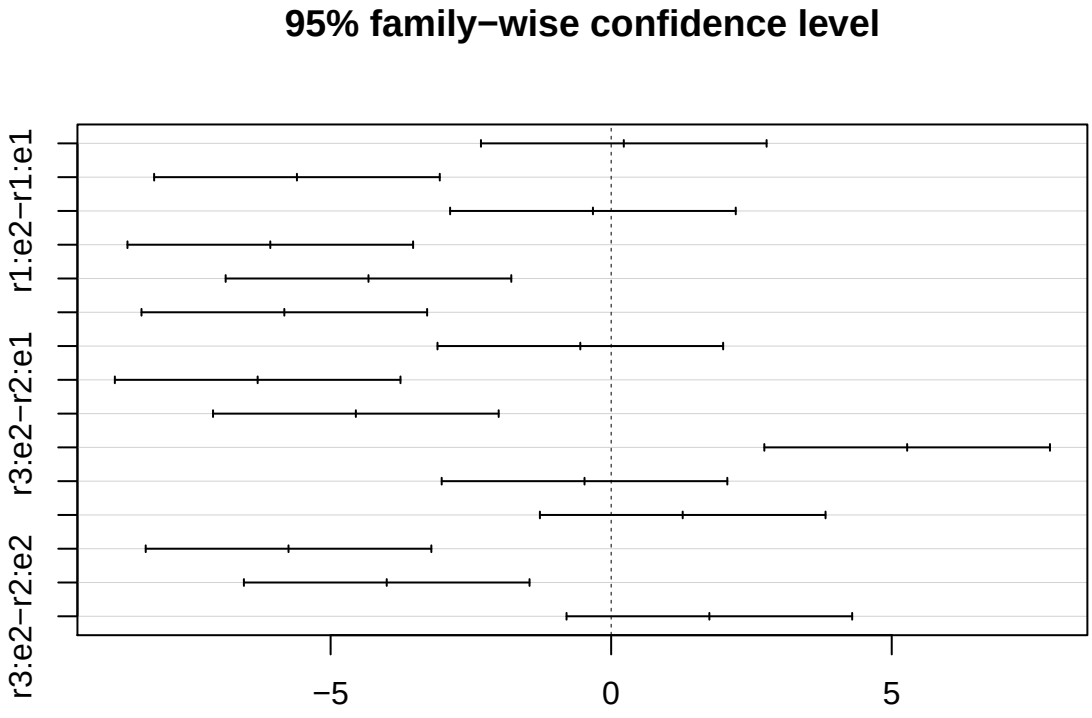
```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## esp             1  19.08   19.08   14.88  0.00116 **
## esp:rec         4 156.62   39.16   30.52 8.44e-08 ***
##   esp:rec: e1   2  87.12   43.56   33.96 7.78e-07 ***
##   esp:rec: e2   2  69.50   34.75   27.09 3.73e-06 ***
## Residuals      18  23.09    1.28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Existem vários testes de comparações múltiplas disponíveis na literatura, e muitos deles são implemen
# R e/ou em pacotes contribuídos. Por exemplo, o pacote multcomp é inteiramente dedicado é implementaçã
# de comparações múltiplas no R. Além disto, procedimentos que não estejam implementados podem ser calc
# usuais do R utilizando os objetos com o ajuste dos modelos. Como ilustração mostramos a seguir duas f
# o Teste de Tukey, a primeira usando uma implementação já disponível com a função TukeyHSD() e uma seg
# necessários passo a passo com operações básicas do R. Para função já disponível simplesmente digitamo
# os resultados podem ser mostrados na forma texto ou gráfica que é produzida com o comando plot(ex04.t

ex04.tk1 <- TukeyHSD(ex04.av)
ex04.tk1
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = resp ~ rec * esp, data = ex04)
##
## $rec
##          diff       lwr        upr     p adj
## r2-r1 -2.7625 -4.207787 -1.3172128 0.0003395
## r3-r1 -4.8000 -6.245287 -3.3547128 0.0000003
## r3-r2 -2.0375 -3.482787 -0.5922128 0.0055472
##
## $esp
##            diff      lwr        upr     p adj
## e2-e1 -1.783333 -2.75476 -0.8119067 0.0011553
##
## $`rec:esp`
##                diff        lwr       upr     p adj
## r2:e1-r1:e1  0.225 -2.3201851  2.770185 0.9997185
## r3:e1-r1:e1 -5.600 -8.1451851 -3.054815 0.0000204
## r1:e2-r1:e1 -0.325 -2.8701851  2.220185 0.9983324
## r2:e2-r1:e1 -6.075 -8.6201851 -3.529815 0.0000068
## r3:e2-r1:e1 -4.325 -6.8701851 -1.779815 0.0004825
## r3:e1-r2:e1 -5.825 -8.3701851 -3.279815 0.0000120
## r1:e2-r2:e1 -0.550 -3.0951851  1.995185 0.9811892
## r2:e2-r2:e1 -6.300 -8.8451851 -3.754815 0.0000041
## r3:e2-r2:e1 -4.550 -7.0951851 -2.004815 0.0002705
## r1:e2-r3:e1  5.275  2.7298149  7.820185 0.0000444
## r2:e2-r3:e1 -0.475 -3.0201851  2.070185 0.9902110
## r3:e2-r3:e1  1.275 -1.2701851  3.820185 0.6135909
## r2:e2-r1:e2 -5.750 -8.2951851 -3.204815 0.0000143
## r3:e2-r1:e2 -4.000 -6.5451851 -1.454815 0.0011258
## r3:e2-r2:e2  1.750 -0.7951851  4.295185 0.2914242
```

```r
plot(ex04.tk1)
```



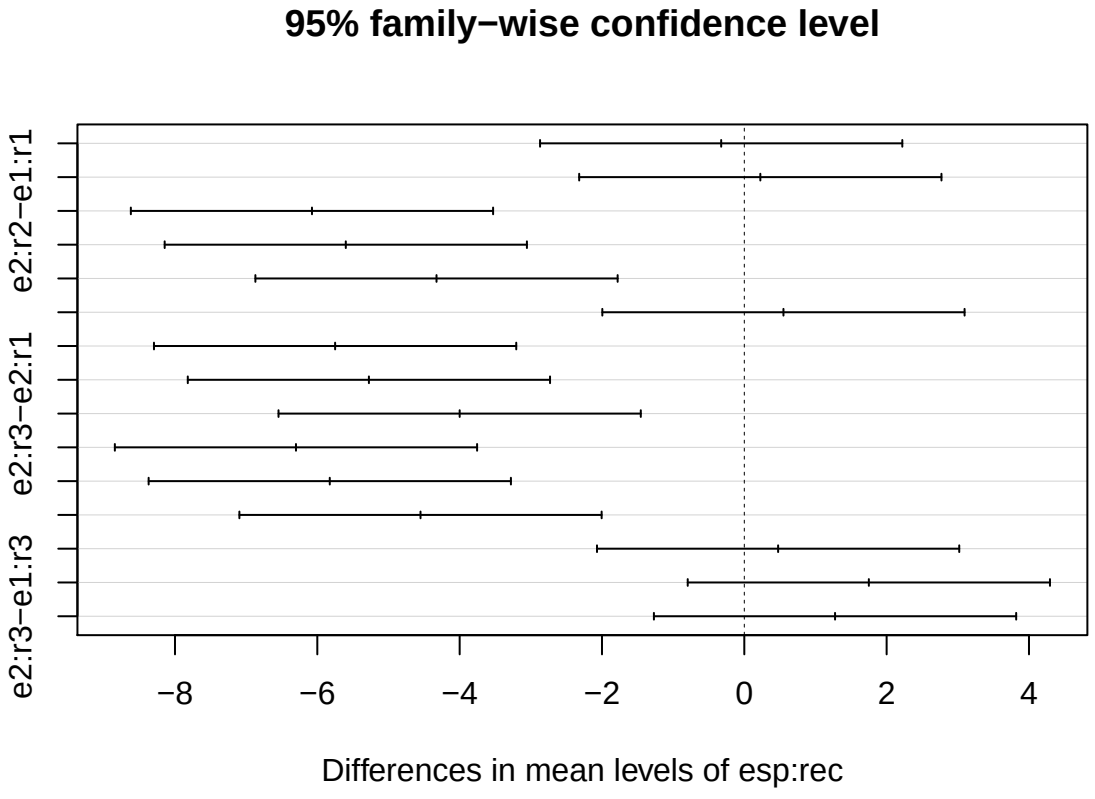



```r
# Esta saída fornece resultados detalhados de várias comparações possíveis entre os níveis dos fatores
# Entretanto, neste caso, nem todos os resultados mostrados nos interessam. Como a interação foi signif
```

```r
# deste experimento a comparação dos níveis fatores principais não nos interessa. Podemos então pedir a
# somente mostre a comparação de médias entre as combinações dos níveis dos fatores e o gráfico com tai
# obtido com plot(ex04.tk2).
 ex04.tk2 <- TukeyHSD(ex04.ave, "esp:rec")
 ex04.tk2
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = resp ~ esp/rec, data = ex04)
##
## $`esp:rec`
##                diff        lwr       upr     p adj
## e2:r1-e1:r1 -0.325 -2.8701851  2.220185 0.9983324
## e1:r2-e1:r1  0.225 -2.3201851  2.770185 0.9997185
## e2:r2-e1:r1 -6.075 -8.6201851 -3.529815 0.0000068
## e1:r3-e1:r1 -5.600 -8.1451851 -3.054815 0.0000204
## e2:r3-e1:r1 -4.325 -6.8701851 -1.779815 0.0004825
## e1:r2-e2:r1  0.550 -1.9951851  3.095185 0.9811892
## e2:r2-e2:r1 -5.750 -8.2951851 -3.204815 0.0000143
## e1:r3-e2:r1 -5.275 -7.8201851 -2.729815 0.0000444
## e2:r3-e2:r1 -4.000 -6.5451851 -1.454815 0.0011258
## e2:r2-e1:r2 -6.300 -8.8451851 -3.754815 0.0000041
## e1:r3-e1:r2 -5.825 -8.3701851 -3.279815 0.0000120
## e2:r3-e1:r2 -4.550 -7.0951851 -2.004815 0.0002705
## e1:r3-e2:r2  0.475 -2.0701851  3.020185 0.9902110
## e2:r3-e2:r2  1.750 -0.7951851  4.295185 0.2914242
## e2:r3-e1:r3  1.275 -1.2701851  3.820185 0.6135909
```

```r
 plot(ex04.tk2)
```



Linear Regression Analysis

I. Let β_1 be

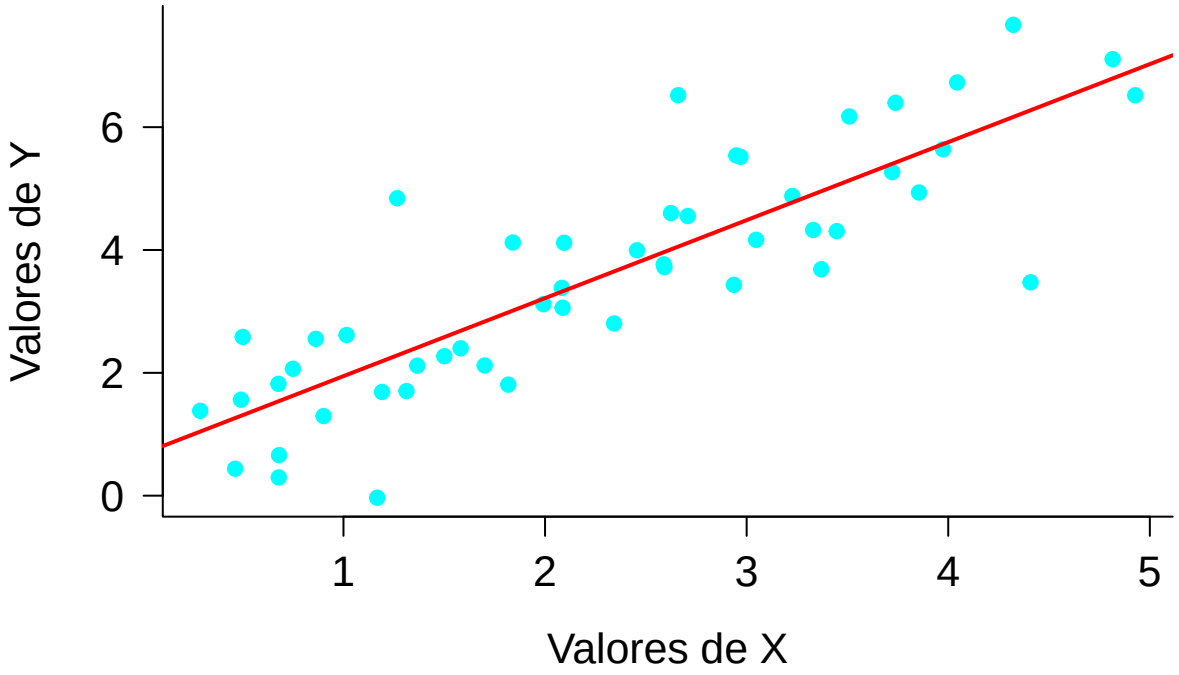
II. $H_0 : \beta_1 = 0$ $H_A : \beta_1 \neq 0$

```
# Análise estatística usando o modelo linear
plot(x,y,
     ylab="Valores de Y", xlab="Valores de X",
     las=1,bty="L",cex.lab=1.3,cex.axis=1.3, pch = 19, col = 5)

m2<-lm(y~x)# fit a linear model of y against x
summary(m2)
```

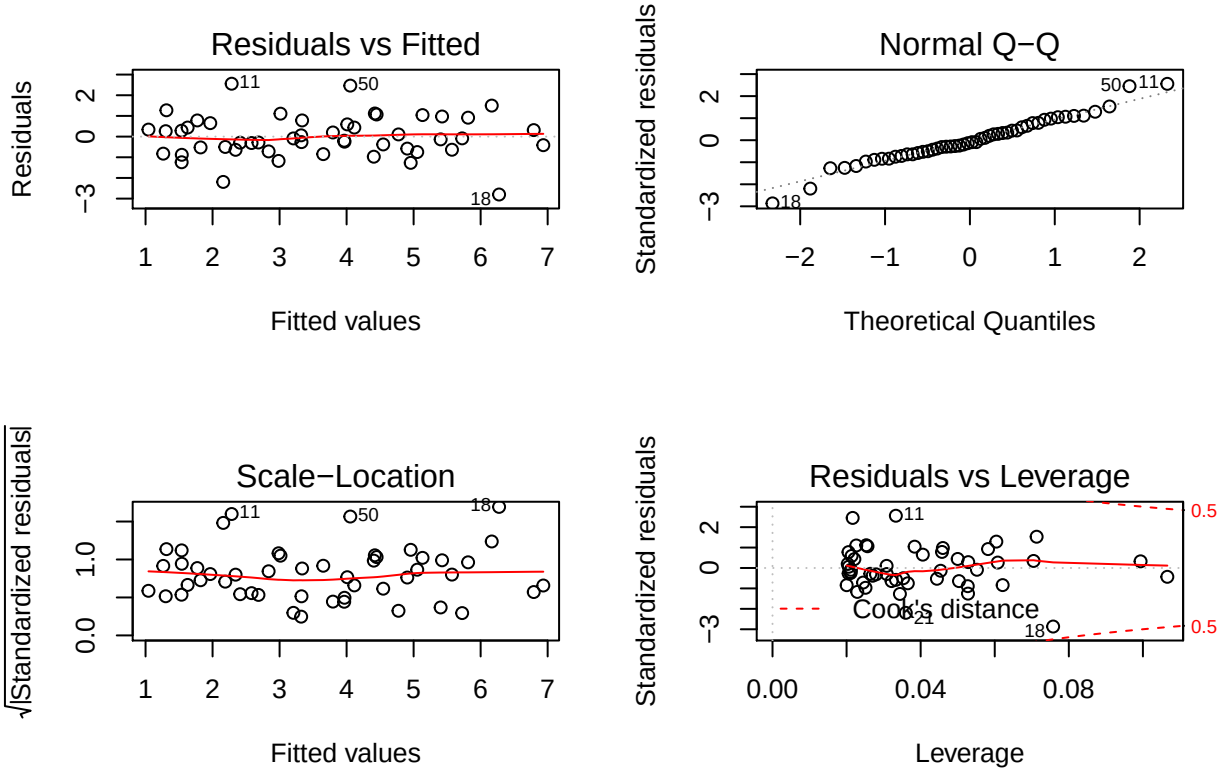
```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8016 -0.6221 -0.1117  0.6355  2.5569
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.6760     0.2983   2.266    0.028 *
## x             1.2701     0.1137  11.166 6.06e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.017 on 48 degrees of freedom
## Multiple R-squared:  0.722,  Adjusted R-squared:  0.7162
## F-statistic: 124.7 on 1 and 48 DF,  p-value: 6.06e-15
```

```
abline(m2, col = 2, lwd = 2, lty = 1)
```




```
par(mfrow=c(2,2))
plot(m2)
```



Pergunta 10: a)Os dados atendem as premissas dos modelos lineares? b) Quanto da variação em Y pode ser explicada pelo modelo? c) Descreva suscintamente os resultados encontrados. d) Qual o valor de Y para a mediana de X?

Regressão não-linear

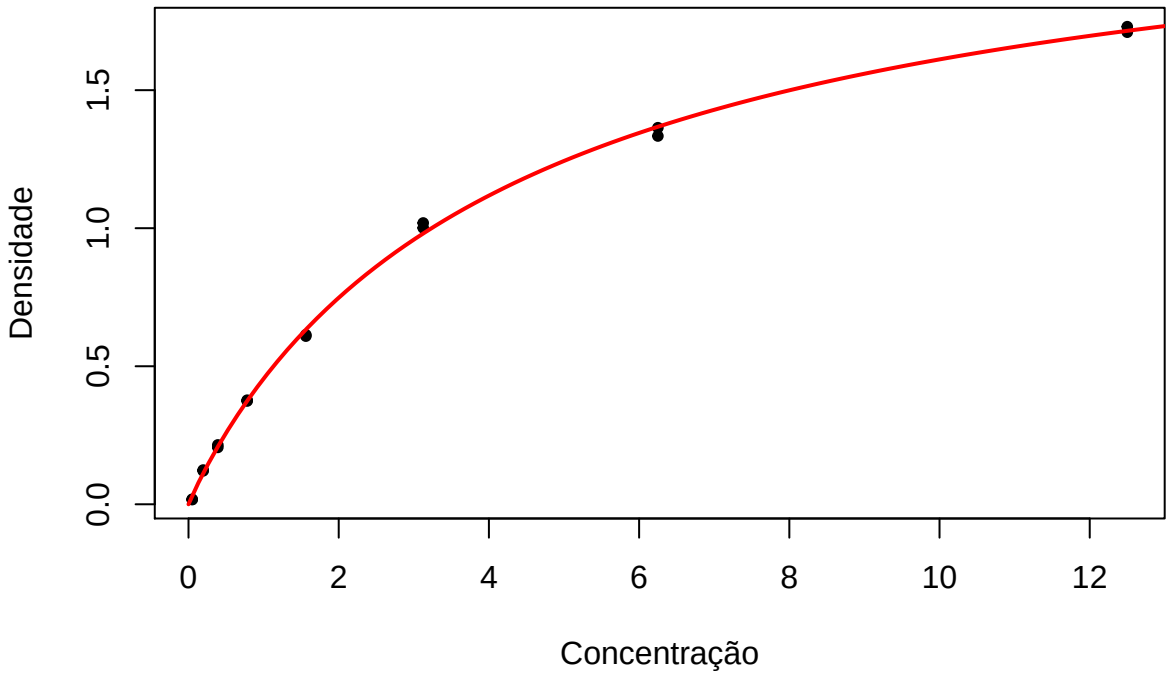
```
my_data <- subset(DNase, Run == 1)
head(my_data)
```

```
##   Run       conc density
## 1   1 0.04882812   0.017
## 2   1 0.04882812   0.018
## 3   1 0.19531250   0.121
## 4   1 0.19531250   0.124
## 5   1 0.39062500   0.206
## 6   1 0.39062500   0.215
```

```
fit1 <- nls(density ~ 1/(1 + exp((xmid - log(conc))/scal)),
            data  = my_data,
            start = list(xmid = 0, scal = 1),
            algorithm = "plinear")
summary(fit1)
```

```
##
## Formula: density ~ 1/(1 + exp((xmid - log(conc))/scal))
##
## Parameters:
##      Estimate Std. Error t value Pr(>|t|)
## xmid  1.48309    0.08135   18.23 1.22e-10 ***
## scal  1.04145    0.03227   32.27 8.51e-14 ***
## .lin  2.34518    0.07815   30.01 2.17e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01919 on 13 degrees of freedom
##
## Number of iterations to convergence: 5
## Achieved convergence tolerance: 1.032e-06
```

```
# plot it:
xs <- data.frame(conc=seq(0,13,len=100))
plot(my_data$conc, my_data$density, ylab="Densidade",xlab="Concentração",pch=20)
lines(xs$conc, predict(fit1,newdata=xs), col="red", lwd=2)
```



Pergunta 11: a) Qual a seria a densidade quando a concentração é 8.345? b) Como os modelos não lineares diferem dos modelos lineares em relação a determinação dos parâmetros?

Regressão logística -

Se a regressão linear serve para prever variáveis Y contínuas, a regressão logística é usada para classificação binária. Se usarmos a regressão linear para modelar uma variável dicotômica (como Y), o modelo resultante pode não restringir o Ys previsto dentro de 0 e 1. Além disso, outras suposições de regressão linear, como a normalidade dos erros, podem ser violadas. Então, em vez disso, modelamos as probabilidades de log do evento $ln\left(\frac{P}{1-P}\right)$, onde, P é a probabilidade de evento.

$$Z_i = ln\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 x_1 + .. + \beta_n x_n$$

A equação acima pode ser modelada usando o glm () definindo o argumento da família como "binomial". Mas estamos mais interessados na probabilidade do evento do que nas probabilidades do evento. Assim, os valores previstos do modelo acima, ou seja, as probabilidades de log do evento, podem ser convertidos para probabilidade de evento da seguinte forma:

$$P_i = 1 - \left(\frac{1}{1+e_i^z}\right)$$

Pergunta 12:a) O que você diria de usar Regressão Logística para o conjunto de dados "Titanic" que vimos em aula. b) Nesse caso, qual seria a variável de resposta?