SME0816 - Planejamento de Experimentos I
Transformação de dados

Profa. Cibele Russo

(Referências: Montgomery (2012), Notas de aula de Roseli Leandro; Clarice Demétrio; Marinho Andrade)

Os modelos

Super parametrizado:

$$y_{ij} = \mu + \tau_i + e_{ij}$$

De médias:

$$y_{ij} = \mu_i + e_{ij}$$

supõem que

$$e_{ij} \sim N(0, \sigma^2)$$

Pressupostos básicos

Os erros

- são normalmente distribuídos
- independentes
- homogêneos (homocedásticos, não heterocedásticos)

Caso uma das pressuposições básicas não seja atendida será necessário transformar os dados.

Transformação de dados

- Necessária quando as pressuposições básicas do modelo não forem atendidas

CONSEQUENCES OF FAILURE TO MEET ASSUMPTIONS UNDERLYING THE FIXED EFFECTS ANALYSES OF VARIANCE AND COVARIANCE

Gene V Glass

Laboratory of Educational Research
University of Colorado

Percy D. Peckham

University of Washington

James R. Sanders

Indiana University

The effects of violating the assumptions underlying the fixed-effects analyses of variance (ANOVA) and covariance (ANCOVA) on Type-I and Type-II error rates have been of great concern to researchers and statisticians. The major effects of violation of assumptions are now well known, after nearly four decades of research. Early summaries and reviews by Hey (1938), Garret and

Cuidado!

- *Entenda* o conjunto de dados
- Faça uma análise explorátoria dos dados
- Converse com o pesquisador (responsável pelos dados)

Transformação de dados

- Necessária quando as pressuposições básicas do modelo não forem atendidas

Recommended Citation
Osborne, Jason (2010) "Improving your data transformations: Applying the Box-Cox transformation," *Practical Assessment, Research, and Evaluation*: Vol. 15 , Article 12.
DOI: https://doi.org/10.7275/qbpc-gk17
Available at: https://scholarworks.umass.edu/pare/vol15/iss1/12

Análise de um conjunto de dados

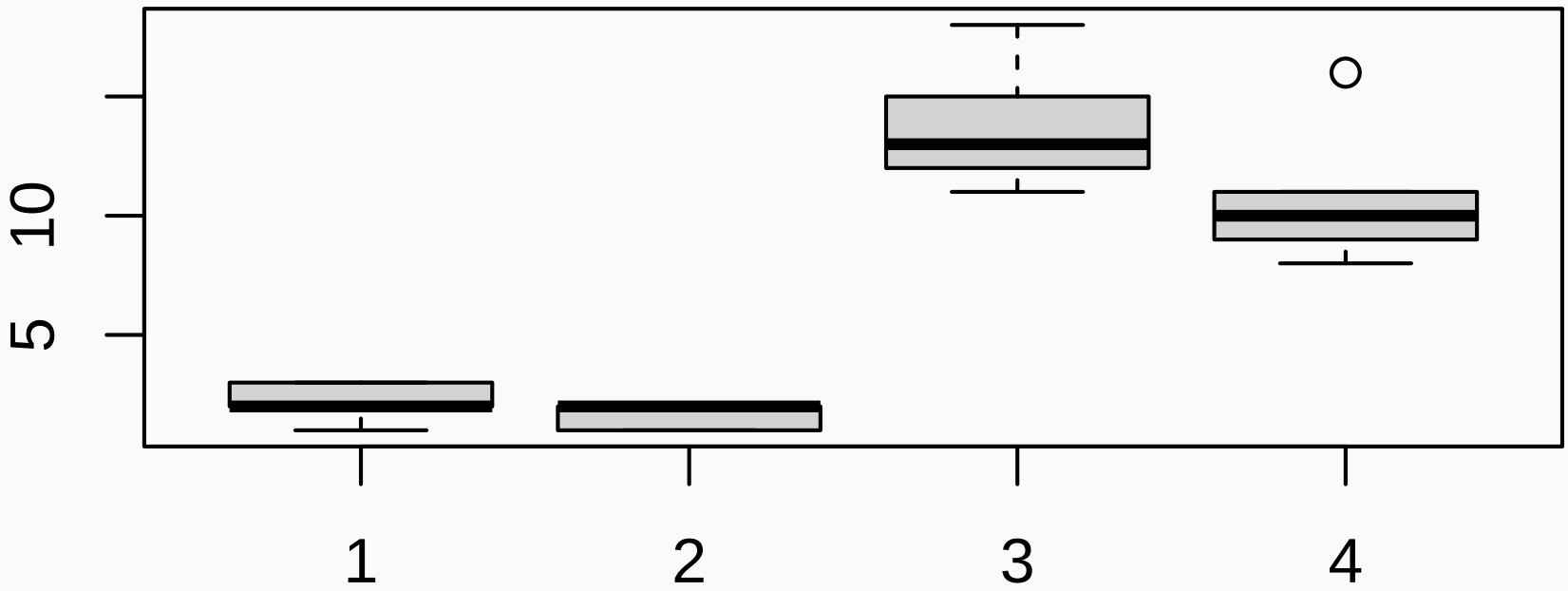
```
y <- c( 3, 3, 2, 2, 1,
        2, 1, 1, 2, 2,
       13,11,15,18,12,
        8,10,16, 9,11)

trat <- as.factor(rep(1:4,each=5))

dad <- data.frame(y=y,trat=trat)
```

Análise descritiva

```
n       <- tapply(dad$y,dad$trat,length)
media   <- tapply(dad$y,dad$trat,mean)
des.pad <- round(tapply(dad$y,dad$trat,sd),4)
cv      <-  round(des.pad/media*100,0)
data.frame(trat=1:4,n=n,media=media,des.pad=des.pad,cv)
```

```
##   trat n media des.pad cv
## 1    1 5   2.2  0.8367 38
## 2    2 5   1.6  0.5477 34
## 3    3 5  13.8  2.7749 20
## 4    4 5  10.8  3.1145 29
```

Análise gráfica



Verificação gráfica das pressuposições básicas

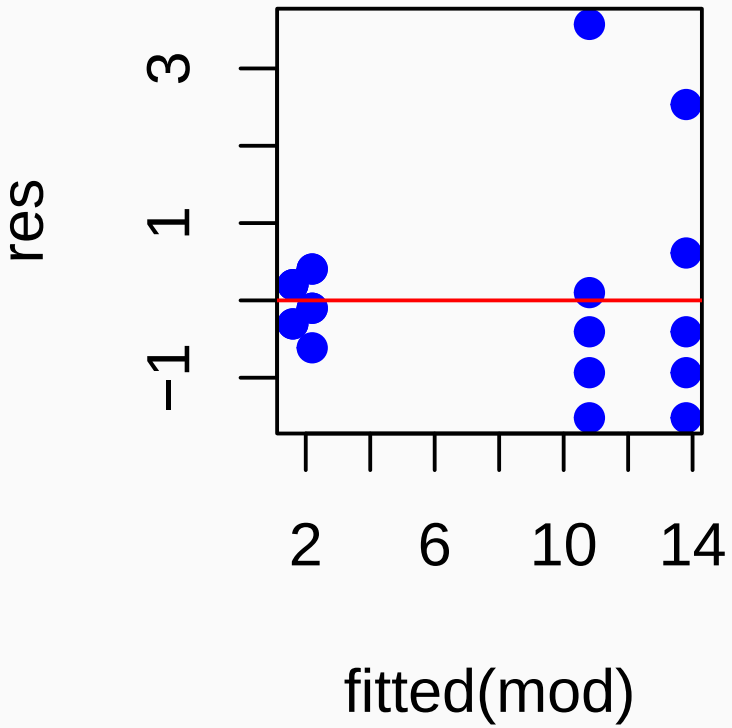
Utilize os resíduos estudentizados

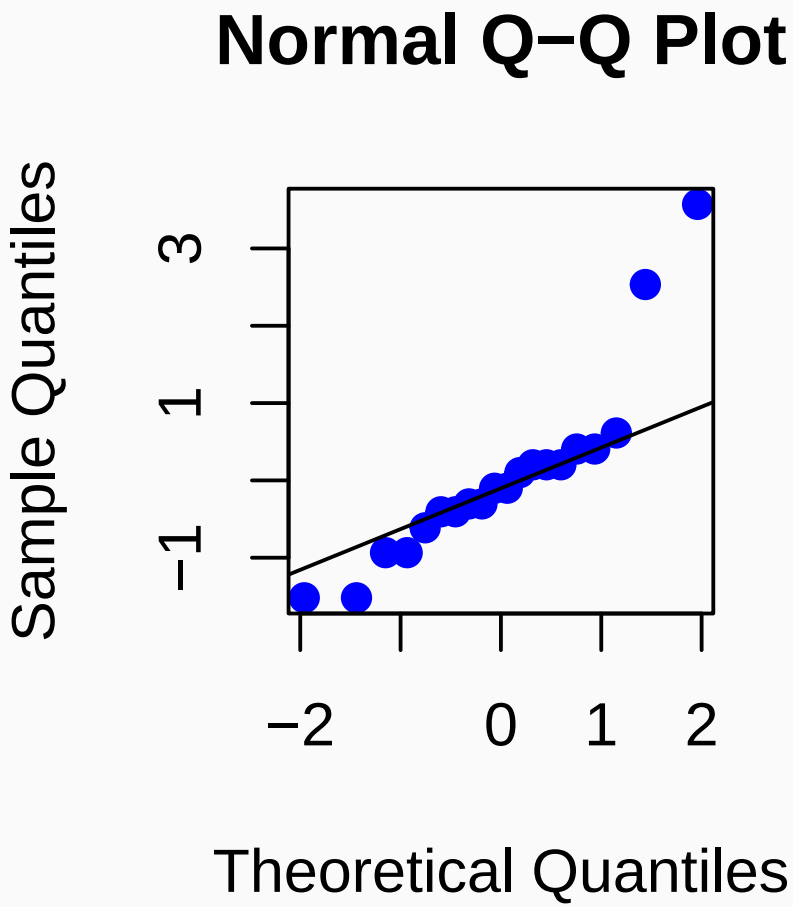
Verificação gráfica dos pressupostos básicos

Análise dos resíduos

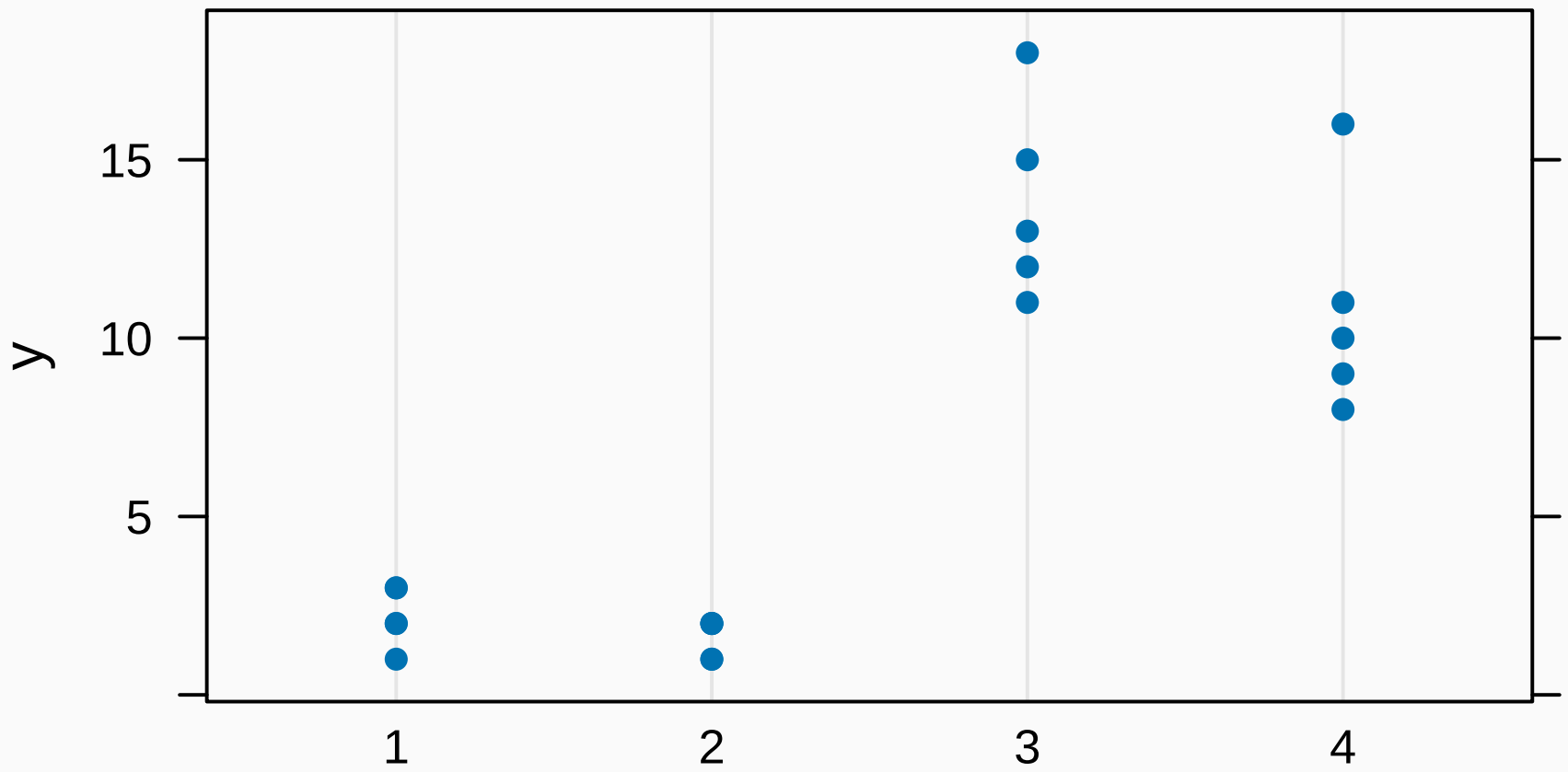
```
mod <- lm(y ~ trat,data=dad)


res <- rstudent(mod)
```

Análise gráfica

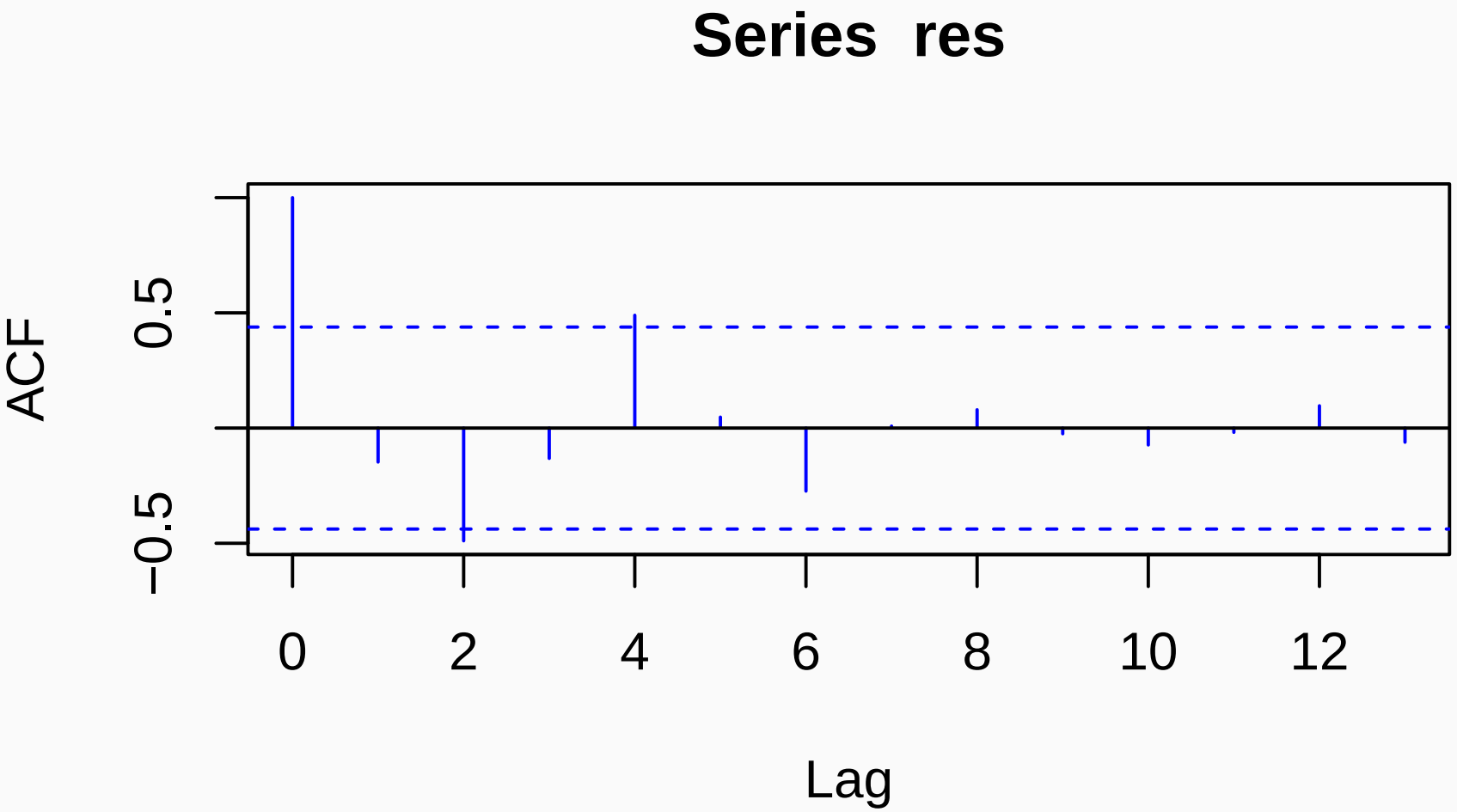




Análise gráfica



Análise gráfica



Verificação das pressuposições básicas

Teste para normalidade

```
shapiro.test(res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.83107, p-value = 0.00261
```

Verificação das pressuposições básicas

Teste para normalidade

```
bartlett.test(y ~ trat, dad)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  y by trat
## Bartlett's K-squared = 12.141, df = 3, p-value = 0.0069
```

Verificação das pressuposições básicas

Teste para normalidade

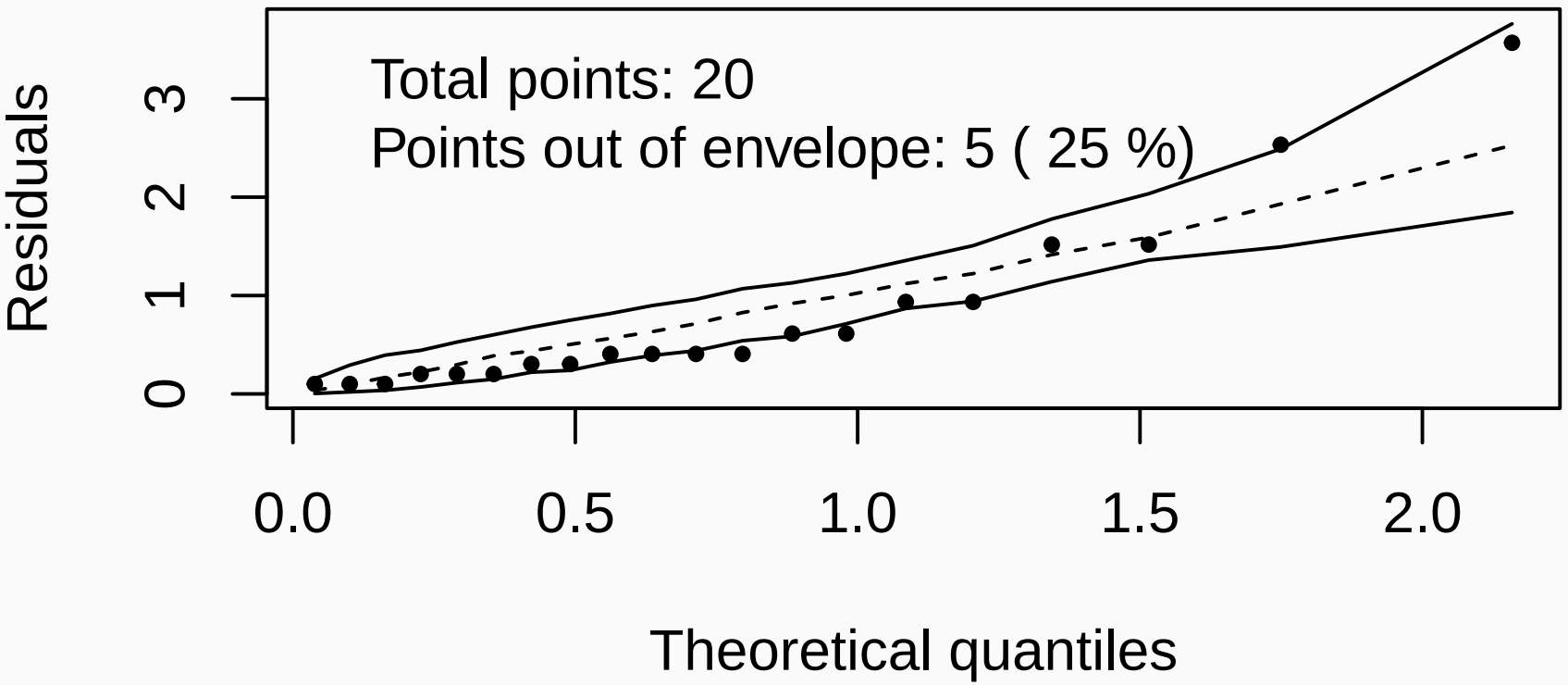
```
car::durbinWatsonTest(mod)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0.1391304       2.269022   0.912
##  Alternative hypothesis: rho != 0
```

- Moral, R. A., Hinde, J., & Demétrio, C. G. B. (2017). Half-Normal Plots and Overdispersed Models in R: The hnp Package. Journal of Statistical Software, 81(10), 1–23. https://doi.org/10.18637/jss.v081.i10

```
hnp::hnp(mod,print.on=T,pch=19,col="blue")
```

```
## Gaussian model (lm object)
```



A transformação Box-Cox

Practical Assessment, Research, and Evaluation

Volume 15 *Volume 15, 2010* Article 12

2010

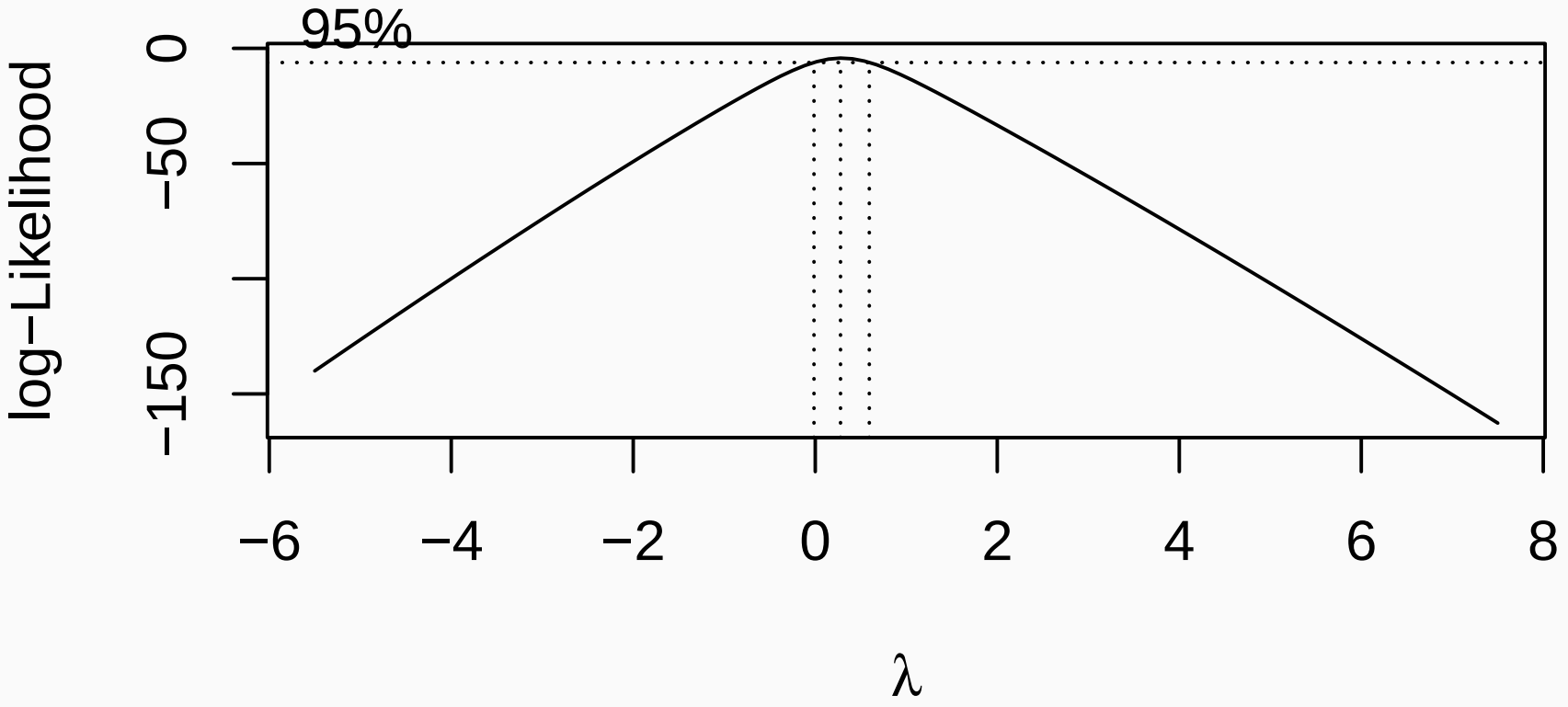
Improving your data transformations: Applying the Box-Cox transformation

Jason Osborne

A transformação Box-Cox

```
require(MASS)
bc <- boxcox(y ~ trat,
       lambda = seq(-5.50, 7.50, length = 100))
(lambda <- bc$x[which.max(bc$y)])

yt <- y^lambda

dad$yt <- yt
```

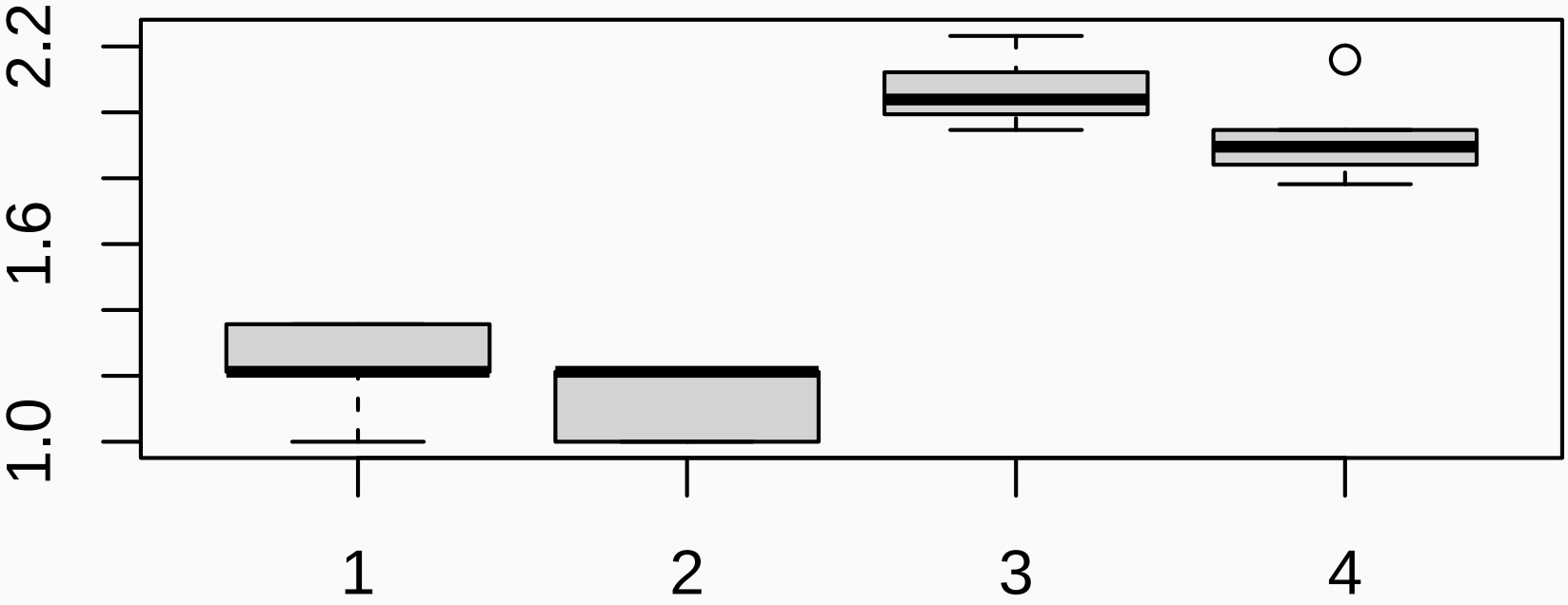
```
## Loading required package: MASS
```




```
## [1] 0.2777778
```

Refazendo a análise com os dados transformados

Análise gráfica



Verificação gráfica das pressuposições básicas

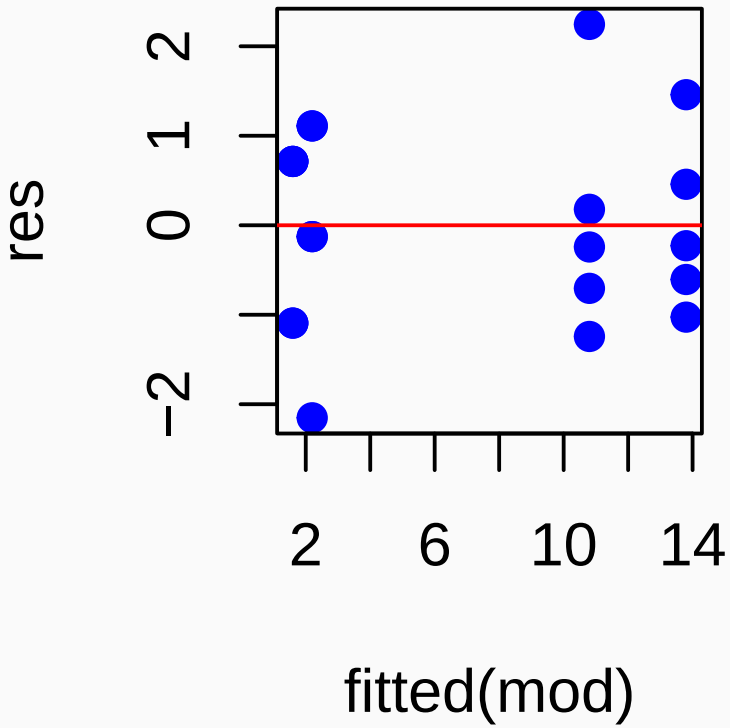
- Utilize os resíduos estudentizados

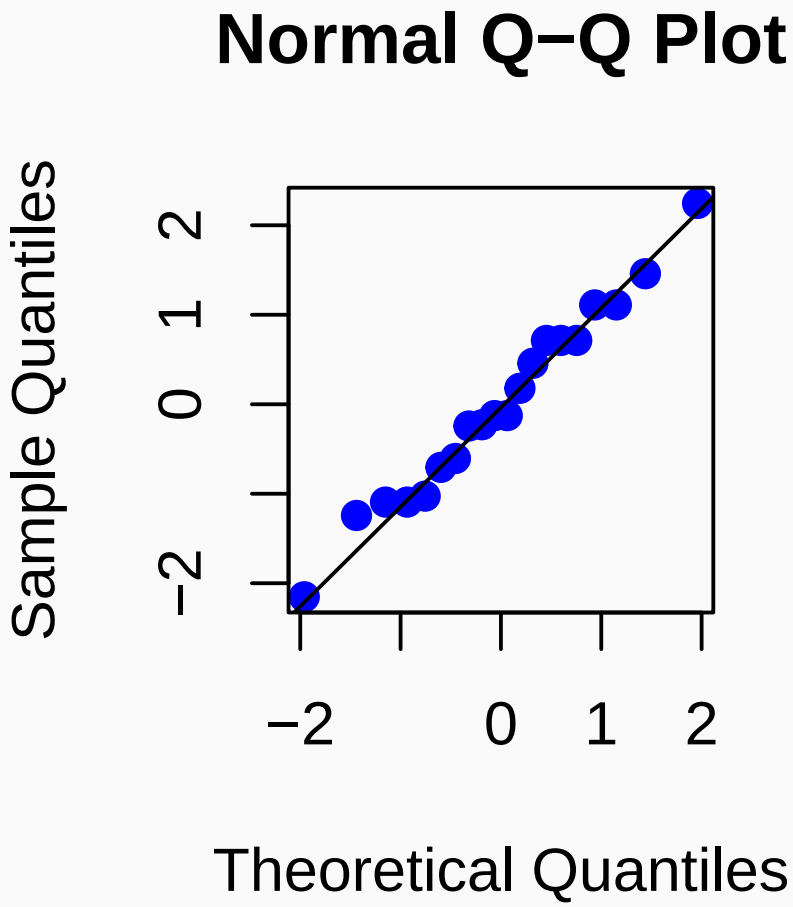
Verificação gráfica das pressuposições básicas

Análise dos resíduos

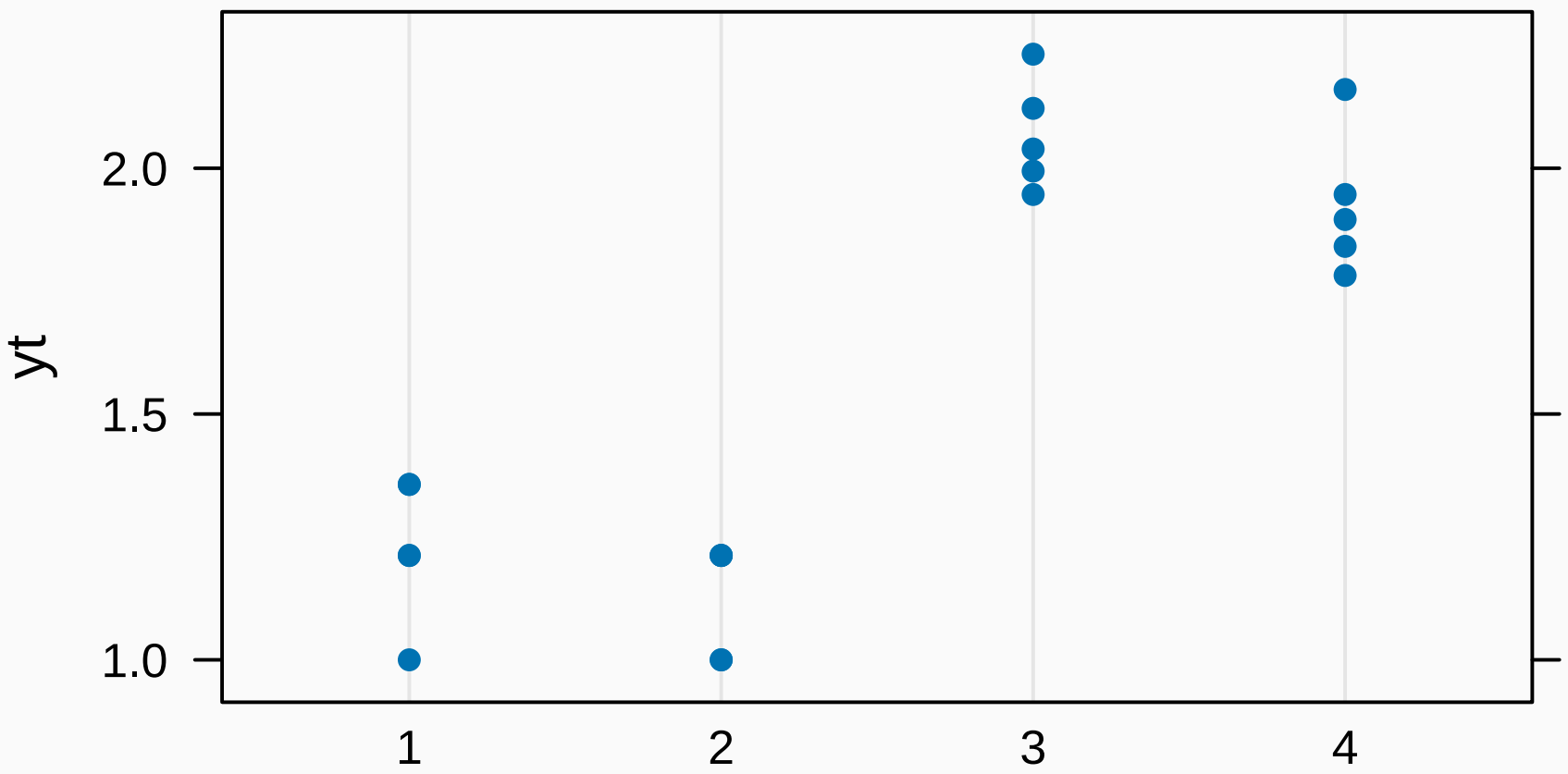
```
mod1 <- lm(yt ~ trat,data=dad)


res  <- rstudent(mod1)
```

Análise gráfica

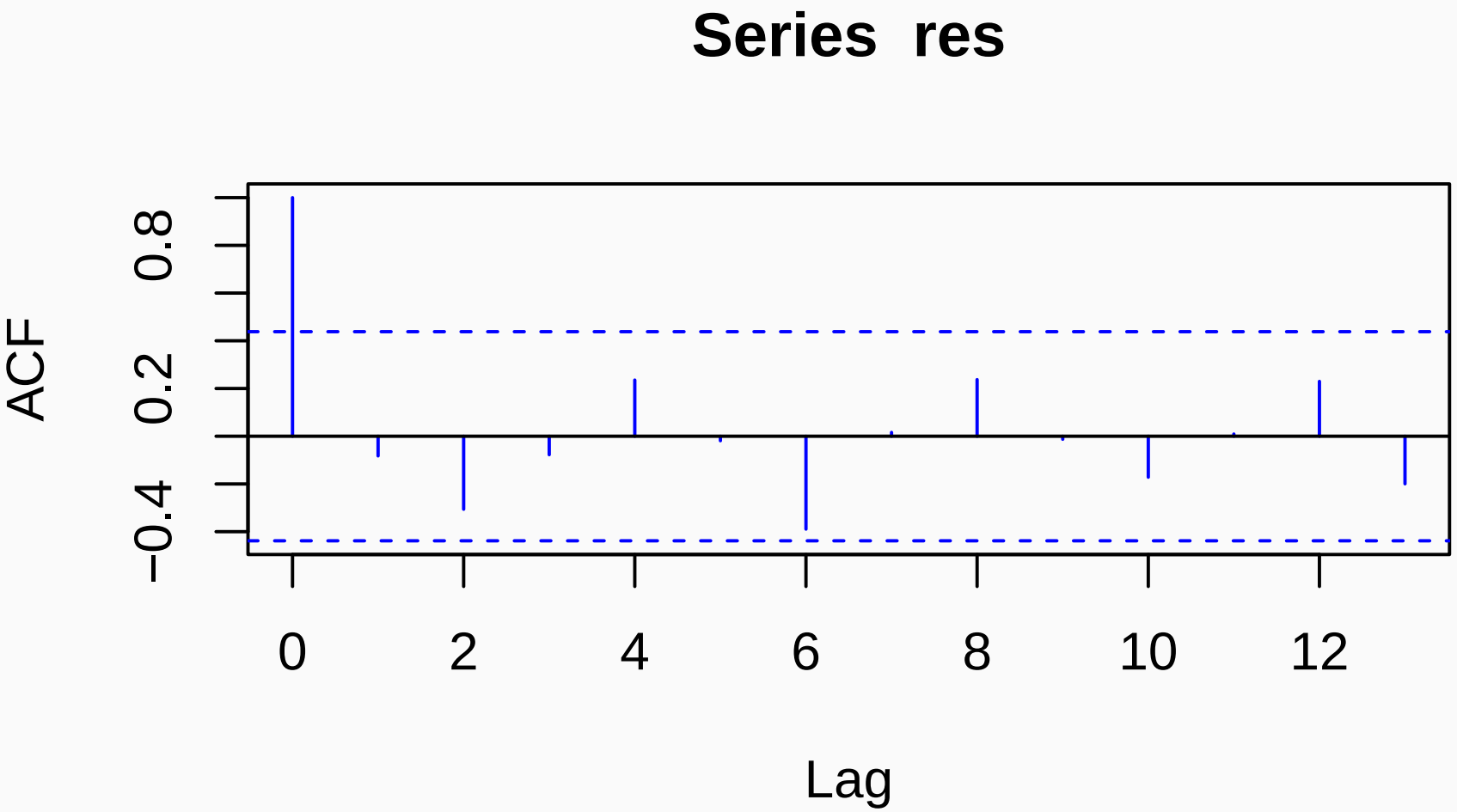




Análise gráfica



Análise gráfica



Verificação das pressuposições básicas

Teste para normalidade

```
shapiro.test(res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.98446, p-value = 0.9781
```

Verificação das pressuposições básicas

Teste para normalidade

```
bartlett.test(yt ~ trat, dad)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  yt by trat
## Bartlett's K-squared = 0.42073, df = 3, p-value = 0.935
```

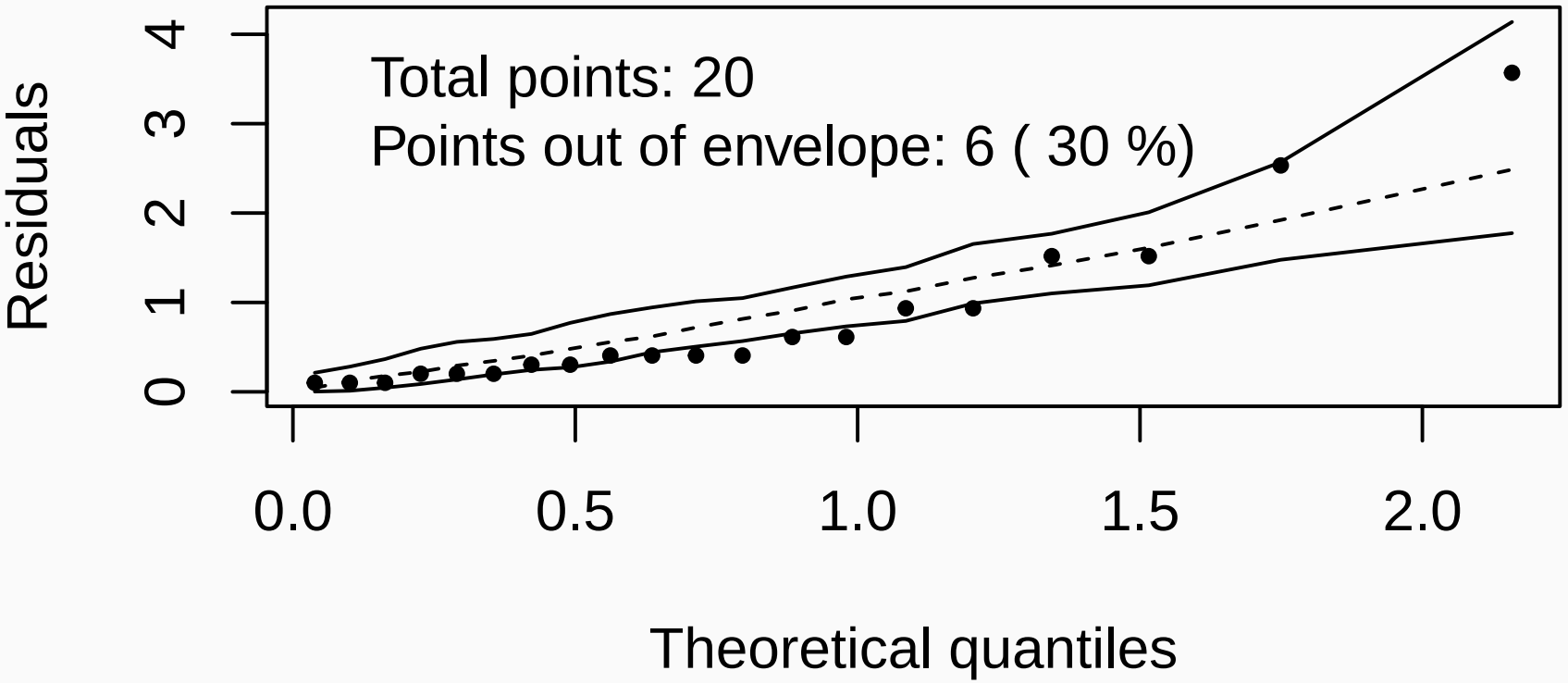
Verificação das pressuposições básicas

Teste para normalidade

```
car::durbinWatsonTest(mod1)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.07819393      2.093954   0.584
##  Alternative hypothesis: rho != 0
```

```
## Gaussian model (lm object)
```



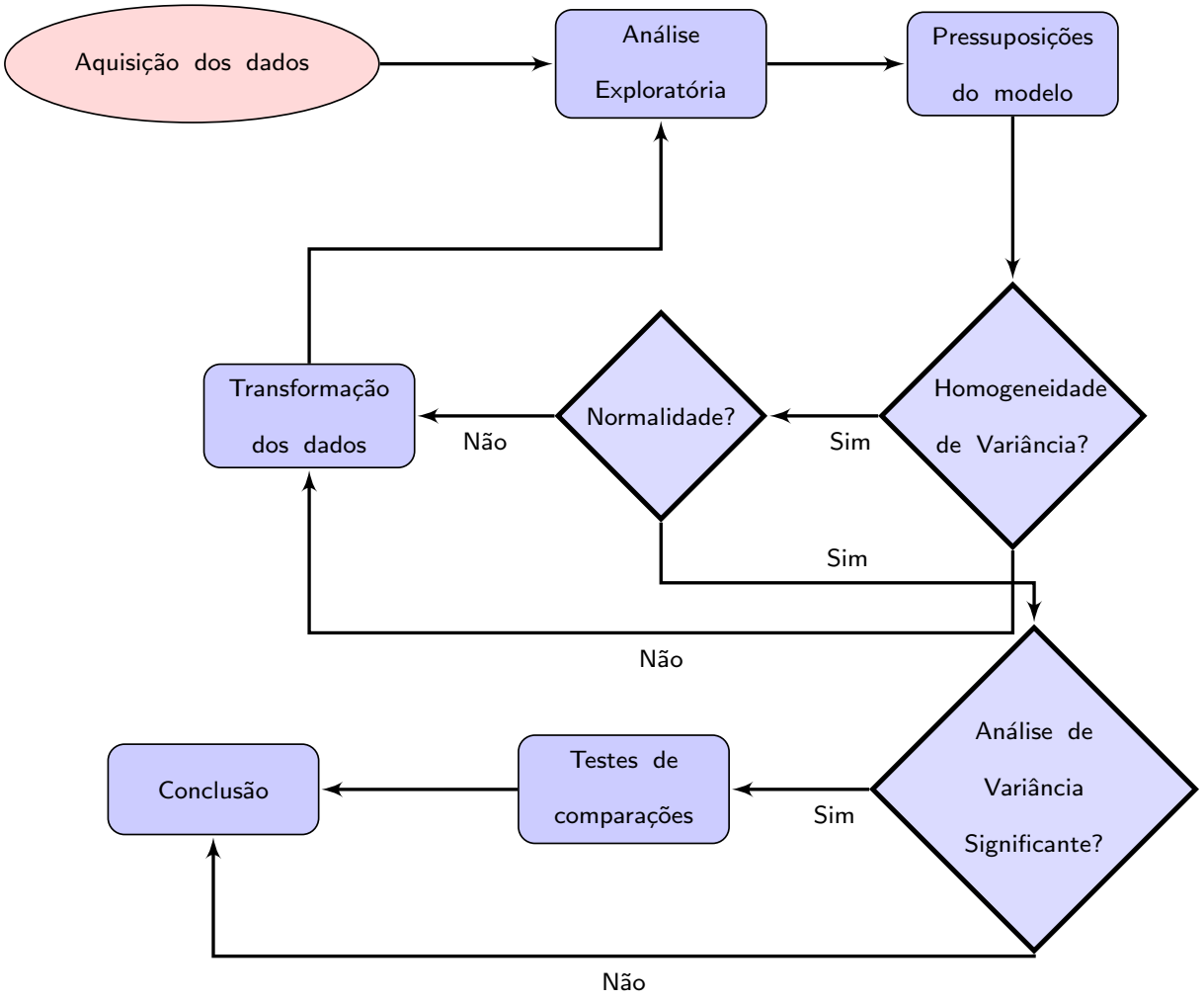
Interpretação do Teste F

```
anova(mod1)
```

```
## Analysis of Variance Table
##
## Response: yt
##           Df Sum Sq Mean Sq F value   Pr(>F)
## trat       3 3.4238 1.14127  66.468 2.99e-09 ***
## Residuals 16 0.2747 0.01717
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Conclusão:

- Existe diferença em pelo menos um contraste de médias (valor $-$ $p = 2,99e-09$ nível de significância $\alpha = 5\%$)


Aquisição dos dados
Análise Exploratória
Pressuposições do modelo
Homogeneidade de Variância?
Sim
Normalidade?
Não
Transformação dos dados
Sim
Não
Análise de Variância Significante?
Sim
Testes de comparações
Conclusão
Não

More on Residuals

For the raw residuals we have

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\theta} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

From the properties of $\mathbf{H}$ ($\mathbf{H}^2 = \mathbf{H}$) it is easy to show that

$\text{var}(e_i) = \sigma^2(1 - h_{ii})$ and $\text{cov}(e_i, e_j) = -\sigma^2 hij \quad i \neq j$

Standardized Residuals

$$r_i = \frac{e_i}{\sqrt{s^2(1 - h_{ii})}}$$

Studentized (Jackknife) Residuals

$$j_i = \frac{r_i}{\left(\frac{n-p-1-r_i^2}{n-p-2}\right)^{1/2}}$$

leverage values – h_{ii} (`%lv`)

$$\mathbf{H} = (h_{ij}) = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$$

Teste de comparações múltiplas

- O teste F (ANOVA) foi significativo?

Não?

- Analise foi finalizada.

Sim?

- É necessário dar continuidade à análise utilizando um teste de comparações múltiplas.

Continuidade da análise

- Quando o teste F é significativo implica que pelo menos um contraste de médias foi significativo.
- Qual a definição de contrastes?
- Quantos contrastes existem?
- Quais foram significativos?

```
anova(mod.lm)
```

```
## Analysis of Variance Table
##
## Response: Etch.rate
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Potencia   3  66871 22290.2  66.797 2.883e-09 ***
## Residuals 16   5339   333.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Rejeita-se H_0 no nível de significância $\alpha = 5\%$

Teste de comparação de médias

Uma variedade de métodos de comparação múltipla estão di

- Teste de Tukey
- Teste de Scheffé
- Teste de Bonferroni
- Teste de Duncan
- Teste de Dunnet

O que é um contraste?

Um contraste é uma combinação linear dos parâmetros da forma:

$$\Gamma = \sum_{i=1}^{a} c_i \mu_i$$

em que as constantes $c_i, i = 1, \ldots, a$ satisfazem

$$\sum_{i=1}^{a} c_i = 0.$$

$H_0 : \mu_3 = \mu_4$ ou equivalentemente, $H_0 : \mu_3 - \mu_4 = 0$

$$c_1 = c_2 = 0, c_3 = 1, c_4 = -1$$

$H_0 : \mu_1 + \mu_2 = \mu_3 + \mu_4$ ou equivalentemente;
$H_0 : \mu_1 + \mu_2 - \mu_3 - \mu_4 = 0$

$$c_1 = c_2 = 1, c_3 = c_4 = -1$$

Teste de Tukey

- é um teste de comparação de médias duas a duas
- Se o número de tratamentos é igual a "a" tem-se

$$m = \frac{a(a-1)}{2}$$

comparações de médias duas a duas.

- $a = 4$, $trat = Potencia = 160, 180, 200, 220$
- $m = \frac{a(a-1)}{2} = \frac{4(4-1)}{2} = 6$ comparações, a saber:

	μ_{160}	μ_{180}	μ_{200}
μ_{180}	$\lvert\mu_{180} - \mu_{160}\rvert$		
μ_{200}	$\lvert\mu_{200} - \mu_{160}\rvert$	$\lvert\mu_{200} - \mu_{180}\rvert$	
μ_{220}	$\lvert\mu_{220} - \mu_{160}\rvert$	$\lvert\mu_{220} - \mu_{180}\rvert$	$\lvert\mu_{220} - \mu_{200}\rvert$

O Teste proposto por Tukey (1953) é também conhecido como:

- teste de Tukey da diferença honestamente significativa (honestly significant difference)(HSD) e,
- teste de Tukey da diferença totalmente significativa (wholly significant difference)(WSD).
- As hipóteses a serem testadas são:

$$\begin{cases} H_0: & \mu_i & = & \mu_j \\ H_1: & \mu_i & \neq & \mu_j \end{cases}, \quad i > j$$

Dados balanceados

O procedimento de Tukey utiliza a estatística da amplitude studentizada

$$Q = \frac{|\bar{Y}_i - \bar{Y}_j|}{\sqrt{\frac{QMRes}{k}}}$$

declara duas médias significativamente diferentes se

$$|\bar{y}_{i\cdot} - \bar{y}_{i\cdot}| > TSD = DMS$$

$$DMS = TSD = q_{\alpha}(a, n-a)\sqrt{\frac{QMRes}{k}},$$

$n - a$ é o número de graus de liberdade do resíduo.
n é o número de unidades experimentais,
e $q(a, n-a)$ é a amplitude studentizada e a é o número de tratamentos

Leia (é interessante!)

- https://www.youtube.com/watch?v=5iPFKW4zC1I
- https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1238&context=pare
- https://scholarworks.umass.edu/pare/vol15/iss1/12/
- https://www.jstor.org/stable/pdf/1169991.pdf
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7714623/
- https://edisciplinas.usp.br/pluginfile.php/1671633/mod_resource/content/2/Aula_05_2016.pdf
- https://rcompanion.org/handbook/I_12.html
- https://docs.ufpr.br/~aanjos/CE213/ce213/node3.html
- https://towardsdatascience.com/beginner-explanation-for-data-transformation-9add3102f3bf