Transformação de dados

Universidade Estadual de Santa Cruz

Ivan Bezerra Allaman

CRONOGRAMA

1. Introdução
2. Tipos de Transformações
3. Exemplos

Introdução

- Quando pelo menos um dos pressupostos da **anova** são violados, uma transformação angular se faz necessário.
- É importante ressaltar que a transformação não garante cumprimento dos pressupostos, e portanto, deve-se fazer nova análise de resíduo para checagem.
- Se nenhuma transformação atender os pressupostos, outra metodologia dever ser utilizada, como por exemplo, a análise não-paramétrica.
- A transformação só serve para fazer inferências, ou seja, testar hipóteses. Logo, a apresentação dos resultados final deve ser feita com a variável na escala original.

- Se for de interesse apresentar o erro padrão da média, não se deve fazer a transformação de volta como é no caso das médias. Neste caso, uma aproximação razoável seria tirar a raiz quadrada da média das variancias dos tratamentos pelo número de repetições (r).

$$EPM = \sqrt{\frac{\frac{\sum_{i=1}^{k} S^2_{trat_A} + S^2_{trat_B} + \cdots + S^2_{trat_i}}{k}}{r}}$$

Tipos de transformações

Transformação raiz quadrada

- É utilizada quando os dados são de contagem e cujo a lei de distribuição é a poisson.
- Como exemplos podemos citar: número de ovos por m^2 em uma pastagem, número de carrapatos em uma determinada área do corpo do animal, número de vezes que um animal vai ao cocho, etc.
- Seja Y uma variável aleatória mensurada na unidade de observação, tem-se a seguinte transformação:

$$Y' = \sqrt{Y}$$

- Se há zeros nos dados, então uma prática seria:

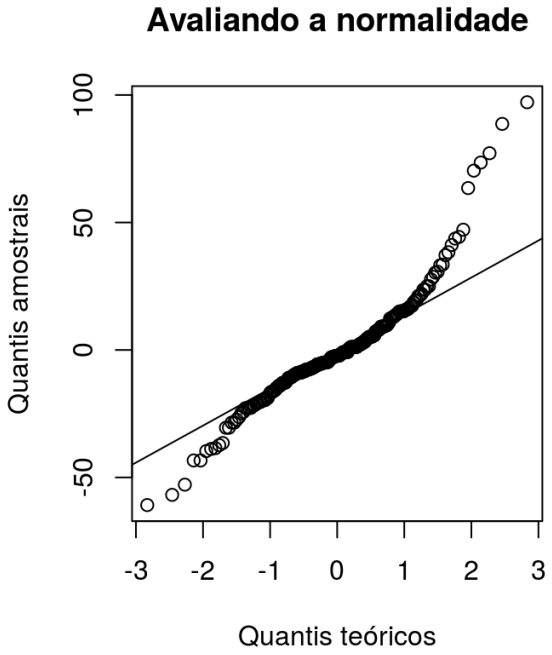
$$Y' = \sqrt{Y + 0,5}$$

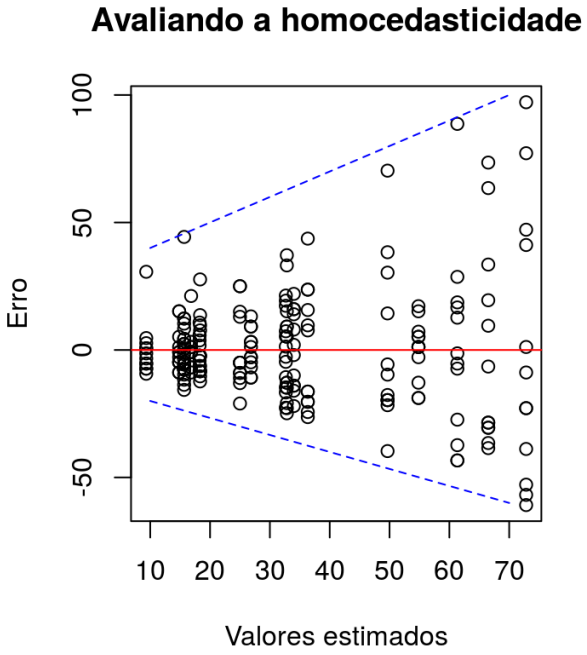
Aplicação

1. Foi feito um estudo com o intuito de avaliar em campo o uso integrado de fungos entomopatogênicos e acaricidas químicos para o controle do carrapato bovino Rhipicephalus (Boophilus) microplus. O delineamento foi o inteiramente ao acaso em uma esquema fatorial 3x6. O número de carrapatos no lado esquerdo do animal foi avaliado e se encontra no link: http://nbcgib.uesc.br/lec/download/R/dados/poly.txt

aplic	grupo	ncarra
1	1	18
1	1	40
1	1	34
1	1	50
1	1	30
1	1	10
1	1	16
1	1	54
1	1	28

Após uma análise de variância, fazendo uma análise de resíduo tem-se o seguintes gráficos:


Avaliando a normalidade
Quantis amostrais
100
50
0
-50
-3
-2
-1
0
1
2
3
Quantis teóricos

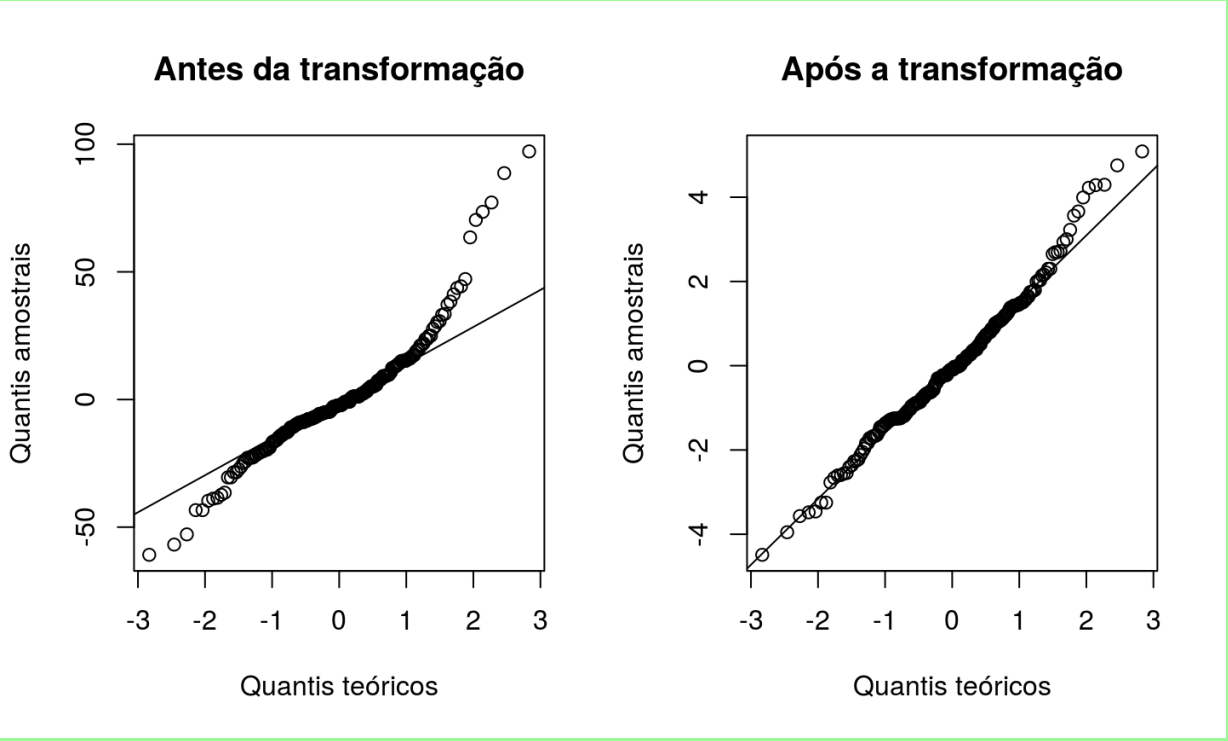

Avaliando a homocedasticidade
Erro
100
50
0
-50
10
20
30
40
50
60
70
Valores estimados

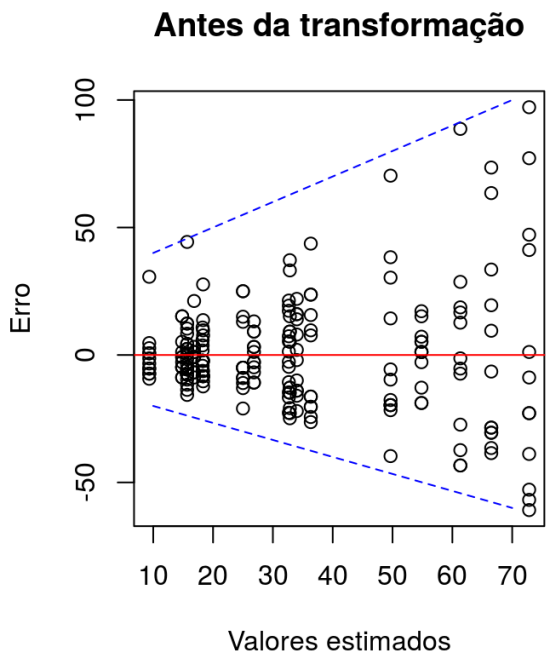
aplic	grupo	ncarra	newvar
1	1	18	4,24
1	1	40	6,32
1	1	34	5,83
1	1	50	7,07
1	1	30	5,48
1	1	10	3,16
1	1	16	4,00
1	1	54	7,35
1	1	28	5,29

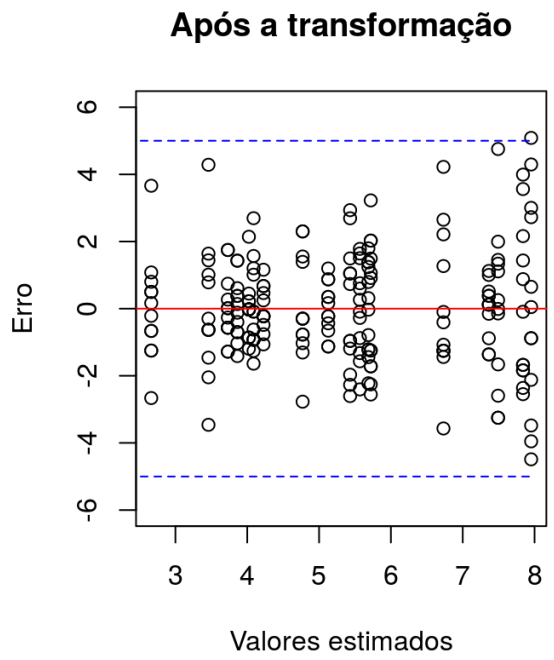
A título didático, se pegarmos o primeiro número dos dados apresentados tem-se:

$$\sqrt{18} = 4,24$$

Portanto, tem-se os gráficos antes e após a transformação.




Antes da transformação
Erro
100
50
0
-50
10
20
30
40
50
60
70
Valores estimados


Após a transformação
Erro
6
4
2
0
-2
-4
-6
3
4
5
6
7
8
Valores estimados

Transformação logarítmica

- Geralmente quando os dados são contínuos e mesmo assim não se aderem a distribuição normal, então a transformação logarítmica pode ser útil.
- Neste caso não importa se a base é 10 ou *e* (logarítmo neperiano).
- Portanto, seja Y uma variável aleatória mensurada na unidade de observação, tem-se a seguinte transformação:

$$Y' = \log(Y) \quad ou \quad Y' = \ln(Y)$$

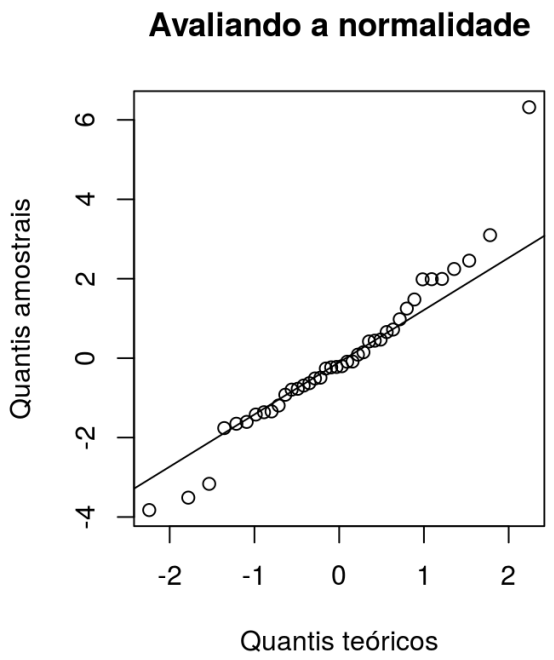
- Se há zeros nos dados, também podemos fazer:

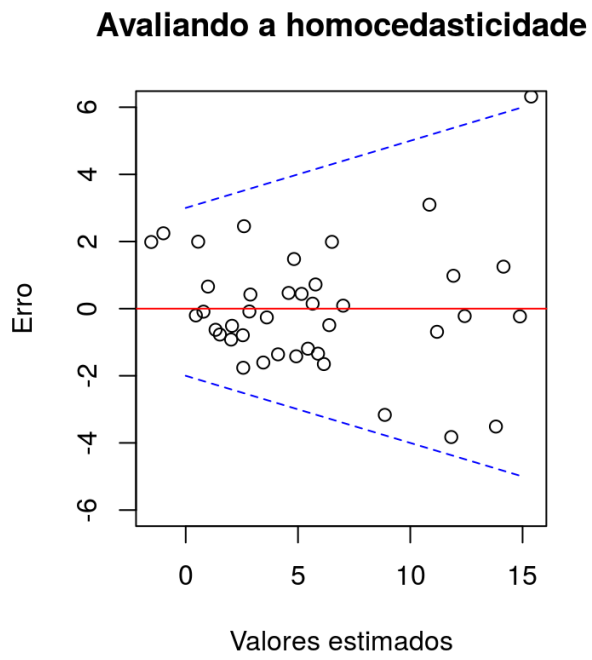
$$Y' = \log(Y + 0,5) \quad ou \quad Y' = \ln(Y + 0,5)$$

Aplicação

2. Um experimento foi feito para avaliar a época de plantio de milho e o tipo de plantio (cova e linha) em um delineamento em blocos ao acaso em esquema fatorial 4x2 (4 épocas e 2 tipos de plantio). A variável analisada foi a produção em pêso de grãos. Os dados estão disponíveis no link:
   http://nbcgib.uesc.br/lec/download/R/dados/con.txt

   Fazendo uma análise de resíduos após análise de variância tem-se:


Avaliando a normalidade
Quantis amostrais
Quantis teóricos
-4
-2
0
2
4
6
-2
-1
0
1
2

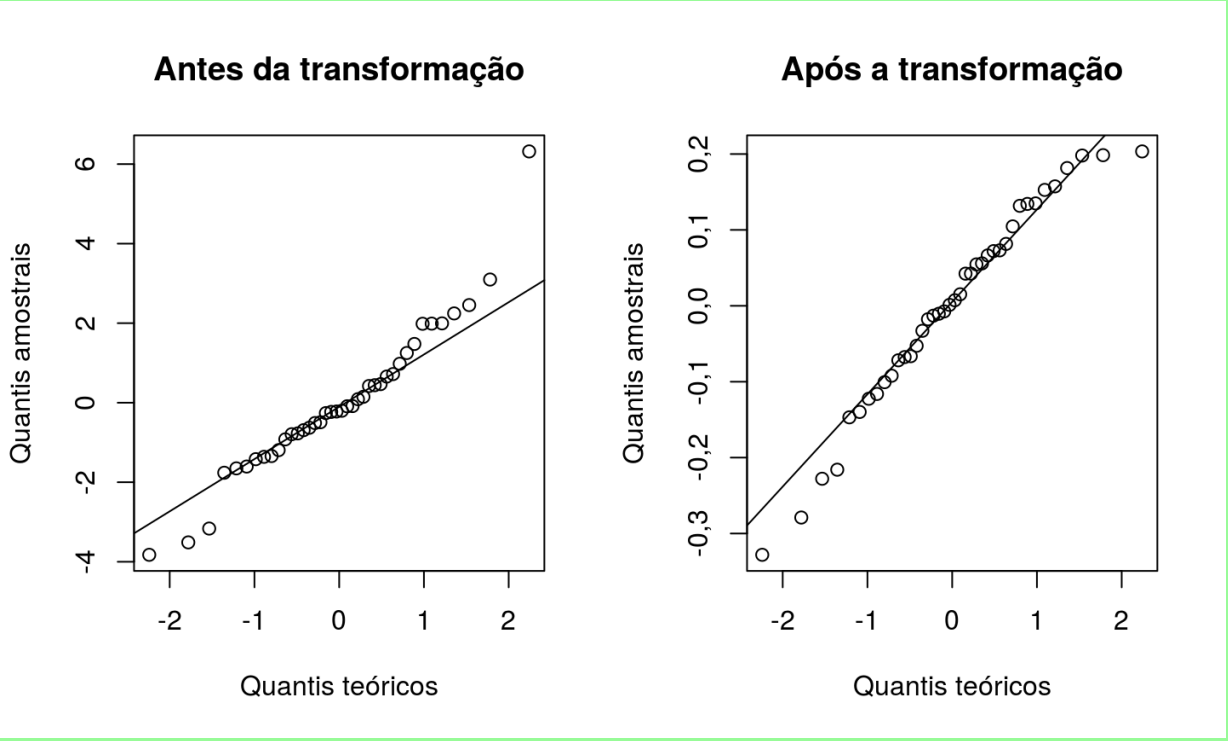

Avaliando a homocedasticidade
Erro
Valores estimados
-6
-4
-2
0
2
4
6
0
5
10
15

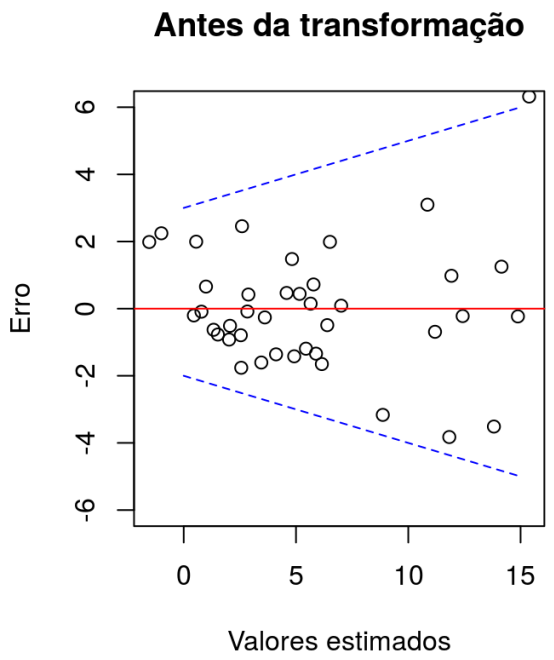
bloco	epoca	plantio	resp	newvar
1	1	cova	6,5	0,8129
1	1	linha	15,4	1,1875
1	2	cova	5,6	0,7482
1	2	linha	10,5	1,0212
1	3	cova	3,3	0,5185
1	3	linha	3,5	0,5441
1	4	cova	0,7	-0,1549
1	4	linha	0,7	-0,1549
2	1	cova	7,1	0,8513

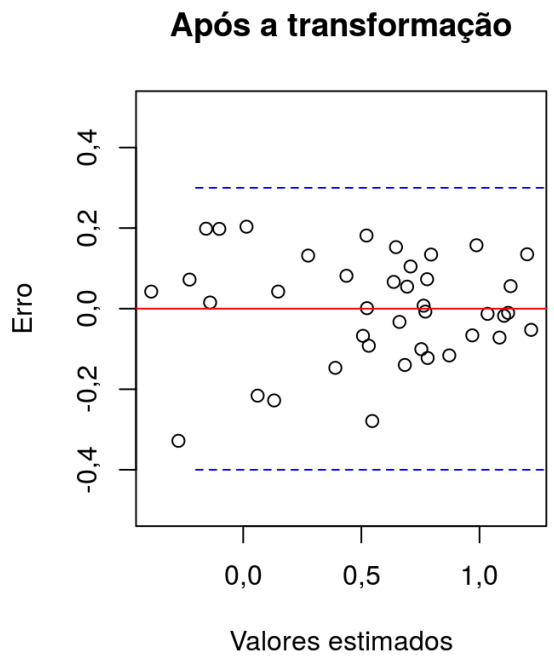
A título didático, se pegarmos o primeiro número dos dados apresentados tem-se:

$$\log 6,5 = 0,8129$$

Portanto, tem-se os gráficos antes e após a transformação.




Antes da transformação
Erro
6
4
2
0
-2
-4
-6
0
5
10
15
Valores estimados


Após a transformação
Erro
0,4
0,2
0,0
-0,2
-0,4
0,0
0,5
1,0
Valores estimados

Transformação arcoseno raiz quadrada

- A função arcoseno é a inversa da função seno com domínio no intervalo $[-\pi/2, \pi/2]$ e imagem no intervalo $[-1, -1]$
- Quando os dados são binominais, proporções ou percentagens podemos utilizar a transformação em questão.
- Logo, seja Y uma variável aleatória mensurada na unidade de observação, tem-se a seguinte transformação:

$$Y' = \operatorname{arcsin} \sqrt{Y} (\text{em termos decimais})$$

ou

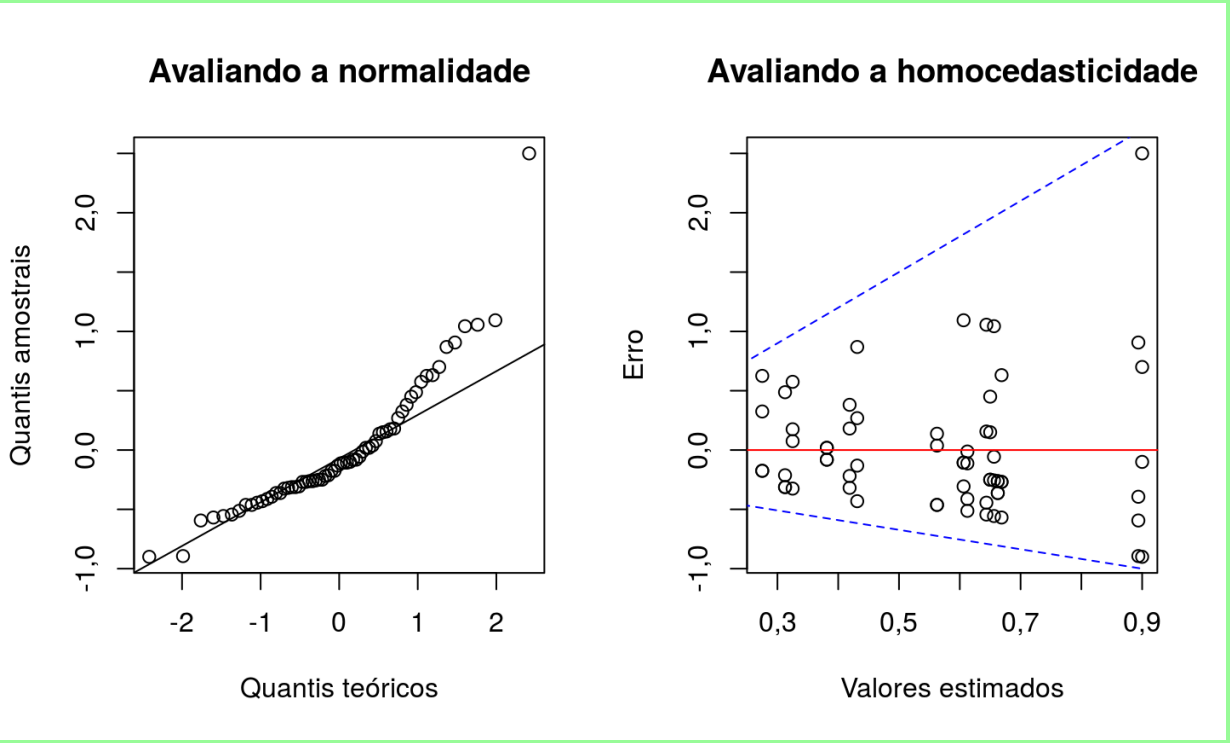
$$Y' = \operatorname{arcsin} \sqrt{Y/100} (\text{em termos de percentagem})$$

Aplicação

3. Foi realizado um estudo para avaliar a utilização do ácido docosahexaenoico (DHA) e Trolox no resfriamento e na congelação de sêmen de garanhão. Uma das variáveis analisadas foi a percentagem de sêmen hiperativo que segue no link: http://nbcgib.uesc.br/lec/download/R/dados/cris.txt

gar	rep	trat	hiperativo
G1	P1	T1	0,4
G2	P1	T1	0,4
G3	P1	T1	0,4
G4	P1	T1	1,3
G1	P2	T1	0,1

Após a análise de variância tem-se a seguinte análise de resíduo:

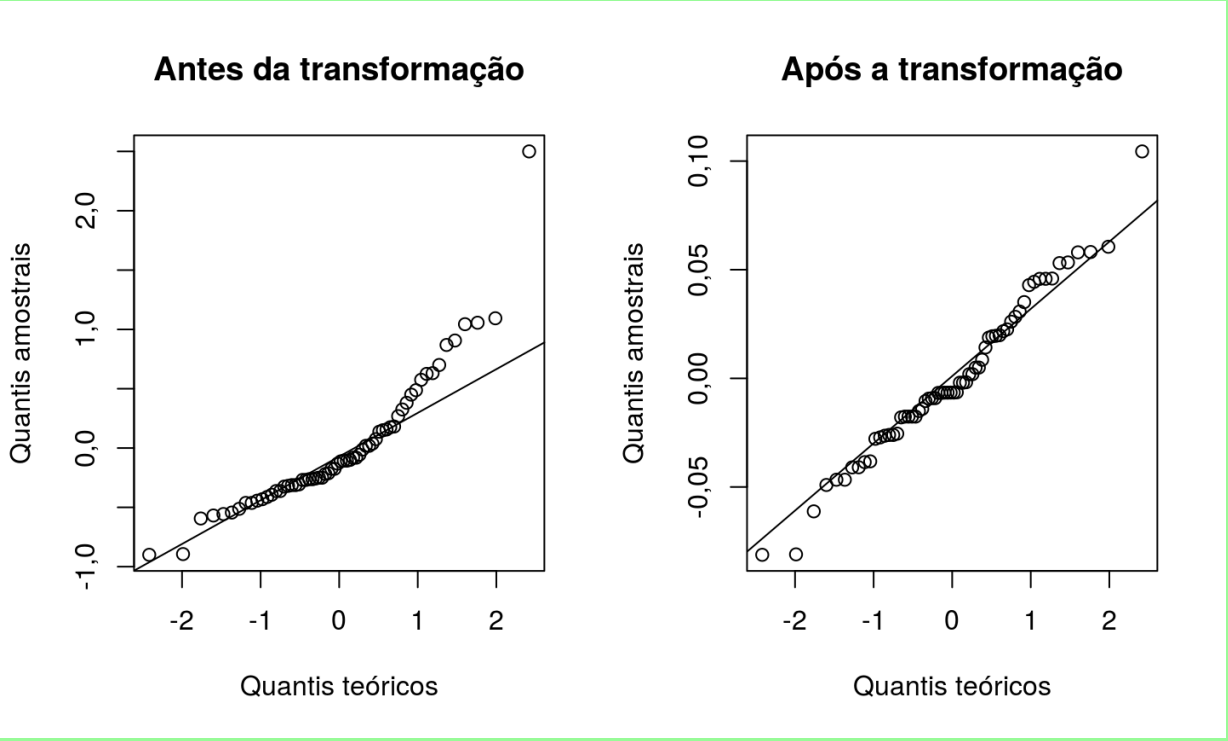


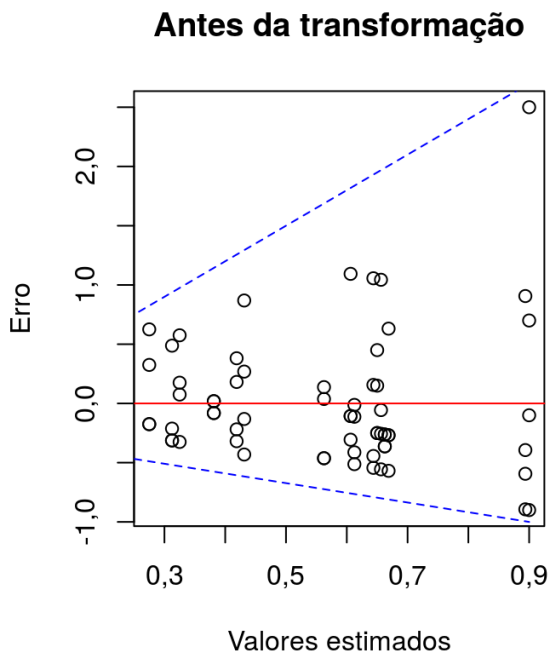
gar	rep	trat	hiperativo	newvar
G1	P1	T1	0,4	0,063
G2	P1	T1	0,4	0,063
G3	P1	T1	0,4	0,063
G4	P1	T1	1,3	0,114
G1	P2	T1	0,1	0,032

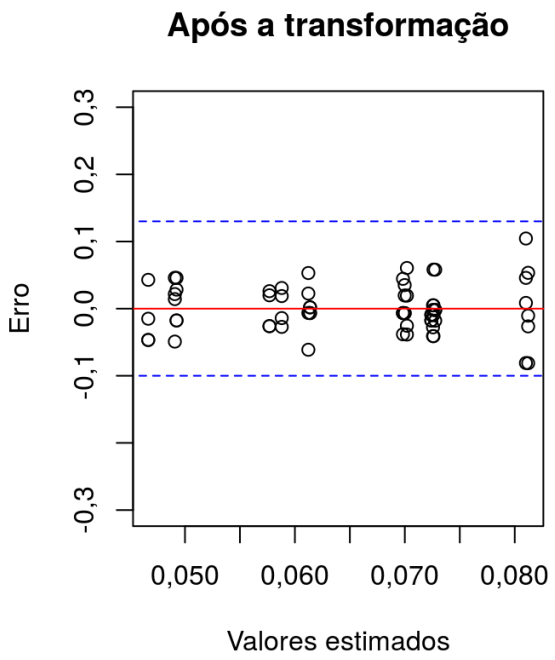
A título didático, se pegarmos o primeiro número dos dados apresentados tem-se:

$$\arcsin\sqrt{0,4/100} = 0,063$$

Portanto, tem-se os gráficos antes e após a transformação.




Antes da transformação
Erro
-1,0
0,0
1,0
2,0
0,3
0,5
0,7
0,9
Valores estimados


Após a transformação
Erro
-0,3
-0,1
0,0
0,1
0,2
0,3
0,050
0,060
0,070
0,080
Valores estimados

Transformação boxcox

- A mais recente das transformações é aplicada a qualquer tipo de variável e geralmente resolve a grande maioria dos casos, principalmente naqueles em que as transformações usuais (raiz quadrada, logarítmica, etc.) não funcionam.
- Dado que Y é uma variável aleatória discreta ou contínua, então a transformação é feita como:

$$Y' = \begin{cases} \ln(Y) & \text{se } \lambda = 0 \\ \frac{Y^{\lambda}-1}{\lambda} & \text{se } \lambda \neq 0 \end{cases}$$

- O parâmetro λ é o que queremos encontrar, pois ele é o que maximiza o logarítmo da função de verossimilhança para uma distribuição normal.

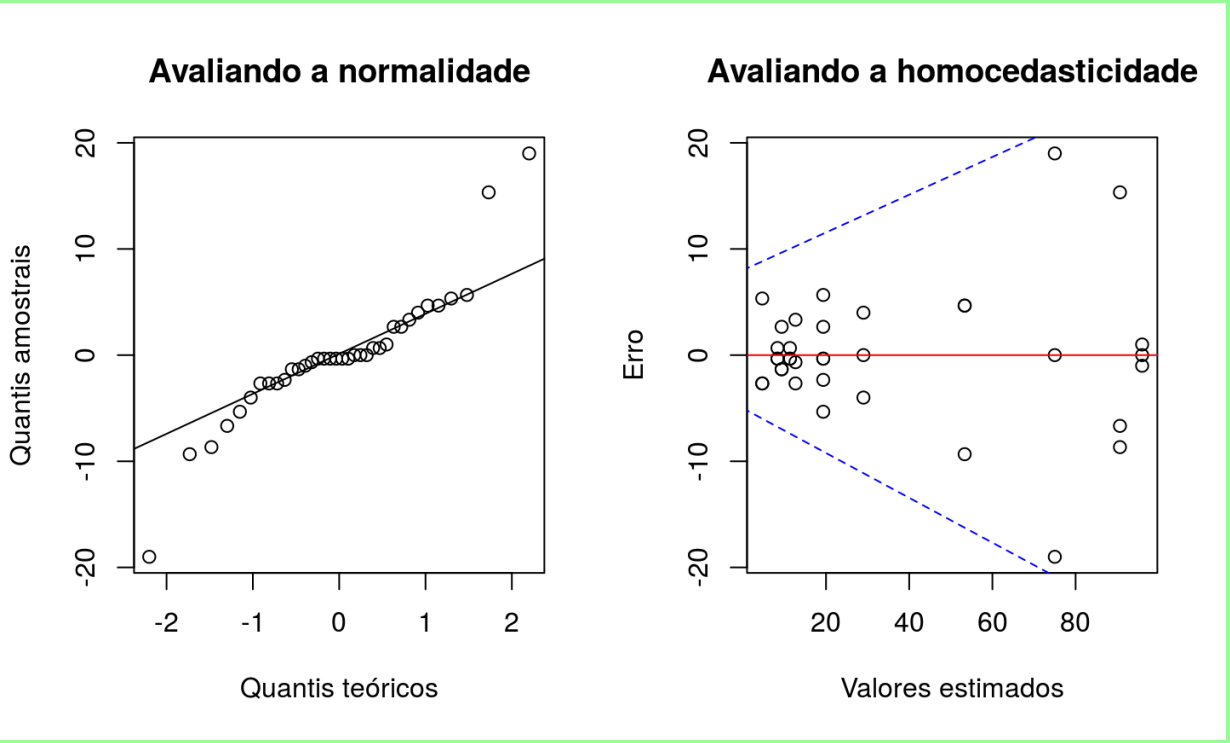
- Para encontrar o valor de λ é necessário o auxílio de um software. No caso do `R`, podemos usar a função `boxcox` do pacote `MASS`.
- Como o método trabalha com o logarítmo da função de verossimilhança, caso tenha zeros nos dados, também podemos acrescentar 0,5 aos dados para que seja possível a transformação.

Aplicação

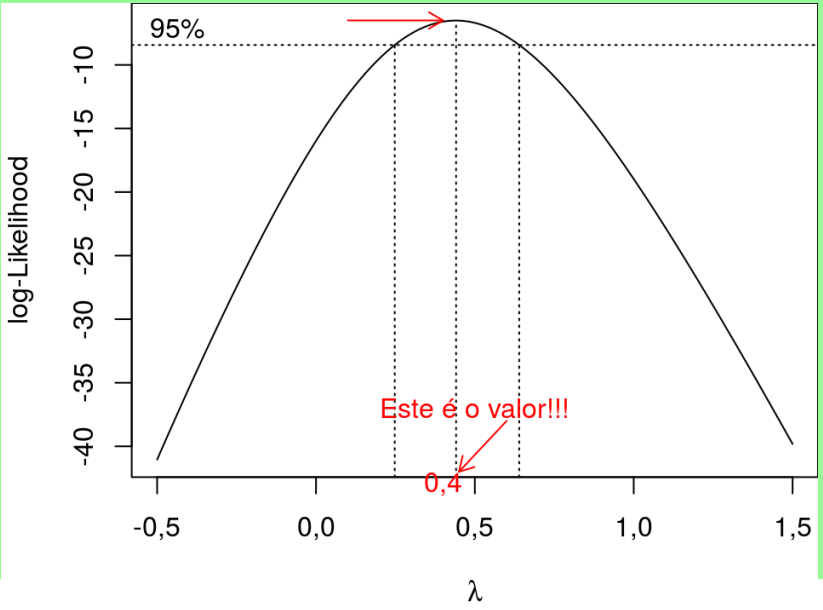
4. Foi realizado um estudo com o intuito de avaliar o impacto de diferentes densidades e tempo de transporte em uma determinada espécie de peixe. Uma das variáveis analisadas foi o cortisol. Os dados estão disponíveis no seguinte link: http://nbcgib.uesc.br/lec/download/R/dados/marc.txt

tempo	den	cortisol
0	98	56
0	98	75
0	98	94
0	146	96
0	146	95
0	146	97

Após uma análise de variância segue a análise de resíduos:



Considerando que para o referido exemplo, o delineamento foi o inteiramente ao acaso em um esquema fatorial 3x4 e que houve violação dos pressupostos, deveremos examinar o seguinte gráfico considerando o modelo estatístico do delineamento citado.



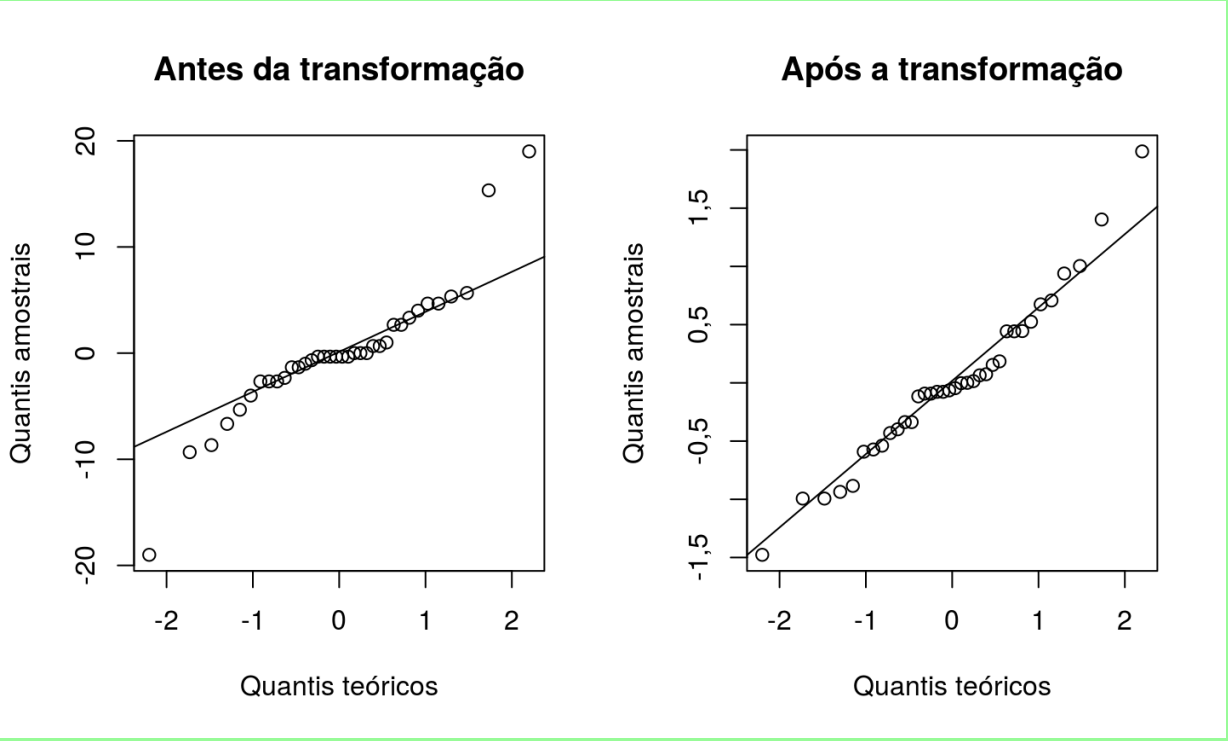
Então temos:

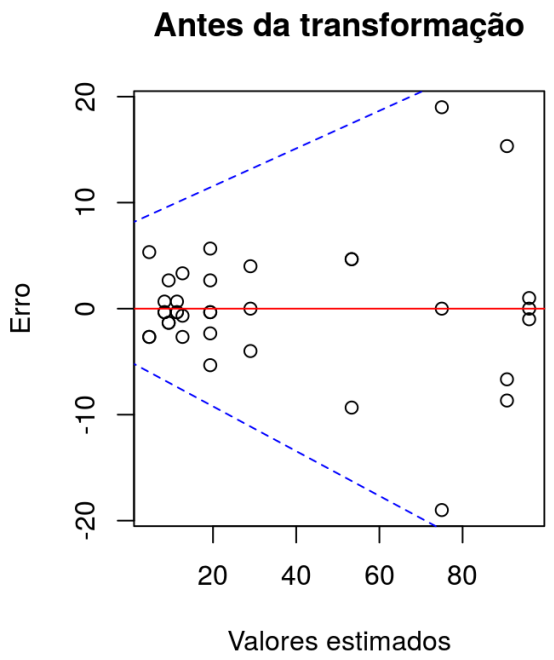
tempo	den	cortisol	cortbox
0	98	56	10,0088
0	98	75	11,5593
0	98	94	12,8883
0	146	96	13,0185
0	146	95	12,9536
0	146	97	13,0829

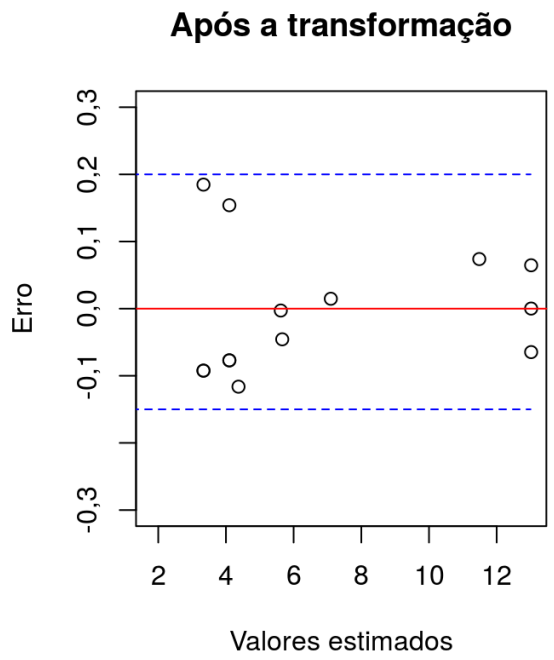
Didadicamente, pegando o primeiro valor dos dados temos:

$$(56^{0,4} - 1)/0,4 = 10,0088$$

Portanto, tem-se os gráficos antes e após a transformação.




Antes da transformação
Erro
20
10
0
-10
-20
20
40
60
80
Valores estimados


Após a transformação
Erro
0,3
0,2
0,1
0,0
-0,1
-0,3
2
4
6
8
10
12
Valores estimados