Modelo linear generalizado

Distribuição gama

2022

Este exemplo se refere a um experimento em que a resistência (em horas) de um determinado tipo de vidro foi avaliada segundo quatro valores de voltagem (200, 250, 300 e 350, em kilovolts) e duas temperaturas (170 e 180, em graus Celsius). O principal interesse consiste em comparar as resistências médias em relação aos níveis de voltagem e temperatura.

Uma descrição do problema encontra-se no livro do Prof. G. A. Paula, pag. 175 (http://www.ime.usp.br/~giapaula/texto_2013.pdf). Os dados estão no arquivo http://www.ime.usp.br/~giapaula/vidros.dat.

A linguagem R é utilizada no exemplo.

A função `read.table` lê uma tabela de dados e gera uma folha de dados (*data frame*). As variáveis `volt` (voltagem) e `temp` (temperatura) são transformadas em qualitativas (da classe *factor*).

```
# Separador decimal nos resultados: ","
options(OutDec = ",")
```

```
dados <- read.table("vidros.dat")
colnames(dados) <- c("resist", "volt", "temp")
```

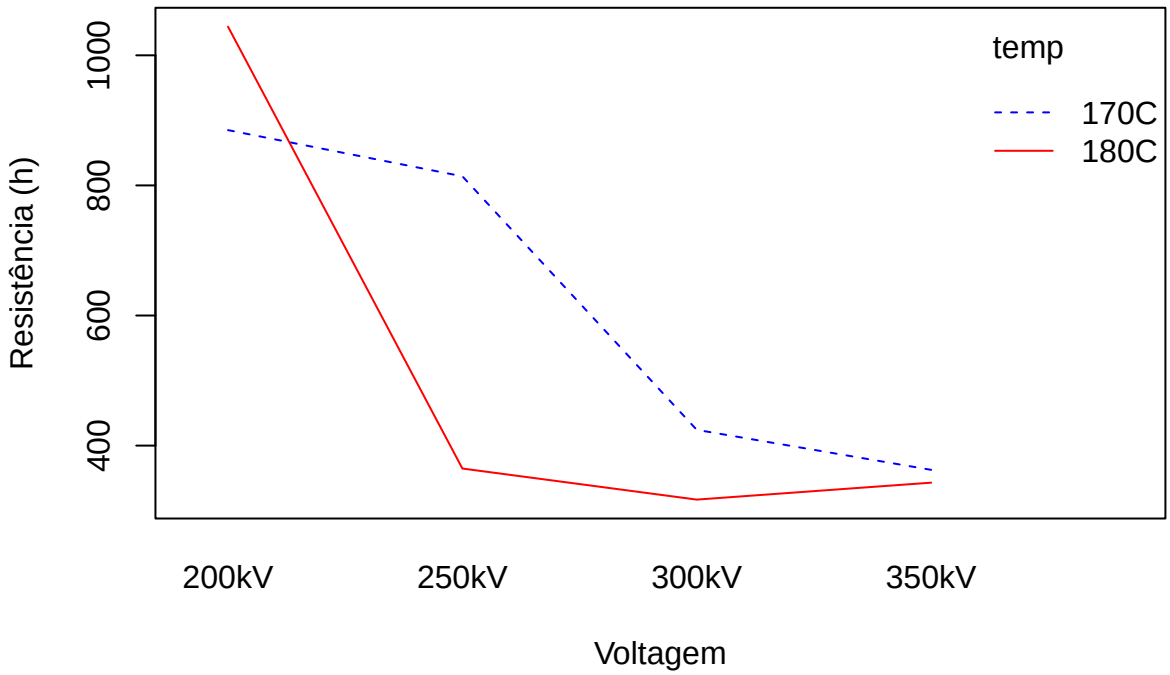
```
cat("\n Tamanho da amostra (n):", n <- nrow(dados))
```

```
##
##  Tamanho da amostra (n): 32
```

```
dados$volt <- factor(dados$volt, labels = c("200kV", "250kV", "300kV", "350kV"))
dados$temp <- factor(dados$temp, labels = c("170C", "180C"))
```

O gráfico de interação abaixo indica uma redução da resistência média (por que "média"?) à medida que aumenta o nível de voltagem. Além disso, há indicação de interação entre voltagem e temperatura (por quê?).

```
# Gráfico de interação
with(dados, interaction.plot(volt, temp, resist, xlab = "Voltagem",
   ylab = "Resistência (h)", col = c("blue", "red")))
```



Nota 1 Você recomendaria um gráfico de caixas (*box plot*) para cada combinação de voltagem e temperatura?

Nota 2. Represente os dados em um gráfico de pontos.

Como a variável resposta é positiva, ajustamos um modelo gama. O resíduo de quantil, implementado na função `gamlss` do pacote `gamlss`, será utilizado para verificar se o modelo faz um bom ajuste aos dados. Neste pacote, a função densidade de probabilidade da distribuição gama é dada por, para $y > 0$,

$$f(y;\mu,\sigma) = \frac{y^{1/\sigma^2-1}\exp\left(-\dfrac{y}{\sigma^2\mu}\right)}{(\sigma^2\mu)^{1/\sigma^2}\Gamma(1/\sigma^2)}, \tag{1}$$

em que $\mu > 0$ e $\sigma > 0$ são as letras usadas no pacote para os dois primeiros parâmetros (de um máximo de 4). Temos que

$$E(Y) = \mu \quad \text{e} \quad \text{Var}(Y) = \sigma^2\mu^2, \tag{2}$$

de modo que o coeficiente de variação é

$$\frac{\sqrt{\text{Var}(Y)}}{E(Y)} = \frac{\sqrt{\sigma^2\mu^2}}{\mu} = \sigma. \tag{3}$$

Dizemos que o modelo gama é um modelo para dados com coeficiente de variação constante (ou seja, o coeficiente de variação não depende da média μ), lembrando que no MLG gama em uma amostra a média μ_i varia com as observações via preditor linear $\eta_i = \boldsymbol{x}_i^\top\boldsymbol{\beta}$ e função de ligação g, em que $\mu_i = g^{-1}(\eta_i)$, $i = 1,\ldots,n$.

Nota 3. Na notação da Tabela 2, pag. 13, do livro Demétrio, C. G. B. (2002), *Modelos Lineares Generalizados em Experimentação Agronômica*, ESALQ (https://docs.ufpr.br/~niveam/micro%20da%20sala/bom/Apostila%20de%20MLG.pdf) temos $\nu = 1/\sigma^2$ e $a(\phi) = 1/\nu = \sigma^2$.

O número de observações e algumas estatísticas descritivas (média, desvio padrão e coeficiente de variação) da variável resposta são apresentadas para cada nível de voltagem e temperatura.

```
myfunction <- function(x) {
   return(c(length(x), mean(x), sd(x), sd(x) / mean(x)))
}
by(dados$resist, dados[, c("volt", "temp")], FUN = myfunction)
```

```
## volt: 200kV
## temp: 170C
## [1]   4,0000000 885,0000000 311,1944730   0,3516322
## ------------------------------------------------------------
## volt: 250kV
## temp: 170C
## [1]   4,0000000 814,0000000 229,6490075   0,2821241
## ------------------------------------------------------------
## volt: 300kV
## temp: 170C
## [1]   4,0000000 424,2500000 147,8769195   0,3485608
## ------------------------------------------------------------
## volt: 350kV
## temp: 170C
## [1]   4,0000000 362,7500000 155,9174461   0,4298207
## ------------------------------------------------------------
## volt: 200kV
## temp: 180C
## [1] 4,000000e+00 1,044000e+03 5,760787e+01 5,517995e-02
## ------------------------------------------------------------
## volt: 250kV
## temp: 180C
## [1]   4,0000000 364,7500000 121,7439252   0,3337736
## ------------------------------------------------------------
## volt: 300kV
## temp: 180C
## [1]   4,0000000 317,0000000  57,6599225   0,1818925
## ------------------------------------------------------------
## volt: 350kV
## temp: 180C
## [1]   4,0000000 343,0000000 118,0621305   0,3442045
```

Pelos resultados acima, o coeficiente de variação amostral varia de 0,055 a 0,43, sendo que dos oito grupos de observações, em apenas um o valor é igual ao mínimo. Ressalte-se que o número de observações em cada grupo é apenas 4. Prosseguimos com o modelo gama. Será utilizada a função de ligação identidade. O primeiro modelo ajustado inclui os efeitos principais e a interação entre voltagem e temperatura.

```
## Modelos
m1 <- glm(resist ~ volt * temp, family = Gamma(link = "identity"), data = dados)
summary(m1)
```

```
##
## Call:
## glm(formula = resist ~ volt * temp, family = Gamma(link = "identity"),
##     data = dados)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0,62791  -0,28368   0,01998   0,22074   0,52524
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)          885,0      137,8   6,424 1,21e-06 ***
## volt250kV            -71,0      187,2  -0,379  0,70780
## volt300kV           -460,8      152,8  -3,016  0,00598 **
```

```
## volt350kV            -522,2      148,9  -3,508  0,00181 **
## temp180C              159,0      213,1   0,746  0,46275
## volt250kV:temp180C   -608,2      254,3  -2,392  0,02496 *
## volt300kV:temp180C   -266,2      228,4  -1,165  0,25529
## volt350kV:temp180C   -178,8      226,8  -0,788  0,43831
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0,09693636)
##
##     Null deviance: 9,4487  on 31  degrees of freedom
## Residual deviance: 2,4277  on 24  degrees of freedom
## AIC: 423,29
##
## Number of Fisher Scoring iterations: 3
```

Nota 4. Levando em conta a função de ligação adotada, os sinais das estimativas estão compatíveis com o gráfico de interação?

Nota 5. A estimativa do parâmetro de dispersão ϕ é obtida de $\widehat{\phi} = X^2/(n-p)$, sendo que X^2 denota a estatística X^2 generalizada de Pearson.

Apenas uma interação apresenta coeficiente significativo a um nível de 5% (mas não a 1%). Ajustamos um modelo sem a interação entre voltagem e temperatura.

```
m2 <- update(m1, . ~ . -volt:temp)
summary(m2)
```

```
##
## Call:
## glm(formula = resist ~ volt + temp, family = Gamma(link = "identity"),
##     data = dados)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0,75440  -0,24250   0,01167   0,15293   0,63561
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1039,94     122,50   8,489 4,21e-09 ***
## volt250kV    -426,49     135,61  -3,145  0,00402 **
## volt300kV    -608,81     126,21  -4,824 4,89e-05 ***
## volt350kV    -612,89     126,04  -4,863 4,40e-05 ***
## temp180C     -117,77      56,43  -2,087  0,04644 *
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0,117714)
##
##     Null deviance: 9,4487  on 31  degrees of freedom
## Residual deviance: 3,4306  on 27  degrees of freedom
## AIC: 428,53
##
## Number of Fisher Scoring iterations: 9
```

Os coeficientes são significativos a um nível de 5%, sendo que no caso da variável temperatura a situação é limítrofe. Em seguida os dois modelos são comparados. Na tabela ANODEV abaixo usamos `test = "F"`

porque no modelo gama o parâmetro de dispersão é estimado.

```
anova(m2, m1, test = "F")
```

```
## Analysis of Deviance Table
##
## Model 1: resist ~ volt + temp
## Model 2: resist ~ volt * temp
##   Resid. Df Resid. Dev Df Deviance      F  Pr(>F)
## 1        27     3,4306
## 2        24     2,4277  3   1,0029 3,4487 0,03246 *
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
```

A diferença entre os dois modelos (`m1` e `m2`) não é significativa a um nível de 1%. Por simplicidade, adotamos o modelo contendo apenas os efeitos principais (modelo `m2`), no qual todos os níveis de voltagem, em relação ao nível de referência (200 kV), apresentam diferença negativa, decrescente e significativa sobre a resistência média. Com base no modelo `m2`, ocorre uma redução de cerca de 426 horas na resistência média quando a voltagem muda de 200 para 250, mantida a temperatura constante.

Abaixo vemos que as estimativas de μ_i, $i = 1, \ldots, n$, são positivas, de modo que a função de ligação identidade para μ gerou valores admissíveis.

```
summary(fitted(m2))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   309,3   398,6   463,4   569,0   690,6  1039,9
```

Nota 6. Procure simplificar ainda mais, excluindo a variável temperatura.

Em seguida apresentamos o gráfico de envelope para o modelo mais simples (modelo sem a interação entre voltagem e temperatura). Os dois modelos (`m1` e `m2`) são ajustados com a função `gamlss`, que fornece estimativas de máxima verossimilhança para todos os parâmetros. A função de ligação identidade é adotada.

```
## Envelope m2
library(gamlss)
m1g <- gamlss(resist ~ volt * temp, family = GA(mu.link = "identity",
           sigma.link = "identity"), data = dados)
```

```
## GAMLSS-RS iteration 1: Global Deviance = 405,2909
## GAMLSS-RS iteration 2: Global Deviance = 405,2909
```

Nota 7. A desviância global (*global deviance*) é dada por $-2\ell(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}; \boldsymbol{y})$.

```
summary(m1g)
```

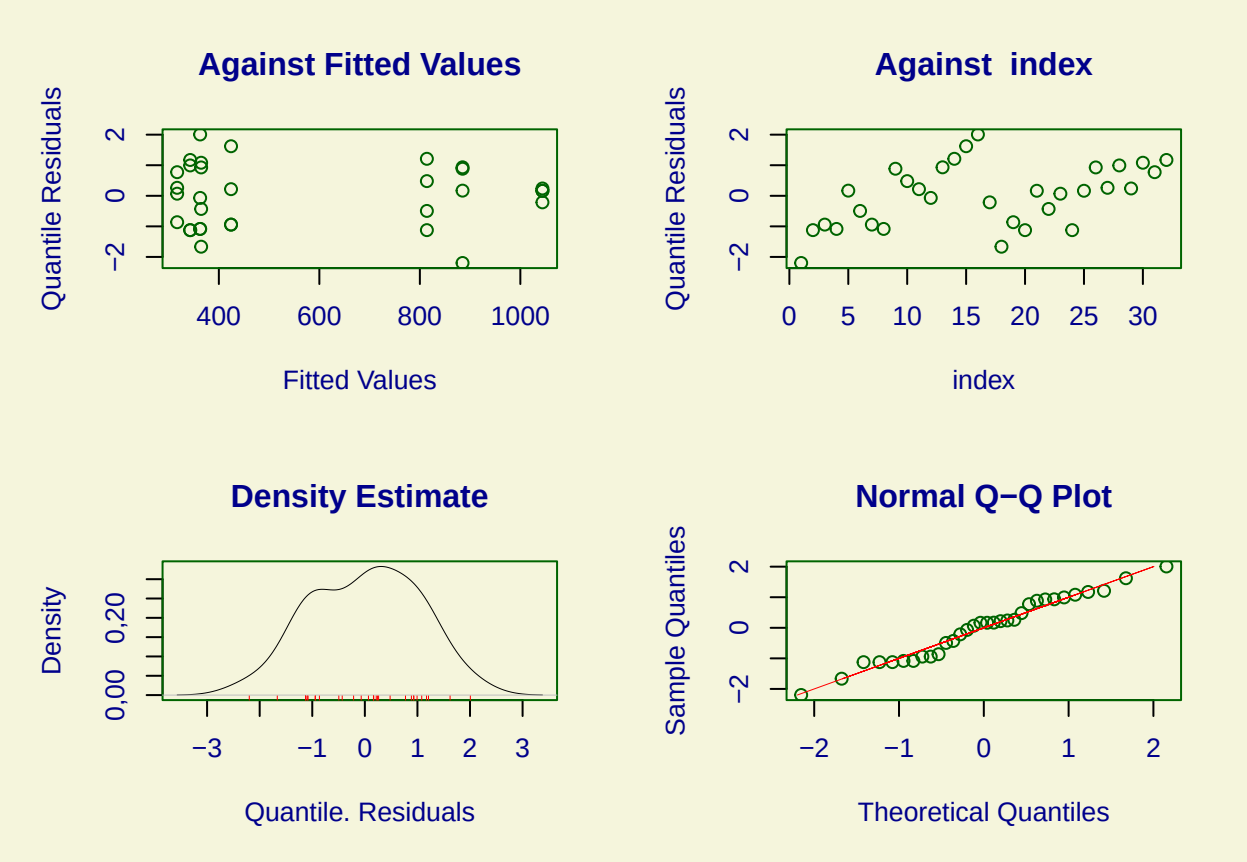
```
## ******************************************************************
## Family:  c("GA", "Gamma")
##
## Call:
## gamlss(formula = resist ~ volt * temp, family = GA(mu.link = "identity",
##     sigma.link = "identity"), data = dados)
##
## Fitting method: RS()
##
## ------------------------------------------------------------------
## Mu link function:  identity
## Mu Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)           885,0      121,0   7,312 1,94e-07 ***
## volt250kV             -71,0      164,5  -0,432 0,669986
## volt300kV            -460,8      134,2  -3,432 0,002273 **
## volt350kV            -522,2      130,8  -3,992 0,000574 ***
## temp180C              159,0      187,2   0,849 0,404405
## volt250kV:temp180C   -608,2      223,4  -2,722 0,012147 *
## volt300kV:temp180C   -266,2      200,7  -1,326 0,197698
## volt350kV:temp180C   -178,8      199,3  -0,897 0,378997
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
##
## ----------------------------------------------------------------------
## Sigma link function:  identity
## Sigma Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0,2737     0,0338     8,1 3,47e-08 ***
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
##
## ----------------------------------------------------------------------
## No. of observations in the fit:  32
## Degrees of Freedom for the fit:  9
##       Residual Deg. of Freedom:  23
##                       at cycle:  2
##
## Global Deviance:     405,2909
##             AIC:     423,2909
##             SBC:     436,4826
## **********************************************************************
```

Nota 8 Nos resultados acima para o parâmetro σ está sendo testada a hipótese $\mathrm{H}_0 : \sigma = 0$ contra $\mathrm{H}_1 : \sigma \neq 0$, que não faz sentido neste exemplo (por quê?).

A função genérica `plot` gera alguns gráficos com os resultados do ajuste. No canto inferior direito vemos o gráfico de quantis (gráfico QQ) dos resíduos de quantil.

```
plot(m1g)
```




```
## ************************************************************************
##        Summary of the Quantile Residuals
##                             mean   =  0,0001889944
##                         variance   =  1,032289
##                 coef. of skewness  =  -0,1101654
##                 coef. of kurtosis  =  2,135939
## Filliben correlation coefficient   =  0,9870134
## ************************************************************************
```

```
m2g <- update(m1g, . ~ . -volt:temp)
```

```
## GAMLSS-RS iteration 1: Global Deviance = 416,521
## GAMLSS-RS iteration 2: Global Deviance = 416,5209
```

```
summary(m2g)
```

```
## ******************************************************************
## Family:  c("GA", "Gamma")
##
## Call:
## gamlss(formula = resist ~ volt + temp, family = GA(mu.link = "identity",
##     sigma.link = "identity"), data = dados)
##
## Fitting method: RS()
##
## ------------------------------------------------------------------
## Mu link function:  identity
```

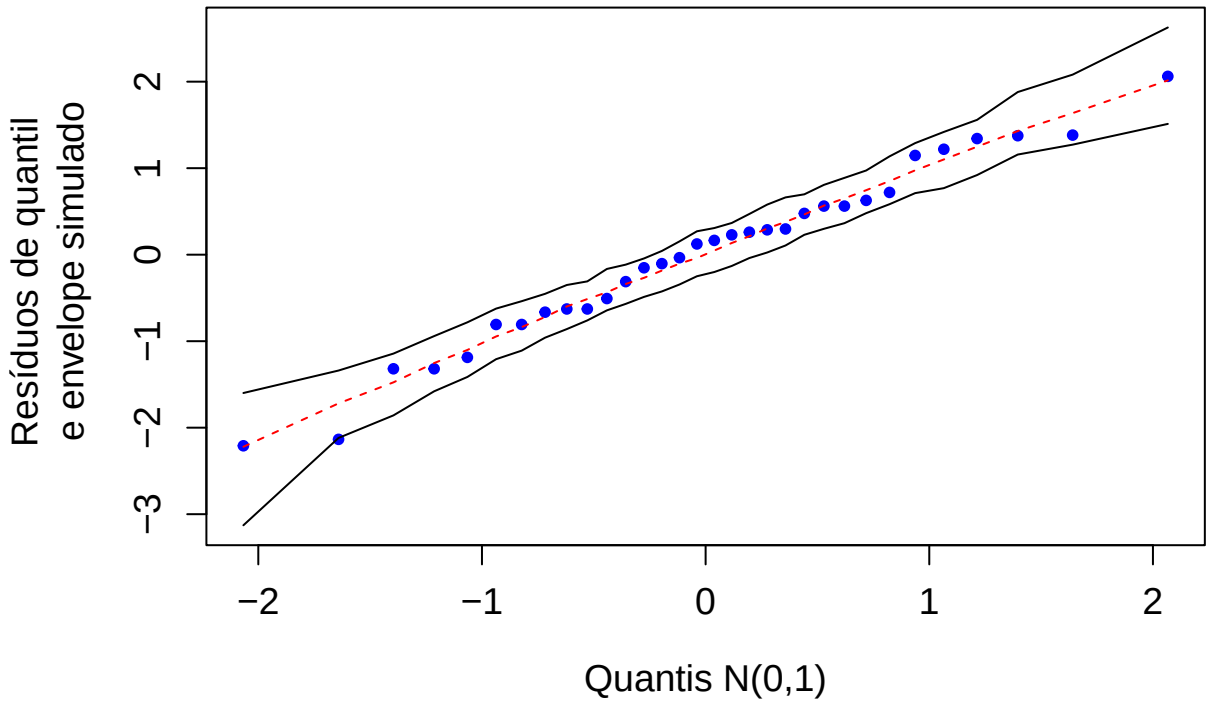
```
## Mu Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1039,84     117,84   8,824 2,68e-09 ***
## volt250kV    -426,45     130,38  -3,271  0,00302 **
## volt300kV    -608,82     118,41  -5,142 2,31e-05 ***
## volt350kV    -612,90     118,07  -5,191 2,03e-05 ***
## temp180C     -117,59      56,65  -2,076  0,04792 *
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
##
## ------------------------------------------------------------------
## Sigma link function:  identity
## Sigma Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0,32459    0,03988   8,139 1,28e-08 ***
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
##
## ------------------------------------------------------------------
## No. of observations in the fit:  32
## Degrees of Freedom for the fit:  6
##       Residual Deg. of Freedom:  26
##                       at cycle:  2
##
## Global Deviance:     416,5209
##             AIC:     428,5209
##             SBC:     437,3153
## ******************************************************************
```

```
## Gráfico de envelope para m2
# Número de simulações
B <- 100

rq <- resid(m2g)
rqo <- sort(rq)

# Simulações
mrq <- matrix(0, B, n)
for (b in 1:B) {
   ysim <- simulate(m2, nsim = 1)[, 1]
   m2s <- gamlss(ysim ~ volt + temp, family = GA(mu.link = "identity",
            sigma.link ="identity"), data = dados)
   rqs <- resid(m2s)
   mrq[b,] <- rqs
}
```

```
mrq <- t(apply(mrq, 1, sort))
Z <- qnorm((1:n - 3/8) / (n + 1/4))
rqm <- apply(mrq, 2, mean)
rq25 <- apply(mrq, 2, function(x) quantile(x, 0.025))
rq975 <- apply(mrq, 2, function(x) quantile(x, 0.975))
mrq <- cbind(Z, rqo, rq25, rqm, rq975)
```

```
# Envelope
par(mai = c(1.2, 1.2, 0.5, 0.1))
```

```
plot(mrq[, 1], mrq[, 2], pch = 20, ylim = range(mrq[, -1]),
     cex.axis = 1.2, cex.lab = 1.2, xlab = "Quantis N(0,1)",
     ylab = "Resíduos de quantil \n e envelope simulado", col = "blue")
lines(mrq[, 1], mrq[, 3])
lines(mrq[, 1], mrq[, 4], lty = 2, col = "red")
lines(mrq[, 1], mrq[, 5])
```



Nota 9. Apresente os resultados do ajuste do modelo exponencial (caso particular do modelo gama). O gráfico de resíduos de quantil com envelope indica que este modelo faz um bom ajuste?

Nota 10. Com a função `gamlss` podemos propor um modelo de regressão em que o coeficiente de variação σ está relacionado com as covariáveis. Assim, temos um modelo com dois preditores lineares e duas funções de ligação. O comando abaixo ajusta um modelo gama com funções de ligação identidade e logaritmo para a média μ e σ, respectivamente, notando que as covariáveis nos preditores lineares não são as mesmas. O modelo abaixo não é um MLG tradicional.

```
m3g <- gamlss(resist ~ volt + temp, family = GA(mu.link = "identity",
       sigma.link ="log"), sigma.formula = ~ volt, data = dados)
```

Nota 11. Refaça o exemplo em linguagem Python.