Modelos para dados de contagem

Poisson e binomial negativa

2023

Os dados se referem ao número de faltas escolares de crianças em um certo ano letivo. Os dados estão disponíveis no objeto `quine` do pacote `MASS` em linguagem R. Modelos serão ajustados com a função `gamlss` do pacote `gamlss` (ou seja, a função `glm` não será usada).

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

```
# Pacotes
library(MASS)
library(gamlss)
```

```
# Descrição dos dados
help(quine)
```

```
quine {MASS}    R Documentation

Absenteeism from School in Rural New South Wales

Description
The quine data frame has 146 rows and 5 columns. Children from Walgett, New South Wales,
Australia, were classified by Culture, Age, Sex and Learner status and the number of days
absent from school in a particular school year was recorded.

Usage
quine

Format
This data frame contains the following columns:

Eth
ethnic background: Aboriginal or Not, ("A" or "N").

Sex
sex: factor with levels ("F" or "M").

Age
age group: Primary ("F0"), or forms "F1," "F2" or "F3".

Lrn
learner status: factor with levels Average or Slow learner, ("AL" or "SL").

Days
days absent from school in the year.

Source
```

S. Quine, quoted in Aitkin, M. (1978) The analysis of unbalanced cross classifications (with discussion). Journal of the Royal Statistical Society series A 141, 195-223.

References

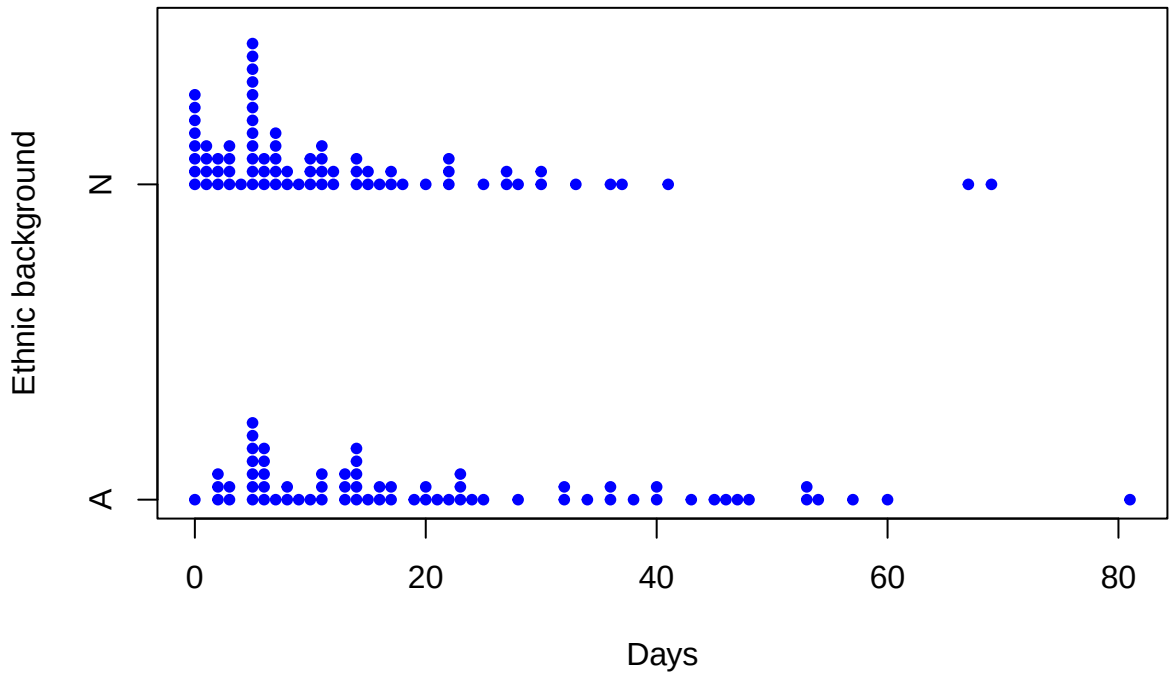
Venables, W. N. and Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth edition. Springer.

```
# Estatísticas descritivas
data(quine)
summary(quine)
```

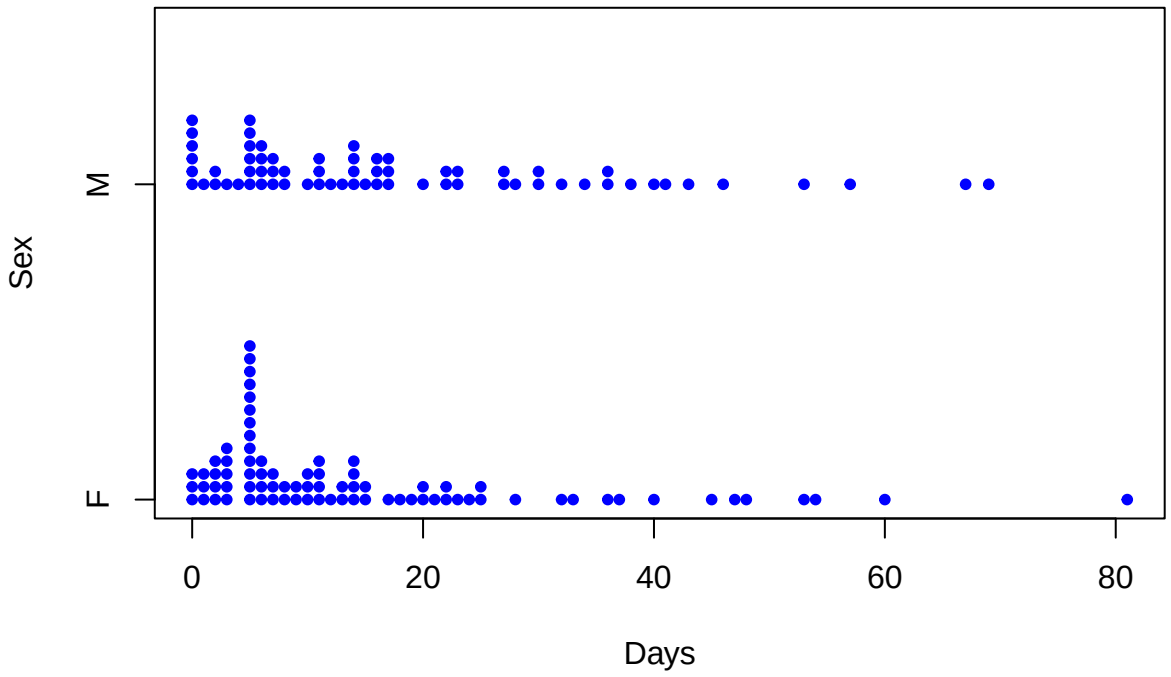
```
##  Eth    Sex    Age     Lrn          Days
##  A:69   F:80   F0:27   AL:83   Min.   : 0,00
##  N:77   M:66   F1:46   SL:63   1st Qu.: 5,00
##                F2:40           Median :11,00
##                F3:33           Mean   :16,46
##                                3rd Qu.:22,75
##                                Max.   :81,00
```

Os gráficos abaixo mostram a relação entre a variável resposta e as covariáveis.

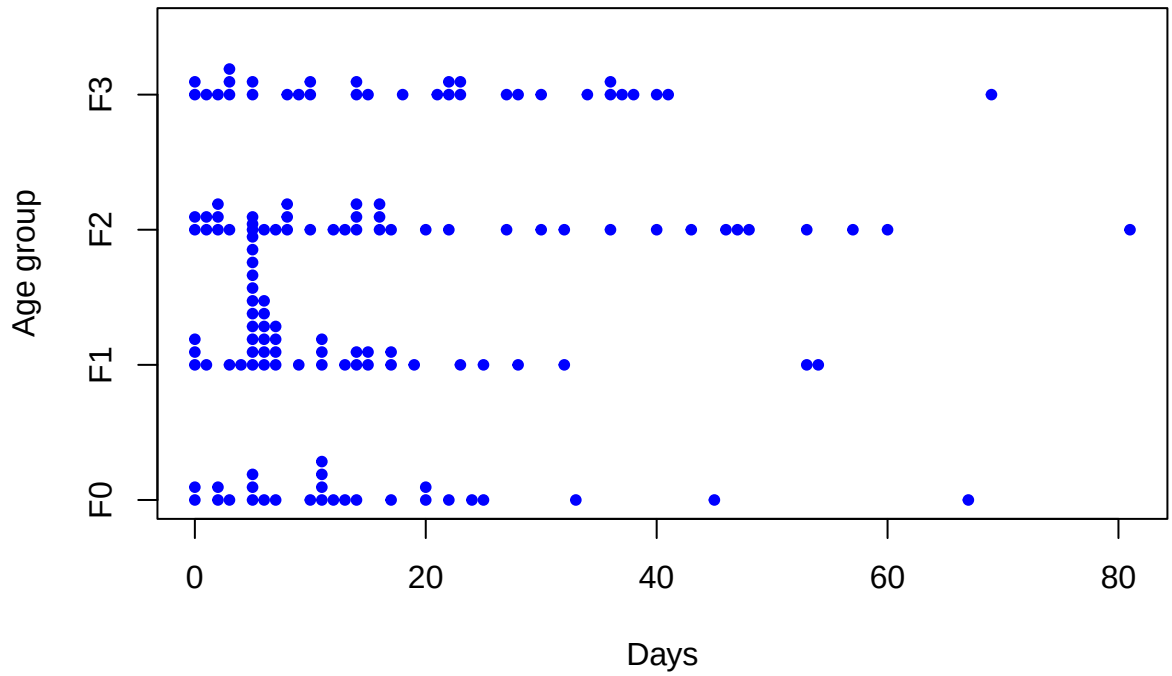
```
with(quine, stripchart(Days ~ Eth, method = "stack", pch = 20,
     ylab = "Ethnic background", col = "blue"))
```



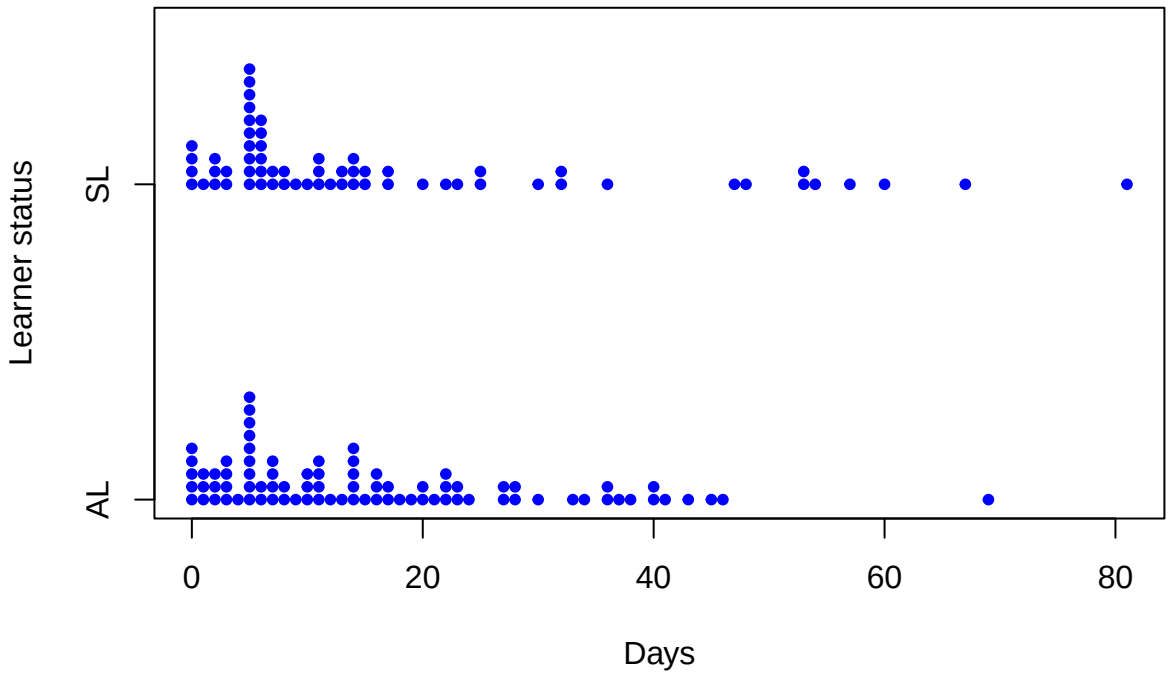

```
with(quine, stripchart(Days ~ Sex, method = "stack", pch = 20,
      ylab = "Sex", col = "blue"))
```




```
with(quine, stripchart(Days ~ Age, method = "stack", pch = 20,
      ylab = "Age group", col = "blue"))
```

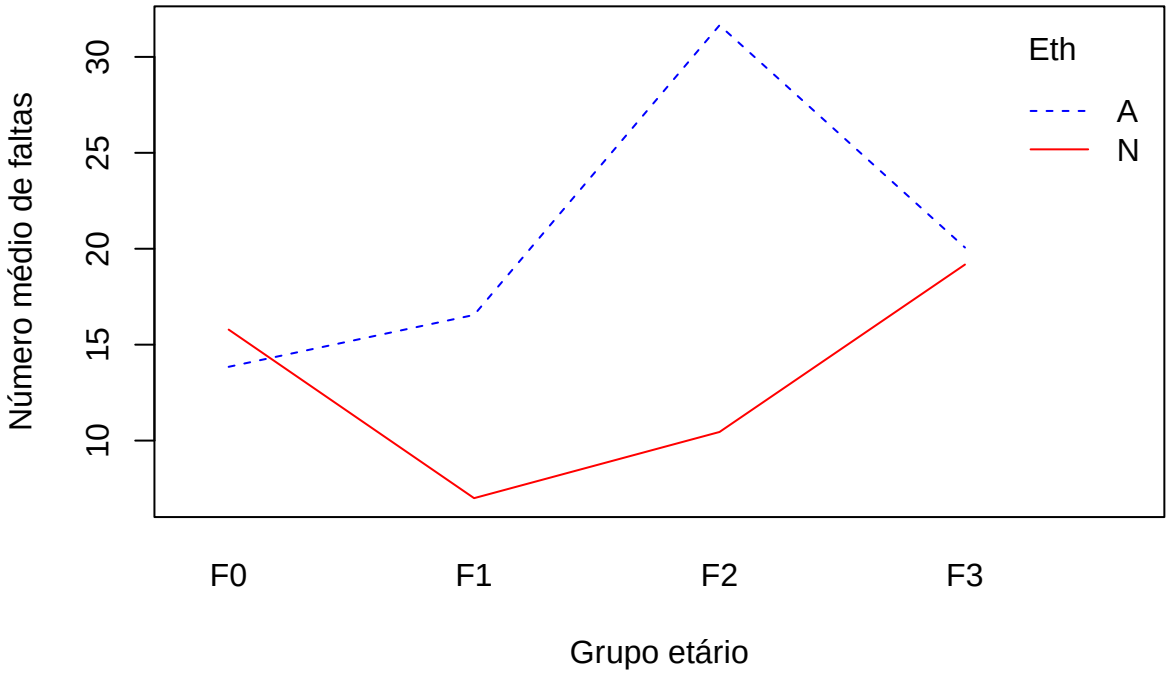



```
with(quine, stripchart(Days ~ Lrn, method = "stack", pch = 20,
      ylab = "Learner status", col = "blue"))
```



O gráfico abaixo sugere que há interação entre etnia e grupo etário.

```
# Gráfico de interação
with(quine, interaction.plot(Age, Eth, Days, xlab = "Grupo etário",
   ylab = "Número médio de faltas", col = c("blue", "red")))
```

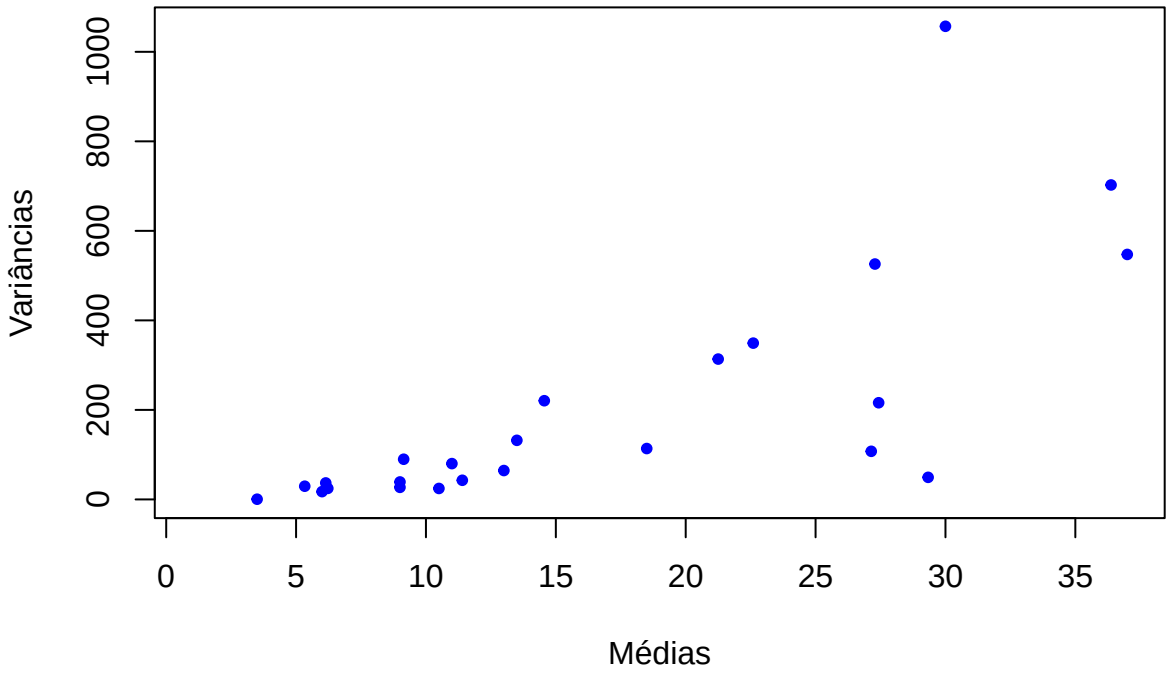


Nota 1. Verifique se há interação entre outras covariáveis.

O modelo Poisson com parâmetro μ é tal que $E(Y) = \mu$ e $\text{Var}(Y) = \mu$, ou seja, $\text{Var}(Y) = E(Y) = \mu$. Dizemos que o modelo Poisson acomoda dados com equidispersão (*equidispersion*).

As quatro covariáveis são qualitativas. Portanto, podemos calcular médias e variâncias amostrais do número de faltas escolares nas $2 \times 2 \times 4 \times 2 = 32$ combinações de níves das covariáveis.

```r
# Relação entre média e variância
medias <- with(quine, tapply(Days, list(Eth, Sex, Age, Lrn), mean))
vars <- with(quine, tapply(Days, list(Eth, Sex, Age, Lrn), var))
plot(medias, vars, pch = 20, col = "blue", xlab = "Médias", ylab = "Variâncias")
```



Nota 2. Se tivermos covariáveis quantitativas, sugira uma forma de avaliar a relação entre variância e média.

O gráfico acima indica que a variância amostral supera a média amostral e é crescente em relação à média amostral de uma forma que não é linear. O modelo binomial negativa é uma alternativa ao modelo Poisson. Usamos a implementação `NBI` do pacote `gamlss`, com função massa de probabilidade

$$f(y;\mu,\sigma) = P(Y = y;\mu,\sigma) = \frac{\Gamma(y+1/\sigma)}{\Gamma(1/\sigma)y!}\frac{(\sigma\mu)^y}{(1+\sigma\mu)^{y+1/\sigma}}I_{0,1,2,\ldots}(y), \tag{1}$$

em que $\mu > 0$ e $\sigma > 0$. Temos $E(Y) = \mu$ e $\text{Var}(Y) = \mu + \sigma\mu^2 = \mu(1+\sigma\mu)$, ou seja, $\text{Var}(Y)$ é uma função quadrática da média μ e $\text{Var}(Y) > E(Y)$. Dizemos que o modelo binomial negativa acomoda dados com sobredispersão (*overdispersion*). Abaixo utilizamos a função de ligação logaritmo (*default*) para os dois parâmetros (ambos são positivos).

```r
## Modelos
# Binomial negativa sem interação Eth:Age
mbn <- gamlss(Days ~ ., family = NBI, data = quine)
```

```
## GAMLSS-RS iteration 1: Global Deviance = 1093,151
## GAMLSS-RS iteration 2: Global Deviance = 1093,151
```

```r
summary(mbn)
```

```
## ******************************************************************
## Family:  c("NBI", "Negative Binomial type I")
##
## Call:  gamlss(formula = Days ~ ., family = NBI, data = quine)
##
## Fitting method: RS()
##
```

```
## -------------------------------------------------------------------
## Mu link function:  log
## Mu Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2,89437    0,22791  12,700  < 2e-16 ***
## EthN        -0,56954    0,15761  -3,614 0,000422 ***
## SexM         0,08257    0,16469   0,501 0,616923
## AgeF1       -0,44862    0,23758  -1,888 0,061089 .
## AgeF2        0,08823    0,24155   0,365 0,715482
## AgeF3        0,35737    0,24662   1,449 0,149589
## LrnSL        0,29234    0,18293   1,598 0,112305
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
##
## -------------------------------------------------------------------
## Sigma link function:  log
## Sigma Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0,2429     0,1263  -1,923   0,0566 .
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
##
## -------------------------------------------------------------------
## No. of observations in the fit:  146
## Degrees of Freedom for the fit:  8
##       Residual Deg. of Freedom:  138
##                       at cycle:  2
##
## Global Deviance:     1093,151
##             AIC:     1109,151
##             SBC:     1133,02
## *******************************************************************
```

Nota 3. Apresente o gráfico de resíduos de quantil com envelope para o modelo acima.

Vemos acima que `Sex` e `Lrn` não são significativas a um nível de 5%. Em seguida excluímos estas duas covariáveis e incluímos a interação entre `Eth` e `Age`.

```
# Binomial negativa com interação Eth:Age
mbn2 <- update(mbn, . ~ . + Eth:Age - Sex - Lrn)
```

```
## GAMLSS-RS iteration 1: Global Deviance = 1084,638
## GAMLSS-RS iteration 2: Global Deviance = 1084,638
```

```
summary(mbn2)
```

```
## *******************************************************************
## Family:  c("NBI", "Negative Binomial type I")
##
## Call:  gamlss(formula = Days ~ Eth + Age + Eth:Age, family = NBI,
##     data = quine)
##
## Fitting method: RS()
##
## -------------------------------------------------------------------
## Mu link function:  log
```

```
## Mu Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2,6280     0,2495  10,535  < 2e-16 ***
## EthN          0,1311     0,3455   0,379  0,70495
## AgeF1         0,1784     0,3195   0,558  0,57755
## AgeF2         0,8267     0,3172   2,606  0,01017 *
## AgeF3         0,3708     0,3337   1,111  0,26844
## EthN:AgeF1   -0,9916     0,4394  -2,257  0,02561 *
## EthN:AgeF2   -1,2392     0,4466  -2,775  0,00629 **
## EthN:AgeF3   -0,1763     0,4636  -0,380  0,70438
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
##
## ----------------------------------------------------------------------
## Sigma link function:  log
## Sigma Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0,3056     0,1282  -2,383   0,0186 *
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
##
## ----------------------------------------------------------------------
## No. of observations in the fit:  146
## Degrees of Freedom for the fit:  9
##       Residual Deg. of Freedom:  137
##                       at cycle:  2
##
## Global Deviance:     1084,638
##             AIC:     1102,638
##             SBC:     1129,49
## **********************************************************************
```

Vemos acima que `Eth` não é significativa a um nível de 5%. Como existem interações significativas entre `Eth` e `Age`, optamos por manter a covariável `Eth` no modelo.

Nota 4. O modelo binomial negativa pode ser ajustado com a função `glm.nb` do pacote `MASS`.

O modelo Poisson com o mesmo preditor linear do modelo binomial negativa também é ajustado aos dados.

```
# Poisson com interação Eth:Age
mpoi2 <- gamlss(Days ~ Eth * Age, family = PO, data = quine)
```

```
## GAMLSS-RS iteration 1: Global Deviance = 2185,585
## GAMLSS-RS iteration 2: Global Deviance = 2185,585
```

```
summary(mpoi2)
```

```
## **********************************************************************
## Family:  c("PO", "Poisson")
##
## Call:  gamlss(formula = Days ~ Eth * Age, family = PO, data = quine)
##
##
## Fitting method: RS()
##
## ----------------------------------------------------------------------
## Mu link function:  log
```

```
## Mu Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2,62801    0,07454  35,258  < 2e-16 ***
## EthN         0,13110    0,10040   1,306  0,19381
## AgeF1        0,17838    0,09261   1,926  0,05615 .
## AgeF2        0,82673    0,08447   9,787  < 2e-16 ***
## AgeF3        0,37084    0,09312   3,983  0,00011 ***
## EthN:AgeF1  -0,99157    0,13637  -7,271 2,41e-11 ***
## EthN:AgeF2  -1,23923    0,12824  -9,664  < 2e-16 ***
## EthN:AgeF3  -0,17627    0,12753  -1,382  0,16915
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
##
## ------------------------------------------------------------------
## No. of observations in the fit:  146
## Degrees of Freedom for the fit:  8
##       Residual Deg. of Freedom:  138
##                       at cycle:  2
##
## Global Deviance:     2185,585
##             AIC:     2201,585
##             SBC:     2225,454
## ******************************************************************
```

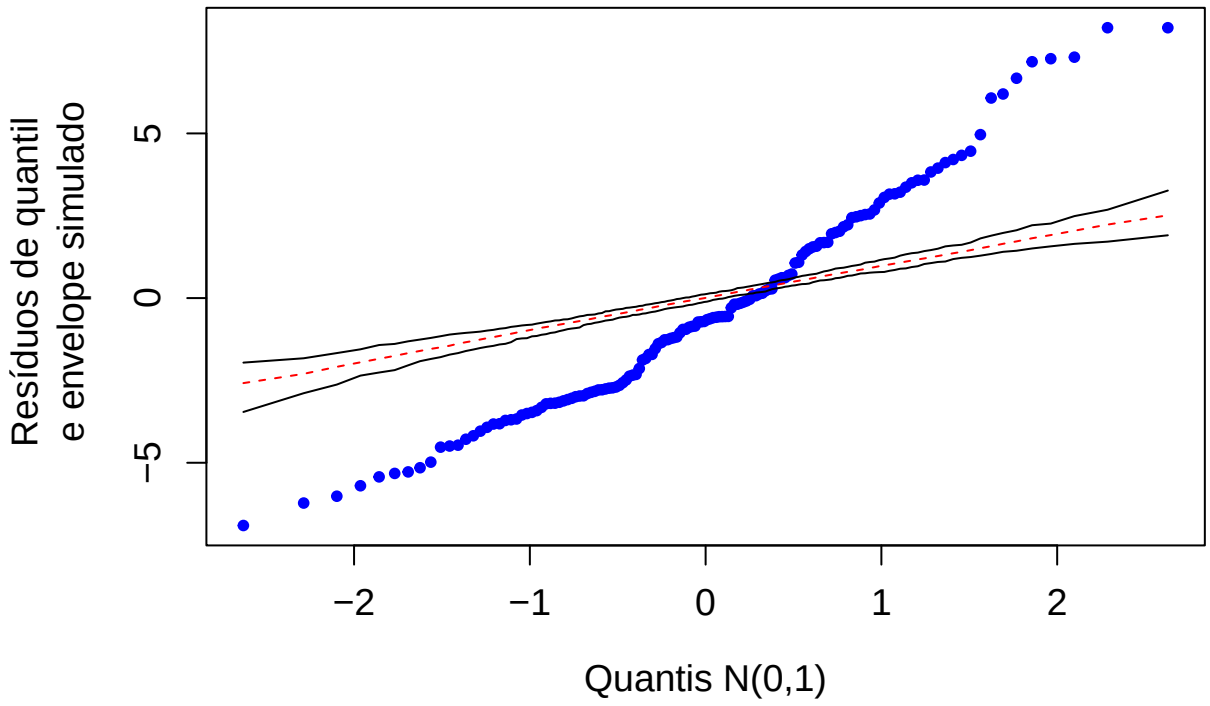
Notamos que ocorre mudança inferencial quando comparamos os coeficientes ($\boldsymbol{\beta}$) dos modelos Poisson e binomial negativa.

Em seguida apresentamos os gráficos de resíduos de quantil com envelope.

```
### Envelopes
B <- 100 # Número de simulações

## Poisson
rq <- resid(mpoi2)
rqo <- sort(rq)

# Simulações
n <- nrow(quine)
set.seed(9810)

mrq <- matrix(0, B, n)
for (b in 1:B) {
   ysim <- rPO(n, mu = mpoi2$mu.fv)
   msim <- gamlss(ysim ~ Eth * Age, family = PO, data = quine)
   rqs <- resid(msim)
   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])
```

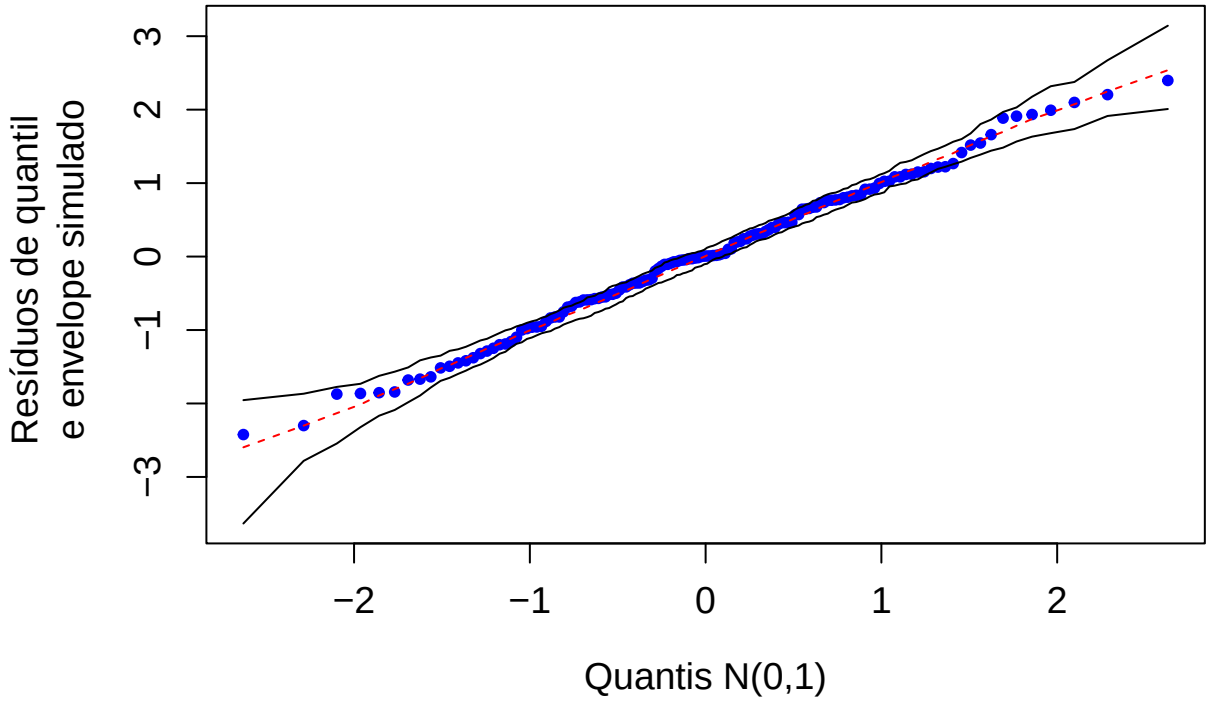


No gráfico acima vemos no eixo vertical que a distribuição dos resíduos de quantil apresenta mais variabilidade do que é possível obter com o modelo Poisson.

```
## Binomial negativa
rq <- resid(mbn2)
rqo <- sort(rq)

# Simulações
mrq <- matrix(0, B, n)
for (b in 1:B) {
   ysim <- rNBI(n, mu = mbn$mu.fv, sigma = mbn$sigma.fv)
   msim <- gamlss(ysim ~ Eth * Age, family = NBI, data = quine)
   rqs <- resid(msim)
   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)
```

```r
# 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 5. O resíduo de quantil normalizado está implementado na função `qresiduals` do pacote `statmod` em R.

Nota 6. Procure interpretar as estimativas dos coeficientes da regressão (β).

Nota 7. Refaça o exemplo em linguagem Python.