Avaliação 2

Regressão II

Paulo Ricardo Seganfredo Campana

9 de março de 2024

Questão 1. O conjunto de dados descrito no arquivo `heartdis.txt` apresenta as variáveis `caso`, número do caso (desconsidere esta variável no modelo proposto) `x1`, pressão sistólica do sangue, `x2`, uma medida de colesterol, `x3`, variável dummy = 1 se há histórico na família de doenças cardíacas, `x4`, uma medida de obesidade, `x5`, idade e `HeartDisease`, se o paciente tem doença cardíaca (variável resposta).

a) Realize o ajuste da regressão logística e selecione as variáveis. O modelo é adequado? Justifique sua escolha.

b) Faça a curva ROC do modelo. O que você pode concluir sobre o ajuste do modelo?

c) Construa um envelope para os resíduos. Há algum ponto que não pertence ao envelope? Se sim, qual(is)?

d) Construa um intervalo de confiança de 90% para os parâmetros do modelo.

e) Interprete o coeficiente β_5 da idade. Mantendo-se as outras varáveis constantes, o acréscimo de um ano na idade do paciente aumenta (ou diminui) em quanto a chance do paciente desenvolver uma doença cardíaca?

```r
data1 <- read.csv("heartdis.txt")

fit1 <- glm(HeartDisease ~ . - caso, family = binomial(link = "logit"), data1)
fit1 <- step(fit1, trace = 0)

summary(fit1)$coefficients
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -4.3518    0.49126   -8.86 8.11e-19
## x2            0.1698    0.05345    3.18 1.49e-03
## x3            0.8820    0.21947    4.02 5.85e-05
## x5            0.0548    0.00908    6.03 1.61e-09
```

A seleção step-wise por AIC do modelo retirou as variáveis `x1` relacionado a pressão sanguínea e `x4` relacionada a obesidade do modelo por serem não significantes, o modelo de regressão logística é dado então por:

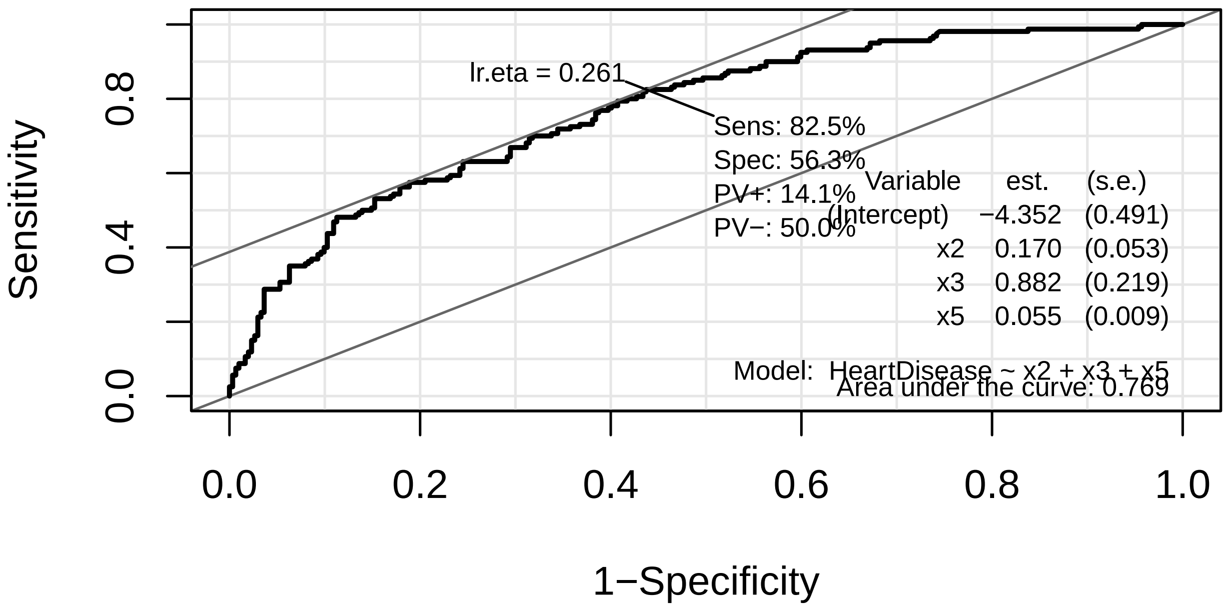
$$\ln\left(\frac{\hat{\mu}}{1-\hat{\mu}}\right) = -4.352 + 0.17x_2 + 0.882x_3 + 0.0548x_5$$

A função desvio do modelo tem valor menor do que o quantil de 95% de uma qui-quadrado, que seria a distribuição desta função sob a hipótese de que o modelo com 4 parâmetros é adequado, assim não rejeitamos esta hipótese e o modelo é aceito.

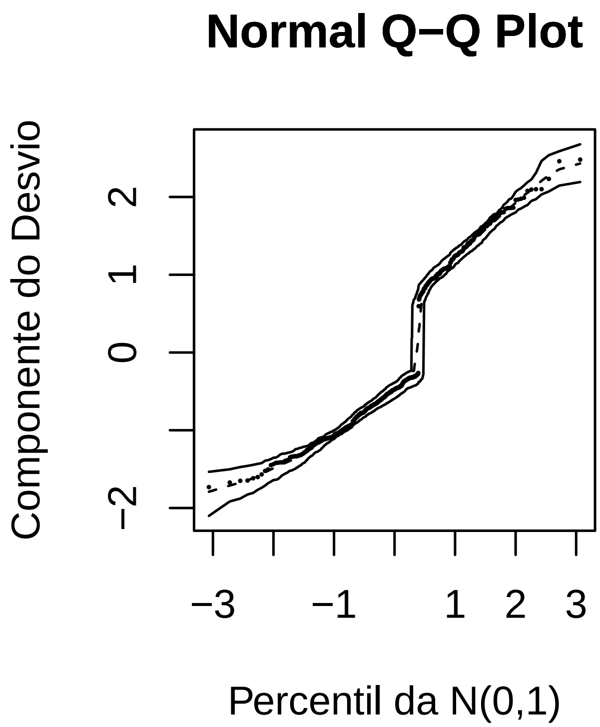
```r
desvio1 <- summary(fit1)$deviance / summary(fit1)$dispersion # 496.2
q.quadr1 <- qchisq(0.95, summary(fit1)$df.residual)          # 508.9
```

O modelo possui uma área sob a curva ROC não tão alta de 0.769, o modelo serve para inferência sobre o efeito das variáveis na presença de doenças cardíacas, porém suas previsões não serão muito confiáveis.

```r
Epi::ROC(form = HeartDisease ~ x2 + x3 + x5, data = data1, plot = "ROC")
```



Não aparentam ter nenhuma observação com resíduo fora do envelope de confiança.



Todos os coeficientes do modelo são altamente significantes, com o intervalo não contendo 0, mesmo assim a variância dos coeficientes é alta.

```
confint(fit1, level = 0.90)
## Waiting for profiling to be done...
##                  5 %   95 %
## (Intercept) -5.1911 -3.573
## x2           0.0830  0.259
## x3           0.5221  1.245
## x5           0.0401  0.070
```

Como o a regressão logística modela o log da razão de chances de forma aditiva, temos que a exponencial do coeficiente da idade representa que a cada ano de vida do paciente, sua razão de chances de possuir doença cardíaca aumenta em aproximadamente 5.6%.

$$
\begin{aligned}
\ln\left(\frac{\hat{\mu}}{1-\hat{\mu}}\right) &= -4.352 + 0.17x_2 + 0.882x_3 + 0.0548x_5 \\
\ln\left(\frac{\hat{\mu}}{1-\hat{\mu}}\right) &= C + 0.0548x_5 \\
\frac{\hat{\mu}}{1-\hat{\mu}} &= e^C e^{0.0548x_5} \\
\frac{\hat{\mu}}{1-\hat{\mu}} &\propto 1.056^{x_5}
\end{aligned}
$$

Questão 2. Considere o banco de dados `Prestige` do pacote `carData` do `R` que fornece 102 observações com seis variáveis das quais iremos utilizar apenas as variáveis: `prestige` (variável resposta) score de prestígio de Pineo-Porter para a ocupação, de uma pesquisa social feita nos meados dos anos 60, `income`, renda média, em dólares em 1971 e `education`, média, em anos, de estudo para a determinada educação.

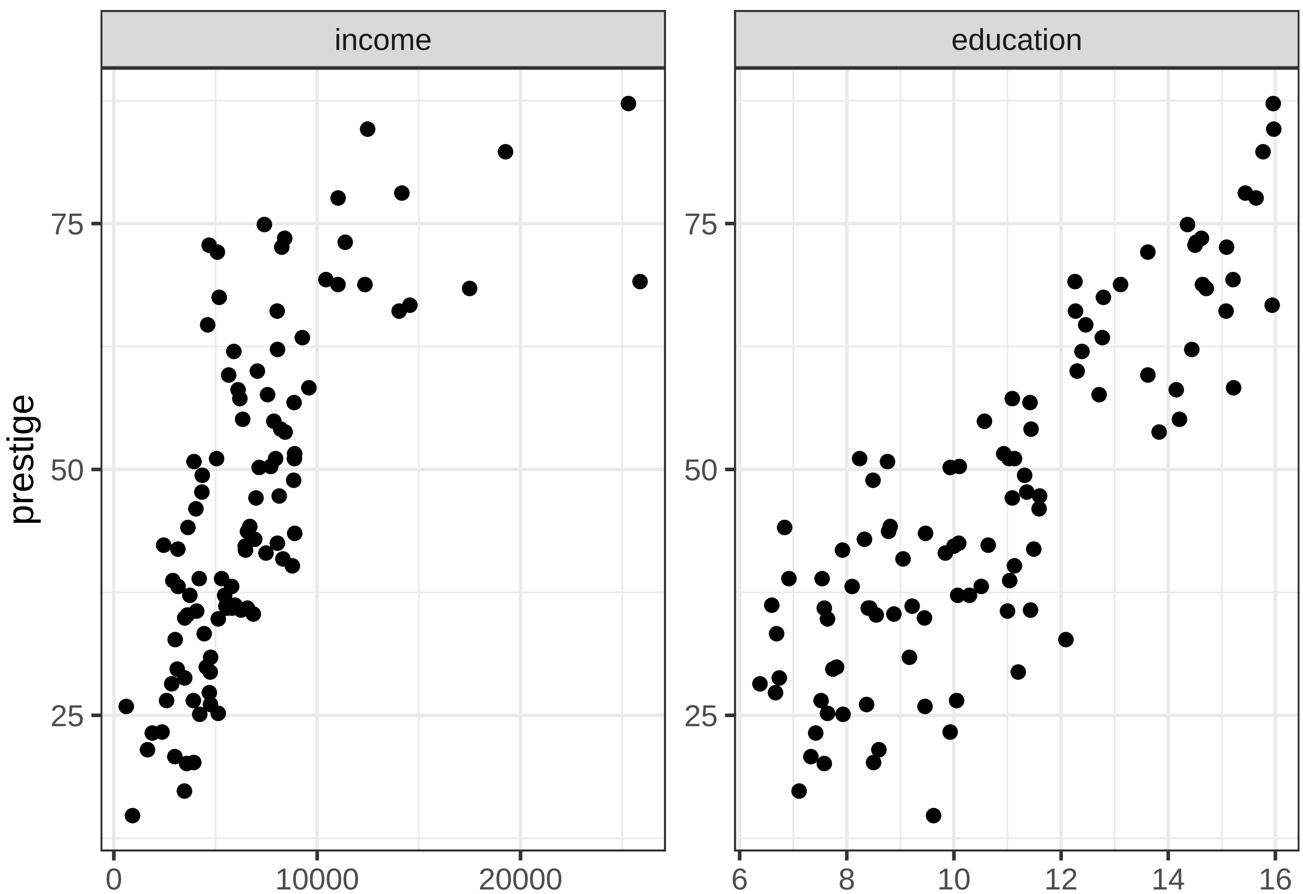
a) Faça o gráfico de dispersão da variável resposta `prestige` pelas variáveis explicativas `income` e `education`.

b) Realize o ajuste de um modelo GAM com a variável resposta `prestige` tendo uma distribuição Normal. Faça o gráfico das funções de suavização.

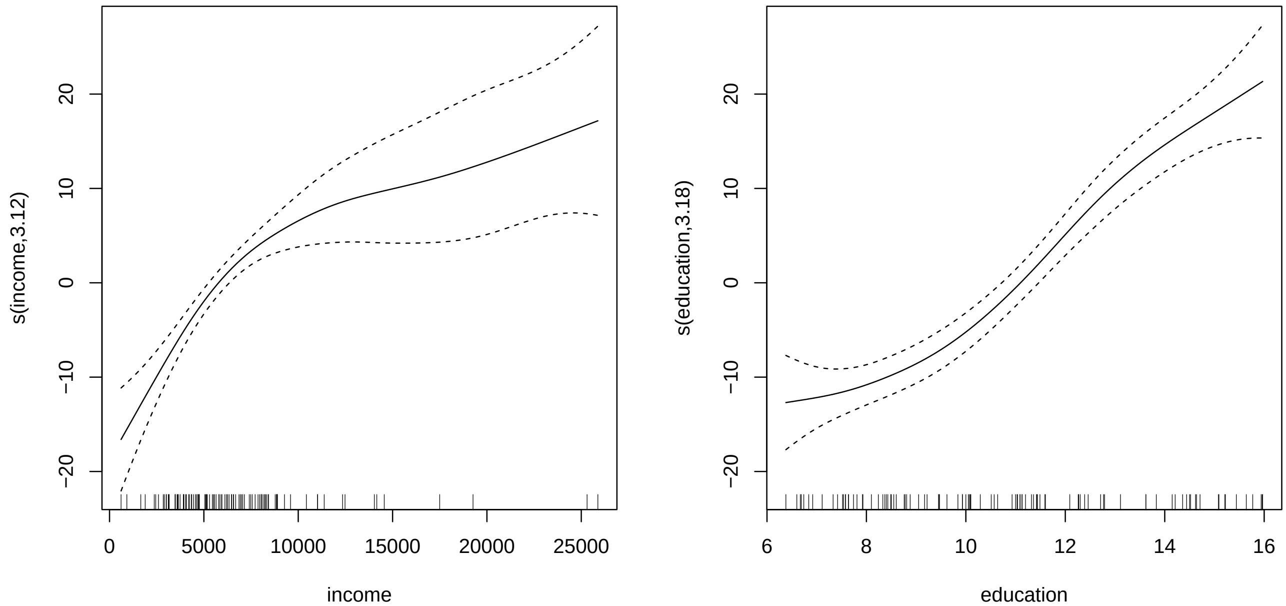
c) Faça uma análise de diagnósticos do modelo escolhido. O que você pode concluir do modelo?

```
data2 <- carData::Prestige
```

O score de prestígio e a renda média parecem ter relações mais fortes porém não linear, com a variável de anos de educação, a relação é mais fraca e parece ser linear.

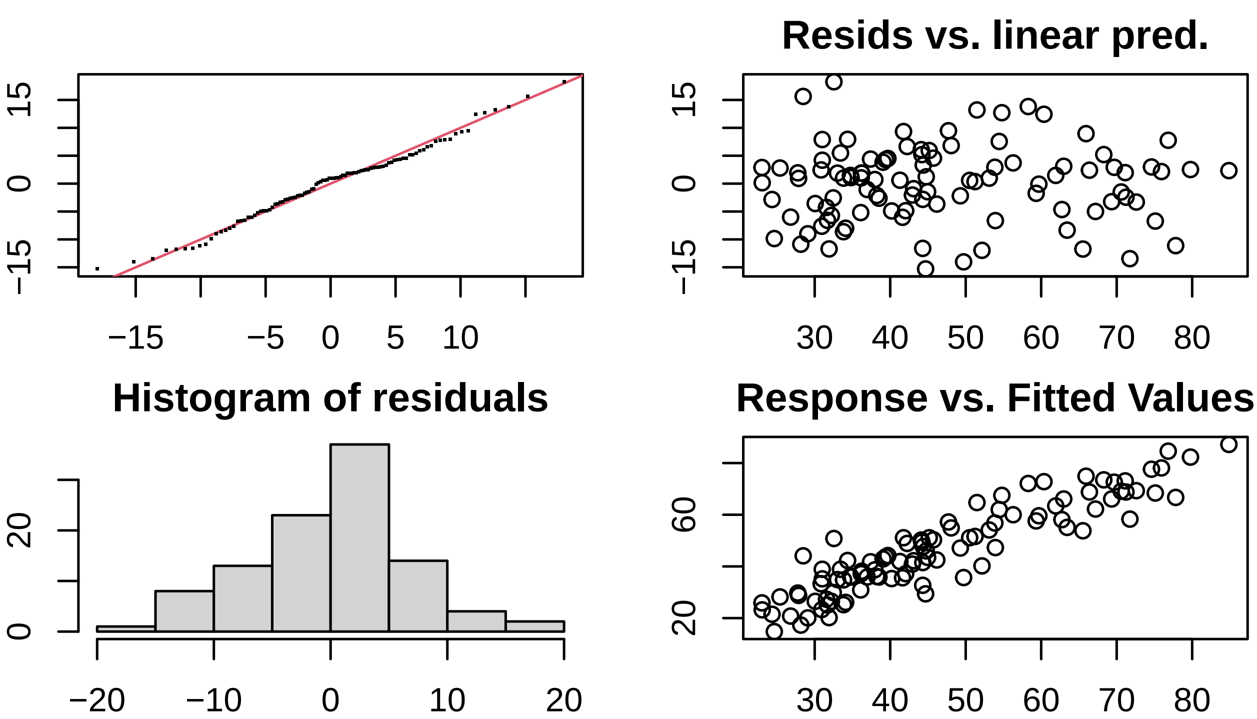



```
library(mgcv)
fit2 <- gam(prestige ~ s(income) + s(education), data = data2)
plot(fit2)
```



As funções estimadas são não lineares, assim a relação entre o score de prestígio e a renda é que o prestígio cresce rápidamente com a renda até por volta de 10 mil doláres, e acima disso ou cresce bem mais devagar ou não exerce influência, essa incerteza é devido a pouca quantidade de profissões observadas com alta renda. A relação com os anos de educação é quase linear, com desvios nas extremidades.

```
par(mar = rep(2, 4))
gam.check(fit2)
```



Os resíduos do modelo aparentam ter distribuição aproximadamente normal, os valores previstos são lineares em relação com os valores observados do score de prestígio e não vemos nenhuma heterocedasticidade no modelo conforme variam os preditores.

Questão 3. Considere o banco de dados `fabric` do pacote `gamlss` do `R`. Em que `y` é o número de falhas em um rolo de tecido e `leng` é o comprimento do tecido. A variável `x`, que é o log de `leng` não usaremos na questão.

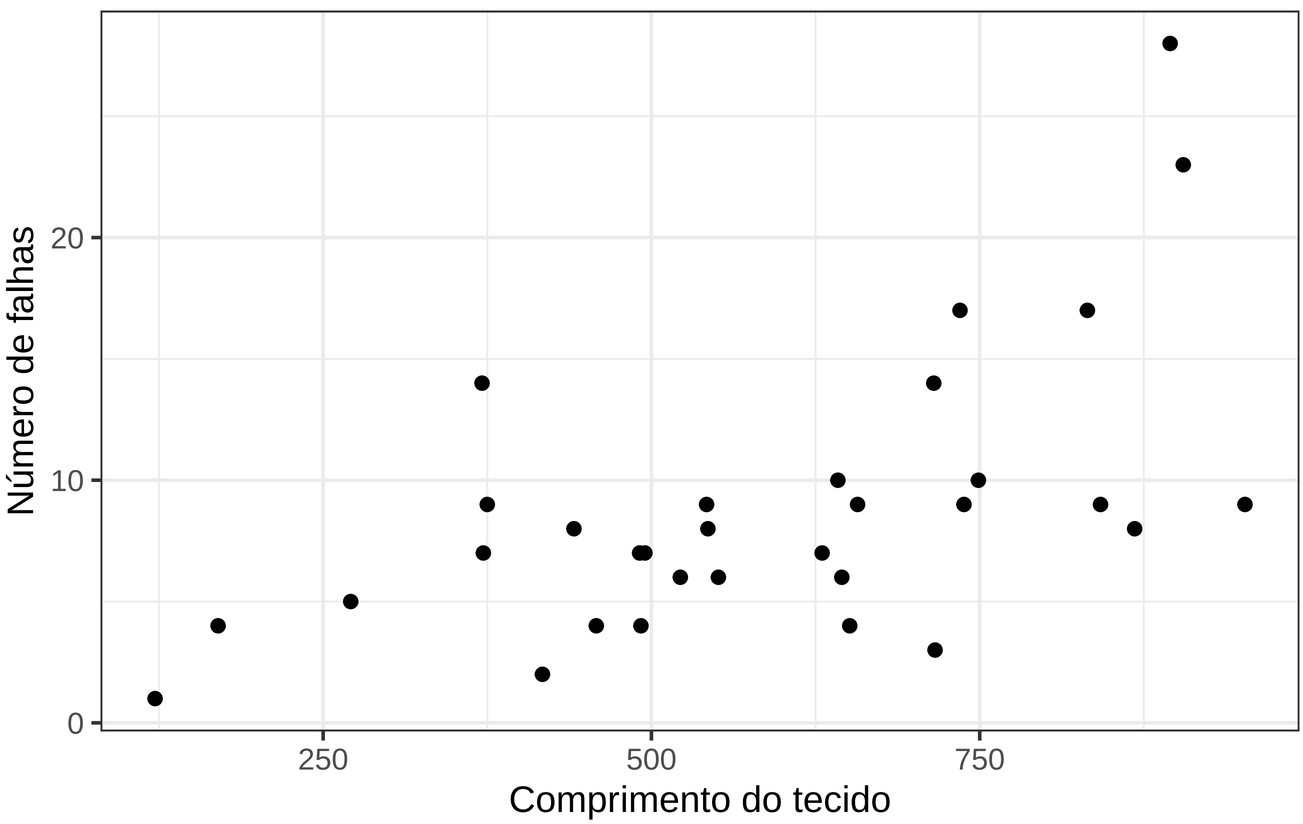
a) Faça o gráfico de dispersão da variável resposta `y` pela variável explicativa (`leng`).

b) Realize o ajuste de um modelo GAMLSS com a variável resposta `R` tendo uma distribuição Poisson.

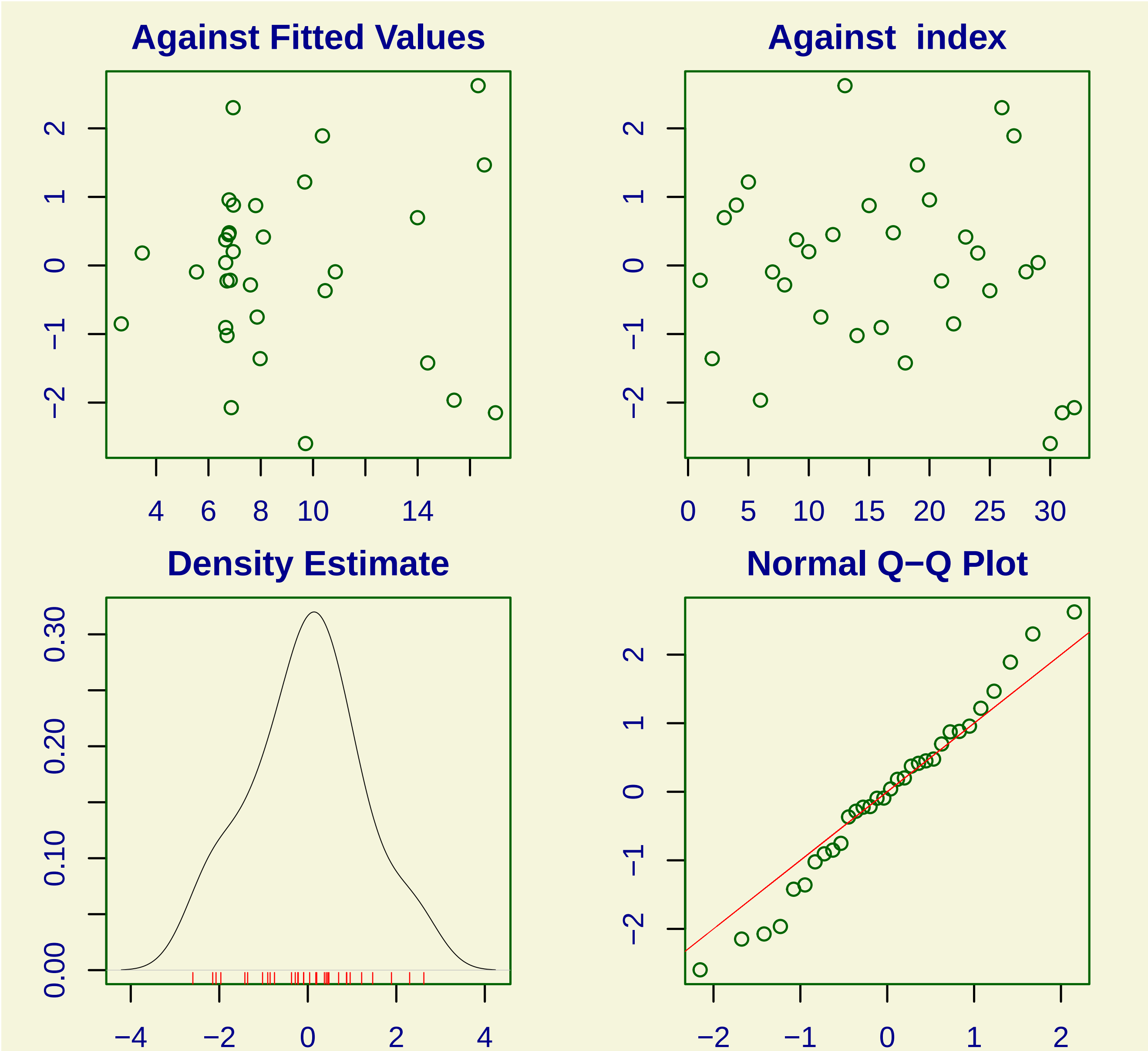
c) Faça uma análise de diagnósticos do modelo escolhido. O que você pode concluir do modelo?

```r
library(gamlss)
data3 <- fabric
```

O comprimento do tecido tem relação não linear com o número de falhas em um rolo e esta relação não é tão forte. A variância do número de falhas também aparenta aumentar com o comprimento.




```r
fit3 <- gamlss(y ~ cs(leng), ~leng, family = PO, data = data3)
par(mar = rep(2, 4))
plot(fit3)
```



Os resíduos do modelo não sofrem alteração na variância para diferentes valores de `y`, eles não são autocorrelacionados e tem distribuição aproximadamente normal.