Segunda avaliação de Regressão II

Gabriel de Jesus Pereira

2024-03-10

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?

```
dados1 <- readr::read_csv("heartdis.txt") |>
  select(-caso)

modelo1 <- glm(HeartDisease ~ + x1 + x2 + x3 + x4 + x5,
               family = binomial(link = "logit"),
               data = dados1)
```

Tabela 1: Tabela dos coeficientes e outras estatísticas do modelo

	Coeficiente	Erro padrão	Estatística	$Pr(>\|z\|)$
Intercepto	-4.313426	0.943928	-4.570	4.89e-06
x_1	0.006435	0.005503	1.169	0.242227
x_2	0.186163	0.056325	3.305	0.000949
x_3	0.903863	0.221009	4.090	4.32e-05
x_4	-0.035640	0.028833	-1.236	0.216433
x_5	0.052780	0.009512	5.549	2.88e-08

Vemos pela tabela acima que boa parte das variáveis acima são significativas, com excessão do x_1, que é a pressão sistólica do sangue e a variável x_4, que é a medida de obesidade. Dessa forma, vamos ajustar outro modelo, mas sem essas variáveis.

```
modelo1 <- glm(HeartDisease ~ x2 + x3 + x5,
               family = binomial(link = "logit"),
               data = dados1)

phi1 <- summary(modelo1)$dispersion
desvio1 <- summary(modelo1)$deviance / phi1
q.quadr1 <- qchisq(0.95, desvio1)

probs_previstas <- predict(modelo1, dados1, type = "response")
classes_previstas <- ifelse(probs_previstas > 0.5, 1, 0)
```

Tabela 2: Tabela dos coeficientes e outras estatísticas do modelo com variáveis significativas

	Coeficiente	Erro padrão	Estatística	$Pr(>\|z\|)$
Intercepto	-4.351833	0.491257	-8.859	2e-16
x_2	0.169796	0.053446	3.177	0.00149
x_3	0.881992	0.219469	4.019	5.85e-05
x_5	0.054755	0.009077	6.033	1.61e-09

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = -4.351 + 0.169x_2 + 0.881x_3 + 0.054x_5$$

A tabela a seguir nos diz que, a um nível de 5% de significância, o modelo é adequado. Chegamos a esse resultados pois o desvio é menor que o quantil χ^2.

	Resultado Final
$Desvio/\phi$	496.1803
χ^2	549.1079

Observando a tabela acima, vemos que o desvio é menor que χ^2, indicando que o modelo é adequado.

Agora observe a matriz confusão abaixo:

```
# matriz de confusão

cm <- confusionMatrix(table(classes_previstas, dados1$HeartDisease))
cm
```

```
Confusion Matrix and Statistics


classes_previstas   0   1
                0 259  80
                1  43  80

               Accuracy : 0.7338
                 95% CI : (0.691, 0.7735)
    No Information Rate : 0.6537
    P-Value [Acc > NIR] : 0.0001366

                  Kappa : 0.3782

 Mcnemar's Test P-Value : 0.0011703

            Sensitivity : 0.8576
            Specificity : 0.5000
         Pos Pred Value : 0.7640
         Neg Pred Value : 0.6504
             Prevalence : 0.6537
         Detection Rate : 0.5606
   Detection Prevalence : 0.7338
      Balanced Accuracy : 0.6788

       'Positive' Class : 0
```

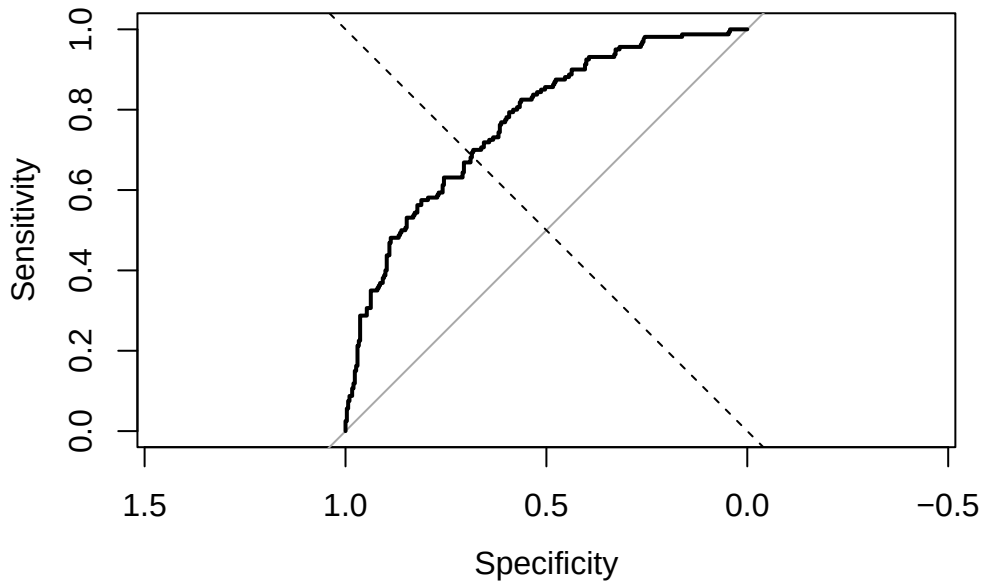
Pela matriz de confusão, podemos ver que 259 pessoas corretamente classificadas como não tendo doença cardíaca e 80 pessoas classificadas corretamente como tendo doença cardíaca. Ainda, vemos que 43 pessoas sem doença cardíaca foram incorretamente classificadas como tendo a doença. 80 pessoas com a doença foram incorretamente classificadas como não tendo a doença cardíaca. Podemos ver também que conseguimos uma acurácia de 73.38%, que é a proporção de predições corretas. Assim, 85.76% dos casos em que a pessoa não tem a doença foram identificadas pelo modelo. Já os casos em que as pessoas tem a doença, 50% foram corretamente identificados.

Agora fica mais claro o porque os casos em que as pessoas não tem a doença são melhores classificados pelo modelo. Na nossa base de dados existem 302 pessoas sem a doença e 160 tem a doença. Dessa forma, os casos em que as pessoas não tem a doença, serão melhor classificados.

b)

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

```
roc_obj <- roc(dados1$HeartDisease, probs_previstas)
plot(roc_obj, main = "")
abline(0, 1, lty = 2, col = "black")
```



```r
auc(roc_obj)
```

```
Area under the curve: 0.7689
```

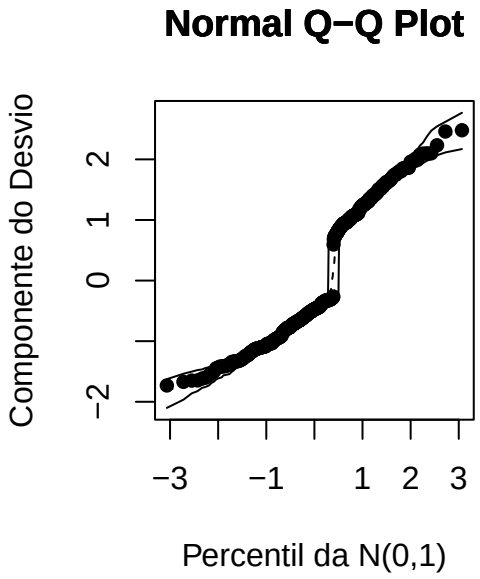
Podemos ver que o AUC é 76.89%, indicando uma performance razoável do classificador.

c)

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

```r
par(mfrow=c(1,1))
X <- model.matrix(modelo1)
n <- nrow(X)
p <- ncol(X)
w <- modelo1$weights
W <- diag(w)
H <- solve(t(X)%*%W%*%X)
H <- sqrt(W)%*%X%*%H%*%t(X)%*%sqrt(W)
h <- diag(H)
td <- resid(modelo1,type="deviance")/sqrt(1-h)
e <- matrix(0,n,100)
#
for(i in 1:100){
dif <- runif(n) - fitted(modelo1)
dif[dif >= 0 ] <- 0
dif[dif<0] <- 1
nresp <- dif
fit <- glm(nresp ~ X, family=binomial)
w <- fit$weights
W <- diag(w)
H <- solve(t(X)%*%W%*%X)
H <- sqrt(W)%*%X%*%H%*%t(X)%*%sqrt(W)
h <- diag(H)
e[,i] <- sort(resid(fit,type="deviance")/sqrt(1-h))}
#
e1 <- numeric(n)
e2 <- numeric(n)
#
for(i in 1:n){
```

```
  eo <- sort(e[i,])
e1[i] <- (eo[2]+eo[3])/2
e2[i] <- (eo[97]+eo[98])/2}
#
med <- apply(e,1,mean)
faixa <- range(td,e1,e2)
par(pty="s")
qqnorm(td,xlab="Percentil da N(0,1)",
ylab="Componente do Desvio", ylim=faixa, pch=16)
#
par(new=T)
#
qqnorm(e1,axes=F,xlab="",ylab="",type="l",ylim=faixa,lty=1)
par(new=T)
qqnorm(e2,axes=F,xlab="",ylab="", type="l",ylim=faixa,lty=1)
par(new=T)
qqnorm(med,axes=F,xlab="", ylab="", type="l",ylim=faixa,lty=2)
```



Não existem pontos fora do envelope, significando que o modelo se ajustou bem aos dados. Existem apenas alguns pontos próximos dos limites do envelope, mas não parecem indicar nenhuma anormalidade no modelo. No entanto, apesar de todos os coeficientes serem significativos, a variância dos coeficientes é bastante alta.

d)

Construa um intervalo de confiaça de 90% para os parâmetros do modelo.

```
confint(modelo1, level = 0.9)
```

```
Waiting for profiling to be done...
```

```
                    5 %        95 %
(Intercept) -5.19111107 -3.57250142
x2           0.08300748  0.25934953
x3           0.52213657  1.24479726
x5           0.04014425  0.07004993
```

e)

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

```
exp(modelo1$coefficients)
```

```
(Intercept)          x2          x3          x5
 0.01288317  1.18506272  2.41570620  1.05628186
```

Mantendo-se as outras variáveis constantes, vemos que as chances de se ter a doença cardíaca aumenta em 5.6% a cada ano.

Questão 2

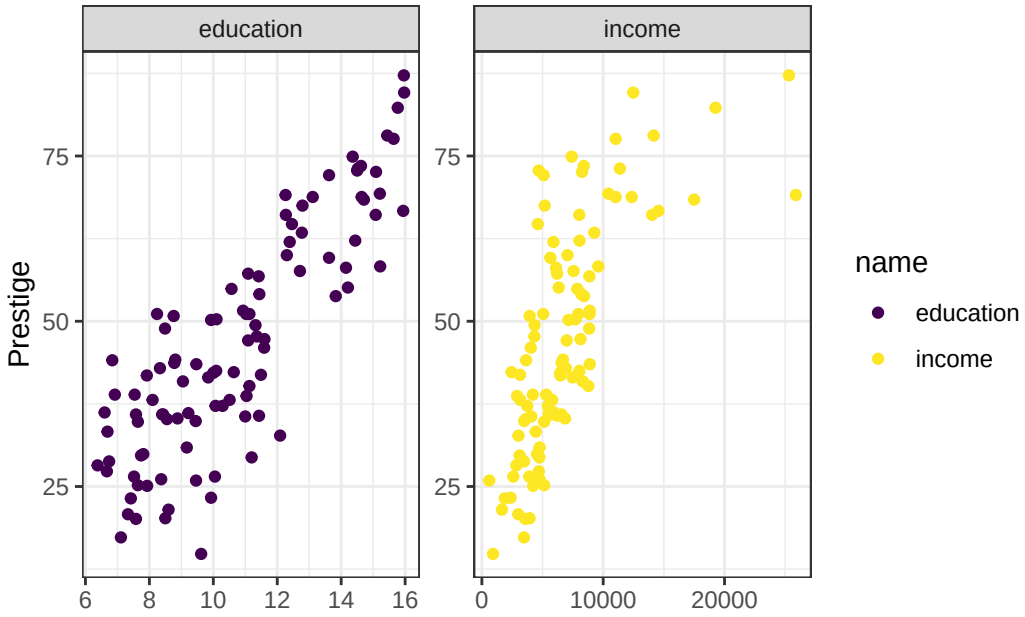
Considere o banco de dados Prestige do pacote carData do R que fornece 102 observações com seis variáveis das quais iremos utilizar apenas as variáveis: **prestige** (variável resposta) score de prestígio de Pineo-Porter para a ocupaço, de uma pesquisa social feita nos meados dos anos 60, **income**, 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**.

```
dados2 <- carData::Prestige

dados2 |>
  pivot_longer(c(income, education)) |>
  ggplot(aes(y = prestige, x = value, color = name)) +
  geom_point() +
  facet_wrap(name~., scales = "free") +
  scale_color_viridis_d() +
  theme_bw() +
  labs(y = "Prestige",
       x = "")
```



Observando os gráficos de dispersão, a variável resposta **prestige** parece ter uma correlação maior com a variável **education** do que com a variável **income**. Além disso, o gráfico à direita parece ter algumas observações extremas para **income** acima de 18000.

b)

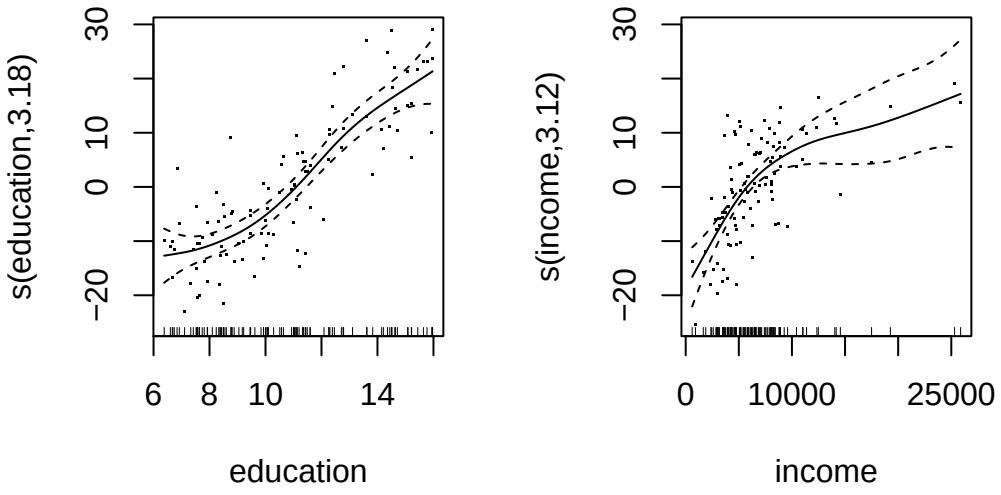
Realize o ajuste de um modelo GAM com a variável resposta **prestige** tendo uma distribuição Normal. Faça o gráfico das funções de suavização.

```
library(mgcv)

modelo2 <- gam(
  prestige ~ s(education) + s(income),
  data = dados2,
  family=gaussian)

par(mfrow = c(1, 2))

plot(modelo2, residuals=TRUE)
```



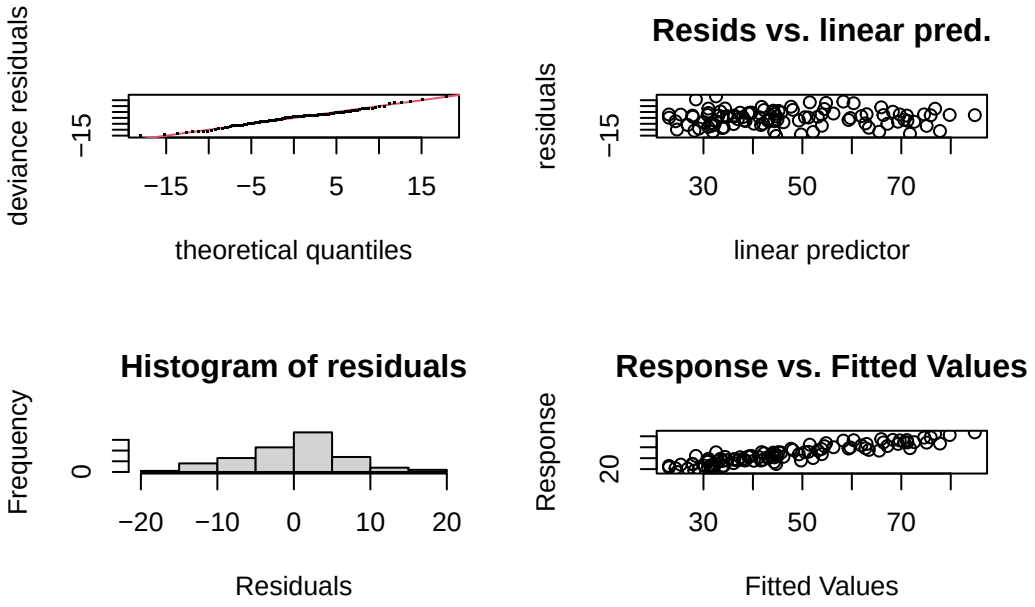
Para um modelo bem ajustado, os resíduos parciais devem estar uniformemente dispersos em torno da curva. Dessa forma, podemos perceber que o caso com a variável **education** parece ter se ajustado melhor. Podemos ver também que ao final do gráfico de **income** o spline não parece estar interpolando bem ao final da distribuição. Isso pode ser devido a poucas observações naquele intervalo de **income**, ou devido a pontos extremos da própria distribuição.

Além disso, podemos ver que o intervalo de credibilidade é mais estreito no gráfico à direita, indicando ter um efeito estatístico significando maior que o gráfico à direita.

c)

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

```
gam.check(modelo2)
```




```
Method: GCV   Optimizer: magic
Smoothing parameter selection converged after 4 iterations.
The RMS GCV score gradient at convergence was 9.783945e-05 .
The Hessian was positive definite.
Model rank =  19 / 19

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

               k'  edf k-index p-value
s(education) 9.00 3.18    1.03    0.61
s(income)    9.00 3.12    0.98    0.43
```

Podemos perceber pelo QQ-Plot que o modelo parece sim seguir uma distribuição aproximadamente normal, pois se assemelham aos quantis teóricos da distribuição normal. Pelo gráfico de resíduos versus estimados, nos mostra que o modelo parece ser homocedástico. Ainda, observando o histograma dos resíduos, podemos ver que os resíduos parecem seguir de fato uma distribuição aproximadamente normal. Além disso, a relação entre os valores observados versus ajustados parecem ser lineares.

Questão 3

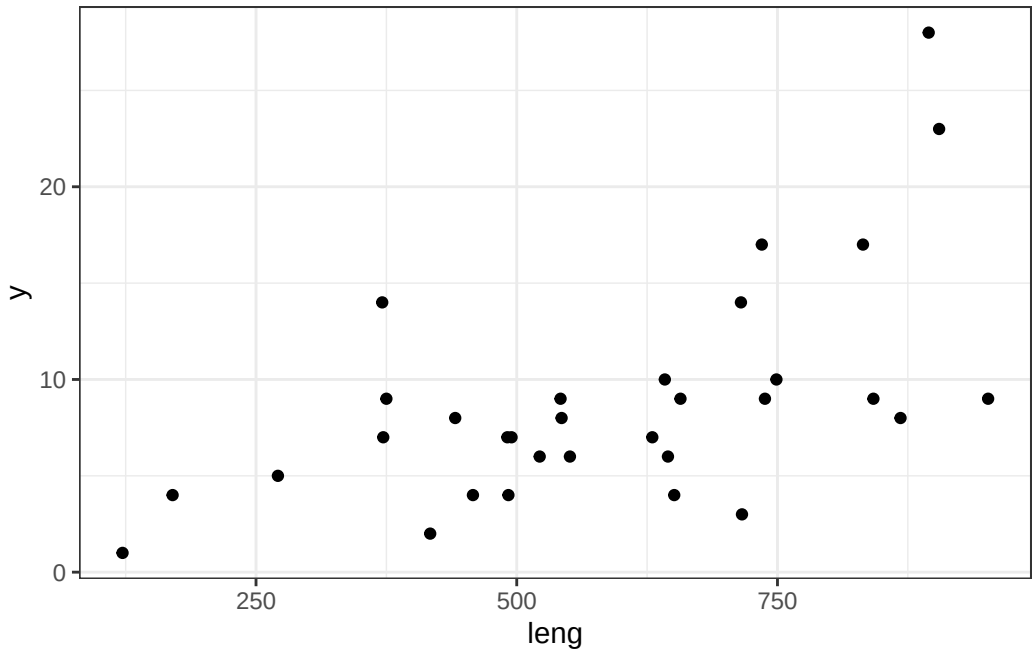
Considere o banco de dados **fabric** do pacote **gamlss** do R. Em que y é o número de falhas em um rolo de tecido e **leng** é o comprimento do tecido. A variável x, que é o log de **leng** não usaremos na questão.

a)

Faça o gráfico de dispersão da variável resposta y pela variável explicativa (x).

```
dados3 <- fabric |>
  select(-x)

dados3 |>
  ggplot(aes(y = y, x = leng)) +
  geom_point() +
  theme_bw()
```



Podemos perceber que a relação entre o número de falhas em um rolo de tecido e o comprimento de tecido é não linear. Dessa forma, talvez o ideial seja utilizar um modelo que consiga capitar essa relação não linear.

b)

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

```
modelo3 <- gamlss(y ~  leng, data = dados3, family = PO)
```

```
GAMLSS-RS iteration 1: Global Deviance = 185.0559
GAMLSS-RS iteration 2: Global Deviance = 185.0559
```

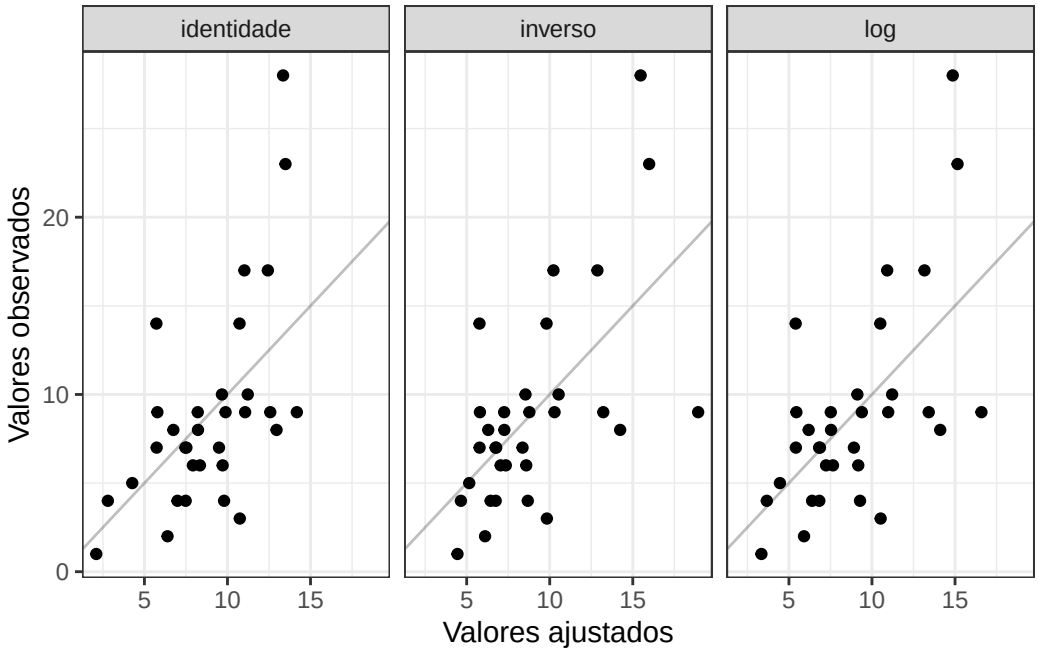
```
modelo3.2 <- gamlss(y ~  leng, data = dados3, family = PO(mu.link = "identity"))
```

```
GAMLSS-RS iteration 1: Global Deviance = 187.7486
GAMLSS-RS iteration 2: Global Deviance = 187.7486
```

```r
modelo3.3 <- gamlss(y ~  leng, data = dados3, family = PO(mu.link = "inverse"))
```

```
GAMLSS-RS iteration 1: Global Deviance = 184.7636
GAMLSS-RS iteration 2: Global Deviance = 184.7636
```

```r
tibble(
  y = dados3$y,
  log = fitted(modelo3),
  identidade = fitted(modelo3.2),
  inverso = fitted(modelo3.3),
) |>
  pivot_longer(-y) |>
  ggplot(aes(x = value, y = y)) +
  facet_wrap(vars(name)) +
  geom_point() +
  geom_abline(alpha = 0.25) +
  labs(x = "Valores ajustados", y = "Valores observados") +
  theme_bw()
```



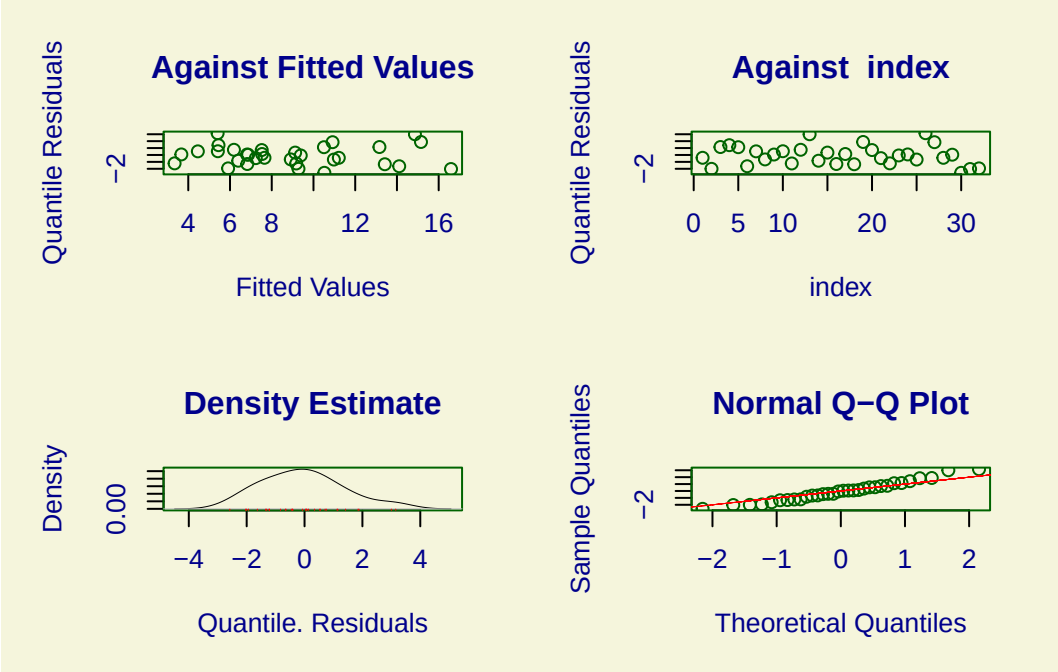
Pelos gráficos de valores observados versus valores ajustados podemos perceber que nenhum dos modelos conseguiram se ajustar muito bem aos dados. No entanto, vemos que o Poisson

com função de ligação **inverso** foi aquele que melhor se ajustou aos dados. O inverso também foi aquele que obteve o menor GAIC (184.764), enquanto o maior foi o de ligação identidade (187.749)

c)

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

```
plot(modelo3)
```




```
****************************************************************
     Summary of the Randomised Quantile Residuals
                           mean   =  -0.06155866
                       variance   =  2.000395
                  coef. of skewness  =  0.3853693
                  coef. of kurtosis  =  2.57788
Filliben correlation coefficient  =  0.989325
****************************************************************
```

Podemos perceber pela densidade dos quantis dos resíduos que os resíduos parecem seguir uma distribuição aproximadamente normal. Podemos perceber também pelos quantis dos resíduos

versus valores ajustados que o modelo parece ser homocedástico. Ainda, no QQ-plot, vemos também que parece seguir aproximadamente uma distribuição normal.