Segunda avaliação de Regressão II

Gabriel de Jesus Pereira

2024-03-09

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)

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 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. Ainda, chegamos no seguinte modelo:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = -4.313 + 0.0064x_1 + 0.186x_2 + 0.903x_3 - 0.035x_4 + 0.052x_5$$

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
$Desvio/\phi$	493.6152
χ^2	546.4087

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
                0 255  76
                1  47  84

               Accuracy : 0.7338
```

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

                  Kappa : 0.3858

 Mcnemar's Test P-Value : 0.0115805

            Sensitivity : 0.8444
            Specificity : 0.5250
         Pos Pred Value : 0.7704
         Neg Pred Value : 0.6412
             Prevalence : 0.6537
         Detection Rate : 0.5519
   Detection Prevalence : 0.7165
      Balanced Accuracy : 0.6847

       'Positive' Class : 0
```

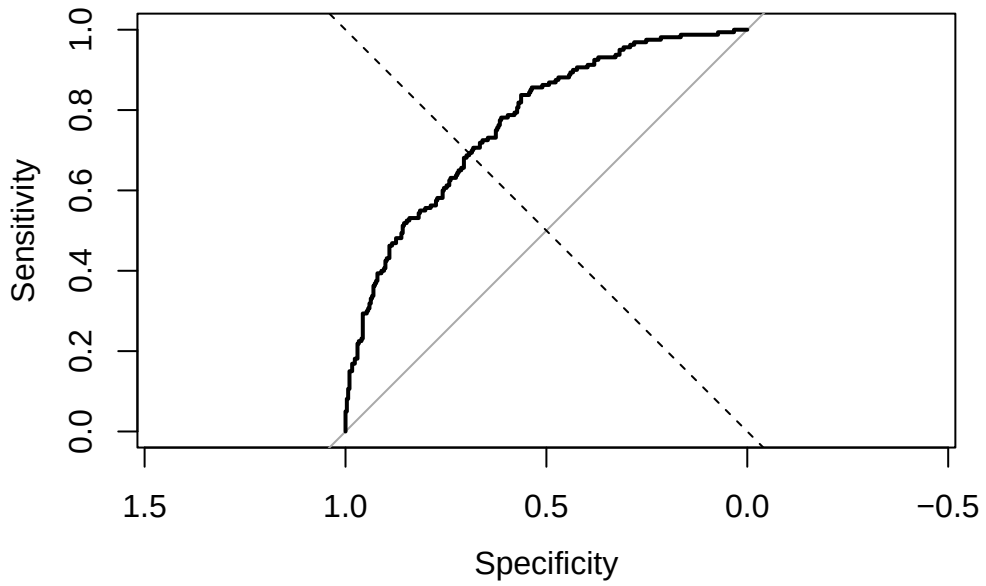
Pela matriz de confusão, podemos ver que 255 pessoas corretamente classificadas como não tendo doença cardíaca e 84 pessoas classificadas corretamente como tendo doença cardíaca. Ainda, vemos que 47 pessoas sem doença cardíaca foram incorretamente classificadas como tendo a doença. 76 pessoas com a doença 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, 84.44% dos casos em que a pessoa não tem a doença foram identificadas pelo modelo. Já os casos em que as pessoas tem a doença, 52.5% 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?

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

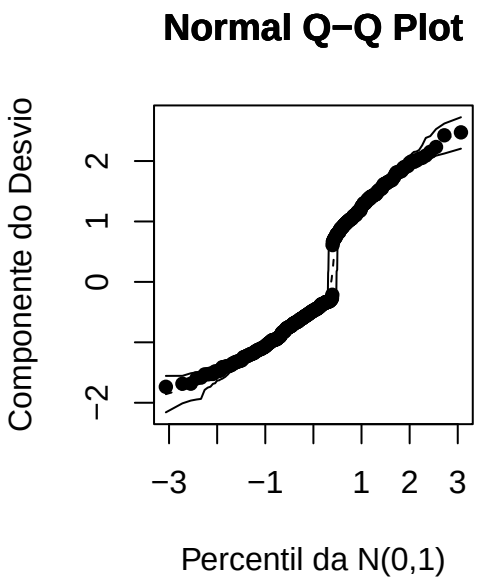


c)

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

```
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.

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) -5.88798718 -2.77748746
x1          -0.00259996  0.01554765
x2           0.09503135  0.28088192
x3           0.54167682  1.26944228
x4          -0.08381711  0.01122148
x5           0.03740383  0.06874928
```

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?

Questão 2

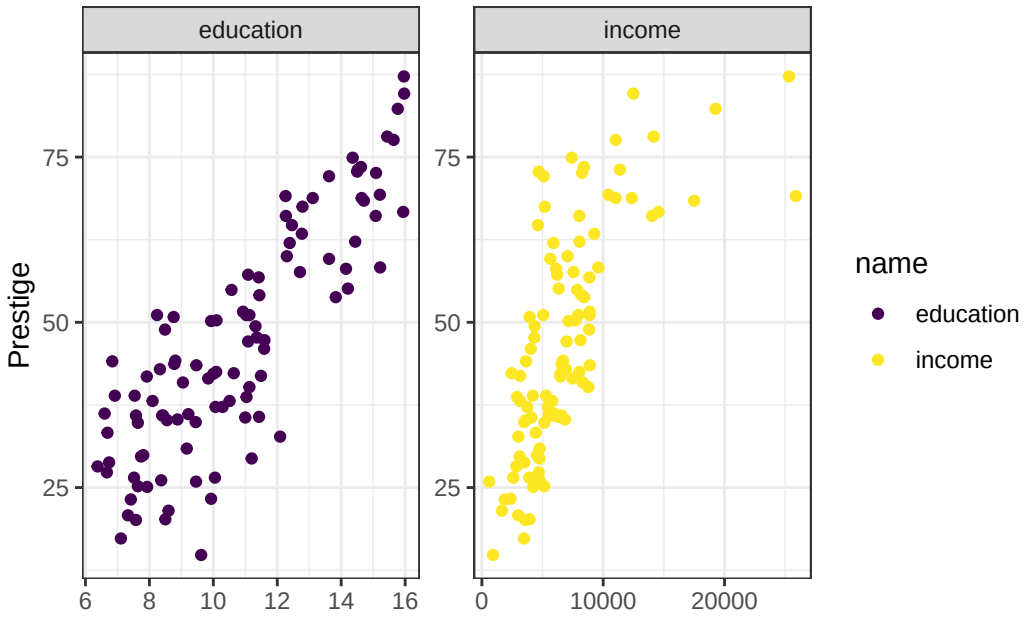
Considere o banco de dados Prestige do pacote carData do R que fornece 102 observações com seis variáveis das quais iremos utilizar apenas as variáveis: **prestige** (variável resposta) score de prestígio de Pineo-Porter para a ocupaço, de uma pesquisa social feita nos meados dos anos 60, **income**,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



b)

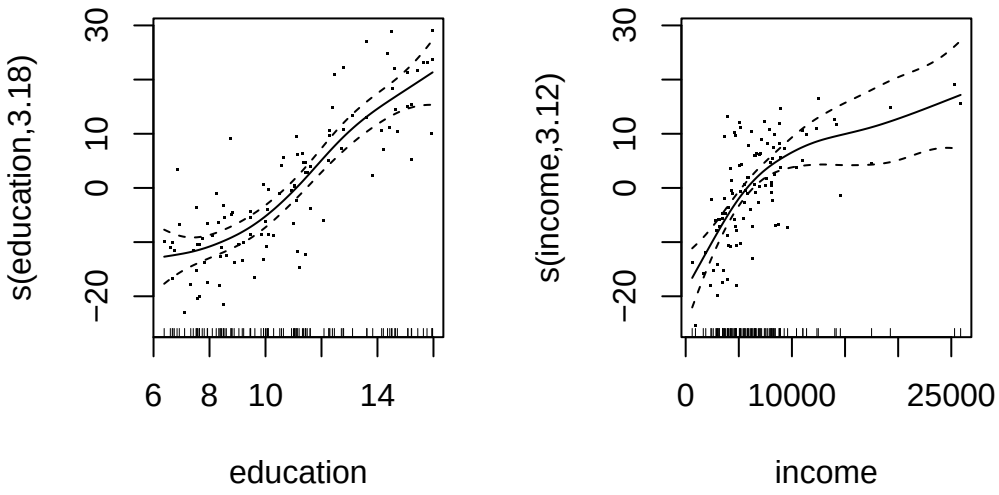
Realize o ajuste de um modelo GAM com a variável resposta **prestige** tendo uma distribuição Normal. Faça o gráfico das funções de suavização.

```
library(mgcv)

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

par(mfrow = c(1, 2))

plot(modelo2, residuals=TRUE)
```



c)

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

Questão 3

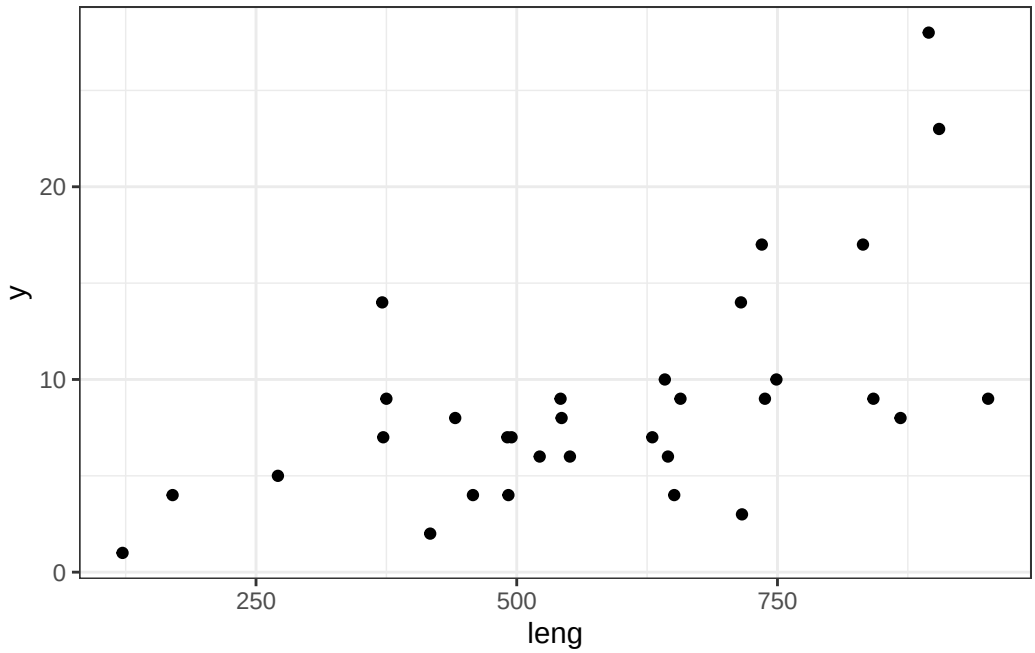
Considere o banco de dados **fabric** do pacote **gamlss** do R. Em que y é o número de falhas em um rolo de tecido e **leng** é o comprimento do tecido. A variável x, que é o log de **leng** não usaremos na questão.

a)

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

```
dados3 <- fabric

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



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(mu.link = "log"))
```

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

c)

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