Untitled

Bruno

#A)

```
dados <- read.csv("covid.csv", header=TRUE, dec=',')

set.seed(10288640)
tabela <- dados[sample(nrow(dados), 2000), ]

write.csv(tabela,file= "baseprincipal.csv",row.names = FALSE)
```

#B)

```
tabela <- read.csv("baseprincipal.csv", header=TRUE, dec=',')

set.seed(10288640)
dt = sort(sample(nrow(tabela), nrow(tabela)*.8))

train<-tabela[dt,]
test<-tabela[-dt,]

write.csv(train,file= "basetreino.csv",row.names = FALSE)
write.csv(test,file= "baseteste.csv",row.names = FALSE)
```

#C)

Iremos modificar os dados da variável dependente y, de 1, 2 e 3 para 0, 1 e NA. Além disso, como nosso objetivo de estudo está relacionado a presença ou não de coronavírus, iremos retirar as linhas que contém o dado NA, ou seja, ainda está esperando o resultado do teste. E assim, o nosso problema se adequa para utilizar o modelo de regressão binária.

```
train <-  read.csv("basetreino.csv", header=TRUE, dec=',')
test <-   read.csv("baseteste.csv", header=TRUE, dec=',')


train <- train[c('sex', 'patient_type', 'intubed', 'pneumonia', 'age', 'pregnancy',
       'diabetes', 'copd', 'asthma', 'inmsupr', 'hypertension',
       'other_disease', 'cardiovascular', 'obesity', 'renal_chronic',
       'tobacco', 'contact_other_covid','covid_res', 'icu')]

test <- test[c('sex', 'patient_type', 'intubed', 'pneumonia', 'age', 'pregnancy',
       'diabetes', 'copd', 'asthma', 'inmsupr', 'hypertension',
       'other_disease', 'cardiovascular', 'obesity', 'renal_chronic',
       'tobacco', 'contact_other_covid','covid_res', 'icu')]
```

```r
train$covid_res[train$covid_res == 1] <- as.integer(0)
train$covid_res[train$covid_res == 2] <- as.integer(1)
train$covid_res[train$covid_res == 3] <- NA

test$covid_res[test$covid_res == 1] <- as.integer(0)
test$covid_res[test$covid_res == 2] <- as.integer(1)
test$covid_res[test$covid_res == 3] <- NA

train <- na.omit(train)
test <- na.omit(test)



train$intubed[train$intubed == 97] <- 3
train$intubed[train$intubed == 98] <- 3
train$intubed[train$intubed == 99] <- 3

train$pregnancy[train$pregnancy == 97] <- 3
train$pregnancy[train$pregnancy == 98] <- 3
train$pregnancy[train$pregnancy == 99] <- 3

train$diabetes[train$diabetes == 97] <- 3
train$diabetes[train$diabetes == 98] <- 3
train$diabetes[train$diabetes == 99] <- 3

train$copd[train$copd == 97] <- 3
train$copd[train$copd == 98] <- 3
train$copd[train$copd == 99] <- 3

train$asthma[train$asthma == 97] <- 3
train$asthma[train$asthma == 98] <- 3
train$asthma[train$asthma == 99] <- 3

train$inmsupr[train$inmsupr == 97] <- 3
train$inmsupr[train$inmsupr == 98] <- 3
train$inmsupr[train$inmsupr == 99] <- 3

train$hypertension[train$hypertension == 97] <- 3
train$hypertension[train$hypertension == 98] <- 3
train$hypertension[train$hypertension == 99] <- 3

train$other_disease[train$other_disease == 97] <- 3
train$other_disease[train$other_disease == 98] <- 3
train$other_disease[train$other_disease == 99] <- 3

train$cardiovascular[train$cardiovascular == 97] <- 3
train$cardiovascular[train$cardiovascular == 98] <- 3
train$cardiovascular[train$cardiovascular == 99] <- 3

train$obesity[train$obesity == 97] <- 3
train$obesity[train$obesity == 98] <- 3
train$obesity[train$obesity == 99] <- 3
```

```
train$renal_chronic[train$renal_chronic == 97] <- 3
train$renal_chronic[train$renal_chronic == 98] <- 3
train$renal_chronic[train$renal_chronic == 99] <- 3

train$tobacco[train$tobacco == 97] <- 3
train$tobacco[train$tobacco == 98] <- 3
train$tobacco[train$tobacco == 99] <- 3

train$contact_other_covid[train$contact_other_covid == 97] <- 3
train$contact_other_covid[train$contact_other_covid == 98] <- 3
train$contact_other_covid[train$contact_other_covid == 99] <- 3

train$icu[train$icu == 97] <- 3
train$icu[train$icu == 98] <- 3
train$icu[train$icu == 99] <- 3
```

```
test$intubed[test$intubed == 97] <- 3
test$intubed[test$intubed == 98] <- 3
test$intubed[test$intubed == 99] <- 3

test$pregnancy[test$pregnancy == 97] <- 3
test$pregnancy[test$pregnancy == 98] <- 3
test$pregnancy[test$pregnancy == 99] <- 3

test$diabetes[test$diabetes == 97] <- 3
test$diabetes[test$diabetes == 98] <- 3
test$diabetes[test$diabetes == 99] <- 3

test$copd[test$copd == 97] <- 3
test$copd[test$copd == 98] <- 3
test$copd[test$copd == 99] <- 3

test$asthma[test$asthma == 97] <- 3
test$asthma[test$asthma == 98] <- 3
test$asthma[test$asthma == 99] <- 3

test$inmsupr[test$inmsupr == 97] <- 3
test$inmsupr[test$inmsupr == 98] <- 3
test$inmsupr[test$inmsupr == 99] <- 3

test$hypertension[test$hypertension == 97] <- 3
test$hypertension[test$hypertension == 98] <- 3
test$hypertension[test$hypertension == 99] <- 3

test$other_disease[test$other_disease == 97] <- 3
test$other_disease[test$other_disease == 98] <- 3
test$other_disease[test$other_disease == 99] <- 3

test$cardiovascular[test$cardiovascular == 97] <- 3
test$cardiovascular[test$cardiovascular == 98] <- 3
test$cardiovascular[test$cardiovascular == 99] <- 3

test$obesity[test$obesity == 97] <- 3
```

```
test$obesity[test$obesity == 98] <- 3
test$obesity[test$obesity == 99] <- 3

test$renal_chronic[test$renal_chronic == 97] <- 3
test$renal_chronic[test$renal_chronic == 98] <- 3
test$renal_chronic[test$renal_chronic == 99] <- 3

test$tobacco[test$tobacco == 97] <- 3
test$tobacco[test$tobacco == 98] <- 3
test$tobacco[test$tobacco == 99] <- 3

test$contact_other_covid[test$contact_other_covid == 97] <- 3
test$contact_other_covid[test$contact_other_covid == 98] <- 3
test$contact_other_covid[test$contact_other_covid == 99] <- 3

test$icu[test$icu == 97] <- 3
test$icu[test$icu == 98] <- 3
test$icu[test$icu == 99] <- 3
```

```
x_train <- subset(train, select = -c(covid_res))
y_train <- subset(train, select = c(covid_res))

x_test <- subset(test, select = -c(covid_res))
y_test <- subset(test, select = c(covid_res))

x_train <- data.frame(scale(x_train))

x_test <- data.frame(scale(x_test))

train_ <- cbind(x_train,y_train)
test_  <- cbind(x_test,y_test)
```

Seja $Y_i, i = 1, \ldots, n$ a variável binária definida por

$$Y_i = \begin{cases} 1, & \textit{tem covid} \\ 0, & \textit{caso contrário} \end{cases}$$

com $n = 1427$ sendo o número de pacientes.

Para a formulação de nosso modelo nós assumimos que esta variável segue uma distribuição de Bernoulli denotada por $Y_i \sim Bernoulli(\mu_i)$, a qual assume dois valores 0 e 1, sendo 1 para caso de cavid e 0 caso contrário, com probabilidade $\mu_i \in [0, 1]$.

Sabemos que para uma resposta binária temos que, $E(Y_i) = \sum_{y=0}^{1} P(Y_i = y_i)y = 1 \times P(Y_i = 1) + 0 \times P(Y_i = 0) = \mu_i \in (0, 1)$. Então, temos interesse em estimar $\widehat{\mu}_i$, em que $y = 1$ significa que o paciente possui covid.

Assim, o modelo de regressão binária diz que,

$$Y_i \sim Bernoulli(\mu_i)$$

com

$$\mu_i = F(\eta_i) = F(x_i^T \beta), \ i = 1, \ldots, n$$

em que

- $x_i = (x_{i1}, x_{i2}, \ldots, x_{ik})^T$ vetor com as variáveis explicativas, tal que x_{i1} é o intercepto
- $\beta = (\beta_1, \ldots, \beta_k)^T$ vetor com k coeficientes de regressão, sendo um a mais do que o número de covariáveis.
- onde $F(.)$ é a função de distribuição acumulada com suporte na reta e $F^{-1}(.)$ é a função de ligação.

O modelo proposto é chamado modelo de regressão binária. Este modelo é um modelo de classificação que faz parte dos chamados modelos lineares generalizados. Existem também outros modelos de classificação no aprendizado supervisionado.

Especificamente, vamos considerar o modelo de regressão binária com uma função de ligação logito.

- Componente aleatório: $y_1, ..., y_{1427}$ é uma amostra aleatória $Y_i \sim Bernoulli(\mu_i)$
- Componente sistemático: $\eta_i = \beta_0 + \beta_1 \cdot sex + \beta_2 \cdot patient + \beta_3 \cdot intubed...$
- Função de ligação $\text{logit}(\mu_i) = ln\left(\frac{\mu_i}{1-\mu_i}\right)$ ou

$$\underbrace{\eta_i = log\Big(\frac{\mu_i}{1-\mu_i}\Big)}_{\text{Função de ligação logito}} = \underbrace{x_i^T\beta}_{\text{Preditor linear}}, \ i = 1, \ldots, 4601$$

Considerando a função distribuição acumulada da distribuição logística, tem-se a função de ligação logito, ou seja, em que o modelo é

$$\mu_i = F(\eta_i) = \frac{exp(\eta_i)}{1 + exp(\eta_i)} = \frac{exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 ...)}{1 + exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 ...)}$$

O nosso interesse é utilizar o chamado modelo de regressão binária para modelar $\mu_i = E(Y_i|X), i = 1, \ldots, n$ e estimar os coeficientes de regressão β associados com as variáveis explicativas considerando uma determinada função de ligação.

```
fit.model<-glm(formula = covid_res ~ sex +patient_type+ intubed +pneumonia +  age + pregnancy+
                         diabetes+ copd +asthma+ inmsupr+ hypertension+
                         other_disease +cardiovascular+ obesity+ renal_chronic+
                         tobacco +contact_other_covid+ icu, family = binomial(link = "logit"), data = tra
summary(fit.model)
```

```
##
## Call:
## glm(formula = covid_res ~ sex + patient_type + intubed + pneumonia +
##     age + pregnancy + diabetes + copd + asthma + inmsupr + hypertension +
##     other_disease + cardiovascular + obesity + renal_chronic +
##     tobacco + contact_other_covid + icu, family = binomial(link = "logit"),
##     data = train_)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1303  -1.2166   0.8439   1.0239   1.9112
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          0.20743    0.05544   3.742 0.000183 ***
```

```
## sex                 -0.29493    0.35505  -0.831 0.406166
## patient_type        -0.13603    0.24496  -0.555 0.578664
## intubed             -0.10029    0.21849  -0.459 0.646230
## pneumonia            0.28313    0.07823   3.619 0.000296 ***
## age                 -0.14512    0.06403  -2.266 0.023436 *
## pregnancy            0.21336    0.35433   0.602 0.547078
## diabetes             0.01546    0.06258   0.247 0.804852
## copd                -0.15548    0.06546  -2.375 0.017545 *
## asthma              -0.02879    0.05861  -0.491 0.623360
## inmsupr             -0.15679    0.06566  -2.388 0.016946 *
## hypertension        -0.03368    0.06504  -0.518 0.604586
## other_disease       -0.09526    0.06032  -1.579 0.114237
## cardiovascular       0.05569    0.06268   0.889 0.374253
## obesity              0.10500    0.05684   1.847 0.064702 .
## renal_chronic       -0.01885    0.06178  -0.305 0.760300
## tobacco             -0.09050    0.05769  -1.569 0.116718
## contact_other_covid -0.20550    0.05793  -3.547 0.000389 ***
## icu                  0.09419    0.22622   0.416 0.677142
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1963.5  on 1426  degrees of freedom
## Residual deviance: 1852.5  on 1408  degrees of freedom
## AIC: 1890.5
##
## Number of Fisher Scoring iterations: 4
```

```
(IC1 <- confint.default(fit.model, level=0.95))
```

```
##                            2.5 %      97.5 %
## (Intercept)          0.098770569  0.31608501
## sex                 -0.990814827  0.40096089
## patient_type        -0.616137918  0.34407179
## intubed             -0.528521534  0.32794505
## pneumonia            0.129798470  0.43646171
## age                 -0.270620136 -0.01961310
## pregnancy           -0.481121564  0.90784183
## diabetes            -0.107193826  0.13811770
## copd                -0.283792985 -0.02717537
## asthma              -0.143668446  0.08609768
## inmsupr             -0.285484318 -0.02809742
## hypertension        -0.161163951  0.09380252
## other_disease       -0.213479422  0.02295173
## cardiovascular      -0.067151856  0.17852995
## obesity             -0.006403195  0.21640135
## renal_chronic       -0.139933174  0.10223704
## tobacco             -0.203569429  0.02257225
## contact_other_covid -0.319047424 -0.09195948
## icu                 -0.349188538  0.53756732
```

Observamos que os intervalos de confiança dos coeficientes para estimar alguns betas contém o valor zero, o que confirma que essas covariáveis não são significativas

Verificando a significância das variáveis utilizando um teste alternativo baseado na análise de deviance usando a estatística qui-quadrado.

```
# Teste chisq para o modelo 1
anova(fit.model,test = 'Chisq')
```

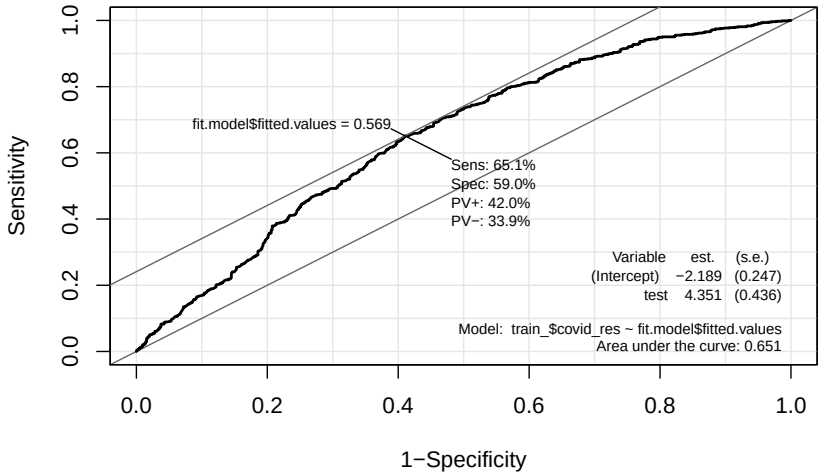
```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: covid_res
##
## Terms added sequentially (first to last)
##
##
##                     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                 1426     1963.5
## sex                  1    4.300      1425     1959.2 0.0381115 *
## patient_type         1   45.369      1424     1913.8 1.632e-11 ***
## intubed              1    0.000      1423     1913.8 0.9912819
## pneumonia            1   18.152      1422     1895.7 2.040e-05 ***
## age                  1    5.473      1421     1890.2 0.0193114 *
## pregnancy            1    0.634      1420     1889.6 0.4258980
## diabetes             1    0.001      1419     1889.5 0.9732036
## copd                 1    7.979      1418     1881.6 0.0047321 **
## asthma               1    0.283      1417     1881.3 0.5949703
## inmsupr              1    6.559      1416     1874.7 0.0104355 *
## hypertension         1    0.005      1415     1874.7 0.9453351
## other_disease        1    1.881      1414     1872.8 0.1702583
## cardiovascular       1    0.728      1413     1872.1 0.3934782
## obesity              1    3.675      1412     1868.5 0.0552471 .
## renal_chronic        1    0.121      1411     1868.3 0.7282748
## tobacco              1    3.125      1410     1865.2 0.0770824 .
## contact_other_covid  1   12.530      1409     1852.7 0.0004004 ***
## icu                  1    0.175      1408     1852.5 0.6760514
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notamos o mesmo que no teste anterior.

análises preditivas

Usando os seguintes comandos obtemos as Curvas ROC (receiver operating characteristic) para o modelo

```
ROC(fit.model$fitted.values, train_$covid_res, plot= "ROC")
```

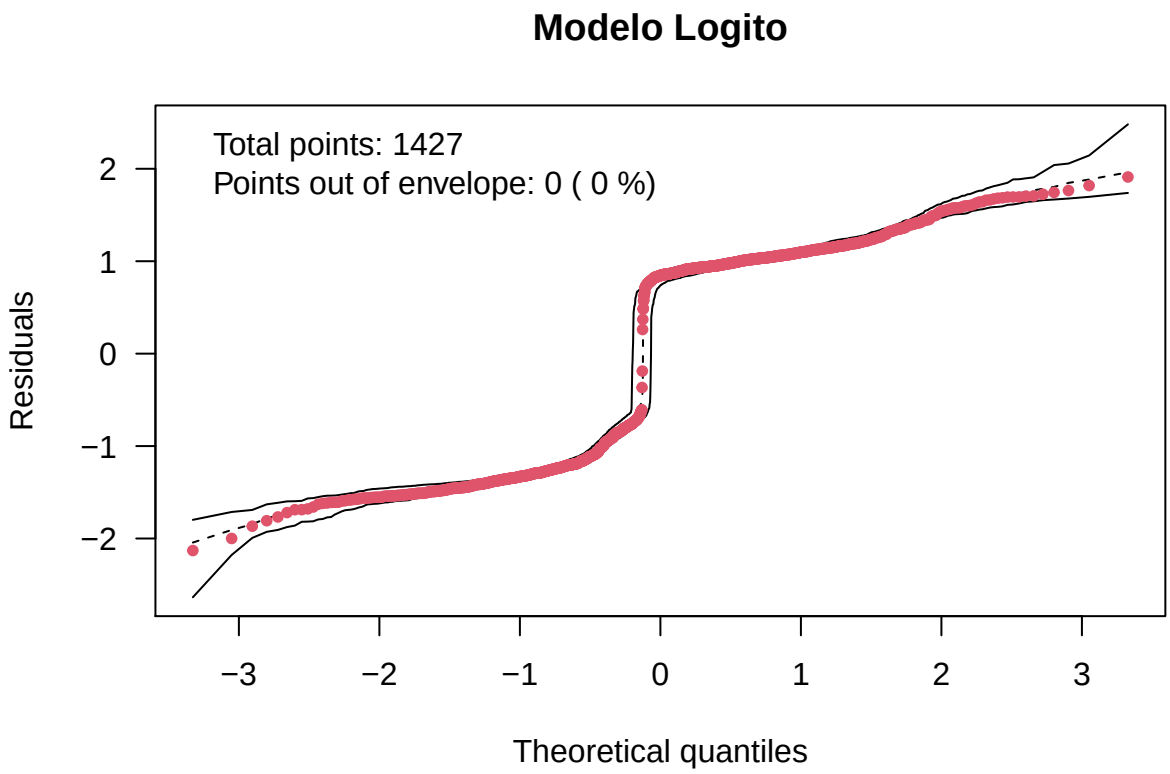


Obtivemos uma área abaixo da curva de 0.646, uma sensibilidade de 68,8% e uma especificidade de 54,8%.

```
hnp.fit.model = hnp(fit.model, print.on=TRUE, plot=FALSE,
halfnormal=F)
```

```
## Binomial model
```

```
plot(hnp.fit.model,main="Modelo Logito",las=1,pch=20,cex=1,col=c(1,1,1,2))
```



Conforme observado no gráfico acima, nenhum ponto está fora dos limites do envelope, o que indica bom ajuste dos dados ao modelo.

Usando outras funções de ligação

```
fit.modelp<- glm(formula = covid_res ~  sex +patient_type+ intubed +pneumonia +  age + pregnancy+
                        diabetes+ copd +asthma+ inmsupr+ hypertension+
                        other_disease +cardiovascular+ obesity+ renal_chronic+
                        tobacco +contact_other_covid+ icu,
    family = binomial(link = "probit"), data = train_)

summary(fit.modelp)
```

```
##
## Call:
## glm(formula = covid_res ~ sex + patient_type + intubed + pneumonia +
##     age + pregnancy + diabetes + copd + asthma + inmsupr + hypertension +
##     other_disease + cardiovascular + obesity + renal_chronic +
##     tobacco + contact_other_covid + icu, family = binomial(link = "probit"),
##     data = train_)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1558  -1.2166   0.8476   1.0277   1.9099
```

```
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          0.12836    0.03408   3.766 0.000166 ***
## sex                 -0.17042    0.21881  -0.779 0.436055
## patient_type        -0.07765    0.14866  -0.522 0.601448
## intubed             -0.06196    0.13286  -0.466 0.640960
## pneumonia            0.17459    0.04801   3.636 0.000277 ***
## age                 -0.08895    0.03934  -2.261 0.023754 *
## pregnancy            0.12037    0.21849   0.551 0.581682
## diabetes             0.01110    0.03843   0.289 0.772655
## copd                -0.09275    0.03908  -2.373 0.017631 *
## asthma              -0.01830    0.03584  -0.511 0.609592
## inmsupr             -0.09235    0.03912  -2.361 0.018231 *
## hypertension        -0.02169    0.04000  -0.542 0.587549
## other_disease       -0.05828    0.03665  -1.590 0.111776
## cardiovascular       0.03216    0.03815   0.843 0.399272
## obesity              0.06332    0.03503   1.807 0.070720 .
## renal_chronic       -0.01331    0.03778  -0.352 0.724610
## tobacco             -0.05410    0.03510  -1.541 0.123246
## contact_other_covid -0.12682    0.03573  -3.549 0.000386 ***
## icu                  0.06430    0.13659   0.471 0.637842
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1963.5  on 1426  degrees of freedom
## Residual deviance: 1852.9  on 1408  degrees of freedom
## AIC: 1890.9
##
## Number of Fisher Scoring iterations: 4
```

```
fit.modelc<- glm(formula = covid_res ~  sex +patient_type+ intubed +pneumonia +  age + pregnancy+
                        diabetes+ copd +asthma+ inmsupr+ hypertension+
                        other_disease +cardiovascular+ obesity+ renal_chronic+
                        tobacco +contact_other_covid+ icu,
    family = binomial(link = "cauchit"), data = train_)

summary(fit.modelc)
```

```
##
## Call:
## glm(formula = covid_res ~ sex + patient_type + intubed + pneumonia +
##     age + pregnancy + diabetes + copd + asthma + inmsupr + hypertension +
##     other_disease + cardiovascular + obesity + renal_chronic +
##     tobacco + contact_other_covid + icu, family = binomial(link = "cauchit"),
##     data = train_)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9694  -1.2118   0.8412   1.0102   1.8580
##
## Coefficients:
```

```
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          0.1762487  0.0495556   3.557 0.000376 ***
## sex                 -0.3654923  0.3144312  -1.162 0.245076
## patient_type        -0.1532621  0.2327423  -0.659 0.510213
## intubed             -0.0878964  0.2055430  -0.428 0.668920
## pneumonia            0.2474539  0.0711192   3.479 0.000502 ***
## age                 -0.1324769  0.0568230  -2.331 0.019732 *
## pregnancy            0.2920372  0.3129071   0.933 0.350663
## diabetes             0.0007159  0.0553920   0.013 0.989689
## copd                -0.1519265  0.0662564  -2.293 0.021848 *
## asthma              -0.0230060  0.0528807  -0.435 0.663523
## inmsupr             -0.1622614  0.0689954  -2.352 0.018684 *
## hypertension        -0.0225802  0.0571415  -0.395 0.692722
## other_disease       -0.0769741  0.0557783  -1.380 0.167586
## cardiovascular       0.0604637  0.0579677   1.043 0.296921
## obesity              0.0985280  0.0490559   2.008 0.044592 *
## renal_chronic       -0.0067092  0.0557264  -0.120 0.904171
## tobacco             -0.0883233  0.0547167  -1.614 0.106486
## contact_other_covid -0.1792520  0.0504594  -3.552 0.000382 ***
## icu                  0.0468036  0.2181863   0.215 0.830148
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1963.5  on 1426  degrees of freedom
## Residual deviance: 1850.5  on 1408  degrees of freedom
## AIC: 1888.5
##
## Number of Fisher Scoring iterations: 9
```

```
fit.modelcl<- glm(formula = covid_res ~  sex +patient_type+ intubed +pneumonia +  age + pregnancy+
                        diabetes+ copd +asthma+ inmsupr+ hypertension+
                        other_disease +cardiovascular+ obesity+ renal_chronic+
                        tobacco +contact_other_covid+ icu,
    family = binomial(link = "cloglog"), data = train_)

summary(fit.modelcl)
```

```
##
## Call:
## glm(formula = covid_res ~ sex + patient_type + intubed + pneumonia +
##     age + pregnancy + diabetes + copd + asthma + inmsupr + hypertension +
##     other_disease + cardiovascular + obesity + renal_chronic +
##     tobacco + contact_other_covid + icu, family = binomial(link = "cloglog"),
##     data = train_)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2096  -1.2120   0.8478   1.0312   1.8178
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -0.24492    0.03815  -6.420 1.36e-10 ***
```

```
## sex                 -0.13908    0.23248  -0.598 0.549660
## patient_type        -0.15125    0.18201  -0.831 0.405973
## intubed             -0.08272    0.16461  -0.503 0.615279
## pneumonia            0.22137    0.05946   3.723 0.000197 ***
## age                 -0.09097    0.04308  -2.112 0.034701 *
## pregnancy            0.08207    0.23237   0.353 0.723961
## diabetes             0.01549    0.04425   0.350 0.726312
## copd                -0.09254    0.03946  -2.345 0.019007 *
## asthma              -0.02161    0.03683  -0.587 0.557364
## inmsupr             -0.08644    0.03969  -2.178 0.029408 *
## hypertension        -0.03303    0.04529  -0.729 0.465835
## other_disease       -0.06013    0.03894  -1.544 0.122612
## cardiovascular       0.03762    0.04627   0.813 0.416198
## obesity              0.07412    0.04010   1.848 0.064549 .
## renal_chronic       -0.01351    0.04224  -0.320 0.749008
## tobacco             -0.04915    0.03666  -1.341 0.179962
## contact_other_covid -0.13648    0.03900  -3.499 0.000467 ***
## icu                  0.01132    0.16300   0.069 0.944613
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1963.5  on 1426  degrees of freedom
## Residual deviance: 1855.4  on 1408  degrees of freedom
## AIC: 1893.4
##
## Number of Fisher Scoring iterations: 10
```

```
# Geradora para a função de ligação loglog
loglog <- function( ) structure(list(
  linkfun = function(mu) -log(-log(mu)),
  linkinv = function(eta)
    pmax(pmin(exp(-exp(-eta)), 1 - .Machine$double.eps),
         .Machine$double.eps),
  mu.eta = function(eta) {
    eta <- pmin(eta, 700)
    pmax(exp(-eta - exp(-eta)), .Machine$double.eps)
  },
  dmu.deta = function(eta)
    pmax(exp(-exp(-eta) - eta) * expm1(-eta),
         .Machine$double.eps),
  valideta = function(eta) TRUE,
  name = "loglog"
), class = "link-glm")
fit.modelll<- glm(formula = covid_res ~  sex +patient_type+ intubed +pneumonia +  age + pregnancy+
                        diabetes+ copd +asthma+ inmsupr+ hypertension+
                        other_disease +cardiovascular+ obesity+ renal_chronic+
                        tobacco +contact_other_covid+ icu,
    family = binomial(link = loglog()), data = train_)

summary(fit.modelll)
```

```
##
```

```
## Call:
## glm(formula = covid_res ~ sex + patient_type + intubed + pneumonia +
##     age + pregnancy + diabetes + copd + asthma + inmsupr + hypertension +
##     other_disease + cardiovascular + obesity + renal_chronic +
##     tobacco + contact_other_covid + icu, family = binomial(link = loglog()),
##     data = train_)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0662  -1.2241   0.8571   1.0208   1.9814
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          0.54327    0.04172  13.020  < 2e-16 ***
## sex                 -0.29129    0.25549  -1.140  0.25423
## patient_type        -0.03305    0.15182  -0.218  0.82766
## intubed             -0.06386    0.13711  -0.466  0.64140
## pneumonia            0.18263    0.05168   3.534  0.00041 ***
## age                 -0.11249    0.04677  -2.405  0.01617 *
## pregnancy            0.23498    0.25411   0.925  0.35511
## diabetes             0.01320    0.04308   0.306  0.75928
## copd                -0.11777    0.05188  -2.270  0.02322 *
## asthma              -0.01866    0.04541  -0.411  0.68115
## inmsupr             -0.12582    0.05184  -2.427  0.01522 *
## hypertension        -0.01748    0.04538  -0.385  0.70006
## other_disease       -0.07206    0.04465  -1.614  0.10659
## cardiovascular       0.03327    0.04090   0.813  0.41604
## obesity              0.06942    0.03995   1.738  0.08225 .
## renal_chronic       -0.02134    0.04394  -0.486  0.62714
## tobacco             -0.07465    0.04426  -1.687  0.09166 .
## contact_other_covid -0.15368    0.04298  -3.575  0.00035 ***
## icu                  0.13004    0.14096   0.923  0.35624
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1963.5  on 1426  degrees of freedom
## Residual deviance: 1850.9  on 1408  degrees of freedom
## AIC: 1888.9
##
## Number of Fisher Scoring iterations: 6
```

```
# Dataframe para verificar o AIC
data.frame(Modelo=c("Modelo logito","Modelo probito","Modelo cauchito","Modelo cloglog","Modelo loglog"),
           AIC = c(AIC(fit.model),AIC(fit.modelp),AIC(fit.modelc),
                   AIC(fit.modelcl), AIC(fit.modelll)))
```

```
##            Modelo      AIC
## 1   Modelo logito 1890.494
## 2  Modelo probito 1890.901
## 3 Modelo cauchito 1888.486
## 4  Modelo cloglog 1893.414
## 5   Modelo loglog 1888.857
```

Portanto, escolhemos o modelo de regressão binária com ligação cauchito por porque seu AIC foi o menor dentre todos os outros modelos.

Seleção de variáveis

vamos fazer uma análise de seleção de variáveis utilizando a função do R stepAIC, que nos ajuda a detectar os melhores preditores.

Utilizando a função stepAIC, temos:

```
# stepAIC
stepAIC(fit.modelc)
```

```
## Start:  AIC=1888.49
## covid_res ~ sex + patient_type + intubed + pneumonia + age +
##     pregnancy + diabetes + copd + asthma + inmsupr + hypertension +
##     other_disease + cardiovascular + obesity + renal_chronic +
##     tobacco + contact_other_covid + icu
##
##                       Df Deviance    AIC
## - diabetes             1   1850.5 1886.5
## - renal_chronic        1   1850.5 1886.5
## - icu                  1   1850.5 1886.5
## - hypertension         1   1850.7 1886.7
## - asthma               1   1850.7 1886.7
## - intubed              1   1850.7 1886.7
## - patient_type         1   1851.0 1887.0
## - pregnancy            1   1851.2 1887.2
## - sex                  1   1851.6 1887.6
## - cardiovascular       1   1851.6 1887.6
## <none>                     1850.5 1888.5
## - other_disease        1   1852.6 1888.6
## - tobacco              1   1853.1 1889.1
## - obesity              1   1854.5 1890.5
## - age                  1   1856.1 1892.1
## - inmsupr              1   1856.9 1892.9
## - copd                 1   1856.9 1892.9
## - contact_other_covid  1   1863.4 1899.4
## - pneumonia            1   1863.7 1899.7
##
## Step:  AIC=1886.49
## covid_res ~ sex + patient_type + intubed + pneumonia + age +
##     pregnancy + copd + asthma + inmsupr + hypertension + other_disease +
##     cardiovascular + obesity + renal_chronic + tobacco + contact_other_covid +
##     icu
##
##                       Df Deviance    AIC
## - renal_chronic        1   1850.5 1884.5
## - icu                  1   1850.5 1884.5
## - hypertension         1   1850.7 1884.7
## - asthma               1   1850.7 1884.7
## - intubed              1   1850.7 1884.7
## - patient_type         1   1851.0 1885.0
```

```
## - pregnancy            1   1851.2 1885.2
## - sex                  1   1851.6 1885.6
## - cardiovascular       1   1851.6 1885.6
## <none>                     1850.5 1886.5
## - other_disease        1   1852.6 1886.6
## - tobacco              1   1853.1 1887.1
## - obesity              1   1854.5 1888.5
## - age                  1   1856.2 1890.2
## - inmsupr              1   1856.9 1890.9
## - copd                 1   1856.9 1890.9
## - contact_other_covid  1   1863.4 1897.4
## - pneumonia            1   1863.8 1897.8
##
## Step:  AIC=1884.5
## covid_res ~ sex + patient_type + intubed + pneumonia + age +
##     pregnancy + copd + asthma + inmsupr + hypertension + other_disease +
##     cardiovascular + obesity + tobacco + contact_other_covid +
##     icu
##
##                       Df Deviance    AIC
## - icu                  1   1850.6 1882.6
## - hypertension         1   1850.7 1882.7
## - intubed              1   1850.7 1882.7
## - asthma               1   1850.7 1882.7
## - patient_type         1   1851.0 1883.0
## - pregnancy            1   1851.2 1883.2
## - sex                  1   1851.6 1883.6
## - cardiovascular       1   1851.7 1883.7
## <none>                     1850.5 1884.5
## - other_disease        1   1852.6 1884.6
## - tobacco              1   1853.1 1885.1
## - obesity              1   1854.5 1886.5
## - age                  1   1856.2 1888.2
## - inmsupr              1   1857.0 1889.0
## - copd                 1   1857.1 1889.1
## - contact_other_covid  1   1863.4 1895.4
## - pneumonia            1   1864.1 1896.1
##
## Step:  AIC=1882.56
## covid_res ~ sex + patient_type + intubed + pneumonia + age +
##     pregnancy + copd + asthma + inmsupr + hypertension + other_disease +
##     cardiovascular + obesity + tobacco + contact_other_covid
##
##                       Df Deviance    AIC
## - intubed              1   1850.7 1880.7
## - hypertension         1   1850.8 1880.8
## - asthma               1   1850.8 1880.8
## - pregnancy            1   1851.3 1881.3
## - patient_type         1   1851.5 1881.5
## - sex                  1   1851.6 1881.6
## - cardiovascular       1   1851.7 1881.7
## <none>                     1850.6 1882.6
## - other_disease        1   1852.6 1882.6
## - tobacco              1   1853.2 1883.2
```

```
## - obesity              1   1854.5 1884.5
## - age                  1   1856.2 1886.2
## - inmsupr              1   1857.0 1887.0
## - copd                 1   1857.2 1887.2
## - contact_other_covid  1   1863.4 1893.4
## - pneumonia            1   1864.1 1894.1
##
## Step:  AIC=1880.71
## covid_res ~ sex + patient_type + pneumonia + age + pregnancy +
##     copd + asthma + inmsupr + hypertension + other_disease +
##     cardiovascular + obesity + tobacco + contact_other_covid
##
##                       Df Deviance    AIC
## - hypertension         1   1850.9 1878.9
## - asthma               1   1850.9 1878.9
## - pregnancy            1   1851.4 1879.4
## - sex                  1   1851.8 1879.8
## - cardiovascular       1   1852.0 1880.0
## <none>                     1850.7 1880.7
## - other_disease        1   1852.8 1880.8
## - tobacco              1   1853.3 1881.3
## - patient_type         1   1853.6 1881.6
## - obesity              1   1854.7 1882.7
## - age                  1   1856.5 1884.5
## - inmsupr              1   1857.0 1885.0
## - copd                 1   1857.3 1885.3
## - contact_other_covid  1   1863.8 1891.8
## - pneumonia            1   1864.1 1892.1
##
## Step:  AIC=1878.91
## covid_res ~ sex + patient_type + pneumonia + age + pregnancy +
##     copd + asthma + inmsupr + other_disease + cardiovascular +
##     obesity + tobacco + contact_other_covid
##
##                       Df Deviance    AIC
## - asthma               1   1851.1 1877.1
## - pregnancy            1   1851.6 1877.6
## - sex                  1   1852.0 1878.0
## - cardiovascular       1   1852.1 1878.1
## <none>                     1850.9 1878.9
## - other_disease        1   1853.0 1879.0
## - tobacco              1   1853.5 1879.5
## - patient_type         1   1853.7 1879.7
## - obesity              1   1854.7 1880.7
## - age                  1   1856.6 1882.6
## - inmsupr              1   1857.5 1883.5
## - copd                 1   1857.7 1883.7
## - contact_other_covid  1   1863.9 1889.9
## - pneumonia            1   1864.3 1890.3
##
## Step:  AIC=1877.11
## covid_res ~ sex + patient_type + pneumonia + age + pregnancy +
##     copd + inmsupr + other_disease + cardiovascular + obesity +
##     tobacco + contact_other_covid
```

```
##
##                       Df Deviance    AIC
## - pregnancy            1   1851.8 1875.8
## - sex                  1   1852.2 1876.2
## - cardiovascular       1   1852.3 1876.3
## <none>                     1851.1 1877.1
## - other_disease        1   1853.2 1877.2
## - tobacco              1   1853.7 1877.7
## - patient_type         1   1853.9 1877.9
## - obesity              1   1854.8 1878.8
## - age                  1   1856.9 1880.9
## - inmsupr              1   1857.7 1881.7
## - copd                 1   1858.0 1882.0
## - contact_other_covid  1   1864.0 1888.0
## - pneumonia            1   1864.6 1888.6
##
## Step:  AIC=1875.81
## covid_res ~ sex + patient_type + pneumonia + age + copd + inmsupr +
##     other_disease + cardiovascular + obesity + tobacco + contact_other_covid
##
##                       Df Deviance    AIC
## - cardiovascular       1   1853.0 1875.0
## <none>                     1851.8 1875.8
## - other_disease        1   1854.0 1876.0
## - sex                  1   1854.1 1876.1
## - tobacco              1   1854.4 1876.4
## - patient_type         1   1854.7 1876.7
## - obesity              1   1855.5 1877.5
## - age                  1   1857.4 1879.4
## - inmsupr              1   1858.4 1880.4
## - copd                 1   1858.7 1880.7
## - contact_other_covid  1   1865.0 1887.0
## - pneumonia            1   1865.2 1887.2
##
## Step:  AIC=1874.96
## covid_res ~ sex + patient_type + pneumonia + age + copd + inmsupr +
##     other_disease + obesity + tobacco + contact_other_covid
##
##                       Df Deviance    AIC
## <none>                     1853.0 1875.0
## - other_disease        1   1855.2 1875.2
## - sex                  1   1855.3 1875.3
## - tobacco              1   1855.5 1875.5
## - patient_type         1   1856.3 1876.3
## - obesity              1   1856.6 1876.6
## - age                  1   1859.1 1879.1
## - copd                 1   1859.1 1879.1
## - inmsupr              1   1859.4 1879.4
## - pneumonia            1   1865.8 1885.8
## - contact_other_covid  1   1866.4 1886.4


##
## Call:  glm(formula = covid_res ~ sex + patient_type + pneumonia + age +
##     copd + inmsupr + other_disease + obesity + tobacco + contact_other_covid,
```

```
##     family = binomial(link = "cauchit"), data = train_)
##
## Coefficients:
##         (Intercept)                  sex         patient_type
##             0.17535             -0.07464             -0.12179
##           pneumonia                  age                 copd
##             0.23744             -0.12858             -0.14312
##             inmsupr        other_disease              obesity
##            -0.16313             -0.08028              0.09254
##             tobacco  contact_other_covid
##            -0.08590             -0.18108
##
## Degrees of Freedom: 1426 Total (i.e. Null);  1416 Residual
## Null Deviance:       1963
## Residual Deviance: 1853  AIC: 1875
```

```
fit.model2<- glm(formula = covid_res ~ sex + patient_type + pneumonia + age +
    copd + inmsupr + other_disease + obesity + tobacco + contact_other_covid,
    family = binomial(link = "cauchit"), data = train_)


summary(fit.model2)
```

```
##
## Call:
## glm(formula = covid_res ~ sex + patient_type + pneumonia + age +
##     copd + inmsupr + other_disease + obesity + tobacco + contact_other_covid,
##     family = binomial(link = "cauchit"), data = train_)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9706  -1.2127   0.8472   1.0106   1.7676
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          0.17535    0.04932   3.555 0.000378 ***
## sex                 -0.07464    0.04896  -1.525 0.127360
## patient_type        -0.12179    0.06759  -1.802 0.071533 .
## pneumonia            0.23744    0.06908   3.437 0.000588 ***
## age                 -0.12858    0.05280  -2.435 0.014881 *
## copd                -0.14312    0.06426  -2.227 0.025923 *
## inmsupr             -0.16313    0.06777  -2.407 0.016074 *
## other_disease       -0.08028    0.05545  -1.448 0.147692
## obesity              0.09254    0.04831   1.916 0.055414 .
## tobacco             -0.08590    0.05398  -1.591 0.111553
## contact_other_covid -0.18108    0.04997  -3.624 0.000290 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1963.5  on 1426  degrees of freedom
## Residual deviance: 1853.0  on 1416  degrees of freedom
## AIC: 1875
```

```
##
## Number of Fisher Scoring iterations: 5
```

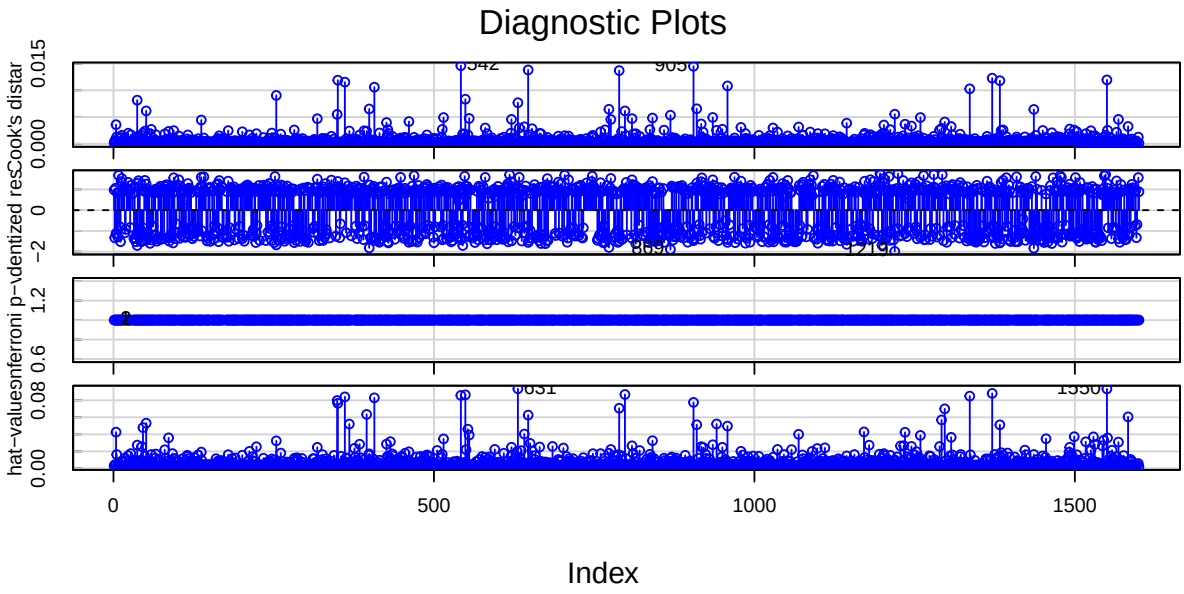
Análise diagnóstica para identificar pontos problemáticos no modelo reduzido

```
#source("diag.bino.txt")
```

Identificação de Pontos problemáticos

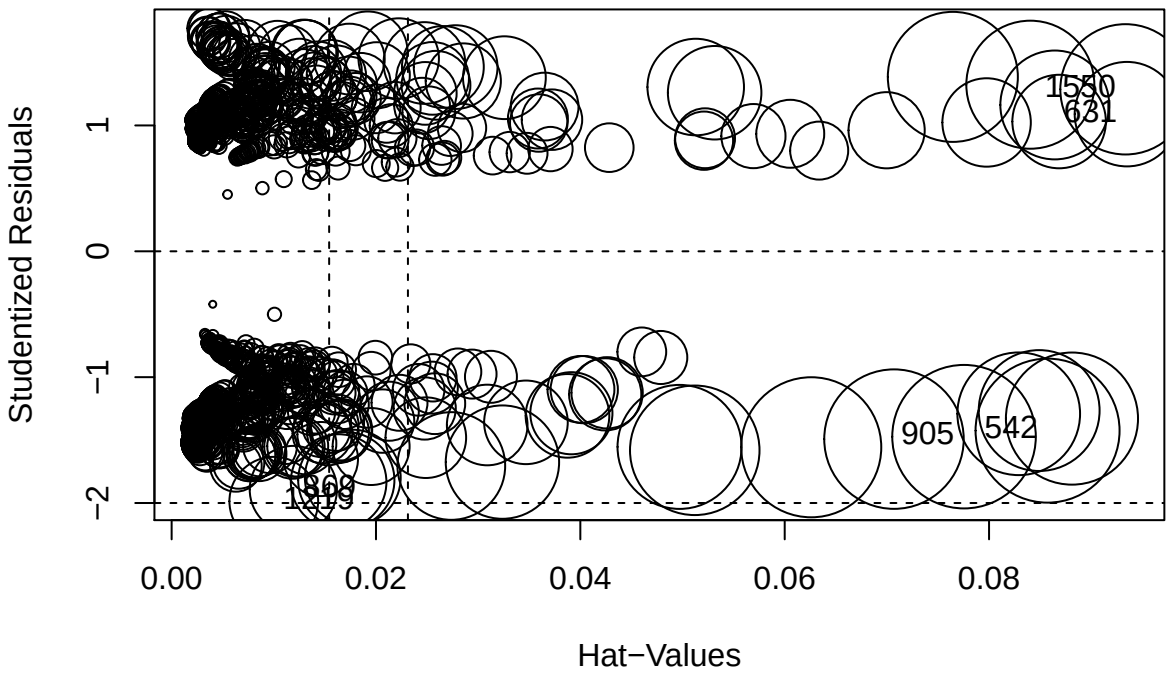
A figura a seguir apresenta diferentes quantidades calculadas para cada uma das observações usando medidas de diagnóstico de pontos influentes usualmente apresentadas nos modelos lineares generalizados. A quantidade "Cook" corresponde a distância de Cook (para detectar pontos influentes), "Studentized" corresponde aos resíduos stutentizados (para detectar homocedasticidade), "Bonf" corresponde aos valores p do teste Bonferroni para outliers e , por fim, "hat" para os valores-hat values (ou pontos de alavanca).

```
# gráfico de influência e alavanca
influenceIndexPlot(fit.model2,col='blue')
```



Para identificar quais são os pontos influentes dentres os apresentados nos 4 gráficos anteriores:

```
influencePlot(fit.model2)
```




```
##        StudRes         Hat       CookD
## 542  -1.424921 0.085726379 0.014588198
## 631   1.089632 0.093495961 0.007680556
## 869  -1.892506 0.011909169 0.005347685
## 905  -1.472888 0.077550829 0.014489825
## 1219 -1.985882 0.009998555 0.005537280
## 1550  1.286957 0.093352110 0.011941960
```

Considerando os valores dos resíduos studentizados, percebemos nenhum ponto se encontram fora do intervalo (-2,2).

Para identificar os pontos influentes, precisamos encontrar aqueles com valor $\widehat{h} > \frac{2p}{n} = \frac{22}{1427} = .0154$, onde $p = 11$ é o número de coeficientes de regressão e $n = 1427$ é o número de observações. Neste caso, identificamos como ponto de alavanca (hat) 542, 631, 905 e 1550, já para os pontos de influência (Distância de Cook): 542, 905 e 1550, e assim, levando em consideração os pontos que têm mais de uma indicação problemática, concluímos que estes pontos requerem uma análise mais detalhada.

```
# Retirada do ponto 542
ajuste1<-glm(formula = covid_res ~  sex + patient_type + pneumonia + age +
    copd + inmsupr + other_disease + obesity + tobacco + contact_other_covid,
                    subset = -c(542),
                    family = binomial(link = "cauchit"), data = train_)

# Retirada do ponto 905
ajuste2<-glm(formula = covid_res ~  sex + patient_type + pneumonia + age +
    copd + inmsupr + other_disease + obesity + tobacco + contact_other_covid,
```

```
                    subset = -c(905),
                    family = binomial(link = "cauchit"), data = train_)

# Retirada do ponto 1550
ajuste3<-glm(formula = covid_res ~  sex + patient_type + pneumonia + age +
    copd + inmsupr + other_disease + obesity + tobacco + contact_other_covid,
                    subset = -c(1550),
                    family = binomial(link = "cauchit"), data = train_)

# Retirada do ponto 542 e 905
ajuste4<-glm(formula = covid_res ~  sex + patient_type + pneumonia + age +
    copd + inmsupr + other_disease + obesity + tobacco + contact_other_covid,
                    subset = -c(542,905),
                    family = binomial(link = "cauchit"), data = train_)
# Retirada do ponto 542 e 1550
ajuste5<-glm(formula = covid_res ~  sex + patient_type + pneumonia + age +
    copd + inmsupr + other_disease + obesity + tobacco + contact_other_covid,
                    subset = -c(542,1550),
                    family = binomial(link = "cauchit"), data = train_)
# Retirada do ponto 905 e 1550
ajuste6<-glm(formula = covid_res ~  sex + patient_type + pneumonia + age +
    copd + inmsupr + other_disease + obesity + tobacco + contact_other_covid,
                    subset = -c(905, 1550),
                    family = binomial(link = "cauchit"), data = train_)

# Retirada do ponto 542, 905 e 1550
ajuste7<-glm(formula = covid_res ~  sex + patient_type + pneumonia + age +
    copd + inmsupr + other_disease + obesity + tobacco + contact_other_covid,
                    subset = -c(542,905, 1550),
                    family = binomial(link = "cauchit"), data = train_)
```

```
compareCoefs(fit.model2,ajuste1, ajuste2, ajuste3, ajuste4, ajuste5,ajuste6,ajuste7)
```

```
## Calls:
## 1: glm(formula = covid_res ~ sex + patient_type + pneumonia + age + copd +
##   inmsupr + other_disease + obesity + tobacco + contact_other_covid, family =
##   binomial(link = "cauchit"), data = train_)
## 2: glm(formula = covid_res ~ sex + patient_type + pneumonia + age + copd +
##   inmsupr + other_disease + obesity + tobacco + contact_other_covid, family =
##   binomial(link = "cauchit"), data = train_, subset = -c(542))
## 3: glm(formula = covid_res ~ sex + patient_type + pneumonia + age + copd +
##   inmsupr + other_disease + obesity + tobacco + contact_other_covid, family =
##   binomial(link = "cauchit"), data = train_, subset = -c(905))
## 4: glm(formula = covid_res ~ sex + patient_type + pneumonia + age + copd +
##   inmsupr + other_disease + obesity + tobacco + contact_other_covid, family =
##   binomial(link = "cauchit"), data = train_, subset = -c(1550))
## 5: glm(formula = covid_res ~ sex + patient_type + pneumonia + age + copd +
##   inmsupr + other_disease + obesity + tobacco + contact_other_covid, family =
##   binomial(link = "cauchit"), data = train_, subset = -c(542, 905))
## 6: glm(formula = covid_res ~ sex + patient_type + pneumonia + age + copd +
##   inmsupr + other_disease + obesity + tobacco + contact_other_covid, family =
##   binomial(link = "cauchit"), data = train_, subset = -c(542, 1550))
## 7: glm(formula = covid_res ~ sex + patient_type + pneumonia + age + copd +
```

```
##   inmsupr + other_disease + obesity + tobacco + contact_other_covid, family =
##   binomial(link = "cauchit"), data = train_, subset = -c(905, 1550))
## 8: glm(formula = covid_res ~ sex + patient_type + pneumonia + age + copd +
##   inmsupr + other_disease + obesity + tobacco + contact_other_covid, family =
##   binomial(link = "cauchit"), data = train_, subset = -c(542, 905, 1550))
##
##                     Model 1 Model 2 Model 3 Model 4 Model 5 Model 6 Model 7
## (Intercept)          0.1754  0.1770  0.1772  0.1754  0.1788  0.1770  0.1772
## SE                   0.0493  0.0494  0.0494  0.0493  0.0495  0.0494  0.0494
##
## sex                 -0.0746 -0.0765 -0.0764 -0.0746 -0.0783 -0.0765 -0.0764
## SE                   0.0490  0.0490  0.0490  0.0490  0.0491  0.0490  0.0490
##
## patient_type        -0.1218 -0.1223 -0.1219 -0.1218 -0.1225 -0.1223 -0.1219
## SE                   0.0676  0.0676  0.0676  0.0676  0.0677  0.0676  0.0676
##
## pneumonia            0.2374  0.2377  0.2379  0.2374  0.2382  0.2377  0.2379
## SE                   0.0691  0.0691  0.0691  0.0691  0.0691  0.0691  0.0691
##
## age                 -0.1286 -0.1292 -0.1296 -0.1286 -0.1302 -0.1292 -0.1296
## SE                   0.0528  0.0529  0.0529  0.0528  0.0529  0.0529  0.0529
##
## copd                -0.1431 -0.1431 -0.1434 -0.1431 -0.1435 -0.1431 -0.1434
## SE                   0.0643  0.0643  0.0643  0.0643  0.0643  0.0643  0.0643
##
## inmsupr             -0.1631 -0.1634 -0.1635 -0.1631 -0.1637 -0.1634 -0.1635
## SE                   0.0678  0.0678  0.0678  0.0678  0.0679  0.0678  0.0678
##
## other_disease       -0.0803 -0.0800 -0.0802 -0.0803 -0.0799 -0.0800 -0.0802
## SE                   0.0555  0.0555  0.0555  0.0555  0.0555  0.0555  0.0555
##
## obesity              0.0925  0.0933  0.0932  0.0925  0.0940  0.0933  0.0932
## SE                   0.0483  0.0483  0.0484  0.0483  0.0484  0.0483  0.0484
##
## tobacco             -0.0859 -0.0857 -0.0858 -0.0859 -0.0856 -0.0857 -0.0858
## SE                   0.0540  0.0540  0.0540  0.0540  0.0541  0.0540  0.0540
##
## contact_other_covid -0.1811 -0.1809 -0.1830 -0.1811 -0.1828 -0.1809 -0.1830
## SE                   0.0500  0.0500  0.0501  0.0500  0.0501  0.0500  0.0501
##
##                     Model 8
## (Intercept)          0.1788
## SE                   0.0495
##
## sex                 -0.0783
## SE                   0.0491
##
## patient_type        -0.1225
## SE                   0.0677
##
## pneumonia            0.2382
## SE                   0.0691
##
## age                 -0.1302
```

```
## SE                   0.0529
##
## copd                -0.1435
## SE                   0.0643
##
## inmsupr             -0.1637
## SE                   0.0679
##
## other_disease       -0.0799
## SE                   0.0555
##
## obesity              0.0940
## SE                   0.0484
##
## tobacco             -0.0856
## SE                   0.0541
##
## contact_other_covid -0.1828
## SE                   0.0501
##
```

Conseguimos perceber que os coeficientes de regressão dos modelos propostos, quando se retiraram os pontos identificados na análise de diagnóstico não mudaram em relação ao modelo com todos os pontos (model1:fit.model2), e as interpretações são mantidas. Assim, mantemos o modelo reduzido como modelo final.

Comparando o AIC dos modelos com retiradas dos pontos:

```
data.frame(
  Modelo= c("Completo", "Removendo 542", "Removendo 905",
            "Removendo 1550", "Removendo 542 e 905","Removendo 542 e 1550",
            "Removendo 905 e 1550","Removendo os 3"),
  AIC = c(AIC(fit.model2),AIC(ajuste1), AIC(ajuste2), AIC(ajuste3),
             AIC(ajuste4), AIC(ajuste5), AIC(ajuste6),
             AIC(ajuste7)))
```

```
##                 Modelo      AIC
## 1             Completo 1874.961
## 2        Removendo 542 1872.976
## 3        Removendo 905 1872.645
## 4       Removendo 1550 1874.961
## 5  Removendo 542 e 905 1870.653
## 6 Removendo 542 e 1550 1872.976
## 7 Removendo 905 e 1550 1872.645
## 8       Removendo os 3 1870.653
```

Ao comparar os AICs, observamos que sempre que removemos algum ponto detectado na análise diagnóstico, obtemos um menor AIC, indicando um melhor modelo, embora a diminuição seja pequena. Nós detectamos que o modelo com menor AIC é aquele que retira todos os três pontos influentes. O AIC do modelo com todos os pontos é 1874.961 e o AIC do modelo removendo os três pontos influentes é 1870.653.

Modelo final e interpretação de parâmetros

Anteriormente, dizemos que o ganho de AIC quando retiramos os 3 pontos problemáticos foi mínimo, dessa forma, escolhemos o modelo reduzido como o mais apropriado para os dados de covid.

Assim o modelo final é

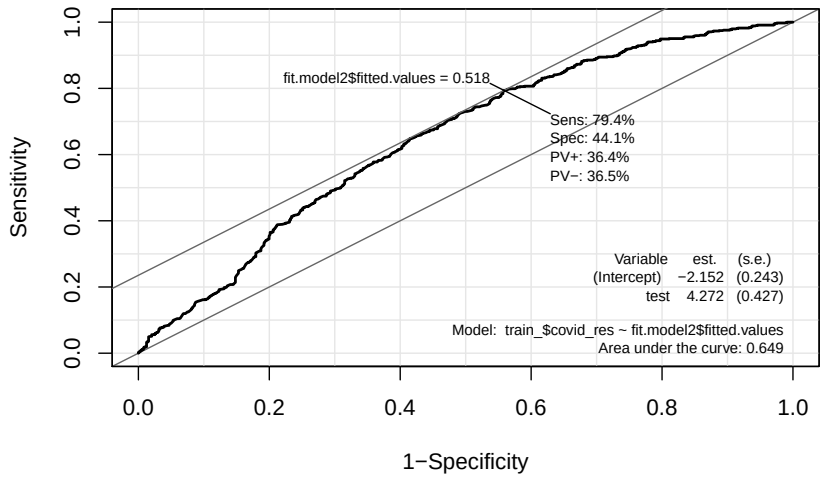
- $y_i|x \overset{\text{iid}}{\sim} Bernoulli(\widehat{\mu}_i)$
- $tan(\pi(\widehat{\mu}_i - 0.5)) = 0.17535301 - 0.07464329 \times sex - 0.12179391 \times patient_type + 0.23743990 \times pneumonia - 0.12858149 \times age - 0.14312439 \times copd - 0.16313370 \times inmsupr - 0.08028254 \times other_disease + 0.09253628 \times obesity - 0.08589672 \times tobacco - 0.18107767 \times contact_other_covid$
  ou
- $\widehat{\mu}_i = 0.5 + \frac{1}{\pi} arctan(0.17535301 - 0.07464329 \times sex - 0.12179391 \times patient_type + 0.23743990 \times pneumonia - 0.12858149 \times age - 0.14312439 \times copd - 0.16313370 \times inmsupr - 0.08028254 \times other_disease + 0.09253628 \times obesity - 0.08589672 \times tobacco - 0.18107767 \times contact_other_covid)$

Notamos que os coeficientes são positivos. Assim, a cada aumento de uma unidade na variável preditiva pneumonia para um efeito zero do resto, há um aumento de $0.5 + atan(0.17535301 + 0.23743990)/pi = 0.625$ na média (probabilidade) da variável resposta (covid).

Em outras palavras podemos dizer que a) isolando o índice de pneumonia, se este se incrementa em uma unidade há uma probabilidade de 62.5% de ser covid. O mesmo raciocínio se aplica às outras variáveis

Uma forma bastante utilizada para determinar o ponto de corte é através da Curva ROC que para o modelo final é

```
ROC(fit.model2$fitted.values,  train_$covid_res, plot= "ROC")
```



Pela análise da curva ROC, escolhemos o ponto de corte referente a combinação da sensibilidade e 1-especificidade que mais se aproxima do canto superior esquerdo do gráfico que neste caso é aproximadamente 0.7.

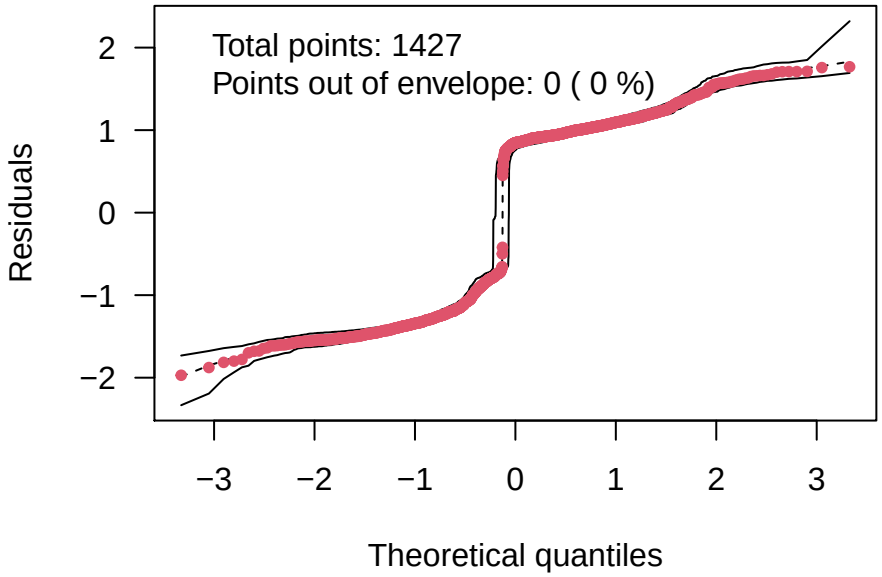
Temos encontrado que área baixo a curva ROC do modelo de 65%, a sensibilidade do modelo é 79.4% (capacidade do modelo classificar um indivíduo com covid dado que realmente ele está com covid) e especificidade de 44.1% (capacidade do modelo predizer um indivíduo sem covid dado que ele realmente não tem covid).

Adicionalmente a análise de Envelope do modelo é .

```
hnp.glm.cauchit <- hnp(fit.model2, print.on=TRUE, plot=FALSE, halfnormal=F)
```

```
## Binomial model
```

```
plot(hnp.glm.cauchit, las=1, pch=20, cex=1, col=c(1,1,1,2))
```



Parece que o modelo está bem ajustado.

#D)

- Concluímos que o melhor modelo de classificação para os dados é o modelo de regressão binária utilizando função de ligação Cauchit sem desconsiderar nenhuma observação. As estatísticas como AIC e curva ROC foram descritas acima.
- Três observações foram identificadas como problemáticas e as análises mostraram que elas podem ser desconsideradas na formulação do modelo porém o ganho em termos de ajuste não foi relevante.

#E)

- Modelo Cauchito

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  60  29
##          1  85 177
##
##                Accuracy : 0.6752
##                  95% CI : (0.6235, 0.724)
##     No Information Rate : 0.5869
##     P-Value [Acc > NIR] : 0.0004088
##
##                   Kappa : 0.2896
##
##  Mcnemar's Test P-Value : 2.588e-07
##
##             Sensitivity : 0.4138
##             Specificity : 0.8592
##          Pos Pred Value : 0.6742
##          Neg Pred Value : 0.6756
##              Prevalence : 0.4131
##          Detection Rate : 0.1709
##    Detection Prevalence : 0.2536
##       Balanced Accuracy : 0.6365
##
##        'Positive' Class : 0
##
```

- Modelo Logístico

```
fit.modell <- glm(formula = covid_res ~  sex + patient_type + pneumonia + age +
    copd + inmsupr + other_disease + obesity + tobacco + contact_other_covid,
                      family = binomial(link = "logit"), data = train_)

pred<-predict (fit.modell, type='response', newdata=test_)
confusionMatrix(as.factor(as.numeric(pred>0.5)),as.factor(test_$covid_res))
```

```
## Confusion Matrix and Statistics
##
```

```
##           Reference
## Prediction   0   1
##          0  58  25
##          1  87 181
##
##                Accuracy : 0.6809
##                  95% CI : (0.6293, 0.7294)
##     No Information Rate : 0.5869
##     P-Value [Acc > NIR] : 0.0001782
##
##                   Kappa : 0.2975
##
##  Mcnemar's Test P-Value : 8.216e-09
##
##             Sensitivity : 0.4000
##             Specificity : 0.8786
##          Pos Pred Value : 0.6988
##          Neg Pred Value : 0.6754
##              Prevalence : 0.4131
##          Detection Rate : 0.1652
##    Detection Prevalence : 0.2365
##       Balanced Accuracy : 0.6393
##
##        'Positive' Class : 0
##
```

O modelo encontrado possui acurácia (67.5%) semelhante ao modelo logístico proposto no Notebook do Kaggle e também ao modelo logístico feito nesta questão.