Análise de dados (EDA)

Cristian Villegas

2023-05-04

Contents

1 Leitura de dados

```
library(tidyverse)
library(hnp)

dados <- read.csv("Arthritis.txt", sep="",
                  stringsAsFactors=TRUE)

dados$id<- 1:nrow(dados)

dados<- tibble(dados)
dados
```

```
## # A tibble: 51 x 9
##    Sex     Age Group Week0 Week1 Week5 Week9 Week13    id
##    <fct> <int> <fct> <int> <int> <int> <int>  <int> <int>
##  1 M        48 A         1     1     1     1      1     1
##  2 M        29 A         1     1     1     1      1     2
##  3 M        59 P         1     1     1     1      1     3
##  4 F        56 P         1     1     1     1      1     4
```

```
##  5 M        33 P         1     1     1     1      1     5
##  6 M        61 P         1     1     0     1      1     6
##  7 M        63 A         0     0     1    NA     NA     7
##  8 M        57 P         1     0     1     1      1     8
##  9 M        47 P         1     1     1     0      1     9
## 10 F        42 A         0     0     1    NA      0    10
## # ... with 41 more rows
```

2 Alguns resumos dos dados

```
dados %>%
  group_by(Sex) %>%
  summarise( n = n())
```

```
## # A tibble: 2 x 2
##   Sex       n
##   <fct> <int>
## 1 F        13
## 2 M        38
```

```
dados %>%
  group_by(Sex) %>%
  summarise(media_Sex = mean(Age))
```

```
## # A tibble: 2 x 2
##   Sex   media_Sex
##   <fct>     <dbl>
## 1 F          51.8
## 2 M          50.2
```

```
dados %>%
  group_by(Group) %>%
  summarise( n = n())
```

```
## # A tibble: 2 x 2
##   Group     n
##   <fct> <int>
## 1 A        27
## 2 P        24
```

```
dados %>%
  group_by(Group) %>%
  summarise( media_Age = mean(Age))
```

```
## # A tibble: 2 x 2
##   Group media_Age
##   <fct>     <dbl>
## 1 A          51.0
## 2 P          50.2
```

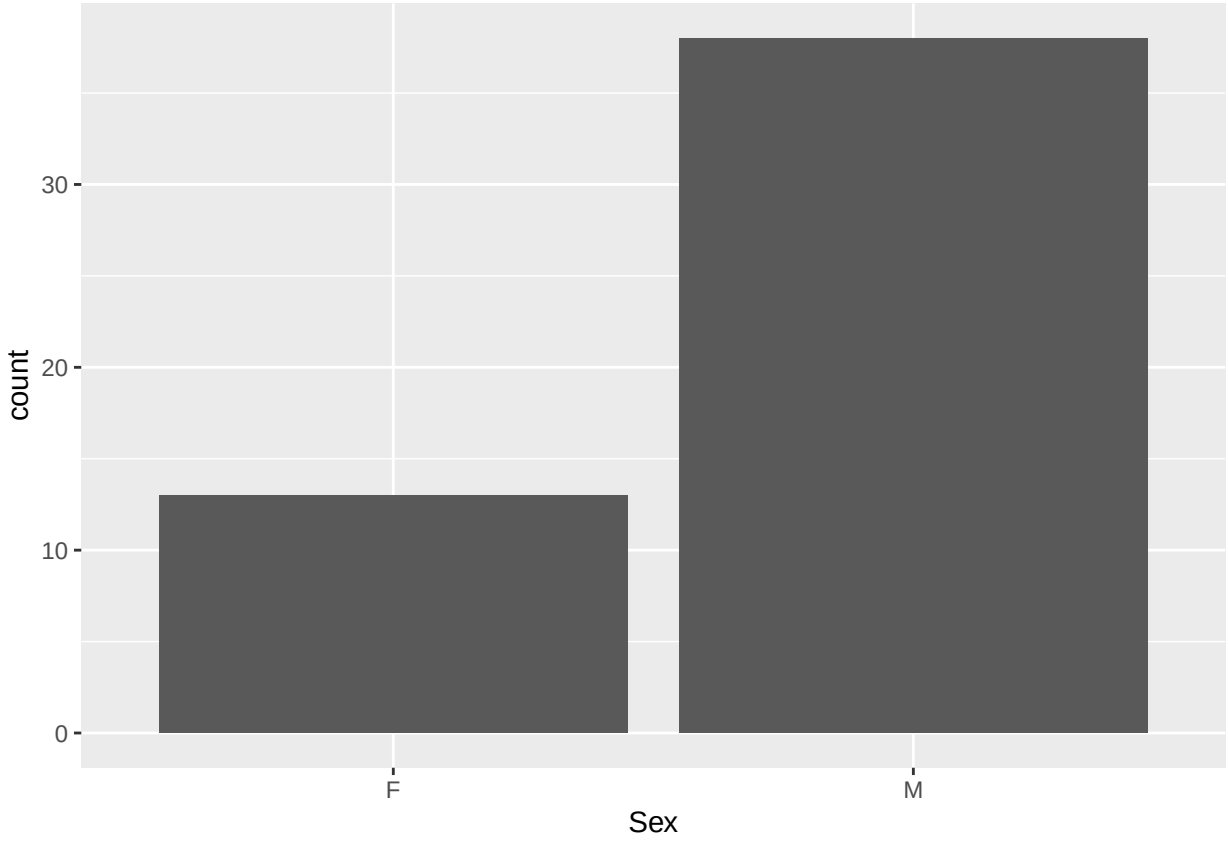
```
dados %>%
  group_by(Sex, Group) %>%
  summarise( n = n())
```

```
## # A tibble: 4 x 3
## # Groups:   Sex [2]
```

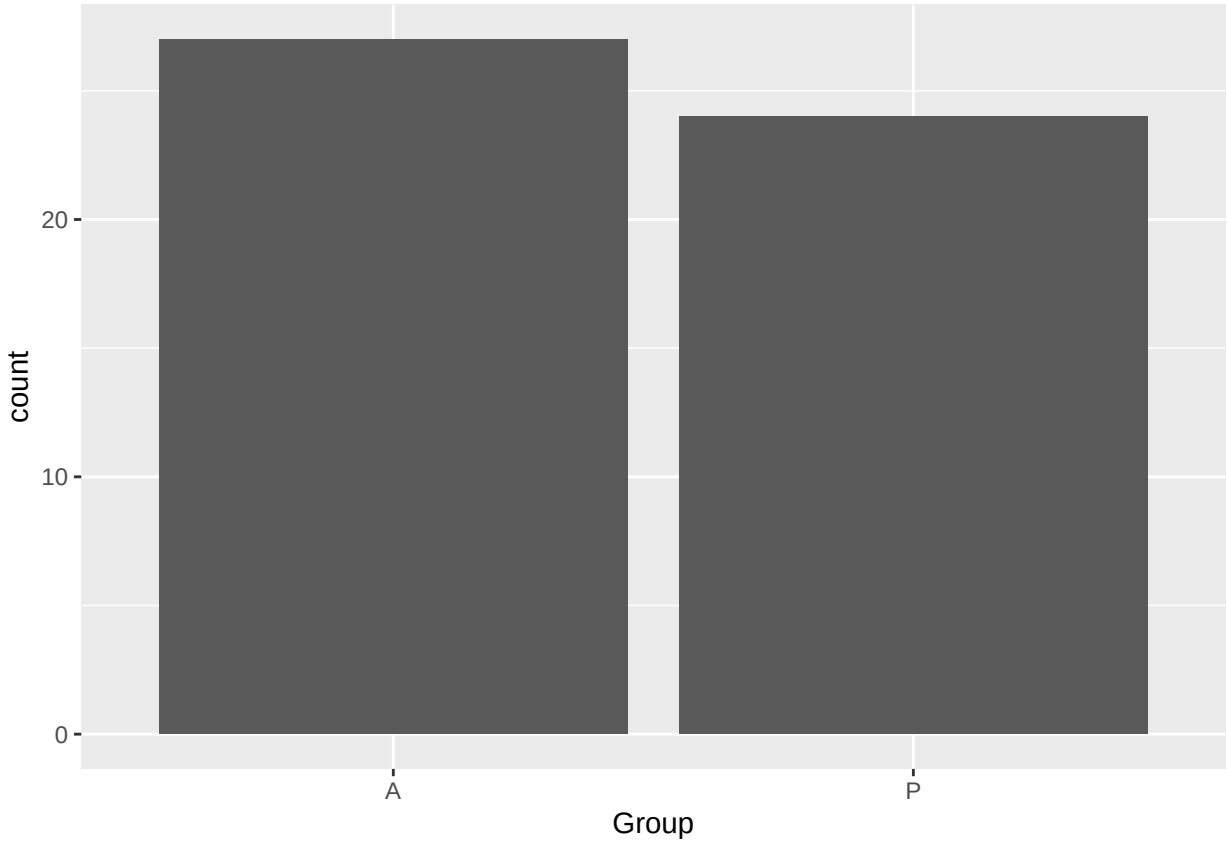
```
##   Sex   Group     n
##   <fct> <fct> <int>
## 1 F     A         7
## 2 F     P         6
## 3 M     A        20
## 4 M     P        18
```

3 Gráficos de interesse

```
ggplot(dados, aes(x = Sex)) +
  geom_bar()
```




```
ggplot(dados, aes(x = Group)) +
  geom_bar()
```

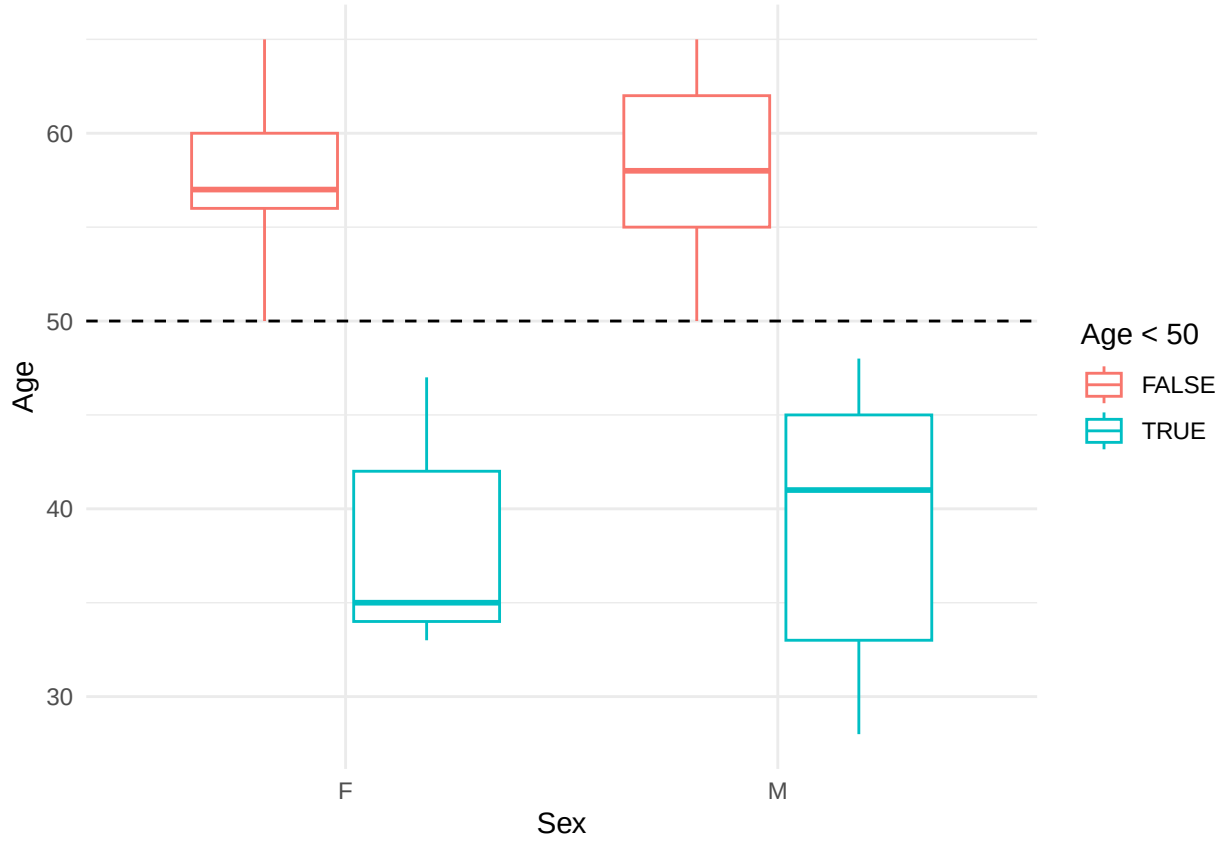


4 Transformando os dados

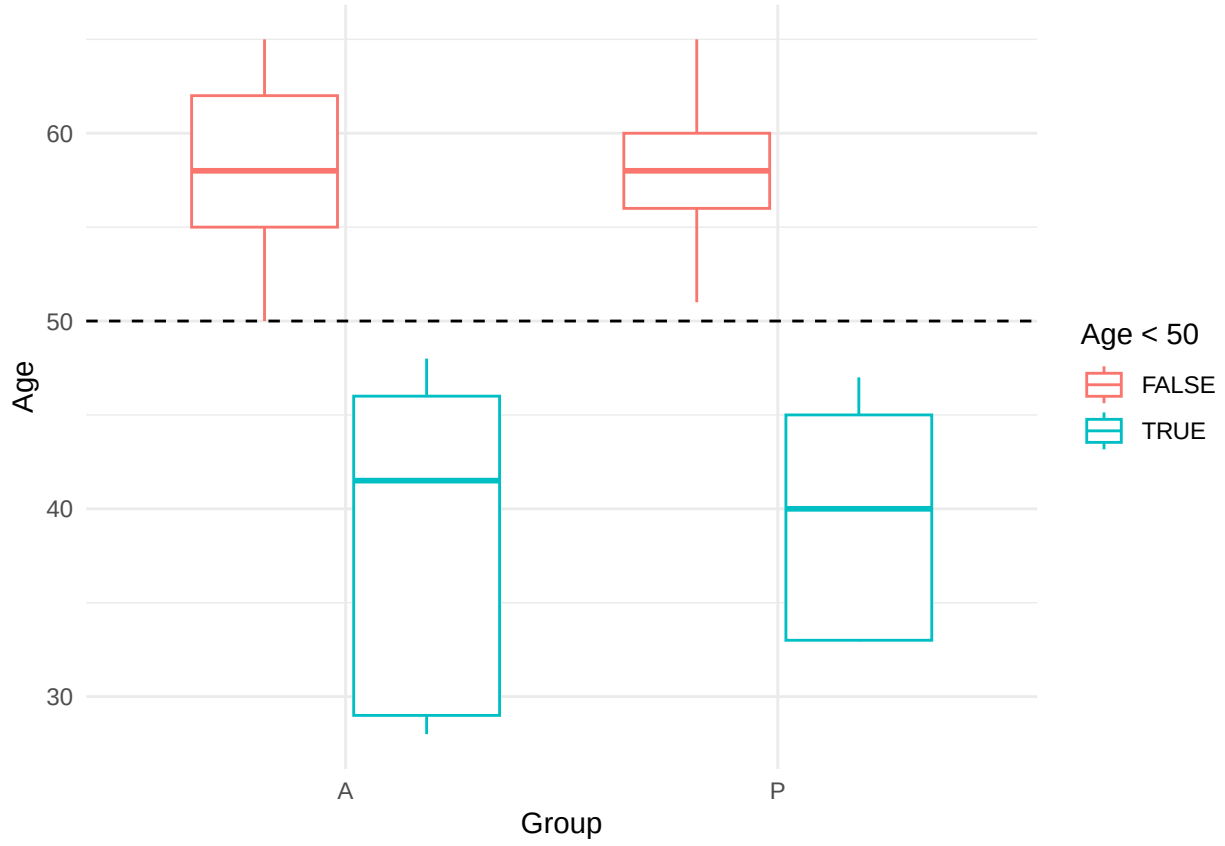
```
dados_longos<- dados %>%
  pivot_longer(
    cols = starts_with("Week"),
    names_to = "week",
    names_prefix = "Week",
    values_to = "Y",
    values_drop_na = TRUE
  )
```

4.1 Gráficos antes de transformar dados

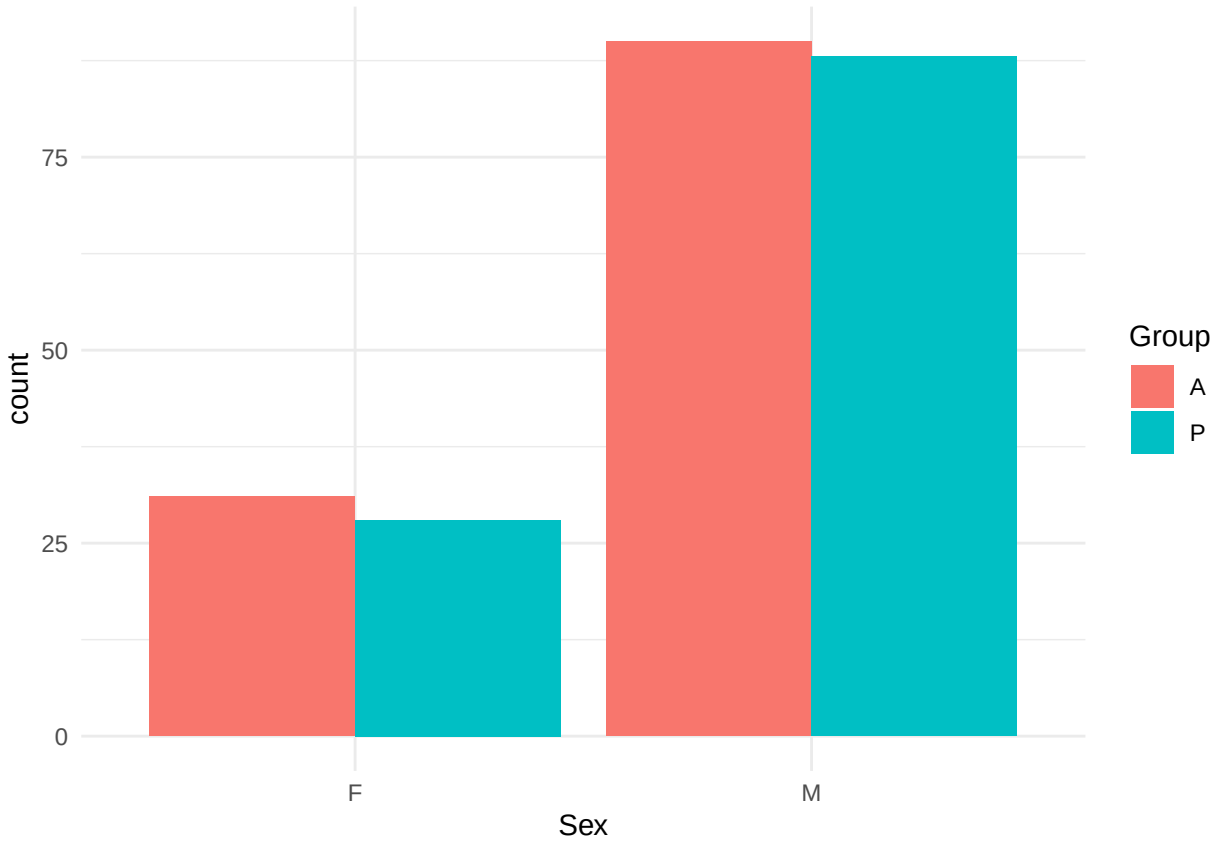
```
ggplot(dados_longos, aes(Sex, Age, col = Age < 50)) +
  geom_boxplot()+
  geom_hline(yintercept = 50, col = "black", linetype = 2)+
  theme_minimal()
```



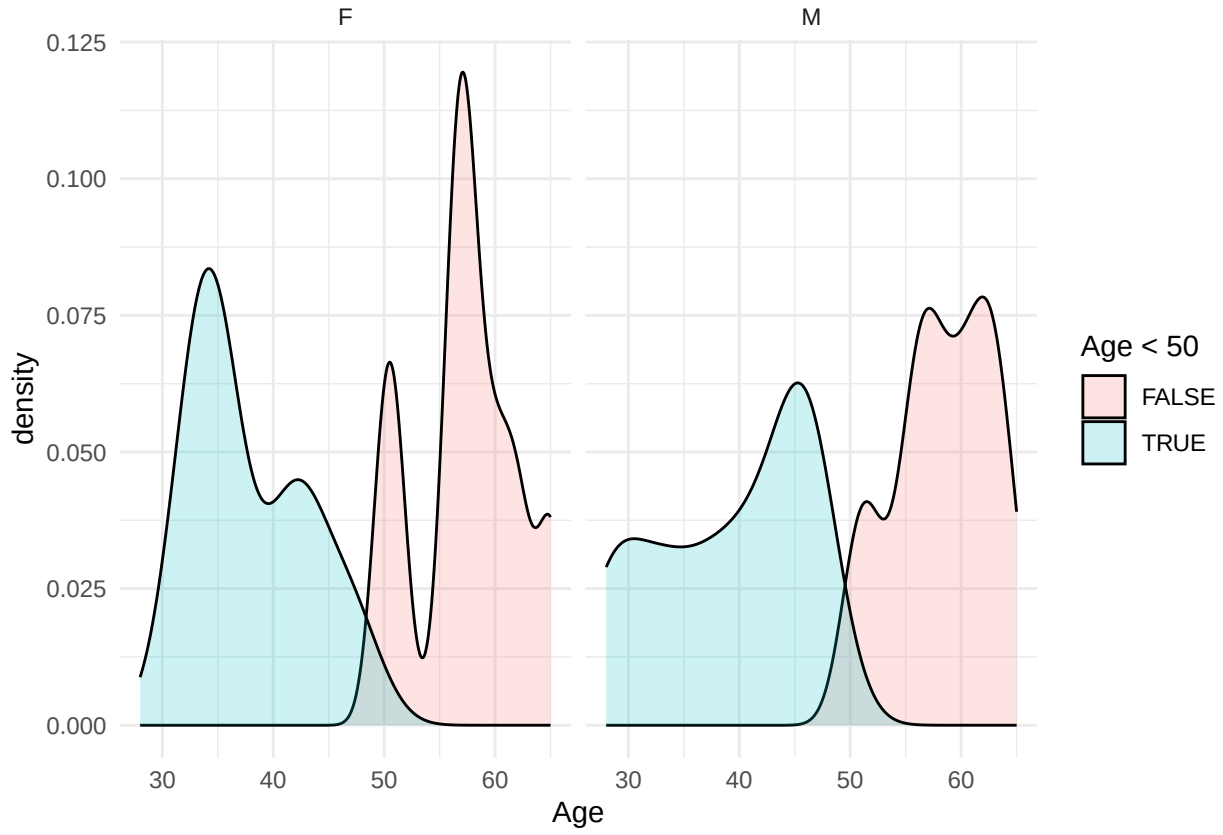

```
ggplot(dados_longos, aes(Group, Age, col = Age < 50)) +
  geom_boxplot()+
  geom_hline(yintercept = 50, col = "black", linetype = 2)+
  theme_minimal()
```




```
ggplot(dados_longos, aes(x = Sex, fill = Group)) +
  geom_bar(position=position_dodge())+
  theme_minimal()
```




```
ggplot(dados_longos, aes(Age, fill = Age < 50)) +
  #geom_histogram(fill = "yellow",
  #               aes(y = after_stat(density)), bins=6)+
  geom_density(alpha=0.2)+
  facet_wrap(~Sex)+
  theme_minimal()
```



4.2 Transformando dados (seguindo o feito pelo Jalmar)

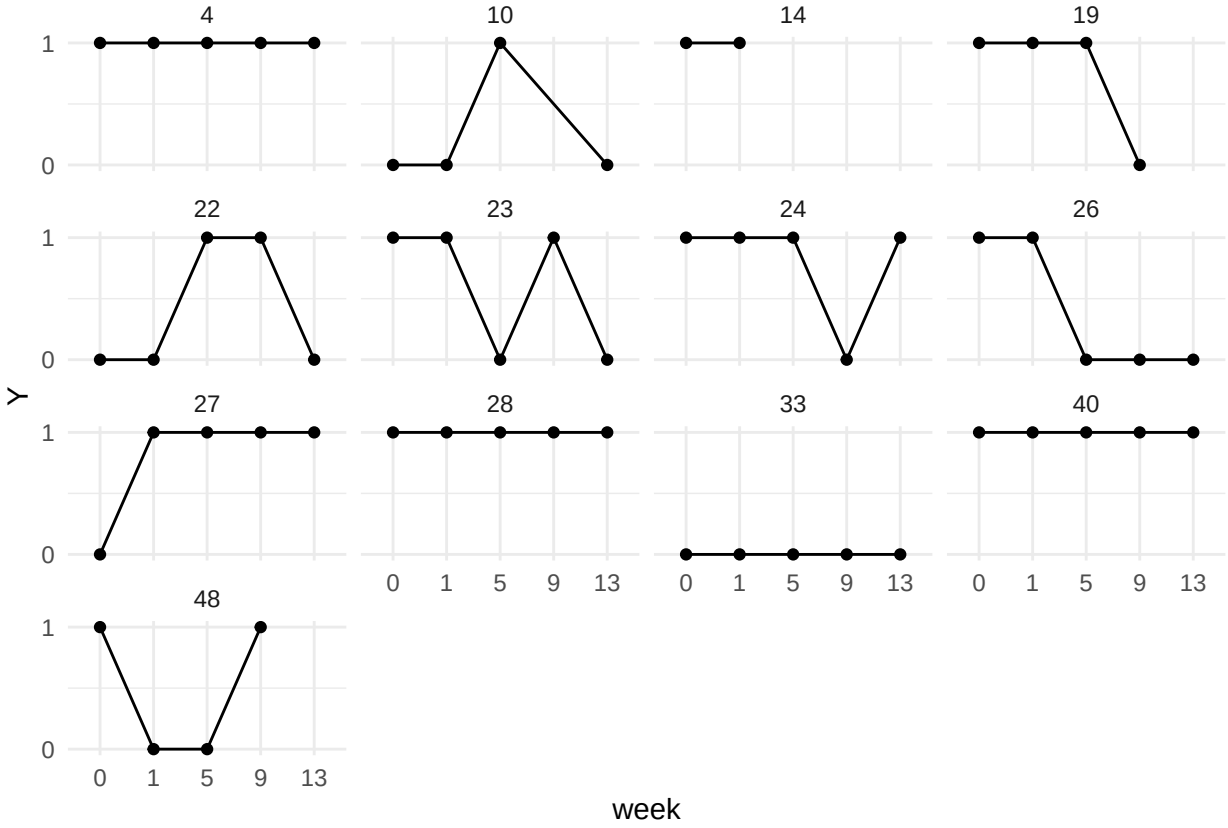
```
dados_longos$Sex<- recode_factor(dados_longos$Sex, `F` = "0", `M` = "1")

dados_longos$Age<- factor(case_when(dados_longos$Age <50  ~ 1,
  dados_longos$Age >=50 ~ 0, .default = dados_longos$Age),
  levels = c(0, 1))

dados_longos$Group<- recode_factor(dados_longos$Group, `P` = "0", `A` = "1")

dados_longos$week<- factor(dados_longos$week,
                           levels = c(0, 1, 5, 9, 13) )

dados_longos %>%
  group_by(week) %>%
  summarise( n = n())
```

```
## # A tibble: 5 x 2
##   week      n
##   <fct> <int>
## 1 0        51
## 2 1        51
## 3 5        48
## 4 9        45
## 5 13       42
```

5 Gráficos de perfis

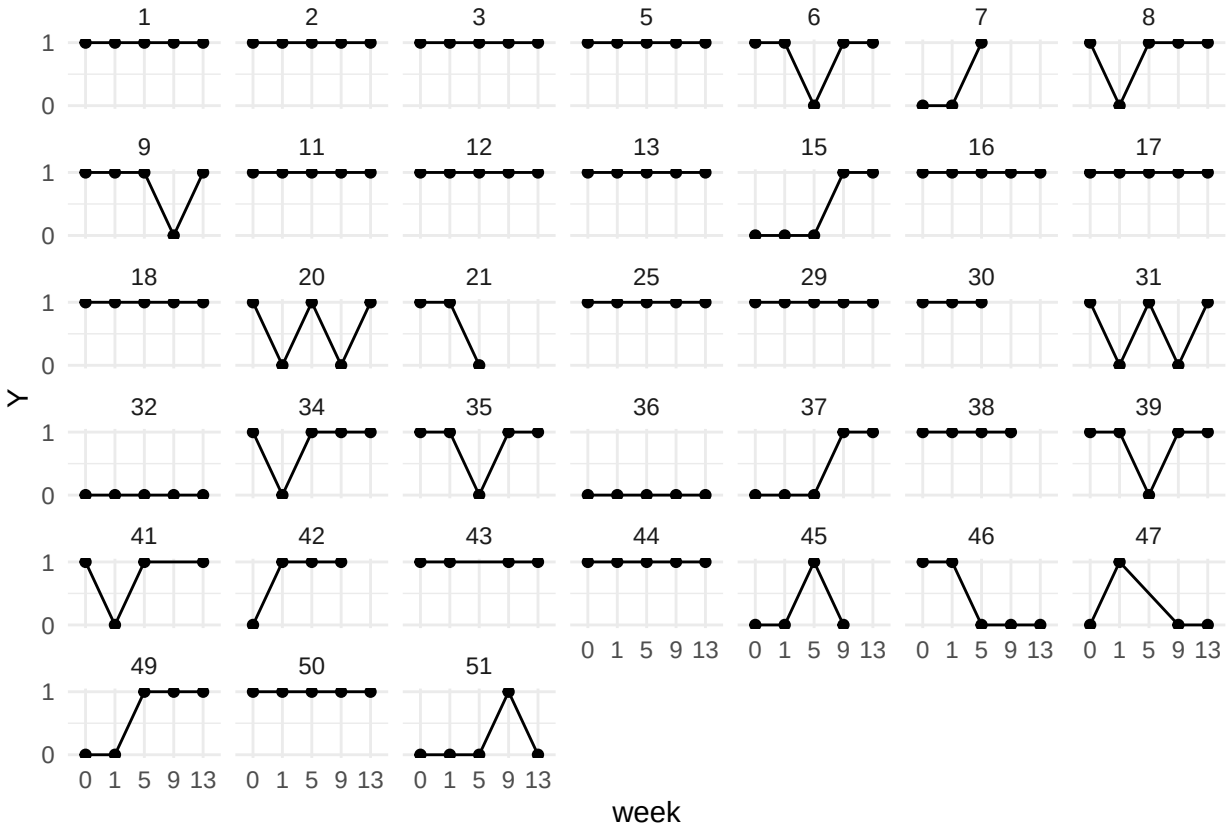
5.1 Sexo Female == 0

```
dados_longos %>% filter(Sex == "0") %>%
  ggplot(aes(week, Y, group = id)) +
  geom_point()+
  geom_line()+
  theme_minimal()+
  scale_y_continuous(breaks = c(0,1))+
  facet_wrap(~id)
```



5.2 Sexo Male == 1

```
dados_longos %>% filter(Sex == "1") %>%
  ggplot(aes(week, Y, group = id)) +
  geom_point()+
  geom_line()+
  theme_minimal()+
  scale_y_continuous(breaks = c(0,1))+
  facet_wrap(~id)
```



6 Ajuste de modelos

6.1 cloglog

```
modelo_cloglog<- glm(Y ~ Sex +
                Age +
                Group +
                as.numeric(week),
              family = binomial(link = "cloglog"),
              data= dados_longos)
modelo_cloglog$family
```

```
##
## Family: binomial
## Link function: cloglog
```

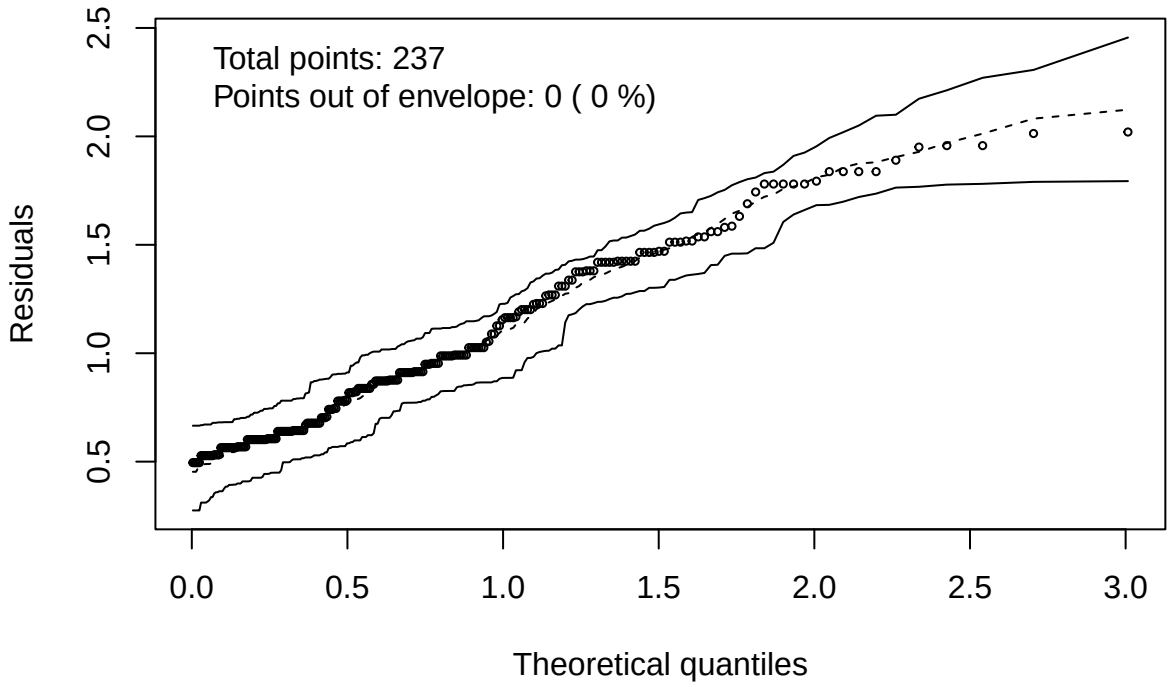
```
summary(modelo_cloglog)
```

```
##
## Call:
## glm(formula = Y ~ Sex + Age + Group + as.numeric(week), family = binomial(link = "cloglog"),
##     data = dados_longos)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0202  -1.2688   0.6394   0.8721   1.2010
```

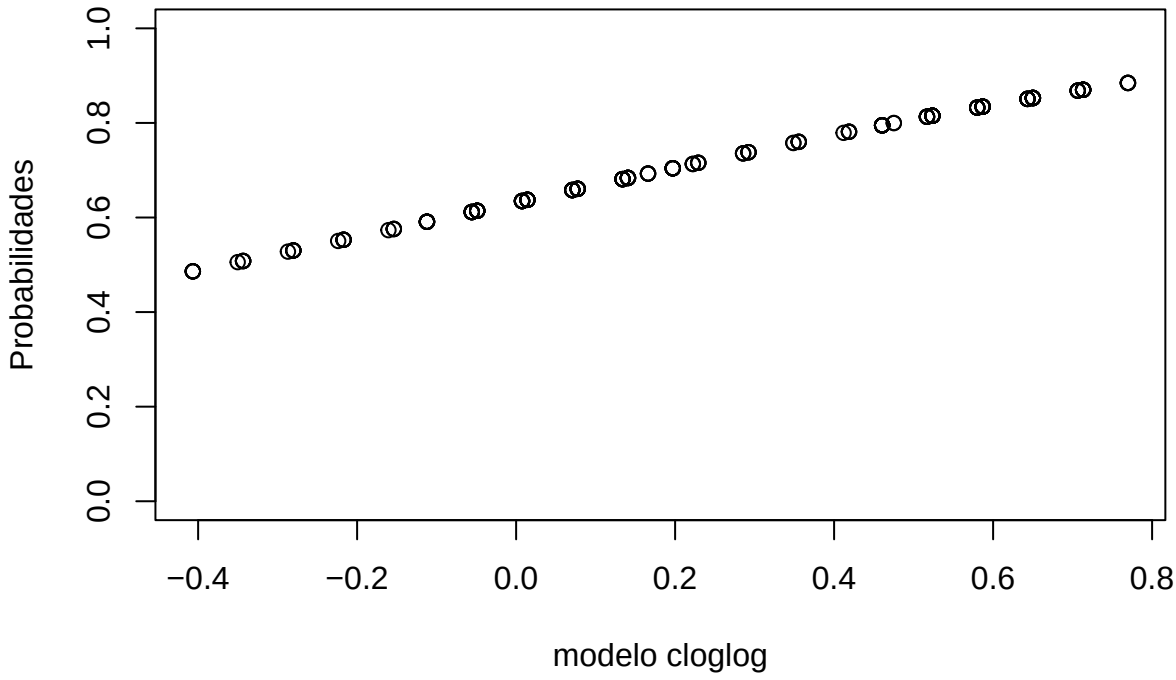
```
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.46985    0.27391  -1.715 0.086281 .
## Sex1              0.29448    0.19859   1.483 0.138115
## Age1              0.05639    0.17402   0.324 0.745898
## Group1            0.57257    0.16874   3.393 0.000691 ***
## as.numeric(week)  0.06321    0.05979   1.057 0.290470
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 282.26  on 236  degrees of freedom
## Residual deviance: 268.18  on 232  degrees of freedom
## AIC: 278.18
##
## Number of Fisher Scoring iterations: 5
```

```
hnp(modelo_cloglog, print.on = TRUE)
```

```
## Binomial model
```




```
plot(predict.glm(modelo_cloglog, type="response")~predict.glm(modelo_cloglog, type="link"),
     ylab = "Probabilidades",
     xlab =  "modelo cloglog",
     ylim=c(0,1))
```



6.2 logit

```
modelo_logit<- glm(Y ~ Sex +
               Age +
               Group +
               as.numeric(week),
                     family = binomial(link = "logit"),
                     data= dados_longos)
modelo_logit$family
```

```
##
## Family: binomial
## Link function: logit
```

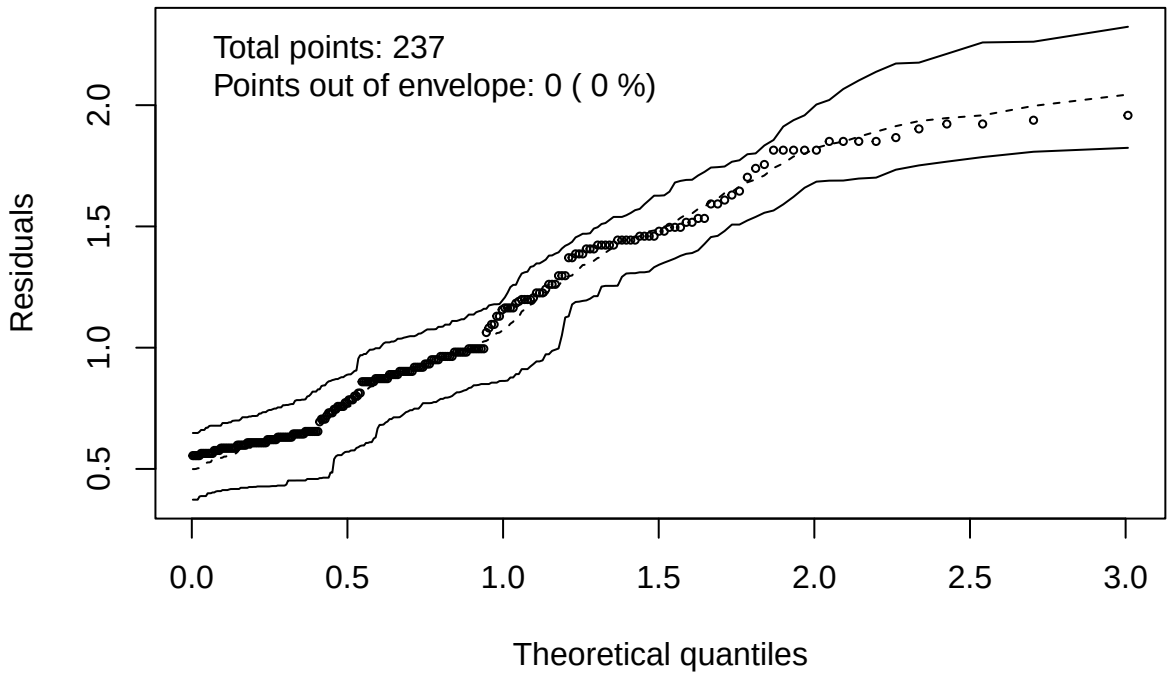
```
summary(modelo_logit)
```

```
##
## Call:
## glm(formula = Y ~ Sex + Age + Group + as.numeric(week), family = binomial(link = "logit"),
##     data = dados_longos)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9579  -1.2612   0.6309   0.8723   1.1986
##
## Coefficients:
```

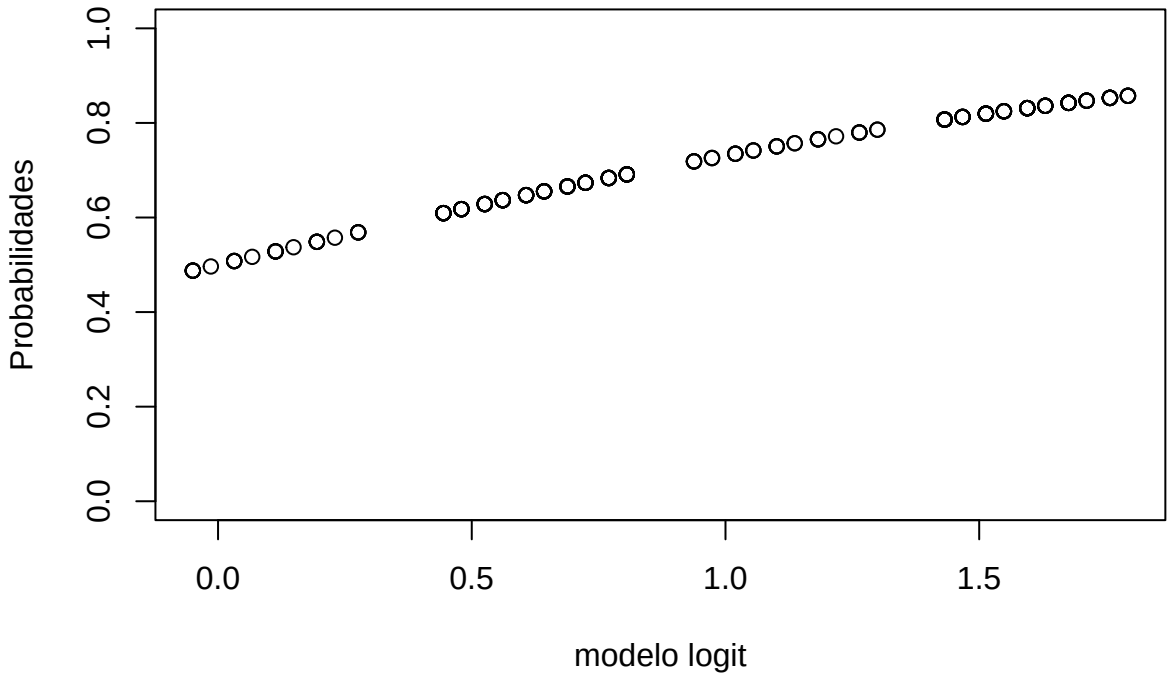
```
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.13114    0.44776  -0.293  0.76962
## Sex1              0.49379    0.33640   1.468  0.14214
## Age1              0.03550    0.31166   0.114  0.90930
## Group1            0.98760    0.30382   3.251  0.00115 **
## as.numeric(week)  0.08148    0.10635   0.766  0.44360
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 282.26  on 236  degrees of freedom
## Residual deviance: 268.79  on 232  degrees of freedom
## AIC: 278.79
##
## Number of Fisher Scoring iterations: 4
```

```
hnp(modelo_logit, print.on = TRUE)
```

```
## Binomial model
```




```
plot(predict.glm(modelo_logit, type="response")~predict.glm(modelo_logit, type="link"),
     ylab = "Probabilidades",
     xlab =  "modelo logit",
     ylim=c(0,1))
```



6.3 probit

```
modelo_probit<- glm(Y ~ Sex +
               Age +
               Group +
               as.numeric(week),
                   family = binomial(link = "probit"),
                   data= dados_longos)
modelo_probit$family
```

```
##
## Family: binomial
## Link function: probit
```

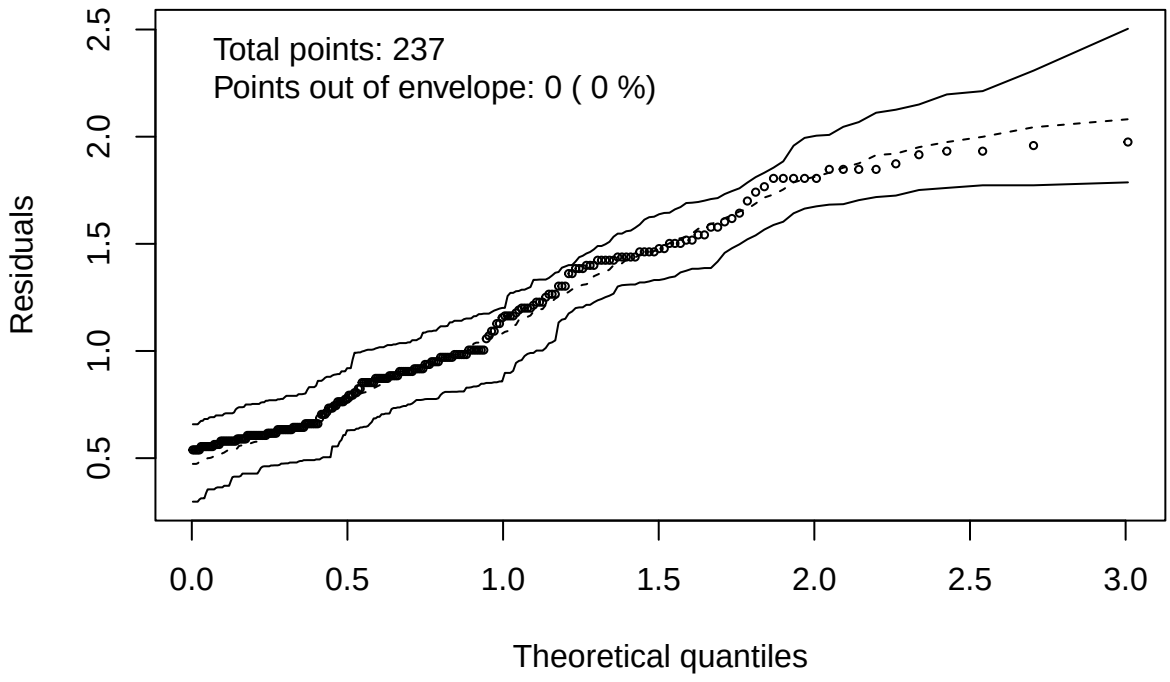
```
summary(modelo_probit)
```

```
##
## Call:
## glm(formula = Y ~ Sex + Age + Group + as.numeric(week), family = binomial(link = "probit"),
##     data = dados_longos)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9750  -1.2650   0.6329   0.8721   1.2001
##
## Coefficients:
```

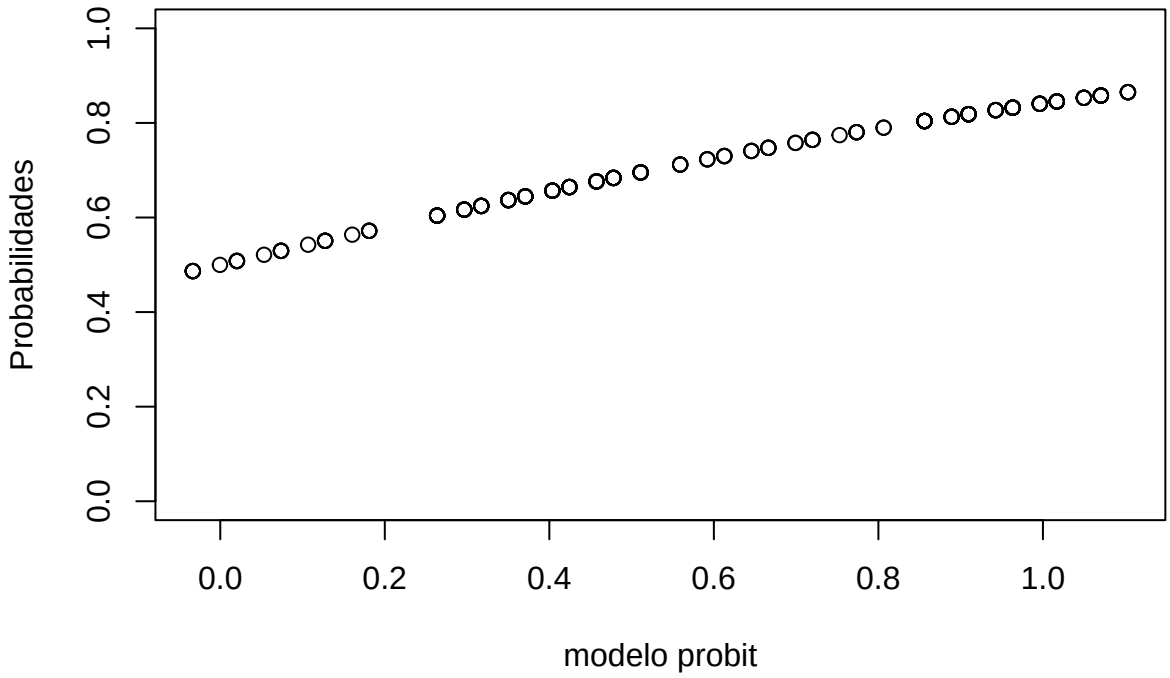
```
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.08693    0.27115  -0.321 0.748510
## Sex1              0.29695    0.20235   1.468 0.142233
## Age1              0.03294    0.18530   0.178 0.858905
## Group1            0.59240    0.17866   3.316 0.000914 ***
## as.numeric(week)  0.05359    0.06329   0.847 0.397121
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 282.26  on 236  degrees of freedom
## Residual deviance: 268.63  on 232  degrees of freedom
## AIC: 278.63
##
## Number of Fisher Scoring iterations: 4
```

```
hnp(modelo_probit, print.on = TRUE)
```

```
## Binomial model
```




```
plot(predict.glm(modelo_probit, type="response")~predict.glm(modelo_probit, type="link"),
     ylab = "Probabilidades",
     xlab = "modelo probit",
     ylim=c(0,1))
```



6.4 cauchit

```
modelo_cauchit<- glm(Y ~ Sex +
               Age +
               Group +
               as.numeric(week),
                    family = binomial(link = "cauchit"),
                    data= dados_longos)
modelo_cauchit$family
```

```
##
## Family: binomial
## Link function: cauchit
```

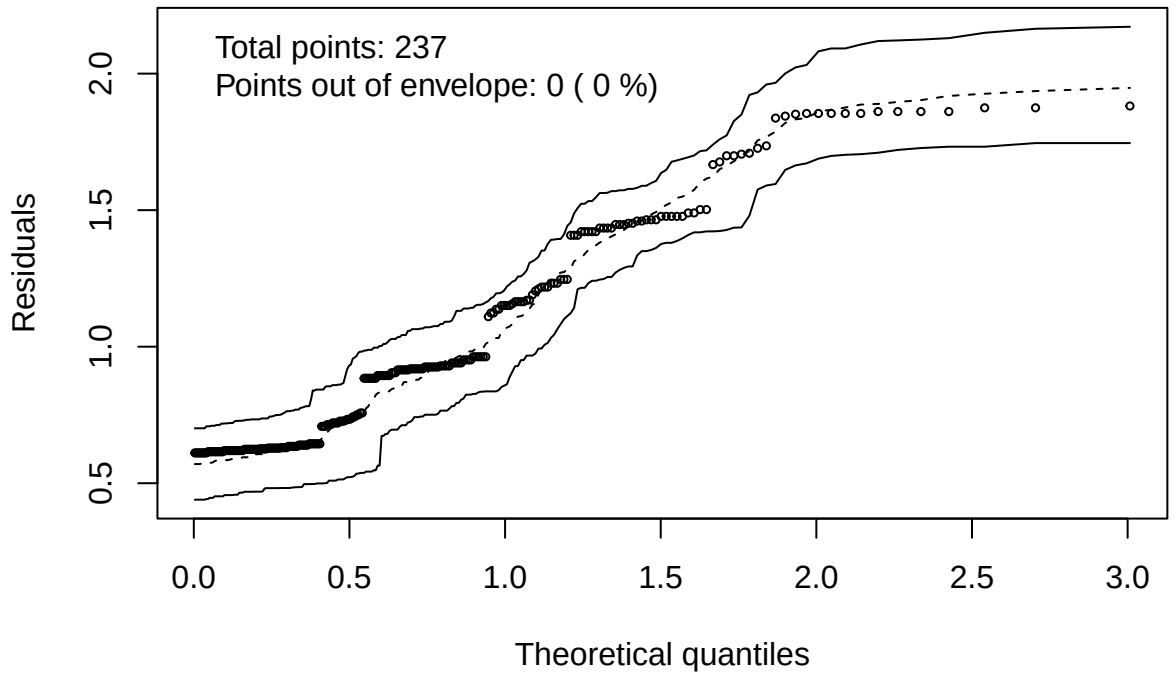
```
summary(modelo_cauchit)
```

```
##
## Call:
## glm(formula = Y ~ Sex + Age + Group + as.numeric(week), family = binomial(link = "cauchit"),
##     data = dados_longos)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8815  -1.2323   0.6286   0.8943   1.2118
##
## Coefficients:
```

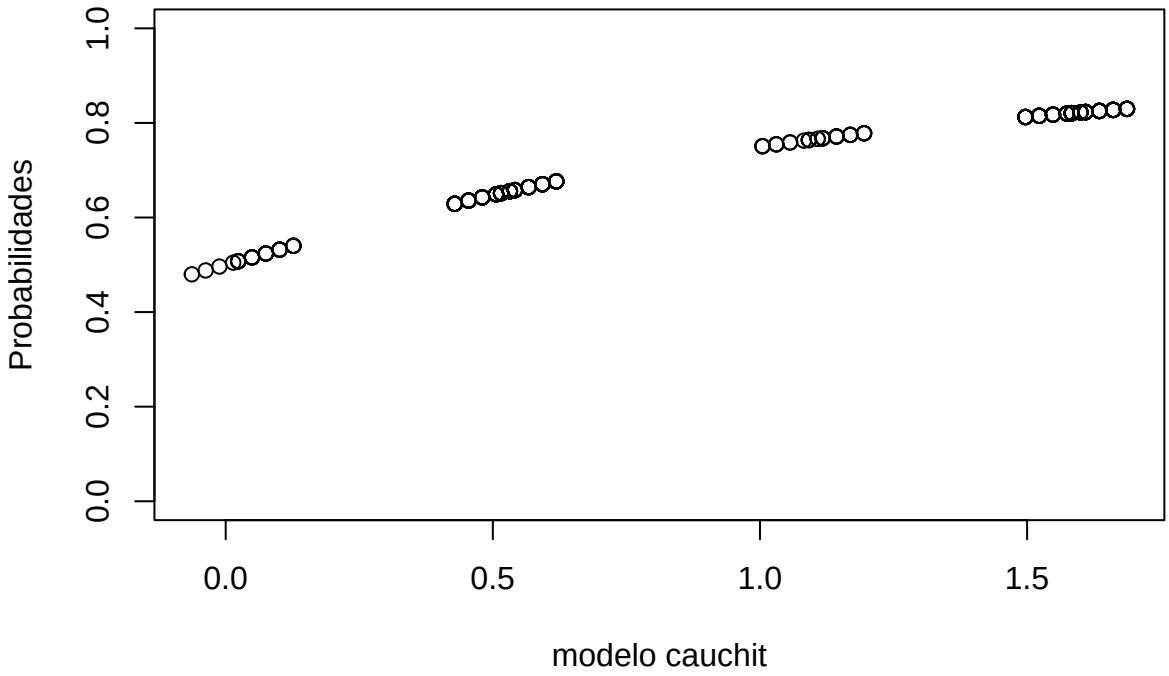
```
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.002309   0.420176  -0.005   0.9956
## Sex1              0.492029   0.334590   1.471   0.1414
## Age1             -0.086834   0.321406  -0.270   0.7870
## Group1            1.068118   0.383931   2.782   0.0054 **
## as.numeric(week)  0.025811   0.107747   0.240   0.8107
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 282.26  on 236  degrees of freedom
## Residual deviance: 269.49  on 232  degrees of freedom
## AIC: 279.49
##
## Number of Fisher Scoring iterations: 6
```

```
hnp(modelo_cauchit, print.on = TRUE)
```

```
## Binomial model
```




```
plot(predict.glm(modelo_cauchit, type="response")~predict.glm(modelo_cauchit, type="link"),
     ylab = "Probabilidades",
     xlab =  "modelo cauchit",
     ylim=c(0,1))
```


Probabilidades
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
modelo cauchit