Introduction aux statistiques

Fabien Haury

2022-07-05

1 Librairies

```
library(plyr)
library(tidyverse)
library(lubridate)
library(naniar)
library(scales)
library(ggthemes)
library(vcd)
library(flextable)
library(crosstable)
library(finalfit)
library(summarytools)
library(rstatix)
```

2 Import et préparation des données

2.1 Import des données

```r
countries.HDI <- read_csv("jeu_de_donnees/countries.HDI.CSV",
                    locale = locale(encoding = "ISO-8859-1"),
                    na = "NA")

effec1_quest_compil <- read_csv("jeu_de_donnees/effec1.quest.compil.csv",
                        locale = locale(encoding = "ISO-8859-1"),
                        na = "NA")

effec2_quest_compil <- read_csv("jeu_de_donnees/effec2.quest.compil.csv",
                        locale = locale(encoding = "ISO-8859-1"),
                        na = "NA")

effec3_quest_compil <- read_csv("jeu_de_donnees/effec3.quest.compil.csv",
                        locale = locale(encoding = "ISO-8859-1"),
                        na = "NA")

usages_effec1 <- read_csv("jeu_de_donnees/usages.effec1.csv",
                    na = "NA")

usages_effec2 <- read_csv("jeu_de_donnees/usages.effec2.csv",
                    na = "NA")

usages_effec3 <- read_csv("jeu_de_donnees/usages.effec3.csv",
                    na = "NA")
```

2.2 Création des variables intermédiaires

```r
quizz <- c("Quizz.1.bin", "Quizz.2.bin", "Quizz.3.bin", "Quizz.4.bin",
         "Quizz.5.bin")

video <- c("S1.L1", "S1.L2", "S1.L3", "S1.L4", "S1.L5", "S1.L6",
         "S2.L1", "S2.L2", "S2.L3", "S2.L4", "S2.L5", "S2.L6",
         "S3.L1.1","S3.L1.2", "S3.L2", "S3.L3", "S3.L4", "S3.L5",
         "S4.L1.1","S4.L1.2", "S4.L2", "S4.L3", "S4.L4", "S4.L5",
         "S5.L1.1","S4.L1.2", "S5.L2", "S5.L3", "S5.L4", "S5.L5")

colonne_cible <- c("Student_ID", "Gender", "HDI", "Total_video", "Total_quizz", "Statut",
               "Iteration", "Exam.bin", "HDI_count")
```

2.3 Préparation des données

```r
theme_set(theme_stata(base_family = "serif"))
theme_update(panel.grid.major = element_blank(),
          panel.background = element_rect(fill = "#CFCFCF"),
          axis.title = element_text(size = 15))

### Catégories apprenants
usages_effec1 <- usages_effec1 %>%
  mutate(Statut = case_when(Exam.bin == 1 ~ "Completer",
                        rowSums(usages_effec1[, quizz]) >= 1 & Assignment.bin == 1 &
                          Exam.bin == 0  ~ "Disengaging learners",
                        rowSums(usages_effec1[, quizz]) == 0 & Assignment.bin == 0 &
                          rowSums(usages_effec1[, video]) >= 6  ~ "Auditing learner",
                        rowSums(usages_effec1[, quizz]) == 0 & Assignment.bin == 0 &
                          rowSums(usages_effec1[, video]) < 6  ~ "Bystanders"))

usages_effec2 <- usages_effec2 %>%
  mutate(Statut = case_when(Exam.bin == 1 ~ "Completer",
                        rowSums(usages_effec2[, quizz]) >= 1 & Assignment.bin == 1 &
                          Exam.bin == 0  ~ "Disengaging learners",
                        rowSums(usages_effec2[, quizz]) == 0 & Assignment.bin == 0 &
                          rowSums(usages_effec2[, video]) >= 6  ~ "Auditing learner",
                        rowSums(usages_effec2[, quizz]) == 0 & Assignment.bin == 0 &
                          rowSums(usages_effec2[, video]) < 6  ~ "Bystanders"))
```

```r
usages_effec3 <- usages_effec3 %>%
  mutate(Statut = case_when(Exam.bin == 1 ~ "Completer",
                            rowSums(usages_effec3[, quizz]) >= 1 & Assignment.bin == 1 &
                              Exam.bin == 0  ~ "Disengaging learners",
                            rowSums(usages_effec3[, quizz]) == 0 & Assignment.bin == 0 &
                              rowSums(usages_effec3[, video]) >= 6  ~ "Auditing learner",
                            rowSums(usages_effec3[, quizz]) == 0 & Assignment.bin == 0 &
                              rowSums(usages_effec3[, video]) < 6  ~ "Bystanders"))


### Merge table
merge_1 <- merge(effec1_quest_compil, usages_effec1, by = "Student_ID", all = TRUE)
merge_2 <- merge(effec2_quest_compil, usages_effec2, by = "Student_ID", all = TRUE)
merge_3 <- merge(effec3_quest_compil, usages_effec3, by = "Student_ID", all = TRUE)

merge_1 <- merge_1 %>% mutate(Iteration = 1)
merge_2 <- merge_2 %>% mutate(Iteration = 2)
merge_3 <- merge_3 %>% mutate(Iteration = 3)

mooc_full_join <- rbind.fill(merge_1, merge_2, merge_3)

mooc_filtered <- mooc_full_join %>%
  group_by(Country_HDI) %>%
  mutate(HDI_count = n(),
         HDI = case_when(Country_HDI == "M" | Country_HDI == "H" ~ "I",
                         Country_HDI == "B" ~ "B",
                         Country_HDI == "TH" ~ "TH")) %>%
  ungroup()%>%
  mutate(Total_video = rowSums(mooc_full_join[, video]),
         Total_quizz = rowSums(mooc_full_join[, quizz]),
         Gender = case_when(Gender == "un homme" ~ "homme",
                            Gender == "une femme" ~ "femme")) %>%
  subset(select = colonne_cible)


### Suppression des dataframes/variables inutiles
rm(effec1_quest_compil)
rm(effec2_quest_compil)
rm(effec3_quest_compil)
rm(usages_effec1)
rm(usages_effec2)
rm(usages_effec3)
rm(merge_1)
rm(merge_2)
rm(merge_3)
rm(colonne_cible)
rm(video)
rm(quizz)
```

3 Description du jeu de données

```r
mooc_cross_table <- crosstable(mooc_filtered, c(Statut, Iteration),
                               by = Iteration, total = "both",
                               percent_pattern = "{n} ({p_col})",
                               showNA = "ifany")%>%
                  as_flextable()

mooc_cross_table
```

label	variable	Iteration			Total
		1	2	3	

label	variable	Iteration			Total
		1	2	3	
Statut	Auditing learner	150 (2.60%)	107 (3.79%)	106 (3.52%)	363 (3.13%)
	Bystanders	3141 (54.49%)	1719 (60.87%)	1981 (65.73%)	6841 (58.96%)
	Completer	20 (0.35%)	878 (31.09%)	843 (27.97%)	1741 (15.01%)
	Disengaging learners	2453 (42.56%)	120 (4.25%)	84 (2.79%)	2657 (22.90%)
	NA	3222	1350	1587	6159
	Total	8986 (49.68%)	4174 (24.34%)	4601 (25.98%)	17761 (100.00%)

4 Chi2 et mosaic plot

```
mooc_hdi_gender <- mooc_filtered %>% select("Gender", "HDI")
```

4.1 Chi2 test, V de Cramer

```
chisq <- chisq.test(table(mooc_hdi_gender$HDI, mooc_hdi_gender$Gender))

chisq
```

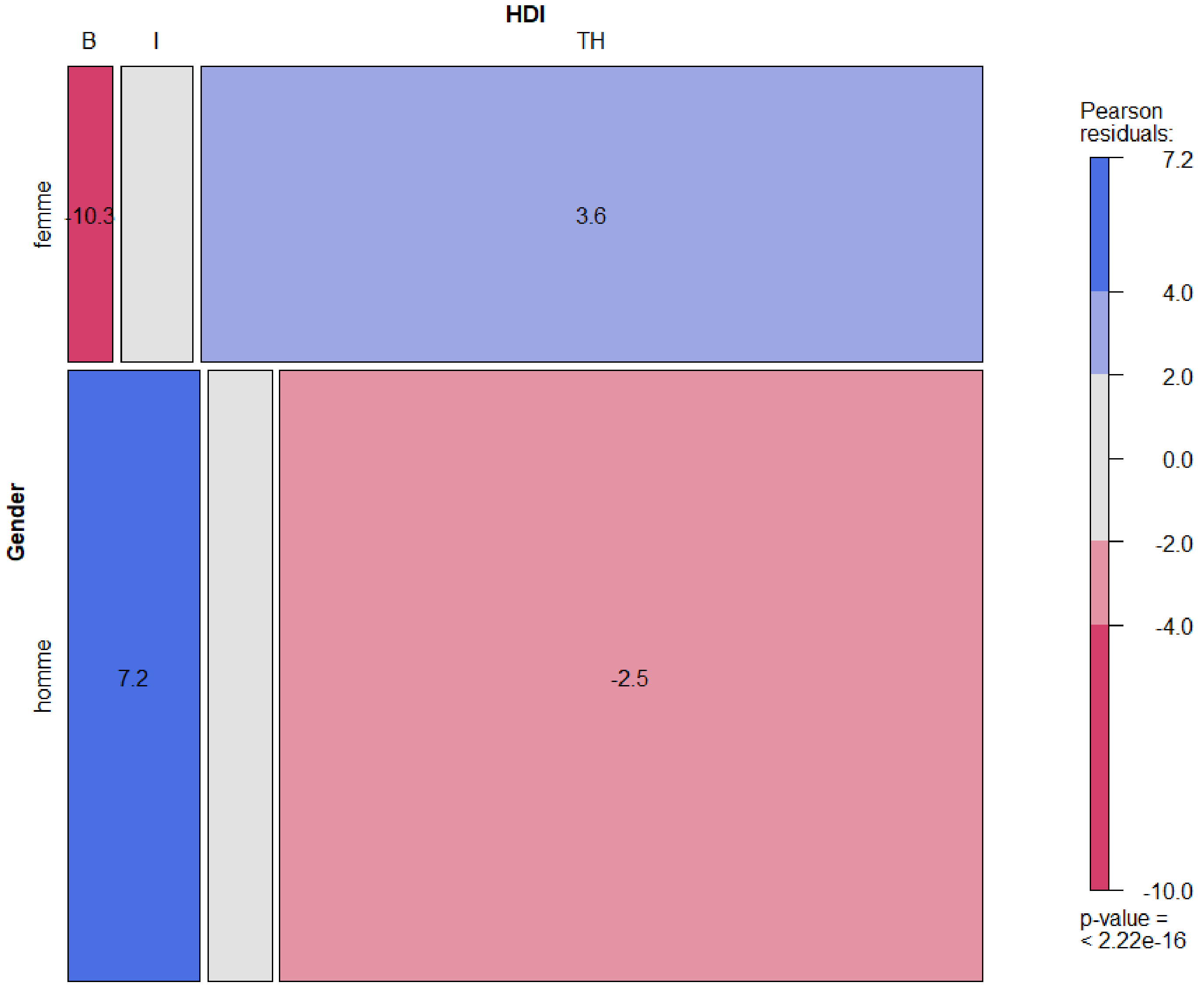
```
##
##  Pearson's Chi-squared test
##
## data:  table(mooc_hdi_gender$HDI, mooc_hdi_gender$Gender)
## X-squared = 179.05, df = 2, p-value < 2.2e-16
```

```
questionr::cramer.v(table(mooc_hdi_gender$Gender, mooc_hdi_gender$HDI))
```

```
## [1] 0.1413849
```

4.2 Mosaic plot

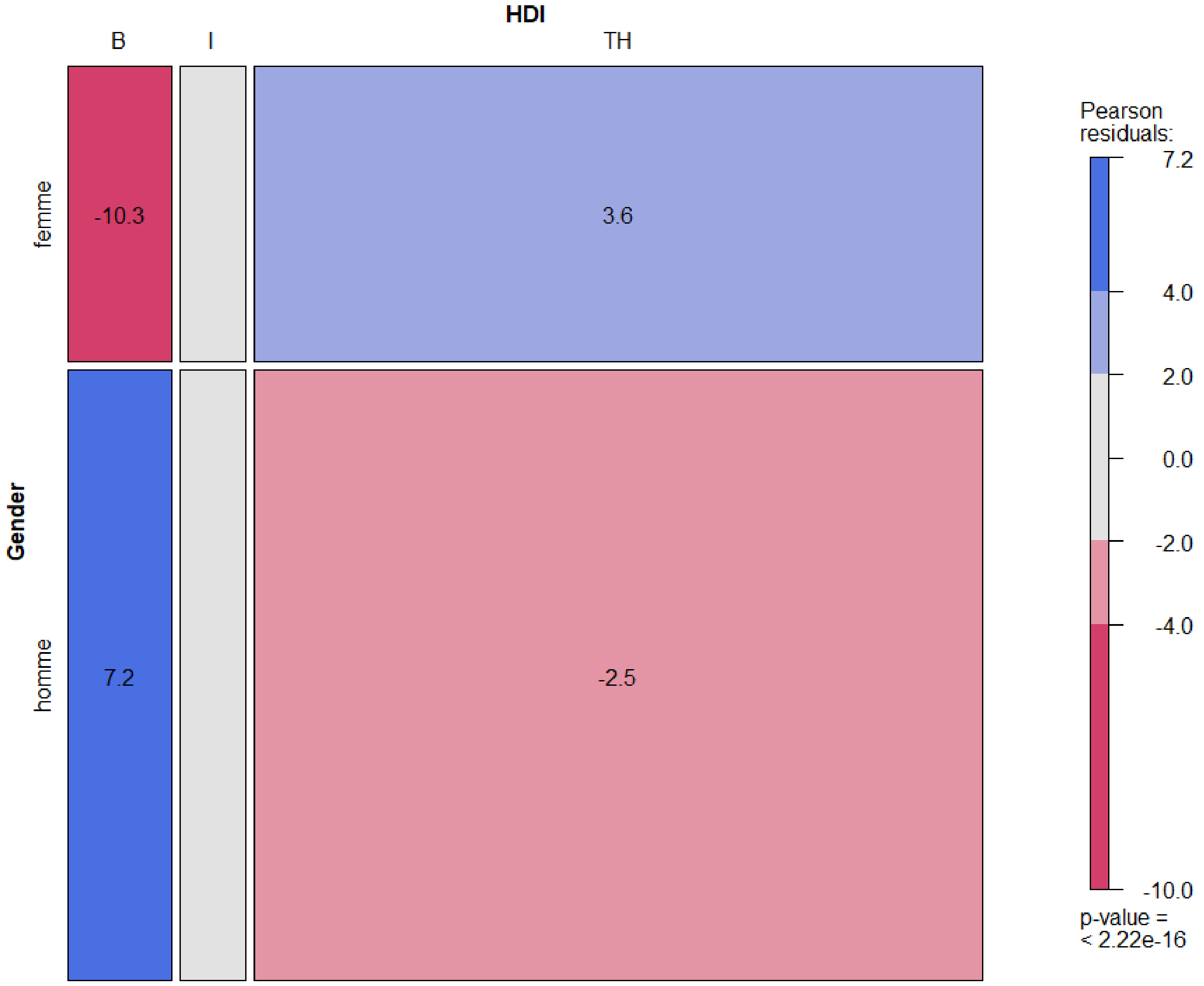
```
mosaic_observed <- mosaic(~ Gender + HDI,
    data = mooc_hdi_gender,
    shade = TRUE, legend = TRUE,
    labeling=labeling_residuals)
```




```
mosaic_observed
```

```
##        HDI    B    I   TH
## Gender
## femme       147  233 2546
## homme       883  432 4716
```

```
mosaic_expected <- mosaic(~ Gender + HDI,
      data = mooc_hdi_gender,
      shade = TRUE, legend = TRUE, type = "expected",
      labeling=labeling_residuals)
```




```
mosaic_expected
```

```
##        HDI         B         I        TH
## Gender
## femme       336.4720  217.2368 2372.2912
## homme       693.5280  447.7632 4889.7088
```

5 Modèle linéaire, tests non-paramétriques

```
mooc_selection <- mooc_filtered %>% select(Total_quizz, Total_video, HDI, Gender)
```

5.1 Test de Student

```
t.test(Total_video ~ Gender, mooc_selection)
```

```
##
##  Welch Two Sample t-test
##
## data:  Total_video by Gender
## t = 3.7589, df = 5872.6, p-value = 0.0001723
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.4752355 1.5112247
## sample estimates:
## mean in group femme mean in group homme
##            14.54214            13.54891
```

5.2 Test non-paramétrique

```
wilcox.test(Total_video ~ Gender, mooc_selection, correct = FALSE)
```

```
##
##  Wilcoxon rank sum test
##
## data:  Total_video by Gender
## W = 9532794, p-value = 0.000481
## alternative hypothesis: true location shift is not equal to 0
```

```
mooc_selection %>% wilcox_effsize(Total_video ~ Gender)
```

```
## # A tibble: 1 x 7
##   .y.         group1 group2 effsize    n1    n2 magnitude
## * <chr>       <chr>  <chr>    <dbl> <int> <int> <ord>
## 1 Total_video femme  homme   0.0366  2991  6108 small
```

5.3 Test de corrélation de Pearson

```
cor.test( ~ Total_quizz + Total_video, mooc_selection, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  Total_quizz and Total_video
## t = 170.96, df = 15644, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8015376 0.8124652
## sample estimates:
##       cor
## 0.8070705
```

5.4 Test de corrélation de Spearman

```
cor.test( ~ Total_quizz + Total_video, mooc_selection, method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  Total_quizz and Total_video
## S = 1.28e+11, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##     rho
## 0.79948
```

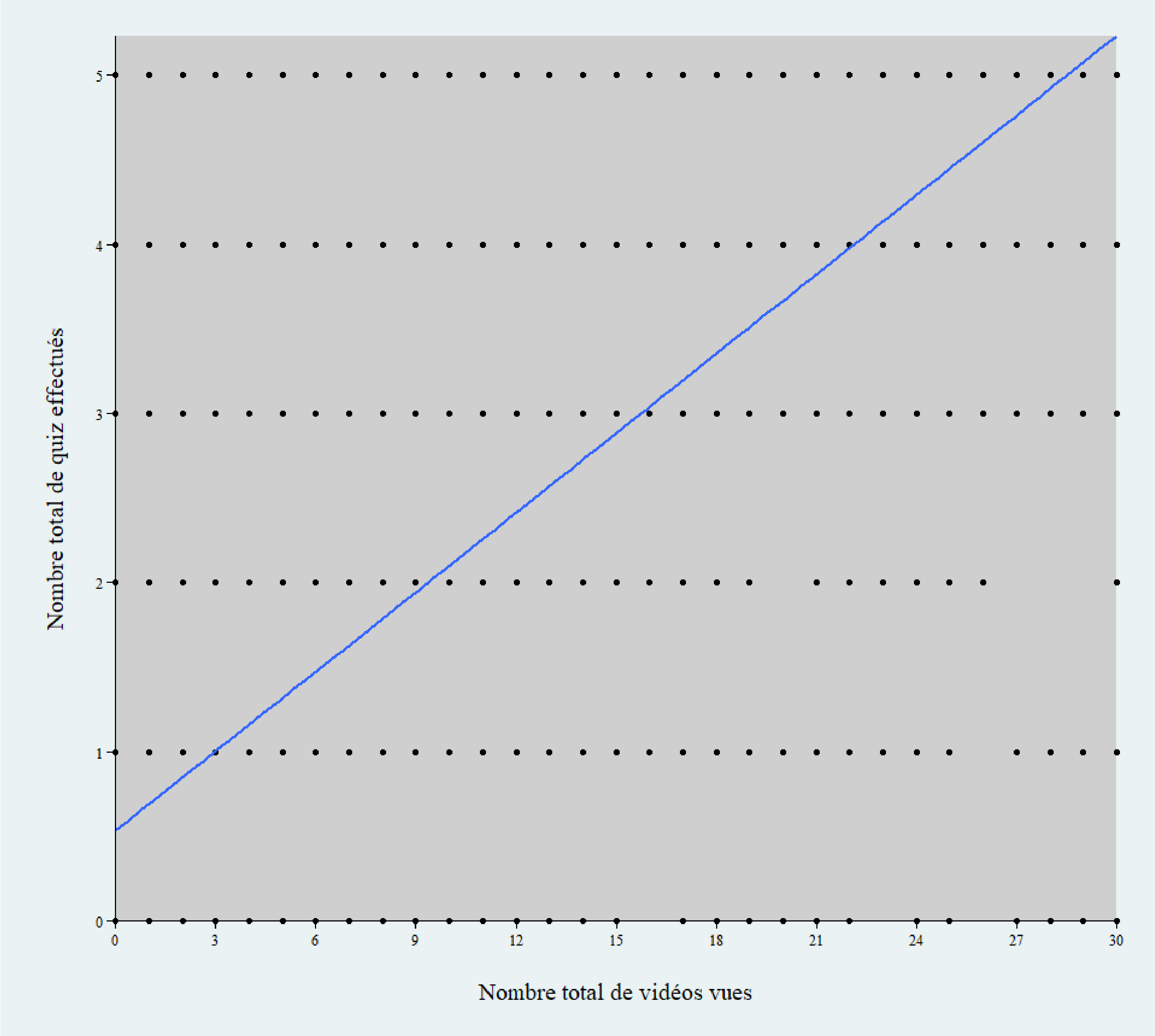
5.5 Modèle linéaire

```
model <- lm(Total_quizz ~ Total_video, mooc_selection)
summary(model)
```

```
##
## Call:
## lm(formula = Total_quizz ~ Total_video, data = mooc_selection)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2326 -0.5344 -0.5344  0.0806  4.4656
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.534369   0.013166   40.59   <2e-16 ***
## Total_video 0.156609   0.000916  170.96   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.29 on 15644 degrees of freedom
##   (2115 observations deleted due to missingness)
## Multiple R-squared:  0.6514, Adjusted R-squared:  0.6513
## F-statistic: 2.923e+04 on 1 and 15644 DF,  p-value: < 2.2e-16
```

5.6 Scatter plot

```
ggplot(mooc_selection, aes(Total_video, Total_quizz)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  scale_x_continuous(breaks = seq(0, 30, 3)) +
  scale_y_continuous(breaks = seq(0, 5, 1)) +
  labs(x = "\nNombre total de vidéos vues",
     y = "Nombre total de quiz effectués\n") +
  theme(axis.text.y = element_text(angle = 0))+
  expand_limits(x = 0, y = 0) +
  coord_cartesian(expand = FALSE, clip = "off")
```



5.7 ANOVA sans statistiques inférentielles

```
mod_1 <- lm(Total_video ~ Gender + HDI, mooc_selection)
anova(mod_1)
```

```
## # A tibble: 3 x 5
##      Df `Sum Sq` `Mean Sq` `F value`   `Pr(>F)`
##   <int>    <dbl>     <dbl>     <dbl>      <dbl>
## 1     1    1867.     1867.      14.3  1.59e-  4
## 2     2   74434.    37217.     285.   1.44e-120
## 3  8947 1169852.      131.      NA   NA
```

```
summary(mod_1)
```

```
##
## Call:
## lm(formula = Total_video ~ Gender + HDI, data = mooc_selection)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.345 -10.175  -3.175  13.655  23.549
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.6200     0.4204  15.747  < 2e-16 ***
## Genderhomme  -0.1694     0.2603  -0.651    0.515
## HDII          4.2555     0.5714   7.448 1.04e-13 ***
## HDITH         8.7246     0.3846  22.686  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.43 on 8947 degrees of freedom
##   (8810 observations deleted due to missingness)
## Multiple R-squared:  0.06123,    Adjusted R-squared:  0.06091
## F-statistic: 194.5 on 3 and 8947 DF,  p-value: < 2.2e-16
```

5.8 ANOVA avec statistiques inférentielles

```
mod_2 <- lm(Total_video ~ Gender + HDI + Gender * HDI, mooc_selection)
anova(mod_2)
```

```
## # A tibble: 4 x 5
##      Df `Sum Sq` `Mean Sq` `F value`   `Pr(>F)`
##   <int>    <dbl>     <dbl>     <dbl>      <dbl>
## 1     1    1867.     1867.      14.3  1.58e-  4
## 2     2   74434.    37217.     285.   1.40e-120
## 3     2     402.      201.       1.54  2.15e-  1
## 4  8945 1169450.      131.      NA    NA
```

```
summary(mod_2)
```

```
##
## Call:
## lm(formula = Total_video ~ Gender + HDI + Gender * HDI, data = mooc_selection)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.417 -10.136  -3.136  13.583  23.606
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         6.9592     0.9431   7.379 1.73e-13 ***
## Genderhomme        -0.5651     1.0185  -0.555   0.5791
## HDII                2.9121     1.2044   2.418   0.0156 *
## HDITH               8.4577     0.9699   8.720  < 2e-16 ***
## Genderhomme:HDII    1.9415     1.3788   1.408   0.1591
## Genderhomme:HDITH   0.2842     1.0567   0.269   0.7879
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.43 on 8945 degrees of freedom
##   (8810 observations deleted due to missingness)
## Multiple R-squared:  0.06155,    Adjusted R-squared:  0.06103
## F-statistic: 117.3 on 5 and 8945 DF,  p-value: < 2.2e-16
```

6 Régression logistique

6.1 Présenter des odds ratios

```
mooc_glm <- mooc_filtered %>% select(HDI, Gender, Exam.bin)
explanatory <- c("HDI", "Gender")
dependent <- c("Exam.bin")
```

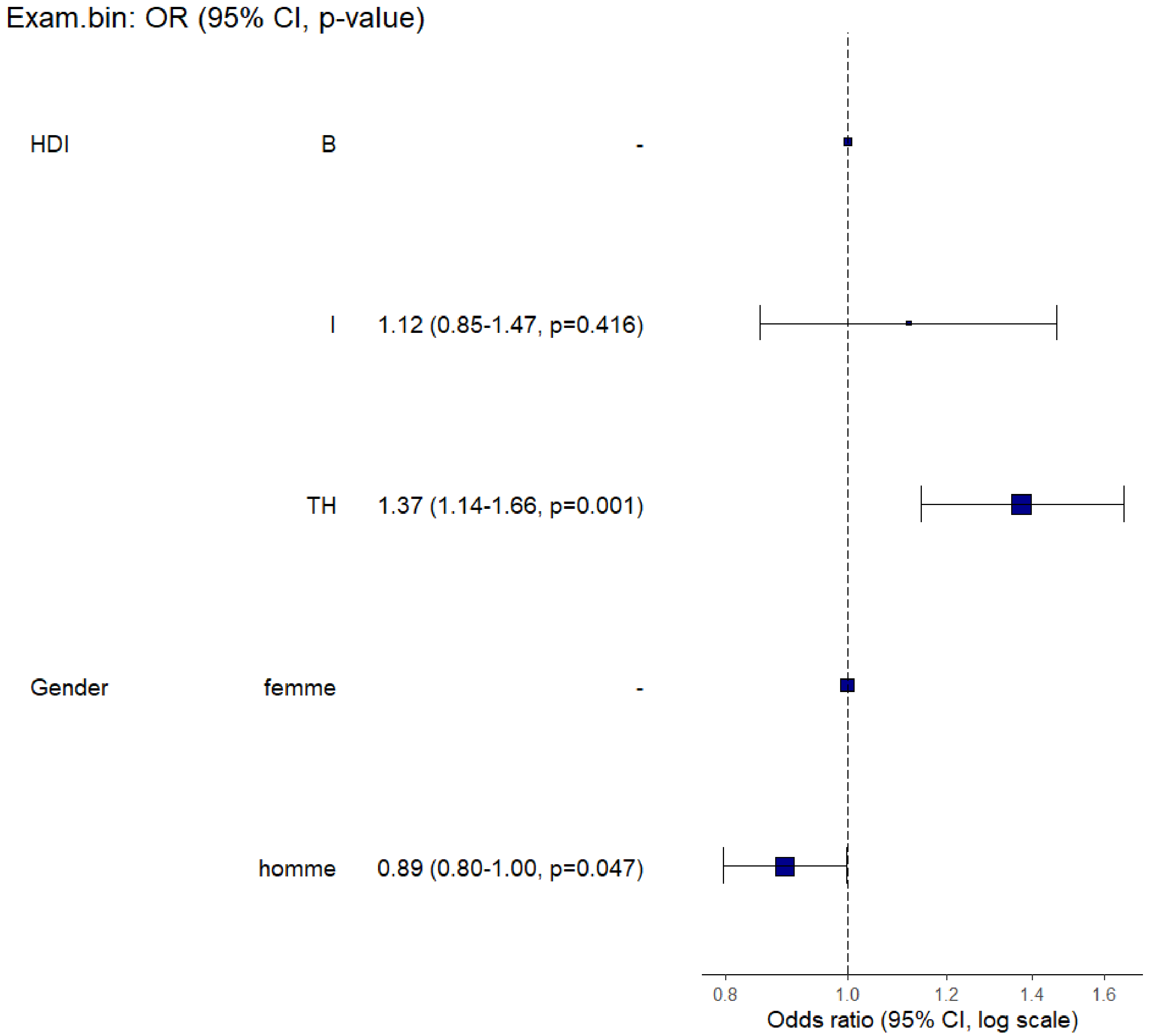
6.2 Calcul des odds ratio

```
glm_test <- glm(Exam.bin ~ HDI + Gender, mooc_glm, family = "binomial")
questionr::odds.ratio(glm_test)
```

```
## # A tibble: 4 x 4
##      OR `2.5 %` `97.5 %`        p
##   <dbl>   <dbl>    <dbl>    <dbl>
## 1 0.187   0.152    0.227 8.48e-62
## 2 1.12    0.852    1.47  4.16e- 1
## 3 1.37    1.14     1.66  7.86e- 4
## 4 0.892   0.796    0.999 4.72e- 2
```

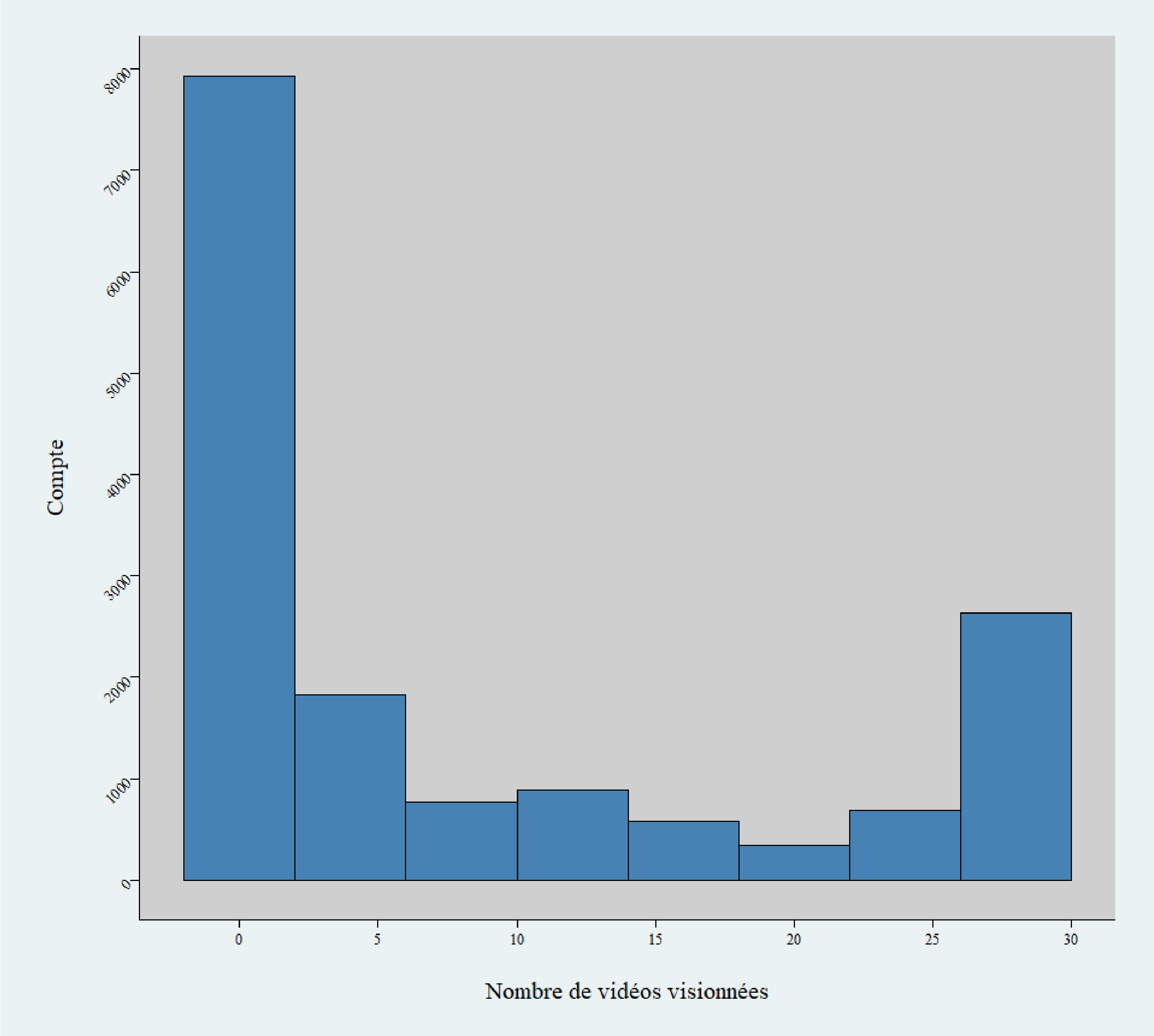
6.3 Forest plot des odds ratio

```
mooc_glm %>% or_plot(dependent, explanatory)
```

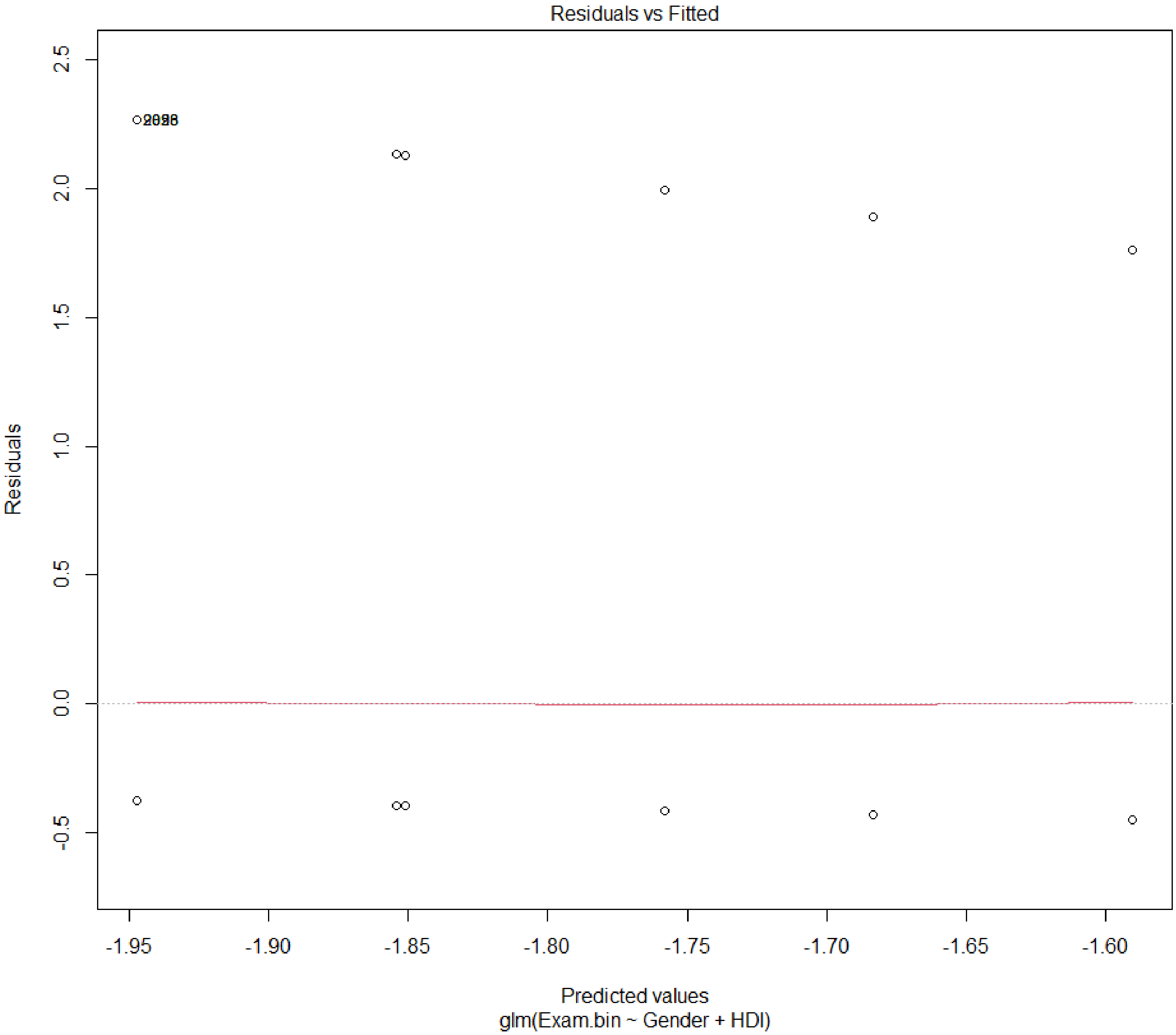


6.4 Données de comptage, loi de Poisson

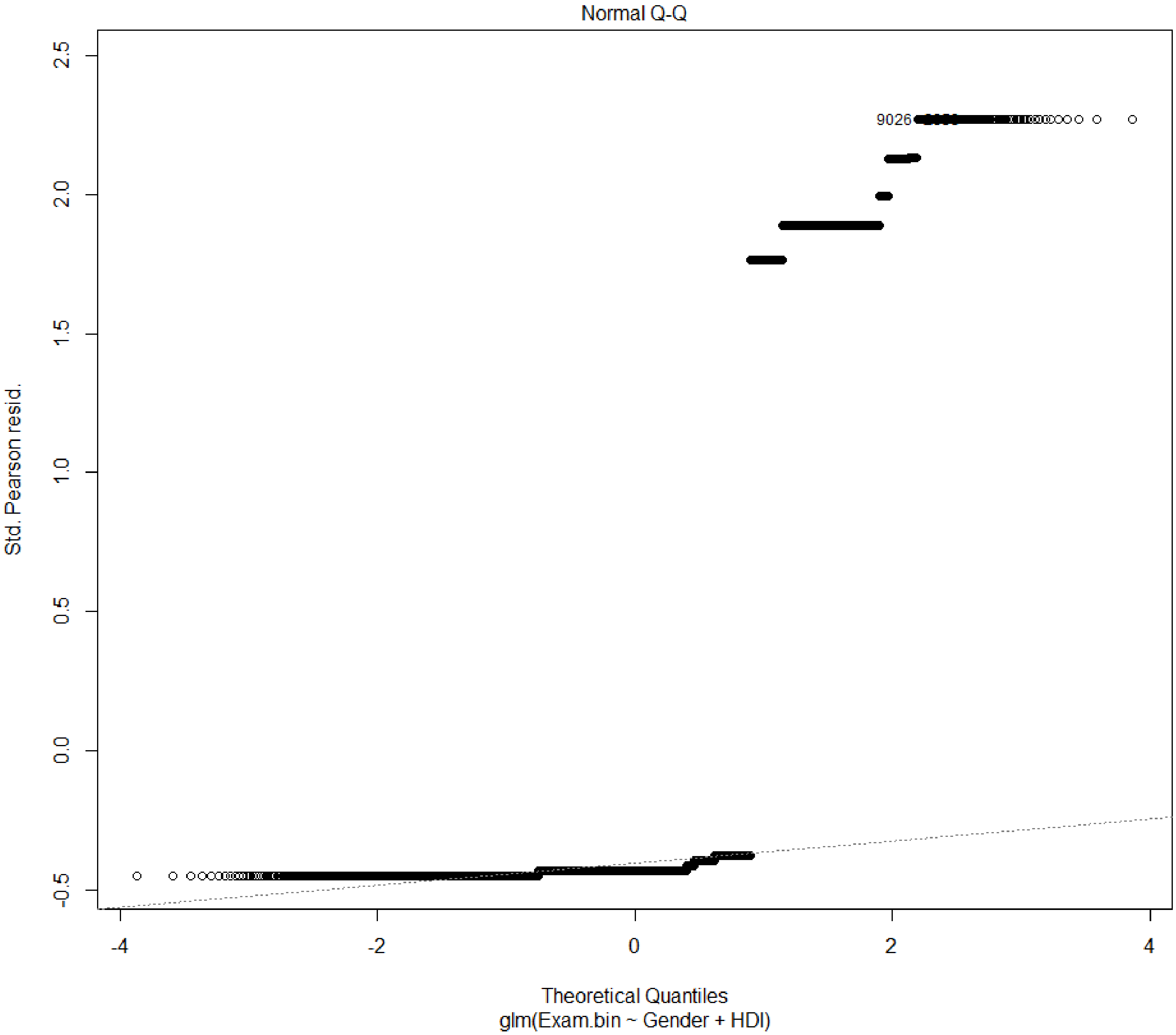
```
ggplot(mooc_selection, aes(Total_video)) +
  geom_histogram(binwidth = 4 ,fill = "steelblue", color = "black") +
  scale_x_continuous(breaks = seq(0, 30, 5)) +
  scale_y_continuous(breaks = seq(0, 8000, 1000)) +
  labs(x = "\nNombre de vidéos visionnées",
       y = "Compte\n") +
  theme(axis.text.y = element_text(angle = 45))
```



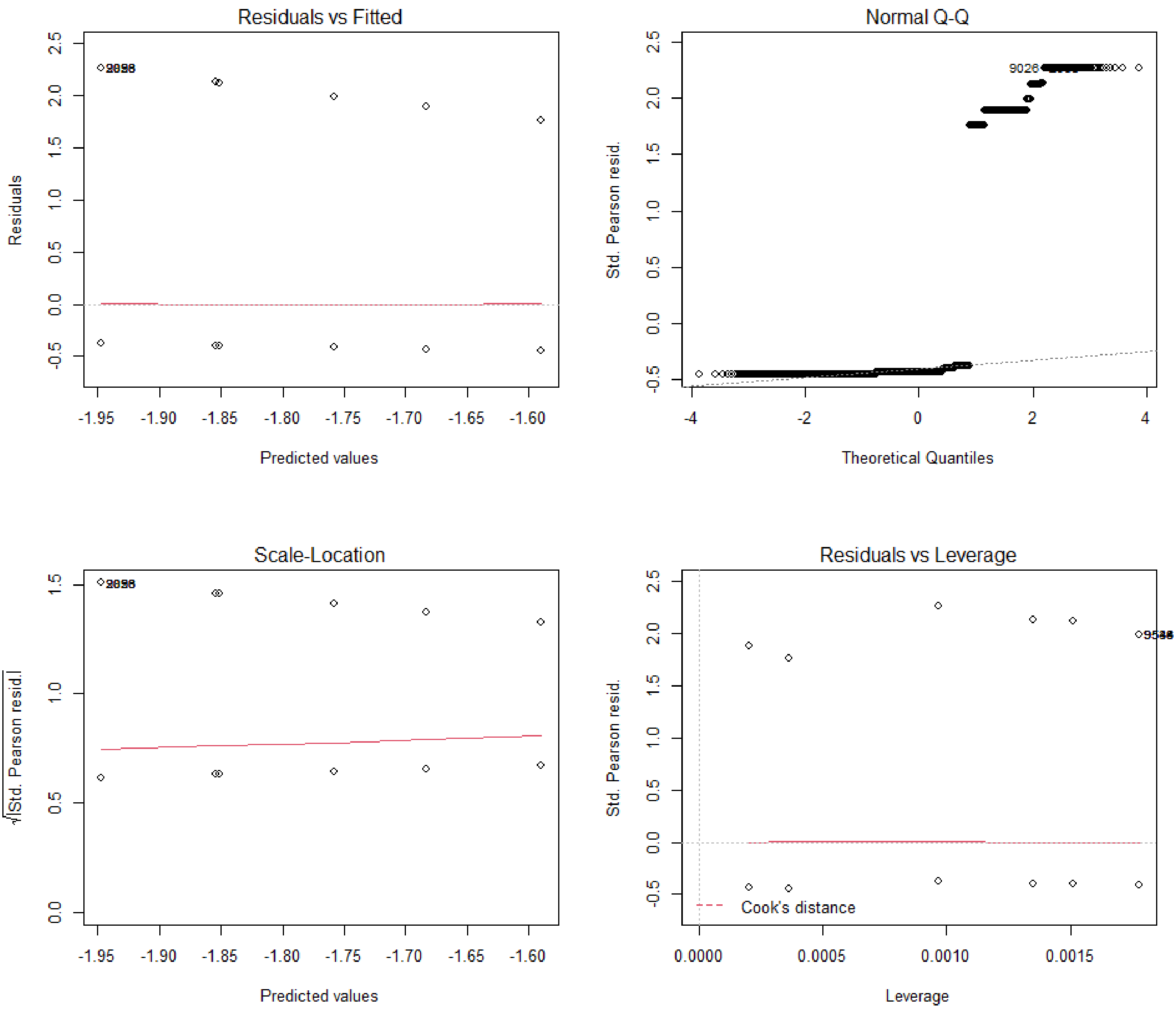

```
mod_3 <- glm(Exam.bin ~ Gender + HDI, mooc_glm, family = "poisson")

plot(mod_3, which = c(1,2))
```



Residuals vs Fitted
Residuals
Predicted values
glm(Exam.bin ~ Gender + HDI)
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.95
-1.90
-1.85
-1.80
-1.75
-1.70
-1.65
-1.60




```r
par(mfrow = c(2, 2))
plot(mod_3)
```


Residuals vs Fitted
Residuals
Predicted values
Normal Q-Q
Std. Pearson resid.
Theoretical Quantiles
9026
Scale-Location
√|Std. Pearson resid.|
Predicted values
Residuals vs Leverage
Std. Pearson resid.
Leverage
Cook's distance