Statistics

MEENOWA Sarvesh

28/11/2021

```
# Import required packages
library(readr)
library(plyr)
library(dplyr)
library(plotly)
library(xtable)
library(tufte)
library(summarytools)
library(dplyr)
library(vcd)
#install.packages("multcomp")
library(multcomp)
library(finalfit)
library(DHARMa)
library(ggplot2)
#install.packages("pscl")
library(pscl) #McFadden , pseudo-R2 library
library(survival)
library(survminer)
library(naniar)
library(broom)
```

3 Data wrangling, feature engineering

```
# Import effec data files
effec1_df <- read_csv("H:/Downloads/Datatsets/effec1.quest.compil.csv",
                      locale = locale(encoding = "ISO-8859-1"))
effec2_df <- read_csv("H:/Downloads/Datatsets/effec2.quest.compil.csv",
                      locale = locale(encoding = "ISO-8859-1"))
effec3_df <- read_csv("H:/Downloads/Datatsets/effec3.quest.compil.csv",
                      locale = locale(encoding = "ISO-8859-1"))
```

```
# rbind(append rows) effec data files
effec_df <- rbind.fill(effec1_df, effec2_df, effec3_df)
```

```
# Import usages_effec data files
usages_effec1_df <- read_csv("H:/Downloads/Datatsets/usages.effec1.csv")
usages_effec2_df <- read_csv("H:/Downloads/Datatsets/usages.effec2.csv")
usages_effec3_df <- read_csv("H:/Downloads/Datatsets/usages.effec3.csv")
```

```
# rbind usages_effec data files
usages_effec_df <- rbind.fill(usages_effec1_df, usages_effec2_df,
                              usages_effec3_df)
```

```
# Merge effec_df and usages_effec_df with Student_ID as key
df_no_HDI <- full_join(effec_df, usages_effec_df, by="Student_ID")
```

```
# Import countries_hdi data file
#Assign headers to each column i.e Country, HDI, and index
countries_HDI_df <- read_csv("H:/Downloads/Datatsets/countries.HDI.csv",
                             locale = locale(encoding = "ISO-8859-1"),
                             col_names = c("Country","HDI", "Index"))
```

```
# Change H and M HDI to I
##Group together, for the HDI variable, the High and Medium level to create a
#new intermediate level.
levels(countries_HDI_df$HDI) <- c(levels(countries_HDI_df$HDI), "I")
countries_HDI_df$HDI[countries_HDI_df$HDI == "M"]  <- "I"
countries_HDI_df$HDI[countries_HDI_df$HDI == "H"]  <- "I"
```

```
# Merge df_no_HDI and countries_HDI_df
full_df <- full_join(df_no_HDI, countries_HDI_df[c("Country","HDI")], by = "Country")
```

```
#export full df as csv
#write.csv(full_df,"H:/Downloads/Datatsets/full_df.csv", row.names = FALSE)

full_df <- read.csv("H:/Downloads/Datatsets/full_df.csv", encoding="utf-8")
```

4 Describing behaviour of the courses

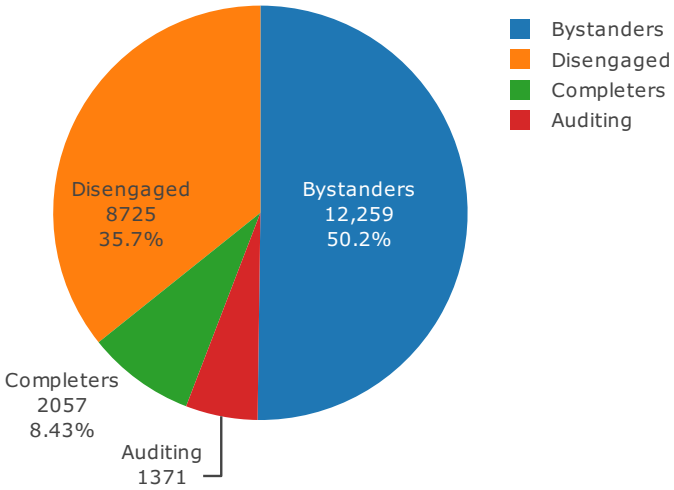
```
#completers , exam bin is used as proxy for completion
completers = nrow(full_df[which(full_df$Exam.bin == 1),])
#get number of videos for each student
full_df$n.videos <- rowSums(full_df[,60:94],na.rm=T)
#auditors
auditing = nrow(full_df %>% filter(Exam.bin == 0  & last.quizz ==0 & Assignment.bin==0&n.videos/35 >0.1
#bystanders
bystanders = nrow(full_df %>% filter(Exam.bin == 0  & last.quizz ==0 & Assignment.bin==0&n.videos/35 <=
#disengaged learners
disengaged = nrow(full_df %>% filter(Exam.bin == 0 & (Quizz.1.bin == 1 | Quizz.2.bin == 1 | Quizz.3.bin
```

```
#adding type of learners to our dataframe to use them later in survival analysis
full_df <- full_df %>%
       mutate(learner = case_when(Exam.bin == 1 ~ "completers",
                                  Exam.bin == 0  & last.quizz ==0 & Assignment.bin==0&n.videos/35 >0.1 ~
                                  Exam.bin == 0  & last.quizz ==0 & Assignment.bin==0&n.videos/35 <=0.1 ~
                                  Exam.bin == 0 & (Quizz.1.bin == 1 | Quizz.2.bin == 1 | Quizz.3.bin == 1
       ))
```

```r
#create dataframe of type of learners and their values
df_prop <- data.frame(first_column=c('Completers','Auditing','Bystanders','Disengaged'),
                      second_column=c(completers,auditing,bystanders,disengaged))
```

```r
#rename columns
colnames(df_prop) <- c("Types","Values")
```

```r
#plot pie chart in plotly
fig <- plot_ly()


fig <- df_prop %>% plot_ly(type='pie', labels=~Types, values=~Values,textinfo="label+percent+value",

              insidetextorientation='radial')

fig
```



5.1 From Student's t-test to two-ways ANOVAs

Compare the number of views of videos between genders.

- Assuming equal variance, var = T

```r
t.test(n.videos ~ Gender,data=full_df,var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  n.videos by Gender
## t = -3.544, df = 9929, p-value = 0.000396
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.5730798 -0.4526372
## sample estimates:
##  mean in group un homme mean in group une femme
##                15.62396                16.63681
```

- Assuming unequal variance , var = F

```r
t.test(n.videos ~ Gender,data=full_df,var.equal=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  n.videos by Gender
## t = -3.5174, df = 6247.4, p-value = 0.000439
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.5773589 -0.4483581
## sample estimates:
##  mean in group un homme mean in group une femme
##                15.62396                16.63681
```

- Which test should you use to assess whether the difference is statistically significant ?
  - comparing two independent groups

Compare the number of views of videos depending on the HDI of the country of origin. Same questions. Which test should you use to assess whether the difference is statistically significant ?

```r
#HDI has more than 2 groups, so we use one-way anova
model1 <- aov(n.videos ~ HDI, data = full_df)
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: n.videos
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## HDI           2 1197321  598660  6836.3 < 2.2e-16 ***
## Residuals 28373 2484641      88
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#get latex table
#print(xtable(model1))
```

- What is the difference between the two tests you just used ?
  - difference between independent t-tests and one way ANOVA

Use Gender, HDI and socioeconomic status as explaining variables (lm command in R, lm(y x1+x2)). Introduce an ANOVA table (anova(model) in R) in your report. (socioeconomic status ==> CSP)

```
model2 <- anova(lm(n.videos~HDI,full_df))
model2
```

```
## Analysis of Variance Table
##
## Response: n.videos
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## HDI           2 1197321  598660  6836.3 < 2.2e-16 ***
## Residuals 28373 2484641      88
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#get latex table of model 2
#print(xtable(model2))
```

```
#Gender and HDI- ind.variables
model3 <- anova(lm(n.videos~Gender+HDI,full_df))
model3
```

```
## Analysis of Variance Table
##
## Response: n.videos
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## Gender       1    2252    2252  13.437  0.000248 ***
## HDI          2  102869   51435 306.961 < 2.2e-16 ***
## Residuals 9833 1647626     168
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#print(xtable(model3))
```

```
#ind var : gender, hdi, csp
model4 <- anova(lm(n.videos~Gender+HDI+CSP,full_df))
model4
```

```
## Analysis of Variance Table
##
## Response: n.videos
##             Df  Sum Sq Mean Sq  F value    Pr(>F)
## Gender       1    2104    2104  12.6321  0.000381 ***
## HDI          2  103062   51531 309.3229 < 2.2e-16 ***
```

```
## CSP         10    8265     826   4.9609 3.293e-07 ***
## Residuals 9748 1623955     167
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#print(xtable(model4))
```

5.2 Model refinement, pairwise comparisons

Update the model, and add an interaction parameter in the it (For instance Gender*HDI in R). Use the summary of the model to see the interaction parameter.

```
model5 <- lm(n.videos~Gender+HDI+Gender*HDI,full_df)
model5
```

```
##
## Call:
## lm(formula = n.videos ~ Gender + HDI + Gender * HDI, data = full_df)
##
## Coefficients:
##           (Intercept)        Genderune femme                   HDII
##                 8.179                  1.608                 5.165
##                 HDITH   Genderune femme:HDII  Genderune femme:HDITH
##                 9.355                 -3.757                -1.458
```

```
print(summary(model5))
```

```
##
## Call:
## lm(formula = n.videos ~ Gender + HDI + Gender * HDI, data = full_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.684 -11.345  -3.535  14.465  26.821
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             8.1794     0.3838  21.313  < 2e-16 ***
## Genderune femme         1.6077     0.9881   1.627  0.10375
## HDII                    5.1653     0.6964   7.418 1.29e-13 ***
## HDITH                   9.3552     0.4250  22.014  < 2e-16 ***
## Genderune femme:HDII   -3.7571     1.3984  -2.687  0.00723 **
## Genderune femme:HDITH  -1.4578     1.0351  -1.408  0.15903
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.94 on 9831 degrees of freedom
##   (18633 observations deleted due to missingness)
## Multiple R-squared:  0.06069,    Adjusted R-squared:  0.06022
## F-statistic:   127 on 5 and 9831 DF,  p-value: < 2.2e-16
```

```r
#print(xtable(summary(model5)))
```

```r
#tukey hsd on interaction parameters
model_interaction <- aov(n.videos~Gender*HDI, data=full_df)



TukeyHSD(model_interaction, conf.level=.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = n.videos ~ Gender * HDI, data = full_df)
##
## $Gender
##                         diff       lwr      upr    p adj
## une femme-un homme 1.023198 0.4761975 1.570198 0.000247
##
## $HDI
##          diff      lwr      upr p adj
## I-B  3.983320 2.603921 5.362720     0
## TH-B 8.960934 8.063119 9.858748     0
## TH-I 4.977613 3.822447 6.132779     0
##
## $`Gender:HDI`
##                                 diff        lwr        upr     p adj
## une femme:B-un homme:B     1.6077092 -1.2086283  4.4240467 0.5804967
## un homme:I-un homme:B      5.1653385  3.1805130  7.1501641 0.0000000
## une femme:I-un homme:B     3.0159828  0.4843342  5.5476314 0.0089909
## un homme:TH-un homme:B     9.3551867  8.1439222 10.5664512 0.0000000
## une femme:TH-un homme:B    9.5050626  8.2019692 10.8081560 0.0000000
## un homme:I-une femme:B     3.5576294  0.4789709  6.6362878 0.0127222
## une femme:I-une femme:B    1.4082736 -2.0482926  4.8648398 0.8552188
## un homme:TH-une femme:B    7.7474775  5.1006365 10.3943185 0.0000000
## une femme:TH-une femme:B   7.8973534  5.2072497 10.5874571 0.0000000
## une femme:I-un homme:I    -2.1493558 -4.9699277  0.6712161 0.2509925
## un homme:TH-un homme:I     4.1898481  2.4538922  5.9258040 0.0000000
## une femme:TH-un homme:I    4.3397241  2.5384929  6.1409552 0.0000000
## un homme:TH-une femme:I    6.3392039  3.9975688  8.6808390 0.0000000
## une femme:TH-une femme:I   6.4890798  4.0986519  8.8795078 0.0000000
## une femme:TH-un homme:TH   0.1498759 -0.7287952  1.0285470 0.9966601
```

```r
#xtable(tidy(TukeyHSD(model_interaction, conf.level=.95)))
```

Use a stepwise algorithm (step command in R) to assess the performance of various versions of the model (use both forward and backward options).

```r
#convert birth year to integer
full_df$birth.year <- as.integer(full_df$birth.year)

#create age groups
full_df$birth.year[full_df$birth.year<1940] <- NA
```

```r
full_df$birth.year[full_df$birth.year>2020]<- NA
#calculate age
full_df$age <- 2020-full_df$birth.year
#create seq
seq_1 = seq(0,90,by=3)
#break age into seq1
full_df$age.group <- cut(full_df$age,seq_1)

head(full_df$age.group)
```

```
## [1] <NA>    <NA>    (33,36] (51,54] (36,39] <NA>
## 30 Levels: (0,3] (3,6] (6,9] (9,12] (12,15] (15,18] (18,21] (21,24] ... (87,90]
```

```r
#remove all Nas in the following variables
full_df_subset = na.omit(full_df[c('Gender','HDI','n.videos','CSP','age.group','CSP.fin')])

model6 <- lm(n.videos~Gender+HDI+CSP+age.group,full_df_subset)

step(model6,direction="both")
```

```
## Start:  AIC=48098.51
## n.videos ~ Gender + HDI + CSP + age.group
##
##             Df Sum of Sq     RSS   AIC
## - Gender     1        25 1563576 48097
## <none>                   1563552 48099
## - CSP       10      7289 1570841 48122
## - age.group 20     11226 1574778 48126
## - HDI        2     77848 1641400 48551
##
## Step:  AIC=48096.65
## n.videos ~ HDI + CSP + age.group
##
##             Df Sum of Sq     RSS   AIC
## <none>                   1563576 48097
## + Gender     1        25 1563552 48099
## - CSP       10      7266 1570842 48120
## - age.group 20     11205 1574781 48124
## - HDI        2     78993 1642569 48555

##
## Call:
## lm(formula = n.videos ~ HDI + CSP + age.group, data = full_df_subset)
##
## Coefficients:
##                                                          (Intercept)
##                                                               3.5979
##                                                                 HDII
##                                                               4.5352
##                                                                HDITH
##                                                               8.9506
##                       CSPArtisans, commerÃ§ants, chefs d'entreprise
```

```
##                                                          3.3687
##                         CSPArtisans, commerçants, chefs d'entreprise
##                                                          1.4515
##                             CSPCadres et professions intellectuelles
##                                                          2.5882
##                                                       CSPEmployés
##                                                          2.9086
##                                           CSPEn recherche d'emploi
##                                                          4.6907
##                                                      CSPEtudiants
##                                                          2.7876
## CSPInactif (autre que étudiant, retraité, ou en recherche d'emploi)
##                                                          5.1031
##                                                       CSPOuvriers
##                                                          5.2653
##                                       CSPProfessions intermédiaires
##                                                          1.1503
##                                                     CSPRetraités
##                                                          5.2261
##                                                 age.group(21,24]
##                                                          -0.3869
##                                                 age.group(24,27]
##                                                          1.7080
##                                                 age.group(27,30]
##                                                          1.1518
##                                                 age.group(30,33]
##                                                          0.6700
##                                                 age.group(33,36]
##                                                          0.5962
##                                                 age.group(36,39]
##                                                          2.4483
##                                                 age.group(39,42]
##                                                          2.6719
##                                                 age.group(42,45]
##                                                          2.1757
##                                                 age.group(45,48]
##                                                          2.5871
##                                                 age.group(48,51]
##                                                          2.9964
##                                                 age.group(51,54]
##                                                          4.3159
##                                                 age.group(54,57]
##                                                          3.0454
##                                                 age.group(57,60]
##                                                          4.3389
##                                                 age.group(60,63]
##                                                          3.0874
##                                                 age.group(63,66]
##                                                          5.0376
##                                                 age.group(66,69]
##                                                          2.1073
##                                                 age.group(69,72]
##                                                          3.4817
##                                                 age.group(72,75]
```

```
##                                                          5.4535
##                                                 age.group(75,78]
##                                                          2.3168
##                                                 age.group(78,81]
##                                                         -15.2746
```

```
#Linear model with 3 ind var : Gender,HDI,csp + interaction parameter(Gender*HDI)
anova(lm(n.videos~Gender+HDI+CSP.fin,full_df_subset))
```

```
## Analysis of Variance Table
##
## Response: n.videos
##             Df  Sum Sq Mean Sq  F value    Pr(>F)
## Gender       1    1677    1677   9.9762  0.001591 **
## HDI          2   90881   45440 270.3316 < 2.2e-16 ***
## CSP.fin      6    6437    1073   6.3826 1.015e-06 ***
## Residuals 9380 1576694     168
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#xtable(anova(lm(n.videos~Gender+HDI+CSP.fin,full_df_subset)))
tidy(TukeyHSD(aov(lm(n.videos~Gender+HDI+CSP.fin,full_df_subset))))
```

```
## # A tibble: 25 x 7
##    term    contrast           null.value estimate conf.low conf.high adj.p.value
##    <chr>   <chr>                   <dbl>    <dbl>    <dbl>     <dbl>       <dbl>
##  1 Gender  une femme-un homme          0    0.902    0.342     1.46      0.00159
##  2 HDI     I-B                        0    4.22     2.76      5.67      0
##  3 HDI     TH-B                       0    9.00     8.04      9.96      0
##  4 HDI     TH-I                       0    4.78     3.58      5.98      0
##  5 CSP.fin Artisans, commerç~         0   -2.07    -5.32      1.17      0.491
##  6 CSP.fin Autre-Artisans, c~         0   -0.830   -4.21      2.55      0.991
##  7 CSP.fin Cadres et profess~         0   -1.04    -4.01      1.93      0.946
##  8 CSP.fin Employés-Artisans~         0   -1.41    -4.59      1.76      0.846
##  9 CSP.fin En recherche d'em~         0    0.491   -2.63      3.61      0.999
## 10 CSP.fin Etudiants-Artisan~         0   -2.23    -5.26      0.796     0.310
## # ... with 15 more rows
```

```
#xtable(tidy(TukeyHSD(aov(lm(n.videos~Gender+HDI+CSP.fin,full_df_subset)))))
```

- Age group is divided into too many parts, so we create a smaller group

```
#create second age group
full_df$Age.group <- cut(full_df$age,c(0,30,50,80,100))

head(full_df$age.group2)
```

```
## NULL
```

```r
full_df_subset = na.omit(full_df[c('Gender','HDI','n.videos','CSP','age.group','Age.group','CSP.fin')])



model7 <- lm(n.videos~Gender+HDI+CSP.fin+Age.group,full_df_subset)

(summary(step(model7,direction="both")))
```

```
## Start:  AIC=48105.64
## n.videos ~ Gender + HDI + CSP.fin + Age.group
##
##             Df Sum of Sq     RSS   AIC
## - Gender     1        14 1572104 48104
## <none>                   1572090 48106
## - CSP.fin    6      5009 1577099 48124
## - Age.group  2      4604 1576694 48129
## - HDI        2     80975 1653065 48573
##
## Step:  AIC=48103.73
## n.videos ~ HDI + CSP.fin + Age.group
##
##             Df Sum of Sq     RSS   AIC
## <none>                   1572104 48104
## + Gender     1        14 1572090 48106
## - CSP.fin    6      4995 1577099 48122
## - Age.group  2      4601 1576705 48127
## - HDI        2     82286 1654390 48579

##
## Call:
## lm(formula = n.videos ~ HDI + CSP.fin + Age.group, data = full_df_subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.700 -11.573  -3.123  14.300  28.008
##
## Coefficients:
##                                                  Estimate Std. Error t value
## (Intercept)                                        9.1419     1.1819   7.735
## HDII                                               4.4496     0.6260   7.108
## HDITH                                              9.0095     0.4242  21.239
## CSP.finArtisans, commerçants, chefs d'entreprise  -2.0213     1.0991  -1.839
## CSP.finAutre                                      -0.8497     1.1450  -0.742
## CSP.finCadres et professions intellectuelles      -0.8904     1.0062  -0.885
## CSP.finEmployés                                   -0.9327     1.0847  -0.860
## CSP.finEn recherche d'emploi                       0.8408     1.0604   0.793
## CSP.finEtudiants                                  -1.6169     1.0852  -1.490
## Age.group(30,50]                                  -0.1283     0.4989  -0.257
## Age.group(50,80]                                   1.7077     0.5893   2.898
##                                                  Pr(>|t|)
## (Intercept)                                      1.14e-14 ***
## HDII                                             1.26e-12 ***
## HDITH                                             < 2e-16 ***
```

```
## CSP.finArtisans, commerçants, chefs d'entreprise  0.06593 .
## CSP.finAutre                                      0.45804
## CSP.finCadres et professions intellectuelles      0.37624
## CSP.finEmployés                                   0.38989
## CSP.finEn recherche d'emploi                      0.42782
## CSP.finEtudiants                                  0.13628
## Age.group(30,50]                                  0.79709
## Age.group(50,80]                                  0.00377 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.95 on 9379 degrees of freedom
## Multiple R-squared:  0.06182,    Adjusted R-squared:  0.06082
## F-statistic:  61.8 on 10 and 9379 DF,  p-value: < 2.2e-16
```

```
#create second age group
full_df$age.group2 <- cut(full_df$age,c(0,30,50,80,100))

head(full_df$age.group2)
```

```
## [1] <NA>    <NA>    (30,50] (50,80] (30,50] <NA>
## Levels: (0,30] (30,50] (50,80] (80,100]
```

```
#create subset for linear model

full_df_subset = na.omit(full_df[c('Gender','HDI','n.videos','CSP.fin','age.group','age.group2','learne
#create linear model for HDI,CSP,
model7 <- lm(n.videos~Gender+HDI+CSP.fin+Age.group,full_df_subset)

(summary(step(model7,direction="both")))
```

```
## Start:  AIC=48082.11
## n.videos ~ Gender + HDI + CSP.fin + Age.group
##
##             Df Sum of Sq     RSS   AIC
## - Gender     1        18 1570928 48080
## <none>                   1570909 48082
## - CSP.fin    6      4969 1575879 48100
## - Age.group  2      4658 1575567 48106
## - HDI        2     81219 1652128 48551
##
## Step:  AIC=48080.22
## n.videos ~ HDI + CSP.fin + Age.group
##
##             Df Sum of Sq     RSS   AIC
## <none>                   1570928 48080
## + Gender     1        18 1570909 48082
## - CSP.fin    6      4951 1575879 48098
## - Age.group  2      4654 1575581 48104
## - HDI        2     82493 1653420 48557

##
```

```
## Call:
## lm(formula = n.videos ~ HDI + CSP.fin + Age.group, data = full_df_subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.712 -11.560  -3.111  14.288  28.020
##
## Coefficients:
##                                                 Estimate Std. Error t value
## (Intercept)                                       9.1255     1.1818   7.722
## HDII                                              4.4490     0.6259   7.108
## HDITH                                             9.0201     0.4242  21.266
## CSP.finArtisans, commerçants, chefs d'entreprise -2.0208     1.0989  -1.839
## CSP.finAutre                                     -0.8497     1.1448  -0.742
## CSP.finCadres et professions intellectuelles     -0.8891     1.0061  -0.884
## CSP.finEmployés                                  -0.9049     1.0847  -0.834
## CSP.finEn recherche d'emploi                      0.8442     1.0602   0.796
## CSP.finEtudiants                                 -1.5860     1.0852  -1.462
## Age.group(30,50]                                 -0.1252     0.4991  -0.251
## Age.group(50,80]                                  1.7218     0.5895   2.921
##                                                 Pr(>|t|)
## (Intercept)                                     1.27e-14 ***
## HDII                                            1.26e-12 ***
## HDITH                                            < 2e-16 ***
## CSP.finArtisans, commerçants, chefs d'entreprise   0.0659 .
## CSP.finAutre                                       0.4580
## CSP.finCadres et professions intellectuelles       0.3769
## CSP.finEmployés                                    0.4042
## CSP.finEn recherche d'emploi                       0.4259
## CSP.finEtudiants                                   0.1439
## Age.group(30,50]                                   0.8019
## Age.group(50,80]                                   0.0035 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.94 on 9375 degrees of freedom
## Multiple R-squared:  0.06194,    Adjusted R-squared:  0.06094
## F-statistic:  61.9 on 10 and 9375 DF,  p-value: < 2.2e-16
```

```
#latex table for figure
print(xtable((summary(step(model7,direction="both")))))
```

```
## Start:  AIC=48082.11
## n.videos ~ Gender + HDI + CSP.fin + Age.group
##
##             Df Sum of Sq     RSS   AIC
## - Gender     1        18 1570928 48080
## <none>                   1570909 48082
## - CSP.fin    6      4969 1575879 48100
## - Age.group  2      4658 1575567 48106
## - HDI        2     81219 1652128 48551
##
## Step:  AIC=48080.22
## n.videos ~ HDI + CSP.fin + Age.group
```

```
##
##             Df Sum of Sq     RSS   AIC
## <none>                    1570928 48080
## + Gender     1        18 1570909 48082
## - CSP.fin    6      4951 1575879 48098
## - Age.group  2      4654 1575581 48104
## - HDI        2     82493 1653420 48557
## % latex table generated in R 4.0.3 by xtable 1.8-4 package
## % Sat Jan 08 15:59:18 2022
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrrrr}
##   \hline
##  & Estimate & Std. Error & t value & Pr($>$$|$t$|$) \\
##   \hline
## (Intercept) & 9.1255 & 1.1818 & 7.72 & 0.0000 \\
##   HDII & 4.4490 & 0.6259 & 7.11 & 0.0000 \\
##   HDITH & 9.0201 & 0.4242 & 21.27 & 0.0000 \\
##   CSP.finArtisans, commerçants, chefs d'entreprise & -2.0208 & 1.0989 & -1.84 & 0.0659 \\
##   CSP.finAutre & -0.8497 & 1.1448 & -0.74 & 0.4580 \\
##   CSP.finCadres et professions intellectuelles & -0.8891 & 1.0061 & -0.88 & 0.3769 \\
##   CSP.finEmployés & -0.9049 & 1.0847 & -0.83 & 0.4042 \\
##   CSP.finEn recherche d'emploi & 0.8442 & 1.0602 & 0.80 & 0.4259 \\
##   CSP.finEtudiants & -1.5860 & 1.0852 & -1.46 & 0.1439 \\
##   Age.group(30,50] & -0.1252 & 0.4991 & -0.25 & 0.8019 \\
##   Age.group(50,80] & 1.7218 & 0.5895 & 2.92 & 0.0035 \\
##    \hline
## \end{tabular}
## \end{table}
```

- Assess the colinearity of all three independant variables of the last model (excluding interaction parameters). To do that, use a chi-test between HDI and Gender, produce a mosaic plot and propose its interpretation (look for residuals below -2 or above 2).
  - referring to the linear model of n.videos ~ Gender + HDI + CSP

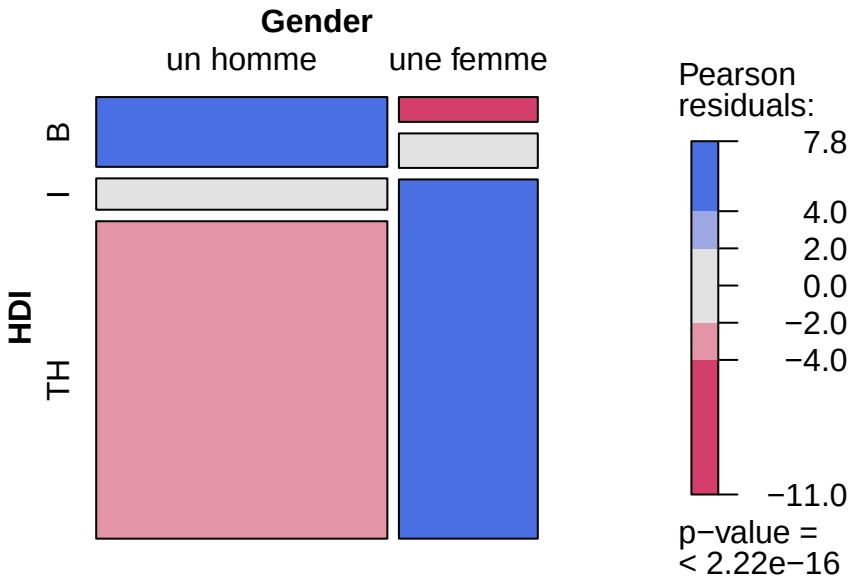
```
#references
#https://statsandr.com/blog/chi-square-test-of-independence-in-r/
#http://www.sthda.com/english/wiki/chi-square-test-of-independence-in-r
#For interpretation purposes
```

```
full_df_subset2 = na.omit(full_df[c('Gender','HDI','n.videos','CSP')])


chisq <- chisq.test(table(full_df_subset2$Gender,full_df_subset2$HDI))
chisq
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(full_df_subset2$Gender, full_df_subset2$HDI)
## X-squared = 215.1, df = 2, p-value < 2.2e-16
```

```r
#install.packages('summarytools')

# fourth method:
full_df_subset2 %$%
  ctable(Gender, HDI,
    prop = "r", chisq = TRUE, headings = FALSE
  ) %>%
  print(
    method = "render",
    style = "rmarkdown",
    footnote = NA
  )
```

```r
mosaic(~ Gender + HDI,
  direction = c("v", "h"),
  data = full_df_subset2,
  shade = TRUE
)
```



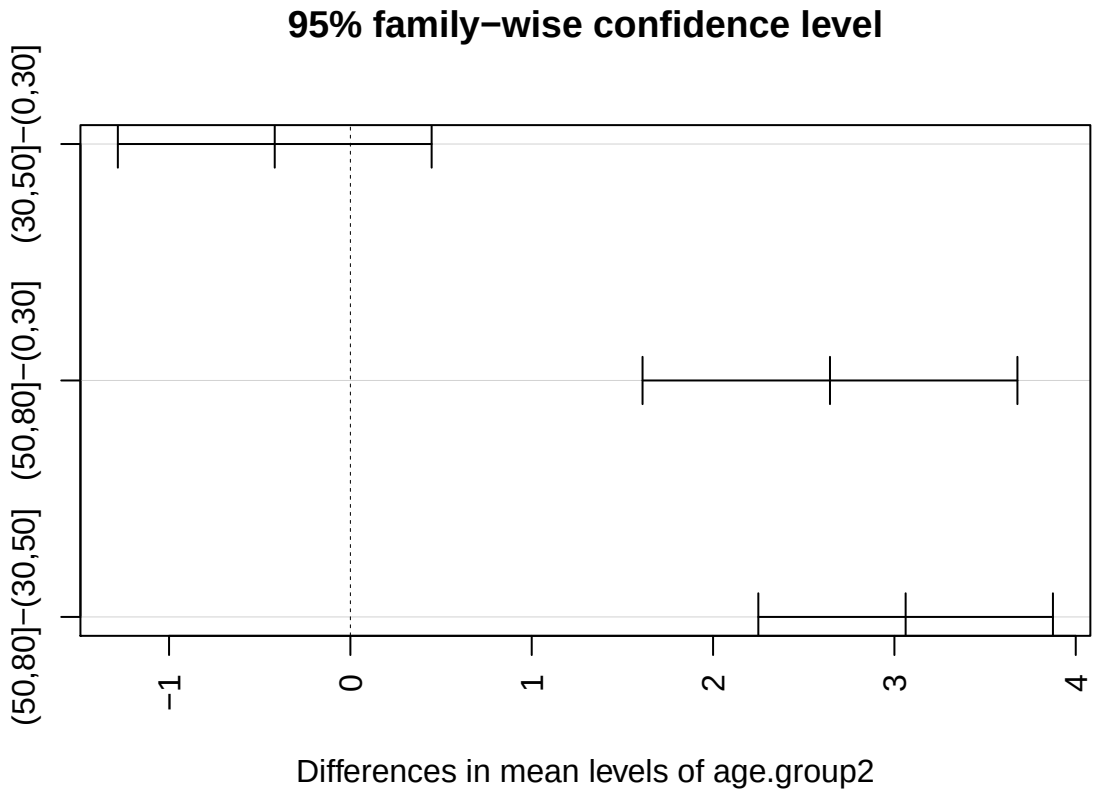
Use Tukey HSD, and propose a table, to see the pairwise differences between learners of different socioeconomic status.

```r
model8 <- aov(n.videos~age.group2, data=full_df_subset)
```

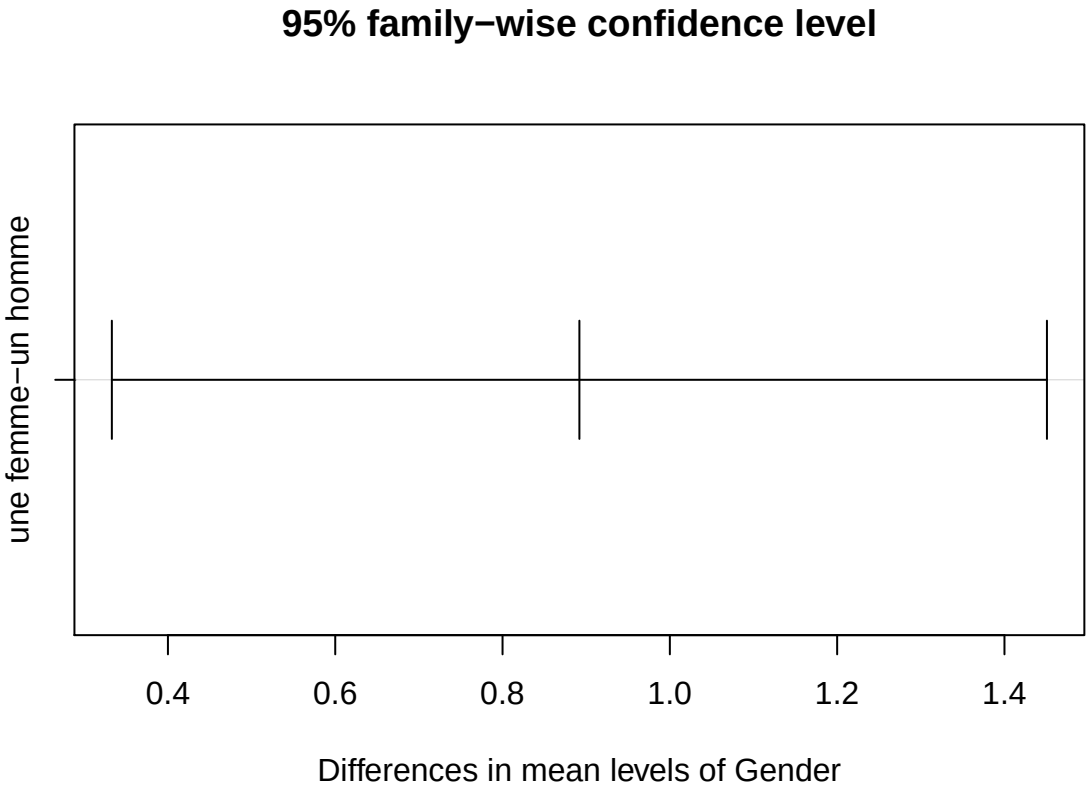
```r
TukeyHSD(model8, conf.level=.95)
```

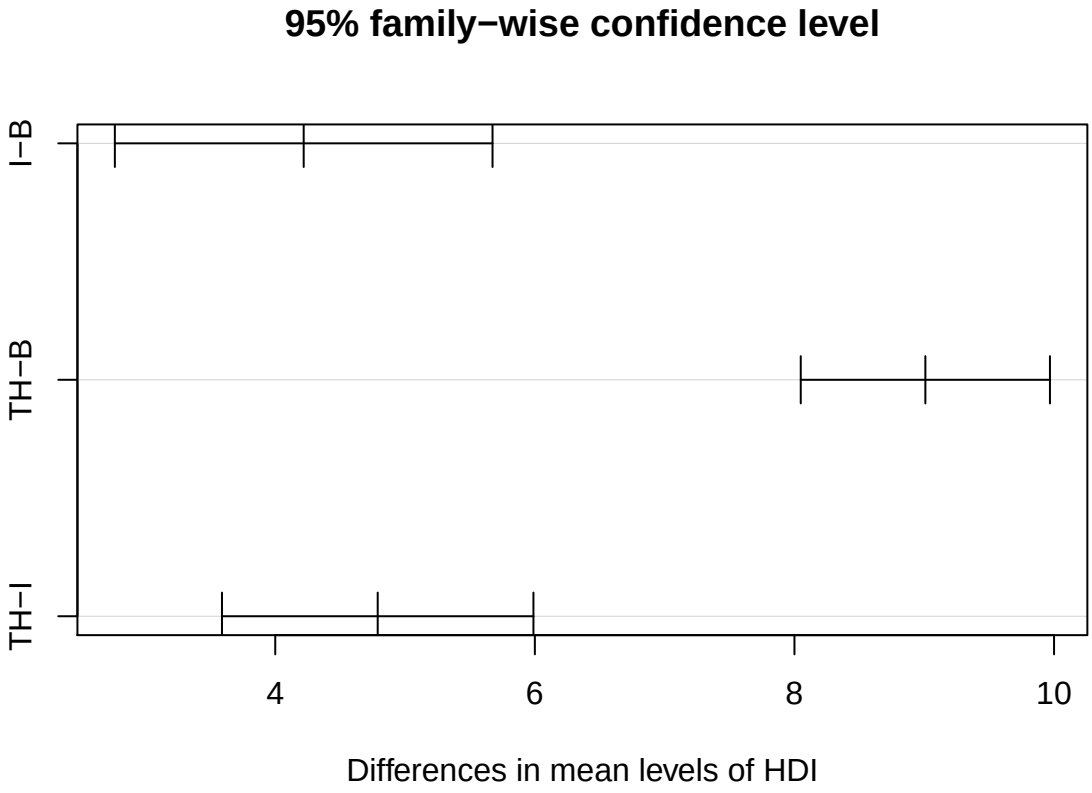
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = n.videos ~ age.group2, data = full_df_subset)
##
## $age.group2
##                       diff       lwr      upr     p adj
## (30,50]-(0,30]  -0.4172011 -1.282373 0.447971 0.4953417
## (50,80]-(0,30]   2.6447011  1.611056 3.678346 0.0000000
## (50,80]-(30,50]  3.0619022  2.249545 3.874260 0.0000000
```

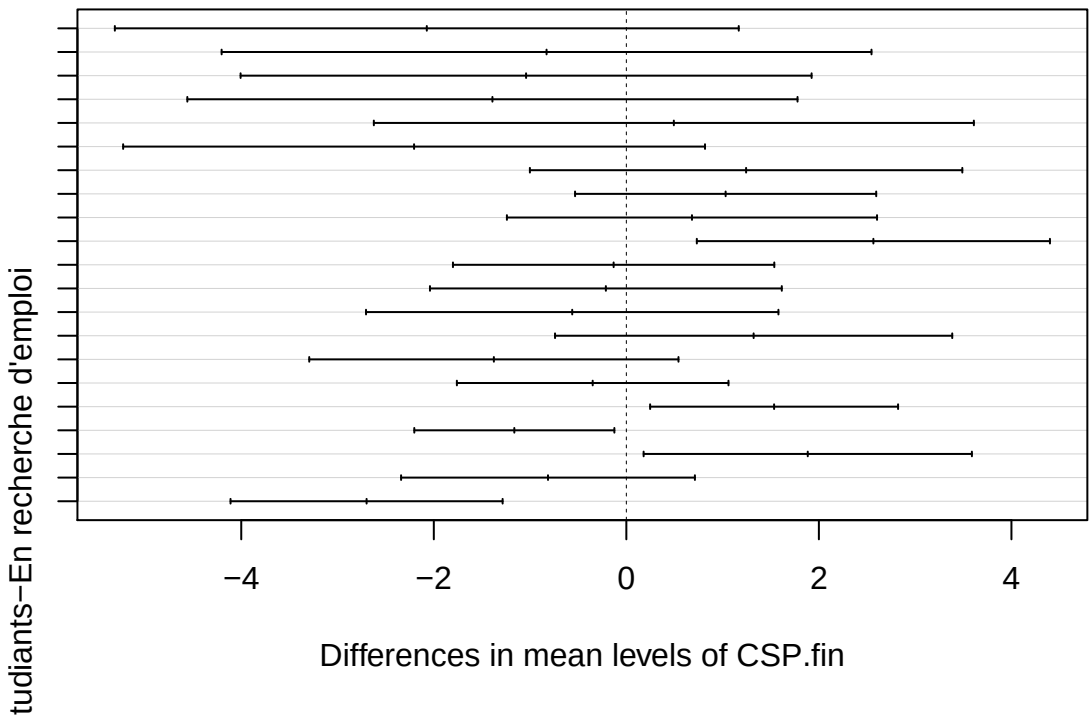
```r
#need to resize plot
plot(TukeyHSD(model8, conf.level=.95),las=3)
```

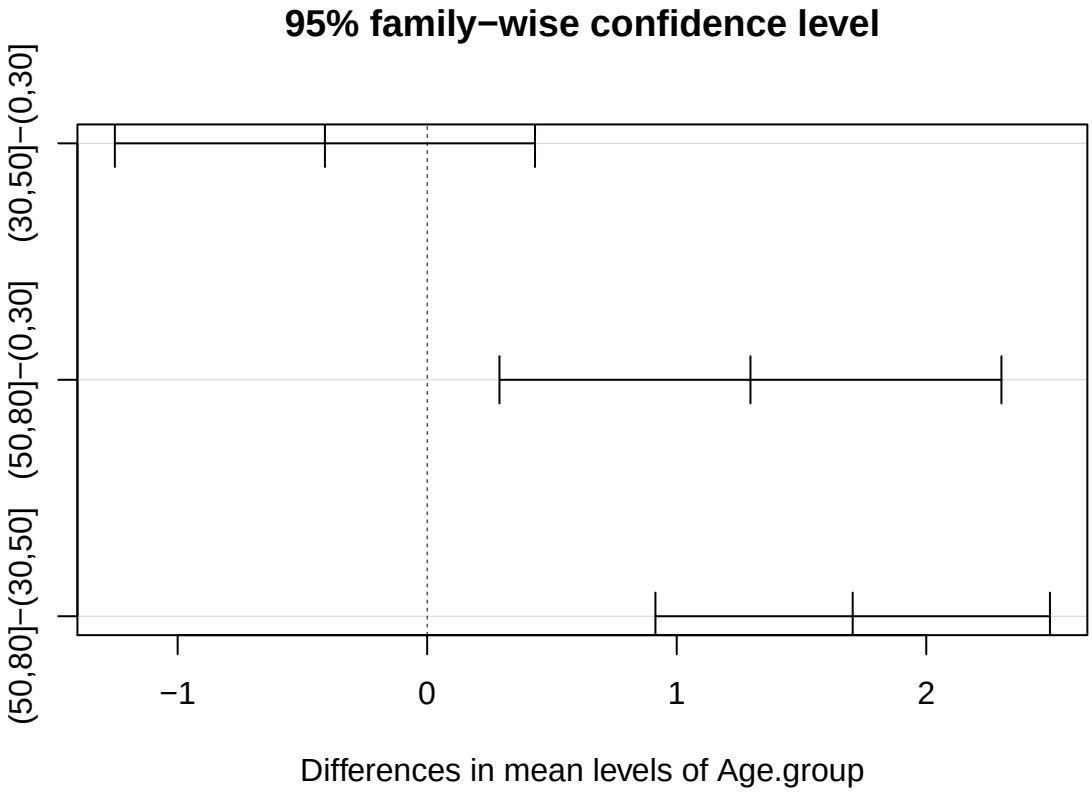



```r
#new model with gender, hdi, csp and age group 2
model9 <- aov(n.videos~Gender+HDI+CSP.fin+Age.group, data=full_df_subset)

#apply tukeyhsd to pairwise comparisons
thsd <- TukeyHSD(model9, conf.level=.95)
#xtable(tidy(thsd))

plot(TukeyHSD(model9, conf.level=.95))
```


95% family−wise confidence level
une femme−un homme
0.4
0.6
0.8
1.0
1.2
1.4
Differences in mean levels of Gender

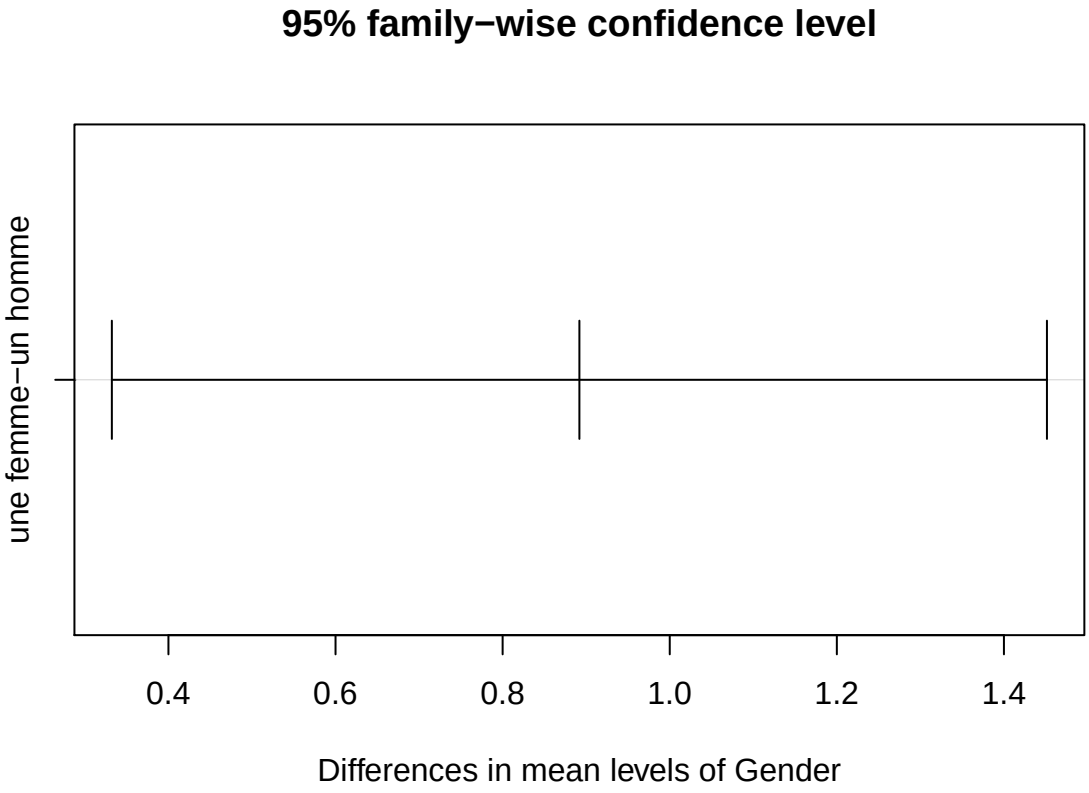

95% family−wise confidence level
I−B
TH−B
TH−I
4
6
8
10
Differences in mean levels of HDI

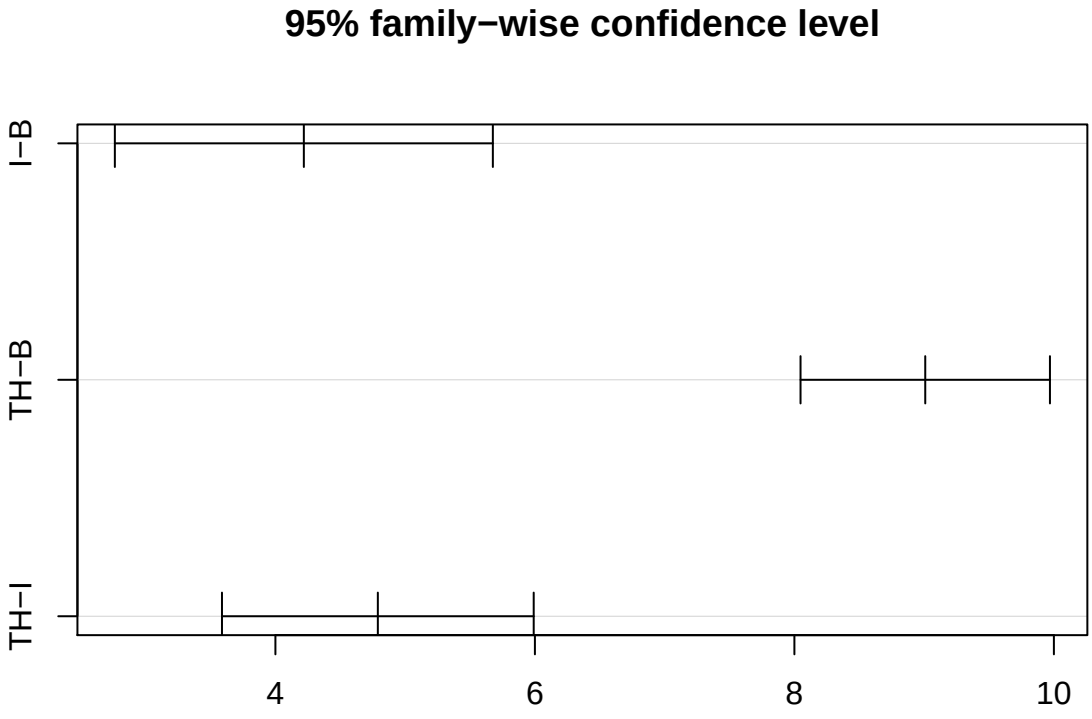

95% family−wise confidence level
tudiants−En recherche d'emploi
−4
−2
0
2
4
Differences in mean levels of CSP.fin

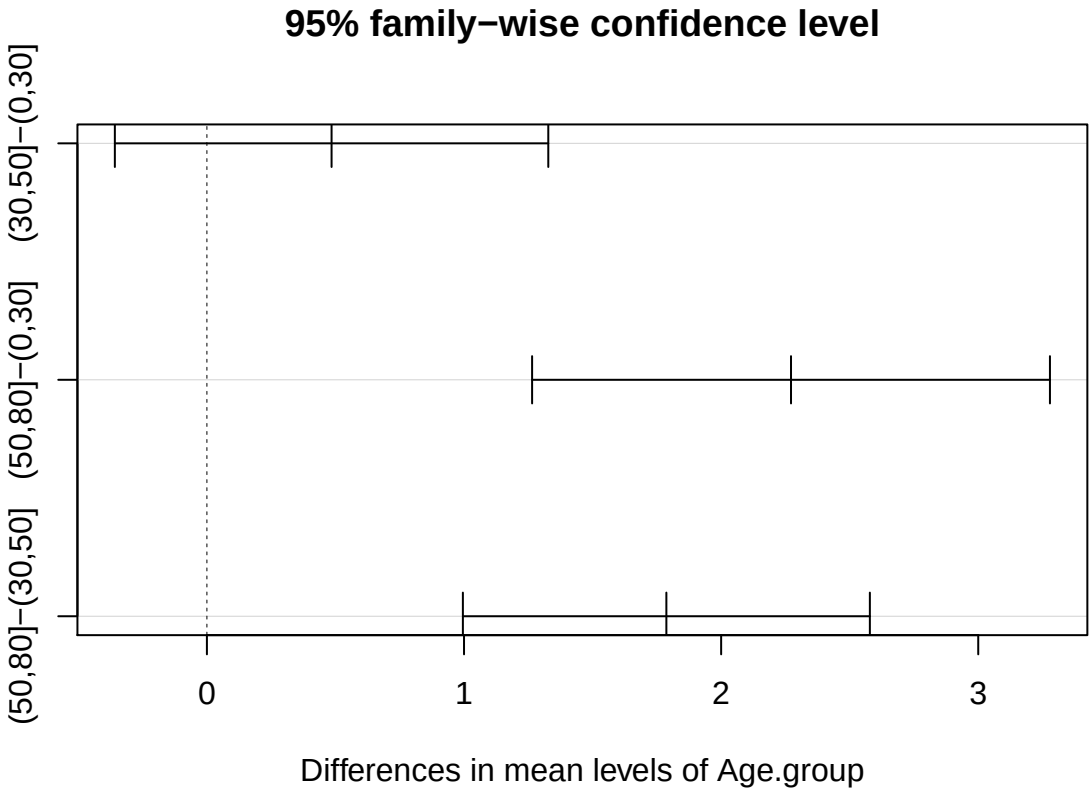


Output is too big for tukeyhsd + CSP, so we can try other alternatives

```
#new model with gender, hdi,learner type and age group 2
model_10 <- aov(n.videos~Gender+HDI+Age.group, data=full_df_subset)

#apply tukeyhsd to pairwise comparisons
thsd <- TukeyHSD(model_10, conf.level=.95)
#thsd
#xtable(tidy(thsd))

plot(TukeyHSD(model_10, conf.level=.95))
```


95% family−wise confidence level
une femme−un homme
0.4
0.6
0.8
1.0
1.2
1.4
Differences in mean levels of Gender


95% family−wise confidence level
I−B
TH−B
TH−I
4
6
8
10
Differences in mean levels of HDI



- In order to get a better understanding of the issue of pairwise comparisons,we designed a dataset with many continuous variables. Use parwise comparisons with the lm model to detect statistically significant relationships between variables. What variables appear to be correlated ? Include a graph in your report and comment it.
- First step/method : use Tukey HSD for pairwise comparisons , we can also use glht method with tukey to produce pairwise comparisons
- Apply bonferroni

```
##### Using glht method

#create model 10

model10 <- lm(n.videos~Gender+HDI+CSP.fin+age.group2,data=full_df_subset)
```

```
# running glht()
post.hoc <- glht(model10)

# displaying the result table with summary()
summary(post.hoc)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = n.videos ~ Gender + HDI + CSP.fin + age.group2,
##     data = full_df_subset)
```

```
##
## Linear Hypotheses:
##                                                        Estimate Std. Error
## (Intercept) == 0                                        9.13112    1.18198
## Genderune femme == 0                                   -0.09607    0.28932
## HDII == 0                                               4.46981    0.62904
## HDITH == 0                                              9.04152    0.42906
## CSP.finArtisans, commerçants, chefs d'entreprise == 0 -2.01898    1.09894
## CSP.finAutre == 0                                      -0.83745    1.14548
## CSP.finCadres et professions intellectuelles == 0      -0.88674    1.00613
## CSP.finEmployés == 0                                   -0.89567    1.08507
## CSP.finEn recherche d'emploi == 0                       0.85677    1.06094
## CSP.finEtudiants == 0                                  -1.57940    1.08542
## age.group2(30,50] == 0                                 -0.12386    0.49914
## age.group2(50,80] == 0                                  1.72400    0.58955
##                                                        t value Pr(>|t|)
## (Intercept) == 0                                         7.725   <0.001 ***
## Genderune femme == 0                                    -0.332   1.0000
## HDII == 0                                                7.106   <0.001 ***
## HDITH == 0                                              21.073   <0.001 ***
## CSP.finArtisans, commerçants, chefs d'entreprise == 0  -1.837   0.3792
## CSP.finAutre == 0                                       -0.731   0.9856
## CSP.finCadres et professions intellectuelles == 0       -0.881   0.9575
## CSP.finEmployés == 0                                    -0.825   0.9704
## CSP.finEn recherche d'emploi == 0                        0.808   0.9738
## CSP.finEtudiants == 0                                   -1.455   0.6505
## age.group2(30,50] == 0                                  -0.248   1.0000
## age.group2(50,80] == 0                                   2.924   0.0278 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

```
#apply bonferroni
summary(post.hoc, test = adjusted("bonferroni"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = n.videos ~ Gender + HDI + CSP.fin + age.group2,
##     data = full_df_subset)
##
## Linear Hypotheses:
##                                                        Estimate Std. Error
## (Intercept) == 0                                        9.13112    1.18198
## Genderune femme == 0                                   -0.09607    0.28932
## HDII == 0                                               4.46981    0.62904
## HDITH == 0                                              9.04152    0.42906
## CSP.finArtisans, commerçants, chefs d'entreprise == 0 -2.01898    1.09894
## CSP.finAutre == 0                                      -0.83745    1.14548
## CSP.finCadres et professions intellectuelles == 0      -0.88674    1.00613
## CSP.finEmployés == 0                                   -0.89567    1.08507
## CSP.finEn recherche d'emploi == 0                       0.85677    1.06094
## CSP.finEtudiants == 0                                  -1.57940    1.08542
## age.group2(30,50] == 0                                 -0.12386    0.49914
```

```
## age.group2(50,80] == 0                                    1.72400    0.58955
##                                                         t value Pr(>|t|)
## (Intercept) == 0                                          7.725 1.49e-13 ***
## Genderune femme == 0                                     -0.332   1.0000
## HDII == 0                                                 7.106 1.54e-11 ***
## HDITH == 0                                               21.073  < 2e-16 ***
## CSP.finArtisans, commerçants, chefs d'entreprise == 0  -1.837   0.7945
## CSP.finAutre == 0                                        -0.731   1.0000
## CSP.finCadres et professions intellectuelles == 0        -0.881   1.0000
## CSP.finEmployés == 0                                     -0.825   1.0000
## CSP.finEn recherche d'emploi == 0                         0.808   1.0000
## CSP.finEtudiants == 0                                    -1.455   1.0000
## age.group2(30,50] == 0                                   -0.248   1.0000
## age.group2(50,80] == 0                                    2.924   0.0415 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- bonferroni method)
```

```
tidy(pairwise.t.test(full_df_subset$n.videos, full_df_subset$CSP.fin , p.adjust="bonferroni"))
```

```
## # A tibble: 21 x 3
##    group1                                  group2                      p.value
##    <chr>                                   <chr>                         <dbl>
##  1 Artisans, commerçants, chefs d'entreprise Artisans, commerÃ§ants, ~   9.72e-1
##  2 Autre                                   Artisans, commerÃ§ants, ~   1   e+0
##  3 Autre                                   Artisans, commerçants, c~   1   e+0
##  4 Cadres et professions intellectuelles   Artisans, commerÃ§ants, ~   1   e+0
##  5 Cadres et professions intellectuelles   Artisans, commerçants, c~   1   e+0
##  6 Cadres et professions intellectuelles   Autre                       1   e+0
##  7 Employés                                Artisans, commerÃ§ants, ~   6.21e-3
##  8 Employés                                Artisans, commerçants, c~   1.91e-1
##  9 Employés                                Autre                       1.12e-3
## 10 Employés                                Cadres et professions in~   8.84e-7
## # ... with 11 more rows
```

6.1 Producing an Odd-Ratios table (Logistic Regression)

Use a logistic regression model (glm in R, binary family) to test whether completion, in the course, is linked to the user characteristics that you studied earlier. Make an odd-ratio table. Signal the odd-ratios that are significant in terms of p-value (with stars). Interpret the results by providing at least two alternative explanations for how socioeconomic status, or human development index, is linked to completion.

```
# if event is rare, odds ratio and relative risk are almost the same
mod_reg1 = glm(Exam.bin ~ Gender + HDI,data=full_df,family='binomial')
aov(mod_reg1)
```

```
## Call:
##    aov(formula = mod_reg1)
##
## Terms:
##                     Gender       HDI Residuals
```

```
## Sum of Squares     0.9824    3.9338 1425.2427
## Deg. of Freedom         1         2      9829
##
## Residual standard error: 0.3807937
## Estimated effects may be unbalanced
## 18637 observations deleted due to missingness
```

```
A=exp(coef(mod_reg1))    # Odd ratios
exp(confint(mod_reg1))  # calculate confidence intervals
```

```
##                     2.5 %    97.5 %
## (Intercept)     0.1230143 0.1698055
## Genderune femme 0.9947384 1.2402111
## HDII            0.9299391 1.5506161
## HDITH           1.2927521 1.8219648
```

```
summary(mod_reg1)
```

```
##
## Call:
## glm(formula = Exam.bin ~ Gender + HDI, family = "binomial", data = full_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6639  -0.6331  -0.6331  -0.5204   2.0330
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.93113    0.08218 -23.498  < 2e-16 ***
## Genderune femme  0.10537    0.05626   1.873   0.0611 .
## HDII             0.18449    0.13032   1.416   0.1569
## HDITH            0.42562    0.08749   4.865 1.15e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9169.8  on 9832  degrees of freedom
## Residual deviance: 9134.2  on 9829  degrees of freedom
##   (18637 observations deleted due to missingness)
## AIC: 9142.2
##
## Number of Fisher Scoring iterations: 4
```

```
anova(mod_reg1)
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Exam.bin
##
```

```
## Terms added sequentially (first to last)
##
##
##        Df Deviance Resid. Df Resid. Dev
## NULL                    9832     9169.8
## Gender  1    6.683      9831     9163.1
## HDI     2   28.860      9829     9134.2
```

```
#OR table with confidenc intervals
exp(cbind(OR = coef(mod_reg1), confint.default(mod_reg1)))
```

```
##                         OR     2.5 %    97.5 %
## (Intercept)     0.1449847 0.1234150 0.1703242
## Genderune femme 1.1111225 0.9951205 1.2406469
## HDII            1.2025995 0.9315220 1.5525618
## HDITH           1.5305356 1.2893605 1.8168225
```

```
#pseudo-R2 , McFadden
pR2(mod_reg1)
```

```
## fitting null model for pseudo-r2
```

```
##           llh       llhNull            G2      McFadden          r2ML
## -4.567117e+03 -4.584888e+03  3.554273e+01  3.876074e-03  3.608113e-03
##          r2CU
##  5.949548e-03
```

```
#optional
# if we want to change the reference
mod_reg2 = glm(Exam.bin ~ HDI +relevel(as.factor(Gender), ref = "une femme"),data=full_df,family='binom
summary(mod_reg2)
```

```
##
## Call:
## glm(formula = Exam.bin ~ HDI + relevel(as.factor(Gender), ref = "une femme"),
##     family = "binomial", data = full_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6639  -0.6331  -0.6331  -0.5204   2.0330
##
## Coefficients:
##                                                      Estimate Std. Error
## (Intercept)                                          -1.82576    0.09434
## HDII                                                  0.18449    0.13032
## HDITH                                                 0.42562    0.08749
## relevel(as.factor(Gender), ref = "une femme")un homme -0.10537    0.05626
##                                                      z value Pr(>|z|)
## (Intercept)                                          -19.354  < 2e-16 ***
## HDII                                                   1.416   0.1569
## HDITH                                                  4.865 1.15e-06 ***
```

```
## relevel(as.factor(Gender), ref = "une femme")un homme  -1.873   0.0611 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9169.8  on 9832  degrees of freedom
## Residual deviance: 9134.2  on 9829  degrees of freedom
##   (18637 observations deleted due to missingness)
## AIC: 9142.2
##
## Number of Fisher Scoring iterations: 4
```

```
#Model 3 , completion ~ Gender + CSP + HDI
mod_reg3 = glm(Exam.bin ~ Gender + HDI + CSP.fin + Age.group,data=full_df,family='binomial')

# ORS + confidence intervals
C = exp(cbind(OR = coef(mod_reg3), confint.default(mod_reg3)))


C
```

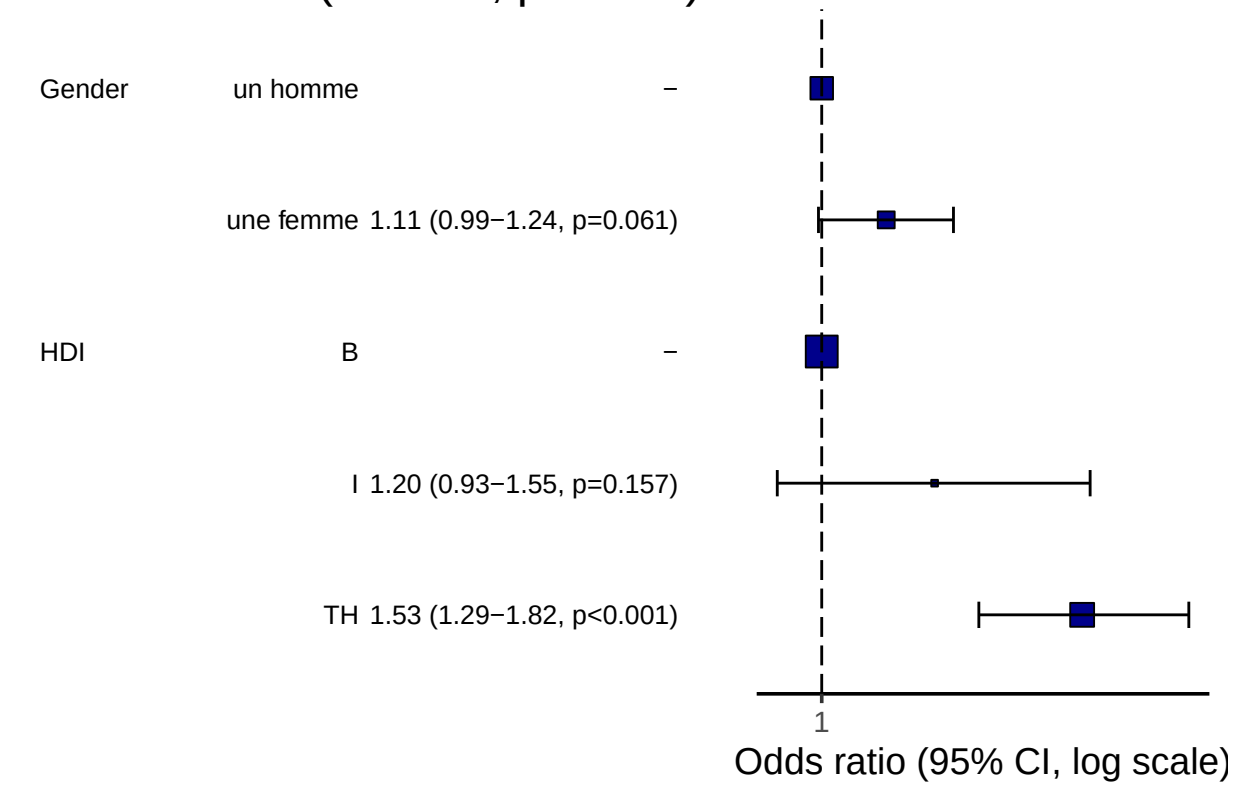
```
##                                                       OR      2.5 %    97.5 %
## (Intercept)                                    0.6995348 0.46624832 1.0495457
## Genderune femme                                1.1308938 1.00973139 1.2665951
## HDII                                           1.1549206 0.88173803 1.5127413
## HDITH                                          1.3792447 1.14276223 1.6646647
## CSP.finArtisans, commerçants, chefs d'entreprise 0.1389167 0.09460615 0.2039809
## CSP.finAutre                                   0.2498646 0.17110070 0.3648864
## CSP.finCadres et professions intellectuelles   0.2728915 0.20001029 0.3723296
## CSP.finEmployés                                0.2362347 0.16548846 0.3372249
## CSP.finEn recherche d'emploi                   0.2990798 0.21378417 0.4184068
## CSP.finEtudiants                               0.2174479 0.15237029 0.3103203
## Age.group(30,50]                               0.8026309 0.65504876 0.9834633
## Age.group(50,80]                               1.0390288 0.82213113 1.3131492
```

```
summary(mod_reg3)
```

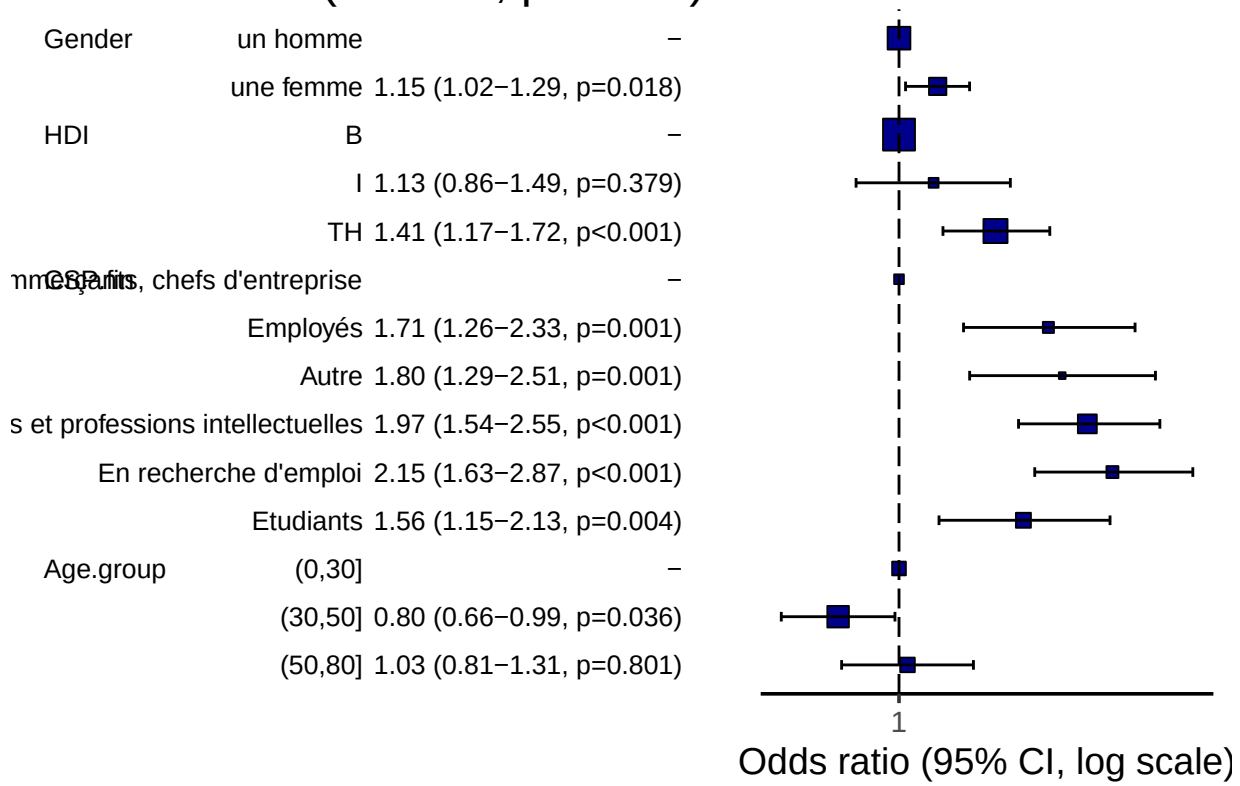
```
##
## Call:
## glm(formula = Exam.bin ~ Gender + HDI + CSP.fin + Age.group,
##     family = "binomial", data = full_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2311  -0.6525  -0.6172  -0.4991   2.2918
##
## Coefficients:
##                                                Estimate Std. Error z value
## (Intercept)                                    -0.35734    0.20699  -1.726
## Genderune femme                                 0.12301    0.05782   2.127
## HDII                                            0.14403    0.13770   1.046
## HDITH                                           0.32154    0.09596   3.351
```

```
## CSP.finArtisans, commerçants, chefs d'entreprise -1.97388    0.19600 -10.071
## CSP.finAutre                                     -1.38684    0.19320  -7.178
## CSP.finCadres et professions intellectuelles     -1.29868    0.15853  -8.192
## CSP.finEmployés                                  -1.44293    0.18160  -7.946
## CSP.finEn recherche d'emploi                     -1.20704    0.17130  -7.046
## CSP.finEtudiants                                 -1.52580    0.18146  -8.409
## Age.group(30,50]                                 -0.21986    0.10367  -2.121
## Age.group(50,80]                                  0.03829    0.11946   0.320
##                                                  Pr(>|z|)
## (Intercept)                                      0.084285 .
## Genderune femme                                  0.033382 *
## HDII                                             0.295578
## HDITH                                            0.000806 ***
## CSP.finArtisans, commerçants, chefs d'entreprise  < 2e-16 ***
## CSP.finAutre                                     7.06e-13 ***
## CSP.finCadres et professions intellectuelles     2.56e-16 ***
## CSP.finEmployés                                  1.93e-15 ***
## CSP.finEn recherche d'emploi                     1.84e-12 ***
## CSP.finEtudiants                                  < 2e-16 ***
## Age.group(30,50]                                 0.033937 *
## Age.group(50,80]                                 0.748597
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8814.1  on 9385  degrees of freedom
## Residual deviance: 8656.9  on 9374  degrees of freedom
##   (19084 observations deleted due to missingness)
## AIC: 8680.9
##
## Number of Fisher Scoring iterations: 4
```

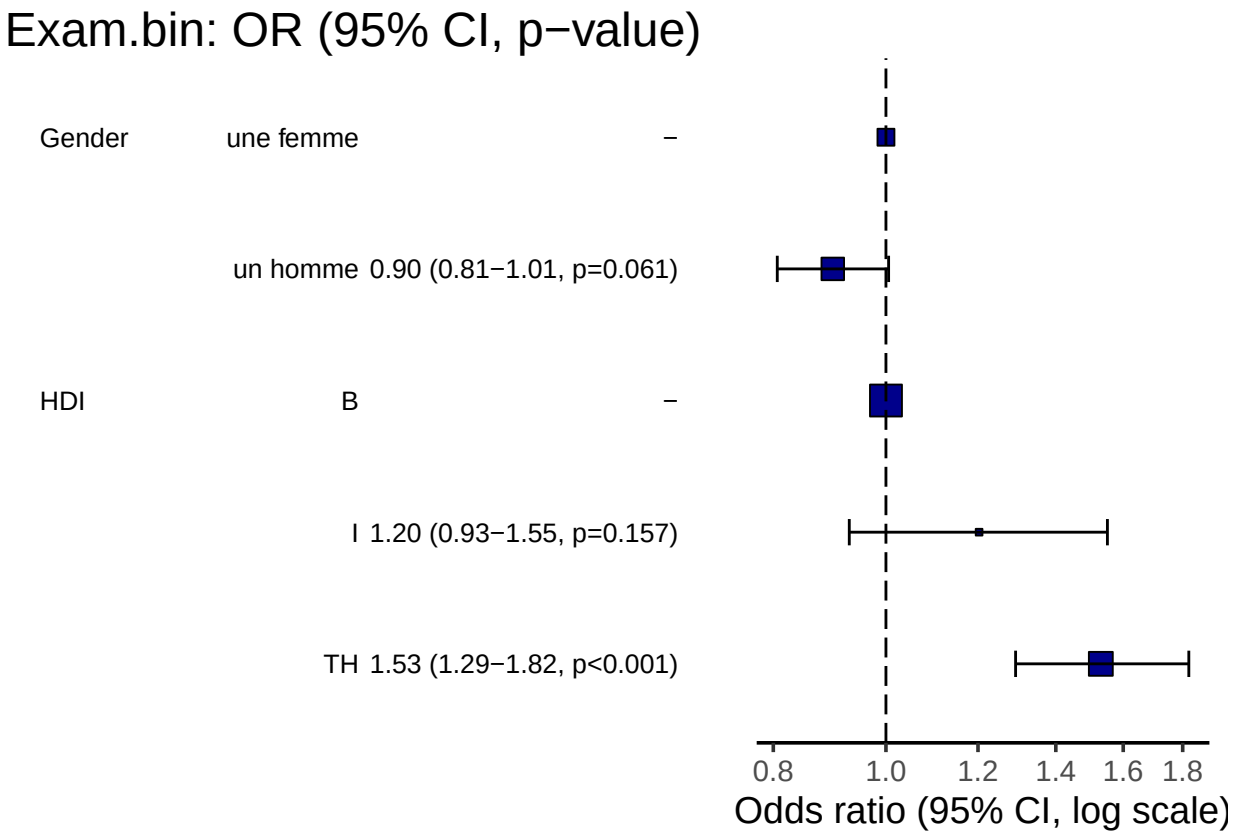
```
#Odds-ratio plot also known as forest plot
full_df %>%   or_plot('Exam.bin', c('Gender','HDI'),
          breaks = c(0.5, 1, 5, 10, 20, 30),
          table_text_size = 3.5)
```




```
#Odds-ratio plot but with csp
# full_df %>%   or_plot('Exam.bin', c('Gender','HDI','CSP.fin'),
#           breaks = c(0.5, 1, 5, 10),
#           table_text_size = 3)

# lev.init=levels(full_df$CSP.fin)
# full_df$CSP.fin2=full_df$CSP.fin
# lev.init
# levels(full_df$CSP.fin2)=c("Autre","Cadres et professions intellectuelles","Employés","En recherche d

#Odds-ratio plot with Gender, HDi and Age group
full_df %>% mutate(CSP.fin=factor(CSP.fin,levels=c("Artisans, commerçants, chefs d'entreprise",'Employé
          breaks = c(0.5, 1, 5, 10),
          table_text_size = 3.5)
```

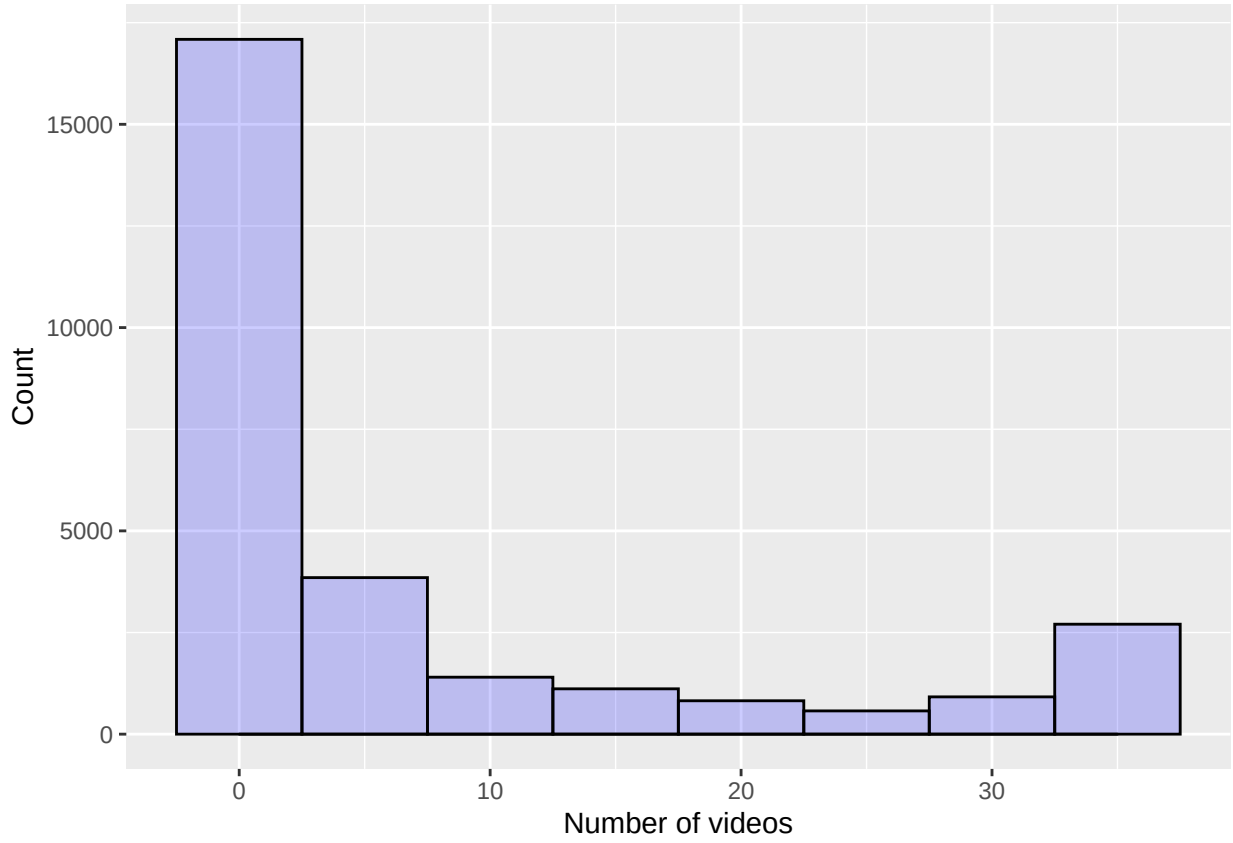



```
#we can see for csp, the variables aren't statistically significant, so we can take gender and HDi only

#Forest OR plot with female as reference instead of male

full_df %>% mutate(Gender=factor(Gender,levels=c('une femme','un homme'))) %>%
   or_plot('Exam.bin', c('Gender','HDI'), table_text_size = 3.5)
```



6.2 Poisson model for count data

```
qplot(full_df$n.videos,
      geom="histogram",
      binwidth = 5,
      xlab = "Number of videos",
      ylab="Count",
      fill=I("blue"),
      col=I("black"),
      alpha=I(.2),
    ) +   geom_density()
```




```
#poisson model <=> family="poisson"
mod_reg4 = glm(n.videos ~ Gender+HDI,data=full_df,family=poisson(link="log"))
#mod_reg4 = glm(n.videos ~ Gender+HDI,data=full_df,family=quasipoisson)

summary(mod_reg4)
```

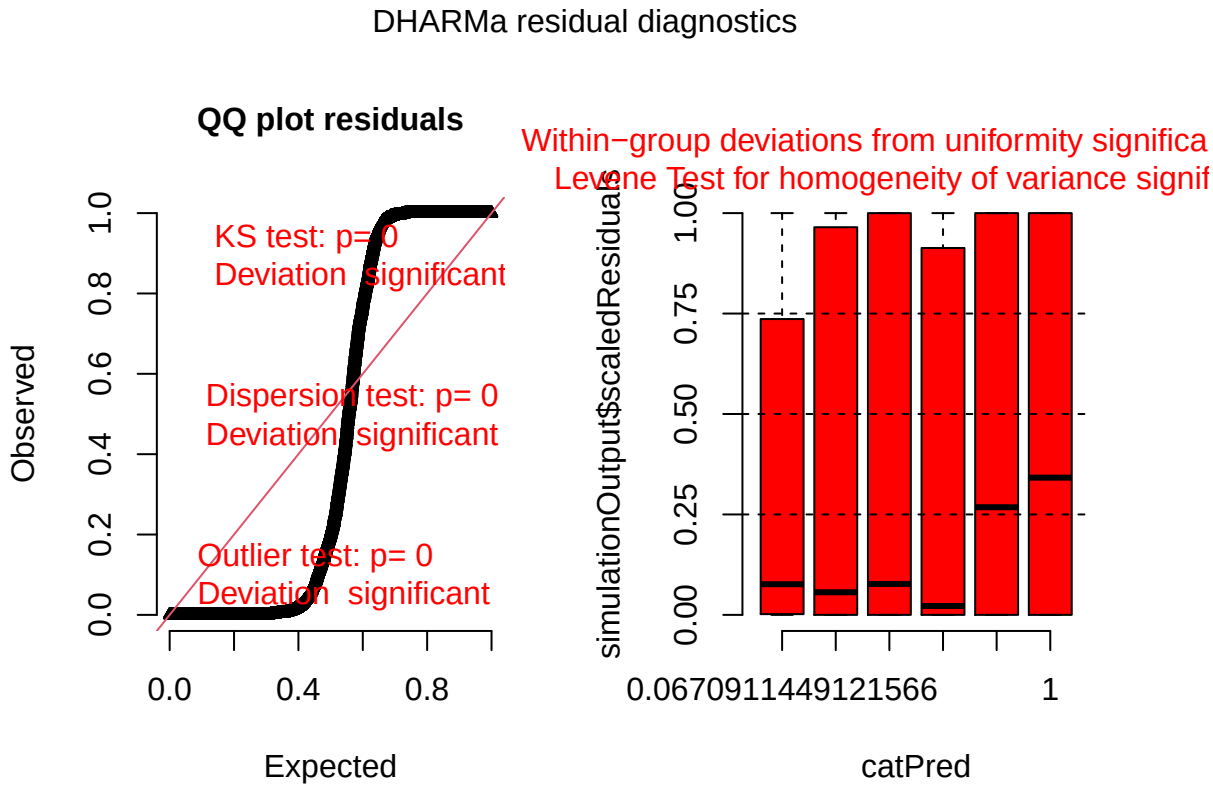
```
##
## Call:
## glm(formula = n.videos ~ Gender + HDI, family = poisson(link = "log"),
##     data = full_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.9404  -3.5607  -0.8802   3.2575   6.8264
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     2.130090   0.009452 225.368   <2e-16 ***
## Genderune femme 0.004977   0.005372   0.926    0.354
## HDII            0.402182   0.013949  28.833   <2e-16 ***
## HDITH           0.735331   0.009858  74.596   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
```

```
##
##     Null deviance: 127057  on 9836  degrees of freedom
## Residual deviance: 119468  on 9833  degrees of freedom
##   (18633 observations deleted due to missingness)
## AIC: 157580
##
## Number of Fisher Scoring iterations: 5
```

```
#latex table
#print(xtable(summary(mod4)))
```

Residual analysis of poisson model * Check homoscedasticity of the residuals i.e residual analysis ==> homoscedasticity assumes the residuals are approximately equal for all predicted dependent variable scores , assumes equal variance

```
#check for homoscedasticity
plot(simulateResiduals(mod_reg4))
```



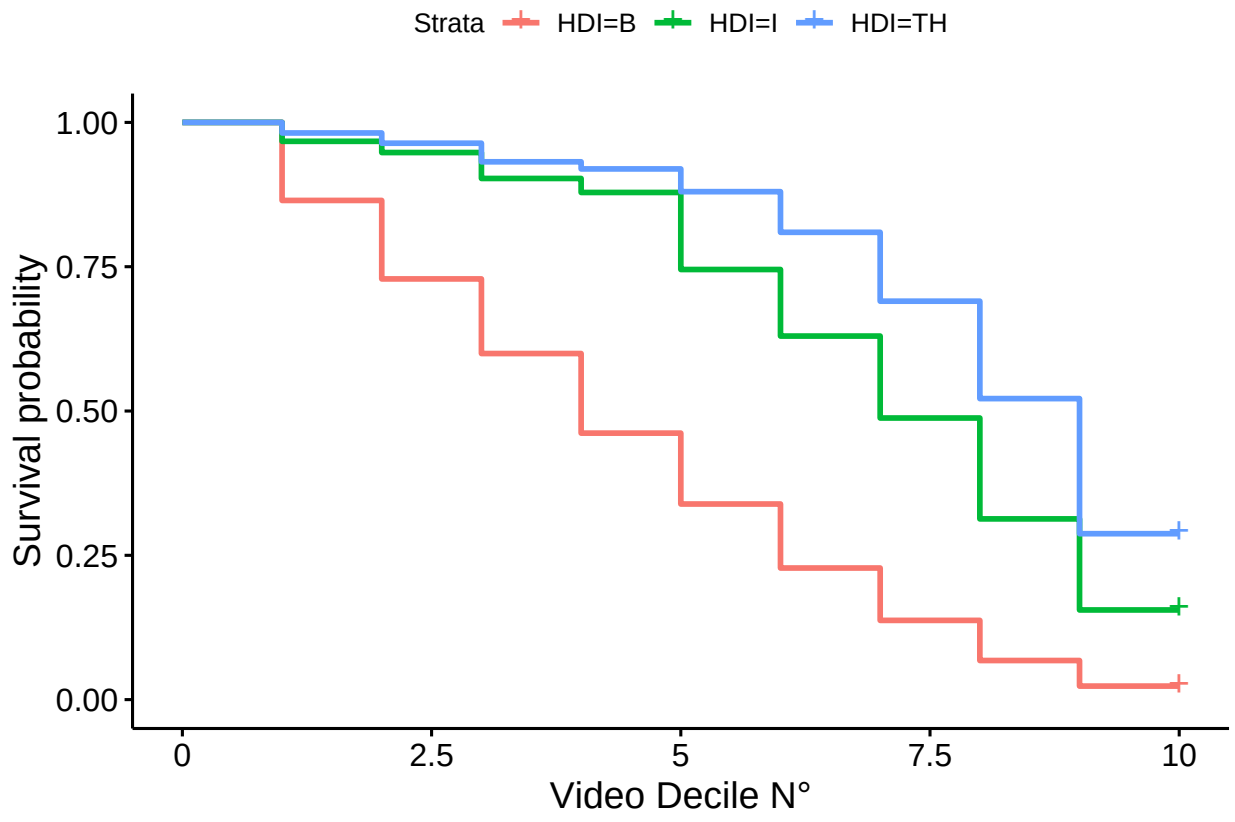
7 Survival Analysis

- You must reason in terms of proportion of the available videos that the learner viewed. Prepare the data so that they are fit for a survival analysis.

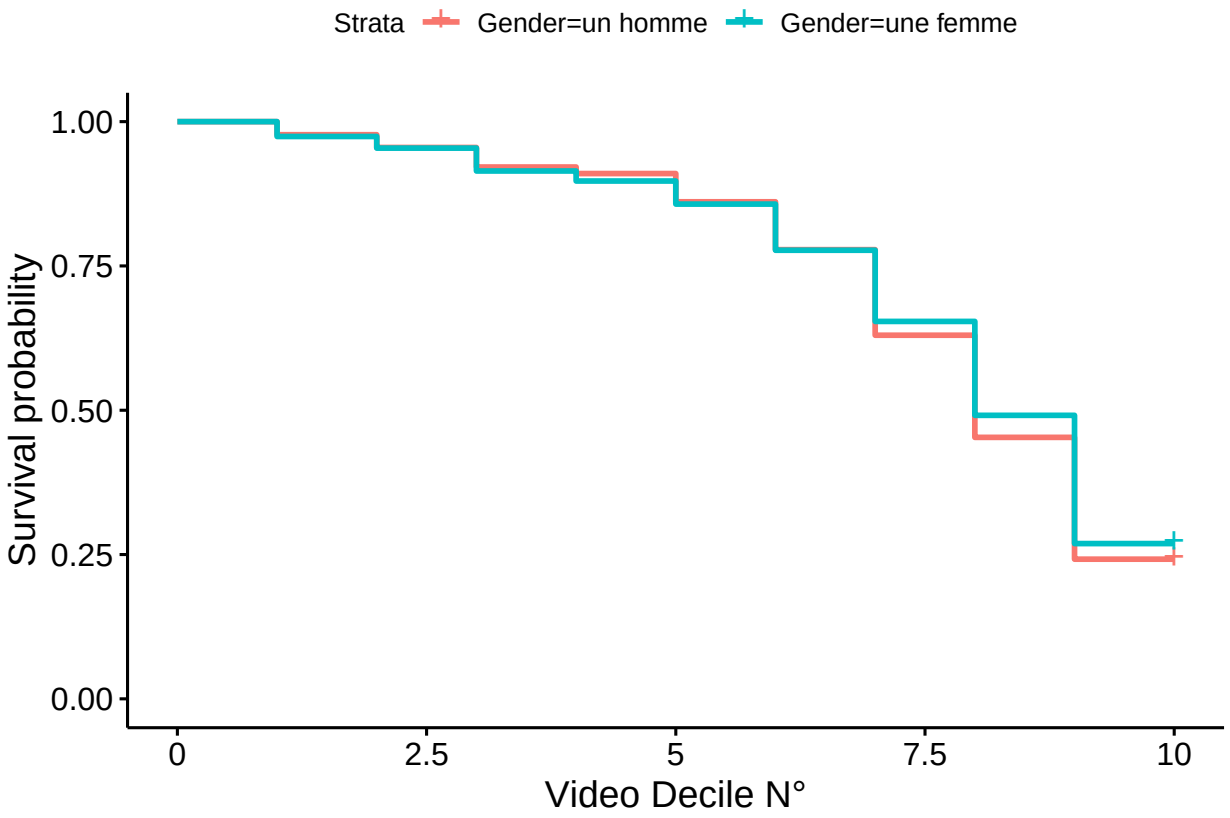
```
#check deciles for number of videos
n.videos_dec = quantile(full_df$n.videos, probs = seq(.1, .9, by = .1))
#add deciles (new column ) for the number of videos
#using mutate method
full_df<-full_df %>%
  mutate(n.videos.decile = ntile(n.videos, 10))

# add status based on deciles
full_df$status.vid=rep(NA, nrow(full_df))
for (i in 1:nrow(full_df)) {
  if (full_df$n.videos.decile[i]<10) {full_df$status.vid[i]=1}
  if (full_df$n.videos.decile[i]==10) {full_df$status.vid[i]=0}
}
```

- Compare video consumption behavior between auditing and disengaging learners, but this time with a survival analysis (and not the linear model like you did earlier).
- plot the survival curve. Where do you see the most significative drop in terms of video consumption ?

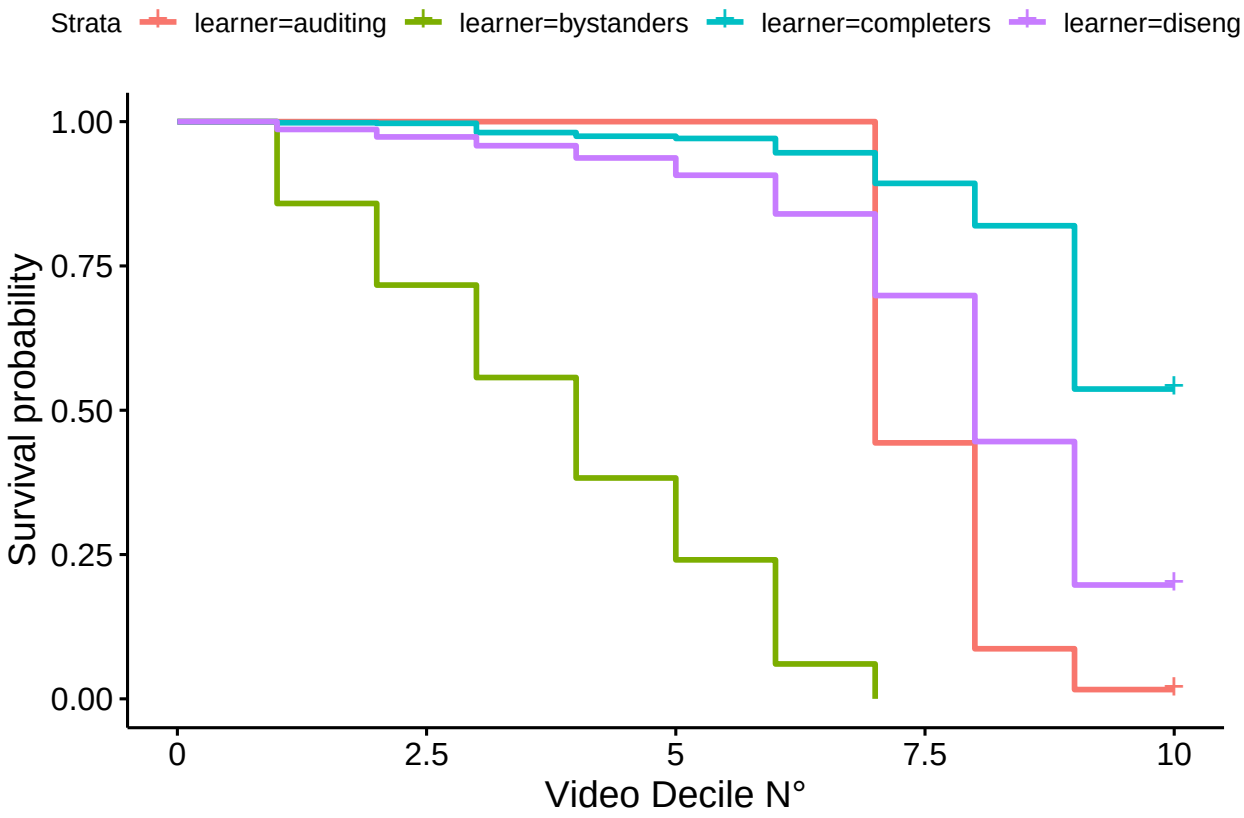
```
#number of videos survival analysis based on HDI
surv_mod1 <- survfit(Surv(n.videos.decile, status.vid) ~ HDI , data=full_df)
ggsurvplot(surv_mod1, data = full_df ,xlab="Video Decile N°")
```




```
#number of videos survival analysis based on Gender
surv_mod2 <- survfit(Surv(n.videos.decile, status.vid) ~ Gender , data=full_df)
ggsurvplot(surv_mod2, data = full_df, xlab="Video Decile N°")
```




```
#number of videos survival analysis based on type of learners(completers, disengaging etc)
surv_mod3 <- survfit(Surv(n.videos.decile, status.vid) ~ learner , data=full_df)
ggsurvplot(surv_mod3, data = full_df,xlab="Video Decile N°")
```



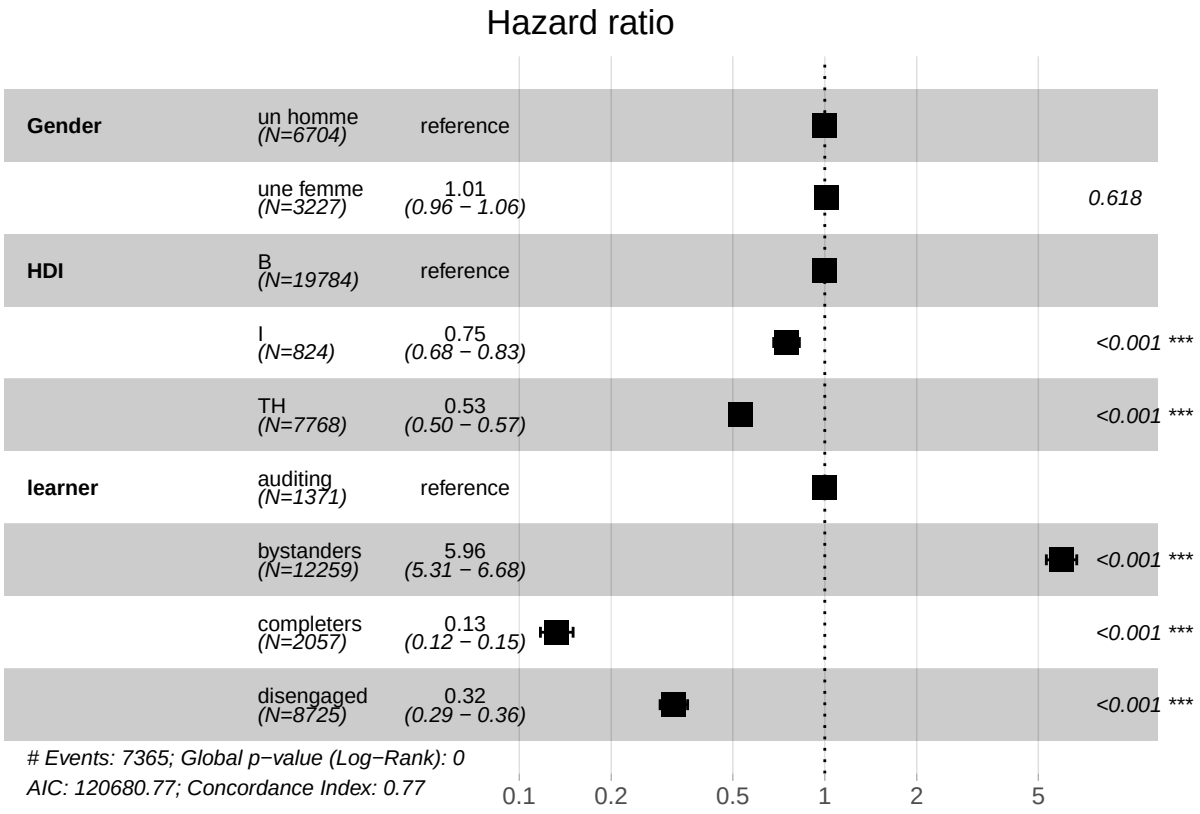
Compute the hazard ratios

```
#Calculate hazard ratios using coxph
mod_cox <- coxph(formula = Surv(n.videos.decile, status.vid) ~ Gender+HDI+learner, data = full_df)

mod_cox
```

```
## Call:
## coxph(formula = Surv(n.videos.decile, status.vid) ~ Gender +
##     HDI + learner, data = full_df)
##
##                         coef exp(coef) se(coef)       z        p
## Genderune femme      0.01266   1.01274  0.02538   0.499    0.618
## HDII                -0.28829   0.74954  0.04934  -5.843 5.13e-09
## HDITH               -0.63348   0.53074  0.03247 -19.509  < 2e-16
## learnerbystanders    1.78456   5.95693  0.05845  30.533  < 2e-16
## learnercompleters   -2.02005   0.13265  0.06284 -32.147  < 2e-16
## learnerdisengaged   -1.13751   0.32062  0.05288 -21.511  < 2e-16
##
## Likelihood ratio test=6846  on 6 df, p=< 2.2e-16
## n= 9833, number of events= 7365
##    (18637 observations deleted due to missingness)
```

References are : Male(for gender), Low(For HDI), auditing (for types of learners)

```
#hazard ratios in forest plot
ggforest(mod_cox,data=full_df)
```



Brief interpretation : people from rich countries tend to disengage much slower from the course than people from poor country(H=0.45,ref=poor,p-value<0.001)