Université Cergy Paris

Diplôme universitaire : Data Analyst

UE 3 : Introduction to statistics

January 6, 2023

Haury Fabien

Contents

Part 1

Description of the dataset

The dataset relates to analytical learnings from the different iterations of a MOOC. In particular, we will focus on learner engagement, video viewing and numbers of quizzes completed, their gender and the Human Development Index (HDI).

Variable		Iteration			Totaux des lignes
		1	2	3	
Statut	Auditing learners	152 (2.64%)	106 (3.75%)	107 (3.55%)	365 (3.15%)
	Bystanders	3139 (54.46%)	1720 (60.91%)	1980 (65.69%)	6839 (58.95%)
	Completers	20 (0.34%)	878 (31.09%)	843 (27.97%)	1741 (15%)
	Disengaging learners	2453 (42.56%)	120 (4.25%)	84 (2.79%)	2657 (22.90%)
	NA	3222	1350	1587	6159
Totaux des colonnes		8986 (49.68%)	4174 (24.34%)	4601 (25.98%)	17761 (100%)

Table 1: Learner status contingency table per iteration with percentage for each iteration

Table 1 represents the proportions of the status of the learners for each iteration with the number of individuals.
The Iteration variable represents the number of times the MOOC has been held. The Status variable corresponds to the attendance of the participant, described as follows:

- Auditing learners : if no quiz has been taken and no homework has been returned, but has viewed more than six videos.
- Bystanders : : if no quiz has been completed and no homework has been submitted, but has viewed less than six videos.
- Completers : : if they passed the exam.
- Disengaging learners : if a quiz has been taken or an assignment has been returned, but the certificate has not been obtained, and the exam has not been taken.

We can see that more than half of the learners are Bystanders, with 58.95%. The first iteration has 8,986 learners, almost half of the total learners (49.68%). The second half is split evenly with 4,171 (24.34%) and 4,601 (25.98%) learners for the second and third iteration respectively. The number of Completers increases between the first and the other two iterations, going from 20 learners for the first iteration to 878 for the second iteration and 843 for the third. Conversely, the number of Disengaging learners decreases over the iterations, going from 2,453 for the first iteration to 120 for the second iteration and finally to 84 for the third.

Part 2

Chi square et mosaic plot

A chi² test (χ^2) of independence makes it possible to verify the absence of a statistical link between two variables X and Y, here Gender and HDI. The two are said to be independent when there is no statistical link between them, in other words, knowledge of X does not in any way allow us to decide on Y.
The null hypothesis H_0 of this test is as follows: the two variables HDI and Gender are independent.
The alternative hypothesis H_1 of this test is the following: the two variables HDI and Gender are not independent.

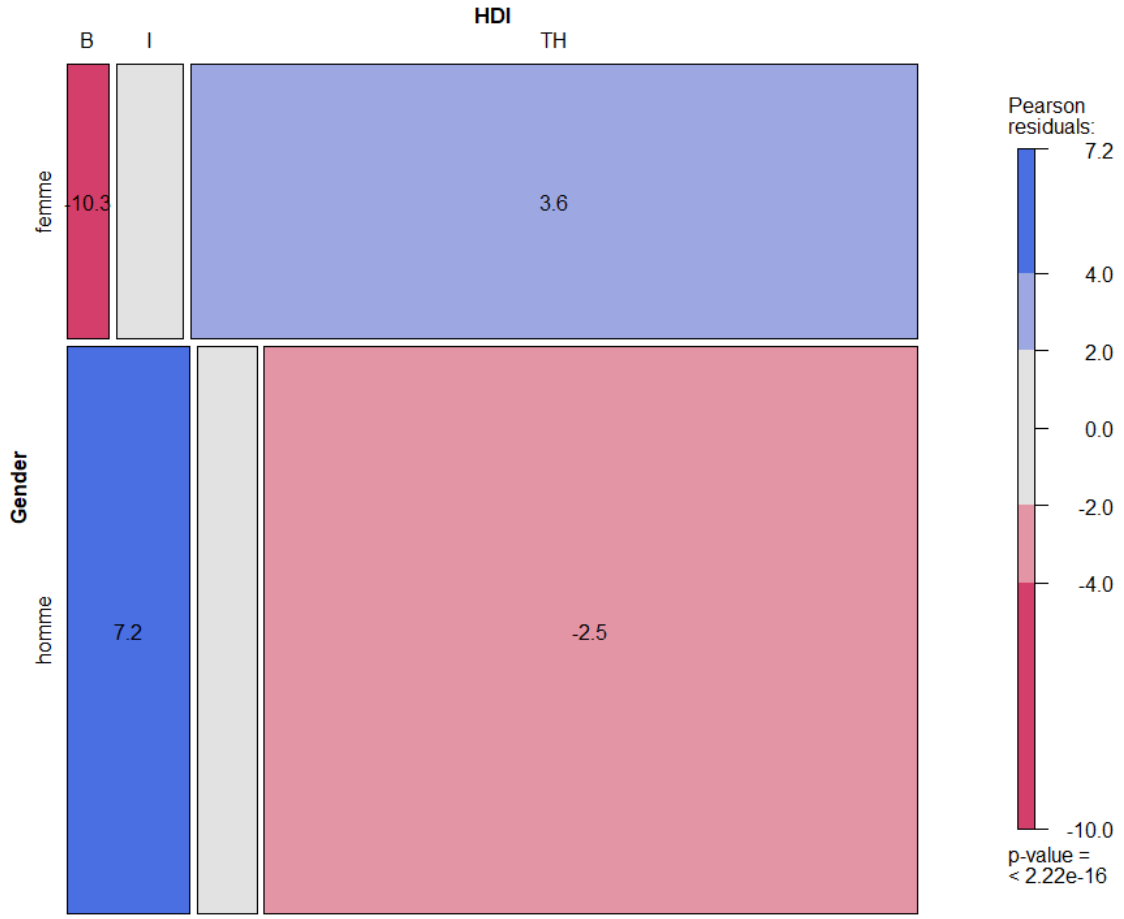

Figure 1: – Mosaic plot of the values obtained from the χ^2 of Gender vs. HDI

A mosaic chart is a square subdivided into rectangular tiles whose area represents the conditional relative frequency of a cell in the contingency table. Each tile is colored to show the deviation from the expected (residual) frequency of a χ^2 test for the level of the residual for that cell/level combination. The legend is presented to the right of the graph. Specifically, blue means that there are more observations in that cell than would be expected with the null model (independence). Red means there are fewer observations than would be expected. This can be read as an indication of which cells contribute to the significance of the chi-square test result.

Figure 1 represents a mosaic plot of the residual values obtained after an χ^2 of Gender in relation to HDI. It can be seen that the gender distribution is as follows, 67.3% are men and 23.7% women. Similarly, the distribution of the HDI is as follows, B: 14.3% are women and 85.7% men, I: 35% are women and 65% are men, TH: 35.1% are women and 64.9%. The result of the χ^2 is equal to 179.05 with a degree of freedom of two, which is calculated as follows: (row - 1)*(col - 1). The p-value is $2.2e^{-16}$ which is below the 5% threshold. Accordingly, we reject the null hypothesis H_0 and admit that the alternative hypothesis H_1 is true. There is therefore a link between gender and the HDI, and Cramer's V test allows us to determine the intensity of the relationship between these two variables. This one with a score of 0.14 indicates a weak intensity.

We can see that for HDI B women are very strongly under-represented and men strongly over-represented. The HDI B representing poor/developing countries, we can imagine several reasons for this, a restriction on education and therefore on MOOCs for women, restricted internet access by men, or others. Conversely, the HDI TH shows a slight over-representation of women and a slight under-representation of men. The TH HDI representing the richest/developed countries, we can imagine several reasons for this, access to education is identical for women and men, or others.

Part 3

Linear model, non-parametric tests

Non-parametric tests

We perform a non-parametric Wilcoxon test on the variables of the total number of videos viewed and gender (male and female), with the following assumptions:
H_0 : the average of the two groups are equal on the number of videos viewed.
H_1 : the average of the two groups are not equal on the number of videos viewed.
T p-value is 0.000481 and therefore below the threshold of 5%. The test is therefore statistically significant, and we reject the null hypothesis H_0 and accept the alternative hypothesis H_1. Women watch more videos than men.

Spearman correlation test

Spearman's correlation is the nonparametric equivalent of Pearson's correlation test. it measures the relationship between two variables. The correlation coefficients between -1 and +1, 0 reflecting a zero relationship between the two variables, a negative value (negative correlation) mean that when one of the variables increases, the other decreases ; while a positive (positive correlation) indicates that the variables vary together in the same direction. We must therefore carry out a hypothesis test.
H_0: no correlation between the two variables: $\rho = 0$
H_1: correlation between the two variables: $\rho \neq 0$
The p-value is $2.2e^{-16}$ which is below the 5% threshold. The test is therefore statistically significant, and we reject the null hypothesis H_0 and accept the alternative hypothesis H_1. The ρ coefficient is equal to à 0.80, which shows a strong positive correlation between the two variables.

Figure 2 below represents a scatter plot of the number of quizzes performed compared to the number of videos viewed with a regression line. We can see that the more videos a learner has watched, the more quizzes they will achieve afterward. The regression line is an affine function with the following equation $y = 0.15x + 0.53$. We also see that the data are not normally distributed, they seem ordered by rank.

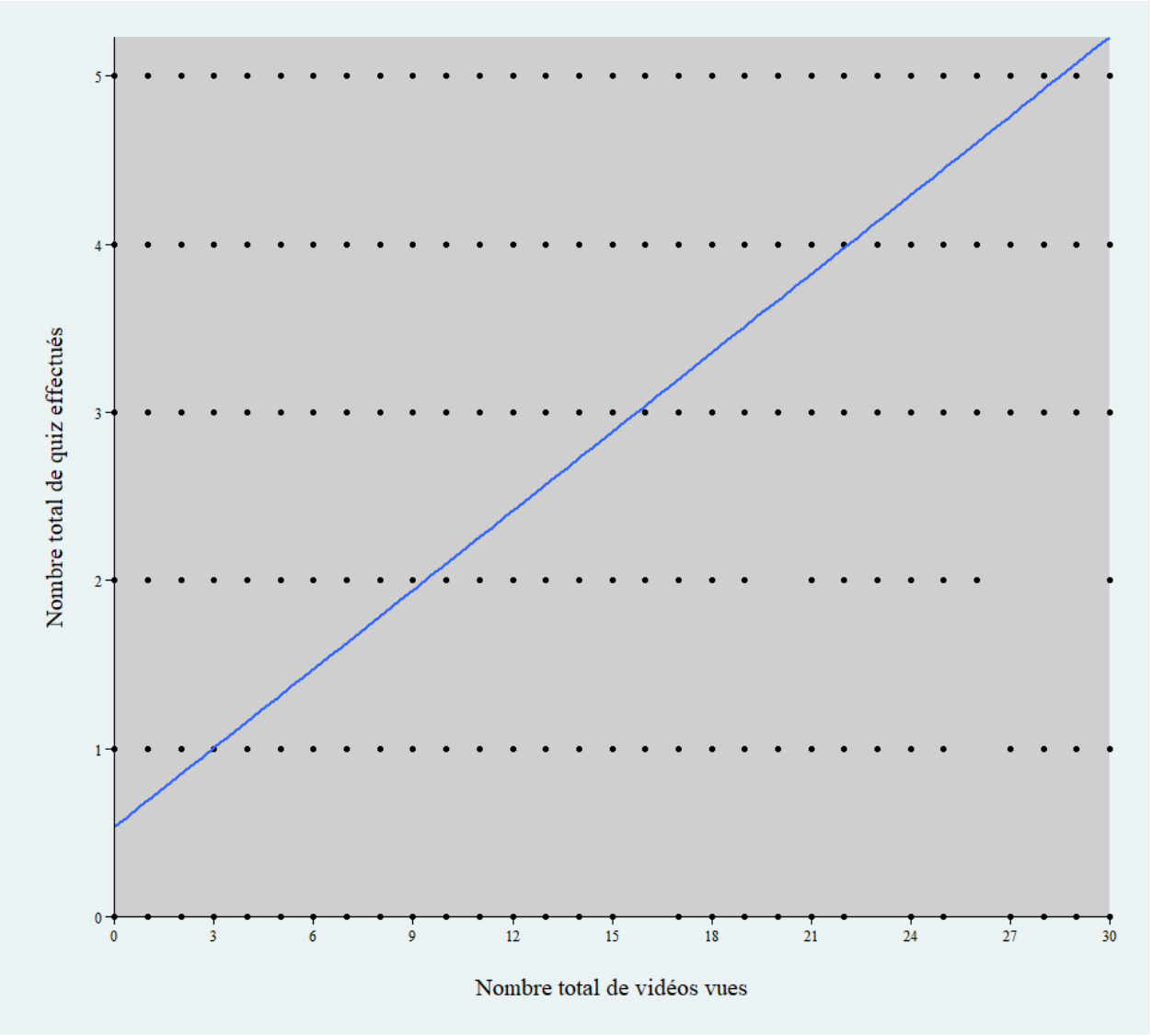

Figure 2: Scatter plot of the number of quizzes performed compared to the number of videos viewed with regression line

ANOVA without interaction

The hypotheses are as follows:
H_0: all gender groups have the same average number of videos watched.
H_1 : not all genre groups have the same average number of videos watched.
H_0: all HDI groups have the same average number of videos watched.
H_1 : not all HDI groups have the same average number of videos watched.

	Df	Sum Sq	Mean Sq	F value	P value
Genre	1	1867.14	1867.14	14.28	$< 1.5e^{-4}$
IDH	2	74433.66	37216.83	284.63	$< 2.2e^{-16}$
Résidus	8947	1169851.64	130.75		

Table 2: ANOVA table of the number of videos viewed in relation to gender and HDI without interaction

Table 2 represents an ANOVA table of the number of videos viewed in relation to Gender and HDI with no interaction between the last two. We found a statistically significant difference in the number of videos watched by the Gender variable ($F(1, 8947) = 14.28$, $p < 0.00$) and by the HDI variable ($F(2, 8947) = 284.67$, $p < 0.00$). We therefore reject the null hypotheses for these two variables and admit their alternative hypotheses.
The degree of freedom (Ddl or Df) designates the maximum number of logically independent values, that is to say values which have the freedom to vary, in the data sample. It is calculated as follows: $D_f = N - 1$ with D_f with Df the degree of freedom and N the size of the sample. For example, the Gender variable contains two samples (male and female) so $N = 2$, so $D_f = 2 - 1 = 1$.

ANOVA with interaction

The hypotheses are as follows:
H_0: all gender groups have the same average number of videos watched.
H_1 : not all genre groups have the same average number of videos watched.
H_0: all HDI groups have the same average number of videos watched.
H_1 : not all HDI groups have the same average number of videos watched.
H_0 : all gender and HDI groups have the same average number of videos watched.
H_1 : not all gender and HDI groups have the same average number of videos watched.

	Df	Sum Sq	Mean Sq	F-value	P-value
Genre	1	1867.14	1867.14	14.28	$< 1.5e^{-4}$
IDH	2	74433.66	37216.83	284.67	$< 2e^{-16}$
Genre : IDH	2	401.95	200.97	1.54	$< 2.24e^{-1}$
Résidus	8945	1169449.70	130.74		

Table 3: ANOVA table of the number of videos viewed in relation to gender and HDI with interaction

Table 3 represents an ANOVA table of the number of videos viewed in relation to gender and HDI, with no interaction between the last two. We found a statistically significant difference in the number of videos watched by the Genre variable ($F(1, 8945) = 14.28$, $p < 1.5e^{-4}$) and by the HDI variable ($F(2, 8945) = 284.67$, $p < 2e^{-16}$). Similarly, we find that there is no statistically significant interaction in the interaction between gender and HDI ($F(2, 8945) = 1.54$, $p < 2.24e^{-1}$). We therefore reject the null hypotheses for these two variables separately and admit their alternative hypotheses and admit the null hypothesis when considering their interaction.

Linear model with interaction

	Estimate	Std. Error	t-value	P-value
Intercept	6.95	0.94	7.39	$1.56e^{-13}$***
Genre homme	-0.58	1.01	-0.58	$5.64e^{-1}$
IDH I	2.90	1.20	2.42	$1.56e^{-2}$*
IDH TH	8.42	0.96	8.70	$< 2e^{-16}$***
Genre homme : IDH I	1.93	1.37	1.40	$1.60e^{-1}$
Genre homme : IDH TH	0.30	1.05	0.29	$7.72e^{-1}$

Table 4: Table of linear model coefficients of total video viewed as a function of HDI gender and the interaction between gender and HDI

Table 4 represents the coefficients of the linear model of total video viewed as a function of HDI gender and the interaction between gender and HDI. Women from a country with HDI B (intercept) watch an average of 6.95 videos (p value = $1.56e^{-13}$), those with HDI I watch an average of 2.90 more videos (p value = $1.56e^{-2}$), and those with HDI TH 8.42 (p value < $2e^{-16}$) more videos. Men from a country with HDI B watch an average of 6.37 more videos (p value = $5.64e^{-1}$), those with HDI I watch an average of 1.93 more videos (p value = $1.60e^{-1}$) and those with HDI TH 0.30 more videos (p value = $7.72e^{-1}$).

Part 4

Logistic regression

4.1 Present odds ratios

	OR	CI : 2.5 %	CI : 97.5 %	P-value
IDH B		ref.		
IDH I	1.12	0.85	1.47	$4.15e^{-1}$
IDH TH	1.37	1.14	1.66	$7.86e^{-4}$***
Genre femme		ref.		
Genre homme	0.89	0.80	1.00	$4.72e^{-2}$*

Table 5: Table of odds ratio of success in the final exam according to HDI and Gender

Table 5 represents the odds ratio table of Exam.bin according to HDI and Gender. The HDI B is used as a reference, so the Odds Ratio of the HDI I and the HDI TH will be calculated according to it. A learner from a country with HDI I is more likely to pass the final exam than a learner from a country with HDI B (OR: 1.12, CI: 0.85-1.47, p-value: $4.15e^{-1}$). A learner from a country with a TH HDI is more likely to pass the final exam than a learner from a country with a B HDI (OR: 1.37, CI: 1.14-1.66, p-value: $7.86e^{-4}$). The Gender female is used as a reference, so the Odds Ratio of Gender male will be calculated according to it. A male learner is less likely to pass the final exam than a female learner (OR: 0.89, CI: 0.80-1.00, p-value: $4.72e^{-2}$). Comparing the results of Table 5 and Table 4, we can see that these results follow the same reasoning. We see that the chances that a person will pass the final exam will increase if he or she is in a country with HDI TH and therefore has the possibility to watch a large number of videos. We also see the fact that men tend to take the final exam less.

Odds Ratio, OR, also called odds ratio, is a non-parametric approach to measuring the association between two variables X^1, X^2 by determining the chance/risk of an event of X^2 occurring given the values of X^1. The Relative Risk, denoted RR, is a non-parametric approach to measuring the association between two variables X^1, X^2 by determining the chance/risk of an event of X^2 occurring in one of the two groups of X^1 compared to the other group of the same variable. The Odds Ratio is the ratio of the odds of an event in the first subgroup to the odds of that same event in the second group, while the Relative Risk is the ratio of the chance of an event occurring in the first subgroup to the chance of that same event occurring in the second subgroup. The Odds Ratio converges to the Relative Risk when the number of events in the Odds Ratio is low.

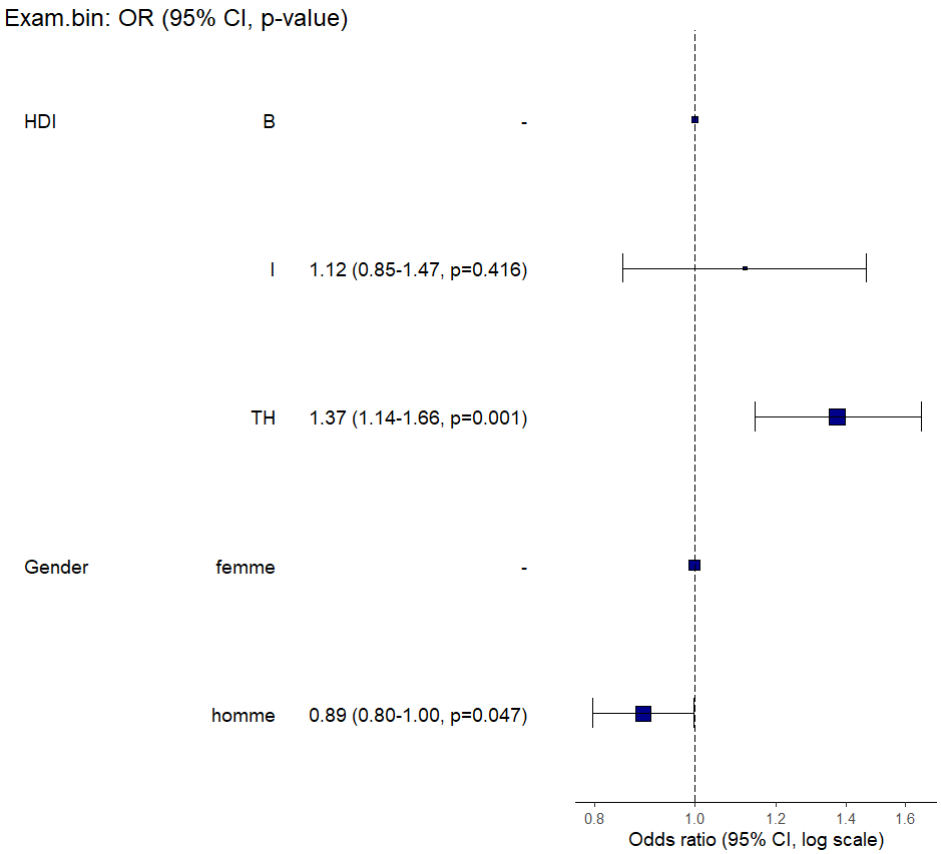

Figure 3: Forest plot of the odds ratio of passing the final exam according to HDI and gender

Figure 3 represents a forest plot of the odds ratio of Exam.bin as a function of HDI and gender. It is the graphical representation of Table 5. The horizontal lines represent the length of the confidence interval, the greater the length, the wider the confidence interval and the less reliable the result. If a point is located to the left of the vertical line, then its Odds Ratio is less than 1. If a point is located on the vertical line, then its Odds Ratio is equal to 1. If a point is located to the right of the vertical line then its Odds Ratio is greater than 1.

4.2 Counting data and Poisson's law

Figure 4 shows the distribution of the number of videos viewed. This is bimodal, with the strongest mode on the left indicating that very few videos are viewed (1-2 videos for a total of about 8,000), the second mode is on the right indicating that a large number of videos are viewed (27–30 videos for a total of about 2,500 videos). These results tell us two trends: either the learner watches few videos, or they will watch a large number of videos.

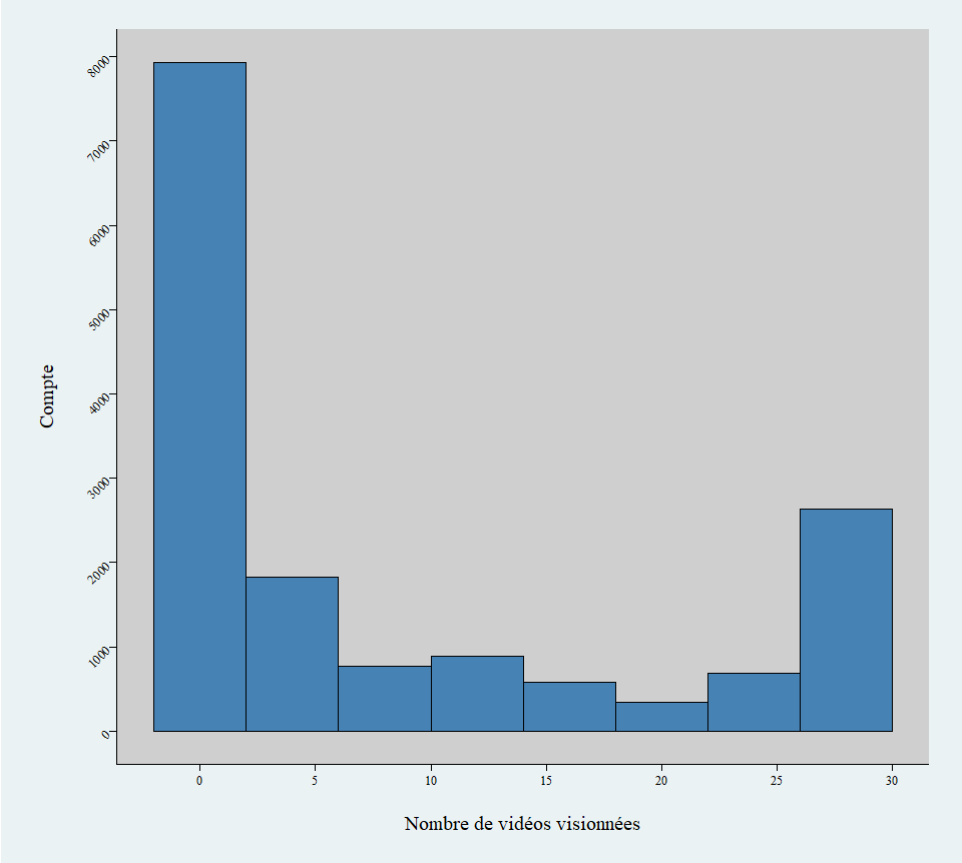

Figure 4: Distribution of the number of videos viewed

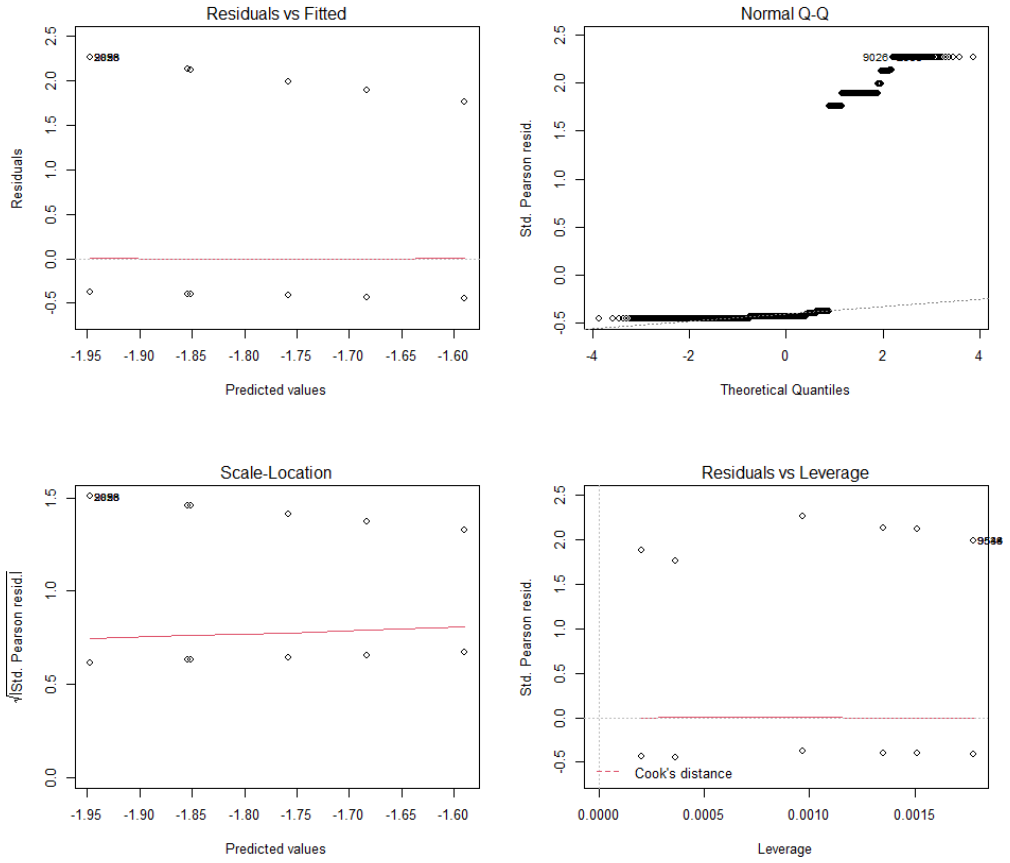

Figure 5: Normality of the distribution of the generalized linear model of final exam status by gender and HDI

Figure 5 is composed of four separate graphs:

- Residuals vs. Fitted : it should be symmetrically distributed with a clustering towards the middle of the graph around zero and no discernible pattern.
- Q-Q plot : a Q-Q plot allows visualizing the distribution of a given distribution with respect to a theoretical model. If the distribution studied is normal, then all the points representing the observations will be on the right-hand side of the theoretical model.
- Scale-Location : used to check the homoscedasticity or the heteroscedasticity.
- Residuals vs. Leverage: used to detect outliers.

Homoscedasticity means that the variance of the errors in the regression is identical. Homoscedasticity is also called homogeneity of variance.
Heteroscedasticity means that the variance of the errors in the regression differs.

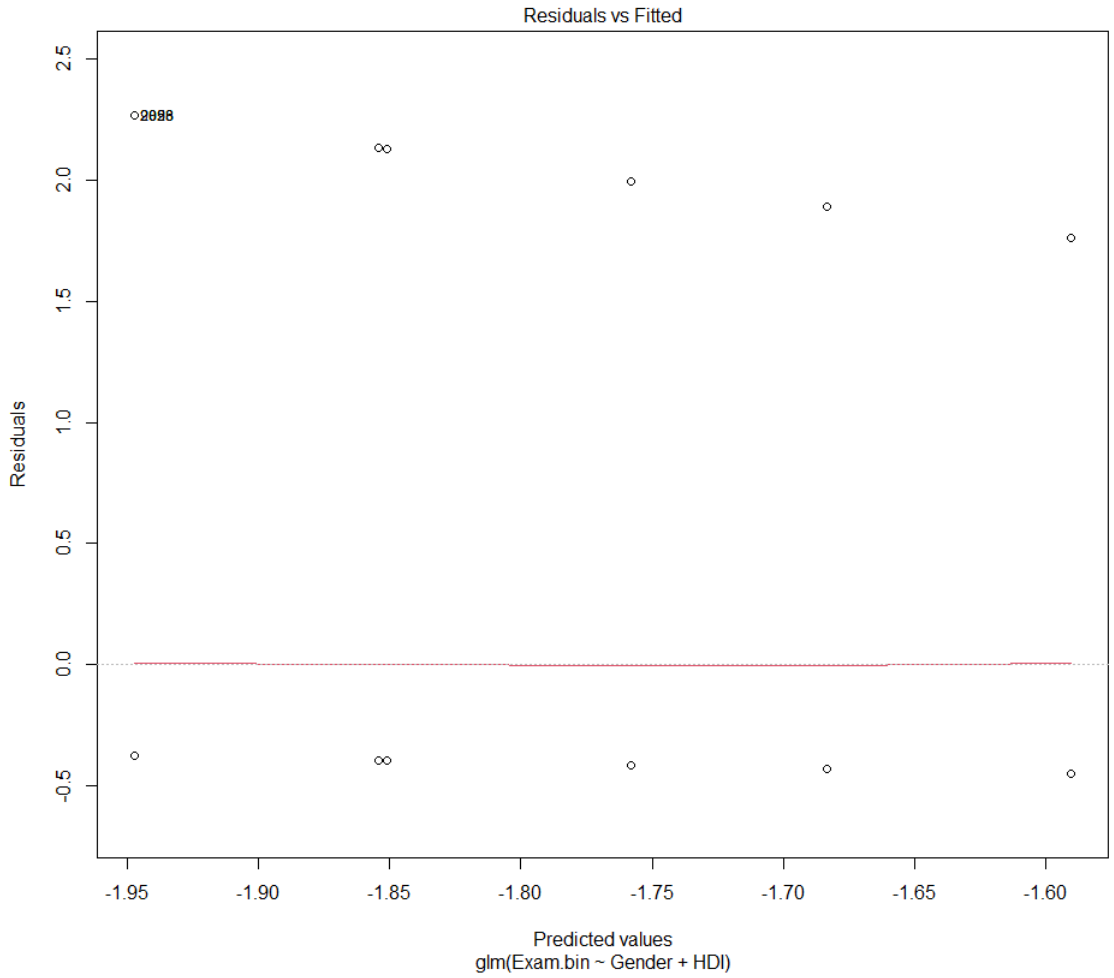

Figure 6: Residuals vs. Fitted from the generalized linear model of final exam status by gender and HDI

Figure 6 represents the residuals as a function of the predicted values for a generalized linear model following a Poisson family. We can see that the distribution of points around the conditional expectation is not symmetrical which indicates a skewed distribution, as shown in Figure 4.

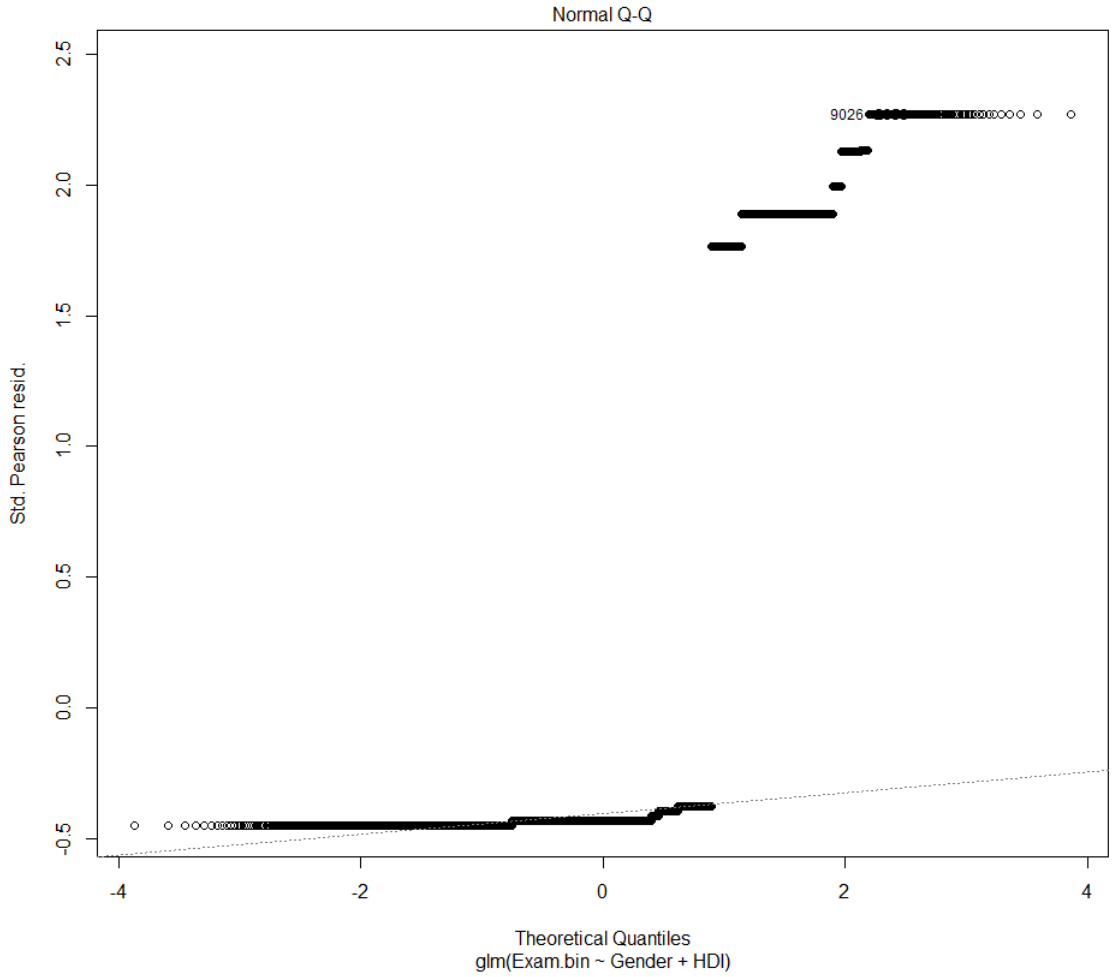

Figure 7: QQ plot of the generalized linear model of final exam status by gender and HDI

7 The Q-Q plot in Figure 7 is divided into three parts. The first part from quantile -4 to quantile -2 shows a spread and a distance of the points from the theoretical line, it means a strongly left biased distribution. The second part from quantile -1 to quantile 1 shows an unbiased distribution because all these points are in the vicinity of the theoretical line. The third part from quantile 1 to quantile 4 shows the presence of a tail and a mode.

List of Figures

List of Tables