Université Cergy Paris

Diplôme universitaire : Data Analyst

UE 3 : Introduction aux statistiques

6 juillet 2022

Haury Fabien

Table des matières

Chapitre 1

Description du jeu de données

Le jeu de données porte sur des learnings analytics issus des différentes itérations d'un mooc. Nous nous concentrerons en particulier sur l'engagement des apprenants, le visionnage de vidéos et les nombres de quiz réalisés, leur genre et l'indice de développement humain (IHD).

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 – Table de contingence du statut des apprenants par itération avec pourcentage pour chaque itération

La Table 1 représente les proportions du statut des apprenants pour chaque itération avec le nombre d'individus.
La variable *Itération* représente le nombre de fois où le mooc a été tenu. La variable *Statut* correspond à l'assiduité du participant, décrite comme suit :

- Auditing learners : si aucun quiz n'a été réalisé et aucun devoir rendu, mais a visualisé plus de six vidéos.
- Bystanders : si aucun quiz n'a été réalisé et aucun devoir rendu, mais a visualisé moins de six vidéos.
- Completers : s'ils ont passé l'examen.
- Disengaging learners : si un quiz a été réalisé ou un devoir rendu, mais le certificat n'a pas été obtenu et l'examen n'a pas été réalisé.

Nous pouvons constater que plus de la moitié des apprenants sont des *Bystanders* avec 58.95%. La première itération comporte 8 986 apprenants, soit presque la moitié du total des apprenants(49.68%). La seconde moitié est répartie de façon équitable avec 4 171 (24.34%) et 4 601 (25.98%) apprenants pour la seconde et troisième itération respectivement. Le nombre de Completers augmente entre la première et les deux autres itérations passant de 20 apprenants pour la première itération à 878 pour la seconde itération et 843 pour la troisième. Inversement, le nombre de Disengaging learners diminue au fil des itérations passant de 2 453 pour la première itération à 120 pour la seconde itération et enfin à 84 pour la troisième.

Chapitre 2

Chi square et mosaic plot

Un test chi² (χ^2) d'indépendance permet de vérifier l'absence de lien statistique entre deux variables X et Y, ici Genre et IDH. Les deux sont dites indépendantes lorsqu'il n'existe aucun lien statistique entre elles, autrement dit, la connaissance de X ne permet en aucune manière de se prononcer sur Y.
L'hypothèse nulle H_0 de ce test est la suivante : les deux variables IDH et Genre sont indépendantes.
L'hypothèse alternative H_1 de ce test est la suivante : les deux variables IDH et Genre ne sont pas indépendantes.

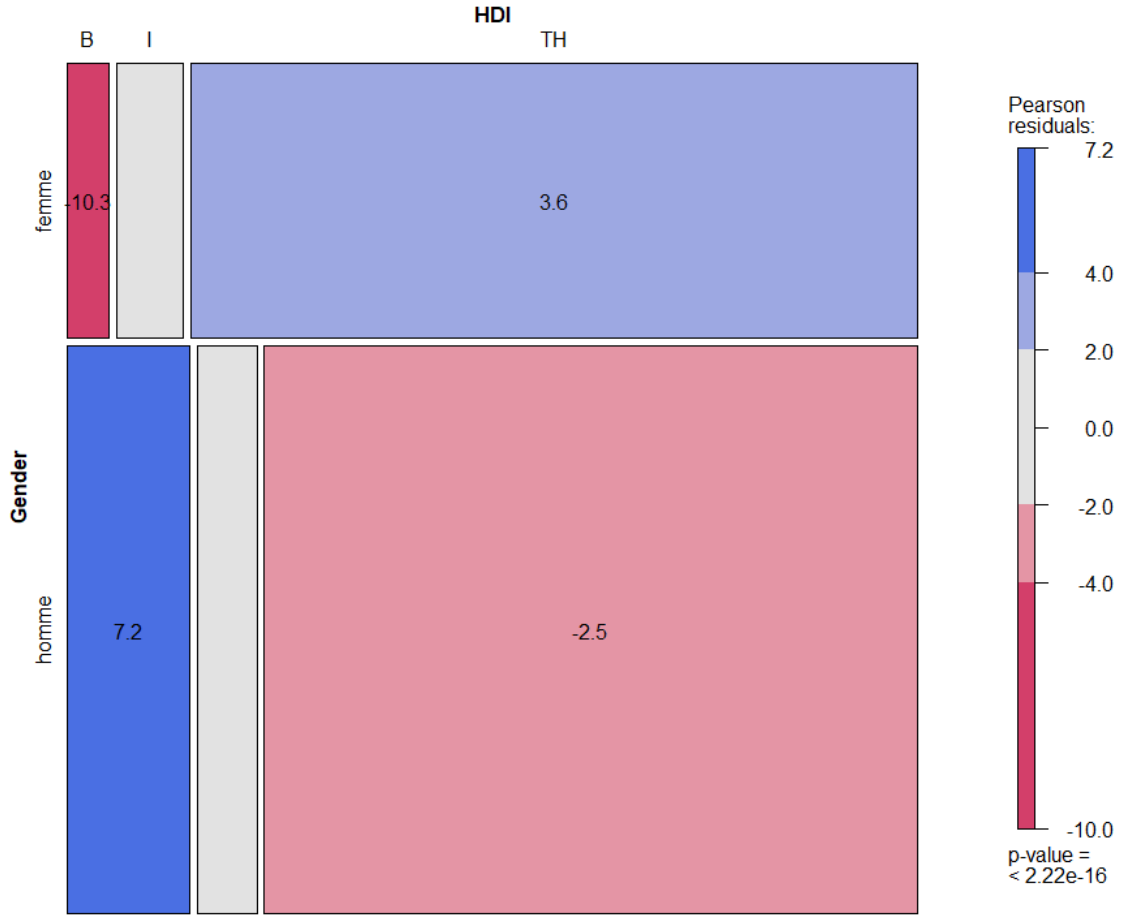

FIGURE 1 – Mosaic plot des valeurs obtenues du test χ^2 de Genre par rapport à IDH

Un graphique en mosaïque est un carré subdivisé en carreaux rectangulaires dont la surface représente la fréquence relative conditionnelle d'une cellule du tableau de contingence. Chaque carreau est coloré pour montrer l'écart par rapport à la fréquence attendue (résiduel) d'un test χ^2 de Pearson. Les couleurs représentent le niveau du résidu pour cette cellule / combinaison de niveaux. La légende est présentée à droite du graphique. Plus précisément, le bleu signifie qu'il y a plus d'observations dans cette cellule que ce à quoi on pourrait s'attendre avec le modèle nul (indépendance). Le rouge signifie qu'il y a moins d'observations que ce qui aurait été attendu. On peut lire cela comme une indication des cellules qui contribuent à la signification du résultat du test du chi².

La Figure 1 représente un mosaic plot des valeurs résiduelles obtenues après un χ^2 de Genre par rapport à IDH. On constate que la répartition de Genre est la suivante, 67.3% sont des hommes et 23.7% des femmes. De même, la répartition de l'IDH est la suivante, B : 14.3% sont des femmes 85.7% des hommes, I : 35% sont des femmes 65% des hommes, TH : 35.1% sont des femmes et 64.9% des hommes. Le résultat du χ^2 est égal à 179.05 avec un degré de liberté égal à 2, qui se calcule comme suit : (#row -1) * (#col -1). La p-value est de $2.2e^{-16}$ ce qui est inférieur au seuil de 5%. En conséquence, nous rejetons l'hypothèse nulle H_0 et admettons que l'hypothèse alternative H_1 est vrai. Il existe donc un lien entre le genre et l'IDH, le test V de Cramer nous permet de déterminer l'intensité de la relation entre ces deux variables. Celui-ci avec un score de 0.14 nous indique une intensité faible.

Nous pouvons voir que pour l'IDH B les femmes sont très fortement sous-représentées et les hommes fortement sur-représentés. L'IDH B représentant les pays pauvres/en voie de développement, nous pouvons imaginer plusieurs raisons à cela, une restriction à l'éducation et donc aux mooc pour les femmes, un accès internet restreint par les hommes, ou autres. A contrario, l'IDH TH montre une légère sur représentation des femmes et une légère sous représentation des hommes. L'IDH TH représentant les pays les pays les plus riches/développés, nous pouvons imaginer plusieurs raisons à cela, l'accès à l'éducation est identique pour les femmes et les hommes, ou autres.

Chapitre 3

Modèle linéaire, tests non paramétriques

Test non paramétrique

Nous réalisons un test non paramétrique dit de Wilcoxon sur les variables du nombre total de vidéos vues et du genre (homme et femme), avec les hypothèses suivantes :
H_0 : la moyenne des deux groupes sont égales sur le nombres de vidéos vues.
H_1 : la moyenne des deux groupes sont différents sur le nombres de vidéos vues.
La p-value est de 0.000481 et donc inférieure au seuil de 5%. Le test est donc statistiquement significatif et nous rejetons l'hypothèse nulle H_0 et admettons l'hypothèse alternative H_1. Les femmes regardent plus de vidéos que les hommes.

Test de corrélation de Spearman

La corrélation de Spearman est l'équivalent non paramétrique de la corrélation de Pearson. Elle mesure le lien entre deux variables. Le coefficient de corrélation varie entre -1 et +1, 0 reflétant une relation nulle entre les deux variables, une valeur négative (corrélation négative) signifiant que lorsqu'une des variables augmente, l'autre diminue ; tandis qu'une valeur positive (corrélation positive) indique que les deux variables varient ensemble dans le même sens. On doit donc réaliser un test d'hypothèse.
H_0 : pas de corrélation entre les deux variables : $\rho = 0$
H_1 : corrélation entre les deux variables : $\rho \neq 0$
La p-value est de $2.2e^{-16}$ et donc inférieure au seuil de 5%. Le test est donc statistiquement significatif et nous rejetons l'hypothèse nulle H_0 et admettons l'hypothèse alternative H_1. Le coefficient ρ est égal à 0.80, ce qui montre une corrélation positive forte entre nos deux variables.

La Figure 2 ci-dessous représente un scatter plot du nombre de quiz effectués par rapport au nombre de vidéos vues avec droite de régression. Nous pouvons voir que plus un apprenant a regardé de vidéos, plus il effectuera de quiz après. La droite de régression est une fonction affine ayant pour équation $y = 0.15x + 0.53$. Nous voyons aussi que les données ne sont pas distribuées normalement. Elles semblent ordonnées par rang.

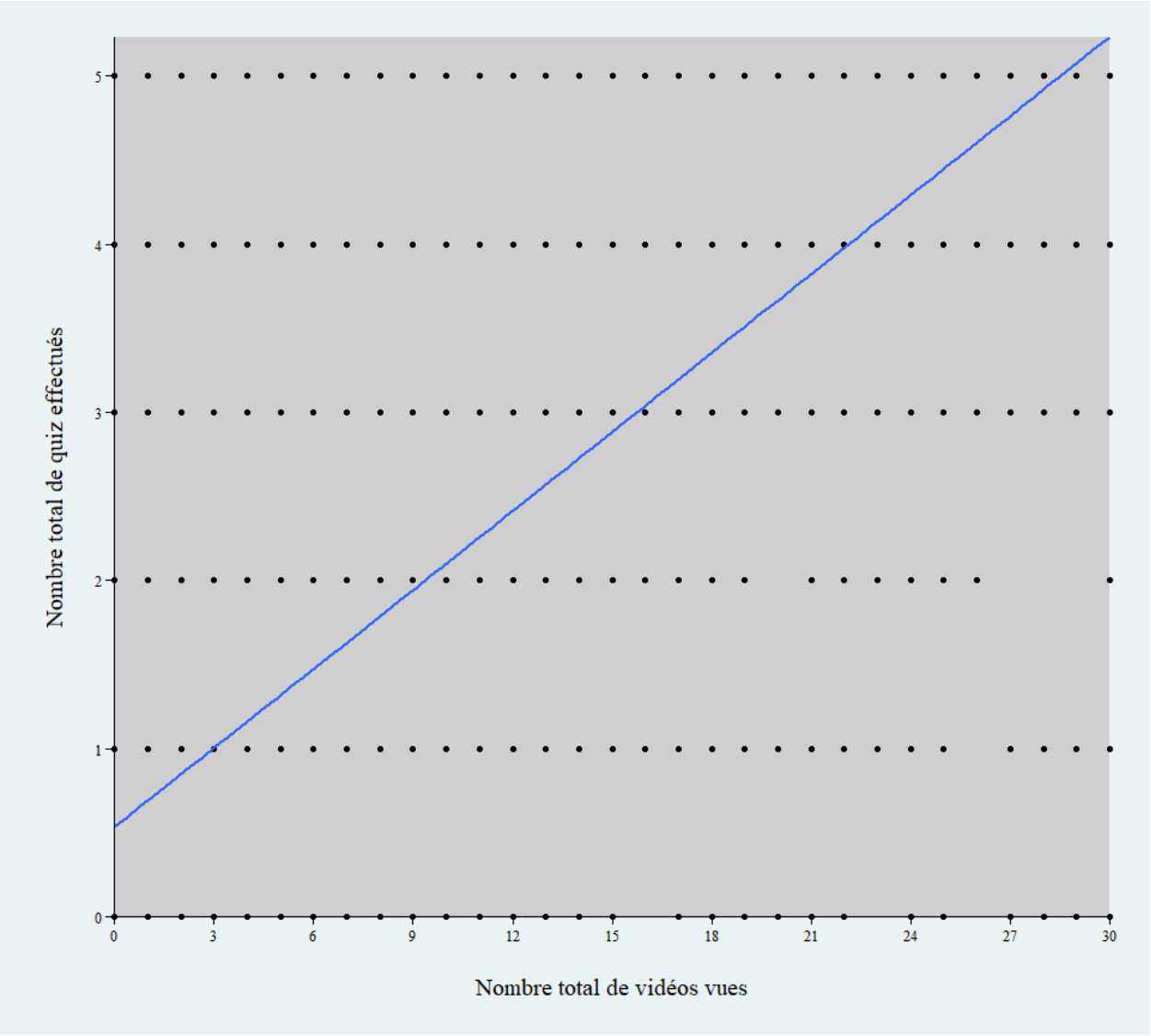

FIGURE 2 – Scatter plot du nombre de quiz effectués par rapport au nombre de vidéo vue avec droite de régression

ANOVA sans interaction

Les hypothèses sont les suivantes :
H_0 : tous les groupes de genre ont la même moyenne de vidéos regardées.
H_1 : tous les groupes de genre n'ont pas la même moyenne de vidéos regardées.
H_0 : tous les groupes IDH ont la même moyenne de vidéos regardées.
H_1 : tous les groupes IDH n'ont pas la même moyenne de vidéos regardées.

	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 – Table ANOVA du nombre de vidéos vues par rapport au genre et IDH sans interaction

La Table 2 représente une table ANOVA du nombre de vidéos vues par rapport au genre et IDH sans interaction entre les deux dernières. Nous avons trouvé une différence statistiquement significative dans le nombre de vidéos visionnées par la variable Genre ($F(1, 8947) = 14.28$, $p < 0.00$) et par la variable IDH ($F(2, 8947) = 284.67$, $p < 0.00$). Nous rejetons donc les hypothèses nulles pour ces deux variables et admettons leurs hypothèses alternatives.
Le degré de liberté ou degrees of freedom (Ddl ou Df) désigne le nombre maximal de valeurs logiquement indépendantes, c'est-à-dire de valeurs qui ont la liberté de varier, dans l'échantillon de données. Il se calcule ainsi : $D_f = N - 1$ avec D_f le degré de liberté et N la taille de l'échantillon. Par exemple, la variable *Genre* contient deux échantillons (homme et femme) donc $N = 2$, ainsi $D_f = 2 - 1 = 1$.

ANOVA avec interaction

Les hypothèses sont les suivantes :
H_0 : tous les groupes de genre ont la même moyenne de vidéos regardées.
H_1 : tous les groupes de genre n'ont pas la même moyenne de vidéos regardées.
H_0 : tous les groupes IDH ont la même moyenne de vidéos regardées.
H_1 : tous les groupes IDH n'ont pas la même moyenne de vidéos regardées.
H_0 : tous les groupes de genre et IDH ont la même moyenne de vidéos regardées.
H_1 : tous les groupes de genre et IDH n'ont pas la même moyenne de vidéos regardées.

	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 – Table ANOVA du nombre de vidéos vues par rapport au genre et IDH avec interaction

La Table 3 représente une table ANOVA du nombre de vidéos vues par rapport au genre et IDH sans interaction entre les deux dernières. Nous avons trouvé une différence statistiquement significative dans le nombre de vidéos visionnées par la variable Genre ($F(1, 8945) = 14.28$, $p < 1.5e^{-4}$) et par la variable IDH ($F(2, 8945) = 284.67$, $p < 2e^{-16}$). De même, nous constatons qu'il n'y a pas d'interaction statistiquement significative dans l'interaction entre le genre et l'IDH ($F(2, 8945) = 1.54$, $p < 2.24e^{-1}$). Nous rejetons donc les hypothèses nulles pour ces deux variables séparément et admettons leurs hypothèses alternatives et admettons l'hypothèse nulle lorsque nous considérons leur interaction.

Modèle linéaire avec 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 des coefficients du modèle linéaire du total de vidéo visionnées en fonction du genre de l'IDH et de l'interaction entre le genre et l'IDH

La Table 4 représente les coefficients du modèle linéaire du total de vidéo visionnées en fonction du genre de l'IDH et de l'interaction entre le genre et l'IDH. Les femmes d'un pays avec l'IDH B (intercept) regardent en moyenne 6,95 vidéos (p value = $1.56e^{-13}$), celles avec un IDH I regardent en moyenne 2.90 vidéos de plus (p value = $1.56e^{-2}$) et celles avec un IDH TH 8.42 (p value < $2e^{-16}$6) vidéos de plus. Les hommes d'un pays avec l'IDH B regardent en moyenne 6,37 vidéos (p value = $5.64e^{-1}$), ceux avec un IDH I regardent en moyenne 1.93 vidéos de plus (p value = $1.60e^{-1}$) et ceux avec un IDH TH 0.30 vidéos de plus (p value = $7.72e^{-1}$).

Chapitre 4

Régression logistique

4.1 Présenter des 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 d'odds ratio de la réussite à l'examen final en fonction de l'IDH et du Genre

La Table 5 représente la table d'odds ratio de Exam.bin en fonction de l'IDH et de Genre. L'IDH B sert de référence, ainsi les Odds Ratio de l'IDH I et de l'IDH TH seront calculés en fonction de celui-ci. Un apprenant venant d'un pays avec un IDH I a plus de chance de passer l'examen final qu'un apprenant venant d'un pays avec un IDH B (OR : 1.12, CI : 0.85-1.47, p-value : $4.15e^{-1}$). Un apprenant venant d'un pays avec un IDH TH a plus de chance de passer l'examen final qu'un apprenant venant d'un pays avec un IDH B (OR : 1.37, CI : 1.14-1.66, p-value : $7.86e^{-4}$). Le Genre femme sert de référence, ainsi les Odds Ratio de Genre homme seront calculés en fonction de celui-ci. Un apprenant de Genre homme a moins de chance de passer l'examen final qu'un apprenant de Genre femme (OR : 0.89, CI : 0.80-1.00, p-value : $4.72e^{-2}$)
En comparant les résultats de la Table 5 et de la Table 4, nous pouvons constater que ces résultats suivent le même raisonnement. Nous voyons que les chances qu'une personne passe l'examen final augmenteront si celle-ci se trouve dans un pays avec IDH TH et donc a la possibilité de visionner un grand nombre de vidéos. Nous constatons aussi le fait que les hommes tendent à moins passer l'examen final.

L'Odds Ratio, noté OR, également appelé rapport des chances ou rapport des cotes, est une approche non paramétrique permettant de mesurer l'association entre deux variables X^1, X^2 en déterminant la chance/le risque qu'un évènement de X^2 se produise sachant les valeurs de X^1. Le Risque Relatif, noté RR, est une approche non paramétrique permettant de mesurer l'association entre deux variables X^1, X^2 en déterminant la chance/le risque qu'un évènement de X^2 se produise dans l'un des deux groupes de X^1 par rapport à l'autre groupe de cette même variable. L'Odds Ratio est le rapport entre la cote d'un évènement chez le premier sous-groupe et la cote de ce même évènement chez le second groupe alors que le Risque Relatif est le rapport entre le risque de voir un évènement se produire chez le premier sous-groupe et le risque de voir ce même évènement se produire chez le second sous-groupe. L'Odds Ratio converge vers le Risque Relatif quand le nombre d'évènements de l'Odds Ratio est faible.

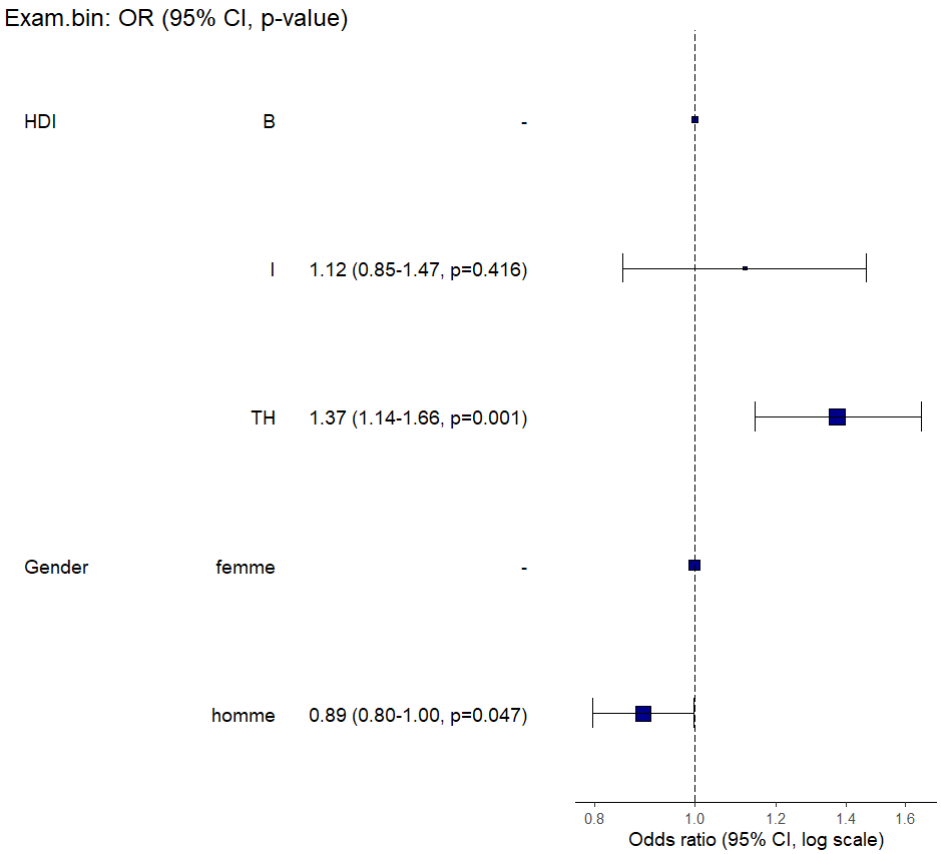

FIGURE 3 – Forest plot des odds ratio de la réussite à l'examen final en fonction de l'IDH et du Genre

La Figure 3 représente un forest plot des odds ratio de Exam.bin en fonction de l'IDH et de Genre. Elle est la représentation graphique de la Table 5. Les lignes horizontales représentent la longueur de l'intervalle de confidence, plus celle-ci est grande, plus large est l'intervalle de confidence et moins le résultat est fiable. Si un point se situe à gauche de la ligne verticale alors son Odds Ratio est inférieur à 1. Si un point se situe sur la ligne verticale alors son Odds ratio est égal à 1. Si un point se situe à droite de la ligne verticale alors son Odds Ratio est supérieur à 1.

4.2 Données de comptage et loi de Poisson

La Figure 4 représente la distribution du nombre de vidéos visionnées. Celle-ci est bimodale, avec la plus forte des mode à gauche indiquant que très peu de vidéos sont visionnées (1-2 vidéos pour un total d'environ 8 000), la seconde mode se situe à droite indiquant qu'un grand nombre de vidéos sont visionnées (27-30 vidéos pour un total d'environ 2 500 vidéos). Ces résultats nous indiquent deux tendances : soit l'apprenant regarde peu de vidéos, soit il regardera un grand nombre de vidéos

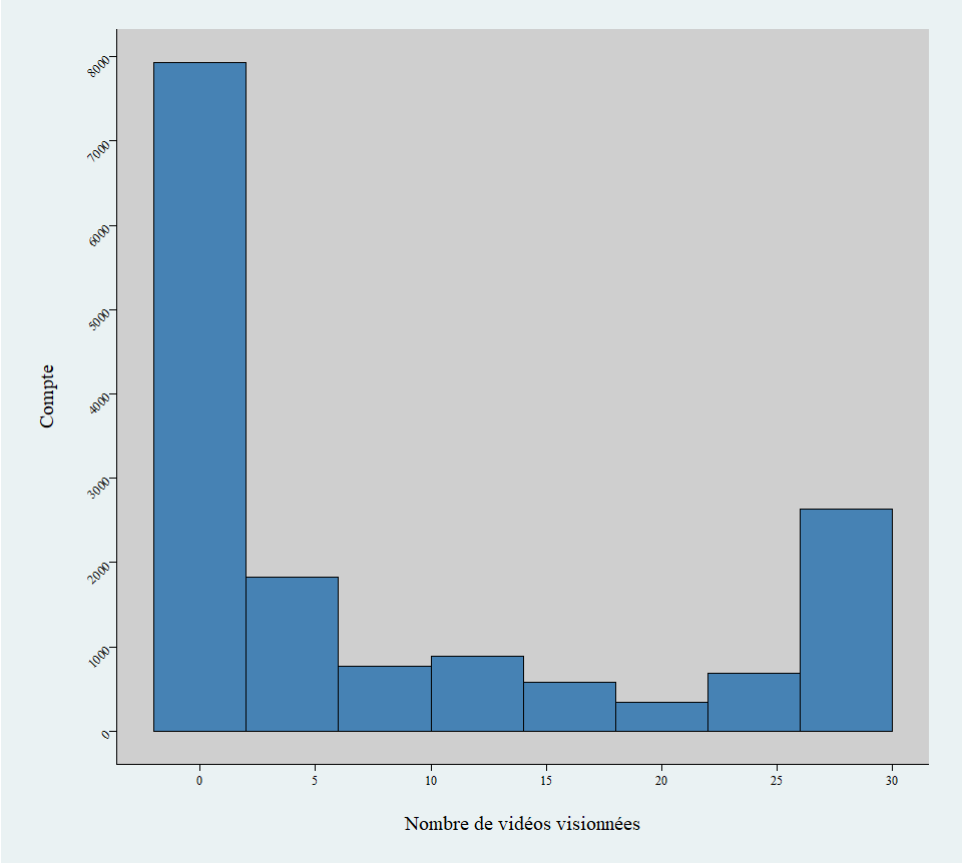

FIGURE 4 – Distribution du nombre de vidéos visionnées

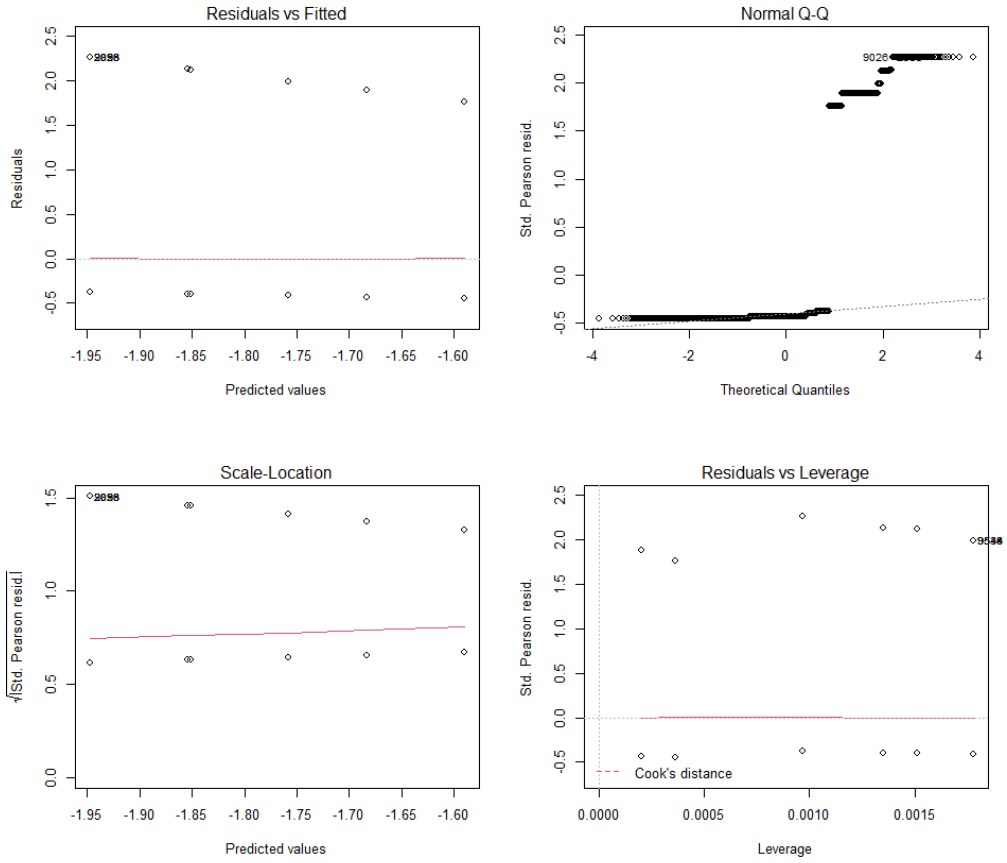

FIGURE 5 – Normalité de la distribution du modèle linéaire généralisé du statut de l'examen final en fonction du genre et de l'IDH

La Figure 5 est composée de quatre graphiques distincts :

- Residuals vs Fitted : il devrait être symétriquement distribué avec un groupement vers le milieu du graphique aux environs de zéro et aucun motif discernable.
- Q-Q plot : un Q-Q plot permet de visualiser la répartition d'une distribution donnée au regard d'un modèle théorique. Si la distribution étudiée est normale alors l'ensemble des points représentant les observations se trouveront sur la droite du modèle théorique.
- Scale-Location : utiliser pour vérifier l'homoscédasticité ou l'hétéroscédasticité.
- Residuals vs Leverage : permet de détecter des outliers.

L'homoscédasticité signifie que la variance des erreurs de la régression est identique . L'homoscédasticité est aussi appelée homogénéité de variance.
L'hétéroscédasticité signifie que la variance des erreurs de la régression diffère.

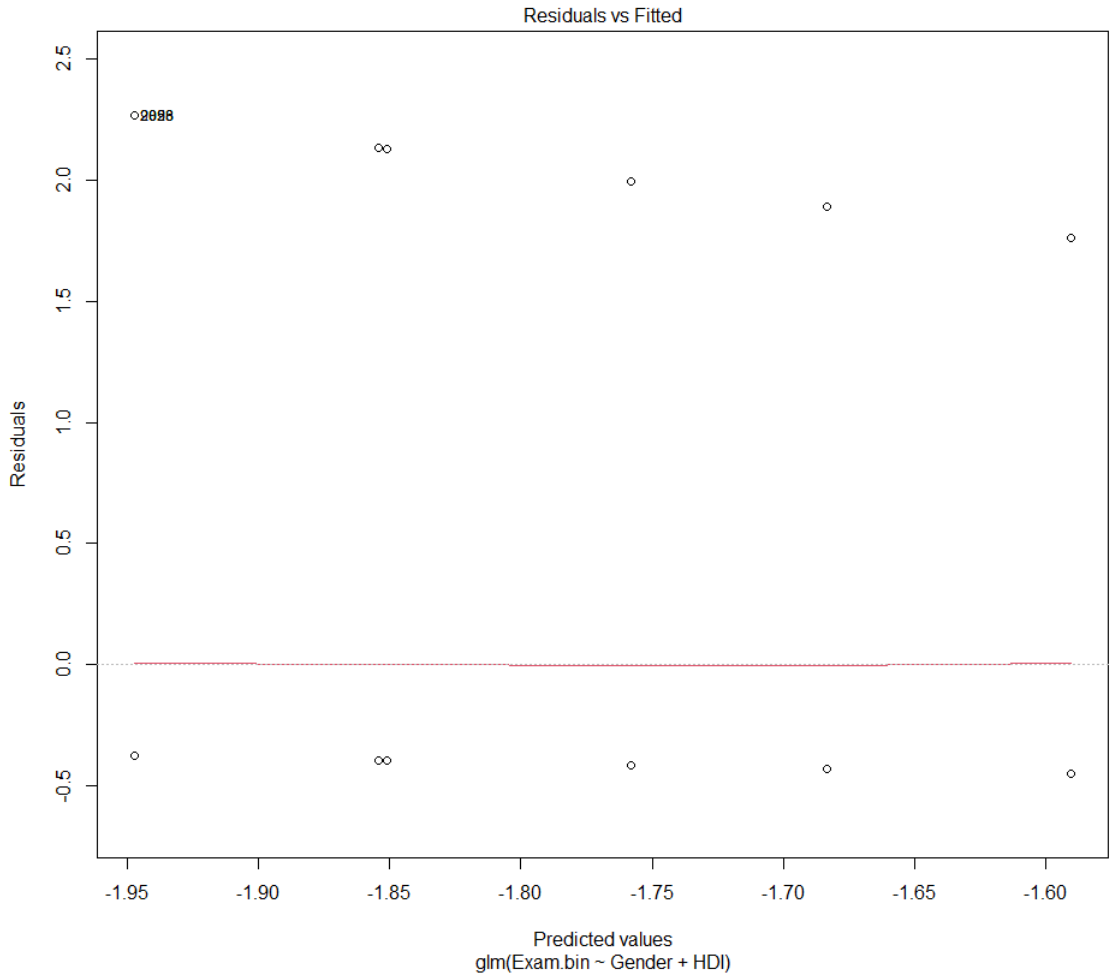

FIGURE 6 – Residuals vs Fitted du modèle linéaire généralisé du statut de l'examen final en fonction du genre et de l'IDH

La Figure 6 représente les résidus en fonction des valeurs prédites pour un modèle linéaire généralisé suivant une famille de Poisson. Nous pouvons voir que la distribution des points autour de l'espérance conditionnelle n'est pas symétrique ce qui indique une distribution biaisée, comme montré à la Figure 4.

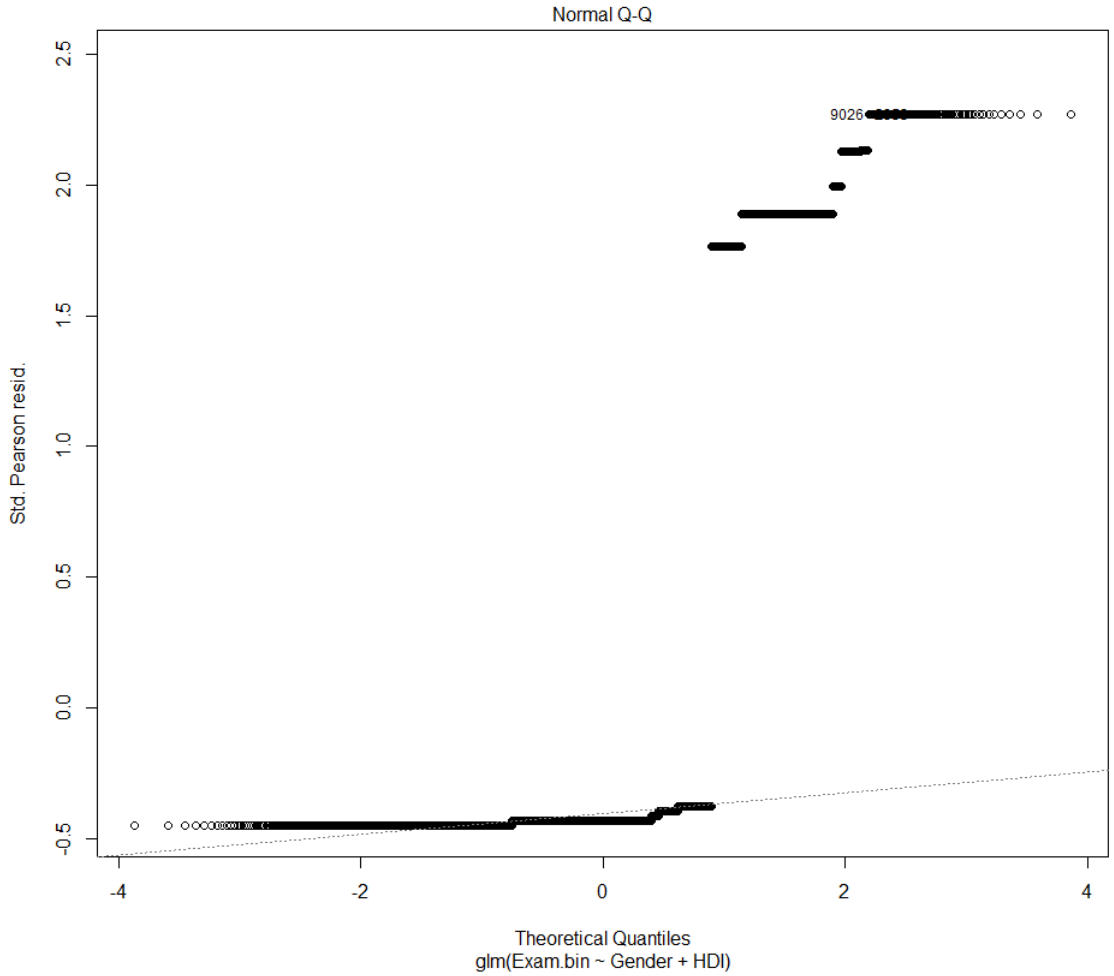

FIGURE 7 – QQ plot du modèle linéaire généralisé du statut de l'examen final en fonction du genre et de l'IDH

Le Q-Q plot de la Figure 7 se découpe en trois parties. La première allant du quantile -4 au quantile -2 montre un étalement et un éloignement des points par rapport à la droite théorique, cela signifie une distribution fortement biaisée à gauche. La seconde partie du quantile -1 au quantile 1 montre une distribution sans biais car l'ensemble de ces points se trouve au voisinage de la droite théorique. La troisième partie allant du quantile 1 au quantile 4 montre la présence d'une queue et d'un mode.

Table des figures

Liste des tableaux