Rapport de Statistique sur les Données des MOOC

RU WANG

Encadrant : DR. MATTHIEU CISEL

8 juillet 2022

Résumé

Ce rapport explore des modèles d'engagement d'apprentissage et des profils dans des jeux de données relatifs à deux MOOCs pertinents. Il apparaît que la majorité des apprenants sont des observateurs. Nous avons également étudié statistiquement comment la consommation de vidéos d'apprentissage est influencée par le genre et le pays de résidence en utilisant des outils d'analyse différents.

Mots-clefs : MOOC, Statistiques analytiques, Régression linéaire

Table des matières

Table des figures

Liste des tableaux

1 Introduction

La formation en ligne existe depuis des décennies, de nombreuses universités proposant des cours en ligne à un public restreint et limité. Ce qui a changé en 2011, c'est l'échelle et la disponibilité, lorsque l'Université de Stanford a proposé trois cours gratuits au public. Après son énorme succès avec des inscriptions d'environ 100 000 apprenants ou plus, plusieurs plateformes en ligne ont cherché à stimuler des initiatives similaires, principalement des établissements d'enseignement supérieur. Avec l'accroissement du nombre de cours et des plateformes de "Massive Open Online Course" (MOOC), l'étude sur les utilisateurs et leurs usage est devenue un domaine de recherche. Dans ce rapport, nous avons analysé des profils d'apprenants à travers deux MOOCS. En utilisant les statistique descriptives et statistiques mathématiques, nous avons montré l'existence d'une corrélation entre des profils d'inscrits, leur engagement et leur comportement.

2 Présentation des Données

Dans ce rapport, nous utilisons les données de 2 MOOCs :

-'Effectuation', donné par Philippe Sibersah, EMLYON Business school, (référencé par la suite comme MOOC1),

-'ABC de la Gestion de Projet', donné par Rémi Bachelet, Ecole Centrale Lille (référencé par la suite comme MOOC2).

Les données sont constitués d'informations liées aux apprenants inscrits sur ces 2 cours pendant 3 sessions. Ces deux MOOCs donnent lieu à un certificat de réussite. Dans le cas du MOOC1, qui dure 5 semaines, il est nécessaire de soumettre un travail de mi-session et de réussir un examen pour obtenir le certificat. Pour le MOOC2, qui dure 4 semaines, réussir un examen est également requis pour l'obtention du certificat.

3 Description de données

Dans les jeux de données relatifs à chaque MOOC, des variables sont binaires (valeur 0 ou 1). Ces variables permettent d'indiquer le respect d'une condition (réussite à l'examen,

consultation d'une vidéo, complétion d'un quizz, etc.). Malgré le caractère binaire des résultats, les données ont été stockées dans un format int64 ou float64. D'autres informations au contraire sont de natures diverses (Student_ID, genre, pays des utilisateurs, etc.). Pour compléter ces éléments, des extraits de données sont fournis en Table 1 et Table 2, consacrés respectivement aux MOOC1 et MOOC2. Une colonne Exemple illustre ce qui est contenu dans chaque variable.

TABLE 1 – Variable de MOOC1(N=14974)

Variable	Type de données	Valeurs manquantes(%)	Exemple
Student_ID	int64	0	28
Exam.Bin	int64	0	0
Assignment.Bin	int64	0	0
Quizze1.Bin	int64	0	0
Quizze2.Bin	int64	0	0
Quizze3.Bin	int64	0	0
Quizze4.Bin	int64	0	0
Quizze5.Bin	int64	0	0
S1.L1	int64	0	1
S1.L2	int64	0	0
S1.L3	int64	0	0
S1.L4	int64	0	0
S1.L5	int64	0	0
S1.L6	int64	0	0
S2.L1	int64	0	0
S2.L2	int64	0	0
S2.L3	int64	0	0
S2.L4	int64	0	0
S2.L5	int64	0	0
S2.L6	int64	0	0
S3.L1.1	int64	0	0

following...

. . . continuing

Variable	Type de données	Valeurs manquantes(%)	Exemple
S3.L1.2	int64	0	0
S3.L2	int64	0	0
S3.L3	int64	0	0
S3.L4	int64	0	0
S3.L5	int64	0	0
S4.L1.1	int64	0	0
S4.L1.2	int64	0	0
S4.L2	int64	0	0
S4.L3	int64	0	0
S4.L4	int64	0	0
S4.L5	int64	0	0
S5.L1.1	int64	0	0
S5.L1.2	int64	0	0
S5.L2	int64	0	0
S5.L3	int64	0	0
S5.L4	int64	0	0

TABLE 2 – Variable de MOOC2(N=16176)

Variable	Type de données	Valeurs manquantes(%)	Exemple
Student_ID	float64	0	68029
Gender	object	90.25	un homme
Country	object	90.26	France
Country_HDI.fin	object	90.34	TH
Certif.bin	float64	76	1

Dans les données sur le MOOC1, les variables Sx.Ly indiquent si les apprenants ont visualisé la vidéo y de la semaine x.(S1.L2 représente la 2ème vidéo de la semaine 1). Les quizz sont associés au numéro de la semaine de formation (Quizz3 correspond au quizz de

fin de la semaine 3). À l'exception de la variable Student_ID, toutes les données relatives au MOCC1 sont binaires. La valeur 1 représente un résultat positif par rapport aux attentes, tandis que la valeur 0 représente une valeur negative. Par exemple, la Table 1 montre que l'étudiant représenté, d'identifiant 28, n'a pas réussi l'examen. Par ailleurs, il n'a pas fourni le travail demandé (Assignment.Bin), n'a fait aucun quizz, et n'a seulement regardé qu'une vidéo sur l'ensemble du programme.

Contrairement aux données du MOOC1, les données recueillies du MOOC2 contiennent de nombreuses lacunes, identifiées par les valeurs très élevées (>75%) de valeurs manquantes. Cependant, beaucoup plus d'informations sur les utilisateurs sont proposées : le genre, le pays, l'indice de développement humain (HDI) du pays. Ces données sont particulièrement intéressantes pour la compréhension des comportements des utilisateurs.

4 Statistiques et Analyse de Données

4.1 Catégorie des apprenants

L'objectif de l'étude est de corréler les profils des utilisateurs aux comportements d'apprentissage. Pour ce faire, nous définissons 4 catégories comportementales, soit 4 types d'apprenants :

- 'Completer', personne qui a obtenu le certificat de réussite,

- 'Disengaging learner', personne qui a répondu à au moins un quiz ou a fourni un devoir demandé, mais n'a pas obtenu le certificat,

- 'Auditing learner', personne qui a visualisé plus de 6 vidéos, mais n'a remis aucune forme de travail demandé (devoirs, quiz, etc.) et n'a pas obtenu le certificat,

- 'Bystander', personne qui a visionné moins de 6 vidéos, qui n'a remis aucune forme de travail demandé (devoirs, quiz, etc.) et n'a pas obtenu le certificat.

La Table 3 montre que le taux de réussite pour le MOOC1 (marqué par le pourcentage de Completer) s'accroît significativement à partir de la session 2. Ce taux reste néanmoins inférieur au quart, mais il faut nuancer ce résultat en relevant que le taux de Bystander augmente jusqu'à 50 % des inscrits.

TABLE 3 – Distribution des 4 categories pour MOOC1 à travers 3 sessions

	Completer	Disengaging learner	Auditing learner	Bystander	Nombre total
Session1	0.25	58.43	1.91	39.41	7965
Session2	23.11	28.80	2.79	45.29	3798
Session3	21.71	24.54	2.76	50.99	3883

4.2 Chi-2 et mosaic plot

Les données relatives au MOOC2 contiennent des informations personnelles sur les utilisateurs, notamment le genre et l'indice de développement humain (HDI) du pays. Ce dernier correspond au niveau de richesse du pays où vit l'apprenant et admet trois valeurs : H (riche), I (Intermédiaire) et M (Modéré). Les données de genre enregistrées sont catégorisées en 2 types : Femme et Homme. On s'intéresse au lien pouvant exister entre le nombre de vidéos visualisées, le genre et le HDI. Il convient dans ce cas d'évaluer l'indépendance des variables genre et HDI.

Le test d'indépendance du chi carré ne peut comparer que des variables catégorielles. Il ne peut pas faire de comparaisons entre variables continues ou entre variables catégorielles et continues. De plus, le test d'indépendance du chi carré n'évalue que les associations entre les variables catégorielles et ne peut fournir aucune inférence sur la causalité. Le test du Chi-2 d'indépendance dont est ici utilisé. Le résultat du Chi-2 de conformité permet de rejeter l'hypothèse d'indépendance si la p-value est inférieure aux seuils de signification α (0.1, 0.05, 0.001) en général. En l'occurrence, la p-value est inférieure à 0,001. On rejette donc largement l'hypothèse d'indépendance. On peut dès lors affirmer avec moins d'une chance sur mille de se tromper qu'il existe un lien statistique entre les deux variables.

Vous trouverez nos données résumées dans le tableau Table 4 pour les apprenants ayant suivi les 2 MOOCs pour chaque session,

La Table 4 montre que, dans les trois sessions, l'indicateur, appelé « p-value », dont la valeur indique si deux variables sont significativement liées entre elles ou non, est significativement inférieur aux seuils de signification, donc les variables de genre et de HDI sont significativement liées entre elles.

Pour regarder plus précisément quelle variable contribue le plus ou le mois à l'attendu, une

TABLE 4 – Tableau de contingence des données des nombres de vidéos vues selon le genre et le HDI

	Session1		Session2		Session3	
HDI	Femme	Homme	Femme	Homme	Femme	Homme
H	1364	2800	683	1114	499	802
I	136	275	58	81	39	76
M	87	579	39	180	21	124
p-value	5.4e-24		1.2e-08		8.6e-08	

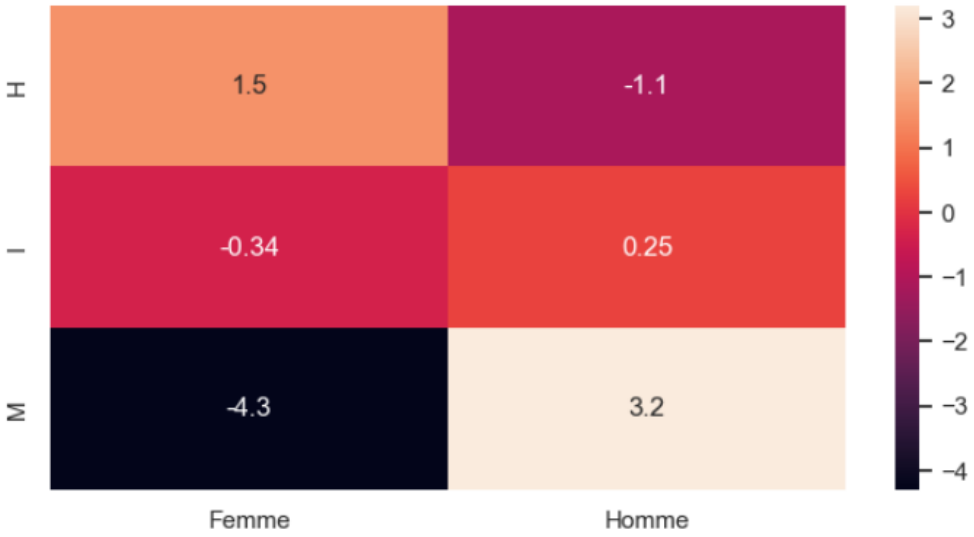


FIGURE 1 – Heatmap de résidus standardisés de Chi-2 du session 3

"heatmap" de résidus normalisés est montrée en Figure 1. Les résidus standardisés sont les résidus bruts (ou la différence entre les comptages observés et les comptages attendus), divisés par la racine carrée des comptages attendus. Le résidu standardisé peut être interprété comme n'importe quel score standard. Les résidus standardisés sont calculés pour chaque cellule du plan. Les valeurs résiduelles normalisées positives indiquent que le nombre de vidéos vus est plus élevé que prévu. Les valeurs résiduelles normalisées négatives indiquent que le nombre de vidéos vus est plus faible que prévu.

La Figure 1 nous montre que dans les pays de ressources modérées, les femmes contribuent aux nombre de vidéos vues significativement moins que nous l'attendions, contrairement aux hommes qui contribuent au nombre de vidéos vues significativement plus que prévu. Toutefois, cette différence entre les femmes et les hommes est le contraire de la situation dans les pays

riches. Ce sont les femmes qui ont vu plus de vidéos que prévu, tandis que les hommes en ont vu moins. De plus, les écarts aux prévisions sont plus importantes dans les pays de ressources modérées que dans les pays riches.

4.3 Modèle linéaire, tests non paramétriques

Selon le test du chi-2, il existe un lien entre le genre et le HDI. Maintenant, on analyse le comportement des apprenants quant au visionnage des vidéos en fonction du genre. Pour évaluer cela, les tests de Student et de Mann-Whitney sont exécutés.

Un test de Student, également connu sous le nom de test T de Student, est un outil permettant d'évaluer si deux groupes diffèrent l'un de l'autre. Ici, l' hypothèse nulle (H0) : les deux moyennes de groupes sont identiques, et l'hypothèse alternative (H1) : les deux moyennes de groupes sont différentes.

Le test U de Mann-Whitney est souvent utilisé comme solution alternative à l'utilisation d'un test de Student (T-test) dans le cas où les données ne sont pas distribuées selon une loi normale ou dans le cas où les données sont peu nombreuses. Il s'agit en effet d'un test non-paramétrique, i.e. un test qui ne repose pas sur une hypothèse de distribution des données.

TABLE 5 – Tableaux de paramètres par deux tests de nombres de vidéos vues selon les genres

	test T de Student		test U de Mann-Whitney	
Session	la statistique	p-value	la statistic	p-value
Session 1	-3.42	0.00	2853195.00	0.00
Session 2	-2.13	0.03	524195.50	0.08
Session 3	0.15	0.88	287942.00	0.83

Selon les résultats obtenus et présentés en Table 5, les results de p-values selon les deux tests sont proches pour chaque session. Les p-values des deux tests pour la session1 sont inférieures au seuil de signification α (0.01), donc l'hypothese nulle est rejectée, c'est-à-dire que les deux groupes dans la session 1 sont significativement differents. Au contaire, en session 3, les p-values des deux tests sont supérieurs à α (0.01), donc, nous échouons à rejeter l'hypothèse nulle. Cela montre que pour la session 3, il n'y a pas de différence significative entre les deux groupes. Pour la session 2, selon le seuil de signification α (0.01), il n'y a pas

non plus de différence significative entre les deux groupes (comme pour la session 3).

En revanche, si $\alpha = 0.1$, l'hypothèse nulle H0 est rejetée et les deux groupes sont significativement différents. Le choix de la valeur de α dépend du besoin et d'informations complémentaires. On voit néanmoins qu'en fonction de la valeur choisie, l'interprétation est différente.

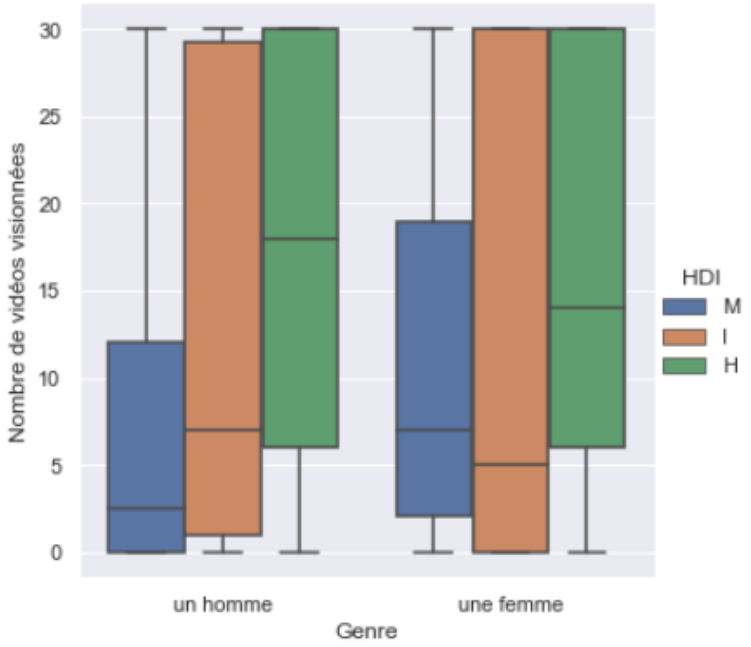


FIGURE 2 – Boxplot les vues de vidéos par selon de genre pour session3

L'étape suivante consiste à regarder s'il existe une corrélation pour les quiz. Une analyse de régression linéaire du nombre de vidéos vues par le nombre de quiz effectués est réalisée. Si le modèle de régression linéaire est validé, cela montre la corrélation entre le nombre de vidéos vues et le nombre de quiz effectués.

La Figure 3 est la représentation graphique du nombre de vidéos vues en fonction du nombre de quiz réalisés.

L'emploi de la méthode classique de comptage, représenté graphiquement par un scatterplot (Figure 3) ne permet que très rarement et pas dans le cas illustré de montrer une corrélation apparente. Toutefois, l'analyse numérique de la régression linéaire, dont le résultat est présenté en Table 6, montre que le nombre de quiz augmente avec le nombre de vidéos vues pour les 3 sessions.

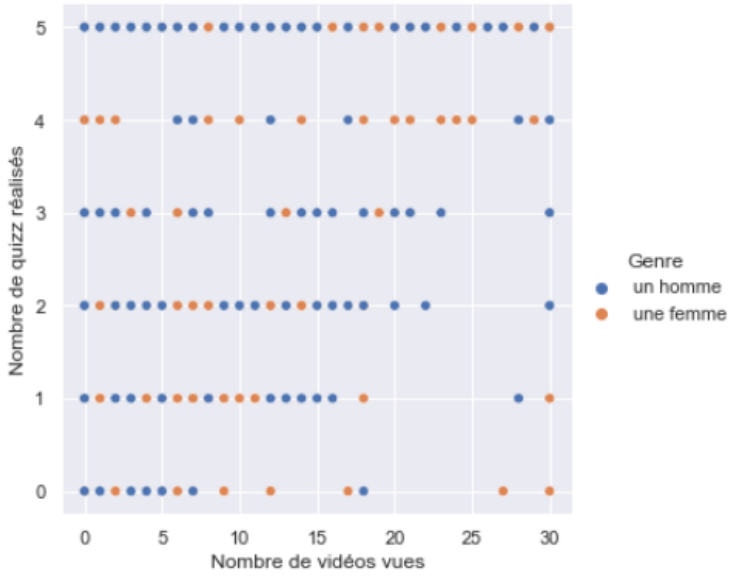


FIGURE 3 – Le nombre de vidéos vues en fonction de nombre de quiz réalisés

Le terme de "slope" indique le changement du nombre de quiz effectués pour une augmentation unitaire du nombre de vidéos vues. Ainsi pour une vidéo supplémentaire regardée, 0.15 quiz est effectué en plus. Dans le tableau récapitulatif (Table 6), nous pouvons voir que la valeur p pour les deux paramètres ("Intercept" et "Slope") est indiquée à 0, pour des raisons de précision numérique relative aux grandeurs statistiques étudiées. De ce résultat, nous pouvons rejeter l'hypothèse nulle à presque tous les niveaux de signification. Et ici l'hypothèse nulle est que la variable nombre de vidéos vues n'a pas d'influence sur le nombre de quiz realisés.

"R-squared" est le coefficient de détermination qui nous indique à quel point la variation en pourcentage de la variable dépendante peut être expliquée par la variable indépendante. Ici, une variation de 66,9 % du nombre de quiz effectués peut être expliquée par le nombre de vidéos vues. La valeur maximale possible de "R-squared" est 1, ce qui signifie que plus la valeur de R-squared est élevée, meilleure est la régression.

Comme montré sur la Figure 3, les données suivent une tendance monotone lorsqu'elles sont représentées graphiquement. Les données monotones se produisent lorsque les valeurs se déplacent dans une certaine direction sur un graphique, mais elles ne le font pas de manière linéaire. Pour illustrer cette tendance, nous avons effectué un calcul de corrélation de Spearman.

TABLE 6 – Tableaux de paramètres de l'analyse de régression linéaire

	Coef		t		p>|t|		
Session	Intercept	Slope	Intercept	Slope	Intercept	Slope	R-squared
Session 1	0.7336	0.1550	35.865	111.816	0.000	0.000	0.611
Session 2	0.3392	0.1573	14.371	96.215	0.000	0.000	0.709
Session 3	0.3465	0.1578	15.627	94.899	0.000	0.000	0.699

Le coefficient de corrélation de Spearman est une mesure statistique de la force d'une relation monotone entre données appariées. Le coefficient de corrélation est inférieur à 1 en valeur absolue, soit $-1 \leq r_s \leq 1$. Les coefficients de corrélation de Spearman dans les 3 sessions sont respectivement 0.794, 0.818, 0.783. Donc, cette analyse aboutit à la meme conclusion que l'analyse de regression linéaire : les deux variables sont fortement corrélées.

La variation de nombre de quiz réalisés augumente avec le nombre de vidéos vues. Or le nombre de vidéos vues varie selon le genre et le HDI. Donc, l'analyse de la variance(ANOVA) à deux facteurs est appliquée, car cela nous permet de savoir comment deux variables indépendantes, combinées, affectent une variable dépendante. Les results sont dans le Table 7.

En faisant apparaître explicitement un paramètre d'interaction du genre et du HDI, nous pouvons voir qu'il y a un effet de l'HDI sur le nombre de vidéos regardées (p-value=0). Mais l'effet d'interaction est significatif (p-value >0.3). Nous ne pouvons donc pas interpréter les effets principaux sans tenir compte de l'effet d'interaction. Cependent, comme la régression linéaire simple et le test de t, l'ANOVA doit respecter condition statistiques pour que les résultats soient valides,

- Distribution normale des résidus,

-Homoscédasticité de la variance des résidus entre les groupes,

ces conditions sont verifiés dans le section 4.5.

4.4 Régression logistique

Semblable à la régression linéaire, la régression logistique est également utilisée pour estimer la relation entre une variable dépendante et une ou plusieurs variables indépendantes,

TABLE 7 – Tableaux de paramètres de ANOVA à deux facteurs

	sum sq			df			F			PR(>F)		
	S.1	S.2	S.3	S.1	S.2	S.3	S.1	S.2	S.3	S.1	S.2	S.3
genre	63.75	141.22	204.11	1	1	1	0.51	1.04	1.46	0.47	0.31	0.23
hdi	50949.53	12732.47	9743.44	2	2	2	204.39	47.07	34.79	0	0	0
gender :hdi	270.68	87.61	151.19	2	2	2	1.09	0.32	0.54	0.34	0.72	0.58
Residual	651978.69	290499.11	217618.87	5231	2148	1554	NAN	NAN	NAN	NAN	NAN	NAN

mais elle est utilisée pour faire une prédiction sur une variable catégorielle par rapport à une variable continue. Une variable catégorielle est une variable binaire, qui peut être vraie ou fausse, oui ou non, 1 ou 0, etc. L'unité de mesure diffère également de la régression linéaire car elle produit une probabilité, mais la fonction logit transforme la courbe en S en ligne droite. Le resultat d'examen est une variable catégorielle, soit réussi noté 1, soit raté noté 0. Nous avons donc utilisé la régression logistique pour explorer l'effet du genre et du HDI sur les résultats de l'examen et du certificat. Les résultats sont montrés dans la Table 8.

TABLE 8 – Tableau de résultats de régression logistique

Réussite	Coef	Odds Ratio	z	P>|z|	[0.025	0.975]
Intercept	0.06	1.06	0.84	0.40	-0.08	0.19
Gener[T.une femme]	-2.1	0.81	-1.93	0.05	-0.42	0.00
HDI[T.I]	0.24	1.27	1.22	0.22	-0.14	0.62
HDI[T.M]	0.18	1.20	1.01	0.31	-0.17	0.53

Dans ce tableau, le modèle de regression logistique, qui donne les estimations du coefficient et de l'ordonnée à l'origine, nous donne l'équation suivante :

$$log(p/(1-p)) = logit(p) = 0.06 + -2.1 * femme + 0.24 * I + 0.18 * M$$

p est la probabilité de réussite. Comme l' "odds ratio" est l'exponentielle des coefficients, lorsque les coefficients sont positifs, les "odds ratios" sont supérieurs à 1. Lorsque les coefficients sont négatifs, les "odds ratios" sont inférieurs à 1.

Tout d'abord, notre coefficient est décrit comme Gener[T.une femme]. Il n'y a aucune mention d'homme, et c'est parce qu'avec les variables explicatives catégorielles, le premier niveau de la variable sera utilisé comme référence. Le groupe masculin dans la variable genre est utilisé comme référence, à laquelle tous les autres niveaux de la variable seront comparés. (Dans ce cas, il n'y a qu'un seul autre niveau - féminin.) Ainsi, notre coefficient de -2,1 est inférieur à zéro, ce qui signifie que les femmes sont plus susceptibles d'avoir échoué que les hommes (le niveau de référence). Plus précisément, le coefficient négatif signifie que les femmes étaient moins susceptibles d'atteindre le résultat positif de notre variable de réponse, c'est-à-dire la reussite.

Bien sûr, cela n'a aucun sens si nous ne regardons pas la valeur p ; mais dans ce cas, notre

valeur p est inférieure à 0,1. Ainsi, en effet, les femmes avaient moins de chances de reussite au MOOC que les hommes. Mais jusqu'à quel point ? Ainsi, avec un intervalle de confiance de 95%, les femmes avaient entre 0.42 et 0 fois plus de chances de réussir le MOOC que les hommes.

C'est similaire pour l'autre variable HDI, les pays riches H sont pris comme référence. Donc, le tableau ne montre que les parametres de HDI intermédiaires (I) et HDI modérés (M). Comme les p-values sont supérieures à 0.1, qui n'indique pas de signification importante. Pour être plus didactique, une représentation des "odds-ratios" via un "forest plot" est donné en Figure 4 :

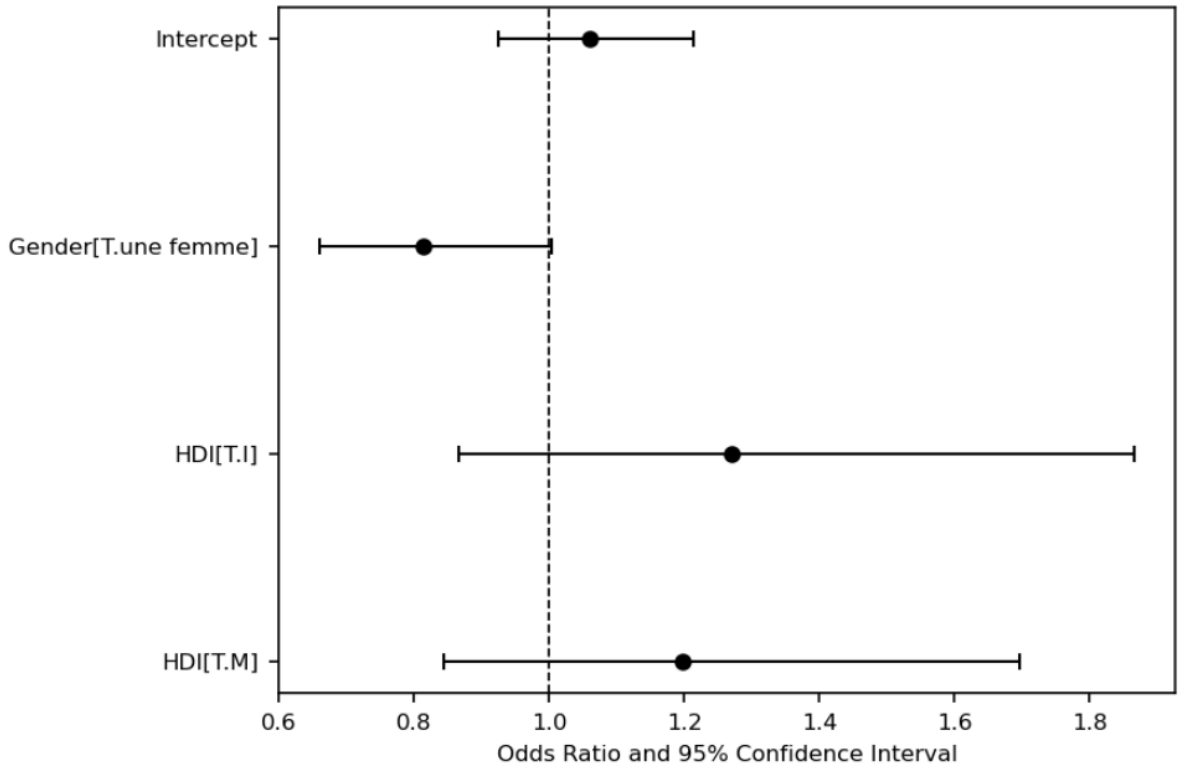
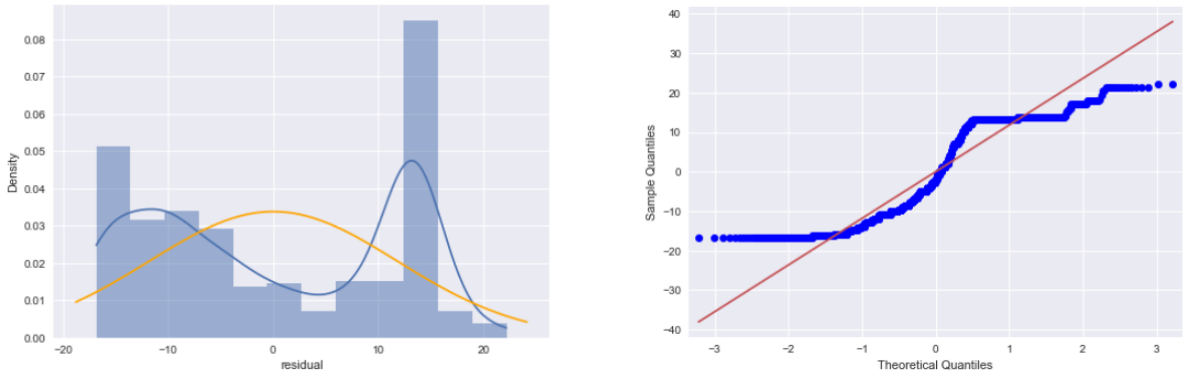


FIGURE 4 – "forest plot" des "odds ratios"

4.5 Données de comptage et loi de Poisson

Nous avons exploré le nombre de video vues selon les genres et les HDI avec la regression linéaire et les résultats de "p-values" sont supérieurs à 0.1. Comme une des hypothèses est que les résidus soient normalement distribués, nous l'avons vérifié en traçant la distribution



(a) Distribution de résidus (b) Q-Q plot de résidus

FIGURE 5 – Résidus normalisés

des résidus et le "Q-Q plot". Un graphique Q-Q, abréviation de graphique quantile-quantile, est un type de graphique que nous pouvons utiliser pour déterminer si les résidus d'un modèle suivent, ou non, une distribution normale. Si les points sur le graphique forment approximativement une ligne diagonale droite, alors l'hypothèse de normalité est satisfaite. Le graphique Q-Q de la Figure 5 (b) montre que les résidus ne suivent pas une distribution normale.

Etant donné que les données ne conviennent pas au modèle de régression lineaire, la variable du nombre de vidéos visionnées est une donnée de comptage, sur laquelle nous appliquons une régression de Poisson.

TABLE 9 – Tableau de result de régression Poission

Nombre de video vues	Coef	std err	z	P>|z|	[0.025	0.975]
Intercept	2.83	0.01	333.59	0.00	2.81	2.84
Gener[T.une femme]	-0.05	0.01	-3.59	0.00	-0.08	-0.02
HDI[T.I]	-0.27	0.03	-10.04	0.00	-0.33	-0.22
HDI[T.M]	-0.68	0.03	-23.08	0.00	-0.74	-0.62

À partir de ce tableau, nous pouvons voir qu'avec la modélisation de la régression de Poisson, toutes les valeurs de p sont inférieures à 0,01. Cela signifie que tous les paramètres ont une importance significative pour le nombre total de vidéos visionnées. Il y a plus d'hommes que de femmes qui regardent les vidéos, et plus le pays est riche, plus les gens regardent les vidéos de MOOC.

5 Discussion

Nous avons exploré les données de deux MOOC. Malgré l'importance du nombre d'inscriptions, la majorité des activités du cours est dirigée par les apprenants très engagés, qui généralement réussissent à obtenir le certificat ou l'examen, mais qui ne représentent qu'une minorité des apprenants. En utilisant différents types de régression : régression linéaire simple, logistique, ANOVA à deux facteurs et la régression de Poisson, nous trouvons que le HDI et le genre ont effets significatifs sur le nombre de vidéos vues des MOOCs. Ces résultats sont basés sur le pays de résidence des apprenants. Un travail supplémentaire peut être fait en fonction de la classe sociale, du poste, des réponses au questionnaire, etc. pour classer les apprenants et affiner les corrélations.

6 Bibliographie

[1] Matthieu Cisel, Mattias Mano, Rémi Bachelet, Philippe Silberzahn. A Tale of Two MOOCs : Analyzing Long-Term Course Dynamics. European Moocs Stakeholders Summit (eMOOCs), May 2015, Mons, Belgium. ⟨hal-01635080⟩

[2] Julien Jacqmin. Why are some Massive Open Online Courses more open than others ? Technovation , Volume 112(102395), April 2022.

[3] https ://libguides.library.kent.edu/spss/chisquare

[4] https ://support.minitab.com/fr-fr/minitab/18/help-and-how-to/statistics/tables/how-to/chi-square-test-for-association/interpret-the-results/all-statistics/

[5] http ://r.qcbs.ca/workshop04/book-fr/