Introduction aux statistiques

Matthieu Cisel - DU Data Analyst

Mai 2023

1 Introduction

Vous trouverez dans Datacamp deux cours d'introduction à la régression, l'un sur R, l'autre sur Python. Vous devez obtenir le certificat du cours. Certaines parties des cours portent sur la prédiction, qui ne sera pas mobilisée dans le cadre de ce projet. Néanmoins, cette perspective peut être utile. Par ailleurs, nous travaillerons avec des régressions multiples, qui sont abordées, sur le plan technique, uniquement dans les cours "Intermediate regression". Si leur suivi n'est pas obligatoire, les vidéos correspondantes pourront vous êtes utiles le cas échéant, pour des explications complémentaires.

Nous allons dans ce cours nous concentrer sur divers aspects du modèle linéaire (régression linéaire simple, test de Student, ANOVA), ainsi que sur la régression logistique. Nous apprendrons par exemple à décrire les résultats d'une table d'ANOVA ou d'odd-ratios. Le jeu de données utilisés dans ce projet porte sur des learning analytics issus de différentes itérations d'un MOOC (le MOOC Effectuation, appelé MOOC1). Il s'agit de retrouver certains des résultats obtenus dans un article intitulé "A Tale of Two MOOCs", publié en 2015. La focale porte sur l'engagement des apprenants, et notamment sur le visionnage de vidéos et la réalisation de quiz.

En premier lieu, nous allons définir des catégories de participants sur la base de leur engagement. Par exemple, s'ils ont passé l'examen ou obtenu le certificat, nous parlons de "Completer". Exam.bin, bin pour binaire, variable booléenne avec 1= obtention. Même logique pour le certificat.

Si un quiz (Quizz.1.bin=1 si le quizz 1 a été fait) a été réalisé ou un devoir soumis (Assignment.bin=1), mais le certificat n'a pas été obtenu / examen pas réalisé, on parle de "disengaging learners". Si aucun quiz n'a été réalisé et aucun devoir soumis (Assignment.bin=0), mais que l'apprenant a visualisé plus de 6 vidéos, nous parlons d'auditeur ("auditing learner"). Les vidéos sont numérotées par semaine et par place dans la séquence. La première vidéo de la semaine 1 est donc S1.L1. Cette variable prend la valeur 1 si la vidéo a été visionnée, 0 sinon. Si en plus de ne pas avoir fait de quiz/devoir, moins de 6 vidéos (strictement moins) ont été visionnées, c'est un "bystander".

2 Préparation du jeu de données

Nous vous fournissons des données sur des questionnaires et des logs issus de différentes itérations de différents MOOC. Votre première mission correspond à reconstituer une base de données rectangulaire avec tous les fragments que nous vous proposons. Par exemple, usages.effec.1 porte sur les logs de la première itération du MOOC Effectuation. Vous allez devoir utiliser des commandes comme merge (base), full-join, rbind ou rbind.fill (pour R), ou leurs équivalents Python. Vous devez commencer par faire un "column bind" pour lier les données de log et les données de questionnaires pour une itération donnée.

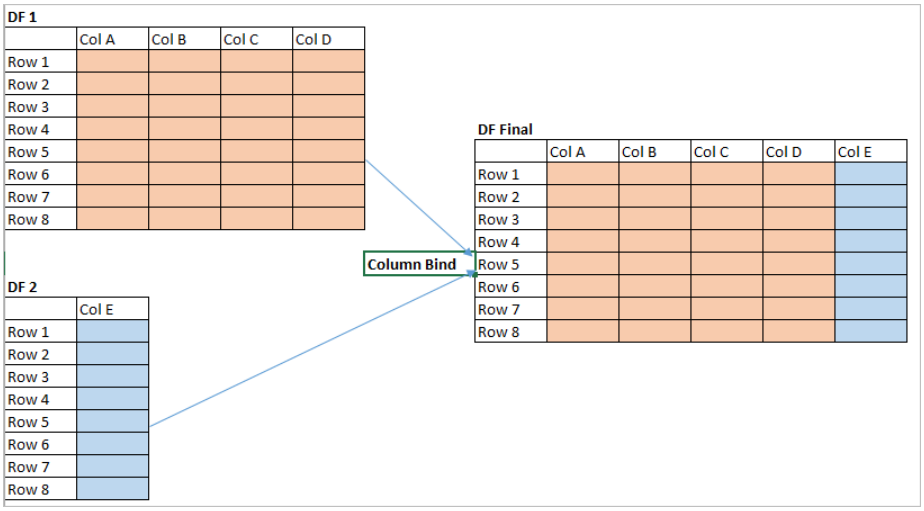

Figure 1: Column Bind

Vous devez ensuite réaliser un "row bind" pour rassembler les données issues de différentes itérations d'un MOOC donné. Avant la création de ce jeu de données global, créez une colonne intitulée itération (avec mutate, etc.) pour garder en mémoire le numéro d'itération correspondant. Vous allez constater que les différents jeux de données n'ont pas le même nombre de colonnes, car le nombre de vidéos ou de quiz a évolué d'une itération à l'autre. Vous allez ensuite simplifier le jeu de données, en ne conservant, comme variables issues des questionnaires, que les seuls éléments que vous allez mobiliser dans les analyses (HDI, et genre). C'est la commande select dans dplyr.

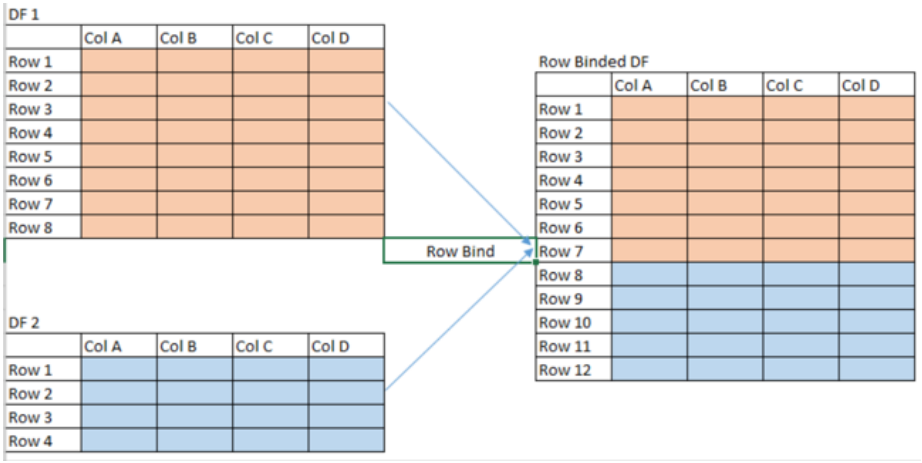

Figure 2: Row Bind

L'intégralité des étapes demandées est résumée dans le schéma ci-dessous. Les numéros sur les flèches représentent l'ordre dans lequel doivent être opérées les opérations de fusion.

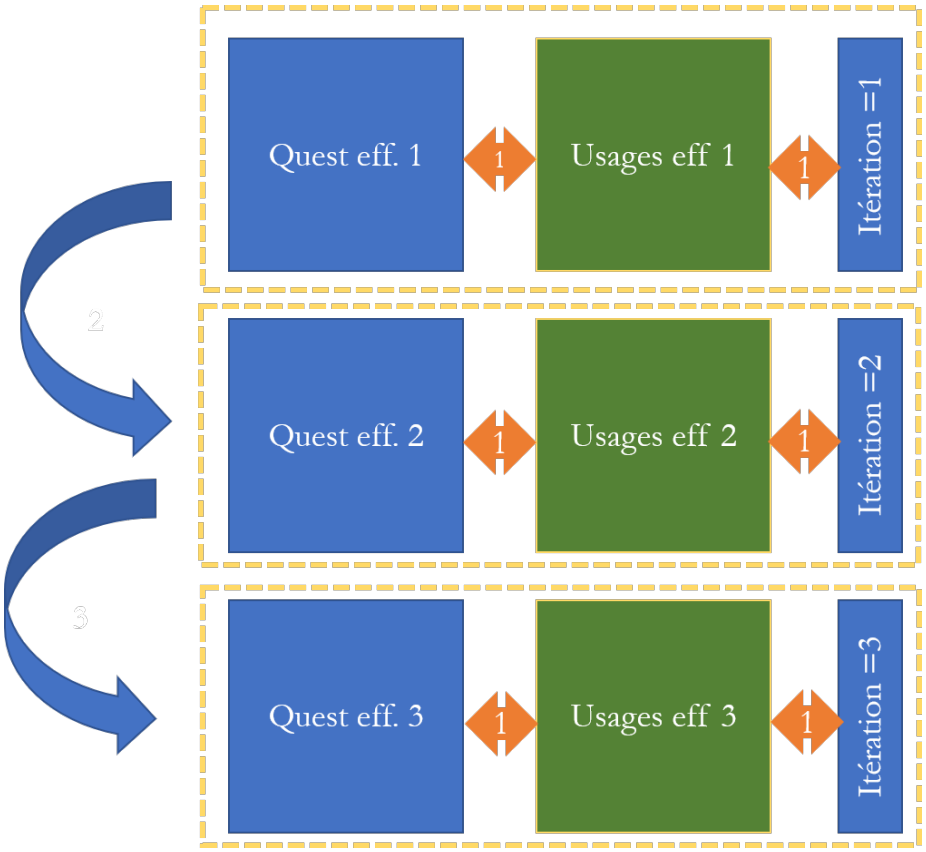

Figure 3: Schématisation de la préparation des données

Créez une nouvelle variable correspondant au nombre de vidéos visionnées pour un apprenant donné, et du nombre de quiz réalisés. En termes d'indices, nous pouvons vous donner les termes colSums ou rowSums (sur R), dont le fonctionnement est illustré dans la colonne ci-dessous. Faites un "count" de chacune des catégories concernant l'HDI (Human Development Index). Nous avons B (Bas), M (Moyen), H (Haut), et TH (Très haut). Créez une nouvelle variable HDI où vous regroupez dans une catégorie intermédiaire (I) les modalités M et H.

```
> myMat
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    6   11   16   21
[2,]    2    7   12   17   22
[3,]    3    8   13   18   23
[4,]    4    9   14   19   24
[5,]    5   10   15   20   25
> colSums(myMat)  # sum of each column
[1]  15  40  65  90 115
```

Figure 4: Faire une somme sur l'ensemble d'une colonne

3 Description du jeu de données

Votre rapport devra suivre de manière linéaire les instructions contenues dans ce polycopié. La différence avec votre notebook Jupyter résidera dans les paragraphes de description et d'interprétation.

Réalisez une table donnant, en lignes les proportions des quatre types d'apprenants que nous avons définis (bystander, auditing, completer, disengaging), et en distinguant en colonne les 3 itérations. Vous devez trouver une approche pour fournir de manière synthétique le nombre de personnes concernées par itération. La table devra ressembler à ce que l'on obtient ci-dessous, sur le plan esthétique.

	Variables	Nombre	Pct. Manquant
1	Date de premiere inscription en doctorat	63976	85,71%
2	Identifiant auteur	317655	29,04%
3	Langue de la these	383879	14,24%
4	Date de soutenance	390898	12,68%
5	Year	390898	12,68%
6	Identifiant etablissement	430559	3,82%
7	Mise a jour dans theses.fr	447467	0,04%
8	Directeur de these	447629	0,00%
9	Directeur de these (nom prenom)	447629	0,00%
10	Titre	447635	0,00%
11	Discipline	447639	0,00%
12	Etablissement de soutenance	447640	0,00%
13	Auteur	447644	0,00%
14	Identifiant directeur	447644	0,00%
15	Statut	447644	0,00%
16	Identifiant de la these	447644	0,00%
17	Accessible en ligne	447644	0,00%
18	Publication dans theses.fr	447644	0,00%

Figure 3 : Taux de données manquantes pour chaque variable

Figure 5: Table satisfaisante sur le plan esthétique

4 Chi2 et mosaic plot

Vous allez dans un premier temps croiser les variables HDI et Gender; vous devez trouver vous-mêmes les commandes nécessaires. Cette recherche autonome fait partie du projet. Réalisez un test d'indépendance fondé sur le chi2. Faites un mosaic plot concernant les résidus de ce chi 2 (cf. figure ci-dessous). Sachez que la représentation des couleurs d'un mosaic plot de chi 2, sur Python, est sous-optimale. Que représentent les couleurs bleues et rouges ? Pourquoi n'est-il pas possible que toutes les cases soient bleues ou rouges ? En Python, la création des échelles de couleurs peut être fastidieuse; cantonnez-vous à représenter un mosaic plot sans résidus. Décrivez en quelques lignes les résultats obtenus, en fournissant les résultats des tests du chi2 (valeur du chi, et p-value). Calculez le V de Cramer. Que représente cette métrique. Proposez ensuite un paragraphe visant à interpréter ces résultats.

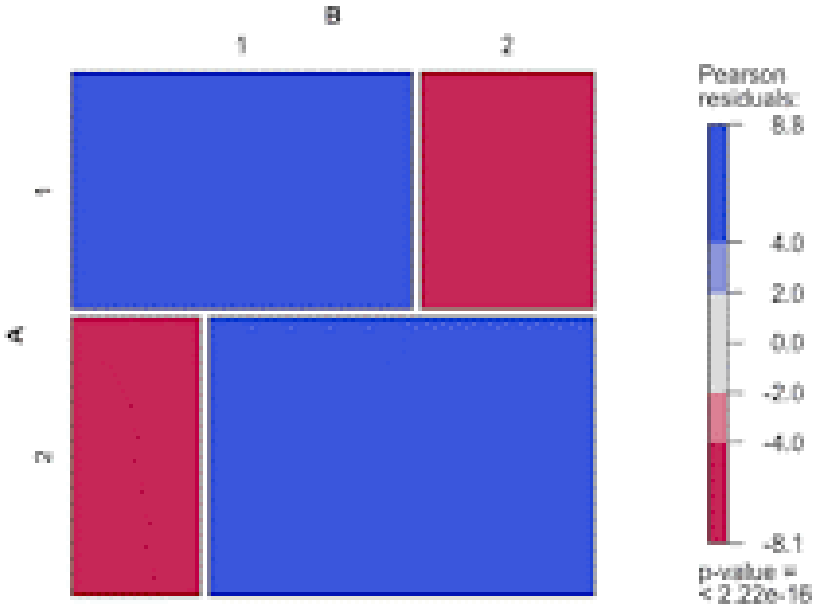

Figure 6: Chi2, variables qualitatives et mosaic plot

5 Modèle linéaire, tests non paramétriques

Nous allons dans un premier temps faire comme si la variable "Nombre de vidéos vues" était normalement distribuée (ce n'est pas le cas), et appliquer d'abord un test de Student, en comparant le nombre de vidéos vues selon les genres. Réalisez le test. Faites la même chose avec un test non-paramétrique. Incorporez les résultats de ce dernier dans le rapport à travers un paragraphe de description, suivi d'un paragraphe d'interprétation.

Utilisez une régression linéaire, avec un test de corrélation de Pearson, puis de Spearman, pour établir le lien entre nombre de quiz réalisés et nombre de vidéos visionnées. Faites un scatterplot pour représenter ce lien. Vous ne présenterez dans le rapport que les résultats du Spearman.

Nous allons maintenant utiliser une ANOVA pour évaluer l'effet, sur le nombre de vidéos vues, de l'HDI et du genre, sans nous intéresser aux interactions entre ces variables dans un premier temps. Vous allez devoir présenter deux

tables d'ANOVA. Dans la première, vous présentez l'ANOVA dans son ensemble, avec les sommes de carrés, le F, comme dans la Figure 6, mais en utilisant un format de table satisfaisant sur le plan esthétique. Dans R, il faut par exemple utiliser la commande mod=lm(y x1+x2) où x1=HDI et x2=Gender, et faire un anova et/ou un summary de mod. Présentez et interprétez les tables, en fournissant les p-values, les valeurs de F avec les ddl correspondant (F(ddl1, ddl2)=...) Expliquez dans un paragraphe à part pourquoi vous obtenez la valeur 1 pour Genre, et la valeur 2 pour HDI, en ce qui concerne les degrés de liberté (ddl, ou df en anglais).

```
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 '.'
```

Figure 7: Table d'ANOVA

Vous devez ensuite présenter les résultats de l'ANOVA avec cette fois-ci les statistiques inférentielles (les estimations des effets associés à une modalité donnée). La Figure 7 vous éclairera à cet égard.

```
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 ***
```

Figure 8: ANOVA et présentation de statistiques inférentielles

Vous devez ensuite introduire dans le modèle un paramètre d'interaction entre le genre et l'HDI dans votre modèle, pour obtenir une table comme dans la Figure 8. Qu'est-il arrivé à l'effet "Femme" sur le nombre de vidéos, comment l'interprétez-vous ?

```
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)
```

Figure 9: Visualiser des interactions entre variables indépendantes

	Df	Sum Sq	Mean Sq	F-Value	P-Value
weld$fgage	2	278.60	139.30	12.74	0
weld$ftime	4	385.53	96.38	8.82	0
weld$fgage : weld$ftime	8	597.07	74.63	6.83	0
Residuals	15	164.00	10.93	-	-

Figure 10: Exemple de table d'ANOVA correctement représentée

6 Régression logistique

6.1 Présenter des odd ratios

Nous allons nous intéresser maintenant à la régression logistique. Nous allons commencer par nous pencher sur une variable booléenne, l'obtention du certificat et/ou la réalisation de l'examen final. Mobiliser les mêmes variables que précédemment (genre et HDI), mais sans prendre en compte l'interaction entre ces variables. Faites une table d'odd-ratios, puis le forest plot correspondant.

Sur R, il vous faudra remplacer le lm par un glm (generalized linear model), en rappelant dans le code quel type de glm vous allez mobiliser (binomial ou Poisson). Pensez à passer les coefficients du modèle à l'exponentielle, le cas échéant. Vous pouvez vous inspirer de la figure ci-dessous. Il est obligatoire de rapporter les p-values sous formes d'astérisque, les intervalles de confiance (non présentés ci-dessous), et la modalité de référence doit apparaître dans la table, comme ci-dessous (avec le label Réf).

	MOOC1 V1	MOOC1 V2	MOOC1 V3
Women		Ref.	
Men	1.11	0.98	0.99
Low management positions		Ref.	
Higher management positions	0.99	0.87	1.66*
Jobseeker	0.98	0.77	1.54
Students	1.25*	0.71	2.33***
Others	0.76	0.73	1.68*
HDI Low		Ref.	
HDI Intermediate	0.85	0.87	0.68
HDI Very High	1.22*	1.57	0.61*

Figure 11: Table d'odds-ratios mettant en évidence les seuls OR

Vous décrirez les résultats du modèle en rapportant dans le texte des odd ratios avec les p-values correspondantes; vous confronterez ces résultats avec ceux des études précédentes sur le nombre de vidéos, et expliquerez dans un paragraphes à part pourquoi un odd-ratio ne correspond pas stricto sensus à un risk relatif (risk ratio) Rappelez dans quelles conditions risk ratios et odd ratios convergent.

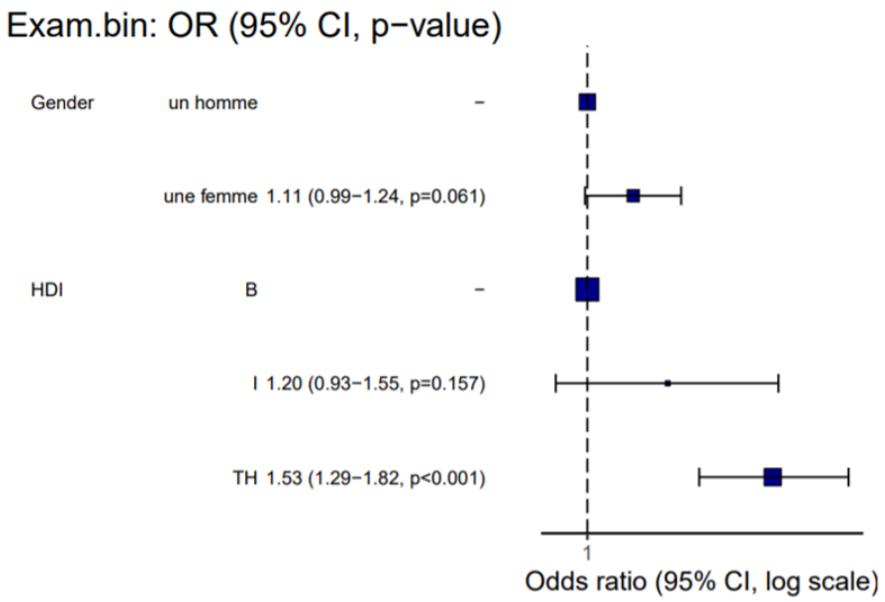

Figure 12: Représentation des odds-ratios via un forest plot

6.2 Données de comptage et loi de Poisson

Nous allons cette fois revenir sur la variable nombre de vidéos vues, mais en mobilisant le modèle correct. Représentez la distribution de la variable, comme ci-dessous. Selon vous, pourquoi la variable ne suit-elle pas tout à fait une loi de Poisson (compte tenu du fait que l'on étudie un MOOC) ?

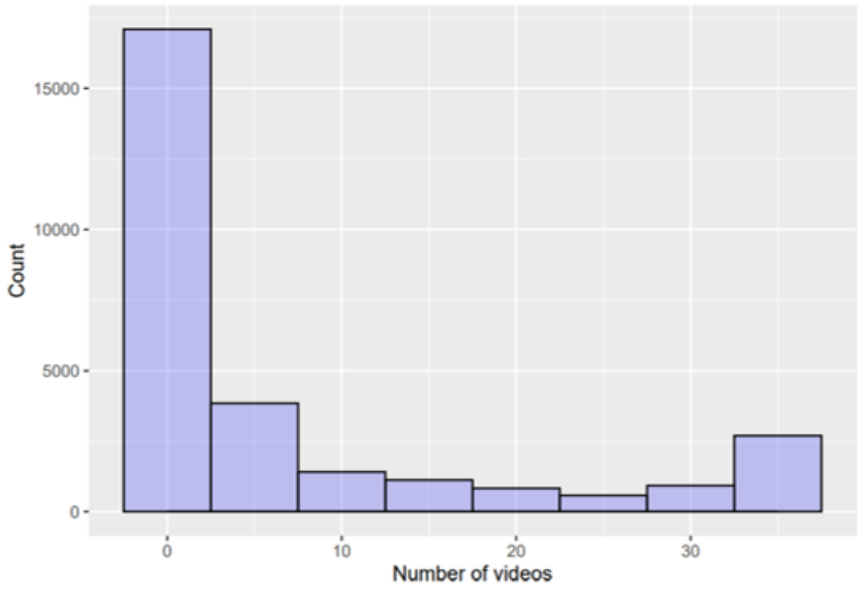

Figure 13: Distribution d'une variable suivant une loi de Poisson

Produisez ensuite les graphes ci-dessous, pour tester la normalité de la variable. Expliquez quelle forme devrait avoir un qqplot, et ce à quoi correspond l'homosedasticité. Quelle forme devrait avoir le scatterplot si la variable était normalement distribuée.

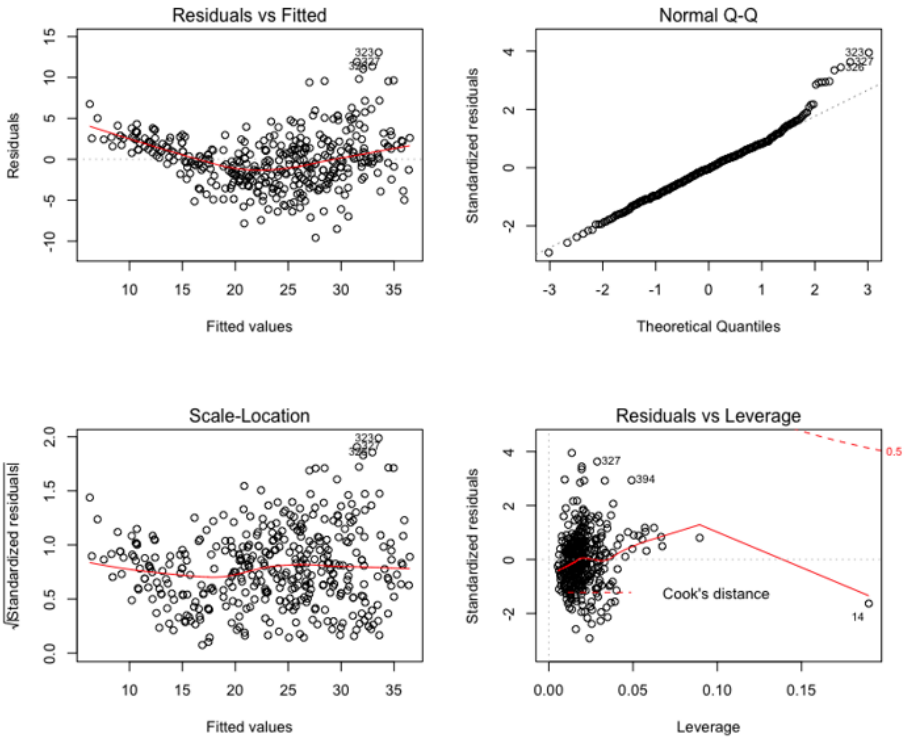

Figure 14: Normalité de la distribution de la variable : quelques graphes

Mobilisez un glm avec une loi de Poisson, en conservant les variables indépendantes mobilisées pour la régression logistique binomiale (Genre et IDH). Décrivez les résultats.