Aix Marseille University

Faculty of Economics

Aix Marseille School of Economics

****

Master 2: Econometrics, Big Data, Statistics

****

Méthodologie des études économétriques et statistiques

Les déterminants de la performance en mathématiques d'un étudiant.

PRESENTE PAR :

Divine TULOMBA

SUPERVISE PAR :

PATRICK SEVESTRE

Année scolaire: 2019-2020

Sommaire

Résumé:

Cette étude cherche à déterminer les caractéristiques spécifiques aux étudiants qui permettrait d'expliquer leur performance en mathématiques. Une étude de statistique descriptive de nos données en coupe transversale auprès de 395 étudiants de deux écoles au Portugal a permis de montrer par exemple qu'il y avait plus des femmes que d'hommes dans notre échantillon. Une analyse de régression logistique à travers la méthode de la sélection pas à pas descendante a été utilisée pour identifier les prédicteurs de cette performance scolaire des étudiants. Les résultats montrent que quatre caractéristiques de ce groupe d'étudiants identifié comme « performant » ont été détecté comme significatifs, à savoir, le niveau d'études de la mère, les heures de travail personnel hebdomadaire, les échecs scolaires antérieures et le soutien pédagogique supplémentaire à la famille. Les échecs scolaires antérieures et le soutien pédagogique supplémentaire en mathématique influencent de façon négative la performance mais Le niveau scolaire de la mère et les heures de travail personnel hebdomadaire influencent positivement la performance de ce groupe d'étudiants dans notre modèle. Finalement, nous avons trouvé que notre modèle était validé à travers le test d'absence de multi colinéarité et test de la validité du modèle.

I. Introduction

L'éducation est le meilleur héritage qu'une nation puisse donner à ses citoyens. En effet, le développement d'une nation ou d'une communauté dépend en grande partie de la qualité de l'éducation de celle-ci. L'éducation aide également les individus à grandir, à se développer, à gagner décemment leur vie dans la société et à contribuer positivement au bien-être de la société.

Chaque pays se distingue grâce à la qualité de son système éducatif et peut produire des individus qualifiés, qui participeront activement à la recherche et au progrès technologique nécessaires à la croissance économique et au maintien de la compétitivité de celui-ci.

Si la qualité du système éducatif de l'enseignement est un enjeu sociétal pour la croissance économique du pays, le développement de la scolarisation et de l'accès à des niveaux d'études toujours plus élevés répond également à une demande sociale celle des individus qui face à la montée du chômage ainsi qu'à la robotisation de la société misent sur l'obtention du diplôme et des compétences pour augmenter leurs chances d'insertion sur le marché du travail.

Nous nous intéressons aux mathématiques car dans la société actuelle, les mathématiques jouent un rôle fondamental dans le développement économique et technologique parce que les problèmes à résoudre dans ces domaines sont écrits en langage mathématique.

Notons d'ailleurs qu'il est bien connu, et depuis longtemps, que les phénomènes de la physique et de la mécanique sont écrits par des formules mathématiques. Il est sûrement moins connu que de nos jours de nombreux problèmes en économie, biologie, santé, communications, énergie, etc. sont aussi écrits par des équations ou des modèles mathématiques. Afin, d'étudier ces problèmes de manière rigoureuse et efficace il faut faire avoir des compétences en mathématiques.

Les mathématiques sont indispensables pour l'étude de problèmes liés au développement des nouvelles technologies et à l'innovation car le design de nouveaux produits industriels est le plus souvent réalisé à l'aide de la modélisation mathématique et de la simulation numérique et non plus avec la réalisation de prototypes bien plus coûteux, ou tout simplement irréalisables.

L'enjeu de la réussite scolaire est donc de taille dans une organisation mondiale régie autour de la croissance économique où la concurrence est internationale.

C'est ainsi que toute cette évolution de la société et les enquêtes récentes menées par le projet d'avis du conseil économique, social et environnemental ont montré l'importance de la performance des étudiants dans les domaines scientifiques.
Nous essayerons de détecter dans cette étude les déterminants de la performance des étudiants en fonction de leur caractéristique personnelle.

Le document sera organisé de la manière suivante : Dans la section 2, nous considérons la revue de la littérature, dans la section 3 nous verrons le modèle économique et de la méthodologie dans la section 4. La section 5 parlera brièvement des procédures de collecte de données, ainsi que de la base des données. La section 6 parlera des statistiques descriptives et des résultats de notre études dans la section 7. Nous aurons les tests enfin de validé les hypothèses du modèle dans la section 8, puis nous aurons la conclusion 9 et l'annexe dans la section 10.

II. Littérature

De nombreux modèles ont été développés pour expliquer la relation entre les caractéristiques d'un individu et ses performances scolaires.

Le modèle développé par Haveman, par ERMISCH ou encore par MURNAME et al. (1981), ERMISCH et FRANCESCONI (2001), BLACK et al. (2005), BLAU (1999) sont quelques-uns des des économistes qui expliquent ce phénomène.

A travers le travail de Haveman, nous avons constaté que le capital humain des parents apparaît comme un facteur fondamental dans l'explication de la réussite scolaire d'un enfant et dans la plupart de ces études cette variable apparaît comme significative.

Par exemple, Haveman et al ont repris plusieurs études en s'appuyant surtout sur le statut socio-économique parental utilisée.

Également, d'autres auteurs ont essayé de déterminer si les deux parents ont le même niveau d'influence dans la réussite scolaire de leur enfant ou si l'un d'entre eux à une plus grande influence. Dans les années 1980-1990 les économistes comme Black ou encore Murname, par exemple, ont essayé d'étudier l'impact du niveau éducationnel des deux parents et si les deux avaient une influence positive dans la réussite de l'étudiant. Dans la plupart d'articles, le capital humain de la mère, défini comme le nombre d'années de scolarité, est plus important pour la réussite académique de l'enfant que celui du père.

Cependant, d'autres caractéristiques des parents ainsi que de la famille ont fait l'objet de nombreuses études chez les économistes des différents pays. Comme par exemple, l'influence de la structure familiale sur la probabilité de la réussite d'un enfant. Vivre dans une famille monoparentale a été définit comme négativement corrélé avec la réussite scolaire et dans la majorité des cas la variable est statistiquement significative.

En 2003, ERMISCH et al ont expliqué l'impact de la structure familiale dans la réussite d'un enfant et ont pu déterminé que le succès de l'enfant était négativement influencé par le fait d'avoir des parents célibataires que par le chômage parental. Dans le même ordre d'idées en 1992, MANSKI et al. ont trouvé que la probabilité de terminer ses études secondaires augmente si l'étudiant vit avec ses deux parents. De plus, en ce qui concerne le sexe de l'étudiant il a été parfois démontré que les garçons avaient de meilleures notes que les filles.

III. Modèle économique

A- Kml

Dans la littérature, on définit la réussite scolaire de 3 manières :

- ➔ La réussite scolaire est traduite par le nombre d'années d'études
- ➔ La réussite scolaire est traduite par la probabilité d'obtenir un diplôme du secondaire
- ➔ La réussite scolaire traduit par les résultats obtenus dans les différents types des tests ou de matière, c'est-à-dire la réussite traduite par la performance de l'étudiant en question.

Nous allons donc nous pencher sur la troisième définition et il est bon de savoir, de comprendre dans notre cas qu'est-ce qu'est la performance :

Pour cela nous allons utiliser l'algorithme KML. Dans les études longitudinales, les variables ne se limitent pas à des mesures uniques mais peuvent être considérées comme des trajectoires variables. KML est un package de R fournissant une implémentation de K-means conçus pour travailler spécifiquement sur des trajectoires. Il donne les conditions de départ en k-moyennes et des critères de qualité pour choisir le meilleur de clusters.

B-Régression logistique

La régression logistique ou modèle logit est un modèle de régression binomiale. Comme pour tous les modèles de régression binomiale, il s'agit de modéliser au mieux un modèle mathématique simple à des observations réelles nombreuses. Y la variable à prédire (variable expliquée) et X=(X1, X2, …, Xj) les variables prédictives ou encore à expliquer. Dans le cadre de la régression logistique binaire, la variable Y prend deux modalités possibles {1,0}. Les variables Xj sont exclusivement continues ou binaires. Nous avons une probabilité telle que : P(Y=1) nous avons ici la probabilité que l'évènement ne se réalise pas, ou que la situation soit comme nous le pensions. P(Y=0) nous avons ici la probabilité que l'évènement ne se réalise pas.

P(Y=1|X), ici nous avons la probabilité que l'évènement se réalise dépendamment des certaines caractéristiques contenues dans X.

IV. Description des variables

La base de données concerne les résultats des élèves e deux écoles portugaises dans l'enseignement secondaire. Ce travail a été construite et présentée par Paulo Cortez dans un papier publié en 2014-2015. Cette base de données comprend 3variables (dont 17 variables explicatives et 16 variables quantitaves) et 395 observations.

Les attributs de données incluent les notes des élèves, les caractéristiques démographiques, sociales et scolaires et ont été collectés à l'aide de rapports et de questionnaires. Cette base nous aidera à apporter une réponse satisfaisante à notre préoccupation principale, celle de déterminer les facteurs qui expliquent la performance d'un étudiant en mathématiques au Portugal. Plus précisément nous nous intéressons aux facteurs micro-économiques que nous jugeons importants aux vues des données disponibles.

Tableau 1 : Description de la base des données

Variable	Nom des variables	les valeurs prises par la variable	Type
Y	note finale	0 à 20	Quantitative
X1	école de l'élève	GP=Gabriel Pareira ou MS=Mousinho da Silveira	Qualitative
X2	sexe de l'étudiant	F = femme et H = homme	Qualitative
X3	Age de l'étudiant	de 15 à 22 ans	Quantitative
X4	type d'adresse du domicile de l'étudiant	U = Urbain et R= Rural	Qualitative
X5	taille de la famille	LE3 = inférieur ou égale à 3 et GT3 supérieur à 3	Qualitative
X6	Statut de cohabitation du parent	T = vivant ensemble ou A = à part	Qualitative
X7	Education de la mère	0= aucun, 1=enseignement primaire, 2 = 5 à 9 année d'étude 3 = etu secondaire et 4= supérieur	Quantitative
X8	Education du père	0= aucun, 1=enseignement primaire, 2 = 5 à 9 année d'étude 3 = etu secondaire et 4= supérieur	Quantitative
X9	Le travail de la mère	Enseignant', 'lié aux soins de santé', 'services', 'à domicile' ou 'autres'	Qualitative
X10	Le travail du père	Enseignant', 'lié aux soins de santé', 'services', 'à domicile' ou 'autres'	Qualitative
X11	Raison sur le choix de l'école	Proche de maison', 'Réputaiton', 'Préférence' ou 'autre'	Qualitative
X12	Tuteur de l'élève	mère', 'père' ou 'autre'	Qualitative
X13	Temps de trajet domicile-école	1 à <15min, 2 à 15 à 30 min, 3 à 30 min à 1 heure ou 4 à>1heure	Quantitative
X14	Heure d'étude hebdomadaire	1-<2 heures, 2-2 à 5 heures, 3-5 à 10 heures ou 4 - > 10 heures	Quantitative
X15	Nombre d'échecs de classe antérieurs	0, 1, 2 3 (fois) sinon 4	Quantitative
X16	Soutien pédagogique supplémataire	Oui ou non	Qualitative
X17	Soutien éducatif à la famille	Oui ou non	Qualitative
X18	Cous supplémentaires en Math	Oui ou non	Qualitative
X19	Activité extra-scolaires	Oui ou non	Qualitative
X20	Fréquenté une crèche	Oui ou non	Qualitative
X21	Veut suivre des études supérieures	Oui ou non	Qualitative
X22	Accèes internet à domicile	Oui ou non	Qualitative
X23	Dans une relation amoureue	Oui ou non	Qualitative
X24	Qualité des relations familiales	De 1 = très mauvais à 5 = excellent	Quantitative
X25	Temps libre après l'école	De 1 = très faible à 5 = élevé	Quantitative
X26	Sortie avec des amis	De 1 = très faible à 5 = élevé	Quantitative
X27	Consommation d'alcool au travail	De 1 = très faible à 5 = élevé	Quantitative
X28	Consommation d'alcool le Week-End	De 1 = très faible à 5 = élevé	Quantitative
X29	Etat de santé actuel	De 1 =très mauvais à 5 = très bon	Quantitative
X30	Nombre d'abscences à l'école	De 0 à 93	Quantitative
X31	Grade de première période	De 0 à 20	Quantitative
X32	Grade de la deuxième période	De 0 à 20	Quantitative

Source : *réalisé sous Excel avec les auteurs du travail*

V. Statistiques descriptives

Les principales statistiques décrivant les grandes tendances des variables sont regroupés dans le tableau suivant :

Tableau 2 : Statistiques descriptives des variables

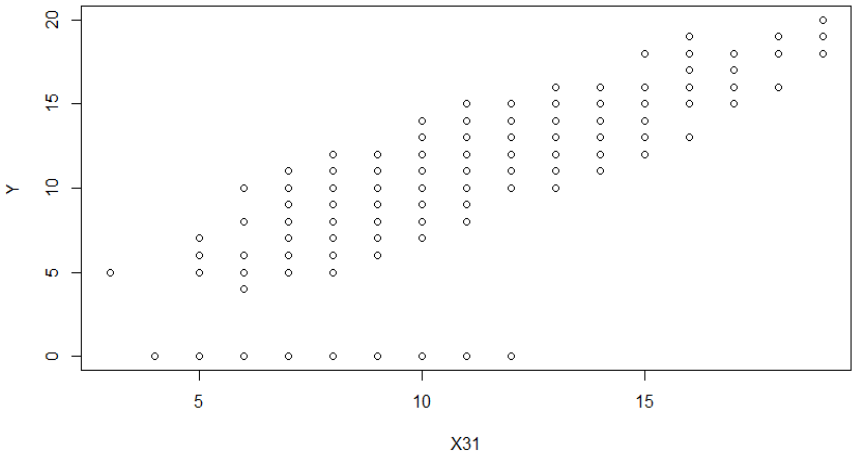
	Obs	Ecart-type	Médiane	Moyenne	Min	Maximum
Y	395	4,58	11	10,42	0	20
X1	395	0,32	1	0,8835	0	1
X2	395	0,49	1	0,5266	0	1
X3	395	1,27	17	16,7	15	22
X7	395	1,09	3	2,749	0	4
X8	395	1,08	2	2,522	0	4
X15	395	0.74	0	0.3342	0	3
X18	395	0.5	0	0.4582	0	1
X23	395	0.47	0	0.3342	0	1
X24	395	0.89	4	3.944	1	5
X29	395	1.39	4	3.554	1	5
X30	395	8	4	5.709	0	75
X31	395	3.32	11	10.91	3	19
X32	395	3.76	11	10.71	0	19

Source : *Réalisé par les auteurs du travail avec le logiciel R*

Dans cet échantillon, nous voyons que la moyenne des notes finales est de 10.42, mais il dispose d'un écart-type assez élevé, donc la distribution est assez dispersée, c'est-à-dire que les valeurs ne sont pas concentrées autour de la moyenne. Nous pouvons également observer que la moyenne d'âge est d'environ 16 ans et demie, les âges vont de 15 ans à 22 ans. Dans notre échantillon, il y a plus des femmes que d'hommes, les femmes constituent environ 52% de l'échantillon.

Nous avons ensuite regardé la corrélation entre les notes de la dernière période et ceux de la première à travers le graphique suivant.

Graphique 1 : la corrélation entre les notes de la 1er période et de la 3ème période

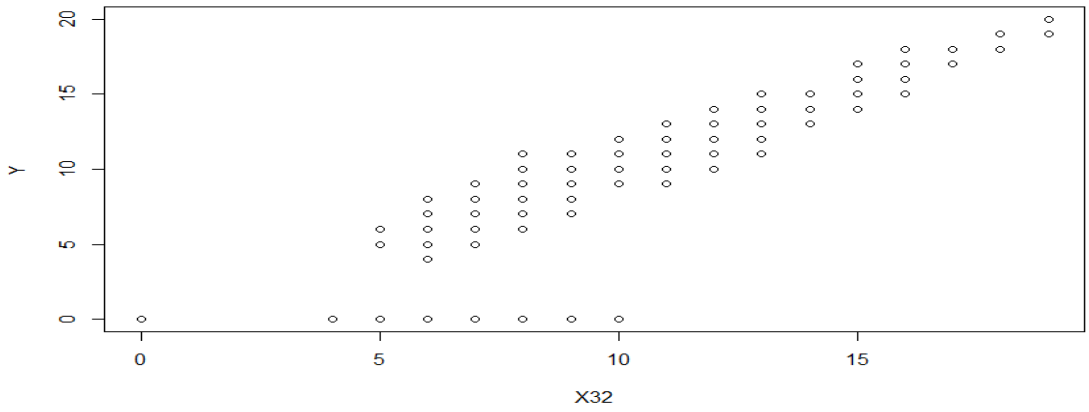


Source : *Réalisé par les auteurs du travail avec le logiciel R*

Nous voyons ici qu'il y a une corrélation positive entre les notes de la dernière période et ceux de la deuxième période. Mais aussi, lorsque nous avons des mauvaises notes à la période une, nous avons aussi des mauvaises notes à la dernière période.

Le graphique suivant : lui représente le lien entre les notes de la dernière période et les notes de la deuxième période. Comme pour la période précédente, nous voyons qu'il y a une relation entre le fait d'avoir une bonne note à la dernière période et une bonne note à la deuxième période.

Graphique 2 : la corrélation entre les notes de la 2ème période et de la 3ème période



Source : *Réalisé par les auteurs du travail avec le logiciel R*

Par exemple, lorsque les individus ont une note supérieure à 15 à la période deux, ils ont aussi une note supérieure à 15 à la période trois.

Tableau 3 : Matrice de corrélation

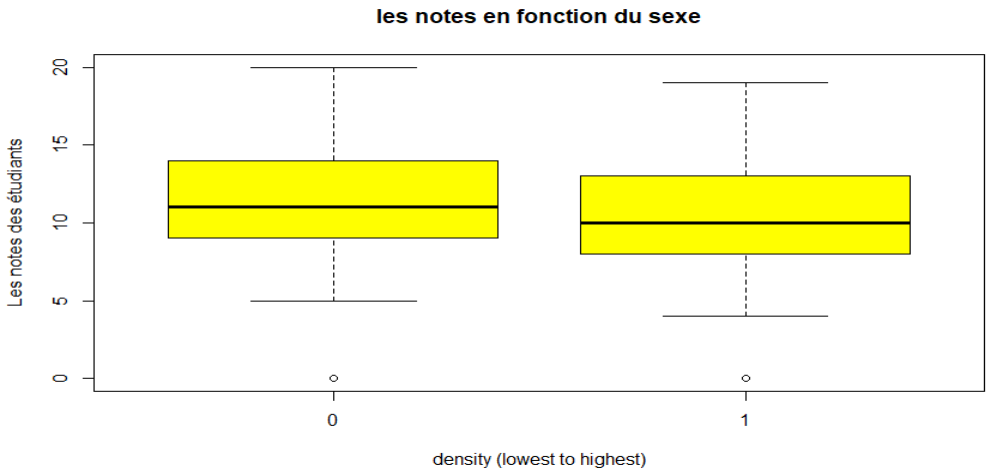
Matrice de corrélation de 3 notes			
	X31	**X32**	**Y**
X31	1.0000000	0.8521181	0.8014679
X32	0.8521181	1.0000000	0.9048680
Y	0.8014679	0.9048680	1.0000000

Source : *Réalisé par les auteurs du travail avec le logiciel R*

Nous pouvons constater que les notes de la période une sont fortement corrélés avec ceux de la période trois, il y a cependant une plus grande corrélation entre les notes de la deuxième période et de la troisième période.

Nous avons fait quelques boxplots, enfin de mieux expliqué la relation entre les notes de la dernière période et certaines variables explicatives.

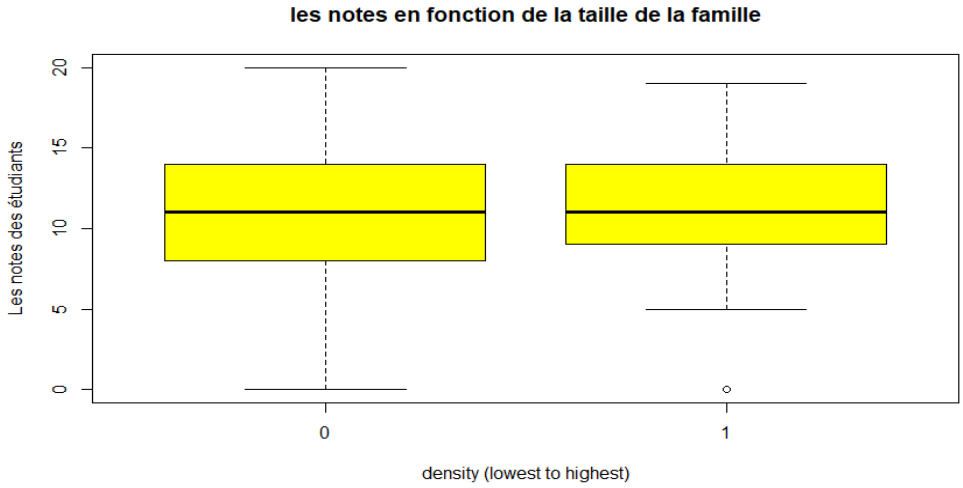
Graphique 3 : Boxplot



Source : *Réalisé par les auteurs du travail avec le logiciel R*

Ici les hommes sont représentés par zéro et les femmes par le chiffre 1, nous pouvons dire de ce boxplot que les hommes ont de meilleures notes en mathématiques que les femmes à la dernière période.

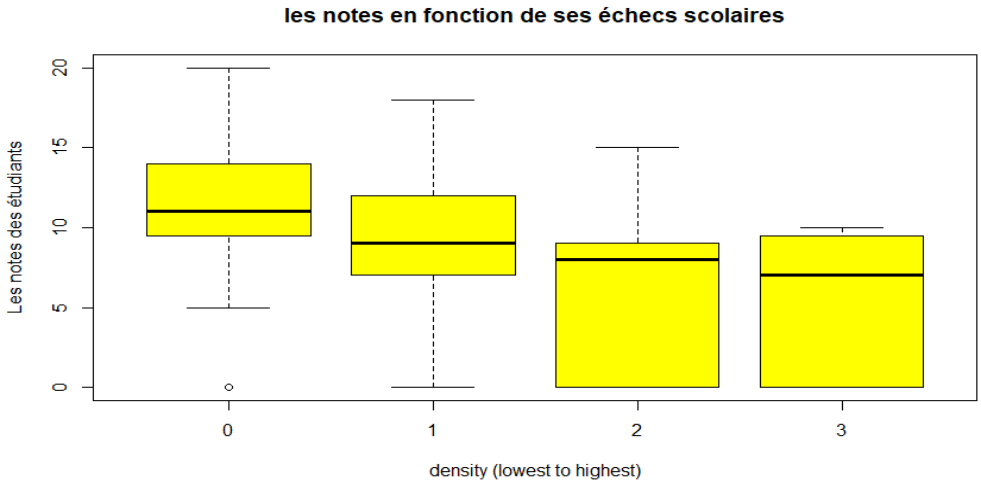
Graphique 4 : Boxplot



Source : *Réalisé par les auteurs du travail avec le logiciel R*

Les notes ont tendance à diminuer lorsque la taille de la famille est supérieure à 3 personnes, alors que les enfants ayant une famille de plus de 3 personnes ont des notes légèrement plus faibles.

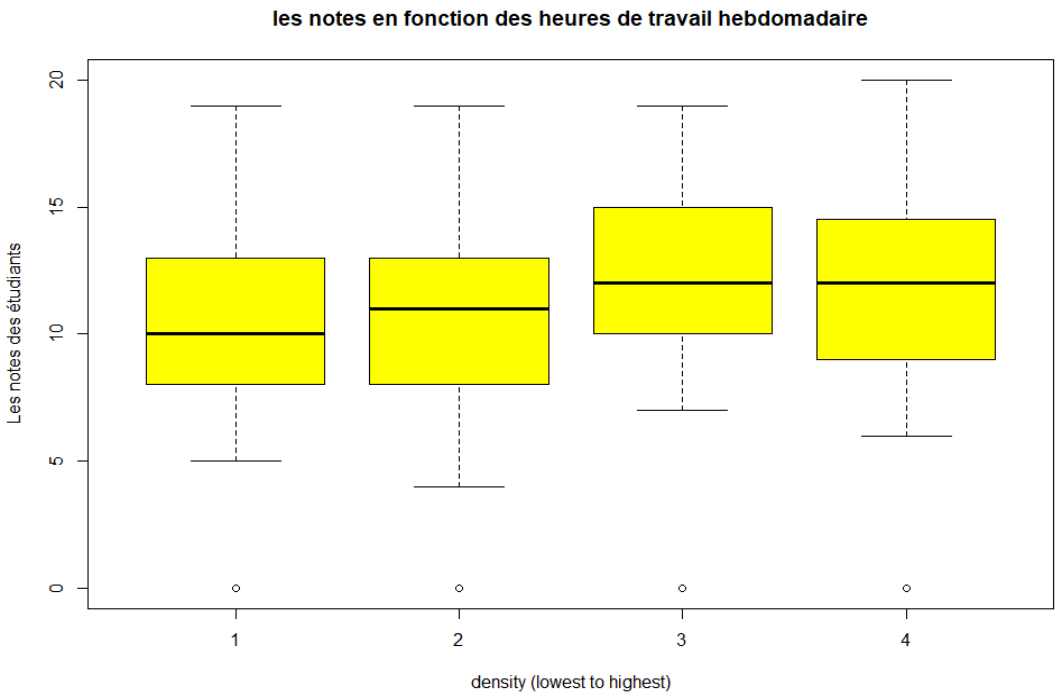
Graphique 5 : Boxplot



Source : *Réalisé par les auteurs du travail avec le logiciel R*

Les notes diminuent fortement à la dernière période lorsque l'élève a redoublé la même classe plus de deux fois.

Graphique 6 : Boxplot



Source : *Réalisé par les auteurs du travail avec le logiciel R*

Ce dernier graphique, nous montre que plus les heures de travail hebdomadaire augmentent, les notes de la dernière période augmentent aussi.

VI. Méthodologie

A- Kml

Nous créerons une nouvelle base des données qui ne contiendra que les trois notes, de 3 périodes respectivement, ensuite nous allons faire exécuter notre algorithme pour créer ce différent groupe selon les notes allant du plus performant au moins performant.

Alors la performance d'un étudiant : traduit dans cette étude le faite pour un élève d'avoir de note excellent ou « élevé » durant les 3 périodes. Il ne s'agit pas alors d'avoir un 15 en période 1 et un 0 en Période 3. Nous cherchons à ce qu'il y ait une robustesse dans nos résultats, nous cherchons à ce que nos résultats soient bons et constants dans le temps.

donc pour tenter d'expliquer la performance nous utilisons cet algorithme dont le but ici est de définir le groupe le plus performant, car par exemple un élève a des notes (10, 12,11) pour les 3 périodes respectivement et un autre qui a (15,16,15), nous n'avons pas le même profil des élèves ni la même performance, et celui qui a environ 15 dans les trois périodes est beaucoup plus « performant » que celui qui a environ 10 ou 11 pendant les trois périodes.

Cette avec l'idée suivante que nous essayerons tout d'abord de déterminer la performance pour ensuite faire une régression logistique linéaire.

Une fois la classification de mes groupes faite, nous créons la variable « trajectoire » dans laquelle nous mettrons toute nos classes, aussi, nous créerons la variable dummy « traj » contenant le groupe des élèves ayant les notes les plus élevés.

B-Régression logit

Un modèle logit binaire a été envisagé pour l'estimation de la performance scolaire de chaque élève. Nous supposons que les résultats scolaires de chaque élève sont dichotomiques, élevés ou faibles. Supposons que la variable binaire Y= (0,1) dénote la performance scolaire de chaque élève.

Soit y=1, si et seulement si l'élève obtient de bons résultats et y=0 dans le cas contraire, c'est-à-dire qu'il fait de mauvais résultats.

La probabilité de résultats scolaires élevés de l'élève est exprimée comme suit :

(1)

$$\text{Pi} = (\{Y = 1|Xi\}) = \frac{\exp{(Xi\beta)}}{1 + \exp{(Xi\beta)}}$$

$$\text{Où } X_i\beta = \beta_i + \beta_2 X_2 \ldots + \beta_k X_k$$

L'équation **(1)** indique que la probabilité de ne pas avoir un meilleur rendement scolaire peut être exprimée comme suit :

(2)

$$1 - p = \frac{1}{1 + \exp{(X_i\beta)}}$$

On peut donc écrire :

(3)

$$\frac{Pi}{1 - \text{Pi}} = \exp{(X_i\beta)}$$

En prenant le logarithme naturel de l'équation **(3),** nous obtenons la fonction logistique

(4)

$$\boldsymbol{L_i = ln\left(\frac{P_i}{1 - P_i}\right) =\propto \beta \qquad X_i + \epsilon_i}$$

Ou ln = logarithmes naturels ; Pi=probabilité d'atteindre des performances académiques élevées, définies en termes de la fonction de probabilité logistique cumulative ; (1-Pi) = probabilité d'obtenir de faibles résultats scolaires ; Xi= vecteur de variables explicatives ; Ɛ= terme de perturbation aléatoire ; [α, β] sont l'ordonnée à l'origine et la pente des paramètres à estimer.

Nos variables explicatives sont composées des variables suivantes :

X2 = le sexe de l'individu, le fait qu'on soit une femme ou un homme.

X4 = Le milieu sociale peut être un élément dans l'explication de la performance, vivre dans un milieu urbain ou rurale peut affecter la performance, selon certains économistes.

X5 = La structure de la famille a été un élément clé dans l'explication de la réussite ou performance, c'est pour cette raison que nous avons décidé d'introduire la variable « taille de la famille », le fait de vivre dans une famille d'une taille inférieur ou supérieur à 3 personnes.

X6 = En 1996, l'économiste Manski a trouvé que la probabilité de terminer ses études secondaires augmente si l'étudiant vit avec ses deux parents et lors d'une étude sur la structure familiale et du chômage au Royaume Uni que le fait de vivre dans une famille monoparentale avait plus d'influence sur la scolarité de l'enfant que le chômage des parents.

X7 & X8 = Le niveau d'études de la mère aurait beaucoup plus d'impact que celui du père selon des nombreuses littératures, c'est la raison pour laquelle, nous avons mis les deux variables concernant le niveau d'éducation de la mère et du père enfin de voir les effets de chacune.

X12 = Dans la même idée que pour les variables précédentes, nous allons essayer de voir si la faite de vivre avec « le père » ou la « mère » peut influencer le niveau de l'éducation de l'élève.

X24 = Le temps passé avec les enfants favorise leur réussite scolaire. Ici nous allons inclure cette variable comme étant « bonne relation » ou « pas bonne relation », et passé plus de temps avec les enfants serait pour nous un signe pour dire la relation est bonne et contrairement passé moins de temps pour nous serait signe que la relation n'est pas bonne.

Les signes attendus des coefficients sont :

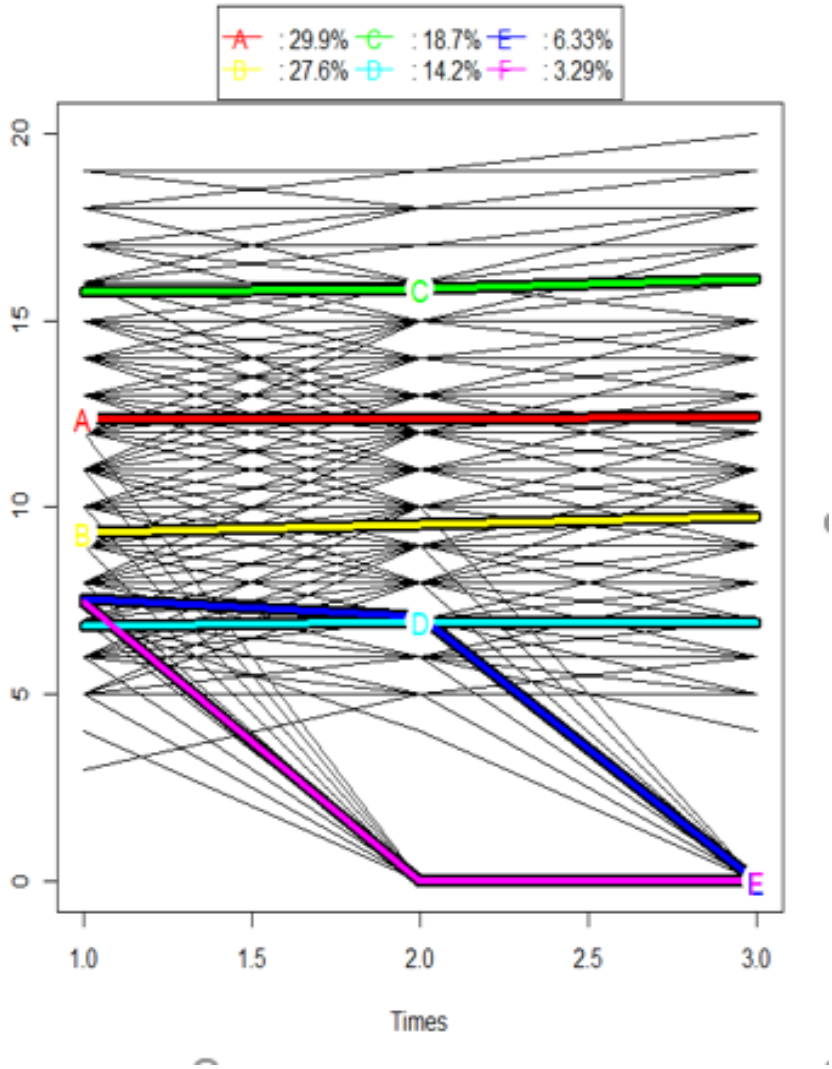
$$\boldsymbol{\beta}1 > 0, \boldsymbol{\beta}2 > 0, \boldsymbol{\beta}3 > 0, \boldsymbol{\beta}4 > 0, \boldsymbol{\beta}5 < 0, \boldsymbol{\beta}6 > 0, \boldsymbol{\beta}7 < 0, \boldsymbol{\beta}8 > 0$$

VII. Les résultats

A- Le résultat avec le package KML

Ce package a généré différents groupes et parmi ces classes, il y a un groupe que nous avons détecté comme plus « performant » que les autres et celle-ci est la classe « C ». Cette classe est particulièrement intéressante car ce groupe d'étudiants ont une moyenne des notes qui est supérieure à 15 durant les 3 périodes simultanément.

Graphique 7 : KML algorithme



Source : *Réalisé par les auteurs du travail avec le logiciel R*

Nous avons classifié nos notes selon les différentes périodes et nous avons obtenu 6 différents groupes qui se dessinent mais nous voyons 5 profils différents car les groupes E en bleu et le groupe F en violet ont à peu près le même profil, alors que les restes des groupes affichent un profil différent.

Ces résultats nous révèlent la pertinence de notre étude car il y a une robustesse dans notre performance, c'est la raison pour laquelle nous avons gardé les trois périodes car la performance pour nous une question de la durée et non d'un instant t.

Les traits noirs que nous voyons sur le graphique c'est la trajectoire des notes pour l'ensemble de 395 individus que nous avons dans notre base des données.
Nous avons créé une variable appelé « trajectoire », trajectoire prends les groupes A, B, C, D, E, F.

A partir de cette variable, nous allons essayer crée notre nouvelle variable en le mettant en variable dummy que nous appellerons ***traj***, la variable ***traj*** sera égale à 1 lorsque le groupe sera performant, nous mettons les autres groupes en muet (ils prendront la valeur 0). Ainsi nous regarderons l'impact des variables explicatives sur le groupe appelé « performant à travers les modèles basés sur la régression logistique.

B- Modèle 1 : Modèle logit

Tableau 4 : Résultat de la régression logistique

Modèle 1 : REGRESSION LOGISTIQUE					
Deviance	**Residuals:**				
Min	**1Q**	**Median**	**3Q**	**Max**	
-0.9809	-0.7091	-0.5434	-0.3818	2.3414	
Coefficients:					
	Estimate	**Std.Error**	**z value**	**Pr(>|z|)**	
(Intercept)	-3.83523	1.09396	-3.506	0.000455	***
X2	-0.38743	0.26899	-1.440	0.149771	
X4	0.61311	0.37360	1.641	0.100778	
X5	0.16700	0.29377	0.568	0.569729	
X6	-0.04344	0.42615	-0.102	0.918815	
X7	0.38209	0.16891	2.262	0.023694	*
X8	-0.01511	0.15799	-0.096	0.923809	
X12mere	0.72610	0.63818	1.138	0.255221	
X12pere	0.59695	0.68265	0.874	0.381869	
X24	0.08159	0.14793	0.552	0.581262	
signif codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1**					
(Dispersion parameter for binomial family taken to be 1)					
Null deviance : 381.05 on 394 degrees of freedom					
Residual deviance : 361,59 on 385 degrees of freedom					
AIC : 381,59					
Number of Fisher scoring iterations : 5					

Source : *Réalisé par les auteurs du travail avec le logiciel R*

C-Modèle 2 : Modèle logit en plus des variables de contrôles

Tableau 5 : Résultat de la régression logistique

Modèle 2 : REGRESSION LOGIT AVEC VARIABLES DES CONTRÔLES					
Deviance	**Residuals:**				
Min	**1Q**	**Median**	**3Q**	**Max**	
-1.6908	-0.6887	-0.3878	-0.1105	2.8376	
Coefficients:					
	Estimate	**Std.Error**	**z value**	**Pr(>|z|)**	
(Intercept)	1.24170	2.79900	0.444	0.65732	
X2	-0.47399	0.31060	-1.526	0.12700	
X3	-0.14198	0.13766	-1.031	0.30237	
X4	0.36732	0.42823	0.858	0.39101	
X5	0.19522	0.31308	0.624	0.53293	
X6	-0.37702	0.47709	-0.790	0.42937	
X7	0.27819	0.18217	1.527	0.01267	*
X8	0.03733	0.17476	0.214	0.83086	
X12mere	-0.17057	0.73833	-0.231	0.81730	
X12pere	-0.18038	0.77688	-0.232	0.81639	
X13	-0.27560	0.26498	-1.040	0.29830	
X14	0.41980	0.17891	2.346	0.01896	*
X15	-1.25493	0.56685	-2.214	0.02684	*
X16	-2.82552	1.04749	-2.697	0.00699	**
X17	-0.47928	0.32474	-1.476	0.13998	
X18	-0.13830	0.30344	-0.456	0.64855	
X23	-0.23148	0.32429	-0.714	0.47535	
X24	0.04007	0.16005	0.250	0.80233	
X29	-0.16242	0.10370	-1.566	0.11728	
X30	-0.04407	0.02934	-1.502	0.13312	
signif codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1**					
(Dispersion parameter for binomial family taken to be 1)					
Null deviance : 381.05 on 394 degrees of freedom					
Residual deviance : 313.10 on 375 degrees of freedom					
AIC : 353.1					
Number of Fisher scoring iterations : 6					

Source : *Réalisé par les auteurs du travail avec le logiciel R*

Les résidus de déviance montrent la distribution des résidus pour les cas individuels utilisés dans le modèle. Notre régression nous montre les coefficients, leurs erreurs standard, la statistique z (parfois appelée statistique z de Wald) et les valeurs P associées.

La déviance nulle et la déviance résiduelle sont utilisées pour tester si les variables indépendantes fournissent une explication statistiquement significative. Un test du chi carré, utilisant la différence entre les deux résidus, indique la signification globale du modèle. Le critère d'information d'Akaike (AIC) est de -2 * log-vraisemblance + 2 * k où k est le nombre de paramètres estimés est réalisé et permet de comparer les deux modèles. Si l'on considère un ensemble de modèles candidats, le modèle choisi est celui qui aura la plus faible valeur d'AIC.

Le fisher scoring indique le nombre optimal d'itérations. Par exemple, au-delà d'un certain nombre d'itérations, il n'y a aucun gain pratique. Vous pouvez penser que cela est analogue à la détermination du nombre maximal de nœuds dans les arbres de décision.

Pour le premier modèle, nous pouvons voir qu'il y a qu'une seule variable significative, le niveau d'études de la mère (X7) est significatif dans l'interprétation de notre modèle. C'est-à-dire que le niveau de l'étude de la mère influence positivement la probabilité pour l'élève d'être performant à l'école. Alors que celui du père influence négativement la probabilité d'être performant pour l'élève. Le fait d'être une femme (X2) diminue la probabilité d'être performant en mathématiques. Pour le milieu sociale (X3), le signe est positif c'est-à-dire que la faite de vivre dans un milieu urbain, augmente la probabilité d'être performant en mathématiques Nous voyons aussi que la probabilité d'être performant pour les élèves augmentent lorsque la taille de la famille est inférieure ou égale à trois personnes (X5) et lorsque les parents vivent ensemble (X6). De plus, une bonne relation familiale augmente la probabilité d'être performant à l'école de 0.08.

Dans le modèle 2, Nous avons ici plus des variables explicatives que dans le premier modèle. Le niveau d'éducation de la mère (X7), l'heure d'étude hebdomadaire (14), nombre d'échecs de classe antérieurs (15) ainsi que le soutien pédagogique supplémentaire (16) sont significatifs pour notre modèle.
Ce qui est surprenant pour nous, nous trouvons que le soutien pédagogique supplémentaire n'augmente pas forcément la probabilité d'être performant. Cependant, nous pouvons voir que les échecs à l'école diminuent la probabilité d'être performant pour les élèves. L'augmentation d'heures de travail personnel augmente la probabilité d'être performant à l'école de 0.41980.

Si votre Null Deviance est vraiment petit, cela signifie que le modèle Null explique assez bien les données. De même avec votre déviance résiduelle.

Nous avons ensuite choisi de faire une sélection pas à pas des variables des plus explicatives dans notre modèle à partir du modèle 2 et de créer un nouveau modèle à partir de cela qui peut potentiellement améliorer notre modèle.
Ainsi de voir si une sélection aléatoire des variables le plus pertinent pour notre modèle permettrait d'améliorer notre modèle.

D-Sélection pas à pas descendante

La sélection pas à pas descendante est le test de suppression se basant sur la probabilité du rapport de vraisemblance calculé à partir d'estimations de paramètres conditionnels.

La sélection pas à pas descendante se base sur la probabilité de la statistique de Wald et supprime au fur et à mesure les variables les moins explicatives dans le modèle.

Tableau 6 : Résultat de la régression logistique

Modèle 3 AVEC LA SELECTION PAS A PAS					
Deviance	**Residuals:**				
Min	**1Q Median**	**3Q**	**Max**		
-1.5801	-0.7194 -0.39	-0.1062	2.7976		
Coefficients	:				
	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**	
constante	-0.93950	0.79427	-1.183	0.2369	
X2	-0.55055	0.30285	-1.818	0.0691	.
X7	0.32465	0.14332	2.265	0.0235	*
X13	-0.37887	0.24344	-1.556	0.1196	
X14	0.35808	0.17091	2.095	0.0362	*
X15	-1.35363	0.55582	-2.435	0.0149	*
X16	-2.56947	1.02853	-2.498	0.0125	*
X17	-0.46099	0.30065	-1.533	0.1252	
X29	-0.16743	0.09968	-1.680	0.0930	.
X30	-0.04802	0.02807	-1.711	0.0872	.
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
(Dispersion parameter for binomial family taken to be 1)					
Null deviance: 381.05 on 394 degrees of freedom					
Residual deviance: 317.69 on 385 degrees of freedom					
AIC: 337.69					
Number of Fisher Scoring iterations: 6					

Source : *Réalisé par les auteurs du travail avec le logiciel R*

La sélection pas à pas descendante nous subjère un nouveau modèle avec les variables expliquant le mieux le modèle. Nous préférons l'utilisation aux deux modèles précédents à cause par exemple de son AIC qui est assez faible.

Nous pouvons dire ici que le fait d'être une femme **(X2)** diminue la probabilité d'être performant en mathématiques de (-0.55). L'absence d'un soutien éducatif **(X17),** un mauvais état de santé **(X29)** mais aussi un nombre d'absences élevés à l'école **(X30)** détériorent la probabilité d'être performant à l'école.

Ce nouveau modèle explique mieux notre variable « traj » qui identifie le groupe le plus performant.

Nous choisissons de conserver ce modèle et de poursuivre avec lui pour le reste de notre étude.

Par la suite nous avons calculé les odds ratio du modèle 3.

Dans le cadre d'un modèle logistique, généralement on ne présente pas les coefficients du modèle mais leur valeur exponentielle, cette dernière correspondant en effet à des *odds ratio*, également appelés rapports des côtes. Les odds ratio se définissent comme le rapport de la cote d'un événement arrivant au groupe le plus performant, avec celle du même événement arrivant à un groupe B d'individus peu performant.

L'odds ratio est proche du risque relatif lorsque le nombre d'événements est faible.

Tableau 7 : Caclucl des odds ratios

Les odds Ratio du modèle 3			
	Odds Ratio	**p**	
constante	0.3908243	0.23687	
X2	0.5766326	0.06908	.
X7	1.3835432	0.02350	*
X13	0.6846314	0.11963	
X14	1.4305751	0.03616	*
X15	0.2583012	0.01488	*
X16	0.0765763	0.01248	*
X17	0.6306586	0.12520	
X29	0.8458319	0.09300	.
X30	0.9531124	0.08717	.

Source : *Réalisé par les auteurs du travail avec le logiciel R*

Un odds ratio de 1 correspond à l'absence d'effet. En cas d'effet négatif du phénomène étudié, l'odds ratio est inférieur à 1 et il est supérieur à 1 en cas d'effet positif sur le phénomène étudié. Plus l'odds ratio est éloigné de 1, plus l'effet est important. A travers les odds ratio, nous pouvons dire ici que les variables X7, X14, X15 et X16 sont admis statistiquement significatifs pour le modèle comme pour le modèle précédent. Cependant comme nous pouvons le voir seuls le niveau d'études de la mère ainsi que le nombres d'heures de travail personnel **(X14)** ont un effet positif sur la performance. Le nombre d'échecs de classe antérieures **(X15)** et le soutien pédagogique ont une mauvaise influence sur la performance **(X16).**

VIII. Les tests de pour la validation du modèle

A- Le test de multi-colinéarité

La multi colinéarité est un problème qui survient lorsque certaines variables de prévision du modèle mesurent le même phénomène. Une multi colinéarité prononcée s'avère problématique, car elle peut augmenter la variance des coefficients de régression et rendre ainsi l'interprétation du modèle.

Nous parlons de multi-colinéarité parfaite lorsqu'une des variables explicatives d'un modèle est une combinaison linéaire d'une ou plusieurs autres variables explicatives introduites dans le même modèle. L'absence de multi colinéarité parfaite est une des conditions requises pour pouvoir estimer un modèle linéaire et par extension un modèle linéaire généralisé dont les modèles de régression logistique.
Pour mesurer cette colinéarité, l'approche la plus classique consiste à examiner les facteurs d'inflation de la variance (FIV) ou variance inflation factor (VIF) en anglais. Les FIV estiment de combien la variance d'un coefficient est « augmentée » en raison d'une relation linéaire avec d'autres prédicteurs.

Si tous les FIV sont égales à 1, il n'existe pas de multi-colinéarité, mais si certains FIV sont supérieures à 1, les prédicteurs sont corrélés. Cependant certains économistes disent qu'il faut s'inquiéter lorsque le vif est supérieur à 2.5 et d'autres à 5. Il n'y a pas vraiment de valeur au-delà de laquelle on doit considérer qu'il y a multi colinéarité.

Tableau 7 : Test de multi colinéarité

Test de multi-colinéarité	
Coéficients	**Vif**
X7	1,006
X13	1,01
X14	1,02
X15	1,019
X29	1,025
X30	1,035

Source : *Réalisé par les auteurs du travail avec le logiciel R*

Tous nos FIV sont proches de 1, il n y a donc pas de problème de colinéarité à explorer.

B-Le test d'Hosmer-Lemeshow

Pour tester la validité du modèle, Hosmer et Lemeshow ont proposé un test du khi-deux de validité du modèle sur un tableau établi à partir des probabilités prédites.
Les hypothèses sont les suivantes :

- **H0 :** La validité du modèle logit
- **H1 :** Rejet du modèle logit

La statistique du test d'Hosner-Lemeshow :

$$G_{HL}^2 = \sum_{j=1}^{10} \frac{(O_j - E_j)^2}{E_j(1 - E_j/n_j)} \sim \chi_8^2$$

Pour évaluer la validité du modèle de régression logistique, on doit vérifier si la p-valeur associée qui utilise la loi du Chi-deux est supérieur à 0.05. Si la p valeur > 0.05, on admet que le modèle est bien adapté aux données et est bien spécifié.

Tableau 8 : Test de la validité du modèle

Test de validité du modèle
Hosmer and Lemeshow goodness of fit (GOF) test
X-squared = 14.16, df = 8, p-value = 0.07769

Source : *Réalisé par les auteurs du travail avec le logiciel R*

Nous avons effectué le test pour le modèle. Le test de Hosmer-Lemeshow a donné une valeur de p qui est égale à environ 0.07 donc mon modèle est bien spécifié et validé.

IX. Conclusion

L'objectif de cette étude était de déterminer les caractéristiques spécifiques aux étudiants qui expliquerait leur performance. Un plan d'étude transversal a été utilisé et le domaine d'étude était les mathématiques. La population cible de cette étude était constituée d'étudiants inscrits qui suivaient des cours dans deux lycées différents au Portugal. Le nombre total d'étudiants utilisé pour cette étude était de 395 individus dont environ 52% des femmes.

Un modèle logit binaire a été utilisé pour estimer les déterminants du groupe le plus performant. A partir du modèle logit binaire, les résultats du premier modèle nous montrent que seul le niveau de l'éducation de la mère **(X7)** pouvait expliquer la performance des élèves.
Ensuite nous avons inséré à ce modèle des variables des contrôles qui avait pour but d'améliorer le modèle. Ces variables des contrôles sont des variables que nous avons jugés importants pour l'explication de la performance. Nous avons trouvé alors trois autres variables explicatives, l'heure d'étude hebdomadaire, les échecs scolaires antérieures mais aussi le soutien pédagogique supplémentaire en mathématiques. A travers la méthode de sélection pas à pas descendante nous avons sélectionné les variables le plus explicatives à partir du modèle 2 et nous avons créé un nouveau modèle avec ces variables explicatives. Nous avons préféré le modèle 3 crée par la sélection pas à pas au modèle 1 &2 puisque l'AIC était plus faible pour celui-ci.
Nos variables explicatives significatifs sont du signe attendu sauf le soutien pédagogique qui n'est pas du signe attendu. Enfin, les résultats de la régression logistique binaire ont révélé que le niveau de l'éducation de la mère, les échecs scolaires antérieures, les heures de travail hebdomadaire et le soutien pédagogique supplémentaire sont les caractéristiques individuelles qui expliquent le mieux la performance de ce groupe d'étudiants qualifié comme performant.

Nous avons finalement fait un test pour tester la multi colinéarité des variables, nous avons trouvé qu'il n y avait pas de problème de colinéarité. Ensuite, nous avons fait un test de Hosmer et Lemeshow enfin de voir si notre modèle était bien spécifié, la p valeur s'est révelé être non significatif, c'est à dire que le modèle logit est validé.

X. ANNEXE

- Cours de machine learning
- Cours de méthodologie
- An econometric assessment of factors that predict academic performance of tertiary students in H0, GHANA (James Dickson Fiagborlo, Ho Polytechnic Etornam K. Kunu, Ho Polytechnic)
- Econometry , edition 16 decembre 2011, William Greene
- https://journals.openedition.org/pyramides/250#tocto1n2
- https://archive.ics.uci.edu/ml/datasets/student+performance