Projet économétrie

GUICHARD Allan

21/03/2021

I - Introduction

Dans le but de proposer sur le site de l'université un système automatisé d'évaluation à destination des étudiants souhaitant intégrer la formation économique, il est nécessaire d'étudier les résultats des examens d'entrée précédents. Ce programme devrait permettre à l'étudiant qui est intéressé par le parcours économique, d'avoir un aperçu de sa probabilité de réussite en fonction de son profil. C'est pour cela que nous devons explorer un maximum de possibilités de modélisation de la solution, afin que la probabilité retournée à l'utilisateur soit la plus proche possible de la réalité.

II - Etude du problème

1. Statistiques descriptives

On commence par détailler les taux de réussite des candidats en fonction de leurs profils.

Nous souhaitons avoir une idée du taux de réussite selon la série de Bac obtenue pour les candidats sur les 6 dernières années.

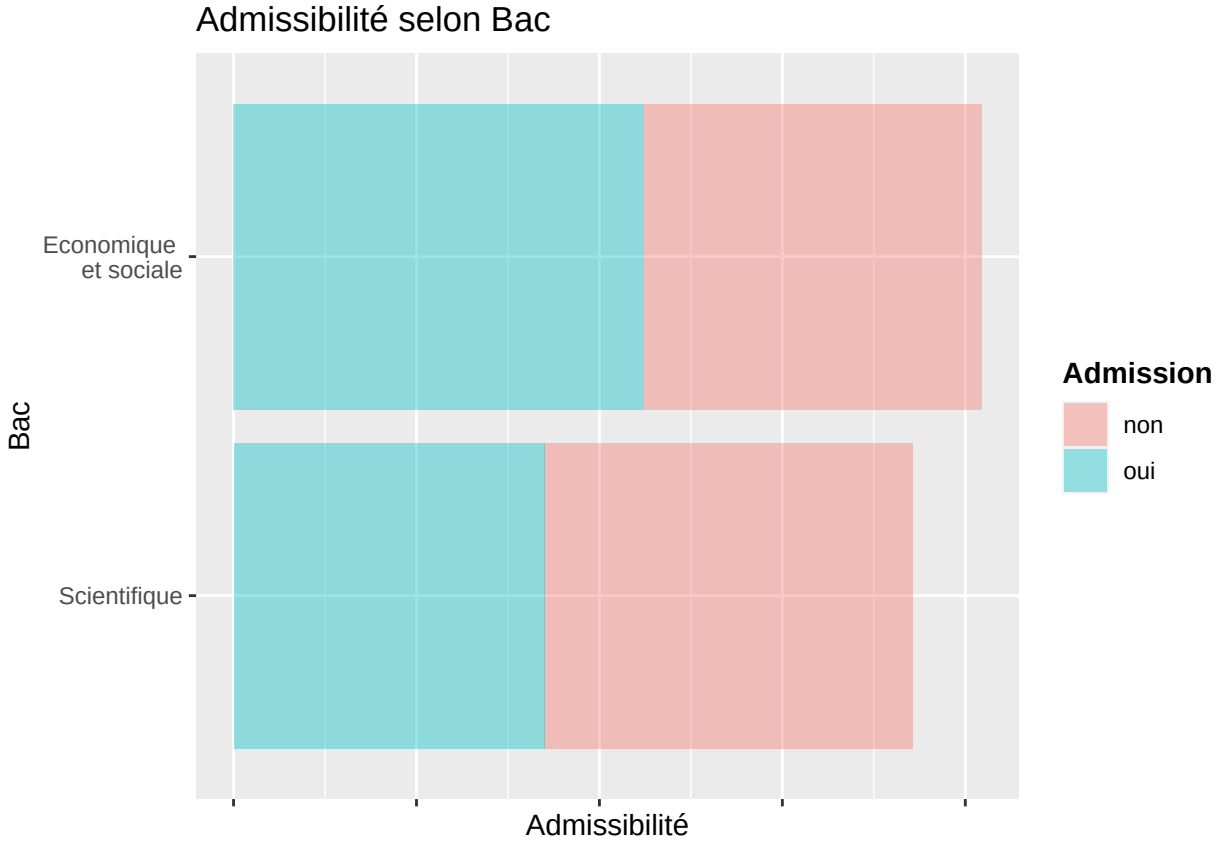


Table 1: Répartition des admissions

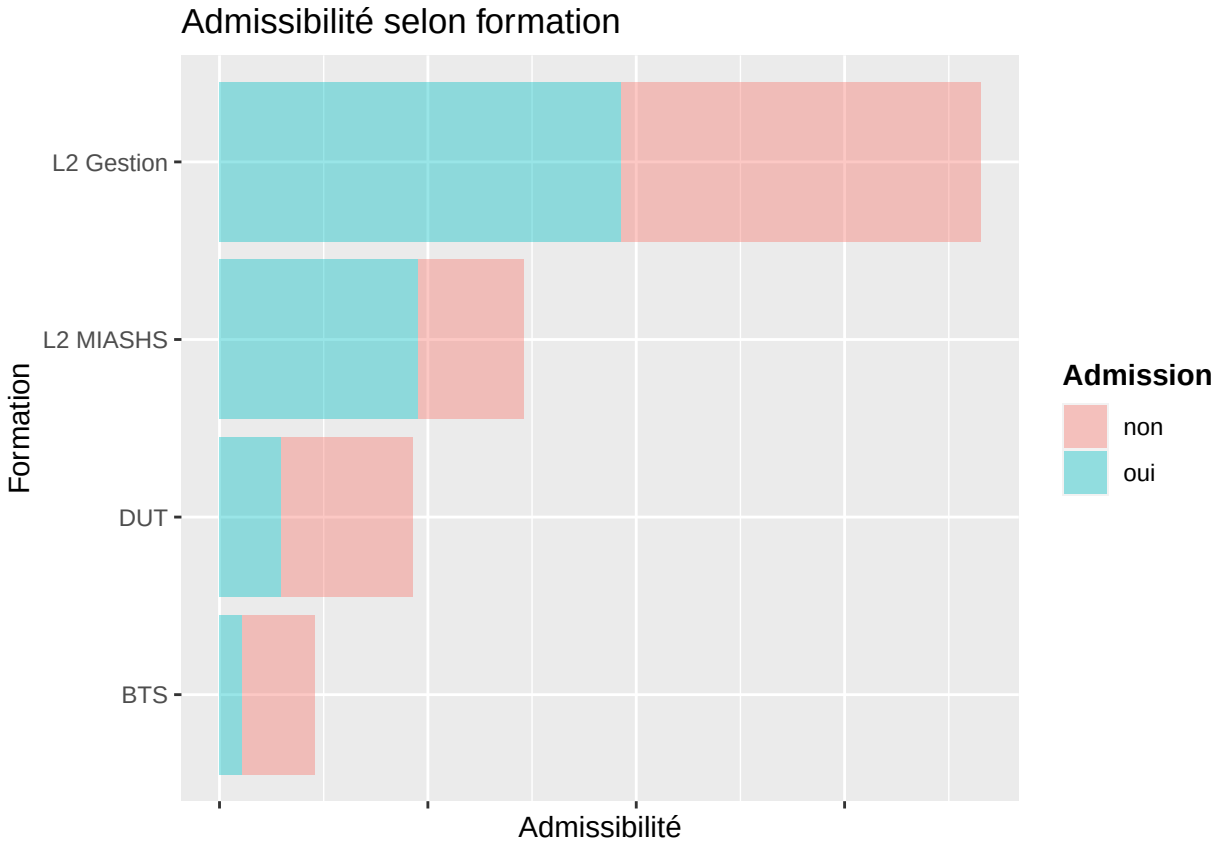
	Non	Oui	Total
Scientifique	503	425	928
Economique et sociale	461	561	1022
Total	964	986	1950

Table 2: Taux de réussite

	Non	Oui
Scientifique	54.20	45.80
Economique et sociale	45.11	54.89

Le taux de réussite à ce concours d'entrée est plus élevé pour les candidats ayant obtenu un Bac économique et social que pour ceux qui ont obtenu un Bac scientifique. Toutefois le rapport reste équilibré, donc la série de Bac n'apparaît pas pour le moment comme une variable discriminante, c'est à dire qu'elle ne nous permet pas à priori d'affirmer que c'est un élément de réussite ou d'échec.

On va donc s'intéresser à la formation suivie post-Bac.



Ici la répartion des candidats par formation suivie est assez déséquilibrée, nous allons donc surtout observer les taux de réussite. Il semble que cette variable soit un peu plus discriminante, par exemple, le taux de réussite pour les candidats provenant d'un BTS est de 24%. Toutefois on peut remarquer que quand le nombre de candidat par catégorie augmente, le rapport des taux de réussite s'équilibre. On ne peut alors conclure sur le caractère discriminant de cette variable.

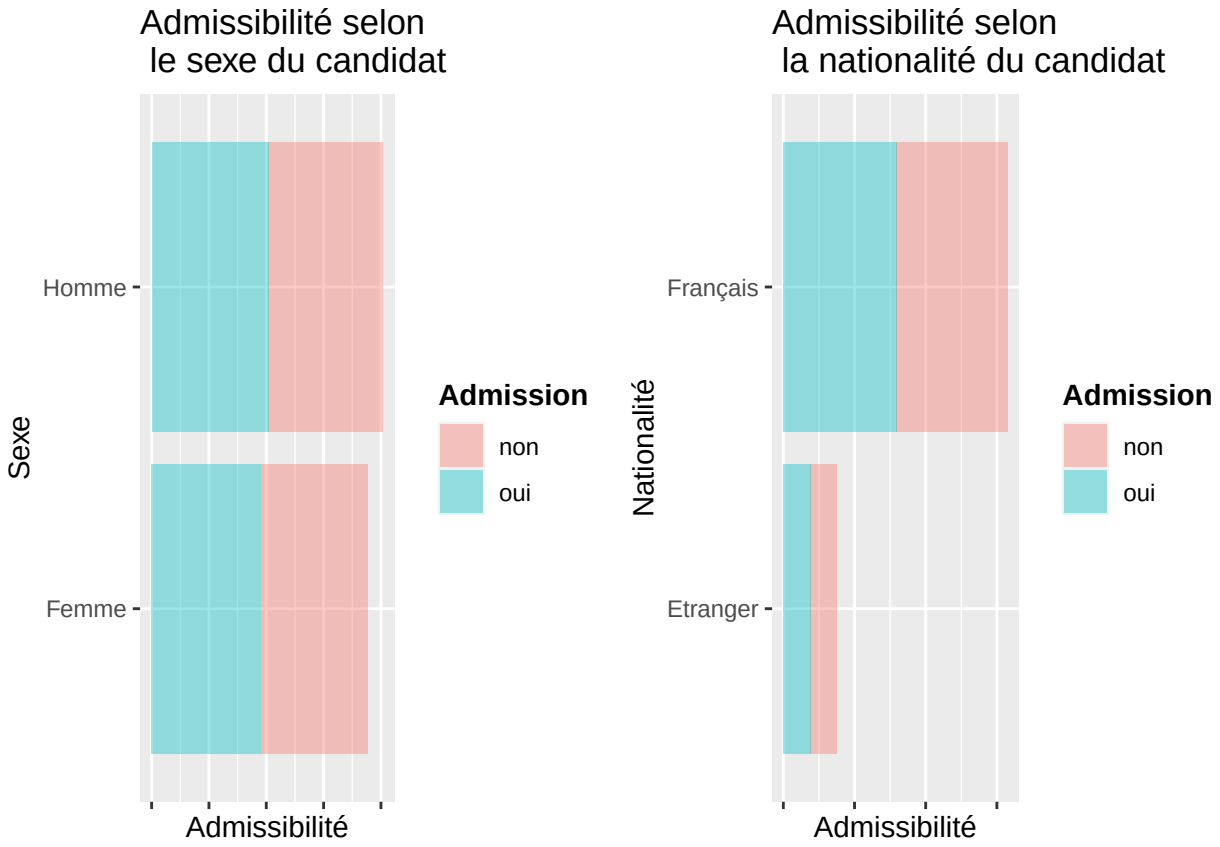
Pour compléter cette étude sur les profils des candidats, nous pouvons étudier le taux de réussite selon le sexe et selon la nationalité du candidat.

Table 3: Répartition

	Non	Oui	Total
BTS	105	33	138
DUT	189	89	278
L2 MIASHS	152	286	438
L2 Gestion	518	578	1096
Total	964	986	1950

Table 4: Taux de réussite

	Non	Oui
BTS	76.09	23.91
DUT	67.99	32.01
L2 MIASHS	34.70	65.30
L2 Gestion	47.26	52.74



Les rapports sont comme on pouvait s'y attendre très équilibré.

L'intérêt pour le département d'économie de mettre en avant ces statistiques, c'est d'abord d'attirer le plus d'étudiant provenant de formations "voisines" possible. En effet, la base de donnée ne comprend que les étudiants provenants des Bac S et ES et donc issus de formations post-Bac plus ou moins proches de la licence d'économie. De plus, avec un taux de réussite total très légèrement supérieur à 50%, cela devrait encourager uniquement les plus motivés à passer l'épreuve d'entrée. Publier ces statistiques aurait aussi un impact sur les potentiels candidats qui ne se retrouveraient pas dans les catégories présenteé par ces statistiques, cela pourrait les décourager à tenter leur chance. D'un autre côté pour ceux qui se sentiraient concernés, cela montre qu'il n'y a pas de profil type de réussite ni d'échec mais que c'est le travail qui fera la différence.

2. Relation entre la note et le profil

Avant de commencer la modélisation, je souhaite modifier la nature de certaines variables afin de simplifier la compréhension et donc l'interprétation des résultats. De plus, nous avons besoin de modifier la variable `Admission` qui doit être `numeric` pour pouvoir réaliser les modèles de prédiction.

On souhaite étudier la relation entre la note obtenue par le candidat à l'épreuve d'entrée et son profil. Pour cela, nous allons créer un modèle qui contient `Note_concours` en variable à expliquer et les variables explicatives sont les caractéristiques de chaque candidat.

En observant le résultat de la régression à la page suivante, on remarque que la grande majorité des variables est très significative. Ce qui n'est pas surprenant c'est que les variables `Sexe`, `Nationalite` et `Serie_Bac` ne semblent pas significatives et cela confirme nos hypothèses par rapport aux visualisations précédentes. On remarque que le type de formation post Bac suivie a un impact moyen positif, par rapport à la catégorie de référence qui est le BTS, sur la note obtenue à l'épreuve d'entrée. Notamment pour le parcours en mathématiques et informatique appliquée aux sciences humaines qui possède le plus gros impact puisque la note d'un candidat issu de ce parcours est en moyenne plus élevée de 1.3 par rapport aux candidats provenant d'un BTS.

Il est intéressant aussi de notifier que pour un étudiant qui va avoir un an de plus par rapport à l'âge normal correspondant à ce niveau d'étude, sa note va diminuer de presque 0.6 points. On peut donc imaginer qu'un étudiant qui tente de se réorienter éprouvera plus de difficultés que les autres pour intégrer la licence d'économie.

Il est utile ici tester la significité de certaines variables. Plus précisemment nous nous intéressons aux variables `Sexe`, `Nationalite`, `Serie_Bac` et `Mention_Bac`:

Après réalisation des test du Khi-deux sur ces variables, il s'est avéré que les variables `Sexe` et `Nationalite` sont bien indépendantes avec le fait d'avoir une note au concours supérieure à 12.00. En revanche nous rejetons l'hypothèse d'indépendance de la série de Bac obtenu. Nous retirerons alors les variables `Sexe` et `Nationalite` du modèle MCO.

Table 5: Modèle MCO

	Dependent variable:
	Note_concours
SexeH	0.004
	(0.045)
Nationalitefrançais	−0.030
	(0.057)
Retard-1	0.580***
	(0.117)
Retard1	−0.432***
	(0.061)
Retard2	−0.925***
	(0.075)
Retard3	−1.594***
	(0.068)
Serie_BacES	0.029
	(0.046)
Mention_BacB	0.160*
	(0.082)
Mention_BacP	−0.261***
	(0.049)
Mention_BacTB	0.106
	(0.165)
Formation_suiviDUT	0.408***
	(0.104)
Formation_suiviMIASHS	1.379***
	(0.100)
Formation_suiviSEG	0.964***
	(0.092)
Mention_FormationB	0.389***
	(0.070)
Mention_FormationP	−0.515***
	(0.051)
Mention_FormationTB	0.442***
	(0.144)
Constant	11.832***
	(0.118)
Observations	1,950
R^2	0.468
Adjusted R^2	0.463
Residual Std. Error	0.989 (df = 1933)
F Statistic	106.123*** (df = 16; 1933)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 6: Test de Breush-Pagan

BP	Degré de liberté	P-value
63.10812	14	0

3. Modèle de probabilité linéaire

L'objectif final de notre étude est de pouvoir donner à un potentiel nouveau candidat la probabilité qu'il réussisse à l'examen d'entrée. Les modèles que nous allons construire sont donc destinés à être utilisés sur de nouvelles données. C'est pourquoi, il me semble judicieux de créer deux échantillons de notre population. Ainsi nous utiliserons un échantillon d'entraînement pour construire les modèles et un échantillon de test pour vérifier et comparer l'efficacité de chaque modèle. L'échantillon d'entraînement est constitué des 2/3 des observations et sera généré aléatoirement. L'échantillon de test sera composé du reste de la population et sera aussi généré de manière aléatoire.

On souhaite maintenant exprimer le fait qu'un candidat soit admis ou non comme une probabilité.

On notera que l'on a retiré du modèle les variables qui se sont révélées non-significatives lors de la régression avec les moindres carrés généralisés. De plus nous remarquons que la variable `Note_concours` pose problème dans un modèle ou l'on souhaite prédire une probabilité de réussite pour un nouveau candidat. C'est une variable qui à elle seule discrimine tous les individus et rend inutilisable les autres variables. De plus, les informations données par cette variable sont incompatibles avec notre objectif qui est d'utiliser nos modèles sur de nouvelles données.

Malheureusement ce modèle est inutilisable car ici nous voulons exprimer une probabilité d'être admis ou non. En effet ici y_i prend la valeur 0 ou 1, ainsi il en va de même pour l'erreur. On en déduis que l'hypothèse de normalité de la distribution de l'erreur n'est plus viable et qu'elle suit maintenant une loi discrète. Comme nous soupçonnons la présence d'hétéroscédasticité, faisons un test pour vérifier :

La valeur de ce test de Brush Pagan est supérieure à la valeur du khi-deux pour 13 degré de liberté et une erreur de premier degré à 0.05. On a : $59.4 > 22.4$, donc on rejette l'hypothèse nulle d'homoscédasticité. Observons la répartition des résidus pour visualiser cette hétéroscédasticité sur les graphiques de la page suivante.

En effet l'hétéroscédasticité apparaît ici, nous observons que les résidus ne collent pas aux prédictions. C'est pour cela que nous observons de l'hétéroscédasticité. En fait l'erreur, comme notre variable à expliquer suit une loi binomiale et ne peut prendre que deux valeurs, 0 ou 1. Il y a donc une incompatibilité entre les résidus et les données du modèle.

La première étape de la correction de l'hétéroscédasticité va être d'établir une contrainte au modèle. On a p_i la probabilité pour un candidat d'être admis. p_i dépend des 9 prédicteurs du modèle. La contrainte est donc la suivante :

$$0 \leq p_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ ... \ + \beta_8 X_8 \leq 1$$

Nous commençons par supprimer les valeurs prédites non comprises entre 0 et 1.

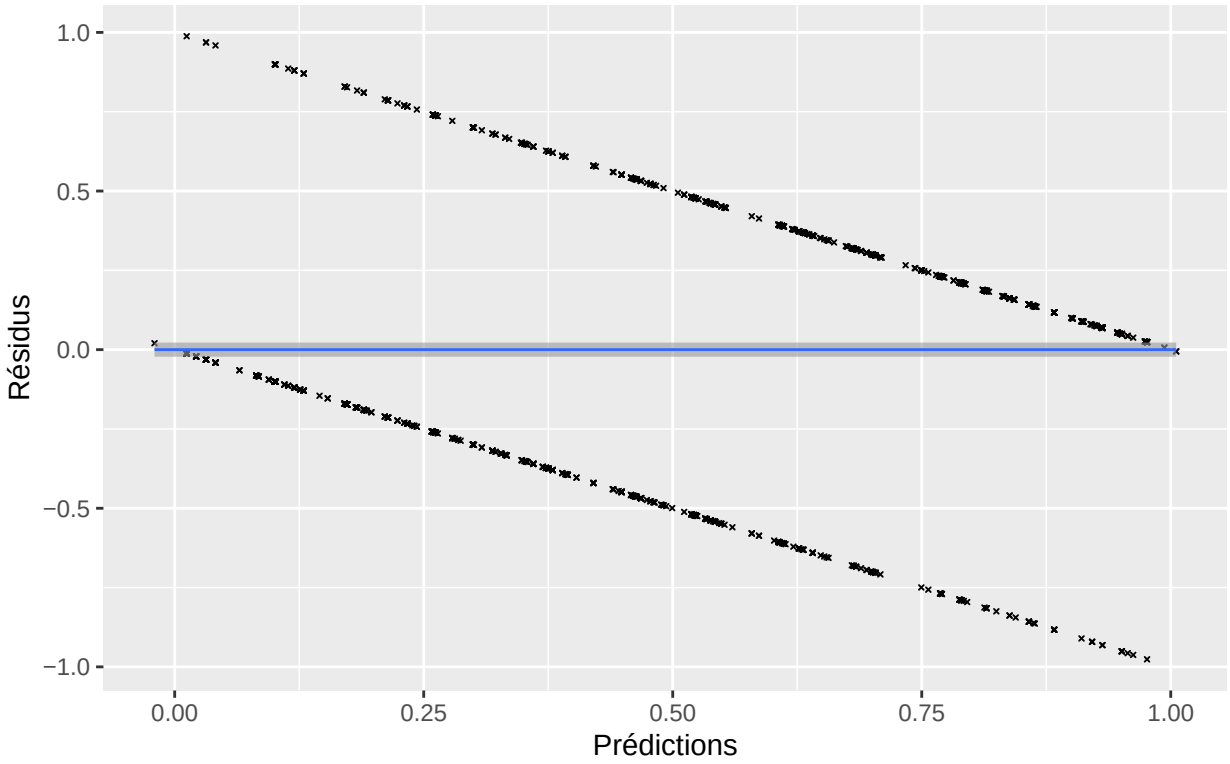
Nous allons corriger les données en appliquant un poids aux données. Pour cela nous devons calculer la variance corrigée :

$$Var_{corr} = p_{i,corr}(1 - p_{i,corr})$$

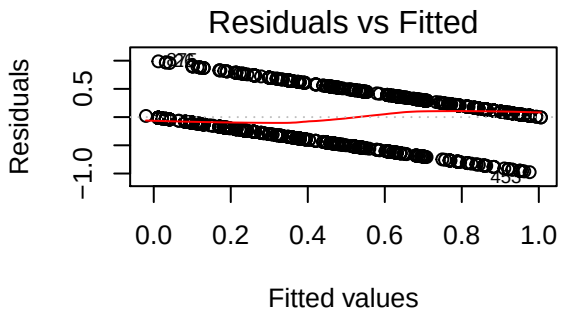
Ensuite nous appliquons le poids :

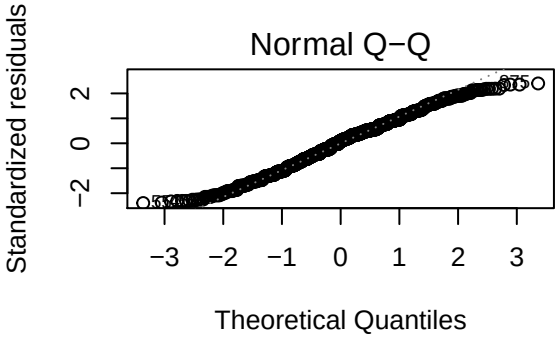
$$weights = \frac{1}{Var_{corr}}$$

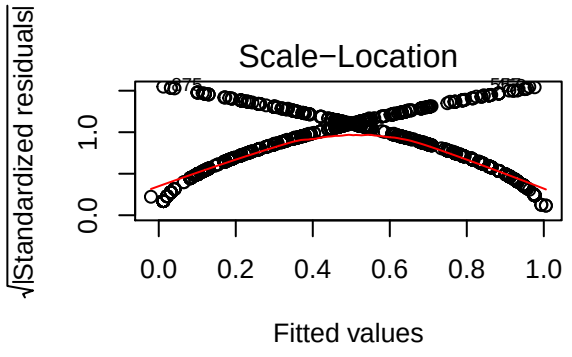
Résidus du modèle à probabilité linaire selon les valeurs prédites



\







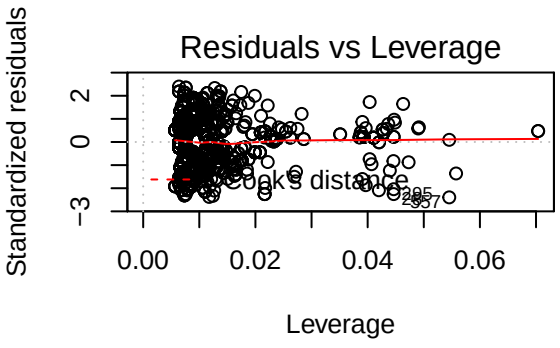
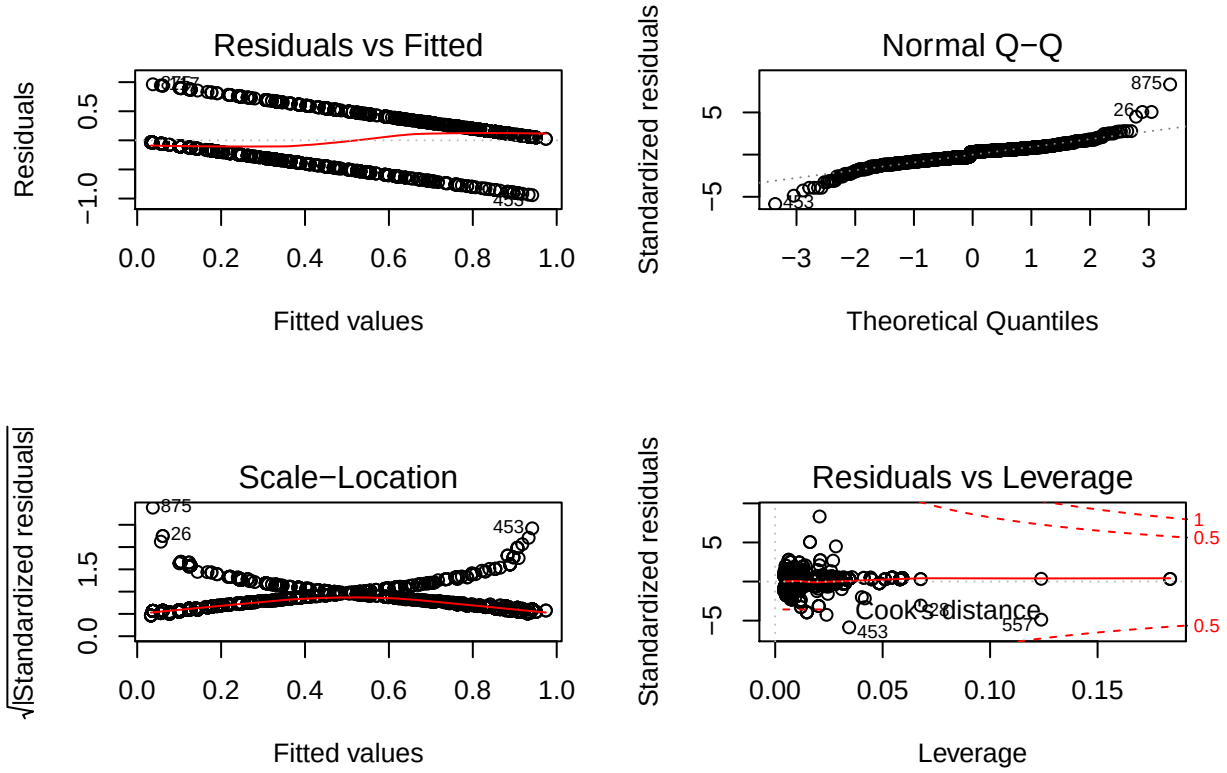


Table 7: Test de Breush-Pagan sur les données corrigées

BP	Degré de liberté	P-value
62.26883	14	0

Résulats du modèle corrigé

L'hétéroscédasticité persiste puisque l'on a vu dans le deuxième test de Breush-Pagan que l'on rejette encore l'hypothèse d'homoscédasticité.



Finalement après application des poids sur les données, on obtient un modèle avec un R^2 plus élevé, présenté à la page suivante, que dans le premier modèle. C'est la seule source de satisfaction étant donné que je ne suis pas parvenu à corriger l'hétéroscédasticité.

Table 8: Comparaison des modèles à probabilités linéaires

	Dependent variable:	
	Admission	
	Standard	Corrigé
	(1)	(2)
Retard-1	0.131**	0.135***
	(0.060)	(0.039)
Retard1	−0.163***	−0.148***
	(0.031)	(0.035)
Retard2	−0.309***	−0.311***
	(0.040)	(0.043)
Retard3	−0.511***	−0.495***
	(0.035)	(0.033)
Serie_BacES	0.019	0.024
	(0.024)	(0.021)
Mention_BacB	0.025	0.035
	(0.041)	(0.032)
Mention_BacP	−0.089***	−0.067***
	(0.025)	(0.023)
Mention_BacTB	0.055	0.059
	(0.083)	(0.076)
Formation_suiviDUT	0.129**	0.140***
	(0.052)	(0.052)
Formation_suiviMIASHS	0.439***	0.428***
	(0.049)	(0.050)
Formation_suiviSEG	0.278***	0.284***
	(0.046)	(0.048)
Mention_FormationB	0.068*	0.091**
	(0.035)	(0.036)
Mention_FormationP	−0.159***	−0.140***
	(0.026)	(0.026)
Mention_FormationTB	0.130	0.103
	(0.080)	(0.068)
Constant	0.492***	0.454***
	(0.053)	(0.054)
Observations	1,300	1,297
R^2	0.324	0.545
Adjusted R^2	0.317	0.540
Residual Std. Error	0.413 (df = 1285)	1.073 (df = 1282)
F Statistic	43.966*** (df = 14; 1285)	109.516*** (df = 14; 1282)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 9: Test de Wald

	Degré de liberté	Statistique du khi-2	P-value
Retard	4	166.4906180	0.0000000
Serie Bac	1	0.4501776	0.5022506
Mention Bac	3	16.0274851	0.0011194
Formation Suivie	3	78.8824308	0.0000000
Mention Formation	3	45.9308633	0.0000000

4. Modèle Logit

Nous savons que le modèle à probabilité linéaire pose toujours un problème d'hétéroscédasticité, heureusement il existe d'autres méthodes pouvant estimer une telle variable à expliquer.

1) Significativité des coefficients

La constante individuelle dans ce modèle ne semble pas significative tout comme les modalités `très bien` et `bien` des variables `Mention_Formation` et `Mention_Bac` par rapport à la modalité `Assez bien`. Nous allons donc réaliser des tests pour confirmer ou infirmer la significativité de ces coefficients.

La variable `Serie_Bac`, ne semble pas significative toutefois, la p-value est très proche du seuil et je fais le choix de conserver cette variable. Néanmoins nous conservons cette information et nous la réutiliserons lors de la comparaison des pseudo-R^2.

2) Signe des coefficients

Comme nous l'avions noté plutôt dans cette étude, le fait d'être en retard par rapport à l'âge normal à ce niveau d'étude semble avoir un impact négatif sur la probabilité d'être admis au concours d'entrée, tout comme les mentions passable par rapport à la mention assez bien. Toutes les formations post-Bac ont un impact positif sur la probabilité de réussite de ce concours par rapport à la formation BTS.

3) Interprétation des rapports de chances

Bien que nous avons interprété les signes des coefficients, nous n'avons pas encore précisé la mesure de l'impact de ces différentes variables. Nous devons alors calculer les ODDs ratios, en fait, en faisant l'exponentielle des coefficients on obtient les rapports de chances directement interprétables.

Nous dirons par exemple qu'un étudiant provenant de licence 2 de mathématiques à environ 9 fois plus de chance d'être admis à ce concours par rapport à un étudiant de BTS.

Aussi un candidat qui est en retard de deux années par rapport à l'âge normal à 0.23 fois plus de chance d'être admis. Cela signifie en fait qu'il a moins de chance d'être admis au concours.

Table 10: Modèle Logit

	Dependent variable:
	Admission
Retard-1	0.855*
	(0.437)
Retard1	−0.797***
	(0.170)
Retard2	−1.456***
	(0.220)
Retard3	−2.621***
	(0.221)
Serie_BacES	0.094
	(0.140)
Mention_BacB	0.190
	(0.265)
Mention_BacP	−0.508***
	(0.147)
Mention_BacTB	0.361
	(0.559)
Formation_suiviDUT	0.604**
	(0.299)
Formation_suiviMIASHS	2.229***
	(0.291)
Formation_suiviSEG	1.317***
	(0.263)
Mention_FormationB	0.323
	(0.209)
Mention_FormationP	−0.819***
	(0.150)
Mention_FormationTB	0.725
	(0.509)
Constant	0.012
	(0.301)
Observations	1,300
Log Likelihood	−663.831
Akaike Inf. Crit.	1,357.662
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 11: Rapport de chance selon chaque modalités

	Rapports de chance
Constante	1.0115964
Retard - 1	2.3514983
Retard + 1	0.4505269
Retard + 2	0.2331575
Retard + 3	0.0727550
Série_BacES	1.0983885
Mention_BacB	1.2087865
Mention_BacP	0.6019815
Mention_BacTB	1.4351023
Formation_suiviDUT	1.8285891
Formation_suiviMIASHS	9.2930001
Formation_suiviSEG	3.7337094
Mention_FormationB	1.3805952
Mention_FormationP	0.4407118
Mention_FormationTB	2.0641441

Table 12: Tableau des effectifs

Réalité	Predictions		
	Non Admis	Admis	Total
Non Admis	240	96	336
Admis	68	246	314
Total	308	342	650

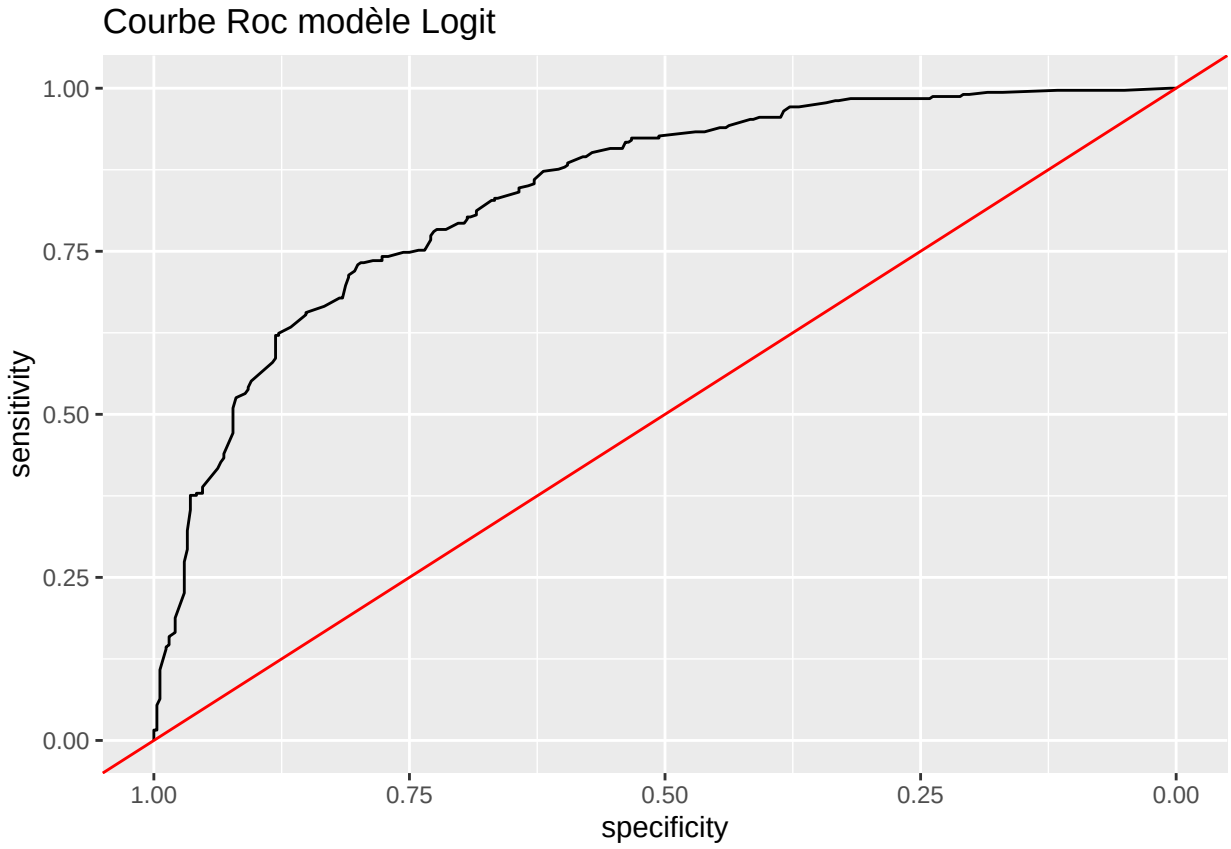
Table 13: Matrice de confusion

Réalité	Prédictions	
	Non Admis	Admis
Non Admis	71.43	28.57
Admis	21.66	78.34

4) Matrice de confusion et courbe ROC

Ce que l'on peut dire du tableau des effectifs, c'est que la population est très équilibrée entre les admis et les non admis, ce qui n'est pas surprenant compte tenu des statistiques descriptives vues plus tôt. Concernant la matrice de confusion, on va d'abord s'intéresser au taux de vrais positifs, en d'autre termes, parmi le nombre de personnes admise en réalité, 78,34% des prédictions sont correctes, on en déduis le taux d'erreur de 21.66%. Cela est honorable mais pas tout à fait satisfaisant. De plus le taux d'erreur global de 25% n'est pas non plus une bonne nouvelle. Nous pourrions toutefois procéder à un ajustement du modèle en augmentant la probablité d'être admis, ainsi, le taux de vrais positifs augmenterais. Malheureusement, cela n'est pas cohérent avec notre objectif car cela signifierait que nous considérerons un étudiant admis s'il a moins de 50% de chance de réussite. Le seuil de 0,5 dans notre modèle prend donc tout son sens.

Afin de conclure sur la qualité de ce modèle, il faut dans un premier temps tracer la courbe ROC et mesurer l'aire sous cette courbe qui va nous indiquer la qualité des précisions du modèle quel que soit le seuil de classification sélectionné.



Nous obtenons une valeur AUC de 0.84, ce qui traduit une bonne qualité du modèle.

5. Modèle Probit

Bien que le modèle Logit se révèle être fiable et efficace, nous devons explorer toute les pistes.

Concernant les signes et la significativité des coefficients, nous pouvons tirer les mêmes conclusions que pour le modèle Logit car on obtient des estimations similaires d'une méthode à l'autre. En fait on obtient les coefficients des estimateurs du modèle probit à partir de ceux du modèle logit :

$$\beta_{probit} \simeq \beta_{logit} \frac{\sqrt{3}}{\pi}$$

Après visualisations des coefficients, nous en déduisons que les rapports de chances eux aussi vont se rapprocher de ceux obtenus avec le modèle logit il est donc inutile de se répeter ici. En revanche nous pouvons introduire un premier élément de comparaison entre le modèle logit et probit, la matrice de confusion.

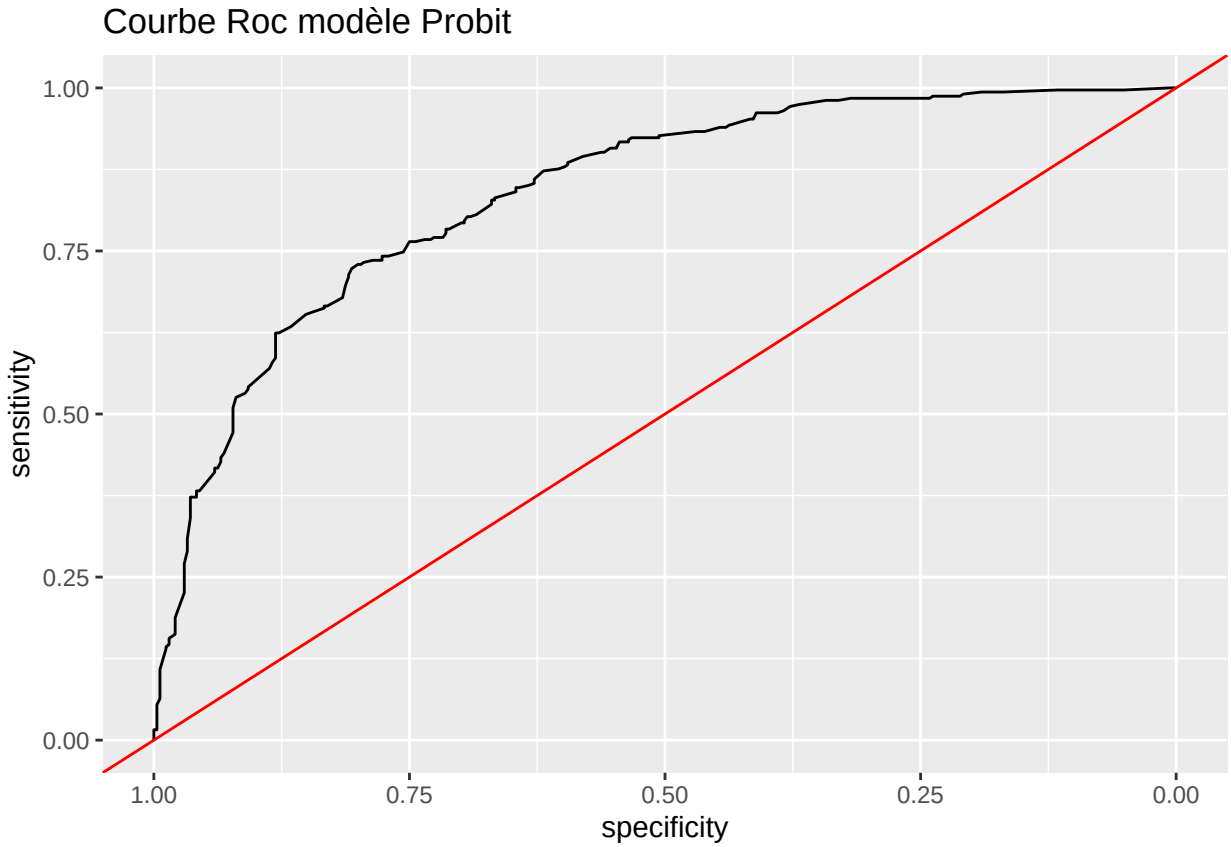
Table 14: Tableau des effectifs

Réalité	Predictions		
	Non Admis	Admis	Total
Non Admis	240	96	336
Admis	68	246	314
Total	308	342	650

Table 15: Matrice de confusion

Réalité	Prédictions	
	Non Admis	Admis
Non Admis	71.43	28.57
Admis	21.66	78.34

Ces résultats ne nous permettent pas de conclure sur le modèle probit par rapport au modèle logit tant les données obtenues sont proches. On obtient d'ailleurs la même erreur globale bien que les erreurs de type 1 et 2 diffèrent. Intéressons nous maitenant à la courbe ROC du modèle probit:



Encore une fois il est difficile de conclure sur un modèle plutôt qu'un autre car la valeur AUC obtenue est de 0.84.

III - Conclusion

Le premier modèle nous donne des éléments d'informations qui confirment nos intuitions des statistiques descriptives. Le modèle à probabilité s'est révélé inefficace, de par la présence d'hétéroscédasticité d'un côté et d'autre part la loi de distribution des estimations n'est pas compatible avec notre objectif qui est d'exprimer une probabilité. Nous avons dû recourir à des modèles dont les estimations suivent une loi de

Table 16: Pseudo R2

R2 Logit	R2 Probit
0.2626941	0.2625551

distribution telle que:

$$\lim_{\alpha+\beta_1 X_1+...+\beta_k X_i \to -\infty} F(\alpha+\beta_1 X_i+...+\beta_k X_i)=0, \lim_{\alpha+\beta_1 X_1+...+\beta_k X_i \to +\infty} F(\alpha+\beta_1 X_i+...+\beta_k X_i)=1$$

C'est pourquoi nous avons fait appel aux modèles logit et probit. C'est sur ces deux modèles que notre choix final va se porter pour répondre à la demande du département d'économie.

1) Comparaison des erreurs

	Erreur Globale	Erreur de type 1	Erreur sur les admissions
Logit	0.252	0.286	0.217
Probit	0.252	0.286	0.217

2) Pseudo R2

Il semblerait que le modèle Logit soit très légèrement supérieur, toutefois les modèles sont très proches, c'est pourquoi en conclusion nous pouvons avancer le fait que les modèles logit et probit fonctionnent avec un pseudo R^2 supérieur à 0.2. Nous avons une très légère préférence pour le modèle logit qui offre un taux d'erreur plus faible sur les admissions. Pour finir nous devrons donc demander aux étudiants voulant intégrer la 3ème année de licence d'économie leur parcours post-bac, leur âge et la mention obtenue au Bac ainsi que celle de leur dernière année pour pouvoir leur fournir leur probabilité de réussite. Il est important de garder en mémoire qu'il n'y a pas de profil type de réussite et que seule la motivation et le travail permettront d'intégrer et de réussir au sein du parcours d'économie.