Marketing & Analyse

LEFAFTA Rémi

Contents

1 Analyse statistique du taux de réussite des épreuves d'admissibilité selon le profil du candidat.

1.1 Sommaire de nos variables

1.1.1 Admissibilité selon l'année

Il est intéréssant de savoir si il y a une incidence temporelle sur l'admissibilité. En effet, peut-être que les examens sont devenus plus faciles ou inversement et que les taux d'admissibilité ont ainsi été changé à travers les années. Cependant, d'après la table ci-dessus, l'année ne semble pas discriminer l'admissibilité.

Table 1: Admissibilité selon l'année

Admissibilité	*Année*					
	-6	-5	-4	-3	-2	-1
Admissible	53.548	55.219	47.712	48.844	52.367	46.459
Non_admissible	46.452	44.781	52.288	51.156	47.633	53.541

1.1.2 Admissibilité selon le retard

Le niveau d'avance ou de retard par rapport à l'âge moyen semble être discrimant. Plus le retard est élevé, plus le taux d'admissibilité est faible. C'est ainsi une variable pertinente dans notre sujet de travail.

Table 2: Admissibilité selon le retard

Admissibilité	*Retard*				
	-1	0	1	2	3
Admissible	85.542	75.071	47.481	30.827	15.303
Non_admissible	14.458	24.929	52.519	69.173	84.697

1.1.3 Admissibilité selon le sexe et la nationalité

Ni le sexe, ni la nationalité ne semble discriminer le fait d'être admissible ou non, ce qui est assez logique même si le fait d'être étranger pourrait intuitivement nous faire penser qu'il y aurait un taux d'admissibilité plus faible.

Table 3: Admissibilité selon la nationalité, le sexe

Admissibilité	*Nationtalité*		*Sexe*	
	Etranger	Français	Femme	Homme
Admissible	51.2	50.413	50.691	50.446
Non admissible	48.8	49.587	49.309	49.554

1.1.4 Admissibilité selon le type de bac et la mention

Les individus ayant une mention B ou TB en ES ont un taux d'admissibilité très élevé, avec respectivement 85,2 et 85,3. Il semble que les bacheliers ES aient à priori, plus de chance d'être admissible. Cette variable semble à son tour avoir la capacité de discriminer.

Table 4: Admissibilité selon le type de bac et la mention

Admissibilité	*ES*				*S*			
	P	AB	B	TB	P	AB	B	TB
Admissible	45.946	53.004	85.217	85.294	38.406	48.848	68	60
Non admissible	54.054	46.996	14.783	14.706	61.594	51.152	32	40

1.1.5 Admissibilité selon études supérieures

A nouveau, la formation suivie semble discriminer nos individus. Ceux ayant fait un BTS ont à priori seulement 24% de chance d'être considérés comme admissibles. A l'inverse, les titulaires d'un MIASHS ont été 65% a avoir été déclaré admissibles. On a ainsi une nouvelle variable permettant de discriminer nos individus concernant l'admissibilité.

Table 5: Admissibilité selon la formation suivie

Admissibilité	*Formation suivie*			
	BTS	DUT	MIASHS	SEG
Admissible	23.913	32.014	65.297	52.737
Non_admissible	76.087	67.986	34.703	47.263

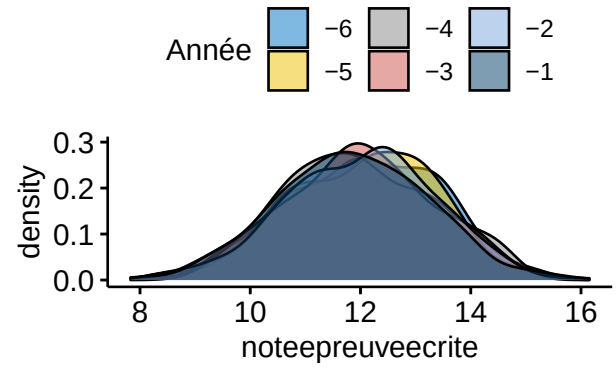
1.1.6 Etude de la variable `notes aux épreuves écrites`

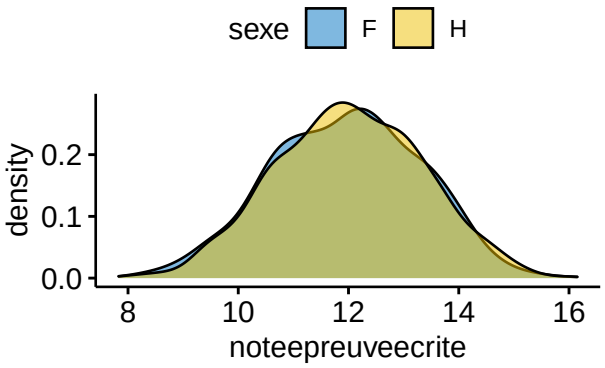
	Moyenne aux épreuves écrites
Minimum	7.830
Maximum	16.150
Moyenne	11.988
Mediane	12.010
EcartType	1.350
Variance	1.821

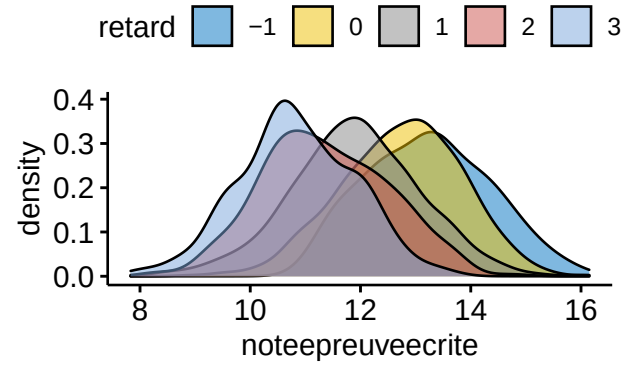
Le sommaire de cette variable nous donne des indices intéréssants quant à la distribution de notre variable `notes aux épreuves écrites`. En sachant qu'un élève est admissible à condition que sa moyenne soit supérieure à 12, la médiane nous indique qu'un peu plus de la moitié de la population est considérée comme admissible. Cette variable permet ainsi de discriminer parfaitement les étudiants admissibles et non admissibles.

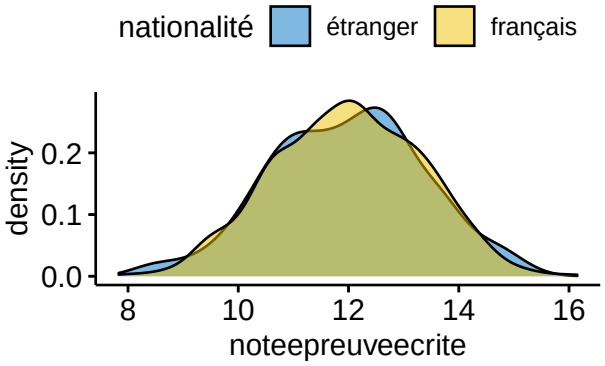
1.1.7 Graphiques

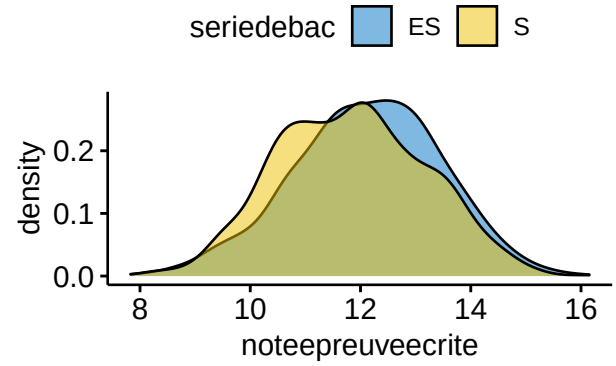
Nous allons représenter la densité de nos variables selon la note à l'épreuve écrite. Comme dit précédemment, une note supérieure à 12 implique la possibilité d'être admis.

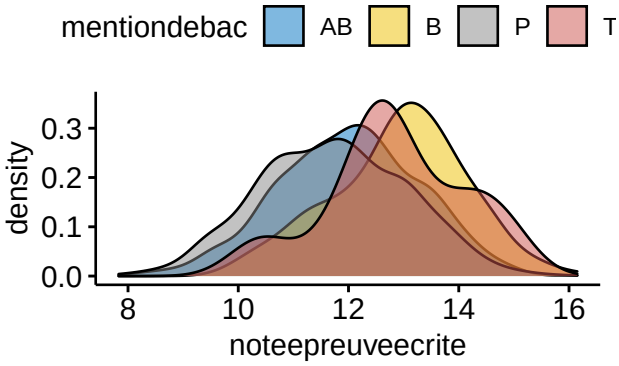

Année
−6
−5
−4
−3
−2
−1
density
noteepreuveecrite

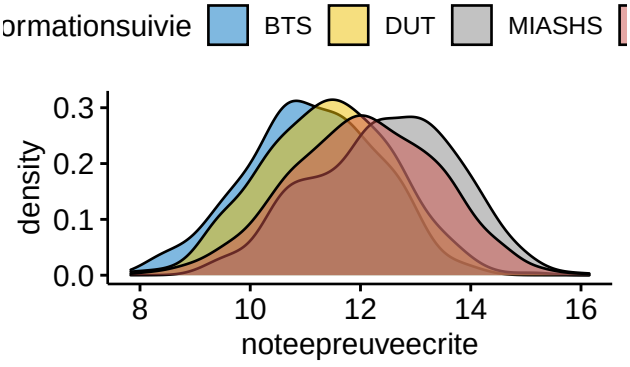

sexe
F
H
density
noteepreuveecrite

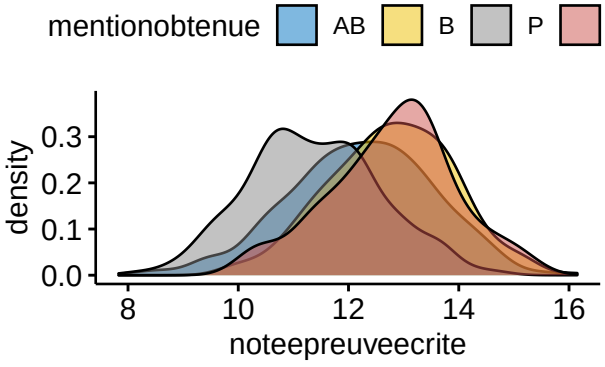

retard
−1
0
1
2
3
density
noteepreuveecrite


nationalité
étranger
français
density
noteepreuveecrite


seriedebac
ES
S
density
noteepreuveecrite

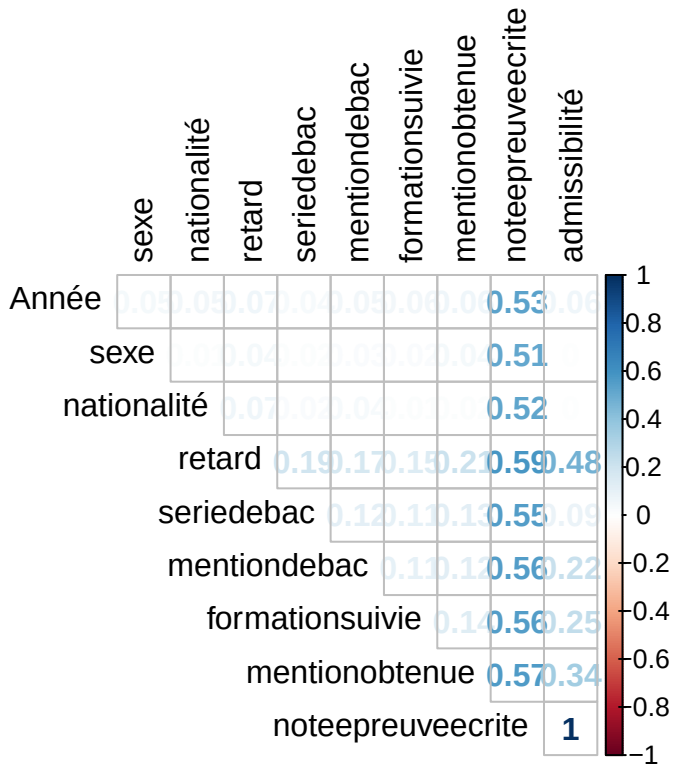

mentiondebac
AB
B
P
T
density
noteepreuveecrite


ormationsuivie
BTS
DUT
MIASHS
density
noteepreuveecrite


mentionobtenue
AB
B
P
density
noteepreuveecrite

1.1.8 Lien entre nos variables

Nous allons utiliser la matrice de Cramer afin de voir le lien entre nos variables.



Comme nous avons pu le dire, l'admissibilité est parfaitement corrélée avec la note à l'épreuve écrite. La note aux épreuvres écrites est partiellemment liée à toutes les autres variables, cependant aucune de ces corrélations ne dépassent 0.8, ce n'est donc pas un lien fort.

1.2 Conclusion

Notre étude statistique nous a permis de mettre en lumière les variables qui semblent à priori discriminer l'admissibilité de nos individus. Il serait intéréssant de publier ces statistiques, afin d'aider et d'orienter au mieux les étudiants voulant continuer leurs études dans cette troisième année de licence et par la suite en master. Cependant, il est possible que la publication de ces statistiques découragent certains étudiants à passer les épreuvres, notamment ceux qui ont à priori une faible chance d'être admissibles. D'une autre part, ceux à priori admissibles, ne doivent pas non plus considérer leur admissibilité comme acquise.

2 Estimation par MCO du lien entre la note des individus et leurs caractéristiques

Notre individu référence est :

- Concours : passé il y a 6 ans
- Sexe : feminin
- Nationtalité : étrangère

- Un an d'avance
- Baccalauréat : ES
- Mention au baccalauréat : AB
- Formation post bac : BTS
- Mention études supérieures : AB

En faisant la régression avec toute les variables, excepté la variable admissibilité, nous remarquons que les résultats statistiques (table 7 en annexe) précédemment énoncés sont adéquats avec nos resultats :

- Les coefficients associés aux différentes années ne sont pas significatifs, on avait déjà pu montrer qu'il n'existait pas de lien entre l'admissibilité (c'est à dire note > 12) et les années, cela confirme notre à priori.
- Le coefficient associé au sexe masculin n'est pas significatif.
- Le coefficient associé au bac S n'est pas non plus significatif.
- A l'inverse, la mention au bac et celle en études supérieures ont des coefficients significatifs, ainsi que le retard. Cela confirme nos analyses préliminaires sur le sujet.

On obtient un R2 de 0.47.

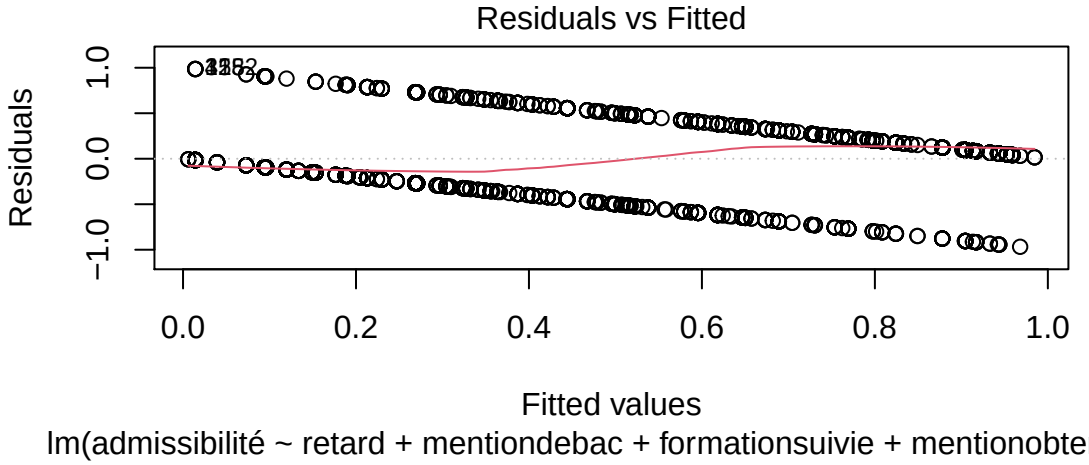
La table 7 présente la nouvelle régression sans nos variables non discriminantes.

Le R2 est légèrement plus faible, cependant notre R2 ajusté est plus élevé ! De plus, tous nos coefficients associés aux modalités sont significatifs exceptés pour la mention de bac TB.

3 Modèle de probabilité linéaire

Nous allons maintenant approcher le problème par le biais de la variable `admissibilité`.

Nous voyons dans la table 8 que les coefficients associés aux modalités mentiondebacB et mentiondebacTB ne sont statistiquement pas significatif alors que les autres le sont. Cependant, on voit graphiquement que l'approximation linéaire est inadaptée.



De plus, par définition, le modèle à probabilité linéaire est hétéroscédastique. La variance du terme d'erreur

est égale à la variance d'une loi de Bernouilli. Nos estimateurs ne sont donc plus BLUE. Par ailleurs, nous devons imposer une contrainte à notre intervalle de probabilité. De base l'intervalle est non borné. Nous ne pouvons donc pas parler de probabilité, il faut qu'on limite notre intervalle de probabilité entre 0 et 1.

Nous réalisons désormais un test de Breusch-Pagan afin de confirmer qu'il y bien de l'hétéroscédasticité dans notre modèle.

```
studentized Breusch-Pagan test
```

data: reg3 BP = 83.961, df = 13, p-value = 1.972e-12

Cela confirme ce que nous avons pu évoquer précédemment. On connait la forme de la variance du terme d'erreur, on peut ainsi appliquer la méthode des MCG, en limitant notre intervalle de probabilité comme énoncé précédemment.

Nos coefficients estimés sont dans la table 8 et sont globalement similaires. On notera que dans le modèle estimé par MCG, le coefficient associé à la modalité `mentionbacTB` est significative à 5% tandis qu'elle n'est pas significative par la méthode des MCO. La différence notable est le R^2 qui est passé de 0.33 à 0.543. On a une qualité de l'ajustement linéaire bien meilleur lorsqu'on estime par la méthode des MCG, grâce au fait que les écarts types estimés soient plus faibles. De plus, la présence d'hétéroscédasticité a des implications sur nos coefficients estimés par la méthode des MCO. Ils restent non biaisés et consistants mais ils ne sont plus efficients.
Nos estimateurs ne sont donc plus BLUE. Nous devons ainsi prendre en compte cette hétéroscédasticité au quel cas nos écarts-types estimés sont incorrects. Les estimateurs trouvés par la méthode des MCG sont quant à eux BLUE. On réalise cependant un nouveau test de Breusch-Pagan afin de voir si nous avons réellement corrigé notre problème d'hétéroscédasticité.

```
studentized Breusch-Pagan test
```

data: reg3_mcg BP = 83.961, df = 13, p-value = 1.972e-12

Les termes d'erreurs sont toujours hétéroscédastiques. Nous n'avons pas réussi à corriger totalement le problème. Nous ne pouvons donc pas considérer que nos estimateurs sont efficients.

4 Modèle Logit

La régression logiste est pertinente dans notre contexte. En effet, elle est plus adaptée afin d'estimer une variable dichotomique.

4.0.1 Signe des paramètres

On peut voir dans la table 8 que les signes associés aux coefficients nous permettent de connaitre la probabilité, en plus ou en moins, d'être admissible par rapport à la catégorie de référence. Prenons deux exemples :

- Les modalités `retard0`, `retard1`, `retard2` et `retard3` ont un signe négatif. Cela veut dire que par rapport à la catégorie de référence `retard-1`, ils ont moins de chance d'être admissibles, toutes choses égales par ailleurs.
- Les modalités `formationsuiviDUT`, `formationsuiviMIASHS` et `formationsuiviSEG` ont un signe positif. Par conséquent, ils ont plus de chance d'être admissibles que la catégorie de référence qui est `formationsuiviBTS`, toutes choses égales par ailleurs.

4.0.2 Significativé

Nos coefficients associés à nos variables sont tous significatifs à 0,1% à l'exception de la constante, de `retard0`, significative à seulement 10%, de `mentionbacB`, `mentionbacTB`, `formationsuiviDUT` significatifs à 5% tout comme `mentionobtenueTB`. On n'interprétera pas les variables non significatives par la suite.

4.1 Rapports des chances

Afin d'interpréter les estimations obtenues, on calcul les rapports des chances présentés dans la table ci-dessous. On interprète toujours par rapport à la catégorie de référence.

	OR	2.5 %	97.5 %
(Intercept)	2.0341914	0.9436836	4.6134635
retard0	0.4424043	0.2172706	0.8330142
retard1	0.2126248	0.1041472	0.4018949
retard2	0.1061408	0.0505769	0.2072421
retard3	0.0333284	0.0157838	0.0653791
mentiondebacB	1.3219522	0.8566661	2.0736532
mentiondebacP	0.6275343	0.4953556	0.7937941
mentiondebacTB	1.7687709	0.7237600	4.9222349
formationsuivieDUT	1.7629040	1.0668971	2.9564365
formationsuivieMIASHS	8.8575422	5.4931575	14.5779563
formationsuivieSEG	4.1455040	2.6776401	6.5581796
mentionobtenueB	1.8118079	1.2851542	2.5728245
mentionobtenueP	0.4105994	0.3227037	0.5216249
mentionobtenueTB	2.2846194	1.1068816	5.0299235

Les personnes en retard de 3 ans (`retard3`) ont 0.03 fois moins de chance d'être admissibles par rapport à la catégorie de référence (`retard-1`). On peut aussi l'interpréter de la manière inverse, c'est à dire que ceux ayant une année d'avance (`retard-1`) ont $1/0.0.3 = 33$ fois plus de chance d'être admissibles que ceux étant dans l'âge normal. On pourra aussi évoquer le fait que ceux ayant effectué une `formationsuiviSEG` ont 4,14 fois plus de chance d'être admissibles par rapport à la catégorie de référence (`formationsuiviBTS`).

4.2 Ajustement du modèle

Le test du rapport de vraisemblance consiste à comparer les vraisemblances de deux modèles emboités, celui contraint et celui non contraint. La statistique de test s'écrit :

$$LR = -2(L_c - (-2ln\ L_nc)) = D_c - D_{nc}$$

avec D_c la déviance du modèle contraint et D_{nc} la déviance du modèle non contraint. On réalise le test du rapport de vraisemblance. On obtient la déviance respective des modèles avec $D_c = 2703.026$ et $D_{nc} = 1967.967$. Ainsi, en soustrayant, le résultat est positif. On considère donc que notre modèle logit est meilleur qu'un modèle avec seulement la constante.

4.3 Matrice de confusion

	Predicted 0	Predicted 1	Total
Actual 0	731	233	964
Actual 1	244	742	986
Total	975	975	1950

On peut caractériser les valeurs qu'on a obtenu :

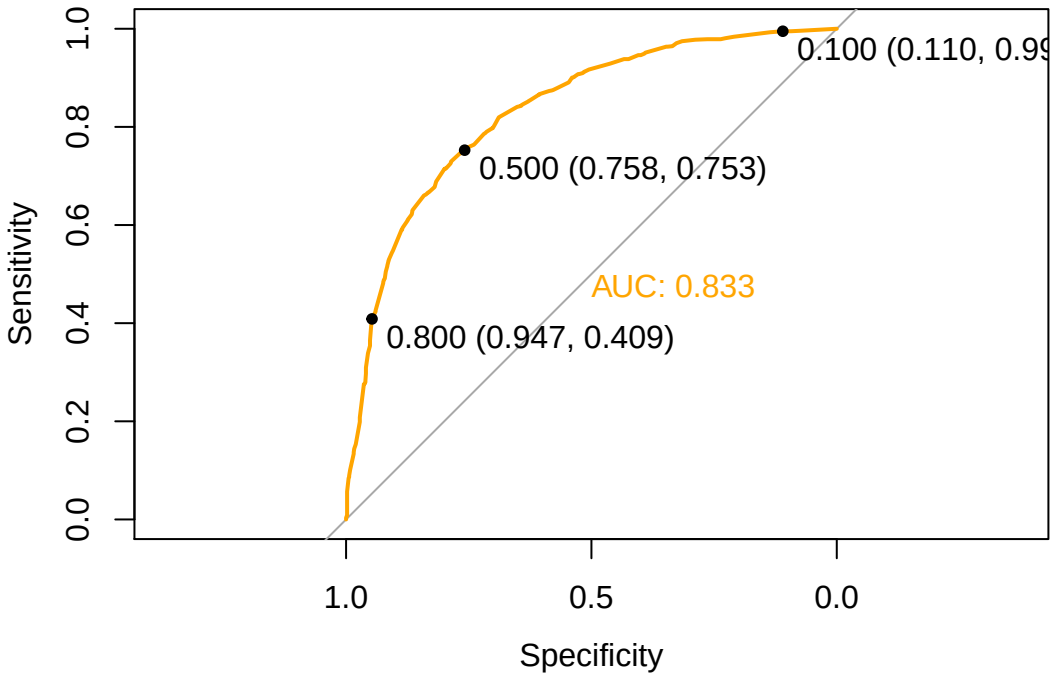
- 731 vrais négatifs (VN) : quand notre modèle prédit correctement 0, c'est à dire non admissible.
- 742 vrais positif (VP) : quand notre modèle prédit correctement 1, c'est à dire admissible.
- 233 faux négatifs (FN) : quand notre modèle prédit de manière incorrecte 1, c'est à dire admissible.
- 244 faux positifs (FP) : quand notre modèle prédit de manière incorrecte 0, c'est à dire non admissible.

De plus, grâce à ces informations, on peut définir des indices permettant de mesurer la perfomance :

- Accuracy $= 1 - (FP + FN)/Total = 0.755$, ainsi le taux d'erreur est de $1 - 0.755 = 0.245$.
- Spécificité $= VN/(VP + FN) = 0.75$, la spécificité mesure ici la proportion de non admissibles classée correctement.
- Sensibilité $= VP/(VP + FN) = 0.76$, la sensisiblité mesure, dans notre cas, la proportion de personnes admissibles qui ont été correctement classé en tant qu'admissibles.

La spécificité et la sensibilité sont des indicateurs complémentaires à l'accuracy. En effet cela permet de mesurer les erreurs spécifiques aux classes.

4.4 Courbe ROC



La valeur de l'AUC est de 0,833. Graphiquement, c'est l'aire sous la courbe ROC. Cela résume plus exactement comment le modèle prédit pour chaque niveau de sueil. Plus l'AUC est haute, meilleur est le modèle. Dans notre cas un AUC de 0,833 est acceptable.

Nous pouvons observer la spécificité et la sensibilité pour différents seuils, dans l'objectif de fournir un pronostic concernant l'admissibilité des candidats selon leurs profils. Il est intéréssant de garder un seuil de 0.5 dans le sens où nous ne voulons pas bien prédire une modalité au détriment d'une autre. Plus exactement, nous ne voulons pas avoir une spécificité faible car nous ne voulons pas classer en tant qu'admissibles, des candidats qui ont une probabilité faible de l'être. A l'inverse, avoir une sensibilité faible va classer beaucoup de nos individus potentiellement admissibles en tant que non admissibles, ce n'est pas non plus souhaitable.

5 Modèle Probit

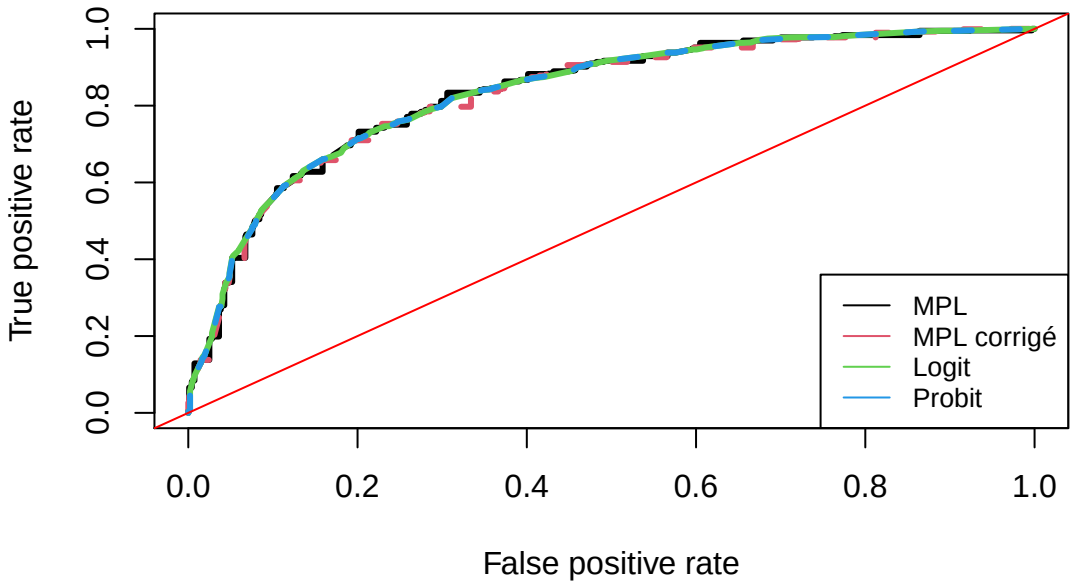
Les résultats du modèle probit sont dans la table 8. Ce modèle très simillaire au modèle logit, à la différence près que le modèle logit va supposer que les termes suivent une loi logistique à l'inverse des termes d'erreurs, qui sont distribués selon une loi normale pour le probit. On retrouve les estimateurs du modèle probit dans la table. On remarque que les paramètres estimés du probit et logit sont très similaires, on a mathémathiquement :

$$\hat{\beta}_{logit} \simeq \hat{\beta}_{probit} \frac{\pi}{\sqrt{3}}$$

.

6 Comparaison de nos modèles

On va à nouveau utiliser la courbe ROC. Nos modèles estimés donnent des résultats en termes de sensibilité et de spécificité assez silimaires.



7 Choix du modèle

Dans un premier, il faut choisir entre le modèle à probabilité linéaire ou les modèles probit et logit. On avait déjà pu évoquer que pour prédire une variable dichotomique, les modèles logistiques étaient plus intéréssants. De plus, la présence d'hétéroscédasticité malgré la correction, nous donne des estimateurs non efficaces. De ce postulat, on choisira soit le modèle probit, soit le modèle logit.

Table 6: Accuracy pour Probit et Logit

	Logit	Probit
Accuracy	0.755	0.755

On se retrouve donc avec deux modèles ayant une accuracy parfaitement identique. On est capable de fournir une probabilité fiable 75% du temps. Proposer cet outil aux étudiants est intéréssant. En effet, même si il est imparfait, il permet d'avoir une idée de la probabilité d'être admissible. On notera que l'utilisation de la variable dichotomique `admissibilité` permet d'avoir une plus faible variabilité dans nos estimateurs à l'inverse de la régression linéaire que nous avons pu faire, qui relie la note aux caractéristiques.

8 Annexes

Table 7: MCO Notes et caractéristiques

	Dependent variable:	
	noteepreuveecrite	
	(1)	(2)
Année-5	−0.054	
	(0.081)	
Année-4	−0.012	
	(0.080)	
Année-3	−0.084	
	(0.078)	
Année-2	−0.015	
	(0.078)	
Année-1	−0.095	
	(0.077)	
sexeH	0.003	
	(0.045)	
nationalitéfrançais	−0.028	
	(0.057)	
retard0	−0.580***	−0.579***
	(0.117)	(0.117)
retard1	−1.011***	−1.013***
	(0.120)	(0.120)
retard2	−1.503***	−1.508***
	(0.128)	(0.127)
retard3	−2.174***	−2.178***
	(0.124)	(0.122)
seriedebacS	−0.032	
	(0.046)	
mentiondebacB	0.162**	0.163**
	(0.082)	(0.082)
mentiondebacP	−0.263***	−0.261***
	(0.049)	(0.049)
mentiondebacTB	0.116	0.117
	(0.165)	(0.164)
formationsuivieDUT	0.404***	0.406***
	(0.104)	(0.104)
formationsuivieMIASHS	1.374***	1.372***
	(0.100)	(0.099)
formationsuivieSEG	0.962***	0.960***
	(0.092)	(0.092)
mentionobtenueB	0.388***	0.390***
	(0.070)	(0.070)
mentionobtenueP	−0.511***	−0.517***
	(0.051)	(0.051)
mentionobtenueTB	0.439***	0.438***
	(0.144)	(0.144)
Constant	12.488***	12.412***
	(0.164)	(0.145)
Observations	1,950	1,950
R^2	0.468	0.467
Adjusted R^2	0.463	0.464
Residual Std. Error	0.989 (df = 1928)	0.988 (df = 1936)
F Statistic	80.892*** (df = 21; 1928)	130.714*** (df = 13; 1936)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 8: Résumé de nos modèles

	Dependent variable:			
	admissibilité			
	MPL	MPL Corrigé	Logit	Probit
	(1)	(2)	(3)	(4)
retard0	-0.135^{***}	-0.122^{***}	-0.816^{**}	-0.476^{**}
	(0.048)	(0.037)	(0.340)	(0.187)
retard1	-0.286^{***}	-0.273^{***}	-1.548^{***}	-0.915^{***}
	(0.050)	(0.041)	(0.342)	(0.189)
retard2	-0.437^{***}	-0.442^{***}	-2.243^{***}	-1.344^{***}
	(0.053)	(0.045)	(0.357)	(0.198)
retard3	-0.636^{***}	-0.611^{***}	-3.401^{***}	-2.019^{***}
	(0.051)	(0.038)	(0.360)	(0.197)
mentiondebacB	0.039	0.037	0.279	0.148
	(0.034)	(0.029)	(0.225)	(0.126)
mentiondebacP	-0.080^{***}	-0.066^{***}	-0.466^{***}	-0.271^{***}
	(0.020)	(0.019)	(0.120)	(0.070)
mentiondebacTB	0.080	0.099^{**}	0.570	0.326
	(0.068)	(0.045)	(0.485)	(0.267)
formationsuivieDUT	0.117^{***}	0.119^{***}	0.567^{**}	0.340^{**}
	(0.043)	(0.037)	(0.260)	(0.151)
formationsuivieMIASHS	0.426^{***}	0.409^{***}	2.181^{***}	1.298^{***}
	(0.041)	(0.037)	(0.249)	(0.144)
formationsuivieSEG	0.291^{***}	0.295^{***}	1.422^{***}	0.854^{***}
	(0.038)	(0.033)	(0.228)	(0.132)
mentionobtenueB	0.109^{***}	0.090^{***}	0.594^{***}	0.350^{***}
	(0.029)	(0.027)	(0.177)	(0.102)
mentionobtenueP	-0.174^{***}	-0.174^{***}	-0.890^{***}	-0.531^{***}
	(0.021)	(0.020)	(0.122)	(0.072)
mentionobtenueTB	0.144^{**}	0.110^{**}	0.826^{**}	0.487^{**}
	(0.060)	(0.048)	(0.384)	(0.218)
Constant	0.613^{***}	0.601^{***}	0.710^{*}	0.414^{*}
	(0.060)	(0.049)	(0.402)	(0.225)
Observations	1,950	1,950	1,950	1,950
R^2	0.334	0.546		
Adjusted R^2	0.330	0.543		
Log Likelihood			-983.984	-983.616
Akaike Inf. Crit.			1,995.967	1,995.232
Residual Std. Error (df = 1936)	0.409	1.073		
F Statistic (df = 13; 1936)	74.735^{***}	178.886^{***}		
Note:				*p<0.1; **p<0.05; ***p<0.01