Projet Marketing et Analyse
Sesmestre 8

Lucas Chaveneau

Contents

Problématique **2**

Statistique descriptive **3**
- La variable `annee` . . . 3
- La variable `sexe` . . . 3
- La variable `nationalite` . . . 4
- La variable `serie_de_bac` . . . 4
- La variable `mention_de_bac` . . . 4
- La variable `retard` . . . 4
- La variable `formation_suivie` . . . 5
- La variable `mention_obtenue` . . . 5
- La variable `note_epreuves_ecrites` . . . 6

Droite de régression linéaire **6**

Modèle à probabilité linéaire. **6**

Modèle logit **7**
- Création de notre modèle : . . . 7
- Rapport de chances . . . 8
- Ajustement du modèle . . . 8
- Courbe ROC . . . 9
- Matrice de confusion . . . 9

Modèle probit **10**
- Courbe ROC . . . 11
- Matrice de confusion . . . 11

Quel modèle choisir ? **11**

Annexe **13**

Problématique

La département d'économie propose un parcours de formation en 3 ans débutant en L3 et conduisant à un Master. Sont éligibles les titulaires d'une deuxième année de licence de sciences économiques (SEG) ou de mathématiques (MIASHS) de même que les détenteurs d'un DUT ou d'un BTS relevant du domaine de l'économie ou de la gestion, sous réserve d'être reçu à un concours comportant des épreuves d'admissibilité et des épreuves orales d'admission.

Les épreuves d'admissibilité comportent une dissertation sur un thème d'actualité, un test de connaissance en économie, ainsi qu'une série d'exercices de mathématiques, de statistiques et de comptabilité. Pour être déclaré admissible, il faut obtenir à ces épreuves une note moyenne supérieure à 12. A peu prés la moitié des 300 à 350 candidats qui se présentent chaque année à ce concours satisfont cette exigence.

La direction du département souhaiterait pouvoir mettre à disposition des étudiants, sur son site web, un système automoatisé d'évaluation synthétique de leurs chances de réussite tenant compte du profil de chacun. Elle se laisse le choix de la manière dont elle formulera ce pronostic sur son site mais vous demande de fournir, pour chaque profil, une estimation de la probabilité d'être déclaré admissible

Statistique descriptive

Nous allons dés à présent étudier les variables de notre base de données selon l'admissibilité au concours.

La variable `annee`

Cette variable représente l'année du concours.

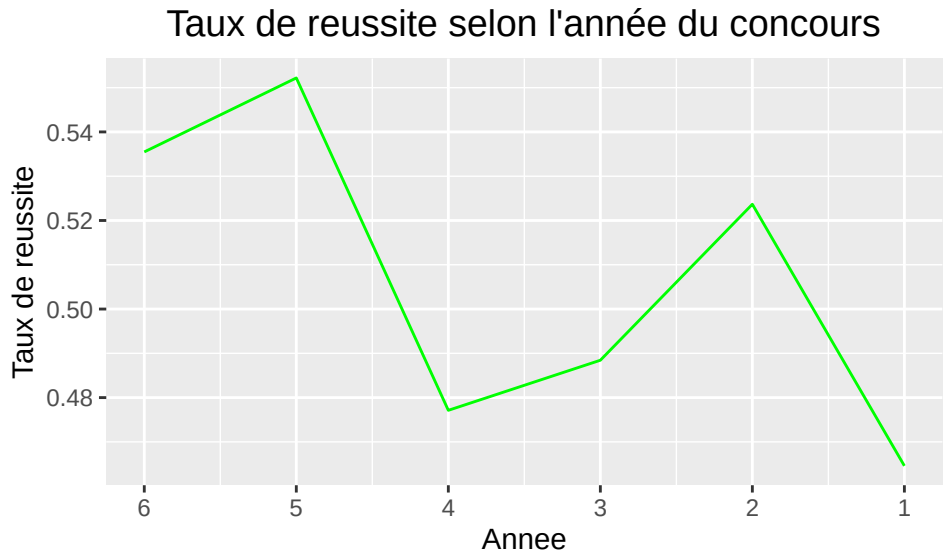
Sur cette partie nous allons tous simplement regarder le taux de réussite selon l'année du concours.

Table 1: Fréquence des réussites au concours

	Concours passé					
	1 ans	2 ans	3 ans	4 ans	5 ans	6 ans
Oui	189	161	177	160	133	144
Non	164	177	169	146	164	166
Total d'inscrit au concours	353	338	346	306	297	310

Nous remarquons que cela ne suit pas une tendance. Le nombre de personnes refusé reste sensiblement le même au fil des années.

Essayons de regarder au plus prêt le taux de réussite des différents concours.



Nous avions un plus haut taux de réussite aux 6éme et 5éme année avant celle-ci. La dernière année assume être celle qui à le plus faible taux de réussite.

La variable `sexe`.

Essayons de regarder si le sexe de l'individu influe sur le taux d'admissibilité.

Table 2: Fréquence des réussites au concours

	Non admis	Admis	Taux de réussite
Femme	464	477	0.507
Homme	500	509	0.504

Nous ne voyons pas de grandes différences entre les taux de réussite des femmes ou des hommes. Le sexe de l'individu n'influe pas empiriquement sur la réussite de l'éxamen

La variable `nationalite`.

Cette variable représente si l'individu est étranger au lieu du concours ou bien s'il est résident au lieu de concours. Essayons de regarder si cette variable a une incidence sur le taux de réussite.

Table 3: Fréquence des réussites au concours

	Non admis	Admis	Taux de réussite
Etranger	183	192	0.512
Français	781	794	0.504

Les français sont beaucoup plus présents que les étrangers à ce concours. Il n'y a pas de différences notables entre les taux de réussites. Le fait d'être étranger ou non, ne change pas empiriquement notre taux de réussite.

La variable `serie_de_bac`

Nous savons pertinemment que les séries de bac les plus représentées dans ce concours sont *S* ou *ES*. Essayons de voir si cela influe sur le taux de réussite de ce concours.

Table 4: Fréquence des réussites au concours

	Non admis	Admis	Taux de réussite
ES	461	561	0.549
S	503	425	0.458

Il semble que les personnes ayant obtenu un bac de type *économique et social* réussisent mieux que les personnes ayant un bac *scientifique*.

La variable `mention_de_bac`

Table 5: Fréquence des réussites au concours

	Non admis	Admis	Taux de réussite
Passable	475	346	0.421
Assez bien	441	459	0.510
Bien	41	149	0.784
très bien	7	32	0.821

Nous pouvons remarquer que le groupe le plus présent sont des personnes ayant eu leur bac avec mention *assez bien*. Bien évidemment, les personnes ayant eu leur bac avec mention *bien* ou *très bien* sont moins présents.

Le taux de réussite augmente si nous avons eu une meilleur mention au bac.

La variable `retard`

Cette variable compte l'écart à l'âge normal pour ce niveau d'étude.

Essayons de voir si plusieurs facteurs qui nous a poussé à redoubler ou à retarder notre passage aux études sont significatifs dans l'admissibilité de ce concours.

Table 6: Fréquence des réussites au concours

	Non admis	Admis	Taux de réussite
3 année de retard	321	58	0.153
2 année de retard	184	82	0.308
1 année de retard	271	245	0.475
Aucun retard	176	530	0.751
Une année d'avance	12	71	0.855

Nous pouvons donc supposer empiriquement que le fait d'avoir pris du retard dans les études nous offre moins de capacités/chances de réussite.

De plus, le taux de réussite est plutôt élevé pour les personnes avec une année d'avance par apport à l'âge normal pour ce niveau d'étude.

La variable `formation_suivie`

Cette variable désigne la formation post-bac de l'individu. Ses modalités sont :

- **BTS** : Brevet de technicien supérieur.
- **DUT** : Diplôme universitaire en technologie.
- **MIASHS** : L2 mathématiques et informatique appliquées aux sciences humaines et sociales.
- **SEG** : L2 économie et gestion.

En voici le descriptif en fonction de l'admissibilité :

Table 7: Fréquence des réussites au concours

Type de diplome			
	Non admis	Admis	Taux de réussite
BTS	105	33	0.239
DUT	189	89	0.320
MIASHS	152	286	0.653
SEG	518	578	0.527

Les personnes en formation *MIASHS* ont le plus haut taux de réussite, suivies par les *SEG*.

Les *BTS* admettent le plus faible taux de réussite.

La variable `mention_obtenue`

Cette variable détaille les mentions obtenues lors de la formation post-bac.

Table 8: Fréquence des réussites au concours

	Non admis	Admis	Taux de réussite
Passable	532	234	0.305
Assez bien	344	483	0.584
Bien	76	229	0.751
très bien	12	40	0.769

Comme pour les mentions du bac, plus nous augmentons notre mention, plus nous réussissons au concours.

La variable `note_epreuves_ecrites`.

Nous n'allons pas décrire les notes des épreuves écrites en fonction de l'admissibilité. Elles sont en effet colinéaires, puisque l'admissibilité dépend de la note aux épreuves écrites.

Droite de régression linéaire

Dans cette partie nous allons réaliser une droite de régression linéaire sur la variable `note_epreuves_ecrites` par la méthode des moindres carrés ordinaire. La variable `admissibilité` est très corrélée à la variable à prédire, nous allons donc l'enlever de nos variables discriminantes. La variable `annee` n'étant pas significatif, nous décidons de l'enlever (l'année du concours n'influe pas sur les notes de l'écrit). De plus, elle ne représente pas les caractéristiques de l'individu, il serait non pertinent de la rajouter.

Le tableau synthétisant les coefficients se trouve dans la **table 12** en annexe.

Les paramètres estimés semblent cohérents avec notre analyse descriptive. Par exemple pour la variable `retard` :

- Le retard -1, décrivant si l'individu est en avance d'un an par apport à l'age normal de l'année d'étude, est la catégorie de référence. Il est donc normal que les coefficients des autres modalités de la variable `retard` soient négatifs. Puisqu'en effet nous avions remarqué un plus haut taux de réussite pour les personnes en avance d'un an.

Les coefficients associés aux variables `formation_suivi`, `mention_obtenue` et `retard` sont significatifs.

Pour la variable `mention_de_bac`, seulement le coefficient associé à la modalité *passable* est significatif. Seulement la modalité passable a une influence sur la note aux épreuves écrites.

Il aurait été pertinent de rajouter une relation entre `mention_obtenue` et `formation_suivi` puis entre `mention_du_bac` et `serie_de_bac` puisque ces variables sont liées. C'est à dire qu'une mention *très bien* pour un *BTS* ne vaut pas la même chose qu'une mention *très bien* dans un *MIASHS*.

Modèle à probabilité linéaire.

A présent nous allons estimer par la méthode des moindres carrés ordinaire, la variable `admissibilite`. La variable `admissibilite` n'est pas une variable quantitative, elle est dichotomique. Ce fait est un problème puisque la méthode des moindres carrés ordinaire n'est pas adaptée pour une variable dépendante dichotomique pour plusieurs points :

- L'hypothèse des normalités des résidus n'est plus tenue.
- Nous sommes en présence d'hétéroscédasticité.
- Les probabilités estimées peuvent être comprise entre $]-\infty;\infty[$.

Comme la question précédente, nous risquons d'avoir des résultats érronnés si nous n'enlevons pas la colinéarité entre `note_epreuves_ecrites` et `admissibilite`. C'est à dire l'omission de `note_epreuves_ecrites` ici.

Nous devons au préalable recoder la variable `admissibilite` en numérique sinon cela ne fonctionnera pas.

Nous avons aussi enlevé la variable `annee` le coefficient associé à cette variable n'était pas significatif.

Les coefficients estimés se trouve dans la **table 13** en annexe.

Les significativités des coefficients sont sensiblement les mêmes qu'à la régression linéaire sur la variable `note_epreuves_ecrites`

Nous devons essayer de corriger l'hétéroscédasticité. Detectons là premièrement, par un test de white ou un test de Breusch Pagan :

Dans le test de Breush pagan :

$$\begin{cases} H_0 : V(\epsilon_i) = \sigma^2 \\ H_1 : V(\epsilon_i) = \sigma_i^2 \end{cases}$$

Notre $p_{value} < 0,05$, l'hypothèse H_0 est rejetée, c'est à dire que nous sommes en présence d'hétéroscédasticité.

Pour corriger l'hétéroscédasticité nous devons passer par le modèle des moindres carrés quasi généralisés (MCQG).

Pour pouvoir faire ce modèle, nous devons pondérer chaque observation par l'inverse de $\sqrt{V(\epsilon_i|X_i)}$.

Cependant nous sommes en présence de certaines valeurs prédites < 0. Ceci est un vrai problème car l'estimation de la variance de l'erreur est : $\widehat{V}(\epsilon_i) = \widehat{p_i}(1 - \widehat{p_i})$. Si $\widehat{p_i} < 0$ alors $\widehat{V}(\epsilon_i) < 0$ et une variance négative est impossible. Il aurait été de même $\forall \widehat{p_i} > 1$ mais nous n'en n'avons pas. Nous les remplacerons par des **NA** et non par des 0 puisqu'il n'est pas possible de calculer $\sqrt{0}$.

Les nouveaux coefficients estimés se trouve dans la **table 14** en annexe.

La variable `serie_de_bac` est devenue significative, ainsi que la modalité *TB* de la variable `mention_de_bac`. Il n'y a pas d'autres différences notables entre la significativité des coefficients.

La valeur des coefficients restent logiques.

Nous avons aussi un meilleur R^2 même s'il n'est pas un excellent indicateur de la qualité du modèle.

Modèle logit

Dans cette partie nous découperons notre base de données en deux :

- L'une pour créer des données d'entrainement : le modèle va apprendre de ces données.
- L'autre pour tester le modèle.

Création de notre modèle :

Nous ferons comme les modèles précédent :

- Nous enlevons la variable `note_epreuves_ecrites` pour cause de colinéarité.
- Nous enlevons la variable `annee` car son coefficient n'est pas significativement différent de 0.

La tableau synthétisant la régression logistique se retouve dans la **table 15** en annexe.

Seules les variables complètes `sexe` et `nationalité` ont des coefficients non significativements différents de 0. Nous pouvions le pressentir dans notre analyse descriptive. Les conclusions sur la significativité des coefficients restent sensiblement les mêmes qu'au modèle à probabilité linéaire.

Les coefficients ont des signes cohérents.

Dans ce modèle **logit** nous pouvons dire que la mention de bac *passable* a une influence négative par apport à la catégorie de référence *assez bien*.

Nous pouvons aussi affirmer que le fait d'avoir pas ou des années de retard par apport à l'age normal de l'année d'étude a une influence négative par rapport au fait d'avoir un an d'avance.

Rapport de chances

Table 9: Rapport de chances et intervalle de confiance

	OR	2.5 %	97.5 %
(Intercept)	1.78	0.46	7.56
sexeH	0.69	0.47	1.01
nationalitefrançais	0.83	0.51	1.34
as.factor(retard)0	0.61	0.19	1.65
as.factor(retard)1	0.40	0.12	1.10
as.factor(retard)2	0.11	0.03	0.34
as.factor(retard)3	0.06	0.02	0.17
serie_de_bacS	0.80	0.54	1.18
mention_de_bacB	2.70	1.13	7.11
mention_de_bacP	0.75	0.50	1.13
mention_de_bacTB	2.14	0.51	11.52
formation_suivieDUT	1.95	0.82	4.87
formation_suivieMIASHS	8.99	3.84	22.32
formation_suivieSEG	4.31	2.00	9.87
mention_obtenueB	2.07	1.15	3.83
mention_obtenueP	0.39	0.25	0.59
mention_obtenueTB	4.08	1.02	21.47

Grâce à ce tableau nous pouvons dire que dans ce modèle :

- Une personne, qui a 3 ans de retard, a 16 fois ($\frac{1}{0.06}$) moins de chances, par apport à quelqu'un qui a un an d'avance, de réussir au concours.
- Une personne, qui a fait un *DUT*, a $1,95$ fois plus de chances de réussir le concours qu'une personne qui a fait un *BTS*.
- Une personne qui a eu la mention *très bien* dans leurs études supérieurs a 4.08 fois plus de chances de réussir au concours qu'une personne qui a eu la mention *assez bien*
- Du fait de la significativité du coefficient associé a la modalité *homme* de la variable `sexe`, nous pouvons dire qu'être un homme ne donne pas plus de chances de réussir le concours.

Ajustement du modèle

Nous devons à présent tester la significativité de notre modèle en prenant un modèle contraint avec juste une constante et un modèle non contraint.

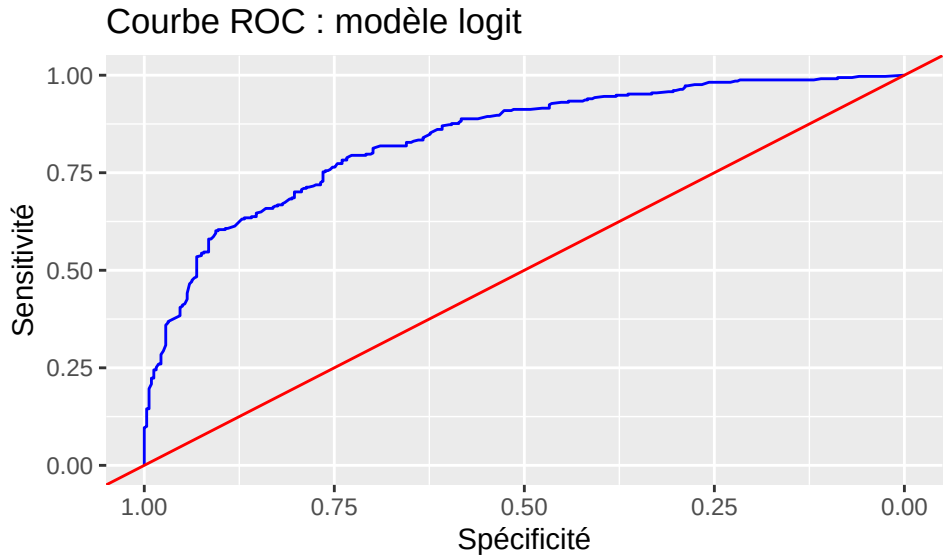
$$\begin{cases} H_0 : \beta_0 = \beta_1 = ... = \beta_j = 0 \\ H_1 : \exists j \ \text{ tel que } \ \beta_j \neq 0 \end{cases}$$

Table 10: Résultats du test de la significativité d'un modèle

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
#Df	2	9.000	11.314	1	5	13	17
LogLik	2	−386.717	90.110	−450.435	−418.576	−354.859	−323.000
Df	1	16.000		16.000	16.000	16.000	16.000
Chisq	1	254.870		254.870	254.870	254.870	254.870
Pr(>Chisq)	1	0.000		0.000	0.000	0.000	0.000

La p_{value} est inférieure à $\alpha = 0.05$, nous rejetons l'hypothèse H_0. Notre modèle est meilleur qu'un modèle contraint.

Courbe ROC

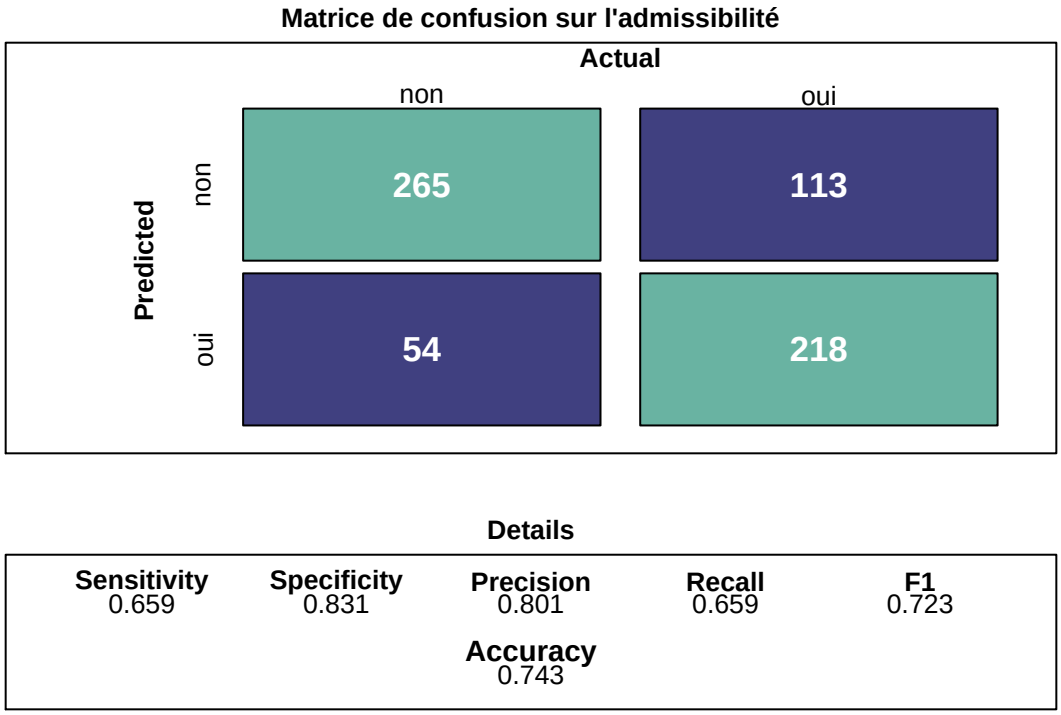


La courbe ROC représente le taux de bonnes prédictions sur la modalité *non* par apport au taux de bonnes prédictions *oui*, si on change le seuil d'acceptation.

La courbe ROC est une bonne mesure de la qualité d'un modèle. L'AUC est l'air sous la courbe ROC, il faut qu'elle soit maximale.

Matrice de confusion

Passons à présent à la matrice de confusion



Notre modèle **logit** admet un taux d'erreur de $1 - 0,743 = 0,257$. Il estime mauvaisement dans l'ordre de 25%

Il a estimé 54 *oui* alors qu'elle était des *non*.

Modèle probit

Le modèle probit suppose que les termes d'erreurs du modèle suivent une loi logistique alors que le modèle probit suppose que ses termes d'erreurs suivent une loi normal.

Les modèles probit et logit sont liés. En effet, les coefficients associés aux variables discriminantes suivent cette relation :

$$\hat{\beta}_j^{logit} = \frac{\pi}{\sqrt{3}} \hat{\beta}_j^{probit}$$

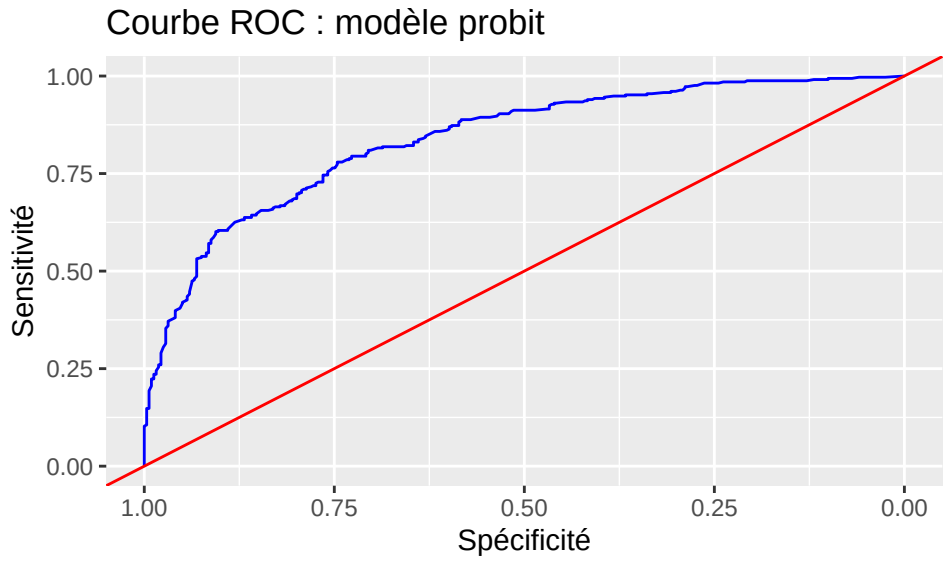
Nous allons à présent estimer la variable `admissibilite` par un modèle probit.

Le tableau synthétisant les résultats du modèle se trouve dans **table 16** en annexe

Nous pouvons remarquer qu'il y a moins de coefficients, associés aux variables discriminantes, significatifs.

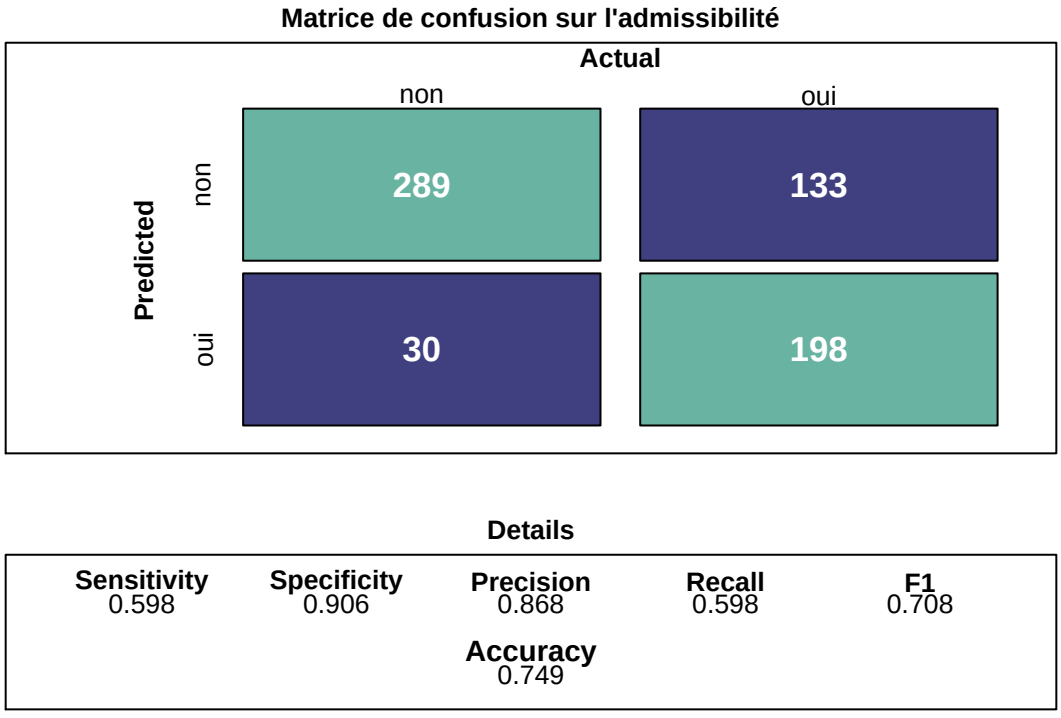
Les signes des coefficients restent logiques.

Courbe ROC



La courbe ROC est sensiblement identique à la courbe ROC du modèle logit.

Matrice de confusion



Quel modèle choisir ?

Dans un premier temps, il ne serait pas pertinent de choisir les modèles à probabilité linéaire, même si nous avons essayé de corriger l'hétéroscedasticité par la méthode des MCQG. En effet, le modèle à probabilité

linéaire va à l'encontre de plusieurs hypothèses vues précédemment.

Nous allons donc comparer les erreurs sur ces deux types de modèles : probit et logit.

Table 11: Erreurs selon le modèle

	Logit	Probit
Erreur total	0.2569231	0.2507692
Erreur sur non	0.1692790	0.0940439
Erreur sur oui	0.3413897	0.4018127

Les deux modèles admettent le même taux d'erreur total. Sur l'ensemble des données ils font le même nombre d'erreur de prédiction. Cependant, ces erreurs se focalisent pas sur la même modalité.

Le modèle logit admet une erreur de 17% sur les *non* alors que probit admet une erreur de 9.5%. Le modèle logit fait plus d'erreur sur la modalité *non*. Au contraire, logit admet une erreur de 34% sur les *oui* contre 40% pour le modèle probit. Notre modèle logit estime mieux les *oui* que notre modèle probit.

Du fait de la quasi égalité de nos erreurs total, le modèle choisi dépendra de ce que nous voulons :

- Si nous voulons bien prédire les *non*, il faudrat privilégier le modèle probit.
- Si nous voulons bien prédire les *oui*, nous choisirons donc le modèle logit.

Annexe

Table 12: modèle de régression linéaire

	Dependent variable:
	note_epreuves_ecrites
sexeH	0.004
	(0.045)
as.factor(retard)0	−0.580***
	(0.117)
as.factor(retard)1	−1.012***
	(0.120)
as.factor(retard)2	−1.505***
	(0.127)
as.factor(retard)3	−2.174***
	(0.123)
nationalitefrançais	−0.030
	(0.057)
serie_de_bacS	−0.029
	(0.046)
mention_de_bacB	0.160*
	(0.082)
mention_de_bacP	−0.261***
	(0.049)
mention_de_bacTB	0.106
	(0.165)
formation_suivieDUT	0.408***
	(0.104)
formation_suivieMIASHS	1.379***
	(0.100)
formation_suivieSEG	0.964***
	(0.092)
mention_obtenueB	0.389***
	(0.070)
mention_obtenueP	−0.515***
	(0.051)
mention_obtenueTB	0.442***
	(0.144)
Constant	12.442***
	(0.156)
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 13: modèle de régression linéaire par MCQG

	Dependent variable:
	admissibilite
sexeH	-0.022
	(0.019)
nationalitefrançais	-0.022
	(0.024)
as.factor(retard)0	-0.137^{***}
	(0.048)
as.factor(retard)1	-0.288^{***}
	(0.050)
as.factor(retard)2	-0.438^{***}
	(0.053)
as.factor(retard)3	-0.637^{***}
	(0.051)
serie_de_bacS	-0.010
	(0.019)
mention_de_bacB	0.037
	(0.034)
mention_de_bacP	-0.080^{***}
	(0.020)
mention_de_bacTB	0.077
	(0.068)
formation_suivieDUT	0.117^{***}
	(0.043)
formation_suivieMIASHS	0.428^{***}
	(0.041)
formation_suivieSEG	0.292^{***}
	(0.038)
mention_obtenueB	0.109^{***}
	(0.029)
mention_obtenueP	-0.174^{***}
	(0.021)
mention_obtenueTB	0.146^{**}
	(0.060)
Constant	0.647^{***}
	(0.065)
Observations	1,950
R^2	0.335
Adjusted R^2	0.330
Residual Std. Error	0.409 (df = 1933)
F Statistic	60.862^{***} (df = 16; 1933)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 14: modèle à probabilité linéaire

	Dependent variable:
	admissibilite
sexeH	0.009
	(0.016)
nationalitefrançais	−0.033
	(0.021)
as.factor(retard)0	−0.121***
	(0.038)
as.factor(retard)1	−0.271***
	(0.042)
as.factor(retard)2	−0.444***
	(0.046)
as.factor(retard)3	−0.586***
	(0.039)
serie_de_bacS	−0.038**
	(0.017)
mention_de_bacB	0.035
	(0.029)
mention_de_bacP	−0.053***
	(0.019)
mention_de_bacTB	0.090**
	(0.046)
formation_suivieDUT	0.157***
	(0.040)
formation_suivieMIASHS	0.421***
	(0.040)
formation_suivieSEG	0.295***
	(0.037)
mention_obtenueB	0.086***
	(0.028)
mention_obtenueP	−0.194***
	(0.021)
mention_obtenueTB	0.129***
	(0.048)
Constant	0.635***
	(0.055)
Observations	1,927
R^2	0.522
Adjusted R^2	0.518
Residual Std. Error	1.117 (df = 1910)
F Statistic	130.457*** (df = 16; 1910)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 15: modèle logit

	Dependent variable:
	admissibilite
sexeH	-0.215^{*}
	(0.114)
nationalitefrançais	-0.117
	(0.143)
as.factor(retard)0	-0.306
	(0.310)
as.factor(retard)1	-0.559^{*}
	(0.316)
as.factor(retard)2	-1.308^{***}
	(0.339)
as.factor(retard)3	-1.720^{***}
	(0.331)
serie_de_bacS	-0.128
	(0.116)
mention_de_bacB	0.591^{**}
	(0.255)
mention_de_bacP	-0.166
	(0.122)
mention_de_bacTB	0.491
	(0.449)
formation_suivieDUT	0.392
	(0.265)
formation_suivieMIASHS	1.313^{***}
	(0.259)
formation_suivieSEG	0.885^{***}
	(0.236)
mention_obtenueB	0.433^{**}
	(0.176)
mention_obtenueP	-0.553^{***}
	(0.129)
mention_obtenueTB	0.807^{*}
	(0.424)
Constant	0.341
	(0.407)
Observations	650
Log Likelihood	-322.565
Akaike Inf. Crit.	679.129
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 16: modèle probit

	Dependent variable:
	admissibilite
sexeH	-0.215^{*}
	(0.114)
nationalitefrançais	-0.117
	(0.143)
as.factor(retard)0	-0.306
	(0.310)
as.factor(retard)1	-0.559^{*}
	(0.316)
as.factor(retard)2	-1.308^{***}
	(0.339)
as.factor(retard)3	-1.720^{***}
	(0.331)
serie_de_bacS	-0.128
	(0.116)
mention_de_bacB	0.591^{**}
	(0.255)
mention_de_bacP	-0.166
	(0.122)
mention_de_bacTB	0.491
	(0.449)
formation_suivieDUT	0.392
	(0.265)
formation_suivieMIASHS	1.313^{***}
	(0.259)
formation_suivieSEG	0.885^{***}
	(0.236)
mention_obtenueB	0.433^{**}
	(0.176)
mention_obtenueP	-0.553^{***}
	(0.129)
mention_obtenueTB	0.807^{*}
	(0.424)
Constant	0.341
	(0.407)
Observations	650
Log Likelihood	-322.565
Akaike Inf. Crit.	679.129
Note:	*p<0.1; **p<0.05; ***p<0.01