

Access to credit

Aybuké BICAT & Corentin DUCLOUX



¶ Table des matières :

Sujet	2
Analyse statistique	2
Lien entre crédit et niveau de revenu	2
Lien entre crédit et genre	3
Lien entre crédit et appartenance à la population active	3
Lien entre crédit et carte de débit/crédit	3
Lien entre crédit & âge	4
Analyse des corrélations	5
Analyse des correspondances multiples	5
Modèle à probabilité linéaire	7
Sans correction d'hétéroscédasticité	7
Avec correction	8
Logit	9
Rapports de chance	10
Ajustement du modèle	11
Courbe ROC	11
Matrice de confusion	12
Probit	12
Courbe ROC	13
Matrice de confusion	13
Comparaison des résultats	14

Sujet

Les données proviennent de la base **Global Findex 2017** de la *Banque Mondiale*. Elles sont collectées dans 148 pays. Dans chaque pays, 1000 individus ont été interrogés.

L'objectif est de repérer les déterminants de l'accès au crédit ~ variable à expliquer `fin19`.

Commençons par importer les données :

```
df_credit <- read_dta("~/R data/findex_Germany.dta")
```

Une analyse de la première variable `economy` permet de remarquer que le seul pays observé pour tous les individus est l'Allemagne.

Analyse statistique

On commence par observer les effectifs par modalité de la **variable à expliquer** :

<i>Effectif des individus ayant un crédit:</i>	
Variable <code>fin19</code>	<i>n</i>
1	228
2	767

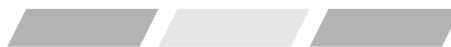
- On remarque que seulement 22.9% des individus de l'échantillon ont un *crédit immobilier*.
- Les proportions des modalités de `fin19` sont déséquilibrées ⇒ **Cela va avoir une incidence lorsque l'on va vouloir estimer la variable avec un modèle.**

Lien entre crédit et niveau de revenu

La variable `inc_q` est découpée en 5 classes proportionnelles (quintiles) déterminant le niveau de revenu des individus ⇒ 1 correspondant aux 20% les plus pauvres et 5 correspondant aux 20% les plus riches.

On peut se demander si faire partie des individus les plus riches influe dans la contraction d'un crédit :

<i>Crédit en fonction du niveau de revenu</i>		
Revenu	Crédit	Pas de crédit
1	23	154
2	49	146
3	44	144
4	43	148
5	69	175



- On remarque que les individus situés dans le quintile 1 des revenus (20% les plus pauvres) ne possèdent pas de crédit dans 87% des cas.
- La plus haute proportion des individus possédant un crédit est située dans le quintile 5 (20% les plus riches) avec 28.3%.

Lien entre crédit et genre

<i>Crédit en fonction du sexe</i>		
Sexe	Crédit	Pas de crédit
1	112	374
2	116	393

Note: 1 si Homme, 2 si Femme

- On remarque qu'il n'y a pas de différence significative dans les proportions entre les hommes et les femmes concernant la contraction d'un crédit (23%) & (22.8%)

Lien entre crédit et appartenance à la population active



<i>Crédit en fonction de la présence sur le marché du travail</i>		
Employé	Crédit	Pas de crédit
0	68	278
1	160	489

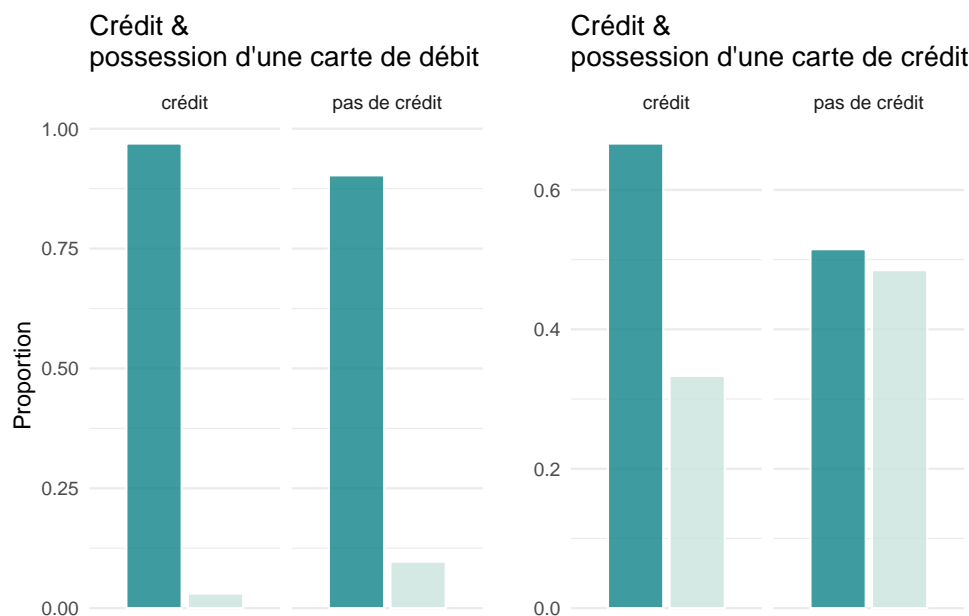
Note: 1 si employé, 0 sinon

- Les individus qui ne sont pas présents sur le marché du travail (chômeurs, mineurs, ou à la retraite) ont un crédit dans 19.7% des cas, soit une proportion de 5% inférieure aux actifs.
- Plusieurs effets peuvent ici intervenir :
 - Si un individu est âgé, il est probable qu'il ait déjà eu des crédits auparavant et qu'il a terminé de les rembourser.
 - Si un individu est mineur, il est logique qu'il ne puisse pas faire de crédit.
 - Si un individu est au chômage, sa capacité d'emprunt est réduite.

Lien entre crédit et carte de débit/crédit

On peut se demander si posséder une **carte de crédit** ou une **carte de débit** influe dans le fait d'avoir contracté un crédit :

- La barre  distingue la proportion d'individus ayant une carte de crédit/débit
- La barre  distingue la proportion d'individus n'ayant pas de carte de crédit/débit



- La proportion d'individus ne possédant pas de crédit est plus conséquente lorsque ceux-ci ne possèdent pas de carte de crédit, tandis qu'il ne semble pas y avoir un lien significatif entre carte de débit & crédit.

Lien entre crédit & âge

Contracter un crédit n'est pas indispensable pour tout le monde ! En effet, lorsqu'un individu est à la retraite, il est très probablement déjà propriétaire, dès lors, il n'a pas besoin de contracter de crédit.

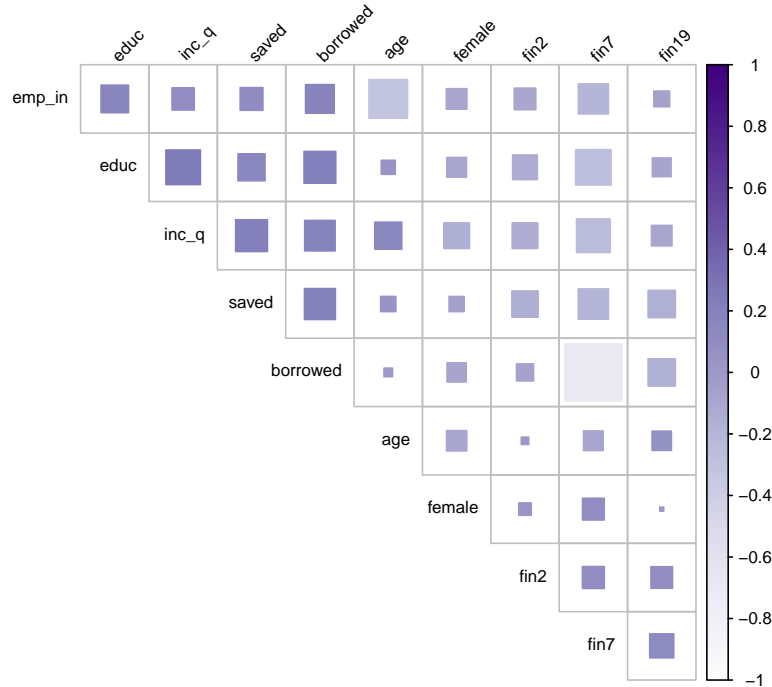
- On découpe la variable **age** en 5 classes égales :

Crédit en fonction de l'appartenance à une classe d'âge

Classe d'âge	Crédit	Pas de crédit
[15,30]	18.5%	81.5%
[30,45]	32%	68%
[45,60]	27.5%	72.5%
[60,75]	14.3%	85.7%
[75,90]	4%	96%

- On remarque effectivement que les classes d'âge les plus concernées par le crédit sont **[30,45]** & **[45,60]**.
- On observe une baisse significative du taux de crédit passé cet âge.

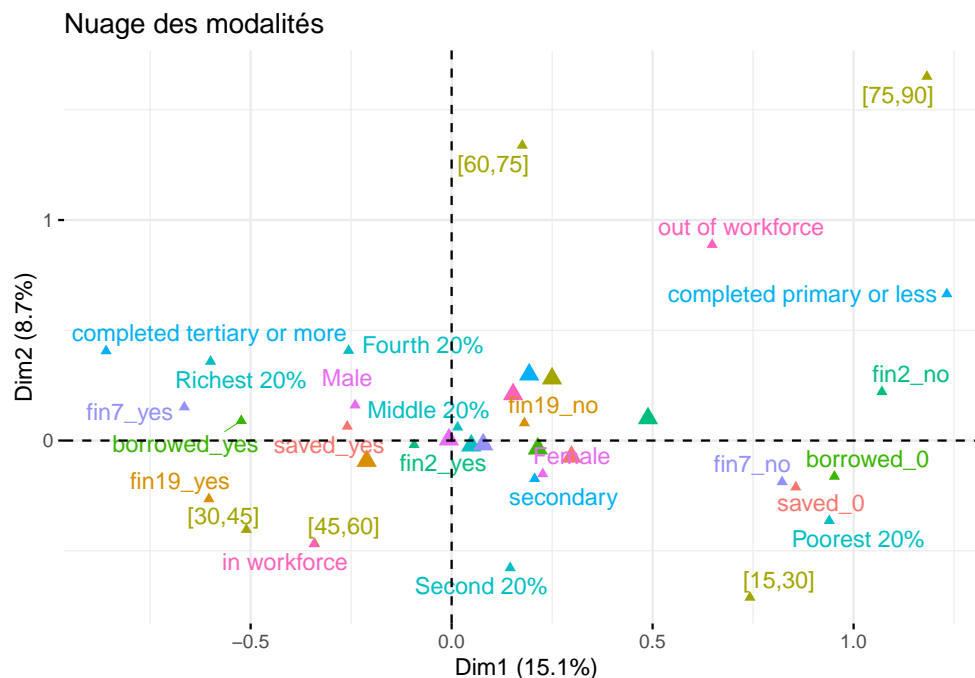
Analyse des corrélations



- La variable **fin7** et la variable **borrowed** sont fortement (négativement) corrélées : **-0.7**.
- **Intuition** : sans carte de crédit, impossible d'emprunter \Rightarrow Les individus concernés n'ont peut-être pas de compte ouvert dans un établissement bancaire.

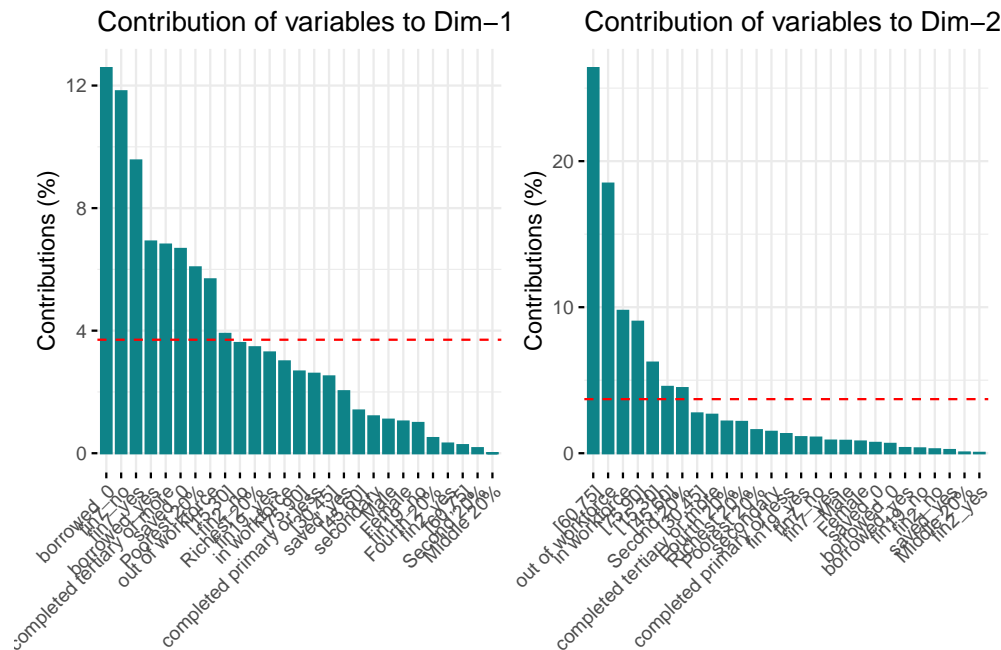
Analyse des correspondances multiples

Remarque : On retire les variables **fin33** & **fin46** de l'analyse car elles ont une proportion de valeurs non-renseignées égales à respectivement 39.6% & 61.1%



D'après le nuage des modalités, on peut voir que 2 groupes distincts apparaissent :

- Le groupe qui a contracté un crédit, où on retrouve un profil de personnes qui sont des hommes entre 30 et 60 ans et qui ont fait des études supérieures. Ce sont aussi les personnes les plus aisées puisqu'elles sont associées aux modalités de revenu les plus élevées soit de **Middle 20%** à **Richest 20%**. Ils ont également une carte de crédit et de débit, et ont déjà épargné.
- Le deuxième groupe se compose majoritairement de femmes, qui se sont arrêtées à l'enseignement secondaire et qui ont un revenu moyen, voire faible (2e et 3e quantile). Dans ce groupe, les individus n'ont pas contracté de crédit mais possèdent une carte de débit.



Pour les contributions des modalités aux axes, on peut voir que:

- Les variables **borrowed**, **fin7**, **emp_in**, la modalité **completed tertiary or more** (soit avoir complété l'enseignement supérieur), la modalité **Poorest 20%** de la variable **inc_q** et enfin la tranche d'âge la plus jeune [15,30] contribuent à la construction de l'axe F_1 .
- Toutes les modalités de l'âge sauf la tranche [30,45] contribuent à la construction de l'axe F_2 , ainsi que la variable **emp_in** et la modalité **Second 20%** venant de **inc_q**.

Modèle à probabilité linéaire

Méthode des moindres carrés ordinaires

L'objectif est désormais d'estimer la variable **fin19** (ie l'accès au crédit) par la méthode des MCO. Cependant, la variable n'est pas *quantitative* mais *dichotomique*.

$$\text{fin19} = \begin{cases} 1 & \text{si l'individu a un crédit} \\ 2 & \text{si l'individu n'a pas de crédit} \end{cases}$$

Cela va impliquer plusieurs violations d'hypothèses :

- Les MCO conduisent à des estimations distribuées entre $] -\infty ; +\infty[\Rightarrow$ L'estimation par MCO est donc biaisée.
- Non-normalité des résidus : $\epsilon_i \not\sim \mathcal{N}(0, \sigma^2)$.
- Les termes d'erreurs ϵ_i n'auront pas la même variance pour toutes les valeurs des X_i : on est en présence d'**hétéroscédasticité**.

Sans correction d'hétéroscédasticité

On a vu précédemment que les variables **fin7** et **borrowed** étaient très corrélées. Nous allons donc retirer la variable **fin7** (risque de multicolinéarité).

TRANSFORMATIONS :

- Nous renommons la variable à expliquer **fin19** en **credit** et changeons les modalités (1, 2) \Rightarrow (0, 1). **Dans ce cas, la modalité 0 correspond à avoir un crédit tandis que la modalité 1 correspond à ne pas avoir de crédit.**
- Nous utiliserons la variable **age** transformée en classes pour capturer les effets spécifiques d'appartenance à une classe d'âge.
- Les variables **female** et **educ** n'ont pas été incluses dans le modèle car leurs coefficients associés ne sont pas significatifs (*ce dont on pouvait se douter grâce à l'analyse descriptive*).
- Pour rendre la lecture des résultats plus claire, nous transformons toutes les variables dichotomiques et catégorielles en *factor* avant la spécification du modèle.

$$\begin{aligned} \text{credit} = & \beta_0 + \beta_1 \text{age}_{[30-45]} + \beta_2 \text{age}_{[45-60]} + \beta_3 \text{age}_{[60-75]} + \beta_4 \text{age}_{[75-90]} + \\ & \beta_5 \text{inc_q}_2 + \beta_6 \text{inc_q}_3 + \beta_7 \text{inc_q}_4 + \beta_8 \text{inc_q}_5 + \\ & \beta_9 \text{employed}_1 + \beta_{10} \text{saved}_1 + \beta_{11} \text{borrowed}_1 + \epsilon \end{aligned}$$

- Les coefficients associés à **saved** et **borrowed** sont significatifs au seuil de 1%, les coefficients associés à **income_quantile5** et à **age[30,45]** sont significatifs au seuil de 5%. Le signe de toutes ces variables est **négatif** : la probabilité de ne pas avoir de crédit diminue lorsque l'individu a emprunté l'année précédente, fait des économies, est situé dans la classe d'âge de 30 à 45 ans ou fait partie des plus riches.

Table 1: Modèle à probabilité linéaire

	Dependent variable:
	credit
age[30,45]	-0.104** (0.041)
age[45,60]	-0.064 (0.039)
age[60,75]	0.116*** (0.044)
age[75,90]	0.191*** (0.069)
income_quintile2	-0.083* (0.043)
income_quintile3	-0.066 (0.044)
income_quintile4	-0.064 (0.044)
income_quintile5	-0.104** (0.043)
employed1	0.094*** (0.033)
saved1	-0.114*** (0.032)
borrowed1	-0.097*** (0.029)
Constant	0.935*** (0.046)
Observations	984
R ²	0.080
Adjusted R ²	0.069
Residual Std. Error	0.407 (df = 972)
F Statistic	7.652*** (df = 11; 972)
Note: *p<0.1; **p<0.05; ***p<0.01	

- Les coefficients associés aux modalités **age[60,75]**, **age[75,90]** et à la variable **employed** sont significatifs au seuil de 1% mais leur signe est **positif** : la probabilité de ne pas avoir de crédit augmente lorsque l'individu fait partie des classes d'âge ci-dessus, ou lorsqu'il est employé.

Avec correction

Nous devons en premier lieu détecter la présence d'hétéroscédasticité : on peut le faire avec le test de *Breusch – Pagan* :

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

La *p – value* du test est inférieure à 0.05, c'est à dire que l'hypothèse H_0 est rejetée et qu'il y a de l'hétéroscédasticité.

Le modèle à probabilité linéaire fournit une variance connue du terme d'erreur à utiliser avec la méthode des **MCP (Moindres Carrés Pondérés)**, c'est à dire $V(\epsilon_i) = p_i(1 - p_i)$. Dans ce cas il faut pondérer chaque observation par $\frac{1}{\sqrt{V(\epsilon_i X_i)}}$.

- Il faut cependant préalablement vérifier qu'aucune des variances estimées ne soit négative : une façon d'éviter les probabilités < 0 ou > 1 est de les limiter à l'intervalle $[0, 1]$.

Table 2: Régression linéaire par méthode des MCP

	<i>Dependent variable:</i>
	credit
age[30,45]	−0.109*** (0.040)
age[45,60]	−0.077** (0.036)
age[60,75]	0.079** (0.031)
age[75,90]	0.092** (0.036)
income_quintile2	−0.138*** (0.031)
income_quintile3	−0.052* (0.031)
income_quintile4	−0.060** (0.029)
income_quintile5	−0.101*** (0.032)
employed1	0.072*** (0.025)
saved1	−0.089*** (0.022)
borrowed1	−0.072*** (0.022)
Constant	0.939*** (0.032)
Observations	984
R ²	0.130
Adjusted R ²	0.120
Residual Std. Error	1.041 (df = 972)
F Statistic	13.240*** (df = 11; 972)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

On remarque cette fois que l'on a beaucoup plus de coefficients très significatifs et avec des valeurs légèrement différentes. Leur signe ne change pas. On remarque aussi que le R^2 a presque doublé par rapport à la régression précédente.

Logit

- En comparaison avec les modèles de régression linéaire, on voit que le coefficient associé à la variable `income_quintile` est non significativement différent de 0 pour le quintile 4.
- Les variables les plus significatives restent `saved1` et `borrowed1`, avec une influence négative sur la probabilité de ne pas avoir un crédit par rapport aux catégories de référence `saved0` et `borrowed0` (c'est à dire une influence *positive* sur la probabilité d'avoir un crédit).
- Les quintiles 2 & 5 sont assez significatifs, avec une influence positive sur la probabilité d'avoir un crédit par rapport à la catégorie de référence (**20% les plus pauvres**).

Table 3: Logit

<i>Dependent variable:</i>	
credit	
age[30,45]	−0.564** (0.251)
age[45,60]	−0.376 (0.245)
age[60,75]	0.776*** (0.300)
age[75,90]	1.998*** (0.765)
income__quintile2	−0.578** (0.292)
income__quintile3	−0.497* (0.298)
income__quintile4	−0.481 (0.301)
income__quintile5	−0.699** (0.288)
employed1	0.575*** (0.205)
saved1	−0.857*** (0.240)
borrowed1	−0.617*** (0.189)
Constant	2.524*** (0.343)
Observations	984
Log Likelihood	−487.288
Akaike Inf. Crit.	998.576

Note: *p<0.1; **p<0.05; ***p<0.01

Rapports de chance

Table 4: Odds Ratio selon les variables

	OR	p-value
(Intercept)	12.477 ***	0
age[30,45]	0.569 **	0.025
age[45,60]	0.687	0.126
age[60,75]	2.172 ***	0.01
age[75,90]	7.377 ***	0.009
income__quintile2	0.561 **	0.047
income__quintile3	0.609 *	0.096
income__quintile4	0.618	0.111
income__quintile5	0.497 **	0.015
employed1	1.778 ***	0.005
saved1	0.424 ***	0
borrowed1	0.54 ***	0.001

- Un individu dans la classe d'âge [30, 45] a $\frac{1}{0.569} \simeq 1.8$ fois plus de chance d'avoir un crédit qu'un individu situé dans la catégorie de référence [15, 30].
- Un individu dans la classe d'âge [60, 75] a 2.17 fois plus de chance de ne pas avoir de crédit qu'un individu situé dans la catégorie de référence [15, 30]. De la même manière, un individu dans la classe d'âge [75, 90] a 7.38 fois plus de chance de ne pas avoir de crédit.



- Par rapport aux **20% les plus pauvres**, les **20% les plus riches** ont $\frac{1}{0.497} \simeq 2$ fois plus de chance d'avoir un crédit.
- Un individu ayant épargné l'année précédente a $\frac{1}{0.424} \simeq 2.4$ fois plus de chance d'avoir un crédit qu'un individu n'ayant pas épargné. La même observation peut être réalisée si l'individu a emprunté en $N - 1$: l'individu aura dans ce cas $\frac{1}{0.54} \simeq 1.85$ fois plus de chance d'avoir un crédit qu'un individu n'ayant pas emprunté l'année précédente.

Ajustement du modèle

On effectue le test d'adéquation de Hosmer-Lemeshow pour évaluer l'adéquation du modèle **Logit** à nos données. Les hypothèses du test sont les suivantes :

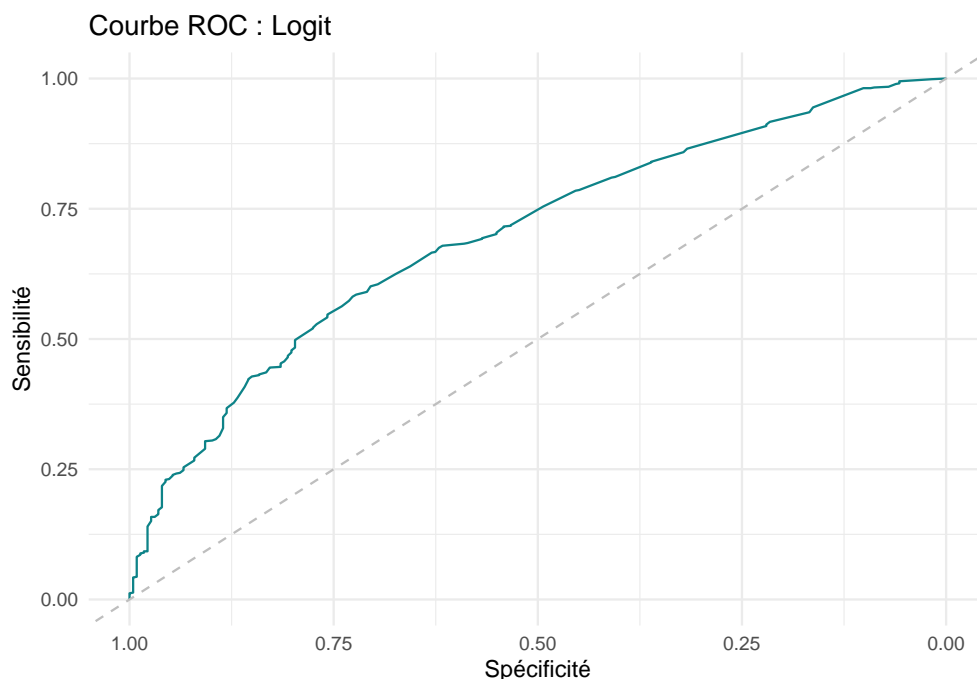
$$\begin{cases} H_0 : \text{Le modèle s'ajuste bien aux données} & | \text{ si } p > 0.05 \\ H_1 : \text{Le modèle s'ajuste mal aux données} & | \text{ si } p < 0.05 \end{cases}$$

- On obtient une $p - \text{value} = 0.9$, celle-ci étant bien supérieure à 0.05, on conserve l'hypothèse H_0 : le modèle est bien ajusté.

Courbe ROC

La courbe ROC est une mesure de la performance du modèle logit à travers le taux de faux positifs & de faux négatifs.

- On recherche généralement une courbe ROC proche du coin supérieur gauche (1, 0), car dans ce cas le modèle ne fait pas d'erreur.



La mesure AUC , l'aire sous la courbe, est une mesure globale de qualité du modèle \Rightarrow plus celle-ci est proche de 1, plus la qualité du modèle est grande.

- On a ici un $AUC = 0.7 \Rightarrow$ Le modèle n'est pas très performant.

Matrice de confusion

Table 5: Matrice de confusion associée au Logit

	$\hat{credit}_i = 0$	$\hat{credit}_i = 1$
$credit_i = 0$	13	214
$credit_i = 1$	7	750

- On retrouve le problème de *classe déséquilibrée* évoqué dans la première partie. En effet, le modèle prédit beaucoup mieux les individus qui n'ont pas de crédit que les individus qui ont effectivement un crédit.

Probit

Spécifications liées au modèle

La différence entre le modèle **Logit** et **Probit** est que dans le modèle **Probit**, les termes d'erreur $\epsilon_i \sim \mathcal{N}(0, 1)$. La fonction de répartition de l'erreur est donc :

$$P_i = \int_{-\infty}^{\beta_0 + \beta_i x_i} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

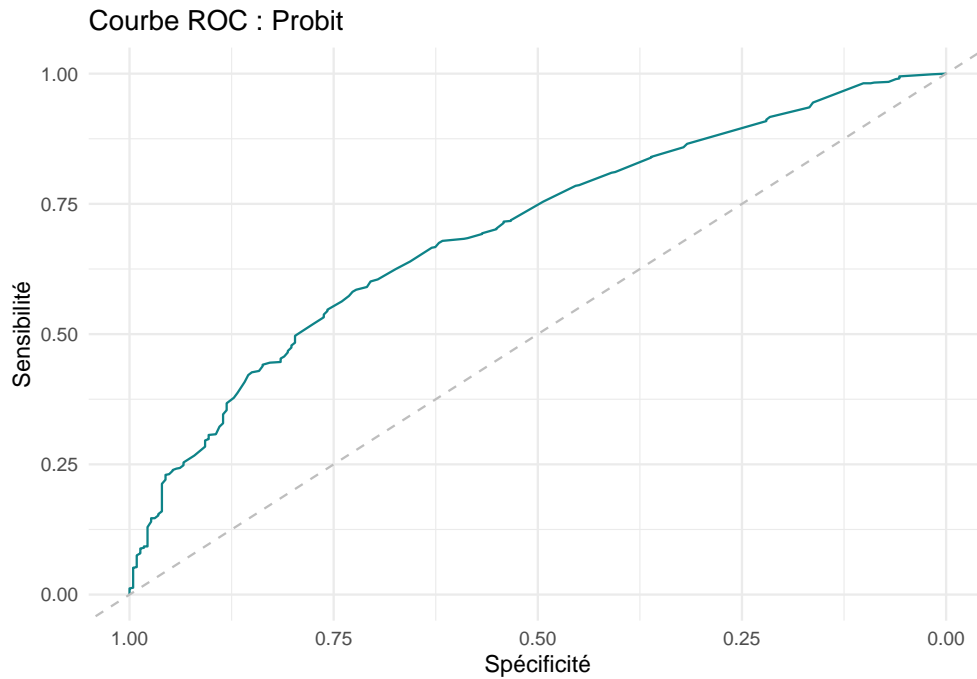
Table 6: Probit

<i>Dependent variable:</i>	
credit	
age[30,45]	-0.330** (0.146)
age[45,60]	-0.217 (0.141)
age[60,75]	0.440*** (0.167)
age[75,90]	0.998*** (0.352)
income_quintile2	-0.349** (0.164)
income_quintile3	-0.293* (0.167)
income_quintile4	-0.289* (0.169)
income_quintile5	-0.410** (0.162)
employed1	0.332*** (0.119)
saved1	-0.474*** (0.129)
borrowed1	-0.352*** (0.106)
Constant	1.474*** (0.186)
Observations	984
Log Likelihood	-487.655
Akaike Inf. Crit.	999.309

Note: *p<0.1; **p<0.05; ***p<0.01

- Les signes et la significativité des coefficients sont sensiblement les mêmes que le modèle **Logit** étudié précédemment.

Courbe ROC



Matrice de confusion

Table 7: Matrice de confusion associée au Probit

	$\hat{credit}_i = 0$	$\hat{credit}_i = 1$
$credit_i = 0$	10	217
$credit_i = 1$	3	754

- Le modèle **Probit** prédit **encore mieux** les individus n'ayant pas de crédit (comparé au modèle **Logit**)... cependant, ce n'est pas ce qui nous intéresse le plus.



Comparaison des résultats

Quelques informations

Avant même de comparer les résultats, on peut exclure les modèles à probabilité linéaire des modèles candidats puisque ceux-ci violent de nombreuses hypothèses, même lorsque l'hétéroscédasticité a été corrigée.

On va utiliser le **critère d'informations d'Akaike (AIC)** pour départager les 4 modèles.

Ce critère mesure la **qualité de prédiction** d'un modèle en comparant son erreur de prédiction aux informations apportées par son nombre de paramètres.

$$AIC = \ln \left(\frac{SCR_e}{T} \right) + \frac{2(p+q)}{T}$$

<i>Comparaison des AIC</i>	
Modèle	AIC
Linéaire	1035.47
Linéaire corrigé	959.88
Logit	998.58
Probit	999.31

Entre le modèle **Logit** & **Probit**, on préfère donc le modèle **Logit** puisque celui-ci a l'*AIC* le plus faible et qu'il classifie mieux les personnes ayant un crédit.

De plus, l'*AUC* du modèle **Logit** est légèrement plus élevé.

Table 8: Comparaison des modèles

	<i>Dependent variable:</i>			
	credit			
	<i>OLS</i>	<i>logistic</i>	<i>probit</i>	
	(1)	(2)	(3)	(4)
age[30,45]	−0.104** (0.041)	−0.109*** (0.040)	−0.564** (0.251)	−0.330** (0.146)
age[45,60]	−0.064 (0.039)	−0.077** (0.036)	−0.376 (0.245)	−0.217 (0.141)
age[60,75]	0.116*** (0.044)	0.079** (0.031)	0.776*** (0.300)	0.440*** (0.167)
age[75,90]	0.191*** (0.069)	0.092** (0.036)	1.998*** (0.765)	0.998*** (0.352)
income_quintile2	−0.083* (0.043)	−0.138*** (0.031)	−0.578** (0.292)	−0.349** (0.164)
income_quintile3	−0.066 (0.044)	−0.052* (0.031)	−0.497* (0.298)	−0.293* (0.167)
income_quintile4	−0.064 (0.044)	−0.060** (0.029)	−0.481 (0.301)	−0.289* (0.169)
income_quintile5	−0.104** (0.043)	−0.101*** (0.032)	−0.699** (0.288)	−0.410** (0.162)
employed1	0.094*** (0.033)	0.072*** (0.025)	0.575*** (0.205)	0.332*** (0.119)
saved1	−0.114*** (0.032)	−0.089*** (0.022)	−0.857*** (0.240)	−0.474*** (0.129)
borrowed1	−0.097*** (0.029)	−0.072*** (0.022)	−0.617*** (0.189)	−0.352*** (0.106)
Constant	0.935*** (0.046)	0.939*** (0.032)	2.524*** (0.343)	1.474*** (0.186)
Observations	984	984	984	984
R ²	0.080	0.130		
Adjusted R ²	0.069	0.120		
Log Likelihood			−487.288	−487.655
Akaike Inf. Crit.			998.576	999.309
Residual Std. Error (df = 972)	0.407	1.041		
F Statistic (df = 11; 972)	7.652***	13.240***		

Note:

*p<0.1; **p<0.05; ***p<0.01