Econométrie des variables qualitatives I

M1 EKAP et M2 CODEME

Muriel Travers
LEMNA, IAE Economie et Management
Université de Nantes

www.univ-nantes.fr

UNIVERSITÉ DE NANTES

Organisation

Cours : 18h – TD : 12 h

Contact :

muriel.travers@univ-nantes.fr ; Bureau 227

Modalités de contrôle :

- Dossier sur cas pratique (groupe de 2 étudiants) : écrit et présentation orale de 20 minutes avec PPT

Prérequis :

- Connaissance de base du logiciel R
- Si possible, installation de Rstudio

Plan et bibliographie

Plan du cours

Introduction

Chapitre 1: Modèle de choix binaire (Logit/Probit)

Chapitre 2 : Modèle multinomial ordonné

Chapitre 3 : Modèle multinomial non ordonné

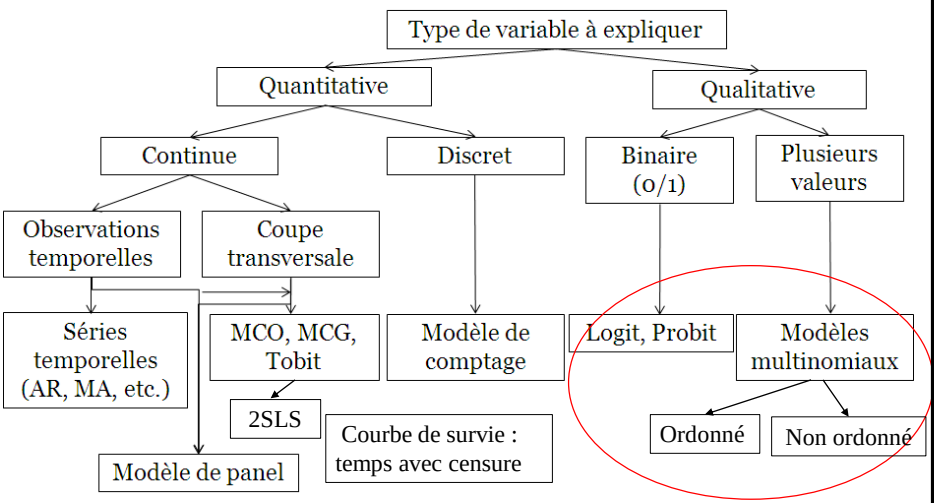
Bibliographie

Logit/Probit : Araujo C., Brun J.F, Combes J.L, *Bréal*, 2008
Logit/Probit : Bourbonnais, *Économétrie*, Dunod, 2005
Green W., *Econometrie*, Pearson, 2005

Introduction

Rappel des différents modèles économétriques

➔ Les chapitres vus dans ce cours sont ceux entourés par le cercle rouge



Intérêt de la modélisation des modèles de choix

➔ Individus amenés dans sa vie à faire des choix à titre personnel ou professionnel

➔ Intéressant pour un décideur de connaître les choix et leurs déterminants afin de prédire les probabilités d'occurrence du phénomène étudié

Exemple de modèle de choix binaire (Logit ou probit) (variable à expliquer : binaire 1/0)

➔ Analyse des facteurs expliquant pourquoi les personnes interrogées ont un CAP nul ou positif pour la mise en place d'une nouvelle ligne de tramway ou de la création d'un nouveau lieu culturel.

Exemples de modèles multinomiaux

Lorsque la variable à expliquer prend plusieurs valeurs différentes ➔ Utilisation de modèles multinomiaux

➔ Lorsque la variable à expliquer est une variable qualitative dont les modalités (alternatives) ne peuvent être classées les unes par rapport aux autres, on parle de modèle multinomial non ordonné

Exemple 1 : / Transport

Enquête menée auprès d'un échantillon d'individus où on leur demande de préciser quel est leur moyen de transport habituel pour se rendre sur leur lieu de travail (bus, tramway, voiture, à pied, vélo)

➔ Déterminer les facteurs expliquant tel ou tel choix de transport

On peut également faire une enquête pour connaître les déterminants du choix de transport entre le train, l'avion, le bus.

➔ Possibilité d'utiliser ces modèles pour ensuite voir quel pourrait être l'impact d'une hausse du prix de l'essence ou à une hausse du prix du ticket du tramway/bus ?
➔ Dans une optique environnementale, connaître la part de l'avion et ses déterminants peuvent être intéressants pour connaître les dépenses énergétiques.

Exemple 2 : / Environnement

➔ Enquête réalisée auprès de pêcheurs à ligne pour connaître la probabilité qu'ils fréquentent tels ou tels sites en fonction :

- des caractéristiques de ces sites,
- du coût pour se rendre aux sites
- le fait que le pêcheur fréquente un site alternatif
- des caractéristiques des pêcheurs (âge, catégorie socio-professionnelle, genre)

Exemple 3 : / Culture

➔ Enquête auprès de personnes d'Angers pour connaître leur distraction au cours de la semaine, afin de connaître les déterminants de tel ou tel choix.

* La télévision
* Le théâtre
* Le cinéma
* La discothèque

Les variables explicatives sont deux types :

- des variables dont les valeurs prises varient d'un individu à l'autre (revenu, âge, genre, etc.)

- des variables variant selon les alternatives (temps de transport, prix du transport, etc.)

➔ Selon la nature des variables explicatives, on parlera de modèles multinomiaux :

- simples : fonction uniquement des caractéristiques des individus
- conditionnels : fonction uniquement des variables variant en fonction des alternatives
- « généraux » : modèle combinant les deux types de variables explicatives

➔ Lorsqu'il existe une hiérarchie naturelle entre les modalités de la variable à expliquer, on parle alors de **modèle (polytomique) ordonné**.

Exemple 1 :

➔ Etude des déterminants du choix de la taille d'un soda (petite, moyenne, grande) commandé dans un fast-food.

Choix fonction du type de sandwich commandé, de la commande ou non de frites, de l'âge du consommateur, etc.

Grande taille > Moyenne taille > Petite taille

Exemple 2 :

➔ Analyse des déterminants de l'obtention de médaille olympique (or > argent > bronze) dans une discipline donnée en fonction du nombre d'heures d'entrainement, du fait de suivre ou non un régime particulier avant l'épreuve, de l'âge de l'athlète, de la popularité de la discipline dans le pays de l'athlète, etc.

Exemple 3 :

➔ Etude des facteurs influençant la possibilité qu'un étudiant de l'université s'inscrive en 2ème cycle

1 : peu plausible
2 : probable
3 : très probable

➔ Choix peut être lié :

- au niveau de scolarité des parents (1 si au moins des parents a un diplôme d'université, 0 sinon)
- si l'établissement actuel de l'étudiant est un établissement privé ou public (1 si établissement public, 0 sinon)
- de la moyenne de ses notes actuelles allant de 1 à 5 (1 la note la plus mauvaise, 5 la meilleure)

Chapitre 1:
Modèle de choix binaire Logit/Probit

1) Objectif

➔Expliquer les valeurs d'une variable qualitative de type binaire Y à partir de variables explicatives $X=(X_1,\ldots,X_p)$ qualitatives et/ou quantitatives.

➔ Cherche à expliquer un phénomène ne pouvant prendre que deux modalités :

$y_i=1$ si le phénomène étudié est observé pour l'individu i

$y_i=0$ dans le cas contraire

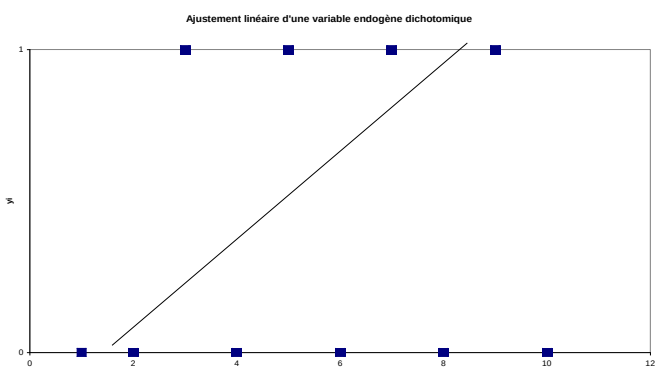
Exemples :

- payer ou non pour un programme d'aménagement du territoire
- voter ou ne pas voter pour un candidat
- être malade ou ne pas l'être
- consommation d'un bien ou non, etc.

2) Spécification du modèle

➔ Méthode des MCO non applicable

- Valeurs prédites peuvent être en dessous de 0 et au-dessus 1.
- Du fait de la nature dichotomique de la variable, le nuage de points se situe sur la droite y=0, soit sur la parallèle y=1.



➔ Modélisation, **non pas de la variable à expliquer mais de la probabilité que celle-ci prenne la valeur 1 ou 0.**

➔ Supposons que la réalisation de $y_i = 1$ soit plutôt associée à des valeurs élevées des variables explicatives x_j, celle de $y_i = 0$ à des valeurs faibles des variables explicatives x_j.

➔ Existe une valeur seuil dépendant de la combinaison linéaire

$$\beta_j x_{ij} \quad \forall i=1,...,n \;\; j=1,...,k$$

Pour modéliser cette probabilité, on suppose qu'il existe une variable latente telle que :

$$\begin{cases} y_i = 1 \;\; si \;\; y_i^* > c \\ y_i = 0 \;\; si \;\; y_i^* \leq c \end{cases}$$

où la variable y^*_i dépend linéairement d'un certain nombre de variables explicatives qui peuvent être quantitatives et/ou qualitatives.

$$y_i^* = cste + \sum_{j=1}^{k} \beta_j x_{ij} + u_i = x_i\beta + u_i \qquad i = 1,...,n$$

➔ La règle de décision est probabiliste :

$$\begin{cases} Prob(y_i = 1) = Prob(x_i\beta + u_i > c) = Prob(u_i > c - x_i\beta) \\ Prob(y_i = 0) = Prob(x_i\beta + u_i \leq c) = Prob(u_i \leq c - x_i\beta) \end{cases}$$

Par convention, c = 0 (puisqu'il existe une constante dans le vecteur β)

$$\begin{cases} Prob(y_i = 1) = Prob(u_i > -x_i\beta) = Prob(u_i \leq x_i\beta) = \text{Fonction cumulative}(u_i) \\ Prob(y_i = 0) = Prob(u_i \leq -x_i\beta) = Prob(u_i > x_i\beta) \end{cases}$$

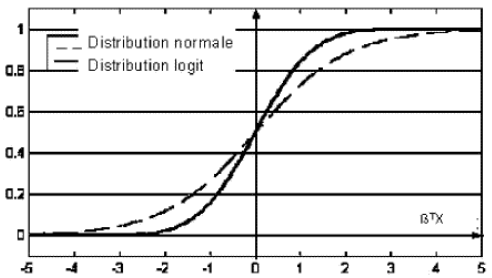
➔ Le type d'estimation à utiliser dépend donc de l'hypothèse faite sur la distribution du terme d'erreur.

➔ Modèle Logit si les termes d'erreur sont supposés suivre une loi logistique
Modèle Probit si ces derniers sont supposés suivre une loi normale

Logit : $\begin{cases} \text{Prob}(y_i = 1) = \text{Prob}(u_i \le x_i\beta) = F(x_i\beta) \\ \text{Prob}(y_i = 0) = 1 - \text{Prob}(u_i \le x_i\beta) = 1 - F(x_i\beta) \end{cases}$

Où : F est la fonction de répartition d'une loi logistique : $F(x_i\beta) = \dfrac{e^{x_i\beta}}{1+e^{x_i\beta}}$

Probit : $F(x_i\beta) = \int_{-\infty}^{x_i\beta} \dfrac{1}{\sqrt{2\pi}} \exp\left(-\dfrac{z^2}{2}\right) dz$



3) Estimation des coefficients du modèle

➔ Estimation des coefficients du modèle obtenus via la méthode du maximum de vraisemblance.

➔ La vraisemblance d'une observation correspond à la probabilité d'observer les données.

Elle s'écrit pour l'ensemble des individus i de la manière suivante :

$$L(Y,\beta) = \prod_{i=1}^{n} \text{Prob}(Y = y_i / X = x_i) = \prod_{i=1}^{n} p_i^{y_i} \left(1 - p_i\right)^{1-y_i}$$

Soit :

$$L(Y,\beta) = \prod_{i=1}^{n} F(x_i\beta)^{y_i} \left(1 - F(x_i\beta)\right)^{1-y_i}$$

D'où le calcul de la log-vraisemblance (*Log-Likelihood*)

$$\text{LogL}(Y,\beta) = \sum_{i=1}^{n} y_i \times \ln\left(F(x_i\beta)\right) + (1-y_i) \times \ln\left(1-F(x_i\beta)\right)$$

➔ Choix des valeurs estimés telle que cette log-vraisemblance soit maximale, c'est-à-dire parmi tous les modèles possibles on choisit celui qui rend l'observation de l'échantillon la plus probable

$$\underset{\beta}{\text{Max}}\ \text{LogL}(Y,\beta) = \frac{\partial \text{LogL}(Y,\beta)}{\partial \beta} = G(\beta) = 0$$

$$\Leftrightarrow \frac{\partial\left(\sum_{i=1}^{n} y_i \times \ln\left(F(x_i\beta)\right) + (1-y_i) \times \ln\left(1-F(x_i\beta)\right)\right)}{\partial \beta} = 0$$

➔ Le vecteur des paramètres à estimer est donc défini par une résolution d'un système de K équations non linéaires en β

➔ Pas possible d'estimer le maximum de la vraisemblance directement d'où l'utilisation d'algorithme d'optimisation numérique pour l'estimation des coefficients des variables explicatives.

Remarque : la condition suffisante pour que le maximum de Log L(Y,β) existe est que Log (L(Y,β)) soit concave.

Or : $$\text{LogL}(Y,\beta) = \sum_{i=1}^{n} y_i \times \ln\left(F(x_i\beta)\right) + (1-y_i) \times \ln\left(1-F(x_i\beta)\right)$$

$$= \sum_{y_{i=1}} \ln\left(F(x_i\beta)\right) + \sum_{y_{i=0}} \ln\left(1-F(x_i\beta)\right)$$

Il suffit alors de montrer que les 2 termes soient concaves

Soit le modèle Logit. Dans ce cas :

$$\ln\left(F\left(x_i\beta\right)\right)=\ln\left(\frac{e^{x_i\beta}}{1+e^{x_i\beta}}\right)=\ln\left(e^{x_i\beta}\right)-\ln\left(1+e^{x_i\beta}\right)=x_i\beta-\ln\left(1+e^{x_i\beta}\right)$$

$$\frac{\partial\ln\left(F\left(x_i\beta\right)\right)}{\partial\beta}=\frac{x_i\beta-\ln\left(1+e^{x_i\beta}\right)}{\partial\beta}=x_i^{'}-\frac{\left(x_i^{'}e^{x_i\beta}\right)}{1+e^{x_i\beta}}=\frac{x_i^{'}\left(1+e^{x_i\beta}\right)-\left(x_i^{'}e^{x_i\beta}\right)}{1+e^{x_i\beta}}$$

$$=\frac{x_i^{'}+x_i^{'}e^{x_i\beta}-\left(x_i^{'}e^{x_i\beta}\right)}{1+e^{x_i\beta}}=\frac{x_i^{'}}{1+e^{x_i\beta}}$$

$$\frac{\partial^2\ln\left(F\left(x_i\beta\right)\right)}{\partial^2\beta}=\frac{\partial\left(\frac{x_i^{'}}{1+e^{x_i\beta}}\right)}{\partial\beta}=-x_i^{'}\frac{x_i^{'}e^{x_i\beta}}{\left(1+e^{x_i\beta}\right)^2}$$

Strictement inférieur à 0 donc $\ln\left(F\left(x_i\beta\right)\right)$ concave

$$\ln\left(1-F\left(x_i\beta\right)\right)=\ln\left(1-\frac{e^{x_i\beta}}{1+e^{x_i\beta}}\right)=\ln\left(\frac{1+e^{x_i\beta}-e^{x_i\beta}}{1+e^{x_i\beta}}\right)=\ln\left(\frac{1}{1+e^{x_i\beta}}\right)$$

$$=\ln 1-\ln\left(1+e^{x_i\beta}\right)=-\ln\left(1+e^{x_i\beta}\right)$$

$$\frac{\partial\left(\ln\left(1-F\left(x_i\beta\right)\right)\right)}{\partial\beta}=-\frac{\partial\left(\ln\left(1+e^{x_i\beta}\right)\right)}{\partial\beta}=-\frac{x_i^{'}e^{x_i\beta}}{1+e^{x_i\beta}}$$

$$\frac{\partial^2\left(\ln\left(1-F\left(x_i\beta\right)\right)\right)}{\partial^2\beta}=\frac{\partial\left(-\frac{x_i^{'}e^{x_i\beta}}{1+e^{x_i\beta}}\right)}{\partial\beta}=-x_i^{'}\frac{\partial\left(\frac{e^{x_i\beta}}{1+e^{x_i\beta}}\right)}{\partial\beta}$$

$$=-x_i^{'}\left(\frac{x_i^{'}e^{x_i\beta}\left(1+e^{x_i\beta}\right)-e^{x_i\beta}\left(x_i^{'}e^{x_i\beta}\right)}{\left(1+e^{x_i\beta}\right)^2}\right)=-x_i^{'}\frac{x_i^{'}e^{x_i\beta}}{\left(1+e^{x_i\beta}\right)^2}$$

Strictement inférieur à 0 donc également concave Log L(Y,β) concave

Les paramètres β peuvent être obtenus par un processus itératif :

Pour cela il faut définir :
- Des valeurs initiales des paramètres β pour amorcer le processus itératif
- Une règle de passage d'un β au suivant
- Une règle d'arrêt du processus (Variation du β entre itération actuelle et précédente < seuil de tolérance)

Dans le cas d'un modèle logit : les valeurs initiales du vecteur β sont les valeurs de β obtenus dans le modèle linéaire

Règles de passage :

- Règle de Newton Raphson : $\hat{\beta}_i = \hat{\beta}_{i-1} - \lambda_{i-1} H\left(\hat{\beta}_{i-1}\right)^{-1} \times G\left(\hat{\beta}_{i-1}\right)$

$$G\left(\hat{\beta}_{i-1}\right) = \frac{\partial \text{LogL}\left(Y, \hat{\beta}_{i-1}\right)}{\partial \hat{\beta}_{i-1}} \qquad \lambda_{i-1} \in [0,1]$$

$$H\left(\hat{\beta}_{i-1}\right) = \frac{\partial G\left(\hat{\beta}_{i-1}\right)}{\partial \hat{\beta}_{i-1}} = \frac{\partial^2 \left(\text{LogL}\left(Y, \hat{\beta}_{i-1}\right)\right)}{\partial \hat{\beta}_{i-1} \partial \hat{\beta}_{i-1}'}$$

Matrice hessienne du log vraisemblance

Règle de la méthode du scoring :

$$\hat{\beta}_i = \hat{\beta}_{i-1} + \left[I\left(\hat{\beta}_{i-1}\right) \right]^{-1} \times G\left(\hat{\beta}_{i-1}\right)$$

$$I\left(\hat{\beta}_{i-1}\right) = -E\left[\frac{\partial^2 \left(\text{LogL}\left(Y, \hat{\beta}_{i-1}\right)\right)}{\partial \hat{\beta}_{i-1} \partial \hat{\beta}_{i-1}'} \right]$$

Matrice d'information de Fisher

Règle des méthodes de quasi newton (ex :=méthode bfgs) : basée sur des matrices hessiennes approximées à chaque itération

Propriété de l'estimateur par maximum de vraisemblance : il est le plus efficace asymptotiquement

à savoir :
- non biaisé asymptotiquement (convergent)
- ayant la plus petite variance

1) Convergent : $\hat{\beta} \xrightarrow[n\to\infty]{} \beta_0$ (vraie valeur)

2) Asymptotiquement efficace : $V\left(\hat{\beta}\right) = \left[I\left(\beta_0\right)\right]^{-1}$

3) De plus, l'EMV est asymptotiquement normalement distribué :

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow[n\to\infty]{} N\left[\beta_0, \left[I\left(\beta_0\right)\right]^{-1}\right] \quad \text{avec} \quad I\left(\beta_0\right) = -E\left[\frac{\partial^2\left(\text{LogL}\left(Y,\beta\right)\right)}{\partial\beta\partial\beta'}\right]_{\beta=\beta_0}$$

Remarque : la variance dépend de la vraie valeur β_0 : mais comme β_0 pas connue, on estime la variance à la valeur estimée de β

$$\hat{V}\left(\hat{\beta}\right) = \left[I\left(\hat{\beta}\right)\right]^{-1}$$

4) Interprétation des résultats

Attention : les valeurs numériques des coefficients n'ont pas d'interprétation directe dans le cas des méthodes d'estimation Logit et Probit.

➔ Valeurs des coefficients dépendent du codage utilisé pour la variable dépendante. Celui-ci est arbitraire : les valeurs obtenues seraient différentes pour tout autre codage.

En revanche, leur signe et le fait qu'ils soient ou non significatifs sont interprétables.

➔ Le signe permet de savoir si la probabilité de y_i est une fonction croissante ou décroissante de la variable explicative correspondante (toutes choses égales par ailleurs).

➔ Vérification au préalable de l'influence des variables explicatives sur les variations de la variable dépendante à expliquer.

$$H_0 : \hat{\beta}_1 = 0, \ldots, \hat{\beta}_k = 0$$

➔ Ce test se réalise à partir de la statistique du **ratio de vraisemblance** définie de la manière suivante :

$$2(LL(\beta)-LL(\beta_c)) \rightarrow \chi^2_{1-\alpha}(k)$$

Où :

k est le nombre de variables explicatives du modèle estimé (non contraint) hors constante
LL(β) : valeur de la log vraisemblance lorsque le modèle inclut l'ensemble des variables explicatives du modèle
$LL(\beta_c)$: valeur de la log vraisemblance lorsque le modèle ne comporte que la constante

Dans le cas où l'on teste l'hypothèse nulle de l'ensemble des coefficients des variables explicatives, on va calculer : (Null Deviance-Residual Deviance)
Car Residual Deviance = $-2LL(\beta)$ et Null Deviance = $-2LL(\beta_c)$ dans le cas des modèles logit/probit

➔ **Si la probabilité est < 0,05** (valeur calculée > valeur théorique) cela signifie que l'hypothèse de nullité de tous les coefficients est refusée au seuil de risque de 5%
➔ Le modèle présenté peut donc être retenu pour le seuil de risque choisi.

* **<u>Significativité des variables explicatives</u>** :

- Utilisation du test de Wald qui se lit de la même manière que le t de Student.

- Si la p-value est < 0,05 (ou 0,1) , on refuse au seuil de risque de 5 % (ou de 10%) la nullité du coefficient du paramètre analysé.

$H_0 : \beta_j = 0$
$H_1 : \beta_j \neq 0$

$$\text{Wald} = z = \frac{\hat{\beta}_j^2}{\hat{\sigma}_j^2} \geq \chi^2_{1-\alpha}(1)$$

➔ Refus de l'hypothèse H_0 pour le seuil de risque choisi

*** Effet marginal d'une variable explicative**

➔ L'effet marginal d'une variation de la variable X_j sur la probabilité de l'événement Y=1 se calcule pour une variable quantitative continue de la manière suivante :

$$\frac{\partial F(x_i\beta)}{\partial Xj} = F'(x_i\beta)\times\beta_j$$

Modèle Logit :

$$\frac{\partial F(x_i\beta)}{\partial X_j} = \frac{\beta_j e^{x_i\beta}}{\left(1+e^{x_i\beta}\right)^2}$$

Modèle Probit :

$$\frac{\partial F(x_i\beta)}{\partial X_j} = \frac{\beta}{\sqrt{2\pi}}\exp\left(-\frac{x_i\beta^2}{2}\right)$$

*** Calcul du odd-ratio**

Possibilité de calculer ce que l'on appelle le **odd-ratio dans le cas du modèle Logit** : calcul du surcroît de chance (ou de risque) d'observer le phénomène étudié lorsque la variable explicative étudiée varie.

Exemple :

Soit le modèle (estimé) cherchant à expliquer le ronflement à partir du genre et du tabagisme.

$$P_{estimé} = -2{,}1972 + 1{,}5863*\text{homme} + 1{,}1856*\text{tabac}$$

➔ Surcroit de risque de ronfler lorsque l'on est un homme : exp(1,5863) = 4,89

Un homme a donc environ 5 fois plus de risque de ronfler qu'une femme

➔ Surcroit de risque de ronfler lorsque l'on est fumeur : exp(1,1856) = 3,27

Un individu qui fume a 3,3 fois plus de chance de ronfler qu'un individu non-fumeur.

*** Qualité d'ajustement du modèle :**

➔ Lorsque les variables explicatives ont une influence sur la variable dépendante, il est intéressant de connaître la **qualité de l'ajustement** du modèle.

Cette qualité est mesurée par le R² de Mc Fadden :

Valeur pour le modèle avec les variables explicatives

$$R^2_{\text{Mc Fadden}} = 1 - \frac{LL(\beta)}{LL(\beta_C)} = 1 - \frac{\text{Residual Deviance}}{\text{Null Deviance}}$$

Valeur pour le modèle avec uniquement la constante.

➔ Cet indicateur est compris entre 0 et 1. Lorsqu'il est proche de 1, le modèle est quasi parfait

***Critère d'information d'Akaike (AIC) :**

* S'applique aux modèles estimés par la méthode du maximum de vraisemblance.

* On peut augmenter la vraisemblance du modèle en ajoutant une variable explicative supplémentaire.

* Le critère d'information d'Akaike pénalise les modèles en fonction du nombre de variables explicatives afin de satisfaire le critère de parcimonie : compromis entre la qualité de l'ajustement et la complexité du modèle, en pénalisant les modèles ayant un grand nombre de variables explicatives.

Soit M un modèle logistique à p variables explicatives et n observations.

$$AIC(M) = -2LL_n(\widehat{\beta_n}) + 2p$$

Le meilleur modèle sera donc celui possédant l'AIC le plus faible

***Critère d'information bayésien (BIC) :**

Ce critère est défini de la manière suivante :

$$BIC(M) = -2LL_n(\widehat{\beta_n}) + pln(n)$$

Le modèle retenu sera celui qui minimise le BIC.

Remarque :

Lorsque $ln(n) > 2$ (donc pour n ≥ 8), le BIC aura tendance à choisir des modèles plus parcimonieux que AIC.

* **Qualité du modèle en termes de prévision : matrice de confusion**

Comparaison des probabilités observées et estimées

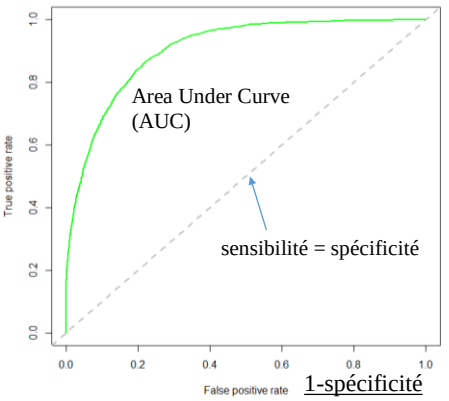
	Probabilité estimée		Total
Probabilité observée (échantillon)	0	1	
0	a (vrai négatif)	b (faux positif)	a+b
1	c (faux négatif)	d (vrai positif)	c+d
Total	a+c	b+d	N=a+b+c+d

Indicateur de sensibilité (capacité à prédire un évènement, taux de vrai positif): d/c+d
Indicateur de spécificité (capacité à prédire un non-événement, taux de vrai négatif): a/a+b
Indicateur de précision : a+d/(a+b+c+d)
Taux de faux positif =b/a+b = 1- spécificité
Taux de faux négatif = c/c+d
Taux d'erreur = c+b/N

➔ Un bon modèle a des valeurs faibles de taux d'erreur et de faux positifs + des valeurs élevées de sensibilité, de précision et de spécificité.

*** Courbe ROC (Receiver Operator Characteristic)**

Sensibilité



1-spécificité

Un bon modèle aura un fort AUC (forte sensibilité et forte spécificité)

Certains graphiques présentent la spécificité (abscisse) : dans ce cas, l'axe va de 1 → 0

* **Comparaison Probit *versus* Logit**

➔ Relativement similaires en termes d'ajustement statistique.

➔ Les normalisations utilisées pour les estimations ne sont pas les mêmes :

Modèle Probit : réalisé sous l'hypothèse : $\sigma = 1$

Modèle Logit : $\sigma = \pi/\sqrt{3}$

➔ Coefficients « Logit » et « Probit » comparés si les coefficients « Probit » sont multipliés par $\pi / 3^{1/2}$ (ou de diviser les coefficients « Logit » par $\pi / 3^{1/2}$).

➔ Amemiya (1981) préconise de multiplier les coefficients du modèle Probit par 1,6 fois pour obtenir ceux du modèle Logit.

➔ Cette approximation fonctionnera d'autant mieux que la valeur moyenne de Y est proche de 0,5.

5) Estimation : Application sous R

Exemple : Quels sont les facteurs permettant d'expliquer l'utilisation d'engrais (ou non) par un agriculteur ?

Enquête réalisée auprès de 398 agriculteurs

ENGRAIS =1 si l'agriculteur utilise des engrais, 0 sinon
DETTE : montant du crédit (par hectare) de l'agriculteur
DISTANCE : distance qui sépare l'exploitation considérée du marché le plus proche
EXPERT : nombre d'heures de discussion entre l'agriculteur et l'expert agricole
IRRIG =1 si des techniques d'irrigation sont utilisées, 0 sinon
PROPR =1 si l'agriculteur est propriétaire de l'exploitation, 0 sinon

➔ Fonction utilisée sous R est **glm()** .

➔ Son utilisation est similaire à la fonction lm(). Il faut juste ajouter le type de loi de probabilité utilisé (logit ou probit)

➔ Ajouter l'option **family=binomial(link="logit")** dans la commande glm pour le modèle Logit et **family=binomial(link="probit")** pour le modèle Probit

cor(Engrais[,c("DISTANCE","DETTE","EXPERT","IRRIG","PROPR")], use="complete.obs",method = c("spearman"))

	DISTANCE	*DETTE*	*EXPERT*	*IRRIG*	*PROPR*
DISTANCE	*1.000000000*	*0.10950362*	*0.009586334*	*-0.05893897*	*0.04251549*
DETTE	*0.109503617*	*1.00000000*	*0.270757804*	*0.30444228*	*-0.03296128*
EXPERT	*0.009586334*	*0.27075780*	*1.000000000*	*0.26225133*	*0.03265904*
IRRIG	*-0.058938973*	*0.30444228*	*0.262251330*	*1.00000000*	*-0.03130397*
PROPR	*0.042515486*	*-0.03296128*	*0.032659036*	*-0.03130397*	*1.00000000*

➔ Pas de corrélation entre les différentes variables explicatives
(Pour les variables IIRIG et PROPR après factorisation, il faut utiliser en toute rigueur les tests du Khi² (entre variables qualitatives et d'égalité des moyennes (Qualitative/quantitative) : cf TD 1)
➔ Factorisation des variables explicatives de type qualitative

Engrais$IRRIG<-as.factor(Engrais$IRRIG)
Engrais$PROPR<-as.factor(Engrais$PROPR)
summary(Engrais)

```
   ENGRAIS          EXPERT           DISTANCE        IRRIG      DETTE          PROPR
Min.   :0.0000   Min.   :  0.000   Min.   : 0.000   0:204   Min.   :   0   0:242
1st Qu.:0.0000   1st Qu.:  0.000   1st Qu.: 1.823   1:194   1st Qu.:   0   1:156
Median :0.0000   Median :  0.000   Median : 3.865           Median :  0
Mean   :0.3995   Mean   :  4.717   Mean   : 5.581           Mean   : 396
3rd Qu.:1.0000   3rd Qu.:  1.375   3rd Qu.: 7.065           3rd Qu.: 400
Max.   :1.0000   Max.   :472.000   Max.   :55.000           Max.   :7900
```

Table <- table(Engrais$ENGRAIS)
round(100*Table/sum(Table), 2)

```
    0     1
60.05 39.95
```

Modèle Logit

modele<-
glm(ENGRAIS~EXPERT+DISTANCE+IRRIG+DETTE+PROPR,data=Engrais,family=
binomial(link="logit"))
summary(modele)


```
Coefficients:
             Estimate   Std. Error   z value   Pr(>|z|)
(Intercept) -1.7517076  0.2542212   -6.890    5.56e-12 ***
EXPERT       0.0299871  0.0164412    1.824    0.0682 .
DISTANCE    -0.0223775  0.0225327   -0.993    0.3207
IRRIG[T.1]   1.9280523  0.2452185    7.863    3.76e-15 ***
DETTE        0.0002692  0.0001338    2.012    0.0442 *
PROPR[T.1]   0.5281763  0.2467122    2.141    0.0323 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 535.55 on 397 degrees of freedom
Residual deviance: 429.46 on 392 degrees of freedom
AIC: 441.46

Number of Fisher Scoring iterations: 6
```

Calcul du VIF des coefficients estimés

```
library(car)
vif(modele)
```

EXPERT	*DISTANCE*	*IRRIG*	*DETTE*	*PROPR*
1.024198	*1.022854*	*1.036460*	*1.046981*	*1.047543*

Intérêt du modèle (existence d'au moins une variable dont le coefficient est non nul)

Valeur calculée : Null Deviance-Residual Deviance
Valeur test : $Khi^2(5)$

```
chi2<- (modele$null.deviance-modele$deviance)
ddl<-modele$df.null-modele$df.residual
pvalue<-pchisq(chi2,ddl,lower.tail=F)
print(pvalue)
```

[1] 2.736826e-21

Interprétation :

➔ La p-value étant inférieure à 0,05, on refuse l'hypothèse de nullité de l'ensemble des coefficients de variables explicatives du modèle. Il y a donc un intérêt à estimer ce modèle au seuil de risque de 5 %

Interprétation des coefficients :

➔ La distance entre l'exploitation et le marché n'influence pas de manière significative l'utilisation d'engrais par l'agriculteur.

➔ La probabilité qu'un agriculteur utilise un engrais augmente avec le niveau d'endettement de l'exploitant : un agriculteur très endetté a besoin de produire et de vendre davantage pour pouvoir rembourser sa dette (p-value 0,0442 ; signe positif du coefficient).

➔ Probabilité augmente avec le temps passé à discuter avec un expert dont le but est d'expliquer les avantages à utiliser de l'engrais (p-value : 0,0682, valeur positive). (significativité supérieure à 5 % mais inférieure à 10 %)

➔ Probabilité augmente si l'agriculteur est propriétaire de la parcelle : il est alors davantage impliqué dans le résultat de la production (p-value : 0,0323, signe +)

➔ L'irrigation favorise fortement l'utilisation de l'engrais par les engrais (significativité inférieure à 1%, signe +)

Calcul des effets marginaux pour chaque variable explicative (/ à leur niveau moyen)

mean(dlogis(predict(modele,type="link"))) * coef(modele)

(Intercept)	*EXPERT*	*DISTANCE*	*IRRIG[T.1]*	*DETTE*
-3.160067e-01	*5.409645e-03*	*-4.036888e-03*	*3.478191e-01*	*4.855856e-05*

PROPR[T.1]
9.528260e-02

➔ Un mètre de plus (par rapport à la valeur moyenne) par rapport au marché diminue de 0,004 la probabilité que l'agriculteur utilise de l'engrais.

➔ Un euro de plus de dette augmente de 0,000049 la probabilité que l'agriculture utilise de l'engrais.

Calcul des odd-ratios :

exp(coef(modele))

(Intercept)	*EXPERT*	*DISTANCE*	*IRRIG[T.1]*	*DETTE*	*PROPR[T.1]*
0.1734775	*1.0304412*	*0.9778710*	*6.8761044*	*1.0002692*	*1.6958368*

➔ Les agriculteurs irriguant leur champ ont environ 6,9 fois plus de chance d'utiliser des engrais par rapport aux agriculteurs n'irriguant pas leur terrain.

➔ Les agriculteurs propriétaires ont environ 1,7 fois plus de chance d'utiliser des engrais par rapport aux agriculteurs locataires.

➔ A partir des coefficients estimés, calculer de odd-ratios plus complexes

Les agriculteurs propriétaires irriguant leurs champs ont 11,7 fois (exp(1.928+0.528)) de chance d'utiliser des engrais par rapport aux agriculteurs locataires n'irriguant pas leur champ.

Calcul de la probabilité estimée pour chaque observation de la base de données :

```
pred.proba<-predict(modele,type="response")
print(pred.proba)
```

```
        1         2         3         4         5         6         7
0.14642796 0.14365313 0.99999778 0.85442519 0.55438962 0.14405271 0.05794078
        8         9        10        11        12        13        14
0.53646887 0.12060943 0.41515042 0.14467468 0.65820165 0.70509736 0.14461931
       15        16        17        18        19        20        21
0.65915753 0.66983357 0.67199525 0.20459867 0.66720484 0.21970426 0.66021252
```

Application de la règle pour chacune des valeurs prédites

```
pred.moda<-factor(ifelse(pred.proba>0.5,"1","0"))
print(pred.moda)
```

```
 1  2  3  4  5  6  7  8  9 10 11  12 13 14 15 16 17 18 19 20 21
 0  0  1  1  1  0  0  1  0  0  0   1  1  0  1  1  1  0  1  0  1 ……
```


Tableau de prédiction

```
mc<-table(Engrais$ENGRAIS,pred.moda)
print(mc)
```

```
   pred.moda
     0   1
  0 169  70
  1  37 122
```

Dans 169+122 = 291 cas sur (169+70+37+122) 398 cas (73%) la valeur est correctement prédite.

Calcul du % d'erreur :

```
err<-(mc[2,1]+mc[1,2])/sum(mc)
print(err)
```

```
[1] 0.2688442
```

Calcul des autres indicateurs de qualité de prévision du modèle:

Sensibilite<-mc[2,2]/(mc[2,1]+mc[2,2])
print(Sensibilite)
[1] 0.7672956
Specificite<-mc[1,1]/(mc[1,1]+mc[1,2])
print(Specificite)
[1] 0.707113
Precision<-mc[2,2]/(mc[1,2]+mc[2,2])
print(Precision)
[1] 0.6354167
Faux_positif<- mc[1,2]/(mc[1,1]+mc[1,2])
print(Faux_positif)
[1] 0.2928870

➔Indicateurs en termes de précision, de sensibilité et de spécificité bons.

➔ Modèle prévoit correctement le phénomène étudié.

➔ La fonction **hitmiss()** permet de calculer directement le tableau mc et le % de cas correctement prédit, l'indicateur de sensibilité et l'indicateur de spécificité

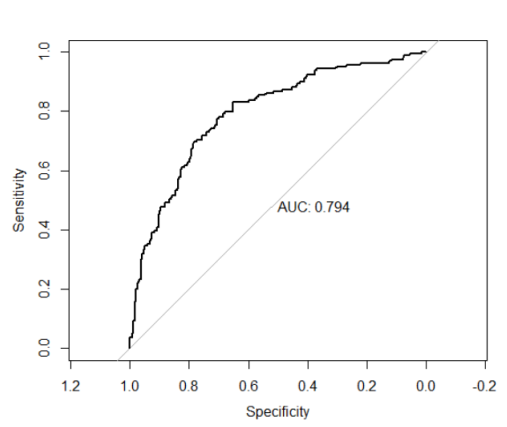
library(pscl)
hitmiss(modele)

Classification Threshold = 0.5
y=0 y=1
yhat=0 169 37
yhat=1 70 122
Percent Correctly Predicted = 73.12%
Percent Correctly Predicted = 70.71%, for y = 0
Percent Correctly Predicted = 76.73% for y = 1

Null Model Correctly Predicts 60.05%

[1] 73.11558 70.71130 76.72956

library(pROC)
pred <-predict(modele)
Test_roc=roc(Engrais$ENGRAIS ~pred, plot=TRUE, print.auc=TRUE)



Calcul du Pseudo R^2

R2_Mc_Fadden<-1-(modele$deviance/modele$null.deviance)
R2_Mc_Fadden

> R2_Mc_Fadden
[1] 0.1981039

➔ Existe des méthodes de sélection des variables dans les modèles Logit et Probit (dans le cas du probit remplacer dans les commandes logit par probit)

Méthode Forward

modele<-glm((ENGRAIS~1),data=Engrais,family=binomial(logit))
modele.forward<-
step(modele,scope=list(lower=~1,upper=~EXPERT+DISTANCE+IRRIG+DETTE+PROPR), data=Engrais, direction ="forward")
summary(modele.forward)

```
Start:  AIC=537.55
ENGRAIS ~ 1

           Df  Deviance    AIC
+ IRRIG     1    449.08  453.08
+ EXPERT    1    514.66  518.66
+ DETTE     1    522.26  526.26
+ DISTANCE  1    532.26  536.26
+ PROPR     1    532.97  536.97
<none>           535.55  537.55

Step:  AIC=453.08
ENGRAIS ~ IRRIG

           Df  Deviance    AIC
+ EXPERT    1    438.09  444.09
+ PROPR     1    444.52  450.52
+ DETTE     1    444.59  450.59
<none>           449.08  453.08
+ DISTANCE  1    448.37  454.37
```



```
Step:  AIC=444.09
ENGRAIS ~ IRRIG + EXPERT

           Df  Deviance    AIC
+ PROPR     1    434.76  442.76
+ DETTE     1    434.88  442.88
<none>           438.09  444.09
+ DISTANCE  1    437.52  445.52

....
....


Step:  AIC=440.53
ENGRAIS ~ IRRIG + EXPERT + PROPR + DETTE

           Df  Deviance    AIC
<none>           430.53  440.53
+ DISTANCE  1    429.46  441.46
```

```
Coefficients:
             Estimate   Std. Error  z value  Pr(>|z|)
(Intercept) -1.8746572   0.2268887  -8.262  < 2e-16 ***
IRRIG[T.1]   1.9536666   0.2443481   7.995  1.29e-15 ***
EXPERT       0.0310167   0.0168107   1.845   0.0650 .
PROPR[T.1]   0.5095427   0.2456535   2.074   0.0381 *
DETTE        0.0002568   0.0001320   1.946   0.0517 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 535.55 on 397 degrees of freedom
Residual deviance: 430.53 on 393 degrees of freedom
AIC: 440.53

Number of Fisher Scoring iterations: 6
```


Méthode Backward

```
modele<-
glm((ENGRAIS~EXPERT+DISTANCE+IRRIG+DETTE+PROPR),data=Engrais,fa
mily=binomial(logit))
modele.backward<-
step(modele,scope=list(lower=~1,upper=~EXPERT+DISTANCE+IRRIG+DETTE+P
ROPR), data=Engrais,direction ="backward")
summary(modele.backward)
```

```
Coefficients:
             Estimate   Std. Error  z value  Pr(>|z|)
(Intercept) -1.8746572   0.2268887  -8.262  < 2e-16 ***
EXPERT       0.0310167   0.0168107   1.845   0.0650 .
IRRIG[T.1]   1.9536666   0.2443481   7.995  1.29e-15 ***
DETTE        0.0002568   0.0001320   1.946   0.0517 .
PROPR[T.1]   0.5095427   0.2456535   2.074   0.0381 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Méthode Both

```
modele<-glm((ENGRAIS~1),data=Engrais,family=binomial(logit))
modele.both<-
step(modele,scope=list(lower=~1,upper=~EXPERT+DISTANCE+IRRIG+DETTE+P
ROPR), data=Engrais,direction ="both")
summary(modele.both)
```

```
Start:  AIC=537.55
ENGRAIS ~ 1

           Df Deviance    AIC
+ IRRIG     1   449.08 453.08
+ EXPERT    1   514.66 518.66
+ DETTE     1   522.26 526.26
+ DISTANCE  1   532.26 536.26
+ PROPR     1   532.97 536.97
<none>          535.55 537.55
```



```
Step:  AIC=453.08
ENGRAIS ~ IRRIG

           Df Deviance    AIC
+ EXPERT    1   438.09 444.09
+ PROPR     1   444.52 450.52
+ DETTE     1   444.59 450.59
<none>          449.08 453.08
+ DISTANCE  1   448.37 454.37
- IRRIG     1   535.55 537.55

Step:  AIC=444.09
ENGRAIS ~ IRRIG + EXPERT

           Df Deviance    AIC
+ PROPR     1   434.76 442.76
+ DETTE     1   434.88 442.88
<none>          438.09 444.09
+ DISTANCE  1   437.52 445.52
- EXPERT    1   449.08 453.08
- IRRIG     1   514.66 518.66
```

```
Step:  AIC=442.76
ENGRAIS ~ IRRIG + EXPERT + PROPR

                Df  Deviance    AIC
+ DETTE          1    430.53 440.53
<none>                434.76 442.76
+ DISTANCE       1    434.00 444.00
- PROPR          1    438.09 444.09
- EXPERT         1    444.52 450.52
- IRRIG          1    513.10 519.10

Step:  AIC=440.53
ENGRAIS ~ IRRIG + EXPERT + PROPR + DETTE

                Df  Deviance    AIC
<none>                430.53 440.53
+ DISTANCE       1    429.46 441.46
- DETTE          1    434.76 442.76
- PROPR          1    434.88 442.88
- EXPERT         1    438.90 446.90
- IRRIG          1    502.57 510.57
```

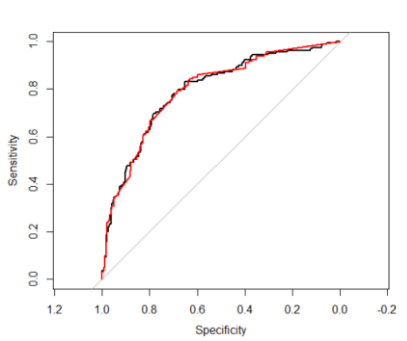


```
Coefficients:
             Estimate   Std. Error  z value  Pr(>|z|)
(Intercept) -1.8746572  0.2268887  -8.262   < 2e-16 ***
IRRIG[T.1]   1.9536666  0.2443481   7.995   1.29e-15 ***
EXPERT       0.0310167  0.0168107   1.845   0.0650 .
PROPR[T.1]   0.5095427  0.2456535   2.074   0.0381 *
DETTE        0.0002568  0.0001320   1.946   0.0517 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 535.55 on 397 degrees of freedom
Residual deviance: 430.53 on 393 degrees of freedom
AIC: 440.53

Number of Fisher Scoring iterations: 6
```

➔ La variable Distance n'est pas retenue pour l'estimation d'après ces 3 méthodes de sélection des variables explicatives,

modele2<-glm(ENGRAIS ~ EXPERT+IRRIG+DETTE+PROPR , data=Engrais, family=binomial(logit))

pred<-predict(modele)
Test_roc=roc(Engrais$ENGRAIS ~pred)
pred2<-predict(modele2)
Test_roc2=roc(Engrais$ENGRAIS ~pred2)
plot(Test_roc2, add=TRUE, col='red')



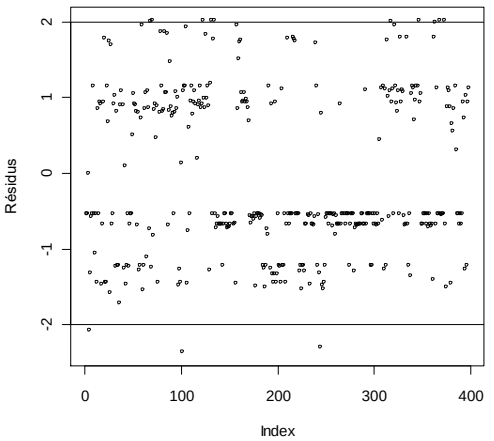
Points pouvant influencés les résultats de l'estimation d'un modèle Logit (même type de procédure pour un modèle Probit)

Il est nécessaire de calculer le résidu pour chacune des observations

➔ On considère qu'une observation influence les résultats si le résidu associé n'est pas comprise dans l'intervalle [-2,2]

Graphique des résidus

plot(rstudent(modele2),type="p",cex=0.5,ylab="Résidus")
abline(h=c(-2,2))



➔ Pour déterminer précisément les observations concernées, il suffit de classer les résidus par ordre croissant :

sort(rstudent(modele2))

101	*244*	*4*	*36*	*26*
-2.364109306	***-2.293277800***	*-2.071953788*	*-1.710523533*	*-1.571624877*

60
-1.534629776

.....

368	*373*
2.018336524	*2.018336524*

➔ A priori, les observations 101 et 244 semblent affecter les résultats de l'estimation

Création d'une nouvelle base de données :

Engrais_nv <-Engrais[-c(101,244),]

Réestimation du modèle 2 avec la nouvelle base de données :

modele3<-
glm(ENGRAIS~EXPERT+IRRIG+DETTE+PROPR,data=Engrais_nv,family=binomial (logit))
summary(modele3)

Coefficients:

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	*-1.9574795*	*0.2328118*	*-8.408*	*< 2e-16 ****
EXPERT	*0.0463105*	*0.0204814*	*2.261*	***0.0238 ****
IRRIG[T.1]	*1.9700710*	*0.2473996*	*7.963*	*1.68e-15 ****
DETTE	*0.0003865*	*0.0001536*	*2.516*	***0.0119 ****
PROPR[T.1]	*0.5462063*	*0.2493430*	*2.191*	*0.0285 **

---
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Null deviance: 533.51 on 395 degrees of freedom
Residual deviance: 419.37 on 391 degrees of freedom
AIC: 429.37

Remarque :

Les variables DETTE et EXPERT deviennent significatives au seuil de 5%

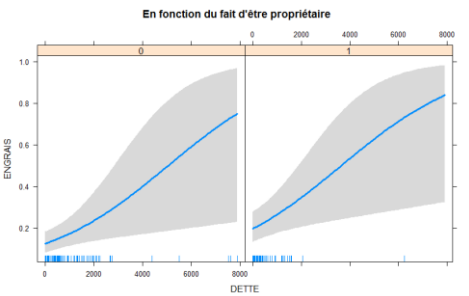
Le R2 de Mac Fadden est meilleur (0,21 au lieu de 0,19)

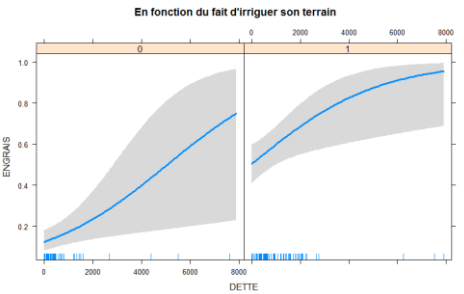
➔ On peut visualiser avec la fonction **visreg()** l' effet d'une variable explicative quantitative pour une variable explicative qualitative donnée sur la probabilité d'utiliser de l'engrais

library(visreg)

visreg(modele3, "DETTE", by="PROPR",scale="response", main="En fonction du fait d'être propriétaire")

visreg(modele3, "DETTE", by="IRRIG",scale="response", main="En fonction du fait d'irriguer son terrain")





Ré-estimation du modèle sur la base initiale (Engrais, modele 2) par un modèle Probit

```
modeleP<-
glm(ENGRAIS~EXPERT+IRRIG+DETTE+PROPR,data=Engrais,family=binomial(probit))
summary(modeleP)
```

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.118e+00	1.260e-01	-8.867	<2e-16	***
EXPERT	1.679e-02	9.170e-03	1.831	0.0670	.
IRRIG1	1.183e+00	1.426e-01	8.292	<2e-16	***
DETTE	1.474e-04	7.647e-05	1.927	0.0539	.
PROPR1	2.967e-01	1.442e-01	2.057	0.0396	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null deviance: 535.55 on 397 degrees of freedom
Residual deviance: 430.84 on 393 degrees of freedom
AIC: 440.84

➔ En termes de significativité des variables : mêmes effets.

40 % des individus ont Y=1 : coefficients Probit : coefficients Logit / 1,6

Effets marginaux dans le cas d'un modèle Probit :

mean(dnorm(predict(modeleP,type="link"))) * coef(modeleP)

(Intercept)	*EXPERT*	*IRRIG1*	*DETTE*	*PROPR1*
-3.422771e-01	*5.143254e-03*	*3.622369e-01*	*4.514188e-05*	*9.085882e-02*

Résultats et comparaison avec ceux obtenus pour le modèle Logit

Pourcentage d'erreur de prévision : *0,274 (Rappel Logit : 0,269)*

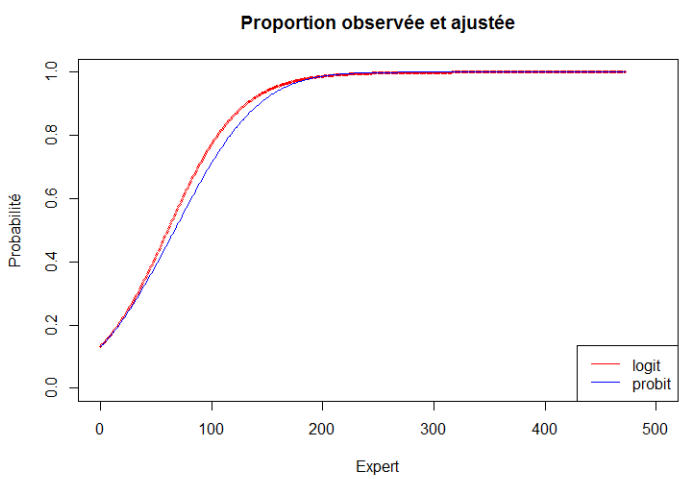
R2 Mac_Fadden *: 0,196 (Rappel Logit : 0,198)*

➔ **Graphique permettant de comparer les modèles probit et logit pour une variable donnée (ici, Expert)**

```
# Permet de récupérer le coefficient estimé de la constante du modèle logit
beta0=coef(modele2)[1]
#Permet de récupérer le coefficient estimé associé à la variable Expert
beta1=coef(modele2)[2]

#Permet de récupérer le coefficient estimé de la constante du modèle Probit
beta0p=coef(modeleP)[1]
#Permet de récupérer le coefficient estimé associé à la variable Expert
beta1p=coef(modeleP)[2]

#Permet de définir l'échelle de l'axe des abscisses (472 étant la valeur maximale associé à la variable Expert)
summary(Engrais$EXPERT)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	0.000	0.000	4.717	1.375	472.000

```
#Graphique
abscisse1=seq(0,472,length=472)
title("Proportion observée et ajustée")
plot(abscisse1,plogis(beta0+beta1*abscisse1),
cex=0.4,lty=1,col="red",xlim=c(0,500),ylim=c(0,1),xlab="Expert",ylab="Probabilité")

lines(abscisse1,pnorm(beta0p+beta1p*abscisse1),cex=0.4,col="blue", lty=1)
legend("bottomright",c("logit","probit"),lwd=1,col=c("red","blue"))
```

6) Prise en compte de l'hétéroscédasticité des erreurs

➔ Dans l'estimation des modèles précédents, nous avons supposé implicitement que les erreurs du modèle estimé étaient homoscédastiques.

Or, dans le cas d'un modèle Probit, il est supposé que les résidus ont une moyenne de 0 et une variance de 1. De même, dans le cas d'un modèle Logit, la variance des erreurs est supposée constante et invariante en fonction des individus.

Par conséquent, l'hétéroscédascité des erreurs peut également exister dans les modèles dont la variable à expliquer est de type qualitatif binaire.

Il faut donc tester l'hypothèse d'homoscédasticité des erreurs et estimer le modèle tenant compte de l'hétéroscédasticité des erreurs si nécessaire

```
library(glmx)
modele3h<-
hetglm(ENGRAIS~EXPERT+IRRIG+DETTE+PROPR|DETTE+EXPERT+IRRIG+
+PROPR, data=Engrais_nv, family=binomial(logit))
summary(modele3h)
```



```
             Estimate  Std. Error  z value    Pr(>|z|)
(Intercept) -2.148445  0.318923   -6.737    1.62e-11 ***
EXPERT       0.041589  0.026597    1.564    0.1179
IRRIG1       1.874987  0.320146    5.857    4.72e-09 ***
DETTE        0.002704  0.001185    2.281    0.0225 *
PROPR1       0.421799  0.274243    1.538    0.1240

Latent scale model coefficients (with log link):

             Estimate   Std. Error  z value Pr(>|z|)
DETTE       1.070e-03   3.558e-04   3.008  0.00263 **
EXPERT      9.261e-05   1.135e-02   0.008  0.99349
IRRIG1     -3.802e-01   5.163e-01  -0.736  0.46150
PROPR1     -8.789e-02   2.262e-01  -0.389  0.69760
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: -200.8 on 9 Df
LR test for homoskedasticity: 17.84 on 4 Df, p-value: 0.001325
```

La variable DETTE a un impact significatif (au seuil de risque de 1%) sur la variance des résidus

P<0,05, donc l'hypothèse d'homoscédasticité des erreurs n'est pas acceptée au seuil de risque de 5%

h1 <- hetglm(ENGRAIS~EXPERT+IRRIG+DETTE+PROPR|DETTE,data=Engrais_nv, family=binomial(logit))
summary(h1)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	*-2.341355*	*0.285507*	*-8.201*	*2.39e-16*	***
EXPERT	*0.055545*	*0.025010*	*2.221*	*0.02636*	*
IRRIG1	*2.040712*	*0.302480*	*6.747*	*1.51e-11*	***
DETTE	*0.003436*	*0.001124*	*3.058*	*0.00223*	** *(au lieu de * /modele 3)*
PROPR1	*0.536786*	*0.297131*	*1.807*	*0.07083*	. *(au lieu de * / modele 3)*

Latent scale model coefficients (with log link):

	Estimate	Std. Error	z value	Pr(>|z|)	
DETTE	*0.0010123*	*0.0003172*	*3.191*	*0.00142*	**

P<0,05, il est donc nécessaire de retenir ce modèle pour l'interprétation des résultats

Log-likelihood: -201.4 on 6 Df
LR test for homoskedasticity: 16.66 on 1 Df, p-value: 4.473e-05

h1c <- hetglm(ENGRAIS~1,data=Engrais_nv,family=binomial(logit))
summary(h1c)

(R2McFAdden<-1-(h1$loglik/h1c$loglik))

[*1] 0.2451605*

exp(coef(h1))

(Intercept)	*EXPERT*	*IRRIG1*	*DETTE*	*PROPR1*	*(scale)_DETTE*
0.09619717	*1.05711646*	*7.69608939*	*1.00344192*	*1.71050061*	*1.00101285*

mean(dlogis(predict(h1,type="link"))) * coef(h1)

(Intercept)	*EXPERT*	*IRRIG1*	*DETTE*	*PROPR1*	*(scale)_DETTE*
-0.3937317028	*0.0093406506*	*0.3431743739*	*0.0005778139*	*0.0902681015*	*0.0001702393*

h1h <-
hetglm(ENGRAIS~EXPERT+IRRIG+DETTE+PROPR|1,data=Engrais_nv,family=binomial(logit))
summary(h1h)

Coefficients (binomial model with logit link):

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	*-1.9574795*	*0.2328121*	*-8.408*	*< 2e-16 ****
EXPERT	*0.0463105*	*0.0204817*	*2.261*	*0.0238 **
IRRIG1	*1.9700710*	*0.2473997*	*7.963*	*1.68e-15 ****
DETTE	*0.0003865*	*0.0001536*	*2.516*	*0.0119 **
PROPR1	*0.5462063*	*0.2493432*	*2.191*	*0.0285 **

Latent scale model coefficients (with log link):
None (constant scale = 1).

Log-likelihood: -209.7 on 5 Df
Dispersion: 1

library(lmtest)
lrtest(h1h,h1)

Likelihood ratio test

Model 1: ENGRAIS ~ EXPERT + IRRIG + DETTE + PROPR | 1
Model 2: ENGRAIS ~ EXPERT + IRRIG + DETTE + PROPR | DETTE

	#Df	LogLik	Df	Chisq	Pr(>Chisq)	
1	*5*	*-209.69*				
2	*6*	*-201.36*	*1*	*16.659*	*4.473e-05*	*****

---
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

➔ p-value <0,05 : par conséquent, au seuil de risque de 5%, le deuxième modèle améliore de manière significative la log vraisemblance

➔ Il est donc conservé pour l'interprétation des résultats d'estimation.