Performance du modèle multinomial logit dans les cas non-linéaires

Marcelin Tworski
2A Ensimag Fillière MMIS
Encadrant : Pierre Lemaire
Laboratoire G-SCOP

Contents

1 Introduction

Largement utilisés dans les sciences économiques, les modèles de choix discrets permettent de prédire les choix des acteurs à partir des données qui influencent ces choix. Les Random Utility Models, un type de modèle à choix discret, attribuent à chaque alternative une utilité, l'alternative ayant l'utilité la plus élevé devenant le choix.

Les différents acteurs étant confrontés à des situations différentes, et étant eux-même différents, leurs choix ne seront pas invariablement les mêmes. Lors d'une enquête sur les modes de transports ou les choix de consommation, on part des choix des acteurs et, à partir des facteurs que l'on estime être pertinents, on essaye de déterminer leur influence.

Le modèle multinomial logit permet d'estimer l'influence des différents facteurs que l'on aura jugés pertinents. Cependant, par exemple, ce modèle fait l'hypothèse d'une élasticité constante sur l'ensemble des valeurs admissible des variable explicatives. Cette hypothèse est en pratique rarement vérifiée.

Nous nous intéresserons ici aux performances du modèle logit pour rendre compte de modèles où ses hypothèses ne seront pas vérifiées. A partir de modèles de choix ad hoc, nous testerons le modèle logit en ne respectant pas ses hypothèses, pour notamment déterminer comment l'adapter.

Comment retrouver la relation réelle dans le cas où l'élasticité n'est pas constante ? Quelles sont les conséquences du manque d'information lorsqu'il manque au modèle des variables explicatives ? Comment le modèle logit peut-il rendre compte, ou au contraire résister aux dépendances fortes (Ex : Bus gratuit entre 16 et 18 ans pour un revenu de X à Y avec X et Y proche). Nous tenterons de fournir entre autres des informations utiles aux concepteurs de questionnaires et à ceux qui les analysent.

Ainsi, après avoir exposé la notion de Random Utility Model, nous verrons comment on peut utiliser le modèle logit dans les cas où l'utilité ne dépend pas linéairement des variables. Nous testerons ensuite notre appoche avec différentes fonctions, en affaiblissant les hypothèses : non-linéarité, non-monotonie, non-dérivabilité.

2 Notions utilisées

2.1 Random Utility Models

Un Random Utility Model décrit un modèle de choix où chaque alternative à une utilité specifique. Un utilisateur faisant un choix chosit l'alternative ayant l'utilité la plus élevée. L'utilité calculé pour chaque alternative dépend de variables explicatives x_i, qui sont des variables influençant ce choix. Elles peuvent dépendre de l'alternative ou de l'utilisateur. McFadden a formalisé ce modèle [5]

Les objectifs peuvent être multiples : Modélisation d'une situation en vue de prévisions, influences sur les choix lorsque les causalités sont connues.

Avec m alternatives, pour tout $k \in [1; n]$, on a l'utilité :

$$U_k = \alpha_0 + \sum_{i=1}^{n} \alpha_{i,k} x_i + \epsilon_k$$

Où ϵ suit une loi normale centrée.

On note également,

$$U_k = V_k + \epsilon_k$$

où V_k est la composante déterministe.

On se réduira par la suite à 2 alternatives, ceci est équivalent à comparer une alternative à toutes les autres $U_2' = \max_{k \neq 1} U_k$

Exemple : Pour un déplacement, plusieurs choix de transport sont possible.

On suppose que l'on a :

$$Utilite = a \times cout + b \times vitesse + \epsilon$$

Le coût et la vitesse du bus et de la voiture sont différents, on aura donc des utilités différentes. Parmi ces deux choix, l'utilisateur choisr

ϵ correspond aux défauts de modélisations. Par exemple le confort influence également la décision , mais on ne dispose pas des données nécessaires.

2.2 Modèle multinomial logistique

Il s'agit d'une reconstitution du RUM à partir des choix des utilisateurs. Ainsi en entrée on a les choix, et en sortie les coefficients $\alpha_{i,k}$ pour k fixé.

Chaque écriture des modèles ne donnent pas nécessairement des modèles différent.

Par exemple, si $\forall k, U'_k = C \times U_k$ Alors les modèles U et U' sont équivalents.

Ce n'est donc pas la valeur de l'utilité d'une alternative qui importe, mais sa valeur par rapport aux utilités des autres alternatives.

Il faut alors connaître quel est le modèle de référence lors de l'execution du multinomial logit.

Le modèle de référence est celui-ci [1][2]: On note p la probabilité que le choix 1 soit fait.

$$\frac{p}{1-p} = \frac{e^{V_1}}{\sum_{k \neq i} e^{V_k}}$$

On obtient donc avec deux alternatives :

$$\frac{p}{1-p} = \frac{e^{V_1}}{e^{V_2}}$$

Ce qui nous donne

$$V_1 = \ln(\frac{p}{1-p}) + V_2$$

.

3 Modèles non-linéaires

L'utilité n'a pas de raison particulière d'être linéaire selon les variables explicatives. Par exemple, on peut considérer que l'utilité d'utiliser les transports en commun décroît au fur et à mesure que l'on grandit et augmente lorsque l'on vieillit. Encore, pour certains articles, on aperçoit un effet de snobisme pour l'acheteur car un produit peu cher sera signe d'une moins grande qualité ou de distinction sociale moindre. On cherche à évaluer la dépendance au prix, qui peut-être positive ou négative. Cette dépendance correspond à la dérivé de f par rapport au prix.

On veut pouvoir être capable de retrouver ces résultats. Il s'agit de :
1)Construire une population qui fera un choix en fonction d'une variable x.
2) Interpréter ces données fictives
3) Comparer la sortie avec l'entrée

3.1 Modèle général des RUMs

Nous nous intéressons ici aux cas non-linéaires des modèles de choix. C'est à dire que l'on suppose seulement les utilités de la forme :

$$U_k = \alpha_0 + \sum_{i=1}^{n} f_{i,k}(x_i) + \epsilon_k$$

3.2 Utilisation du modèle logit

Nous voulons savoir comment utiliser le modèle logit pour analyser des données. Nous allons donc l'utiliser en violant explicitement les hypothèses. La linéarité n'est alors vrai que localement, suivant les fonction considérées.

On s'intéresse aux cas où il n'y a qu'une variable explicative.

On a donc

$$U_1 = V_1 + \epsilon_1$$
$$U_2 = V_2 + \epsilon_2$$

avec

$V_1 = f_1(x)$ et $V_2 = f_2(x)$

Que l'on simplifiera sans perte de généralité en $V_1 = f(x)$ et $V_2 = 0$

avec $f = f_1 - f_2$
Soit $I = [a, b]$ l'intervalle des valeurs admissibles de x,
Soit

$$A = \frac{1}{|I|} \int_I f'(x)dx$$

Le modèle logit interprète l'utilité du choix 1 comme $V_1 = f(x) = C + Ax$

où C est une constante.

Ainsi, on récupère une utilité qui est une fonction affine.

On aimerait pouvoir analyser ces données de façon plus précise.

3.2.1 Catégorisation

On décide donc de séparer les données en catégories. Pour nos expériences on suppose les données uniformes sur I. On fera donc des catégories uniformes. On fera donc tourner N modèle logit avec pour chacun les données où, $x \in [a + (i-1)\frac{b-a}{N}; a + i\frac{b-a}{N}] = I_i$ pour $i \in [1; N]$ On note $x_i = [a + i\frac{b-a}{N}]$ pour $i \in [1; N]$ et $x'_i = \frac{x_{i-1}+x_i}{2}$ pour $i \in [1; N-1]$

Afin de vérifier que notre méthode fonctionne, on construit des RUM ad-hoc afin de pouvoir comparer la sortie du logit avec les entrées.

3.2.2 Adaptation au modèle de référence

Il faut maintenant faire correspondre notre modèle ad-hoc, qui ne respectera pas nécessairement les conditions de référence, avec le modèle pris en référence. Il faut donc calculer la probabilité p. On notera U_1' et U_2' les utilités de notre modèle ad-hoc et V_1' et V_2' leurs composantes déterministes.

On a :

$$\begin{aligned} p &= P(U_1' > U_2') \\ &= P(V_1' + \epsilon_1 > V_2' + \epsilon_2) \\ &= P(\epsilon_1 - \epsilon_2 > V_2' - V_1') \end{aligned}$$

On pose σ_1 et σ_2 les écarts types de ϵ_1 et ϵ_2. $\epsilon_1 - \epsilon_2$ suit donc une loi normale de moyenne 0 et d'écart type $\sigma' = \sqrt{\sigma_1^2 + \sigma_2^2}$

On a donc

$$\begin{aligned} p &= P(\epsilon_1 - \epsilon_2 > V_2' - V_1') \\ &= 1 - F_{(0,\sigma')}(V_2' - V_1') \end{aligned}$$

Avec F La fonction de répartition de la loi normale.

On a dorénavant tous les outils pour évaluer les performances de notre méthode.

3.3 Objectifs des tests de la méthode.

Le but est de proposer une méthode d'analyse de données réelles. La première limitation est le nombre de personnes interrogées : Le prix d'un sondage est souvent proportionnel à ce nombre. Il est donc utile de connaître quelle est l'évolution en précision de la méthode en fonction du nombre de sondés. À partir de quel seuil peut on considérer la méthode assez efficace ? La seconde question est le nombre N de catégories à constituer. Plus on augmentera le nombre de catégories, plus le nombre de sondés à l'intérieur de cette catégorie diminuera rendant la simulation moins précise. En revanche, plus on augmente le nombre de catégories, plus on pourra rendre compte des changements abruptes.

On cherche à savoir si cette approche est efficace, suivant les différentes hypothèses que l'on fera sur les fonctions d'utilités.

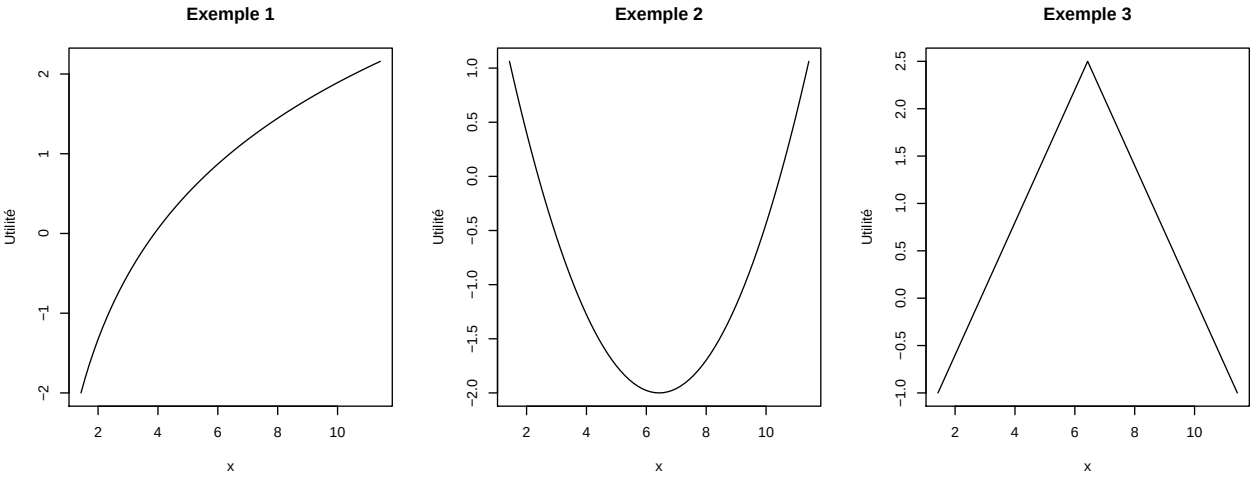
4 Expériences

Nous mènerons nos expériences sur trois exemples, avec notamment $\sigma_1 = \sigma_2$. On choisit $I = [10, 80]$
1) $V_1 = a\log(bx) - K$
2) $V_1 = a(x - b) - K$
3) $V_1 = -2a\chi_{(x>b)}(x)(x - b) + ax - K$

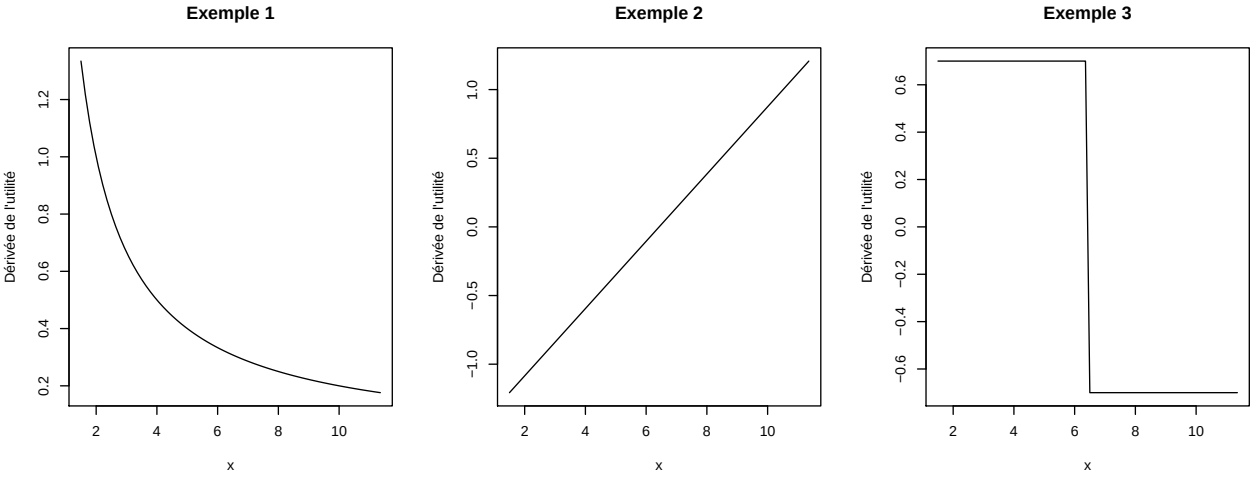
Note : Les paramètres a, b, et K ne sont pas les mêmes d'une expérience à l'autre.

4.1 Courbes théoriques

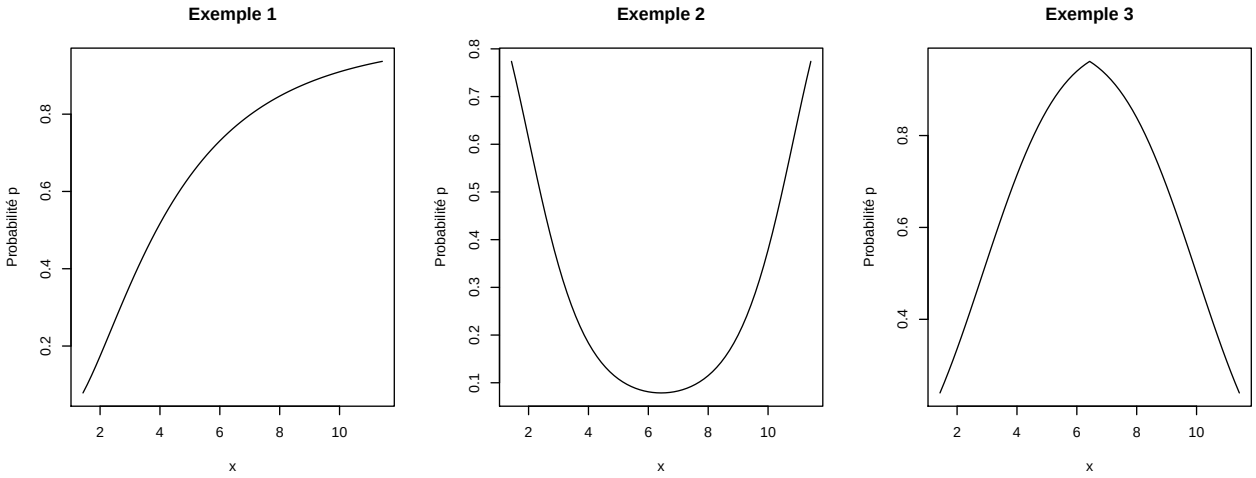
4.1.1 Utilités



4.1.2 Les dérivées

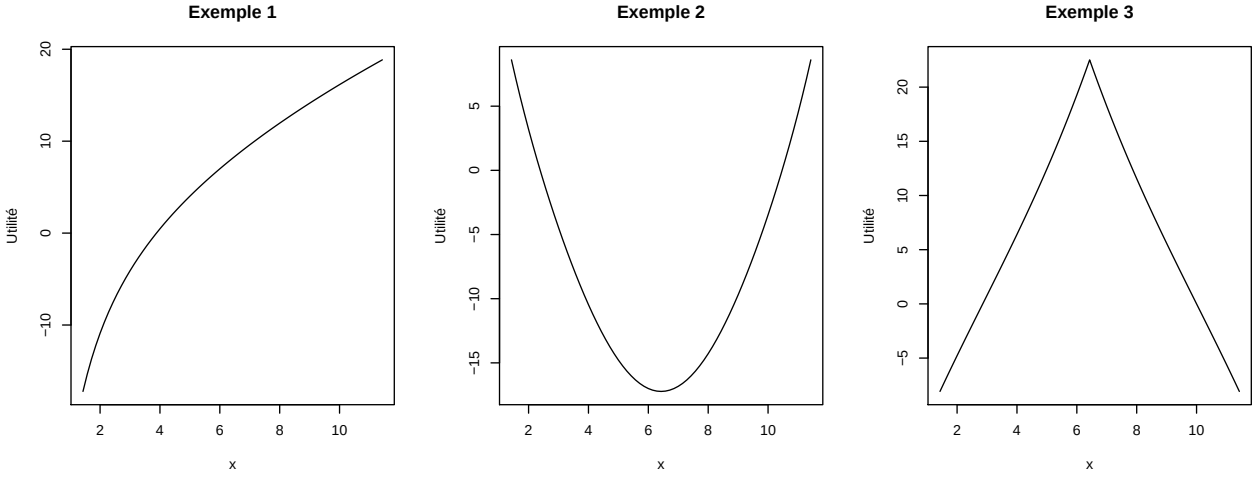


4.1.3 Probabilités du choix 1

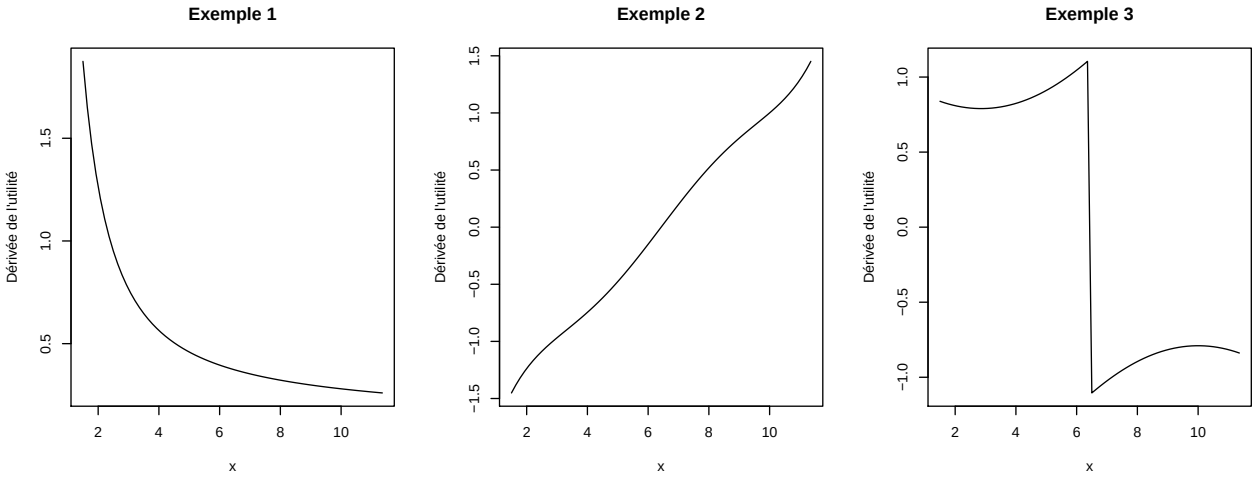


Comme vu précédemment, le modèle logit ne permettra pas de retrouver le même modèle RUM et donc la même utilité, mais une utilité qui donnera un modèle équivalent, vérifiant $V_1 = \ln(\frac{p}{1-p})$. C'est ce modèle là qui sera par la suite estimé.

4.1.4 Utilités telles qu'interprétées par le modèle logit

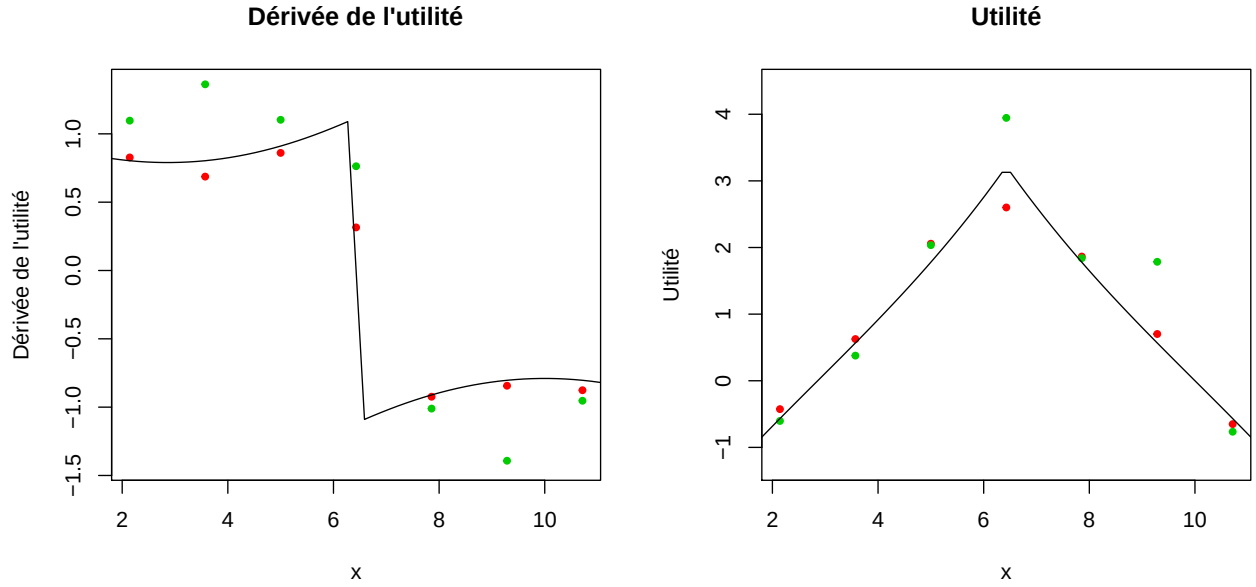


4.1.5 Dérivées de ces utilités



4.2 Exemple d'une simulation

Tout d'abord voici le résultat d'une simulation sur le cas 3, pour une population de 2 000 en vert et une population 20 000 en rouge. La courbe théorique est la courbe continue.



5 Efficacité en fonction de la taille de la population

On va réaliser 100 de simulations et évaluer leur précision suivant la taille de la population. La taille de la population tiré sur une échelle logarithmique pour des soucis de temps de simulation. Chaque point correspond à une simulation.

5.1 Mesure de l'erreur

Le modèle approxime localement autour de $f(x_i)$ par $C + f'(x_i')x_i'$ La méthode des catégories ne peut pas, à l'instar de quand le modèle logit tourne avec toutes les données, prévoir des fluctuations à l'intérieur de cette catégorie. Il peut donc au mieux prédire

$$A_i = \frac{1}{|I_i|} \int_{I_i} f'(x)dx$$

On peut retrouver $V_1(x_i) = C_i + A_i x_i'$ En pratique on cherche surtout à estimer f' donc les A_i Notre méthode est donc moins efficace quand $|f'_{x_{i'}} - A_i|$ est grand.
La constante C_i fournise par le modèle est l'utilité en zéro du la fonction linéaire du modèle logit approximation local de f.

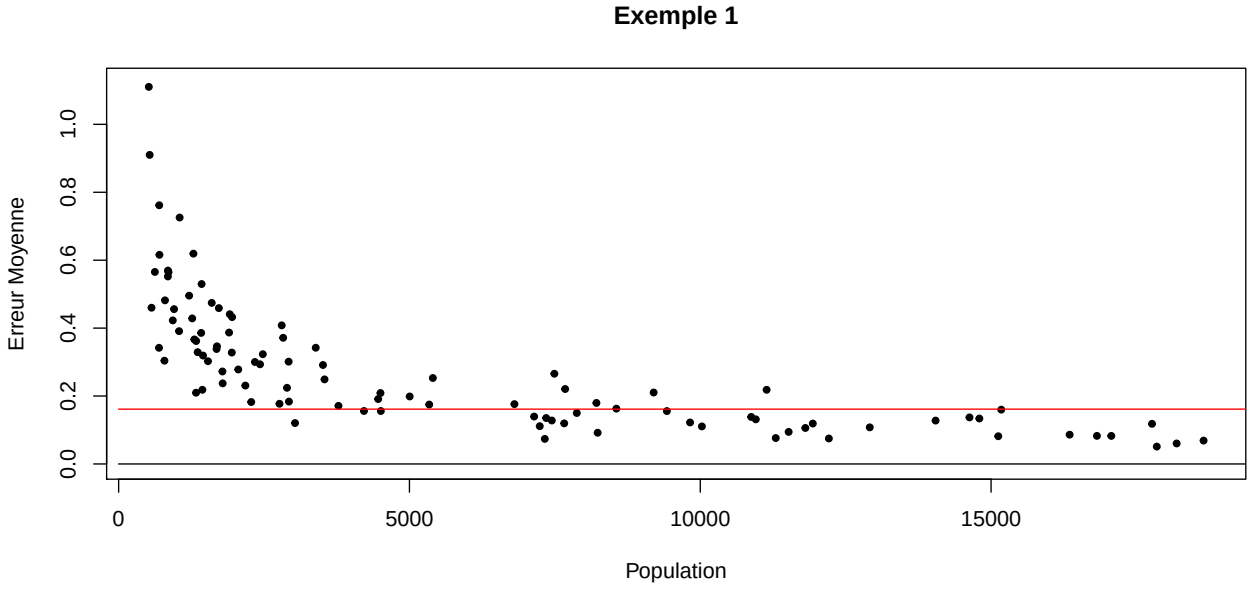
Note : On a

$$\frac{1}{|I_i|} \int_{I_i} f'(x)dx \underset{N\to\infty}{\longrightarrow} f'(x_i')$$

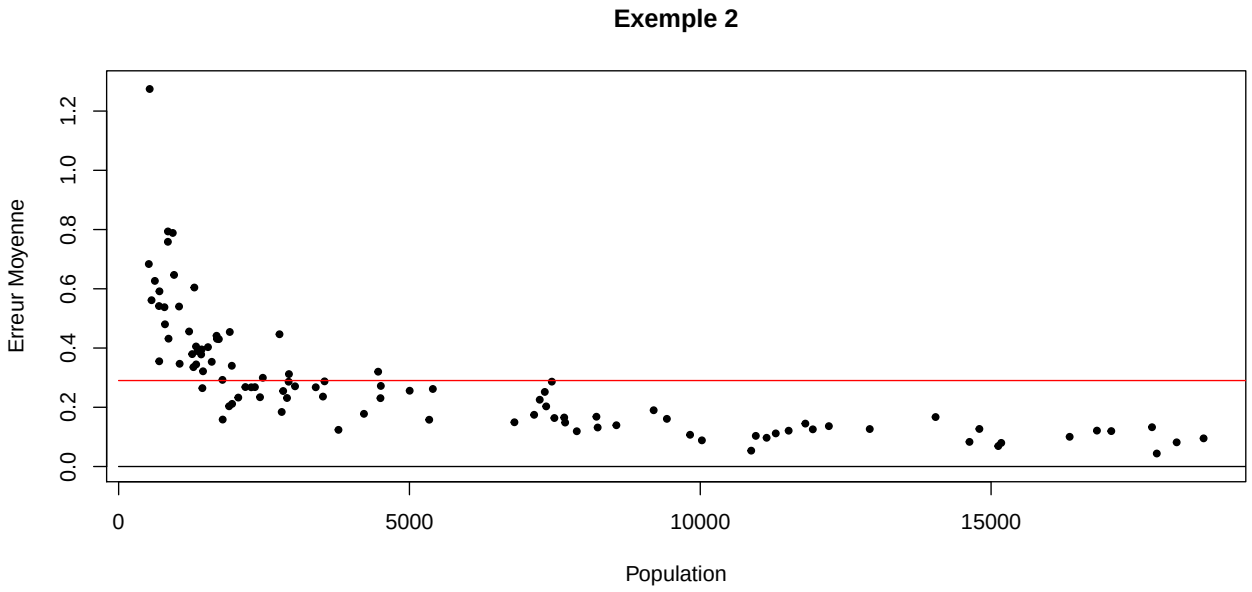
sous hypothèses de dérivabilités et continuités.
$|f'_{x_{i'}} - A_i|$ etant directement évaluable pour nos fonctions, on se contente d'étudier la différence entre A_i et la valeur obtenue lorsque l'on étudiera le nombre de sondés.
Erreur Moyenne : $\frac{1}{N} \sum_i |\hat{f'}_i - A_i|$
Seuil d'acceptabilité (assez arbitraire) : $\frac{1}{10}|f'(I)|$

On prend 7 catégories.

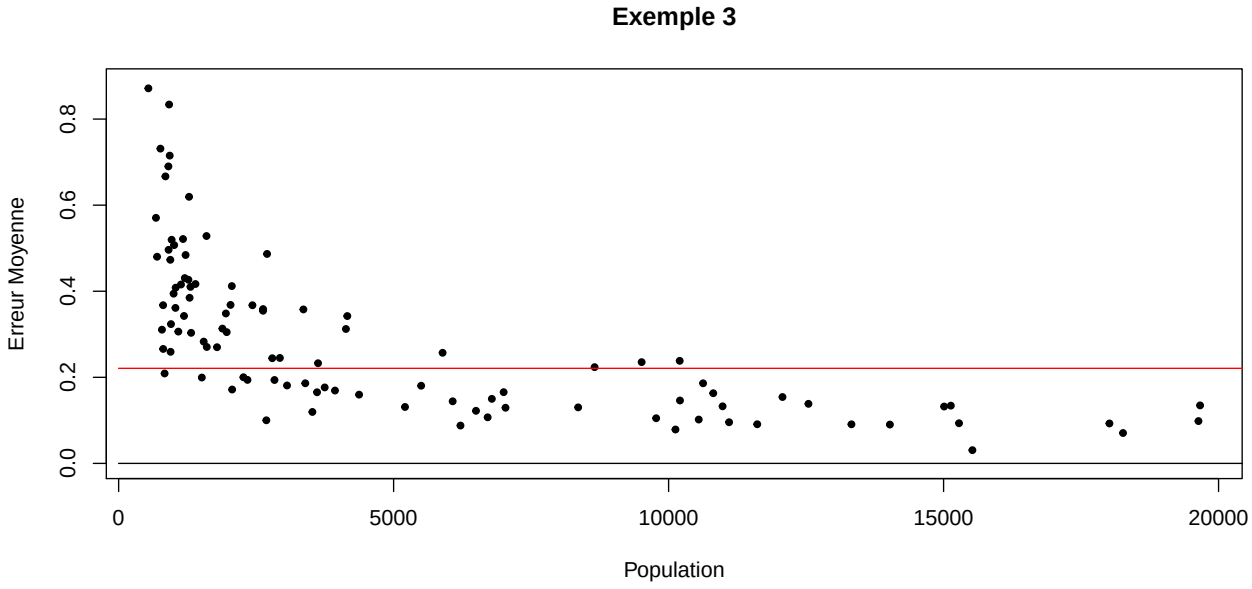
5.2 Exemple 1



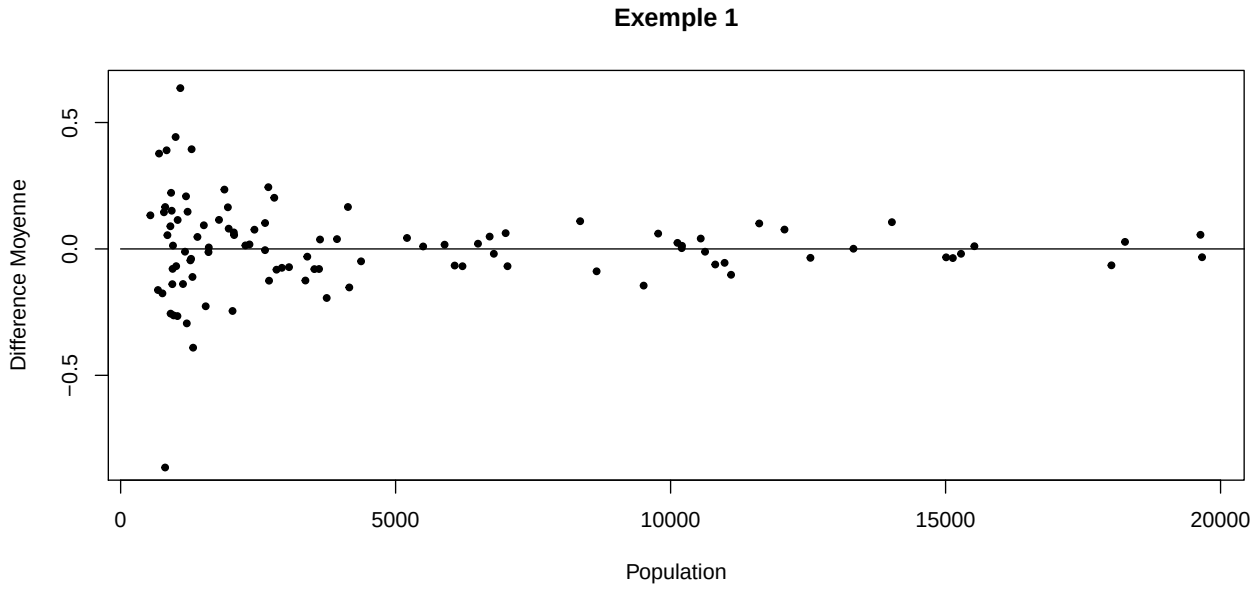
5.3 Exemple 2



5.4 Exemple 3



5.5 Biais



5.6 Interprétation

Malgré une variance très élevée, il ne semble pas il y avoir de biais supplémentaire (On a déjà supprimé le biais correspondant à $|f'_{x_{i'}} - A_i|$). L'erreur semble converger vers 0 pour tous les exemples, comme l'indique la loi des grands nombres pour un estimateur sans biais. L'erreur devient acceptable vers une population de 5000, ce qui est certes beaucoup. L'exemple 2 semble être meilleur, on peut l'expliquer par :
-Une dérivée simple
-Une grande amplitude (de probabilité, d'utilité)

6 En fonction du nombre de catégories

On cherche maintenant à déterminer combien de catégories doiventêtres faîtespour une population fixée.

6.1 Comment reconstituer f' avec les résultats obtenus ?

Si on a une idée de quoi ressemble cette fonction, on peut la reconstituer suivant ce modèle. En l'absence d'idée, il semble raisonnable de considérer f' affine par morceaux. Il existe plusieurs approches :

- Soit on relie les $f'(x'_i)$.
- Soit on relie des moyennes flottantes en prenant un poids pour les voisins.

La première approche est adaptée pour reconstituer dans les cas où il y a peu de catégories : Les points sont plus précis et les modifier n'as pas de sens.

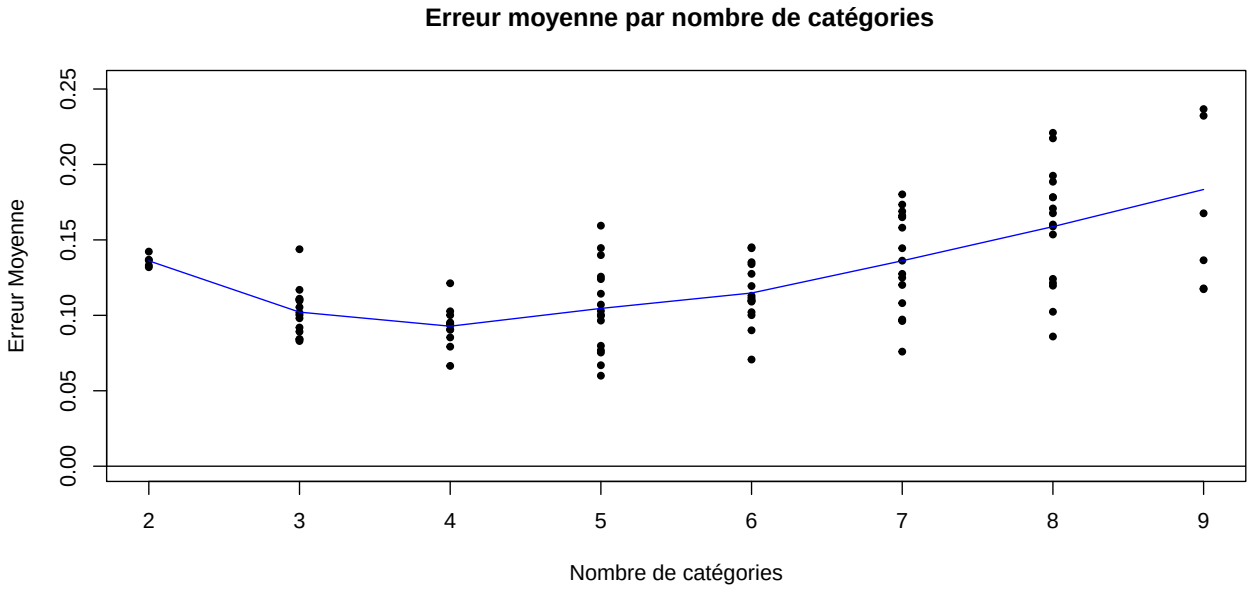
En revanche, dans le cas où il y a beaucoup de catégories, les points étant moins précis, on a davantage intêret à prendre une moyenne flottante, surtout si l'erreur est supposé sans biais et f' continue.

6.2 Mesure de l'erreur

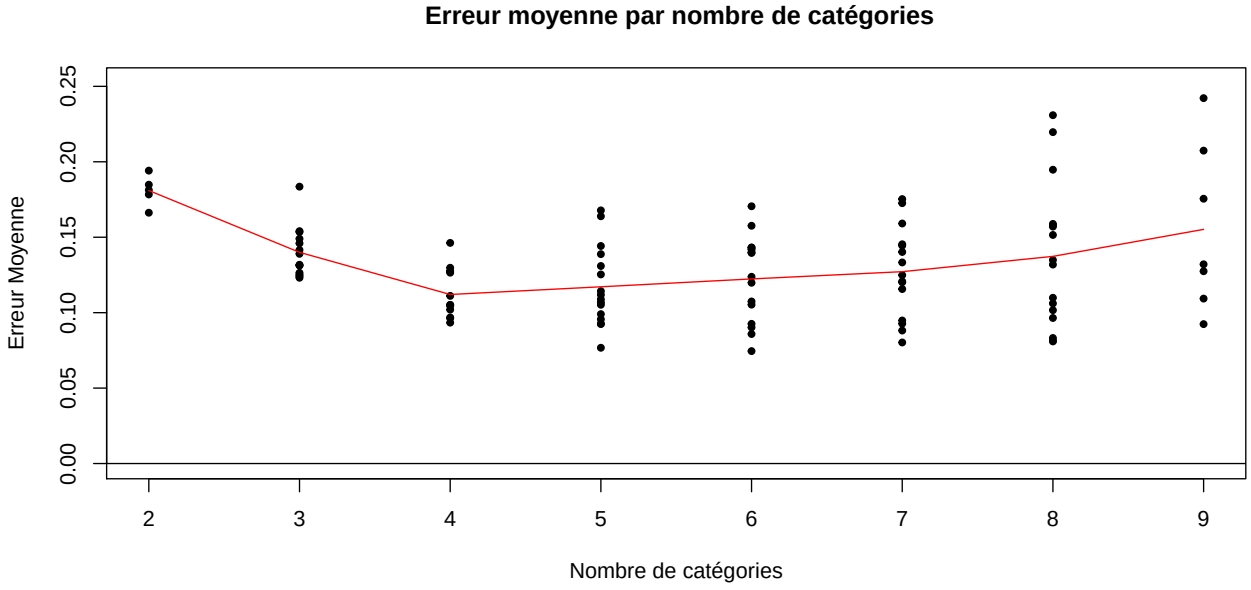
L'erreur sera donc la moyenne intégrale de la valeur absolue de la différence entre la fonction reconstitué
Avec une population de 10 000 individus :

6.3 Exemple 1

6.3.1 Mesure 1

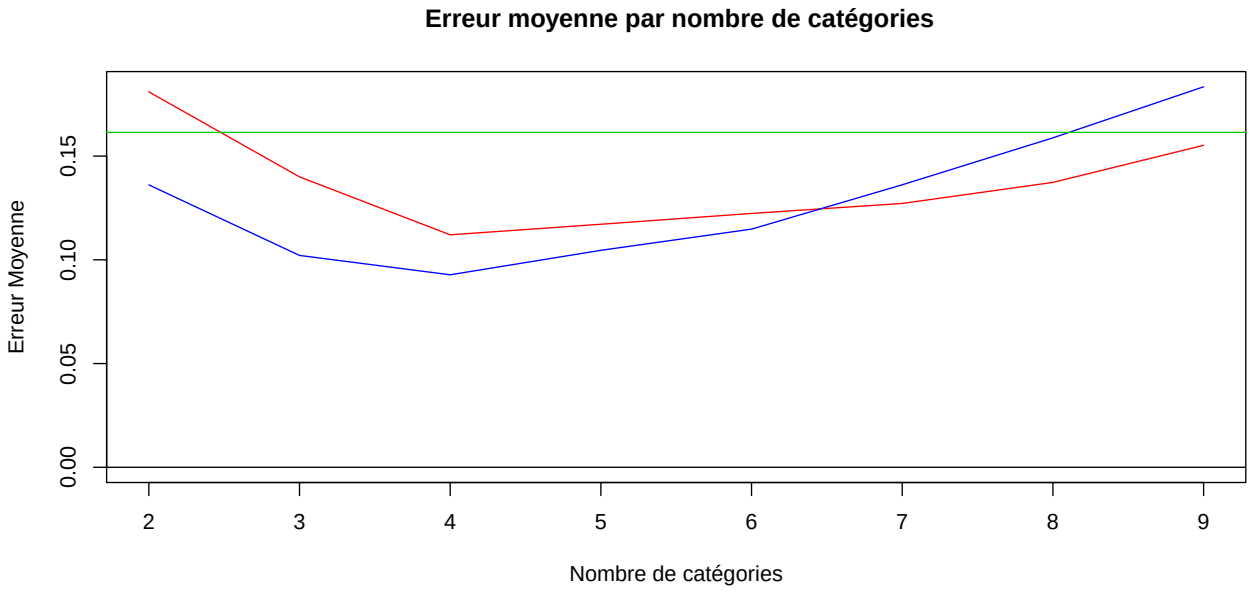


6.3.2 Mesure 2

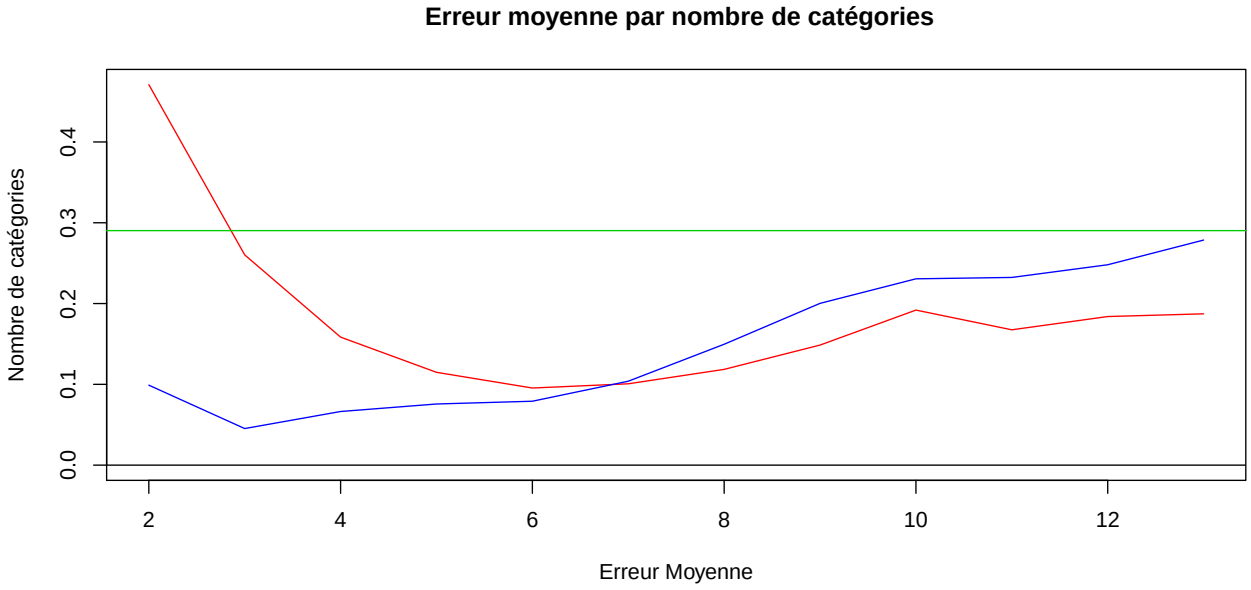


On représente le seuil d'acceptabilité précedemment défini. Il s'agit de la courbe verte.

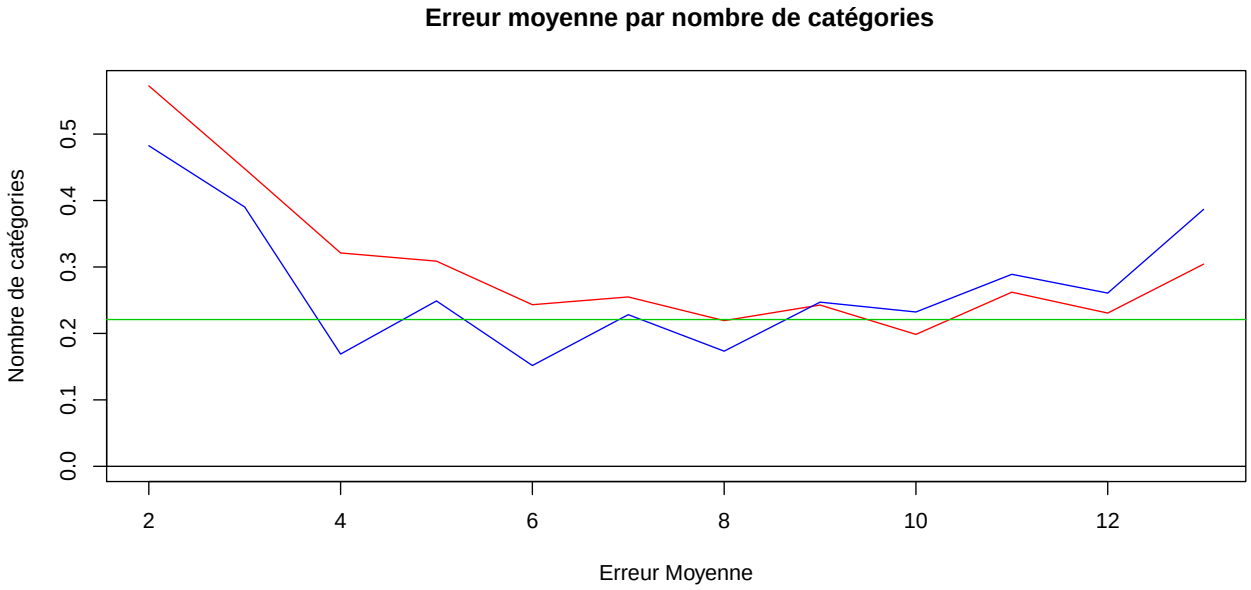
6.3.3 Comparaison



6.4 Exemple 2



6.5 Exemple 3



6.6 Interpretation

Il est donc pertinent de faire des catégories. Le nombre de catégories optimales pour les différentes fonctions va de 2 à 7, pour une population de 10000. Ce résultat doit être ramené à la population par catégorie. On remarque que le nombre de catégories doit être pair pour l'exemple 3, ce qui s'explique facilement avec la forme de la fonction. Changer les catégories légèrement peut donc être utile. Egalement, on remarque, en comparant avec les résultats des simulations de la partie précédente, que la courbe est mal reconstruite et on obtient donc un erreur assez élevé pour l'exemple 3.

7 Conclusion

On a donc une méthode qui permet d'extraire davantage d'information de données représentant un choix binaire lorsque l'utilité n'est pas une fonction linéaire des variables explicatives. La méthode nécessite d'avoir une population conséquente et de bien réflechir à comment reconstituer la fonction recherché avec les résultats obtenus. La régularité de la fonction recherché est également importante : Dans le cas où on n

7.1 Extension de la démarche.

Cette méthode est aussi adaptée lorsque l'on cherche à expliquer une variation de la dépendance à une variable explicative par rapport à une autre variable explicative indépendante.

Exemple : Une variation de confort dans le domaine des transports attirera davantage les plus agés que les plus jeunes ?

On modélise par : $V_1 = g(age) \times confort$ où f le facteur d'appreciation du confort en fonction de l'âge Le confort est indépendant de l'âge.

Note : Ici g correspond à f' de tout à l'heure.

Plus généralement avec $V_1 = f(x) \times y$ on applique la même méthode pour estimer les $f(x_i)$. On catégorise, puis on regarde le résultats du logit par rapport à y.

7.2 Cadre général d'utilisation

De manière générale, on peut toujours prendre qu'une partie des données pour mener notre analyse.
Dans le cas où on a d'autres catégories pertinentes à apporter, il faut le faire. (par exemple : retraité, chômeur, étudiant, salarié ...)
Si on a pas de catégories à faire, on peut également faire du clustering, c'est à dire rassembler nos variables par catégories. Cela nécessite une distance, qui n'est pas facile à déterminer. Ensuite il faut être capable d'interpréter les catégories former par notre algorithme (exemple algorithme k-moyennes).

Il faut faire attention à :
1) Faire des catégories qui ont du sens et que l'on pourra interpréter
2) Etre vigilant sur la population de chaque catégorie et donc sur le nombre de catégories, si elles sont de même taille etc. Suffisamment grande pour avoir un modèle performant, suffisamment précise pour détecter les changements.

On retiendra donc comme critère empirique que les catégories doivent comprendre au moins en moyenne 700 personnes si on veut une précision moyenne, environ 2000 personnes si on veut une bonne précision. Dans le cas où les catégories on été faîtes naïvement, Il est pertinent de reconduire l'analyse en les perturbants afin de voir si ces catégories précisément ne cachaient pas un phénomène.
3) Si l'on s'attends à un brusque changement de comportement autour d'une valeur, augmenter le nombre de sondés proche cette valeur si on veut justement l'estimer correctement.

8 Bibliographie

[1]Yves Croissant *Estimation of multinomial logit in R*
[2]http://en.wikipedia.org *Multinomial logistic regression*
[3]Chandra R. Bhat. *Random Uility-Based for Discrete Choice Models for Public Transport*
[4] William Green. *Discrete Choice*
[5] Daniel McFadden. *Conditional logit analysis of qualitative choice behaviour*

9 Remerciements

Merci à Pierre Lemaire de m'avoir encadré pendant la durée de ce projet et ses conseils judicieux.