Lien entre le modèle à rapport de côtes proportionnelles et le modèle logit ordinal hétéroscédastique

Christian Derquenne

Recherche et Développement

Plan

1 Contexte – motivations – objectif

2 Modèle logit hétéroscédastique

3 Lien entre le modèle à rapport de côtes proportionnelles et le modèle logit ordinal hétéroscédastique

4 Application à des données marketing

5 Apports – applications – voies futures

Contexte – motivations – objectif

(i) Utilisation de modèle logit ordinal dans de nombreux domaines (économétrie, essais thérapeutiques, génétiques, marketing, …)

(ii) Variables « explicatives » catégorielles :

→ données groupées (croisement des modalités)

→ coefficients de position = niveau moyen/groupe

(iii) Deux groupes avec le même niveau ⇒ même comportement/variable réponse

Contexte – motivations – objectif

(iv) Problème : dispersions différentes

(v) Solution raisonnable :

→ modèle logit ordinal hétéroscédastique (MLOH) [Mc. Cullagh, 1980, Derquenne, 1995, 1996, Foulley, 1995, 1996]

→ coefficients de dispersion en plus

(vi) Modèle logit ordinal = modèle à rapport de côtes proportionnelles (MRCP)

→ utilisation d'un test (statistique du score)

→ rejet de $H_0 \Rightarrow$ ex : modèle logit généralisé (MLG)

Contexte – motivations – objectif

(vii) Problème du MLG :

→ plus difficile à interpréter

→ plus gourmand en nombre de paramètres

(viii) Objectifs du papier :

→ MRCP ⇔ modèle logit ordinal homoscédastique

→ utilisation d'un MLOH à la place d'un modèle logit généralisé

→ introduction d'un système de re-pondération des individus pour revenir à un MRCP

Plan

1 Contexte – motivations – objectif

2 Modèle logit hétéroscédastique

3 Lien entre le modèle à rapport de côtes proportionnelles et le modèle logit ordinal hétéroscédastique

4 Application à des données marketing

5 Apports – applications – voies futures

Modèle logit hétéroscédastique

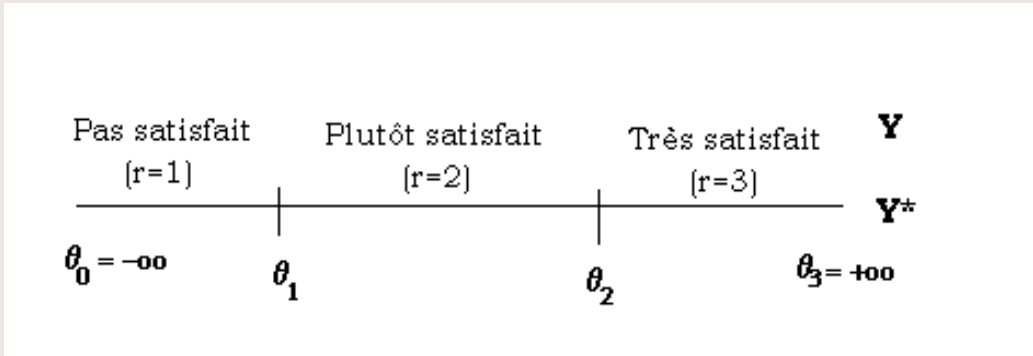
Modèle logit ordinal homoscédastique :

$$Y^* = \mu_0 + X\beta + \varepsilon \qquad (1)$$

Distribution logistique sous-jacente :

$$\Pr(Y \le r / X) = \Pr(Y^* \le \alpha_r / X) = \frac{\exp^{(\alpha_r - X\beta)}}{1 + \exp^{(\alpha_r - X\beta)}} = F(\alpha_r - X\beta) \qquad (2)$$

où $\alpha_r = \theta_r - \mu_0$



Modèle logit hétéroscédastique

Données groupées :

$$Y_t^* = \mu_0 + \sum_{j=1}^{p} \sum_{i/j=1}^{k_j} \beta_{i/j} x_{i/j,t} + \varepsilon \qquad (3)$$

Pour un groupe m :

$$g_m = \mathbf{x}'_m \beta \qquad (4)$$

où $\mathbf{x}_m$ = vec-col de 0/1 du groupe m, de taille $\sum_{j=1}^{p} k_j - p$

et β = vecteur-paramètres de position

Modèle logit hétéroscédastique

Modèle logit ordinal hétéroscédastique :

$$Y^{*}=\mu_0+X\beta+\varepsilon\sigma \tag{5}$$

Distribution logistique sous-jacente :

$$\Pr(Y \le r / X)=\Pr(Y^{*} \le \alpha_r / X)=\frac{\exp^{((\alpha_r-X\beta)/\sigma)}}{1+\exp^{((\alpha_r-X\beta)/\sigma)}} \tag{6}$$

Pour un groupe *m* :

$$\Pr(Y \le r / m)=\Pr(Y^{*} \le \alpha_r / X)=\frac{\exp^{((\alpha_r-g_m)/\sigma_m)}}{1+\exp^{((\alpha_r-g_m)/\sigma_m)}} \tag{7}$$

où $\sigma_m=\exp(\mathbf{x}'_m\upsilon)$

avec ν = vecteur du log des coefficients de dispersion

Plan

1 Contexte – motivations – objectif

2 Modèle logit hétéroscédastique

3 Lien entre le modèle à rapport de côtes proportionnelles et le modèle logit ordinal hétéroscédastique

4 Application à des données marketing

5 Apports – applications – voies futures

Lien entre MRCP et MLOH

Modèle à rapport de côtes proportionnelles :

$$\ln\left[\frac{\Pr[Y \le r / X = x]}{1 - \Pr[Y \le r / X = x]}\right] = \alpha_r - X\beta \quad \forall r = 1, R-1 \qquad (8)$$

Pour deux valeurs de $X = (x_1, x_2)$:

$$\left[\frac{\Pr[Y \le r / X = x_1]}{1 - \Pr[Y \le r / X = x_1]}\right] \Big/ \left[\frac{\Pr[Y \le r / X = x_2]}{1 - \Pr[Y \le r / X = x_2]}\right] = \exp\left(-\left(x_1 - x_2\right)\beta\right) \qquad (9)$$

Ce rapport est indépendant de la réponse r

Lien entre MRCP et MLOH

Données groupées, pour deux groupes *m* et *m'* :

$$\frac{OR_m^{(r)}}{OR_{m'}^{(r)}} = \frac{OR_m^{(r')}}{OR_{m'}^{(r')}} \quad \Leftrightarrow \quad \exp\left(g_m^{(r)} - g_{m'}^{(r)}\right) = \exp\left(g_m^{(r')} - g_{m'}^{(r')}\right) \quad (10)$$

où $OR_m^{(r)} = \exp\left(\alpha_r - g_m^{(r)}\right)$ = rapport de chances

alors : $g_m^{(r)} - g_m^{(r')} = g_{m'}^{(r)} - g_{m'}^{(r')} \Leftrightarrow 0 = 0$

car les coefficients de position pour deux réponses *r* et *r'* d'un groupe *m* sont égaux par hypothèse du modèle RCP

Lien entre MRCP et MLOH

Problème : si cette hypothèse ne tient pas

→ Utilisation du test du score ("pentes parallèles")

→ Mise en œuvre du modèle logit généralisé (MLG)

$$\Pr(Y = r / m) = \frac{\exp\left(\alpha_r - g_m^{(r)}\right)}{1 + \sum_{k=1}^{R-1} \exp\left(\alpha_k - g_m^{(r)}\right)} \qquad \Pr(Y = R / m) = \frac{1}{1 + \sum_{k=1}^{R-1} \exp\left(\alpha_k - g_m^{(r)}\right)} \tag{11}$$

Rapport de chances avec des coefficients de dispersion

$$OR_m^{(r)} = \exp\left(\frac{\alpha_r - g_m^{(r)}}{\sigma_m}\right) \tag{12}$$

→ modèle logit ordinal hétéroscédastique

Lien entre MRCP et MLOH

Théorème 1 :

Soit Y une variable se distribuant sur une échelle ordinale à R catégories, soient $X_1,...,X_j,...,X_p$, p prédicteurs catégoriels permettant de générer M groupes d'individus et soit enfin un modèle logit ajustant la variable Y à l'aide des p variables X_j, alors ce modèle sera à rapport de côtes proportionnelles, si et seulement si le modèle logit ordinal est homoscédastique, c'est-à-dire $\forall$ (m,m'), $\sigma_m = \sigma_{m'}$ où σ_m et $\sigma_{m'}$ sont les écarts-types des groupes m et m'

Lien entre MRCP et MLOH

Démonstration :

Grâce à (10) et (12), on peut écrire :

$$\exp\left(\frac{(\sigma_m-\sigma_{m'})\alpha_r}{\sigma_m\sigma_{m'}}-\left(\frac{g_m^{(r)}}{\sigma_m}-\frac{g_{m'}^{(r)}}{\sigma_{m'}}\right)\right)=\exp\left(\frac{(\sigma_m-\sigma_{m'})\alpha_{r'}}{\sigma_m\sigma_{m'}}-\left(\frac{g_m^{(r')}}{\sigma_m}-\frac{g_{m'}^{(r')}}{\sigma_{m'}}\right)\right) \quad (13)$$

$$\Leftrightarrow$$

$$(\sigma_m-\sigma_{m'})\alpha_r-\sigma_{m'}(g_m^{(r)}-g_m^{(r')})=(\sigma_m-\sigma_{m'})\alpha_{r'}-\sigma_m(g_{m'}^{(r)}-g_{m'}^{(r')})$$

Par conséquent si $\sigma_m=\sigma_{m'}$ alors $(g_m^{(r)}-g_m^{(r')})=(g_{m'}^{(r)}-g_{m'}^{(r')})$

Lien entre MRCP et MLOH

Corollaire :

Si le modèle logit ordinal est homoscédastique, alors le test du score du rapport de côtes proportionnelles (hypothèse des pentes parallèles dans le cas de variables X numériques) sera mécaniquement non significatif. Par conséquent, ce test permettra aussi de détecter l'hétéroscédasticité dans les données ajustées en postulant un modèle logit ordinal

Démonstration :

Elle découle directement du théorème précédent

Lien entre MRCP et MLOH

Théorème 2 :

Si le modèle logit ordinal est hétéroscédastique, alors il est possible de trouver une transformation adéquate du système de pondération des individus pour revenir à un modèle à rapport de côtes proportionnelles

Démonstration :

Pour chaque groupe *m*, la version observée de la probabilité cumulée théorique pour la réponse r (7) a la forme suivante :

$$\tilde{\Pr}(Y \le r \,/\, \text{groupe } m) = \frac{\sum_{k=1}^{r} n_{mk}}{n_m} = \frac{\exp\left(\left(\tilde{\alpha}_r - \tilde{g}_m\right)/\tilde{\sigma}_m\right)}{1+\exp\left(\left(\tilde{\alpha}_r - \tilde{g}_m\right)/\tilde{\sigma}_m\right)} \qquad (14)$$

Lien entre MRCP et MLOH

Démonstration (suite) :

Le rapport de chances empirique a la forme suivante :

$$\exp\left(\left(\tilde{\alpha}_r - \tilde{g}_m\right)/\tilde{\sigma}_m\right) = \sum_{k=1}^{r} n_{mr} \Big/ \left(n_m - \sum_{k=1}^{r} n_{mr} \right)$$

$$\Leftrightarrow$$

$$\exp\left(\tilde{\alpha}_r - \tilde{g}_m\right) = \left(\sum_{k=1}^{r} n_{mr} \Big/ \left(n_m - \sum_{k=1}^{r} n_{mr} \right) \right)^{\tilde{\sigma}_m}$$

alors :

$$\frac{\exp\left(\tilde{\alpha}_r - \tilde{g}_m\right)}{1+\exp\left(\tilde{\alpha}_r - \tilde{g}_m\right)} = \left(\sum_{k=1}^{r} n_{mk} \right)^{\tilde{\sigma}_m} \Big/ \left[\left(n_m - \sum_{k=1}^{r} n_{mk} \right)^{\tilde{\sigma}_m} + \left(\sum_{k=1}^{r} n_{mk} \right)^{\tilde{\sigma}_m} \right] \quad (15)$$

Lien entre MRCP et MLOH

Démonstration (suite) :

Pseudo proportion pour une réponse r d'un groupe m :

$$\tilde{\pi}_{mr} = \frac{\left(\sum_{k=1}^{r} n_{mk}\right)^{\tilde{\sigma}_m}}{\left(n_m - \sum_{k=1}^{r} n_{mk}\right)^{\tilde{\sigma}_m} + \left(\sum_{k=1}^{r} n_{mk}\right)^{\tilde{\sigma}_m}} - \frac{\left(\sum_{k=1}^{r-1} n_{mk}\right)^{\tilde{\sigma}_m}}{\left(n_m - \sum_{k=1}^{r-1} n_{mk}\right)^{\tilde{\sigma}_m} + \left(\sum_{k=1}^{r-1} n_{mk}\right)^{\tilde{\sigma}_m}} \quad (16)$$

En remplaçant $\tilde{\sigma}_m$ par $\hat{\sigma}_m$ alors on obtient une valeur estimée du nombre de réponses r pour le groupe m : $\tilde{n}_{mr} = n_m \tilde{\pi}_{mr}$

Le nouveau poids pour un individu t est : $\tilde{w}_t = w_t \frac{\tilde{n}_{mr}}{n_{mr}}$

Plan

1 Contexte – motivations – objectif

2 Modèle logit hétéroscédastique

3 Lien entre le modèle à rapport de côtes proportionnelles et le modèle logit ordinal hétéroscédastique

4 Application à des données marketing

5 Apports – applications – voies futures

Application à des données marketing

Enquête de satisfaction sur 3500 clients

Objectif : modéliser une réponse ordinale (3 niveaux) à l'aide de quatre variables candidates à l'explication contenant respectivement (2,2,5,5) catégories

Modèles utilisés :

→ logit ordinal (rapport de côtes proportionnelles)
→ logit ordinal hétéroscédastique
→ logit généralisé (nominal)
→ logit ordinal avec poids hétéroscédastique

Application à des données marketing

Mise en œuvre :

→ 99 groupes d'individus (non vides)

→ 12 paramètres pour le logit ordinal (MRCP)

→ 22 paramètres pour le logit ordinal hétéroscédastique (MLOH)

→ 22 paramètres pour le logit généralisé (MLG)

→ 12 paramètres pour le logit ordinal repondéré

Application à des données marketing

Modèle logit ordinal (MRCP)

p-valeurs : **score** = 0,0311 et **déviance** = 0,4950

Parameter	Effect	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
alpha1	Fixed	-1.4523	0.1556	87.0798	<.0001
alpha2	Fixed	1.9486	0.1616	145.4364	<.0001
< 35 ans	Fixed	-0.7917	0.1438	30.3183	<.0001
35-44ans	Fixed	-0.7408	0.1459	25.7888	<.0001
45-54ans	Fixed	-0.5054	0.1499	11.3646	0.0007
55-64ans	Fixed	-0.2171	0.1532	2.0088	0.1564
> 65 ans	Fixed	0	.	.	.
Locatair	Fixed	-0.4575	0.0691	43.7738	<.0001
Propriet	Fixed	0	.	.	.
Collect	Fixed	0.3060	0.0681	20.2138	<.0001
Individ	Fixed	0	.	.	.
Autre	Fixed	0.5284	0.1844	8.2142	0.0042
Bois	Fixed	0.1969	0.1369	2.0678	0.1504
Electric	Fixed	1.9372	0.0951	414.9034	<.0001
Fioul	Fixed	-0.3952	0.1426	7.6767	0.0056
Gaz v r	Fixed	0	.	.	.

Application à des données marketing

Parameter	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
alpha1	0.1635	0.3119	0.2748	0.6002
alpha2	1.4460	0.1885	58.8376	<.0001
18_34ans	1.3890	0.2956	22.0749	<.0001
18_34ans	0.8009	0.1585	25.5409	<.0001
35_44ans	1.3033	0.2997	18.9066	<.0001
35_44ans	0.7343	0.1597	21.1400	<.0001
Pas satis	0.9352	0.3063	9.3213	0.0023
45_54ans	0.3730	0.1608	5.3799	0.0204
55_64ans	0.3923	0.3204	1.4996	0.2207
55_64ans	0.2088	0.1616	1.6685	0.1965
__65_ans	0	.	.	.
__65_ans	0	.	.	.
Locatair	0.8438	0.1194	49.9129	<.0001
Locatair	0.2747	0.0850	10.4564	0.0012
Propriet	0	.	.	.
Propriet	0	.	.	.
Collect	-0.5089	0.1170	18.9326	<.0001
Collect	-0.3304	0.0849	15.1541	<.0001
Individ	0	.	.	.
Individ	0	.	.	.
Autre	-0.8129	0.3112	6.8225	0.0090
Autre	-0.3631	0.2789	1.6954	0.1929
Bois	-0.3174	0.2559	1.5387	0.2148
Bois	-0.1331	0.2414	0.3040	0.5814
Electric	-3.2383	0.1662	379.8164	<.0001
Electric	-1.3604	0.1308	108.1331	<.0001
Fioul	0.7744	0.3348	5.3496	0.0207
Fioul	0.4793	0.3287	2.1265	0.1448
Gaz v r	0	.	.	.
Gaz v r	0	.	.	.

Modèle logit nominal (MLG)

p-valeur : déviance = 0,7038

Application à des données marketing

parameter	Effect	Estimate	Standard Error	Wald Chi-square	Pr> ChiSq
alpha1	Fixed	-1.5329	0.2159	50.4009	0.0000
alpha2	Fixed	2.0884	0.2776	56.6002	0.0000
18-34ans	Fixed	-0.8014	0.1616	24.5948	0.0000
35-44ans	Fixed	-0.7705	0.1625	22.4729	0.0000
45-54ans	Fixed	-0.5136	0.1585	10.4993	0.0012
55-64ans	Fixed	-0.2161	0.1539	1.9722	0.1602
> 65 ans	Fixed	0.0000	.	.	.
18-34ans	Random	-0.0119	0.1009	0.0138	0.9063
35-44ans	Random	0.0345	0.1024	0.1138	0.7359
45-54ans	Random	0.0931	0.1057	0.7763	0.3783
55-64ans	Random	0.0345	0.1095	0.0993	0.7527
> 65 ans	Random	0.0000	.	.	.
Locatair	Fixed	-0.4878	0.0878	30.8601	0.0000
Propriet	Fixed	0.0000	.	.	.
Locatair	Random	0.0927	0.0436	4.5121	0.0337
Propriet	Random	0.0000	.	.	.
Collect	Fixed	0.3007	0.0777	14.9794	0.0001
Individ	Fixed	0.0000	.	.	.
Collect	Random	0.1107	0.0437	6.4272	0.0112
Individ	Random	0.0000	.	.	.
Autre	Fixed	0.5604	0.2160	6.7309	0.0095
Bois	Fixed	0.1998	0.1535	1.6936	0.1931
Electric	Fixed	2.0756	0.2361	77.2850	0.0000
Fioul	Fixed	-0.4193	0.1580	7.0411	0.0080
Gaz v r	Fixed	0.0000	.	.	.
Autre	Random	0.0144	0.1014	0.0203	0.8868
Bois	Random	-0.0159	0.0827	0.0370	0.8475
Electric	Random	-0.1316	0.0514	6.5647	0.0104
Fioul	Random	-0.0906	0.1015	0.7978	0.3718
Gaz v r	Random	0.0000	.	.	.

Modèle logit ordinal hétéroscédastique

p-valeur(dév) = 0,7401

Application à des données marketing

Modèle logit ordinal avec la pondération hétéroscédastique

p-valeurs : **score** = 0,9913 et **déviance** = 0,8348

Parameter	Effect	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
alpha1	Fixed	-1.5327	0.1492	105.5757	<.0001
alpha2	Fixed	2.0706	0.1562	175.6715	<.0001
18_34ans	Fixed	0.8285	0.1376	36.2368	<.0001
35_44ans	Fixed	0.7604	0.1396	29.6810	<.0001
45_54ans	Fixed	0.5335	0.1434	13.8383	0.0002
55_64ans	Fixed	0.2197	0.1464	2.2521	0.1334
__65_ans	Fixed	0	.	.	.
Locatair	Fixed	0.4835	0.0664	53.0149	<.0001
Propriet	Fixed	0	.	.	.
Collect	Fixed	-0.3210	0.0653	24.1894	<.0001
Individ	Fixed	0	.	.	.
Autre	Fixed	-0.5636	0.1786	9.9571	0.0016
Bois	Fixed	-0.1964	0.1316	2.2276	0.1356
Electric	Fixed	-2.0585	0.0939	480.8215	<.0001
Fioul	Fixed	0.4259	0.1362	9.7734	0.0018
Gaz v r	Fixed	0	.	.	.

Plan

1 Contexte – motivations – objectif

2 Modèle logit hétéroscédastique

3 Lien entre le modèle à rapport de côtes proportionnelles et le modèle logit ordinal hétéroscédastique

4 Application à des données marketing

5 Apports – applications – voies futures

Apports – applications – voies futures

(i) Equivalence d'un modèle logit ordinal homoscédastique et d'un modèle à rapport de côtes proportionnelles

(ii) Un modèle logit ordinal hétéroscédastique n'est pas un modèle à rapport de côtes proportionnelles

(iii) Si l'hypothèse nulle du rapport de côtes proportionnelles est rejetée, alors il est préférable d'ajuster un modèle logit ordinal hétéroscédastique à la place d'un modèle logit nominal :

→ plus coûteux ($R > 3$)

→ résultats plus difficilement interprétables

Apports – applications – voies futures

(iv) Introduction d'une transformation des poids des individus construite à partir des coefficients de dispersion du modèle hétéroscédastique pour obtenir un modèle à rapport de côtes proportionnelles

(v) Avec l'application de cette transformation, le test du score associé devient mécaniquement fortement non significatif, ce qui permet de travailler directement avec un modèle rapport de côtes proportionnelles

(vi) Démarche très fructueuse dans différents domaines d'applications (marketing, épidémiologie, génétique, économétrie, sensiométrie, ...)

(vii) Extension du modèle logit ordinal hétéroscédastique à des modèles plus complexes : modèles à équations structurelles (variables observées sur des échelles ordinales)

Bibliographie

Derquenne Ch. (1995), Heteroskedastic Logit Model, *50th Session of the International Statistical Institute*, Bejing - China

Derquenne Ch. (1996), Modèle Logit Dichotomique Hétéroscédastique, *XXVIIIèmes Journées de Statistique*, Laval - Québec

Derquenne Ch. (1996), Heteroskedastic Ordinal Logit Model, *4th World Congress of the Bernoulli Society*, Vienna - Austria

Derquenne Ch. (2006), *Traitements statistiques de données catégorielles : Recherche exploratoire de structures et modélisation de phénomènes*, thèse en Mathématiques Appliquées et Sciences Sociales, Université Paris-Dauphine, Paris, France

Foulley J-L. & Qaas RL., (1995), Heterogeneous variances in Gaussian Linear mixed models, *Genet Sel Evol*, **27**, 211-228

Bibliographie

Foulley J-L. & Gianola D., (1996), Statistical analysis of ordered categorical data via a structural heteroskedastic threshold model, Genet Sel Evol, **28**, 249-273

Mc Cullagh P., (1980), Regression Models for Ordinal Data, *J.R. Stat. Soc.,* B **42**, 109-142