Comportement asymptotique

Fabian Bastin

Avril 2011

Les paramètres sont estimés par maximum de vraisemblance, et les propriétés asymptotiques en dérivent directement. Elles sont pourtant souvent mal connues. . .

Une référence clé :
Newey, Whitney K., and Daniel McFadden. ”Large Sample Estimation and Hypothesis Testing.” In Handbook of Econometrics. Vol. 4. Edited by Daniel McFadden and Robert Engle. Amsterdam, The Netherlands : Elsevier, North-Holland, 1986, chapter 36, pp. 2111-2245.

Contexte général

Nous considérons un estimateur statistique du vecteur de paramètres θ_0, défini come une fonction de la forme

$$h(\cdot) : \mathcal{X} \to \Theta \subseteq \mathcal{R}^K,$$

où $\mathcal{X}$ est l'espace d'échantillonnage (ici la population et les situation de choix), et Θ est l'espace de paramètres (i.e. l'ensemble des valeurs que l'estimateur peut prendre).

Nous écrirons aussi l'estimateur comme

$$\hat{\boldsymbol{\theta}} = h(X_1, X_2, \ldots, X_N) := h(X).$$

Une valeur particulière de cette estimateur, basé sur une réalisation d'échantillonnage particulière $x := (x_1, x_2, \ldots, x_N)$, est appelé un estimé de θ, dénoté par

$$\hat{\theta} = h(x_1, x_2, \ldots, x_N) := h(x).$$

Contexte général (suite)

La distribution échantionnale de $\hat{\theta}$ est défini comme la distribution de $h(X)$, et nous dénotons sa fonction de densité par

$$f(\hat{\theta}; x_1, x_2, \ldots, x_N) := \frac{1}{N} f(\hat{\theta}; x).$$

Nous dénoterons aussi la distribution de l'échantillon par $f(x; \theta_0)$, la distribution jointe de $X := (X_1, X_2, \ldots, X_N)$, sous le paramètre θ_0.

On ne connaît que rarement la forme f, et on utilisera une approximation $\hat{f}$.

Maximum de vraisemblance

Un cas particulier est l'estimateur de maximum de vraisemblance :

$$\hat{\theta}_{ML} = \arg\max_{\tilde{\theta}} \hat{f}(\tilde{\theta}; x).$$

Si les X_i. $i = 1, \ldots, N$, sont i.i.d., nous avons

$$\hat{f}(\tilde{\theta}; x) = \frac{1}{N} \prod_{i=1}^{N} \hat{f}(\tilde{\theta}; x_i).$$

Il est cependant habituellement préférable de résoudre le programme

$$\max_{\tilde{\theta}} \frac{1}{N} \sum_{i=1}^{N} \ln \hat{f}(\tilde{\theta}; x_i).$$

Conditions de régularité

Un modèle de probabilité est dit regulier si la distribution de l'échantillon $\hat{f}(x;\theta) = \hat{f}(x_1, x_2, \ldots, x_N;\theta)$ satisfait les conditions de régularités suivantes.

- R1 L'espace de paramètres Θ est un sous-ensemble ouvert de $\mathcal{R}^K$, $K < N$,
- R2 Le support de la distribution $\mathcal{X}_0 := \{x \mid \hat{f}(x;\theta) > 0\}$ est le même pour tout $\theta \in \Theta$.
- R3 La fonction de score

$$s(\theta) := \nabla_\theta \ln \hat{f}(x;\theta),$$

  existe et est finie $\forall\, \theta \in \Theta$, $x \in \mathcal{X}_0$.
- R4 Considérant l'estimateur $h(x)$, nous pouvons permuter la différentiation par rapport à θ et l'intégration par rapport à x, i.e.

$$\nabla_\theta \int h(x)\hat{f}(x;\theta)dx = \int h(x)\nabla_\theta \hat{f}(x;\theta)dx < \infty.$$

Matrice d'information

Pour de tels modèles réguliers, nous définissons la matrice d'information de Fisher pour l'échantillon $(x_1, \ldots, x_N)$ comme

$$I(\theta_0) = E[s(x;\theta_0)s(x;\theta_0)^T],$$

qui peut être estimé comme

$$I_N(\hat{\theta}) = \frac{1}{N}\sum_{n=1}^{N} s(x_m;\hat{\theta})s(x_m;\hat{\theta})^T.$$

Normalité asymptotique

Soit $L(\theta; x) = \ln \hat{f}(\theta); x)$. Sous **R1**, nous avons :

$$0 = \frac{1}{N} \sum_{i=1}^{N} \nabla_\theta \ln f(\hat{\theta}_{ML}; x).$$

Nous exigeons de plus

- R5 $L(\cdot; x) : \Theta \to [0, \infty)$ est continue en tout $\theta \in \Theta$,
- R6 Pour tout $\theta_1 \neq \theta_2$ where $\theta_1 \in \Theta$, $\theta_2 \in \Theta$, $f(x; \theta_1) \neq f(x; \theta_2)$.
- R7 $E[\ln f(X; \theta)]$ exists, and $\theta_0 = \arg\max_{\tilde{\theta}} E[\ln f(\theta; X)]$.
- R8 $\frac{1}{N} L(\theta; X)$ converge presque sûrement vers $E[\ln f(x; \theta)]$ pour tout $\theta \in \Theta$.
- R9 $\ln L(\theta; x)$ est deux fois continûment différentiable dans un intervalle ouvert autour de θ.
- R10 θ_0 et $\hat{\theta}_N = \arg\max_\theta \frac{1}{N} L(\theta; X)$ (qu'on suppose exister et unique) appartiennent à un certain sous-ensemble comptact de Θ, presque sûrement, pour N assez grand.

Normalité asymptotique (suite)

Ces conditions assurent que l'estimateur $\hat{\theta}_N$ est consistant, i.e. $\hat{\theta}_N \to \theta_0$, même si la distribution $\hat{f}$ est mal spécifiée ($\hat{f} \neq f$).

Alors, du théorème de Slutzky,

$$\sqrt{n}\left(\hat{\theta}_{ML} - \theta_0\right) \xrightarrow{d} N(0, H^{-1}(\theta_0)I(\theta_0)H^{-1}(\theta_0)). \tag{1}$$

Supposons de plus

R12 Les opérations d'intégration et de differentiation peuvent être échangée pour the second derivatives of $f(x;\theta)$ when considering $h(x)$, i.e.

$$\nabla^2_{\theta\theta} \int h(x)f(x;\theta)dx = \int h(x)\nabla^2_{\theta\theta} f(x;\theta)dx.$$

Normalité asymptotique (suite)

Sous **R1–R12**, on peut montrer que

$$I(\theta_0) = -E\left[\nabla^2_{\theta\theta} \ln f(X;\theta_0)\right], \tag{2}$$

i.e. la matrice d information est égale à l'opposé du hessien de la log-vraisemblance, évaluée en θ_0.

Cette égalite appelée égalité de la matrice d information.

Elle est souvent utilisée, comme la matrice de variance-covariance matrix peut alors être réduite à $H^{-1}(\theta_0)$.

Violations des hypothèses de normalité

R12 ne tient pas en présence de mauvaise spécification ; dans ce cas, l'égalité de la matrice d'information n'est pas valide.

Pour des spécifications consistantes, , $H^{-1}(\theta_0)I(\theta_0)H^{-1}(\theta_0)$ Donne toujours la matrice de variance-covariance, qui peut être estimée par

$$H_N^{-1}(\hat{\theta}_N)I_N(\hat{\theta})H_N^{-1}(\hat{\theta}_N)$$

Un important cas de mauvaise spécification est quand les observations $X_1, \ldots, X_N$ sont dépendentes, aussi

$$f(\theta; X_1, \ldots, X_N) \neq \prod_{i=1}^{N} f(\theta; X_i).$$

Un tel cas arrive quand nous estimons un logit sur base de données panel data, comme les choix d'un même individu sont typiquement correlées.

Beaucoup d'études se sont attachées à démontrer la consistance de l'estimateur, sous la fausse hypothèse d'indépendance (mais pour un modèle sinon correctement spécifié). Dès lors, l'etimateur

$$H_N^{-1}(\hat{\theta}_N) I_N(\hat{\theta}) H_N^{-1}(\hat{\theta}_N)$$

est encore valide, mais pas $H_N^{-1}(\hat{\theta}_N)$, qui ne peut plus être vu quand approximation de la matrice d'information.

Test statistique des estimateurs

Sous l'hypothèse que l'estimateur θ_i est normalement distribué avec la variance σ_i^2, la statistique

$$t_{\theta_i} = \frac{\hat{\theta}_i - 0}{\sigma_i},$$

suit une distribution de Student, de moyenne 0 et variance 1, qui peut être approximée par une $\mathcal{N}(0, 1)$.
En pratique, le nombre d'observations est habituellement assez grand pour rendre cette approximation fiable.

Nous pouvons construire un intervalle de confiance de niveau α comme suit :

$$(\hat{\theta}_i - Z_{\alpha/2}\sigma_i, \hat{\theta}_i + Z_{\alpha/2}\sigma_i),$$

$Z_{\alpha/2}$ est le quantile $1 - \alpha/2$ d'une $\mathcal{N}(0, 1)$.

Test statistique des estimateurs (suite)

Considérons le test statistique

$$H_0 : \theta_i = 0, \quad H_1 : \theta_i \neq 0,$$

pour un certain i dans $\{1, \ldots, K\}$. Nous rejetons H_0 si 0 n'appartient pas à l'intervalle $(\hat{\theta}_i - Z_{\alpha/2}\sigma_i, \hat{\theta}_i + Z_{\alpha/2}\sigma_i)$.

Exemple numérique : design optimal

Puisque la matrice d'information fournit indirectement l'estimation de la matrice de variance covariance, plusieurs auteurs capitalisent dessus pour concevoir des expériences de préférences déclarées.

En particulier, Bliemer et Rose (2010) proposent d'utiliser la mesure d'erreur D dans la conception d'expérience pour des modèles de logit mélangé.

Etant donné certains a priori, ils cherchent à miniser l'erreur D, définie comme

$$\Delta(\text{Cov}(\hat{\theta}))^{1/K},$$

où $\Delta(A)$ est le déterminant de A, et ils estiment $\text{Cov}(\hat{\theta})$ as $H_N^{-1}(\hat{\theta})$.

Exemple numérique : design optimal (suite)

Une autre mesure populaire le mesure d'erreur A, calculée comme

$$\frac{tr(\text{Cov}(\hat{\theta}))}{K},$$

où tr tient pour trace.

Les erreur D et A diffèrent de manière significative vu que l'erreur A repose seulement sur la diagonale de la matrice, i.e. les variances, mais pas les covariances, tandis que l'erreur D tient compte de toute la matrice de variance-covariance.

Dès lors, un design D-optimal design peut créer des corrélations entre les facteurs étudiés afin de diminuer l'erreur totale.

Illustration

Nous reproduisons le cas d'étude 1 de Bliemer et Rose, qui supposent un modèle mixed logit, avec une formulation panel, développent l'expression du hessien, et cherchent un design minimisant l'erreur D.

L'expérience suppose de plus que chaque répondant face face à deux alternatives, décrites pas quatre attributs, chacun avec trois niveaux de variations, et doit repondre à 9 situations de choix.

Les niveaux des quatre attributs pour chaque alternative dans les neuf situations de choix sont présentés ci-après.

Niveaux

s	j	x_{j1}	x_{j2}	x_{j3}	x_{j4}
1	1	3	3	1	2
	2	1	1	2	2
2	1	2	1	1	1
	2	2	3	3	3
3	1	3	3	3	2
	2	1	1	1	1
4	1	3	2	2	3
	2	1	2	2	2
5	1	1	2	2	3
	2	3	2	3	1
6	1	1	1	3	2
	2	3	3	1	2
7	1	2	2	1	1
	2	2	2	3	3
8	1	2	1	3	1
	2	2	3	1	3
9	1	1	3	2	3
	2	3	1	2	1

Niveaux

Les a priori suivant ont été employés pour dériver le design optimal :

$$\begin{aligned}
\beta_1 &\sim \mathcal{N}(0.6, 0.04), \\
\beta_2 &\sim \mathcal{N}(-0.9, 0.04), \\
\beta_3 &= -0.2, \\
\beta_4 &= 0.8.
\end{aligned}$$

En suivant les recommandations de Bliemer et Rose, Bastin et Cirillo (2011) ont généré une population de 100 individuals pour le modèle mixed logit, et aussi peu que 23 individuals pour la version multinomial logit, obtenue imposant aux β's d'être constants entre individus.

Paramètres estimés

Variable	MNL (23 ind.)	MNL (100 ind.)	MMNL Lattice
β_1	0.642 (4.61)	0.598 (8.59)	0.624 (8.45)
σ_1	-	-	0.217 (2.16)
β_2	-0.970 (5.74)	-0.905 (10.70)	-0.951 (10.64)
σ_2	-	-	0.263 (3.21)
β_3	-0.261 (2.73)	-0.162 (3.62)	-0.172 (3.64)
β_4	0.722 (4.41)	0.841 (10.68)	0.872 (10.85)

Notes : 16381 tirs ont été utilisés pour le modèle mixed logit, en utilisant une lattice perturbée aléatoirement.

Observations

Le modèle logit reproduit assez bien les valeurs moyennes des coefficients normalement distribués, et raisonnablement bien les paramètres constants.

Mais une taille de 23 individus peut être vue comme trop petite pour utiliser adéquatement l'analyse asymptotique (valide pour$N \to \infty$).

Une analyse par bootstrap (sur les individus) permet de vérifier ce point : la matrice de variance-covariance n'est pas stable pour une si petite population. Les résultats avec 100 individus sont bien meilleurs.

Observations (MMNL)

MNL (23 ind.)	β_1	β_2		β_3	β_4
β_1	1.000	-0.667		-0.008	0.548
β_2	-0.667	1.000		0.240	-0.812
β_3	-0.008	0.240		1.000	-0.242
β_4	-0.548	-0.812		-0.242	1.000
MNL (100 ind.)	β_1	β_2		β_3	β_4
β_1	1.000	-0.700		-0.300	0.627
β_2	-0.700	1.000		0.387	-0.777
β_3	-0.300	0.387		1.000	-0.469
β_4	0.627	-0.777		-0.469	1.000

Observations (MNL)

MMNL	β_1	σ_1	β_2	σ_2	β_3	β_4
β_1	1.000	0.196	-0.683	0.106	-0.281	0.658
σ_1	0.196	1.000	-0.250	0.102	-0.097	0.219
β_2	-0.655	-0.250	1.000	-0.152	0.354	-0.790
σ_2	0.106	0.102	-0.152	1.000	-0.068	0.082
β_3	-0.281	-0.097	0.354	-0.068	1.000	-0.377
β_4	0.658	0.219	-0.790	0.082	-0.377	1.000
MMNL - robust	β_1	σ_1	β_2	σ_2	β_3	β_4
β_1	1.000	0.214	-0.684	-0.031	-0.275	0.622
σ_1	0.214	1.000	-0.196	0.180	0.046	0.070
β_2	-0.683	-0.196	1.000	-0.030	0.352	-0.763
σ_2	-0.031	0.180	-0.030	1.000	0.020	-0.015
β_3	-0.275	0.046	0.352	0.020	1.000	-0.439
β_4	0.622	0.070	-0.763	-0.015	-0.439	1.000

Observations (suite)

Présence d'importantes corrélations.

MNL : la norme de Frobenius de la différence entre la matrice d'information et l'opposion du hessien aux paramètres estimés vaut 2.35 et 2.64, avec 23 and 100 individus respectivement.
MMNL : 1.39 (prise en compte partielle de la dynamique de choix).

	MNL (23 ind.)	MNL (100 ind.)	MMNL	MMNL - robust
A-error	0.482	0.506	0.696	0.651
D-error	0.291	0.283	0.462	0.432

Exemple simulé

Panel synthétique de 1000 individus, 10 observations par individus.
5 alternatives, chaque avec trois attributs suivant une $N(0, 1)$.

Un coefficient est supposé constant ($\beta_1 = -0.4$), et deux sont supposés normalement distribués ($\beta_2 = \mathcal{N}(0.2, 1.0)$, $\beta_3 = \mathcal{N}(0.8, 0.25)$).

Estimation

Variable	MNL	MMNL MC	MMNL Lattice
β_1	-0.326 (26.44)	-0.400 (33.30)	-0.400 (33.20)
β_2	0.171 (6.86)	0.255 (4.69)	0.254 (4.64)
σ_2	-	0.978 (16.73)	0.980 (16.41)
β_3	0.810 (24.91)	0.811 (24.58)	0.811 (24.91)
σ_3	-	0.486 (15.49)	0.486 (15.71)

MNL	β_1	β_2	β_3
β_1	1.000	-0.040	-0.083
β_2	-0.040	1.000	0.053
β_3	-0.083	0.053	1.000

$$-H(\hat{\theta}) = \begin{pmatrix} 6.956 & 0.1348 & 0.5778 \\ 0.1348 & 7.1911 & -0.2547 \\ 0.5778 & -0.2547 & 0.6147 \end{pmatrix},$$

$$I(\hat{\theta}) = \begin{pmatrix} 7.2966 & 0.2069 & 0.9563 \\ 0.2069 & 32.0420 & -0.6355 \\ 0.9563 & -0.6355 & 11.6705 \end{pmatrix}.$$