Chapitre 8

La Méthode du Maximum de Vraisemblance

8.1 Introduction

Les techniques d'estimation dont nous avons discuté jusqu'ici – moindres carrés et variables instrumentales – sont applicables uniquement aux modèles de régression. Mais tous les modèles ne peuvent pas s'écrire comme une égalité entre la variable dépendante et une fonction de régression plus un terme d'erreur, ou de telle sorte qu'un ensemble de variables dépendantes, sous la forme d'un vecteur, soit égal à un vecteur de fonctions de régression plus un vecteur d'aléas (Chapitre 9). Dans ces cas, les moindres carrés et les variables instrumentales ne sont tout simplement pas appropriés. Dans ce chapitre, nous introduisons par conséquent une troisième méthode d'estimation, qui est beaucoup plus largement applicable que les techniques dont nous avons discuté jusqu'ici, mais qui nécessite également d'assez fortes hypothèses. Il s'agit de l'estimation par la méthode du **maximum de vraisemblance**, ou **ML**.

A titre d'exemple du manque de pertinence des moindres carrés, considérons le modèle

$$y_t^\gamma = \beta_0 + \beta_1 x_t + u_t, \quad u_t \sim \text{IID}(0, \sigma^2), \tag{8.01}$$

qui ressemble presque à un modèle de régression. Ce modèle a du sens tant que le membre de droite de (8.01) demeure toujours positif, et il peut même être un modèle attrayant dans certains cas.[1] Par exemple, supposons que les observations portant sur y_t soient inclinées à droite mais que celles portant sur x_t ne le soient pas. Alors un modèle de régression conventionnel pourrait réconcilier ces deux faits uniquement si les aléas u_t étaient inclinés à droite, ce que l'on ne voudrait probablement pas supposer et qui rendrait l'utilisation des moindres carrés douteuse. D'un autre côté, le modèle (8.01) avec $\gamma < 1$

[1] A proprement parler, il est impossible, naturellement, de garantir que le membre de droite de (8.01) soit toujours positif, mais ce modèle peut être considéré comme une très bonne approximation si $\beta_0 + \beta_1 x_t$ est toujours plus grand que σ.

pourrait bien être capable de réconcilier ces faits tout en permettant aux aléas d'avoir une distribution symétrique.

Si γ était connu, (8.01) serait un modèle de régression. Mais si γ doit être estimé, (8.01) *n'est pas* un modèle de régression. Par conséquent, il ne peut pas être raisonnablement estimé par moindres carrés. La fonction somme-des-carrés est

$$SSR(\boldsymbol{\beta}, \gamma) = \sum_{t=1}^{n} \left(y_t^{\gamma} - \beta_0 - \beta_1 x_t\right)^2,$$

et si, par exemple, tous les y_t étaient plus grands que l'unité, il est clair que cette fonction pourrait être arbitrairement construite proche de zéro simplement en laissant tendre γ vers moins l'infini et en posant β_0 et β_1 égaux à zéro. Par conséquent, personne ne pourrait *jamais* obtenir des estimations sensées de (8.01) en utilisant les moindres carrés ordinaires. Cependant, ce modèle peut être estimé très facilement en utilisant la méthode du maximum de vraisemblance qui sera expliquée dans la Section 8.10.

L'idée fondamentale de l'estimation par maximum de vraisemblance est, comme le nom l'implique, de trouver un ensemble d'estimations de paramètres, appelé $\hat{\boldsymbol{\theta}}$, telles que la **vraisemblance** d'avoir obtenu l'échantillon que nous utilisons soit maximisée. Nous signifions par là que la densité de probabilité jointe pour le modèle que l'on estime est évaluée aux valeurs observées de la (des) variable(s) dépendante(s) et traitée comme une fonction de paramètres du modèle. Le vecteur $\hat{\boldsymbol{\theta}}$ des estimations ML donne alors le maximum de cette fonction. Ce principe d'estimation est très largement applicable: si nous pouvons écrire la densité jointe de l'échantillon, nous pouvons en principe utiliser le maximum de vraisemblance, soumis bien sûr à certaines conditions de régularité. Par ailleurs, il a un nombre de propriétés extrêmement commodes, dont nous discuterons brièvement dans ce qui suit et plus en détail dans le reste de ce chapitre. Il possède également quelques propriétés peu pratiques, et pour cela, le praticien doit parfois être méfiant.

La manière la plus simple de saisir l'idée fondamentale de l'estimation par ML est de considérer un exemple simple. Supposons que chaque observation y_t soit générée par la densité

$$f(y_t, \theta) = \theta e^{-\theta y_t}, \quad y_t > 0, \;\; \theta > 0, \tag{8.02}$$

et soit indépendante de toutes les autres y_t. Il s'agit de la densité de la **distribution exponentielle**.[2] Il y a un seul paramètre inconnu θ que nous

[2] La distribution exponentielle est utile pour l'analyse des phénomènes tels que les files d'attente ou les durées du chômage. Consulter n'importe quel ouvrage de statistique de niveau avancé, tel que Cox et Hinkley (1974) ou Hogg et Craig (1978). Pour des traitements plus précis, consulter, entre autres, Cox et Oakes (1984), Lawless (1982), et Miller (1981). Voir Kiefer (1988) et Lancaster (1990) pour des applications économiques.

désirons estimer, et nous disposons de n observations avec lesquelles nous allons travailler. La densité jointe des y_t sera désignée sous le nom de **fonction de vraisemblance** et notée $L(\boldsymbol{y}, \theta)$; pour toute valeur de θ, cette fonction nous renseigne sur la probabilité que nous aurions eue d'observer l'échantillon $\boldsymbol{y} \equiv [y_1 \vdots \cdots \vdots y_n]$.

Comme les y_t sont indépendants, leur densité jointe est simplement le produit de leurs densités marginales. Ainsi, la fonction de vraisemblance s'écrit

$$L(\boldsymbol{y}, \theta) = \prod_{t=1}^{n} \theta e^{-\theta y_t}. \tag{8.03}$$

Dans le cas d'échantillons de grande taille, (8.03) peut devenir extrêmement importante ou extrêmement petite, et prendre des valeurs qui sont bien au-delà des possibilités des nombres à virgule flottante que les ordinateurs manipulent. Pour cette raison, parmi d'autres, il est d'usage de maximiser le *logarithme* de la fonction de vraisemblance plutôt que la fonction de vraisemblance elle-même. Bien évidemment, nous obtiendrons la même réponse en procédant ainsi, car la **fonction de logvraisemblance** $\ell(\boldsymbol{y}, \theta) \equiv \log\big(L(\boldsymbol{y}, \theta)\big)$ est une fonction monotone croissante de $L(\boldsymbol{y}, \theta)$; si $\hat{\theta}$ maximise $\ell(\boldsymbol{y}, \theta)$, il doit aussi maximiser $L(\boldsymbol{y}, \theta)$. Dans le cas de (8.03), la fonction de logvraisemblance est

$$\ell(\boldsymbol{y}, \theta) = \sum_{t=1}^{n} \big(\log(\theta) - \theta y_t\big) = n \log(\theta) - \theta \sum_{t=1}^{n} y_t. \tag{8.04}$$

La maximisation de la fonction de logvraisemblance par rapport au seul paramètre inconnu θ, est une procédure directe. Différentier l'expression la plus à droite de (8.04) par rapport à θ et poser la dérivée à zéro donne la condition du premier ordre

$$\frac{n}{\theta} - \sum_{t=1}^{n} y_t = 0, \tag{8.05}$$

et nous trouvons pour la résolution de l'estimateur ML $\hat{\theta}$ que

$$\hat{\theta} = \frac{n}{\sum_{t=1}^{n} y_t}. \tag{8.06}$$

Dans ce cas, il n'est pas nécessaire de se soucier des multiples solutions de (8.05). La dérivée seconde de (8.04) est toujours négative, ce qui nous permet de conclure que $\hat{\theta}$ défini par (8.06) est *l'unique* estimateur ML. Notons que cela ne sera pas toujours le cas; pour certains problèmes les conditions du premier ordre peuvent mener à des solutions multiples.

Dès à présent, nous pourrions à juste titre poser certaines questions relatives aux propriétés de $\hat{\theta}$. Est-ce dans tous les sens du terme un bon estimateur à utiliser? Est-il biaisé? Est-il convergent? Comment est-il distribué? Et

ainsi de suite. Nous pourrions certainement étudier ces questions pour ce cas particulier. Mais une grande part de cette investigation se révèlerait inutile, car le fait que $\hat{\theta}$ soit un estimateur ML nous renseigne immédiatement sur un grand nombre de ses propriétés. C'est, en effet, une des caractéristiques les plus attrayantes de l'estimation ML: parce que beaucoup d'éléments sur les propriétés des estimateurs ML sont généralement connus, nous n'avons pas toujours besoin de pratiquer une étude particulière dans tous les cas.

Deux propriétés attrayantes majeures des estimateurs ML sont la **convergence** et la **normalité asymptotique**. Celles-ci sont des propriétés que nous avons déjà longuement étudiées dans le contexte des moindres carrés, et à ce titre nous n'avons pas besoin de les présenter davantage. Une troisième propriété attrayante est l'**efficacité asymptotique**. Ceci est vrai dans un sens plus fort pour les estimateurs ML que pour ceux des moindres carrés; comme nous n'avions pas formulé de fortes hypothèses sur la distribution des aléas lorsque nous discutions des moindres carrés, nous ne pouvions qu'affirmer que les estimations par moindres carrés non linéaires étaient asymptotiquement efficaces à l'intérieur d'une classe d'estimateurs assez limitée. Comme la méthode du maximum de vraisemblance nous force à expliciter en partie les hypothèses de distribution des aléas, nous serons capables de prouver des résultats plus forts.

Le fait que la matrice de covariance des estimations des paramètres résultant de l'estimation par ML puisse être estimée sans difficulté de différentes façons est étroitement lié à ces propriétés. Plus loin, comme nous le verrons dans la Section 8.9, la procédure ML conduit naturellement à plusieurs statistiques de test asymptotiquement équivalentes, dont au moins une d'entre elles peut être calculée aisément. Les estimations ML en elles-mêmes sont directement calculables, parce que la maximisation, même la maximisation non linéaire, est une procédure très bien comprise et, au moins conceptuellement, facile à effectuer. Ainsi une des qualités les plus appréciables de l'estimateur ML est son **calcul**: les estimations ML, aussi bien que les écarts types estimés et les statistiques de test, peuvent généralement être calculés de manière directe, bien que parfois coûteuse.

Une cinquième propriété souhaitable des estimateurs ML est l'**invariance**, terme par lequel nous signifions l'invariance à la reparamétrisation du modèle. Ceci est facile à illustrer à travers l'exemple que nous considérons jusqu'ici. Supposons que nous ayons paramétrisé la densité de y_t comme

$$f'(y_t, \phi) = (1/\phi)e^{-y_t/\phi}, \tag{8.07}$$

où $\phi \equiv 1/\theta$. Il est facile de décrire la relation entre $\hat{\phi}$ et $\hat{\theta}$. La logvraisemblance dans la paramétrisation en ϕ est

$$\ell'(\boldsymbol{y}, \phi) = \sum_{t=1}^{n} \Big(-\log(\phi) - \frac{y_t}{\phi} \Big) = -n \log(\phi) - \frac{1}{\phi} \sum_{t=1}^{n} y_t.$$

La condition de premier ordre pour un maximum de ℓ' est alors

$$-\frac{n}{\phi}+\frac{1}{\phi^2}\sum_{t=1}^{n}y_t=0,$$

et l'estimation ML décrite par $\hat{\phi}$ est donc

$$\hat{\phi}=\frac{1}{n}\sum_{t=1}^{n}y_t=\frac{1}{\hat{\theta}}.$$

Nous constatons que la relation entre $\hat{\phi}$ et $\hat{\theta}$ est exactement la même que celle établie entre ϕ et θ. Alors, dans ce cas, l'estimation ML est **invariante à la reparamétrisation**. En fait, ceci est une propriété générale du maximum de vraisemblance. Tout spécialement dans les cas où la reparamétrisation est plus ou moins arbitraire, elle peut être une de ses propriétés les plus attrayantes.

Les propriétés du ML ne sont pas toutes enviables. Une caractéristique indésirable majeure concerne la dépendance aux hypothèses explicites de distribution des aléas, que le chercheur ressent souvent comme étant trop forte. Ceci n'est pas toujours un problème aussi sérieux que ce qu'il peut paraître. Bien qu'*en général* les propriétés asymptotiques des estimateurs ML soient valables seulement lorsque le modèle est correctement spécifié à tous les égards, nombreux sont les cas où une ou plusieurs de ces propriétés restent valides malgré quelques spécifications douteuses. Par exemple, l'estimateur des moindres carrés non linéaires correspond à l'estimateur par maximum de vraisemblance lorsque le modèle est un modèle de régression non linéaire à aléas normaux et indépendants (consulter la Section 8.10) et, comme nous l'avons vu, la convergence et la normalité asymptotique des NLS ne nécessitent pas l'hypothèse de normalité des aléas. Ainsi lorsque les aléas ne sont pas normaux, l'estimateur des moindres carrés non linéaires est un exemple de l'**estimateur quasi-ML**, ou **estimateur QML**, c'est-à-dire un estimateur ML appliqué à une situation pour laquelle il n'est pas entièrement valable; voir White (1982) et Gouriéroux, Monfort, Trognon (1984). Les estimateurs QML sont aussi parfois appelés **estimateurs pseudo-ML**.

L'autre caractéristique majeure indésirable du ML est que ses propriétés avec des échantillons finis peuvent être très différentes de ces propriétés asymptotiques. Bien qu'elles soient convergentes, les estimations des paramètres ML sont typiquement biaisées, et les estimations de la matrice de covariance ML peuvent être sérieusement trompeuses. Parce qu'en pratique les propriétés avec des échantillons finis sont souvent inconnues, le chercheur doit décider (souvent sans beaucoup d'information) comment se fier aux propriétés asymptotiques connues. Ceci introduit un facteur d'imprécision dans les efforts fournis pour établir des inférences par ML quand la taille de l'échantillon n'est pas extrêmement importante.

Dans le reste de ce chapitre, nous discuterons des propriétés les plus importantes du maximum de vraisemblance. La relation entre les moindres carrés et le maximum de vraisemblance sera introduite à la Section 8.10 et sera aussi un des thèmes abordés dans le Chapitre 9, qui s'intéresse principalement aux moindres carrés généralisés et à leur relation avec le ML. Des exemples d'estimation par maximum de vraisemblance en économétrie seront fournis dans la suite du livre. Des exemples complémentaires peuvent être trouvés chez Cramer (1986).

8.2 Concepts Fondamentaux et Notation

L'estimation par maximum de vraisemblance repose sur la notion de **vraisemblance** d'un ensemble donné d'observations relatives à un modèle, ou ensemble de DGP. Un DGP, en tant que processus stochastique, peut être caractérisé de plusieurs manières. Nous développons maintenant la notation à partir de laquelle nous pouvons promptement exprimer une telle caractérisation qui est particulièrement utile pour nos objectifs immédiats. Nous supposons que chaque observation pour tout échantillon de taille n est une réalisation d'une variable aléatoire y_t, $t = 1, \ldots, n$, prenant des valeurs dans $\mathbb{R}^m$. Bien que la notation y_t passe sous silence la possibilité que l'observation est en général un vecteur, il est plus commode de laisser la notation vectorielle $\boldsymbol{y}$ (ou $\boldsymbol{y}^n$ si nous désirons faire explicitement référence à la taille de l'échantillon) désigner l'échantillon entier

$$\boldsymbol{y}^n = [y_1 \vdots y_2 \vdots \cdots \vdots y_n].$$

Si chaque observation est un scalaire, $\boldsymbol{y}$ est un vecteur de dimension n, tandis que si chaque observation est un vecteur de dimension m, $\boldsymbol{y}$ est une matrice de dimension $n \times m$. Le vecteur ou la matrice $\boldsymbol{y}$ peut posséder une densité de probabilité, c'est-à-dire la densité jointe de ses éléments compte tenu du DGP. Cette densité, si elle existe, est une application dont l'ensemble d'arrivée est la droite réelle et dont l'ensemble de départ est un ensemble de réalisations possibles de $\boldsymbol{y}$, ensemble que nous noterons $\mathcal{Y}^n$ et qui sera en général un sous-ensemble de $\mathbb{R}^{nm}$ choisi arbitrairement. Il sera nécessaire de porter toute notre attention sur la définition de la densité dans certains cas, mais il suffit pour l'instant de supposer qu'il s'agit de la densité ordinaire par rapport à la mesure de Lebesgue sur $\mathbb{R}^{nm}$.[3] Quand d'autres possibilités existent, il se trouve que le choix parmi celles-ci se révèle non pertinent pour nos propos.

Nous pouvons à présent définir formellement la fonction de vraisemblance associée à un modèle donné pour un échantillon $\boldsymbol{y}$ donné. Cette fonction dépend d'une part des paramètres du modèle et d'autre part, de l'ensemble

[3] De cette manière, nous avons exclu les modèles à variables dépendantes qualitatives et les modèles dans lesquels la distribution de la variable dépendante a des atomes, car dans ces cas une densité par rapport à la mesure de Lebesgue n'existe pas. Voir le Chapitre 15.

d'observations donné par $\boldsymbol{y}$; sa valeur correspond exactement à la densité associée au DGP caractérisé par le vecteur paramétrique $\boldsymbol{\theta} \in \Theta$, évaluée au point d'échantillon $\boldsymbol{y}$. L'ensemble Θ désigne ici l'**espace paramétrique** dans lequel $\boldsymbol{\theta}$ prend ses valeurs; nous supposerons que c'est un sous-ensemble de $\mathbb{R}^k$. Nous désignerons la fonction de vraisemblance par: $L : \mathcal{Y}^n \times \Theta \to \mathbb{R}$ et sa valeur pour $\boldsymbol{\theta}$ et $\boldsymbol{y}$ par $L(\boldsymbol{y}, \boldsymbol{\theta})$. Dans bien des cas pratiques, tel que celui examiné à la section précédente, les observations y_t sont indépendantes et chaque y_t a une densité de probabilité $L_t(y_t, \boldsymbol{\theta})$. La fonction de vraisemblance pour ce cas spécial est alors

$$L(\boldsymbol{y}, \boldsymbol{\theta}) = \prod_{t=1}^{n} L_t(y_t, \boldsymbol{\theta}). \tag{8.08}$$

La fonction de vraisemblance (8.03) de la section précédente est évidemment un cas particulier de ce cas présent. Quand chacune des observations y_t est identiquement distribuée selon une densité $f(y_t, \boldsymbol{\theta})$, comme dans cet exemple, $L_t(y_t, \boldsymbol{\theta})$ est égale à $f(y_t, \boldsymbol{\theta})$ pour tout t.

Même lorsque la fonction de vraisemblance ne peut pas s'écrire sous la forme de (8.08), il est toujours possible (du moins en théorie) de factoriser $L(\boldsymbol{y}, \boldsymbol{\theta})$ en une série de **contributions**, chacune provenant d'une seule observation. Supposons que les observations individuelles y_t, $t = 1, \ldots, n$, soient *ordonnées* d'une certaine manière, comme par exemple suivant un ordre chronologique dans les séries temporelles. Or, cette factorisation peut être accomplie comme suit. Nous commençons par la densité marginale ou non conditionnelle[4] de la première observation y_1, que l'on peut appeler $L_1(y_1)$, en supprimant la dépendance par rapport à $\boldsymbol{\theta}$ pour le moment. Puis, la densité marginale des deux premières observations jointes peut être écrite comme le produit de $L_1(y_1)$ par la densité de y_2 conditionellement à y_1, et nous la notons $L_2(y_2 \,|\, y_1)$. Si maintenant, nous prenons les trois premières observations ensemble, leur densité jointe est le produit de la densité non conditionnelle des deux premières prises simultanément, par la densité de la troisième conditionnellement aux deux premières, et ainsi de suite. Le résultat pour l'échantillon

[4] Nous utilisons le terme "non conditionnel" par commodité. Certains statisticiens considèrent *toutes* les distributions ou *toutes* les densités comme conditionnelles à une chose ou à une autre, et nous ne voulons pas dire que nous excluons ce point de vue. Les distributions, les densités, ou espérances auxquelles nous nous référons comme non conditionnelles devraient être comprises comme étant *seulement* conditionnées aux variables véritablement exogènes, c'est-à-dire, les variables pour lesquelles le DGP est assez indépendant du DGP de $\boldsymbol{y}$. Les Bayesiens peuvent souhaiter considérer les paramètres du DGP comme des variables conditionnantes, et cette conception n'est pas non plus écartée par notre traitement.

entier des observations est

$$\begin{aligned} L(\boldsymbol{y}) &= L_1(y_1)L_2(y_2 \mid y_1)L_3(y_3 \mid y_2, y_1) \cdots L_n(y_n \mid y_{n-1}, \ldots, y_1) \\ &= \prod_{t=1}^{n} L_t(y_t \mid y_{t-1}, \ldots, y_1). \end{aligned} \tag{8.09}$$

Notons que ce résultat est parfaitement général et peut être appliqué à n'importe quelle densité ou fonction de vraisemblance. L'ordre des observations est habituellement l'ordre naturel, comme pour les séries temporelles, mais même si aucun ordre naturel n'existe, (8.09) demeure vraie pour un classement arbitraire.

Comme nous l'indiquions dans la dernière section, on utilise dans la pratique la fonction de logvraisemblance $\ell(\boldsymbol{y}, \boldsymbol{\theta})$ plutôt que la fonction de vraisemblance $L(\boldsymbol{y}, \boldsymbol{\theta})$. La décomposition de $\ell(\boldsymbol{y}, \boldsymbol{\theta})$ en contributions provenant d'observations individuelles résulte de (8.09). Elle peut être écrite comme suit, en supprimant la dépendance par rapport à $\boldsymbol{\theta}$ pour alléger les notations:

$$\ell(\boldsymbol{y}) = \sum_{t=1}^{n} \ell_t(y_t \mid y_{t-1}, \ldots, y_1), \tag{8.10}$$

où $\ell_t(y_t \mid y_{t-1}, \ldots, y_1) \equiv \log L_t(y_t \mid y_{t-1}, \ldots, y_1)$.

Nous sommes à présent en position de donner la définition de l'**estimation par maximum de vraisemblance**. Nous disons que $\hat{\boldsymbol{\theta}} \in \Theta$ est une estimation par maximum de vraisemblance, une **estimation ML**, ou une **MLE**, pour les données $\boldsymbol{y}$ si

$$\ell(\boldsymbol{y}, \hat{\boldsymbol{\theta}}) \geq \ell(\boldsymbol{y}, \boldsymbol{\theta}) \quad \forall \boldsymbol{\theta} \in \Theta. \tag{8.11}$$

Si l'inégalité est stricte, alors $\hat{\boldsymbol{\theta}}$ est l'unique MLE. Une MLE peut ne pas exister en général, à moins que la fonction de logvraisemblance ℓ ne soit continue par rapport aux paramètres $\boldsymbol{\theta}$, et que l'ensemble Θ ne soit *compact* (c'est-à-dire fermé et borné). C'est pourquoi il est d'usage, dans les traitements formels de l'estimation par maximum de vraisemblance, de supposer que Θ est en effet compact. Nous ne désirons pas formuler cette hypothèse, parce qu'elle s'accorde en effet très mal avec la pratique standard, pour laquelle une estimation est valable partout dans $\mathbb{R}^k$. Mais cela signifie que nous devons vivre avec la possible non existence de la MLE.

Il est souvent commode d'utiliser une autre définition de la MLE, qui n'est pas équivalente en général. Si la fonction de vraisemblance atteint un maximum *intérieur* à l'espace paramétrique, alors elle, ou de façon équivalente la fonction de logvraisemblance, doit satisfaire les conditions du premier ordre pour un maximum. Ainsi une MLE peut se *définir* comme une solution aux **équations de vraisemblance**, qui correspondent précisément aux conditions du premier ordre suivantes:

$$\boldsymbol{g}(\boldsymbol{y}, \hat{\boldsymbol{\theta}}) \equiv \mathbf{0}, \tag{8.12}$$

où le **vecteur gradient**, ou **vecteur score**, $\boldsymbol{g} \in \mathbb{R}^k$ est défini par

$$\boldsymbol{g}^\top(\boldsymbol{y}, \boldsymbol{\theta}) \equiv D_\theta \ell(\boldsymbol{y}, \boldsymbol{\theta}) = \sum_{t=1}^{n} D_\theta \ell_t(\boldsymbol{y}, \boldsymbol{\theta}). \tag{8.13}$$

Puisque $D_\theta \ell$ est un vecteur ligne, $\boldsymbol{g}$ est le vecteur *colonne* des dérivées partielles de la fonction de logvraisemblance ℓ par rapport aux paramètres $\boldsymbol{\theta}$. Nous avons écrit $\ell_t(\boldsymbol{y}, \boldsymbol{\theta})$, et non $\ell_t(y_t, \boldsymbol{\theta})$, parce qu'en général ℓ_t peut dépendre de valeurs "passées" de la variable dépendante, $y_{t-1}, y_{t-2}, \ldots$. Elle ne dépend pas des valeurs "futures" bien entendu, mais l'utilisation de la notation vectorielle est encore le moyen le plus simple de nous rappeler de la dépendance par rapport à d'autres éléments que y_t.

Comme il peut arriver que plus d'une valeur de $\boldsymbol{\theta}$ satisfasse les équations de vraisemblance (8.12), la définition nécessite par ailleurs que l'estimation $\hat{\boldsymbol{\theta}}$ soit associée à un *maximum* local de ℓ et que

$$\operatorname*{plim}_{n\to\infty}\big(n^{-1}\ell(\boldsymbol{y}, \hat{\boldsymbol{\theta}})\big) \geq \operatorname*{plim}_{n\to\infty}\big(n^{-1}\ell(\boldsymbol{y}, \boldsymbol{\theta}^*)\big),$$

où $\boldsymbol{\theta}^*$ est n'importe quelle autre solution des équations de vraisemblance. Cette seconde définition de la MLE est souvent associée à Cramér, dans sa célèbre preuve de convergence (Cramér, 1946). Dans la pratique, la nécessité que $\operatorname{plim}\big(n^{-1}\ell(\boldsymbol{y}, \hat{\boldsymbol{\theta}})\big) \geq \operatorname{plim}\big(n^{-1}\ell(\boldsymbol{y}, \boldsymbol{\theta}^*)\big)$ est à l'évidence impossible à vérifier en général. Le problème vient du fait que l'on ne connaît pas le DGP et que par conséquent, le calcul analytique des limites en probabilité est impossible. Si pour un échantillon donné il existe deux racines ou plus aux équations de vraisemblance, celle qui est associée à la valeur la plus haute de $\ell(\boldsymbol{y}, \boldsymbol{\theta})$ pour cet échantillon peut ne pas converger vers celle qui est associée à la valeur la plus haute asymptotiquement. Dans la pratique, s'il existe plus d'une solution pour les équations de vraisemblance, l'on sélectionne celle qui est associée à la valeur la plus haute de la fonction de logvraisemblance. Malgré tout, s'il y a deux ou plusieurs solutions pour lesquelles les valeurs correspondantes de $\ell(\boldsymbol{y}, \boldsymbol{\theta})$ sont très proches, il est fort possible de sélectionner la mauvaise.

Nous insistons sur le fait que ces deux définitions de la MLE ne sont pas équivalentes. En conséquence, il est parfois nécessaire de parler des **MLE du Type 1** quand nous faisons référence à celles obtenues par la maximisation de $\ell(\boldsymbol{y}, \boldsymbol{\theta})$ sur Θ, et des **MLE de Type 2** quand nous faisons référence à celles obtenues comme solutions des équations de vraisemblance. Bien que dans la plupart des cas, en pratique, chacune pourrait être utilisée et que dans certains cas, les deux types de MLE coïncident, il existe des situations où seul un des deux types de MLE est réalisable. En particulier, il existe des modèles où $\ell(\boldsymbol{\theta})$ est non bornée dans certaines directions, et la définition de l'estimateur de Type 1 ne peut donc pas être utilisée, mais néanmoins il existe un $\hat{\boldsymbol{\theta}}$ qui est une racine convergente des équations de vraisemblance; consulter

Kiefer (1978) pour un modèle de ce genre. D'un autre côté, la définition de l'estimateur de Type 2 ne s'applique pas au problème standard de l'estimation d'un ou de deux points terminaux d'une distribution uniforme, parce que les équations de vraisemblance ne sont jamais satisfaites.

Il est utile d'étudier le problème de l'estimation des points terminaux d'une distribution uniforme. Supposons que pour tout t la densité de y_t soit

$$f(y_t) = \begin{cases} 1/\alpha & \text{si } 0 \leq y_t \leq \alpha \\ 0 & \text{sinon.} \end{cases}$$

Ici, on sait qu'une borne de la distribution uniforme est zéro, mais il faut estimer α, l'autre borne. Les fonctions de vraisemblance et de logvraisemblance sont respectivement,

$$L(\boldsymbol{y}, \alpha) = \begin{cases} \alpha^{-n} & \text{si } 0 \leq y_t \leq \alpha \text{ pour tout } y_t \\ 0 & \text{sinon} \end{cases}$$

et

$$\ell(\boldsymbol{y}, \alpha) = \begin{cases} -n \log(\alpha) & \text{si } 0 \leq y_t \leq \alpha \text{ pour tout } y_t \\ -\infty & \text{sinon.} \end{cases} \tag{8.14}$$

L'équation de vraisemblance obtenue en dérivant $\ell(\boldsymbol{y}, \alpha)$ par rapport à α et en annulant la dérivée est

$$-\frac{n}{\alpha} = 0.$$

Comme cette équation n'a pas de solution finie, il n'existe aucune estimation ML de Type 2. Cependant, il est clair que nous pouvons trouver une estimation ML de Type 1. De (8.14), il est évident que pour maximiser $\ell(\boldsymbol{y}, \alpha)$ nous devons rendre $\hat{\alpha}$ aussi petite que possible. Comme $\hat{\alpha}$ ne peut pas être plus petite que la plus grande valeur de y_t observée, l'estimation ML de Type 1 doit simplement être

$$\hat{\alpha} = \max_t(y_t).$$

Par le terme **estimateur du maximum de vraisemblance** nous désignerons la variable aléatoire qui associe à chaque occurrence aléatoire possible $\boldsymbol{y}$ la MLE correspondante.[5] La distinction entre une **estimation** et un **estimateur** a été établie dans la Section 5.2. Nous pouvons rappeler qu'un estimateur, une variable aléatoire, est représenté comme une fonction (implicite ou explicite) des ensembles possibles d'observations, alors qu'une estimation est une valeur que peut prendre cette fonction pour un ensemble d'observations bien spécifié.

[5] Dans les cas de non existence de la MLE dans certains échantillons, l'estimateur peut être défini comme une variable aléatoire appropriée en lui assignant une valeur arbitrairement, telle que $-\infty$, pour ces échantillons où la MLE n'existe pas.

Tout comme il existe deux définitions possibles des estimations ML, il existe également deux définitions possibles d'un estimateur ML. Les définitions suivantes montrent clairement que l'estimateur est une variable aléatoire, qui dépend des valeurs observées de l'échantillon $\boldsymbol{y}$. L'**estimateur de Type 1**, correspondant à la définition standard (8.11) de la MLE, est $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ défini par

$$L\big(\boldsymbol{y}, \hat{\boldsymbol{\theta}}(\boldsymbol{y})\big) > L(\boldsymbol{y}, \boldsymbol{\theta}) \quad \forall \boldsymbol{\theta} \in \Theta \text{ tel que } \boldsymbol{\theta} \neq \hat{\boldsymbol{\theta}}(\boldsymbol{y}). \tag{8.15}$$

L'**estimateur de Type 2**, correspondant à la définition (8.12) de Cramér, est $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ défini par:

$$\boldsymbol{g}\big(\boldsymbol{y}, \hat{\boldsymbol{\theta}}(\boldsymbol{y})\big) = \boldsymbol{0}, \tag{8.16}$$

où $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ donne un maximum local de ℓ, et

$$\operatorname*{plim}_{n\to\infty}\Big(n^{-1}\ell\big(\boldsymbol{y}, \hat{\boldsymbol{\theta}}(\boldsymbol{y})\big)\Big) \geq \operatorname*{plim}_{n\to\infty}\Big(n^{-1}\ell(\boldsymbol{y}, \boldsymbol{\theta}^*(\boldsymbol{y}))\Big) \tag{8.17}$$

pour n'importe quelle autre solution $\boldsymbol{\theta}^*(\boldsymbol{y})$ des équations de vraisemblance.

Nous concluons cette section par une variété de définitions qui seront utilisées dans le reste du chapitre et plus généralement dans le reste du livre. En utilisant la décomposition (8.10) de la fonction de logvraisemblance $\ell(\boldsymbol{y}, \boldsymbol{\theta})$, nous pouvons définir une matrice $\boldsymbol{G}(\boldsymbol{y}, \boldsymbol{\theta})$ de dimension $n \times k$ dont l'élément type est

$$G_{ti}(\boldsymbol{y}, \boldsymbol{\theta}) \equiv \frac{\partial \ell_t(\boldsymbol{y}, \boldsymbol{\theta})}{\partial \theta_i}. \tag{8.18}$$

Nous appellerons $\boldsymbol{G}(\boldsymbol{y}, \boldsymbol{\theta})$ la **matrice des contributions au gradient**, ou **matrice CG** pour faire court. Cette matrice est intimement reliée au vecteur gradient $\boldsymbol{g}$, qui est juste $\boldsymbol{G}^\top \boldsymbol{\iota}$, où comme d'habitude $\boldsymbol{\iota}$ désigne un vecteur de taille n pour lequel chaque élément est égal à 1. La $t^{\text{ième}}$ ligne de $\boldsymbol{G}$, qui mesure la contribution au gradient de la $t^{\text{ième}}$ observation, sera noté $\boldsymbol{G}_t$.

La **matrice Hessienne** associée à la fonction de logvraisemblance $\ell(\boldsymbol{y}, \boldsymbol{\theta})$ est la matrice $\boldsymbol{H}(\boldsymbol{y}, \boldsymbol{\theta})$ de dimension $k \times k$ dont l'élément type est

$$H_{ij}(\boldsymbol{y}, \boldsymbol{\theta}) \equiv \frac{\partial^2 \ell(\boldsymbol{y}, \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}. \tag{8.19}$$

Nous définissons l'**espérance de la Hessienne moyenne** pour un échantillon de taille n comme

$$\mathcal{H}^n(\boldsymbol{\theta}) \equiv E_\theta\big(n^{-1}\boldsymbol{H}(\boldsymbol{y}, \boldsymbol{\theta})\big).$$

La notation E_θ signifie que l'espérance est calculée en utilisant le DGP caractérisé par le vecteur paramétrique $\boldsymbol{\theta}$ plutôt que par le DGP qui pourrait réellement avoir généré un quelconque échantillon particulier donné. Ainsi, un DGP différent est implicitement utilisé pour calculer l'espérance pour chaque

$\boldsymbol{\theta}$. La **limite de la Hessienne** ou **Hessienne asymptotique**, si elle existe, est définie comme

$$\mathcal{H}(\boldsymbol{\theta}) \equiv \lim_{n\to\infty} \mathcal{H}^n(\boldsymbol{\theta}).$$

Cette quantité, qui est une matrice symétrique, et en général semi-définie négative, apparaîtra un grand nombre de fois dans la théorie asymptotique de l'estimation ML.

Nous définissons l'**information contenue dans l'observation $\boldsymbol{t}$** par $\boldsymbol{I}_t(\boldsymbol{\theta})$, la matrice de dimension $k \times k$ dont l'élément type est

$$\big(\boldsymbol{I}_t(\boldsymbol{\theta})\big)_{ij} \equiv E_\theta\big(G_{ti}(\boldsymbol{\theta})G_{tj}(\boldsymbol{\theta})\big). \tag{8.20}$$

Le fait que $\boldsymbol{I}_t(\boldsymbol{\theta})$ soit une matrice symétrique, en général semi-définie positive, et qu'elle soit définie positive à condition qu'il existe une relation linéaire entre les composantes du vecteur aléatoire $\boldsymbol{G}_t$ est une conséquence immédiate de cette définition. La **matrice d'information moyenne** pour un échantillon de taille n est définie par

$$\mathcal{I}^n(\boldsymbol{\theta}) \equiv \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{I}_t(\boldsymbol{\theta}) = n^{-1}\boldsymbol{I}^n, \tag{8.21}$$

et la **matrice d'information à la limite** ou **matrice d'information asymptotique**, si elle existe, est définie par

$$\mathcal{I}(\boldsymbol{\theta}) \equiv \lim_{n\to\infty} \mathcal{I}^n(\boldsymbol{\theta}). \tag{8.22}$$

La matrice $\boldsymbol{I}_t(\boldsymbol{\theta})$ mesure la quantité *espérée* d'information contenue dans la $t^{\text{ième}}$ observation et $\boldsymbol{I}^n \equiv n\mathcal{I}^n$ mesure la quantité espérée d'information contenue dans l'échantillon entier. Les matrices d'information $\mathcal{I}^n$ et $\mathcal{I}$ sont, comme $\boldsymbol{I}_t$, symétriques, et en général semi-définies positives. La matrice d'information moyenne $\mathcal{I}^n$ et l'espérance de la Hessienne moyenne $\mathcal{H}^n$ ont été définies telles qu'elles soient $O(1)$ quand $n \to \infty$. Elles sont donc très pratiques à utiliser lors de l'analyse asymptotique. La terminologie dans ce domaine n'est pas entièrement unifiée. Certains auteurs utilisent simplement le terme "matrice d'information" pour se référer à $\mathcal{I}^n$, tandis que d'autres l'utilisent pour se référer à n fois $\mathcal{I}^n$, ce que nous avons appelé $\boldsymbol{I}^n$.

8.3 Transformations et Reparamétrisations

Dans cette section et dans les suivantes, nous développons la théorie classique de l'estimation par maximum de vraisemblance et, en particulier, nous démontrons les propriétés qui font que cette théorie produit une méthode d'estimation qui possède de nombreux avantages. Nous démontrerons aussi que dans certaines circonstances ces propriétés font défaut. Comme nous en

avons discuté dans la Section 8.1, les principales caractéristiques enviables des estimateurs ML sont l'**invariance**, la **convergence**, la **normalité asymptotique**, l'**efficacité asymptotique**, et la **calculabilité**. Dans cette section, nous discuterons de la première de celles-ci, l'invariance des estimateurs ML à la reparamétrisation du modèle.

L'idée d'invariance est un concept important dans l'analyse économétrique. Notons $\mathbb{M}$ le modèle qui nous intéresse. Une **paramétrisation** du modèle $\mathbb{M}$ est une application, disons λ, dont l'espace de départ est un espace paramétrique Θ et qui va vers $\mathbb{M}$. Il existera en général une infinité de paramétrisations pour tout modèle $\mathbb{M}$ donné. Après tout, peu de contraintes portent sur l'espace paramétrique Θ, en dehors de sa dimension. Il est possible de construire une application bijective et dérivable partant d'un sous-ensemble de $\mathbb{R}^k$ vers pratiquement n'importe quel autre sous-ensemble de $\mathbb{R}^k$ par des procédés tels que la translation, la rotation, la dilatation, et bien d'autres encore, et n'importe lequel de ces autres sous-ensembles peut donc faire office d'espace paramétrique pour le modèle $\mathbb{M}$. C'est justement à cause de ces possibilités, que l'on désire que les estimateurs possèdent la propriété d'invariance. Le terme d'"invariance" est compris dans ce contexte comme l'invariance au type de transformation dont nous avons discuté, et que nous appelons formellement **reparamétrisation**.

Pour illustrer le fait que n'importe quel modèle peut être paramétrisé un nombre infini de fois, considérons le cas d'une distribution exponentielle, dont nous avons discuté dans la Section 8.1. Nous avons vu que la fonction de vraisemblance pour un échantillon de réalisations indépendantes obéissant à cette distribution était (8.03). Si nous posons $\theta \equiv \delta^\alpha$, nous pouvons définir une famille entière de paramétrisations indexées par α. Nous pouvons choisir α comme étant n'importe quel nombre fini non nul. La fonction de vraisemblance correspondant à cette famille de paramétrisations est

$$L(\boldsymbol{y}, \delta) = \prod_{t=1}^{n} \delta^\alpha e^{-\delta^\alpha y_t}.$$

Evidemment, le cas $\alpha = 1$ correspond à la paramétrisation en θ de (8.02) et le cas $\alpha = -1$ correspond à la paramétrisation en ϕ de (8.07).

Il est facile de voir que les estimateurs ML sont invariants aux reparamétrisations du modèle. Définissons par $\boldsymbol{\eta} : \Theta \to \Phi \subseteq \mathbb{R}^k$ une application régulière qui transforme le vecteur $\boldsymbol{\theta}$ en un unique vecteur $\boldsymbol{\phi} \equiv \boldsymbol{\eta}(\boldsymbol{\theta})$. La fonction de vraisemblance pour le modèle $\mathbb{M}$ en termes des nouveaux paramètres $\boldsymbol{\phi}$, disons L', est définie par la relation

$$L'(\boldsymbol{y}, \boldsymbol{\phi}) = L(\boldsymbol{y}, \boldsymbol{\theta}) \quad \text{où } \boldsymbol{\phi} = \boldsymbol{\eta}(\boldsymbol{\theta}). \tag{8.23}$$

L'équation (8.23) suit immédiatement des faits que la fonction de vraisemblance est la densité d'un processus stochastique et que $\boldsymbol{\theta}$ et $\boldsymbol{\phi} = \boldsymbol{\eta}(\boldsymbol{\theta})$

décrivent le même processus stochastique. Définissons $\hat{\boldsymbol{\phi}}$ comme $\boldsymbol{\eta}(\hat{\boldsymbol{\theta}})$ et $\boldsymbol{\phi}^*$ comme $\boldsymbol{\eta}(\boldsymbol{\theta}^*)$. Alors si

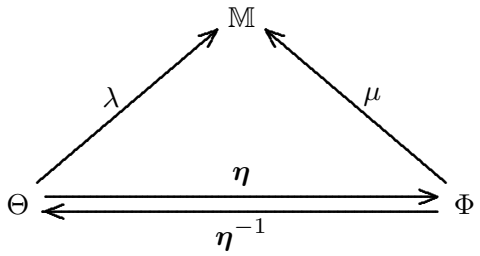
$$L(\boldsymbol{y}, \hat{\boldsymbol{\theta}}) > L(\boldsymbol{y}, \boldsymbol{\theta}^*) \ \forall \boldsymbol{\theta}^* \neq \hat{\boldsymbol{\theta}},$$

il s'ensuit que

$$L'(\boldsymbol{y}, \hat{\boldsymbol{\phi}}) = L'\big(\boldsymbol{y}, \boldsymbol{\eta}(\hat{\boldsymbol{\theta}})\big) = L(\boldsymbol{y}, \hat{\boldsymbol{\theta}}) > L(\boldsymbol{y}, \boldsymbol{\theta}^*) = L'(\boldsymbol{y}, \boldsymbol{\phi}^*) \ \forall \boldsymbol{\phi}^* \neq \hat{\boldsymbol{\phi}}.$$

Ainsi nous obtiendrons les estimations ML $\hat{\boldsymbol{\theta}}$ si nous maximisons $L(\boldsymbol{\theta})$ et les estimations ML $\hat{\boldsymbol{\phi}}$ si nous maximisons $L'(\boldsymbol{\phi})$. Mais ces deux séries d'estimations sont équivalentes, parce qu'elles caractérisent le même DGP, car $L(\hat{\boldsymbol{\theta}}) = L'(\hat{\boldsymbol{\phi}})$.

Une fois que l'on a choisi une paramétrisation d'un modèle, disons $\lambda : \Theta \rightarrow \mathbb{M}$, et que l'on dispose d'une application bijective régulière $\boldsymbol{\eta} : \Theta \rightarrow \Phi$ qui transforme le premier vecteur de paramètres $\boldsymbol{\theta}$ en un second $\boldsymbol{\phi}$, il est possible de reparamétriser le modèle en construisant une application du second espace paramétrique Φ vers le premier Θ à l'aide de $\boldsymbol{\eta}^{-1}$ (qui existe nécessairement puisque $\boldsymbol{\eta}$ est bijective) et de revenir à $\mathbb{M}$ à l'aide de λ. Ainsi, formellement, la nouvelle paramétrisation est une application $\mu \equiv \lambda \circ \boldsymbol{\eta}^{-1}$, qui va de Φ vers $\mathbb{M}$, bijective et régulière. Il peut être utile pour l'intuition de garder à l'esprit le diagramme suivant de commutation:



L'invariance est en général une propriété enviable, car elle assure que (peut-être arbitrairement) les changements dans la manière dont nous retranscrivons le modèle n'auront aucun effet sur les estimations que nous obtiendrons. Mais cette propriété implique néanmoins que les estimateurs ML des *paramètres* ne peuvent pas, en général, être sans biais. Supposons qu'il existe une paramétrisation dans laquelle l'estimateur ML de $\boldsymbol{\theta}$ soit sans biais. Nous pouvons écrire cette propriété comme

$$E_0(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}_0,$$

où E_0 indique que nous calculons les espérances par rapport au DGP caractérisé par le vecteur paramétrique $\boldsymbol{\theta}_0$. Alors, si la fonction $\boldsymbol{\eta}(\boldsymbol{\theta})$ qui offre une nouvelle paramétrisation est non linéaire, comme cela sera le cas en général, cela doit être le cas que

$$E_0(\hat{\boldsymbol{\phi}}) = E_0\big(\boldsymbol{\eta}(\hat{\boldsymbol{\theta}})\big) \neq \boldsymbol{\phi}_0$$

parce que, pour une fonction non linéaire $\boldsymbol{\eta}(\boldsymbol{\theta})$,

$$E_0\big(\boldsymbol{\eta}(\hat{\boldsymbol{\theta}})\big) \neq \boldsymbol{\eta}\big(E_0(\hat{\boldsymbol{\theta}})\big) = \boldsymbol{\eta}(\boldsymbol{\theta}_0) = \boldsymbol{\phi}_0.$$

Ceci suggère que, bien que la paramétrisation que nous choisissons n'ait pas d'importance pour l'estimation du DGP, elle peut avoir un effet substantiel sur les propriétés de nos estimations paramétriques avec des échantillons finis. En choisissant la paramétrisation appropriée, nous pouvons dans certains cas assurer que nos estimations sont sans biais, ou proches d'être sans biais, et que leurs distributions sont proches de leurs distributions asymptotiques. Par contraste, si nous choisissons une paramétrisation inappropriée, nous pourrions par inadvertance rendre nos estimations sévèrement biaisées et dont les distributions sont éloignées de leurs distributions asymptotiques.

8.4 La Convergence

Une des raisons pour lesquelles l'estimation par maximum de vraisemblance est largement utilisée est que les estimateurs ML sont, sous des conditions assez générales, convergents. Dans cette section, nous expliquons pourquoi c'est le cas. Nous nous intéressons premièrement à l'estimateur ML de Type 1, bien que nous proposions aussi certaines discussions au sujet de l'estimateur de Type 2. Nous commençons en posant la définition:

$$\bar{\ell}(\boldsymbol{\theta};\boldsymbol{\theta}_0) \equiv \operatorname*{plim}_{n\to\infty 0}\big(n^{-1}\ell^n(\boldsymbol{y}^n,\boldsymbol{\theta})\big), \tag{8.24}$$

où la notation "plim_0" signifie comme d'habitude que la limite en probabilité est calculée sous le DGP caractérisé par $\boldsymbol{\theta}_0$. La fonction $\bar{\ell}(\boldsymbol{\theta};\boldsymbol{\theta}_0)$ est la valeur limite de n^{-1} fois la fonction de logvraisemblance, quand les données sont générées par un cas particulier du modèle avec $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Une condition de régularité importante qui doit être satisfaite afin qu'un estimateur ML soit convergent est que le modèle soit asymptotiquement identifié. Par définition, ceci sera le cas si le problème

$$\max_{\boldsymbol{\theta}\in\Theta} \bar{\ell}(\boldsymbol{\theta};\boldsymbol{\theta}_0) \tag{8.25}$$

ne comporte qu'une unique solution. Cette définition implique que n'importe quel DGP appartenant au modèle génèrera des échantillons qui, s'ils sont suffisamment grands, identifieront le modèle. L'interprétation est la même que dans le contexte du modèle de régression.

Nous désirons maintenant démontrer que $\bar{\ell}(\boldsymbol{\theta};\boldsymbol{\theta}_0)$ est maximale en $\boldsymbol{\theta}_0$, la valeur de $\boldsymbol{\theta}$ qui caractérise le DGP. Nous désignons par $\hat{\boldsymbol{\theta}} \equiv \hat{\boldsymbol{\theta}}(\boldsymbol{y})$ le maximum global de la fonction de vraisemblance $L(\boldsymbol{y},\boldsymbol{\theta})$, et réclamons que cette fonction soit *continue* en $\boldsymbol{\theta}$, et nous désignons par $\boldsymbol{\theta}^*$ n'importe quel autre vecteur de paramètres (non stochastique) dans Θ, et réclamons que cet espace soit

compact. Ces deux exigences signifient qu'il n'y a aucun problème sur la possible non existence de la MLE. Nous désignerons les espérances calculées par rapport au DGP par $E_0(\cdot)$. Alors, grâce à l'inégalité de Jensen (consulter l'Annexe B), on montre que

$$E_0\left(\log\left(\frac{L(\boldsymbol{\theta}^*)}{L(\boldsymbol{\theta}_0)}\right)\right) \leq \log\left(E_0\left(\frac{L(\boldsymbol{\theta}^*)}{L(\boldsymbol{\theta}_0)}\right)\right), \tag{8.26}$$

car le logarithme est une fonction concave. Plus loin, (8.26) deviendra une inégalité stricte à chaque fois que $L(\boldsymbol{\theta}^*)/L(\boldsymbol{\theta}_0)$ sera une variable aléatoire non dégénérée. Une dégénérescence se produira seulement s'il existe $\boldsymbol{\theta}' \neq \boldsymbol{\theta}_0$ tel que $L(\boldsymbol{\theta}')/L(\boldsymbol{\theta}_0)$ soit identiquement unitaire; $\ell(\boldsymbol{\theta}') - \ell(\boldsymbol{\theta}_0)$ serait alors identiquement égale à zéro. Mais la condition d'identification asymptotique (8.25) élimine cette possibilité pour des tailles d'échantillon assez grandes, puisque, si elle est vérifiée, $\boldsymbol{\theta}' \neq \boldsymbol{\theta}_0$ implique que $L(\boldsymbol{\theta}') \neq L(\boldsymbol{\theta}_0)$.

En utilisant le fait que $L(\boldsymbol{\theta}_0)$ est la densité jointe de $\boldsymbol{y}$, nous voyons que l'espérance à l'intérieur du logarithme dans le membre de droite de (8.26) est

$$E_0\left(\frac{L(\boldsymbol{\theta}^*)}{L(\boldsymbol{\theta}_0)}\right) = \int_{y^n} \frac{L(\boldsymbol{\theta}^*)}{L(\boldsymbol{\theta}_0)} L(\boldsymbol{\theta}_0)\, d\boldsymbol{y} = \int_{y^n} L(\boldsymbol{\theta}^*)\, d\boldsymbol{y} = 1.$$

Nous gérerons la nullité éventuelle de $L(\boldsymbol{\theta}_0)$ en définissant la seconde intégrale ci-dessus comme nulle lorsque $L(\boldsymbol{\theta}_0)$ l'est aussi. Comme le logarithme de 1 est 0, il suit de (8.26) que

$$E_0\left(\log\left(\frac{L(\boldsymbol{\theta}^*)}{L(\boldsymbol{\theta}_0)}\right)\right) < 0,$$

qui peut être récrit comme

$$E_0\big(\ell(\boldsymbol{\theta}^*)\big) - E_0\big(\ell(\boldsymbol{\theta}_0)\big) < 0. \tag{8.27}$$

Ainsi, l'espérance de la fonction de logvraisemblance lorsqu'elle est évaluée avec le véritable vecteur paramétrique, $\boldsymbol{\theta}_0$, est strictement supérieure à l'espérance évaluée avec n'importe quel autre vecteur de paramètres, $\boldsymbol{\theta}^*$.

La prochaine étape consiste à montrer que ce qui est vrai pour les espérances mathématiques dans (8.27), l'est aussi, à la limite lorsque $n \to \infty$, pour l'analogue correspondant à l'échantillon. Cette expression analogue correspondant à l'échantillon est

$$\frac{1}{n}\big(\ell(\boldsymbol{\theta}^*) - \ell(\boldsymbol{\theta}_0)\big) = \frac{1}{n}\sum_{t=1}^{n} \ell_t(\boldsymbol{y}, \boldsymbol{\theta}^*) - \frac{1}{n}\sum_{t=1}^{n} \ell_t(\boldsymbol{y}, \boldsymbol{\theta}_0). \tag{8.28}$$

Maintenant, il est nécessaire de supposer que les sommes dans (8.28) satisfont certaines conditions de régularité suffisantes pour qu'une loi des grands

nombres leur soit appliquée. Comme nous l'avons vu dans le Chapitre 4, celles-ci nécessitent que les ℓ_t soient indépendantes ou du moins, qu'elles ne manifestent pas trop fortement une dépendance; qu'elles possèdent une sorte d'espérance (bien qu'elles puissent ne pas posséder une espérance habituelle); et qu'elles possèdent des variances bornées supérieurement; pour tous les détails, consulter la Section 4.7. Nous pouvons donc réclamer, parce que cela est pratique, que pour tout $\boldsymbol{\theta} \in \Theta$, $\{\ell_t(\boldsymbol{\theta})\}_{t=1}^{\infty}$ satisfait la condition WULLN de la Section 4.7 pour le DGP caractérisé par $\boldsymbol{\theta}_0$. Nous pouvons alors utiliser (8.27) pour affirmer que

$$\operatorname*{plim}_{n\to\infty}{}_0\big(n^{-1}\ell(\boldsymbol{\theta}^*)\big) - \operatorname*{plim}_{n\to\infty}{}_0\big(n^{-1}\ell(\boldsymbol{\theta}_0)\big) < 0, \tag{8.29}$$

où les deux limites en probabilité existent. En fait, grâce à la définition (8.24),

$$\operatorname*{plim}_{n\to\infty}{}_0\big(n^{-1}\ell(\boldsymbol{\theta}^*)\big) = \bar{\ell}(\boldsymbol{\theta}^*; \boldsymbol{\theta}_0),$$

ce qui démontre l'existence de la fonction $\bar{\ell}(\boldsymbol{\theta}^*; \boldsymbol{\theta}_0)$. Il reste à démontrer que l'inégalité dans (8.29) est stricte, car la *limite* des inégalités strictes (8.27) n'est pas nécessairement une inégalité stricte. Cependant, la condition d'identification asymptotique (8.25) peut encore être invoquée pour rétablir l'inégalité stricte.

Avec l'hypothèse d'identification asymptotique donnée et le résultat (8.29), il est maintenant facile de voir pourquoi $\hat{\boldsymbol{\theta}}$ doit être convergente. Nous savons que

$$n^{-1}\ell(\hat{\boldsymbol{\theta}}) \geq n^{-1}\ell(\boldsymbol{\theta}_0), \tag{8.30}$$

pour tout n, parce que $\hat{\boldsymbol{\theta}}$ maximise la fonction de logvraisemblance. Clairement (8.29) et (8.30) ne peuvent pas toutes deux être vraies à moins que

$$\operatorname*{plim}_{n\to\infty}{}_0\big(n^{-1}\ell(\hat{\boldsymbol{\theta}})\big) = \operatorname*{plim}_{n\to\infty}{}_0\big(n^{-1}\ell(\boldsymbol{\theta}_0)\big). \tag{8.31}$$

Mais si le modèle est asymptotiquement identifié, la valeur $\hat{\boldsymbol{\theta}}$ qui maximise (8.24) doit être unique. Alors, (8.31) ne peut pas être vérifiée à moins que $\operatorname{plim}_0(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}_0$.[6]

Nous pouvons maintenant énoncer le théorème suivant, que l'on doit à Wald (1949):

Théorème 8.1. Théorème de Convergence de Wald.

L'estimateur ML (8.15) pour un modèle représenté par la famille paramétrique des fonctions de logvraisemblance $\ell(\boldsymbol{\theta})$ dans lesquelles $\boldsymbol{\theta}$ est contraint à résider dans un espace paramétrique compact, est convergent si les contributions $\{\ell_t(\boldsymbol{\theta})\}_{t=1}^{\infty}$ satisfont les conditions de

[6] Parce que $\hat{\boldsymbol{\theta}}$ est stochastique, cet argument n'est pas rigoureux.

régularité WULLN et si, en plus, le modèle est asymptotiquement identifié.

Notons que le résultat a été démontré uniquement pour des espaces paramétriques *compacts*, car autrement nous ne pourrions pas être sûr que $\hat{\boldsymbol{\theta}}$ existe pour tout n. Il existe des modèles, par exemple certains appelés modèles de régime endogène, dans lesquels le fait qu'une variance ne puisse tendre vers zéro pour une densité de probabilité qui a de bonnes propriétés, conduit à une défaillance de la compacité de l'espace paramétrique (puisqu'en excluant une variance nulle, on crée une borne ouverte partiellement dans cet espace). Par exemple, il peut ne pas exister de MLE de Type 1 avec une limite en probabilité; consulter Kiefer (1978).

Il existe deux ensembles majeurs de circonstances dans lesquelles les estimations ML peuvent ne pas être convergentes. Le premier survient quand le nombre de paramètres n'est pas fixe mais augmente avec n. Cette possibilité n'est même pas considérée dans le théorème précédent, où $\boldsymbol{\theta}$ est indépendant de n. Mais il n'est pas surprenant que cela engendre des problèmes, car si le nombre de paramètres n'est pas fixe, il est loin d'être évident que la quantité d'information que l'échantillon nous donne à propos de chacun d'eux augmentera suffisamment rapidement lorsque $n \to \infty$. Il est en fait possible de laisser le nombre de paramètres augmenter, mais le taux d'accroissement doit être modéré (par exemple, comme $n^{1/4}$). De tels problèmes sont bien au-delà des objectifs de cet ouvrage; consulter, entre d'autres, Neyman et Scott (1948), Kiefer et Wolfowitz (1956), et Kalbfleisch et Sprott (1970).

Les cas d'absence de convergence les plus fréquemment rencontrés sont ceux dans lesquels le modèle n'est pas identifié asymptotiquement. Ceci peut arriver même quand il *est* identifié par n'importe quel échantillon fini. Par exemple, considérons le modèle de régression

$$y_t = \alpha \frac{1}{t} + u_t, \quad u_t \sim \text{NID}(0, 1),$$

considéré à l'origine dans la Section 5.2. Nous avons déjà vu que des modèles de ce type ne peuvent pas être estimés de manière convergente par les moindres carrés, et c'est un exercice simple de monter que de tels modèles ne peuvent pas non plus être estimés de manière convergente par le maximum de vraisemblance. Une manière de concevoir ce type de problème est d'observer que, lorsque n augmente, chaque observation nouvelle porte de moins en moins d'information au sujet de α. Ainsi, bien que la matrice d'information d'échantillon fini $\boldsymbol{I}^n$ soit toujours de plein rang (de un dans ce cas), la matrice d'information asymptotique $\mathcal{I}$ ne l'est pas (elle converge vers zéro dans ce cas). Dans ce cas habituel où l'estimateur ML est convergent, chaque nouvelle observation additionne approximativement la même quantité d'information et $\mathcal{I}$, étant la limite de la moyenne des $\boldsymbol{I}_t$, sera alors de plein rang.

Dans la plupart des situations, la seule chose que nous aurons besoin de connaître sera la convergence de l'estimateur ML de Type 1. Cependant, on

trouve des cas dans lesquels seul l'estimateur de Type 2 existe. Dans le reste de cette section, nous esquissons alors la preuve de la convergence de l'estimateur ML de Type 2, tel qu'il est défini par (8.16) et (8.17). Pour que cet estimateur existe, il est bien sûr nécessaire que les contributions ℓ_t pour la fonction de logvraisemblance $\ell(\boldsymbol{y}, \boldsymbol{\theta})$ soient dérivables par rapport aux paramètres $\boldsymbol{\theta}$, et aussi supposerons-nous qu'elles sont continûment différentiables au moins une fois. Grâce à cette hypothèse, l'argument qui suit n'est plus utile dans de nombreux ensembles de circonstances: si l'espace paramétrique Θ est compact et le vecteur paramétrique $\boldsymbol{\theta}_0$ associé au DGP est à l'intérieur de Θ, alors pour des échantillons assez importants, la probabilité que la maximum de ℓ soit réalisé en un point intérieur de Θ devient arbitrairement proche de l'unité. Quand cela arrive, les estimateurs de Type 1 et de Type 2 coïncideront asymptotiquement. D'un autre côté, si $\boldsymbol{\theta}_0$ est sur la frontière de Θ, il y aura une probabilité positive, pour des échantillons arbitrairement grands, que l'estimateur de Type 2 n'existe pas. Dans un tel cas, la question de sa convergence éventuelle ne se pose pas.

La situation est plus délicate dans le cas d'un espace paramétrique non compact. Nous remarquons tout d'abord que si $\boldsymbol{\theta}_0$ se situe sur la frontière de Θ, il y aura une probabilité positive pour que l'estimateur de Type 2 n'existe pas, mais ce n'est pas la compacité qui est en cause. Nous supposons donc que $\boldsymbol{\theta}_0$ est à l'intérieur de Θ. Nous supposerons ensuite que la condition de la définition suivante est satisfaite:

Définition 8.1.

Le modèle caractérisé par la fonction de logvraisemblance ℓ est **identifiée asymptotiquement sur un espace paramétrique Θ non compact** si le modèle est asymptotiquement identifié et si, de plus, il n'existe aucune séquence $\{\boldsymbol{\theta}^n\}$ ne comportant aucun point limite qui satisfasse

$$\bar{\ell}(\boldsymbol{\theta}^n; \boldsymbol{\theta}_0) \longrightarrow \bar{\ell}(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0); \quad \bar{\ell}(\boldsymbol{\theta}^n; \boldsymbol{\theta}_0) < \bar{\ell}(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0). \tag{8.32}$$

L'identification asymptotique semble écarter l'existence de telles séquences, mais il n'en est rien. Pour que la séquence n'ait aucun point limite, elle doit diverger à l'infini dans certaines directions, ou autrement, converger vers un point qui n'appartient pas à l'espace paramétrique non compact, tel qu'un point de variance nulle. Ainsi, le fait que $\bar{\ell}(\boldsymbol{\theta}^n; \boldsymbol{\theta}_0)$ tende vers la limite $\bar{\ell}(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0)$ n'implique pas l'existence d'un point dans Θ, disons $\boldsymbol{\theta}^\infty$, pour lequel $\bar{\ell}(\boldsymbol{\theta}^\infty; \boldsymbol{\theta}_0) = \bar{\ell}(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0)$. En effet, l'existence de $\boldsymbol{\theta}^\infty$ devrait contredire l'identification asymptotique dans son sens habituel. Mais pour que l'on puisse interpréter l'identification asymptotique dans son sens habituel dans un espace paramétrique non compact, l'existence de suites satisfaisant (8.32) doit être éliminée, même si elles n'ont pas de point limite.

Retournons maintenant au cas des estimateurs de Type 2. Considérons un voisinage *compact* Θ_0 de $\boldsymbol{\theta}_0$. Nous pourrions définir un autre estimateur ML comme le point qui donne le maximum de ℓ dans Θ_0. Grâce au

Théorème de convergence de Wald (Théorème 8.1) ce nouvel estimateur serait convergent. Deux cas possibles semblent alors exister. Le premier est celui pour lequel il existe une probabilité positive asymptotiquement que cet estimateur soit sur la *frontière* du voisinage Θ_0 et le second est celui pour lequel cette probabilité est nulle. Dans le second cas, le nouvel estimateur et l'estimateur de Type 2 coïncident asymptotiquement, compte tenu de la condition d'identification asymptotique pour un ensemble non compact Θ, et ce dernier est donc convergent. Mais en fait le premier cas ne peut pas survenir. Pour un Θ_0 fixé, $\boldsymbol{\theta}_0$ est à une distance positive de la frontière de Θ_0, et la convergence du nouvel estimateur exclut toute probabilité positive asymptotiquement concentrée sur une région fermée éloignée de $\boldsymbol{\theta}_0$. Ainsi nous concluons que lorsque l'espace paramétrique est non compact, à condition que le DGP reste à l'*intérieur* de cet espace et que le modèle soit asymptotiquement identifié sur son espace paramétrique non compact, l'estimateur de Type 2 est convergent. Ces résultats sont résumés dans le théorème suivant:

Théorème 8.2. Second Théorème de Convergence.

Soit un modèle représenté par une famille paramétrique de fonctions de logvraisemblance $\ell(\boldsymbol{\theta})$ au moins une fois continûment différentiables dans laquelle $\boldsymbol{\theta}$ est contraint d'appartenir à un espace paramétrique non nécessairement compact. Alors, pour les DGP qui se situent à l'intérieur de cet espace paramétrique, l'estimateur ML défini par (8.16) et (8.17) est convergent si les contributions $\{\ell_t(\boldsymbol{\theta})\}_{t=1}^{\infty}$ satisfont les conditions de régularité WULLN et si de plus l'espace paramétrique est compact et le modèle est asymtotiquement identifié, ou si l'espace paramétrique est non compact et le modèle est asymptotiquement identifié au sens de la Définition 8.1.

8.5 La Distribution Asymptotique de l'Estimateur ML

Nous commençons notre analyse en démontrant un résultat simple mais fondamental concernant le gradient $\boldsymbol{g}$ et la matrice $\boldsymbol{G}$ de CG:

$$E_\theta\big(G_{ti}(\boldsymbol{\theta})\big) \equiv E_\theta\left(\frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \theta_i}\right) = 0. \tag{8.33}$$

Ce résultat indique que, sous le DGP caractérisé par $\boldsymbol{\theta}$, l'espérance de chaque élément de la matrice CG, évaluée en $\boldsymbol{\theta}$, est zéro. Ceci implique que

$$E_\theta\big(\boldsymbol{g}(\boldsymbol{\theta})\big) = \mathbf{0} \quad \text{et} \quad E_\theta\big(\boldsymbol{G}(\boldsymbol{\theta})\big) = \mathbf{0}.$$

C'est un résultat très important pour plusieurs raisons. En particulier, il nous permettra d'appliquer un théorème de la limite centrale à la quantité

$n^{-1/2}\boldsymbol{g}(\boldsymbol{\theta}_0)$. La démonstration est comme suit:

$$\begin{aligned} E_\theta\big(G_{ti}(y_t,\boldsymbol{\theta})\big) &= \int \frac{\partial \log L_t(y_t,\boldsymbol{\theta})}{\partial\theta_i} L_t(y_t,\boldsymbol{\theta})\,dy_t \\ &= \int \frac{1}{L_t(y_t,\boldsymbol{\theta})} \frac{\partial L_t(y_t,\boldsymbol{\theta})}{\partial\theta_i} L_t(y_t,\boldsymbol{\theta})\,dy_t \\ &= \int \frac{\partial L_t(y_t,\boldsymbol{\theta})}{\partial\theta_i}\,dy_t \qquad (8.34) \\ &= \frac{\partial}{\partial\theta_i}\int L_t(y_t,\boldsymbol{\theta})\,dy_t \\ &= \frac{\partial}{\partial\theta_i}(1) = 0. \end{aligned}$$

L'avant dernière étape est simplement une conséquence de la normalisation de la densité $L_t(y_t,\boldsymbol{\theta})$. L'étape précédente, dans laquelle les ordres de différentiation et d'intégration sont interchangés, est valide sous une variété de conditions de régularité, parmi lesquelles la plus simple est que le domaine d'intégration, disons $\mathcal{Y}_t$, soit indépendant de $\boldsymbol{\theta}$. De façon alternative, si cette hypothèse n'est pas vraie, alors il suffit que $L_t(y_t,\boldsymbol{\theta})$ s'annule sur la frontière du domaine $\mathcal{Y}_t$ et que $\partial\ell_t(y_t,\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ soit uniformément bornée; consulter l'Annexe B.

Les résultats simples concernant la distribution asymptotique des estimations ML sont obtenus le plus facilement dans le contexte de l'estimateur de Type 2, défini par (8.16) et (8.17). Par conséquent, nous limiterons notre attention à ce cas et nous supposerons que $\hat{\boldsymbol{\theta}}$ est une racine des équations de vraisemblance (8.12). Il est alors relativement simple de montrer que $\hat{\boldsymbol{\theta}}$ possède la propriété de **normalité asymptotique**, dont nous avons discuté dans le Chapitre 5. Pour un DGP caractérisé par $\boldsymbol{\theta}_0$, le vecteur des estimations paramétriques $\hat{\boldsymbol{\theta}}$ tend vers la limite non stochastique $\boldsymbol{\theta}_0$. Cependant, si nous multiplions la différence $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ par $n^{1/2}$, la quantité résultante $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ aura une limite en probabilité qui est une variable aléatoire avec une distribution normale multivariée. Comme dans le cas des NLS, nous pouvons occasionnellement y faire référence de façon peu formelle comme à la distribution asymptotique de $\hat{\boldsymbol{\theta}}$, bien que cela ne soit pas correct techniquement.

Maintenant, nous esquissons une démonstration de normalité asymptotique de la MLE de Type 2. Nous commençons par le développement de Taylor des équations de vraisemblance (8.12) autour de $\boldsymbol{\theta}_0$, pour obtenir

$$\mathbf{0} = \boldsymbol{g}(\hat{\boldsymbol{\theta}}) = \boldsymbol{g}(\boldsymbol{\theta}_0) + \boldsymbol{H}(\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \qquad (8.35)$$

où $\bar{\boldsymbol{\theta}}$ est une combinaison convexe de $\boldsymbol{\theta}_0$ et $\hat{\boldsymbol{\theta}}$, qui peut être différente pour chaque ligne de l'équation vectorielle. Si nous résolvons (8.35) par rapport à

$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ et si nous récrivons tous les facteurs de manière à les rendre $O(1)$, nous obtenons

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -\big(n^{-1}\boldsymbol{H}(\bar{\boldsymbol{\theta}})\big)^{-1}\big(n^{-1/2}\boldsymbol{g}(\boldsymbol{\theta}_0)\big), \tag{8.36}$$

dans laquelle nous voyons que $n^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$ égale une matrice de dimension $k\times k$ fois un vecteur de dimension k. La matrice s'avèrera être asymptotiquement non aléatoire, et le vecteur s'avèrera être asymptotiquement normal, ce qui implique que $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ doit être asymptotiquement normal.

Nous voulons en premier lieu montrer que $n^{-1}\boldsymbol{H}(\bar{\boldsymbol{\theta}})$ tend vers une certaine matrice limite non stochastique quand $n \to \infty$. Souvenons-nous que le $ij^{\text{ième}}$ élément de $n^{-1}\boldsymbol{H}(\bar{\boldsymbol{\theta}})$ est

$$\frac{1}{n}\sum_{t=1}^{n} \frac{\partial^2 \ell_t(\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}, \tag{8.37}$$

évalué en $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}$. Nous ferons en sorte que la condition WULLN s'applique à la série dont l'élément type est (8.37). Pour que cela soit réalisable, $n^{-1}\boldsymbol{H}(\bar{\boldsymbol{\theta}})$ doit tendre vers $\mathcal{H}(\bar{\boldsymbol{\theta}})$ quand $n \to \infty$. Mais comme $\hat{\boldsymbol{\theta}}$ est convergent pour $\boldsymbol{\theta}_0$ et que $\bar{\boldsymbol{\theta}}$ reste entre $\hat{\boldsymbol{\theta}}$ et $\boldsymbol{\theta}_0$, il est clair que $n^{-1}\boldsymbol{H}(\bar{\boldsymbol{\theta}})$ doit également tendre vers $\mathcal{H}(\boldsymbol{\theta}_0)$. De plus, si le modèle est fortement asymptotiquement identifié, la matrice $\mathcal{H}(\boldsymbol{\theta}_0)$ doit être définie négative, et nous supposerons que c'est effectivement le cas.

En utilisant cet argument et (8.36), nous voyons que

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{=} -\mathcal{H}^{-1}(\boldsymbol{\theta}_0)\big(n^{-1/2}\boldsymbol{g}(\boldsymbol{\theta}_0)\big). \tag{8.38}$$

Le seul élément stochastique dans le membre de droite de (8.38) est

$$n^{-1/2}\boldsymbol{g}(\boldsymbol{\theta}_0), \tag{8.39}$$

dont un élément type est

$$n^{-1/2}\sum_{t=1}^{n} \left.\frac{\partial \log L_t(y_t, \boldsymbol{\theta})}{\partial\theta_i}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = n^{-1/2}\sum_{t=1}^{n} G_{ti}(\boldsymbol{\theta}_0).$$

Ainsi (8.39) est $n^{-1/2}$ fois une somme de n quantités. D'après le résultat (8.33), nous savons que chacune de ces quantités a une espérance égale à zéro. Il semble alors plausible qu'un théorème de la limite centrale s'y applique. Dans une démonstration formelle, on devrait commencer par les conditions de régularité appropriées et les utiliser pour démontrer qu'un CLT particulier s'applique en effet à (8.39), mais nous omettrons cette étape. Si nous supposons que (8.39) est asymptotiquement normal, il suit immédiatement de (8.38) que $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ doit l'être également.

La **matrice de covariance asymptotique** de $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ est simplement l'espérance asymptotique de $n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top$. En utilisant (8.38), cette quantité est égale à

$$\big(-\mathcal{H}^{-1}(\boldsymbol{\theta}_0)\big)\Big(\frac{1}{n}E_0\big(\boldsymbol{g}(\boldsymbol{\theta}_0)\boldsymbol{g}^\top(\boldsymbol{\theta}_0)\big)\Big)\big(-\mathcal{H}^{-1}(\boldsymbol{\theta}_0)\big).$$

Un élément type de l'espérance dans le facteur central est

$$\tfrac{1}{n}E_0\Bigg(\Big(\sum_{t=1}^{n} G_{ti}(\boldsymbol{\theta}_0)\Big)\Big(\sum_{s=1}^{n} G_{sj}(\boldsymbol{\theta}_0)\Big)\Bigg). \tag{8.40}$$

Ceci est n^{-1} fois l'espérance du produit de deux sommes. Si nous devions développer explicitement le produit, nous verrions que chacun des termes dans la sommation des n^2 termes dans (8.40) serait de la forme

$$G_{ti}(\boldsymbol{\theta}_0)G_{sj}(\boldsymbol{\theta}_0) = \frac{\partial \log(L_t)}{\partial \theta_i}\frac{\partial \log(L_s)}{\partial \theta_j}.$$

Tous ces termes doivent avoir une espérance égale à zéro, sauf quand $t = s$. Supposons sans perte de généralité que $t > s$. Alors

$$\begin{aligned} E_0\big(G_{ti}(\boldsymbol{\theta}_0)G_{sj}(\boldsymbol{\theta}_0)\big) &= E_0\Big(E\big(G_{ti}(\boldsymbol{\theta}_0)G_{sj}(\boldsymbol{\theta}_0)\,|\,\boldsymbol{y}^s\big)\Big) \\ &= E_0\Big(G_{sj}(\boldsymbol{\theta}_0)E\big(G_{ti}(\boldsymbol{\theta}_0)\,|\,\boldsymbol{y}^s\big)\Big) = 0. \end{aligned}$$

La dernière égalité provient du fait que $E_0\big(G_{ti}(\boldsymbol{\theta}_0)\,|\,\boldsymbol{y}^s\big) = 0$, qui est elle-même vraie parce que la preuve du résultat général (8.33) s'applique aussi bien à l'espérance conditionnelle qu'à l'espérance non conditionnelle.

Comme $E_0\big(G_{ti}(\boldsymbol{\theta}_0)G_{sj}(\boldsymbol{\theta}_0)\big) = 0$ pour tout $t \neq s$,

$$\tfrac{1}{n}E_0\Bigg(\Big(\sum_{t=1}^{n} G_{ti}(\boldsymbol{\theta}_0)\Big)\Big(\sum_{s=1}^{n} G_{sj}(\boldsymbol{\theta}_0)\Big)\Bigg) = \tfrac{1}{n}E_0\Big(\sum_{t=1}^{n} G_{ti}(\boldsymbol{\theta}_0)G_{tj}(\boldsymbol{\theta}_0)\Big). \tag{8.41}$$

De (8.20) et (8.21) nous voyons que le membre de droite de (8.41) correspond simplement à $\mathcal{I}^n(\boldsymbol{\theta}_0)$, la matrice d'information moyenne pour un échantillon de taille n. En utilisant le fait que $\mathcal{I}(\boldsymbol{\theta}_0)$ est la limite de $\mathcal{I}^n(\boldsymbol{\theta}_0)$ quand $n \to \infty$, nous concluons que la matrice de covariance asymptotique de $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ est

$$\boldsymbol{V}^{\infty}\big(n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\big) = \mathcal{H}^{-1}(\boldsymbol{\theta}_0)\mathcal{I}(\boldsymbol{\theta}_0)\mathcal{H}^{-1}(\boldsymbol{\theta}_0). \tag{8.42}$$

Dans la prochaine section, nous verrons que cette expression peut être grandement simplifiée.

Nous pouvons à présent établir les résultats précédents comme suit:

Théorème 8.3. Théorème de Normalité Asymptotique.

L'estimateur ML de Type 2, $\hat{\boldsymbol{\theta}}$, pour un modèle fortement identifié asymptotiquement représenté par la famille paramétrique des fonctions de logvraisemblance $\ell(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, quand il existe et est convergent pour le vecteur paramétrique $\boldsymbol{\theta}_0$ qui caractérise le DGP, est asymptotiquement normal si

(i) les contributions $\ell_t(\boldsymbol{y}, \boldsymbol{\theta})$ à ℓ sont au moins deux fois continûment différentiables en $\boldsymbol{\theta}$ pour presque tout $\boldsymbol{y}$ et tout $\boldsymbol{\theta} \in \Theta$,

(ii) les séries composantes de $\{D^2_{\theta\theta}\,\ell_t(\boldsymbol{y}, \boldsymbol{\theta})\}_{t=1}^{\infty}$ satisfont la condition WULLN sur Θ, et

(iii) les séries composantes de $\{D_\theta\,\ell_t(\boldsymbol{y}, \boldsymbol{\theta})\}_{t=1}^{\infty}$ satisfont la condition CLT.

Par le terme de normalité asymptotique, nous signifions que la série de variables aléatoires $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ a une limite en probabilité qui est une variable aléatoire de l'ordre de l'unité, normalement distribuée d'espérance nulle et de matrice de covariance (8.42).

8.6 L'Egalité de la Matrice d'Information

Dans cette section, nous établirons un résultat important qui permet une simplification substantielle de l'expression (8.42) de la matrice de covariance asymptotique de l'estimateur ML. Ce résultat, qui, comme l'annonce le titre de la section, est connu sous le nom de l'**égalité de la matrice d'information**, est

$$\mathcal{H}(\boldsymbol{\theta}_0) = -\mathcal{I}(\boldsymbol{\theta}_0). \tag{8.43}$$

Littéralement, la matrice d'information Hessienne asymptotique est l'opposé de la matrice d'information asymptotique. Un résultat analogue est vrai pour des observations individuelles:

$$E_0\big(D^2_{\theta\theta}\,\ell_t(\boldsymbol{y}, \boldsymbol{\theta}_0)\big) = -E_0\big(D_\theta^\top \ell_t(\boldsymbol{y}, \boldsymbol{\theta}_0) D_\theta\,\ell_t(\boldsymbol{y}, \boldsymbol{\theta}_0)\big). \tag{8.44}$$

Le dernier résultat implique clairement le premier, étant données les hypothèses qui permettent l'application d'une loi des grands nombres aux séries $\{D^2_{\theta\theta}\,\ell_t(\boldsymbol{y}, \boldsymbol{\theta}_0)\}_{t=1}^{\infty}$ et $\{D_\theta^\top \ell_t(\boldsymbol{y}, \boldsymbol{\theta}_0) D_\theta\,\ell_t(\boldsymbol{y}, \boldsymbol{\theta}_0)\}_{t=1}^{\infty}$.

Le résultat (8.44) est démontré à l'aide d'un argument très similaire à celui utilisé au début de la dernière section pour monter que l'espérance de la matrice CG est égale zéro. Du fait que

$$\frac{\partial \ell_t}{\partial \theta_i} = \frac{1}{L_t}\frac{\partial L_t}{\partial \theta_i},$$

nous obtenons après une différentiation supplémentaire:

$$\frac{\partial^2 \ell_t}{\partial \theta_i \partial \theta_j} = \frac{1}{L_t}\frac{\partial^2 L_t}{\partial \theta_i \partial \theta_j} - \frac{1}{L_t^2}\frac{\partial L_t}{\partial \theta_i}\frac{\partial L_t}{\partial \theta_j}.$$

En conséquence,

$$\frac{\partial^2 \ell_t}{\partial \theta_i \partial \theta_j} + \frac{\partial \ell_t}{\partial \theta_i}\frac{\partial \ell_t}{\partial \theta_j} = \frac{1}{L_t}\frac{\partial^2 L_t}{\partial \theta_i \partial \theta_j}. \tag{8.45}$$

Maintenant, si nous calculons l'espérance de (8.45) pour le DGP caractérisé par la même valeur du vecteur paramétrique $\boldsymbol{\theta}$ que celle avec laquelle les fonctions ℓ_t et L_t sont évaluées (que nous désignerons comme d'habitude par E_θ), nous trouvons que

$$\begin{aligned} E_\theta\left(\frac{\partial^2 \ell_t}{\partial\theta_i\partial\theta_j} + \frac{\partial\ell_t}{\partial\theta_i}\frac{\partial\ell_t}{\partial\theta_j}\right) &= \int L_t \frac{1}{L_t}\frac{\partial^2 L_t}{\partial\theta_i\partial\theta_j}\, dy_t \\ &= \frac{\partial^2}{\partial\theta_i\partial\theta_j}\int L_t\, dy_t = 0, \end{aligned} \tag{8.46}$$

à condition que, comme pour (8.34), la permutation de l'ordre de différentiation et d'intégration puisse être justifiée. Alors, le résultat (8.46) établit (8.44), puisqu'il implique que

$$E_\theta\left(\frac{\partial^2 \ell_t}{\partial\theta_i\partial\theta_j}\right) = 0 - E_\theta\left(\frac{\partial\ell_t}{\partial\theta_i}\frac{\partial\ell_t}{\partial\theta_j}\right) = -E_\theta\left(\frac{\partial\ell_t}{\partial\theta_i}\frac{\partial\ell_t}{\partial\theta_j}\right).$$

Afin d'établir (8.43), rappelons que, à partir de (8.19) et de la loi des grands nombres,

$$\begin{aligned} \mathcal{H}(\boldsymbol{\theta}) &= \lim_{n\to\infty}\left(\frac{1}{n}\sum_{t=1}^{n} E_\theta\left(\frac{\partial^2\ell_t(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right)\right) \\ &= -\lim_{n\to\infty}\left(\frac{1}{n}\sum_{t=1}^{n} E_\theta\left(\frac{\partial\ell_t(\boldsymbol{\theta})}{\partial\theta_i}\frac{\partial\ell_t(\boldsymbol{\theta})}{\partial\theta_j}\right)\right) \\ &= -\mathcal{I}(\boldsymbol{\theta}), \end{aligned}$$

où la dernière ligne provient directement de la définition de la matrice d'information asymptotique, (8.22). Alors ceci donne (8.43).

En substituant soit $-\mathcal{H}(\boldsymbol{\theta}_0)$ à $\mathcal{I}(\boldsymbol{\theta}_0)$ soit $\mathcal{I}(\boldsymbol{\theta}_0)$ à $-\mathcal{H}(\boldsymbol{\theta}_0)$ dans (8.42), il est facile de conclure que la matrice de covariance asymptotique de l'estimateur ML est donnée par l'une ou l'autre des deux expressions équivalentes $-\mathcal{H}(\boldsymbol{\theta}_0)^{-1}$ et $\mathcal{I}(\boldsymbol{\theta}_0)^{-1}$. Formellement, nous pouvons écrire

$$\boldsymbol{V}^\infty\big(n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\big) = \mathcal{I}^{-1}(\boldsymbol{\theta}_0) = -\mathcal{H}^{-1}(\boldsymbol{\theta}_0).$$

Afin d'effectuer une quelconque inférence statistique, il est nécessaire de pouvoir *estimer* $\mathcal{I}^{-1}(\boldsymbol{\theta}_0)$ ou $-\mathcal{H}^{-1}(\boldsymbol{\theta}_0)$. L'estimateur qui vient immédiatement à l'esprit est $\mathcal{I}^{-1}(\hat{\boldsymbol{\theta}})$, c'est-à-dire l'inverse de la matrice d'information asymptotique évaluée avec la MLE, $\hat{\boldsymbol{\theta}}$. Notons que la fonction matricielle $\mathcal{I}(\boldsymbol{\theta})$ *n'est pas* un objet dépendant de l'échantillon. Elle peut, en principe, être calculée théoriquement comme une fonction matricielle des paramètres du modèle à partir (de la série) des fonctions de logvraisemblance ℓ^n. Pour certains modèles, c'est un calcul entièrement réalisable, et cela donne alors ce qui est souvent l'estimateur préféré de la matrice de covariance asymptotique.

Mais pour certains modèles, le calcul, même s'il était réalisable, serait excessivement laborieux, et dans ces cas, il est commode de disposer d'autres estimateurs convergents de $\mathcal{I}(\boldsymbol{\theta}_0)$ et en conséquence de la matrice de covariance asymptotique.

Un estimateur commun est l'opposé de ce que l'on nomme **matrice Hessienne empirique**. Cette matrice est définie comme

$$\hat{\mathcal{H}} \equiv \mathcal{H}^n(\boldsymbol{y}, \hat{\boldsymbol{\theta}}) = \frac{1}{n} \sum_{t=1}^{n} D^2_{\theta\theta} \ell_t(\boldsymbol{y}, \hat{\boldsymbol{\theta}}); \tag{8.47}$$

elle correspond simplement à $\mathcal{H}^n(\boldsymbol{y}, \boldsymbol{\theta})$ évaluée en $\hat{\boldsymbol{\theta}}$. La loi des grands nombres et la convergence de $\hat{\boldsymbol{\theta}}$ elle-même garantissent immédiatement la convergence de (8.47) pour $\mathcal{H}(\boldsymbol{\theta}_0)$. Quand la matrice Hessienne empirique est directement disponible, comme cela sera le cas si les programmes de maximisation qui utilisent les dérivées secondes sont employés, l'opposé de son inverse peut fournir une manière très commode d'estimer la matrice de covariance de $\hat{\boldsymbol{\theta}}$. Cependant, la matrice Hessienne est souvent difficile à calculer, et si elle n'est pas déjà calculée pour d'autres fins, il est probablement insensé de la calculer uniquement pour estimer une matrice de covariance.

Un autre estimateur de la matrice de covariance communément utilisé est connu sous le nom d'**estimateur produit-extérieur-du-gradient**, ou **estimateur OPG**. Il est basé sur la définition

$$\mathcal{I}(\boldsymbol{\theta}) \equiv \lim_{n\to\infty} \left(\frac{1}{n} \sum_{t=1}^{n} E_\theta \big(D_\theta^\top \ell_t(\boldsymbol{\theta}) D_\theta \ell_t(\boldsymbol{\theta}) \big) \right).$$

L'estimateur OPG est

$$\hat{\mathcal{I}}_{\text{OPG}} \equiv \frac{1}{n} \sum_{t=1}^{n} D_\theta^\top \ell_t(\boldsymbol{y}, \hat{\boldsymbol{\theta}}) D_\theta \ell_t(\boldsymbol{y}, \hat{\boldsymbol{\theta}}) = \frac{1}{n} \boldsymbol{G}^\top(\hat{\boldsymbol{\theta}}) \boldsymbol{G}(\hat{\boldsymbol{\theta}}), \tag{8.48}$$

et sa convergence est garantie une fois de plus par la condition CLT, qui inclut une loi des grands nombres pour la somme dans (8.48).

L'estimateur OPG de la matrice d'information a été préconisé par Berndt, Hall, Hall, et Hausman (1974) dans un article célèbre et on s'y réfère parfois sous le nom de l'estimateur BHHH. Ils ont aussi suggéré son utilisation comme partie d'un système général pour la maximisation de fonctions de logvraisemblance, analogue aux systèmes basés sur la régression de Gauss-Newton dont nous avons discuté dans la Section 6.8. Malheureusement, l'estimateur (8.48) passe pour être plutôt bruité dans la pratique, ce qui limite son utilité à l'estimation des matrices de covariance et à la maximisation numérique.[7]

[7] Il y aura quelques discussions supplémentaires dans le Chapitre 13 sur les manières alternatives d'estimer la matrice de covariance. Pour une discussion de la performance de l'estimateur OPG dans le système d'estimation BHHH, consulter Belsley (1980).

Alors que dans $\mathcal{I}(\hat{\boldsymbol{\theta}})$ le seul élément stochastique est la MLE $\hat{\boldsymbol{\theta}}$ elle-même, à la fois la matrice Hessienne empirique et l'estimateur OPG dépendent explicitement de l'échantillon réalisé $\boldsymbol{y}$, et cette dépendance leur transmet un bruit additionnel qui rend les inférences basées sur ces estimateurs moins fiables que l'on ne le souhaiterait. Souvent l'estimateur OPG semble être particulièrement pauvre, comme nous en discuterons dans le Chapitre 13.

Dans certains cas, il est possible de trouver des estimateurs quelque part entre l'estimateur (habituellement) préféré $\mathcal{I}(\hat{\boldsymbol{\theta}})$ et l'estimateur OPG, pour lequel on peut calculer les espérances de certains des termes apparaissant dans (8.48) mais pas de tous. Ceci semble être une bonne procédure à suivre pour la qualité de l'inférence statistique que l'on peut obtenir à partir des distributions asymptotiques des estimateurs ou des statistiques de test. L'estimateur Gauss-Newton de la matrice de covariance est de ce type chaque fois que le modèle contient des variables dépendantes retardées, car la matrice $n^{-1}\boldsymbol{X}^{\top}(\hat{\boldsymbol{\beta}})\boldsymbol{X}(\hat{\boldsymbol{\beta}})$ dépendra alors de valeurs retardées de $\boldsymbol{y}$ aussi bien que de $\hat{\boldsymbol{\beta}}$. Beaucoup plus d'exemples de ce type d'estimateur apparaîtront plus tard dans ce livre, plus particulièrement dans les Chapitres 14 et 15.

La discussion précédente n'a peut-être pas rendu clair un point qui est de la plus haute importance pratique quand on essaie de pratiquer des inférences concernant un ensemble d'estimations ML $\hat{\boldsymbol{\theta}}$. Tout ce qui se rattache à la théorie de la distribution asymptotique se note en terme de $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$, mais en pratique nous voulons en fait utiliser $\hat{\boldsymbol{\theta}}$ pour réaliser des inférences à propos de $\boldsymbol{\theta}$. Ceci signifie que nous devons baser nos inférences *non pas* sur des quantités qui estiment $\mathcal{I}(\boldsymbol{\theta}_0)$ mais plutôt sur des quantités qui estiment $n\mathcal{I}(\boldsymbol{\theta}_0)$. Alors les trois estimateurs qui peuvent être utilisés en pratique pour estimer $\boldsymbol{V}(\hat{\boldsymbol{\theta}})$ sont l'inverse de l'opposé de la matrice Hessienne numérique,

$$\big(-\boldsymbol{H}(\hat{\boldsymbol{\theta}})\big)^{-1}, \tag{8.49}$$

l'inverse de l'estimateur OPG de la matrice d'information $O(n)$,

$$\big(\boldsymbol{G}^{\top}(\hat{\boldsymbol{\theta}})\boldsymbol{G}(\hat{\boldsymbol{\theta}})\big)^{-1}, \tag{8.50}$$

et l'inverse de la matrice d'information $O(n)$ elle-même,

$$\big(n\mathcal{I}(\hat{\boldsymbol{\theta}})\big)^{-1} \equiv \big(\boldsymbol{I}^n(\hat{\boldsymbol{\theta}})\big)^{-1}. \tag{8.51}$$

En plus de (8.49), (8.50) et (8.51), qui sont très largement applicables, il y a des estimateurs hybrides variés pour certaines classes de modèles, tels que les estimateurs basés sur les régressions de Gauss-Newton et sur d'autres régressions artificielles. Notons que tous ces estimateurs de matrice de covariance seront n fois plus petits que les estimateurs de la matrice de covariance $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$, tels que (8.47) et (8.48), dont nous avons discuté jusqu'ici.

Bien qu'il soit commun de calculer autant d'espérances que possible quand on estime la matrice de covariance de $\hat{\boldsymbol{\theta}}$, il n'est pas évident que cela

soit toujours une bonne chose. Considérons l'exemple suivant. Supposons que $y_t = \beta x_t + u_t$, où x_t est une variable binaire dont nous savons qu'elle prend la valeur 1 avec une probabilité p et la valeur 0 avec la probabilité $1 - p$. Supposons de plus (pour simplifier) que la variance de u_t soit connue et égale à l'unité. Alors la matrice d'information, qui est simplement un scalaire dans ce cas, est $E\left(n^{-1}\sum_{t=1}^{n} x_t^2\right) = p$. Ainsi l'estimation usuelle de la variance de $\hat{\beta}$ basée sur la matrice d'information est simplement $(np)^{-1}$.

Il devrait être évident que, quand np est petit, $(np)^{-1}$ pourrait être une estimation très trompeuse de la variance réelle de $\hat{\beta}$ conditionnelle à l'échantillon particulier qui a été observé. Supposons, par exemple, que n soit 100 et p soit .02. L'estimation habituelle de la variance serait $\frac{1}{2}$. Mais il pourrait survenir qu'aucun des x_t de l'échantillon ne soit égal à 1; ceci arriverait avec une probabilité .133. Alors cet échantillon particulier n'identifierait pas du tout β, et la variance de $\hat{\beta}$ serait infinie. De façon contraire, il peut survenir qu'un seul des x_t dans l'échantillon soit égal à 1. Alors β serait identifié, mais $\frac{1}{2}$ serait à l'évidence une sous-estimation de la variance réelle de $\hat{\beta}$. D'un autre côté, si plus de deux des x_t étaient égaux à 1, $\hat{\beta}$ aurait une variance plus petite que $(np)^{-1}$. L'estimation de la variance asymptotique ne correspondrait à la véritable variance de $\hat{\beta}$ conditionnelle à l'échantillon observé que dans le cas où np était égal à sa valeur espérée, 2.

Cet exemple est très spécial, mais le phénomène qu'il illustre est assez général. A chaque fois que nous calculons la matrice de covariance d'un certain vecteur d'estimations paramétriques, nous nous soucions vraisemblablement de la précision de cet ensemble particulier d'estimations. Cela dépend de la quantité d'information qui a été fournie par l'échantillon dont nous disposons plutôt que de la quantité d'information qui serait fournie par un échantillon type de la même taille. Désormais, dans un sens très concret, c'est la matrice d'information *observée* plutôt que la matrice d'information *attendue* qui devrait nous intéresser. Pour une discussion beaucoup plus étendue sur ce point, consulter Efron et Hinkley (1978).

8.7 La Fonction de Logvraisemblance Concentrée

Il arrive souvent que les paramètres dont dépend une fonction de logvraisemblance puissent être partitionnés en deux ensembles de façon à rendre facile l'écriture de l'estimateur ML d'un groupe de paramètres comme une fonction des valeurs de l'autre groupe. Nous rencontrerons un exemple de ceci, en connexion avec l'estimation ML des modèles de régression, dans la Section 8.10, et d'autres exemples dans le Chapitre 9. Dans cette situation, il peut être très pratique de **concentrer** la fonction de logvraisemblance en l'écrivant comme une fonction d'un seul des deux groupes de paramètres. Supposons que nous puissions écrire la fonction de logvraisemblance $\ell(\boldsymbol{y}, \boldsymbol{\theta})$ comme $\ell(\boldsymbol{y}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. Les conditions du premier ordre qui définissent les estimateurs ML (de Type 2)

$\hat{\boldsymbol{\theta}}_1$ et $\hat{\boldsymbol{\theta}}_2$ sont

$$D_1\ell(\boldsymbol{y},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)=\mathbf{0} \quad \text{et} \quad D_2\ell(\boldsymbol{y},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)=\mathbf{0},$$

où, comme d'habitude, $D_i\ell$ désigne le vecteur ligne des dérivées partielles $\partial\ell/\partial\boldsymbol{\theta}_i$ pour $i=1,2$. Supposons qu'il soit possible de résoudre le second ensemble de conditions du premier ordre, afin de pouvoir écrire

$$\boldsymbol{\theta}_2=\boldsymbol{\tau}(\boldsymbol{y},\boldsymbol{\theta}_1).$$

Ceci implique alors que, identiquement en $\boldsymbol{\theta}_1$,

$$D_2\ell\big(\boldsymbol{y},\boldsymbol{\theta}_1,\boldsymbol{\tau}(\boldsymbol{y},\boldsymbol{\theta}_1)\big)=\mathbf{0}. \tag{8.52}$$

En substituant $\boldsymbol{\tau}(\boldsymbol{y},\boldsymbol{\theta}_1)$ à $\boldsymbol{\theta}_2$ dans $\ell(\boldsymbol{y},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$, nous obtenons la **fonction de logvraisemblance concentrée**

$$\ell^c(\boldsymbol{y},\boldsymbol{\theta}_1)\equiv\ell\big(\boldsymbol{y},\boldsymbol{\theta}_1,\boldsymbol{\tau}(\boldsymbol{y},\boldsymbol{\theta}_1)\big).$$

Si $\hat{\boldsymbol{\theta}}_1$ maximise celle-ci, nous pouvons alors obtenir $\hat{\boldsymbol{\theta}}_2$ grâce à $\boldsymbol{\tau}(\boldsymbol{y},\hat{\boldsymbol{\theta}}_1)$, et il est évident que $[\hat{\boldsymbol{\theta}}_1 \vdots \hat{\boldsymbol{\theta}}_2]$ maximisera $\ell(\boldsymbol{y},\boldsymbol{\theta})$. Dans certains cas, cette stratégie peut réduire substantiellement la quantité d'efforts nécessaires à l'obtention des estimations ML.

Il est évident que $\ell^c(\boldsymbol{y},\hat{\boldsymbol{\theta}}_1)$ sera identique à $\ell(\boldsymbol{y},\hat{\boldsymbol{\theta}})$. Cependant, il n'est pas évident que nous puissions calculer une matrice de covariance estimée pour $\hat{\boldsymbol{\theta}}_1$ basée sur $\ell^c(\boldsymbol{y},\boldsymbol{\theta}_1)$ de la même manière que celle que nous calculons lorsque nous nous basons sur $\ell(\boldsymbol{y},\boldsymbol{\theta})$. En fait, à condition d'utiliser comme estimateur l'inverse de l'opposée de la matrice Hessienne empirique, on dispose d'un estimateur évident. La raison est que, en vertu de la manière dont ℓ^c est construite, l'inverse de sa matrice Hessienne par rapport à $\boldsymbol{\theta}_1$ est égale au bloc $(\boldsymbol{\theta}_1,\boldsymbol{\theta}_1)$ de l'inverse de la matrice Hessienne de $\ell(\boldsymbol{y},\boldsymbol{\theta})$ par rapport au vecteur paramétrique entier $\boldsymbol{\theta}$. Ceci provient du théorème de l'enveloppe et des résultats standards sur les matrices partitionnées, comme nous allons le démontrer à présent.

Grâce aux conditions du premier ordre (8.52), le gradient de ℓ^c par rapport à $\boldsymbol{\theta}_1$ est

$$\begin{aligned} D_1\ell^c(\boldsymbol{\theta}_1) &= D_1\ell\big(\boldsymbol{\theta}_1,\boldsymbol{\tau}(\boldsymbol{\theta}_1)\big)+D_2\ell\big(\boldsymbol{\theta}_1,\boldsymbol{\tau}(\boldsymbol{\theta}_1)\big)D\boldsymbol{\tau}(\boldsymbol{\theta}_1)\\ &= D_1\ell\big(\boldsymbol{\theta}_1,\boldsymbol{\tau}(\boldsymbol{\theta}_1)\big), \end{aligned}$$

où la dépendance explicite à $\boldsymbol{y}$ a été supprimée. Ce résultat est simplement le théorème de l'enveloppe appliqué à ℓ^c. Ainsi la matrice Hessienne de $\ell^c(\boldsymbol{\theta}_1)$ est

$$D_{11}\ell^c(\boldsymbol{\theta}_1)=D_{11}\ell\big(\boldsymbol{\theta}_1,\boldsymbol{\tau}(\boldsymbol{\theta}_1)\big)+D_{12}\ell\big(\boldsymbol{\theta}_1,\boldsymbol{\tau}(\boldsymbol{\theta}_1)\big)D\boldsymbol{\tau}(\boldsymbol{\theta}_1). \tag{8.53}$$

Afin d'exprimer le membre de droite de (8.53) en termes uniquement des blocs de la matrice Hessienne de ℓ, nous dérivons (8.52) par rapport à $\boldsymbol{\theta}_1$, et obtenons

$$D_{21}\ell\big(\boldsymbol{\theta}_1, \boldsymbol{\tau}(\boldsymbol{\theta}_1)\big) + D_{22}\ell\big(\boldsymbol{\theta}_1, \boldsymbol{\tau}(\boldsymbol{\theta}_1)\big) D\boldsymbol{\tau}(\boldsymbol{\theta}_1) = \mathbf{0}.$$

En résolvant cette équation pour $D\boldsymbol{\tau}(\boldsymbol{\theta}_1)$ et en substituant le résultat dans (8.53), l'expression de la matrice Hessienne de ℓ^c, nous aboutissons à

$$D_{11}\ell^c = D_{11}\ell - D_{12}\ell(D_{22}\ell)^{-1}D_{21}\ell, \tag{8.54}$$

expression dans laquelle les arguments de ℓ et ℓ^c ont été omis pour simplifier l'écriture. La matrice Hessienne de ℓ peut être écrite sous forme partitionnée comme

$$D_{\theta\theta}\,\ell = \begin{bmatrix} D_{11}\ell & D_{12}\ell \\ D_{21}\ell & D_{22}\ell \end{bmatrix}.$$

Les résultats standards sur les matrices partitionnées (consulter l'Annexe A) nous apprennent que le bloc $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1)$ de l'inverse de cette matrice Hessienne est

$$\big(D_{11}\ell - D_{12}\ell(D_{22}\ell)^{-1}D_{21}\ell\big)^{-1},$$

dont l'inverse est précisément l'expression pour $D_{11}\ell^c$ dans (8.54).

L'utilisation des fonctions de logvraisemblance concentrées comporte certains désavantages. La fonction de logvraisemblance originelle peut dans la plupart des cas être écrite de manière commode comme

$$\ell(\boldsymbol{y}, \boldsymbol{\theta}) = \sum_{t=1}^{n} \ell_t(y_t, \boldsymbol{\theta}). \tag{8.55}$$

Ceci n'est cependant généralement pas exact pour la fonction de logvraisemblance concentrée. L'équivalent de (8.55) est

$$\ell^c(\boldsymbol{y}, \boldsymbol{\theta}_1) = \sum_{t=1}^{n} \ell_t\big(y_t, \boldsymbol{\theta}_1, \boldsymbol{\tau}(\boldsymbol{y}, \boldsymbol{\theta}_1)\big),$$

et il est évident qu'en raison de la dépendance de $\boldsymbol{\tau}(\cdot)$ au vecteur entier $\boldsymbol{y}$, il n'y a pas en général de manière simple d'écrire $\ell^c(\boldsymbol{y}, \boldsymbol{\theta}_1)$ comme une somme des contributions de chacune des observations. Cela signifie que l'estimateur OPG de la matrice d'information n'est généralement pas disponible pour les fonctions de logvraisemblance concentrées. On peut bien sûr utiliser $\ell^c(\boldsymbol{y}, \boldsymbol{\theta}_1)$ pour l'estimation et se reporter ensuite vers $\ell(\boldsymbol{y}, \boldsymbol{\theta})$ quand vient l'heure d'estimer la matrice de covariance des estimations.

8.8 L'Efficacité Asymptotique de l'Estimateur ML

Dans cette section, nous démontrerons l'**efficacité asymptotique** de l'estimateur ML ou, à proprement parler, de l'estimateur ML de Type 2. La convergence asymptotique signifie que la variance de la distribution asymptotique de n'importe quel estimateur convergent des paramètres diffère de celle d'un estimateur efficace asymptotiquement par une matrice semi-définie positive; voir la Définition 5.6. On parle d'*un* estimateur efficace asymptotiquement plutôt que de *l'*estimateur efficace asymptotiquement parce que la propriété d'efficacité asymptotique est une propriété de la distribution asymptotique seulement; il peut exister de nombreux estimateurs (et il en existera effectivement) qui diffèrent avec des échantillons finis mais qui ont la même distribution asymptotique efficace. Un exemple de modèle de régression non linéaire peut être pris, dans lequel, comme nous le verrons dans la Section 8.10, l'estimation NLS est équivalente à l'estimation ML si nous supposons la normalité des aléas. Comme nous l'avons vu dans la Section 6.6, il existe des modèles non linéaires qui correspondent exactement à des modèles linéaires auxquels on impose certaines contraintes non linéaires. Dans de tels cas nous avons vu que l'estimation en une étape qui commence à partir des estimations de modèle linéaire était asymptotiquement équivalente à l'estimation NLS, et par conséquent asymptotiquement efficace. L'estimation en une étape est aussi possible dans le contexte général du maximum de vraisemblance et peut souvent fournir un estimateur efficace qui est plus facile à calculer que l'estimateur ML lui-même.

Nous commençons notre démonstration de l'efficacité asymptotique de l'estimateur ML par une discussion applicable à n'importe quel estimateur convergent, au taux $n^{1/2}$ et asymptotiquement sans biais, des paramètres du modèle représenté par la fonction de logvraisemblance $\ell(\boldsymbol{y}, \boldsymbol{\theta})$. Notons que la convergence en elle-même n'implique pas l'absence de biais asymptotiquement sans l'imposition de diverses conditions de régularité. Puisque tout estimateur convergent et intéressant au sens économétrique que nous connaissons est en fait asymptotiquement sans biais, nous ne traiterons ici que de tels estimateurs. Désignons un tel estimateur par $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$, avec une notation qui insiste sur le fait que l'estimateur est une variable aléatoire qui dépend de l'échantillon $\boldsymbol{y}$ réalisé. Notons que nous avons changé ici de notation, car $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ n'est pas en général l'estimateur ML. Au lieu de cela, ce dernier sera noté $\tilde{\boldsymbol{\theta}}(\boldsymbol{y})$; la nouvelle notation est conçue pour être cohérente, à travers l'ouvrage, avec notre traitement des estimateurs contraints et non contraints, puisque dans un sens profond l'estimateur ML correspond aux premiers de ces estimateurs et l'estimateur convergent arbitraire $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ correspond aux seconds.

Comme $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ est supposé être asymptotiquement sans biais, nous avons

$$\lim_{n\to\infty} E_\theta\big(\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}\big) = \mathbf{0}.$$

Avec une notation plus explicite, ceci devient:

$$\lim_{n\to\infty}\left(\int_{\mathcal{Y}^n} L^n(\boldsymbol{y}^n,\boldsymbol{\theta})\,\hat{\boldsymbol{\theta}}^n(\boldsymbol{y}^n)\,d\boldsymbol{y}^n - \boldsymbol{\theta}\right) = \mathbf{0}, \tag{8.56}$$

où, comme précédemment, $\mathcal{Y}^n$ désigne le sous-espace de $\mathbb{R}^{nm}$ sur lequel le vecteur échantillon $\boldsymbol{y}^n$ peut varier en conservant une taille n. Les prochaines étapes impliquent la différentiation de la relation (8.56) par rapport aux éléments de $\boldsymbol{\theta}$, en permutant l'ordre des opérations de différentiation et d'intégration, et en calculant la limite quand $n \to \infty$. Nous omettons la discussion sur les conditions de régularité nécessaires pour que ceci soit admissible et poursuivons en écrivant directement le résultat de la différentiation du $j^{\text{ième}}$ élément de (8.56) par rapport au $i^{\text{ième}}$ élément de $\boldsymbol{\theta}$:

$$\lim_{n\to\infty}\int_{\mathcal{Y}^n} L^n(\boldsymbol{y}^n,\boldsymbol{\theta})\frac{\partial \ell^n(\boldsymbol{y}^n,\boldsymbol{\theta})}{\partial\theta_i}\,\hat{\theta}_j(\boldsymbol{y}^n)\,d\boldsymbol{y}^n = \delta^i_j. \tag{8.57}$$

Le membre de droite de cette équation est le delta de Kronecker, égal à 1 quand $i = j$ et égal à 0 sinon. L'équation (8.57) peut être récrite comme

$$\lim_{n\to\infty} E_\theta\left(n^{-1/2}\frac{\partial \ell^n(\boldsymbol{y}^n,\boldsymbol{\theta})}{\partial\theta_i}\,n^{1/2}\big(\hat{\theta}_j - \theta_j\big)\right) = \delta^i_j, \tag{8.58}$$

où nous avons introduit certaines puissances de n pour s'assurer que les quantités qui apparaissent dans l'expression possèdent des limites en probabilité de l'ordre de l'unité. Nous avons aussi retranché θ_j à $\hat{\theta}_j$; ceci a été possible parce que $E_\theta\big(D_\theta\,\ell(\boldsymbol{\theta})\big) = \mathbf{0}$, et désormais le produit de θ_j par $E_\theta\big(D_\theta\,\ell(\boldsymbol{\theta})\big)$ est également nul.

L'expression (8.58) peut être écrite sans aucune opération à la limite si nous utilisons les distributions asymptotiques du gradient $D_\theta\,\ell$ et le vecteur $n^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})$. Introduisons une notation supplémentaire dans le but de discuter des variables aléatoires asymptotiques. Nous posons les définitions

$$\boldsymbol{s}^n(\boldsymbol{\theta}) \equiv n^{-1/2}\boldsymbol{g}(\boldsymbol{y}^n,\boldsymbol{\theta}), \quad \boldsymbol{s}(\boldsymbol{\theta}) \equiv \operatorname*{plim}_{n\to\infty,\theta}\boldsymbol{s}^n(\boldsymbol{\theta}), \tag{8.59}$$

$$\hat{\boldsymbol{t}}^n(\boldsymbol{\theta}) \equiv n^{1/2}\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big), \text{ et } \hat{\boldsymbol{t}}(\boldsymbol{\theta}) \equiv \operatorname*{plim}_{n\to\infty,\theta}\hat{\boldsymbol{t}}^n(\boldsymbol{\theta}). \tag{8.60}$$

Ainsi $\boldsymbol{s}(\boldsymbol{\theta})$ et $\hat{\boldsymbol{t}}(\boldsymbol{\theta})$ sont des vecteurs de dimension k dont les éléments types respectifs sont $s_i(\boldsymbol{\theta})$ et $\hat{t}_j(\boldsymbol{\theta})$. Le premier est la valeur à la limite de $n^{-1/2}$ fois un élément type du gradient de $\ell(\boldsymbol{y},\boldsymbol{\theta})$, tandis que le second est la valeur à la limite de $n^{1/2}$ fois un élément type de la différence entre $\hat{\boldsymbol{\theta}}$ et $\boldsymbol{\theta}$. La notation a été conçue dans l'intention d'être mnémotechnique, $\boldsymbol{s}(\boldsymbol{\theta})$ correspondant au vecteur *score* et $\hat{\boldsymbol{t}}(\boldsymbol{\theta})$ correspondant au *thêta chapeau*. Grâce à cette nouvelle notation commode, l'expression (8.58) devient

$$E_\theta\big(\hat{\boldsymbol{t}}(\boldsymbol{\theta})\boldsymbol{s}^\top(\boldsymbol{\theta})\big) = \mathbf{I}_k, \tag{8.61}$$

où $\mathbf{I}_k$ est simplement la matrice identité de dimension $k \times k$.

Il n'est pas en général exact pour *n'importe quel* estimateur convergent que la limite en probabilité dans (8.60) existe ou, si elle existe, qu'elle soit non nulle. La classe des estimateurs pour lesquels celle-ci existe et n'est pas nulle est appelée la classe des **estimateurs convergents au taux** $n^{1/2}$. Ainsi que nous en avons discuté dans le Chapitre 5, ceci signifie que le taux de convergence, quand $n \to \infty$, de l'estimateur $\hat{\boldsymbol{\theta}}$ vers la véritable valeur $\boldsymbol{\theta}$ est le même que le taux de convergence de $n^{-1/2}$ vers zéro. L'existence d'une limite en probabilité non nulle dans (8.60) implique clairement cette propriété, et nous avons déjà montré que l'estimateur ML est convergent au taux $n^{1/2}$. La convergence de $\hat{\boldsymbol{\theta}}$ implique également que l'espérance de la variable aléatoire à la limite $\hat{\boldsymbol{t}}(\boldsymbol{\theta})$ est égale à zéro.

Pour la partie suivante de l'argumentation, nous considérons en premier lieu le cas simple dans lequel $k = 1$. Alors à la place de (8.61) nous avons la relation scalaire

$$E_\theta\big(\hat{t}(\theta)s(\theta)\big) = \mathrm{Cov}_\theta\big(\hat{t}(\theta), s(\theta)\big) = 1. \tag{8.62}$$

Ici nous avons utilisé le fait que les espérances aussi bien de $\hat{t}(\theta)$ que de $s(\theta)$ sont zéro. Le résultat (8.62) implique l'inégalité bien connue de Cauchy-Schwartz:

$$1 = \Big(\mathrm{Cov}_\theta\big(\hat{t}(\theta), s(\theta)\big)\Big)^2 \leq \mathrm{Var}_\theta\big(\hat{t}(\theta)\big)\mathrm{Var}_\theta\big(s(\theta)\big) = \mathrm{Var}_\theta\big(\hat{t}(\theta)\big)\mathcal{I}(\theta), \tag{8.63}$$

où la dernière égalité provient de la définition (8.59) de $s(\theta)$ et de la définition de la matrice d'information asymptotique $\mathcal{I}(\theta)$, qui est dans ce cas un scalaire. L'inégalité (8.63) implique que

$$\mathrm{Var}_\theta\big(\hat{t}(\theta)\big) \geq \frac{1}{\mathcal{I}(\theta)}. \tag{8.64}$$

Ce résultat établit, dans ce cas à une dimension, que la variance asymptotique de n'importe quel estimateur convergent à un taux $n^{1/2}$ ne peut pas être inférieure à l'inverse de ce qu'il semble être logique d'appeler le scalaire d'information. Comme le membre de droite de (8.64) est précisément la variance asymptotique de l'estimateur ML, l'efficacité asymptotique de ce dernier est aussi établie par ce résultat. Notons que (8.64) élimine n'importe quel estimateur pour lequel la limite en probabilité de $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ est égale à zéro. Un tel estimateur serait naturellement *plus* efficace asymptotiquement que l'estimateur ML, car il devrait converger plus rapidement vers la véritable valeur de $\boldsymbol{\theta}$.

Le résultat général analogue à (8.64) pour le cas $k \geq 1$ peut maintenant être établi en ajoutant un tout petit peu plus de travail. Considérons la matrice entière de covariance de tous les éléments de $\hat{\boldsymbol{t}}$ et de $\boldsymbol{s}$, c'est-à-dire la

matrice de covariance de $[\hat{\boldsymbol{t}}(\boldsymbol{\theta}) \vdots \boldsymbol{s}(\boldsymbol{\theta})]$. Notons $\boldsymbol{V}$ la matrice de covariance de $\hat{\boldsymbol{t}}$. Alors (8.61) et le fait que $\mathrm{Var}_\theta\big(\boldsymbol{s}^\top(\boldsymbol{\theta})\big) = \mathcal{I}(\boldsymbol{\theta})$ signifient que la matrice de covariance de $[\hat{\boldsymbol{t}}(\boldsymbol{\theta}) \vdots \boldsymbol{s}(\boldsymbol{\theta})]$ peut être écrite comme

$$\mathrm{Var}(\hat{\boldsymbol{t}}, \boldsymbol{s}) = \begin{bmatrix} \boldsymbol{V} & \mathbf{I}_k \\ \mathbf{I}_k & \mathcal{I} \end{bmatrix}.$$

Comme il s'agit d'une matrice de covariance, celle-ci doit être semi-définie positive. Ainsi, pour n'importe quel vecteur $\boldsymbol{a}$ de dimension k, l'expression suivante est non négative:

$$\big[\boldsymbol{a}^\top - \boldsymbol{a}^\top \mathcal{I}^{-1}\big] \begin{bmatrix} \boldsymbol{V} & \mathbf{I}_k \\ \mathbf{I}_k & \mathcal{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{a} \\ -\mathcal{I}^{-1}\boldsymbol{a} \end{bmatrix} = \boldsymbol{a}^\top\big(\boldsymbol{V} - \mathcal{I}^{-1}\big)\boldsymbol{a}.$$

Mais ceci implique, comme $\boldsymbol{a}$ est arbitraire, que la matrice $\big(\boldsymbol{V} - \mathcal{I}^{-1}\big)$ est semi-définie positive, ce qui correspond à ce que nous avons voulu prouver.

Ce résultat constitue un cas particulier de la **borne inférieure de Cramér-Rao**, suggérée à l'origine par Fisher (1925) dans un de ses premiers articles classiques sur l'estimation ML et énoncé sous sa forme moderne par Cramér (1946) et Rao (1945). Celle-ci est spéciale parce qu'il s'agit d'une version asymptotique du résultat d'origine. La borne inférieure de Cramér-Rao s'applique en fait à *n'importe quel* estimateur sans biais sans tenir compte de la taille de l'échantillon. Cependant, comme les estimateurs ML ne sont pas en général sans biais, seul le résultat de la version asymptotique représente un intérêt dans le contexte de l'estimation ML, et aussi avons-nous restreint notre attention au cas asymptotique.

Le fait que l'estimateur ML atteigne asymptotiquement la borne inférieure de Cramér-Rao implique que n'importe quel estimateur convergent au taux $n^{1/2}$ peut être écrit comme la somme de l'estimateur ML et d'un autre vecteur aléatoire qui est asymptotiquement indépendant du premier. Ce résultat fournit une manière révélatrice de réfléchir à la relation entre les estimateurs efficaces et non efficaces. Pour l'établir, nous commençons par poser les définitions

$$\begin{aligned} &\tilde{\boldsymbol{t}}^n(\boldsymbol{\theta}) \equiv n^{1/2}\big(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\big), \quad \tilde{\boldsymbol{t}}(\boldsymbol{\theta}) \equiv \operatorname*{plim}_{n\to\infty}\big(\tilde{\boldsymbol{t}}^n(\boldsymbol{\theta})\big), \\ &\boldsymbol{v}^n \equiv \hat{\boldsymbol{t}}^n(\boldsymbol{\theta}) - \tilde{\boldsymbol{t}}^n(\boldsymbol{\theta}), \text{ et } \boldsymbol{v} \equiv \hat{\boldsymbol{t}}(\boldsymbol{\theta}) - \tilde{\boldsymbol{t}}(\boldsymbol{\theta}). \end{aligned} \tag{8.65}$$

Comme on peut le voir à partir des définitions (8.60) et (8.65), $\boldsymbol{v}^n$ et $\boldsymbol{v}$ ne dépendent pas directement de $\boldsymbol{\theta}$.

Nous souhaitons montrer que la matrice de covariance de $\boldsymbol{v}$ et $\tilde{\boldsymbol{t}}$ est une matrice égale à zéro. Cette matrice de covariance est

$$\begin{aligned} \mathrm{Cov}_\theta\big(\boldsymbol{v}, \tilde{\boldsymbol{t}}(\boldsymbol{\theta})\big) &= E_\theta\big(\boldsymbol{v}\tilde{\boldsymbol{t}}^\top(\boldsymbol{\theta})\big) \\ &= E_\theta\Big(\big(\hat{\boldsymbol{t}}(\boldsymbol{\theta}) - \tilde{\boldsymbol{t}}(\boldsymbol{\theta})\big)\tilde{\boldsymbol{t}}^\top(\boldsymbol{\theta})\Big) \\ &= E_\theta\big(\hat{\boldsymbol{t}}(\boldsymbol{\theta})\tilde{\boldsymbol{t}}^\top(\boldsymbol{\theta})\big) - \mathcal{I}^{-1}(\boldsymbol{\theta}). \end{aligned} \tag{8.66}$$

En utilisant l'égalité de la matrice d'information, le résultat (8.38) peut être écrit comme

$$n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{=} \big(\mathcal{I}(\boldsymbol{\theta})\big)^{-1}\big(n^{-1/2}\boldsymbol{g}(\boldsymbol{\theta})\big).$$

Dans la notation de (8.59) et (8.60), ceci devient

$$\tilde{\boldsymbol{t}}(\boldsymbol{\theta}) = \mathcal{I}^{-1}(\boldsymbol{\theta})\boldsymbol{s}(\boldsymbol{\theta}).$$

Ainsi, en continuant à partir de la dernière ligne de (8.66), nous obtenons

$$\begin{aligned}\mathrm{Cov}_\theta\big(\boldsymbol{v}, \tilde{\boldsymbol{t}}(\boldsymbol{\theta})\big) &= E_\theta\big(\hat{\boldsymbol{t}}(\boldsymbol{\theta})\boldsymbol{s}^\top(\boldsymbol{\theta})\mathcal{I}^{-1}(\boldsymbol{\theta})\big) - \mathcal{I}^{-1}(\boldsymbol{\theta}) \\ &= E_\theta\big(\hat{\boldsymbol{t}}(\boldsymbol{\theta})\boldsymbol{s}^\top(\boldsymbol{\theta})\big)\mathcal{I}^{-1}(\boldsymbol{\theta}) - \mathcal{I}^{-1}(\boldsymbol{\theta}) \\ &= \mathcal{I}^{-1}(\boldsymbol{\theta}) - \mathcal{I}^{-1}(\boldsymbol{\theta}) = \mathbf{0}.\end{aligned}$$

Le résultat fondamental (8.61) a été utilisé pour obtenir ici la dernière ligne.

Ainsi, nous concluons que

$$\hat{\boldsymbol{t}}(\boldsymbol{\theta}) = \tilde{\boldsymbol{t}}(\boldsymbol{\theta}) + \boldsymbol{v}, \tag{8.67}$$

où $\boldsymbol{v}$ est asymptotiquement non corrélé avec $\tilde{\boldsymbol{t}}$. Si $\hat{\boldsymbol{t}}$ et $\tilde{\boldsymbol{t}}$ sont asymptotiquement normaux, cette corrélation asymptotiquement nulle implique par la suite une indépendance asymptotique. Une autre manière d'écrire le résultat (8.67) est

$$\hat{\boldsymbol{\theta}} \stackrel{a}{=} \tilde{\boldsymbol{\theta}} + n^{-1/2}\boldsymbol{v}^n.$$

Ceci montre clairement qu'un estimateur $\hat{\boldsymbol{\theta}}$ non efficace mais convergent peut toujours être décomposé, asymptotiquement, en la somme d'un estimateur ML $\tilde{\boldsymbol{\theta}}$ asymptotiquement efficace et d'une autre variable aléatoire, qui tend vers zéro quand $n \to \infty$ et est asymptotiquement non corrélée avec l'estimateur efficace. Evidemment, tout l'éventail des estimateurs asymptotiquement normaux et convergents peut être généré à partir de l'estimateur ML $\tilde{\boldsymbol{\theta}}$ en lui additionnant des variables aléatoires multivariées normales d'espérances nulles indépendantes de $\tilde{\boldsymbol{\theta}}$. On peut imaginer que celles-ci soient des bruits parasitant le signal efficace émis par $\tilde{\boldsymbol{\theta}}$. L'interprétation du résultat de Cramér-Rao est assez évidente à présent: comme la variance de la somme de deux variables aléatoires indépendantes est la somme de leurs variances respectives, la matrice semi-définie positive qui correspond à la différence entre les matrices de covariance de $\hat{\boldsymbol{\theta}}$ et $\tilde{\boldsymbol{\theta}}$ est précisément la matrice de covariance (peut-être dégénérée) du vecteur des variables de bruit $n^{-1/2}\boldsymbol{v}$.

Ces résultats pour les estimateurs ML sont similaires, mais beaucoup plus forts que les résultats obtenus pour les moindres carrés non linéaires dans la Section 5.5. Nous y avions vu que n'importe quel estimateur convergent mais non efficace qui est asymptotiquement linéaire pour les aléas peut être écrit comme la somme de l'estimateur efficace et d'une variable aléatoire (ou vecteur) qui est asymptotiquement non corrélée avec l'estimateur efficace. La démonstration du Théorème de Gauss-Markov fournissait également un résultat similaire.

8.9 Les Trois Statistiques de Test Classiques

Une des caractéristiques attrayantes de l'estimation ML est que les statistiques de test basées sur les trois principes dont nous avons discuté pour la première fois dans le Chapitre 3 — le principe du rapport de vraisemblance, le principe du multiplicateur de Lagrange et le principe de Wald — sont toujours disponibles et sont souvent faciles à calculer. Ces trois principes de test d'hypothèse furent énoncés pour la première fois dans le contexte de l'estimation ML, et certains auteurs utilisent encore les termes de "rapport de vraisemblance", "multiplicateur de Lagrange", et "Wald" dans le seul contexte des tests basés sur les estimations ML. Dans cette section, nous fournissons une introduction à ce que l'on désigne souvent sous le nom des **trois tests classiques**. Ces trois statistiques de test possèdent la même distribution asymptotique sous l'hypothèse nulle; s'il y a r contraintes d'égalité, elles sont distribuées suivant une distribution du $\chi^2(r)$. En effet, elles tendent réellement vers la même variable aléatoire asymptotiquement, à la fois sous l'hypothèse nulle et sous la série des DGP qui sont proches de l'hypothèse nulle dans un certains sens. Un traitement approprié de ces résultats importants nécessite plus de développements que nous n'en disposons dans cette section. Ainsi, nous remettons celui-ci au Chapitre 13, qui fournit une discussion beaucoup plus détaillée des trois statistiques de test classiques.

Conceptuellement, le plus simple des trois tests classiques est le **rapport de vraisemblance**, ou test **LR**. La statistique de test est simplement deux fois la différence entre les valeurs contrainte et non contrainte de la fonction de logvraisemblance,

$$2\big(\ell(\hat{\boldsymbol{\theta}}) - \ell(\tilde{\boldsymbol{\theta}})\big), \tag{8.68}$$

où $\hat{\boldsymbol{\theta}}$ désigne l'estimation ML non contrainte de $\boldsymbol{\theta}$, $\tilde{\boldsymbol{\theta}}$ désigne l'estimation ML soumise aux r contraintes distinctes, et où la dépendance de ℓ à $\boldsymbol{y}$ a été supprimée pour simplifier la notation. Le nom de la statistique LR provient du fait que (8.68) est égale à

$$2\log\left(\frac{L(\hat{\boldsymbol{\theta}})}{L(\tilde{\boldsymbol{\theta}})}\right),$$

ou deux fois le logarithme du rapport des fonctions de vraisemblance. Elle est très facile à calculer lorsqu'à la fois les estimations contraintes et les non contraintes sont disponibles, et c'est une de ses caractéristiques les plus attrayantes.

Pour dériver la distribution asymptotique de la statistique LR, il faut calculer un développement en série de Taylor au second ordre de $\ell(\tilde{\boldsymbol{\theta}})$ autour de $\hat{\boldsymbol{\theta}}$. Bien que nous ne terminerons pas la construction de cette statistique dans cette section, il est révélateur de parcourir les premières étapes. Le résultat du développement en série de Taylor est

$$\ell(\tilde{\boldsymbol{\theta}}) \cong \ell(\hat{\boldsymbol{\theta}}) + \tfrac{1}{2}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})^\top \boldsymbol{H}(\hat{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}). \tag{8.69}$$

Ici, il n'y a pas de terme du premier ordre parce que $\boldsymbol{g}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ grâce aux conditions du premier ordre (8.12). En ordonnant les termes de (8.69) nous obtenons

$$\begin{aligned} 2\big(\ell(\hat{\boldsymbol{\theta}}) - \ell(\tilde{\boldsymbol{\theta}})\big) &\cong -(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})^\top \boldsymbol{H}(\hat{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) \\ &\overset{a}{=} \big(n^{1/2}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})\big)^\top \mathcal{I}(\hat{\boldsymbol{\theta}})\big(n^{1/2}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})\big). \end{aligned} \tag{8.70}$$

Cet exercice permet d'expliquer la provenance du facteur de 2 dans la définition de la statistique LR. La prochaine étape consisterait à remplacer $n^{1/2}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})$ dans (8.70) par

$$n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) - n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$$

et d'utiliser ensuite le résultat (8.38), simultanément avec un résultat analogue pour les estimations contraintes que nous obtiendrons sous peu, pour établir la distribution asymptotique de la statistique LR. Nous réaliserons ceci dans le Chapitre 13.

Nous portons maintenant notre attention sur le **multiplicateur de Lagrange**, ou test **LM**. En effet, cette statistique de test porte deux noms et prend deux formes différentes, qui s'avèrent être numériquement identiques si la même estimation de la matrice d'information est utilisée pour les calculer. Une forme, proposée à l'origine par Rao (1948), est appelée la **forme score du test LM**, ou simplement le **test score**, et est calculée en utilisant le gradient ou le vecteur score du modèle non contraint évalué avec les estimations contraintes. L'autre forme, qui donne au test son nom, a été proposée par Aitchison et Silvey (1958, 1960) et Silvey (1959). Cette dernière forme est calculée en utilisant le vecteur des multiplicateurs de Lagrange qui émerge si on maximise la fonction de vraisemblance soumise aux contraintes au moyen d'un Lagrangien. Les économètres utilisent généralement le test LM sous sa forme score mais insistent néanmoins pour le nommer test LM, peut-être parce que les multiplicateurs de Lagrange sont aussi largement utilisés en économétrie. Les références sur les tests LM en économétrie sont Breusch et Pagan (1980) et Engle (1982a, 1984). Buse (1982) fournit une discussion intuitive des relations entre les tests LR, LM, et Wald.

Une manière de maximiser $\ell(\boldsymbol{\theta})$ soumise aux contraintes exactes

$$\boldsymbol{r}(\boldsymbol{\theta}) = \mathbf{0}, \tag{8.71}$$

où $\boldsymbol{r}(\boldsymbol{\theta})$ est un vecteur de dimension r avec $r \leq k$, consiste à maximiser simultanément le Lagrangien

$$\ell(\boldsymbol{\theta}) - \boldsymbol{r}^\top(\boldsymbol{\theta})\boldsymbol{\lambda}$$

par rapport à $\boldsymbol{\theta}$ et à le minimiser par rapport au vecteur de dimension r $\boldsymbol{\lambda}$ des multiplicateurs de Lagrange. Les conditions du premier ordre qui caractérisent la solution de ce problème sont

$$\begin{aligned} \boldsymbol{g}(\tilde{\boldsymbol{\theta}}) - \boldsymbol{R}^\top(\tilde{\boldsymbol{\theta}})\tilde{\boldsymbol{\lambda}} &= \mathbf{0} \\ \boldsymbol{r}(\tilde{\boldsymbol{\theta}}) &= \mathbf{0}, \end{aligned} \tag{8.72}$$

où $\boldsymbol{R}(\boldsymbol{\theta})$ est une matrice de dimension $r \times k$ avec comme élément type $\partial r_i(\boldsymbol{\theta})/\partial\theta_j$.

Nous sommes intéressés par la distribution de $\tilde{\boldsymbol{\lambda}}$ sous l'hypothèse nulle, aussi supposerons-nous que le DGP satisfait (8.71) avec le vecteur paramétrique $\boldsymbol{\theta}_0$. La valeur du vecteur $\boldsymbol{\lambda}$ des multiplicateurs de Lagrange si $\tilde{\boldsymbol{\theta}}$ était égal à $\boldsymbol{\theta}_0$ devrait être égale à zéro. Ainsi, il semble naturel de prendre un développement en série de Taylor au premier ordre des conditions du premier ordre (8.72) autour du point $(\boldsymbol{\theta}_0, \mathbf{0})$. Ceci donne

$$\begin{aligned} \boldsymbol{g}(\boldsymbol{\theta}_0) + \boldsymbol{H}(\bar{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) - \boldsymbol{R}^\top(\bar{\boldsymbol{\theta}})\tilde{\boldsymbol{\lambda}} &= \mathbf{0} \\ -\boldsymbol{R}(\ddot{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) &= \mathbf{0}, \end{aligned}$$

où $\bar{\boldsymbol{\theta}}$ et $\ddot{\boldsymbol{\theta}}$ désignent les valeurs de $\boldsymbol{\theta}$ qui se situent entre $\tilde{\boldsymbol{\theta}}$ et $\boldsymbol{\theta}_0$. Ces équations peuvent être récrites comme

$$\begin{bmatrix} -\boldsymbol{H}(\bar{\boldsymbol{\theta}}) & \boldsymbol{R}^\top(\bar{\boldsymbol{\theta}}) \\ \boldsymbol{R}(\ddot{\boldsymbol{\theta}}) & \mathbf{0} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \\ \tilde{\boldsymbol{\lambda}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{g}(\boldsymbol{\theta}_0) \\ \mathbf{0} \end{bmatrix}. \tag{8.73}$$

Si nous multiplions $\boldsymbol{H}(\bar{\boldsymbol{\theta}})$ par n^{-1}, $\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ par $n^{1/2}$, $\boldsymbol{g}(\boldsymbol{\theta}_0)$ par $n^{-1/2}$, et $\tilde{\boldsymbol{\lambda}}$ par $n^{-1/2}$, nous ne changeons pas l'égalité dans (8.73), et nous transformons toutes les quantités qui y apparaissent en des quantités $O(1)$. Les lecteurs peuvent vouloir vérifier que ces facteurs de n sont en effet les plus appropriés et, en particulier, que $\tilde{\boldsymbol{\lambda}}$ doit être multiplié par $n^{-1/2}$. En utilisant le fait que $\tilde{\boldsymbol{\theta}}$ et par conséquent $\bar{\boldsymbol{\theta}}$ et $\ddot{\boldsymbol{\theta}}$ sont convergents, en appliquant une loi des grands nombres convenable à $n^{-1}\boldsymbol{H}(\bar{\boldsymbol{\theta}})$, et en résolvant les équations du système résultant, nous obtenons

$$\begin{bmatrix} n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ n^{-1/2}\tilde{\boldsymbol{\lambda}} \end{bmatrix} \stackrel{a}{=} \begin{bmatrix} -\mathcal{H}_0 & \boldsymbol{R}_0^\top \\ \boldsymbol{R}_0 & \mathbf{0} \end{bmatrix}^{-1} \begin{bmatrix} n^{-1/2}\boldsymbol{g}(\boldsymbol{\theta}_0) \\ \mathbf{0} \end{bmatrix}, \tag{8.74}$$

où $\mathcal{H}_0$ désigne $\mathcal{H}(\boldsymbol{\theta}_0)$ et $\boldsymbol{R}_0$ désigne $\boldsymbol{R}(\boldsymbol{\theta}_0)$.

Le système des équations (8.74) est, pour le cas contraint, l'équivalent de l'équation (8.38) pour le cas non contraint. La première chose à noter, le concernant, est que les k éléments de $n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ et les r éléments de $n^{-1/2}\tilde{\boldsymbol{\lambda}}$ dépendent tous du vecteur de dimension k aléatoire $n^{-1/2}\boldsymbol{g}(\boldsymbol{\theta}_0)$. Nous avons déjà vu que, sous des conditions de régularité standards, ce dernier est asymptotiquement normalement distribué avec un vecteur d'espérances nulles et une matrice de covariance $\mathcal{I}(\boldsymbol{\theta}_0)$. Ainsi à partir de (8.74) nous voyons qu'à la fois $n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ et $n^{-1/2}\tilde{\boldsymbol{\lambda}}$ doivent être asymptotiquement normalement distribués. Observons que le vecteur de dimension $(k + r)$ dans le membre de gauche de (8.74) doit avoir une matrice de covariance singulière, car son rang ne peut pas excéder k, qui est le rang de $\mathcal{I}(\boldsymbol{\theta}_0)$.

En inversant analytiquement la matrice partitionnée et en multipliant ensuite les deux facteurs du membre de droite de (8.74), il est possible d'obtenir

assez facilement, bien que cela soit quelque peu ennuyeux, les expressions de $n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ et de $n^{-1/2}\tilde{\boldsymbol{\lambda}}$. Celles-ci sont

$$n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{=} -\mathcal{H}_0^{-1}\big(\mathbf{I} - \boldsymbol{R}_0^\top(\boldsymbol{R}_0\mathcal{H}_0^{-1}\boldsymbol{R}_0^\top)^{-1}\boldsymbol{R}_0\mathcal{H}_0^{-1}\big)\big(n^{-1/2}\boldsymbol{g}(\boldsymbol{\theta}_0)\big)$$

et

$$n^{-1/2}\tilde{\boldsymbol{\lambda}} \stackrel{a}{=} \big(\boldsymbol{R}_0\mathcal{H}_0^{-1}\boldsymbol{R}_0^\top\big)^{-1}\boldsymbol{R}_0\mathcal{H}_0^{-1}\big(n^{-1/2}\boldsymbol{g}(\boldsymbol{\theta}_0)\big).$$

À partir de la seconde de ces expressions, de la normalité asymptotique de $n^{-1/2}\boldsymbol{g}(\boldsymbol{\theta}_0)$, et de l'égalité de la matrice d'information, il est facile de voir que

$$n^{-1/2}\tilde{\boldsymbol{\lambda}} \stackrel{a}{\sim} N\big(\mathbf{0},\, (\boldsymbol{R}_0\,\mathcal{I}_0^{-1}\boldsymbol{R}_0^\top)^{-1}\big). \tag{8.75}$$

Maintenant, il est simple de dériver le test du multiplicateur de Lagrange sous sa forme LM. La statistique de test est simplement une forme quadratique du vecteur de dimension r $n^{-1/2}\tilde{\boldsymbol{\lambda}}$:

$$\big(n^{-1/2}\tilde{\boldsymbol{\lambda}}\big)^\top\big(\tilde{\boldsymbol{R}}\tilde{\mathcal{I}}^{-1}\tilde{\boldsymbol{R}}^\top\big)\big(n^{-1/2}\tilde{\boldsymbol{\lambda}}\big) = \frac{1}{n}\tilde{\boldsymbol{\lambda}}^\top\tilde{\boldsymbol{R}}\tilde{\mathcal{I}}^{-1}\tilde{\boldsymbol{R}}^\top\tilde{\boldsymbol{\lambda}}. \tag{8.76}$$

Ici, $\tilde{\mathcal{I}}$ peut être n'importe quelle matrice qui utilise les estimations contraintes $\tilde{\boldsymbol{\theta}}$ pour estimer $\mathcal{I}(\boldsymbol{\theta}_0)$ de manière convergente. Différentes variantes de la statistique LM utiliseront différentes estimations de $\mathcal{I}(\boldsymbol{\theta}_0)$. Il est évident à partir de (8.75), que sous les conditions de régularité standards cette statistique de test sera asymptotiquement distribuée suivant une $\chi^2(r)$ sous l'hypothèse nulle.

La statistique LM (8.76) est numériquement égale à un test basé sur le vecteur score $\boldsymbol{g}(\tilde{\boldsymbol{\theta}})$. Du premier ensemble des conditions du premier ordre (8.72), $\boldsymbol{g}(\tilde{\boldsymbol{\theta}}) = \boldsymbol{R}^\top\tilde{\boldsymbol{\lambda}}$. Si l'on substitue $\boldsymbol{g}(\tilde{\boldsymbol{\theta}})$ à $\boldsymbol{R}^\top\tilde{\boldsymbol{\lambda}}$ dans (8.76) nous aboutissons à la forme score du test LM,

$$\frac{1}{n}\tilde{\boldsymbol{g}}^\top\tilde{\mathcal{I}}^{-1}\tilde{\boldsymbol{g}}. \tag{8.77}$$

Dans la pratique, cette forme score est souvent plus utile que la forme LM parce que, comme les estimations contraintes sont rarement obtenues via un Lagrangien, $\tilde{\boldsymbol{g}}$ est généralement facilement disponible alors que typiquement $\tilde{\boldsymbol{\lambda}}$ ne l'est pas. Cependant, la construction du test via les multiplicateurs de Lagrange est révélatrice, car elle montre clairement la provenance des r degrés de liberté.

Le troisième des trois tests classiques est le **test de Wald**. Ce test est très facile à dériver. Il consiste à savoir si le vecteur des contraintes, évaluées à l'aide des estimations non contraintes est suffisamment proche du vecteur nul pour que les contraintes soient plausibles. Dans le cas des contraintes (8.71), le test de Wald est basé sur le vecteur $\boldsymbol{r}(\hat{\boldsymbol{\theta}})$, qui devrait tendre asymptotiquement vers un vecteur nul si les contraintes sont valables. Comme nous l'avons vu dans les Sections 8.5 et 8.6,

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{\sim} N\big(\mathbf{0},\, \mathcal{I}^{-1}(\boldsymbol{\theta}_0)\big).$$

Un développement en série de Taylor de $\boldsymbol{r}(\hat{\boldsymbol{\theta}})$ autour de $\boldsymbol{\theta}_0$ donne $\boldsymbol{r}(\hat{\boldsymbol{\theta}}) \cong \boldsymbol{R}_0(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. Ainsi,

$$\boldsymbol{V}\big(n^{1/2}\boldsymbol{r}(\hat{\boldsymbol{\theta}})\big) \stackrel{a}{=} \boldsymbol{R}_0\, \mathcal{I}_0^{-1}\boldsymbol{R}_0^\top.$$

Il s'ensuit qu'une statistique de test appropriée est

$$n\boldsymbol{r}^\top(\hat{\boldsymbol{\theta}})\big(\hat{\boldsymbol{R}}\hat{\mathcal{I}}^{-1}\hat{\boldsymbol{R}}^\top\big)^{-1}\boldsymbol{r}(\hat{\boldsymbol{\theta}}), \tag{8.78}$$

où $\hat{\mathcal{I}}$ désigne n'importe quelle estimation de $\mathcal{I}(\boldsymbol{\theta}_0)$ basée sur les estimations non contraintes $\hat{\boldsymbol{\theta}}$. Différentes variantes du test de Wald utiliseront différentes estimations de $\mathcal{I}(\boldsymbol{\theta}_0)$. Il est facile de voir qu'étant données les conditions de régularité adéquates, la statistique de test (8.78) sera asymptotiquement distribuée suivant une $\chi^2(r)$ sous l'hypothèse nulle.

La propriété fondamentale des trois statistiques des test classiques est que sous l'hypothèse nulle, quand $n \to \infty$, elles tendent toutes vers la même variable aléatoire, qui est distribuée suivant une $\chi^2(r)$. Nous prouverons ce résultat au cours du Chapitre 13. La conséquence est que, avec de grands échantillons, le choix parmi les trois importe peu. Si à la fois $\hat{\boldsymbol{\theta}}$ et $\tilde{\boldsymbol{\theta}}$ sont faciles à calculer, il est intéressant d'utiliser le test LR. Si $\tilde{\boldsymbol{\theta}}$ est facile à calculer mais que $\hat{\boldsymbol{\theta}}$ ne l'est pas, comme cela est souvent le cas pour les tests de spécification de modèle, alors le test LM devient attrayant. Si d'un autre côté $\hat{\boldsymbol{\theta}}$ est facile à calculer mais $\tilde{\boldsymbol{\theta}}$ ne l'est pas, comme cela peut être le cas quand nous sommes intéressés par les contraintes non linéaires imposées à un modèle linéaire, alors le test de Wald devient attrayant. Quand la taille de l'échantillon n'est pas grande, un choix pertinent parmi les trois tests est compliqué par le fait qu'ils peuvent avoir des propriétés avec des échantillons finis très différentes, qui peuvent par la suite différer formidablement selon les variantes alternatives des tests LM et Wald. Ceci rend le choix des tests plutôt plus compliqué en pratique que ce que la théorie asymptotique ne le suggère.

8.10 Les Modèles de Régression Non Linéaire

Dans cette section, nous discutons des possibilités de l'usage de la méthode du maximum de vraisemblance pour l'estimation des modèles de régression univarié non linéaire. Quand les aléas sont supposés être normalement et indépendamment distribués avec une variance constante, l'estimation ML de ces modèles est, du moins en ce qui concerne l'estimation des paramètres de la fonction de régression, numériquement identique à l'estimation NLS. L'exercice présente néanmoins un intérêt. Tout d'abord, il fournit une illustration concrète de la manière d'utiliser la méthode du maximum de vraisemblance. Deuxièmement, il fournit une matrice de covariance asymptotique pour les estimations de $\boldsymbol{\beta}$ et σ conjointement, alors que les NLS ne la calculent que pour les estimations de $\boldsymbol{\beta}$. Finalement, en considérant certaines extensions du modèle de régression normal, nous sommes capables de démontrer la puissance de l'estimation ML.

La classe des modèles que nous considérerons est

$$\boldsymbol{y} = \boldsymbol{x}(\boldsymbol{\beta}) + \boldsymbol{u}, \quad \boldsymbol{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I}), \tag{8.79}$$

où la fonction de régression $\boldsymbol{x}(\boldsymbol{\beta})$ satisfait les conditions pour les Théorèmes 5.1 et 5.2, et les données sont supposées avoir été générées par un cas particulier de (8.79). Le vecteur paramétrique $\boldsymbol{\beta}$ est supposé être de longueur k, ce qui implique qu'il y a $k+1$ paramètres à estimer. La notation "$\boldsymbol{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$" signifie que le vecteur des aléas $\boldsymbol{u}$ est supposé être distribué suivant une loi normale multivariée de vecteur d'espérance zéro et de matrice de covariance $\sigma^2\mathbf{I}$. Ainsi, les aléas individuels u_t sont indépendants, chacun étant distribué suivant la $N(0, \sigma^2)$. La fonction de densité de u_t est

$$f(u_t) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{u_t^2}{2\sigma^2}\right).$$

Afin de construire la fonction de vraisemblance, nous avons besoin de la fonction de densité de y_t plutôt que de celle de u_t. Ceci nous demande d'utiliser un résultat standard en statistique qui est établi dans l'Annexe B.

Le résultat en question indique que si une variable aléatoire x_1 a une fonction de densité $f_1(x_1)$ et si une autre variable aléatoire x_2 lui est reliée par

$$x_1 = h(x_2),$$

où la fonction $h(\cdot)$ est monotone et continûment différentiable, alors la fonction de densité de x_2 est donnée par

$$f_2(x_2) = f_1\big(h(x_2)\big) \left|\frac{\partial h(x_2)}{\partial x_2}\right|.$$

Ici, le second facteur est la valeur absolue du Jacobien de la transformation. Dans de nombreux cas, comme nous le verrons plus tard, sa présence fait apparaître les **termes Jacobiens** dans les fonctions de logvraisemblance. Cependant, dans ce cas, la fonction qui relie u_t à y_t est

$$u_t = y_t - x_t(\boldsymbol{\beta}).$$

Le facteur Jacobien $|\partial u_t/\partial y_t|$ est alors égal à l'unité. Ainsi, nous concluons que la fonction de densité de y_t est

$$\frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{\big(y_t - x_t(\boldsymbol{\beta})\big)^2}{2\sigma^2}\right). \tag{8.80}$$

La contribution à la fonction de logvraisemblance apportée par la $t^{\text{ième}}$ observation est le logarithme de (8.80),

$$\ell_t(y_t, \boldsymbol{\beta}, \sigma) = -\tfrac{1}{2}\log(2\pi) - \log(\sigma) - \frac{1}{2\sigma^2}\big(y_t - x_t(\boldsymbol{\beta})\big)^2.$$

Comme toutes les informations sont indépendantes, la fonction de logvraisemblance elle-même correspond précisément à la somme des contributions $\ell_t(y_t, \boldsymbol{\beta}, \sigma)$ sur tout t, ou

$$\begin{aligned}\ell(\boldsymbol{y}, \boldsymbol{\beta}, \sigma) &= -\frac{n}{2}\log(2\pi) - n\,\log(\sigma) - \frac{1}{2\sigma^2}\sum_{t=1}^{n}\big(y_t - x_t(\boldsymbol{\beta})\big)^2 \\ &= -\frac{n}{2}\log(2\pi) - n\,\log(\sigma) - \frac{1}{2\sigma^2}\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big)^\top\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big).\end{aligned} \tag{8.81}$$

La première étape dans la maximisation de $\ell(\boldsymbol{y}, \boldsymbol{\beta}, \sigma)$ consiste à la concentrer par rapport à σ, comme cela fut expliqué dans la Section 8.7. La différentiation de la seconde ligne de (8.81) par rapport à σ et l'égalisation de la dérivée à zéro donnent

$$\frac{\partial \ell(\boldsymbol{y}, \boldsymbol{\beta}, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big)^\top\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big) = 0,$$

et la résolution de cette équation produit le résultat

$$\hat{\sigma}(\boldsymbol{\beta}) = \Big(\frac{1}{n}\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big)^\top\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big)\Big)^{1/2}.$$

Ici la notation $\hat{\sigma}(\boldsymbol{\beta})$ signifie que l'estimation ML de σ est maintenant une fonction de $\boldsymbol{\beta}$. Notons que nous avons divisé par n plutôt que par $n - k$. Si nous pouvions évaluer $\hat{\sigma}^2(\boldsymbol{\beta})$ à la véritable valeur $\boldsymbol{\beta}_0$, nous obtiendrions une estimation non biaisée de σ^2. Cependant, nous l'évaluons en fait à l'estimation ML $\hat{\boldsymbol{\beta}}$, qui, comme nous le voyons, est égale à l'estimation NLS. Ainsi, comme nous l'avons vu dans la Section 3.2, $\hat{\sigma}^2$ doit être biaisée vers le bas en tant qu'estimateur de σ^2.

La substitution de $\hat{\sigma}(\boldsymbol{\beta})$ dans la seconde ligne de (8.81) permet de construire la fonction de logvraisemblance concentrée

$$\begin{aligned}\ell^c(\boldsymbol{y}, \boldsymbol{\beta}) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\Big(\frac{1}{n}\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big)^\top\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big)\Big) - \frac{n}{2} \\ &= C - \frac{n}{2}\log\Big(\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big)^\top\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big)\Big),\end{aligned} \tag{8.82}$$

où C est un terme constant. Le second terme dans (8.82) est moins $n/2$ fois le logarithme de la somme des résidus au carré. Ainsi, nous voyons que *maximiser* la fonction de logvraisemblance concentrée est équivalent à *minimiser* $SSR(\boldsymbol{\beta})$. Les estimations ML $\hat{\boldsymbol{\beta}}$ seront simplement les estimations NLS avec lesquelles nous sommes déjà familiers.

Le terme constant dans (8.82) est en fait

$$\frac{n}{2}\big(\log(n) - 1 - \log(2\pi)\big).$$

Comme cette expression ne dépend pas de $\boldsymbol{\beta}$, elle peut être ignorée dans toutes les utilisations sauf en fait pour le calcul de la valeur de $\ell(\boldsymbol{y}, \boldsymbol{\beta}, \sigma)$. De telles constantes sont souvent complètement ignorées dans un travail théorique et sont même parfois ignorées par des programmes informatiques, et le résultat de tout ceci est que les valeurs des fonctions de logvraisemblance pour un même modèle et un même ensemble de données reportées par différents programmes peuvent parfois différer.

Le fait que l'estimateur ML $\hat{\boldsymbol{\beta}}$ pour la classe des modèles (8.79) corresponde exactement à l'estimateur NLS comporte une importante implication. Comme nous l'avons vu dans la Section 8.8, les estimateurs ML sont asymptotiquement efficaces. Ainsi, l'estimateur NLS sera asymptotiquement efficace à chaque fois que les aléas sont normalement et indépendamment distribués avec une variance constante. Cependant, si les aléas ont une quelqu'autre distribution connue, l'estimateur ML diffèrera en général de celui des NLS et sera plus efficace que ce dernier (voir plus loin pour un exemple extrême). Ainsi, bien que l'estimateur NLS soit convergent sous de très faibles conditions sur la distribution des aléas, comme nous l'avons vu dans la Section 5.3, et soit efficace dans la classe des estimateurs asymptotiquement linéaires qui sont applicables sous ces conditions peu restrictives, il ne coïncide avec l'estimateur ML efficace que si les aléas sont supposés être normalement distribués. La signification de tout ceci est la suivante. Si la seule hypothèse que l'on veut formuler concernant les aléas est qu'ils satisfassent les conditions de régularité pour les NLS, alors l'estimateur NLS est asymptotiquement efficace dans la classe des estimateurs asymptotiquement linéaires et convergents des paramètres de la fonction de régression. Cependant, si l'on est prêt à fournir l'effort de spécifier la véritable distribution des aléas, alors l'estimateur ML sera en général plus efficace, *à condition que la spécification présumée des aléas soit correcte*. L'estimateur ML ne sera pas plus efficace dans le cas où les aléas sont supposés être normaux, puisqu'alors les estimateurs ML et NLS seront équivalents.

Dans la Section 8.6, nous avons vu que si $\hat{\boldsymbol{\theta}}$ est un vecteur d'estimations ML, alors le vecteur $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ est asymptotiquement normalement distribué avec un vecteur d'espérance zéro et une matrice de covariance égale à l'inverse de la matrice d'information asymptotique $\mathcal{I}(\boldsymbol{\theta}_0)$. Ce résultat signifie qu'il est presque toujours intéressant de calculer $\mathcal{I}(\boldsymbol{\theta})$ pour n'importe quel modèle qui est estimé par maximum de vraisemblance. Nous avons vu qu'il y a en général deux manières de procéder. L'une consiste à trouver l'opposée de la limite en probabilité de n^{-1} fois la matrice Hessienne, et l'autre consiste à trouver la limite en probabilité de n^{-1} fois $\boldsymbol{G}^\top(\boldsymbol{\theta})\boldsymbol{G}(\boldsymbol{\theta})$, où $\boldsymbol{G}(\boldsymbol{\theta})$ est la matrice CG. Ces deux méthodes entraîneront la même réponse, s'il est tout à fait faisable de calculer $\mathcal{I}(\boldsymbol{\theta})$, bien qu'une approche puisse être plus facile que l'autre dans certaines situations données.

Pour le modèle de régression non linéaire (8.79), le vecteur paramétrique $\boldsymbol{\theta}$ est le vecteur $[\boldsymbol{\beta} \vdots \sigma]$. Nous calculons à présent la matrice d'information

asymptotique $\mathcal{I}(\boldsymbol{\beta}, \sigma)$ pour ce modèle en utilisant la seconde méthode, basée sur la matrice CG, qui ne nécessite que les dérivées premières. Il s'agit d'un bon exercice que de répéter la construction en utilisant la matrice Hessienne, qui nécessite les dérivées secondes, et de vérifier que cela produit les mêmes résultats. La dérivée première de $\ell_t(y_t, \boldsymbol{\beta}, \sigma)$ par rapport à β_i est

$$\frac{\partial \ell_t}{\partial \beta_i} = \frac{1}{\sigma^2}\big(y_t - x_t(\boldsymbol{\beta})\big)X_{ti}(\boldsymbol{\beta}) = \frac{1}{\sigma^2}\, e_t(\boldsymbol{\beta})X_{ti}(\boldsymbol{\beta}), \tag{8.83}$$

où $e_t(\boldsymbol{\beta}) \equiv y_t - x_t(\boldsymbol{\beta})$ et, comme d'habitude, $X_{ti}(\boldsymbol{\beta}) \equiv \partial x_t(\boldsymbol{\beta})/\partial \beta_i$. La dérivée première de $\ell_t(y_t, \boldsymbol{\beta}, \sigma)$ par rapport à σ est

$$\frac{\partial \ell_t}{\partial \sigma} = -\frac{1}{\sigma} + \frac{\big(y_t - x_t(\boldsymbol{\beta})\big)^2}{\sigma^3} = -\frac{1}{\sigma} + \frac{e_t^2(\boldsymbol{\beta})}{\sigma^3}. \tag{8.84}$$

Les expressions (8.83) et (8.84) sont tout ce dont nous avons besoin pour calculer la matrice d'information en utilisant la matrice CG. La colonne de cette matrice qui correspond à σ aura l'élément type (8.84), tandis que les k colonnes restantes, qui correspondent aux β_i, auront l'élément type (8.83).

L'élément de $\mathcal{I}(\boldsymbol{\beta}, \sigma)$ correspondant à β_i et β_j est

$$\mathcal{I}(\beta_i, \beta_j) = \operatorname*{plim}_{n\to\infty}\left(\frac{1}{n}\sum_{t=1}^{n}\frac{e_t^2(\boldsymbol{\beta})}{\sigma^4}X_{ti}(\boldsymbol{\beta})X_{tj}(\boldsymbol{\beta})\right).$$

Comme $e_t^2(\boldsymbol{\beta})$ a une espérance de σ^2 sous le DGP caractérisé par $(\boldsymbol{\beta}, \sigma)$ et est indépendant de $\boldsymbol{X}(\boldsymbol{\beta})$, nous pouvons le remplacer ici par σ^2 pour obtenir

$$\mathcal{I}(\beta_i, \beta_j) = \operatorname*{plim}_{n\to\infty}\left(\frac{1}{n}\sum_{t=1}^{n}\frac{1}{\sigma^2}X_{ti}(\boldsymbol{\beta})X_{tj}(\boldsymbol{\beta})\right).$$

Ainsi, nous voyons que le bloc entier $(\boldsymbol{\beta}, \boldsymbol{\beta})$ de la matrice d'information asymptotique est

$$\frac{1}{\sigma^2}\operatorname*{plim}_{n\to\infty}\left(\frac{1}{n}\boldsymbol{X}^\top(\boldsymbol{\beta})\boldsymbol{X}(\boldsymbol{\beta})\right). \tag{8.85}$$

L'élément de $\mathcal{I}(\boldsymbol{\beta}, \sigma)$ correspondant à σ est

$$\begin{aligned}\mathcal{I}(\sigma, \sigma) &= \operatorname*{plim}_{n\to\infty}\left(\frac{1}{n}\sum_{t=1}^{n}\left(\frac{1}{\sigma^2} + \frac{e_t^4(\boldsymbol{\beta})}{\sigma^6} - \frac{2e_t^2(\boldsymbol{\beta})}{\sigma^4}\right)\right) \\ &= \frac{1}{n}\left(\frac{n}{\sigma^2} + \frac{3n\sigma^4}{\sigma^6} - \frac{2n\sigma^2}{\sigma^4}\right) \\ &= \frac{2}{\sigma^2}.\end{aligned} \tag{8.86}$$

Ici, nous avons utilisé les faits que, sous le DGP caractérisé par $(\boldsymbol{\beta}, \sigma)$, $E\big(e_t^2(\boldsymbol{\beta})\big) = \sigma^2$ et $E\big(e_t^4(\boldsymbol{\beta})\big) = 3\sigma^4$, la dernière égalité étant une propriété bien connue de la distribution normale (consulter la Section 2.6 et l'Annexe B).

Finalement, l'élément de $\mathcal{I}(\boldsymbol{\beta}, \sigma)$ correspondant à β_i et σ est

$$\begin{aligned}\mathcal{I}(\beta_i, \sigma) &= \operatorname*{plim}_{n\to\infty}\left(\frac{1}{n}\sum_{t=1}^{n}\left(-\frac{e_t(\boldsymbol{\beta})X_{ti}(\boldsymbol{\beta})}{\sigma^3} + \frac{e_t^3(\boldsymbol{\beta})X_{ti}(\boldsymbol{\beta})}{\sigma^5}\right)\right) \\ &= 0.\end{aligned} \tag{8.87}$$

Les éléments sont nuls parce que, sous le DGP caractérisé par $(\boldsymbol{\beta}, \sigma)$, $e_t(\boldsymbol{\beta})$ est indépendant de $\boldsymbol{X}(\boldsymbol{\beta})$, et le fait que les aléas soient normalement distribués implique que $E\big(e_t(\boldsymbol{\beta})\big) = E\big(e_t^3(\boldsymbol{\beta})\big) = 0$.

En collectant les résultats (8.85), (8.86), et (8.87), nous concluons que

$$\mathcal{I}(\boldsymbol{\beta}, \sigma) = \frac{1}{\sigma^2}\begin{bmatrix}\operatorname{plim}\big(n^{-1}\boldsymbol{X}^\top(\boldsymbol{\beta})\boldsymbol{X}(\boldsymbol{\beta})\big) & \mathbf{0} \\ \mathbf{0}^\top & 2\end{bmatrix}. \tag{8.88}$$

Nos résultats sur la distribution asymptotique des estimateurs ML (Sections 8.5 et 8.6) nous permettent de conclure que

$$\begin{bmatrix}n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\ n^{1/2}(\hat{\sigma} - \sigma_0)\end{bmatrix} \overset{a}{\sim} N\left(\mathbf{0},\ \begin{bmatrix}\sigma_0^2 \operatorname{plim}\big(n^{-1}\boldsymbol{X}_0^\top\boldsymbol{X}_0\big)^{-1} & \mathbf{0} \\ \mathbf{0}^\top & \sigma_0^2/2\end{bmatrix}\right), \tag{8.89}$$

où $\boldsymbol{\beta}_0$ et σ_0 désignent les valeurs de $\boldsymbol{\beta}$ et σ sous le DGP, et $\boldsymbol{X}_0$ désigne $\boldsymbol{X}(\boldsymbol{\beta}_0)$. Parce que la matrice d'information (8.88) est bloc-diagonale entre le bloc $(\boldsymbol{\beta}, \boldsymbol{\beta})$ et le bloc (σ, σ) (qui est un scalaire), son inverse est simplement la matrice qui se compose de chaque bloc inversé séparément. Comme nous le verrons dans le Chapitre 9, ce type de bloc-diagonalité est une propriété très importante des modèles de régression avec erreurs normales.

A partir de (8.89), nous voyons que la matrice de covariance de $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ est la même matrice de covariance asymptotique préalablement établie pour les estimations NLS des paramètres d'une fonction de régression, ce qui n'est pas surprenant car $\hat{\boldsymbol{\beta}}$ est simplement un vecteur d'estimations NLS. Mais ici nous l'avons dérivée comme un cas particulier des résultats généraux de la Section 8.6 sur la distribution asymptotique des estimateurs ML. Le résultat selon lequel la variance asymptotique de $n^{1/2}(\hat{\sigma} - \sigma_0)$ est $\sigma_0^2/2$ est nouveau. Comme nous l'avons vu dans le Chapitre 5, la méthode des moindres carrés non linéaires ne produit pas directement une estimation de σ bien qu'il soit facile d'en construire plusieurs estimations, une fois que le vecteur $\hat{\boldsymbol{\beta}}$ a été obtenu. La méthode du maximum de vraisemblance, couplée avec l'hypothèse de normalité, produit directement une estimation de σ et aussi une mesure de la variabilité de cette estimation. Cependant, cette dernière n'est en général

valide que sous l'hypothèse de normalité. De plus, comme nous en avons discuté plus tôt, l'estimation ML $\hat{\sigma}^2 = n^{-1}SSR(\hat{\boldsymbol{\beta}})$ est biaisée vers le bas, et en pratique il peut alors être préférable d'utiliser $s^2 = (n-k)^{-1}SSR(\hat{\boldsymbol{\beta}})$.

Dans la dérivation de (8.88) et (8.89), nous avons choisi d'écrire la matrice d'information en termes de $\boldsymbol{\beta}$ et de σ. De nombreux auteurs choisissent de l'écrire en termes de $\boldsymbol{\beta}$ et de σ^2. Le résultat équivalent à (8.89) dans cette paramétrisation alternative est

$$\begin{bmatrix} n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\ n^{1/2}(\hat{\sigma}^2 - \sigma_0^2) \end{bmatrix} \overset{a}{\sim} N\left(\mathbf{0}, \begin{bmatrix} \sigma_0^2 \operatorname{plim}\left(n^{-1}\boldsymbol{X}_0^\top \boldsymbol{X}_0\right)^{-1} & \mathbf{0} \\ \mathbf{0}^\top & 2\sigma_0^4 \end{bmatrix}\right). \tag{8.90}$$

Ce résultat et (8.89) sont tous deux corrects. Cependant, avec n'importe quel échantillon fini, l'intervalle de confiance pour σ basé sur (8.89) sera différent de l'intervalle de confiance basé sur (8.90). Comme nous en discuterons dans le Chapitre 13, le premier intervalle de confiance sera généralement plus précis, parce que la distribution de $n^{1/2}(\hat{\sigma} - \sigma_0)$ sera plus proche de la distribution normale avec des échantillons finis que celle de $n^{1/2}(\hat{\sigma}^2 - \sigma_0^2)$. Il est alors préférable de paramétriser le modèle en termes de σ plutôt que de σ^2.

Dans la pratique, naturellement, nous sommes intéressés par $\hat{\boldsymbol{\beta}}$ et $\hat{\sigma}$ plutôt que par $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ et $n^{1/2}(\hat{\sigma} - \sigma_0)$. Ainsi, au lieu d'utiliser (8.88), nous devrions en fait réaliser des inférences basées sur la matrice de covariance estimée

$$\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}, \hat{\sigma}) = \begin{bmatrix} \hat{\sigma}^2(\hat{\boldsymbol{X}}^\top \hat{\boldsymbol{X}})^{-1} & \mathbf{0} \\ \mathbf{0}^\top & \hat{\sigma}^2/2n \end{bmatrix},$$

dont le bloc supérieur gauche de dimension $k \times k$ est l'estimateur NLS habituel de la matrice de covariance pour $\hat{\boldsymbol{\beta}}$.

Dans la Section 8.1, nous avons considéré un exemple simple, (8.01), qui ne pouvait pas être estimé par moindres carrés. Si nous formulons l'hypothèse additionnelle que les aléas sont normalement distribués, ce modèle devient

$$y_t^\gamma = \beta_0 + \beta_1 x_t + u_t, \quad u_t \sim \text{NID}(0, \sigma^2), \tag{8.91}$$

qui ressemble presque à un modèle de régression, excepté que la variable dépendante est soumise à une transformation non linéaire.

La fonction de logvraisemblance correspondant à (8.91) est

$$\begin{aligned} \ell(\boldsymbol{\beta}, \gamma, \sigma) = &-\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{t=1}^{n}\left(y_t^\gamma - \beta_0 - \beta_1 x_t\right)^2 \\ &+ n\log|\gamma| + (\gamma - 1)\sum_{t=1}^{n}\log(y_t). \end{aligned} \tag{8.92}$$

Les trois premiers termes constituent exactement la fonction de logvraisemblance que nous obtiendrions si nous traitions y_t^γ comme la variable dépendante. Les quatrième et cinquième termes ne représentent en fait qu'un seul

terme, un terme Jacobien. Ce terme apparaît parce que $\partial u_t/\partial y_t = \gamma y_t^{\gamma-1}$. Par conséquent la contribution à la fonction de vraisemblance apportée par observation t doit inclure le facteur Jacobien $|\gamma y_t^{\gamma-1}|$, qui est la valeur absolue de $\partial u_t/\partial y_t$. En sommant sur tous les t et opérant le logarithme nous obtenons le terme qui apparaît dans (8.92).

En concentrant la fonction de logvraisemblance par rapport à σ nous aboutissons à

$$\begin{aligned}\ell^c(\boldsymbol{\beta},\gamma) = C - n\log\Bigg(\sum_{t=1}^{n}\big(y_t^{\gamma} - \beta_0 - \beta_1 x_t\big)^2\Bigg) \\ + n\log|\gamma| + (\gamma-1)\sum_{t=1}^{n}\log(y_t).\end{aligned} \tag{8.93}$$

La maximisation de cette quantité par rapport à γ et $\boldsymbol{\beta}$ est simple. Si un programme d'optimisation non linéaire convenable n'est pas disponible, on peut simplement faire une recherche à une dimension sur γ, en calculant β_0 et β_1 conditionnels à γ à l'aide des moindres carrés, afin de trouver la valeur $\hat{\gamma}$ qui maximise (8.93). Naturellement, on ne peut pas utiliser la matrice de covariance OLS obtenue de cette manière, car elle traite l'estimation $\hat{\gamma}$ comme fixée. La matrice d'information *n'est pas* bloc-diagonale entre $\boldsymbol{\beta}$ et les autres paramètres de (8.91), aussi doit-on calculer et inverser la matrice d'information entière pour obtenir une matrice de covariance estimée.

L'estimation ML s'applique dans ce cas à cause du terme Jacobien qui apparaît dans (8.92) et (8.93). Il disparaît quand $\gamma = 1$ mais joue un rôle extrêmement important pour toutes les autres valeurs de γ. Nous avons vu dans la Section 8.1 que si l'on appliquait les NLS à (8.01) et si tous les y_t étaient supérieurs à l'unité, on aboutirait à une estimation de γ infiniment grande et négative. Cela n'arrivera pas si l'on utilise le maximum de vraisemblance, parce que le terme $(\gamma-1)\sum_{t=1}^{n}\log(y_t)$ ne tendra pas vers moins l'infini quand $\gamma \to \infty$ beaucoup plus vite que le logarithme du terme de la somme des carrés ne tend vers plus l'infini. Cet exemple illustre l'utilité de l'estimation ML pour traiter des modèles de régression modifiés dans lesquels la variable dépendante est soumise à une transformation. Nous rencontrerons d'autres problèmes de ce type dans le Chapitre 14.

L'estimation ML peut aussi être très utile lorsque l'on croit que les aléas sont non normaux. Comme exemple extrême, considérons le modèle suivant:

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + \alpha\varepsilon_t, \quad f(\varepsilon_t) = \frac{1}{\pi(1+\varepsilon_t^2)}, \tag{8.94}$$

où $\boldsymbol{\beta}$ est un vecteur de dimension k et $\boldsymbol{X}_t$ est la $t^{\text{ième}}$ ligne d'une matrice de dimension $n \times k$. La densité de ε_t est ici la densité de Cauchy (consulter la Section 4.6) et ε_t n'a donc pas de moments finis. Le paramètre α est

simplement un paramètre d'échelle, et *non pas* l'écart type des aléas; comme la distribution de Cauchy n'a pas de moments, les aléas *n'ont pas* d'écart type.

Si nous écrivons ε_t comme une fonction de y_t, nous trouvons que

$$\varepsilon_t = \frac{y_t - \boldsymbol{X}_t\boldsymbol{\beta}}{\alpha}.$$

Ainsi, la densité de y_t est

$$f(y_t) = \frac{1}{\pi\alpha}\left(1 + \frac{(y_t - \boldsymbol{X}_t\boldsymbol{\beta})^2}{\alpha^2}\right)^{-1},$$

le facteur $1/\alpha$ étant un facteur Jacobien. La contribution à la fonction de logvraisemblance de la $t^{\text{ième}}$ observation est ainsi

$$-\log(\pi) - \log(\alpha) - \log\left(1 + \frac{(y_t - \boldsymbol{X}_t\boldsymbol{\beta})^2}{\alpha^2}\right),$$

et la fonction de logvraisemblance elle-même est

$$\ell(\boldsymbol{\beta}, \alpha) = -n\log(\pi) - n\log(\alpha) - \sum_{t=1}^{n}\log\left(1 + \frac{(y_t - \boldsymbol{X}_t\boldsymbol{\beta})^2}{\alpha^2}\right). \tag{8.95}$$

Les conditions du premier ordre pour $\hat{\beta}_i$ peuvent être écrites comme

$$-2\hat{\alpha}^{-2}\sum_{t=1}^{n}\left(1 + \frac{(y_t - \boldsymbol{X}_t\hat{\boldsymbol{\beta}})^2}{\hat{\alpha}^2}\right)^{-1}(y_t - \boldsymbol{X}_t\hat{\boldsymbol{\beta}})X_{ti} = 0. \tag{8.96}$$

L'expression équivalente pour l'estimation ML avec des erreurs normales (c'est-à-dire OLS) est

$$-\hat{\sigma}^{-2}\sum_{t=1}^{n}(y_t - \boldsymbol{X}_t\hat{\boldsymbol{\beta}})X_{ti} = 0. \tag{8.97}$$

La différence entre les équations de vraisemblance (8.96) et (8.97) est frappante. La dernière indique qu'une somme *non pondérée* des résidus fois chacun des régresseurs doit être égale à zéro. La première indique qu'une somme *pondérée* des mêmes quantités doit être égale à zéro, avec des poids inversement reliés à la taille des résidus. La raison de ceci est que la distribution de Cauchy génère de nombreuses valeurs extrêmes. Il y aura en général de nombreux aléas très importants, et afin d'éviter qu'ils n'influencent trop les estimations, la procédure ML d'estimation de $\hat{\boldsymbol{\beta}}$ leur attribue beaucoup moins de poids que ne le font les OLS. Ces estimations ML possèdent toutes les propriétés habituelles de convergence, de normalité asymptotique, et ainsi de suite. Par contraste, si l'on appliquait simplement les OLS au modèle

(8.94), les aléas extrêmement grands fréquemment générés par la distribution de Cauchy feraient en sorte que les estimations ne soient même pas convergentes. Le théorème de convergence habituel pour les moindres carrés ne s'applique pas ici parce que les ε_t n'ont pas de moments finis.

Parce que les équations de vraisemblance (8.96) dépendent des résidus, la valeur $\hat{\alpha}$ affecte la valeur $\hat{\boldsymbol{\beta}}$ qui les résoud. Ainsi, il est nécessaire de les résoudre conjointement pour $\hat{\boldsymbol{\beta}}$ et $\hat{\alpha}$. Malheureusement, il existe en général de multiples solutions à ces équations; voir Reeds (1985). Ainsi, une grande quantité d'efforts doit être consacrée à localiser le maximum global de la fonction de logvraisemblance (8.95).

8.11 Conclusion

Ce chapitre a fourni une introduction à toutes les caractéristiques majeures de l'estimation par maximum de vraisemblance et des tests de spécification, que nous utiliserons à travers le reste de ce livre. Le Chapitre 9 de Cox et Hinkley (1974) fournit un traitement plus détaillé sur certains des sujets que nous avons couverts. Une autre référence utile est Rothenberg (1973). Dans les deux prochains chapitres, nous utiliserons certains résultats de ce chapitre, avec les résultats antérieurs des estimateurs NLS et IV, pour traiter des sujets variés qui préoccupent les économètres. Le Chapitre 9 traite de la méthode des moindres carrés généralisés que l'on considère à la fois comme un exemple d'estimation ML et comme une extension des moindres carrés. Le Chapitre 10 traite ensuite du sujet très important de corrélation en série. Le Chapitre 13 fournira un traitement beaucoup plus détaillé sur les trois statistiques de test classiques que ne le fit la Section 8.9 et introduira une régression artificielle, comparable à la régression de Gauss-Newton, que l'on pourra utiliser avec des modèles estimés par ML.

Termes et Concepts