

# Chapitre 14

## Transformation de la Variable Dépendante

### 14.1 INTRODUCTION

Quand nous avons introduit le concept d'une fonction de régression dans le Chapitre 2, nous l'avons défini comme la fonction qui détermine la moyenne d'une variable dépendante  $y_t$  conditionnelle à un ensemble d'information  $\Omega_t$ . Avec cette définition, nous pouvons toujours écrire

$$y_t = x_t(\beta) + u_t \quad (14.01)$$

et affirmer que  $u_t$  a une moyenne nulle conditionnelle à  $\Omega_t$ , à condition que  $x_t(\beta)$  ait été correctement spécifiée. Cependant, quelle que soit la façon correcte dont  $x_t(\beta)$  a été spécifiée, nous ne pouvons pas affirmer que  $u_t$  soit i.i.d. ou possède d'autres propriétés souhaitables. En particulier, il n'y existe aucune raison pour que  $u_t$  soit normalement distribué, homoscédastique, ou même symétrique. Cependant nous avons besoin que  $u_t$  soit homoscédastique pour que les estimations NLS  $\hat{\beta}$  soient efficaces et pour que les inférences basées sur l'estimateur habituel des moindres carrés de la matrice de covariance soient valides.<sup>1</sup> Nous avons aussi besoin que  $u_t$  soit symétrique (et normalement distribué de préférence ou proche de l'être) pour que les résultats asymptotiques fournissent une bonne indication sur les propriétés des estimateurs en échantillon fini. De plus, si nous désirons prédire  $y_t$  conditionnelle à  $\Omega_t$  et construire un quelconque intervalle de prévision, nous devons connaître (ou du moins être capable d'estimer) la distribution de  $u_t$ .

Si nous pouvons trouver la moyenne de  $y_t$  conditionnelle à  $\Omega_t$ , alors nous pouvons probablement tout aussi bien trouver la moyenne conditionnelle de n'importe quelle fonction monotone lisse de  $y_t$ , disons  $\tau(y_t)$ . Par exemple,  $\tau(y_t)$  pourrait être  $\log y_t$ ,  $y_t^{1/2}$ , ou  $y_t^2$ . Si nous écrivons

$$\tau(y_t) = E(\tau(y_t) | \Omega_t) + v_t \quad (14.02)$$

<sup>1</sup> Comme nous l'avons vu dans la Section 11.6 et comme nous en discuterons prochainement dans le Chapitre 16, il est possible de réaliser des inférences asymptotiquement valides même en présence d'une forme inconnue d'hétéroscédasticité. Cependant les inférences en échantillon fini seront presque toujours plus précises si les aléas sont homoscédastiques au départ.

pour un quelconque  $\tau(\cdot)$  non linéaire, alors l'aléa  $v_t$  ne peut pas être normalement et indépendamment distribué, ou n.i.d., si  $u_t$  est n.i.d. dans (14.01). Inversement, si  $v_t$  est n.i.d. dans (14.02),  $u_t$  ne peut pas être n.i.d. dans (14.01).

Considérons maintenant un exemple concret, et très réaliste. Supposons que nous estimions le modèle (14.01) quand le DGP pour  $y_t$  est réellement

$$\log y_t = \log(m_t) + v_t, \quad (14.03)$$

où  $m_t$  est dans l'ensemble d'information  $\Omega_t$ , et l'aléa  $v_t$  est  $\text{NID}(0, \sigma^2)$ . Il s'ensuit que

$$y_t = \exp(\log(m_t) + v_t) = m_t \exp(v_t) \cong m_t(1 + v_t) = m_t + m_t v_t,$$

où l'approximation  $\exp(v_t) \cong 1 + v_t$  qui est utilisée ici sera satisfaisante lorsque  $\sigma$  est petit. Si  $m_t = x_t(\beta_0)$  pour un quelconque  $\beta_0$ , la régression non linéaire (14.01) est au moins approximativement valide pour la moyenne conditionnelle de  $y_t$ , bien que ceci ne soit pas nécessairement le cas si la transformation dans (14.03) n'était pas logarithmique. Mais les aléas  $u_t$  qui adhèrent à  $m_t$  ne peuvent pas être n.i.d. En effet, ils seront hétéroscédastiques, avec une variance proportionnelle au carré de  $x_t(\beta_0)$ . Ils seront aussi quelque peu asymétriques à droite, particulièrement si  $\sigma$  n'est pas très petit, parce que le fait que, pour  $a > 0$ ,  $e^a - 1 > |e^{-a} - 1|$ . Ceci implique que toute valeur positive de  $v_t$  se transforme en un  $u_t$  dont la valeur absolue est plus grande que celle apportée par  $-v_t$ . Comme  $v$  est symétrique,  $u$  doit alors être asymétrique à droite.

Cet exemple démontre que, même quand la variable dépendante avait été réellement générée par un DGP à erreurs n.i.d., l'utilisation de la mauvaise transformation de la variable dépendante comme régressande fournira en général une régression à aléas ni homoscedastiques ni symétriques. Ainsi, quand nous rencontrerons l'hétéroscédasticité et l'asymétrie dans les résidus d'une régression, une façon possible de les éliminer consistera à estimer un modèle de régression différent dans lequel la variable dépendante a été soumise à une **transformation non linéaire**. Il s'agit en fait d'une approche déjà grandement utilisée en économétrie et en statistique, et dont nous discutons plus en détail dans ce chapitre. Cependant, nous devrions insister dès à présent que dans n'importe quel cas donné il peut ne pas exister de transformation de la variable dépendante qui fournisse des résidus symétriques et homoscedastiques. Il est également possible qu'une certaine forme des moindres carrés pondérés fonctionnera mieux qu'un modèle qui comporte une transformation de la variable dépendante. Ainsi les techniques qui seront discutées dans ce chapitre ne seront pas utiles pour chaque cas.

Il existe de nombreuses manières où les transformations de la variable dépendante peuvent être employées dans un modèle de régression. Désignons  $\tau(x, \lambda)$  une transformation non linéaire de  $x$  avec comme paramètre scalaire

$\lambda$  qui peut ou pas avoir été estimé. La transformation la plus communément répandue est la **transformée de Box-Cox**, qui a été proposée par Box et Cox (1964) dans un très célèbre article; elle sera discutée dans la prochaine section. Une classe de modèle qui utilise une telle transformation est celle suggérée à l'origine par Box et Cox:

$$\tau(y_t, \lambda) = x_t(\boldsymbol{\beta}) + u_t, \quad (14.04)$$

où la transformée s'applique seulement à la variable dépendante. Cette classe de modèles a été très courante en statistique mais beaucoup moins en économétrie. Une seconde classe de modèles est

$$\tau(y_t, \lambda) = \tau(x_t(\boldsymbol{\beta}), \lambda) + u_t, \quad (14.05)$$

dans laquelle la transformation  $\tau(x, \lambda)$  est appliquée à la fois à la variable et à la fonction de régression. Les modèles de ce type avaient été préconisés par Carroll et Ruppert (1984, 1988), qui les ont appelés modèles de "transformation des deux cotés". Ces modèles ont aussi été très largement utilisés en statistique et dans une moindre mesure en économétrie; un exemple précoce est Leech (1975).

Une troisième classe de modèles est

$$\tau(y_t, \lambda) = \sum_{i=1}^k \beta_i \tau(X_{ti}, \lambda) + \sum_{j=1}^l \gamma_j Z_{tj} + u_t, \quad (14.06)$$

où  $X_{ti}$  et  $Z_{tj}$  désignent à la fois les observations sur les variables indépendantes, la distinction étant que les  $X_{ti}$  sont soumis à la transformée et que les  $Z_{tj}$  ne le sont pas. Il s'agit de l'approche qui a été généralement prise en économétrie, avec la transformation  $\tau(x, \lambda)$  étant inmanquablement la transformée de Box-Cox.<sup>2</sup> La classe de modèles (14.06) est plus générale que (14.04), au moins si  $x_t(\boldsymbol{\beta})$  dans ce modèle est restreint à être linéaire, et d'une certaine manière elle est aussi plus générale que (14.05). Elle peut aussi être généralisée par la suite en permettant à la valeur de  $\lambda$  utilisée pour transformer  $y_t$  d'être différent de la valeur (ou des valeurs) utilisée pour transformer les  $X_{ti}$  (consulter la Section 14.7).

Notons que les modèles (14.04) et (14.05) sont principalement concernés par l'obtention de résidus homoscédastiques et symétriques, alors que la forme fonctionnelle de la fonction de régression est considérée comme une donnée. En revanche, dans le modèle (14.06), la forme fonctionnelle dépend explicitement de  $\lambda$ . Peut-être qu'en conséquence de ceci, la majeure partie de la littérature des débuts de l'économétrie s'est principalement intéressée

<sup>2</sup> Les études qui utilisent ou discutent cette approche incluent Zarembka(1968, 1974), White (1972), Heckman et Polachek (1974), Savin et White (1978), et Spitzer (1976, 1978, 1982a, 1982b, 1984).

à déterminer la forme fonctionnelle de la fonction de régression, en restant très peu concernée par les propriétés des résidus. Ce manque d'intérêt a été mal placé, parce que la caractéristique clé de tout modèle comprenant une transformation de la variable dépendante est que la transformation affecte directement les propriétés des résidus.

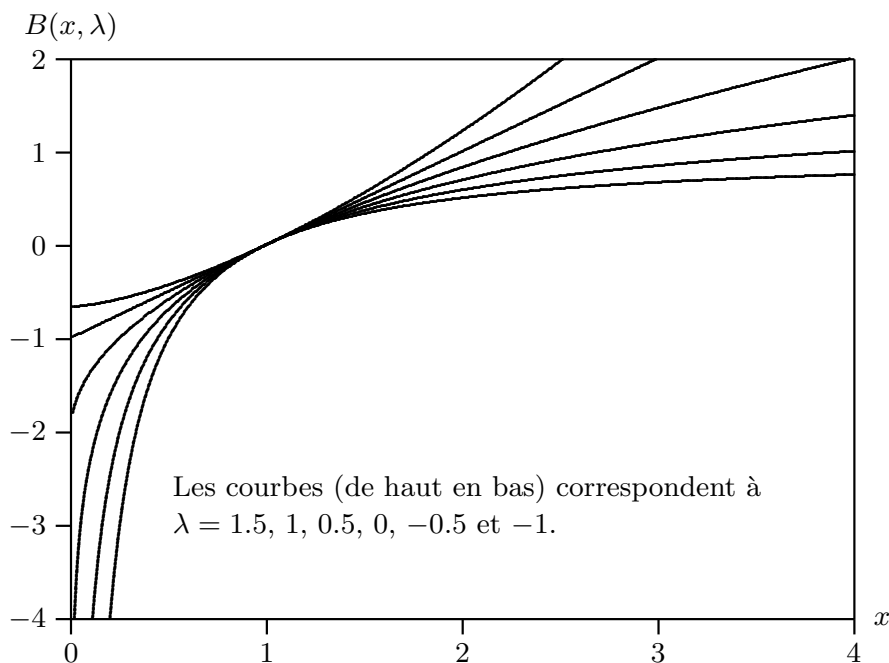
Les modèles (14.04), (14.05), et (14.06) ne peuvent pas être qualifiés de modèles de régression, parce que la variable dépendante n'est pas simplement égale à la somme d'une fonction de régression et d'un aléa. Bien que ces modèles soient différents, et puissent fournir des résultats très différents en pratique, ils comportent tous un élément en commun, à savoir, que la variable dépendante est soumise à une transformation non linéaire avec le paramètre  $\lambda$ . Si  $\lambda$  était connu, ces modèles pourraient tous être estimés par moindres carrés non linéaires et testés en utilisant la régression de Gauss-Newton. Mais tant que  $\lambda$  est inconnu et qu'il doit être estimé, la méthode NLS est clairement inappropriée. Dans la plupart des cas, un algorithme par moindres carrés choisirait simplement  $\lambda$  de façon à rendre  $\tau(y_t, \lambda)$  aussi petit que possible afin de rendre la somme des résidus au carré aussi petite que possible. Ainsi, cela fournirait inévitablement des résultats insensés, comme nous en avons discuté dans le Chapitre 8 en connexion avec le modèle (8.01).

Dans la prochaine section, nous discutons de la transformée de Box-Cox et de l'estimation des modèles de régression où la variable dépendante a été soumise à cette transformation. L'estimation par maximum de vraisemblance se trouve être très facile parce que la fonction de logvraisemblance incorpore un terme Jacobien qui empêche  $\lambda$  de devenir trop petit. Dans la Section 14.3, nous réalisons une légère digression pour discuter de certaines des propriétés utiles des termes Jacobiens dans l'estimation ML. Dans la section 14.4, nous discutons alors d'une nouvelle classe de régressions artificielles appelée **régressions artificielles à longueur double** et, dans la Section 14.5, nous montrons comment celles-ci peuvent être utilisées pour l'estimation et le test des modèles comprenant la transformée de Box-Cox. Dans la Section 14.6, nous discutons de la façon dont il est possible de tester la spécification de linéarité ou de loglinéarité d'un modèle contre une alternative Box-Cox ou autre. Finalement, dans la Section 14.7, nous traitons brièvement de certains modèles qui comprennent des généralisations ou des alternatives de la transformée de Box-Cox.

## 14.2 LA TRANSFORMÉE DE BOX-COX

La transformée de Box-Cox est la transformation non linéaire de loin la plus rencontrée en statistique et en économétrie. Elle est définie comme

$$B(x, \lambda) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{quand } \lambda \neq 0; \\ \log(x) & \text{quand } \lambda = 0, \end{cases}$$



**Figure 14.1** Transformées de Box-Cox pour des valeurs diverses de  $\lambda$

où l'argument  $x$  doit être positif. D'après la règle de l'Hôpital,  $\log x$  est la limite de  $(x^\lambda - 1)/\lambda$  quand  $\lambda \rightarrow 0$ . La Figure 14.1 montre la transformée de Box-Cox pour différentes valeurs de  $\lambda$ . En pratique,  $\lambda$  s'étend généralement d'une valeur inférieure à 0 à une valeur supérieure à 1. Il peut être montré que  $B(x, \lambda') \geq B(x, \lambda'')$  pour  $\lambda' \geq \lambda''$ , et cette inégalité est évidente sur la figure. Ainsi la valeur de courbure de la transformée de Box-Cox augmente quand  $\lambda$  s'éloigne de 1 dans l'une ou l'autre direction.

Il existe trois variétés de modèle de Box-Cox. Nous nous référerons à (14.04) et (14.05) avec  $\tau(\cdot)$  donné par la transformée de Box-Cox, du **modèle de Box-Cox simple** et du **modèle de Box-Cox transformé des deux côtés** respectivement. Nous nous référerons à (14.06) du **modèle de Box-Cox conventionnel**, parce qu'il s'agit du plus communément utilisé en économétrie.

Une des raisons de la popularité de la transformée de Box-Cox est qu'elle incorpore à la fois la possibilité d'aucune transformation (quand  $\lambda = 1$ ) et la possibilité d'une transformation logarithmique (quand  $\lambda = 0$ ). Sous réserve que les régresseurs incluent un terme constant, soumettre la variable dépendante à la transformée de Box-Cox  $\lambda = 1$  est équivalent à n'effectuer aucune transformation. Soumettre la variable dépendante à la transformée de Box-Cox avec  $\lambda = 0$  est équivalent à utiliser  $\log y_t$  comme régressande. Comme ces deux transformations sont deux cas spécifiques très plausibles, il est très séduisant d'utiliser une transformation qui tienne compte des deux à la fois. Même quand le modèle conventionnel de Cox-Box n'est pas considéré comme véritablement plausible, ce dernier fournit une alternative commode

à partir de laquelle il est possible de tester la spécification des modèles de régression linéaire et non linéaire; consulter la Section 14.6.

Cependant, la transformée de Box-Cox n'est pas sans sérieux inconvénients. Considérons le modèle de Box-Cox simple

$$B(y_t, \lambda) = x_t(\boldsymbol{\beta}) + u_t, \quad u_t \sim \text{NID}(0, \sigma^2). \quad (14.07)$$

Pour la plupart des valeurs de  $\lambda$  (mais pas pour  $\lambda = 0$  ou  $\lambda = 1$ ) la valeur de  $B(y_t, \lambda)$  est bornée soit inférieurement soit supérieurement; de manière spécifique, quand  $\lambda > 0$ ,  $B(y_t, \lambda)$  ne peut pas être inférieur à  $-1/\lambda$  et, quand  $\lambda < 0$ ,  $B(y_t, \lambda)$  ne peut pas être supérieur à  $-1/\lambda$ . Cependant, si  $u_t$  est normalement distribué, le membre de droite de (14.07) n'est pas borné et pourrait, du moins en principe, prendre des valeurs positives ou négatives arbitrairement grandes. Ainsi, à strictement parler, (14.07) est logiquement impossible en tant que modèle pour  $y_t$ . Ceci reste vrai si nous remplaçons  $x_t(\boldsymbol{\beta})$  par une fonction de régression qui dépend de  $\lambda$ .

Une manière de traiter ce problème est de supposer que les données sur  $y_t$  sont observées seulement quand les bornes ne sont pas enfreintes, comme dans Poirier (1978) et Poirier et Ruud (1979). Ceci nous conduit à discuter des fonctions de logvraisemblance similaires à celles discutées dans la Section 15.6.<sup>3</sup> Cependant, rien ne justifie le fait que les données doivent toujours être générées de cette manière, et de plus, à la fois l'estimation et le test deviennent très compliqués quand on prend en compte cette sorte de troncature d'échantillon. Une seconde manière de traiter de ce problème consiste simplement à l'ignorer. Cette application du bien célèbre "algorithme autruche" prend tout son sens si  $\lambda$  est non négatif (ou du moins pas inférieur à zéro) et  $y_t$  est positive est relativement grande par rapport à  $\sigma$  pour toutes les observations dans l'ensemble d'information. Quand ces deux conditions sont satisfaites, nous pouvons être sûrs que  $u_t$  sera plus petit relativement à  $B(y_t, \lambda)$  et  $x_t(\boldsymbol{\beta})$ ; par conséquent, la probabilité que le membre de droite de (14.07) n'enfreigne la borne du membre de gauche sera très petite.

Nous adopterons cette seconde approche, parce que les modèles de Box-Cox comportant des valeurs négatives de  $\lambda$  ne sont pas très intéressants, et que, dans de nombreux cas pratiques, la moyenne conditionnelle de  $y_t$  est toujours relativement grande par rapport à n'importe quelle variation autour de la moyenne conditionnelle. Dans de tels cas, il semble assez raisonnable d'utiliser des modèles dans lesquels la variable dépendante est soumise à la transformée de Box-Cox. Cependant, dans d'autres cas, il peut ne pas être correct d'utiliser un modèle de Box-Cox; consulter la Section 14.7.

Considérons maintenant la façon d'obtenir des estimations convergentes de  $\lambda$  et  $\boldsymbol{\beta}$  dans (14.07). Il s'agit du cas le plus simple à discuter, mais tout ce

<sup>3</sup> Une approche différente, mais similaire, avait été proposée par Amemiya et Powell (1981).

que nous dirons s'appliquera également, avec quelques légères et évidentes modifications, aussi bien aux modèles transformés des deux côtés qu'aux modèles de Box-Cox conventionnels, dans lesquels le paramètre de transformation  $\lambda$  apparaît également dans la fonction de régression. Puisque, clairement, les moindres carrés ne serviront pas dans ce cas, il est naturel de se tourner vers le maximum de vraisemblance. Comme nous avons supposé que les  $u_t$  sont normalement et indépendamment distribués, nous pouvons facilement écrire la fonction de logvraisemblance pour ce modèle. Il s'agit de

$$\begin{aligned} \ell(\mathbf{y}, \boldsymbol{\beta}, \lambda, \sigma) = & -\frac{n}{2} \log(2\pi) - n \log \sigma \\ & - \frac{1}{2\sigma^2} \sum_{t=1}^n (B(y_t, \lambda) - x_t(\boldsymbol{\beta}))^2 + (\lambda - 1) \sum_{t=1}^n \log y_t. \end{aligned} \quad (14.08)$$

Ici le dernier terme est la somme sur toutes les observations du logarithme de

$$\frac{\partial B(y_t, \lambda)}{\partial y_t} = \frac{\partial}{\partial y_t} \left( \frac{y_t^\lambda - 1}{\lambda} \right) = y_t^{\lambda-1},$$

qui est le Jacobien de la transformation de  $y_t$  vers  $u_t$ .

Le rôle de ce terme Jacobien est crucial. Afin d'éviter d'avoir plus d'un cas, supposons pour simplifier que tous les  $y_t$  sont plus grands que 1. Puisque

$$\text{plim}_{\lambda \rightarrow -\infty} B(x, \lambda) = 0$$

pour  $x > 1$ , en laissant  $\lambda \rightarrow -\infty$  alors  $B(y_t, \lambda) \rightarrow 0$  pour tout  $t$ . Ainsi, à condition qu'il existe une certaine valeur de  $\boldsymbol{\beta}$  qui fait que la fonction de régression  $x_t(\boldsymbol{\beta})$  égale zéro pour tout  $t$ , la somme des résidus au carré,

$$\sum_{t=1}^n (B(y_t, \lambda) - x_t(\boldsymbol{\beta}))^2,$$

devient arbitrairement petite, si on laisse  $\lambda$  tendre vers moins l'infini. Si nous concentrons (14.08) par rapport à  $\sigma$ , la fonction de logvraisemblance devient

$$\ell^c(\mathbf{y}, \boldsymbol{\beta}, \lambda) = C - \frac{n}{2} \log \left( \sum_{t=1}^n (B(y_t, \lambda) - x_t(\boldsymbol{\beta}))^2 \right) + (\lambda - 1) \sum_{t=1}^n \log y_t, \quad (14.09)$$

où  $C$  est une constante qui ne dépend ni de  $\boldsymbol{\beta}$  ni de  $\lambda$ . Ainsi, nous voyons que lorsque nous maximisons la fonction de logvraisemblance, la valeur de  $\lambda$  affectera deux éléments: un terme somme-des-carrés et un terme Jacobien. Le terme Jacobien empêche l'estimation ML de  $\lambda$  de tendre vers moins l'infini, puisque ce terme tend vers moins l'infini tout comme  $\lambda$ .

La maximisation de (14.09) n'est pas très difficile. La meilleure approche, si le logiciel approprié est commode, consiste à utiliser une procédure convenable pour la maximisation non linéaire; consulter la Section 14.5. Une seconde approche consiste à employer une procédure par balayage dans laquelle on recherche les valeurs de  $\lambda$  et estime  $\beta$  par moindres carrés conditionnellement à  $\lambda$ . Une troisième approche consiste à utiliser une astuce qui permet à (14.09) d'être minimisée en utilisant n'importe quel algorithme de moindres carrés non linéaires. Il existe en fait deux manières de réaliser ceci. La plus simple est de noter que si tous les  $y_t$  sont divisés par leur moyenne géométrique  $\dot{y}$ , le terme Jacobien dans (14.09) est alors identiquement égal à zéro, parce que

$$n \log \dot{y} = \sum_{t=1}^n \log y_t.$$

Ainsi, n'importe quelle régression dont les résidus sont  $B(y_t/\dot{y}, \lambda) - x_t(\beta)$  fournira des estimations valides de  $\beta$  et  $\lambda$ . Par exemple, nous pourrions définir la régressande comme un vecteur de zéros et la fonction de régression comme  $B(y_t/\dot{y}, \lambda) - x_t(\beta)$  et alors utiliser n'importe quel algorithme. Cette approche a été usitée pendant de nombreuses années mais comporte le désavantage qu'il faut modifier l'échelle des  $y_t$ ; comme nous le verrons plus loin, cette procédure n'est pas toujours totalement neutre dans le contexte des modèles de Box-Cox.

Une seconde manière d'utiliser un programme NLS a été proposée par Carroll et Ruppert (1988). Nous pouvons récrire (14.09) comme

$$\ell^c(\mathbf{y}, \beta, \lambda) = C^* - \frac{n}{2} \log \left( \sum_{t=1}^n \left( \frac{B(y_t, \lambda) - x_t(\beta)}{\dot{y}^\lambda} \right)^2 \right),$$

où  $C^*$  ne dépend ni de  $\beta$  ni de  $\lambda$ . Comme cette version de la fonction de logvraisemblance n'a seulement qu'un terme en somme-des-carrés, elle peut être maximisée par la minimisation de la somme des résidus au carré:

$$\sum_{t=1}^n \left( \frac{B(y_t, \lambda) - x_t(\beta)}{\dot{y}^\lambda} \right)^2.$$

Nous pouvons procéder ainsi en utilisant une procédure NLS en définissant la régressande comme un vecteur de zéros et la fonction de régression comme  $(B(y_t, \lambda) - x_t(\beta))/\dot{y}^\lambda$ .

Bien que toutes les techniques précédemment décrites fournissent des estimations ML  $\hat{\lambda}$  et  $\hat{\beta}$ , aucune des méthodes basées sur les moindres carrés ne fournit une estimation valide de la matrice de covariance  $\hat{\lambda}$  et  $\hat{\beta}$ . La raison est, comme nous le verrons dans la Section 14.5, la matrice d'information pour les modèles de Box-Cox n'est pas bloc-diagonale en  $\beta$ ,  $\lambda$ , et  $\sigma$ . Les méthodes de grille de valeurs qui estiment  $\beta$  conditionnellement à  $\lambda$  fournissent des matrices de covariance invalides parce qu'elles ignorent le fait que  $\hat{\lambda}$  est elle-même une

estimation. Les méthodes qui amènent un programme NLS à estimer  $\hat{\lambda}$  et  $\hat{\beta}$  conjointement fournissent également des estimations de matrices de covariance invalides parce qu'elles supposent implicitement que la matrice de covariance est bloc-diagonale entre  $\sigma$  et les autres paramètres, ce qui n'est pas le cas pour les modèles de Box-Cox. Comme il est très tentant d'utiliser les estimations incorrectes de l'écart type affichées par le progiciel des moindres carrés, nous recommandons que les procédures basées sur moindres carrés soient seulement usitées pour estimer les modèles de Box-Cox quand le logiciel le plus approprié est indisponible.

Nous pouvons, naturellement, obtenir une matrice de covariance estimée valide de différentes façons en inversant différentes estimations de la matrice d'information. La régression OPG fournit probablement la manière la plus simple d'obtenir une estimation de matrice de covariance, mais ses propriétés en échantillon fini ne sont pas très bonnes, et des techniques plus spécialisées plus appropriées sont disponibles; consulter Spitzer (1984). Dans la Section 14.4 et 14.5, nous discuterons d'une classe de régressions artificielles qui peut être utilisée pour traiter d'une large classe de modèles et semble très bien fonctionner pour les modèles de Box-Cox. Comme toutes les régressions artificielles, ces régressions à longueur double, tel est leur nom, peuvent être employées pour des estimations, inférences, et tests de spécification.

Nous avons remarqué plus tôt que la modification d'échelle de la variable dépendante peut ne pas être neutre dans un modèle de Box-Cox. Dans un modèle transformé des deux côtés, la renormalisation de la variable dépendante a exactement le même effet que s'il n'y avait eu aucune transformation, parce qu'à la fois la variable dépendante et la fonction de régression sont transformées de la même façon. Ainsi, si  $x_t(\beta)$  est linéaire, tous les coefficients seront simplement multipliés par le facteur utilisé pour renormaliser la variable dépendante. Si  $x_t(\beta)$  est non linéaire, la renormalisation de  $y_t$  peut très bien avoir une influence sur  $\beta$  de façon plus délicate et peut même affecter la façon dont le modèle s'ajuste, mais cela se réalisera seulement si la renormalisation affecte l'ajustement du modèle même si aucune transformation n'est impliquée. Cependant, dans les deux autres types de modèle de Box-Cox, les choses ne sont pas aussi simples.

Il existe un important résultat d'invariance pour les modèles de Box-Cox conventionnels et simples. Ce résultat est que, sous certaines conditions, l'estimation de  $\lambda$  est invariante par rapport à l'échelle de la variable dépendante. Supposons que nous multiplions  $y_t$  par une constante  $\alpha$  de sorte que la variable dépendante devienne  $\alpha y_t$ . La transformée de Box-Cox de  $\alpha y_t$  est

$$B(\alpha y_t, \lambda) = \alpha^\lambda B(y_t, \lambda) + B(\alpha, \lambda).$$

Ici le second terme est juste une constante. Pourvu qu'il y ait un terme constant (ou l'équivalent) dans la fonction de régression, l'estimation de la constante s'ajustera toujours automatiquement pour s'y accommoder. Si la fonction de régression est linéaire, toutes les estimations paramétriques sauf

la constante seront simplement multipliées par  $\alpha^\lambda$ , tout comme les résidus et  $\hat{\sigma}$ . Pour le modèle de Box-Cox, la renormalisation est plus compliquée, mais l'effet net est que les résidus sont encore multipliés par  $\alpha^\lambda$ . Ceci est également vrai pour certaines autres fonctions de régression  $x_t(\boldsymbol{\beta})$ , mais pas pour toutes. Pourvu que la renormalisation  $y_t$  soit équivalente à celle des résidus de cette manière, le terme somme-des-carrés dans (14.08), évalué pour un  $\lambda$  fixé arbitrairement au  $\hat{\boldsymbol{\beta}}$  qui minimise la somme des résidus au carré et le correspondant  $\hat{\sigma}^2$ , est invariant sous la renormalisation. Le second terme de (14.08),  $-n \log \sigma$ , devient  $-n \log \sigma - n\lambda \log \alpha$ . Le dernier terme, le Jacobien, devient

$$(\lambda - 1) \sum_{t=1}^n \log y_t + n(\lambda - 1) \log \alpha.$$

Ainsi l'opération entière additionne  $-n \log \alpha$ , une quantité indépendante de tous les autres paramètres, à la fonction de logvraisemblance concentrée par rapport à  $\boldsymbol{\beta}$  et  $\sigma^2$ . Par conséquent il est clair que, pourvu que la normalisation de  $y_t$  est équivalente à celle des résidus, l'estimation ML  $\hat{\lambda}$  ne changera pas quand nous renormalisons  $y_t$ . Ce résultat a été, pour l'essentiel, démontré à l'origine par Schlesselman (1971).

Même quand  $\hat{\lambda}$  est invariant à l'échelle, les autres paramètres ne le seront généralement pas. Dans le modèle de Box-Cox conventionnel, les effets de la renormalisation de  $y_t$  dépendent de la valeur de  $\lambda$ . Quand  $\lambda = 1$ , de telle sorte qu'il s'agisse réellement d'un modèle de régression linéaire, la multiplication de  $y_t$  par  $\alpha$  change simplement tous les coefficients estimés par un facteur de  $\alpha$  et n'a aucun effet sur les  $t$  de Student. Quand  $\lambda = 0$ , de telle sorte qu'il s'agisse réellement d'un modèle de régression loglinéaire, la multiplication de  $y_t$  par  $\alpha$  signifie l'addition d'un  $\log \alpha$  constant à la régressande, qui affecte le terme constant mais aucun des autres coefficients. Mais à l'exception de ces deux cas, tous les autres coefficients changeront généralement quand la variable dépendante est renormalisée. De plus, en raison du manque d'invariance des tests de Wald aux reparamétrisations non linéaires, tous les  $t$  de Student sur les  $\beta_i$  changeront de la sorte; consulter Spitzer (1984). En fait, il est très possible qu'un Student soit hautement significatif pour une normalisation de  $y_t$  et complètement non significatif pour une autre. Ceci implique naturellement que, quelle que soit la normalisation de  $y_t$ , nous ne devons pas faire confiance aux Student (ou à n'importe quelle sorte de test de Wald) dans le contexte des modèles de Box-Cox.

### 14.3 LE RÔLE DES TERMES JACOBIENS DANS L'ESTIMATION ML

Des termes Jacobiens sont apparus dans les fonctions de logvraisemblance dans une variété de contextes dans les Chapitres 8, 9, et 10. Nous avons vu qu'à chaque fois que la variable dépendante est soumise à une transformation non linéaire, la fonction de logvraisemblance contient nécessairement au moins

un terme Jacobien. Dans cette section, nous étudions plus en détails le rôle joué par les termes Jacobiens dans l'estimation ML. Nous continuerons notre discussion des modèles de Box-Cox dans les sections suivantes.

Rappelons que si la densité de probabilité d'une variable aléatoire  $x_1$  est  $f_1(x_1)$  et qu'une autre variable aléatoire  $x_2$  lui y est reliée par  $x_1 = \tau(x_2)$ , où la fonction  $\tau(\cdot)$  est continuellement différentiable et monotone, alors la densité de  $x_2$  est donnée par

$$f_2(x_2) = f_1(\tau(x_2)) \left| \frac{\partial \tau(x_2)}{\partial x_2} \right|. \quad (14.10)$$

Le second facteur ici est la valeur absolue du Jacobien de la transformation, et est alors souvent désigné sous le nom de **facteur Jacobien**. Dans le cas multivarié, où  $\mathbf{x}_1$  et  $\mathbf{x}_2$  sont des vecteurs de dimension  $m$  et  $\mathbf{x}_1 = \boldsymbol{\tau}(\mathbf{x}_2)$ , l'analogie de (14.10) est

$$f_2(\mathbf{x}_2) = f_1(\boldsymbol{\tau}(\mathbf{x}_2)) |\det \mathbf{J}(\mathbf{x}_2)|,$$

où  $|\det \mathbf{J}(\mathbf{x}_2)|$  est la valeur absolue du déterminant de la matrice Jacobienne  $\mathbf{J}(\mathbf{x}_2)$  avec comme élément type

$$J_{ij}(\mathbf{x}_2) \equiv \frac{\partial \tau_i(\mathbf{x}_2)}{\partial x_{2j}}.$$

Ces résultats sont discutés dans l'Annexe B.

Des facteurs Jacobiens dans les fonctions de densité induisent des termes Jacobiens dans les fonctions de logvraisemblance. Ceux-ci peuvent se produire à chaque fois que la transformation de la variable(s) dépendante(s) observée(s) vers les aléas comporte une matrice Jacobienne qui n'est pas la matrice identité. Si les aléas sous-jacents sont supposés être normalement distribués, la présence de ces termes Jacobiens est souvent la seule chose qui fait que la fonction de logvraisemblance soit autre chose qu'une banale transformation de la somme des résidus au carré.

Cependant, il existe des circonstances dans lesquelles la fonction de logvraisemblance ne contient aucun terme Jacobien, même si la matrice Jacobienne n'est pas une matrice identité. Nous avons rencontré une classe de modèles pour lesquels ceci est le cas dans le Chapitre 10. Si nous oublions la première observation, la matrice Jacobienne pour un modèle de régression à erreurs AR(1) est triangulaire inférieure, avec des éléments diagonaux égaux à 1. Puisque le déterminant d'une matrice triangulaire est le produit des éléments de la diagonale, le facteur Jacobien pour de tels modèles est simplement l'unité, et le terme Jacobien est par conséquent zéro.

Dans cette section, naturellement, nous traitons de nombreux autres cas dans lesquels les termes Jacobiens apparaissent dans des fonctions de logvraisemblance. Leur apparition comporte plusieurs conséquences. Tout

d'abord, elle signifie que les moindres carrés non linéaires et la régression de Gauss-Newton ne sont pas applicables à de tels modèles. Des astuces telles que celle que nous avons utilisée dans la section précédente peuvent permettre aux NLS d'être utilisées pour l'estimation, mais elles ne permettront pas à l'inférence d'être basée comme d'habitude sur les estimations NLS. La régression OPG sera applicable, ainsi que des régressions artificielles plus spécialisées telles que la régression à longueur double qui sera introduite dans la prochaine section.

Ensuite, la présence des termes Jacobiens assure que nous ne pouvons jamais obtenir des estimations en des points dans l'espace paramétrique où le Jacobien de la transformation provenant de la variable(s) dépendente(s) vers les aléas sous-jacents est singulier. En de tels points, il ne sera pas du tout possible de réaliser cette transformation. Quand le vecteur de paramètres se rapproche d'un tel point, le déterminant de la matrice Jacobienne tend vers zéro, et le logarithme de ce déterminant tend par conséquent vers moins l'infini. Nous avons vu un exemple de ce phénomène dans la Section 10.6, où la fonction de logvraisemblance pour un modèle à erreurs AR(1) tendait vers moins l'infini quand  $|\rho| \rightarrow 1$ . La transformation pour la première observation,

$$(1 - \rho^2)^{1/2}(y_1 - x_1(\beta)) = \varepsilon_1,$$

ne peut pas être effectuée lorsque  $|\rho| = 1$ , et la fonction de logvraisemblance reflète ce fait en prenant la valeur de moins l'infini.

Cette propriété des fonctions de logvraisemblance est une des plus désirables, parce qu'elle nous empêche d'obtenir des estimations insensées. Cependant, elle implique que les fonctions de logvraisemblance pour de tels modèles doivent avoir des maxima multiples. Par exemple, dans le cas le plus simple dans lequel la singularité divise l'espace paramétrique en deux régions, il doit y avoir au moins un maximum dans chacune de ces régions. Ainsi, si nous commençons l'erreur de débiter l'algorithme de maximisation dans la mauvaise région, l'algorithme peut bien manquer de croiser la singularité et ainsi nous trouverons un maximum local qui n'est pas le maximum global; voir MacKinnon (1979). Nous rencontrerons des exemples complémentaires de singularités dans les fonctions de logvraisemblance dans le Chapitre 18 quand nous discuterons de l'utilisation du maximum de vraisemblance pour l'estimation des modèles à équations simultanées.

La troisième conséquence majeure de la présence de termes Jacobiens dans les fonctions de logvraisemblance, et une des plus intéressantes pour nous dans ce chapitre, est que l'estimation par maximum de vraisemblance, à la différence des moindres carrés, n'est pas gênée par les transformations de la variable dépendante, parce que, comme nous l'avons vu dans la dernière section, la présence d'une transformation occasionne la présence d'un terme Jacobien dans la fonction de logvraisemblance. Un problème courant dans les travaux économétriques appliqués consiste à déterminer la transformation la

plus appropriée de la variable dépendante. Par exemple, la théorie économique pourrait admettre les trois spécifications suivantes:

$$H_1: y_t = \alpha_1 + \beta_1 x_t + u_t, \quad (14.11)$$

$$H_2: \log y_t = \alpha_2 + \beta_2 \log x_t + u_t, \quad \text{et} \quad (14.12)$$

$$H_3: \frac{y_t}{z_t} = \alpha_3 \frac{1}{z_t} + \beta_3 \frac{x_t}{z_t} + u_t, \quad (14.13)$$

où  $z_t$  et  $x_t$  sont des observations sur des variables exogènes ou prédéterminées. Ici, les fonctions de régression sont délibérément très simples, parce que la manière dont elles sont spécifiées est hors de propos du principal argument.

Il n'est clairement pas approprié de comparer les sommes des résidus au carré ou les  $R^2$  issus de (14.11), (14.12), et (14.13). Néanmoins, si nous voulons supposer la normalité, il est très facile de comparer les valeurs des fonctions de logvraisemblance provenant des trois modèles en compétition. Ces fonctions de logvraisemblance, concentrées par rapport au paramètre de variance, sont, respectivement,

$$\ell_1^c(\mathbf{y}, \boldsymbol{\beta}_1) = C - \frac{n}{2} \log \left( \sum_{t=1}^n (y_t - \alpha_1 - \beta_1 x_t)^2 \right), \quad (14.14)$$

$$\ell_2^c(\mathbf{y}, \boldsymbol{\beta}_2) = C - \frac{n}{2} \log \left( \sum_{t=1}^n (\log y_t - \alpha_2 - \beta_2 \log x_t)^2 \right) - \sum_{t=1}^n \log y_t, \quad (14.15)$$

et

$$\ell_3^c(\mathbf{y}, \boldsymbol{\beta}_3) = C - \frac{n}{2} \log \left( \sum_{t=1}^n \left( \frac{y_t}{z_t} - \alpha_3 \frac{1}{z_t} - \beta_3 \frac{x_t}{z_t} \right)^2 \right) - \sum_{t=1}^n \log z_t, \quad (14.16)$$

où la constante  $C$  est la même pour les trois spécifications.

Ce qui rend possible la comparaison de ces trois fonctions de logvraisemblance est la présence des termes Jacobiens dans (14.15) et (14.16). Ils surviennent parce que

$$\frac{\partial \log y_t}{\partial y_t} = \frac{1}{y_t} \quad \text{et} \quad \frac{\partial (y_t/z_t)}{\partial y_t} = \frac{1}{z_t}.$$

Ainsi, si nous souhaitons décider lequel de (14.11), (14.12), et (14.13) s'ajuste le mieux, nous avons simplement à estimer chacun d'entre eux par NLS (ou peut-être par OLS), à retrouver les valeurs des fonctions de logvraisemblance données par le progiciel de régression, à soustraire  $\sum \log y_t$  dans le cas de (14.12) et  $\sum \log z_t$  dans le cas de (14.13), et à comparer les valeurs qui résultent de  $\ell_1$ ,  $\ell_2$ , et  $\ell_3$ . Notons que, pour la plupart des progiciels de régression, les valeurs de  $\ell$  pour (14.12) et (14.13) seront incorrectes quand  $y_t$

(plutôt que  $\log y_t$  ou  $y_t/z_t$ ) est vraiment la variable dépendante. Comme le progiciel ne sait pas que la régressande a été soumise à une transformation, les valeurs qu'il donne omettront les termes Jacobiens dans (14.15) et (14.16).

Cette sorte de procédure peut en fait être réalisée pour tester, et peut-être rejeter, un ou plusieurs des modèles en compétition. Il est possible de voir que chaque paire de  $H_1$ ,  $H_2$ , et  $H_3$  peut être imbriquée dans un modèle plus général comprenant un paramètre supplémentaire. Par exemple, le modèle

$$\frac{y_t}{z_t^\phi} = \alpha \frac{1}{z_t^\phi} + \beta \frac{x_t}{z_t^\phi} + u_t$$

se réduit à  $H_1$  quand  $\phi = 0$  et à  $H_3$  quand  $\phi = 1$ . De façon similaire, le modèle de Box-Cox

$$B(y_t, \lambda) = \alpha + \beta B(x_t, \lambda) + u_t \quad (14.17)$$

se réduit à  $H_1$  quand  $\lambda = 1$  et à  $H_2$  quand  $\lambda = 0$ . Supposons que nous estimions  $H_1$  et  $H_2$ , et que les valeurs de  $\ell_1$  et  $\ell_2$  soient  $-523.4$  et  $-520.7$ , respectivement. Puisque nous connaissons que le modèle emboîtant (14.17) doit s'ajuster au moins aussi bien que celui de  $H_1$  et  $H_2$  qui s'ajuste le mieux, le maximum non contraint de la fonction de logvraisemblance doit être supérieur ou égal à  $-520.7$ . Ainsi une statistique de test LR de  $H_1$  contre le modèle emboîtant doit être supérieure à

$$2(-520.7 - (-523.4)) = 2(523.4 - 520.7) = 5.4.$$

Puisque 5.4 excède la valeur critique à 5% pour un test à un degré de liberté, nous pouvons conclure que le modèle linéaire  $H_1$  sera rejeté à un niveau inférieur à 5% s'il est testé contre le modèle emboîtant, même si nous n'avons pas estimé le dernier ou calculé une statistique de test formelle.

Cet exemple illustre une caractéristique des tests LR qui peut être très commode, à savoir, que nous pouvons parfois mettre une borne inférieure à la statistique de test LR sans réellement estimer le modèle non contraint. Cette caractéristique a été notée par Sargan (1964) dans le contexte du choix entre modèles linéaire et non linéaire; elle est très largement utilisée dans les travaux appliqués, et elle a récemment été proposée comme une base pour la sélection de modèles par Pollak et Wales (1991). La procédure fonctionne seulement dans une direction, naturellement. Ainsi, le fait qu'une bonne performance de  $H_2$  nous permette de rejeter  $H_1$  dans cet exemple ne nous dit rien concernant  $H_2$ .  $H_2$  pourrait très bien être rejetée également si nous l'avions en fait testée contre le modèle emboîtant (consulter la Section 14.6).

## 14.4 RÉGRESSIONS ARTIFICIELLES À LONGUEUR DOUBLE

Pour tous les modèles discutés dans les Sections 14.1 et 14.2, la fonction de logvraisemblance est égale à une somme de contributions pour chacune des  $n$  observations; (14.08) fournit un exemple. Ainsi, la régression OPG pourrait clairement être utilisée pour l'estimation et le test de ces modèles. Cependant, étant donné la performance généralement pauvre en échantillon fini des quantités calculées au moyen de la régression OPG, nous ne préférons pas y baser les inférences. Heureusement, une autre régression artificielle est disponible. Appelée la **régression artificielle à longueur double**, ou **DLR**, elle aussi peut être utilisée avec ces modèles et elle fonctionne vraiment beaucoup mieux que la régression OPG en échantillons finis. Dans cette section, nous fournirons une brève introduction à la DLR. Dans la prochaine section, nous montrons comment elle peut être utilisée dans l'estimation et le test des modèles de Box-Cox. Les principales références à ce sujet sont Davidson et MacKinnon (1984a, 1988). Davidson et MacKinnon (1983a, 1985c), Bera et McKenzie (1986), Godfrey, McAleer, et McKenzie (1988), et MacKinnon et Magee (1990) fournissent des évidences Monte Carlo qui suggèrent que les tests basés sur la DLR ont des performances bien meilleures qu'en ont ceux basés sur la régression OPG en échantillons finis.

La classe des modèles à laquelle s'applique la DLR peut être écrite comme

$$f_t(y_t, \boldsymbol{\theta}) = \varepsilon_t, \quad t = 1, \dots, n, \quad \varepsilon_t \sim \text{NID}(0, 1), \quad (14.18)$$

où chaque  $f_t(\cdot)$  est une fonction lisse qui dépend de la variable aléatoire  $y_t$ , d'un vecteur de paramètres  $\boldsymbol{\theta}$  de dimension  $k$ , et (implicitement) de certaines variables exogènes et/ou prédéterminées. Comme la fonction  $f_t(\cdot)$  peut aussi dépendre des valeurs retardées de  $y_t$ , les modèles dynamiques sont permis. Ceci peut paraître à première vue être une classe de modèles plutôt restrictive, mais elle est en fait très générale. Par exemple, un modèle transformé des deux côtés, comme (14.05) peut, si les aléas sont supposés être  $\text{NID}(0, \sigma^2)$ , être écrit sous la forme de (14.18) en posant les définitions

$$f_t(y_t, \boldsymbol{\theta}) \equiv \frac{1}{\sigma} \left( \tau(y_t, \lambda) - \tau(x_t(\boldsymbol{\beta}), \lambda) \right) \quad \text{et} \quad \boldsymbol{\theta} \equiv [\boldsymbol{\beta} \ ; \ \lambda \ ; \ \sigma].$$

De la même manière, beaucoup d'autres modèles comportant des transformations de la variable dépendante peuvent être mis sous la forme (14.18). Il est même possible de mettre certains modèles multivariés sous cette forme; consulter Davidson et MacKinnon (1984a).

Pour un modèle de la classe à laquelle la DLR s'applique, la contribution de la  $i^{\text{ième}}$  observation à la fonction de logvraisemblance  $\ell(\mathbf{y}, \boldsymbol{\theta})$  est

$$\ell_t(y_t, \boldsymbol{\theta}) = -\frac{1}{2} \log(2\pi) - \frac{1}{2} f_t^2(y_t, \boldsymbol{\theta}) + k_t(y_t, \boldsymbol{\theta}),$$

où

$$k_t(y_t, \boldsymbol{\theta}) \equiv \log \left| \frac{\partial f_t(y_t, \boldsymbol{\theta})}{\partial y_t} \right|$$

est un terme Jacobien. Maintenant établissons les définitions

$$F_{ti}(y_t, \boldsymbol{\theta}) \equiv \frac{\partial f_t(y_t, \boldsymbol{\theta})}{\partial \theta_i} \quad \text{et} \quad K_{ti}(y_t, \boldsymbol{\theta}) \equiv \frac{\partial k_t(y_t, \boldsymbol{\theta})}{\partial \theta_i}$$

et définissons  $\mathbf{F}(\mathbf{y}, \boldsymbol{\theta})$  et  $\mathbf{K}(\mathbf{y}, \boldsymbol{\theta})$  comme les matrices de dimension  $n \times k$  d'éléments type  $F_{ti}(y_t, \boldsymbol{\theta})$  et  $K_{ti}(y_t, \boldsymbol{\theta})$ . De façon similaire, soit  $\mathbf{f}(\mathbf{y}, \boldsymbol{\theta})$  le vecteur de dimension  $n$  d'élément type  $f_t(y_t, \boldsymbol{\theta})$ . Il est facile de voir que le gradient de  $\ell(\mathbf{y}, \boldsymbol{\theta})$  est

$$\mathbf{g}(\mathbf{y}, \boldsymbol{\theta}) = -\mathbf{F}^\top(\mathbf{y}, \boldsymbol{\theta})\mathbf{f}(\mathbf{y}, \boldsymbol{\theta}) + \mathbf{K}^\top(\mathbf{y}, \boldsymbol{\theta})\boldsymbol{\nu}, \quad (14.19)$$

où  $\boldsymbol{\nu}$  désigne un vecteur de dimension  $n$  dont chaque élément est égale à 1.

Le résultat fondamental qui rend possible la DLR est que, pour cette classe de modèle, la matrice d'information  $\mathcal{J}(\boldsymbol{\theta})$  satisfait l'égalité

$$\mathcal{J}(\boldsymbol{\theta}) = \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} (\mathbf{F}^\top(\mathbf{y}, \boldsymbol{\theta})\mathbf{F}(\mathbf{y}, \boldsymbol{\theta}) + \mathbf{K}^\top(\mathbf{y}, \boldsymbol{\theta})\mathbf{K}(\mathbf{y}, \boldsymbol{\theta})) \right) \quad (14.20)$$

et ainsi elle peut être estimée de façon convergente par

$$\frac{1}{n} (\mathbf{F}^\top(\mathbf{y}, \ddot{\boldsymbol{\theta}})\mathbf{F}(\mathbf{y}, \ddot{\boldsymbol{\theta}}) + \mathbf{K}^\top(\mathbf{y}, \ddot{\boldsymbol{\theta}})\mathbf{K}(\mathbf{y}, \ddot{\boldsymbol{\theta}})), \quad (14.21)$$

où  $\ddot{\boldsymbol{\theta}}$  est un estimateur quelconque convergent de  $\boldsymbol{\theta}$ . Nous nous sommes intéressés aux implications de (14.20) plutôt qu'à sa provenance. La dérivation fait appel à certaines propriétés plutôt spéciales de la distribution normale et peut être trouvée dans Davidson et MacKinnon (1984a).

La principale implication de (14.20) est qu'une certaine régression artificielle, que nous appelons la DLR, comporte toutes les propriétés que nous attendons obtenir d'une régression artificielle. La DLR peut être écrite comme

$$\begin{bmatrix} \mathbf{f}(\mathbf{y}, \boldsymbol{\theta}) \\ \boldsymbol{\nu} \end{bmatrix} = \begin{bmatrix} -\mathbf{F}(\mathbf{y}, \boldsymbol{\theta}) \\ \mathbf{K}(\mathbf{y}, \boldsymbol{\theta}) \end{bmatrix} \mathbf{b} + \text{résidus}. \quad (14.22)$$

Cette régression artificielle comporte  $2n$  **observations artificielles**. La régres-sande est  $f_t(y_t, \boldsymbol{\theta})$  pour l'observation  $t$  et l'unité pour l'observation  $t+n$ , et les régresseurs correspondant au  $\boldsymbol{\theta}$  sont  $-\mathbf{F}_t(\mathbf{y}, \boldsymbol{\theta})$  pour l'observation  $t$  et  $\mathbf{K}_t(\mathbf{y}, \boldsymbol{\theta})$  pour l'observation  $t+n$ , où  $\mathbf{F}_t$  et  $\mathbf{K}_t$  désignent respectivement, les  $t^{\text{ièmes}}$  lignes de  $\mathbf{F}$  et de  $\mathbf{K}$ . De façon intuitive, la raison pour laquelle nous avons besoin ici d'une régression à longueur double est que chaque observation d'origine réalise deux contributions à la fonction de logvraisemblance : un terme de somme-des-carrés  $-\frac{1}{2}f_t^2$  et un terme Jacobien  $k_t$ . Nous savons comme résultat, qu'à la fois le gradient et la matrice d'information comprennent chacun deux parties, et la manière de tenir compte des deux à la fois consiste à incorporer deux observations artificielles dans la régression artificielle pour chaque observation d'origine.

Pourquoi (14.22) constitue-t-elle une régression artificielle valide? Comme nous l'avons noté lorsque nous discutons de la régression OPG dans la Section 13.7, il existe deux conditions principales qu'une régression artificielle doit satisfaire. Il est utile d'énoncer clairement ces conditions de manière quelque peu plus formelle ici.<sup>4</sup> Désignons  $\mathbf{r}(\mathbf{y}, \boldsymbol{\theta})$  la régressande pour une quelconque régression artificielle et  $\mathbf{R}(\mathbf{y}, \boldsymbol{\theta})$  la matrice des régresseurs. Soit  $n^*$  le nombre de lignes à la fois de  $\mathbf{r}(\mathbf{y}, \boldsymbol{\theta})$  et de  $\mathbf{R}(\mathbf{y}, \boldsymbol{\theta})$ , qui sera généralement soit  $n$ , soit un multiple entier de  $n$ . La régression de  $\mathbf{r}(\mathbf{y}, \boldsymbol{\theta})$  sur  $\mathbf{R}(\mathbf{y}, \boldsymbol{\theta})$  aura les propriétés d'une régression artificielle si

$$\mathbf{R}^\top(\mathbf{y}, \boldsymbol{\theta}) \mathbf{r}(\mathbf{y}, \boldsymbol{\theta}) = \rho(\boldsymbol{\theta}) \mathbf{g}(\mathbf{y}, \boldsymbol{\theta}) \quad \text{et} \quad (14.23)$$

$$\text{plim}_{\boldsymbol{\theta}} \left( \frac{1}{n^*} \mathbf{R}^\top(\mathbf{y}, \ddot{\boldsymbol{\theta}}) \mathbf{R}(\mathbf{y}, \ddot{\boldsymbol{\theta}}) \right) = \rho(\boldsymbol{\theta}) \mathcal{J}(\boldsymbol{\theta}), \quad (14.24)$$

où  $\ddot{\boldsymbol{\theta}}$  désigne un estimateur convergent quelconque de  $\boldsymbol{\theta}$ . La notation  $\text{plim}_{\boldsymbol{\theta}}$  indique, comme d'habitude, que la limite en probabilité est donnée sous le DGP caractérisé par le vecteur de paramètres  $\boldsymbol{\theta}$ , et  $\rho(\boldsymbol{\theta})$  est un scalaire défini comme

$$\rho(\boldsymbol{\theta}) \equiv \text{plim}_{\boldsymbol{\theta}} \left( \frac{1}{n^*} \mathbf{r}^\top(\mathbf{y}, \boldsymbol{\theta}) \mathbf{r}(\mathbf{y}, \boldsymbol{\theta}) \right).$$

Parce que  $\rho(\boldsymbol{\theta})$  est égal à l'unité pour la régression OPG et pour la DLR, ces deux régressions artificielles satisfont des conditions plus simples

$$\mathbf{R}^\top(\mathbf{y}, \boldsymbol{\theta}) \mathbf{r}(\mathbf{y}, \boldsymbol{\theta}) = \mathbf{g}(\mathbf{y}, \boldsymbol{\theta}) \quad \text{et} \quad (14.25)$$

$$\text{plim}_{\boldsymbol{\theta}} \left( \frac{1}{n^*} \mathbf{R}^\top(\mathbf{y}, \ddot{\boldsymbol{\theta}}) \mathbf{R}(\mathbf{y}, \ddot{\boldsymbol{\theta}}) \right) = \mathcal{J}(\boldsymbol{\theta}), \quad (14.26)$$

aussi bien que les conditions d'origines (14.23) et (14.24). Cependant, ces conditions plus simples ne sont pas satisfaites par la GNR et sont donc de toute évidence trop simples en général.

Maintenant il est facile de voir que la DLR (14.21) satisfait les conditions (14.25) et (14.26). Pour la première de celles-ci, un simple calcul montre que

$$\begin{bmatrix} -\mathbf{F}(\mathbf{y}, \boldsymbol{\theta}) \\ \mathbf{K}(\mathbf{y}, \boldsymbol{\theta}) \end{bmatrix}^\top \begin{bmatrix} \mathbf{f}(\mathbf{y}, \boldsymbol{\theta}) \\ \iota \end{bmatrix} = -\mathbf{F}^\top(\mathbf{y}, \boldsymbol{\theta}) \mathbf{f}(\mathbf{y}, \boldsymbol{\theta}) + \mathbf{K}^\top(\mathbf{y}, \boldsymbol{\theta}) \iota,$$

qui par (14.19) est égal au gradient  $\mathbf{g}(\mathbf{y}, \boldsymbol{\theta})$ . Pour la seconde, nous voyons que

$$\begin{bmatrix} -\mathbf{F}(\mathbf{y}, \boldsymbol{\theta}) \\ \mathbf{K}(\mathbf{y}, \boldsymbol{\theta}) \end{bmatrix}^\top \begin{bmatrix} -\mathbf{F}(\mathbf{y}, \boldsymbol{\theta}) \\ \mathbf{K}(\mathbf{y}, \boldsymbol{\theta}) \end{bmatrix} = \mathbf{F}^\top(\mathbf{y}, \boldsymbol{\theta}) \mathbf{F}(\mathbf{y}, \boldsymbol{\theta}) + \mathbf{K}^\top(\mathbf{y}, \boldsymbol{\theta}) \mathbf{K}(\mathbf{y}, \boldsymbol{\theta}).$$

<sup>4</sup> Pour un traitement plus complet sur ce sujet, consulter Davidson et MacKinnon (1990).

Le membre de droite est juste l'expression qui apparaît dans le résultat fondamental (14.20). En conséquence, il est clair que la DLR doit satisfaire (14.26). Toute cette discussion suppose, naturellement, que les matrices  $\mathbf{F}(\mathbf{y}, \boldsymbol{\theta})$  et  $\mathbf{K}(\mathbf{y}, \boldsymbol{\theta})$  satisfont les conditions de régularité appropriées, qui peuvent ne pas être toujours facilement vérifiables en la réalité; consulter Davidson et MacKinnon (1984a).

La DLR peut être utilisée de toutes les manières que la GNR et la régression OPG peuvent être utilisées. En particulier, elle peut être utilisée pour

- (i) vérifier que les conditions du premier ordre pour un maximum d'une fonction de logvraisemblance sont satisfaites de façon suffisamment exactes,
- (ii) calculer les matrices de covariance estimées,
- (iii) calculer les statistiques de test,
- (iv) calculer les estimations efficaces en une étape, et
- (v) comme partie clé des procédures d'estimation ML.

L'utilisation de (i) a été discutée dans le contexte de la GNR dans la Section 6.1; celle de (ii) a été abordée dans les Sections 6.2, 10.4, et 13.7; l'emploi de (iii) a été beaucoup usité tout au long du livre, et notamment au début du Chapitre 6; et les utilisations de (iv) et (v) ont été discutées, dans le contexte de la GNR, dans les Sections 6.6 et 6.8. Tout ce qui a été dit, ou presque, concernant les utilisations de la GNR et de la régression OPG s'applique également aussi bien à la DLR et ne sera pas, par conséquent, répété ici.

De nombreuses statistiques de test différentes peuvent être programmées en utilisant la même régression artificielle à longueur double. Dans sa forme score, la statistique LM est

$$\tilde{\mathbf{g}}^\top (n\tilde{\mathbf{J}})^{-1} \tilde{\mathbf{g}}, \quad (14.27)$$

où  $\tilde{\mathbf{g}} \equiv \mathbf{g}(\mathbf{y}, \tilde{\boldsymbol{\theta}})$  est le gradient évalué en un ensemble d'estimations contraintes  $\tilde{\boldsymbol{\theta}}$ . Si nous lançons la DLR (14.22) avec les quantités  $\mathbf{f}(\mathbf{y}, \boldsymbol{\theta})$ ,  $\mathbf{F}(\mathbf{y}, \boldsymbol{\theta})$ , et  $\mathbf{K}(\mathbf{y}, \boldsymbol{\theta})$  évaluées en  $\tilde{\mathbf{f}} \equiv \mathbf{f}(\mathbf{y}, \tilde{\boldsymbol{\theta}})$ ,  $\tilde{\mathbf{F}} \equiv \mathbf{F}(\mathbf{y}, \tilde{\boldsymbol{\theta}})$ , et  $\tilde{\mathbf{K}} \equiv \mathbf{K}(\mathbf{y}, \tilde{\boldsymbol{\theta}})$ , la somme expliquée des carrés sera

$$(-\tilde{\mathbf{f}}^\top \tilde{\mathbf{F}} + \boldsymbol{\nu}^\top \tilde{\mathbf{K}})(\tilde{\mathbf{F}}^\top \tilde{\mathbf{F}} + \tilde{\mathbf{K}}^\top \tilde{\mathbf{K}})^{-1}(-\tilde{\mathbf{F}}^\top \tilde{\mathbf{f}} + \tilde{\mathbf{K}}^\top \boldsymbol{\nu}). \quad (14.28)$$

Ceci a clairement la même forme que la statistique LM (14.27). A partir de (14.19), nous voyons que  $\tilde{\mathbf{g}} = -\tilde{\mathbf{F}}^\top \tilde{\mathbf{f}} + \tilde{\mathbf{K}}^\top \boldsymbol{\nu}$ . A partir de (14.20), nous voyons que  $\mathbf{J}(\boldsymbol{\theta})$  est estimée de façon convergente par  $n^{-1}(\tilde{\mathbf{F}}^\top \tilde{\mathbf{F}} + \tilde{\mathbf{K}}^\top \tilde{\mathbf{K}})$  quand les restrictions sont vraies. Ainsi, la somme expliquée des carrés à partir de la DLR, expression (14.28), fournira une statistique de test valide asymptotiquement. Comme d'habitude, les statistiques pseudo- $F$  et pseudo- $t$  seront également valides.

L'expression générale d'une DLR, (14.22), est d'une simplicité trompeuse. Il peut être alors intéressant de voir ce qui se passe si nous utilisons une DLR

dans un cas simple que nous savons déjà traiter. Considérons un modèle de régression non linéaire univarié

$$y_t = x_t(\boldsymbol{\beta}) + u_t, \quad u_t \sim \text{NID}(0, \sigma^2).$$

Lorsqu'il est écrit sous la forme de (14.18), ce modèle devient

$$f_t(y_t, \boldsymbol{\theta}) \equiv \frac{1}{\sigma}(y_t - x_t(\boldsymbol{\beta})) = \varepsilon_t. \quad (14.29)$$

Si  $\boldsymbol{\beta}$  est un vecteur de dimension  $k$ ,  $\boldsymbol{\theta}$  sera un vecteur de dimension  $(k + 1)$ . Considérons maintenant la façon dont nous pourrions tester les restrictions sur  $\boldsymbol{\beta}$  en utilisant une DLR. La nature et le nombre des restrictions sont non pertinents pour notre propos; pour faire simple, nous pouvons supposer qu'on a  $r \leq k$  restrictions de nullité. Les quantités désignées par  $\sim$  sont évaluées en des estimations ML (par exemple, NLS) soumises à ces restrictions.

En calculant  $\mathbf{f}(\mathbf{y}, \boldsymbol{\theta})$ ,  $\mathbf{F}(\mathbf{y}, \boldsymbol{\theta})$ , et  $\mathbf{K}(\mathbf{y}, \boldsymbol{\theta})$  pour le modèle (14.29), en les évaluant aux estimations contraintes  $\tilde{\boldsymbol{\theta}}$ , et en substituant les résultats dans (14.22), cela fournit la DLR

$$\begin{bmatrix} \tilde{\boldsymbol{\varepsilon}} \\ \boldsymbol{\iota} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{X}}/\tilde{\sigma} & \tilde{\boldsymbol{\varepsilon}}/\tilde{\sigma} \\ \mathbf{0} & -\boldsymbol{\iota}/\tilde{\sigma} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ s \end{bmatrix} + \text{résidus}. \quad (14.30)$$

Ici  $\tilde{\boldsymbol{\varepsilon}} \equiv \tilde{\mathbf{u}}/\tilde{\sigma}$  désigne un vecteur de dimension  $n$  de résidus normalisés et  $\tilde{\mathbf{X}}$  désigne une matrice de dimension  $n \times k$  d'élément type  $\partial x_t(\boldsymbol{\beta})/\partial \beta_i$ , évalué en  $\tilde{\boldsymbol{\beta}}$ . Les  $k$  premiers régresseurs dans (14.30) correspondent aux éléments de  $\boldsymbol{\beta}$ , tandis que le dernier correspond à  $\sigma$ ; Ils ont respectivement les coefficients  $\mathbf{b}$  et  $s$ . Il est évident que le dernier régresseur est orthogonal à la régressande. Il est aussi orthogonal à tous les régresseurs qui correspondent aux éléments de  $\boldsymbol{\beta}$  qui ont été estimés sans restriction (par les conditions du premier ordre) et, sous l'hypothèse nulle, il devrait être non corrélé avec les régresseurs restant. Ainsi il est asymptotiquement valide de laisser tomber ce dernier régresseur. Mais quand il est tombé, la seconde moitié de la DLR devient non pertinente, puisque dans la seconde moitié, tous les régresseurs qui restent sont nuls. Si les facteurs de  $1/\tilde{\sigma}$  sont ignorés, il nous reste la régression artificielle

$$\tilde{\mathbf{u}} = \tilde{\mathbf{X}}\mathbf{b} + \text{résidus}, \quad (14.31)$$

qui est simplement la régression de Gauss-Newton. Comme la régressande n'est pas divisée par  $\tilde{\sigma}$ , il est maintenant nécessaire de diviser la somme expliquée des carrés de (14.31) par une estimation de  $\sigma^2$  quand nous calculons la statistique de test.

Le fait que la DLR est équivalente à la GNR quand cette dernière est valide fait sens. Supposons que  $\text{ESS}_{\text{DLR}}$  désigne la somme expliquée des carrés provenant de (14.30) et que  $\text{ESS}_{\text{GN}}$  désigne la somme expliquée des carrés provenant de la GNR modifiée obtenue à partir de (14.31) en remplaçant  $\tilde{\mathbf{u}}$

par  $\tilde{\varepsilon}$ . Il peut être montré que ces deux statistiques de test sont toutes deux des fonctions de la même variable aléatoire. Cependant, elles *ne seront pas* numériquement identiques, la relation exacte entre elles étant

$$\text{ESS}_{\text{DLR}} = \frac{\text{ESS}_{\text{GN}}}{1 - \text{ESS}_{\text{GN}}/(2n)}.$$

Comme  $\text{ESS}_{\text{DLR}}$  sera toujours plus grande que  $\text{ESS}_{\text{GN}}$ , la DLR sera toujours quelque chose de plus enclin à rejeter l'hypothèse nulle que la régression de Gauss-Newton. La différence entre elles sera habituellement très petite, à moins que  $n$  ne soit très petit ou que  $\text{ESS}_{\text{GN}}$  ne soit très grande. Si, à la place de la somme expliquée des carrés, les statistiques  $t$  ou  $F$  sont utilisées, il peut être montré que la DLR et les régressions de Gauss-Newton fournissent des résultats numériquement identiques, sauf pour des corrections légèrement différentes pour des degrés de liberté.

Il est inutile naturellement, d'utiliser une DLR quand une GNR s'applique, c'est-à-dire quand à la fois l'hypothèse nulle et l'hypothèse alternative sont des modèles de régression. Mais quand la variable dépendante est soumise à une transformation non linéaire qui dépend de paramètres inconnus, la GNR n'est pas applicable. Dans la prochaine section, nous montrons comment la DLR peut être utilisée avec les modèles de Box-Cox et d'autres modèles qui comportent des transformations de la variable dépendante.

## 14.5 LA DLR ET LES VARIABLES TRANSFORMÉES

Il est facile de déterminer la forme spécifique que prend la DLR pour chacun des modèles (14.04), (14.05), et (14.06) pour n'importe quelle transformation spécifiée  $\tau(y_t, \lambda)$ . Considérons (14.04) tout d'abord. Nous pouvons écrire

$$f_t(y_t, \boldsymbol{\beta}, \lambda, \sigma) \equiv \frac{1}{\sigma} (\tau(y_t, \lambda) - x_t(\boldsymbol{\beta})).$$

A partir de (14.22), nous voyons que la régressande pour la DLR est

$$\mathbf{r}(\boldsymbol{\theta}) = \begin{bmatrix} f_t(y_t, \boldsymbol{\beta}, \lambda, \sigma) \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma} (\tau(y_t, \lambda) - x_t(\boldsymbol{\beta})) \\ 1 \end{bmatrix},$$

où les quantités supérieures et inférieures à l'intérieur des grands crochets désignent, respectivement, le  $i^{\text{ième}}$  élément et le  $(t+n)^{\text{ième}}$  de la régressande. Nous utiliserons cette notation de façon extensive quand nous discuterons des DLR.

Pour les trois modèles — (14.04), (14.05), et (14.06) — le terme Jacobien pour la  $t^{\text{ième}}$  observation est

$$k_t \equiv \log \left( \frac{\partial f_t(y_t, \boldsymbol{\beta}, \lambda, \sigma)}{\partial y_t} \right) = \log(\tau_y(y_t, \lambda)) - \log \sigma,$$

où  $\tau_y(y_t, \lambda)$  désigne  $\partial\tau(y_t, \lambda)/\partial y_t$ . Ainsi, la matrice des régresseurs pour la DLR qui correspond à (14.04) est

$$\mathbf{R}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\sigma} \mathbf{X}_t(\boldsymbol{\beta}) & -\frac{1}{\sigma} \tau_\lambda(y_t, \lambda) & \frac{\tau(y_t, \lambda) - x_t(\boldsymbol{\beta})}{\sigma^2} \\ \mathbf{0} & \frac{\tau_{y\lambda}(y_t, \lambda)}{\tau_y(y_t, \lambda)} & -\frac{1}{\sigma} \end{bmatrix}, \quad (14.32)$$

où  $\tau_\lambda(y_t, \lambda)$  désigne  $\partial\tau(y_t, \lambda)/\partial\lambda$ , et  $\tau_{y\lambda}(y_t, \lambda)$  désigne  $\partial\tau_y(y_t, \lambda)/\partial\lambda$ . Les deux quantités dans la première colonne de (14.32) désignent les  $t^{\text{ième}}$  et  $(t+n)^{\text{ième}}$  lignes des  $k$  colonnes de la matrice des régresseurs qui correspond à  $\boldsymbol{\beta}$ . De façon similaire, les deux quantités dans chacune des deuxième et troisième colonnes désignent les éléments de la matrice des régresseurs qui correspondent à  $\lambda$  et  $\sigma$ , respectivement.

Quand la transformation  $\tau$  est la transformée de Box-Cox,

$$\tau_\lambda(y, \lambda) = \frac{\lambda y^\lambda \log y - y^\lambda + 1}{\lambda^2} \quad \text{et}$$

$$\frac{\tau_{y\lambda}(y, \lambda)}{\tau_y(y, \lambda)} = \frac{y^{\lambda-1} \log(y)}{y^{\lambda-1}} = \log(y).$$

Par conséquent, la DLR pour le modèle de Box-Cox simple (14.04) avec  $\tau(y_t, \lambda)$  donnée par la transformée de Box-Cox, est

$$\begin{bmatrix} \frac{1}{\sigma} u_t(y_t, \boldsymbol{\beta}, \lambda) \\ 1 \end{bmatrix} \quad (14.33)$$

$$= \begin{bmatrix} \frac{1}{\sigma} \mathbf{X}_t(\boldsymbol{\beta}) & \frac{-(\lambda y_t^\lambda \log y_t - y_t^\lambda + 1)}{\sigma \lambda^2} & \frac{u_t(y_t, \boldsymbol{\beta}, \lambda)}{\sigma^2} \\ \mathbf{0} & \log y_t & -\frac{1}{\sigma} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ a \\ s \end{bmatrix} + \text{résidus},$$

où  $\mathbf{b}$  est un vecteur de dimension  $k$  des coefficients qui correspondent au  $\boldsymbol{\beta}$ ,  $a$  et  $s$  sont des coefficients scalaires qui correspondent à  $\lambda$  et à  $\sigma$ , et

$$u_t(y_t, \boldsymbol{\beta}, \lambda) \equiv B(y_t, \lambda) - x_t(\boldsymbol{\beta}).$$

Si la DLR (14.33) est évaluée en des estimations ML non contraintes  $\hat{\boldsymbol{\theta}} \equiv (\hat{\boldsymbol{\beta}}, \hat{\lambda}, \hat{\sigma})$ , tous les coefficients estimés seront nuls. Puisque les conditions du premier ordre pour  $\sigma$  impliquent que

$$\hat{\sigma} = \left( \frac{1}{n} \sum_{t=1}^n \hat{u}_t^2 \right)^{1/2},$$

la somme totale des carrés à partir de la régression artificielle sera  $2n$ . Ainsi, l'estimation de la matrice de covariance OLS sera simplement  $(2n/(2n - k - 2))(\hat{\mathbf{R}}^\top \hat{\mathbf{R}})^{-1}$ , où  $\hat{\mathbf{R}}$  désigne la matrice des régresseurs qui apparaît dans (14.33), évaluée aux estimations ML. D'après le résultat fondamental (14.20), cette matrice de covariance OLS fournit une estimation valide de la matrice de covariance asymptotique de l'estimateur ML  $\hat{\boldsymbol{\theta}}$ .

Il est clair à partir de (14.33) que cette matrice de covariance asymptotique n'est pas bloc diagonale entre  $\boldsymbol{\beta}$  et les autres paramètres. En formant la matrice  $\mathbf{R}^\top \mathbf{R}$ , en divisant par  $n$ , et en prenant les limites en probabilité, nous voyons que le bloc  $(\boldsymbol{\beta}, \boldsymbol{\beta})$  de la matrice d'information  $\mathcal{J}(\boldsymbol{\theta})$  est simplement

$$\sigma^{-2} \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{X}^\top(\boldsymbol{\beta}) \mathbf{X}(\boldsymbol{\beta}) \right), \quad (14.34)$$

comme cela serait le cas s'il s'agissait d'un modèle de régression non linéaire. L'élément  $(\sigma, \sigma)$  est simplement  $2/\sigma^2$ , qui est encore ce qu'il serait s'il y avait un modèle de régression non linéaire. Mais  $\mathcal{J}(\boldsymbol{\theta})$  contient aussi un élément  $(\lambda, \lambda)$  un élément  $(\lambda, \sigma)$ , une ligne et une colonne  $(\boldsymbol{\beta}, \lambda)$ , chacun d'entre eux étant clairement non nul. Par exemple, l'élément qui correspond à  $\beta_i$  et  $\lambda$  est

$$- \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n\sigma^2\lambda^2} \sum_{t=1}^n X_{ti}(\boldsymbol{\beta}) (\lambda y_t^\lambda \log y_t - y_t^\lambda + 1) \right).$$

Les éléments  $(\lambda, \lambda)$  et  $(\lambda, \sigma)$  peuvent aussi être obtenus d'une manière directe et ils sont manifestement différents de zéro.

Comme  $\mathcal{J}(\boldsymbol{\theta})$  n'est pas bloc diagonale entre  $\boldsymbol{\beta}$  et les deux autres paramètres, le bloc  $(\boldsymbol{\beta}, \boldsymbol{\beta})$  de son inverse ne sera pas égal à l'inverse de (14.34). Ainsi, comme nous l'avons précisé dans la Section 14.2, il est incorrect d'établir des inférences en utilisant la matrice de covariance estimée NLS pour  $\boldsymbol{\beta}$  conditionnelle à  $\lambda$ . De façon similaire, comme l'élément  $(\lambda, \sigma)$  de  $\mathcal{J}(\boldsymbol{\theta})$  est non nul, nous pouvons trouver l'inverse du bloc de dimension  $(k+1) \times (k+1)$  de la matrice d'information qui correspond à  $\boldsymbol{\beta}$  et  $\lambda$  conjointement sans inverser complètement la matrice d'information. La matrice de covariance estimée obtenue en employant une application NLS en donnant des estimations ML sera donc incorrecte.

Il devrait être clair que ce dont nous venons de parler concernant le modèle de Box-Cox simple s'applique également au modèle transformé des deux côtés et au modèle conventionnel, puisque le Jacobien de la transformation est le même pour tous ces modèles. Il est facile d'établir les DLR pour les deux autres modèles. Dans les deux cas, la régressande a la même forme que la régressande de (14.33), sauf pour le modèle transformé des deux côtés

$$u_t(y_t, \boldsymbol{\beta}, \lambda) \equiv B(y_t, \lambda) - B(x_t(\boldsymbol{\beta}), \lambda)$$

et pour le modèle de Box-Cox conventionnel

$$u_t(y_t, \boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda) \equiv B(y_t, \lambda) - \sum_{i=1}^k \beta_i B(X_{ti}, \lambda) - \sum_{j=1}^l \gamma_j Z_{tj}.$$

Le régresseur qui correspond à  $\sigma$  a aussi la même forme que celle qui apparaît dans (14.33).

Pour le modèle transformé des deux côtés, le régresseur qui correspond à  $\beta_i$  est

$$\begin{bmatrix} \frac{1}{\sigma} (x_t(\boldsymbol{\beta}))^{\lambda-1} X_{ti}(\boldsymbol{\beta}) \\ \mathbf{0} \end{bmatrix},$$

et le régresseur qui correspond à  $\lambda$  est

$$\begin{bmatrix} \frac{1}{\sigma \lambda^2} \left( (\lambda (x_t(\boldsymbol{\beta}))^\lambda \log(x_t(\boldsymbol{\beta})) - (x_t(\boldsymbol{\beta}))^\lambda + 1) - (\lambda y_t^\lambda \log y_t - y_t^\lambda + 1) \right) \\ \log y_t \end{bmatrix}.$$

Pour le modèle de Box-Cox conventionnel, les régresseurs qui correspondent à  $\beta_i$  et  $\gamma_j$ , respectivement, sont

$$\begin{bmatrix} \frac{1}{\sigma} B(X_{ti}, \lambda) \\ 0 \end{bmatrix} \quad \text{et} \quad \begin{bmatrix} \frac{1}{\sigma} Z_{tj} \\ 0 \end{bmatrix}, \quad (14.35)$$

et le régresseur qui correspond à  $\lambda$  est

$$\begin{bmatrix} \frac{1}{\sigma \lambda^2} \left( \sum_{i=1}^k \beta_i (\lambda X_{ti}^\lambda \log X_{ti} - X_{ti}^\lambda + 1) - (\lambda y_t^\lambda \log y_t - y_t^\lambda + 1) \right) \\ \log y_t \end{bmatrix}. \quad (14.36)$$

Nous avons maintenant obtenu des DLR pour les trois types les plus communs des modèles de Box-Cox. Des DLR pour d'autres types de modèles qui comprennent des transformations de la variable dépendante peuvent être dérivées de façon similaire. Toutes ces DLR peuvent être utilisées comme une partie clé des algorithmes pour estimer les modèles auxquels ils s'appliquent, exactement de la même manière que les GNR peuvent être utilisées comme une partie des algorithmes pour estimer les modèles de régression non linéaire; consulter la Section 6.8. Etant donnée une quelconque valeur de  $\lambda$  (très probablement 0 ou 1), il est facile d'obtenir des estimations initiales des paramètres restant du modèle par OLS ou NLS. Ceci fournit alors un ensemble complet d'estimations paramétriques, disons  $\boldsymbol{\theta}^{(1)}$ , auquel la DLR peut être évaluée au début. Les estimations des coefficients à partir de la DLR, disons  $\boldsymbol{t}^{(1)}$ , peuvent alors être utilisées pour déterminer la direction dans laquelle mettre à jour les estimations paramétriques, et le processus entier peut être répété autant de fois que nécessaire jusqu'à ce qu'une règle d'arrêt soit satisfaite.

La règle d'actualisation de l'algorithme de maximisation a la forme

$$\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)} + \alpha^{(j)} \mathbf{t}^{(j)}. \quad (14.37)$$

Ici  $\boldsymbol{\theta}^{(j)}$  et  $\boldsymbol{\theta}^{(j+1)}$  désignent les vecteurs des estimations sur les  $j^{\text{ième}}$  et  $(j+1)^{\text{ième}}$  itérations de l'algorithme de maximisation,  $\mathbf{t}^{(j)}$  désigne le vecteur des coefficients estimés à partir de la DLR, et  $\alpha^{(j)}$  désigne la longueur de pas, qui peut être choisie de diverses manières par l'algorithme. Cette règle d'actualisation ressemble à celle de la régression de Gauss-Newton discutées dans la Section 6.8, et fonctionne pour exactement la même raison. Un algorithme basé sur la méthode de Newton (avec une longueur de pas variable  $\alpha$ ) utiliserait la règle d'actualisation

$$\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)} - \alpha^{(j)} (\mathbf{H}(\boldsymbol{\theta}^{(j)}))^{-1} \mathbf{g}(\boldsymbol{\theta}^{(j)}). \quad (14.38)$$

La DLR à pas  $j$  fournit le vecteur de coefficients

$$\mathbf{t}^{(j)} = (\mathbf{R}^\top(\boldsymbol{\theta}^{(j)}) \mathbf{R}(\boldsymbol{\theta}^{(j)}))^{-1} \mathbf{R}^\top(\boldsymbol{\theta}^{(j)}) \mathbf{r}(\boldsymbol{\theta}^{(j)}).$$

D'après la propriété de toutes les régressions artificielles,  $\mathbf{t}^{(j)}$  est asymptotiquement égale à moins l'inverse du Hessian fois le gradient. Par conséquent, il est sensé remplacer  $-(\mathbf{H}(\boldsymbol{\theta}^{(j)}))^{-1} \mathbf{g}(\boldsymbol{\theta}^{(j)})$  dans (14.38) par  $\mathbf{t}^{(j)}$ . Ce qui donne (14.37), qui est la règle d'actualisation basée sur la DLR. La règle d'arrêt devrait normalement être fondée sur une certaine mesure du pouvoir explicatif de la DLR, cela a été discuté dans la Section 6.8.

La DLR peut, naturellement, être utilisée pour la mise en œuvre de tests d'hypothèse de n'importe lequel des modèles dont nous avons discuté. Puisque pour ces modèles la somme des carrés de la régressande est toujours  $2n$ , la quantité  $2n - \text{SSR}$  égalera toujours la somme expliquée des carrés, et elle fournit une statistique de test valide asymptotiquement qui est très facile à calculer. Comme d'habitude, les statistiques pseudo- $F$  et pseudo- $t$  basées sur la régression artificielle sont également valides asymptotiquement. Nous n'élaborerons pas ces sujets ici, puisque qu'il n'y a rien de nouveau pour en discuter; un cas spécifique sera discuté dans la prochaine section.

Il est peut être bon d'interjeter une petite mise en garde sur ce point. Si la régressande ou certains régresseurs dans une DLR qui est utilisée pour tester les hypothèses est construite de façon incorrecte, il est possible, et en effet très probable, que la régression fournira une statistique de test calculée grande et dénuée de sens. Contrôler la plupart des calculs en lançant tout d'abord la DLR *sans* ces régresseurs constitue alors une très bonne idée; ces derniers correspondent aux paramètres qui ont été testés. Cette régression, tout comme les régressions artificielles utilisées pour calculer les matrices de covariance, devrait n'avoir aucun pouvoir explicatif si tout a été construit correctement. Malheureusement, nous ne pouvons pas vérifier les régresseurs

de test de cette manière, et une erreur dans leur construction peut facilement conduire à des résultats incohérents. Par exemple, si nous rajoutons par inadvertance un terme constant à la DLR, il aura presque certainement un pouvoir explicatif substantiel sur la régressande, parce que la seconde moitié de cette dernière est simplement un vecteur composé de 1.

## 14.6 TEST DE MODÈLES LINÉAIRE ET LOGLINÉAIRE

Dans de nombreuses applications, la variable dépendante est toujours positive. Les économètres doivent alors décider si un modèle de régression devrait tenter d'expliquer l'espérance de la variable originale ou de son logarithme. Les deux types de modèles sont souvent plausibles *a priori*. Dans cette section, nous discutons de techniques de sélection entre modèles dans lesquels la régressande est le niveau ou le logarithme de la variable dépendante, et de techniques de test de leur spécification. Les tests basés sur la DLR s'avèrent très utiles pour ce propos.

Supposons initialement que les deux modèles sont linéaires en leurs paramètres. Ainsi, les deux modèles qui se font concurrence sont

$$y_t = \sum_{i=1}^k \beta_i X_{ti} + \sum_{j=1}^l \gamma_j Z_{tj} + u_t, \quad u_t \sim \text{NID}(0, \sigma^2), \quad \text{et} \quad (14.39)$$

$$\log y_t = \sum_{i=1}^k \beta_i \log X_{ti} + \sum_{j=1}^l \gamma_j Z_{tj} + u_t, \quad u_t \sim \text{NID}(0, \sigma^2), \quad (14.40)$$

où la notation, non pas par coïncidence, est la même que pour le modèle de Box-Cox conventionnel. Après l'estimation des deux modèles, il peut être possible de conclure que l'un d'eux devrait être rejeté simplement en comparant les valeurs de leurs fonctions de logvraisemblance, comme cela a été discuté dans la Section 14.3. Cependant, une telle procédure ne peut rien nous dire concernant la validité du modèle qui s'ajuste le mieux. Si les deux modèles sont raisonnables, il est important de les tester tous les deux avant de tenter d'en accepter un autre.

Il existe de nombreuses manières de tester la spécification des modèles de régression linéaire et non linéaire comme (14.39) et (14.40). Les tests les plus communément usités sont basés sur le fait que ceux-ci sont tous les deux des cas spéciaux du modèle de Box-Cox conventionnel

$$B(y_t, \lambda) = \sum_{i=1}^k \beta_i B(X_{ti}, \lambda) + \sum_{j=1}^l \gamma_j Z_{tj} + u_t, \quad u_t \sim \text{NID}(0, \sigma^2). \quad (14.41)$$

De façon conceptuelle, la manière la plus simple de tester (14.39) et (14.40) contre (14.41) consiste à estimer les trois modèles et d'employer un test LR,

comme cela fut suggéré à l'origine par Box et Cox (1964) dans contexte du modèle de Box-Cox simple. Cependant, comme l'estimation de (14.41) peut demander un certain effort, il peut être plus intéressant d'utiliser un test LM à sa place.

Plusieurs manières d'implémenter ce test LM sont valables. Nous ne mentionnerons seulement que celles basées sur les régressions artificielles, puisque celles-ci sont les plus simples à calculer, et, si un test LM n'est pas facile à calculer, il n'a aucun avantage sur le test correspondant LR. Il est évidemment possible de construire des tests LM de (14.39) et (14.40) contre (14.41) en utilisant soit la régression OPG soit la DLR. Les premiers tests tirent leur origine de Godfrey et Wickens (1981) et les derniers de Davidson et MacKinnon (1985c). Les derniers auteurs ont fourni des évidences Monte Carlo que les tests basés sur la DLR sont beaucoup plus performants en échantillons finis que ceux basés sur la régression OPG, une avancée confirmée plus tard par Godfrey, McAleer, et McKenzie (1988).

Il est intéressant de discuter à quoi ressemble la DLR pour tester les régressions linéaire et loglinéaire. Quand nous testons le modèle linéaire (14.39), l'hypothèse nulle est que  $\lambda = 1$ . Dans ce cas, la régressande de la DLR comporte le  $i^{\text{ième}}$  élément  $\hat{u}_t/\hat{\sigma}$  et le  $(t+n)^{\text{ième}}$  élément 1, où  $\hat{u}_t$  désigne le  $t^{\text{ième}}$  résidu issu du modèle linéaire et  $\hat{\sigma}$  désigne l'estimation ML de  $\sigma$ . Le  $t^{\text{ième}}$  et le  $(t+n)^{\text{ième}}$  éléments des régresseurs sont alors

pour  $\beta_i$  :  $X_{ti} - 1$  et 0;

pour  $\gamma_j$  :  $Z_{tj}$  et 0;

pour  $\sigma$  :  $\hat{u}_t/\hat{\sigma}$  et  $-1$ ;

pour  $\lambda$  :  $\sum_{i=1}^k \hat{\beta}_i (X_{ti} \log X_{ti} - X_{ti} + 1) - (y_t \log y_t - y_t + 1)$  et  $\hat{\sigma} \log y_t$ .

Ces régresseurs ne correspondent pas à ceux auxquels l'on pourrait s'attendre à avoir à partir de (14.33), (14.35), et (14.36), parce qu'ils ont tous été multipliés par  $\hat{\sigma}$ , quelque chose qui est sans effet parce qu'il ne change pas le sous-espace engendré par les colonnes de la matrice de régresseurs. Pour la même raison, si un des  $Z_{tj}$  est un terme constant, comme cela sera typiquement le cas, il n'est pas nécessaire de soustraire 1 des  $X_{ti}$ .

Quand nous testons le modèle loglinéaire (14.40), l'hypothèse nulle est que  $\lambda = 0$ . Dans ce cas, la régressande de la DLR comprend le  $i^{\text{ième}}$  élément  $\tilde{u}_t/\tilde{\sigma}$  et le  $(t+n)^{\text{ième}}$  élément 1, où  $\tilde{u}_t$  désigne le  $i^{\text{ième}}$  résidu provenant du modèle loglinéaire et  $\tilde{\sigma}$  désigne l'estimation ML de  $\sigma$ . Le  $i^{\text{ième}}$  et le  $(t+n)^{\text{ième}}$

éléments des régresseurs sont alors

pour  $\beta_i$  :  $\log X_{ti}$  et 0;

pour  $\gamma_j$  :  $Z_{tj}$  et 0;

pour  $\sigma$  :  $\tilde{u}_t/\tilde{\sigma}$  et  $-1$ ;

pour  $\lambda$  :  $\frac{1}{2} \sum_{i=1}^k \tilde{\beta}_i (\log X_{ti})^2 - \frac{1}{2} (\log y_t)^2$  et  $\tilde{\sigma} \log y_t$ .

Cette fois tous les régresseurs ont été multipliés par  $\tilde{\sigma}$ . Le régresseur pour  $\lambda$  avait trouvé son origine à l'aide de la Règle de l'Hôpital :

$$\lim_{\lambda \rightarrow 0} \left( \frac{\lambda x^\lambda \log x - x^\lambda + 1}{\lambda^2} \right) = \frac{1}{2} (\log x)^2.$$

Un test, qui est parfois confondu avec les tests LM tels que ceux dont nous venons juste de discuter, est un test proposé par Andrews (1971) et modifié par Godfrey et Wickens (1981) de telle sorte qu'il s'applique au modèle de Box-Cox conventionnel. L'idée est de prendre une approximation du premier ordre de (14.41) autour de  $\lambda = 0$  ou  $\lambda = 1$ , de réarranger les termes de telle sorte que seul  $\log y_t$  ou  $y_t$  apparaisse sur le membre de gauche, et alors remplacer  $y_t$  à chaque fois qu'il apparaît sur le membre de droite par les valeurs ajustées provenant de la régression sous test. Le résultat est quelque chose qui ressemble à la régression originale qui a été testée, avec l'addition d'un régresseur supplémentaire. Pour l'hypothèse nulle linéaire, ce régresseur supplémentaire est

$$\hat{y}_t \log \hat{y}_t - \hat{y}_t + 1 - \sum_{i=1}^k \hat{\beta}_i (X_{ti} \log X_{ti} - X_{ti} + 1),$$

et pour l'hypothèse loglinéaire il est

$$\frac{1}{2} \left( (\log \tilde{y}_t)^2 - \sum_{i=1}^k \tilde{\beta}_i (\log X_{ti})^2 \right),$$

où  $\hat{y}_t$  et  $\tilde{y}_t$  désignent les valeurs ajustées de  $y_t$  issues des modèles linéaire et loglinéaire respectivement. La statistique de test est simplement le  $t$  de Student sur le régresseur supplémentaire.

Le test de Andrews comporte une propriété plutôt remarquable. Si les  $X_{ti}$  et les  $Z_{tj}$  peuvent être traités comme exogènes, et si les aléas sont réellement normalement distribués, la statistique de test aura réellement la distribution de Student en échantillons finis. Ceci provient du fait que les régresseurs de test dépendent de  $y_t$  seulement au travers des estimations  $\hat{\beta}$  et  $\hat{\gamma}$  (ou  $\tilde{\beta}$  et  $\tilde{\gamma}$ ). L'argument est similaire à celui utilisé dans la Section 11.3 pour montrer que le test  $J_A$  est exact. Il s'ensuit à partir des mêmes résultats de Milliken et de Graybill (1970).

Cependant, le test d'Andrews n'est pas véritablement un test contre la même alternative que les tests LM. De façon implicite, il teste une direction de régression, contre une alternative qui est aussi un modèle de régression. Mais le modèle de Box-Cox (14.41) n'est pas un modèle de régression. Le test d'Andrews doit avoir par conséquent moins de puissance que les tests classiques des modèles linéaire et loglinéaire contre (14.41) quand le dernier a réellement généré les données. En utilisant des techniques similaires à celles discutées dans le Chapitre 12, il a été montré dans Davidson et MacKinnon (1985c) que, quand  $\sigma \rightarrow 0$ , le paramètre de non centralité pour le test d'Andrews s'approche de celui des tests classiques, tandis que, quand  $\sigma \rightarrow \infty$ , il s'approche de zéro. Ainsi, sauf quand  $\sigma$  est petit nous devrions nous attendre à ce que le test d'Andrews manque sérieusement de puissance, et les résultats Monte Carlo confirment cela. Cependant, un avantage possible du test d'Andrews devrait être noté. Contrairement aux tests LM dont nous avons discuté, il n'est pas sensible, asymptotiquement, aux caractéristiques de l'hypothèse de normalité, parce qu'il teste simplement une direction de régression.

Bien que les tests basés sur la transformée de Box-Cox soient très populaires, une seconde approche pour tester les modèles linéaire et loglinéaire mérite aussi d'être mentionnée. Elle traite les deux modèles comme des hypothèses non emboîtées, de la même manière que cela a été fait pour les tests discutés dans la Section 11.3. Cette approche non emboîtée permet de traiter des types plus généraux de modèle que l'approche basée sur la transformée de Box-Cox, parce qu'il n'est pas nécessaire que les deux modèles aient le même nombre de paramètres, ou qu'ils se ressemblent d'une quelconque manière, et il n'est pas nécessaire non plus qu'ils soient linéaires par rapport aux variables ou aux paramètres. Nous pouvons écrire les deux modèles en compétition comme

$$H_1: y_t = x_t(\beta) + u_{1t}, \quad u_{1t} \sim \text{NID}(0, \sigma_1^2), \quad \text{et} \quad (14.42)$$

$$H_2: \log y_t = z_t(\gamma) + u_{2t}, \quad u_{2t} \sim \text{NID}(0, \sigma_2^2). \quad (14.43)$$

Ici, la notation est similaire à celle qui a été utilisée pour le test d'hypothèses non emboîtées dans la Section 11.3 et devrait s'expliquer d'elle-même. Notons que l'hypothèse selon laquelle les aléas suivent une loi normale, dont nous n'avons pas besoin dans notre discussion précédente, est ici nécessaire.

Il existe deux manières évidentes de construire des tests non emboîtés pour les modèles comme (14.42) et (14.43). La première est de tenter d'implémenter les idées de Cox (1961, 1962), comme dans Aneuryn-Evans et Deaton (1980). Malheureusement, elle s'avère plutôt difficile. La seconde approche, beaucoup plus facile, consiste à les baser sur une certaine sorte d'emboîtement artificiel. Considérons (quelque peu arbitrairement) le modèle composite artificielle

$$H_C: (1 - \alpha) \left( \frac{y_t - x_t(\beta)}{\sigma_1} \right) + \alpha \left( \frac{\log y_t - z_t(\gamma)}{\sigma_2} \right) = \varepsilon_t, \quad (14.44)$$

où les hypothèses sur  $u_{1t}$  et  $u_{2t}$  impliquent que  $\varepsilon_t$  est  $N(0, 1)$ . Comme les modèles composites artificiels introduits dans la Section 11.3, ce dernier modèle ne peut pas être estimé parce que de nombreux paramètres seront en général non identifiés. Cependant, en suivant la procédure utilisée pour obtenir les tests en  $J$  et  $P$ , nous pouvons remplacer les paramètres du modèle qui n'est pas testé par des estimations. Ainsi, si nous désirons tester  $H_1$ , nous pouvons remplacer  $\gamma$  et  $\sigma_2$  par les estimations ML  $\hat{\gamma}$  et  $\hat{\sigma}_2$  de telle sorte que  $H_C$  devienne

$$H'_C: (1 - \alpha) \left( \frac{y_t - x_t(\boldsymbol{\beta})}{\sigma_1} \right) + \alpha \left( \frac{\log y_t - z_t(\hat{\gamma})}{\hat{\sigma}_2} \right) = \varepsilon_t.$$

Il est simple de tester  $H_1$  contre  $H'_C$  au moyen de la DLR:

$$\begin{bmatrix} \frac{(y_t - \hat{x}_t)}{\hat{\sigma}_1} \\ 1 \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{X}}_t & \frac{(y_t - \hat{x}_t)}{\hat{\sigma}_1} & \hat{z}_t - \log y_t \\ \mathbf{0} & -1 & \hat{\sigma}_1/y_t \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ s \\ a \end{bmatrix} + \text{résidus}, \quad (14.45)$$

où  $\hat{x}_t \equiv x_t(\hat{\boldsymbol{\beta}})$ ,  $\hat{\mathbf{X}}_t \equiv \mathbf{X}_t(\hat{\boldsymbol{\beta}})$ , et  $\hat{z}_t \equiv z_t(\hat{\gamma})$ . La DLR (14.45) est en fait une version simplifiée de la DLR que l'on obtient initialement. Tout d'abord,  $\hat{\sigma}_1$  fois le régresseur d'origine pour  $\sigma_1$  a été additionné au régresseur d'origine pour  $\alpha$ . Alors les régresseurs qui correspondent à  $\boldsymbol{\beta}$  et à  $\sigma_1$  ont été multipliés par  $\hat{\sigma}_1$ , et les régresseurs qui correspondent à  $\alpha$  ont été multipliés par  $\hat{\sigma}_2$ . Aucune des ces modifications n'affecte le sous-espace engendré par les régresseurs, et par conséquent, aucun d'entre eux n'affecte les statistiques de test qu'on obtient. La dernière colonne de la matrice des régresseurs dans (14.45) est celle qui correspond à  $\alpha$ . Les autres colonnes devraient être orthogonales à la régressande par construction.

De façon similaire, si nous espérons tester  $H_2$ , nous pouvons remplacer  $\boldsymbol{\beta}$  et  $\sigma_1$  par les estimations ML  $\hat{\boldsymbol{\beta}}$  et  $\hat{\sigma}_1$  de telle sorte que  $H_C$  devienne

$$H''_C: (1 - \alpha) \left( \frac{y_t - x_t(\hat{\boldsymbol{\beta}})}{\hat{\sigma}_1} \right) + \alpha \left( \frac{\log y_t - z_t(\gamma)}{\sigma_2} \right) = \varepsilon_t.$$

Il est alors simple de tester  $H_2$  contre  $H''_C$  au moyen de la DLR

$$\begin{bmatrix} \frac{\log y_t - \hat{z}_t}{\hat{\sigma}_2} \\ 1 \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{Z}}_t & \frac{\log y_t - \hat{z}_t}{\hat{\sigma}_2} & \hat{x}_t - y_t \\ \mathbf{0} & -1 & \hat{\sigma}_2 y_t \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ s \\ a \end{bmatrix} + \text{résidus}. \quad (14.46)$$

Encore une fois, ceci est une version simplifiée de la DLR que l'on obtient au début, et la dernière colonne de la matrice du régresseur est celle qui correspond à  $\alpha$ .

Les tests dont nous venons juste de discuter se généralisent bien évidemment très facilement aux modèles qui comprennent n'importe quelle sorte de

transformation de la variable dépendante, y compris les modèles de Box-Cox et d'autres modèles dans lesquels la transformation dépend d'un ou plusieurs paramètres inconnus. Pour plus de détails, consulter Davidson et MacKinnon (1984a). Il devrait être bien précisé que le modèle composite artificiel (14.44) est très arbitraire. Contrairement au modèle qui semble similaire pour les modèles de régression qui a été employé dans la Section 11.3, il ne fournit pas des tests asymptotiquement équivalents aux tests de Cox. De plus, peu de choses sont connues concernant les propriétés en échantillons finis des tests basés sur les DLR comme (14.45) et (14.46).

Une procédure finale qu'il est bon de mentionner est le **test  $P_E$**  suggéré par MacKinnon, White, et Davidson (1983). Il part aussi du modèle composite artificiel (14.44) mais ensuite il suit essentiellement l'approche du test d'Andrews de façon à obtenir une GNR qui teste seulement une direction de régression. Les régressions de test de Gauss-Newton pour le test  $P_E$  sont

$$y_t - \hat{x}_t = \hat{\mathbf{X}}_t \mathbf{b} + a(\hat{z}_t - \log \hat{x}_t) + \text{résidu} \quad (14.47)$$

pour le test de  $H_1$  et

$$\log y_t - \hat{z}_t = \hat{\mathbf{Z}}_t \mathbf{c} + d(\hat{x}_t - \exp \hat{z}_t) + \text{résidu} \quad (14.48)$$

pour le test de  $H_2$ . Les statistiques de test les plus simples à utiliser sont les  $t$  de Student pour  $a = 0$  dans (14.47) et  $d = 0$  dans (14.48). Comme le test d'Andrews, le test  $P_E$  manque très probablement de puissance, sauf quand la variance du DGP est très petite. Son tout premier avantage est que, contrairement aux tests basés sur la DLR, il sera asymptotiquement insensible aux caractéristiques de l'hypothèse de normalité.

## 14.7 LES AUTRES TRANSFORMATIONS

Les modèles basés sur la transformée de Box-Cox ne fonctionneront pas de façon adéquate à chaque fois. En particulier, le modèle de Box-Cox conventionnel n'est pas souvent très satisfaisant, pour des raisons que nous allons discuter. Dans cette section, nous discutons brièvement d'un nombre d'autres transformations qui peuvent être utiles dans certains cas. Nous n'en dirons pas beaucoup concernant les méthodes d'estimation et d'inférence pour ces modèles, sauf de noter qu'elles peuvent toutes être estimées par maximum de vraisemblance, en utilisant la DLR comme une partie de l'algorithme de maximisation, et que la DLR peut toujours être utilisée pour calculer des matrices de covariance et des statistiques de tests.

Un problème majeur avec le modèle de Box-Cox conventionnel est que le paramètre de transformation  $\lambda$  joue deux rôles différents: il modifie les propriétés des résidus, et change aussi la forme fonctionnelle de la fonction de régression. Par exemple, supposons que le DGP soit réellement un modèle de

régression linéaire à erreurs hétéroscédastiques qui ont une variance proportionnelle au carré de l'espérance de la variable dépendante :

$$y_t = \mathbf{X}_t \boldsymbol{\beta}_0 + u_t, \quad u_t \sim N(0, \sigma_0^2 (\mathbf{X}_t \boldsymbol{\beta}_0)^2), \quad (14.49)$$

où  $\sigma_0$  et  $\boldsymbol{\beta}_0$  désignent des valeurs sous le DGP. Si nous estimions un modèle de Box-Cox conventionnel en utilisant les données générées de cette manière, nous obtiendrions presque certainement une estimation de  $\lambda$  qui serait inférieure à l'unité, parce que ceci réduirait l'hétéroscédasticité dans les résidus. Ainsi, nous pourrions conclure de façon incorrecte qu'une spécification linéaire était inappropriée ou même qu'une spécification loglinéaire était inappropriée.

Le problème est que le paramètre de transformation dans le modèle de Box-Cox conventionnel affecte à la fois la forme de la fonction de régression et l'hétéroscédasticité des résidus. Une solution évidente est de permettre de façon explicite l'hétéroscédasticité, comme dans le modèle

$$B(y_t, \lambda) = \sum_{i=1}^k \beta_i B(X_{ti}, \lambda) + \sum_{j=1}^l \gamma_j Z_{tj} + u_t, \quad u_t \sim N(0, \sigma^2 h(\mathbf{w}_t \boldsymbol{\delta})),$$

où  $h(\cdot)$  est une fonction scédastique,  $\mathbf{w}_t$  est un vecteur des observations des variables dépendantes, et  $\boldsymbol{\delta}$  est un vecteur des paramètres à estimer. Si on est tout d'abord intéressé par l'hétéroscédasticité de la forme qui apparaît dans (14.49), la fonction scédastique  $h(\mathbf{w}_t \boldsymbol{\delta})$  pourrait être spécifiée comme  $(\mathbf{X}_t \boldsymbol{\beta})^2$ . Consulter, parmi d'autres, Gaudry et Dagenais (1979), Lahiri et Egy (1981), et Tse (1984).

Une autre possibilité est de permettre qu'il y ait plus d'un paramètre de transformation, comme dans les modèles

$$B(y_t, \lambda) = \sum_{i=1}^k \beta_i B(X_{ti}, \phi) + \sum_{j=1}^l \gamma_j Z_{tj} + u_t \quad \text{et} \quad (14.50)$$

$$B(y_t, \lambda) = B\left(\left(\sum_{i=1}^k \beta_i B(X_{ti}, \phi) + \sum_{j=1}^l \gamma_j Z_{tj}\right), \lambda\right) + u_t, \quad (14.51)$$

où, dans les deux cas,  $u_t$  est supposé  $N(0, \sigma^2)$ . Le premier de ces modèles est une généralisation évidente du modèle de Box-Cox conventionnel, et a été utilisé un certain nombre de fois en économétrie, parfois avec plus d'un paramètre  $\phi$ . Le second combine le modèle de Box-Cox conventionnel avec le modèle transformé des deux côtés, et n'a été utilisé dans aucun domaine. Dans les deux cas, le paramètre  $\phi$  affecte tout d'abord la forme fonctionnelle de la fonction de régression, tandis que le paramètre  $\lambda$  affecte tout d'abord les propriétés des aléas. Naturellement, il est loin d'être clair de savoir lequel des modèles (14.50), (14.51), et le modèle de Box-Cox conventionnel, sera le plus performant dans un cas donné.

Comme nous l'avons vu, la transformée de Box-Cox ne peut pas être appliquée aux variables qui peuvent prendre une valeur nulle ou négative. De nombreux auteurs, incluant John et Draper (1980) et Bickel et Doksum (1981), ont proposé des manières pour l'étendre de telle sorte qu'elle puisse être utilisée dans de tels cas. Par exemple, la proposition de Bickel-Doksum consiste à utiliser la transformation

$$\frac{\text{sign}(y)|y|^\lambda - 1}{\lambda} \quad (14.52)$$

à la place de la transformée de Box-Cox. Il est logiquement possible d'appliquer (14.52) aux variables qui peuvent prendre de petites valeurs (mais non nulles) et à celles qui prennent des valeurs négatives. Cependant, cette transformation ne comporte pas des propriétés particulièrement attrayantes; consulter Magee (1988). Quand  $\lambda$  est petit, elle a une pente extrêmement forte pour  $y$  proche de zéro. En plus, quand  $y < 0$ , (14.52) n'a pas de limite quand  $\lambda \rightarrow 0$ .

Il n'existe pas de raison de restreindre l'attention aux versions modifiées de la transformée de Box-Cox, puisque d'autres transformations peuvent bien être plus appropriées pour certains types de données. Par exemple, quand  $y_t$  est contrainte à rester comprise entre zéro et un, un modèle comme

$$y_t = \mathbf{X}_t(\boldsymbol{\beta}) + u_t, \quad u_t \sim N(0, \sigma^2),$$

ne comporte réellement pas de sens, parce qu'il existe toujours une chance que  $u_t$  soit si grand que  $y_t$  tombe en dehors de l'intervalle 0-1. Dans un tel cas, il peut être souhaitable d'employer la transformation

$$\tau(y) = \log\left(\frac{y}{1-y}\right),$$

puisque  $\tau(y)$  peut varier entre moins l'infini et plus l'infini. Cette transformation ne comporte pas de paramètre inconnu, et ainsi ne nécessite pas que l'on quitte le cadre des modèles de régression; consulter Cox (1970).

Une famille intéressante de transformations a été étudiée par Burbidge, Magee, et Robb (1988) et MacKinnon et Magee (1990). Ces transformations ont la forme  $\theta(\alpha y)/\alpha$ , où la fonction  $\theta(\cdot)$  est croissante en ses arguments et possède les propriétés suivantes :

$$\theta(0) = 0; \quad \theta'(0) = 1; \quad \theta''(0) \neq 0. \quad (14.53)$$

Contrairement à la transformée de Box-Cox, cette transformation peut être appliquée aux variables d'un autre signe et aux variables nulles. Certaines fonctions  $\theta(\cdot)$  possèdent les propriétés (14.53). Une des plus simples est la fonction  $y + y^2$ , pour laquelle la transformation serait

$$\frac{\theta(\alpha y)}{\alpha} = y + \alpha y^2. \quad (14.54)$$

Evidemment, celle-ci sera une fonction convexe de  $y$  quand  $\alpha$  est une fonction positive et concave quand  $\alpha$  est négative. N'importe quelle transformation de la forme  $\theta(\alpha y)/\alpha$  qui satisfait (14.53) sera localement équivalente à (14.54), et ainsi nous voyons qu'un test de  $\alpha = 0$  peut être interprété comme un test contre n'importe quelle forme de non linéarité quadratique locale.

Pour cette famille de transformation, le modèle (14.04) deviendrait

$$\frac{\theta(\alpha y_t)}{\alpha} = x_t(\boldsymbol{\beta}) + u_t, \quad u_t \sim \text{NID}(0, \sigma^2).$$

Il est facile de tester l'hypothèse nulle que  $\alpha = 0$  en utilisant une DLR très similaire à celle utilisée pour tester l'hypothèse nulle selon laquelle  $\lambda = 1$  dans un modèle de Box-Cox simple (14.04); consulter MacKinnon et Magee (1990) pour plus de détails. Ce test est sensible à plusieurs formes communes de mauvaises spécifications de modèle, qui incluent la non linéarité dans la fonction de régression, l'hétéroscédasticité, et l'asymétrie. Cela tend à être étroitement relié au test bien connu RESET; consulter la Section 6.5. On obtiendrait le test RESET si la transformation s'appliquait à  $x_t(\boldsymbol{\beta})$  à la place de  $y_t$ , comme dans le modèle

$$y_t = \frac{\theta(\alpha x_t(\boldsymbol{\beta}))}{\alpha} + u_t, \quad u_t \sim \text{NID}(0, \sigma^2).$$

Puisqu'il s'agit simplement d'un modèle de régression non linéaire, un test pour  $\alpha = 0$  peut être basé sur une GNR. Il s'agit simplement du  $t$  de Student pour  $a = 0$  dans

$$y_t - x_t(\hat{\boldsymbol{\beta}}) = \mathbf{X}_t(\hat{\boldsymbol{\beta}})\mathbf{b} + a x_t^2(\hat{\boldsymbol{\beta}}) + \text{résidu},$$

qui est une forme du test RESET.

## 14.8 CONCLUSION

À l'exception du modèle de Box-Cox conventionnel, les modèles qui comprennent des transformations de la variable dépendante ont été plutôt rarement utilisés en économétrie. Ceci est surprenant, parce qu'ils fournissent souvent une manière simple et peu coûteuse d'obtenir un modèle pour lequel les résidus se comportent bien et il existe une importante littérature les concernant en statistique, comprenant des ouvrages de McCullagh et Nelder (1983), Atkinson (1985), et Carroll et Ruppert (1988).

Nous avons vu dans ce chapitre qu'il n'est pas du tout difficile de traiter des modèles de ce type. Pourvu que l'on veuille supposer la normalité — et une telle hypothèse de distribution semble être nécessaire du moment que l'on quitte le cadre des modèles de régression — il est simple de les estimer par maximum de vraisemblance. La régression artificielle à longueur double

est extrêmement utile dans le contexte de ces modèles. Tout ce que l'on peut faire avec la régression de Gauss-Newton pour les modèles de régression non linéaire peut être réalisé avec la DLR pour les modèles qui comprennent des transformations de la variable dépendante. La régression OPG peut être utilisée à la place de la DLR, mais sera généralement moins performante.

## TERMES ET CONCEPTS

algorithmes de maximisation utilisant la DLR	régression artificielle (formulation générale)
facteurs Jacobiens	régressions linéaire contre loglinéaire
modèles de Box-Cox : conventionnel, simple, et transformé des deux côtés	termes Jacobiens
modèles autres que les régressions	tests non emboîtés
observations artificielles (pour DLR)	test $P_E$
régression artificielle à longueur double (DLR)	test RESET
	transformée de Box-Cox
	transformation non linéaire