The Stata Journal (2008)
8, Number 4, pp. 574–578

Stata par la pratique : statistiques, graphiques et éléments de programmation par Éric Cahuzac et Christophe Bontemps

Antoine Terracol
EQUIPPE, Universités de Lille
CES, Université Paris 1
Paris, France
terracol@univ-paris1.fr

Résumé. Cet article propose une revue de *Stata par la pratique* par Éric Cahuzac et Christophe Bontemps

Keywords: gn0042, apprentissage de Stata, Cahuzac, Bontemps, splp

1 Introduction

Avec *Stata par la pratique : statistiques, graphiques et éléments de programmation*, Éric Cahuzac et Christophe Bontemps (2008) offrent au lecteur francophone l'occasion de se familiariser avec Stata dans leur langue maternelle. À ma connaissance, le seul autre ouvrage disponible en français est celui de Bocquier (1998). La spectaculaire évolution de Stata au cours de ces dix dernières années rend particulièrement bienvenue la parution du livre de Cahuzac et Bontemps.

Ce livre s'adresse à des lecteurs déjà familiers avec les méthodes statistiques abordées dans les divers chapitres. Les éléments théoriques sont donc limités au strict nécessaire, les auteurs se concentrant sur la mise en œuvre de ces méthodes à l'aide de Stata. Les praticiens ainsi que les étudiants de second ou troisième cycle trouveront dans ce livre un guide utile pour la prise en main du logiciel, ainsi que de nombreux exemples de résolution de problèmes pratiques.

Le livre s'organise en huit chapitres, commençant par la prise en main du logiciel avant d'aborder la manipulation et la description des données, la modélisation et l'inférence puis l'analyse graphique. Un chapitre est consacré à la gestion des tableaux de résultats, puis des éléments de programmation sont introduits. Le chapitre final propose une série de solutions à des problèmes couramment rencontrés dans la pratique.

2 Contenu de l'ouvrage

Dans l'avant-propos, les auteurs indiquent que les huit chapitres du livre se veulent indépendants, et permettent de répondre à des problèmes types. Cependant, le novice sera bien avisé de suivre la progression des chapitres pour aborder au fur et à mesure des problèmes de plus en plus complexes (même si, par exemple, les chapitres 4 et 5 peuvent être intervertis sans dommages pour la compréhension).

Le premier chapitre présente rapidement le logiciel, les différentes fenêtres (le lecteur regrettera peut-être l'absence de copie d'écran présentant l'espace de travail) et les boutons les plus utiles (*Viewer*, *Browser*, *Do-file Editor*, etc.). Une brève description des fichiers que l'utilisateur de Stata sera amené à manipuler, des indications sur la mise à jour du logiciel, sur l'installation de nouvelles commandes et sur quelques raccourcis utiles sont ensuite proposés. Ce chapitre est celui qui m'a semblé le moins abouti du livre. Le lecteur débutant aurait en effet besoin de plus d'informations afin de pouvoir pleinement comprendre les commandes présentées dans les chapitres suivants, ou de pouvoir facilement utiliser d'autres commandes. Il manque en particulier une explication générale de la structure des commandes de Stata (afin, par exemple, de savoir que l'on doit placer les options après une virgule[1]) ainsi qu'un guide de lecture des diagrammes de syntaxe des fichiers d'aide (afin de pouvoir distinguer les options « obligatoires » des options « optionnelles »). La discussion des fichiers do et de leur utilité (reléguée au chapitre 7) aurait gagné à être déplacée au chapitre 1. Le lecteur débutant risquant de croire que, étant présentés dans un des derniers chapitres, les fichiers do sont réservés à un usage plus avancé ; et que la ligne de commandes est la façon la plus appropriée pour lui d'interagir avec Stata. De même, la présentation des fichiers Stata Markup and Control Language et log aurait gagnée à être placée dans ce premier chapitre, et non dans l'avant dernier.

Le second chapitre traite de la manipulation des données et aborde l'importation de bases de données (par exemple au format ASCII), leur description et leur modification (incluant la fusion et la superposition de bases de données, ansi que la modification de bases à l'aide de `collapse` ou `reshape`.). Il couvre l'essentiel des commandes utiles aux praticiens, et même plus. Les commandes sont groupées par thématique et une description de chacune, parfois avec un exemple d'output, est fournie. Le lecteur se voit présenter des concepts utiles pour un usage plus avancé (scalaires, macros et matrices) de façon assez claire.[2] Quelques exemples relativement simples permettent ensuite au lecteur de se familiariser avec leur utilisation. Quelques détails éditoriaux nuisent cependant à la lisibilité du chapitre. Par exemple, l'opérateur d'égalité (`==`) est utilisé pour la première fois quelque pages avant le tableau présentant la liste des opérateurs.

1. Une option est ainsi utilisée pour la première fois au chapitre 2 en page 9, mais le fait que l'on doive séparer les options du reste de la commande avec une virgule n'est précisé qu'en page 125.
2. Dans leur comparaisons entre scalaires et macros, les auteurs omettent de préciser que les scalaires stockent les valeurs numériques de façon un peu plus précise que les macros, et que leur utilisation est donc recommandée pour le stockage de valeurs intermédiaires. Voir [U] **18.5 Scalars and matrices**.

Le chapitre 3 aborde les commandes de statistique descriptive, de tests usuels et d'analyse des données. Le lecteur se voit tout d'abord présenter les commandes permettant l'exploration de données univariées, discrètes et continues, puis les tests usuels de normalité et des tests d'association entre variables : corrélations, χ^2, comparaisons de distributions, etc. La seconde partie de ce chapitre traite de l'analyse multivariée : analyse de la variance et de la covariance, analyse en composantes principales, analyse factorielle, classification automatique, etc. Alors que la première partie du chapitre laisse parfois l'impression d'être un catalogue de commandes, cette seconde partie rentre beaucoup plus dans le détail et propose d'intéressantes discussions sur les diverses méthodes présentées, ainsi que des exemples d'analyse détaillés et commentés. Le titre de l'ouvrage prend alors tout son sens, puisque le lecteur peut effectivement apprendre Stata *par la pratique* en étant guidé par l'interprétation détaillée des résultats des commandes proposées.

Le chapitre 4 s'intéresse aux commandes de modélisation et d'estimation. Il aborde en premier lieu la régression linéaire, puis se concentre sur les modèles à variable dépendante limitée et aux données de comptage. Tout comme la seconde partie du chapitre 3, ce chapitre présente une analyse détaillée des commandes, de l'interprétation des résultats obtenus, et des procédures de diagnostics post-estimation (tests, analyse des résidus, etc.). La présentation des modèles probit, logit et multinomiaux, en particulier, est extrêmement complète et, plus qu'une simple introduction, permet au lecteur de se familiariser avec les outils nécessaires à une analyse fine de ce type de données et à leur modélisation. On pourra toutefois regretter l'absence d'information sur la structure générale des commandes d'estimation et sur les outils d'analyse post-estimation disponibles après toutes les commandes d'estimation. On remarquera également que les auteurs traitent de la commande `estimates store` dans ce chapitre, mais n'évoquent la très proche commande `estimates save` qu'au chapitre 7.

Le chapitre 5, qui traite des graphiques, évite les écueils des chapitres précédant en présentant d'emblée une structure générale de la syntaxe des commandes de graphiques. Le lecteur pourra ainsi plus facilement créer des graphiques différents de ceux présentés dans ce chapitre. Une fois la logique de la syntaxe expliquée, les auteurs présentent les options les plus courantes permettant de légender les axes, de changer les symboles et les styles de ligne, etc. Le chapitre couvre ensuite les graphiques univariés et permettant de représenter certaines statistiques descriptives ; puis se concentre sur les graphiques bi-dimensionnels (`twoway`). Enfin, la superposition de graphes ainsi que des fonctions plus avancées comme la gestion séparée des axes des ordonnées, ou la représentation de régressions paramétriques ou non paramétriques sont abordées. La description de chaque commande ou option est complétée par un exemple de graphique que l'on peut obtenir de cette façon. Mon expérience d'enseignement de Stata m'a appris qu'il était difficile de présenter clairement les vastes possibilités graphiques de Stata de façon concise mais complète ; et je pense que Cahuzac et Bontemps ont réussi cet exercice périlleux.

Le chapitre 6 est le plus original de l'ouvrage. Il présente au lecteur les diverses commandes (le plus souvent écrites par des utilisateurs) permettant d'automatiser la création de tableaux de statistiques descriptives et de résultats d'estimation. Les auteurs se concentrent principalement sur les sorties au format LaTeX largement utilisé dans la

communauté scientifique, mais ne négligent pas pour autant les formats de logiciels commerciaux. La lecture de ce chapitre permettra aux utilisateurs de gagner un temps précieux lors de la rédaction de leurs rapports, et leur évitera de fastidieux exercices de copier/coller.

Dans le chapitre 7, les auteurs présentent les principaux éléments permettant de se familiariser avec la programmation sous Stata. Après un certain nombre de conseils généraux sur la façon de programmer efficacement, le chapitre aborde la façon de créer et de manipuler les macros locales et globales, les boucles et le préfixe `bysort`, puis l'utilisation des résultats sauvegardés par les commandes antérieures. Enfin, la manipulation de fichiers do et de programmes (au sens strict) sous la forme de fichiers ado sont explicitées. Si la première partie du chapitre est claire et contient des informations fort utiles[3], la seconde partie m'a laissé une impression plus mitigée. La section concernant les procédures (fichiers ado) est à mon avis trop succincte et aurait mérité un plus ample développement. Par exemple, l'utilisation de `syntax` n'est pas du tout explicitée, bien qu'utilisée dans les exemples. De même, la notion de classe (e-class et r-class) est présentée, mais les auteurs n'expliquent pas comment créer un programme d'une classe donnée, ni comment sauvegarder des résultats dans des macros `e()` ou `r()`. Le lecteur risquant au final de se trouver incapable de créer une commande ado fonctionnelle.[4] Le chapitre 7 fournit cependant au lecteur des bases suffisamment solides pour pouvoir écrire des fichiers do efficients et compacts, et pour pouvoir comprendre, adapter ou modifier les exemples présentés au chapitre suivant.

Finalement, le chapitre 8 propose une série de petits programmes commentés permettant de résoudre des problèmes couramment rencontrés dans la pratique. L'étude attentive de ces programmes permettra au lecteur de comprendre la bonne façon d'aborder les problèmes similaires qu'il ne manquera pas de rencontrer dans ses recherches.

3 Conclusion

L'ouvrage d'éric Cahuzac et Christophe Bontemps se donne pour objectif de permettre au lecteur d'acquérir une autonomie rapide et de lui donner les principes de base d'une utilisation aisée du logiciel. Cet objectif est très largement atteint, et la communauté des utilisateurs francophones de Stata dispose d'un nouvel outil de référence, d'apprentissage et d'enseignement qui faisait défaut. Cahuzac et Bontemps parviennent à présenter un grand nombre de concepts et de commandes sans pour autant noyer le lecteur sous un déluge d'information. Au contraire, le débutant trouvera des explications claires et rigoureuses (sans être techniques) sur la façon d'interpréter et d'analyser ses résultats. Les exemples du chapitre 8 permettront également aux novices de comprendre comment les commandes vues aux chapitres précédents permettent de résoudre efficacement un grand nombre de problèmes. Enfin, l'accent est souvent mis sur des commandes téléchargeables, soulignant ainsi une des grandes forces de Stata : son ouverture

3. Y compris pour un utilisateur qui, comme moi, utilise Stata depuis presque 10 ans.
4. D'autant plus que les auteurs ne précisent pas qu'il faut impérativement donner au fichier ado le même nom que la procédure qui y est définie.

sur la toile. Bien entendu, nul ouvrage n'est parfait, et certains choix d'organisation du contenu peuvent sembler discutables. Par exemple, certains éléments qui me semblent former un groupe cohérent sont parfois discutés dans des chapitres distincts. Ainsi, les macros `_n` et `_N` sont introduites au chapitre 2, le préfixe `bysort` au chapitre 7, et leur utilisation conjointe (qui est une combinaison particulièrement puissante) n'apparat que dans un exemple au chapitre 8. Malgré certains défauts, *Stata par la pratique*, je n'en doute pas, trouvera sa place sur les étagères de nombreux praticiens.

4 Références

Bocquier, P. 1998. *L'essentiel de Stata.* Paris : Ritme Informatique - Global Design.

Cahuzac, É., et C. Bontemps. 2008. *Stata par la pratique : statistiques, graphiques et éléments de programmation.* College Station, TX : Stata Press.

À propos de l'auteur

Antoine Terracol est Maître de confrences en économie à l'Université Charles-de-Gaulle – Lille 3. Ses thmes de recherche incluent l'économie du travail, l'évaluation des politiques publiques, ainsi que l'économie comportementale. Il utilise Stata dans le cadre de ses recherches depuis 1999 et l'enseigne occasionnellement.