Revista de Economía Aplicada E A *Número 28 (vol. X), 2002, págs. 163 a 174*

Stata 7.0 para windows*

MARÍA E. ROCHINA BARRACHINA
JUAN A. SANCHIS LLOPIS
Universitat de València

Stata es un programa estadístico muy potente para analizar, manejar y representar gráficamente datos[1]. Permite realizar un elevado número de operaciones que van desde la simple manipulación de los mismos hasta la aplicación de técnicas de estimación complicadas que posibilitan, por ejemplo, la estimación de modelos con datos de series temporales, datos de corte transversal y datos de panel.

En la actualidad Stata es un programa ampliamente extendido tanto en ámbito geográfico como por tipo de usuario (instituciones académicas, centros de investigación, organismos públicos y empresas). Además, existen libros recientes que utilizan el programa Stata para ilustrar sus aplicaciones empíricas. Entre ellos cabe destacar Deaton (1997), Hamilton (1997), Hardin y Hilbe (2001), Long (1997) y Rabe-Heskth y Everitt (2000). La principal ventaja sobre sus más cercanos competidores (por ejemplo Limdep, SAS, SPSS, TSP y Eviews) reside en su facilidad de uso y en su velocidad en el proceso de datos. La rapidez en el manejo de datos se debe, por una parte, a una programación eficiente e inteligente y, por otra, a que mantiene los datos en memoria RAM mientras se está trabajando con ellos. Este sistema también actúa como un sistema de seguridad que impide reemplazar una base de datos hasta que de un modo explícito se decida grabar sobre ella. Aunque Stata reserva por defecto un MB de memoria para los datos, el programa es enormemente flexible y permite tanto cambios en los límites de memoria (para manejar bases de datos muy grandes) como el uso de formatos extremadamente comprimidos para almacenar datos.

Además de la velocidad y la facilidad de manejo, otra gran ventaja de Stata es que cubre todas las etapas de la investigación empírica: edición y manipulación de datos, ejecución de órdenes de estimación de modelos econométricos y de contrastes de hipótesis y, por último, creación de gráficos y tablas.

(*) Agradecemos los comentarios y sugerencias de Enrique Sentana.

(1) El programa Stata lo desarrolla STATA Corporation, 4905 Lakeway Drive, College Station, Texas 77845, USA. Tel. 1-979-696-4600, fax 1-976-696-4601, e-mail stata@stata.com, http://www.stata.com. En España Stata es distribuido por Timberlake Consulting S.L., C/ Méndez Núñez número 1, 3.º, 41001 Sevilla. Tel./fax 95 422 0648, e-mail timberlake@zoom.es. Los precios varían según el tipo de usuario (único o múltiple) y la modalidad del programa (Small Stata, Intercooled Stata y Stata/SE). En la página *web* de Stata *http://www.stata.com/info/order* se dispone de información actualizada sobre precios.

Requerimientos de los equipos

El programa Stata está disponible para diversos entornos (UNIX, Windows y Macintosh). Las versiones para los tres son esencialmente idénticas. La instalación de la versión para Windows ME/98/95/2000/NT (que es la que nosotros utilizamos) es muy sencilla y se realiza mediante un CD de instalación. Existen tres "variedades" de Stata para windows, Small Stata, Intercooled Stata y Stata/SE (Special Edition). Small Stata requiere ordenadores con un mínimo de 16 MB de memoria RAM, Intercooled Stata precisa de 32 MB y Stata/SE de 64 MB.

Intercooled Stata y Stata/SE representan las modalidades profesionales de Stata, las más rápidas y potentes. Small Stata es una modalidad limitada. La diferencia fundamental entre las tres variedades es el tamaño de la base de datos que permiten analizar. Mientras que Small Stata permite manejar un número pequeño de variables (99) y de observaciones (1000), Intercooled Stata permite manejar 2.047 variables, Stata/SE 32.766, y el límite en el número de observaciones de estas dos últimas variedades viene determinado por la cantidad de memoria RAM del equipo. Otra diferencia entre las modalidades es el número máximo de variables independientes que se permite incorporar en las órdenes de estimación (11.000 en Stata/SE, 800 en Intercooled Stata y 38 en Small Stata).

Incorporaciones al programa hasta su versión actual

El uso de Stata se ha extendido de un modo tan sustancial que ha permitido la aparición de nuevas versiones, más completas en contenido y más potentes en velocidad y límites. Dada la rapidez con la que se actualiza el programa, la probabilidad de que el usuario necesite cambiar a otro programa es decreciente. En 1994 se utilizaba la versión de Stata 3.1 (para el sistema operativo DOS). Esta versión de Stata fue revisada por Ferral (1994) y por Mestre (1994). En 1995 se introdujo la versión 4.0, que ya incluía una variante para Windows. Esta versión fue revisada por Banks (1996). En 1997 aparece Stata 5.0, versión que consolida la opción de Stata para Windows como el producto estable y estándar para PCs. Una revisión de esta versión la encontramos en Rees (1998). En 1999 aparece en el mercado Stata 6.0 (en esta versión desaparece ya la modalidad de Stata para el entorno DOS) y en 2001 la versión que está actualmente vigente, Stata 7.0. A continuación presentamos de forma breve las sucesivas incorporaciones a Stata hasta su versión actual.

Antiguas versiones de Stata fueron criticadas por ser "excesivamente para datos de corte transversal" y por no incorporar programadas algunas técnicas econométricas sofisticadas. Por ejemplo, hasta la versión 3.1 no se incorporó la estimación de ecuaciones simultáneas ni la estimación por máxima verosimilitud. Aunque la versión 4.0 incorporaba un conjunto de procedimientos de estimación para series temporales el balance econométrico del programa estaba todavía fuertemente sesgado hacia las técnicas micro-econométricas para datos de corte transversal. Sin embargo, como respuesta a la segunda de las críticas, la versión 4.0 incluyó nuevas técnicas de estimación como, por ejemplo, la estimación de funciones de densidad a través de una *kernel*, mínimos cuadrados no lineales, mínimos cuadrados generalizados, modelos de efectos fijos y aleatorios, regresión

por cuantiles, nuevas órdenes mejoradas para *bootstrap* y experimentos de Monte Carlo, así como una ampliación de los contrastes de hipótesis posteriores a las órdenes de estimación. Además, la versión 4.0 era entre un diez y un cincuenta por ciento más rápida, dependiendo de la orden que se ejecutara, que la versión 3.1.

Los cambios más importantes en la versión 5.0 consistieron en una expansión sustancial de las órdenes para el tratamiento de datos de encuesta y de datos de panel. Esto permitió, por ejemplo, la posibilidad de estimar modelos *probit* con efectos aleatorios y modelos para datos de panel robustos a problemas de heteroscedasticidad y autocorrelación. También se extendieron e incorporaron órdenes para la estimación de modelos de duración para datos con variación temporal, datos truncados por la izquierda, datos censurados por la derecha y eventos recurrentes, además de nuevas órdenes para estimar modelos *Cox* de riesgo proporcional y de regresión *Weibull* y exponencial. La incorporación de todos estos elementos en la versión 5.0 aumentó su atractivo para los microeconómetras, aunque era todavía un programa para datos de corte transversal.

La versión de Stata 6.0 incorporó distintas novedades entre las que destacan las que se enumeran a continuación. Por primera vez el programa permitía la interacción con la web (el uso de bases de datos de la web, acceso a actualizaciones de Stata mientras no aparece la nueva versión del mismo, etc.). Se añadieron componentes específicos de series temporales, como operadores para diferenciar y retardar variables, que podían usarse tanto en expresiones como en los listados de variables de muchas órdenes, se proporcionaron nuevos formatos para datos de series temporales y se introdujeron nuevas órdenes para estimar modelos *arima* y modelos de la familia de los *arch* (*arch, garch, egarch, arch-in-mean*, etc.), así como para tabular y representar gráficamente autocorrelaciones globales y parciales, correlaciones transversales y periodogramas. También se incorporaron en esta versión contrastes de raíces unitarias (por ejemplo, el contraste de Phillips-Perron) y de ruido blanco. Se ampliaron las posibilidades de los análisis de duración tanto por la inclusión de cuatro nuevos estimadores paramétricos para modelos *log-normal*, *log-logistic*, *Gompertz* y *log-gamma* generalizada como por la posibilidad de realizar contrastes estadísticos y gráficos del supuesto de riesgo proporcional tras la estimación *Cox*. En datos de panel se ampliaron a doce los estimadores. Por ejemplo, se incluyeron estimadores de efectos aleatorios para los modelos *logit*, *tobit*, *poisson*, la regresión por intervalos, la regresión binomial negativa y regresiones *log-log* complementarias. También se incorporaron estimadores de efectos fijos para el modelo *poisson* (*poisson* condicional) y el binomial negativo. Se ampliaron de nuevo las órdenes para el tratamiento de datos de encuesta. Se mejoró la orden de maximización de funciones de verosimilitud convirtiéndola en más fácil de usar, más rápida y robusta. Se incorporaron nuevos estimadores como el *probit* bivariante, con selección de muestra, o heteroscedástico. Se permitió una implementación general del contraste de Hausman, lo que posibilitaba, por ejemplo, contrastar la independencia de alternativas irrelevantes tras la estimación de un modelo *logit* multinomial o la regresión logística condicional, así como la realización de contrastes de exogeneidad o de restricciones de sobreidentificación para mínimos cuadrados en dos y tres etapas. Finalmente, Stata 6.0 era un cinco por ciento más rápido que la versión anterior.

Versión actual del programa (Stata 7.0)

Stata 7.0 es de un ocho a un doce por ciento más rápido que Stata 6.0 e incluye 91 nuevas órdenes. A continuación presentamos un resumen del contenido actual del programa Stata indicando aquello que es novedoso para los usuarios de Stata 6.0. En un primer bloque vamos a enumerar algunos de los modelos econométricos que Stata tiene como órdenes pre-programadas:

1. Modelos de regresión lineal con estructura simple de los errores: mínimos cuadrados ordinarios, variables instrumentales, *tobit* para datos censurados, regresión por intervalos, mínimos cuadrados ordinarios no lineales, regresión por cuantiles y, como novedad en esta versión, una orden específica de estimación para modelos de regresión con datos truncados. En la estimación de estos modelos se permite imponer restricciones lineales en los parámetros.

2. Modelos de regresión lineal con errores heteroscedásticos.

3. Modelos de regresión lineal con sistemas de ecuaciones (errores correlacionados): mínimos cuadrados bietápicos y trietápicos donde se permite imponer restricciones en los parámetros, modelos *sure* y regresión multivariante.

4. Modelos para variable dependiente cualitativa binaria: *probit*, *logit*, modelo *log-log*, *probit* bivariante y *probit* heteroscedástico (con especificación de la varianza del término de error).

5. Modelos para variable dependiente cualitativa de respuesta múltiple: *logit* y *probit* ordenados, *logit* multinomial, condicional y anidado. Los modelos *logit* anidados se introducen en la nueva versión del programa. Estos modelos permiten superar el problema de independencia de las alternativas irrelevantes de los modelos *logit* multinomial o condicional.

6. Modelos para variable dependiente tipo *count data*: *poisson*, regresión binomial negativa y *poisson* y regresión binomial negativa con exceso de ceros.

7. Modelos con selección de muestra: modelo de regresión con selección de muestra, *probit* con selección de muestra y una nueva orden, en esta versión, que permite la estimación de modelos de tratamiento.

8. Modelos con datos de series temporales: *arima, arch, arma* y regresión con errores AR(1).

9. Modelos para datos de panel:

a) Modelos de regresión lineal para datos de panel: en esta versión, aparte de los estimadores tradicionales de efectos fijos (intragrupos e intergrupos) y de efectos aleatorios (mínimos cuadrados generalizados), aparecen tres novedades. La primera, estimadores de efectos fijos y aleatorios con variables instrumentales. La segunda novedad es que se pueden estimar modelos para datos de panel de efectos fijos y aleatorios suponiendo que el término de error sigue un AR(1), con la posibilidad de contrastar la autocorrelación de los errores. Por último, se incorpora el estimador de Arellano y Bond para modelos con datos de panel y variables explicativas predeterminadas, lo cual permite la estimación de modelos dinámicos. Para estos modelos se incluyen los contrastes de Sargan y el de autocorrelación de Arellano y Bond. En todos los nuevos estimadores para datos de panel las órdenes siguen siendo robustas a la existencia de un panel incompleto.

b) Modelos de regresión lineal con datos de panel censurados: *tobit* y regresión por intervalos de efectos aleatorios.

c) Modelos para variable dependiente cualitativa con datos de panel: *logit*, *probit*, *log-log* con efectos aleatorios y *logit* condicional (con efectos fijos).

d) Modelos para variable dependiente tipo *count data* para datos de panel: *poisson* y regresión binomial negativa con efectos aleatorios, modelos condicionales de efectos fijos para *poisson* y la regresión binomial negativa.

e) Modelos de coeficientes aleatorios con datos de panel.

10. Modelos de duración: modelos *Cox* de riesgo proporcional, modelos de duración paramétricos (exponencial, *Weibull*, *Gompertz*, *log-normal*, *log-logistic*, y *gamma* generalizada). Estos modelos permiten censura por la derecha, truncamiento por la izquierda, huecos en la historia de los individuos y regresores que varían con el tiempo (en los modelos *Cox* es una novedad de esta versión que se permitan regresores continuos que varíen con el tiempo). Todas las distribuciones paramétricas utilizadas en los modelos de duración se pueden estimar ahora permitiendo la existencia de heteroscedasticidad individual inobservable. Además, todos los parámetros de las distribuciones pueden hacerse depender de las variables explicativas en esta nueva versión de Stata.

11. Análisis multivariante: componentes principales y análisis factorial.

12. El análisis *cluster* es otra novedad en esta versión.

Aparte de las nuevas órdenes de estimación que la versión 7.0 incorpora frente a la anterior, el programa también incluye nuevas prestaciones que se detallan a continuación:

1. Stata 7.0 permite el cálculo de efectos marginales y elasticidades tras la estimación de cualquier modelo. Stata 6.0 sólo incorporaba esta utilidad para determinados modelos. Esta mejora afecta de un modo muy favorable a la interpretación de los resultados que se obtienen en la estimación de modelos con variable dependiente discreta y censurada, como por ejemplo, los modelos *probit*, *tobit* y los multinomiales.

2. Casi todos los comandos de estimación que utilizan como método la máxima verosimilitud acomodan ahora restricciones lineales (por ejemplo en *heckman*, *poisson*, etc.).

3. Después de la estimación por *probit, logit, poisson*, binomial negativa y muchos otros estimadores para modelos de corte transversal por máxima verosimilitud se pueden calcular errores estándar y matrices de varianzas-covarianzas robustas a problemas de heteroscedasticidad y autocorrelación.

4. Se incorpora la posibilidad de elegir para los números el formato decimal europeo.

5. Por último, los nombres de variables, programas, matrices, ficheros de resultados, etc., pueden ser hasta de 32 caracteres, mientras que hasta ahora el límite era de 8 caracteres.

Compatibilidad con versiones anteriores

Al igual que ocurría con versiones anteriores, Stata 7.0 es compatible al cien por cien con la versión previa. En la práctica esto significa que el nuevo programa

lee automáticamente las bases de datos en formatos previos a la versión actual, aunque si se deseara utilizar un programa escrito en Stata 6.0 habría que añadir al principio del programa *version 6.0*. Si una base de datos en versión Stata 7.0 quiere usarse en la versión 6.0 hay que grabarla especificando la opción *save's old*.

Elementos básicos de funcionamiento

El programa Stata está organizado en torno a varias ventanas que contienen resultados, gráficos, una lista de variables, la historia reciente de las órdenes y las órdenes actuales. Para aquellos a los que les gusta examinar los datos de cerca sin la ayuda del análisis estadístico se puede acceder a un visor/editor de datos similar a una hoja de cálculo. La estructura de trabajo estándar en Stata se compone de una base de datos que se carga en memoria y sobre la que se ejecutan programas, o bien desde el editor de programas (que contiene una lista de órdenes) o mediante órdenes aisladas ejecutadas de forma interactiva desde la ventana de órdenes. La sintaxis de todas las órdenes aparece recogida en las ayudas de Stata y cumple los objetivos de ser lógica, clara y predecible. Por último, los resultados de una sesión de trabajo se pueden enviar a un fichero de resultados.

El lenguaje de Stata es además plenamente programable. Por tanto, usuarios avanzados pueden generar nuevas órdenes o procedimientos no incluidos en el programa. Para ello, Stata dispone de un manual (que forma parte de sus manuales básicos) dedicado a la programación en general. También existen manuales específicos de programación. Por ejemplo, para escribir órdenes propias de estimación de funciones de máxima verosimilitud se recomienda el libro *Maximum-Likelihood Estimation with Stata* (que se puede encontrar, junto con otros manuales relacionados con Stata, en la página *web* http://www.stata.com/bookstore).

Uno de los puntos más fuertes de Stata es su flexibilidad en el manejo de datos. Los datos pueden, de una manera muy sencilla, editarse, cargarse, fusionarse, añadirse, comprimirse, ordenarse, generar con ellos nuevas variables y decodificarse de unos formatos a otros. Adicionalmente, se pueden añadir etiquetas y títulos a las variables, a las bases de datos, etc. El programa permite utilizar datos en formatos distintos a Stata (Excel, Gauss, SPSS, etc.). Esto se puede hacer de dos formas. Una, copiando y pegando los datos desde una hoja de cálculo al editor de datos de Stata. Otra, a través del programa Stat/Transfer[2] que convierte archivos de datos en cualquier otro formato (de una lista especificada) a formato de datos Stata. Además, los datos pueden leerse directamente en formato ASCII con el uso de las órdenes y los procedimientos de Stata adecuados para esta acción.

La mayoría de las órdenes en Stata se pueden ejecutar sobre determinadas submuestras que cumplen ciertas condiciones (condiciones de rango y/o lógicas). En definitiva, podemos decir que Stata es un programa dominado por la escritura de órdenes más que un programa de uso del teclado para la ejecución de las mismas. Este aspecto es esencial puesto que permite una mayor versatilidad.

(2) Stat/Transfer es un programa complementario de Stata que es también distribuido en España por Timberlake Consulting S.L.

Manuales y acceso a las ayudas de Stata

La calidad de la documentación que proporciona Stata es excelente. Se puede acceder a la ayuda de Stata por tres vías: ayuda en manuales, ayuda en pantalla (una vez dentro del programa) y ayuda a través de la página web de Stata. Existen cuatro volúmenes de órdenes de Stata ordenadas alfabéticamente (*Reference Manual*), donde se recoge tanto la sintaxis de las órdenes como la teoría econométrica detrás de las técnicas de estimación, así como referencias bibliográficas sobre las mismas. Se incluye también un manual para crear y editar gráficos (*Graphics Manual*). Si se pretende obtener una visión global de la estructura del programa y de cuestiones específicas tales como funciones, expresiones, etc., existe una guía del usuario (*User's Guide*). Stata incluye una guía básica para Windows (*Getting Started*) que proporciona una rápida introducción sobre cómo iniciarse en el manejo de Stata. Estos dos últimos manuales son muy útiles como introducción a Stata y es muy aconsejable leerlos antes de adentrarse en el manejo del programa. Por último, para usuarios más avanzados Stata dispone de un manual de programación (Programming Manual).

La ayuda en pantalla es muy amplia y proporciona de nuevo la sintaxis de las órdenes, sus opciones y una serie de órdenes relacionadas, aunque siempre es más completa la información de los manuales. Si se desconoce el nombre de la orden que en Stata corresponde a una determinada acción podemos aproximarnos al tema en cuestión por medio de una búsqueda en pantalla, para la que se debe utilizar inglés formal y terminología estadística.

Finalmente, Stata también proporciona ayuda en su página *web* http://www.stata.com. La mitad de la página está dedicada a ayuda al usuario, e incluye, entre otras, información de utilidad, respuestas a las preguntas más comunes de los usuarios (FAQ), modos de interactuar con otros usuarios y actualizaciones oficiales de Stata (aparecen seis veces al año coincidiendo con cada uno de los meses impares y son previas a una nueva versión del programa). También es posible suscribirse a Statalist, un servidor dedicado a discusiones sobre Stata así como a discusiones estadísticas. En la página web de Stata se puede encontrar, además, información sobre cursos interactivos de Stata por internet y sobre reuniones oficiales de usuarios en distintos lugares del mundo, de las cuales dos han sido en España. Finalmente, para acceder e instalar nuevos procedimientos de Stata escritos por los propios usuarios podemos acudir al *Stata Technical Bulletin* (STB) que se publica en su web y para el cual existe una subscripción gratuita de un año.

Edición y creación de gráficos

La creación de gráficos es una prestación importante del programa Stata. Hay que destacar tanto su calidad, como la sencillez y el fácil acceso en la edición de gráficos. Las posibilidades de etiquetado y anotaciones de los gráficos dentro de Stata son muy extensas y hace que en muchos casos no sea necesaria la reedición de los gráficos en otros programas.

En el manual dedicado exclusivamente a la creación y edición de gráficos (*Graphics Manual*) se puede comprobar la gran multitud de opciones que permite

Stata en la edición de los mismos. Por ejemplo, se pueden crear gráficos de barras, histogramas, gráficos donde se representa una variable, gráficos de doble entrada, gráficos de quesos, etc. Además, en todos ellos podemos cambiar los títulos de los ejes, re-escalar los ejes, cambiar los símbolos que representan cada observación, conectar los puntos de las observaciones, acompañar con líneas de tendencia, cambiar el estilo de las líneas de conexión, etc.

Entre las novedades de esta versión se destaca la posibilidad de especificar el tamaño del gráfico a la hora de reprensentarlo o exportarlo (aunque todavía existen limitaciones a la hora de manipular estos gráficos dentro de otros programas del entorno Windows) y la opción de imprimirlo directamente desde la pantalla. Stata 7.0 también ha simplificado la manera de exportar un gráfico a otros formatos mediante una orden que permite incorporar múltiples opciones. Y por último, se permite la selección del tipo de línea en los gráficos.

Fiabilidad numérica de Stata *versus* otros programas

Los economistas generalmente eligen un programa estadístico, aparte de por el precio del mismo, por la facilidad de uso, por la velocidad de ejecución de órdenes o por alguna característica especial de dicho programa. Sin embargo, un programa estadístico no sólo debe valorarse por la lista de opciones que ofrece sino también por su fiabilidad numérica. Los estudios sobre este último aspecto han puesto de manifiesto que distintos programas estadísticos proporcionan distintas respuestas a un mismo problema [McCullough y Vinod (1999)]. Pequeñas diferencias en la precisión numérica de los dígitos decimales de un número pueden producir grandes cambios en el resultado de un procedimiento posterior que lo incorpora. Estas diferencias hacen que cada vez sea más importante comparar la fiabilidad numérica entre distintos programas como un elemento importante a la hora de decidir cuál es el que más se ajusta a nuestras necesidades y/o exigencias. Una forma estándar de evaluar la fiabilidad numérica de los progamas estadísticos aparece recogida en McCullough (1998) y se centra en tres aspectos: estimación, generación de números aleatorios y funciones de distribución. Por lo que respecta a la estimación, el National Institute of Standards and Technology (NIST) proporciona bases de datos (Statistical Reference Datasets –StRD–, disponibles en http://www.itl.nist.gov/div898/strd) y valores certificados que sirven para una evaluación objetiva de los programas en cuatro áreas: estadística descriptiva univariante (media, desviación típica y coeficiente de autocorrelación de primer orden), regresión lineal (coeficientes, errores estándar y suma de cuadrados de los residuos), análisis de varianza (estadístico F) y regresión no lineal (coeficientes, errores estándar y suma de cuadrados de los residuos). Los valores certificados obtenidos con una alta fiabilidad pueden entonces compararse con los valores estimados por los distintos programas. La fiabilidad de un generador de números aleatorios (GNA) se comprueba a través de los contrastes de aleatoriedad del programa DIEHARD de Marsaglia (1996). De la calidad de los GNA depende la validez de técnicas como, por ejemplo, el *bootstrapping* y los experimentos de Monte Carlo. Los programas de Knüsel (1989) o de Brown (1998) se utilizan para contrastar la fiabilidad de las funciones de distribución, necesarias para el cálculo de *p-values* y valores críticos para distintos niveles de significatividad.

A continuación presentamos los resultados de los estudios de fiabilidad numérica disponibles para Stata 7.0, así como para distintos programas estadísticos comparables con Stata.

El análisis de Stata 7.0 aparece en http://www.stata.com/support/cert/ y no incluye las funciones de distribución. Stata supera todos los contrastes NIST-StRD para estadísticos descriptivos univariantes, análisis de varianza y regresión no lineal. No supera el contraste de Filippelli para regresión lineal por detectar regresores altamente colineales, como ocurre para la mayoría de programas estadísticos. De la aplicación del contraste DIEHARD se extrae que el GNA de Stata funciona muy bien. Sin embargo, no se debe olvidar que incluso los programas que pasan el DIEHARD pueden no ser válidos para aplicaciones a gran escala del GNA, debido a un periodo insuficiente[3]. Para que los resultados del muestreo o de las simulaciones sean precisos se le ha de exigir al GNA un periodo muy largo [Ripley (1990), Knuth (1997), Gentle (1998) y Hellekalek (1998)]. Dado que el GNA de la mayor parte de programas estadísticos es de aproximadamente 2^{31}, dichos programas no son adecuados para técnicas que precisan de la aplicación a gran escala del GNA. El periodo del GNA de Stata es de $2^{126} \approx 10^{38}$, muy por encima del periodo típico en la mayoría de programas. Quienes precisan de un GNA de alta calidad deberían considerar los paquetes estadísticos que funcionan mejor en los contrastes DIEHARD, como son Stata o Eviews, y que describen sus algoritmos y el periodo de su GNA, como Stata [Altman y McDonald (1999) y McCullough (1999a)].

El análisis de TSP 4.5 aparece en http://www.tspintl.com/products/tsp/benchmarks/ y no incluye los GNA ni las funciones de distribución. Este programa supera todos los contrastes para estadísticos descriptivos univariantes. Sin embargo, falla en tres de los contrastes del análisis de varianza, en uno de regresión no lineal y en uno de regresión lineal, el de Filippelli, para el que detecta singularidad por regresores altamente colineales. Un análisis del GNA de su versión anterior (4.4) se encuentra en McCullough (1999a). Dicha versión del programa no supera siete de los contrastes DIEHARD sobre el GNA (con periodo aproximado de 2^{31}).

El análisis de la fiabilidad numérica de SAS 8 ha sido eleborado por el Statistical R&D Staff del SAS Institute Inc. (2000) y aparece en http://www.sas.com/rnd/app/papers/abstracts/statisticalaccuracy.html. Este programa supera todos los contrastes para la media y la desviación típica, pero el coeficiente de autocorrelación de primer orden no se puede calcular en uno de los casos. Por otra parte, falla en uno de los contrastes de regresión lineal (el de Filippelli) y en tres del análisis de varianza. Sin embargo, supera todos los contrastes para regresión no lineal. Por último, falla en uno de los contrastes DIEHARD de GNA (periodo de $2^{31} - 1$).

El análisis de SPSS 7.5 aparece en McCullough (1999 b). Este programa supera todos los contrastes para la media y la desviación típica, pero el coeficiente de autocorrelación de primer orden no se puede calcular en uno de los casos, ade-

(3) El periodo de un GNA es el número de valores que produce el GNA antes de empezar a repetirse a sí mismo. Si p es la longitud del periodo y n el número utilizado de números aleatorios, Knuth (1997) recomienda que p>1000n, mientras Hellekalek (1998) y Ripley (1990) recomiendan que $p > 200n^2$. La razón de que sólo deba usarse una fracción del periodo de un GNA se debe a que la aleatoriedad de los valores que produce decrece conforme n tiende a p [L'Ecuyer (1999)].

más de que se falla en cinco de sus contrastes. En realidad no se puede determinar la precisión de SPSS para este coeficiente, dado que el usuario ni puede acceder a este estadístico ni puede controlar el número de dígitos mostrados por el programa para este coeficiente. Este programa no supera uno de los contrastes de regresión lineal (el de Filippelli), uno de los contrastes para regresión no lineal ni tampoco seis de los contrastes para el análisis de varianza y, además, no permite calcular el estadístico F en uno de los casos. Por último, falla en uno de los contrastes DIEHARD de GNA (periodo aproximado de 2^{31}). Puesto que SPSS 7.5 no es la versión actual del programa, la nueva versión puede haber remediado alguna de las deficiencias previas.

El estudio de la fiabilidad numérica de LIMDEP 7.0 para windows se puede encontrar en McCullough (1999 a). De este análisis podemos concluir que se obtienen buenos resultados tanto para los contrastes de los estadísticos descriptivos univariantes como para los contrates de regresión lineal. Sin embargo, el programa falla en uno de los contrastes para la estimación no lineal. No se dispone de información de los contrastes de varianza. Por último, se obtiene un fallo en uno de los contrastes DIEHARD de GNA. El periodo de su GNA es desconocido y no aparece en la documentación del programa.

El análisis de Eviews 3.0 también se recoge en McCullough (1999 a). Esta versión del programa supera todos los contrastes para la estimación lineal (excepto el contraste de Filippelli) y para los estadísticos descriptivos univariantes. En cambio, no supera diez de los contrastes planteados para la estimación no lineal. No se dispone de información de los contrastes de varianza. En cuanto a los contrastes DIEHARD para GNA no se detecta ningún fallo. El periodo de su GNA es desconocido y no aparece en la documentación del programa.

Tras esta revisión de la fiabilidad numérica de cinco programas estadísticos comparables a Stata podemos concluir que Stata 7.0 es el programa que mejor se comporta entre sus más cercanos competidores. En la precisión de estadísticos univariantes se equipara a TSP 4.5, LIMDEP 7.0 y Eviews 3.0, y supera a SAS 8 y SPSS 7.5. En el análisis de varianza supera a TSP 4.5, SAS 8 y SPSS 7.5, mientras que se carece de información para compararlo con LIMDEP 7.0 ó Eviews 3.0. Se puede confiar en todos los programas analizados para la obtención de resultados de regresión lineal puesto que el fallo de todos ellos en el contraste de Filippelli (a excepción de LIMDEP 7.0) se debe a la existencia de regresores altamente colineales. En regresión no lineal Stata 7.0 y SAS 8 van parejos y superan a los demás programas. Destaca el mal comportamiento diferencial de Eviews 3.0 en este campo particular. En cuanto al GNA Stata 7.0 y Eviews 3.0 son los mejores, siendo el peor el GNA de TSP 4.5. Además, Stata 7.0 es el único que ofrece un GNA con un periodo (teóricamente) lo suficientemente largo como para ser usado en simulaciones a gran escala.

Conclusiones

Para concluir vamos a enumerar tanto las limitaciones como las cualidades de Stata 7.0. En primer lugar, hay que decir que Stata no es el mejor programa estadístico para realizar análisis macroeconómicos y financieros con técnicas sofis-

ticadas de series temporales, aunque hay que resaltar su importante esfuerzo en este campo, ya visible desde la versión 6.0. Otro elemento mejorable en el programa es la escasa posibilidad de manipular gráficos de Stata dentro de otros programas del entorno Windows (como Word o Excel). Por último, habría que destacar la falta del estimador del Método Generalizado de Momentos (MGM), aunque tiene una serie de estimadores basados en el principio del MGM, como el estimador de Arellano y Bond para modelos dinámicos de datos de panel.

En cuanto a las cualidades más importantes de Stata 7.0 podemos resaltar las siguientes. Por lo que respecta a su contenido, una de las mayores ventajas de Stata es su estimador de funciones de máxima verosimilitud [Gould y Sribney (1999)], que permite que los usuarios puedan escribir sus propias funciones y hacer que Stata las maximice. A esto hay que añadir que dispone de muy buenas rutinas de maximización para máxima verosimilitud. También hay que destacar que el usuario puede proveer primeras y segundas derivadas para la estimación por máxima verosimilitud, lo cual también se ha comprobado que aumenta la precisión numérica. Stata es un programa excelente para el análisis de datos de panel y de encuesta, de variable dependiente discreta y limitada, de duración y supervivencia, de regresión y contrastes relacionados. Por lo que concierne a la rapidez del programa podemos decir que a pesar de ser un programa basado en operaciones sobre observaciones individuales, éstas son muy rápidas debido a que Stata almacena todos los datos en memoria RAM, aunque también puede usar la memoria virtual. También su rapidez y eficiencia se ve reforzada por el hecho de no ser un programa totalmente orientado al uso de menús. Otro aspecto a destacar como virtud de Stata es que es muy estable, es decir, raramente se bloquea. En cuanto a la precisión y fiabilidad numérica podemos decir que Stata destaca en este campo frente a sus más cercanos competidores. Otro elemento a destacar de Stata es que se desarrolla y actualiza rápidamente tanto de un modo formal como informal y que proporciona una excelente asistencia al usuario, que proviene tanto de la propia empresa como de su comunidad de usuarios. Por último, si a todas estas ventajas le añadimos el hecho de ser un programa barato dentro de los programas estadísticos y que además proporciona precios todavía más asequibles para la comunidad académica, podemos concluir que Stata se convierte en un programa altamente recomendable para la investigación aplicada, especialmente con datos de corte transversal, de panel o de encuesta.

REFERENCIAS BIBLIOGRÁFICAS

Altman, M. y M. Mcdonald (1999): "The robustness of statistical abstractions: a look *under the hood* of statistical models and software", presentado en la Summer Political Methodology Conference Texas A&M University, College Station, 14 de julio. Disponible en http://data.fas.harvard.edu/numerical_stability/.

Banks, J.W. (1996): "STATA 4.0 (DOS), STATA FOR WINDOWS", *Economic Journal*, vol. 106, págs. 748-752.

Brown, B.W. (1998): *DCDFLIB v1.1.* (Double precision cumulative distribution function library). Disponible en http://odin.mdacc.tmc.edu/pub/source.

Deaton, A. (1997): *The analysis of household surveys*, Johns Hopkins University Press.

Ferral, C. (1994): "A review of Stata 3.1", *Journal of Applied Econometrics*, vol. 9, págs. 469-477.

Gentle, J.E. (1998): *Random number generation and Monte Carlo methods*, Springer-Verlag, New York.

Gould, W. y W. Sribney (1999): *Maximum likelihood estimation with Stata,* Stata Press, College Station, Texas.

Hamilton, L. (1997): *Statistics with Stata 5,* Duxbury Press.

Hardin, J. y J. Hilbe (2001): *Generalised linear models and extensions*, Stata Press.

Hellekalek, P. (1998): "Good random number generators are (not so) easy to find", *Mathematics and Computers in Simulations,* vol. 46, págs. 485-505.

Knüsel, L. (1989): "Computergestützte Berechnung Statistischer Verteilungen"Oldenburg, München-Wien. Programa y documentación en inglés disponible en *http://www.statmuenchen.de/~knuesel/elv*.

Knuth, D.E. (1997): *The art of computer programming*, (3.ª Edición), Addison-Wesley, Reading, MA.

L'Ecuyer, P. (1999): "Random number generation", en *Handbook on Simulation,* Ed. J. Banks, Wiley, New York, págs. 93-138.

Long, S.J. (1997): *Regression models for categorical and limited dependent variables*, SAGE.

Marsaglia, G. (1996): "DIEHARD: a battery of tests for random numbers generators", CD-ROM, Department of Statistics and Supercomputer Computations Research Institute, Florida State University (htpp://stat.fsu.edu/~geo).

McCullough, B.D. (1998): "Assessing the reliability of statistical software: part I", *The American Statistician,* vol. 52, págs. 358-366.

McCullough, B.D. (1999 a): "Econometric software reliability: Eviews, LIMDEP, SHAZAM and TSP", *Journal of Applied Econometrics,* vol. 14, págs. 191-202.

McCullough, B.D. (1999 b): "Assessing the reliability of statistical software: part II", *The American Statistician,* vol. 53, págs. 149-159.

McCullough, B.D. y H.D. Vinod (1999): "The numerical reliability of econometric software", *Journal of Economic Literature,* vol. XXXVII, págs. 633-665.

Mestre, R. (1994): "Stata, un paquete sencillo pero potente", *Revista de Economía Aplicada*, vol. 2, págs. 163-172.

Rabe-Hesketh y B. Everitt (2000): *A handbook of statistical analyses using Stata*, segunda edición, Chapman y Hall.

Rees, H. (1998): "STATA 5.0", *Economic Journal*, vol. 108, págs.252-257.

Ripley, B.D. (1990): "Thoughts on pseudorandom number generators", *Journal of Computationals Applied Math,* vol. 31, págs. 153-163.

Statistical R&D Staff, SAS Institute Inc. (2000): "Answering the numerical accuracy of SAS software", disponible en http://www.sas.com/rnd/app/papers/abstracts/statisticalaccuracy.html.