Ejercicios del curso *Introducción al análisis de datos con el software* `Stata`*

José-Ignacio Antón

La evaluación del curso se basa en la entrega de uno de los tres siguientes conjuntos de ejercicios que se presentan a continuación. La primera opción propone una explotación del Latinobarómetro. La segunda alternativa explora una encuesta de hogares tradicional, en este caso, la Encuesta Continua de Hogares de Uruguay. La tercera de las opciones consiste en un análisis de una base de microdatos elegida por el estudiante.

Las soluciones de los ejercicios deben contener, como mínimo, los siguientes elementos:

— Archivo de sintaxis `do`, con comentarios del estudiante que permitan seguir al profesor el análisis.

— Archivo log (e.g., `smcl` o `txt`), que permita al profesor visualizar el resultado de tablas, análisis estádisticos, etc.

— Resto de resultados que no se puedan incluir en los archivos anteriores, como las figuras o los archivos `docx` o `pdf` (resultado de los archivos TEX).

Otras opciones pueden ser admisibles. Por ejemplo, los usuarios más avanzados pueden preferir archivos `.html` o `.pdf` generados con Markdown o TEX.

En los ejercicios se valorará su correcta resolución, la aportación de comentarios detallados que permitan seguir el análisis y la presentación. La superación del curso requiere la entrega de al menos el 70 % de los ejercicios de la alternativa elegida adecuadamente resueltos antes de las 23:59 del día 10 de octubre de 2021.

*Para cualquier consulta o duda, podéis contactar conmigo a través del correo electrónico (janton@usal.es)

Opción 1

Ejercicios asociados al Latinobarómetro

Ejercicio 1

El Latinobarómetro es una encuesta anual que se realiza en la mayor parte de los países de América Latina y el Caribe y que recoge información sobre la percepción de los ciudadanos de la democracia, la economía y la sociedad. Realiza las siguientes tareas:

a) Visita la página web de la encuesta y descarga la base de datos en formato `.dta` y el diccionario de variables correspondientes al año 2017.

b) Recodifica la variable `sexo` como una variable que tome el valor 0 para hombres y 1 para mujeres. Etiqueta esta variable y sus valores.

c) Crea una nueva variable, que puedes llamar `educg`, que recoja el nivel educativo en tres niveles: estudios bajos (primaria e inferior), medios (secundaria) y altos (universitario). Etiqueta la variable y sus valores. De forma similar, crea una variable que recoja la edad del entrevistado, que puedes llamar `edadg` en 3 intervalos (menores 30, entre 30 y 64 años y 65 y más años).

d) Explora la variable *Escala izquierda-derecha* y recodifica aquellas respuestas distintas a la escala 0-10 (ninguno, no sabe, no contesta) como valores perdidos.

e) Guarda un archivo con nombre `latbar2017.dta` con variables que contengan el país de la entrevista, año de la entrevista, sexo, edad, edad en tres grupos educación en tres grupos y escala izquierda-derecha (solo las personas con valores 0-10) y el ponderador `wt`. Excluye de la muestra a España, la población de 65 y más años y todas las personas con valores perdidos en cualquiera de las seis variables.

Ejercicio 2

a) Descarga de la misma página los datos del Latinobarómetro correspondientes a 1997 y 2007. Aplica filtros similares a los del ejercicio anterior y guarda los archivos resultantes.

b) Une los tres archivos empleando la instrucción `append` y guarda el archivo resultante como `latbar19972017.dta`.

c) Visita la web del Fondo Monetario Internacional y descarga del World Economic Outlook Database April 2020 un archivo que contenga la tasa de desempleo de 1997 a 2017 de todos los países del continente. Obtendrás un archivo `.xls`. Investiga cómo crear un archivo en formato `.dta` a partir del mismo. Guarda el archivo con el nombre `weo.dta`.

d) Con la instrucción `merge`, fusiona el archivo `latbar19972017.dta` con el archivo `weo.dta` usando como variables enlace el país y el año. Comprueba qué países aparecen en ambas bases de datos y cuáles no y quédate únicamente con los primeros. Con el comando `collapse`, crea una base de datos resumida que contenga la posición ideológica media por año y país. Reorganiza la base de datos con el comando `reshape` de forma que contenga una columna con la posición ideológica media para cada año y genera una nueva variable que incluya el cambio porcentual de la posición ideológica media entre 1997 y 2017.

Ejercicio 3

a) A partir de la base de datos `latbar19972017.dta`, elabora una tabla que nos indique la distribución porcentual de la educación por sexo (qué porcentaje de hombres y mujeres, respectivamente, tienen niveles educativos bajo, medio y superior) y la distribución porcentual del sexo por nivel educativo (qué porcentajes de personas con estudios bajos, medios o superiores son hombres y mujeres).

b) Elabora una tabla donde muestres la posición ideológica media por sexo y nivel educativo.

c) Realiza un análisis de varianza de la posición ideológica por año, para deter-

minar si la posición ideológica es diferente en cada uno de los tres años de la muestra.

d) Realiza un análisis de regresión en el que explores la relación entre la posición ideológica (variable dependiente) y el sexo, edad, edad al cuadrado y nivel

Ejercicio 4

a) A partir de la base de datos `latbar19972017.dta`, elabora un histograma de la posición ideológica de los encuestados.

b) Elabora un gráfico de barras horizontales en el que muestres la posición ideológica media por nivel educativo. Para ello, puedes emplear, antes de realizar el gráfico, el comando `collapse`, aunque no es estrictamente necesario.

c) Crea un gráfico de líneas en el que muestres la evolución por sexo de la posición ideológica media desde 1997 a 2017. Nuevamente, `collapse` puede ser de utilidad.

d) Realiza un gráfico de dispersión en el que explores si existe relación entre el porcentaje de personas del país con un nivel educativo alto y el porcentaje de personas con posiciones ideológicas de izquierda (definida como un valor en la escala ideológica de 0 a 3) en un año determinado.

e) Elabora un mapa con la posición ideológica media de los países disponible de la base de datos en 2017. Para ello, descarga los mapas vectoriales correspondiente de la web, por ejemplo, desde este enlace. Con la ayuda de `collapse` crea una base de datos con la posición ideológica media por país. Para elaborar el mapa de forma correcta, debes asegurarte que la identificación del país en el Latinobarómetro y en los archivos `.dta` asociados al mapa que generan durante el proceso coinciden.

Ejercicio 5

Realiza las tres regresiones de la posición ideológica sobre los siguientes grupos de variables:

— Sexo, edad y edad al cuadrado.

— Sexo, edad, edad al cuadrado y educación.

— Sexo, edad, edad al cuadrado, educación y año de la encuesta.

Con ayuda de `esttab` o `putexcel`, crea una tabla en Microsoft Word (o un procesador de textos similar) o TEX con la tabla de regresión.

Opción 2

Ejercicios asociados a la Encuesta Continua de Hogares del Uruguay

Ejercicio 1

La principal fuente de información estadística para los análisis del mercado de trabajo y pobreza en el Uruguay es la *Encuesta Continua de Hogares* (ECH), a cargo del Instituto Nacional de Estadística (INE). Visita la web del INE y realiza las siguientes tareas:

a) Descarga las bases de datos de persona y hogar en el formato `.dbf` y el diccionario de variables (en formato `.xlsx`) correspondientes al año 2019. Importa la base de datos de personas `P1_2019_Terceros.dbf` desde `Stata`.

b) Guarda la base de datos como un archivo `.dta` con el nombre `P1_2019.dta`. Repite el mismo procedimiento con la base de datos `P2_2019.dbf`.

c) Fusiona (horizontalmente) `P1_2019.dta` y `P2_2019.dta` con el comando `merge`, empleando como variables de enlace `numero` y `nper`.

d) Recodifica la variable `e26`, sexo del entrevistado, como un variable que tome el valor 0 para los hombres y 1 para las mujeres. Renombra la variable como `sexo` y etiqueta la variable y sus valores. Renombre la variable `e27`, edad del entrevistado, como `edad` y etiqueta la variable.

e) Crea una nueva variable, `edadg` que recoja la edad del entrevistado en tres intervalos (menores 15, entre 15 y 59 años y 60 y más años). Asimismo, a partir de las variables relevantes (`e48`, `e197_1`, `e201_1`, `e212_1`, `e215_1` `e218_1`, `e221_1` y `e224_1`), crea una variable `educ` que recoja el mayor nivel educativo que el individuo ha finalizado en cuatro grupos (sin educación formal completada, primaria completa, media completa y terciaria completa). Etiqueta ambas variables y sus valores.

f) A partir de la variable `pobpcoac`, que recoge al condición de actividad económica, crea una variable que se denomine actividad `actividad` con las siguientes categorías: menores de 14 años, ocupados, desempleados, retirados o jubilados y otros inactivos. Etiqueta la variable y sus valores.

g) Guarda la base de datos con las variables `numero`, `anio`, `dpto`, `nomdpto`, `barrio`, `nombarrio`, `pesoano` y las variables creadas `sexo`, `edad`, `edadg`, `educ` y `actividad`. Puedes utilizar como nombre `ecv2019_persona.dta`.

Nota: La importación de archivos `.dbf` solo es posible de forma directa para las versiones de Stata 15 y posteriores. Como posibles soluciones, podemos abrir cada archivo en Microsoft Excel y guardarlo en otro formato (e.g., `.csv`), como se explica aquí fácilmente importable con Stata. Otras posibilidades son el formato `.dat` e importar los datos archivo de texto delimitado y la extensión `.sav`, referida a SPSS. En este último caso, aunque es posible importar los datos directamente en Stata 16, para versiones anteriores es posible descargar un paquete como `usespss`, como se explica en este enlace.

Ejercicio 2

a) Importa la base de datos del hogar del año 2019 desde `Stata`, `H_2019_Terceros.dbf`. Quédate con las variables `numero`, `anio`, `pobre06` y la variable `d21_16`, que recoge si el hogar tiene conexión a internet. A partir de esta última variable, crea una variable que puedes denominar `internet`, que tome los valores 0 (si no existe acceso a internet) y 1 (si el hogar tiene conexión). Guarda la base de datos con las variables `anio`, `numero`, `internet` y `pobre06` con el nombre `ecv2019_hogar.dta`.

b) Realiza una fusión (horizontal) de los archivos `ecv2019_persona.dta` y `ecv2019_hogar.dta` a partir de la variable `numero`. Guarda el archivo resultante como `ech2019.dta`.

c) Selecciona al menos otras dos olas entre 2006 y 2019 y repite el mismo procedimiento, buscando las variables adecuadas. En la mayor parte de los casos, cada ola cuenta dos archivos de personas (salvo 2006, que incluye tres) y un archivo de hogares.

d) Fusiona (verticalmente) los archivos de las tres olas seleccionadas con el comando `append` y denomina el archivo resultante como `ech.dta`. Para realizar correctamente esta fusión, las variables deben tener un formato similar en todas las olas. Asegúrate de que la variable `anio`, `numero` y `nper` se han co-

dificado de forma numérica. Si no es así, conviértelas en variables numéricas con el comando `destring`.

e) Crea una base de datos resumida, con el comando `collapse` que contenga la proporción de personas pobres por departamento y año.

f) Reorganiza la base de datos con el comando `reshape` de forma que contenga una columna con la proporción de personas pobres cada año y crea una nueva variable que incluya el cambio en la proporción de personas pobres por departamento entre el primer y el último año de tu base de datos.

Ejercicio 3

a) A partir de la base de datos `ech2019.dta`, elabora una tabla que nos indique la distribución porcentual de la educación por sexo (qué porcentaje de hombres y mujeres, respectivamente, tienen niveles educativos bajo, medio y superior) y la distribución porcentual del sexo por nivel educativo (qué porcentajes de personas con estudios bajos, medios o superiores son hombres y mujeres).

b) Elabora una tabla donde muestres la edad media por nivel educativo y sexo.

c) Realiza un contraste hipótesis en el que analices si la proporción de personas pobres por nivel educativo es la misma.

d) Lleva a cabo un contraste de hipótesis en el que explores si la edad media de las personas en Montevideo y en el resto del país es la misma. Para ello, crea primero una variable, que puedes llamar `mvd`, que tome el valor 1 si el departamento es Montevideo y 0 en otro caso.

e) Lleva a cabo un análisis *logit* en el que explores la relación entre la pobreza (variable dependiente) y el sexo, edad, edad al cuadrado y nivel educativo. Calcula los coeficientes, los *odds ratios*, la media de los efectos marginales para toda la muestra, los efectos marginales para los valores medios en todas las variables y los efectos marginales para los valores medianos en todas las variables.

Ejercicio 4

a) A partir de la base de datos `ech.dta`, elabora un gráfico de líneas en el que muestres la evolución temporal del porcentaje de personas con acceso a internet por grupos de edad (una línea para cada grupo de edad). Puedes emplear el comando `collapse` para reducir la base de datos.

b) Con la base de datos `ech2019.dta`, elabora un gráfico de barras horizontales en el que muestres el porcentaje de personas con acceso a internet por nivel educativo.

c) Limita el análisis a los barrios de Montevideo y al año 2019. Mediante el comando `collapse`, calcula el porcentaje de personas con conexión a internet y el porcentaje de personas pobres por barrio en 2019. Elabora un gráfico de dispersión para ver si existe relación entre el porcentaje de personas con conexión a internet (en el eje vertical) y el porcentaje de personas pobres (en el eje horizontal).

d) Descarga de la web del INE web del INE los mapas vectoriales correspondientes al Censo 2011. Con ayuda del comando `collapse`, crea una base de datos con el porcentaje de niños entre 5 y 15 años con acceso a internet en 2019 en los 62 barrios de Montevideo. Elabora un mapa con colores para ilustrar esta variable. Para ello, es preciso emplear los paquetes de mapas correspondientes y los archivos de mapas `ine_barrios_mvd_nbi85`.

Ejercicio 5

A partir de la base de datos de la ECH `ech.dta`, considerando al menos dos olas, debéis elaborar un sencillo perfil de pobreza de las personas de 25 o más años a través de un modelo *probit*. Se sugiere presentar la media de los efectos marginales para toda la muestra de tres modelos que incluyan las siguientes variables:

— Sexo, edad y edad al cuadrado.
— Sexo, edad, edad al cuadrado y educación.
— Sexo, edad, edad al cuadrado, actividad y año de la encuesta.

Con ayuda de `esttab` o `putexcel`, crea una tabla en Microsoft Word (o un procesador de textos similar) o TEX con la tabla de regresión.

Opción 3

Ejercicios con base de datos elegida por el estudiante

El estudiante tiene la posibilidad de seleccionar una base de datos microdatos y realizar un trabajo como el descrito en las dos opciones anteriores. En particular, este estudio debe contener:

a) Ejemplos de importación de datos y manejo de variables (recofidicación, etiquetado de variables y valores, creación de nuevas variables, etc.).
b) Ejemplos de fusión de bases de datos con `merge` y `append` y aplicaciones de `collapse` y `reshape`.
c) Elaboración de varias tablas de contigencia y realización de un análisis econométrico multivariante.
d) Realización de gráficos, uno de los cuales deberá ser un mapa.
e) Elaboración de una tabla que recoja el resultado de varios análisis econométricos y exportación de la misma a Microsoft Word o TEX.

A continuación, podéis encontrar una lista de páginas, a modo de ejemplo, donde existen datos en abierto (algunas de ellas requieren un registro previo gratuito):

— Instituto Nacional de Estadística de España: https://www.ine.es/prodyser/microdatos.htm.
— UK Data Service: https://ukdataservice.ac.uk/.
— IPUMS International: https://international.ipums.org/international/.
— Banco de Datos del CIS: http://www.cis.es/cis/opencms/ES/index.html.
— European Social Survey: https://www.europeansocialsurvey.org/.
— Instituto Nacional de Estadística del Uruguay: https://ine.gub.uy/web/guest/microdatos.
— Living Standards Measurement Study: https://microdata.worldbank.org/index.php/catalog/lsms.
— Observatorio Social del Ministerio de Desarrollo Social y Familia de Chile: http://observatorio.ministeriodesarrollosocial.gob.cl/inicio.
— Instituto Nacional de Estadística y Geográfia de México: https://www.inegi.org.mx/datos/#Microdatos.